Evaluating Multimodal Large Language Model Narrative Interpretation Through the Lens of Appraisal Theory

Jayant Teotia[1], Xiaowei Wang[2], Xulang Zhang[1], Rui Mao[1], Erik Cambria[1]
[1] Nanyang Technological University, Singapore
[2] Universiti Kebangsaan Malaysia, Malaysia
jayant002@e.ntu.edu.sg, {xulang.zhang,rui.mao,cambria}@ntu.edu.sg
p143526@siswa.ukm.edu.my

Abstract

Narrative interpretation is an essential aspect of human cognition, enabling individuals to comprehend complex sequences of events, form emotional connections, and engage in nuanced social reasoning. At the heart of this interpretive ability lies emotional understanding, which cognitive scientists often frame through Appraisal Theory, a model that views emotions as the outcome of subjective evaluations of events in relation to goals, values, and beliefs. In this study, we explore whether multimodal large language models (MLLMs) are able to replicate aspects of this human-like narrative and emotional reasoning. Specifically, we examine how well MLLMs interpret visual narratives, with a focus on their ability to identify and appraise emotional content within scenes. We also investigate whether these models can utilize additional narrative descriptions generated by them to enhance their emotional recognition capabilities, as humans often do. To probe these questions, we conducted a series of experiments using two publicly available datasets, EMOTIC and HECO. Contrary to our expectations, our results reveal a consistent and noteworthy pattern: rather than improving the models' performance, the inclusion of supplementary narrative or contextual information frequently diminishes their ability to accurately recognize emotions. This counterintuitive finding suggests that current MLLMs face significant challenges in integrating multimodal information in a coherent, context-sensitive way. These findings underscore key limitations in the emotional and narrative reasoning capabilities of existing MLLMs and highlight a critical gap between human cognitive processes and current AI approaches.

Keywords: Emotion Recognition, MLLMs, Cognitive Science, Appraisal Theory.

1. Introduction

Artificial intelligence (AI) research has been seeking to imitate human cognitive capacities for decades, with varying degrees of success (Cambria et al., 2026). Recent advancements in large language models (LLMs) and multimodal large language models (MLLMs) have demonstrated remarkable performance across a variety of benchmarks, including natural language understanding (Mao et al., 2024), visual recognition (Zhang et al., 2024), and strategic decision-making (Xu et al., 2024). While these models often achieve or exceed human-level accuracy on narrowly defined tasks, such performance should not be conflated with human-level intelligence. A deeper evaluation, grounded in cognitive science, is essential to assess whether such models possess the underlying mechanisms required for generalizable and context-sensitive reasoning (Mao et al., 2025b).

Human narrative interpretation relies on a complex set of cognitive tools, including the ability to infer implicit information, understand character motivations, and appreciate thematic nuances (Mao et al., 2025a). Similarly, emotion recognition is often explained through Appraisal Theory (Scherer et al., 2001), which posits that emotional responses arise from cognitive evaluations of situations.

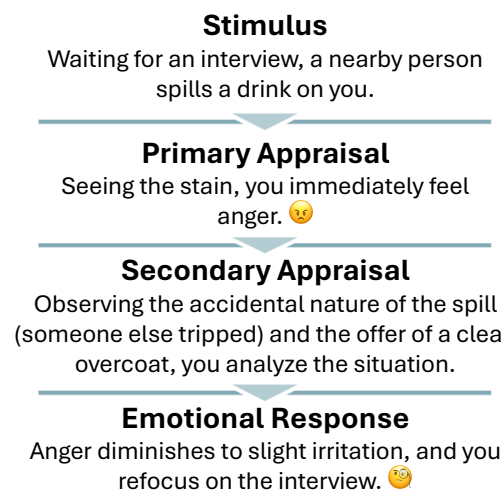

Figure 1: This example demonstrates the sequential cognitive evaluation process central to Appraisal Theory. A triggering event (a spilled drink before an interview) initiates a primary appraisal, eliciting an immediate emotional reaction (anger). Upon further reflection in the secondary appraisal, the individual considers the context (the accident and the offered remedy), leading to a moderated emotional response (from anger to slight irritation).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11417–11426
11–16 May 2026.

Figure 1 illustrates this process where the emotional shifts are not direct reactions to external stimuli but are shaped by the individual's subjective interpretation and reappraisal of the event. In this example, the individual experiences an emotional shift following a sequence of cognitive evaluations. The initial stimulus where a drink being accidentally spilled triggers a primary appraisal and the person reacts with immediate anger upon noticing the stain. However, through secondary appraisal, the individual considers additional contextual information, such as the accidental nature of the spill and the apologetic gesture of offering a clean overcoat. This re-evaluation leads to a moderated emotional response, where the initial anger subsides into mild irritation, and attention is redirected toward the interview. This process underscores how human emotional reactions are dynamic and context-dependent, shaped by layered cognitive interpretations rather than simple stimulus-response associations. It provides a basis for evaluating whether MLLMs can perform emotion appraisal using visual and narrative context.

This work investigates whether multimodal large language models (MLLMs) possess the capacity for narrative interpretation, a core cognitive function essential for understanding human emotional responses, particularly in the context of visual emotion recognition, which has traditionally emphasized facial features. Narrative interpretation requires several high-level cognitive faculties, including the ability to infer intent, integrate temporally and causally related events, and assess the social and emotional relevance of a situation. This is a crucial inquiry, probing the boundaries of MLLM competence and exploring whether these models can progress beyond superficial pattern recognition to deeper emotion appraisals. To investigate this, we design and conduct a series of cognition-inspired experiments that systematically vary the availability of visual and narrative information. The scientific novelty of this work includes: 1) it introduces cognition-inspired, novel experiments to evaluate the emotion-perception capabilities of state-of-the-art MLLMs; 2) it presents empirical findings that reveal the models' limitations in mapping visual and narrative cues to accurate emotional interpretations. We aim at two key research questions:

R1) Do visual contexts enhance MLLM performance in emotion recognition tasks?

