Customizing Visual Emotion Evaluation for MLLMs: An Open-vocabulary, Multifaceted, and Scalable Approach

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, Multimodal Large Language Models (MLLMs) have achieved exceptional performance across diverse tasks, continually surpassing previous expectations regarding their capabilities. Nevertheless, their proficiency in perceiving emotions from images remains debated, with studies yielding divergent results in zero-shot scenarios. We argue that this inconsistency stems partly from constraints in existing evaluation methods, including the oversight of plausible responses, limited emotional taxonomies, neglect of extra-visual factors, and labor-intensive annotations. To facilitate customized visual emotion evaluation for MLLMs, we propose an Emotion Statement Judgment task that overcomes these constraints. Complementing this task, we devise an automated pipeline that efficiently constructs emotion-centric statements with minimal human effort. Through systematically evaluating prevailing MLLMs, our study showcases their stronger performance in emotion interpretation and context-based emotion judgment, while revealing relative limitations in direct determination of sentiment polarity and personalized emotion prediction. When compared to humans, even top-performing MLLMs like GPT-4o demonstrate remarkable performance gaps, underscoring key areas for future improvement. By developing a fundamental evaluation framework and conducting a comprehensive MLLM assessment, we hope this work contributes to advancing emotional intelligence in MLLMs. Codes and data will be released.

1 Introduction

Perceiving emotional signals from visual stimuli is essential for humans to improve decision-making and build effective communication [1, 2]. To computationally model this capability, Affective Image Content Analysis (AICA) has emerged as a key research direction in computer vision [3], focusing on emotion perception through visual features [4]. Over the decades, advances in this field have given rise to various applications, including opinion mining [5], customized advertising [6], and mental health care [7]. Recently, the advent of Multimodal Large Language Models (MLLMs) [8, 9] has revolutionized image understanding tasks [10]. However, their effectiveness in AICA remains contested. Divergent findings underscore a paradox: while some studies [11, 12] demonstrate MLLMs' poor zero-shot emotion recognition performance, others successfully employ them as emotion annotators for training data augmentation [13, 14]. We attribute this discrepancy to the partial incompatibility of conventional emotion evaluation approaches with MLLMs.

Specifically, current evaluation approaches can be broadly categorized into emotion classification and emotion interpretation, as illustrated in Fig. 1 (a,b). In emotion classification, models are required to assign the emotional state of an input image to a predefined set of emotion categories, with most

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

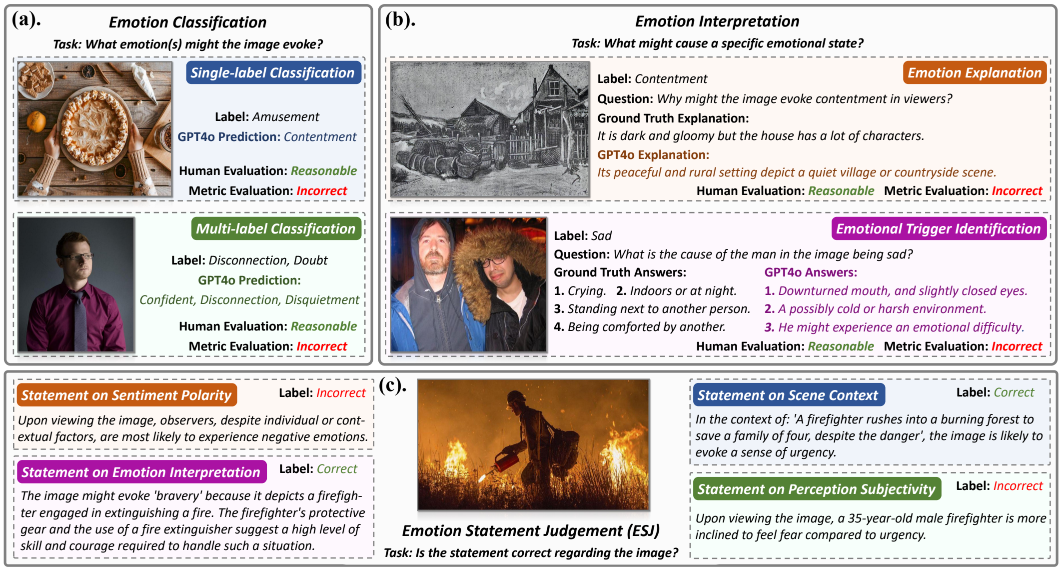

Figure 1: Illustration of different visual emotion evaluation approaches. Compared to current evaluation approaches, emotion statement judgement adopts a deterministic label while maintaining extensibility to evaluation depth and breadth.

benchmarks [15, 16, 17, 18, 19] providing a single label per image, and a few [20, 21] incorporating multiple labels. In contrast, emotion interpretation focuses on understanding the underlying causes of emotions in images. It encompasses two primary sub-tasks: explaining the causes of emotional states [22, 23] and identifying salient visual elements that contribute to the emotional response [24].

We identify four primary limitations when applying these evaluation approaches to MLLMs. *Firstly*, their adoption of fixed ground-truth answers for open-ended questions imposes structural constraints that exclude other plausible responses. Emotion perception is inherently subjective [25, 26], as the same image may evoke divergent reactions across individuals, and emotional states permit varied interpretations. As demonstrated in Fig. 1 (a,b), responses generated by GPT-4o that seem reasonable to humans are judged as inaccurate under rigid evaluation metrics. *Secondly*, they are mostly constructed upon emotion theories with limited emotional taxonomies. Popular emotion classification and interpretation benchmarks, such as FI [16] and Artemis [22], comprise only eight emotion categories. Such taxonomic granularity fails to capture fine-grained affective variations between images. *Thirdly*, they focus solely on intrinsic image attributes while neglecting extrinsic contextual factors. According to recent studies [4, 27], emotion perception can also be shaped by extra-visual factors, including the scene context where the image takes place, as well as the identity and personality of the viewer [25]. *Fourthly*, they predominantly rely on majority voting mechanisms to ensure label reliability in crowdsourced annotations [28], which is labor-intensive, particularly for fine-grained annotation tasks. EMOTIC [20], for instance, requires coordination with 23,788 annotators to label 18,316 images. This operational burden severely constrains dataset scalability in magnitude and generalization capacity across image domains.

To facilitate customized visual emotion evaluations for MLLMs, we propose a dual-component solution that addresses these limitations: the Emotion Statement Judgment (ESJ) task, complemented by the INSETS (**IN**telligent Vi**S**ual **E**motion **T**agger and **S**tatement Constructor) pipeline for efficient annotation. As a pioneering effort, we prioritize a precise over a complex evaluation design, aiming to establish a fundamental offline standard. With this purpose, **ESJ reformulates visual emotion evaluation by requiring MLLMs to validate emotion-centric statements for a given image**. It effectively mitigates ambiguity in open-ended questions while being highly extensible for evaluation depth and diversity. Meanwhile, INSETS annotates images with multiple open-vocabulary emotion labels, significantly refining the emotional taxonomies. These labels are then utilized to construct multifaceted emotion-centric statements, covering intrinsic image attributes like sentiment polarity and emotion interpretation, as well as extrinsic contextual factors like scene context and perception subjectivity. Crucially, only minimal human intervention is required, ensuring a high scalability of the approach. An example of ESJ is illustrated in Fig. 1 (c).

Leveraging INSETS, we introduce two ESJ benchmarks: a large-scale INSETS-462K and its human-refined subset INSETS-3K. Systematic evaluation demonstrates that recent MLLMs exhibit non-trivial visual emotion perception capabilities, yet maintain non-negligible performance gaps compared to humans, particularly in discerning sentiment polarity and comprehending perception subjectivity. In summary, the contributions of this paper are three-fold:

- This paper constitutes a pioneer effort to identify limitations in existing visual emotion evaluations for MLLMs and address them with a customized ESJ task.
- Complementing the ESJ task, this paper designs the INSETS pipeline, providing a reliable approach to annotating images with multiple open-vocabulary emotion labels and constructing multifaceted emotion-centric statements with minimal human effort.
- Utilizing the ESJ task and the INSETS pipeline, this paper conducts a systematic evaluation of recent MLLMs in visual emotion understanding, offering insights and fostering further developments of emotional intelligence in MLLMs.

