Can Large Language Models Help Multimodal Language Analysis? MMLA: A Comprehensive Benchmark

Hanlei Zhang[1*] **Zhuohang Li**[1*] **Hua Xu**[1†] **Yeshuang Zhu**[2†]
Peiwu Wang[1] **Haige Zhu**[3] **Jie Zhou**[2] **Jinchao Zhang**[2]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China [3]Kennesaw State University

Abstract

Multimodal language analysis is a rapidly evolving field that leverages multiple modalities to enhance the understanding of high-level semantics underlying human conversational utterances. Despite its significance, little research has investigated the capability of multimodal large language models (MLLMs) to comprehend cognitive-level semantics. In this paper, we introduce MMLA, a comprehensive benchmark specifically designed to address this gap. MMLA comprises over 61K multimodal utterances drawn from both staged and real-world scenarios, covering six core dimensions of multimodal semantics: intent, emotion, dialogue act, sentiment, speaking style, and communication behavior. We evaluate eight mainstream branches of LLMs and MLLMs using three methods: zero-shot inference, supervised fine-tuning, and instruction tuning. Extensive experiments reveal that even fine-tuned models achieve only about 60%$\sim$70% accuracy, underscoring the limitations of current MLLMs in understanding complex human language. We believe that MMLA will serve as a solid foundation for exploring the potential of large language models in multimodal language analysis and provide valuable resources to advance this field. The datasets and code are open-sourced at `https://github.com/thuiar/MMLA`.

1 Introduction

Multimodal language analysis has emerged as a prominent research area [11], utilizing various modalities to decode cognitive-level semantics in human utterances (e.g., emotion and intent). This analysis is crucial for understanding psychological and behavioral motivations, and it has broad applications in virtual assistants [58], recommender systems [9], and social behavior analysis [40].

This field has attracted significant attention, with early works focusing on annotating sentiment intensity from social media videos [68, 71] and conversations from various TV shows or movies [66, 37]. Additionally, researchers have provided emotion categories for TV shows [45] based on Ekman's six universal emotions [15]. Building on these resources, numerous methods have been developed to learn complementary information and alleviate the challenges posed by the heterogeneous nature of different modalities [53, 23, 26, 78]. In addition to sentiment and emotion, researchers have investigated other linguistic properties such as sarcasm [5, 80] and humor [21, 8], with multimodal fusion methods specifically designed for binary classification tasks [22, 46]. More recently, studies have focused on analyzing coarse-grained and fine-grained intents using new datasets and taxonomies [49, 73, 74], although this area is still in its early stages [52, 85].

*Equal contribution. Work done during the internship of Hanlei Zhang at WeChat AI.
†Corresponding authors. Email: xuhua@tsinghua.edu.cn, yshzhu@tencent.com.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

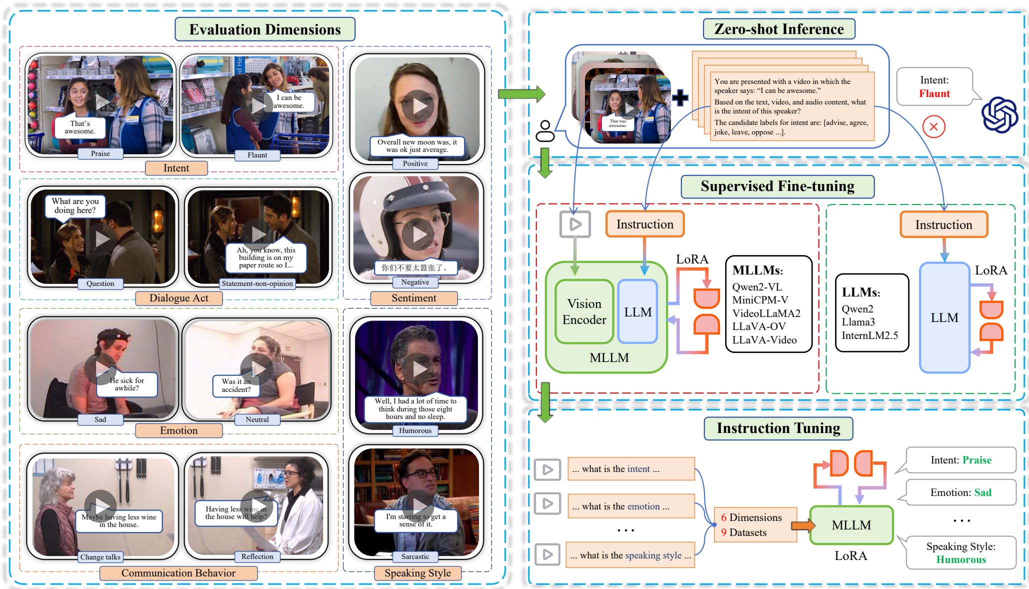

Figure 1: Overview of the MMLA benchmark. The left side shows examples from six evaluation dimensions and nine datasets. The right side displays three methods for evaluating both LLMs and MLLMs: (1) zero-shot inference (top right), which generates predictions from task-specific prompts; (2) supervised fine-tuning (middle right), which trains on each supervised task; and (3) instruction tuning (bottom right), which trains on multiple tasks simultaneously. Both (2) and (3) utilize LoRA to efficiently adapt foundation models.

Despite these advances, existing methods predominantly rely on fusion techniques built on lightweight neural networks [48, 67, 6], which show limited performance on more complicated reasoning tasks [65]. The advent of MLLMs [33, 35, 84, 56] reveals the huge potential for emergent cross-modal reasoning capabilities through scalable model parameters [65]. However, existing MLLM benchmarks mainly focus on low-level perceptual semantics, such as scene and procedure understanding [32], instance location [38], and elementary cognitive-level tasks like video content analysis [17] and commonsense reasoning [65]. These benchmarks fail to address high-level semantics in conversations. Other benchmarks in this field include only a few semantic dimensions, such as emotion and intent [63, 36], or are incapable of evaluating LLMs [34].

To address these challenges, we propose MMLA, the first comprehensive benchmark for multimodal language analysis, aimed at evaluating foundation models. Figure 1 provides an overview of MMLA. In this benchmark, we introduce six representative semantic dimensions for evaluation: *intent*, *emotion*, *dialogue act*, *sentiment*, *speaking style*, and *communication behavior*. These dimensions cover the most important cognitive-level semantic aspects of multimodal conversational interactions. We then collect nine publicly available multimodal language datasets, totaling over 61K multimodal utterances across more than 76 hours of video, with each utterance containing text, video, and audio modalities. These datasets span various sources, including both staged scenarios (e.g., TV series, films, TED talks) and real-world settings (e.g., spontaneous social media videos and motivational interviews). Detailed speaker demographics for these publicly available datasets are collected and reported in the appendix tables 3. Next, we evaluate state-of-the-art (SOTA) LLMs and MLLMs on MMLA. In particular, five branches of MLLMs are employed to leverage both language and video modalities. Additionally, three branches of LLMs that process only text are used for comparison with MLLMs, to assess the effect of non-verbal modalities. The sizes of these models range from 0.5B to 72B parameters. We apply zero-shot inference, supervised fine-tuning, and instruction tuning as evaluation methods.

Extensive experiments demonstrate that existing MLLMs show limited performance in understanding high-level cognitive semantics. Supervised fine-tuning can significantly enhance the multimodal capabilities and achieve new SOTA performance on most tasks. In particular, smaller-scale models show great potential with performance comparable to that of larger-scale models. Through instruction tuning, foundation models can successfully handle multiple tasks with a unified model, achieving performance comparable to supervised fine-tuning. However, even after tuning, these models still exhibit a significant limitation on these tasks, with average accuracy scores below 70%.

Our contributions are summarized as follows: (1) We propose MMLA, a large-scale multimodal language analysis benchmark containing 61K multimodal utterances drawn from over 76 hours of video. MMLA spans six core dimensions that are crucial for understanding high-level cognitive semantics. (2) To the best of our knowledge, MMLA is the first to comprehensively assess the capabilities of foundation models in multimodal language analysis, by evaluating nine mainstream models across three strategies. (3) Extensive experiments reveal new insights into foundation models for multimodal language analysis. MMLA also pushes the limits of existing MLLMs, providing a solid basis and promising new directions for further research. The code is publicly available, and the data are released under their respective licenses (see Appendix A for details).

2 Related Works

Multimodal Language Datasets. With the boom in multimodal language analysis, many significant tasks have emerged alongside the development of benchmark datasets. For example, early research focused on multimodal sentiment analysis and emotion recognition, and there are numerous datasets designed to analyze multilingual opinion sentiment [44, 68, 66, 37]. Zadeh et al. [71] constructed the first large-scale dataset in this field and additionally annotated emotion labels following Ekman's taxonomies [15]. However, these datasets only involve individual opinions and lack conversational interactions among multiple speakers. Busso et al. [3] introduced a dataset that records conversations between two speakers and annotates each utterance with emotion labels from nine categories in multimodal contexts. Nonetheless, dyadic sessions pose limitations when dealing with real-world multi-party scenarios. To address this, Poria et al. [45] provided sentiment and emotion labels for conversations taken from a TV series involving multiple speakers. These abundant resources have led to extensive research on designing effective multimodal fusion methods, including tensor operation-based [69, 39, 70] and transformer-based approaches [53, 48, 23, 20, 26, 75].

Beyond the relatively shallow semantics of sentiment or emotion, researchers have begun to explore more diverse and complex intent semantics in utterances, resulting in substantial new resources. Early work in this field analyzed authors' intents on social media platforms. For instance, Kruk et al. [30] proposed a taxonomy of eight intents based on rhetorical classes, and Zhang et al. [72] introduced four intent classes related to metaphor. However, these intents differ from those found in conversational scenarios. Saha et al. [49] annotated 12 dialogue acts drawn from the Switchboard[19] tag set for two multimodal emotion recognition datasets [3, 45]. Nevertheless, these dialogue acts are coarse-grained communicative intents that are not directly applicable to real-world applications [73]. To address this issue, Zhang et al. [73] proposed the first hierarchical intent taxonomy specifically designed for multimodal contexts and introduced the first multimodal conversational intent recognition dataset. Zhang et al. [74] subsequently extended this dataset into a larger-scale version that accommodates multi-party interactions and includes out-of-scope utterances, reflecting real-world conditions. In addition, some research has focused on individual speaking styles, such as humor [21] and sarcasm [5], which are driven by particular human intents, such as joking or mocking. Recently, Wu et al. [60] investigated more complex communication behaviors between clients and therapists through motivational interviewing in counseling scenarios.

Benchmarks. There are also multimodal benchmarks related to this work. For example, MultiBench [34] constructs a large-scale multimodal learning benchmark spanning various areas, such as healthcare and robotics. Nevertheless, it only covers dimensions related to affective computing and evaluates traditional multimodal machine learning methods without incorporating powerful MLLMs. To investigate the capability of MLLMs, numerous benchmarks have been proposed in recent years. However, most benchmarks focus on perceptual-level or elementary cognitive-level tasks such as visual recognition [32], optical character recognition [17], multimodal question answering [41], video content analysis [1, 16], scientific calculation [38], and visual reasoning [35]. While previous benchmarks cover diverse domains and tasks, none specifically target large-scale multimodal language analysis. MMLA is the first benchmark designed to advance this field in foundation models.

Multimodal Large Language Models. Multimodal large language models have emerged as a new paradigm in multimodal learning due to their superior scalability and cross-modal reasoning capabilities. For example, VideoLLaMA2 [7] introduces a Spatio-Temporal Convolution (STC) connector

Dimensions	Datasets	#C	#U	#Train	#Val	#Test	Video Hours	Source	#Video Length avg. / max.	#Text Length avg. / max.	Language
Intent	MIntRec	20	2,224	1,334	445	445	1.5	TV series	2.4 / 9.6	7.6 / 27.0	English
	MIntRec2.0	30	9,304	6,165	1,106	2,033	7.5	TV series	2.9 / 19.9	8.5 / 46.0	
Dialogue Act	MELD	12	9,989	6,992	999	1,998	8.8	TV series	3.2 / 41.1	8.6 / 72.0	English
	IEMOCAP	12	9,416	6,590	942	1,884	11.7	Improvised scripts	4.5 / 34.2	12.4 / 106.0	
Emotion	MELD	7	13,708	9,989	1,109	2,610	12.2	TV series	3.2 / 305.0	8.7 / 72.0	English
	IEMOCAP	6	7,532	5,237	521	1,622	9.6	Improvised scripts	4.6 / 34.2	12.8 / 106.0	
Sentiment	MOSI	2	2,199	1,284	229	686	2.6	Youtube	4.3 / 52.5	12.5 / 114.0	English
	CH-SIMS v2.0	3	4,403	2,722	647	1,034	4.3	TV series, films	3.6 / 42.7	1.8 / 7.0	Mandarin
Speaking Style	UR-FUNNY-v2	2	9,586	7,612	980	994	12.9	TED	4.8 / 325.7	16.3 / 126.0	English
	MUStARD	2	690	414	138	138	1.0	TV series	5.2 / 20.0	13.1 / 68.0	
Communication Behavior	Anno-MI (client)	3	4,713	3,123	461	1,128	10.8	YouTube	8.2 / 600.0	16.3 / 266.0	English
	Anno-MI (threapist)	4	4,773	3,161	472	1,139	12.1	& Vimeo	9.1 / 1316.1	17.9 / 205.0	

Table 1: Dataset statistics for each dimension in the MMLA benchmark. #C, #U, #Train, #Val, and #Test represent the number of label classes, utterances, training, validation, and testing samples, respectively. *avg.* and *max.* refer to the average and maximum lengths.

that excels in capturing spatiotemporal dynamics for audio-visual tasks. LLaVA-OneVision [31] pioneers cross-modal transfer learning, excelling in zero-shot video understanding despite being trained only on image datasets. LLaVA-Video [79] introduces a new video representation technique that allows maximizing the sampling of video frames, and a high-quality dataset is constructed to promote its instruction-following capability. Qwen2-VL [55] leads in vision-language understanding and generation, performing well in both zero-shot and few-shot settings. MiniCPM-V [64] innovates in model compression to enable efficient mobile deployment without compromising performance. Although current MLLMs perform well on various tasks, no benchmark evaluates their ability to handle complex multimodal language analysis.