R2) How effectively can MLLMs interpret the narratives embedded in visual inputs?

Appraisal Theory suggests that relevant visual context and accurate narrative interpretation, derived from that context, improve human understanding of emotional states. Inspired by this, we designed three controlled experiments: E1) MLLMs are presented with an individual in isolation (no visual context); E2) MLLMs are presented with the complete image (individual within its visual context); E3) MLLMs are presented with the complete image and a narrative interpretation generated by the corresponding MLLM. These designs allow us to isolate the impact of visual context (R1) and textual narrative interpretation (R2) on MLLM emotion recognition. Comparing five MLLMs across these conditions on two public datasets, Emotic (Kosti et al., 2019) and HECO (Yang et al., 2022), yields the following key findings:

F1) Performance in E2 is frequently lower than in E1 across the five MLLMs, suggesting that, as opposed to appraisal theory, visual context does not consistently enhance MLLM emotion recognition. Essentially, MLLMs perform better with a focused object, without background complexity. This suggests they struggle to infer useful narratives from visual context.

F2) Performance in E3 is also frequently lower (on HECO) or comparable (on Emotic) to E1 performance. This indicates that explicitly providing a textual interpretation of the narrative does not consistently improve MLLM emotion recognition. Thus, MLLMs appear unable to generate or leverage useful narratives for emotion detection.

The contribution of this work lies in its investigation of MLLM narrative interpretation and emotion recognition from a cognitive science perspective.

2. Related Work

2.1. Appraisal Theory

The term "Appraisal" was coined by Arnold (1960) to refer to the automatic cognitive process. Numerous appraisal theories seek to elucidate emotional responses from a cognitive science perspective. The Cognitive-Motivational-Relational Theory (CMRT) of Lazarus (1991) integrates cognitive, motivational, and relational processes to explain emotions. Cognitive appraisal involves evaluating a situation's impact on well-being and coping resources. Motivational appraisal ties emotions to personal goals, with threats or support eliciting different reactions. Relational appraisal emphasizes emotions as arising from interactions between individuals and their environment, highlighting situational context over internal experiences alone. Component Process Model from Scherer (1987) and Dimensional Appraisal Theory from Ellsworth and Smith (1988) align with this perspective. Scherer (1987)'s model views emotions as dynamic processes shaped by multiple appraisal dimensions, while Ellsworth and Smith (1988)'s theory identifies specific factors like agency, control, and pleasantness as key determinants of emotional responses. Both emphasize the role of cognitive evaluation in shaping emotions.

Schwarz (1996) argued that affective states, such as moods and emotions, serve as informational heuristics, simplifying decision-making and judgment processes. Rather than analyzing every detail, individuals often rely on their feelings as shortcuts to evaluate situations. Positive emotions tend to result in favorable assessments and risk-taking behaviors, while negative emotions encourage caution, analytical thinking, and critical judgments. Unlike other theories that focus on complex appraisals leading to specific emotions, this approach highlights the practical role of emotions as cognitive shortcuts.

Together, these theories underscore the importance of cognitive appraisal in emotional understanding. Guided by these insights, our study investigates whether and how multimodal large language models (MLLMs) can replicate aspects of human appraisal processes when tasked with recognizing emotions. Our experimental design explicitly targets the model's ability to reason about emotion in ways that mirror human cognitive-emotional processing, across both straightforward and ambiguous emotional scenarios.

2.2. Visual Reasoning

Prior emotion recognition research has highlighted persistent challenges such as ambiguity, subjectivity, and context dependence in affect annotation and prediction (Mao et al., 2023; Fan et al., 2024; Xie and Mao, 2025; Xiao et al., 2025). These challenges suggest that emotion understanding cannot be reduced to simple perceptual classification alone, and instead requires higher-level interpretive reasoning. This motivates recent efforts to evaluate MLLMs not only on basic emotion labels, but also on emotionally ambiguous and narrative-rich scenarios. Cao et al. (2024) studied abstract visual reasoning tasks and found that MLLMs struggle with summarizing visual patterns. Li et al. (2024) discovered a reversed cognitive developmental trajectory in MLLMs, contrasting with human cognitive growth.

Teotia et al. (2024) studied the performance of MLLMs in emotion recognition and engagement-disengagement behaviors in classroom setting. They found that it is easier for the models to classify simple emotions. The performance drops significantly when the model has to infer the classroom dynamics to classify the student behavior and their engagement-disengagement. Hong et al. (2024) evaluated MLLMs in recognizing ambiguous emotions and found that their performance was suboptimal compared to their performance on non-ambiguous emotions. Lu et al. (2024) evaluated GPT-4V and found that while it excels in recognizing facial action units, it struggles with general facial expression recognition.

Dataset	# samples	Task	Metric	# cls
Emotic	7279	M	Macro F1	26
HECO	4000	S	Macro F1	6

Table 1: Datasets and evaluation details. M denotes multi-label prediction. S denotes single-label prediction.

