AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models

Zheng Lian [1] **Haoyu Chen** [2] **Lan Chen** [1] **Haiyang Sun** [3] **Licai Sun** [2] **Yong Ren** [1] **Zebang Cheng** [4] **Bin Liu** [1]
Rui Liu [5] **Xiaojiang Peng** [6] **Jiangyan Yi** [7] **Jianhua Tao** [7][8]

Abstract

The emergence of multimodal large language models (MLLMs) advances multimodal emotion recognition (MER) to the next level—from naive discriminative tasks to complex emotion understanding with advanced video understanding abilities and natural language description. However, the current community suffers from a lack of large-scale datasets with intensive, descriptive emotion annotations, as well as a multimodal-centric framework to maximize the potential of MLLMs for emotion understanding. To address this, we establish a new benchmark for MLLM-based emotion understanding with a novel dataset (MER-Caption) and a new model (AffectGPT). Utilizing our model-based crowd-sourcing data collection strategy, we construct the largest descriptive emotion dataset to date (by far), featuring over 2K fine-grained emotion categories across 115K samples. We also introduce the AffectGPT model, designed with pre-fusion operations to enhance multimodal integration. Finally, we present MER-UniBench, a unified benchmark with evaluation metrics tailored for typical MER tasks and the free-form, natural language output style of MLLMs. Extensive experimental results show AffectGPT's robust performance across various MER tasks. We have released both the code and the dataset to advance research and development in emotion understanding: https://github.com/zeroQiaoba/AffectGPT.

---

[1]Institute of Automation, Chinese Academy of Sciences [2]CMVS, University of Oulu [3]Shanghai Jiao Tong University [4]Shenzhen University [5]Inner Mongolia University [6]Shenzhen Technology University [7]Department of Automation, Tsinghua University [8]Beijing National Research Center for Information Science and Technology, Tsinghua University. Correspondence to: Zheng Lian <lianzheng2016@ia.ac.cn>, Jianhua Tao <jhtao@tsinghua.edu.cn>.

1. Introduction

Emotions encapsulate human intentions, and accurately recognizing emotional states is essential for enhancing human-computer interaction (Minsky, 1988). Emotions can be conveyed through various human behaviors in different forms (Chen et al., 2023; Pfister et al., 2011), giving rise to the task of multimodal emotion recognition (MER), which integrates multimodal information to evaluate human emotional states. As a critical area in artificial intelligence, MER has broad applications, ranging from education (Schutz, 2007) and psychological counseling (Liu et al., 2021a) to empathic embodied robots (Spezialetti et al., 2020).

Traditional methods primarily rely on discriminative models that map human emotions to the most likely categories from predefined emotion taxonomies. The most widely used taxonomy is Ekman's theory (Ekman & Keltner, 1970), which classifies all emotions into six basic categories: *sadness*, *happiness*, *fear*, *anger*, *surprise*, and *disgust*. However, such categorical frameworks exhibit some limitations in modeling human affective states. For example, our emotional expressions are diverse and nuanced due to culture-specific idioms (Matsumoto, 2001), context-dependent metaphors (Kövecses, 2003), and highly personalized behavioral patterns (Izard et al., 1993). Current closed-set classification paradigms fail to capture the rich diversity of emotional expressions in real-world scenarios (Plutchik, 1980). Meanwhile, the rigid emotion taxonomies oversimplify the continuous spectrum of emotional experiences by forcing discrete labels (e.g., *anger* or *surprise*) onto nuanced affective states that often coexist (Cowen & Keltner, 2017). Illustrations are provided in Figure 1, where the diverse and coexisting issues are presented in Figs. (a) and (b).

Recent advances in multi-modal large language models (MLLMs) enable emotion understanding to move beyond traditional discriminative approaches, embracing a more generative framework (Liang et al., 2024). This shift allows models to describe complex, coexisting emotional states in natural language. With the vast vocabulary, MLLMs can generate diverse, descriptive emotion categories beyond basic emotions, offering new opportunities for emotional understanding. However, recent research highlights

(a) Diversity

Description: The sentence might be the woman's comment or reaction to someone nearby. Given the audio cues of a steady tone with humor and the woman's smile and glance to the left, we can infer that her words carry a *mocking* or *sarcastic* tone. Thus, this sentence likely expresses the woman's *mockery* of the other person's lack of understanding or recognition of something, conveyed in a *humorous* manner.

(b) Coexistence

Description: In the video, his eyes are wide open and his mouth is also open, indicating a *surprised* facial expression. In the audio, the character speaks with a stutter, which usually expresses feelings of *nervousness*, *anxiety*, or *unease*. Combined with the text content, the character seems to be *unhappy* and *angry* due to the prejudice of the people around him.

Figure 1: **Emotion complexity analysis**. Human emotions are often diverse and coexist simultaneously. Such complex emotional states are difficult to describe using discriminative frameworks. However, MLLMs can generate emotional descriptions, offering new possibilities for complex emotion modeling. Since the original videos contain real people, to address copyright concerns, we first use DemoAI to remove personal information and then proceed with visualization.

that MLLMs still face limitations in emotion understanding (Lian et al., 2024a;d). To address these challenges, this paper aims to advance emotional understanding from two key perspectives: the dataset and the model. Finally, we establish a unified benchmark tailored to the free-form, natural language output style of MLLMs.

Dataset. The current community still suffers from a lack of large-scale datasets with intensive, descriptive emotion annotations to realize the potential of MLLMs. The annotation strategies for constructing descriptive emotion datasets can be classified into three types: *model-based*, *human-based*, and *human-model collaborative* strategies. The *human-based* strategy is the most common way to construct emotion datasets with rich descriptive annotations. However, it's costly to conduct crowd-sourcing to scale up the dataset size with this purely manual annotation manner. Besides, humans tend to focus on main cues, resulting in brief and incomplete descriptions (Liu et al., 2022a). Thus, researchers propose *model-based* automatic annotation approaches. However, due to the lack of human proofreading, this approach may result in insufficient label quality (Cheng et al., 2024). Recently, Lian et al. (2024a) propose a *human-model collaborative* strategy, in which models provide pre-labeled cues and humans conduct multiple rounds of checks, which can be seen as a *human-led, model-assisted* strategy. Although this approach offers more comprehensive descriptions, it is costly and difficult to scale the dataset. To balance label quality and dataset size, we introduce a novel annotation strategy that conducts model-based crowd-sourcing labeling with human priors, named *model-led human-assisted*, to construct a large-scale emotion descriptive dataset with diverse emotional categories.

Models. Existing MLLMs typically consist of three key components: a modality encoder that converts audio and video into low-level hidden features, a connector that transforms these features into a format more suitable for LLMs, and an LLM-based generator that produces responses based on the given instructions. While the results of MLLMs are promising, existing models generally leave everything of multimodal fusion to LLMs, which is insufficient for MER that emphasizes multimodal characteristics. This paper introduces the AffectGPT model, designed with a pre-fusion operation to emphasize multimodal integration.

Benchmark. Although it's desirable to generate emotional descriptions in a free-form, natural language style (see Appendix D), this poses challenges for quantitative comparison. To address this, we propose metrics specifically designed for this output style. Additionally, to ensure fair and comprehensive evaluation, we introduce MER-UniBench, a benchmark that incorporates three typical tasks: fine-grained emotion recognition, basic emotion recognition, and sentiment analysis. We believe this work can enhance the emotion understanding capabilities of MLLMs and open possibilities for complex emotion modeling. The main contributions of this paper are summarized as follows:

- We construct a large-scale emotional description dataset **MER-Caption**, which adopts a model-led, human-assisted annotation strategy to strike a balance between label quality and dataset size.

- We develop **AffectGPT**, which uses additional pre-fusion operations to enhance multimodal integration, thereby improving emotion understanding.

Table 1: **Dataset comparison.** "I", "A", "V", and "T" stand for image, audio, video, and text, respectively. We observe that descriptive datasets contain more diverse labels, providing the potential for modeling complex emotions.

	Dataset	Modality	# Samples	Description	# Emotions	Annotation Manner
Categorical Dataset	RAF-DB (Li et al., 2017)	I	29,672	✗	7	Human
	AffectNet (Mollahosseini et al., 2017)	I	450,000	✗	8	Human
	EmoDB (Burkhardt et al., 2005)	A	535	✗	7	Human
	MSP-Podcast (Lotfian & Busso, 2017)	A	73,042	✗	8	Human
	DFEW (Jiang et al., 2020)	V	11,697	✗	7	Human
	FERV39k (Wang et al., 2022)	V	38,935	✗	7	Human
	MER2023 (Lian et al., 2023)	A,V,T	5,030	✗	6	Human
	MELD (Poria et al., 2019)	A,V,T	13,708	✗	7	Human
Descriptive Dataset	EmoVIT (Xie et al., 2024)	I	51,200	✓	988	Model
	MERR-Coarse (Cheng et al., 2024)	A,V,T	28,618	✓	113	Model
	MAFW (Liu et al., 2022a)	A,V,T	10,045	✓	399	Human
	OV-MERD (Lian et al., 2024a)	A,V,T	332	✓	236	Human-led+Model-assisted
	MERR-Fine (Cheng et al., 2024)	A,V,T	4,487	✓	484	Human-led+Model-assisted
	MER-Caption	A,V,T	115,595	✓	2,932	Model-led+Human-assisted
	MER-Caption+	A,V,T	31,327	✓	1,972	Model-led+Human-assisted

- We build **MER-UniBench**, which encompasses typical MER tasks with tailored metrics. This benchmark can offer comprehensive evaluation results for MLLM-based emotion understanding.

- Extensive experiments demonstrate the effectiveness of AffectGPT, which achieves over a 9% performance improvement compared to existing MLLMs.

2. MER-Caption: Dataset Construction

Table 1 summarizes existing emotion datasets, which can be broadly classified into categorical and descriptive datasets. The former directly provides emotion labels (e.g., *happy*), while the latter offers textual descriptions related to emotions. We first conduct preliminary experiments to extract emotion labels from descriptive datasets (see Appendix E). As shown in Table 1, descriptive datasets contain more diverse labels, offering the potential to capture complex emotions. Thus, this paper focuses on descriptive datasets.