3 MMLA Benchmark

3.1 Evaluation Dimensions

To comprehensively evaluate the complexity and diversity of human interactions, we select six representative dimensions across various linguistic and interactional levels: *intent*, *dialogue act*, *emotion*, *sentiment*, *speaking style*, and *communication behavior*. These dimensions collectively encapsulate the core aspects of multimodal language analysis [50, 18]. In particular, *intent* captures the ultimate purpose or goal of human communication, such as requesting information or making decisions [47]. In contrast, *dialogue act* is a more coarse-grained type of intent [49]. It typically focuses on the dynamic progression of communication, such as questioning or stating opinions [51]. Nonverbal signals (e.g., gaze shifts, gestures, and facial expressions) provide valuable clues to resolve ambiguities in both perspectives [73, 74].

Sentiment, *emotion*, and *speaking style* are three significant aspects often accompanying communicative interactions. *Sentiment* refers to the polarity (e.g., positive or negative) of subjective opinions [11], *emotion* conveys the speaker's internal psychological state (e.g., happiness, anger) [14], and *speaking style* refers to individual expressive variations in communication (e.g., sarcasm, humor) [43]. Multimodal cues (e.g., facial expressions and gestures) play a crucial role in inferring these communicative characteristics [62, 22, 42]. *Communication behavior* explores the interaction behaviors between individuals (e.g., sustain, change, and reflection), which facilitate the progression of conversations and exhibit social properties within groups [60]. Non-verbal signals (e.g., eye contact and gestures) can help uncover these behaviors and offer insights into modeling social cohesion [54]. Detailed information about the labels used for each dimension in each dataset can be found in Appendix B.

3.2 Data Sources

We collect nine typical publicly available multimodal language datasets corresponding to the evaluation dimensions. Detailed statistics for these datasets are provided in Table 1. We further summarize additional details on the release timeline of each dataset and on the overlap of speakers and scenes across data splits in Appendix F.1. For the *intent* dimension, we use two pioneering multimodal intent datasets, MIntRec [73] and MIntRec2.0 [74], which cover up to 30 intent classes commonly occurring in daily life. For the *emotion* dimension, we utilize two widely used multimodal emotion recognition datasets, MELD [45] and IEMOCAP [3], both containing Ekman's six universal emotion categories as suggested in [26]. Additionally, MELD includes a *neutral* class. For the *dialogue act* dimension, we use curated versions of the MELD and IEMOCAP datasets, with annotations provided by EmoTyDA [49]. These annotations consist of 12 commonly occurring classes selected from the SwitchBoard tag set [19]. For the *sentiment* dimension, we use two popular multimodal sentiment analysis datasets: MOSI [68] and CH-SIMS v2.0 [37]. Both are annotated with sentiment intensity values in the range of [-3, 3]. Following [67], we convert these annotations into polarity-based two-class and three-class labels for evaluation. For the *speaking style* dimension, we focus on two properties that play significant roles in social interactions: humor and sarcasm. We use UR-FUNNY-v2 [21] and MUStARD [5] for binary classification tasks, respectively. For the *communication behavior* dimension, we employ the Anno-MI [60] dataset, which involves motivational interviewing (MI) in counseling dialogues. This dataset is divided into two subsets, each analyzing three or four typical behaviors exhibited by clients and therapists. Details of the annotation quality assurance for each dataset are provided in Appendix C.

These datasets contain a wide variety of characters, scenes, and background contexts in both English and Mandarin. They are sourced from popular TV series (e.g., *Friends*, *The Big Bang Theory*, *Superstore*, etc.), films, online video-sharing platforms (e.g., YouTube, Vimeo, Bilibili), idea-sharing platforms (e.g., TED), and scripted dyadic sessions. We perform necessary cleaning and corrections to ensure the quality of each multimodal sample, aligning transcriptions, raw videos, and audio data. The datasets in the benchmark consist of 61,016 high-quality multimodal samples, totaling 76.6 hours of video, with 12,093 samples reserved for testing.

3.3 Method Overview

Zero-shot Inference. We leverage the generalization capabilities of foundation models for zero-shot inference. Specifically, for LLMs, the prompt template includes the transcribed utterances of speakers as text information, followed by a task-specific query with candidate labels. The LLM generates a response by predicting the next token in an autoregressive manner [2], which corresponds to the most appropriate label. For MLLMs, we extend this template by adding the special token <video> at the beginning of the instruction, with its number aligned to the number of videos. This ensures structured alignment across modalities, enhancing the model's capacity to process multimodal input. Details of the prompt templates used for inference can be found in Appendix D.

Supervised Fine-tuning (SFT). We further optimize foundation models to enhance their instruction-following capabilities using SFT techniques while employing the same instruction templates as used during inference. Fine-tuning is performed by minimizing the cross-entropy loss between the model's autoregressively predicted token probabilities and the ground-truth tokens corresponding to the labels. Let the input sequence be $x = (x_1, x_2, \dots, x_n)$ and the target sequence be $y = (y_1, y_2, \dots, y_m)$. The cross-entropy loss $\mathcal{L}_{\text{CE}}$ is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{m} \log P(y_t | x, y_{<t}; \theta),$$

where $P(y_t | x, y_{<t}; \theta)$ is the probability of token y_t given the input x, the previous tokens $y_{<t}$, and the model parameters θ. To ensure training stability and reduce computational cost, we adopt the Low-Rank Adaptation (LoRA) [25] technique, which significantly reduces the number of parameters to be fine-tuned while preserving the model's generalization capabilities.

Instruction Tuning (IT). Since SFT addresses only the single-task scenario, we further explore the generalization ability of foundation models on multiple tasks. We first combine the training data from all datasets of each task for training, then we use the same template as the other two strategies, with the difference being that the task is not limited to one. The optimization objective follows SFT and uses the candidate labels of each task as supervised targets.

4 Experiments

Evaluation Metrics. We employ six commonly used metrics: accuracy (ACC), weighted F1-score (WF1), weighted precision (WP), macro F1-score (F1), recall (R), and precision (P) for evaluation, as suggested in the literature [74, 45, 83, 68, 22]. In particular, we report the primary results of ACC in this paper, with additional results for the remaining metrics provided in the Appendices.

Evaluation Baselines. We apply the three evaluation methods as described in Section 3.3 on advanced LLMs and MLLMs as baselines. We also compare the foundation models with SOTA multimodal machine learning (MML) methods.

- **LLMs**. Three series of different parameter scales of unimodal foundation models are included: Llama-3 [13] (8B), InternLM-2.5 (7B) [4], and Qwen2 [61] (0.5B, 1.5B, and 7B).
- **MLLMs**. Five series of different parameter scales of multimodal foundation models are included: VideoLLaMA2 [7] (7B), Qwen2-VL [55] (7B and 72B), LLaVA-Video [79] (7B and 72B), LLaVA-OneVision (LLaVA-OV) [31] (7B and 72B), and MiniCPM-V-2.6 [64] (8B). For language decoding, the first series use Mistral [28], and the last four series use Qwen2 with the same parameter scale as the MLLM. We follow the same vision encoders as those in the corresponding released open-source models. We also apply zero-shot inference on one closed-source MLLM, GPT-4o [27] as a baseline.
- **MML Methods**. We collect open-source MML methods with SOTA performance for each dataset for a detailed comparison. Specifically, for MIntRec: MIntOOD [76], MIntRec2.0: MulT [53], MELD and IEMOCAP: UniMSE [26], MELD-DA: TCL-MAP [83], IEMOCAP-DA: MIntOOD [76], MOSI: MMML [59], CH-SIMS v2.0: ALMT [77], UR-FUNNY-v2 and MUStARD: SimMMDG [12]. The results are reported as they appear in the corresponding papers.

Experimental Setup. We mostly follow the original data splits for training, validation, and testing for each dataset, as detailed in Table 1. Each sample consists of text and video data aligned at the utterance level for speakers. For SFT and IT methods, we utilize LLaMAFactory [82] for all LLMs and Qwen2-VL, SwiFT [81] for MiniCPM-V-2.6, LLaVA-NeXT [1] for LLaVA-OV and LLaVA-Video, and VideoLLaMA2 using its own public code [2], respectively. We employ FlashAttention-2 [10] to optimize the attention modules of transformers, reducing memory and time costs. Besides, we leverage the DeepSpeed library for distributed training (e.g., using ZeRO-3 for memory optimization) and parallel computation. The precision type is set to BF16, offering reduced computational costs compared to FP16 or FP32. The learning rates range from 2e-5 to 1e-3, and a cosine learning rate scheduler with warmup ratios from 0.1 to 0.3 is applied. The training batch sizes are chosen from $\{4, 8, 16, 24\}$. The rank and α parameters of the LoRA module are set to $\{8, 16, 64, 128\}$ and $\{16, 32, 128, 256\}$, respectively. All experiments are conducted on NVIDIA A100 GPUs. We monitor model accuracy on the validation set to select the best checkpoint for inference. Details of the used hyperparameters and the full experimental results are shown in Appendix E. To quantify run-to-run variability, we repeat Qwen2-VL-7B under SFT and IT with three random seeds and report mean±std across all six metrics in Appendix H (Table 5). We report evaluation efficiency as inference time in Appendix G.

5 Results and Discussion

5.1 Main Results

To clearly illustrate the performance differences between foundation models on the MMLA benchmark, we present the ranking statistics of the average accuracy (ACC) across all combined testing sets. Specifically, the zero-shot inference performance is shown in Figure 2, while the performance after SFT and IT is shown in Figure 3. We find some interesting and new insights as below.

RANK	MODELS	TYPE	ACC
1	GPT-4o	MLLM	52.60
2	Qwen2-VL-72B	MLLM	52.55
3	LLaVA-OV-72B	MLLM	52.44
4	LLaVA-Video-72B	MLLM	51.64
5	InternLM2.5-7B	LLM	50.28
6	Qwen2-7B	LLM	48.45
7	Qwen2-VL-7B	MLLM	47.12
8	Llama3-8B	LLM	44.06
9	LLaVA-Video-7B	MLLM	43.32
10	VideoLLaMA2-7B	MLLM	42.82
11	LLaVA-OV-7B	MLLM	40.65
12	Qwen2-1.5B	LLM	40.61
13	MiniCPM-V-2.6-8B	MLLM	37.03
14	Qwen2-0.5B	LLM	22.14

Figure 2: Rank of foundation models after zero-shot inference.

Comparable Performance between LLMs and MLLMs in Zero-shot Inference. As shown in Figure 2, the closed-source GPT-4o achieves the best performance, and the three open-source 72B MLLMs occupy the remaining positions in the top four. This is unsurprising, as these models contain more parameters and therefore exhibit stronger generalization and reasoning capabilities, consistent with the scaling laws [29]. However, we note that the much smaller-scale LLM

[1] https://github.com/LLaVA-VL/LLaVA-NeXT
[2] https://github.com/DAMO-NLP-SG/VideoLLaMA2

InternLM2.5-7B achieves comparable performance, within approximately a 2% difference. Furthermore, among models with a similar scale (7B or 8B parameters), most MLLMs exhibit lower performance than LLMs. For example, InternLM2.5 and Qwen2 outperform most MLLMs (e.g., LLaVA-Video, VideoLLaMA2) by 5∼8%. These results indicate that existing MLLMs have significant limitations in leveraging non-verbal information to capture complex high-level semantics without supervision from domain-specific data.

Small MLLMs Rival Large Ones After SFT and IT. Although MLLMs exhibit substantial performance gaps in zero-shot inference, parameter size matters far less once they're trained with SFT or IT. For example, as shown in Figure 3, 7B MLLMs trained with SFT achieve 67.47∼68.30% ACC, while their 72B counterparts reach 68.44∼69.18%, a performance gap of only 1∼2%. Specially, the 8B MiniCPM-V-2.6 after SFT attains second place with 68.88%, only 0.3% lower in ACC than the top model, and surpasses several much larger MLLMs. 7B, 8B, and 72B MLLMs trained with IT also achieve ACC scores within 2% of each other (i.e., 67.25∼68.87%). These results show that small-scale well-trained MLLMs can capture the cognitive semantics underlying human language, suggesting lightweight foundation models are feasible and significantly reduce costs.

Comparable Performance of MLLMs Between SFT and IT. Although SFT offers the advantage of task-specific fine-tuning to boost individual task performance, we observe that MLLMs trained via IT can achieve comparable results or even surpass those of SFT when evaluated on multiple tasks. For example, in Figure 3, the 72B LLaVA-Video ranks third and outperforms its SFT counterpart by 0.43%. Qwen2-VL shows only a slight performance drop after IT (0.54% for 72B and 0.26% for 7B). In contrast, MiniCPM-V-2.6 suffers a more pronounced decline of 1.63% compared with its SFT counterpart, and some models (e.g., LLaVA-Video and LLaVA-OV) are omitted because they exhibit severe hallucinations, producing irrelevant outputs following IT. However, the strong performance of certain MLLMs highlights the potential of training a unified model to excel across diverse tasks without the overhead of maintaining multiple models, thereby demonstrating the robust generalization capabilities of MLLMs on this task.

RANK	MODELS	TYPE	ACC
1	Qwen2-VL-72B (SFT)	MLLM	69.18
2	MiniCPM-V-2.6-8B (SFT)	MLLM	68.88
3	LLaVA-Video-72B (IT)	MLLM	68.87
4	LLaVA-OV-72B (SFT)	MLLM	68.67
5	Qwen2-VL-72B (IT)	MLLM	68.64
6	LLaVA-Video-72B (SFT)	MLLM	68.44
7	VideoLLaMA2-7B (SFT)	MLLM	68.30
8	Qwen2-VL-7B (SFT)	MLLM	67.60
9	LLaVA-OV-7B (SFT)	MLLM	67.54
10	LLaVA-Video-7B (SFT)	MLLM	67.47
11	Qwen2-VL-7B (IT)	MLLM	67.34
12	MiniCPM-V-2.6-8B (IT)	MLLM	67.25
13	Llama3-8B (SFT)	LLM	66.18
14	Qwen2-7B (SFT)	LLM	66.15
15	InternLM2.5-7B (SFT)	LLM	65.72
16	Qwen2-7B (IT)	LLM	64.58
17	InternLM2.5-7B (IT)	LLM	64.41
18	Llama3-8B (IT)	LLM	64.16
19	Qwen2-1.5B (SFT)	LLM	64.00
20	Qwen2-0.5B (SFT)	LLM	62.80

Figure 3: Rank of foundation models after SFT and IT.