EIBench (Lin et al., 2025) introduces the task of Emotion Interpretation (EI), which shifts focus from emotion labeling to understanding the causal factors, explicit or implicit, behind emotional responses. Unlike traditional emotion recognition datasets, EIBench requires rationale-driven reasoning, highlighting the gap in current model capabilities for context-aware affective analysis. DEEMO (Li et al., 2025) introduces a privacy-preserving framework for emotion understanding using de-identified multimodal inputs, focusing on non-facial body language and identity-free cues. Their DEEMO-LLaMA model achieves strong performance in both emotion recognition and reasoning, highlighting the potential of ethical, identity-sensitive affective computing. However, DEEMO does not explicitly address the causal interpretation of emotional states or the cognitive mechanisms underlying emotion appraisal. While Li et al. (2023) examines visual hallucination in MLLMs, their analysis does not extend to cognitive interpretations of this limitation. To the best of our knowledge, prior work has not yet explored the narrative comprehension capacities of MLLMs in visual tasks.

3. Methodology

To investigate the narrative interpretation of MLLMs, we designed three sequential experiments with increasing levels of dependent information. In the first experiment, we studied the emotion of the cropped out subject present in the image. The next experiment utilizes the whole image and analyzes the emotion of the person in it. In the last experiment, various narrative interpretations are extracted from the MLLMs and subsequently, the whole image, along with these interpretations are utilized to analyze the emotion of the individual present in the image. These experiments are further elaborated below.

E1: Isolated Subject. The datasets used in this study consist of images depicting individuals situated in a variety of environmental and social contexts. Each image is annotated with bounding box coordinates that localize the target individual, along with corresponding emotion label(s) that reflect the emotional state(s) of the concerned individual. For the purpose of this task, the region of the image defined by the bounding box, focusing on the individual of interest, is cropped and provided as input

Experiment 1 No visual Context	**Prompt** Suppose you are an emotional analyst. Given the following image, Classify the emotion of the person in the following image from the given emotion labels - Surprise, Excitement, Happiness, Peace, Disgust, Anger, Fear, Sadness. You should assign only one label that appropriately describes the person in the image. Give the answer in JSON format	
Experiment 2 With Visual Context	**Prompt** Suppose you are an emotional analyst. Given the following image, Classify the emotion of the person in the following image from the given emotion labels - Surprise, Excitement, Happiness, Peace, Disgust, Anger, Fear, Sadness. You should assign only one label that appropriately describes the person in the image. Give the answer in JSON format	
Experiment 3 Visual Context + Narrative interpretation	**Prompts** 1) What is happening in the image? Describe the setting and the context. Identify the environment, objects, or actions that provide context to the situation. 2) What is the person in the green bounding box doing? Analyze their posture, activity, emotion, or interaction with the surrounding. 3) Given the image description and the description of the person in the green bounding box, what is the locus of causality of the emotions showed by the person in the green bounding box? 4) Suppose you are an emotional analyst. Given the image, image description, the description of the person in green bounding box and the locus of causality of the emotions of the person in the green bounding box, classify the person in the green bounding box in the image according to the given emotion labels - Surprise, Excitement, Happiness, Peace, Disgust, Anger, Fear, Sadness. You should assign only one label that appropriately describes the person in the bounding box of the given image. Give the answer in JSON format.	

Figure 2: Overview of the three experimental conditions used to evaluate MLLM emotion recognition. **E1**: The model receives only the cropped image of the individual (bounding box). **E2**: The full image is provided with the individual highlighted using a green bounding box. **E3**: The full image is shown along with narrative interpretations derived from three guiding questions, offering contextual insights.

to a multimodal large language model (MLLM) for emotion classification. In this setting, the MLLM is tasked with inferring the emotional state of the individual based solely on facial expressions and body posture, without access to broader contextual elements of the scene. The prompts used to guide the MLLM in this classification task are illustrated in Figure 2. Notably, the prompts differ slightly between the two datasets to reflect their respective classification formats: the HECO dataset is designed for single-label emotion classification, while the EMOTIC dataset supports multi-label classification.

The EMOTIC dataset presents a more nuanced and complex task, as each individual can be associated with a set of 1 to 5 co-occurring emotion labels. This reflects the fact that emotional states are often ambiguous, overlapping, or context-dependent. In contrast, the HECO dataset offers a more straightforward setup by assigning only one discrete emotion label per individual. These two datasets were intentionally selected to evaluate the breadth of the MLLMs' emotional reasoning capabilities. By comparing performance across both single-label and multi-label tasks, we aim to assess whether the models can handle both clear-cut emotional expressions and more ambiguous, multifaceted emotional situations. This dual evaluation reveals how well current MLLMs approximate human-like emotional appraisal under varying complexity and uncertainty.

E2: Additional Visual Context. In this task, the entire image is provided as input to a multimodal large language model (MLLM) for the purpose of emotion recognition. Unlike in the previous experiment, where only the cropped region containing the individual was used, the full scene, including the individual and their surrounding context, is presented to the model. The individual under observation is visually distinguished within the image by a green bounding box, allowing the model to localize the subject of interest.

To guide the MLLM in focusing on the correct individual, the prompts used in this experiment closely follow those from Experiment 1, with a crucial modification: they explicitly instruct the model to identify the emotion of the individual in the green bounding box. This additional directive is intended to anchor the model's attention to the appropriate subject, ensuring that emotion recognition is based on the target individual rather than other figures or contextual elements in the scene. This experimental setup allows us to examine whether access to full-scene visual information, such as background, social interactions, or environmental cues, enhances or hinders the model's ability to accurately interpret emotional states, particularly when provided with clear referential guidance.

