HumanSense: From Multimodal Perception to Empathetic Context-Aware Responses Through Reasoning MLLMs

Zheng Qin[1], Ruobing Zheng[2*] Yabing Wang[1], Tianqi Li[2], Yi Yuan[2], Jingdong Chen[2], Le Wang[1†]

[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
[2]Ant Group, China

Abstract

While Multimodal Large Language Models (MLLMs) show immense promise for achieving truly human-like interactions, progress is hindered by the lack of fine-grained evaluation frameworks for human-centered scenarios, encompassing both the understanding of complex human intentions and the provision of empathetic, context-aware responses. Here we introduce **HumanSense**, a comprehensive benchmark designed to evaluate the human-centered perception and interaction capabilities of MLLMs, with a particular focus on deep understanding of extended multimodal contexts and the formulation of rational feedback. Our evaluation reveals that leading MLLMs still have considerable room for improvement, particularly for advanced interaction-oriented tasks. Supplementing visual input with audio and text information yields substantial improvements, and Omni-modal models show advantages on these tasks. Furthermore, grounded in the observation that appropriate feedback stems from a contextual analysis of the interlocutor's needs and emotions, we posit that reasoning ability serves as the key to unlocking it. We devise a multi-stage, modality-progressive reinforcement learning approach, resulting in **HumanSense-Omni-Reasoning**, which substantially enhances performance on higher-level understanding and interactive tasks. Additionally, we observe that successful reasoning processes appear to exhibit consistent thought patterns. By designing corresponding prompts, we also enhance the performance of non-reasoning models in a training-free manner.

Code — https://github.com/antgroup/HumanSense.git

Datasets — https://huggingface.co/datasets/antgroup/HumanSense_Benchmark

Extended version — https://arxiv.org/abs/2508.10576

Introduction

Science fiction (Quantic Dream 2018) often portrays a future where artificial intelligence serves not merely as a tool for task execution, but as a human companion offering social support and emotional connection. The fundamental evolution from narrow, task-oriented systems to Artificial General Intelligence is predicated on the capacity to comprehend human intentions from speech, expressions, and body language, thereby enabling appropriate responses.

Multimodal Large Language Models (MLLMs) (Xu et al. 2025; Hurst et al. 2024; Anthropic 2024; Team et al. 2023) represent a promising pathway toward realizing this vision. Their ability to holistically process visual, auditory, and textual information enables a comprehensive understanding of users and environments. MLLMs also have the potential to deeply analyze perceived information (Guo et al. 2025) and subsequently plan appropriate feedback, which is not limited to textual responses, but can include suitable emotions, tones, and gesture labels in temporal sequences. Such outputs can be further integrated with video generation (Meng et al. 2025; Qin et al. 2025), speech synthesis (Wang et al. 2023; Suno AI 2023), and talking head (Li et al. 2024c,b; Zhang et al. 2024c; Zheng et al. 2021) methods to provide a highly anthropomorphic interactive experience.

Achieving this goal first requires defining the necessary capabilities, evaluating model performance, and then applying optimization. However, existing benchmarks (Qi et al. 2025; Lin et al. 2024; Hu et al. 2025) lack targeted, fine-grained evaluation for these human-centered scenarios. To address this gap, we first define the primary capabilities needed for MLLMs in such scenarios: 1) multi-modal perception, 2) contextual understanding of implicit information, and 3) appropriate responses in multi-turn interactions. For interactive scenarios, we consider both response content and response strategies.

Based on the above considerations, we propose the HumanSense benchmark. This benchmark comprises 15 progressively challenging tests, totaling 3,882 questions derived from real-world records. In interactive tests, MLLMs are tasked with assuming the role of one party in the interaction and generating responses, which are then compared to real human records. We conduct a comprehensive evaluation of current leading MLLMs, including Vision-Language Models (Zhu et al. 2025; **?**), Omni models (Hurst et al. 2024; Xu et al. 2025), and Audio-Language Models (Chu et al. 2023). The results reveal significant room for improvement in human-centered scenarios, particularly in advanced interaction-oriented tasks. Modality ablation studies show that visual, auditory, and textual information all play crucial roles in high-level tasks, and omni models capable of jointly processing audio, video, and text exhibit a clear advantage.

*Co-first author. Project lead.
†Corresponding author.