MLLMs Still Face Challenges on the MMLA Benchmark. From Figures 2 and 3, we observe that the best MLLM in zero-shot inference (GPT-4o) achieves only 52.6% ACC, and the best model after training with supervised data (72B Qwen2-VL) reaches just 69.18% ACC, still exhibiting huge limitations. These findings underscore the difficulty and importance of the MMLA benchmark, pushing the boundaries of existing MLLMs and laying a solid foundation for future related research.

5.2 Fine-grained Performance on Different Dimensions

To further investigate fine-grained performance across different dimensions, we present the results of three methods for each MLLM and LLM on every dataset, as shown in Figures 4 and 5.

Foundation Models Struggle with Zero-Shot Inference. As shown in Figure 4, zero-shot performance is substantially limited, with ACC scores below 60% on many challenging semantic dimensions (e.g., *Intent*, *Emotion*, *Dialogue Act*, and *Communication Behavior*). This shortcoming arises because these dimensions typically involve numerous categories with nuanced differences. In contrast, performance on the *Sentiment* and *Speaking Style* dimensions is generally higher because these tasks are simpler, requiring only two or three classes to be distinguished. GPT-4o achieves the best results in several dimensions, such as *Intent*, *Dialogue Act*, and *Sentiment*, highlighting its strong ability to leverage multiple modalities for reasoning. However, it still struggles with tasks like sarcasm detection, emotion recognition, and communication behavior recognition, likely due to interference from scene context, background, and characters. Finally, while LLMs show performance comparable to or better than MLLMs of the same parameter scale, their scores remain below 60% in most cases, underscoring the significant limitations of current foundation models on our benchmark.

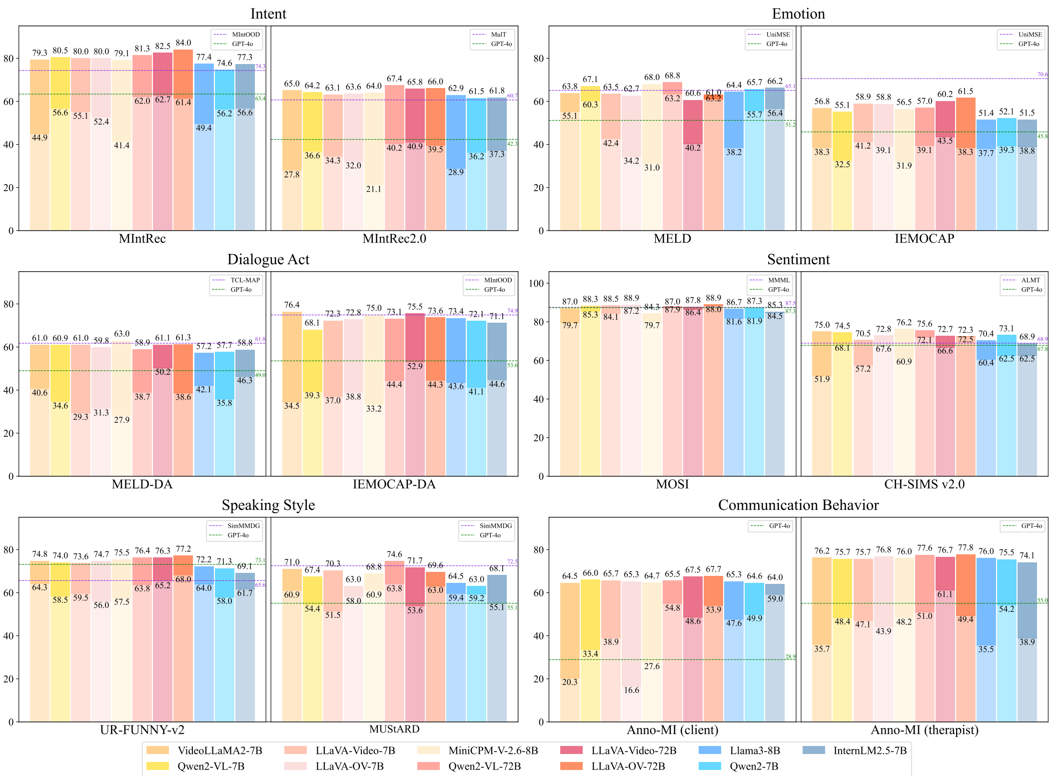

Figure 4: Fine-grained zero-shot inference and SFT performance (ACC). Within each bar, the light-colored lower segment corresponds to zero-shot inference performance, while the darker upper segment represents the additional gains from SFT. The performance of SOTA MML methods (if available) and GPT-4o are indicated with purple and green dashed lines, respectively.

Foundation Models Significantly Improve After SFT. As shown in Figure 4, foundation models exhibit a notable performance boost after SFT. For example, ACC scores increase by 20∼40% on *Intent*, 10∼40% on *Dialogue Act*, 4∼20% on *Speaking Style*, and 5∼50% on *Communication Behavior*. Specifically, MiniCPM-V-2.6 achieves improvements of over 30% across most dimensions. These results demonstrate that training with supervised instruction data effectively helps MLLMs and LLMs distinguish complex semantic categories. Moreover, although both MLLMs and LLMs benefit from SFT, MLLMs consistently outperform LLMs (see Figure 3), despite showing similar zero-shot performance. This suggests that SFT not only aligns modalities better to activate multimodal reasoning, but also that incorporating non-verbal information reduces hallucinations more effectively than using text alone. Finally, MLLMs after SFT set new state-of-the-art results on most datasets except IEMOCAP and MUStARD, highlighting their great potential in multimodal language analysis.

Foundation Models Master Multiple Tasks After IT. As shown in Figure 5, MLLMs after IT can simultaneously match or surpass previous SOTA methods on most datasets. In particular, *72B Qwen2-VL is the first to exceed human performance on MIntRec [73] (86.3% vs. 85.5%)*, marking remarkable progress toward human-level semantic comprehension. 72B LLaVA-Video improves over the SOTA method by 6.3% and approaches human performance on MIntRec2.0 [74]. Similarly, most MLLMs exhibit superior results on sentiment analysis (Ch-sims-v2), humor detection (UR-FUNNY-v2), and emotion recognition (MELD). We also observe that the small-scale MLLM (i.e., 8B MiniCPM-V-2.6) outperforms SOTA on seven datasets across five dimensions and achieves the best score on Ch-sims-v2. Moreover, small-scale MLLMs outperform LLMs on nearly every dataset and task, underscoring that IT enhances multimodal reasoning and demonstrating the potential of training a unified MLLM to tackle multiple complex multimodal language tasks.

5.3 Scalability of Foundation Models on MMLA

To examine the scalability of foundation models [29], we analyze the effect of parameter scale using Qwen2 and Qwen2-VL, presenting both zero-shot inference and SFT results in Figure 6.

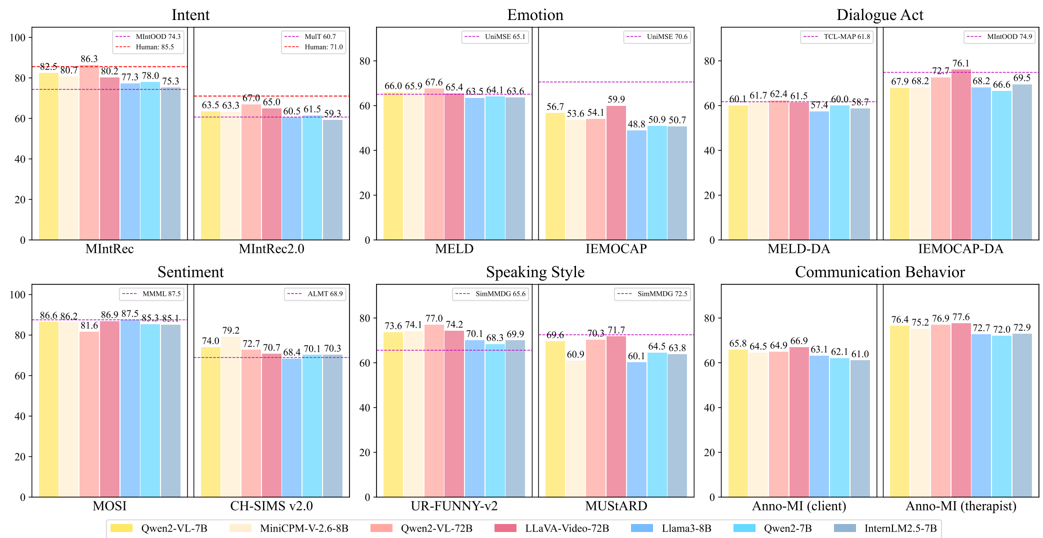

Figure 5: Fine-grained performance (ACC) of instruction-tuned MLLMs and LLMs on each dataset across six dimensions. The performance of SOTA MML methods and humans are indicated with dashed lines, if available.

Scaling Performance of Zero-Shot Inference. In zero-shot inference, scaling Qwen2 from 0.5B to 1.5B parameters achieves significant improvements across all dimensions except *Communication Behavior*. When scaling from 1.5B to 7B, performance gains accelerate on *Intent* and *Communication Behavior*, slow down on *Emotion* and *Dialogue Act*, and even slightly decrease on *Sentiment* and *Speaking Style*. This phenomenon indicates that larger gains occur with smaller scale changes. When moving from Qwen2 to Qwen2-VL, performance is comparable or better in all dimensions except for *Communication Behavior*, which shows a dramatic drop. However, scaling Qwen2-VL from 7B to 72B yields substantial improvements, further validating the scalability of MLLMs.

Scaling Performance of SFT. After SFT, scaling Qwen2 from 0.5B to 7B yields modest improvements of 3$\sim$5% on the *Intent*, *Sentiment*, *Speaking Style*, and *Emotion* dimensions, with limited gains of less than 2% on *Communication Behavior* and *Dialogue Act*. Besides, scaling Qwen2-VL from 7B to 72B achieves substantial improvements of over 5% on *Speaking Style* and *Intent* dimensions, while yielding under 2% gains in *Sentiment*, *Communication Behavior*, and *Dialogue Act*. These results suggest that simply enlarging model parameters provides little benefit for analyzing complex multimodal language semantics when using supervised instructions as prior knowledge. They also highlight the significant challenge posed by this benchmark and underscore the need to design appropriate architectures and curate high-quality data for learning high-level cognitive semantics.

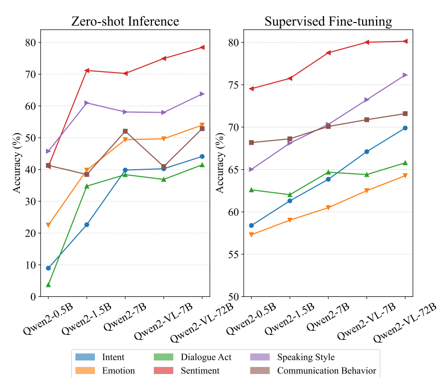

Figure 6: Scalability of Qwen2 and Qwen2-VL on the MMLA benchmark.

6 Conclusions

This paper proposes MMLA, the first large-scale benchmark for evaluating foundation models on multimodal language analysis. It covers six core semantic dimensions across more than 61,000 utterances from nine diverse datasets spanning text, audio, and video modalities. We evaluate five branches of MLLMs and three branches of LLMs, ranging from 0.5B to 72B parameters, using three methods to provide a comprehensive analysis. This benchmark yields several new insights. First, existing MLLMs exhibit poor capabilities and offer no advantage over LLMs in zero-shot inference. Second, supervised fine-tuning (SFT) effectively activates MLLMs, enabling them to leverage non-verbal modalities to understand cognitive-level semantics, and achieves substantial improvements over LLMs. Third, instruction tuning (IT) can further fine-tune a unified model to achieve comparable or better performance on all SFT tasks. Interestingly, we find that smaller MLLMs, after both SFT and IT, demonstrate enormous potential, achieving performance comparable to much larger models while significantly reducing computational costs. Finally, existing MLLMs still face significant challenges, with an average accuracy below 70%, underscoring the importance and difficulty of the proposed benchmark. MMLA establishes a rigorous foundation for advancing multimodal language understanding and cognitive-level human–AI interaction. Details of the limitations and broader societal impacts appear in Appendices I and J.

Acknowledgments and Disclosure of Funding

This research was supported by the National Natural Science Foundation of China (Grant No. 62173195). The authors would like to express their sincere gratitude to Minghao Xiao and Shihao Zhou for their valuable assistance in the early stages of this work, particularly in data collection and experimental setup.

References

[1] Kirolos Ataallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding. *arXiv preprint arXiv:2406.19875*, 2024.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901, 2020.

[3] 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.

[4] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

[5] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, 2019.

[6] Yifan Chen, Kuntao Li, Weixing Mai, Qiaofeng Wu, Yun Xue, and Fenghuan Li. D2R: Dual-branch dynamic routing network for multimodal sentiment detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3536–3547, 2024.

[7] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[8] Lukas Christ, Shahin Amiriparian, Alexander Kathan, Niklas Müller, Andreas König, and Björn W. Schuller. Towards multimodal prediction of spontaneous humor: A novel dataset and first results. *IEEE Transactions on Affective Computing*, 2024.

[9] Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. User attention-guided multimodal dialog systems. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 445–454, 2019.

[10] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

[11] Ringki Das and Thoudam Doren Singh. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):270–307, 2023.

[12] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. Simmmdg: A simple and effective framework for multi-modal domain generalization. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 78674–78695, 2023.

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[14] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[15] Paul Ekman, Wallace V Freisen, and Sonia Ancoli. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125, 1980.

[16] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024.

[17] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[18] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

[19] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520, 1992.

[20] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, 2021.

[21] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2046–2056, 2019.

[22] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12972–12980, 2021.

[23] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020.

[24] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7037–7041, 2022.

[25] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[26] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, 2022.

[27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[28] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[29] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[30] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4622–4632, 2019.

[31] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[32] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742, 2023.

[34] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Proceedings of the Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 34892–34916, 2023.

[36] Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751*, 2024.