2 Related Works

2.1 AICA Benchmarks

Psychological researchers conceptualize emotion representation through two principal frameworks: the Categorical Emotion Space (CES), which discretizes affective states into predefined taxonomies, and the Dimensional Emotion Space (DES), which maps emotions onto continuous 2D/3D coordinate systems (e.g., valence-arousal-dominance (VAD) axes [29]). For simplicity and better interpretability, most benchmarks adopt emotion classification evaluations based on discrete CES emotion taxonomies. This category encompasses both early small-scale benchmarks, such as IAPSa [30] and Abstract [15], as well as later larger-scale benchmarks like FI [16] and WebEmo [18]. Over time, benchmarks with enriched metadata have also been developed. Notable examples include EMOTIC [20], which integrates multiple emotion categories, VAD values, and human-related bounding boxes, and EmoSet [19], which employs describable emotion attributes that cover different levels of visual information.

Some other benchmarks adopt emotion interpretation evaluations by extending CES-based taxonomies with additional emotional explanations, such as Artemis [22] and Affection [23]. EIBench [24] diverges slightly, shifting focus on identifying and extracting visual emotional triggers. Based on these benchmarks, numerous expert models [31, 32, 33] have been developed, demonstrating strong performance under the fine-tuning and testing paradigm. In contrast to them, MLLMs are commonly pre-trained on web-scale data, without explicitly aligning with benchmark-specific knowledge. This discrepancy introduces multiple constraints when applying conventional benchmarks to MLLMs, necessitating customized visual emotion evaluation approaches that account for their generalized knowledge structures.

2.2 Evaluation of MLLMs

Recent years have witnessed surging academic and industrial interest in MLLMs. Unlike earlier models that are limited to specific domains, MLLMs demonstrate versatile competence across diverse tasks [34, 35], fueling expectations about their trajectory toward Artificial General Intelligence [36]. To evaluate MLLMs, various benchmarks have been established to examine their capabilities in areas such as perception [37, 9], reasoning [38, 39], ethics [40, 41], and specialized domains [42, 43]. However, emotional intelligence remains conspicuously underexplored. In existing efforts, FABA-Bench [44] evaluates MLLMs' comprehension of facial expressions and actions, MM-BigBench [45] simply aggregates mainstream image-text benchmarks, and EmoBench [46] is confined to solely language modality. To fill this gap, we propose the ESJ task and corresponding benchmarks INSETS-462K and INSETS-3K to advance more comprehensive visual emotion evaluation of MLLMs.

3 Emotion Statement Judgement

ESJ aims to evaluate the proficiency of MLLMs in perceiving emotions from visual content. In each evaluation trial, MLLMs receive an image and a paired emotion-centric statement. MLLMs are then required to judge the correctness of the statement in relation to the image by responding with **Correct** or **Incorrect**. To ensure both depth and breadth in evaluation, we systematically design emotion-centric statements from four dimensions:

Sentiment Polarity Statements require MLLMs to decide sentiment polarities without any additional clues. They assess MLLMs' proficiency in directly identifying the basic emotional tone.

Emotion Interpretation Statements ask MLLMs to verify the consistency between affective explanations and corresponding emotional states. They measure MLLMs' affective reasoning capability given specific emotional triggers.

Scene Context Statements probe MLLMs' comprehension of the dynamic interplay between the external scene context where the image takes place, and image-evoked emotional responses.

Perception Subjectivity Statements task MLLMs to predict the personalized emotional responses under assumptions of specific viewer identities, examining whether MLLMs can recognize how subjectivity shapes emotional perceptions.

Collectively, these dimensions establish a holistic visual emotion evaluation framework for MLLMs. They cover both intrinsic image attributes emphasized in existing benchmarks and underexplored extrinsic contextual factors critical for human emotional perception [47, 48].

4 Annotation Pipeline: INSETS

Complementing the ESJ task, we design an automated pipeline for constructing emotion-centric statements, termed INSETS (**IN**telligent Vi**S**ual **E**motion **T**agger and **S**tatement Constructor). It operates through two primary stages: open-vocabulary emotion tagging and emotion statement construction. The prompts and templates used in the process are listed in Appendix Table 5.

4.1 Preliminary: Parrott's Hierarchical Model

We first introduce a well-established emotion model, which provides essential context for understanding the subsequent stages. Parrott's Hierarchical model [49, 50] is a tree-structured emotion taxonomy, comprising 6 primary emotions, 25 secondary emotions, and 113 tertiary emotions. The primary category contains three positive emotions (joy, love, and surprise) and three negative emotions (anger, fear, and sadness). Secondary emotions offer more diverse emotional states, each categorized under a corresponding primary emotion, while tertiary emotions further refine secondary emotions into more specific affective states. The complete taxonomy is presented in Appendix Table 7.

4.2 Open-vocabulary Emotion Tagging

At this stage, INSETS aims to assign open-vocabulary emotion labels for images, laying a solid foundation for constructing meaningful emotion-centric statements, with its procedure depicted in Fig. 2. According to Cheng *et al.* [14], MLLMs demonstrate promising capabilities in generating emotional descriptions from visual content and extracting underlying emotions from these descriptions. However, challenges such as hallucinations [51], trustworthiness issues [52], and inherent limitations in emotional perception can lead to inaccuracies in the extracted emotions. To enhance reliability, we devise an ensemble-based majority voting mechanism, aggregating outputs from multiple MLLMs to cross-validate and refine emotion label assignments.

Given an image sample, we first extract its potential open-vocabulary emotions from multiple MLLMs. MLLMs are prompted to analyze the emotions evoked by the image (with #1 prompt in Table 5, abbreviated as "#1" in the following) and then extract emotions applicable to the image (#2) [Fig. 2 (a)]. This process is iteratively applied to all images in the dataset, aggregating potential emotions into an emotion pool. Next, we refine this pool by filtering out words unsuitable as emotion descriptors (#3), using GPT-4 [53] as the judge due to its superior linguistic emotional perception [46] [Fig. 2 (b)]. Once the filtered emotion pool is obtained, we attach the remaining emotions to Parrott's hierarchical emotion model [Fig. 2 (c)]. Specifically, GPT-4 is prompted to categorize each open-vocabulary emotion into the closest tertiary emotion in Parrott's model (#4), followed by manual refinement by a human expert. This process results in an extended version of Parrott's model, which we refer to as the Parrott-based Open-vocabulary Hierarchical Model (POM) [Fig. 2 (d)]. This unified framework

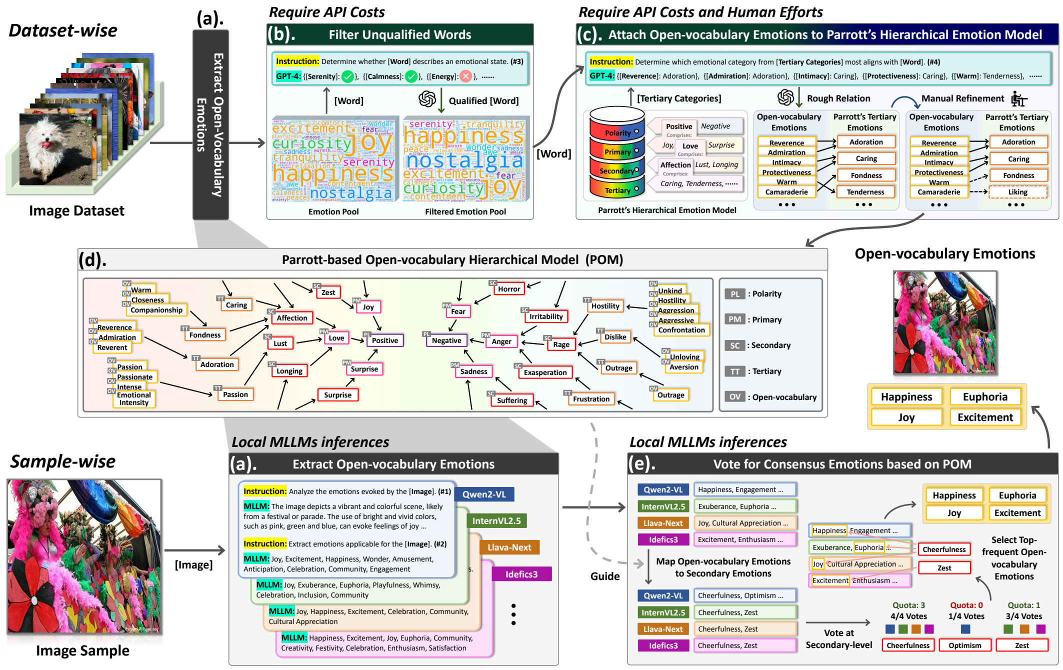

Figure 2: Illustration of the open-vocabulary emotion tagging stage. We first extract all potential open-vocabulary emotions from the image dataset (a) and then attach these emotions to a well-established emotion model (b,c). Through this model (d), we identify and select open-vocabulary emotions consistently recognized by multiple MLLMs as the labels of each image (e).