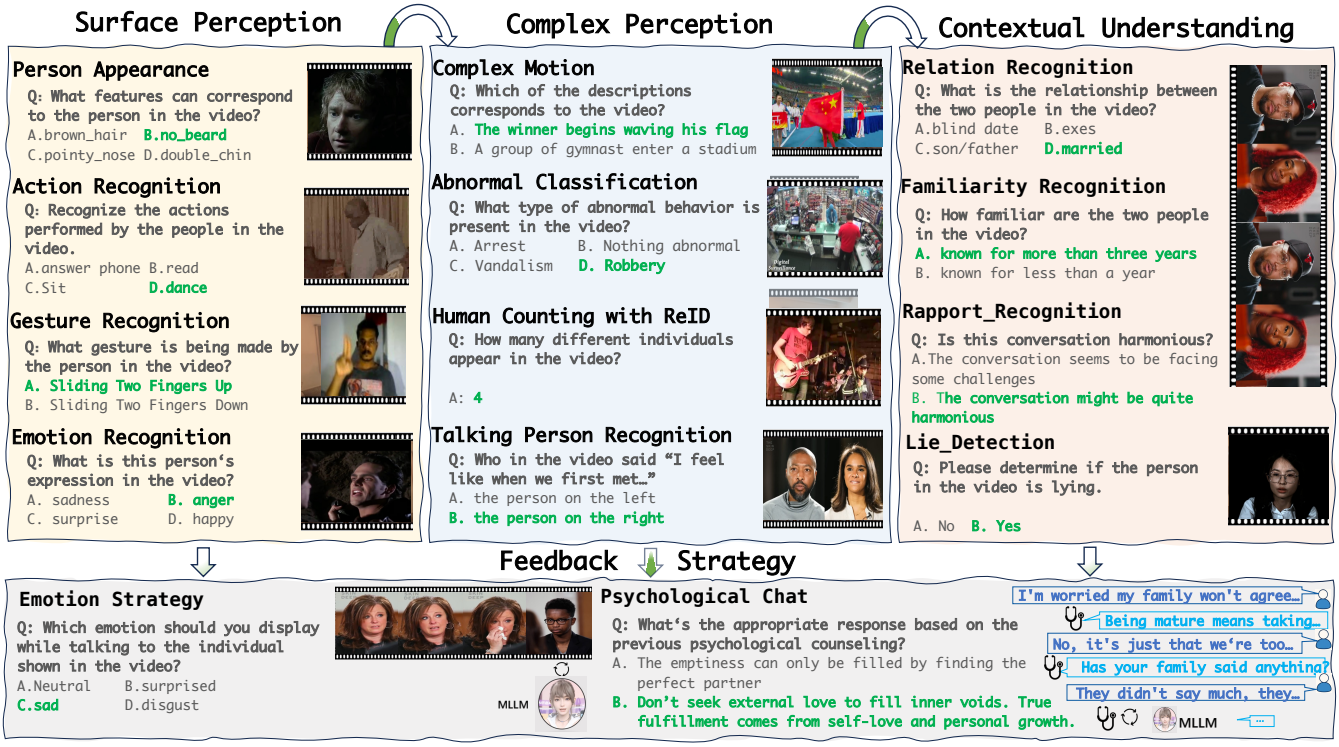

Figure 1: HumanSense benchmark is structured hierarchically to evaluate the end-to-end, human-centric process of perception, understanding, and feedback through a series of demonstrated tasks and questions.

Building on these findings, we propose that omni-modal reasoning can enhance the cognitive and interactive capabilities of MLLMs. This assertion stems from our observation that appropriate feedback in communication relies on thorough consideration of omni-modal context, the interlocutor's needs, emotions, and personal characteristics. Accordingly, we employ a multi-stage, omni-modal reinforcement learning approach to build a reasoning omni model, resulting in substantial improvements in evaluations. Furthermore, we observe that successful reasoning processes exhibit highly consistent patterns. By designing corresponding prompts, we also enhance the performance of non-reasoning models in a training-free manner.

The path to Artificial General Intelligence requires long-term, multifaceted exploration. With this work, we aim to inspire the community to explore the potential of Omni MLLMs for improving human-centered AI interactions and helps shape this emerging direction.

Related Works

Multimodal Large Language Models. LLMs (Bai et al. 2023; Mann et al. 2020; Radford et al. 2018) have been extensively adopted for human behavior and sentiment analysis, facilitating applications including dialogue simulation, behavior prediction, and textual sentiment classification. These capabilities support diverse scenarios ranging from social media monitoring to automated customer support systems. However, textual information alone is often insufficient, as these models lack support for visual cues such as facial expressions and body language, which are essential for comprehensive human behavior analysis. Visual MLLMs (Lin et al. 2023; Chen et al. 2024; Li et al. 2024a; Wang et al. 2024b; Team et al. 2024; Yao et al. 2024) demonstrate strong capabilities in visual understanding, accurately recognizing emotions and behaviors through analysis of facial expressions and body language. However, a critical limitation of these models is their inability to process audio information, resulting in the loss of crucial auditory cues such as dialogue content, vocal intonations, and ambient sounds. This limitation creates significant gaps and introduces biases in their understanding of complex real-world scenarios. Omni models (Xu et al. 2025; Liu et al. 2025; Zhang et al. 2024b; Fu et al. 2025b; Fang et al. 2024; Zhang et al. 2023; Hurst et al. 2024), by contrast, integrate multiple modalities—including vision, language, and audio—to provide comprehensive simulation of complex human interactions. These models can process dialogue content while simultaneously analyzing visual cues, enabling more nuanced and accurate understanding of human communication dynamics.

MLLM Benchmarks. With the advancement of multimodal large models, several evaluation benchmarks (Chen et al. 2024; Fu et al. 2025a; Wang et al. 2024c; Li et al. 2024d; Zhang et al. 2024a) have emerged, most of which focus on assessing video understanding capabilities. Additionally, StreamingBench (Lin et al. 2024) specializes in

streaming video understanding. However, few benchmarks evaluate large models from a human-centered perspective, which is crucial for the practical deployment of such models in real-world scenarios. HumanOmniV2 (Yang et al. 2025) focuses on deciphering intentions, interpreting emotions, and detecting potential deception in an omni-modal manner. While HumanOmniV2 offers valuable insights into human-centered video understanding, it lacks the evaluation of the response planning or interactive capabilities.

HumanSense-Bench

Overview

We aim to systematically assess human-centered capabilities of MLLM through HumanSense framework: 1) Human-centered multi-modal perception, 2) Contextual understanding of implicit information, and 3) Response strategy in interactive scenarios, as Figure 1. The evaluation tasks are organized into a four-tier pyramid structure (L1–L4) according to increasing levels of difficulty, as shown in Figure 2 (Left):

- **Perception** (L1 & L2): The base layer focuses on unimodal, surface perception tasks in L1, and L2 addresses multi-modal, long-duration complex perception tasks. Together, they form the foundational capabilities of intelligence.
- **Understanding** (L3): Built upon perception, this layer evaluates whether the model can uncover implicit information embedded in conversations.
- **Response** (L4): As the pinnacle of capabilities, this layer assesses the model's ability to generate appropriate and rational responses in various interactive scenarios.