E3: Visual Context and Textual Narratives. In this task, each input sample consists of an image containing a marked individual (indicated by

Emotion	Count	Percentage
Peace	1063	14.60%
Affection	977	13.42%
Esteem	993	13.64%
Anticipation	3424	47.04%
Engagement	5630	77.35%
Confidence	3453	47.44%
Happiness	3490	47.95%
Pleasure	2110	28.99%
Excitement	3490	47.95%
Surprise	445	6.11%
Sympathy	616	8.46%
Doubt/Confusion	1057	14.52%
Disconnection	1130	15.52%
Fatigue	430	5.91%
Embarrassment	137	1.88%
Yearning	471	6.47%
Disapproval	330	4.53%
Aversion	221	3.04%
Annoyance	355	4.88%
Anger	178	2.45%
Sensitivity	213	2.93%
Sadness	388	5.33%
Disquietment	905	12.43%
Fear	233	3.20%
Pain	143	1.96%
Suffering	255	3.50%

Table 2: Emotion distribution of multi-label EMOTIC dataset

Emotion	Count	Percentage
Peace	1383	34.58%
Happiness	1333	33.32%
Disgust	378	9.45%
Excitement	351	8.77%
Sadness	172	4.30%
Anger	167	4.17%
Fear	134	3.35%
Surprise	82	2.05%

Table 3: Emotion distribution of single-label HECO dataset

a bounding box), accompanied by textual narrative interpretations generated in response to three distinct guiding questions (see Prompts 1–3 in Experiment 3, Figure 2). These narrative descriptions are designed to capture contextual nuances surrounding the individual, such as inferred intentions, social dynamics, or situational background, thereby enriching the input with interpretive cues that go beyond raw visual data. The goal of incorporating these textual narratives is to direct the MLLM's visual encoder toward a more contextually grounded understanding of the scene.

By anchoring attention to human-like inferences derived from structured prompts, the task encourages the model to process the image not merely in terms of objects or poses, but in relation to the underlying narrative and emotional implications. Given this multimodal input consisting of the full image, the bounding box highlighting the subject, and the associated narrative interpretations, the MLLM is tasked with predicting emotion labels for the identified individual. For the multi-label EMOTIC dataset, we specifically asked the model to select the top five most relevant emotions expressed by the person in the image.

This numeric constraint was introduced after observing that, in the absence of such a specification, MLLMs often produced highly inconsistent outputs. These included over-generating labels (e.g., listing all 26 possible emotion categories), assigning arbitrary numerical scores to categories, or producing free-form responses that deviated from the expected label format. By explicitly requesting the top five emotion labels, we standardized the output format, reduced ambiguity, and improved the comparability of results across model outputs.

4. MLLMs and Datasets

MMLMs. In this study, we evaluate the performance of several state-of-the-art multimodal large language models: DeepSeek Janus Pro 7B (Chen et al., 2025), LLaVA-v1.6-Qwen-7B, LLaVA-v1.6-Mistral-7B, LLaVA-v1.6-Vicuna-7B (Liu et al., 2023), and GPT-4o Mini (Achiam et al., 2023). These models were selected based on a balance between their performance capabilities and our available computational resources. All models, with the exception of GPT-4o Mini, are built upon 7-billion-parameter (7B) architectures, providing strong multimodal reasoning while staying feasible within a limited compute budget.

Datasets. For evaluating MLLM's on our tasks, we have used two datasets (Table 1): Emotic (Kosti et al., 2019), a multi-label dataset with 26 emotion categories, and HECO (Yang et al., 2022), a single-label dataset with 6 emotion categories. The emotional distribution of both these datasets are shown in Table 2 and Table 3. These datasets were selected to cover both multi-label and single-label emotion classification settings, allowing for a comprehensive assessment of MLLMs' emotional reasoning capabilities. To ensure consistency and comparability across models and tasks, we use macro-F1 score as our primary evaluation metric. This metric accounts for class imbalance by computing the unweighted mean of F1 scores across all emotion categories, thereby providing a balanced measure of model performance across both frequent and rare classes.