[37] Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make Acoustic and Visual Cues Matter: CH-SIMS v2.0 Dataset and AV-Mixup Consistent Module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 247–258, 2022.

[38] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision*, pages 216–233, 2025.

[39] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 2247–2256, 2018.

[40] Trisha Mittal, Sanjoy Chowdhury, Pooja Guhan, Snikitha Chelluri, and Dinesh Manocha. Towards determining perceived audience intent for multimodal social media posts using the theory of reasoned action. *Scientific Reports*, 14:10606, 2024.

[41] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023.

[42] Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh, Hunar Singh, and Vinay P. Namboodiri. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision*, pages 576–585, 2021.

[43] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577, 2003.

[44] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 973–982, 2013.

[45] 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 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019.

[46] Shraman Pramanick, Aniket Roy, and Vishal M. Patel Johns. Multimodal learning using optimal transport for sarcasm and humor detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 546–556, 2022.

[47] Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. A survey on spoken language understanding: Recent advances and new frontiers. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 4577–4584, 2021.

[48] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, 2020.

[49] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, 2020.

[50] Tobias Schröder, Terrence C Stewart, and Paul Thagard. Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive Science*, 38(5):851–880, 2014.

[51] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

[52] Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26963–26973, June 2024.

[53] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019.

[54] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[56] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 121475–121499, 2024.

[57] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7216–7223, 2019.

[58] Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Babak Damavandi, Nanyun Peng, and Seungwhan Moon. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 6273–6291, 2023.

[59] Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. Multimodal multi-loss fusion network for sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3588–3602, 2024.

[60] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-mi: A dataset of expert-annotated counselling dialogues. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6177–6181, 2022.

[61] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[62] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651, 2022.

[63] Qu Yang, Mang Ye, and Bo Du. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*, 2024.

[64] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[65] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[66] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 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*, pages 3718–3727, 2020.

[67] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10790–10797, 2021.

[68] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[69] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[70] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5634–5641, 2018.

[71] 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 56th Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246, 2018.

[72] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3214–3225, 2021.

[73] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1688–1697, 2022.

[74] Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Wenrui Li, Yanting Chen, et al. Mintrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *Proceedings of the the 12th International Conference on Learning Representations*, 2024.

[75] Hanlei Zhang, Hua Xu, Fei Long, Xin Wang, and Kai Gao. Unsupervised multimodal clustering for semantics discovery in multimodal utterances. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 18–35, 2024.

[76] Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. Multimodal classification and out-of-distribution detection for multimodal intent understanding. *arXiv preprint arXiv:2412.12453*, 2024.

[77] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv preprint arXiv:2310.05804*, 2023.

[78] Xinxin Zhang, Jun Sun, Simin Hong, and Taihao Li. Amanda: Adaptively modality-balanced domain adaptation for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics: ACL*, pages 14448–14458, 2024.

[79] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[80] Wenye Zhao, Qingbao Huang, Dongsheng Xu, and Peizhi Zhao. Multi-modal sarcasm generation: dataset and solution. In *Findings of the Association for Computational Linguistics: ACL*, pages 5601–5613, 2023.

[81] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 29733–29735, 2025.

[82] Yaowei Zheng, Richong Zhang, Junhao Zhang, YeYanhan YeYanhan, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 400–410, 2024.

[83] Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17114–17122, 2024.

[84] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *Proceedings of the the 12th International Conference on Learning Representations*, 2024.

[85] Zhihong Zhu, Xuxin Cheng, Zhaorun Chen, Yuyan Chen, Yunyan Zhang, Xian Wu, Yefeng Zheng, and Bowen Xing. Inmu-net: advancing multi-modal intent detection via information bottleneck and multi-sensory processing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 515–524, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We make the main claims in abstract and introduction, which clearly state our contributions and research scope.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our works in Appendix I.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This work provides a comprehensive benchmark for evaluating foundation models in multimodal language analysis and does not include any new theoretical derivations or analytical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All implementation details required for replication are provided in Section 4, in Appendices D and E, and in Tables 9–12.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All videos and text utterances for each dataset are available under their respective licenses on Hugging Face[3] and Google Drive[4], and the complete code for all methods of both LLMs and MLLMs is publicly available on GitHub[5] to facilitate reproducibility.

[3]https://huggingface.co/datasets/THUIAR/MMLA-Datasets

[4]https://drive.google.com/drive/folders/1nCkhkz72F6ucseB73XVbqCaDG-pjhpSS

[5]https://github.com/thuiar/MMLA

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The data splits are detailed in Section 1. Core hyperparameters and experimental settings are described in Section 4, and the full hyperparameters are listed in Tables 9–12.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars are reported only for Qwen2-VL-7B in the appendix, as repeating all experiments to estimate errors for every model would be computationally and temporally prohibitive.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As stated in Section 4, all experiments are run on NVIDIA A100 GPUs, each with 40 GB of memory, using approximately eight GPUs for most experiments and one GPU for zero-shot inference of LLMs.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We affirm that all research presented in this paper fully adheres to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader societal impact of our works in Appendix J.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not present any such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All data were gathered under the terms of their respective licenses, which are detailed in Appendix A.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the complete and systematic benchmark code—including running scripts, configuration files, and frameworks—and leaderboards for both zero-shot inference and fine-tuning of each foundation model in MMLA, available in the main paper and online.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or experiments with human subjects, and includes only publicly available datasets.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or experiments with human subjects, and includes only publicly available datasets.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Section 3.3 presents three methods for evaluating LLMs and MLLMs. Section 4 describes the LLMs and MLLMs employed, and Appendix D details the prompts used for both model types.

A License

This benchmark uses nine datasets, each of which is employed strictly in accordance with its official license and exclusively for academic research purposes. We fully respect the datasets' copyright policies, license requirements, and ethical standards. For those datasets whose licenses explicitly permit redistribution, we release the original video data (e.g., MIntRec[6], MIntRec2.0[7], MELD[8], UR-FUNNY-v2[9], MUStARD[10], MELD-DA[11], CH-SIMS v2.0[12], and Anno-MI[13]). For datasets that restrict video redistribution, users should obtain the videos directly from their official repositories (e.g., MOSI[14], IEMOCAP and IEMOCAP-DA[15]). In compliance with all relevant licenses, we also provide the original textual data unchanged, together with the specific dataset splits used in our experiments. This approach ensures reproducibility and academic transparency while strictly adhering to copyright obligations and protecting the privacy of individuals featured in the videos.

B Used Labels for Each Dataset

- **Intent**. The MIntRec dataset uses 20 predefined intent categories derived from two coarse-grained classes (i.e., *achieve goals* and *express emotions and attitudes*), as described in [73]. These categories are: *complain*, *praise*, *apologize*, *thank*, *criticize*, *agree*, *taunt*, *flaunt*, *joke*, *oppose*, *comfort*, *care*, *inform*, *advise*, *arrange*, *introduce*, *leave*, *prevent*, *greet*, and *ask for help*. MIntRec2.0 adds 10 more labels (i.e., *doubt*, *acknowledge*, *refuse*, *warn*, *emphasize*, *ask for opinions*, *confirm*, *explain*, *invite*, and *plan*) to the original 20. We use the in-scope portion of this dataset for intent recognition.
- **Dialogue Act**. The MELD-DA and IEMOCAP-DA datasets select the 12 most frequent dialogue-act tags in everyday conversation, based on the 42 acts defined in [51]. The chosen tags are: *greeting*, *question*, *answer*, *statement-opinion*, *statement-non-opinion*, *apology*, *command*, *agreement*, *disagreement*, *acknowledge*, *backchannel*, and *others*.
- **Emotion**. The IEMOCAP dataset adopts Ekman's six universal emotions (as in prior work [57, 24]): *angry*, *happy*, *sad*, *neutral*, *frustrated*, and *excited*. The MELD dataset uses seven emotion classes [45]: *neutral*, *surprise*, *fear*, *sadness*, *joy*, *anger*, and *disgust*.
- **Sentiment**. For the MOSI and CH-SIMS v2.0 datasets, sentiment intensity scores range from -3 to 3 and are mapped to two- or three-way polarity classes (e.g., *positive*, *neutral*, *negative*), as recommended in [68, 37].
- **Speaking Style**. The UR-FUNNY-v2 and MUStARD datasets both perform binary classification tasks: humor detection (*humorous* vs. *serious*) and sarcasm detection (*sarcastic* vs. *sincere*), respectively.
- **Communication Behavior**. The Anno-MI dataset is split into two parts for counseling dialogue analysis. The first part contains four therapist communication skills: *question*, *input*, *reflection*, and *other*. The second part contains three client talk types: *change*, *neutral*, and *sustain*.

C Assurance of Annotation Quality

We employ rigorous procedures to select datasets with high-quality annotations. Quality is ensured through the following strategies and statistical measures for each dataset:

- **MIntRec** [73] and **MIntRec2.0** [74]. Intent labels are assigned by majority voting (three of five and two of three annotators, respectively). Fleiss's kappa values of 0.88 for MIntRec and 0.69 for MIntRec2.0 indicate excellent and substantial agreement, respectively.
- **MELD** [45] and **IEMOCAP** [3]. Emotion labels are determined by three-annotator majority voting, yielding Fleiss's kappa values of 0.43 (MELD) and 0.40 (IEMOCAP), reflecting acceptable reliability for emotion annotation.
- **MELD-DA** and **IEMOCAP-DA** [49]. Dialogue-act labels are annotated by three experts, achieving over 80% inter-annotator agreement.
- **MUStARD** [5]. Three annotators achieved a Cohen's kappa of 0.588 for sarcasm detection.

[6] https://github.com/thuiar/MIntRec
[7] https://github.com/thuiar/MIntRec2.0
[8] https://github.com/declare-lab/MELD
[9] https://github.com/ROC-HCI/UR-FUNNY
[10] https://github.com/soujanyaporia/MUStARD
[11] https://github.com/sahatulika15/EMOTyDA
[12] https://github.com/thuiar/ch-sims-v2
[13] https://github.com/uccollab/AnnoMI
[14] https://github.com/matsuolab/CMU-MultimodalSDK
[15] https://sail.usc.edu/iemocap

- **UR-FUNNY-v2** [21]. The original UR-FUNNY was annotated based on direct laughter markers in punchlines; noisy and overlapping instances were removed to form the second version, which we use in our benchmark.
- **MOSI** [68]. Five master workers (approval rate > 95%) annotated sentiment intensity, with a Krippendorff's alpha of 0.77.
- **Anno-MI** [60]. Ten therapists from the International Organization of Authoritative Motivational-Interviewing Trainers annotated communication behavior labels, with Fleiss's kappa values of 0.74 (therapist) and 0.47 (client), indicating substantial and moderate agreement, respectively.
- **CH-SIMS v2.0** [37]. This version corrects potential errors in the original CH-SIMS [66]. Seven well-trained annotators rated sentiment intensity: the highest and lowest scores were removed, and the average of the remaining five was mapped to discrete sentiment labels. We use this latest release, which also addresses potential misalignment issues.

D Used Prompts

Zero-shot inference, supervised fine-tuning (SFT), and instruction tuning (IT) employ the following template:

```
You are presented with a video in which the speaker says: <context>.
Based on the textual, visual, and audio content, what is the <dimension>
of this speaker?
The candidate labels for <dimension> are: <the list of labels>.
Respond in the following format: <dimension>: label.
Only one label should be provided.
```

Here, <*context*> denotes the the speaker's utterance, while <*dimension*> refers to one of the six evaluation dimensions described in Section 3.1. The placeholder <*the list of labels*> corresponds to the specific label set for each dimension (cf. Appendix B). In SFT, the ground-truth dimension and label are provided for supervised training. In IT, neither the queried dimension nor its label set is fixed to a single dataset, unlike in SFT.

E Detailed Experimental Results

Due to space constraints, the main paper presents only a subset of the results. Here, we provide the complete results for zero-shot inference, supervised fine-tuning, and instruction-tuning across six evaluation metrics (ACC, WF1, WP, F1, P, R) in Table 6, Table 7, and Table 8. For both LLMs and MLLMs, the best and second-best results are highlighted in bold and underline, respectively.

The results align with the discussions and conclusions in the main paper. For zero-shot inference, multimodal models with 72B parameters achieve the best overall performance across all datasets. However, when comparing LLMs and MLLMs of the same scale, LLMs often exhibit competitive or even superior performance. After supervised fine-tuning, MLLMs show significant improvements and surpass LLMs on almost all datasets across all six metrics, underscoring the importance of incorporating non-verbal modalities for cognitively demanding tasks. After instruction-tuning, both the 7B and 72B MLLMs achieve excellent performance on all tasks, with results comparable to or better than those from supervised fine-tuning, indicating the potential of small-scale MLLMs to solve multiple tasks simultaneously. Moreover, under this evaluation protocol, MLLMs also outperform LLMs, further confirming the benefit of leveraging non-verbal modalities.

We also evaluate one powerful closed-source MLLM, GPT-4o. Specifically, we use OpenAI's GPT-4o API for zero-shot inference with the same prompts as in Appendix D. During inference, we find that GPT-4o can be overly cautious with certain videos. For example, it sometimes fails to select a label from the candidate list and instead outputs responses like: *I'm unable to determine the dimension based on the given information*. To address this, we iteratively modify the prompts based on such outputs until GPT-4o consistently chooses a label from the list. The final results, shown in Table 6, demonstrate that GPT-4o achieves the best or second-best performance on most metrics across datasets, highlighting its effectiveness on this challenging task.

Details of the hyperparameters used for supervised fine-tuning (SFT) and instruction-tuning (IT) of all LLMs and MLLMs are provided in Table 9, Table 10, Table 11, and Table 12.

F Additional dataset details

F.1 Dataset collection timeline and speaker/scene overlap

The MMLA Benchmark aggregates nine widely used multimodal language analysis datasets to evaluate the generalization ability of large-scale models in this field. As shown in Table 2, five out of nine datasets contain no overlap of speakers or scenes, enabling a more rigorous assessment of model generalization. The remaining datasets exhibit partial overlap, which helps isolate semantic understanding by reducing extraneous variability. Together, these datasets provide a comprehensive and balanced evaluation setting. This information will be further updated in the final version. Redistribution of the benchmark is strictly limited to academic research and subject to the original dataset licenses.