This design ensures the systematic nature of the evaluation: the model must first possess a solid perceptual foundation before advancing to deep understanding, ultimately enabling it to make informed feedback decisions at the top level. See Figure 2 for an overview of HumanSense tasks and dataset statistics on task quantity distribution and video length.

Task Definition

In human communication, diverse information is conveyed through different modalities. For example, visual expressions such as facial expressions and gestures can transmit emotional or semantic information; sound can directly express content or indirectly convey emotions. Perceiving these fundamental pieces of information is essential for interaction. We design the following multi-modal perception tasks.

L1: Surface Perception

- **Person Appearance (PA)** evaluates the model's fine-grained perception of facial appearance, as appearance constitutes a fundamental aspect of person identification. We leverage the data annotations from *CelebV-HQ* (Zhu et al. 2022) to design a series of multiple-choice questions. Each question asks whether the person in the video possesses a specific attribute, such as *"Male"*, *"Young"*, *"Chubby"*, *"Rosy cheeks"*, *"Oval face"*, or *"straight hair"*.
- **Action Recognition (AR)** aims to evaluate the model's ability to recognize atomic human actions. We ask MLLMs to identify the current action from a movie clip which comes from the *AVA (Atomic Visual Actions)* (Gu et al. 2018) dataset. The actions include individual behaviors (e.g., *"walk"*, *"sleep"*), interactions with objects (e.g., *"Open a window"*, *"Row a boat"*), and interactions between people (e.g., *"Kiss a person"*, *"talk to a person"*).
- **Gesture Recognition (GR)** aims to evaluate the model's ability to recognize hand gestures, which convey rich semantic information during communication. We construct single-choice questions using *Jester* (Materzynska et al. 2019) dataset, with gestures such as *"Rolling Hand Forward"* and *"Sliding Two Fingers Up"*.
- **Emotion Recognition (ER)** examines the recognition of facial expressions, as they are the primary means of conveying emotions. Based on *CelebV-HQ* (Zhu et al. 2022) dataset, we generate single-choice questions using its built-in labels, asking the model to identify the emotion in videos such as *"Happy"*,*"Sad"*, *"Disgust"*, *"Anger"*, *etc.*

L2: Complex Perception

- **Complex Motion (CM)** examines the description of extended complex action sequences, which relates to the understanding of target behaviors. We utilize the captions from *ActivityNet* dataset (Caba Heilbron et al. 2015) to construct single-choice questions, where the model must identify the correct description of the action performed in the long video clip.
- **Abnormal Classification (AC)** evaluates the detection of abnormal human behaviors. We formulate single-choice questions based on the *UCF-Crime100* dataset (Sultani, Chen, and Shah 2018). The abnormal events include *"Stealing"*, *"Robbery"*, or *"Fighting"*, *etc.*
- **Human Counting with ReID (HC)** evaluates the model's ability to recognize and remember individuals. We use the tracking dataset *TAO* (Dave et al. 2020) to calculate the number of individuals and ask the model about the total count of distinct persons appearing throughout a video. Some challenging questions involve camera transitions and the intermittent appearance and disappearance of humans.
- **Talking Person Recognition (TR)** evaluates the ability to make judgments by integrating visually and auditorily perceived information. Based on *RealTalk* dataset (Geng et al. 2023), we extract video clips and formulate single-choice questions, where the model is required to identify which person is speaking a specific content.

The L1 and L2 tasks assess the perceptual abilities of MLLMs in relation to "seeing" and "hearing". Achieving harmonious communication requires deep thinking about contextual content and providing appropriate responses that correspond to the interlocutor's emotions. Accordingly, L3 examines the model's capacity to "understand" implicit

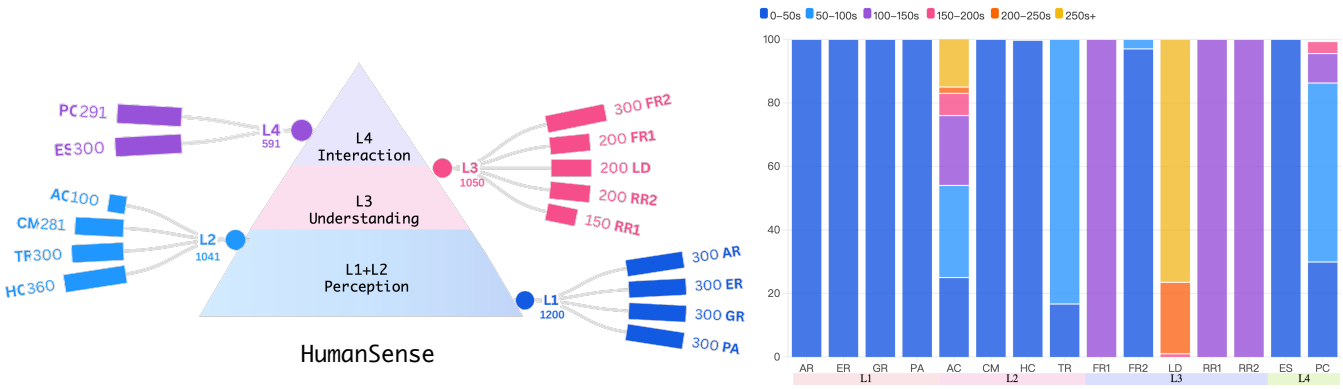

Figure 2: Left: The structure and counts of evaluation tasks in HumanSense. Right: Benchmark statistics on video length.

information during interactions, while L4 evaluates the model's "response" abilities across different scenarios.

