Assessment of Emotional Capability of Multimodal Large Language Models in Dynamic Social Interaction Scenarios

Authors: Zhou Zisen, Huang Qi, Tan Zehong, Liu Rui, Cao Ziheng, Mu Fangman, Yachun Fan, Shaozheng Qin, Fan Yachun, Qin Shaozheng

Date: 2025-09-10T14:56:58+00:00

Abstract

Multimodal Large Language Models (MLLMs) can process and integrate multimodal information including images and text, providing a powerful tool for understanding human psychology and cognitive behavior. Integrating classic emotion psychology paradigms, this study compared the performance of two mainstream MLLMs and human subjects in emotion recognition and emotion reasoning under dynamic social interaction scenarios, isolating the distinct roles of visual features of character dialogues (images) and dialogue content (text) in recognizing and reasoning about the emotions of relevant characters. Results indicate that MLLMs' emotion recognition and emotion reasoning performance based on character dialogue images and dialogue content exhibited correlations of moderate or lower magnitude with human subjects' performance. Although a significant gap remains, MLLMs have preliminarily demonstrated emotion recognition and emotion reasoning abilities similar to human subjects in dyadic interactions. Using human subjects' performance as a benchmark, we further compared MLLMs' emotion recognition and emotion reasoning performance under three conditions: relying solely on character dialogue images, relying solely on dialogue content, and relying on their combination. We found that visual features of character dialogues somewhat constrained MLLMs' performance in basic emotion recognition but effectively facilitated complex emotion recognition, while no significant effect was observed on emotion reasoning. By comparing two mainstream MLLMs and their different versions (GPT-4-vision/turbo vs. Claude-3-haiku), we found that innovations in technical frameworks and principles are more crucial for improving MLLMs' emotion recognition and reasoning abilities in social interactions than simply expanding the scale of training data. The findings of this study hold significant scientific value and implications for understanding the psychological mechanisms of emotion recognition and reasoning in social interactions and for inspiring human-like affective computing

and intelligent algorithms.

Full Text

Evaluation of Emotional Capabilities of Multimodal Large Language Models in Dynamic Social Interaction Scenarios

Zisen Zhou[1], Qi Huang[1], Zehong Tan[2], Rui Liu[3], Ziheng Cao[4], Fangman Mu[5], Yachun Fan[2], Shaozheng Qin[1]*

[1]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China
[2]School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
[3]School of Business Administration, Inner Mongolia University of Finance and Economics, Hohhot 010070, China
[4]Alibaba Group, Hangzhou 310020, China
[5]School of Mathematics and Computer Science, Chuxiong Normal University, Chuxiong 675000, China

Received: 2024-06-23
*This work was supported by the National Natural Science Foundation of China (32130045)
†Correspondence: Yachun Fan, E-mail: fanyachun@bnu.edu.cn; Shaozheng Qin, E-mail: szqin@bnu.edu.cn

Abstract

Multimodal Large Language Models (MLLMs), capable of processing and integrating multimodal data such as images and text, provide powerful tools for understanding human psychology and cognitive behavior. Combining classic emotional psychology paradigms, this study compares the performance of two mainstream MLLMs and human participants in emotion recognition and emotion inference within dynamic social interaction scenarios, aiming to disentangle the distinct roles of visual features of character dialogues (images) and dialogue content (text) in recognizing and inferring emotions. Results indicate that MLLMs' emotion recognition and inference performance based on character dialogue images and content shows moderate or weaker correlations with human participants. Despite a noticeable gap, MLLMs have begun to demonstrate emotion recognition and inference capabilities similar to humans in dyadic interactions. Using human performance as a benchmark, we further compare MLLMs' performance across three conditions: using only character dialogue images, only dialogue content, and both combined. Visual features of character dialogues somewhat constrain basic emotion recognition but effectively facilitate complex emotion recognition, while showing no significant impact on emotion inference. By comparing two mainstream MLLMs and their different versions (GPT-4-vision/turbo vs. Claude-3-haiku), we find that innovations in technical

frameworks are more important than simply scaling training data for enhancing MLLMs' emotional capabilities in social interactions.

These findings hold significant scientific value for understanding the psychological mechanisms of emotion recognition and inference in social interactions and for inspiring human-like affective computing and intelligent algorithms.

Keywords: multimodal large language model, social interaction, emotion recognition, emotion inference
Classification: B842

Emotion plays a critical role in helping individuals adapt to natural and social environments, cope with various stressors, and maintain mental health. Therefore, a deep understanding of emotion generation mechanisms is essential for uncovering human psychological functions. Emotion generation involves a series of complex physiological, psychological, and cognitive processes, including the appraisal and judgment of external emotional stimuli (Lazarus, 1991) and the production and regulation of emotional expressions such as facial expressions and body movements (Ekman, 1993). Emotion is not merely an individual internal phenomenon; it also emerges through interactions with others, regulated by social norms and goals, expressed in social contexts, and influences others (Van Kleef & Côté, 2022). Consequently, in social interactions, individuals require at least two emotional capabilities: emotion recognition—the ability to accurately identify and judge others' internal emotional states based on their emotional expressions to efficiently evaluate external emotional stimuli—and emotion inference—the ability to infer and anticipate how one's own emotional expressions affect others, thereby strategically regulating and managing one's expressions. The coordinated operation of these capabilities enables individuals to achieve more effective adaptation and regulation in complex and dynamic social interaction contexts.