Based on the annotation manner, descriptive datasets can be categorized into *model-based*, *human-based*, and *human-model collaborative* strategies (see Table 1). Although the model-based approach makes it easy to expand the dataset size, it mainly relies on experience to select models and lacks human intervention, resulting in insufficient label quality (Cheng et al., 2024). To enhance label quality, Liu et al. (2022a) relied on human annotators to generate emotion descriptions. However, humans tend to focus on primary clues, easily leading to incomplete descriptions. To this end, Lian et al. (2024a) proposed a *human-led, model-assisted* strategy. Specifically, the model first provides pre-labeled descriptions, and then multiple annotators perform multi-round checks. Although this strategy produces more comprehensive descriptions, it comes with high annotation costs

and faces challenges in scaling the dataset. In this paper, we review these annotation methods and introduce a *model-led, human-assisted* strategy. As shown in Figure 2, we leverage human priors to guide description generation and sample filtering, ultimately achieving automatic annotation for unlabeled data. Using this strategy, we construct the MER-Caption dataset, which includes 115K coarse-labeled samples and 31K fine-labeled samples, making a significant contribution to current descriptive datasets. The raw data in MER-Caption is sourced from the unlabeled portions of MER2024 (Lian et al., 2024b), with explicit permission from the dataset owners. Therefore, this paper does not involve the collection of new data but provides additional annotations for existing datasets. Appendix H provides more comparisons with existing datasets.

2.1. Description Generation

The choice of base models is critical for generating accurate descriptions. Unlike previous work that relied solely on experience (Cheng et al., 2024), we guide model selection using *human priors*. Specifically, we first select a small subset of samples for preliminary experiments. In this phase, we annotate fine-grained labels for each sample, allowing annotators to assign any emotions they deem appropriate, thus providing more diverse and precise labels. Based on the results in preliminary experiments (see Appendix F), we employ SALMONN (Tang et al., 2023) as the audio LLM (ALLM) to generate audio cues, Chat-UniVi (Jin et al., 2024) as the video LLM (VLLM) to extract visual cues, and GPT-3.5 (OpenAI, 2022) ("gpt-3.5-turbo-16k-0613") to merge the audio and video cues with text content. Then, to further reduce annotation costs, we experimented with replacing GPT-3.5 with other open-source LLMs but observed a drop in performance. The primary reason is that multimodal fusion in MER is inherently complex, often en-
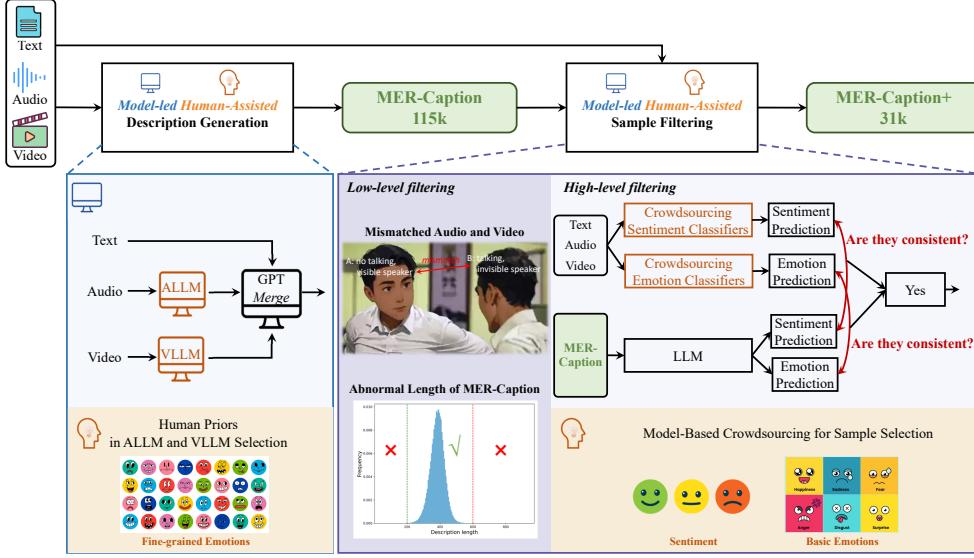
Figure 2: **Dataset construction pipeline.** To create a large-scale dataset with guaranteed label quality, we propose a *model-led, human-assisted* annotation strategy. In this approach, we leverage human priors to guide description generation and sample filtering, ultimately achieving automatic annotation for unlabeled data.

countering issues such as modality conflict, where inconsistencies or contradictions arise between different modalities (see Appendix G). This places high demands on the LLM's reasoning capabilities. Then, we adopt the above strategy for automatic annotation and create the MER-Caption dataset.

2.2. Sample Filtering

Since the descriptions generated by the above process have not been manually verified, MER-Caption inevitably contains some errors. To this end, we implement a two-level filtering process to enhance the label quality.

Low-level Filtering. First, we observe that some samples contain mismatched audio and video. As shown in Figure 2, the visible person is not speaking, while the audio comes from an invisible person. This setup differs from our task, where we aim to analyze a person's emotions based on their audio, video, and text content. Mismatched data complicates this task, shifting the focus to understanding how the interlocutor's actions may influence the target person's emotions. Therefore, we remove such data and plan to address this issue in future work. To automatically determine whether the visible person is speaking, we use TalkNet (Tao et al., 2021). Preliminary experiments indicate that this tool achieves over 90% accuracy in identifying the speaking individual. Then, we remove samples with mismatched audio and video. Second, the length distribution of the generated descriptions roughly follows a Gaussian distribution (see Figure 2). Preliminary experiments reveal that descriptions at both ends of the distribution are more likely to contain

errors. For instance, when ALLM and VLLM (in Section 2.1) fail to generate responses, the resulting descriptions tend to be short. As a result, we further remove descriptions located at both ends of the distribution.

High-level Filtering. In addition to low-level filtering, we propose a *model-based crowdsourcing* technique for high-level filtering. Specifically, we train multiple multimodal emotion and sentiment classifiers using human-annotated categorical datasets. Guided by MERBench (Lian et al., 2024c), we use CLIP ViT-L (Radford et al., 2021) as the visual encoder and HUBERT-L (Hsu et al., 2021) as the acoustic encoder, followed by an attention-based fusion strategy to make final emotion and sentiment predictions. These pre-trained models are then used to predict labels for unlabeled data, generating multiple predictions for each sample. To mitigate potential prediction errors, we apply majority voting to determine the final label, ensuring more reliable results. We refer to this process as *model-based crowdsourcing*. Alternatively, emotions and sentiments can also be predicted based on the descriptions using the strategy outlined in Appendix E. If the labels extracted from the descriptions differ from those obtained through *model-based crowdsourcing*, we consider these descriptions to be of low quality and remove them. Through this process, we can extract knowledge from multiple human-based datasets to guide sample selection. After applying multi-level filtering, we obtain the *MER-Caption+* dataset. Table 1 presents detailed comparisons between our dataset and existing ones, highlighting that our dataset is the largest multimodal emotion description dataset with diverse emotion categories.
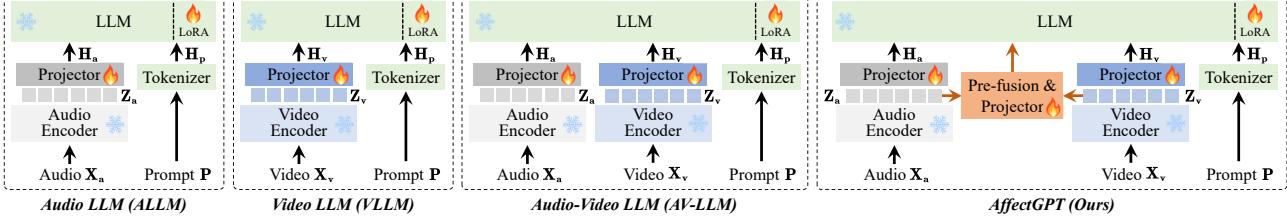
Figure 3: **Model comparison.** ALLM and VLLM primarily use modality-specific encoders and align them with the LLM through projection layers. AV-LLM mainly facilitates cross-modal interaction within the language model. In AffectGPT, we move the cross-modal interaction outside the language model and use a pre-fusion operation to enhance multimodal integration. In these figures, $\mathbf{P}$ can be determined based on the requirement of whether to include $\mathbf{X_t}$.

3. AffectGPT: Model Design

Our primary goal is to map audio-video-text inputs to emotion-related descriptions. In this section, we first review the current mainstream architectures. We then introduce AffectGPT, a model specifically designed to highlight multimodal characteristics in emotion understanding.

3.1. Mainstream Architecture

MLLM aims to understand multimodal input and generate appropriate responses based on the input and user instructions. Unlike pure-text LLMs, the primary challenge for MLLMs lies in enabling the model to perceive multimodal input, i.e., providing the model with "eyes" and "ears". In existing models, the most common approach is to first extract modality-specific embeddings and then align them with the LLM through projection layers. For audio-video joint tasks, Audio-Video LLMs (AV-LLMs) typically facilitate cross-modal interaction within the language model. Figure 3 illustrates the current mainstream architecture.

Formally, for each sample $\mathbf{X}$, we represent its video, audio, and text content as $\mathbf{X_v}$, $\mathbf{X_a}$, and $\mathbf{X_t}$, respectively. Given an instruction $\mathbf{Q}$, the goal is to output the correct response $\mathbf{R}$. For the visual input $\mathbf{X_v}$, we use a video expert to encode it into a latent space $\mathbf{Z_v}$, then apply a projector $G_v(\cdot)$ to generate visual tokens $\mathbf{H_v} = G_v(\mathbf{Z_v})$. Similarly, for the acoustics input $\mathbf{X_a}$, we use an audio expert and a projector to generate the acoustic embeddings $\mathbf{Z_a}$ and tokens $\mathbf{H_a}$. For the instruction $\mathbf{Q}$ and text content $\mathbf{X_t}$, we use a template to merge them into a prompt $\mathbf{P}$, and then map them to the corresponding tokens through the tokenizer and embedding layer in the language model. After obtaining these tokens, we concatenate them and feed them into the LLM decoder. The primary objective is to maximize the likelihood of the target response $\mathbf{R}$, conditioned on multimodal content ($\mathbf{X_v}$, $\mathbf{X_a}$, $\mathbf{X_t}$) and user instruction $\mathbf{Q}$:

$$\max P\left(\mathbf{R}|\mathbf{X_v}, \mathbf{X_a}, \mathbf{X_t}, \mathbf{Q}\right). \tag{1}$$

The above formula is optimized in an autoregressive manner, consistent with the objective function of LLMs. We

represent the response as $\mathbf{R} = \{r_i\}_{i=1}^{L_r}$, where L_r is the number of tokens. Then, Eq. 1 is transformed into:

$$\max \prod_{l=1}^{L_r} P\left(r_l|\mathbf{X_v}, \mathbf{X_a}, \mathbf{X_t}, \mathbf{Q}, \mathbf{R}_{<l}\right). \tag{2}$$

In this equation, r_l is the current token to be predicted, and $\mathbf{R}_{<l} = \{r_i\}_{i=1}^{l-1}$ is the previously generated tokens, which serve as additional conditioning during training.