L3: Contextual Understanding

- **Familiarity Recognition (FR)** evaluates the model's ability to understand human interpersonal interactions and perceive interpersonal closeness. Based on the duration of acquaintance between conversational participants in *The Skin Deep* channel, we construct single-choice questions to have the model determine the familiarity level between individuals in the videos.
- **Rapport Recognition (RR)** evaluates the model's ability to perceive whether the communication atmosphere is harmonious. We annotate video chat content from *The Skin Deep* channel across multiple dimensions, including interaction frequency, communication atmosphere, and degree of viewpoint conflict, to construct single-choice questions. The evaluation encompasses whether the interactive atmosphere is pleasant and whether conflicts exist in the conversational content.
- **Relation Recognition (RG)** evaluates the model's ability to determine human relationships through multimodal context. Based on the relationships between conversational participants in *The Skin Deep* channel, we construct single-choice questions to have the model predict the relationship type between individuals in the videos, such as *"Married"*,*"Siblings"*,*"Son/Mother"*, *etc.*. The model must integrate information from visual appearance, age differences, and dialogue content to make a judgment.
- **Lie Detection (LD)** evaluates the model's ability to detect human lying. We formulated single-choice questions based on the *SEUMLD* dataset. The model is required to determine whether the speaker is lying in each video segment through analysis of visual and audio cues.
- **Fraud Recognition (FG)** evaluates the detection of specific fraudulent behaviors through approximately one-minute phone recordings, derived from the *Telecom Fraud Texts* dataset (Li, Zhang, and Jiang 2024). The fraud types include "*Loan and credit card agency fraud*", "*Impersonation of public security, judiciary, and government agencies*" and *"Impersonation of leaders or acquaintances"*, *etc.*

L4: Feedback Strategy

- **Emo Strategy (ES)** evaluates the MLLM's ability to provide appropriate facial expression feedback during communication, creating an empathetic interaction experience. Based on the *RealTalk* dataset (Geng et al. 2023), we extract video clips using the speaker as input for questions, and annotated the listener's facial expressions as answers to construct single-choice questions.
- **Psychological Chat (PC)** evaluates the ability of MLLMs to generate appropriate responses in complex, long-context interactions. Here we use a professional online psychological dataset, *Emotional First Aid* dataset (Wang, Wu, and Lang 2020), from which we constructed single-choice questions. The model is expected to select an appropriate response based on the previous multi-turn dialogue.

Data Construction

Question-Answer Generation. HumanSense consists of 3,291 video-based questions and 591 audio-based questions. The related data is sourced from existing open-source datasets and YouTube videos. We construct Question-Answer (QA) pairs using templates, leveraging annotations and built-in labels from the existing datasets. We also use various off-the-shelf modules, including emotion recognition, Large Language Models(LLMs), and Optical Character Recognition (OCR), to analyze source data and extract task-relevant information. Detailed information on the construction for each task is available in the appendix.

Question-Answer Augmentation. To improve evaluation generalizability and avoid evaluation bias, we augment both questions and answers. For each questions, we design diverse candidate templates for random selection. For answers, we balance the distribution of correct options and equalize the lengths of correct and incorrect choices. This QA augmentation also avoid reward hacking (Weng 2024) during the following Reinforcement Learning experiments.

Models	Avg.	Avg.*	AR	ER	GR	PA	AC	CM	HC	TR	FR	FG	LD	RR	RG	ES	PC
			L1				L2				L3					L4	
								HumanSense (tiny) Perf.									
Human Level †	0.875	0.874	0.917	0.933	0.767	0.933	0.967	0.967	0.889	0.900	0.900	0.800	0.533	0.967	0.833	0.867	0.933
GPT-4o†	0.552	-	0.583	0.233	0.700	0.517	0.733	0.767	0.522	0.400	0.833	-	0.300	0.467	0.467	0.667	-
InternVL3-8B†	0.558	-	0.417	0.467	0.533	0.433	0.833	0.767	0.567	0.333	0.733	-	0.667	0.433	0.433	0.633	-
Qwen2.5-Omni-7B†	0.578	0.572	0.467	0.500	0.300	0.383	0.633	0.800	0.467	0.600	0.733	0.700	0.567	0.767	0.600	0.700	0.367
Qwen2-Audio-7B†	-	-	-	-	-	-	-	-	-	-	-	0.333	-	-	-	-	0.333
								HumanSense Perf.									
Proprietary Models (API)																	
GPT-4o	0.557	-	0.548	0.282	**0.620**	0.517	**0.750**	**0.776**	0.536	0.570	0.735	-	0.310	0.480	0.535	0.587	-
Vl-Model																	
LLaVA-Next-Video-7B	0.479	-	0.500	0.480	0.263	0.583	0.440	0.413	0.264	0.487	0.665	-	0.505	0.560	0.500	0.577	-
Qwen2-VL-7B	0.507	-	0.473	0.470	0.307	0.322	0.600	0.591	0.424	0.537	0.665	-	0.515	0.627	0.495	0.570	-
Qwen2.5-VL-7B	0.512	-	0.540	0.497	0.267	0.448	0.550	0.644	0.461	0.523	0.480	-	0.545	0.627	0.485	0.590	-
VideoLLaMA3-7B	0.520	-	0.543	0.463	0.323	0.517	0.530	0.694	0.561	0.493	0.610	-	0.515	0.587	0.400	0.530	-
LLaVA-OneVision-7B	0.521	-	0.545	0.510	0.400	**0.592**	0.620	0.676	0.268	0.503	0.600	-	0.530	0.560	0.430	0.543	-
InternVL3-8B	**0.561**	-	0.393	0.483	0.387	0.547	0.670	0.751	**0.630**	0.567	0.735	-	0.555	0.527	0.490	0.557	-
Audio-Model																	
Qwen2-Audio-7B	-	-	-	-	-	-	-	-	-	-	-	0.437	-	-	-	-	0.399
Omni-Model																	
Ola-7B	0.525	0.539	**0.557**	0.463	0.263	0.573	0.320	0.420	0.371	0.597	**0.785**	0.733	**0.565**	0.653	0.640	0.577	**0.567**
IXC2.5-OmniLive-7B	0.544	0.494	0.508	0.467	0.257	0.338	0.660	0.584	0.544	0.533	0.780	0.415	0.500	0.560	0.470	0.463	0.324
Qwen2.5-Omni-7B	0.554	**0.559**	0.473	**0.513**	0.303	0.350	0.600	0.630	0.438	**0.600**	0.770	**0.740**	0.550	**0.713**	**0.650**	**0.607**	0.399