Emotion recognition and inference depend on emotional expressions, which arise from the organic coordination of nonverbal and verbal information. Nonverbal cues such as facial expressions and body movements can directly convey emotional states, motivations, and intentions (Ekman, 1993; Mehrabian, 2017), while verbal information provides necessary supplementation and refined representation of complex emotional content and contextual details (Buck, 1985). Although traditional psychological experiments have revealed interactive effects at the perceptual and judgmental levels by manipulating the consistency or conflict between these modalities (McGurk & MacDonald, 1976; De Gelder & Vroomen, 2000), these paradigms, constrained by static stimuli and highly controlled experimental settings, struggle to systematically simulate and predict complex cross-modal information integration processes.

To overcome these limitations, this study employs Multimodal Large Language Models (MLLMs) to investigate the roles of visual features and content in emotional expression. MLLMs can simultaneously process multiple modalities of data, including images and text, providing a powerful computational framework

for studying the interaction between nonverbal and verbal information (Zhang et al., 2024). By integrating information from different modalities, MLLMs can capture complex emotional expressions and social cues, thereby achieving a more comprehensive understanding of human emotional cognition. Moreover, MLLMs offer researchers a flexible tool to systematically manipulate and control different modalities without laboratory constraints.

Furthermore, to address the challenge that traditional psychological experiments often rely on static images and text, making it difficult to capture the complexity of real social interactions (Schilbach et al., 2013), this study constructs a dynamic social interaction scenario dataset that integrates visual features of character dialogues (images) and dialogue content (text). This dataset enables investigation of how different modalities contribute to emotional expression. Given the difficulties in collecting dynamic social interaction data—such as obtaining authentic natural interaction scenes and privacy concerns (Vinciarelli et al., 2009)—this study utilizes film and television materials to construct the evaluation dataset. Such materials offer high ecological validity in terms of emotional expression richness and social interaction authenticity (Busso et al., 2008). Based on theories of cultural specificity in emotional expression (Matsumoto et al., 2008), we selected Chinese-language film materials to ensure the applicability and effectiveness of the evaluation tool within a specific cultural context. Regarding scenario selection, we focused on dyadic dialogue fragments from these materials. As the most basic unit of social interaction, this design preserves core interaction features such as turn-taking and nonverbal synchronization while avoiding information overload that might occur in multi-person scenarios (Clark & Schaefer, 1989). Additionally, we employed emotion soft-labeling, representing emotional states as probability distributions. Compared to single-category emotion labels, soft labels can more precisely capture subtle differences in emotional expression, revealing the multidimensional and complex nature of emotions in interactive contexts and thereby enhancing ecological validity (Fayek et al., 2016; Sridhar et al., 2021).

Building upon this foundation, this study focuses on two emotion-related capabilities closely linked to social interaction: the ability to recognize speakers' emotions as listeners (Li & Deng, 2020) and the ability to infer listeners' emotions as speakers (Zhao et al., 2021; Pollmann & Finkenauer, 2009). Using dyadic dialogue fragments from Chinese-language films, we construct a dynamic social interaction evaluation dataset that integrates visual features and dialogue content. Combining cognitive-behavioral experimental designs, we first compare the emotion recognition and inference performance of multiple MLLMs and human participants to explore whether MLLMs possess human-like capabilities in dynamic social interactions. We then analyze MLLMs' performance across text-only, image-only, and dual-modality conditions to examine how visual features and dialogue content influence emotional expression. Additionally, by comparing MLLMs with different technical principles (GPT-4-vision/turbo vs. Claude-3-haiku) and different training data scales (GPT-4-vision vs. GPT-4-turbo), we reveal the roles of technical architecture and data scale in emotional capability

development. Methodologically, we primarily compare different MLLMs and modalities through zero-shot performance (Wang et al., 2019), then verify the stability of zero-shot comparisons through repeated measurements to achieve accurate assessment of MLLMs' emotion recognition and inference capabilities in dynamic social interaction scenarios.