3.2. Pre-fusion Operation

Mainstream AV-LLMs leave everything of cross-modal interaction to the LLMs (in Figure 3), which is insufficient for handling MER with multimodal characteristics. To address this, we propose a pre-fusion operation that moves the cross-modal interaction outside the LLMs, further enhancing multimodal integration. We refer to this model as AffectGPT. This paper introduces two types of pre-fusion operations: Q-Former-based and attention-based pre-fusion. By default, we apply this operation to $\mathbf{Z_v} \in \mathbb{R}^{t_v \times d}$ and $\mathbf{Z_a} \in \mathbb{R}^{t_a \times d}$. We also experimented with $\mathbf{H_v}$ and $\mathbf{H_a}$, but this choice led to a decrease in performance.

Q-Former. In this module, we preserve the temporal information in the vision features $\mathbf{Z_v}$ and audio features $\mathbf{Z_a}$, and utilize Q-Former (Li et al., 2023b) for multimodal fusion. Specifically, to compress the multimodal content, we first create K learnable query tokens $\mathbf{Z_q} \in \mathbb{R}^{K \times d}$. Then, we interact $\mathbf{Z_q}$ with the concatenated $\mathbf{Z_v}$ and $\mathbf{Z_a}$ through cross-attention, thereby distilling the knowledge from the multimodal content into the query tokens. Formally, this process can be represented as:

$$\mathbf{Z_{av}} = \text{Concat}\left(\mathbf{Z_a}, \mathbf{Z_v}\right), \tag{3}$$

$$\mathbf{Z_f} = \text{Q-Former}\left(\mathbf{Z_q}, \mathbf{Z_{av}} + \text{PE}\left(\mathbf{Z_{av}}\right)\right), \tag{4}$$

where $\mathbf{Z_{av}} \in \mathbb{R}^{(t_a + t_v) \times d}$, with the concatenation operation applied along the temporal dimension. Here, $\mathbf{Z_f} \in \mathbb{R}^{K \times d}$, and $\text{PE}(\cdot)$ represents the positional encoding.

Attention. Unlike Q-Former which preserves temporal information, we propose a simpler architecture that directly compresses temporal information and applies attention mechanisms for multimodal fusion. This simplified module is inspired by MERBench (Lian et al., 2024c), which proves that in MER tasks, features with temporal information do not always lead to better performance than compressed features. Formally, we first apply average pooling to compress unimodal features. Then, we calculate the attention weights to emphasize important modalities:

$$\hat{\mathbf{Z}}_{\mathbf{a}} = \text{Pooling}\left(\mathbf{Z}_a\right), \hat{\mathbf{Z}}_{\mathbf{v}} = \text{Pooling}\left(\mathbf{Z}_v\right), \quad (5)$$

$$\hat{\mathbf{Z}}_{\mathbf{av}} = \text{Concat}\left(\hat{\mathbf{Z}}_{\mathbf{a}}, \hat{\mathbf{Z}}_{\mathbf{v}}\right), \quad (6)$$

$$\mathbf{Z}_{\mathbf{f}} = \hat{\mathbf{Z}}_{\mathbf{av}}^T\left(\mathbf{W} \cdot \text{Flatten}\left(\hat{\mathbf{Z}}_{\mathbf{av}}\right)\right), \quad (7)$$

where $\hat{\mathbf{Z}}_{\mathbf{a}} \in \mathbb{R}^d$, $\hat{\mathbf{Z}}_{\mathbf{v}} \in \mathbb{R}^d$, $\hat{\mathbf{Z}}_{\mathbf{av}} \in \mathbb{R}^{2 \times d}$, and $\mathbf{W} \in \mathbb{R}^{2 \times 2d}$. Finally, we obtain the fused features $\mathbf{Z}_{\mathbf{f}} \in \mathbb{R}^d$.

Regarding computational efficiency, the pre-fusion operation relies on Q-Former or attention mechanisms, which are significantly less computationally intensive than LLMs. Theoretically, the Q-Former enables cross-modal interaction by distilling multimodal content into query tokens, whereas the attention mechanism achieves this by dynamically computing attention weights based on multimodal inputs.

4. MER-UniBench: Evaluation Benchmark

We introduce MER-UniBench, a comprehensive evaluation benchmark designed to cover typical MER tasks. Given the free-form, natural language output style of MLLMs (see Appendix D), we also design specialized evaluation metrics. More details can be found in Appendix J.

Fine-grained Emotion Recognition. This task enables the prediction of fine-grained emotions, extending beyond basic categories. OV-MERD (Lian et al., 2024a) is a typical dataset for this task. To improve the reliability of the evaluation results, we expand its dataset size, referring to it as OV-MERD+. For the evaluation metrics, we draw inspiration from previous work (Lian et al., 2024a) and calculate results in two steps: eliminating the impact of synonyms and using set-level metrics. First, we apply a three-level grouping strategy to mitigate the impact of synonyms:

- **Level 1.** We map different forms of emotion words to their base form. For example, we map *happier* and *happiness* to *happy*. This function is denoted as $F_{l_1}(\cdot)$.

- **Level 2.** We map synonyms to a unified label. For example, we map *happy* and *joyful* to *happy*. This mapping function is represented as $F_{l_2}(\cdot)$.

- **Level 3.** Emotion wheel provides natural grouping information, with core emotions displayed in the inner part and more nuanced labels in the outer part (Plutchik, 1980). Since there is no consensus on the emotion wheel, we use K emotion wheels (see Appendix K). For each sector of the emotion wheel $w_k, k \in [1, K]$, we map all outer labels to the corresponding inner labels. This mapping function is denoted as $F_{l_3}^{w_k}(\cdot)$.

The above grouping functions can be summarized as:

$$G_{w_k}(\cdot) = F_{l_3}^{w_k}(F_{l_2}(F_{l_1}(\cdot))), k \in [1, K]. \quad (8)$$

For each sample, the number of labels is variable. Therefore, we define a set-based evaluation metric. Specifically, suppose the dataset contains N samples. For sample x_i, the true labels are $\mathbf{Y}_i = \{y_i^j\}_{j=1}^{n_i}$, and the predicted labels are $\hat{\mathbf{Y}}_i = \{\hat{y}_i^j\}_{j=1}^{\hat{n}_i}$. The evaluation metric is defined as follows:

$$\text{Precision}_{\text{S}}^k = \frac{1}{N}\sum_{i=1}^N \frac{\left|G_{w_k}(\mathbf{Y}_i) \cap G_{w_k}(\hat{\mathbf{Y}}_i)\right|}{\left|G_{w_k}(\hat{\mathbf{Y}}_i)\right|}, \quad (9)$$

$$\text{Recall}_{\text{S}}^k = \frac{1}{N}\sum_{i=1}^N \frac{\left|G_{w_k}(\mathbf{Y}_i) \cap G_{w_k}(\hat{\mathbf{Y}}_i)\right|}{|G_{w_k}(\mathbf{Y}_i)|}, \quad (10)$$

$$\text{F}_{\text{S}}^k = 2 \times \frac{\text{Precision}_{\text{S}}^k \times \text{Recall}_{\text{S}}^k}{\text{Precision}_{\text{S}}^k + \text{Recall}_{\text{S}}^k}. \quad (11)$$

Finally, we compute the average results across different emotion wheels for ranking. Take F_{S} as an example:

$$\text{F}_{\text{S}} = \frac{1}{K}\sum_{k=1}^K \text{F}_{\text{S}}^k. \quad (12)$$

Here, $\text{Precision}_{\text{S}}$ indicates the number of correctly predicted labels, and Recall_{S} indicates whether the prediction covers all ground truth. F_{S} is a harmonic mean of two metrics. Since F_{S} considers both accuracy and completeness, we use it as the primary metric, with $\text{Precision}_{\text{S}}$ and Recall_{S} serving as secondary metrics. To extract the predicted emotions $\hat{\mathbf{Y}}_i$, we employ the strategy mentioned in Appendix E.

Basic Emotion Recognition. This task is a key branch of MER, whose main goal is to select the most likely label from a fixed set of basic emotions. For this task, we select four widely used benchmark datasets: MER2023 (Lian et al., 2023), MER2024 (Lian et al., 2024b), IEMOCAP (Busso et al., 2008), and MELD (Poria et al., 2019). However, the output of MLLMs, $\hat{\mathbf{Y}}_i = \{\hat{y}_i^j\}_{j=1}^{\hat{n}_i}$, contains a variable number of labels, while the dataset only provides one true label y_i. In this case, traditional metrics (such as *accuracy*) are not suitable for performance evaluation. To address this, we propose a new metric, *hit rate*, which is set to 1

Table 2: **Main results.** This table presents the results for the primary metrics, with Section 4 outlining the primary metrics for each task. The values for other metrics can be found in Appendix L. In this table, "MOSI", "MOSEI", "SIMS", and "SIMS v2" refer to CMU-MOSI, CMU-MOSEI, CH-SIMS, and CH-SIMS v2, respectively. The last column shows the dataset-wise mean score, i.e., the average score across all datasets.