Table 1: Evaluation on HumanSense. † Indicates results on the HumanSense (tiny) set, for comparison with human-level performance. While GPT-4o is designed as an omni-modal, the absence of audio input support in its current API precluded its evaluation on audio-related tasks (FG, PC). We report two overall average scores: Avg., which excludes the two audio tasks to ensure a fair cross-model comparison, and Avg.*, which is the average across all tasks.

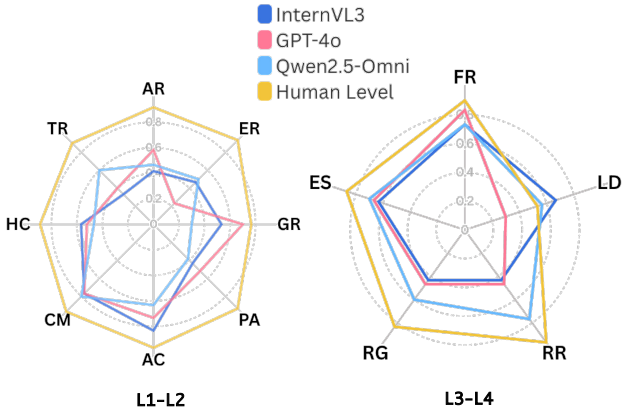

Figure 3: Performance Radar Charts on HumanSense (mini). The results include human-level performance with several state-of-the-art multimodal models.

Quality Control. We conduct manual inspection on a twenty percent sample of all QA pairs, with particular focus on L3 and L4 tasks. For example, in the "Psychological Chat", we ensure that context contains sufficient multi-turn dialogue. We also rigorously validate the correct options to ensure they provide professional and appropriate advice. In the"Rapport Recognition" task, we design scoring dimensions including interaction frequency, communication atmosphere, and content harmony for LLMs evaluation, followed by manual quality inspection of the results.

Evaluation on HumanSense

We conduct a comprehensive evaluation of leading Multimodal Large Language Models (MLLMs) with sizes up to 10B, including: (1) Visual LLMs, which represent the most mainstream branch of MLLMs today; (2) Audio LLMs; and (3) Omni-modal LLMs that are natively designed for integrating vision, audio, and text. For Visual LLMs, we assess models such as Qwen2.5-VL (Bai et al. 2025), Qwen2-VL (Wang et al. 2024a), LLaVA-OneVision (Li et al. 2024a), LLaVANeXT-Video (Liu et al. 2024), VideoLLaMA3 (Zhang et al. 2025) and InternVL3 (Zhu et al. 2025). In the omni-modal models, we test Qwen2.5-Omni (Xu et al. 2025), IXC2.5OmniLive (Zhang et al. 2024b) and Ola (Liu et al. 2025). For audio LLMs, we evaluate Qwen2-Audio (Chu et al. 2023). Additionally, we test the powerful omni-model GPT-4o (Hurst et al. 2024). All evaluations are conducted in a zero-shot setting, employing the default prompts provided for each model. For video processing, we adhere to the respective official configurations for key parameters, including frame extraction methods, frames per second (FPS), and resolution. We release our data and code to the community, aiming to facilitate broader evaluation across a diverse range of models.

To accommodate the characteristics of each model, Visual LLMs are exclusively evaluated on video tasks. Similarly, Audio LLMs are assessed solely on audio tasks. In contrast, omni-models, which are designed for multimodal processing, are required to complete all tasks.

Human Level Performance. To establish a human performance benchmark, we curate a new evaluation set, dubbed HumanSense (tiny), comprising 450 questions randomly sampled across our tasks (30 per task). We first establish a human baseline by having evaluators independently answer each question. Subsequently, we compare the performance of several leading models—including GPT-4o (Hurst et al. 2024), Intern3-VL (Zhu et al. 2025), Qwen2-Audio (Chu et al. 2023), and Qwen2.5-Omni (Xu et al. 2025) with hu-

Models	Avg_{L1} L1	Avg_{L2} L2	FR1	FR2	LD	RR1	RR2	ES	PC
			L3					L4	
Baseline	0.410	0.567	0.770	0.740	0.550	0.713	0.650	0.607	0.399
+ Stg.1	0.555	0.548	0.720	0.557	0.540	0.707	0.620	0.593	0.540
+ Stg.1-2	0.554	0.565	0.775	0.687	0.545	0.693	0.625	0.593	**0.625**
+ Stg.1-3	**0.563**	**0.603**	0.780	0.687	**0.555**	**0.720**	**0.690**	**0.620**	0.619
+ PE	-	-	**0.790**	**0.763**	0.523	**0.720**	0.625	0.600	0.436

Table 2: Evaluation of Multi-Stage Omni-Modal Reinforcement Learning and Training-Free Prompt Enhancement. We report the average scores for L1 and L2, as well as detailed scores for each high-level task. PE represents Prompt Enhancement.