enables multi-level tracing of affective states for each open-vocabulary emotion, facilitating more accurate and interpretable emotion tagging.

Subsequently, leveraging POM, the ensemble-based majority voting mechanism selects reliable open-vocabulary emotion labels for images [Fig. 2 (e)]. First, open-vocabulary emotions extracted from multiple MLLMs are mapped to secondary emotion categories, with model voting determining the quota of open-vocabulary emotions for each secondary category. Second, within each secondary category, the open-vocabulary candidates are ranked based on their frequencies, and the top-ranked ones are selected according to the allocated quota. This process guarantees label reliability while preserving the open-vocabulary nature of the emotion labels, thereby achieving synergistic optimization between annotation precision and semantic coverage.

4.3 Emotional Statement Construction

Building upon the assigned emotion labels, we construct both correct and incorrect emotion-centric statements from four dimensions: sentiment polarity, emotion interpretation, scene context, and perception subjectivity. Its procedure is illustrated in Fig. 3.

The construction pipeline initiates with prototype statement generation [Fig. 3 (a)]. For each emotion label, we trace it back to the MLLM that extracts it, prompting the MLLM to generate three prototype statements: **[1]**. *prototype interpretation* of the emotion by inquiring about the cause of the emotion (#5); **[2]**. *prototype context* that aligns with the emotion by requesting a background story (#6); and **[3]**. *prototype character* who would experience the emotion by questioning the possible identity of the viewer (#7). From the dataset perspective, the prototype generation is distributed across multiple MLLMs, ensuring diversity in the subsequent statement construction.

Sentiment Polarity Statement Construction [Fig. 3 (b)]: Under the guidance of POM, we derive sentiment polarity by mapping the labels to primary emotions. Each image's sentiment polarity is classified into three mutually exclusive categories: **1). Fully Positive** when all labels reside in the positive spectrum; **2). Fully Negative** when all labels reside in the negative spectrum; **3). Mixed** when positive and negative labels both exist. Next, the ground truth correctness of three predefined statements on sentiment polarity (#8,9,10) is determined accordingly.

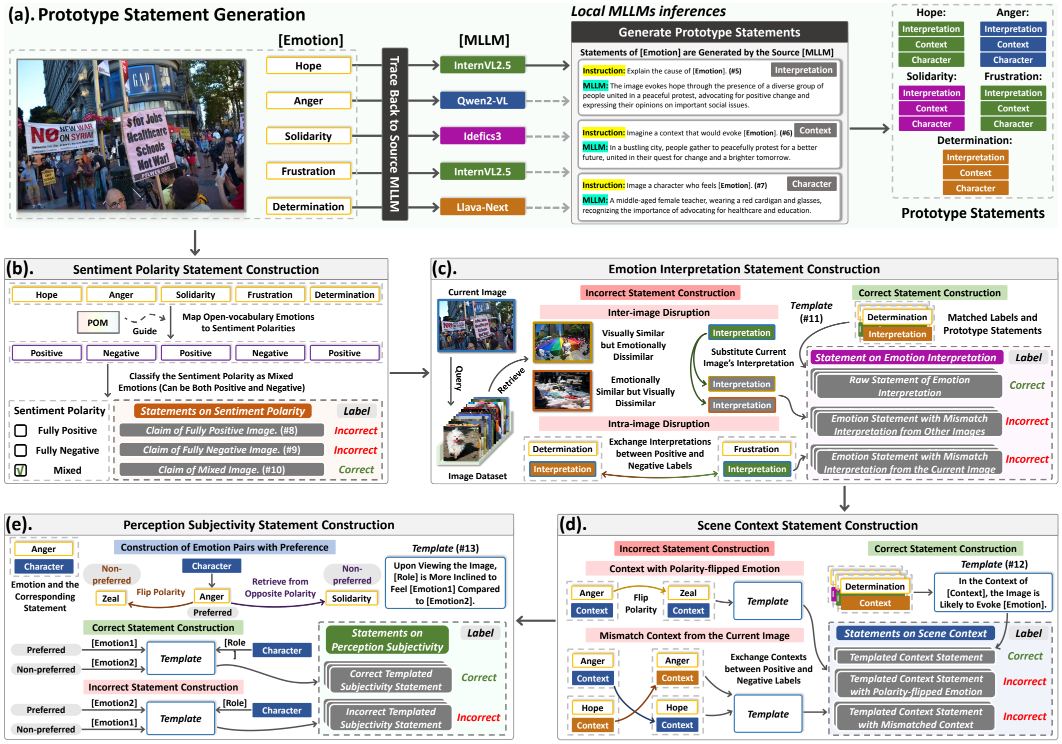

Figure 3: Illustration of the emotional statement construction stage. It begins with prototype statement generation (a) for each emotion label, which is distributed across multiple MLLMs. Then, based on the assigned emotion labels and the corresponding prototype statements, correct and incorrect emotion-centric statements are constructed from four dimensions: sentiment polarity (b), emotion interpretation (c), scene context (d), and perception subjectivity (e).

Emotion Interpretation Statement Construction [Fig. 3 (c)]: We combine a prototype interpretation with an emotional state (#11) to construct an emotion interpretation statement. Matched labels and prototype statements are assigned as correct statements, while unmatched ones are considered incorrect. To construct mismatched pairs, we design two strategies: inter-image and intra-image disruption. The former retrieves two images from the dataset - one exhibiting visual similarity but emotional dissimilarity to test whether MLLMs can comprehend the affective gap [54], and the other demonstrating emotional similarity but visual dissimilarity to evaluate whether MLLMs can identify the emotional triggers in images. Visual similarity is quantified using CLIP-score [55], whereas emotional similarity is determined by the correspondence at tertiary emotions of POM. The latter strategy exchanges interpretations between emotion labels of contrasting polarity within identical images, aiming to assess whether MLLMs can establish precise causal linkages between emotional triggers and specific emotions.

Scene Context Statement Construction [Fig. 3 (d)]: We combine a prototype context and an emotional conclusion (#12) to form a scene context statement, where the construction of correct statements mirrors the previous case. The strategy for incorrect statements differs slightly: in addition to exchanging prototype contexts between emotion labels of contrasting polarity within identical images, we devise a flip polarity operation. Specifically, the emotional label is substituted with a tertiary emotion randomly sampled from the opposing polarity spectrum in POM.

Perception Subjectivity Statement Construction [Fig. 3 (e)]: We combine a prototype character and their inclination toward one of two candidate emotions (#13) to form a perception subjectivity statement. We first create emotion pairs with preference bias. For each prototype character, its preferred emotion is the corresponding emotion label. The non-preferred emotion is derived by either retrieving opposite polarity emotions within the image or the flip polarity operation for the preferred emotion. Next, correct and incorrect statements are constructed by configuring emotion pairs into canonical or anomalous orders, respectively.