2.1 Evaluation Dataset

The dynamic social interaction evaluation dataset consists of dyadic dialogue fragments selected from 15 Chinese-language films. Each dialogue lasts no less than 30 seconds, comprises at least 3 turns, and contains no fewer than 6 utterances (see Table 1). The dataset also provides location, character images, and relationship information for each dyadic dialogue. Considering that pretrained MLLMs might recognize characters through their names, all character names in the dataset are replaced with character codes (see Table 2). For each utterance in the dialogue fragments, the dataset includes: (1) the dialogue content text with context (from the beginning of the fragment to the current turn), and (2) three uniformly sampled frames from the video's beginning, middle, and end (character images containing facial expressions and body language information). These materials are used to evaluate emotion recognition and inference capabilities in both MLLMs and human participants.

2.2 Design and Measurement of Emotion Recognition and Inference

This study evaluates two emotional capabilities: emotion recognition and emotion inference. For each dialogue segment in the dyadic scenarios, both MLLMs and human participants can adopt different roles to evaluate emotions: as listeners recognizing the speaker's emotions, or as speakers inferring the listener's emotions (see Figure 1 [Figure 1: see original paper]).

During emotion recognition and inference tasks, participants are provided with 16 selectable emotion labels: four basic emotions (Amusement, Anger, Sadness, Surprise) and twelve complex emotions (Awe, Concentration, Confusion, Contempt, Contentment, Desire, Disappointment, Doubt, Elation, Interest, Pain, Triumph). While these 16 labels do not cover all possible emotions, they possess facial movement patterns that can be effectively recognized by deep neural networks (Cowen et al., 2021), have been preserved across multiple cultures (Cordaro et al., 2018; Cowen et al., 2019; Cordaro et al., 2020), and can explain emotional dimensions such as valence, arousal, and avoidance (Cowen & Keltner, 2020; Cowen et al., 2019).

2.3 Evaluation Methods

We used G*Power 3.1 to estimate the required sample size, which yielded N = 23 (Effect size f = 0.25; $\alpha = 0.05$, 1 - $\beta = 0.80$, single-factor two-level repeated-measures design).

2.3.1 Human Participant Evaluation Method We collected valid responses from 36 participants (21 females, 15 males; mean age = 25.33 years, SD = 3.57). All participants completed the experiment through the Wenjuanxing platform, evaluating all dyadic dialogue fragments from both character perspectives, and received corresponding compensation. As shown in Figure 2 [Figure 2: see original paper], before each dialogue segment, participants were presented with character images, codes, relationships, and location information, and selected one character's perspective for the task. During the experiment, dialogue text and corresponding three sampled frames were presented sequentially. Previous text remained visible, while new text (the target utterance) was highlighted in red. When the target utterance was spoken by the selected character, the instruction read, "What emotion do you think {the other character} felt while listening to your last sentence?" When spoken by the other character, it read, "What emotion do you think {the other character} felt when saying the last sentence?" Participants selected 1-3 emotion labels from the 16 options and ranked them by relevance. After completing the dialogue, participants switched to the other character's perspective to repeat the evaluation. No answers were designated as correct or incorrect; participants were instructed to respond based on their understanding of the scenario and characters.

2.3.2 MLLM Evaluation Method To batch-obtain MLLM emotion recognition and inference results, we used API interface calls, which are fundamentally equivalent to direct user interface queries using prompts. We called the model API once to obtain zero-shot emotion recognition and inference results for analyzing and comparing performance across different MLLMs and modalities. We then called the API 25 times to obtain repeated measurement results from representative MLLMs for testing the stability of zero-shot performance.

The dual-modality zero-shot prompt for MLLMs includes character images, codes, relationships, location, contextual dialogue content, and three sampled frames for the current utterance, asking the model to adopt either a listener role for emotion recognition or a speaker role for emotion inference. The model outputs probabilities for the 16 emotion labels (see Figure S1 for recognition example, Figure S2 for inference example). For image-only modality, dialogue text is removed; for text-only modality, images and sampled frames are removed. For repeated measurements, the output requirement is modified to selecting 1-3 emotion labels ranked by relevance (see Figures S3-S4).

To enhance emotional analysis, we set the system role as "research assistant" at the prompt's beginning and included instructions such as "clear film-related memories," "feel free to comment on characters in the images," and "comments must be based on characters' emotions" at the end. For occasional MLLM response errors, zero-shot results were manually reviewed to ensure emotion labels and probabilities were returned, though not strictly limited to the provided 16 labels. Nevertheless, over 90% of dialogue scenes returned correctly formatted

responses. For repeated measurement results, automated scripts ensured only 1-3 labels from the 16 options were returned.

2.4 Statistical Analysis

2.4.1 Analysis Based on Emotion Label-Dyadic Dialogue Scene Probability Distribution Matrices For MLLMs returning probability distributions of 16 emotion labels, we combined all dialogue scenes to generate emotion label-dyadic dialogue scene probability distribution matrices $(16\$\times149, see Figure S5) for both emotion recognition and inference performance. For human participants and M$ $3 ranked emotion labels, we applied a weighted scoring system: (1) selecting 1 label: +6 weight; (2) selecting 2 labels: +4 for first, +2 for second; (3) selecting 3 labels: +3 for first, +2 for second, +1 for third. These weights were used to compute probability distributions for each dial$ $dyadic dialogue scene probability distribution matrices (16\times\$149$, see Figure 3A [Figure 3: see original paper]A). We then performed Spearman correlation analyses between MLLM and human matrices to obtain correlation coefficients, enabling comparison across different MLLMs and modalities via Fisher's Z tests.