Emotion Labels	DeepSeek			GPT			Llava-mis			Llava-qwen			Llava-vic		
	E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3
Peace	0.39	0.00	0.89	0.00	0.00	0.00	0.00	0.00	0.76	0.00	0.21	1.62	0.00	0.00	0.19
Affection	15.08	5.93	20.05	41.90	40.21	40.00	6.71	15.42	7.18	9.63	5.79	5.09	22.02	22.04	24.51
Esteem	0.00	0.00	2.76	12.43	8.24	20.18	6.02	6.25	0.97	2.88	3.24	2.42	20.13	19.55	15.19
Anticipation	28.71	12.92	40.09	60.86	62.10	57.87	21.95	41.78	43.02	1.10	3.05	13.83	46.98	46.30	47.31
Engagement	87.70	87.17	87.79	81.14	85.52	87.50	60.47	62.26	71.56	42.13	53.03	36.02	82.14	78.83	80.11
Confidence	65.28	64.73	69.36	65.29	65.21	64.13	47.67	35.61	34.25	64.08	61.83	56.54	63.49	61.48	59.44
Happiness	69.92	66.34	67.11	68.29	66.24	66.72	36.06	41.37	46.28	62.11	59.29	57.31	35.93	46.42	56.42
Pleasure	46.53	47.80	47.38	41.81	35.25	46.63	3.85	13.26	23.34	44.19	41.15	41.18	11.12	22.86	36.27
Excitement	65.15	65.36	67.88	68.99	66.06	64.09	28.21	42.12	47.41	15.41	26.63	42.56	52.88	53.87	57.98
Surprise	8.10	6.54	8.45	22.73	14.57	14.71	2.37	9.85	8.71	10.97	11.55	10.31	11.16	10.73	10.19
Sympathy	0.33	1.12	8.75	24.24	22.89	19.44	2.19	7.40	6.58	7.78	5.50	12.49	8.62	6.92	13.74
Doubt/Confusion	0.00	0.00	0.00	32.56	29.72	18.09	8.05	15.64	9.90	15.34	3.99	1.39	3.19	4.59	6.11
Disconnection	2.65	1.48	3.80	44.10	27.56	22.11	8.53	16.79	12.33	7.95	4.08	8.57	15.72	11.15	17.71
Fatigue	13.84	17.57	11.17	22.40	25.00	25.49	3.46	3.75	6.97	8.76	2.79	6.72	9.67	7.42	11.21
Embarrassment	0.00	0.00	0.00	3.57	5.00	0.00	2.83	2.34	0.86	4.05	3.60	3.70	0.00	0.00	0.00
Yearning	14.40	14.01	8.94	0.00	0.00	0.00	1.05	7.96	9.79	8.94	8.86	6.70	0.00	0.00	0.42
Disapproval	15.25	16.55	13.62	33.68	21.95	24.66	0.69	7.25	5.60	5.14	3.63	3.17	5.64	8.43	10.07
Aversion	4.10	9.91	2.86	5.56	0.00	10.53	0.00	1.07	1.39	0.00	0.00	0.00	0.00	0.00	0.00
Annoyance	9.16	7.05	6.22	25.34	29.91	15.91	0.69	5.48	2.68	6.67	4.89	4.23	0.00	0.00	1.10
Anger	17.59	10.65	10.62	38.10	26.32	30.77	2.55	1.47	2.73	5.31	3.46	4.86	5.15	5.31	5.65
Sensitivity	3.57	1.10	1.52	3.17	4.76	3.64	0.00	0.00	0.89	0.91	0.00	3.28	6.37	6.73	5.21
Sadness	11.79	5.05	17.36	38.27	41.22	46.81	6.38	12.80	3.89	10.68	5.49	6.62	8.70	10.95	20.36
Disquietment	1.33	3.49	3.72	32.63	26.72	24.51	0.76	5.44	4.71	4.12	0.92	4.89	0.87	2.30	5.80
Fear	0.88	1.02	5.41	18.60	18.18	26.67	5.11	4.55	5.03	0.00	0.91	1.53	1.37	2.55	9.77
Pain	8.81	6.30	13.04	34.78	36.84	27.03	0.00	4.51	5.11	5.47	1.75	4.26	0.00	2.15	8.91
Suffering	8.43	1.96	13.99	37.93	43.14	40.74	0.00	4.55	1.36	0.00	0.00	0.00	0.00	0.00	8.84
Macro F1	19.19	17.46	20.49	33.02	30.87	30.70	9.83	14.19	13.97	13.22	12.14	13.05	15.81	16.56	19.71

Table 4: Evaluation results for the Emotic dataset, measured by macro-F1 score. The best performance of a model under a specific experimental setup (E) is in green; The second best is in grey; The worst is in red.

Emotion Labels	DeepSeek			GPT			Llava-mis			Llava-qwen			Llava-vic		
	E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3
Happiness	66.76	42.34	52.73	75.15	57.71	59.61	74.44	71.26	60.80	73.23	59.07	49.96	72.42	63.90	55.74
Surprise	18.72	8.00	2.22	24.69	13.11	5.00	7.27	5.74	8.29	12.97	9.82	2.92	5.63	5.74	6.71
Anger	10.87	6.30	10.22	31.91	14.93	20.00	2.33	4.60	1.14	9.48	5.83	0.00	0.00	0.00	5.48
Fear	0.00	0.00	0.00	29.23	19.51	16.00	1.21	0.00	3.87	2.88	1.42	0.00	1.46	0.00	0.00
Peace	59.41	24.23	45.34	34.10	1.37	21.56	24.28	9.70	41.82	58.33	26.15	32.80	3.39	0.57	16.72
Sadness	2.26	0.00	1.13	32.79	20.69	18.52	20.27	15.48	2.21	24.33	16.39	5.18	2.26	1.15	4.26
Disgust	4.21	5.24	0.00	23.14	6.74	0.00	7.27	1.01	1.54	9.07	0.51	0.52	0.00	0.00	0.00
Excitement	35.24	21.51	29.66	40.72	23.68	24.47	39.92	34.95	26.06	7.65	5.51	13.89	23.72	25.40	26.55
Macro F1	24.68	13.45	15.70	32.41	17.53	18.35	19.68	17.84	4.86	21.99	15.59	11.70	13.61	12.10	5.77

Table 5: Evaluation results for the HECO dataset measured by macro-F1 score.

5. Results and Discussion

5.1. Quantitative Analysis

Tables 4 and 5 show emotion recognition results for the five MLLMs across three experimental conditions: isolated subject (E1), complete image with visual context (E2), and complete image with narrative interpretation (E3). On the Emotic dataset, the addition of visual context and narrative interpretation provided minimal benefit to MLLM performance; in some cases (e.g., GPT, Llava-qwen), performance even decreased with added information. GPT, despite its higher overall performance, achieved its best results without any context, the intigration of context or narrative interpretation reduced its accuracy by nearly half.