Table 1: Statistics of the MLLMs employed in INSETS. For each MLLM, we report the number of parameters, the average extracted emotions per image, the number selected as emotion labels, and the proportion of prototype statements it generates.

MLLMs	#P (B)	Extracted Emotion	Selected Emotion	Generated Statement
LLaVa-1.6 [56]	7.6	8.3	2.4	9.8%
Mantis [57]	8.5	12.6	2.9	13.1%
mPLUG-Owl3 [58]	8.1	9.2	2.7	11.2%
Idefics3 [59]	8.5	10.0	2.9	12.5%
Phi-3.5-Vision [60]	4.1	9.9	2.8	11.7%
Qwen2-VL [61]	8.3	8.8	2.7	10.9%
Llama-3.2-Vision [62]	10.7	7.2	2.3	9.3%
Molmo [63]	8.0	10.8	2.7	12.0%
InternVL2.5 [64]	8.3	8.5	2.3	9.5%

Table 2: Statistics of emotion labels and statements in INSETS-462K and INSETS-3K.

INSETS-462K	
Number of Images	17,716
Number of Statements	462,369
Emotion Labels Per Image	4.9
Distinct Emotion Labels	751
Statements Per Image	26.1
Average Length of Statements	39.0
INSETS-3K	
Number of Images	3,086
Number of Statements	3,086
Emotion Labels Per Image	5.2
Distinct Emotion Labels	424
Statements Per Image	1.0
Average Length of Statements	37.0

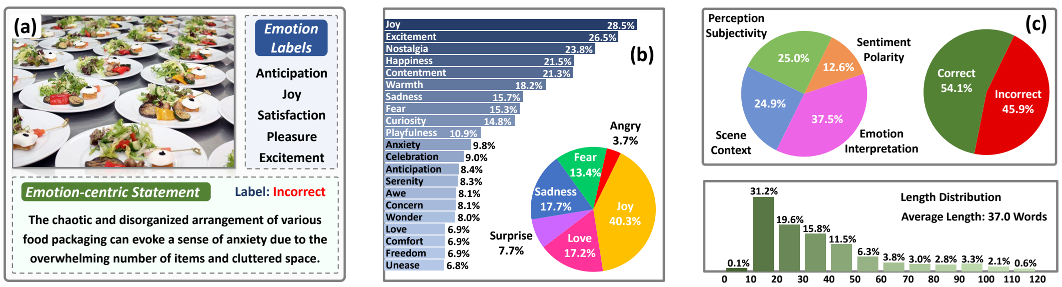

Figure 4: A closer gaze at INSETS-3K. Illustrations of a sample (a), the distribution of emotion labels (b), and the distribution of emotion-centric statements (c).

4.4 Evaluation Benchmarks

In this process, INSETS begins with an image dataset, relying primarily on local MLLM inferences to assign reliable open-vocabulary emotion labels and construct meaningful emotion-centric statements. For open-vocabulary emotion extraction and prototype statement generation, we employ nine recent popular MLLMs with impressive performance [65]. As reported in Table 1, the final assigned emotion labels and prototype statements are evenly sourced across the MLLMs, ensuring diversity in the constructed benchmark. Given the high quality of Emoset [19], we select it as the image source. From a subset of 17,716 images, we construct 462K emotion-centric statements, referred to as INSETS-462K.

To further enhance accuracy, we sample 3,164 distinct image-statement pairs from INSETS-462K for human validation. Among them, annotators judge 218 pairs (6.9%) as inaccurate, 2,868 pairs (90.6%) as accurate, and 79 pairs (2.5%) as ambiguous. Accurate pairs are retained, incorrect labels are corrected, and ambiguous ones are discarded, yielding a high-quality subset of 3,086 distinct image-statement pairs, which we name INSETS-3K. Although the open-vocabulary emotion labels are not required in the ESJ task, we retain them in both INSETS-462K and INSETS-3K to enhance interpretability and facilitate future development of the benchmarks. The statistics of emotion labels and statements are reported in Table 2.

5 Analysis and Evaluation

5.1 Details of INSETS-3K

To gain deeper insights into the properties of INSETS-3K, we provide detailed statistics in Fig. 4. A sample is shown in Fig. 4 (a), which includes five emotion labels and an emotion-centric statement. Fig. 4 (b) illustrates the distribution of popular emotion labels, where the most frequent labels include Joy, Excitement, Nostalgia, Happiness, and Contentment. When mapped to the primary emotions in Parrott's model, Joy is the most dominant category (40.3%), followed by Sadness (17.7%),

Table 3: Evaluation results on INSETS-3K. The MLLMs involved in constructing INSETS-3K are listed in the upper part of the table, while other results are listed in the lower part. The highest values in each section are marked in **bold**.

MLLMs	#Param	Accuracy					Positive Ratio	Give-up Ratio
		Sentiment Polarity	Emotion Interpretation	Scene Context	Perception Subjectivity	Total		
LLaVa-1.6 [56]	7.6B	66.4	69.7	55.3	49.7	60.2	18.4	0
Mantis [57]	8.5B	61.2	65.9	67.2	61.2	64.4	84.4	0.1
mPLUG-Owl3 [58]	8.1B	73.9	79.3	81.7	**75.0**	**78.1**	67.3	0
Idefics3 [59]	8.5B	75.4	78.6	75.5	62.6	73.4	49.5	0.2
Phi-3.5-Vision [60]	4.1B	74.7	72.5	82.6	74.8	75.9	64.1	0
Qwen2-VL [61]	8.3B	70.7	75.0	**86.1**	72.8	76.6	65.7	0
Llama-3.2-Vision [62]	10.7B	68.7	75.9	85.2	72.0	76.3	71.2	0.2
Molmo [63]	8.0B	61.4	76.0	79.2	59.4	70.7	38.1	0
InternVL2.5 [64]	8.3B	**75.7**	**80.2**	79.4	61.3	74.7	52.9	0.2
BLIP2 [66]	7.7B	51.1	52.8	55.4	52.5	53.2	96.8	2.5
InstructBLIP [67]	7.9B	29.8	40.5	33.9	37.8	36.8	43.8	37.5
Otter [68]	8.2B	32.6	21.4	32.1	27.2	27.0	9.9	52.1
DeepSeek-VL [69]	7.3B	68.7	70.8	81.1	**73.2**	73.7	73.1	0
Paligemma [70]	2.9B	50.6	46.3	49.3	45.7	47.4	49.4	5.5
MiniCPM [71]	8.7B	70.4	78.4	81.9	70.5	76.2	66.0	0
Qwen2.5-VL [72]	8.3B	63.2	81.5	**83.9**	66.3	75.9	45.9	0
GPT4o-mini [53]	–	62.5	80.0	78.9	71.8	75.4	49.5	0
GPT4o [53]	–	**72.5**	**84.3**	81.6	69.2	**78.3**	65.0	1.6

Love (17.2%), Fear (13.4%), Surprise (7.7%), and Anger (3.7%). This distribution suggests a rich representation of emotions, ensuring coverage of diverse affective states. Fig. 4 (c) presents statistics on the statements, which exhibit a natural length distribution and a well-balanced distribution across the four evaluation dimensions as well as correct/incorrect labels.

5.2 Evaluation Preparations

We evaluate MLLMs through the ESJ task. Specifically, we provide each MLLM with an image-statement pair and prompt it to determine the correctness of the statement. The prompt is formulated as: *"Based on the provided image and emotional statement, please determine whether the statement aligns with the content of the image. If it does, respond with* ***Correct****. If it does not, respond with* ***Incorrect****."* Each image-statement pair is queried three times per MLLM, and the most frequent response is selected as the final decision. After collecting responses for all images, we adopt accuracy as the primary evaluation metric. As identified in prior work [73], some MLLMs may exhibit a strong bias toward either positive or negative responses, which may compromise accuracy-based evaluation validity. To mitigate this, we introduce two diagnostic metrics: *Positive Ratio* calculates the proportion of positive responses out of all responses; *Give-up Ratio* measures the proportion of cases where the MLLM neither provides a positive nor a negative response.