Dataset	Release Time	Overlap of Speakers and Scenes
MIntRec	2022/10	Yes
MIntRec2.0	2024/01	Yes
MELD	2019/05	Yes
MELD-DA and IEMOCAP-DA	2020/07	–
IEMOCAP	2008/12	No
MOSI	2016/06	No
CH-SIMS v2.0	2022/09	Yes
UR-FUNNY v2	2019/11	No
MUStARD	2019/07	Yes
Anno-MI	2023/03	No

Table 2: Summary of datasets included in the MMLA Benchmark.The table lists each dataset's release time and whether there exists an overlap of speakers or scenes between the training, validation, and test splits.

F.2 Speaker Demographics across Datasets

We summarize the speaker demographics of the datasets used in the MMLA benchmark in Table 3. For datasets derived from scripted or publicly released video materials, including *MIntRec*, *MIntRec2.0*, *MELD*, *MELD-DA*, *MUStARD*, and *IEMOCAP/IEMOCAP-DA*, we provide detailed statistics on key characters, actors, proportions, gender, ethnicity, and age. These datasets primarily originate from well-known television series such as *Friends*, *The Big Bang Theory*, and *Superstore*, as well as other scripted settings involving professional or trained actors. For datasets without explicit speaker-level metadata, we briefly describe their composition below.

The **MOSI** dataset includes 89 distinct speakers (41 female and 48 male), most aged between 20 and 30. Speakers represent diverse ethnic backgrounds (e.g., Caucasian, African American, Hispanic, Asian), and all recordings are in English, sourced from YouTube videos originating in the United States and the United Kingdom.

The **UR-FUNNY-v2** dataset contains 1,866 videos from 1,741 distinct TED speakers covering 417 topics, reflecting a broad range of speaking styles, contexts, and presentation settings.

Overall, the benchmark incorporates both professionally acted and real-world spontaneous speech data, enabling a comprehensive evaluation of multimodal language understanding models across diverse demographic and communicative conditions.

G Model Evaluation Efficiency

To evaluate the scalability of our benchmark, we measured the inference time required for various foundation models on the test split (12,093 samples). All experiments were conducted using four NVIDIA A800-80G GPUs. The remaining data are reserved for training and validation.

Overall, most models with 7B parameters completed inference within two hours, while larger variants (72B) typically required 5–7 hours. These results suggest that large-scale evaluation across 61K multimodal utterances is computationally feasible for current foundation models.

H Performance Variability

To further examine the stability of model performance, we conducted additional statistical analyses on the Qwen2-VL-7B model under both supervised fine-tuning (SFT) and instruction tuning (IT) settings across all datasets in the MMLA benchmark. Each experiment was repeated three times with different random seeds, and the mean ± standard deviation (std) of six evaluation metrics was computed. The results (see Table 5) are consistent with the main findings reported in the paper: the overall performance trends remain unchanged, and the model demonstrates stable behavior across most datasets, confirming the robustness of our conclusions in the main text.

Table 3: Speaker demographics across datasets in the MMLA benchmark. For datasets lacking specific fields (e.g., IEMOCAP), missing values are denoted by "–".

	Character	Actor	Prop.	Gender	Race / Ethn.	Age	Series
MIntRec & MIntRec2.0	Rachel Green	Jennifer Aniston	13.5%	F	White	24–34 (1994–2004)	Friends
	Monica Geller	Courteney Cox	14.6%	F	White	30–40 (1994–2004)	Friends
	Phoebe Buffay	Lisa Kudrow	12.2%	F	White	31–41 (1994–2004)	Friends
	Joey Tribbiani	Matt LeBlanc	13.5%	M	White	26–36 (1994–2004)	Friends
	Chandler Bing	Matthew Perry	16.0%	M	White	24–34 (1994–2004)	Friends
	Ross Geller	David Schwimmer	15.8%	M	White	27–37 (1994–2004)	Friends
	Leonard Hofstadter	Johnny Galecki	17.4%	M	White	32–44 (2007–2019)	TBBT
	Sheldon Cooper	Jim Parsons	22.7%	M	White	34–46	TBBT
	Penny Hofstadter	Kaley Cuoco	11.5%	F	White	21–32	TBBT
	Howard Wolowitz	Simon Helberg	6.0%	M	White	26–38	TBBT
	Raj Koothrappali	Kunal Nayyar	4.0%	M	Indian (S. Asian)	26–38	TBBT
	Amy Farrah Fowler	Mayim Bialik	31.5%	F	White	31–42	TBBT
	Bernadette Wolowitz	Melissa Rauch	0.9%	F	White	27–36	TBBT
	Jonah Simms	Ben Feldman	16.1%	M	White	35–41	Superstore
	Dina Fox	Lauren Ash	14.2%	F	White	32–38	Superstore
	Garrett McNeill	Colton Dunn	8.9%	M	African American	45–51	Superstore
	Cheyenne Thompson	Nichole Sakura	6.7%	F	Asian	27–33	Superstore
	Glenn Sturgis	Mark McKinney	18.3%	M	White	56–62	Superstore
IEMOCAP & DA	Male-1	–	10.87%	M	–	–	–
	Female-1	–	8.70%	F	–	–	–
	Male-2	–	8.70%	M	–	–	–
	Female-2	–	8.70%	F	–	–	–
	Male-3	–	10.87%	M	–	–	–
	Female-3	–	8.70%	F	–	–	–
	Male-4	–	8.70%	M	–	–	–
	Female-4	–	10.87%	F	–	–	–
	Male-5	–	10.87%	M	–	–	–
	Female-5	–	13.04%	F	–	–	–
MELD & DA	Rachel Green	Jennifer Aniston	13.5%	F	White	24–34 (1994–2004)	Friends
	Monica Geller	Courteney Cox	14.0%	F	White	30–40 (1994–2004)	Friends
	Phoebe Buffay	Lisa Kudrow	13.0%	F	White	31–41 (1994–2004)	Friends
	Joey Tribbiani	Matt LeBlanc	16.0%	M	White	26–36 (1994–2004)	Friends
	Chandler Bing	Matthew Perry	11.0%	M	White	24–34 (1994–2004)	Friends
	Ross Geller	David Schwimmer	14.0%	M	White	27–37 (1994–2004)	Friends
MUStARD	Rachel Green	Jennifer Aniston	4.64%	F	White	25–34	Friends
	Monica Geller	Courteney Cox	4.20%	F	White	30–40	Friends
	Phoebe Buffay	Lisa Kudrow	5.07%	F	White	31–41	Friends
	Joey Tribbiani	Matt LeBlanc	5.22%	M	White	26–36	Friends
	Chandler Bing	Matthew Perry	22.90%	M	White	24–34	Friends
	Ross Geller	David Schwimmer	5.51%	M	White	27–37	Friends
	Sheldon Cooper	Jim Parsons	12.90%	M	White	34–46	TBBT
	Leonard Hofstadter	Johnny Galecki	4.93%	M	White	32–44	TBBT
	Raj Koothrappali	Kunal Nayyar	3.77%	M	Indian (S. Asian)	26–38	TBBT
	Howard Wolowitz	Simon Helberg	6.81%	M	White	26–38	TBBT
	Amy Farrah Fowler	Mayim Bialik	2.46%	F	White	31–42	TBBT
	Bernadette Wolowitz	Melissa Rauch	1.59%	F	White	27–36	TBBT
	Penny Hofstadter	Kaley Cuoco	4.93%	F	White	21–32	TBBT
	Dorothy Zbornak	Bea Arthur	5.65%	F	White	63–70	Golden Girls
	Rose Nylund	Betty White	0.14%	F	White	63–70	Golden Girls

Model	Parameters	Type	Inference Time (h)
Qwen2	7B	LLM	1.9
LLaMA3	8B	LLM	4.8
InternLM2.5	7B	LLM	6.7
VideoLLaMA2	7B	MLLM	7.3
Qwen2-VL	7B	MLLM	1.6
Qwen2-VL	72B	MLLM	6.4
LLaVA-Video	7B	MLLM	1.5
LLaVA-Video	72B	MLLM	5.3
LLaVA-OneVision	7B	MLLM	1.6
LLaVA-OneVision	72B	MLLM	5.4
MiniCPM-V2.6	8B	MLLM	1.9

Table 4: Inference time comparison across foundation models using 4×A800-80G GPUs.

Table 5: Mean and standard deviation (mean ± std) of six evaluation metrics across datasets for two fine-tuning configurations conducted using the Qwen2-VL-7B model.

	Dataset	ACC	F1	Precision	Recall	WF1	WP
SFT	MIntRec	79.33 ± 0.34	75.96 ± 1.43	76.97 ± 1.50	76.04 ± 1.51	79.01 ± 0.33	79.45 ± 0.40
	MIntRec2.0	63.68 ± 0.26	59.18 ± 0.30	64.02 ± 0.68	58.32 ± 0.52	62.72 ± 0.39	64.87 ± 0.27
	MELD	67.57 ± 0.31	51.52 ± 0.46	58.14 ± 1.17	48.99 ± 1.02	66.30 ± 0.45	66.72 ± 0.49
	IEMOCAP	55.06 ± 0.27	51.23 ± 2.66	54.29 ± 3.48	51.03 ± 2.68	54.97 ± 0.32	57.63 ± 0.77
	MELD-DA	61.53 ± 0.31	50.45 ± 0.54	58.21 ± 0.56	49.35 ± 1.63	59.87 ± 0.42	61.55 ± 0.16
	IEMOCAP-DA	68.99 ± 0.45	65.57 ± 2.87	71.68 ± 3.16	63.27 ± 3.74	68.53 ± 0.72	70.32 ± 0.83
	MOSI	88.44 ± 0.10	88.43 ± 0.10	88.45 ± 0.09	88.44 ± 0.09	88.44 ± 0.10	88.46 ± 0.09
	CH-sims	75.41 ± 1.00	60.41 ± 3.93	60.76 ± 5.25	61.69 ± 3.05	73.55 ± 1.92	72.76 ± 3.35
	MUStARD	67.39 ± 0.42	66.45 ± 0.34	67.43 ± 0.39	66.46 ± 0.38	66.94 ± 0.14	67.43 ± 0.40
	UR-FUNNY	74.18 ± 0.53	73.86 ± 0.71	75.64 ± 0.78	74.28 ± 0.50	73.83 ± 0.73	75.71 ± 0.83
	Annomi-client	63.95 ± 1.01	48.39 ± 3.08	64.60 ± 0.87	47.26 ± 2.57	58.65 ± 2.19	65.07 ± 0.19
	Annomi-therapist	75.81 ± 0.11	74.78 ± 0.16	75.07 ± 0.39	75.36 ± 0.32	75.98 ± 0.11	77.14 ± 0.61
IT	MIntRec	82.10 ± 0.45	80.20 ± 0.31	82.76 ± 0.91	80.46 ± 0.33	82.04 ± 0.46	84.07 ± 0.61
	MIntRec2.0	64.35 ± 0.63	58.40 ± 0.80	65.96 ± 1.52	56.90 ± 1.11	63.41 ± 0.68	67.05 ± 1.47
	MELD	66.23 ± 0.67	50.71 ± 0.54	60.19 ± 1.17	48.57 ± 0.25	65.27 ± 0.50	67.40 ± 0.79
	IEMOCAP	49.08 ± 0.37	47.24 ± 0.28	57.17 ± 3.98	51.66 ± 1.01	47.76 ± 0.21	62.61 ± 4.84
	MELD-DA	57.54 ± 2.57	48.64 ± 1.15	58.73 ± 3.17	47.89 ± 1.43	55.31 ± 2.16	60.41 ± 3.24
	IEMOCAP-DA	64.40 ± 1.12	58.72 ± 0.76	68.83 ± 4.73	58.05 ± 1.16	62.58 ± 1.02	67.93 ± 1.79
	MOSI	84.99 ± 2.77	77.11 ± 10.65	78.47 ± 9.31	76.14 ± 11.63	86.43 ± 1.34	88.44 ± 0.67
	CH-sims	75.00 ± 0.78	61.28 ± 2.26	65.56 ± 0.09	61.94 ± 2.39	73.23 ± 0.46	73.69 ± 0.76
	MUStARD	65.46 ± 0.24	65.28 ± 0.13	65.79 ± 0.47	65.46 ± 0.24	65.28 ± 0.13	65.79 ± 0.47
	UR-FUNNY	72.50 ± 0.47	72.35 ± 0.51	72.98 ± 0.25	72.50 ± 0.41	72.35 ± 0.52	72.99 ± 0.24
	Annomi-client	62.53 ± 1.40	48.96 ± 2.18	60.84 ± 1.54	50.16 ± 3.95	59.64 ± 1.06	65.77 ± 3.57
	Annomi-therapist	75.05 ± 0.83	73.88 ± 0.87	74.53 ± 0.68	74.47 ± 0.87	75.22 ± 0.88	76.75 ± 0.68

I Limitations

Potential Noise in Real-world Scenarios. The datasets included in this benchmark provide full videos and text utterances from human speakers. However, in real-world scenarios, there might be noise in the videos or text utterances. Future work can explore the generalization capability of foundation models when encountering noisy data and focus on improving the design of more robust foundation models.

Better Optimization of Foundation Models. Although we have made efforts to optimize the hyperparameters using LoRA techniques to achieve the best performance of foundation models ranging from 7 to 72B parameters for each task and dimension, there is still potential for better optimization with more diverse strategies. The provided results establish baseline standards for our benchmark and offer valuable references for future work.

Foundation Models Updates. Given the rapid advancements in foundation model research, we have done our best to use the state-of-the-art models for building baselines. However, new models may have been released in the past one or two months. We will continuously update and maintain our leaderboard and open-source repositories, incorporating the latest powerful models as soon as possible.

J Broader Societal Impact

J.1 Positive Impacts

Lower Deployment Cost. Parameter-efficient MLLMs (e.g., 7B parameters) achieve performance close to much larger models, reducing computation and energy budgets and making multimodal semantic technology more affordable for real-world applications.

Human-centred Applications. Accurate recognition of intent, emotion, and dialogue acts can power more empathetic virtual assistants, enrich accessibility services (e.g., adaptive captions), and support well-being tools that provide real-time behavioural feedback.