	Modality			Basic				Sentiment				Fine-grained	Mean
	A	V	T	MER2023	MER2024	MELD	IEMOCAP	MOSI	MOSEI	SIMS	SIMS v2	OV-MERD+	
OneLLM	√	×	√	25.52	17.21	28.32	33.44	64.01	54.09	63.39	61.98	22.25	41.14
SECap	√	×	√	40.95	52.46	25.56	36.92	55.76	54.18	59.51	57.41	36.97	46.64
PandaGPT	√	×	√	33.57	39.04	31.91	36.55	66.06	61.33	62.93	58.88	31.33	46.84
Qwen-Audio	√	×	√	41.85	31.61	49.09	35.47	70.09	46.90	70.73	65.26	32.36	49.26
SALMONN	√	×	√	55.53	45.38	45.62	46.84	81.00	67.03	68.69	65.93	45.00	57.89
AffectGPT	√	×	√	**72.94**	**73.41**	**56.63**	**55.68**	**83.46**	**80.74**	**82.99**	**83.75**	**59.98**	**72.18**
Otter	×	√	√	16.41	14.65	22.57	29.08	52.89	50.44	57.56	53.12	16.63	34.82
Video-LLaVA	×	√	√	36.93	30.25	30.73	38.95	56.37	61.64	53.28	57.45	34.00	44.40
PandaGPT	×	√	√	39.13	47.16	38.33	47.21	58.50	64.25	62.07	65.25	35.07	50.77
Video-ChatGPT	×	√	√	44.86	46.80	37.33	56.83	54.42	63.12	64.82	65.80	39.80	52.64
VideoChat2	×	√	√	33.67	54.50	36.64	48.70	66.84	54.32	69.49	70.66	39.21	52.67
LLaMA-VID	×	√	√	50.72	57.60	42.75	46.02	61.78	63.89	69.35	67.48	45.01	56.07
VideoChat	×	√	√	48.73	57.30	41.11	48.38	65.13	63.61	69.52	72.14	44.52	56.71
Chat-UniVi	×	√	√	57.62	65.67	45.61	52.37	54.53	63.18	68.15	66.36	48.00	57.94
mPLUG-Owl	×	√	√	56.86	59.89	49.11	55.54	72.40	72.91	72.13	75.00	48.18	62.45
AffectGPT	×	√	√	**74.58**	**75.29**	**57.63**	**62.19**	**82.39**	**81.57**	**87.20**	**86.29**	**61.65**	**74.31**
PandaGPT	√	√	√	40.21	51.89	37.88	44.04	61.92	67.61	68.38	67.23	37.12	52.92
Emotion-LLaMA	√	√	√	59.38	73.62	46.76	55.47	66.13	67.66	78.32	77.23	52.97	64.17
AffectGPT	√	√	√	**78.54**	**78.80**	**55.65**	**60.54**	**81.30**	**80.90**	**88.49**	**86.18**	**62.52**	**74.77**

when $y_i \in \hat{\mathbf{Y}}_i$ and 0 otherwise. Considering that $\hat{\mathbf{Y}}_i$ is in free-form and y_i belongs to basic emotions $\mathcal{Y}$, we may encounter cases where $\hat{y}_i \notin \mathcal{Y}$. To this end, we use the mapping function $G_{w_k}(\cdot)$ and define the metric as follows:

$$\text{HIT} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I} \left[G_{w_k}(y_i) \in G_{w_k}(\hat{\mathbf{Y}}_i) \right], \qquad (13)$$

where $\mathbb{I}[\cdot]$ is an indicator function. The motivation for this metric stems from the fact that basic emotion recognition tasks typically provide majority-voted labels y_i, which are generally reliable. However, emotion descriptions produce free-form outputs $\hat{\mathbf{Y}}_i$ that may contain multiple labels, including fine-grained ones beyond basic emotions. Therefore, we use the *hit rate* as the metric, ensuring that the basic label y_i should be at least in $\hat{\mathbf{Y}}_i$.

During the design of this metric, we also explored the possibility of evaluating potentially incorrect labels in $\hat{\mathbf{Y}}_i$. However, the labels in $\hat{\mathbf{Y}}_i$ that differ from the basic label y_i are not necessarily incorrect - they may represent some fine-grained emotions not covered by basic categories. Since basic emotion recognition tasks lack fine-grained reference labels, we have not yet established appropriate evaluation metrics for this purpose. This remains an important research direction for our future work.

Sentiment Analysis. This task is more fundamental than the two tasks mentioned above, aiming to predict the sentiment polarity. For this task, we select four benchmark

datasets: CMU-MOSI (Zadeh et al., 2017), CMU-MOSEI (Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), and CH-SIMS v2 (Liu et al., 2022b). For these benchmark datasets, the original labels are floating-point values, ranging from $[-1, 1]$ or $[-3, 3]$. We map scores of < 0 to negative sentiment and scores of > 0 to positive sentiment. To extract sentiment labels from the MLLM's output, we employ the strategy outlined in Appendix E. Following previous work (Zadeh et al., 2017; 2018), we evaluate performance using accuracy (ACC) and weighted average F-score (WAF). Due to the inherent label imbalance, we choose WAF as the primary metric and ACC as the secondary metric.

5. Results and Discussion

In this section, we present the experimental results and provide an in-depth analysis. Detailed implementation information can be found in Appendix B.

Main Results. We compare the performance of Affect-GPT with other MLLMs on MER-UniBench. Since our inputs include audio, video, and text content, we only select MLLMs that support at least audio or video. For models that support both audio and video, we test different modality combinations. Model cards are provided in Appendix C. To ensure a fair comparison, we use their official weights and input corresponding multimodal content, asking them to infer the emotional state. In Table 2, AffectGPT significantly outperforms existing MLLMs. This can be attributed to the

Table 3: **Dataset comparison**. We only change the training dataset, keeping all other aspects consistent. This table reports the mean score across all datasets in MER-UniBench.

	Dataset	Filtering	MER-UniBench
	MiniGPT4	×	31.74
		√	35.53
	VideoChat	×	37.16
		√	37.63
General Instruction	LLaVA	×	46.69
		√	46.27
	WavCaps	×	21.65
		√	37.91
	EmoVIT	–	51.05
	MAFW	–	58.16
Emotion Description	MERR-Coarse	–	49.85
	MERR-Fine	–	64.55
	MER-Caption	–	68.91
	MER-Caption+	–	**74.77**

Table 4: **Necessity of filtering.** Besides the results on MER-UniBench, we also provide task-level results. "E" and "S" are abbreviations for emotion and sentiment, respectively.

Low	High	Fine-grained	Basic	Sentiment	MER-UniBench
×	×	58.42	64.36	76.09	68.91
√	×	58.61	62.78	79.04	69.54
√	E	61.72	67.83	82.74	73.78
√	S	61.00	66.33	**85.04**	74.05
√	E+S	**62.52**	**68.38**	84.22	**74.77**

Table 5: Role of pre-fusion operation.

Pre-fusion	Fine-grained	Basic	Sentiment	MER-UniBench
×	61.21	66.28	82.57	72.95
Q-Former	**62.65**	66.89	**84.30**	74.16
Attention	62.52	**68.38**	84.22	**74.77**

fact that current instruction datasets pay little attention to MER tasks. Additionally, existing models place the entire multimodal fusion within the LLM, which is insufficient for MER tasks that require effective multimodal integration. By leveraging our newly proposed dataset and model, we provide a promising approach to enhancing emotion understanding capability in MLLMs. Meanwhile, for different datasets, increasing the input modality does not always improve performance, as it may also introduce irrelevant information that interferes with emotional understanding.

Effectiveness of MER-Caption. Table 3 compares the performance of MER-Caption with existing datasets. For a fair comparison, we use the same model architecture and experimental settings and only change the training data. For general instruction datasets, we further conduct filtering experiments to remove samples without emotion-related content, emphasizing emotion-related subsets. Specifically, we use the prompt in Appendix E and extract emotion labels from each instruction-answer pair. Samples yielding empty emotion outputs are removed.

In Table 3, the excellent performance of MER-Caption proves the limitations of current datasets in addressing MER. On the one hand, general instruction datasets pay insufficient attention to emotion-related tasks. On the other hand, emotional description datasets often suffer from inadequate dataset scales or insufficient annotation quality. Therefore, our dataset can serve as an important complement to existing datasets. Meanwhile, for the general instruction datasets, the filtering approach is less effective on the LLaVA and VideoChat datasets. We hypothesize that the detailed descriptions in non-emotion subsets may also provide valuable cues for inferring emotional states in some scenarios.

Furthermore, we would like to acknowledge that MER-Caption+ may contain inaccurate descriptions due to the use

of an automatic annotation strategy without manual checks. However, the experimental results in Table 3 show that MER-Caption+ achieves significantly better performance than the manually annotated MAFW dataset. The main reason is that humans tend to focus on major clues, which can easily lead to incomplete descriptions. These results confirm that, despite the lack of manual checks in MER-Caption+, we can still ensure the quality of the labels. In the future, we will investigate other post-filtering techniques to further improve MER-Caption+'s annotation quality.

Ablation Study on MER-Caption. As shown in Table 4, compared to the results without filtering or with only low-level filtering, our two-level filtering leads to a performance improvement, further verifying the effectiveness of our filtering technique. These findings underscore that dataset quality is as critical as quantity, and fewer training samples do not necessarily lead to worse performance. Please see Appendix M for more details.

Ablation Study on Model. Table 5 compares different architectures and examines the impact of pre-fusion operations. Our results show that pre-fusion operations generally improve performance. This highlights the importance of treating cross-modal interactions as separate modules to more effectively capture multimodal characteristics.

Analysis of Input Impact. Table 6 reveals the impact of different inputs. The distinction between "face" and "frame" lies in whether an additional face extractor is used to extract faces from frames. We observe a general trend: multimodal results outperform unimodal results. These findings suggest that humans express emotions through multiple modalities, and integrating them leads to improved performance. Additionally, face inputs slightly outperform frame inputs, and their combination does not result in further improvement. This suggests that current MER datasets mainly

Table 6: **Input impact analysis.** The difference between "face" and "frame" is whether an additional face extractor is used to extract faces from frames.

	Input				MER-UniBench
	audio	face	frame	text	
unimodal	√	×	×	×	60.08
	×	√	×	×	60.47
	×	×	√	×	59.47
	×	×	×	√	67.44
multimodal	√	√	×	√	74.77
	√	×	√	√	73.39
	√	√	√	√	74.60

Table 7: User study.

Dataset	MER-Caption+		
	Wins	Losses	Ties
MERR-Coarse	0.86	0.04	0.10
MERR-Fine	0.59	0.27	0.14

focus on people, with limited emotional information conveyed through the environment. As a result, in this paper, we default to using audio, face, and text as the inputs.

User Study. We conduct a user study to evaluate the quality of our proposed dataset. Since MERR-Fine and MERR-Coarse (Cheng et al., 2024) share some samples with our dataset, we randomly select 20 overlapping samples. We then hire four expert annotators and present them with two descriptions for each sample: one from our dataset and one from the other datasets. The annotators are asked to watch the video and select the more accurate description. As shown in Table 7, our dataset provides more accurate descriptions than both the model-based MERR-Coarse and the human-filtered MERR-Fine, thereby validating the effectiveness of our proposed annotation strategy.

Choice of LLMs. This paper adopts Qwen2.5 as the default LLM. In Figure 4(a), we further explore the impact of different LLMs. Experimental results show that the performance difference brought by LLM is limited. These results verify that the superior performance of AffectGPT over the existing MLLMs does not come from LLM but from our proposed emotion description dataset and model.

