EMOCARE: EMPOWERING VISUAL EMOTION COMPREHENSION VIA SUBJECTIVE ADAPTATION AND CANONICAL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual Emotion Comprehension (VEC) aims to infer the emotional responses that visual stimuli may evoke in human observers, which requires reasoning beyond visual content to its affective implications. Although chain-of-thought (CoT) reasoning has shown promise for complex reasoning in multimodal tasks, its application to VEC remains challenging, as existing multimodal large language models (MLLMs) often generate unreliable affective reasoning chains. In this work, we systematically investigate CoT reasoning for VEC and reveal that the major bottleneck lies in the quality of self-generated affective reasoning rather than the CoT paradigm itself. Through fine-grained error analysis, we identify four predominant reasoning failure modes and further highlight the challenge posed by subjective emotion ambiguity. To mitigate this, we introduce **VECBench-Subjective**, a benchmark with subjective emotion annotations that captures the inherent diversity of human affective judgments. Building on these findings, we propose **EmoCARE** , which leverages **C**anonical **A**ffective **R**ubric r**E**finement to enhance the quality of affective reasoning. By transforming fine-grained reasoning errors into instance-specific corrective supervision, EmoCARE enables MLLMs to produce more faithful affective reasoning and achieve improved emotion prediction. Extensive experiments on both VECBench and VECBench-Subjective demonstrate that EmoCARE consistently achieves state-of-the-art performance across multiple MLLMs, validating its effectiveness for both conventional and subjective VEC.

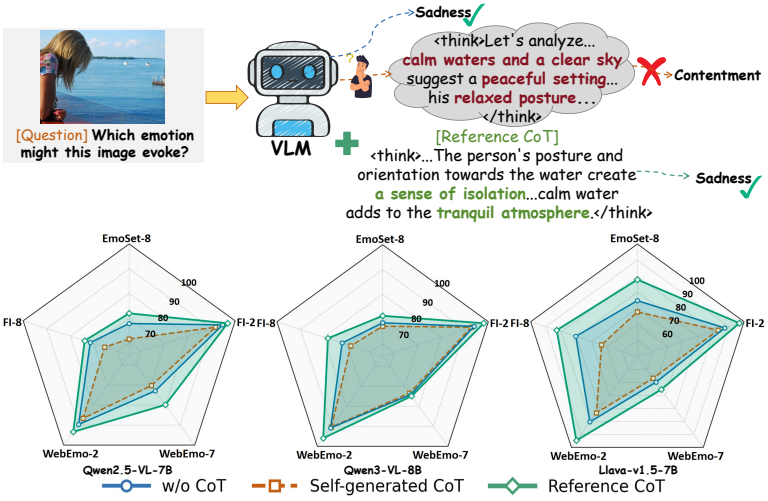

Figure 1: Accuracy of the three reasoning paradigms on five benchmarks across three MLLMs.

1 INTRODUCTION

Visual Emotion Comprehension (VEC) aims to predict viewer-side emotional responses elicited by visual stimuli Zhang et al. (2026b), which requires reasoning about subtle affective cues, contextual semantics, and human-centered implications beyond conventional object or scene recognition Zhou

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

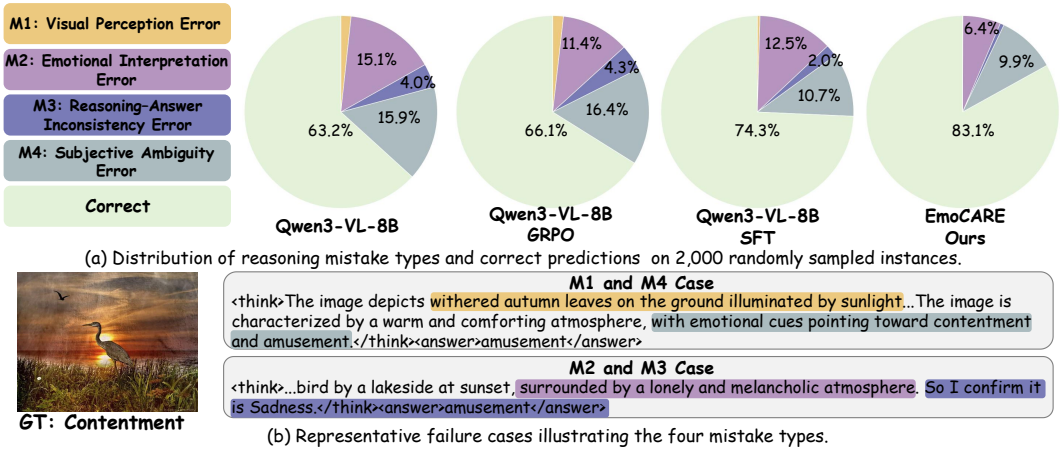

Figure 2: Fine-grained error analysis of *Self-generated CoT* reasoning.

et al. (2017); Zhang et al. (2019). Recent advances in multimodal large language models (MLLMs) have achieved impressive performance in visual perception and language understanding Yin et al. (2024). However, effectively applying these capabilities to VEC remains challenging Lian et al. (2024b), as models must reason about the underlying affective relationships between visual stimuli and subjective emotional responses.