2.4.2 Analysis Based on Mean Emotion Label Probability Distributions Each row of the probability distribution matrix represents the distribution of a specific emotion label across 149 dialogue scenes. The mean probability for each label indicates the likelihood of recognizing or inferring that emotion across the entire dataset. Higher mean probability suggests greater tendency to identify or infer that emotion. We used independent samples t-tests to compare mean probability distributions between MLLMs and humans for the 16 emotion labels, analyzing differences in emotion label preferences to evaluate MLLM performance.

3 Results

Analysis of human participant data (see Section 2.3.1) yielded emotion label-dyadic dialogue scene probability distribution matrices and mean probability distributions for emotion recognition and inference, shown in Figures 3 [Figure 3: see original paper] and 4 [Figure 4: see original paper]. To assess data reliability, we calculated internal consistency (Cronbach's α), revealing high reliability for both emotion recognition ($\alpha = 0.98$) and emotion inference ($\alpha = 0.98$).

Meanwhile, analysis of MLLM data (see Section 2.3.2) collected zero-shot results from GPT-4-vision (image-only, text-only, dual-modality), GPT-4-turbo (dual-modality), and Claude-3-haiku (dual-modality), generating corresponding emotion label-dyadic dialogue scene probability distribution matrices (Figures S5-S14).

3.1 Spearman Correlation Analysis Between MLLM Zero-Shot Performance and Human Participants

Using the method described in Section 2.4.1, we compared the overall similarity between MLLM zero-shot and human probability distribution matrices. Emotion recognition results are shown in Figure 5 [Figure 5: see original paper]; emotion inference results in Figure 6 [Figure 6: see original paper].

Comparing GPT-4-vision across three modalities (image-only/text-only/dual-modality) against humans:

For basic emotion recognition, GPT-4-vision dual-modality showed significantly higher correlation with humans (Spearman's rho: 0.48, 95% CI [0.41, 0.55], Fisher's Z = 0.52, $p < 0.001$) than image-only (Spearman's rho: 0.26, 95% CI [0.19, 0.34], Fisher's Z = 0.27, $p < 0.001$) ($z = 4.32$, $p < 0.001$). Text-only correlation (Spearman's rho: 0.42, 95% CI [0.35, 0.49], Fisher's Z = 0.45, $p < 0.001$) was also significantly higher than image-only ($z = 3.04$, $p = 0.002$). No significant difference existed between dual-modality and text-only ($z = 1.28$, $p = 0.201$).

For complex emotion recognition, dual-modality correlation (Spearman's rho: 0.48, 95% CI [0.45, 0.52], Fisher's Z = 0.53, $p < 0.001$) significantly exceeded image-only (Spearman's rho: 0.35, 95% CI [0.30, 0.39], Fisher's Z = 0.36, $p < 0.001$) ($z = 4.99$, $p < 0.001$) and text-only (Spearman's rho: 0.41, 95% CI [0.37, 0.45], Fisher's Z = 0.44, $p < 0.001$) ($z = 2.64$, $p = 0.008$). Text-only also significantly exceeded image-only ($z = 2.34$, $p = 0.019$).

For basic emotion inference, dual-modality correlation (Spearman's rho: 0.41, 95% CI [0.34, 0.48], Fisher's Z = 0.44, $p < 0.001$) significantly exceeded image-only (Spearman's rho: 0.21, 95% CI [0.13, 0.28], Fisher's Z = 0.21, $p < 0.001$) ($z = 3.98$, $p < 0.001$). Text-only correlation (Spearman's rho: 0.45, 95% CI [0.39, 0.52], Fisher's Z = 0.49, $p < 0.001$) also significantly exceeded image-only ($z = 4.80$, $p < 0.001$). No significant difference existed between dual-modality and text-only ($z = -0.82$, $p = 0.410$).

For complex emotion inference, dual-modality correlation (Spearman's rho: 0.47, 95% CI [0.43, 0.50], Fisher's Z = 0.51, $p < 0.001$) did not significantly differ from image-only (Spearman's rho: 0.42, 95% CI [0.38, 0.46], Fisher's Z = 0.45, $p < 0.001$) ($z = 1.84$, $p = 0.066$). Text-only correlation (Spearman's rho: 0.43, 95% CI [0.39, 0.47], Fisher's Z = 0.46, $p < 0.001$) also did not significantly differ from image-only ($z = 0.34$, $p = 0.737$). No significant difference existed between dual-modality and text-only ($z = 1.50$, $p = 0.133$).