Choice of Audio and Video Encoders. In Figures 4(b) and 4(c), the choice of audio and video encoder has a minimal impact on performance. This underscores that AffectGPT's exceptional performance is primarily driven by our proposed high-quality, large-scale dataset and effective framework, rather than the specific acoustic or visual encoders used. For audio encoders (Figure 4(b)), ImageBind exhibits slightly inferior performance compared to other audio encoders. This may be attributed to the fact that other audio encoders are predominantly utilized in audio

(a) LLM (b) Audio Encoder (c) Video Encoder
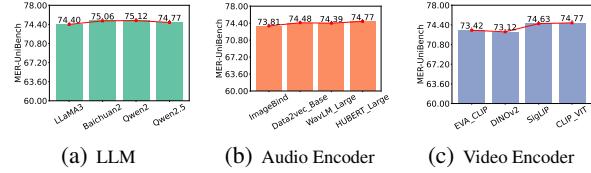
Figure 4: Ablation studies on LLMs, audio encoders, and video encoders.

Table 8: **Impact of rank value in LoRA.** In this table, we count the increase in trainable parameters when using LoRA for the LLM branch. The first row represents the model without the LoRA module.

Rank	# Increased Parameters	MER-UniBench
−	0	73.30
8	20,185,088	74.65
16	40,370,176	74.77
32	80,740,352	74.92

content understanding tasks (e.g., ASR), where audio content plays a critical role in emotion recognition. Similarly, for video encoders (Figure 4(c)), CLIP_VIT marginally outperforms EVA_CLIP and DINOv2, aligning with findings from MERBench (Lian et al., 2024c), a unified benchmark for traditional categorical frameworks. These results suggest that insights derived from traditional categorical frameworks, such as encoder selection, may also be applicable to MLLM-based descriptive frameworks.

Role of LoRA in LLMs. In Table 8, we count the increase in trainable parameters when using LoRA (Hu et al., 2022) for the LLM branch. The first row represents the model without the LoRA module. In Table 8, fine-tuning the LLM with LoRA improves performance compared to models without LoRA. However, increasing the rank for LoRA-based models does not yield significant performance gains and instead increases computational costs.

6. Conclusion

This paper aims to enhance the emotional understanding of MLLMs from three aspects: (1) the dataset MER-Caption, which uses a model-led human-assisted strategy to create a large-scale dataset with guaranteed quality; (2) the model AffectGPT, which enhances multimodal fusion by moving cross-modal interactions outside of the LLM; and (3) the benchmark, which provides comprehensive evaluation metrics tailored to the free-form, natural language output style of MLLMs. Extensive experiments validate the effectiveness of our model and dataset. This work lays the foundation for building MLLMs with emotional understanding, contributing to the advancement of emotion AI.

Acknowledgments

Impact Statement

Social Impact. Emotion plays an important role in human communication, conveying human intentions and deep thoughts. As Minsky (Minsky, 1988) stated: *The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions.* The development of MER technology can enhance the human-computer interaction experience.

Ethics Statement. This paper does not involve the collection of new data. The original data comes from the unlabeled part of MER2024 (Lian et al., 2024b), with permission from the dataset owners. The annotation process does not involve hiring external annotators, and no ethical issues are associated with this process. Additionally, we restrict the use of this dataset under the license of CC BY-NC 4.0, requiring researchers to use our dataset responsibly. Therefore, no ethical concerns are raised in this paper.

References

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. A database of german emotional speech. In *Interspeech*, volume 5, pp. 1517–1520, 2005.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 2008.

Chen, H., Liu, X., Li, X., Shi, H., and Zhao, G. Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8. IEEE, 2019.

Chen, H., Shi, H., Liu, X., Li, X., and Zhao, G. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346–1366, 2023.

Cheng, Z., Cheng, Z.-Q., He, J.-Y., Wang, K., Lin, Y., Lian, Z., Peng, X., and Hauptmann, A. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Cowen, A. S. and Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114 (38):E7900–E7909, 2017.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, 2020.

Du, S., Tao, Y., and Martinez, A. M. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014.

Ekman, P. and Keltner, D. Universal facial expressions of emotion. *California mental health research digest*, 8(4): 151–158, 1970.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. Challenges in representation learning: A report on three machine learning contests. In *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*, pp. 117–124. Springer, 2013.

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., and Marsic, I. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2225–2235, 2018.

Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., and Yue, X. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Hazarika, D., Zimmermann, R., and Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

Hsu, C.-C., Chen, S.-Y., Kuo, C.-C., K. Huang, T.-H., and Ku, L.-W. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp. 1597–1601, 2018.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Huang, Z., Zhao, J., and Jin, Q. Ecr-chain: Advancing generative language models to better emotion-cause reasoners through reasoning chains. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pp. 6288–6296, 2024.

Izard, C. E., Libero, D. Z., Putnam, P., and Haynes, O. M. Stability of emotion experiences and their relations to traits of personality. *Journal of personality and social psychology*, 64(5):847, 1993.

Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., and Liu, J. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2881–2889, 2020.

Jin, P., Takanobu, R., Zhang, W., Cao, X., and Yuan, L. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13700–13710, 2024.

Kövecses, Z. *Metaphor and emotion: Language, culture, and body in human feeling.* Cambridge University Press, 2003.

Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., and Liu, Z. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 1–13, 2023b.

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023c.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., and Qiao, Y.

Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Li, S., Deng, W., and Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2852–2861, 2017.

Li, Y., Wang, C., and Jia, J. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024b.

Lian, Z., Sun, H., Sun, L., Chen, K., Xu, M., Wang, K., Xu, K., He, Y., Li, Y., Zhao, J., et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9610–9614, 2023.

Lian, Z., Sun, H., Sun, L., Chen, L., Chen, H., Gu, H., Wen, Z., Chen, S., Zhang, S., Yao, H., et al. Open-vocabulary multimodal emotion recognition: Dataset, metric, and benchmark. *arXiv preprint arXiv:2410.01495*, 2024a.

Lian, Z., Sun, H., Sun, L., Wen, Z., Zhang, S., Chen, S., Gu, H., Zhao, J., Ma, Z., Chen, X., et al. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 41–48, 2024b.

Lian, Z., Sun, L., Ren, Y., Gu, H., Sun, H., Chen, L., Liu, B., and Tao, J. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *arXiv preprint arXiv:2401.03429*, 2024c.

Lian, Z., Sun, L., Sun, H., Chen, K., Wen, Z., Gu, H., Liu, B., and Tao, J. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367, 2024d.

Liang, Z., Xu, Y., Hong, Y., Shang, P., Wang, Q., Fu, Q., and Liu, K. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pp. 405–409, 2024.

Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., and Yuan, L. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., and Huang, M. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting*

of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 3469–3483, 2021a.

Liu, X., Shi, H., Chen, H., Yu, Z., Li, X., and Zhao, G. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10631–10642, 2021b.

Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., and Shan, S. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 24–32, 2022a.

Liu, Y., Yuan, Z., Mao, H., Liang, Z., Yang, W., Qiu, Y., Cheng, T., Li, X., Xu, H., and Gao, K. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the International Conference on Multimodal Interaction*, pp. 247–258, 2022b.

Lotfian, R. and Busso, C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 2017.

Maaz, M., Rasheed, H., Khan, S., and Khan, F. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Matsumoto, D. Culture and emotion. *The handbook of culture and psychology*, pp. 171–194, 2001.

Minsky, M. *Society of mind*. Simon and Schuster, 1988.

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

OpenAI. Chatgpt, 2022. URL https://openai.com/blog/chatgpt.

Pfister, T., Li, X., Zhao, G., and Pietikäinen, M. Recognising spontaneous facial micro-expressions. In *2011 international conference on computer vision*, pp. 1449–1456. IEEE, 2011.

Plutchik, R. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1, 1980.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pp. 873–883, 2017.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pp. 527–536, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Schutz, P. A. Emotion in education, 2007.

Spezialetti, M., Placidi, G., and Rossi, S. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:532279, 2020.

Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., and Cai, D. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants*, pp. 11–23, 2023.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., MA, Z., and Zhang, C. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Tao, R., Pan, Z., Das, R. K., Qian, X., Shou, M. Z., and Li, H. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3927–3935, 2021.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pp. 6558–6569, 2019.

Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., Ge, W., and Zhang, W. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20922–20931, 2022.

Xie, H., Peng, C.-J., Tseng, Y.-W., Chen, H.-J., Hsu, C.-F., Shuai, H.-H., and Cheng, W.-H. Emovit: Revolutionizing emotion insights with visual instruction tuning. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26596–26605, 2024.

Xu, Y., Chen, H., Yu, J., Huang, Q., Wu, Z., Zhang, S.-X., Li, G., Luo, Y., and Gu, R. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19323–19331, 2024.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., and Yang, K. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

Zhang, H., Li, X., and Bing, L. 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*, pp. 543–553, 2023.

A. Related Works

This paper focuses on constructing datasets and designing models to enhance the emotional understanding capability of MLLMs. In this section, we mainly review related work in these two aspects.

A.1. Emotion Dataset

Emotion datasets are the foundation for building MER systems (Wang et al., 2022; Chen et al., 2023; Liu et al., 2021b; Chen et al., 2019). Most research has focused on building categorical datasets, where basic emotions are first defined, and annotators are asked to select the most likely one (Goodfellow et al., 2013) or multiple (Li et al., 2017) labels from basic emotions. However, emotions are often diverse (Demszky et al., 2020) and can coexist (Du et al., 2014), making it challenging for categorical datasets to fully capture these complex emotions.

To address this, recent studies have shifted from categorical datasets to descriptive datasets, as emotion descriptions provide greater flexibility and enable the description of complex emotions in natural language. To construct such datasets, Liu et al. (2022a) used a human-based annotation strategy to capture the environment, body movements, facial expressions, and other emotion-related cues. However, the high annotation cost limits the scalability of these datasets. With the development of MLLMs, Cheng et al. (2024) used a more cost-effective automatic annotation method, where MLLMs are used to extract emotion-related descriptions from audio, facial expressions, and visual objects. However, they lacked pre-experimentation on MLLM selection, relying on empirical model choices, leading to insufficient label quality. In this paper, we propose a solution to balance label quality and dataset size. By leveraging high-quality human-based datasets to guide description generation and sample filtering, we achieve a quality-assured automatic annotation process and ultimately construct MER-Caption.

A.2. Emotion Models

Emotion models are closely related to the training corpus. For categorical datasets, researchers often build classifiers to map multimodal human information to corresponding emotion labels. Apart from choosing the architecture (such as CNN, RNN, or Transformers), most research focuses on how to align and fuse multimodal information. For example, Hazarika et al. (2020) introduced a decomposition module to split features into modality-specific and modality-invariant representations. Gu et al. (2018) aligned different modalities at the word level and then learned time-dependent cross-modal interactions. Tsai et al. (2019) proposed using cross-attention to align features in the latent space. More recently, Lian et al. (2024c) conducted a fair comparison of various fusion and alignment strategies, showing that temporal-preserving features do not always outperform time-compressed features, suggesting that MER may be more suitable to solve from a global perspective.