J.2 Negative Impacts

Workforce Shift. Automation of multimodal analysis may shift certain roles—such as manual annotation or routine customer-interaction monitoring—toward machine assistance. Although large-scale displacement is unlikely, affected workers may need reskilling, and organisations should plan for a gradual transition to human-in-the-loop workflows.

Privacy Risk. Fine-grained inference from audio–video streams could be repurposed for intrusive surveillance; as researchers, we explicitly discourage commercial misuse and urge developers to embed consent mechanisms and favour on-device processing to limit abuse.

Table 6: Full experimental results on the MMLA benchmark using zero-shot inference.

	Datasets	MIntRec						MIntRec2.0						MELD					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-0.5B	18.88	15.08	26.81	9.03	10.03	16.06	6.79	4.08	11.33	3.15	5.49	9.48	21.03	12.89	52.20	11.79	19.00	37.51
	Llama-3.2-1B	23.60	18.60	26.71	15.08	16.06	25.55	13.63	9.93	23.98	9.56	12.98	22.73	22.34	18.60	47.81	18.30	23.33	25.34
	Qwen2-1.5B	33.26	24.79	29.97	17.29	21.16	21.48	20.31	16.59	34.48	15.36	18.79	28.24	43.56	45.35	51.16	30.97	33.85	32.05
	Llama-3.2-3B	37.98	33.62	42.04	30.99	35.60	36.84	28.58	26.34	42.81	23.27	26.76	34.87	38.24	38.56	59.46	30.60	34.59	34.35
	Qwen2-7B	56.18	**56.06**	62.22	**50.16**	**53.18**	54.36	36.25	**36.08**	43.31	**32.37**	35.73	37.11	55.67	**57.96**	**63.46**	**40.16**	**40.87**	**41.76**
	Llama-3-8B	49.44	48.57	55.22	41.16	43.98	47.12	28.92	28.50	**43.77**	24.58	27.96	35.30	38.20	41.47	59.17	31.95	34.20	34.59
	Llama-3.1-8B	50.56	50.53	58.96	44.57	46.70	49.88	34.58	34.96	43.26	31.17	35.29	35.28	39.04	40.39	61.66	34.32	37.46	37.80
	Internlm-2.5-7B	**56.63**	51.99	**62.51**	43.92	47.92	**61.03**	**37.28**	34.15	43.54	31.33	**35.90**	**40.50**	**56.36**	56.30	59.67	37.19	37.45	41.18
MLLMs	VideoLLaMA2-7B	44.94	39.71	51.18	33.01	38.22	45.75	27.84	24.32	39.82	20.61	21.22	36.53	55.06	53.59	57.71	32.94	32.92	39.10
	Qwen2-VL-7B	56.40	56.22	62.26	51.00	52.43	56.43	35.27	34.59	44.32	31.24	33.91	38.54	60.31	56.55	59.70	35.91	32.62	47.34
	LLaVA-Video-7B	55.06	53.75	61.36	45.63	49.25	54.33	34.33	32.93	37.79	29.74	33.56	32.84	42.38	41.14	59.14	32.02	36.64	41.97
	LLaVA-OV-7B	52.36	51.09	61.91	43.95	47.07	54.02	32.02	31.47	43.18	27.47	28.91	39.11	34.18	30.66	60.44	26.15	32.63	42.68
	MiniCPM-V-2.6-8B	41.35	43.72	60.00	37.20	35.27	54.39	21.10	24.62	41.75	21.51	19.08	36.26	31.03	36.70	62.26	28.64	27.00	38.88
	Qwen2-VL-72B	61.35	60.89	65.44	55.47	60.14	57.55	39.50	39.91	51.32	36.25	39.84	42.64	**63.18**	**60.06**	63.23	**41.17**	37.72	**50.21**
	LLaVA-Video-72B	62.70	62.42	68.83	55.78	59.27	60.93	40.88	40.28	48.25	36.18	41.36	41.49	40.23	37.37	62.94	33.69	**42.73**	35.99
	LLaVA-OV-72B	62.92	62.46	67.96	55.88	59.17	59.89	**43.78**	43.12	48.60	**38.80**	**42.63**	**42.97**	41.30	39.68	**63.44**	34.59	40.73	41.25
	GPT-4o	**63.37**	**63.68**	**69.69**	**58.37**	**61.46**	**62.13**	42.32	**43.60**	**52.98**	37.49	42.38	42.01	51.17	52.80	61.45	21.32	23.99	21.05

	Datasets	IEMOCAP						MELD-DA						IEMOCAP-DA					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-0.5B	24.91	12.40	41.94	9.53	16.82	30.62	4.75	2.06	8.00	4.31	8.87	7.28	2.65	2.92	19.23	4.49	13.73	9.53
	Llama-3.2-1B	24.17	23.21	41.64	19.17	21.45	35.61	11.41	8.74	13.69	9.86	13.01	12.93	18.21	13.09	16.81	10.74	18.45	12.60
	Qwen2-1.5B	33.79	29.16	49.46	24.17	26.56	41.62	31.43	27.45	37.87	21.87	27.10	25.37	38.32	32.71	35.87	21.93	26.21	22.16
	Llama-3.2-3B	35.02	35.47	42.31	29.34	28.98	35.67	33.08	33.65	45.47	25.61	34.62	28.06	33.92	34.88	44.73	21.81	31.54	23.71
	Qwen2-7B	**39.27**	37.35	48.95	31.11	31.54	41.75	35.84	35.77	46.81	29.72	35.78	33.05	41.08	40.13	50.25	28.92	33.65	32.85
	Llama-3-8B	37.73	**39.01**	43.55	**32.77**	31.41	36.67	42.09	40.19	44.32	**30.59**	36.56	32.96	43.58	**42.09**	52.59	**30.30**	34.17	35.45
	Llama-3.1-8B	35.82	36.93	44.32	31.29	29.65	38.00	39.34	38.96	**47.48**	29.07	**39.52**	29.23	40.34	40.59	**53.26**	27.33	**34.76**	30.47
	Internlm-2.5-7B	38.84	35.17	**50.73**	29.87	**32.29**	**41.97**	**46.30**	**41.39**	46.93	26.61	30.31	**37.00**	**44.59**	40.68	52.98	25.65	29.36	**38.94**
MLLMs	VideoLLaMA2-7B	38.29	33.38	50.01	31.72	35.24	50.93	40.59	38.65	39.14	26.21	33.05	24.45	34.50	32.14	39.03	21.94	33.61	22.82
	Qwen2-VL-7B	32.24	24.16	**58.41**	23.65	27.61	**59.96**	34.78	32.42	47.01	27.83	33.30	33.21	38.75	32.82	38.68	27.74	31.57	32.99
	LLaVA-Video-7B	41.18	36.63	51.15	30.20	35.50	42.69	29.33	24.30	46.13	26.86	38.42	35.81	37.00	30.57	40.89	25.47	34.41	30.04
	LLaVA-OV-7B	39.09	35.20	48.97	33.19	**40.93**	47.03	31.33	26.16	34.38	26.83	32.78	29.06	38.75	30.89	41.71	26.44	29.89	37.88
	MiniCPM-V-2.6-8B	31.94	33.12	52.09	25.99	23.91	46.01	27.93	24.93	42.23	25.01	33.40	29.05	33.17	31.05	38.79	23.13	28.11	24.25
	Qwen2-VL-72B	38.29	34.03	55.34	33.33	34.00	55.48	38.59	34.58	53.49	34.40	40.22	41.47	44.32	36.16	**55.51**	34.17	35.85	**45.28**
	LLaVA-Video-72B	43.53	41.81	48.96	35.32	38.44	41.19	**50.25**	**47.22**	50.85	**38.62**	39.79	45.22	52.87	50.11	54.29	**36.84**	38.05	42.04
	LLaVA-OV-72B	43.46	41.59	49.79	34.67	35.95	43.05	48.95	47.10	55.59	37.64	41.78	**47.71**	52.71	**50.64**	55.37	35.60	38.68	38.01
	GPT-4o	**45.81**	**43.38**	55.47	**40.26**	40.77	58.40	49.00	46.50	**55.90**	36.70	**46.24**	40.77	**53.56**	49.99	54.47	33.80	**40.01**	34.59

	Datasets	MOSI						CH-SIMS v2.0						UR-FUNNY-v2					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-0.5B	61.22	67.27	74.76	44.83	40.79	49.84	27.47	25.21	68.65	25.24	38.77	54.10	45.27	47.11	49.27	31.42	30.21	32.84
	Llama-3.2-1B	70.99	72.57	78.10	48.36	47.23	52.15	65.28	61.12	58.60	34.55	36.63	33.36	55.43	55.08	55.48	55.02	55.31	55.48
	Qwen2-1.5B	80.17	81.50	84.15	54.35	53.52	56.06	65.18	66.94	69.64	43.02	43.75	43.45	62.37	61.23	63.93	40.75	41.42	42.66
	Llama-3.2-3B	81.20	82.52	86.63	54.98	54.03	57.81	66.73	65.28	64.29	39.44	39.79	39.78	66.80	66.79	66.80	**66.78**	**66.78**	**66.80**
	Qwen2-7B	81.92	85.52	**89.49**	57.01	54.60	**59.66**	62.48	66.30	**74.48**	43.55	46.37	46.20	57.95	50.23	**69.23**	33.30	38.26	46.29
	Llama-3-8B	81.63	82.63	86.04	55.07	54.33	57.42	60.35	64.85	70.21	41.62	39.39	44.38	63.98	66.15	68.50	44.08	42.63	45.66
	Llama-3.1-8B	84.26	84.72	86.44	56.46	56.10	57.67	**67.41**	**68.18**	69.20	**57.49**	**57.75**	**57.56**	**67.10**	**67.74**	68.82	45.17	44.78	45.85
	Internlm-2.5-7B	**84.55**	**86.28**	88.26	**57.51**	**56.34**	58.86	62.48	65.34	71.11	42.53	44.26	44.11	61.67	58.30	66.50	58.13	61.27	66.62
MLLMs	VideoLLaMA2-7B	79.74	79.44	81.83	79.47	79.88	81.74	51.90	53.52	69.98	30.67	30.30	39.06	64.29	61.91	68.43	61.77	63.94	68.55
	Qwen2-VL-7B	86.59	86.54	87.03	86.53	86.52	87.07	**73.91**	69.30	65.27	**52.30**	**55.87**	49.20	57.65	50.96	65.42	50.70	57.13	65.57
	LLaVA-Video-7B	84.11	84.22	85.20	56.16	56.13	56.76	57.25	61.09	68.24	34.65	32.76	38.30	59.46	54.68	65.38	54.47	59.00	65.51
	LLaVA-ov-7B	87.17	87.49	87.86	58.32	58.10	58.58	67.58	66.94	68.37	38.01	38.70	38.49	56.04	46.41	68.82	46.09	55.44	69.03
	MiniCPM-V-2.6-8B	79.74	82.72	86.37	55.14	53.12	57.61	60.91	62.76	64.98	35.64	34.70	36.78	57.55	55.67	64.98	36.99	38.11	43.39
	Qwen2-VL-72B	88.05	88.24	88.53	58.83	58.71	59.00	72.45	68.26	66.37	38.65	41.41	37.27	68.01	67.03	70.06	66.95	67.77	70.15
	LLaVA-Video-72B	86.44	86.50	86.59	57.66	57.62	57.73	66.60	67.81	70.24	38.45	38.00	39.58	65.19	62.73	69.99	62.59	64.83	70.13
	LLaVA-OV-72B	**88.48**	**88.60**	**88.78**	59.07	58.97	59.20	65.34	67.89	71.98	38.57	37.36	40.62	63.68	60.13	70.39	59.96	63.27	70.55
	GPT-4o	87.32	87.22	88.36	**87.20**	**87.22**	**88.43**	67.80	**71.30**	**79.16**	47.40	51.04	**49.76**	**73.11**	**73.11**	**73.12**	**73.11**	**73.11**	**73.11**

	Datasets	MUStARD						Anno-MI (client)						Anno-MI (therapist)					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-0.5B	50.00	33.50	25.18	22.33	33.33	16.79	57.04	49.22	49.63	25.14	27.58	31.72	25.70	25.55	27.59	18.76	19.27	19.71
	Llama-3.2-1B	51.45	37.61	62.87	37.61	51.45	**62.87**	28.34	30.10	45.86	26.30	34.79	34.64	22.19	22.89	30.10	16.97	16.70	21.86
	Qwen2-1.5B	52.90	52.90	52.90	52.90	52.90	52.90	36.67	40.45	52.43	25.72	28.48	30.84	40.18	37.14	47.12	37.73	41.17	45.99
	Llama-3.2-3B	57.25	52.16	62.61	52.16	57.25	62.61	43.93	46.77	52.03	28.33	26.70	32.97	43.60	43.35	49.07	34.43	34.98	38.05
	Qwen2-7B	57.25	56.45	58.63	37.63	38.16	39.09	49.87	49.25	**54.96**	30.44	32.99	34.63	54.21	52.63	66.77	42.22	43.93	51.52
	Llama-3-8B	52.90	54.19	**62.94**	36.13	35.27	41.96	47.65	49.98	52.92	30.49	29.64	31.91	35.53	28.09	48.90	25.00	31.07	40.06
	Llama-3.1-8B	54.35	54.72	60.77	36.48	36.23	40.51	32.42	34.66	52.83	24.10	25.99	30.30	**57.98**	**58.53**	**67.16**	**47.06**	**47.51**	**51.74**
	Internlm-2.5-7B	**61.59**	**61.35**	61.90	**61.35**	**61.59**	61.90	**58.99**	**51.52**	51.63	**35.99**	**38.53**	**41.16**	38.86	33.63	36.03	35.64	41.57	36.28
MLLMs	VideoLLaMA2-7B	60.87	60.86	60.88	60.86	60.87	60.88	20.29	16.67	51.14	18.43	35.57	41.69	35.71	35.37	40.14	37.00	38.85	40.91
	Qwen2-VL-7B	56.52	55.93	56.89	55.93	56.52	56.89	31.38	30.97	54.69	25.15	38.30	43.39	49.69	46.81	55.70	40.26	43.27	44.76
	LLaVA-Video-7B	51.45	47.69	52.03	47.69	51.45	52.03	38.92	42.33	49.41	**34.90**	36.00	37.78	47.06	37.98	32.12	34.18	42.12	28.99
	LLaVA-ov-7B	57.97	51.02	68.45	51.02	57.97	68.45	16.58	10.84	46.19	15.24	32.56	35.46	43.90	35.63	59.17	31.81	40.47	43.79
	MiniCPM-V-2.6-8B	60.87	59.71	69.01	39.81	40.58	46.01	27.64	28.56	**55.50**	18.85	27.65	33.13	48.25	41.87	57.49	37.06	42.25	45.65
	Qwen2-VL-72B	**63.04**	**62.33**	64.11	**62.33**	**63.04**	64.11	**53.90**	47.65	52.17	34.81	37.06	**43.62**	49.43	46.40	62.68	37.64	42.58	46.81
	LLaVA-Video-72B	53.62	45.24	59.35	45.24	53.62	59.35	48.58	**49.00**	50.99	30.02	30.62	31.21	61.11	**61.33**	**65.21**	**48.24**	**48.59**	**50.86**
	LLaVA-OV-72B	57.97	53.68	62.66	53.68	57.97	62.66	44.33	45.17	50.79	27.10	29.59	30.21	**61.81**	60.93	62.11	47.56	47.94	49.34
	GPT-4o	55.07	44.60	**70.80**	44.60	55.07	**70.80**	28.90	26.75	54.96	29.63	**42.76**	43.46	55.05	53.62	61.56	43.56	46.24	47.08

Table 7: Full experimental results on the MMLA benchmark using supervised fine-tuning.