Comparing three MLLMs (GPT-4-vision/GPT-4-turbo/Claude-3-haiku) against humans:

For basic emotion recognition, GPT-4-vision dual-modality correlation significantly exceeded Claude-3-haiku dual-modality (Spearman's rho: 0.29, 95% CI [0.21, 0.37], Fisher's Z = 0.30, $p < 0.001$) ($z = 3.82$, $p < 0.001$). GPT-4-turbo

dual-modality correlation (Spearman's rho: 0.44, 95% CI [0.36, 0.50], Fisher's Z = 0.47, p < 0.001) also significantly exceeded Claude-3-haiku (z = 2.91, p = 0.004). No significant difference existed between GPT-4-vision and GPT-4-turbo dual-modality (z = 0.92, p = 0.360).

For complex emotion recognition, GPT-4-vision dual-modality correlation significantly exceeded both Claude-3-haiku (Spearman's rho: 0.23, 95% CI [0.18, 0.27], Fisher's Z = 0.23, p < 0.001) (z = 8.92, p < 0.001) and GPT-4-turbo dual-modality (Spearman's rho: 0.42, 95% CI [0.38, 0.46], Fisher's Z = 0.45, p < 0.001) (z = 2.30, p = 0.022). GPT-4-turbo also significantly exceeded Claude-3-haiku (z = 6.63, p < 0.001).

For basic emotion inference, GPT-4-vision dual-modality correlation significantly exceeded Claude-3-haiku (Spearman's rho: 0.12, 95% CI [0.05, 0.20], Fisher's Z = 0.12, p = 0.003) (z = 5.48, p < 0.001). GPT-4-turbo dual-modality correlation (Spearman's rho: 0.43, 95% CI [0.36, 0.49], Fisher's Z = 0.46, p < 0.001) also significantly exceeded Claude-3-haiku (z = 5.78, p < 0.001). No significant difference existed between GPT-4-vision and GPT-4-turbo dual-modality (z = -0.30, p = 0.764).

For complex emotion inference, GPT-4-vision dual-modality correlation significantly exceeded Claude-3-haiku (Spearman's rho: 0.29, 95% CI [0.24, 0.33], Fisher's Z = 0.30, p < 0.001) (z = 6.34, p < 0.001). GPT-4-turbo dual-modality correlation (Spearman's rho: 0.43, 95% CI [0.39, 0.47], Fisher's Z = 0.46, p < 0.001) also significantly exceeded Claude-3-haiku (z = 4.84, p < 0.001). No significant difference existed between GPT-4-vision and GPT-4-turbo dual-modality (z = 1.50, p = 0.134).

To test the stability of zero-shot comparisons, we conducted 25 repeated measurements on GPT-4-vision dual-modality and GPT-4-turbo dual-modality, which showed significant differences only in complex emotion recognition. Results (Figures S15, S16) confirmed that repeated measurement findings were consistent with zero-shot results.

3.2 Independent Samples t-Tests Between MLLM Zero-Shot Performance and Human Participants

Using the method described in Section 2.4.2, we further compared differences in mean probability distributions for recognizing and inferring four basic and twelve complex emotions between MLLMs and humans. Emotion recognition results are shown in Figure 7 [Figure 7: see original paper]; emotion inference results in Figure 8 [Figure 8: see original paper]. Emotion labels without significant differences from humans are boxed (all p-values corrected for multiple comparisons).

GPT-4-vision image-only modality showed no significant differences from humans in recognizing 2 basic emotions (Sadness, Surprise) and 4 complex emotions (Desire, Disappointment, Interest, Pain), and no differences in inferring

1 basic emotion (Sadness) and 4 complex emotions (Disappointment, Elation, Pain, Triumph) (see Supplementary Tables S7, S8).

GPT-4-vision text-only modality showed no significant differences from humans in recognizing 4 basic emotions (Amusement, Anger, Sadness, Surprise) and 6 complex emotions (Contempt, Desire, Disappointment, Elation, Interest, Pain), and no differences in inferring 2 basic emotions (Amusement, Anger) and 6 complex emotions (Contempt, Contentment, Disappointment, Elation, Interest, Triumph) (see Supplementary Tables S9, S10).

GPT-4-vision dual-modality showed no significant differences from humans in recognizing 2 basic emotions (Amusement, Surprise) and 7 complex emotions (Concentration, Contempt, Desire, Disappointment, Elation, Interest, Pain), and no differences in inferring 2 basic emotions (Anger, Sadness) and 7 complex emotions (Concentration, Contempt, Disappointment, Elation, Interest, Pain, Triumph) (see Supplementary Tables S3, S4).