For descriptive datasets, due to their natural language style output, the framework needs to shift from traditional discriminative methods to generative methods. With the development of LLMs and MLLMs, researchers have started to build models based on them. For example, Huang et al. (2024) used Vicuna as the language model, jointly training emotion labels and descriptions. Xie et al. (2024) used the instruction-aware Q-Former module to learn the mapping between input images and emotional descriptions. Cheng et al. (2024) integrated different encoders to understand multimodal inputs and used LLaMA-2 as an LLM decoder. However, current models either only focus on unimodal information (Huang et al., 2024; Xie et al., 2024) or leave all cross-modal interactions to the LLM (Cheng et al., 2024), which is insufficient for solving MER tasks with multimodal characteristics. To this end, we introduce the AffectGPT model in this paper.

B. Implementation Details

Our choice of unimodal encoders is guided by previous research (Lian et al., 2024c), using CLIP ViT-L (Radford et al., 2021) as the visual encoder and HUBERT-L (Hsu et al., 2021) as the acoustic encoder. Given the remarkable performance of Qwen-2.5 (Yang et al., 2024), we choose it as the LLM. To ensure training efficiency, we only fine-tune an extra LoRA module (in the LLM), projector, and pre-fusion branch, while freezing the weights of the LLM and unimodal encoders (see Figure 3). We default to setting the rank in the LoRA module to 16. This approach reduces GPU memory usage and speeds up training. Additionally, through preliminary experiments, we found that pre-training on other instruction datasets followed by a second-stage training on MER-Caption did not lead to performance improvements. The primary reason is the large scale of our dataset and the limited focus on MER in current instruction datasets. Therefore, we do not perform multi-stage training in our experiments. All models are implemented in PyTorch and conducted training and inference on 80GB NVIDIA Tesla A100 GPU. During training, we set the maximum number of epochs to 60, each epoch contains 5000 iterations, and the

batch size of each iteration is 3. To optimize all trainable parameters, we use the AdamW optimizer and set the learning rate to 1e-5. For more implementation details, please refer to the code provided in https://github.com/zeroQiaoba/AffectGPT.

C. Details about MLLMs

Table 9 provides model cards for different MLLMs, including reference papers, supported modalities, and links to pre-trained weights.

Table 9: Model cards for MLLMs.

	Supported Modality	Link
Otter (Li et al., 2023a)	Video, Text	https://github.com/Luodian/Otter
VideoChat (Li et al., 2023c)	Video, Text	https://github.com/OpenGVLab/Ask-Anything/tree/main/video_chat
VideoChat2 (Li et al., 2024a)	Video, Text	https://github.com/OpenGVLab/Ask-Anything/tree/main/video_chat2
Video-LLaVA (Lin et al., 2024)	Video, Text	https://github.com/PKU-YuanGroup/Video-LLaVA
Video-LLaMA (Zhang et al., 2023)	Video, Text	https://github.com/DAMO-NLP-SG/Video-LLaMA
Video-ChatGPT (Maaz et al., 2024)	Video, Text	https://github.com/mbzuai-oryx/Video-ChatGPT
LLaMA-VID (Li et al., 2024b)	Video, Text	https://github.com/dvlab-research/LLaMA-VID
mPLUG-Owl (Ye et al., 2023)	Video, Text	https://github.com/X-PLUG/mPLUG-Owl
Chat-UniVi (Jin et al., 2024)	Video, Text	https://github.com/PKU-YuanGroup/Chat-UniVi
SALMONN (Tang et al., 2023)	Audio, Text	https://github.com/bytedance/SALMONN
Qwen-Audio (Chu et al., 2023)	Audio, Text	https://github.com/QwenLM/Qwen-Audio
SECap (Xu et al., 2024)	Audio, Text	https://github.com/thuhcsi/SECap
OneLLM (Han et al., 2024)	Audio, Video, Text	https://github.com/csuhan/OneLLM
PandaGPT (Su et al., 2023)	Audio, Video, Text	https://github.com/yxuansu/PandaGPT
Emotion-LLaMA (Cheng et al., 2024)	Audio, Video, Text	https://github.com/ZebangCheng/Emotion-LLaMA

D. Visualization of MLLM Outputs

Figure 5 provides an example to visualize the outputs of different MLLMs. These outputs contain varying numbers of emotions, with emotion labels that are open-ended and not restricted to any predefined taxonomy. Therefore, traditional classification metrics, such as accuracy and F1 score, are not suitable for performance evaluation. In this paper, we propose evaluation metrics specifically tailored for the free-form, natural language output style of MLLMs.

Subtitle: Why are you all looking at me like that? So, as long as it's a woman, does she have to have a relationship with me?

Chat-UniVi-main: In the video, a man is seen speaking to the camera while making facial expressions and gestures. The man's body language and expressions suggest that he might be trying to convey a message or share his thoughts with the audience. The environment appears to be a simple setting, with no significant background details. The man's emotional state is not entirely clear from the video alone, as it depends on the context of his words and actions. However, based on his facial expressions and body movements, we can infer that he might be feeling *passionate*, *expressive*, or even slightly *nervous* while addressing the camera. The video's content and the man's demeanor invite the viewer to interpret his emotions and message further.

LLaMA-VID-main: The man in the video appears to be in a state of *confusion* or *uncertainty*, as he looks around and appears to be unsure of what to do. His facial expressions and body movements suggest that he is trying to process information or make a decision. The environment in which he is speaking, which is a laboratory, may also contribute to his emotional state, as it may suggest that he is dealing with a difficult or complex problem. Overall, it is difficult to predict the exact emotional state of the individual based on the information provided, but it is clear that he is experiencing some level of confusion or uncertainty.

SALMONN: It is difficult to predict the emotional state of the individual in the audio without more context. However, based on the given information, it is possible that the individual is feeling *frustrated* or *misunderstood*. They may be feeling like they are being judged or scrutinized by others, and this could be causing them to feel *upset* or *angry*. Alternatively, they may be feeling *defensive* or *protective* of themselves, and this could be causing them to become *aggressive* or *confrontational*. Ultimately, without more information, it is impossible to accurately predict the emotional state of the individual in the audio.

Video-LLaVA: The man in the video is sitting in front of a camera and talking. He appears to be in a *relaxed* state, with his hands resting on his lap. His facial expressions are *calm* and *composed*, and his body language suggests that he is *comfortable* in front of the camera. The environment around him is well-lit, and there are no visible signs of distractions or interruptions. The subtitle content of the video does not provide any specific information about the individual's emotions or the context of the video. However, based on the visual cues, it can be inferred that the man is likely in a *positive* emotional state.

Video-ChatGPT: In the video, a man is seen talking to the camera while a woman looks on. The man is wearing a white shirt and a tie, and he appears to be engaged in a conversation with the woman. The woman's facial expression is not clearly visible, but her body language suggests that she is listening *attentively* to the man. The environment appears to be an office setting, with a desk and a computer visible in the background. The acoustic information, such as the sound of the man speaking, is not provided. The subtitles in the video indicate that the man is asking the woman if she has a relationship with him, and the woman responds by saying that she is not interested in him. Based on these visual and auditory cues, it is difficult to determine the exact emotional state of the individual. However, it is clear that the man is engaged in a conversation with the woman, and the woman is listening attentively to him.

VideoChat2: The video doesn't seem to have any information related to the given subtitle content. It would be best to provide more details or context to help me understand the context and answer the question accurately.

SECap: feeling *astonished*

Figure 5: Visualization of MLLM outputs.

E. Prompt for Label Extraction

To extract emotion labels from MLLM outputs, we use Qwen2.5 and apply the following prompt:

Please assume the role of an expert in the field of emotions. We provide clues that may be related to the emotions of the characters. Based on the provided clues, please identify the emotional states of the main character. Please separate different emotional categories with commas and output only the clearly identifiable emotional categories in a list format. If none are identified, please output an empty list.

For sentiment analysis, we use the multi-step prediction process. Specifically, we first extract emotion labels using the prompt above, and then apply the following prompt for sentiment analysis:

Please act as an expert in the field of emotions. We provide a few words to describe the emotions of a character. Please choose the most likely sentiment from the given candidates: [positive, negative, neutral].

F. Choice of Description Generation Strategy

This section aims to determine the optimal strategy for generating descriptions. In Table 10, we present the results of preliminary experiments. First, we evaluate the performance of different ALLMs and VLLMs. Then, we investigate whether combining these models leads to improved performance. To do this, we use GPT-3.5 to integrate audio and video cues, extracted by the ALLM and VLLM, with text content. As shown in Table 10, we observe that these combinations generally outperform the use of either ALLM or VLLM alone. Based on these findings, we select SALMONN as the ALLM for generating audio cues, Chat-UniVi as the VLLM for generating visual cues, and GPT-3.5 to combine the audio, video, and text cues, resulting in the final descriptions.

We would like to clarify that, in this paper, we do not use *combined results* for model selection. Instead, we rely on the performance of *individual models*. Specifically, for VLLM, Chat-UniVi outperforms mPLUG-Owl and Video-ChatGPT; for ALLM, SALMONN outperforms SECap. As a result, we employ the combination of Chat-UniVi and SALMONN for description generation. The combination experiments are primarily designed to demonstrate that integrating multimodal cues can enhance performance. In future work, we will conduct additional experiments where *combined results* are used for model selection. For example, leveraging the combination of SALMONN and Chat-UniVi for description generation.

Table 10: **Preliminary experiments.** We choose F_S as the primary metric, as this metric considers both accuracy and completeness.