To conduct a comprehensive evaluation, we adopt a diverse range of MLLMs, including both open-source and closed-source ones. Besides the MLLMs used to construct the benchmarks, we also incorporate the following MLLMs: BLIP-2 [66], InstructBLIP [67], Otter [68], Deepseek-VL [69], Paligemma [70], MiniCPM [71], Qwen2.5-VL [72], GPT4o-mini, and GPT4o [53]. All experiments are conducted on NVIDIA GeForce RTX 4090 GPUs.

5.3 Results and Findings

The evaluation results on INSETS-3K are reported in Table 3. Overall, more recent MLLMs outperform earlier MLLMs. The latter suffers from severe response biases or instructional failures. This indicates that advancements in general visual tasks also enhance emotional perception. However, no MLLM achieves optimal performance across all tasks. While InternVL2.5 and GPT4o demonstrate superior performance in identifying the basic emotional tone and performing affective reasoning, they exhibit comparative deficiencies in contextual and personalized emotion prediction. This underscores the multifaceted challenges of visual emotion understanding. Further evaluation on INSETS-462K is reported in Appendix Table 6.

Table 4: Evaluation results of MLLMs and humans on a subset of INSETS-3K containing 300 image-statement pairs.

MLLMs	#Param	Accuracy					Positive Ratio	Give-up Ratio
		Sentiment Polarity	Emotion Interpretation	Scene Context	Perception Subjectivity	Total		
mPLUG-Owl3 [58]	8.1B	74.6	**80.4**	82.9	**77.2**	**79.5**	67.3	0
Phi-3.5-Vision [60]	4.1B	75.4	72.9	**83.9**	73.3	76.1	64.0	0
InternVL2.5 [64]	8.1B	**77.2**	79.5	79.3	63.2	75.1	52.1	0
DeepSeek-VL [69]	7.3B	70.2	70.5	80.2	**73.7**	73.7	73.5	0
MiniCPM [71]	8.7B	70.2	78.9	82.4	72.4	77.1	65.4	0
Qwen2.5-VL [72]	8.3B	64.0	81.5	**83.3**	68.0	76.4	47.4	0
GPT4o-mini [53]	-	64.0	79.2	77.5	71.3	74.9	49.8	0
GPT4o [53]	-	**73.7**	**84.5**	81.2	71.1	**79.0**	64.6	0.6
Human Average	–	92.3	90.1	95.3	89.6	91.6	53.4	0
Human Best	–	97.4	95.8	98.7	94.7	95.2	–	–

Comparison with Human Performance: We sample 300 image-statement pairs from INSETS-3K and evaluate 25 human participants alongside leading MLLMs. As shown in Table 4, humans achieve near-perfect accuracy. In contrast, MLLMs exhibit notable performance gaps, particularly in determining sentiment polarity and understanding perception subjectivity. Given their comparatively high accuracy on emotion interpretation statements, we suggest that the affective reasoning from emotional clues to emotional states is essential for MLLMs to perceive emotions. Regarding perception subjectivity, we speculate that MLLMs may lack sufficient awareness of individual differences. Overall, ESJ is a fundamental task format, and the performance gap between MLLMs and humans highlights the considerable potential for improvement in MLLMs' visual emotional intelligence.

6 Limitations and Discussion

Several limitations in this work can be further improved. *First*, our evaluation primarily focuses on MLLMs with parameters under 10B due to computational constraints imposed by hardware. Although this covers practical deployment scenarios, it excludes larger-scale open-source MLLMs that may exhibit superior visual emotion perception capabilities. *Second*, the current implementation is limited to monolingual evaluation. Yet we highlight that adapting INSETS for multilingual construction would require relatively limited engineering effort, primarily involving adjustments in MLLM selection, prompt design, and template configuration. Moreover, while we explored basic reasoning strategies, advanced strategies such as in-context learning and chain-of-thought prompting remained underexplored. Systematically investigating these techniques can potentially reveal deeper insights into MLLMs' visual emotion perception mechanisms. *Third*, due to the lack of human validation, the INSETS-462K benchmarks inevitably incorporate certain noises and inaccuracies. Nevertheless, the validation process of the INSETS-3K benchmark suggests that over 85% image-statement pairs retain high accuracy. This can be attributed to our efforts in ensuring reliability, including the ensemble-based majority voting mechanism and necessary human interventions.

Future work includes expanding MLLM coverage, extending INSETS to support multilingual construction and further refinement of the human-AI collaborative annotation process.

7 Conclusion

In this paper, we propose the ESJ task and a complemented automated pipeline, INSETS, to advance open-vocabulary, multifaceted, and scalable visual emotion evaluation in MLLMs. Through them, we provide a nuanced and comprehensive evaluation framework that covers sentiment polarity, emotion interpretation, scene context, and perception subjectivity. Our evaluation reveals that while MLLMs exhibit certain capabilities in visual emotion perception, they still lag behind humans, highlighting the necessity of developing emotion-oriented training objectives. We hope that INSETS-3K and INSETS-462K can serve as reliable benchmarks to advance future research, fostering the development of MLLMs with improved emotional reasoning and understanding.

References

[1] N. S. Schutte, J. M. Malouff, C. Bobik, T. D. Coston, C. Greeson, C. Jedlicka, E. Rhodes, and G. Wendorf, "Emotional intelligence and interpersonal relations," *The Journal of Social Psychology*, vol. 141, no. 4, pp. 523–536, 2001.

[2] J. Yang, J. Li, L. Li, X. Wang, and X. Gao, "A circular-structured representation for visual emotion distribution learning," in *CVPR*, 2021, pp. 4237–4246.

[3] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *WIREs Data Mining Knowl. Discov.*, vol. 8, no. 4, 2018.

[4] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: Two decades review and new perspectives," *Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6729–6751, 2022.

[5] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T. Chua, "Predicting personalized emotion perceptions of social images," in *MM*, 2016, pp. 1385–1394.

[6] A. Shukla, S. S. Gullapuram, H. Katti, M. S. Kankanhalli, S. Winkler, and R. Subramanian, "Recognition of advertisement emotions with application to computational advertising," *Trans. Affect. Comput.*, vol. 13, no. 2, pp. 781–792, 2022.

[7] G. N. Yannakakis, "Enhancing health care via affective computing," 2018.

[8] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," 2023.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[10] Z. Yang, L. Li, K. Lin, J. Wang, C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v(ision)," *CoRR*, vol. abs/2309.17421, 2023.

[11] H. Xie, C. Peng, Y. Tseng, H. Chen, C. Hsu, H. Shuai, and W. Cheng, "Emovit: Revolutionizing emotion insights with visual instruction tuning," in *CVPR*, 2024, pp. 26 586–26 595.

[12] S. Bhattacharyya and J. Z. Wang, "Evaluating vision-language models for emotion recognition," *arXiv preprint arXiv:2502.05660*, 2025.

[13] Z. Lian, H. Sun, L. Sun, J. Yi, B. Liu, and J. Tao, "Affectgpt: Dataset and framework for explainable multimodal emotion recognition," *CoRR*, vol. abs/2407.07653, 2024.

[14] Z. Cheng, Z. Cheng, J. He, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. G. Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," in *NeurIPS*, 2024.

[15] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *MM*, 2010, pp. 83–92.

[16] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *AAAI*, 2016, pp. 308–314.

[17] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *ICCV Workshops*, 2017, pp. 308–317.

[18] R. Panda, J. Zhang, H. Li, J. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: Understanding and overcoming dataset bias," in *ECCV*, vol. 11206, 2018, pp. 594–612.

[19] J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, and H. Huang, "Emoset: A large-scale visual emotion dataset with rich attributes," in *ICCV*, 2023, pp. 20 326–20 337.

[20] R. Kosti, J. M. Álvarez, A. Recasens, and À. Lapedriza, "EMOTIC: emotions in context dataset," in *CVPR Workshops*, 2017, pp. 2309–2317.

[21] Z. Wei, J. Zhang, Z. Lin, J.-Y. Lee, N. Balasubramanian, M. Hoai, and D. Samaras, "Learning visual emotion representations from web data," in *CVPR*, 2020, pp. 13 106–13 115.