GPT-4-turbo dual-modality showed no significant differences from humans in recognizing 4 basic emotions (Amusement, Anger, Sadness, Surprise) and 7 complex emotions (Concentration, Contempt, Desire, Disappointment, Elation, Interest, Pain), and no differences in inferring 2 basic emotions (Anger, Sadness) and 6 complex emotions (Contempt, Contentment, Disappointment, Interest, Pain, Triumph) (see Supplementary Tables S5, S6).

Claude-3-haiku dual-modality showed no significant differences from humans in recognizing 0 basic emotions and 4 complex emotions (Concentration, Contentment, Desire, Doubt), and no differences in inferring 0 basic emotions and 3 complex emotions (Concentration, Disappointment, Doubt) (see Supplementary Tables S1, S2).

Integrating overall similarity and mean probability distribution consistency between MLLMs and humans reveals that all MLLMs show moderate or weaker correlations with human performance, with more than half of the emotion labels showing probability distribution differences from humans. Comparing modalities, GPT-4-vision dual-modality outperformed image-only but not text-only in basic emotion recognition; dual-modality outperformed both image-only and text-only in complex emotion recognition; dual-modality outperformed image-only and equaled text-only in both basic and complex emotion inference.

Comparing different MLLMs, GPT-4-vision dual-modality outperformed Claude-3-haiku dual-modality in both emotion recognition and inference. Comparing different training scales, GPT-4-vision dual-modality underperformed GPT-4-turbo dual-modality in basic emotion recognition but outperformed it in complex emotion recognition, with no differences in emotion inference.

Discussion

This study utilized dyadic dialogue fragments from Chinese-language films to construct a dynamic social interaction evaluation dataset integrating visual fea-

tures and dialogue content. We compared emotion recognition and inference performance between two mainstream MLLMs and human participants, revealing that MLLMs have developed preliminary human-like capabilities while also identifying distinct roles of visual features and dialogue content in emotional expression. Furthermore, we found that technical innovation is more critical than data scaling for enhancing MLLMs' emotional capabilities in dynamic social interactions.

The moderate or weaker correlations between MLLMs and humans suggest that while MLLMs can process multiple modalities simultaneously, their integration mechanisms differ from human cognition. MLLMs' strength lies in their ability to consider emotional vocabulary in text, facial expressions, and body language in images, providing a comprehensive understanding of emotions. Previous research demonstrates that emotional vocabulary, facial expressions, and body language play crucial roles in emotion recognition (Ekman & Friesen, 1978; Mehrabian, 2017). Beyond multimodal information processing, MLLMs combine contextual understanding to grasp the background and motivation behind emotional expressions, enabling more accurate recognition and inference in complex social interactions (Lazarus, 1991; Strack & Deutsch, 2004).

Comparisons across GPT-4-vision modalities reveal that verbal information plays a vital role in emotional expression. Ekman (1992) noted that while facial expressions and body language provide initial emotional signals, these nonverbal cues are often ambiguous without dialogue content. A smiling face might convey happiness, but in a sarcastic context, the expressed emotion could be entirely different. Verbal information provides emotional sources, event descriptions, and linguistic tones that help clarify the true intent behind emotional expressions, offering necessary context for more precise interpretation. Thus, understanding emotional expression depends not only on the emotion itself but also on the nature and context of verbal content.

Moreover, visual features play distinct roles in emotion recognition versus inference. Visual features interfere with basic emotion recognition but facilitate complex emotion recognition, suggesting that complex emotional expression relies more heavily on visual features. Basic emotions, due to their directness and universality, can be conveyed through straightforward vocabulary and sentence structures without extensive cognitive processing, as in expressions like "I am happy" or "I am angry" (Lindquist et al., 2016). When visual features and verbal information convey inconsistent basic emotions, cognitive conflict typically arises, disrupting interpretation. Complex emotions, due to their complexity and diversity, cannot be accurately conveyed through vocabulary alone and require more nuanced language combined with visual features to express emotions accurately (Russell, 2003). Conversely, visual features have minimal impact on emotion inference, indicating that verbal information plays a stronger role in regulating others' emotions. Clear verbal expressions help others accurately understand emotional content and intentions (Ekman & Friesen, 2003) and provide new perspectives or frames for interpreting emotion-eliciting events (Gross,

2015). For instance, when a friend is frustrated over exam failure, a hug might be ambiguously interpreted as friendly or perfunctory, whereas statements like "I understand how you feel" or "This is just a small setback" clearly convey support and empathy while guiding cognitive reframing to reduce negative impact.