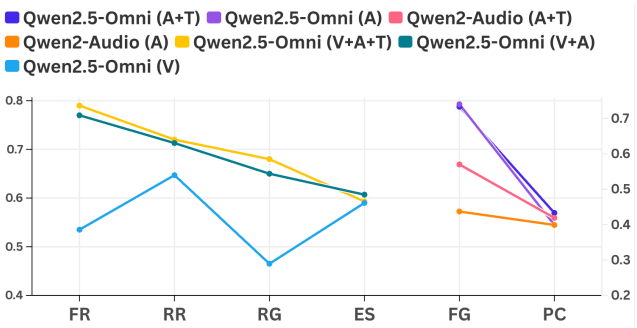

Figure 4: Modality Ablation Across Different Tasks. We visualize the contributions of different modalities across tasks ranging from perception to interaction. The left 4 tasks are video-based questions, and the right 2 tasks are audio-based questions. Note that ASR-transcribed text (T) was used exclusively for this ablation study, not in Table 1.

man standard.

Results

Human Level Performance. Human evaluators achieve an average accuracy of 87.5% on our benchmark, outperforming the best-performing model by a margin of 29.7%. As shown in Figure 3, a substantial performance gap still exists between all models and human-level performance, especially in complex L3-L4 tasks, highlighting a significant need for improvement in the capabilities of current MLLMs on human-centered tasks.

Visual LLMs. As Table 1, Intern3-VL shows the most advantages in terms of average performance, excelling in both L1-L2 perceptual tasks and L3-L4 high-level tasks. Specifically, LLaVA-OneVision achieves the highest metrics in L1 series tasks, reflecting its exceptional visual perception ability on basic tasks. InternVL3 stands out in L2 and some L3 tasks, demonstrating strong long-video memory and contextual understanding, which is consistent with its outstanding performance in video-related benchmarks (Zhu et al. 2025). Notably, in high-level L3-L4 tasks, all models exhibit metrics ranging from 40 to 60, with little variation, suggesting that visual modality input alone is insufficient to provide adequate information for these tasks.

Omni-models and Audio LLMs. The inclusion of audio grants omni-models a significant edge over visual-only LLMs in high-level tasks (L3, L4), such as Rapport Recognition and Lie Detection. This cross-modal synergy is further underscored in the Fraud Recognition task, where the Qwen2.5-Omni (0.74) outperforms its specialized audio counterpart. However, this perceptual advantage diminishes in more complex tasks like Psychological Chat. This highlights a limitation for current leading omni-models: the primary bottleneck is not low-level perception but rather a deficiency in high-level, long-context reasoning, which is essential for truly human-centered understanding.

The charts in Figure 3 visually demonstrate the omni-models, particularly Qwen2.5-Omni, show a marked advantage over the visual-only LLM in higher-level tasks, highlighting the critical role of multimodal, including auditory, perception.

Modality Ablation

To conduct a fine-grained analysis of the importance of different modalities, we performed a series of ablation studies using the Qwen2.5-Omni and Qwen2-Audio models. For clarity, we denote video, audio, and the ASR-transcribed text from video as V, A, and T, respectively. Our experiments focused on six challenging tasks from our L3-L4 difficulty tiers, comprising four video-based and two audio-based tasks. We designed two specific experimental settings to probe modal contributions: (1) augmenting both models with ASR-transcribed text (T) as an additional input, and (2) evaluating the Omni-model in a visual-only setting, removing the audio input.

Figure 4 demonstrates the results for our modality ablation study. For video tasks, audio input (A) serves as a powerful supplement to video (V) for the omni-model, significantly enhancing performance on high-level tasks such as rapport and relationship recognition. Incorporating ASR text (T) provides minimal additional benefit, indicating that raw audio already delivers substantial semantic information for this advanced model. In contrast, for audio-based tasks, the audio-only model exhibits a clear dependence on textual input, highlighting its limited speech comprehension and reliance on explicit semantic support. However, the omni-model gains little from text inputs, demonstrating the advantage of comprehensive multi-modal training. Furthermore, all models perform worse on response-related tasks compared to perceptual tasks, highlighting the importance of improving the response capabilities of MLLMs in interactive scenarios.

From Perception to Responses

The above evaluations confirm that visual, auditory, and textual information all play important roles in high-level tasks. Through data observation, we also find that appropriate feedback in communication relies on thorough consideration of omni-modal context and insight into the interlocutor's needs, emotions, and personal characteristics. Therefore, we believe that reasoning capabilities based on omni-modal inputs are key to enhancing the cognitive and interactive capabilities of MLLMs. In the following sections, we employ a multi-stage, omni-modal reinforcement learning approach to build a reasoning omni model.