	Datasets	MIntRec						MIntRec2.0						MELD					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-0.5B	70.11	70.01	71.37	66.97	66.90	68.97	55.83	55.04	55.74	48.08	47.03	51.22	63.95	62.73	62.15	45.04	43.43	47.30
	Llama-3.2-1B	73.26	73.45	73.93	71.19	71.30	71.50	59.37	59.06	60.63	53.49	53.17	56.33	61.57	60.84	60.83	44.91	43.99	47.90
	Qwen2-1.5B	74.61	74.71	75.88	72.85	72.30	74.96	58.39	57.78	59.06	52.74	51.82	56.25	64.71	63.74	63.57	47.16	45.96	49.89
	Llama-3.2-3B	**77.30**	77.17	77.97	**74.40**	73.04	**77.02**	60.45	59.74	60.27	55.28	56.52	56.07	64.41	63.40	63.11	47.59	46.17	50.20
	Qwen2-7B	74.61	75.10	77.87	69.90	70.26	71.89	61.49	60.91	62.70	53.72	54.29	56.61	65.71	65.11	**65.12**	**50.86**	**49.28**	54.90
	Llama-3-8B	**77.30**	**77.26**	**77.99**	**74.40**	**73.82**	75.95	**62.91**	**62.58**	**63.70**	57.03	57.38	58.91	64.41	63.84	63.64	48.45	47.50	50.29
	Llama-3.1-8B	74.61	74.67	76.73	71.99	73.13	73.04	62.86	62.40	63.28	57.12	**57.52**	58.97	65.02	64.16	63.95	49.40	47.84	52.27
	Internlm-2.5-7B	**77.30**	76.41	76.57	72.32	72.28	74.01	61.83	61.12	62.06	**57.14**	56.87	**60.00**	**66.25**	**65.18**	65.11	48.78	46.66	**55.56**
MLLMs	VideoLLaMA2-7B	79.33	78.82	78.98	76.29	76.45	77.07	65.03	64.64	65.10	58.29	57.08	61.09	63.79	63.17	62.80	43.27	42.21	44.81
	Qwen2-VL-7B	80.45	80.28	80.48	78.55	78.32	79.22	64.19	63.51	64.86	59.68	58.64	63.51	67.09	65.41	65.75	51.45	48.06	57.26
	LLaVA-Video-7B	80.00	79.71	80.85	77.89	77.87	79.87	63.11	62.93	63.66	57.19	56.49	59.28	63.49	62.44	62.44	40.67	38.90	46.05
	LLaVA-OV-7B	80.00	80.02	80.97	78.05	78.03	79.39	63.60	63.04	63.98	56.44	55.36	59.83	62.72	62.90	63.85	43.90	43.92	46.17
	MiniCPM-V-2.6-8B	79.10	78.88	79.78	77.44	79.70	76.29	63.99	63.28	66.15	57.43	58.05	62.21	68.05	66.26	66.92	48.83	46.85	58.22
	Qwen2-VL-72B	81.35	81.00	81.78	75.30	73.97	78.59	**67.39**	**66.98**	**68.34**	**63.65**	**63.12**	**67.04**	**68.81**	**66.68**	**68.01**	**53.43**	49.27	**61.81**
	LLaVA-Video-72B	82.47	82.41	83.34	79.19	79.33	80.69	65.81	65.82	66.23	62.38	61.70	63.69	60.65	61.43	64.27	48.88	50.79	49.22
	LLaVA-OV-72B	**84.04**	**84.00**	**84.94**	**81.70**	**81.62**	**83.04**	66.01	66.15	66.80	61.01	60.36	62.51	61.00	61.79	65.06	49.94	**50.86**	51.84

	Datasets	IEMOCAP						MELD-DA						IEMOCAP-DA					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-0.5B	46.61	45.90	47.25	43.61	43.44	46.54	57.31	56.11	56.26	48.98	47.79	53.58	68.21	68.00	68.84	65.01	61.75	**75.87**
	Llama-3.2-1B	49.08	48.46	49.38	45.99	46.46	47.64	58.36	57.25	57.21	50.09	48.81	53.16	69.11	68.90	69.39	63.55	62.00	66.01
	Qwen2-1.5B	49.88	49.46	51.22	47.61	48.31	49.21	57.21	56.62	56.66	49.56	47.74	**54.17**	67.14	66.42	67.78	58.91	60.83	58.58
	Llama-3.2-3B	50.00	49.24	51.04	47.28	48.28	48.89	58.71	57.36	56.90	49.70	47.80	53.59	71.39	70.95	71.31	67.71	68.66	67.62
	Qwen2-7B	52.10	50.96	52.20	48.31	49.84	49.85	57.71	56.98	57.17	49.25	46.87	53.22	72.08	72.00	72.94	65.16	65.36	65.69
	Llama-3-8B	51.42	50.32	51.84	47.92	49.18	49.48	57.21	56.74	57.29	48.63	47.05	51.35	73.41	73.14	73.74	71.05	69.17	74.16
	Llama-3.1-8B	**52.47**	**52.04**	**53.45**	**50.44**	**51.00**	**52.08**	56.61	55.30	55.59	47.78	45.51	51.87	**74.31**	**74.06**	**74.60**	**72.09**	**70.46**	74.33
	Internlm-2.5-7B	38.41	34.18	51.24	28.64	31.25	42.74	**58.81**	**58.00**	**58.45**	**50.22**	**50.22**	52.28	72.08	71.80	72.35	68.57	68.65	69.44
MLLMs	VideoLLaMA2-7B	56.84	56.93	59.16	47.63	47.75	49.62	61.01	60.11	59.86	47.67	46.23	50.71	**76.43**	**76.28**	**76.58**	68.39	66.62	70.65
	Qwen2-VL-7B	55.12	55.08	56.27	45.99	45.74	47.40	60.91	60.15	61.26	49.70	49.13	57.26	68.10	67.09	68.72	59.85	56.91	66.75
	LLaVA-Video-7B	58.94	59.01	60.17	49.64	48.76	51.36	61.01	60.09	60.98	46.78	43.80	54.29	72.29	71.92	73.77	68.19	64.93	73.72
	LLaVA-OV-7B	58.82	58.88	60.01	49.41	49.18	50.47	59.81	59.16	59.36	48.63	47.51	52.21	72.82	72.36	73.25	68.83	66.86	72.13
	MiniCPM-V-2.6-8B	56.47	56.64	57.94	56.19	57.39	56.35	**63.01**	**61.59**	**62.23**	**54.06**	51.00	**61.48**	74.95	74.90	75.64	71.68	**71.88**	73.70
	Qwen2-VL-72B	56.97	56.50	59.84	54.76	55.65	58.50	58.86	55.39	59.85	47.99	48.98	55.37	73.14	72.38	74.88	72.22	70.81	75.80
	LLaVA-Video-72B	60.17	60.10	62.21	58.87	58.26	61.68	61.11	60.59	60.47	52.71	51.58	54.80	75.53	75.31	76.09	**72.25**	69.93	**76.95**
	LLaVA-OV-72B	**61.53**	**61.36**	**63.89**	**60.07**	**59.52**	**63.45**	61.26	60.73	61.55	53.37	**52.10**	56.75	73.62	73.36	74.50	71.86	69.74	76.10

	Datasets	MOSI						CH-SIMS v2.0						UR-FUNNY-v2					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-0.5B	82.94	82.90	83.44	82.91	83.01	83.40	68.96	68.61	68.29	57.73	57.63	57.90	65.59	65.54	65.63	65.51	65.54	65.64
	Llama-3.2-1B	82.80	82.79	82.95	82.79	82.83	82.92	64.70	65.00	65.70	54.54	54.57	54.88	69.62	69.60	69.72	69.61	69.65	69.69
	Qwen2-1.5B	85.42	85.42	85.43	85.42	85.41	85.44	69.34	69.49	69.77	59.36	59.39	59.44	68.81	68.81	68.85	68.81	68.83	68.84
	Llama-3.2-3B	86.44	86.44	86.46	86.44	86.43	86.46	66.25	66.09	65.94	55.19	55.18	55.22	71.43	71.43	71.46	71.43	71.44	71.44
	Qwen2-7B	87.32	87.32	87.32	87.32	87.31	87.32	**73.11**	**71.71**	71.36	**61.41**	60.62	**64.51**	71.33	71.33	71.35	71.33	71.34	71.34
	Llama-3-8B	86.73	86.74	86.74	86.73	86.74	86.74	70.41	68.67	67.83	56.22	56.31	57.73	72.23	72.20	72.44	72.21	72.29	72.41
	Llama-3.1-8B	**87.61**	**87.61**	**87.65**	**87.61**	**87.63**	**87.64**	68.47	69.17	69.98	58.90	59.21	58.85	**72.64**	**72.60**	**72.83**	**72.61**	**72.69**	**72.80**
	Internlm-2.5-7B	84.99	84.95	85.42	84.96	85.05	85.37	70.21	70.78	**71.70**	60.44	**60.73**	60.61	54.43	43.86	65.01	43.51	53.82	65.18
MLLMs	VideoLLaMA2-7B	87.03	87.02	87.17	87.02	87.06	87.15	74.98	74.82	75.12	64.46	64.24	65.04	74.75	74.75	74.77	74.75	74.76	74.76
	Qwen2-VL-7B	88.34	88.34	88.34	88.34	88.33	88.34	74.49	69.71	66.19	52.69	55.99	50.27	74.04	73.99	74.17	73.97	73.98	74.19
	LLaVA-Video-7B	88.48	88.48	88.49	88.48	88.49	88.48	70.50	71.18	72.00	60.88	61.21	60.83	73.64	73.60	73.73	73.58	73.59	73.75
	LLaVA-OV-7B	**88.92**	**88.92**	**88.92**	**88.92**	**88.92**	**88.92**	72.83	72.82	72.82	63.20	63.17	63.23	74.65	74.64	74.66	74.62	74.62	74.67
	MiniCPM-V-2.6-8B	84.26	84.10	85.87	84.12	84.38	85.78	**76.24**	**76.68**	**77.47**	**67.11**	**67.44**	**67.17**	75.45	75.26	76.50	75.28	75.58	76.43
	Qwen2-VL-72B	87.03	86.98	87.63	86.99	87.10	87.58	75.56	70.95	68.27	53.60	57.52	51.22	76.36	75.78	**79.54**	75.83	76.58	**79.40**
	LLaVA-Video-72B	87.76	87.76	87.76	87.75	87.76	87.76	72.74	74.54	77.28	64.91	66.48	65.27	76.26	76.21	76.56	76.23	76.32	76.52
	LLaVA-OV-72B	**88.92**	**88.92**	**88.92**	**88.92**	**88.92**	**88.92**	72.35	74.17	77.04	65.03	66.91	65.28	**77.16**	**77.10**	77.59	**77.12**	**77.24**	77.54

	Datasets	MUStARD						Anno-MI (client)						Anno-MI (therapist)					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-0.5B	60.87	60.87	60.84	60.84	60.87	60.91	62.98	**60.96**	61.87	**52.69**	**51.37**	56.95	73.33	73.33	74.51	71.34	71.47	72.47
	Llama-3.2-1B	65.94	65.94	65.53	65.53	65.94	66.73	61.29	55.87	58.20	42.10	42.98	54.74	74.47	74.50	75.17	72.77	73.04	73.13
	Qwen2-1.5B	63.04	63.04	63.03	63.03	63.04	63.07	62.89	58.90	61.09	48.33	46.96	58.27	74.30	74.44	75.28	72.91	73.50	72.96
	Llama-3.2-3B	63.04	62.60	63.69	62.6	63.04	63.69	64.75	60.16	**64.96**	50.28	48.38	**65.06**	73.33	73.71	75.93	72.48	73.46	73.04
	Qwen2-7B	63.04	63.04	63.05	63.04	63.04	63.05	64.57	60.88	63.18	50.56	48.87	61.41	75.53	75.80	77.49	74.67	75.39	75.16
	Llama-3-8B	64.49	64.48	64.52	64.48	64.49	64.52	**65.28**	60.70	64.70	49.47	48.09	64.33	**76.05**	**76.41**	**77.98**	**75.48**	**76.56**	**75.50**
	Llama-3.1-8B	60.14	59.67	60.65	59.67	60.14	60.65	39.15	41.84	53.19	37.18	40.87	41.38	75.61	76.03	77.85	74.83	75.68	75.15
	Internlm-2.5-7B	**68.12**	**68.11**	**68.13**	**68.11**	**68.12**	**68.13**	62.71	59.50	61.19	49.83	48.39	56.72	73.42	73.78	76.02	72.50	73.03	73.56
MLLMs	VideoLLaMA2-7B	71.01	71.01	71.03	71.01	71.01	71.03	64.54	61.97	62.87	53.17	51.27	59.64	76.25	76.25	76.34	75.12	74.95	75.37
	Qwen2-VL-7B	67.39	66.28	70.02	66.28	67.39	70.02	65.98	63.03	65.44	54.54	52.38	62.88	75.73	75.78	75.97	74.50	74.72	74.42
	LLaVA-Video-7B	70.29	70.02	71.04	70.02	70.29	71.04	65.69	63.65	64.60	56.35	54.27	61.86	75.68	75.80	76.76	74.92	75.93	74.69
	LLaVA-OV-7B	63.04	62.89	63.27	62.89	63.04	63.27	65.34	64.21	64.63	57.66	56.63	60.19	76.82	77.15	78.80	76.11	77.39	76.11
	MiniCPM-V-2.6-8B	68.84	68.84	68.84	68.84	68.84	68.84	64.66	60.70	64.41	51.87	49.69	63.91	75.96	76.17	78.05	74.99	75.57	75.87
	Qwen2-VL-72B	**74.64**	**74.60**	**74.77**	**74.60**	**74.64**	**74.77**	65.53	59.57	**70.30**	48.68	47.21	**73.38**	77.60	77.90	**79.56**	61.55	62.34	61.67
	LLaVA-Video-72B	71.74	71.62	72.12	71.62	71.74	72.12	67.46	65.90	66.55	59.10	57.12	63.52	76.65	76.70	77.38	75.66	76.60	75.39
	LLaVA-OV-72B	69.57	69.57	69.40	69.40	69.57	69.40	**67.73**	**65.99**	67.40	**59.71**	**58.14**	64.43	**77.79**	**77.95**	78.74	**77.03**	**78.05**	**76.71**

Table 8: Full experimental results on the MMLA benchmark using instruction tuning.