Comparing GPT-4-vision with Claude-3-haiku (different technical principles) and GPT-4-turbo (different training scale) reveals that technical innovation enhances emotional capabilities more than data scaling. The Transformer framework's self-attention mechanism effectively captures dependencies between different positions in input sequences, allowing models to attend to distant relevant information rather than just neighboring elements, thereby more effectively processing complex emotional signals and contextual information (Vaswani et al., 2017). While larger datasets expose models to more diverse emotional expressions, enabling richer pattern learning for basic emotion recognition (Goodfellow et al., 2016; Poria et al., 2017), complex emotions involve mixed basic emotions and sophisticated contextual understanding that are difficult to standardize (Plutchik, 1980; Barrett, 2006). Without deep modeling of context and complex emotional relationships, models struggle to understand complex emotions despite increased data (Barrett et al., 2011; Kosti et al., 2017). Emotion inference involves even more complex cognitive and affective processes, including empathy mechanisms that require combining training data, pretrained models, contextual modeling, perspective-taking, reinforcement learning, and other techniques (Su et al., 2016; Ghosal et al., 2019), making data scaling alone insufficient for improvement.

Future psychological and cognitive neuroscience research should increasingly integrate MLLMs to provide more precise and comprehensive perspectives. Psychological research faces challenges from data diversity and complexity, particularly in cognition, emotion, social interaction, and individual differences. MLLMs can effectively integrate and process these diverse information sources, revealing multidimensional features of human psychological processes. By analyzing interactions between modalities, researchers can uncover brain mechanisms underlying multimodal information integration and investigate emotion-cognition interactions, providing new frameworks for cognitive neuroscience. Conversely, psychological research can provide theoretical support for model development in cognitive abilities, emotional intelligence, social interaction, and personalized services. Cognitive neuroscience can guide model design through research on multimodal integration, attention mechanisms, learning, memory, and decision-making. The synergistic development of psychology, cognitive neuroscience, and artificial intelligence will not only advance AI but also provide powerful tools for understanding human behavior and brain mechanisms.

This study has several limitations. First, due to uniform temporal sampling, some frames may include content from both current and subsequent speakers, potentially affecting emotion recognition and inference. Second, although human participants selected a perspective, they had to immediately switch to the other character's perspective after completing each dialogue, which might cause

judgment biases due to prior knowledge of content and plot, with some participants potentially unable to adapt quickly to role-switching. Third, while GPT-4 and Claude-3 differ in technical principles and training data, both rely heavily on massive internet text data with substantial overlap, making it difficult to fully attribute performance differences to technical principles alone. Finally, during zero-shot evaluation, we did not strictly screen for responses where MLLMs failed to adopt specified roles, though manual review ensured response formats were correct.

In summary, this study constructed a dynamic social interaction evaluation dataset using Chinese-language film dyadic dialogues, compared two mainstream MLLMs with human participants, and found preliminary human-like capabilities while revealing distinct roles of visual features and dialogue content in emotional expression. Technical innovation proves more critical than data scaling for enhancing MLLMs' emotional capabilities. Future research should increasingly integrate MLLMs to advance both AI development and understanding of human behavior and brain mechanisms.

References

Barrett, L. F. (2006). Are emotions natural kinds?. *Perspectives on Psychological Science*, *1*(1), 28−58.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, *20*(5), 286−290.

Buck, R. (1985). Prime theory: An integrated view of motivation and emotion. *Psychological Review*, *92*(3), 389.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., … & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*, 335−359.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, *13*(2), 259−294.

Cordaro, D. T., Sun, R., Kamble, S., Hodder, N., Monroy, M., Cowen, A., … & Keltner, D. (2020). The recognition of 18 facial-bodily expressions across nine cultures. *Emotion*, *20*(7), 1292.

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, *18*(1), 75.

Cowen, A. S., & Keltner, D. (2020). What the face displays: Mapping 28 emotions conveyed by naturalistic expression. *American Psychologist*, *75*(3), 349.

Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2019). Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, *74*(6),

698.

Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H., & Prasad, G. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature*, *589*(7841), 251−257.

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, *3*(4), 369−382.

De Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, *14*(3), 289−311.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3−4), 169−200.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System (FACS)* [Database record]. APA PsycTests.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues* (Vol. 10). Ishk.

Fayek, H. M., Lech, M., & Cavedon, L. (2016, July). Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 566−570). IEEE.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry*, *26*(1), 1−26.

Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2017). Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1667−1675).

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, *13*(3), 1195−1215.

Lindquist, K. A., Barrett, L. F., Bliss-Moreau, E., & Russell, J. A. (2006). Language and the perception of emotion. *Emotion*, *6*(1), 125.

Matsumoto, D., Yoo, S. H., & Nakagawa, S. (2008). Culture, emotion regulation, and adjustment. *Journal of Personality and Social Psychology*, *94*(6), 925.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746−748.

Mehrabian, A. (2017). *Nonverbal communication*. Routledge.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3−33). Academic press.

Pollmann, M. M., & Finkenauer, C. (2009). Empathic forecasting: How do we predict other people's feelings?. *Cognition and Emotion*, *23*(5), 978−1001.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98−125.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393−414.

Sridhar, K., Lin, W. C., & Busso, C. (2021, September). Generative approach using soft-labels to learn uncertainty in predicting emotional attributes. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1−8). IEEE.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220−247.