[22] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas, "Artemis: Affective language for visual art," in *CVPR*, 2021, pp. 11 569–11 579.

[23] P. Achlioptas, M. Ovsjanikov, L. J. Guibas, and S. Tulyakov, "Affection: Learning affective explanations for real-world visual data," in *CVPR*, 2023, pp. 6641–6651.

[24] E. Authors, “Eibench: Assessing the emotion interpretation ability of vision large language models,” https://github.com/Lum1104/EIBench, 2024.

[25] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T. Chua, “Predicting personalized emotion perceptions of social images,” in *MM*, 2016, pp. 1385–1394.

[26] J. Yang, M. Sun, and X. Sun, “Learning visual sentiment distributions via augmented conditional probability neural network,” in *AAAI*, 2017, pp. 224–230.

[27] S. Zhao, X. Hong, J. Yang, Y. Zhao, and G. Ding, “Toward label-efficient emotion and sentiment analysis,” *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1159–1197, 2023.

[28] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in *CVPR*, 2017, pp. 2584–2593.

[29] H. Schlosberg, “Three dimensions of emotion.” *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[30] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, “Emotional category data on images from the international affective picture system,” *Behavior Research Methods*, vol. 37, pp. 626–630, 2005.

[31] L. Xu, Z. Wang, B. Wu, and S. Lui, “MDAN: multi-level dependent attention network for visual emotion analysis,” in *CVPR*. IEEE, 2022, pp. 9469–9478.

[32] G. Jia and J. Yang, “S^2-ver: Semi-supervised visual emotion recognition,” in *ECCV*, ser. Lecture Notes in Computer Science, vol. 13697, 2022, pp. 493–509.

[33] T. Feng, J. Liu, and J. Yang, “Probing sentiment-oriented pretraining inspired by human sentiment perception mechanism,” in *CVPR*, 2023, pp. 2850–2860.

[34] J. Li and W. Lu, “A survey on benchmarks of multimodal large language models,” *CoRR*, vol. abs/2408.08632, 2024.

[35] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, “Eyes wide shut? exploring the visual shortcomings of multimodal llms,” in *CVPR*. IEEE, 2024, pp. 9568–9578.

[36] Y. Maruyama, “The conditions of artificial general intelligence: Logic, autonomy, resilience, integrity, morality, emotion, embodiment, and embeddedness,” in *AGI*, ser. Lecture Notes in Computer Science, vol. 12177, 2020, pp. 242–251.

[37] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, “Seed-bench: Benchmarking multimodal llms with generative comprehension,” *CoRR*, vol. abs/2307.16125, 2023.

[38] J. Nie, G. Zhang, W. An, Y. Tan, A. C. Kot, and S. Lu, “Mmrel: A relation understanding dataset and benchmark in the MLLM era,” *CoRR*, vol. abs/2406.09121, 2024.

[39] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S. Zhu, “RAVEN: A dataset for relational and analogical visual reasoning,” in *CVPR*, 2019, pp. 5317–5327.

[40] Y. Qian, H. Zhang, Y. Yang, and Z. Gan, “How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts,” *CoRR*, vol. abs/2402.13220, 2024.

[41] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, and et al., “Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models,” in *CVPR*, 2024, pp. 14 375–14 385.

[42] P. Chen, J. Ye, G. Wang, Y. Li, Z. Deng, and et al., “Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical AI,” in *NeurIPS*, 2024.

[43] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y. Jiang, “Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario,” in *AAAI*, 2024, pp. 4542–4550.

[44] Y. Li, A. Dao, W. Bao, Z. Tan, T. Chen, H. Liu, and Y. Kong, “Facial affective behavior analysis with instruction tuning,” in *ECCV*, ser. Lecture Notes in Computer Science, vol. 15076, 2024, pp. 165–186.

[45] X. Yang, W. Wu, S. Feng, M. Wang, D. Wang, and et al., “Mm-bigbench: Evaluating multimodal models on multimodal content comprehension tasks,” 2023.

[46] S. Sabour, S. Liu, Z. Zhang, J. M. Liu, J. Zhou, and et al., “Emobench: Evaluating the emotional intelligence of large language models,” in *ACL*, 2024, pp. 5986–6004.

[47] G. Stemmler and J. Wacker, "Personality, emotion, and individual differences in physiological responses," *Biological psychology*, vol. 84, no. 3, pp. 541–551, 2010.

[48] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[49] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: Further exploration of a prototype approach." *Journal of personality and social psychology*, vol. 52, no. 6, p. 1061, 1987.

[50] W. G. Parrott, *Emotions in Social Psychology: Essential Readings*. Psychology Press, 2001.

[51] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou, "Hallucination of multimodal large language models: A survey," *CoRR*, vol. abs/2404.18930, 2024.

[52] X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, and Y. Qiao, "Mm-safetybench: A benchmark for safety evaluation of multimodal large language models," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 15114, 2024, pp. 386–403.

[53] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023.

[54] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, 2006.

[55] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, and et. al., "Learning transferable visual models from natural language supervision," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 8748–8763.

[56] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *CVPR*, 2024, pp. 26 286–26 296.

[57] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen, "MANTIS: interleaved multi-image instruction tuning," *CoRR*, vol. abs/2405.01483, 2024.

[58] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, and et al., "mplug-owl3: Towards long image-sequence understanding in multi-modal large language models," *CoRR*, vol. abs/2408.04840, 2024.

[59] H. Laurençon, A. Marafioti, V. Sanh, and L. Tronchon, "Building and better understanding vision-language models: Insights and future directions," *CoRR*, vol. abs/2408.12637, 2024.

[60] M. I. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, and et al., "Phi-3 technical report: A highly capable language model locally on your phone," *CoRR*, vol. abs/2404.14219, 2024.

[61] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, and et al., "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *CoRR*, vol. abs/2409.12191, 2024.

[62] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, and et al., "The llama 3 herd of models," *CoRR*, vol. abs/2407.21783, 2024.

[63] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, and et al., "Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models," *CoRR*, vol. abs/2409.17146, 2024.

[64] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, and et al., "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *CoRR*, vol. abs/2412.05271, 2024.

[65] O. Contributors, "Opencompass: A universal evaluation platform for foundation models," https://github.com/open-compass/opencompass, 2023.

[66] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 19 730–19 742.

[67] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, and et al., "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *NeurIPS*, 2023.

[68] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *CoRR*, vol. abs/2305.03726, 2023.

[69] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, and et al., "Deepseek-vl: Towards real-world vision-language understanding," *CoRR*, vol. abs/2403.05525, 2024.

[70] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, and et al., “Paligemma: A versatile 3b VLM for transfer,” *CoRR*, vol. abs/2407.07726, 2024.

[71] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, and et al., “Minicpm-v: A GPT-4V level MLLM on your phone,” *CoRR*, vol. abs/2408.01800, 2024.

[72] Q. Team, “Qwen2.5-vl,” https://qwenlm.github.io/blog/qwen2.5-vl/, 2025.

[73] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J. Wen, “Evaluating object hallucination in large vision-language models,” in *EMNLP*, 2023, pp. 292–305.

A Links of MLLMs

We provide the links to the model cards of the MLLMs we evaluated in the experiments.

LLaVa-1.6 [56]

https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

Mantis [57]

https://huggingface.co/TIGER-Lab/Mantis-8B-siglip-llama3

mPLUG-Owl3 [58]

https://huggingface.co/mPLUG/mPLUG-Owl3-7B-241101

Idefics3 [59]

https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3

Phi-3.5-Vision [60]

https://huggingface.co/microsoft/Phi-3.5-vision-instruct

Qwen2-VL [61]

https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

Llama-3.2-Vision [62]

https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

Molmo [63]

https://huggingface.co/allenai/Molmo-7B-D-0924

InternVL2.5 [64]

https://huggingface.co/OpenGVLab/InternVL2_5-8B

BLIP2 [66]

https://huggingface.co/Salesforce/blip2-opt-6.7b-coco

InstructBLIP [67]

https://huggingface.co/Salesforce/instructblip-vicuna-7b

Otter [68]

https://huggingface.co/luodian/OTTER-Image-LLaMA7B-LA-InContext

DeepSeek-VL [69]

https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat

Paligemma [70]

https://huggingface.co/google/paligemma-3b-pt-448

MiniCPM [71]

https://huggingface.co/openbmb/MiniCPM-o-2_6

Qwen2.5-VL [72]

https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

B Prompts and Statement Templates

The prompts and statement templates used in the INSETS pipeline are presented in Table 5.