This trend is amplified on the HECO dataset, where performance peaked consistently with the isolated subject input. On HECO, visual context significantly diminished performance across all MLLMs, and narrative interpretations substantially reduced performance for Llava-mis, Llava-qwen, and Llava-vic. A consistent challenge across both datasets was the recognition of nuanced emotions. Several emotion categories yielded near-zero macro-F1 scores for various MLLMs, indicating a profound difficulty in distinguishing subtle emotional states. For example, DeepSeek struggled with *Peace, Esteem, Doubt/Confusion*, and *Embarrassment* on Emotic, while GPT had issues with *Peace, Embarrassment*, and *Yearning*. Similar patterns of difficulty with specific emotions were observed for the other Llava models.

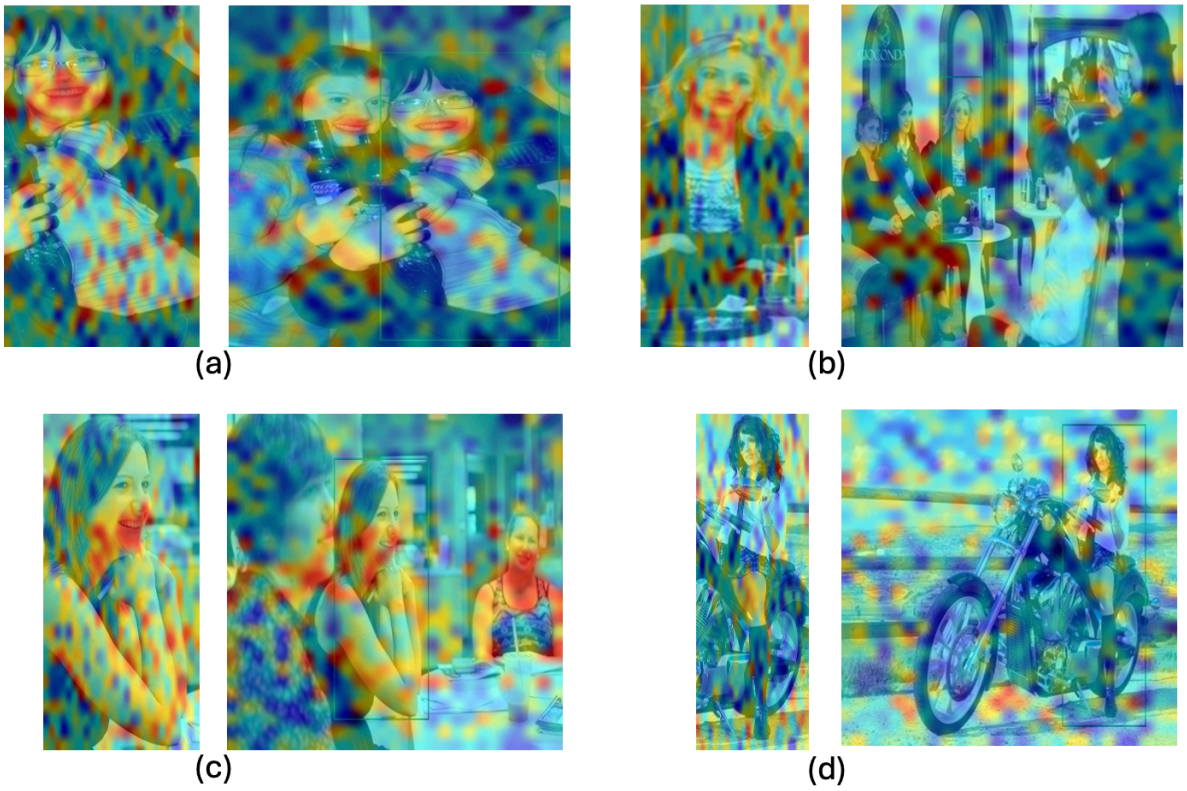

Figure 3: Grad-CAM results for input images with and without context for HECO dataset. Warmer colors indicate higher attention regions by the model.