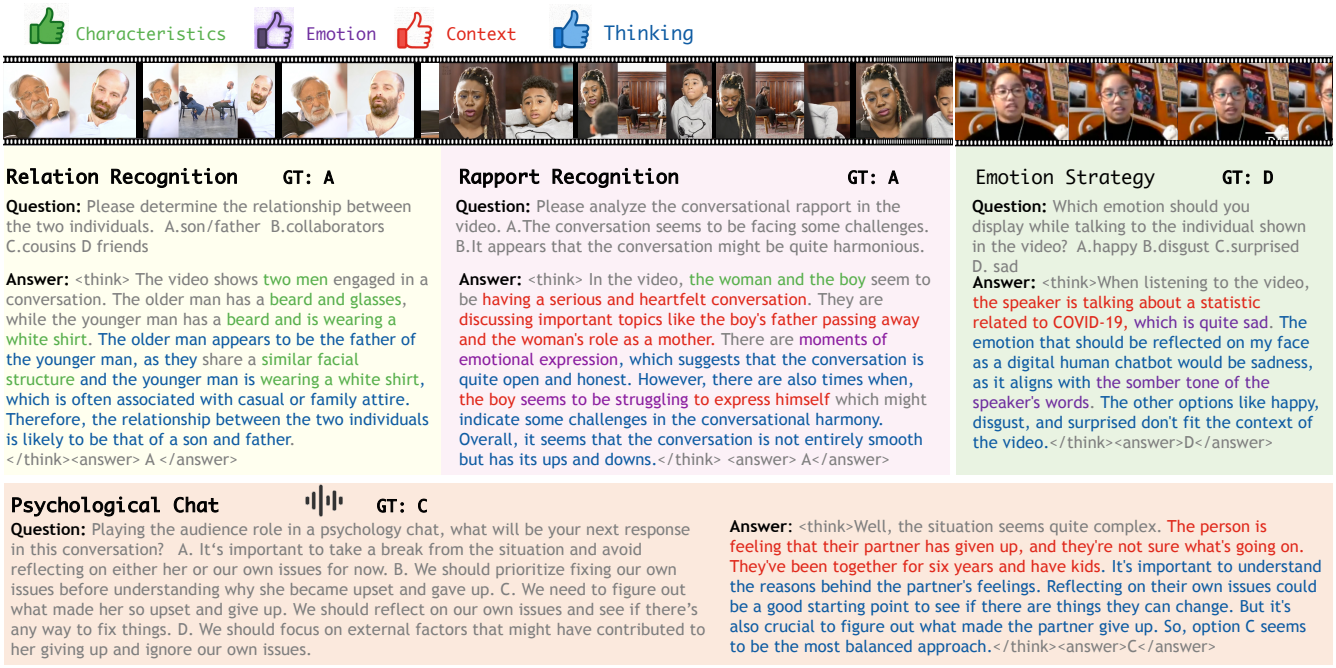

Figure 5: Examples of Successful Reasoning. These cases cover four high-level perception and interaction tasks, including both video-based and audio-based questions. The reasoning processes all demonstrate thinking that integrates characteristics, emotions, and context, and then provides appropriate feedback.

Omni-modal Reinforcement Learning We construct a training set using data independent from the benchmark. We apply Group Relative Policy Optimization (GRPO) (Shao et al. 2024) and design a multi-stage, omni-modal training approach that exposes the Qwen2.5-Omni-7B to all modalities during the Reinforcement Learning with Verifiable Reward (RLVR), enhancing training stability and progressively strengthening perceptual capabilities.

Specifically, in the first stage, we train using pure video frames and textual Question-Answer(QA) pairs to establish reasoning capabilities that integrate visual perception. In the second stage, we train with audio-based QA to develop reasoning that incorporates auditory perception. Finally, we utilize complete video-audio QA to reinforce reasoning that combines both visual and auditory perception. **Further details, such as reward functions, data sizes, and key hyperparameters, are provided in the appendix. The complete training configurations are available in our released code.**

As shown in Table 2, for vision-centric L1 tasks, stage-1 training already yields significant improvements, indicating that reasoning grounded in visual perception can enhance performance on such tasks. For the audio-related tasks PC and FG, stage-2 training leads to notable gains over stage-1, revealing the success of incorporating auditory reasoning. Most tasks achieve optimal performance after completing all 3 training stages. We sample correctly answered examples and find that the model is indeed capable of deep thinking by integrating characteristics, emotions, and contextual information, as shown in Figure 5.

Training-Free Prompt Enhancement We observe that the successful reasoning processes elicited by RL training exhibit a consistent thought pattern: perceiving key characteristics, emotion, and context, followed by thinking and then response. Inspired by this, we believe there exists a training-free approach that can improve existing MLLMs' performance through prompt enhancement. To this end, we design and test the following prompt template, and find that it also boosts performance on high-level tasks, as shown in Table 2.

```
When analyzing the video or audio, focus on
identifying:
   • Key Characteristics: Recognize notable
     features, actions, emotions, or behaviors
     of people.
   • Emotion: Identify the expressed or inferred
     emotional states of individuals
   • Context: Extract relevant dialogue or
     spoken words.
For the following tasks, base your reasoning on
the above elements to draw conclusions.
```

Conclusion

We introduce the HumanSense benchmark to explore MLLMs' capabilities in complex human-centered perception and interaction scenarios. We propose that omni-modal reasoning can enhance MLLMs' performance on such tasks. We aim to inspire the community to recognize the potential of MLLMs in advancing AI interaction experiences.

Acknowledgments

This work was supported in part by the National Key Research and Development Project under Grant 2024YFB4708100, National Natural Science Foundation of China under Grants 62088102, U24A20325 and 12326608, and Key Research and Development Plan of Shaanxi Province under Grant 2024PT-ZCK-80 and Ant Group Research Intern Program.

References

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic. Multimodal AI assistant models.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Tang, Z.; Yuan, L.; et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37: 19472–19495.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *arXiv preprint arXiv:2311.07919*.

Dave, A.; Khurana, T.; Tokmakov, P.; Schmid, C.; and Ramanan, D. 2020. Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision*, 436–454. Springer.