Model	$F_S(\uparrow)$	$\text{Precision}_S(\uparrow)$	$\text{Recall}_S(\uparrow)$
SECap (Xu et al., 2024)	$45.72_{\pm 0.09}$	$54.52_{\pm 0.15}$	$39.37_{\pm 0.05}$
SALMONN (Tang et al., 2023)	$47.96_{\pm 0.04}$	$50.20_{\pm 0.04}$	$45.92_{\pm 0.04}$
Video-ChatGPT (Maaz et al., 2024)	$50.52_{\pm 0.06}$	$54.03_{\pm 0.04}$	$47.44_{\pm 0.07}$
mPLUG-Owl (Ye et al., 2023)	$52.73_{\pm 0.13}$	$54.54_{\pm 0.13}$	$51.04_{\pm 0.13}$
Chat-UniVi (Jin et al., 2024)	$53.08_{\pm 0.01}$	$53.68_{\pm 0.00}$	$52.50_{\pm 0.02}$
SECap + mPLUG-Owl	$56.69_{\pm 0.03}$	$50.05_{\pm 0.23}$	$65.38_{\pm 0.33}$
SECap + Video-ChatGPT	$56.90_{\pm 0.08}$	$52.03_{\pm 0.04}$	$62.79_{\pm 0.14}$
SECap + Chat-UniVi	$57.34_{\pm 0.16}$	$48.85_{\pm 0.29}$	$69.41_{\pm 0.13}$
SALMONN + Video-ChatGPT	$58.19_{\pm 0.23}$	$53.16_{\pm 0.17}$	$64.26_{\pm 0.31}$
SALMONN + Chat-UniVi	$58.43_{\pm 0.06}$	$51.62_{\pm 0.00}$	$67.31_{\pm 0.15}$
SALMONN + mPLUG-Owl	$\mathbf{58.70}_{\pm 0.04}$	$51.77_{\pm 0.01}$	$67.76_{\pm 0.11}$

G. Prompt for Clue Merge

To merge multimodal clues, we use GPT-3.5 and apply the following prompt:

Please act as an expert in the field of emotions. We provide acoustic and visual clues that may be related to the character's emotional state, along with the original subtitle of the video. Please analyze which parts can infer the emotional state and explain the reasons. During the analysis, please integrate the textual, audio, and visual clues.

Even when modality conflicts exist (i.e., the emotions conveyed by audio, video, and text are not the same, as shown in Figure 6), GPT-3.5 can provide reasonable responses, primarily due to its powerful reasoning ability.

Text: Why are you all looking at me like that? So, as long as it's a woman, does she have to have a relationship with me?

Visual Clue: In the video, the screen shows a male character in an indoor setting. At the beginning of the video, his eyes are wide open and his mouth is also open, indicating a *surprised* facial expression. In the following scenes, he looks around, seemingly explaining or narrating something to the people around him. Overall, his emotions are *not positive or optimistic*.

Acoustic Clue: In the audio, the character speaks with a stutter, which usually expresses feelings of *nervousness*, *anxiety*, or *unease*.

> **Merged Clue:** The subtitle expresses the male character's dissatisfaction and anger towards the people around him. Based on the surprised and negative facial expression of the male character in the video clues, as well as the stuttering speech in the audio clues, we can infer that the male character is expressing a feeling of *dissatisfaction* and *anger* in this sentence. He may feel *troubled* by the prejudice of the people around him and is *unhappy* with this unfair treatment.

Figure 6: Example of modality conflict.

H. Dataset Comparison

Figure 7 compares the distribution of description lengths and the number of emotions per sample. We observe that our dataset provides detailed descriptions and rich emotion labels for each sample.
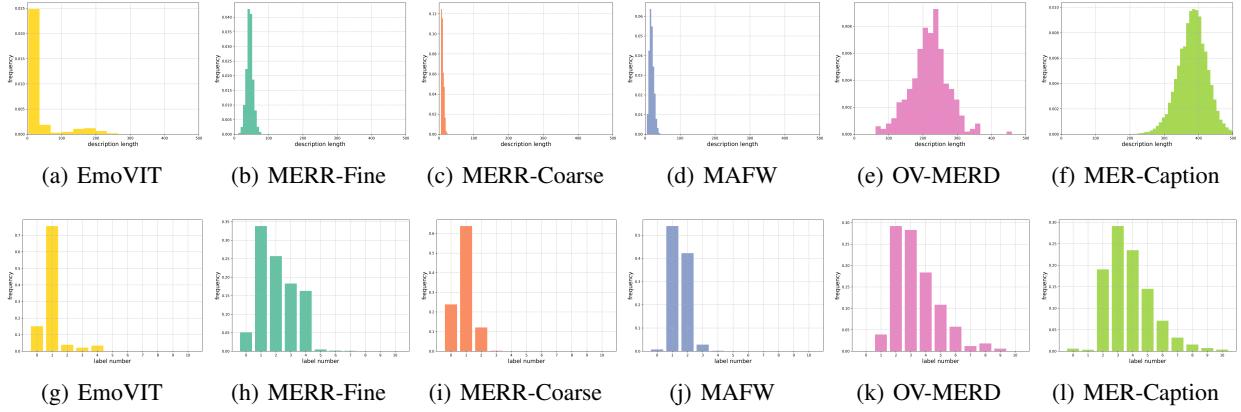
Figure 7: **Dataset comparison**. The first row compares the lengths of the descriptions, while the second row compares the number of labels per sample.

I. Video Duration Distribution

Figure 8 presents the video duration distribution of the MER-Caption dataset. We observe that the majority of samples have durations ranging from 2 to 5 seconds.
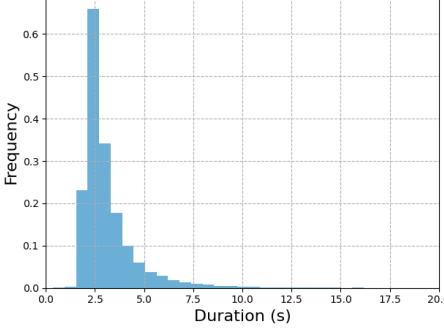
Figure 8: Video duration distribution.

J. MER-UniBench Details

MER-UniBench is a comprehensive evaluation benchmark covering three typical tasks in MER, including fine-grained emotion recognition, basic emotion recognition, and sentiment analysis. Different tasks involve different datasets, and we provide their statistical information in Table 11. In this paper, we intentionally focus on single-person videos, as this allows us to eliminate interference from other speakers and reduce task difficulty. Multi-person MER belongs to another research topic and will be addressed in our future work.

Table 11: Dataset statistics in MER-UniBench. All datasets in our study focus on single-person videos.

	Dataset	Chosen Set	# Samples	Label Description	Data Source
Fine-grained Emotion	OV-MERD+	All	532	unfixed categories and diverse number of labels per sample	movies, TV series
Basic Emotion	MER2023	MER-MULTI	411	most likely label among six candidates	movies, TV series
	MER2024	MER-SEMI	1,169	most likely label among six candidates	movies, TV series
	IEMOCAP	Session5	1,241	most likely label among four candidates	actor's performance
	MELD	Test	2,610	most likely label among seven candidates	"Friends" TV series
Sentiment Analysis	CMU-MOSI	Test	686	sentiment intensity, ranging from [-3, 3]	opinion videos in YouTube
	CMU-MOSEI	Test	4,659	sentiment intensity, ranging from [-3, 3]	opinion videos in YouTube
	CH-SIMS	Test	457	sentiment intensity, ranging from [-1, 1]	movies, TV series, and shows
	CH-SIMS v2	Test	1,034	sentiment intensity, ranging from [-1, 1]	movies, TV series, and shows

OV-MERD+ is our newly collected dataset, an extended version of the previous OV-MERD (Lian et al., 2024a). Unlike traditional datasets, which select a single label from basic emotions, OV-MERD is a fine-grained emotion dataset that allows each sample to have a variable number of emotions, using any emotion not restricted to predefined taxonomies. OV-MERD initially contains 332 samples, and we further expand its dataset size, obtaining OV-MERD+.

MER2023 (Lian et al., 2023) and **MER2024** (Lian et al., 2024b) are widely used in Chinese MER research, with MER2024 being an extended version of MER2023. The original data in both datasets comes from movies and TV shows. They use various techniques to segment video clips, ensuring that each clip has only one person, with their speech content being relatively complete. To ensure annotation quality, they hire multiple annotators, each selecting the most likely label from six candidate emotions: *worry*, *happy*, *neutral*, *angry*, *surprised*, and *sad*. The final label is determined through majority voting.

IEMOCAP (Busso et al., 2008) is one of the most widely used emotion datasets. It contains five sessions, each with a male and a female actor in a laboratory environment. The dataset includes the following emotion labels: *anger*, *happiness*, *sadness*, *neutral*, *excitement*, *frustration*, *fear*, *surprise*, and *others*. Following previous research (Poria et al., 2017), we choose the last session for testing, and use the first four emotions, and merge *surprise* and *happiness* into *happiness*.

MELD (Poria et al., 2019) is an extension of the text-centered EmotionLines dataset (Hsu et al., 2018), adding audio and video content. The raw data is derived from the Friends TV series. The dataset has seven emotion labels, and each sample is assigned to one of the most likely labels: *anger*, *joy*, *sadness*, *neutral*, *disgust*, *fear*, and *surprise*.

CMU-MOSI (Zadeh et al., 2017) and **CMU-MOSEI** (Zadeh et al., 2018) consist of opinion videos collected from online platforms. CMU-MOSEI is an extended version of CMU-MOSI, with more samples and a wider range of topics. In these

datasets, each sample is labeled with a sentiment intensity score ranging from -3 to +3, where -3 represents extremely negative emotion and +3 represents extremely positive emotion.

CH-SIMS (Yu et al., 2020) and **CH-SIMS v2** (Liu et al., 2022b) differ from the English-centered CMU-MOSI and CMU-MOSEI by focusing on emotions within the Chinese culture. The original data comes from movies, TV series, and shows. Similar to CMU-MOSI and CMU-MOSEI, these datasets also annotate sentiment intensity, but with a different range $[-1, 1]$, where -1 represents extremely negative emotion and +1 represents extremely positive emotion.

K. Emotion Wheel

Since there is no universal definition of the emotion wheel, we follow previous work (Lian et al., 2024a) and use five emotion wheels in this paper.