Table 5: Prompts and statement templates employed in the INSETS pipeline.

	Prompts and Statement Templates
#1	You are an Emotional Perception Expert. Please analyze the emotions that might be evoked by the given image. Your analysis should explore a wide range of visual attributes, such as brightness, colorfulness, depicted scenes, objects, human actions, and facial expressions. Additionally, provide detailed explanations linking these attributes to the emotions they may trigger. If applicable, discuss any potential cultural or psychological factors influencing these emotional responses.
#2	You are an Emotional Perception Expert. Your task is to extract all applicable emotions as comprehensively as possible based on the image description. Focus on distinct emotions such as happiness, sadness, fear, anger, etc. Keep the list concise, with a maximum of 10 distinct emotions.
#3	You are tasked with determining whether the word "**[word]**" describes a specific emotional state. An emotional state is a psychological condition involving feelings and reactions triggered by internal or external events. Respond with "Yes" if the word aligns with this definition, or "No" otherwise. The output format should be {"word": "response"}.
#4	You are tasked with assigning the word "**[word]**" to the most closely related emotional category from the following 115 predefined options: "**[categories]**". Consider broader semantic connections and possible emotional nuances when making your judgment. If the word cannot reasonably fit any category, respond with "not applicable". Do not create or assign new categories outside of the provided list. Do not provide any explanations or reasons for your choice. The output format should be {"word": "response"}.
#5	Briefly explain why this image might evoke "**[emotion]**" in viewers, without mentioning any other emotions.
#6	Imagine a background story for the image that would evoke a sense of "**[emotion]**" in viewers. Respond in one sentence. Do not mention the content in the image.
#7	Imagine a character who would feel "**[emotion]**" when viewing this image. Include details such as their age, gender, profession, and other relevant traits. Describe the character in one concise sentence without further explanation.
#8	Upon viewing this image, observers, despite various individual or contextual factors, are most likely to experience positive emotions.
#9	Upon viewing this image, observers, despite various individual or contextual factors, are most likely to experience negative emotions.
#10	Upon viewing this image, observers are equally likely to experience either positive or negative emotions, depending on individual or contextual factors.
#11	Therefore, the image might evoke "**[emotion]**" in viewers.
#12	In the context of: "**[context]**", the image is likely to evoke a sense of "**[emotion]**".
#13	Upon viewing the image, "**[role]**" is more inclined to feel "**[emotion1]**" compared to "**[emotion2]**".

Table 6: Evaluation results of MLLMs on INSETS-462K. (SP: Sentiment Polarity, EI: Emotion Interpretation, SC: Scene Context, PS: Perception Subjectivity)

MLLMs	#P	Accuracy				
	(B)	SP	EI	SC	PS	Total
mPLUG-Owl3 [58]	8.1	64.3	**78.1**	80.5	**78.3**	**77.4**
Phi-3.5-Vision [60]	4.1	66.7	63.7	82.2	76.6	70.3
InternVL2.5 [64]	8.1	**71.4**	69.9	**82.8**	75.0	73.5
DeepSeek-VL [69]	8.7	58.3	60.2	86.2	77.0	68.7
MiniCPM [71]	8.7	**64.3**	77.1	**86.8**	**83.6**	**79.3**
Qwen2.5-VL [72]	8.3	58.3	**81.3**	76.4	63.9	74.3

C Details of Parrott's Hierarchical Model

We present the complete emotion taxonomy of Parrott's hierarchical model in Table 7.

D Further evaluation on INSETS-462K

Despite dataset noise, obvious task-specific disparities persist: MiniCPM surpasses Qwen2.5-VL in judging scene context statements (86.8% vs. 76.4%) but trails in judging emotion interpretation statements (77.1% vs. 81.3%). This finding reinforces the need for emotion-oriented training objectives in the future MLLM developments.

Table 7: Emotion taxonomy of Parrott's hierarchical model.

Primary Emotion	Secondary Emotion	Tertiary Emotion
Love	Affection	Adoration, Fondness, Liking, Attraction, Caring, Tenderness, Compassion, Sentimentality
	Lust	Desire, Passion, Infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement, Bliss, Gaiety, Glee, Jolliness, Joviality, Joy, Delight, Enjoyment, Gladness, Happiness, Jubilation, Elation, Satisfaction, Ecstasy, Euphoria
	Zest	Enthusiasm, Zeal, Excitement, Thrill, Exhilaration
	Contentment	Pleasure
	Pride	Triumph
	Optimism	Eagerness, Hope
	Enthrallment	Enthrallment, Rapture
	Relief	Relief
Surprise	Surprise	Amazement, Astonishment
Anger	Irritability	Aggravation, Agitation, Annoyance, Grouchy, Grumpy, Crosspatch
	Exasperation	Frustration
	Rage	Anger, Outrage, Fury, Wrath, Hostility, Ferocity, Bitterness, Hatred, Scorn, Spite, Vengefulness, Dislike, Resentment
	Disgust	Revulsion, Contempt, Loathing
	Envy	Jealousy
	Torment	Torment
Sadness	Suffering	Agony, Anguish, Hurt
	Sadness	Depression, Despair, Gloom, Glumness, Unhappiness, Grief, Sorrow, Woe, Misery, Melancholy
	Disappointment	Dismay, Displeasure
	Shame	Guilt, Regret, Remorse
	Neglect	Alienation, Defeatism, Dejection, Embarrassment, Homesickness, Humiliation, Insecurity, Insult, Isolation, Loneliness, Rejection
	Sympathy	Pity, Mono no aware, Sympathy
Fear	Horror	Alarm, Shock, Fear, Fright, Horror, Terror, Panic, Hysteria, Mortification
	Nervousness	Anxiety, Suspense, Uneasiness, Apprehension, Worry, Distress, Dread

E Visualization of INSETS-3K

More samples from INSETS-3K are visualized in Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10, Fig. 11, Fig. 12.

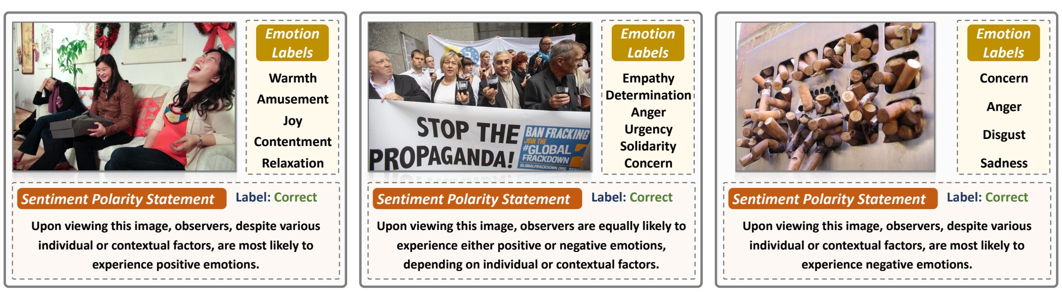

Figure 5: Correct sentiment polarity statements.