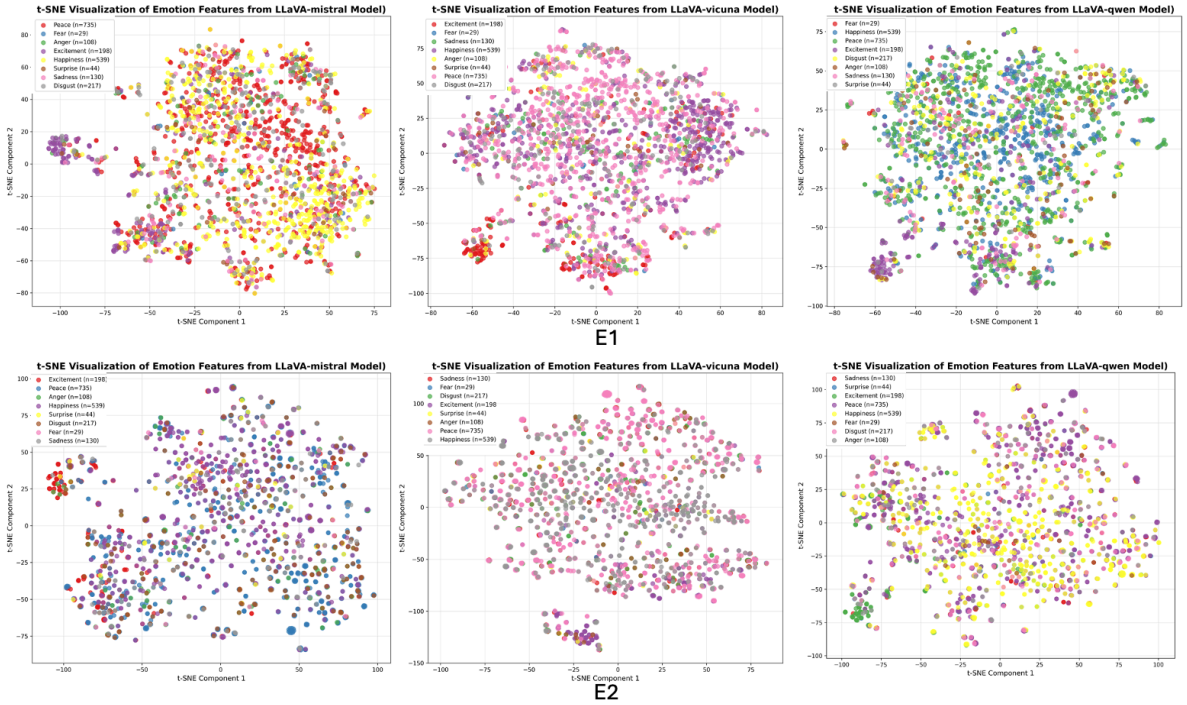

Figure 4: t-SNE plot for input images with and without context in HECO dataset. E1 corresponds to the first experiment with cropped images. E2 corresponds to the second experiment with the whole image and a bounding box on the concerned individual.

On HECO, *Fear* posed a significant challenge for DeepSeek and the Llava models, while *Sadness* and *Disgust* were problematic for DeepSeek and Llava-vic, respectively.

5.2. Qualitative Analysis

Grad-CAM results. Grad-CAM (Gradient-weighted Class Activation Mapping) is a visualization technique used to interpret neural network decisions by highlighting the regions of an input image that most influence the model's prediction. Warmer colors (e.g., red and yellow) indicate higher attention or importance assigned by the model, while cooler colors (blue) indicate less relevance. In Figure 3, each subfigure (a–d) shows pairs of images: the left image presents the target individual in isolation, and the right includes the full social or environmental context. Grad-CAM is intended to reveal where the model is focusing to infer emotional state, these results demonstrate several trends.

First, across all examples, the heatmaps show diffuse and scattered attention patterns. Rather than strongly attending to emotionally salient regions, such as faces, eyes, or body posture, the models often assign equal or even greater importance to irrelevant background elements. Second, the inclusion of visual context appears to dilute attention toward the target individual. In right-hand images, the model's focus shifts away from the boxed individual and onto nearby people or background areas, suggesting that the model lacks mechanisms for disentangling social relevance or establishing a coherent center of emotional analysis. Example (d) is an outlier in the case that the model focuses on the face more in the right image, but it predicts the wrong label. It suggests that even if the attention of the model is in right areas, they might not be able to make correct judgements. Overall, these Grad-CAM results reinforce the broader pattern seen in performance metrics: MLLMs struggle to identify and prioritize emotion-relevant cues, especially when full context is included. This lack of focused, contextually grounded attention limits their ability to emulate human-like emotion perception.

t-SNE results. Figure 4 displays t-SNE visualizations of high-dimensional emotion representations extracted from three MLLMs (LLaVA-mistral, LLaVA-vicuna, and LLaVA-qwen) under two experimental conditions: E1 (top row), where the model is shown an individual in isolation, and E2 (bottom row), where the individual appears within full visual context. Each point corresponds to a sample and is color-coded by its ground-truth emotion label. Overall, the t-SNE plots reveal that the emotion feature spaces are poorly structured across all models and settings.

Most emotion categories appear heavily entangled, with no clear or consistent clustering, suggesting that the models do not learn robust or linearly separable emotion representations. Excitement category is the exception under both the experimental settings.

Visual Context and Emotion Recognition (R1). The counterintuitive finding that less context often yields better performance for MLLMs raises concerns about the nature of their "understanding". It suggests that these models may be relying on superficial cues and statistical correlations rather than comprehensive comprehension of the emotional interactions with the environments. MLLMs may be overwhelmed by the additional context, struggling to distinguish relevant cues from irrelevant noise. This is potentially because of the fundamental limitation in their algorithms: MLLMs lack appropriate cognitive mechanisms, such as interaction-oriented attention (as opposed to task-oriented attention in Transformers (Vaswani et al., 2017)), needed to correctly relate visual context to expressed emotional states. Humans perceive visual scenes not as collections of isolated objects but as structured configurations of affordances, actions, and social interactions. MLLMs, on the other hand, may be processing visual information in a more fragmented and atomistic way, focusing on individual objects or features without fully grasping their relationship. This could explain why adding visual context often hinders their performance.

Narrative Interpretation and Emotion Recognition (R2). The consistent failure of narrative interpretations to improve performance points to a fundamental representational mismatch. Humans appraise current situation and then trigger emotional reactions. MLLMs, conversely, fail to grasp the underlying connection of the visual scene, even when the visual scene is explained as narratives in their prompts. This misalignment hinders the capacities of MLLMs from deliberate emotion analysis (similar to the slow thinking of humans), yielding the biased predictions and near-zero scores for nuanced emotions. Accurate differentiation of emotions like *Embarrassment* (from *Sadness*) or *Aversion* (from *Anger*) often requires understanding the social context and narrative trajectory that gives rise to these feelings. MLLMs may be able to identify certain keywords or phrases in the narrative description, but they fail to connect these linguistic cues to the visual context in a meaningful way. This suggests a lack of true understanding of the narratives either from implicit visual context or explicit prompts.

6. Conclusion

We evaluated the emotion recognition capabilities of Multimodal Large Language Models (MLLMs) through a series of cognition-inspired experiments that varied visual context and narrative information. Our findings reveal that current MLLMs do not exhibit robust emotion appraisal abilities aligned with human-like cognition. Specifically, we observed that the addition of visual context (E2) often degrades performance, indicating that MLLMs may rely on superficial features and lack mechanisms to effectively disambiguate emotional cues from contextual noise. Furthermore, narrative interpretations (E3) failed to enhance performance, highlighting a representational misalignment between language-based explanations and visual comprehension.