Fang, Q.; Guo, S.; Zhou, Y.; Ma, Z.; Zhang, S.; and Feng, Y. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24108–24118.

Fu, C.; Lin, H.; Wang, X.; Zhang, Y.-F.; Shen, Y.; Liu, X.; Cao, H.; Long, Z.; Gao, H.; Li, K.; et al. 2025b. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Geng, S.; Teotia, R.; Tendulkar, P.; Menon, S.; and Vondrick, C. 2023. Affective faces for goal-driven dyadic communication. *arXiv preprint arXiv:2301.10939*.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6047–6056.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hu, K.; Wu, P.; Pu, F.; Xiao, W.; Zhang, Y.; Yue, X.; Li, B.; and Liu, Z. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, J.; Zhang, C.; and Jiang, L. 2024. Innovative Telecom Fraud Detection: A New Dataset and an Advanced Model with RoBERTa and Dual Loss Functions. *Applied Sciences*, 14(24): 11628.

Li, T.; Zheng, R.; Li, B.; Zhang, Z.; Wang, M.; Chen, J.; and Yang, M. 2024b. LokiTalk: Learning Fine-Grained and Generalizable Correspondences to Enhance NeRF-based Talking Head Synthesis. *arXiv preprint arXiv:2411.19525*.

Li, T.; Zheng, R.; Yang, M.; Chen, J.; and Yang, M. 2024c. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. *arXiv preprint arXiv:2411.19509*.

Li, Y.; Zhang, G.; Ma, Y.; Yuan, R.; Zhu, K.; Guo, H.; Liang, Y.; Liu, J.; Wang, Z.; Yang, J.; et al. 2024d. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Lin, J.; Fang, Z.; Chen, C.; Wan, Z.; Luo, F.; Li, P.; Liu, Y.; and Sun, M. 2024. StreamingBench: Assessing the Gap for MLLMs to Achieve Streaming Video Understanding. *arXiv preprint arXiv:2411.03628*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. Llavanext: Improved reasoning, ocr, and world knowledge.

Liu, Z.; Dong, Y.; Wang, J.; Liu, Z.; Hu, W.; Lu, J.; and Rao, Y. 2025. Ola: Pushing the Frontiers of Omni-Modal Language Model with Progressive Modality Alignment. *arXiv preprint arXiv:2502.04328*.

Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1(3): 3.

Materzynska, J.; Berger, G.; Bax, I.; and Memisevic, R. 2019. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 0–0.

Meng, R.; Wang, Y.; Wu, W.; Zheng, R.; Li, Y.; and Ma, C. 2025. EchoMimicV3: 1.3 B Parameters are All You Need for Unified Multi-Modal and Multi-Task Human Animation. *arXiv preprint arXiv:2507.03905*.

Qi, Y.; Zhao, Y.; Zeng, Y.; Bao, X.; Huang, W.; Chen, L.; Chen, Z.; Zhao, J.; Qi, Z.; and Zhao, F. 2025. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*.

Qin, Z.; Zheng, R.; Wang, Y.; Li, T.; Zhu, Z.; Zhou, S.; Yang, M.; and Wang, L. 2025. Versatile Multimodal Controls for Expressive Talking Human Animation. *arXiv preprint arXiv:2503.08714*.

Quantic Dream. 2018. Detroit: Become Human.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479–6488.

Suno AI. 2023. Bark: Text-Prompted Generative Audio Model. GitHub repository. Generative TTS with emotion, tone, and prosody control.

Team, G.; Anil, R.; Borgeaud, S.; et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2301.02111*. Zero-shot TTS with emotion and speaker adaptation.

Wang, H.; Wu, Z.; and Lang, J. 2020. Pate Psychology: Psychological Counseling Question and Answer Database.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024a. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Gu, X.; Huang, S.; Xu, B.; Dong, Y.; et al. 2024c. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.

Weng, L. 2024. Reward Hacking in Reinforcement Learning. *lilianweng.github.io*.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*.

Yang, Q.; Yao, S.; Chen, W.; Fu, S.; Bai, D.; Zhao, J.; Sun, B.; Yin, B.; Wei, X.; and Zhou, J. 2025. HumanOmniV2: From Understanding to Omni-Modal Reasoning with Context. *arXiv preprint arXiv:2506.21277*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Zhang, H.; Wang, Y.; Tang, Y.; Liu, Y.; Feng, J.; Dai, J.; and Jin, X. 2024a. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*.

Zhang, P.; Dong, X.; Cao, Y.; Zang, Y.; Qian, R.; Wei, X.; Chen, L.; Li, Y.; Niu, J.; Ding, S.; Guo, Q.; Duan, H.; Chen, X.; Lv, H.; Nie, Z.; Zhang, M.; Wang, B.; Zhang, W.; Zhang, X.; Ge, J.; Li, W.; Li, J.; Tu, Z.; He, C.; Zhang, X.; Chen, K.; Qiao, Y.; Lin, D.; and Wang, J. 2024b. InternLM-XComposer2.5-OmniLive: A Comprehensive Multimodal System for Long-term Streaming Video and Audio Interactions.

Zhang, Z.; Zheng, R.; Li, B.; Han, C.; Li, T.; Wang, M.; Guo, T.; Chen, J.; Liu, Z.; and Yang, M. 2024c. Learning dynamic tetrahedra for high-quality talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5209–5219.

Zheng, R.; Zhu, Z.; Song, B.; and Ji, C. 2021. A neural lip-sync framework for synthesizing photorealistic virtual news anchors. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 5286–5293. IEEE.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In *ECCV*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.