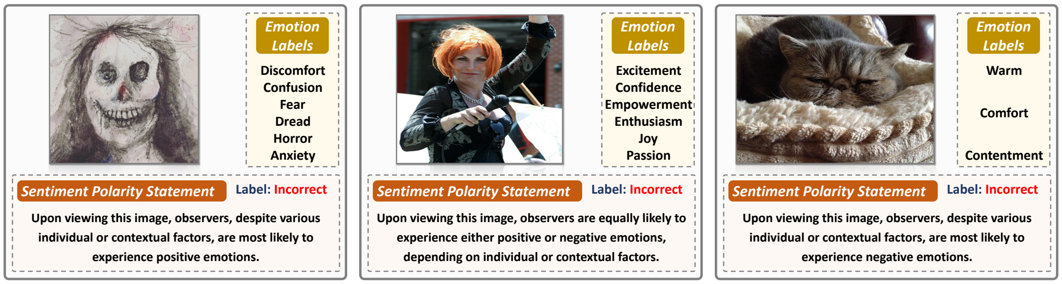

Figure 6: Incorrect sentiment polarity statements.

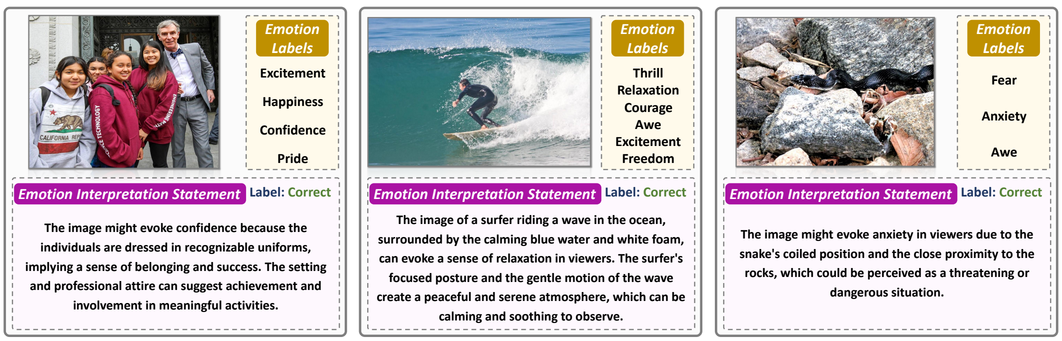

Figure 7: Correct emotion interpretation statements.

Figure 8: Incorrect emotion interpretation statements.

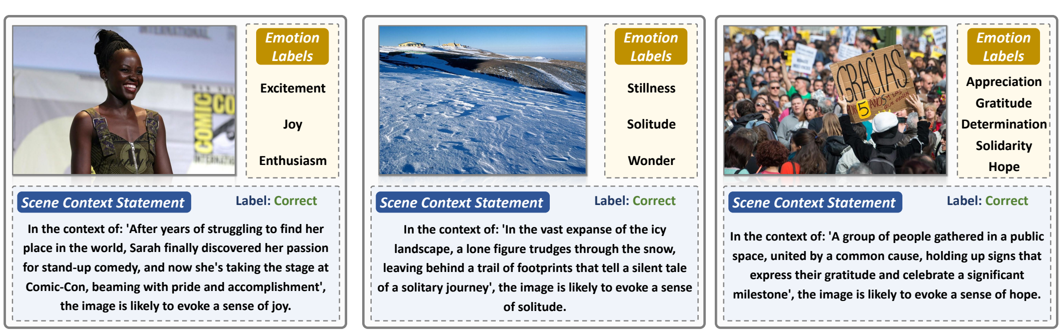

Figure 9: Correct scene context statements.

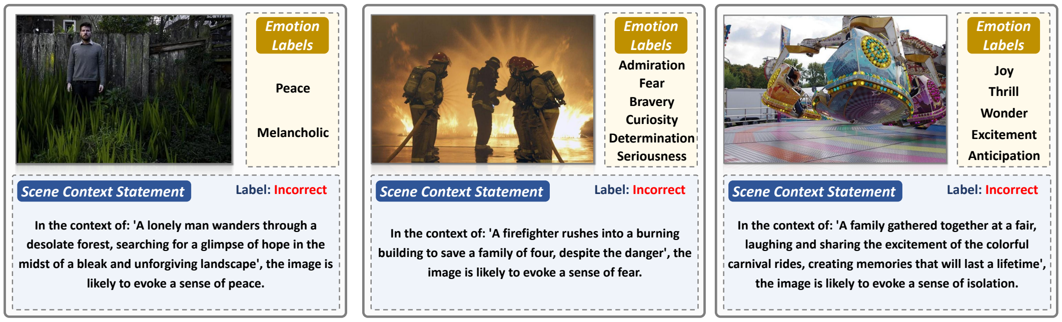

Figure 10: Incorrect scene context statements.

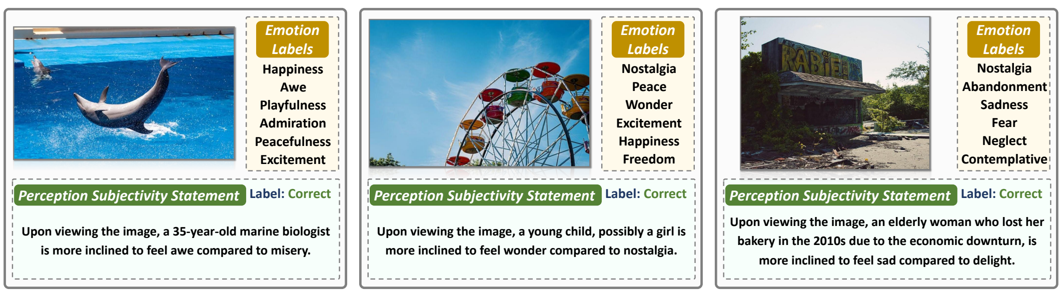

Figure 11: Correct perception subjectivity statements.

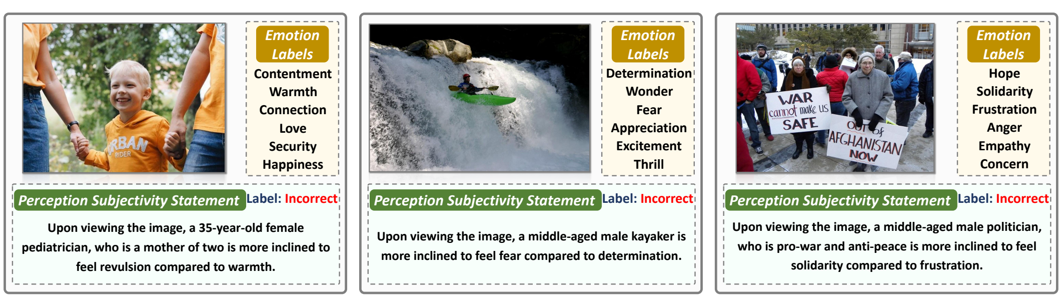

Figure 12: Incorrect perception subjectivity statements.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims include a new ESJ task for visual emotion evaluation of MLLMs and a complemented INSETS pipeline for efficient annotation. It accurately reflects the paper's contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The potential limitations are discussed before the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We present the proposed annotation framework step-by-step, and all prompts and statement templates are listed in the Appendix. We will also release data and code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: We provide INSETS-3K benchmark in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Most of the experiments are evaluations of MLLMs. We follow their default settings provided by the model card, which are listed in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Answer: [No]

Justification: We have not reported error bars. However, to ensure the reliability of results, in each evaluation trial, we queried three times per MLLM and selected the most frequent response as the final decision.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: All experiments are conducted on NVIDIA GeForce RTX 4090 GPUs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have carefully checked the NeurIPS Code of Ethics and conformed in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The main positive societal impacts of this paper include the promotion of emotional intelligence in MLLMs. Since this paper mainly focuses on MLLMs' evaluation, there are limited negative societal impacts.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: This appropriately cites the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes]

Justification: We provide details about the benchmarks we constructed. We will also provide documentation alongside the benchmarks upon releasing them.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [Yes]

Justification: In benchmark construction, human interventions are performed by the authors of the paper. In human testing, volunteers are provided the same instructions as MLLMs, which are provided in the manuscript.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

Answer: [Yes]

Justification: Study participants are asked to judge whether emotion-centric statements are accurate in relation to images. All potential risks are disclosed to the volunteers.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: This paper focuses on the visual emotion evaluation of MLLMs and adopts MLLMs as tools for efficient annotations.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.