(a) w1   (b) w2   (c) w3

(d) w4   (e) w5
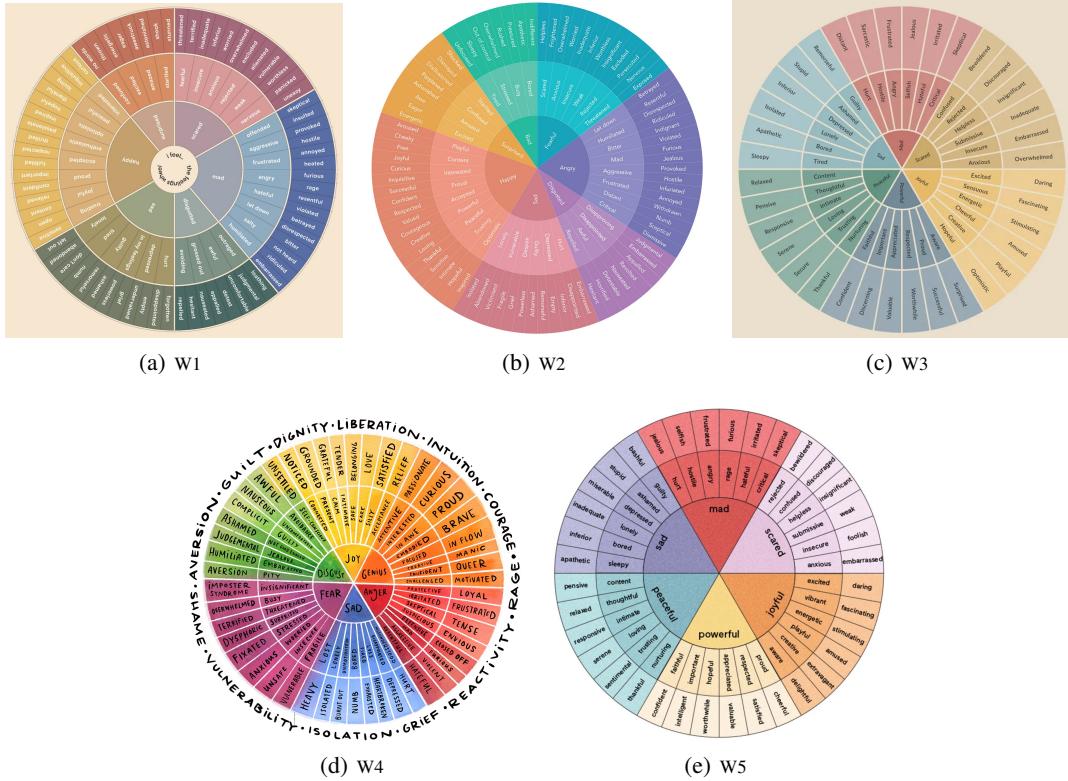
Figure 9: **Emotion wheel**. We use five emotion wheels, all of which are derived from previous research (Lian et al., 2024a).

L. Main Results

Table 12 reports the complete results, with several metrics for each dataset, and the primary metrics are highlighted in gray. In the last column, we report the average value of the primary metrics. These results verify the effectiveness of our AffectGPT in multimodal emotion understanding.

M. Ablation Study on MER-Caption

Table 13 compares the performance across different datasets. To ensure a fair comparison, we keep the model architecture and experimental setup unchanged, only altering the training dataset. Experimental results in Table 13 demonstrate the effectiveness of our MER-Caption dataset for emotion understanding. It addresses the issue of existing datasets either giving insufficient attention to emotion tasks or lacking high-quality emotion descriptions.

Table 12: **Main results**. In this table, "A", "V", and "T" represent audio, video, and text, indicating the input information used by each MLLM during inference. The gray-highlighted columns represent the primary metric for each dataset, while the "Mean" column reports the average score of the primary metrics across all datasets.

	A	V	T	MER2023 HIT(↑)	MER2024 HIT(↑)	MELD HIT(↑)	IEMOCAP HIT(↑)	CMU-MOSI WAF(↑)	CMU-MOSI ACC(↑)	CMU-MOSEI WAF(↑)	CMU-MOSEI ACC(↑)	CH-SIMS WAF(↑)	CH-SIMS ACC(↑)	CH-SIMS v2 WAF(↑)	CH-SIMS v2 ACC(↑)	OV-MERD+ F_s(↑)	OV-MERD+ Precision$_s$(↑)	OV-MERD+ Recall$_s$(↑)	Mean
Otter	×	√	√	16.41	14.65	22.57	29.08	52.89	54.27	50.44	50.77	57.56	60.57	53.12	56.20	16.63	17.67	15.74	34.82
OneLLM	√	×	√	25.52	17.21	28.32	33.44	64.01	64.48	54.09	54.18	63.39	63.92	61.98	62.46	22.25	24.49	20.41	41.14
Video-LLaVA	×	√	√	36.93	30.25	30.73	38.95	56.37	57.62	61.64	64.20	53.28	54.64	57.45	59.28	34.00	36.48	31.86	44.40
SECap	√	×	√	40.95	52.46	25.56	36.92	55.76	56.71	54.18	53.85	59.51	62.89	57.41	60.92	36.97	43.51	32.17	46.64
PandaGPT	√	×	√	33.57	39.04	31.91	36.55	66.06	65.85	61.33	60.73	62.93	62.37	58.88	58.84	31.33	33.08	29.77	46.84
Qwen-Audio	√	×	√	41.85	31.61	49.09	35.47	70.09	71.49	46.90	51.16	70.73	73.45	65.26	68.17	32.36	38.52	27.91	49.26
PandaGPT	×	√	√	39.13	47.16	38.33	47.21	58.50	60.21	64.25	65.55	62.07	61.60	65.25	65.31	35.07	37.86	32.67	50.77
Video-ChatGPT	×	√	√	44.86	46.80	37.33	56.83	54.42	57.77	63.12	65.66	64.82	64.43	65.80	66.85	39.80	43.12	36.97	52.64
VideoChat2	×	√	√	33.67	54.50	36.64	48.70	66.84	67.23	54.32	54.82	69.49	69.59	70.66	71.13	39.21	42.85	36.16	52.67
PandaGPT	√	√	√	40.21	51.89	37.88	44.04	61.92	62.80	67.61	68.82	68.38	67.78	67.23	67.40	37.12	39.64	34.91	52.92
LLaMA-VID	×	√	√	50.72	57.60	42.75	46.02	61.78	62.65	63.89	66.21	69.35	68.81	67.48	67.73	45.01	46.83	43.32	56.07
VideoChat	×	√	√	48.73	57.30	41.11	48.38	65.13	65.09	63.61	63.02	69.52	69.33	72.14	72.12	44.52	44.55	44.49	56.71
SALMONN	√	×	√	55.53	45.38	45.62	46.84	81.00	81.25	67.03	66.90	68.69	69.85	65.93	67.07	45.00	43.57	46.61	57.89
Chat-UniVi	×	√	√	57.62	65.67	45.61	52.37	54.53	57.62	63.18	67.47	68.15	67.78	66.36	67.18	48.00	48.20	47.81	57.94
mPLUG-Owl	×	√	√	56.86	59.89	49.11	55.54	72.40	72.26	72.91	73.17	72.13	71.65	75.00	74.97	48.18	47.91	48.47	62.45
Emotion-LLaMA	√	√	√	59.38	73.62	46.76	55.47	66.13	66.31	67.66	67.25	78.32	78.61	77.23	77.39	52.97	54.85	51.22	64.17
AffectGPT	√	√	√	**78.54**	**78.80**	**55.65**	**60.54**	**81.30**	**81.25**	**80.90**	**80.68**	**88.49**	**88.40**	**86.18**	**86.17**	**62.52**	**62.21**	**63.00**	**74.77**

Table 13: Dataset comparison.

	Dataset	Filtering	MER2023	MER2024	MELD	IEMOCAP	CMU-MOSI	CMU-MOSEI	CH-SIMS	CH-SIMS v2	OV-MERD+	Mean
General Instruction	MiniGPT4	×	11.56	12.91	18.89	16.06	53.57	45.98	57.66	55.16	13.86	31.74
		√	17.57	16.65	22.60	30.18	52.58	56.50	52.36	51.19	20.16	35.53
	VideoChat	×	24.87	22.42	21.56	32.91	50.13	56.17	50.07	51.71	24.56	37.16
		√	27.70	24.73	27.66	39.46	45.45	56.86	43.68	47.05	26.09	37.63
	LLaVA	×	42.21	41.54	32.97	49.96	54.48	56.42	52.04	54.80	35.75	46.69
		√	41.56	42.30	32.61	46.21	52.82	57.72	52.78	53.44	36.96	46.27
	WavCaps	×	5.75	7.71	4.35	4.99	45.59	22.76	53.04	45.68	4.95	21.65
		√	23.72	26.97	23.39	27.30	54.67	49.54	58.12	55.93	21.57	37.91
Emotion Description	EmoVIT	–	39.31	50.24	32.36	48.24	53.40	61.53	69.72	66.53	38.09	51.05
	MAFW	–	52.67	55.99	40.85	57.60	66.11	62.27	75.20	70.02	42.75	58.16
	MERR-Coarse	–	35.34	36.60	29.37	36.94	65.10	63.27	75.12	73.76	33.18	49.85
	MERR-Fine	–	69.00	72.84	47.38	54.49	66.21	60.03	79.90	78.54	52.56	64.55
	MER-Caption	–	72.12	74.21	54.69	56.41	75.10	70.97	80.21	78.06	58.42	68.91
	MER-Caption+	–	**78.54**	**78.80**	**55.65**	**60.54**	**81.30**	**80.90**	**88.49**	**86.18**	**62.52**	**74.77**

N. Impact of Sampling Frames in Video Branch

This paper defaults to sampling 8 frames per video. But if we change the number of sampled frames, will it significantly impact performance? To answer this, we conducted additional experiments in this section. Specifically, we compare two types of inputs: (1) face-only and (2) face-text combinations, and evaluate model performance across different sampling frame counts, ranging from 2 to 64. In Figure 10, we observe that using too few frames (e.g., fewer than 2) results in a noticeable decline in performance, indicating that insufficient frames lead to information loss. However, further increasing the number of sampling frames (e.g., more than 16) does not yield significant performance improvements. This can be attributed to the fact that MER tasks typically use short-duration videos with relatively stable facial expressions.

(a) Face-only Input   (b) Face-text Input
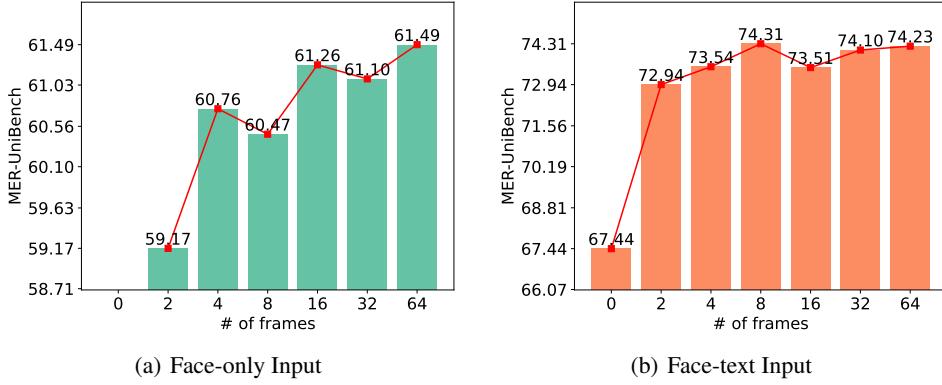
Figure 10: Impact of sampling frames.