	Datasets	MIntRec						MIntRec2.0					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-7B	77.53	77.78	79.08	74.66	75.42	75.57	**61.49**	61.07	62.56	**57.48**	**56.55**	**61.02**
	Llama-3-8B	**78.20**	**78.14**	**79.94**	**76.62**	**76.46**	**79.43**	61.44	**61.33**	**63.12**	57.05	56.28	60.05
	Internlm-2.5-7B	75.96	76.06	78.46	73.11	71.81	76.71	60.06	59.35	62.27	54.13	52.69	58.83
MLLMs	Qwen2-VL-7B	82.92	82.79	83.78	80.44	81.08	81.53	64.19	63.31	64.39	59.96	59.06	63.31
	MiniCPM-V-2.6-8B	80.67	80.56	82.19	77.75	78.80	79.70	63.31	61.78	65.75	55.88	55.64	**65.71**
	Qwen2-VL-72B	**86.29**	**86.09**	**86.75**	**85.17**	**84.64**	**86.76**	**66.99**	**66.63**	**67.45**	**62.96**	**62.05**	65.24
	LLaVA-Video-72B	80.22	79.94	80.63	77.77	76.97	80.29	64.98	64.72	65.40	60.86	59.97	63.20

	Datasets	MELD						IEMOCAP					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-7B	64.64	64.09	65.15	48.83	48.05	52.81	48.09	47.00	52.49	45.53	47.99	49.46
	Llama-3-8B	**66.02**	**64.79**	**65.19**	**49.86**	48.77	**53.82**	**52.34**	**51.28**	**56.88**	**51.04**	**53.07**	**54.89**
	Internlm-2.5-7B	62.84	62.93	63.71	49.76	**49.90**	51.54	50.80	50.21	54.10	49.56	50.75	52.31
MLLMs	Qwen2-VL-7B	67.55	**66.15**	66.26	51.78	48.60	57.86	49.88	49.10	51.13	46.98	45.82	51.21
	MiniCPM-V-2.6-8B	65.86	65.62	**68.69**	48.80	50.39	56.36	53.58	51.91	**61.66**	52.24	55.18	58.79
	Qwen2-VL-72B	**67.59**	65.08	66.67	50.71	45.59	**61.56**	54.07	53.33	57.92	50.95	52.81	56.19
	LLaVA-Video-72B	65.36	65.24	65.49	**51.93**	**51.18**	53.78	**59.93**	**59.86**	60.16	**58.63**	**58.40**	**59.19**

	Datasets	MELD-DA						IEMOCAP-DA					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-7B	58.71	**57.46**	**57.54**	**48.81**	**48.48**	52.23	65.50	64.91	66.69	60.28	63.67	63.60
	Llama-3-8B	**58.91**	57.44	57.26	47.53	45.89	**52.29**	**69.06**	**68.36**	**69.38**	**62.65**	**64.99**	**64.40**
	Internlm-2.5-7B	57.51	55.80	54.94	45.16	44.79	47.29	67.57	66.91	67.60	59.69	61.16	61.67
MLLMs	Qwen2-VL-7B	60.71	59.00	61.08	**51.98**	**51.43**	**56.90**	62.47	61.04	64.42	57.80	59.33	60.50
	MiniCPM-V-2.6-8B	61.71	59.00	59.71	48.10	45.12	56.61	68.15	66.15	70.01	62.29	61.48	66.88
	Qwen2-VL-72B	**62.36**	**60.80**	**61.72**	51.20	51.31	53.38	72.66	72.02	73.52	66.68	67.44	67.34
	LLaVA-Video-72B	61.46	60.55	60.24	50.34	49.48	52.07	**76.06**	**75.96**	**76.22**	**73.17**	**73.94**	**73.54**

	Datasets	MOSI						CH-SIMS v2.0					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-7B	87.17	87.17	87.20	87.17	87.18	87.18	**70.50**	69.59	68.88	58.09	57.97	58.69
	Llama-3-8B	**87.32**	**87.31**	**87.35**	**87.31**	**87.30**	**87.36**	69.54	70.04	70.62	**60.50**	**60.86**	**60.32**
	Internlm-2.5-7B	86.01	86.00	86.02	86.00	85.99	86.03	69.63	**70.06**	**70.69**	59.45	59.57	59.58
MLLMs	Qwen2-VL-7B	79.45	83.78	**89.74**	55.82	52.89	59.86	75.85	73.05	73.07	59.54	59.82	65.70
	MiniCPM-V-2.6-8B	86.15	86.08	87.06	86.09	86.24	**86.99**	**79.15**	**74.99**	**79.88**	58.95	61.00	**81.29**
	Qwen2-VL-72B	81.63	81.31	84.25	81.34	81.79	84.14	72.65	68.26	67.52	51.56	55.53	50.45
	LLaVA-Video-72B	**86.88**	**86.88**	86.88	**86.88**	**86.88**	86.88	70.69	71.64	73.20	**61.49**	**62.12**	61.77

	Datasets	UR-FUNNY-v2						MUStARD					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-7B	70.22	70.22	70.22	70.21	70.21	70.22	63.04	61.34	65.84	61.34	63.04	65.84
	Llama-3-8B	**72.13**	**72.12**	**72.15**	**72.10**	**72.10**	**72.16**	57.97	55.25	60.53	55.25	57.97	60.53
	Internlm-2.5-7B	69.11	68.99	69.57	69.01	69.21	69.53	**65.94**	**64.59**	**68.82**	**64.59**	**65.94**	**68.82**
MLLMs	Qwen2-VL-7B	73.34	73.28	73.46	73.26	73.28	73.48	65.22	65.21	65.23	65.21	65.22	65.23
	MiniCPM-V-2.6-8B	74.14	74.15	74.15	74.14	74.15	74.14	60.87	60.84	60.91	60.84	60.87	60.91
	Qwen2-VL-72B	**76.96**	**76.35**	**80.48**	**76.40**	**77.19**	**80.33**	70.29	68.49	**76.31**	68.49	70.29	**76.31**
	LLaVA-Video-72B	74.25	74.14	74.54	74.11	74.16	74.57	**71.74**	**71.62**	72.12	**71.62**	**71.74**	72.12

	Datasets	Anno-MI (client)						Anno-MI (therapist)					
	Models	ACC	WF1	WP	F1	R	P	ACC	WF1	WP	F1	R	P
LLMs	Qwen2-7B	64.04	61.00	62.87	52.00	**50.43**	58.88	72.28	71.61	74.19	69.37	69.89	72.78
	Llama-3-8B	**66.16**	**61.34**	**68.87**	**52.60**	50.38	**69.72**	**75.26**	**75.40**	**76.26**	**74.21**	**75.00**	**74.06**
	Internlm-2.5-7B	62.18	54.40	64.20	40.80	41.92	67.05	71.23	71.42	73.42	69.54	69.25	71.73
MLLMs	Qwen2-VL-7B	64.54	60.34	62.21	48.41	47.60	58.61	76.44	76.65	77.58	75.14	75.26	75.68
	MiniCPM-V-2.6-8B	64.48	57.16	71.31	42.60	43.75	**78.95**	75.18	75.27	77.83	74.00	74.96	75.30
	Qwen2-VL-72B	64.89	57.99	**72.12**	46.01	45.35	76.80	76.93	76.82	77.73	75.46	75.88	76.11
	LLaVA-Video-72B	**66.93**	**65.06**	66.28	**58.04**	**56.09**	63.26	**77.61**	**77.64**	**78.05**	**76.65**	**77.29**	**76.39**

Table 9: Primary hyperparameters for SFT and IT on the MIntRec, MIntRec2.0, and MELD datasets.

Datasets	Methods	Models	Learning rate	Warmup ratio	Training Batch size	Epochs	Rank	α
MIntRec	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.00011	0.1	8	60	8	16
		LLaVA-Video-7B	0.0002	0.3	4	10	128	256
		LLaVA-OV-7B	0.0001	0.1	6	15	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.0001	0.1	4	20	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16
MIntRec2.0	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.00012	0.3	16	30	8	16
		LLaVA-Video-7B	0.0002	0.3	6	5	128	256
		LLaVA-OV-7B	0.0001	0.1	6	10	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.0001	0.1	4	20	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16
MELD	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.00011	0.1	16	30	8	16
		LLaVA-Video-7B	0.0001	0.2	4	8	64	128
		LLaVA-OV-7B	0.0001	0.1	6	10	128	256
		VideoChat2-7B	0.00001	0	1	10	16	32
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.00011	0.4	4	10	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16

Table 10: Primary hyperparameters for SFT and IT on the IEMOCAP, MELD-DA, and IEMOCAP-DA datasets.

Datasets	Methods	Models	Learning rate	Warmup ratio	Training Batch size	Epochs	Rank	α
IEMOCAP	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.0001	0.2	2	15	8	16
		LLaVA-Video-7B	0.0002	0.2	6	8	128	256
		LLaVA-OV-7B	0.0001	0.1	6	10	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.0001	0.2	2	7	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16
MELD-DA	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.0003	0.1	8	50	8	16
		LLaVA-Video-7B	0.0001	0.2	4	8	64	128
		LLaVA-OV-7B	0.0001	0.1	6	10	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.00011	0.5	4	10	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16
IEMOCAP-DA	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.0003	0.1	12	50	8	16
		LLaVA-Video-7B	0.0001	0.2	4	8	64	128
		LLaVA-OV-7B	0.0001	0.1	4	10	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.0003	0.1	8	5	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16

Table 11: Primary hyperparameters for SFT and IT on the MOSI, CH-SIMS v2.0, and UR-FUNNY-v2 datasets.

Datasets	Methods	Models	Learning rate	Warmup ratio	Training Batch size	Epochs	Rank	α
MOSI	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.0001	0.1	8	20	8	16
		LLaVA-Video-7B	0.0002	0.3	6	5	128	256
		LLaVA-OV-7B	0.0001	0.1	6	20	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.0001	0.2	2	10	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16
CH-SIMS v2.0	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.00011	0.1	12	30	8	16
		LLaVA-Video-7B	0.0002	0.3	6	5	128	256
		LLaVA-OV-7B	0.0001	0.1	6	15	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.0001	0.2	4	10	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16
UR-FUNNY-v2	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.001	0.2	12	10	8	16
		LLaVA-Video-7B	0.0001	0.2	4	8	64	128
		LLaVA-OV-7B	0.0001	0.1	6	10	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.0003	0.1	4	3	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16

Table 12: Primary hyperparameters for SFT and IT on the MUStARD, Anno-MI (client), and Anno-MI (therapist) datasets.

Datasets	Methods	Models	Learning rate	Warmup ratio	Training Batch size	Epochs	Rank	α
MUStARD	SFT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.001	0.2	2	10	8	16
		LLaVA-Video-7B	0.0002	0.4	2	10	128	256
		LLaVA-OV-7B	0.0001	0.1	6	15	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.0004	0.3	2	10	8	16
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16
		LLaVA-OV-72B	0.0001	0.2	1	20	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16
Anno-MI (client)	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.0003	0.1	8	50	8	16
		LLaVA-Video-7B	0.0001	0.2	4	8	64	128
		LLaVA-OV-7B	0.0001	0.1	6	10	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.00011	0.5	4	10	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16
Anno-MI (therapist)	SFT	Llama3-8B	0.0001	0.1	2	8	8	16
		Qwen2-7B	0.0001	0.1	2	8	8	16
		InternLM2.5-7B	0.0001	0.1	2	8	8	16
		VideoLLaMA2-7B	0.00002	0.5	1	10	128	256
		Qwen2-VL-7B	0.0003	0.1	12	50	8	16
		LLaVA-Video-7B	0.0001	0.2	4	8	64	128
		LLaVA-OV-7B	0.0001	0.1	4	10	128	256
		MiniCPM-V-2.6-8B	0.0001	0.05	1	8	16	32
		Qwen2-VL-72B	0.0003	0.1	8	5	8	16
		LLaVA-Video-72B	0.0001	0.3	1	5	8	16
		LLaVA-OV-72B	0.0001	0.3	1	5	8	16
	IT	Llama3-8B	0.00005	0.1	1	8	16	32
		Qwen2-7B	0.00005	0.1	1	8	16	32
		InternLM2.5-7B	0.00005	0.1	1	8	16	32
		Qwen2-VL-7B	0.0001	0.3	8	5	64	128
		MiniCPM-V-2.6-8B	0.0001	0.05	1	5	16	32
		Qwen2-VL-72B	0.0001	0.3	2	3	64	128
		LLaVA-Video-72B	0.0001	0.3	1	3	8	16