Su, P. H., Gasic, M., Mrksic, N., Rojas-Barahona, L., Ultes, S., Vandyke, D., … & Young, S. (2016). On-line active reward learning for policy optimisation in spoken dialogue systems. *arXiv preprint arXiv:1605.07669*.

Van Kleef, G. A., & Côté, S. (2022). The Social Effects of Emotions. *Annual review of psychology*, *73*, 629–658.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743−1759.

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1−37.

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024). Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Zhao, S., Yao, X., Yang, J., Jia, G., Ding, G., Chua, T. S., … & Keutzer, K. (2021). Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6729−6751.

Supplementary Tables

Table S1 Claude-3-haiku zero-shot and human participant emotion recognition independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.08|[0.03, 0.08]||*Anger*|0.13±0.15|[0.01, 0.09]||*Sadness*|0.11±0.17|[−	

Note: All p-values corrected for multiple comparisons; same below

Table S2 Claude-3-haiku zero-shot and human participant emotion inference independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.07|[0.01, 0.06]||*Anger*|0.09±0.13|[0.05, 0.11]||*Sadness*|0.09±0.13|[−	

Table S3 GPT-4-vision zero-shot and human participant emotion recognition independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.08|[−0.03, 0.01]||*Anger*|0.04±0.08|[−0.07, −0.01]||*Sadness*|0.11±0.	

Table S4 GPT-4-vision zero-shot and human participant emotion inference independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.07|[−0.04, −0.02]||*Anger*|0.01±0.06|[−0.07, −0.02]||*Sadness*|0.09±	

Table S5 GPT-4-turbo zero-shot and human participant emotion recognition independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.08|[−0.04, 0.00]||*Anger*|0.04±0.08|[−0.06, 0.01]||*Sadness*|0.11±0.17	

Table S6 GPT-4-turbo zero-shot and human participant emotion inference independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.07|[−0.04, −0.01]||*Anger*|0.00±0.05|[−0.04, 0.01]||*Sadness*|0.09±0.	

Table S7 GPT-4-vision-image zero-shot and human participant emotion recognition independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.08|[−0.05, −0.01]||*Anger*|0.04±0.07|[−0.10, −0.04]||*Sadness*|0.11±	

Table S8 GPT-4-vision-image zero-shot and human participant emotion inference independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.07|[−0.04, −0.02]||*Anger*|0.03±0.05|[−0.07, −0.02]||*Sadness*|0.09±	

Table S9 GPT-4-vision-text zero-shot and human participant emotion recognition independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.08|[−0.03, 0.01]||*Anger*|0.00±0.08|[−0.06, 0.01]||*Sadness*|0.11±0.17	

Table S10 GPT-4-vision-text zero-shot and human participant emotion inference independent samples t-test results

Emotion Label	Cohen's d	95% CI
Amusement	0.07$±0.07|[−0.02, 0.01]||*Anger*|0.01±0.07|[−0.04, 0.01]||*Sadness*|0.09±0.13	

Supplementary Figures

Figure S1 Example of MLLM zero-shot emotion recognition evaluation prompt
Figure S2 Example of MLLM zero-shot emotion inference evaluation prompt
Figure S3 Example of MLLM repeated measurement emotion recognition evaluation prompt
Figure S4 Example of MLLM repeated measurement emotion inference

evaluation prompt
Figure S5 Claude-3-haiku zero-shot emotion recognition emotion label-dyadic dialogue scene probability distribution matrix
Figure S6 Claude-3-haiku zero-shot emotion inference emotion label-dyadic dialogue scene probability distribution matrix
Figure S7 GPT-4-vision zero-shot emotion recognition emotion label-dyadic dialogue scene probability distribution matrix
Figure S8 GPT-4-vision zero-shot emotion inference emotion label-dyadic dialogue scene probability distribution matrix
Figure S9 GPT-4-turbo zero-shot emotion recognition emotion label-dyadic dialogue scene probability distribution matrix
Figure S10 GPT-4-turbo zero-shot emotion inference emotion label-dyadic dialogue scene probability distribution matrix
Figure S11 GPT-4-vision-image zero-shot emotion recognition emotion label-dyadic dialogue scene probability distribution matrix
Figure S12 GPT-4-vision-image zero-shot emotion inference emotion label-dyadic dialogue scene probability distribution matrix
Figure S13 GPT-4-vision-text zero-shot emotion recognition emotion label-dyadic dialogue scene probability distribution matrix
Figure S14 GPT-4-vision-text zero-shot emotion inference emotion label-dyadic dialogue scene probability distribution matrix
Figure S15 MLLM repeated measurement emotion recognition Spearman correlation analysis and comparison
Figure S16 MLLM repeated measurement emotion inference Spearman correlation analysis and comparison

Note: Figure translations are in progress. See original paper for figures.