These outcomes suggest that MLLMs process multimodal information in a fragmented, object-centric manner and lack interaction-oriented attention or narrative reasoning capabilities critical for accurate emotion understanding (Zhang et al., 2026). Addressing these limitations will require rethinking model architectures and training paradigms to incorporate socially grounded, context-aware reasoning mechanisms (Ong et al., 2025). This work contributes a framework for probing these deficiencies and outlines key directions for the development of appraisal-informed MLLMs for emotion recognition.

Acknowledgments

This work is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL).

7. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Magda Blondiau Arnold. 1960. Emotion and personality. *Psychological Aspects*, 1.

Erik Cambria, Rui Mao, Amir Hussain, Keith Oatley, and Geoffrey Hinton. 2026. Artificial intelligence as the fourth decentering revolution: From cosmic, biological, and psychological displacement to cognitive decentering. *Cognitive Computation*, 18(20):1–13.

Xu Cao, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Jianguo Cao, and James M Rehg. 2024. What is the visual cognition gap between humans and multimodal LLMs? *arXiv preprint arXiv:2406.10424*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Phoebe C Ellsworth and Craig A Smith. 1988. From appraisal to emotion: Differences among unpleasant feelings. *Motivation and emotion*, 12(3):271–302.

Chunxiao Fan, Jie Lin, Rui Mao, and Erik Cambria. 2024. Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion*, 106:102306.

Xin Hong, Yuan Gong, Vidhyasaharan Sethu, and Ting Dang. 2024. AER-LLM: Ambiguity-aware emotion recognition leveraging large language models. *arXiv preprint arXiv:2409.18339*.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766.

Richard S Lazarus. 1991. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8):819.

Deng Li, Bohao Xing, Xin Liu, Baiqiang Xia, Bihan Wen, and Heikki Kälviäinen. 2025. Deemo: De-identity multimodal emotion recognition and reasoning. *arXiv preprint arXiv:2504.19549*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yijiang Li, Qingying Gao, Haoran Sun, Haiyun Lyu, Dezhi Luo, and Hokin Deng. 2024. CogDevelop2K: Reversed cognitive development in multimodal large language models. *arXiv preprint arXiv:2410.10855*.

Yuxiang Lin, Jingdong Sun, Zhi-Qi Cheng, Jue Wang, Haomin Liang, Zebang Cheng, Yifei Dong, Jun-Yan He, Xiaojiang Peng, and Xian-Sheng Hua. 2025. Why we feel: Breaking boundaries

in emotional reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5196–5206.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Hao Lu, Xuesong Niu, Jiyao Wang, Yin Wang, Qingyong Hu, Jiaqi Tang, Yuting Zhang, Kaishen Yuan, Bin Huang, Zitong Yu, et al. 2024. GPT as psychologist? preliminary evaluations for GPT-4v on visual affective computing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 322–331.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A survey on assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866, Torino, Italia. ELRA and ICCL.

Rui Mao, Mengshi Ge, Sooji Han, Wei Li, Kai He, Luyao Zhu, and Erik Cambria. 2025a. A survey on pragmatic processing techniques. *Information Fusion*, 114:102712.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

Rui Mao, Qian Liu, Xiao Li, Erik Cambria, and Amir Hussain. 2025b. Bridging minds and machines: Toward an integration of ai and cognitive science. *arXiv preprint arXiv:2508.20674*.

Keane Ong, Wei Dai, Carol Li, Dewei Feng, Hengzhi Li, Jingyao Wu, Jiaee Cheong, Rui Mao, Gianmarco Mengaldo, Erik Cambria, and Paul Pu Liang. 2025. Human behavior atlas: Benchmarking unified psychological and social behavior understanding.

Klaus R Scherer. 1987. Toward a dynamic theory of emotion: The component process model of affective states. *Geneva studies in Emotion and Communication*, 1:1–98.

Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press.

Norbert Schwarz. 1996. Feelings and phenomenal experiences. *Social psychology: Handbook of basic principles/Guilford*.

Jayant Teotia, Xulang Zhang, Rui Mao, and Erik Cambria. 2024. Evaluating vision language models in detecting learning engagement. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 496–502. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Luwei Xiao, Rui Mao, Shuai Zhao, Qika Lin, Yanhao Jia, Liang He, and Erik Cambria. 2025. Exploring cognitive and aesthetic causality for multimodal aspect-based sentiment analysis. *IEEE Transactions on Affective Computing*.

Yunhe Xie and Rui Mao. 2025. PGIF: A personality-guided iterative feedback graph network for multimodal conversational emotion recognition. *IEEE Transactions on Computational Social Systems*, 12(5):3583–3595.

Xinrun Xu, Yuxin Wang, Chaoyi Xu, Ziluo Ding, Jiechuan Jiang, Zhiming Ding, and Börje F Karlsson. 2024. A survey on game playing agents and large models: Methods, applications, and challenges. *arXiv preprint arXiv:2403.10249*.

Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. 2022. Emotion recognition for multiple context awareness. In *European Conference on Computer Vision*, pages 144–162. Springer.

Dong Zhang, Yuansheng Ma, Linqin Li, Shoushan Li, Erik Cambria, and Guodong Zhou. 2026. Training-free and zero-shot regeneration for hallucination mitigation in MLLMs: Representation understanding perspective. *Expert Systems with Applications*, 309:131102.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.