Given the reasoning-intensive nature of VEC, chain-of-thought (CoT) provides a natural interface for exposing intermediate affective reasoning and improving prediction interpretability Lian et al. (2024a); Zhao et al. (2026). While CoT prompting has been shown to benefit complex reasoning in language and multimodal tasks Rashid et al. (2026); Zhang et al. (2026a); Guo et al. (2025b), we observe a counterintuitive phenomenon in the context of VEC: vanilla CoT prompting consistently degrades emotion prediction accuracy compared to direct reasoning, in stark contrast to its established benefits on other reasoning-intensive tasks.

To substantiate this observation, we conduct a controlled preliminary study on 3rd-stage of the VEC-CoT Wu et al. (2026a) dataset, which contains 14,346 image–label–CoT triplets covering diverse visual emotion scenarios. As illustrated in Figure 1, we conduct controlled evaluations on Qwen2.5-VL-7B, Qwen3-VL-8B, and LLaVA-v1.5-7B under different prompting settings. Specifically, we first compare direct prediction without CoT and *Self-generated CoT*, where the model is prompted to produce intermediate reasoning steps before predicting the final emotion label. Counterintuitively, *Self-generated CoT* consistently underperforms *w/o CoT* across different models, raising an important question: *Is the performance degradation caused by the CoT reasoning paradigm itself, or by the inability of current MLLMs to generate reliable affective reasoning chains?*

To answer this question, we further adopt a *Reference CoT* setting, in which the model is provided with high-quality reasoning traces from the dataset. To avoid answer leakage, we remove any reasoning segments that directly reveal the final emotion label. Details on leakage prevention and scalability analysis are provided in Appendix A.1 and Appendix A.2, respectively. The results reveal a consistent pattern: *Reference CoT* achieves the best performance, followed by *w/o CoT*, while *Self-generated CoT* performs the worst. This finding suggests that the performance degradation does not originate from the CoT paradigm itself, but rather from the limited quality of self-generated affective reasoning chains. When MLLMs are capable of generating proper and informative reasoning trajectories, CoT can effectively improve VEC performance. These observations motivate us to investigate its underlying failure modes and develop a more effective mechanism for improving affective reasoning quality.

To characterize such reasoning failures, we define four types of failure modes that arise at different stages of affective reasoning, as illustrated in Figure 2(b), which provides the representative failure cases. Specifically, MLLMs exhibit a Visual Perception Error when misinterpreting or hallucinating visual elements, and an Emotional Interpretation Error when possessing correct visual grounding but drawing flawed affective interpretation. Furthermore, a Reasoning-Answer Inconsistency Error is identified when final answers contradict reasoning trajectories. Lastly, a Subjective Ambiguity Error captures logical mispredictions purely due to the inherent subjectivity of VEC.

Among these failure modes, we identify two fundamental challenges for reliable VEC reasoning. First, Subjective Ambiguity Error arises from the inherent subjectivity of human emotional responses, where identical visual stimuli can legitimately evoke diverse interpretations depending on personal and cultural contexts Masuda et al. (2008); Zhao et al. (2018). Consequently, rigid single-label evaluations may unfairly penalize valid alternative interpretations. To mitigate this limitation, we propose **VECBench-Subjective**, which constructs multi-candidate emotion annotations via MLLM-based candidate generation, hit-and-vote filtering, and human verification. This enables fairer subjectivity-aware evaluation with validated multi-label emotional targets.

Second, Emotional Interpretation Error and the remaining reasoning failures still represent major bottlenecks in VEC. Their diverse and instance-dependent patterns make it challenging to provide effective corrective supervision with a unified criterion. Fine-grained adaptive rubrics provide a potential solution by enabling failure-aware guidance tailored to individual reasoning trajectories. Recent advances have established rubrics as a promising reward signal for reinforcement learning (RL) beyond verifiable tasks Gunjal et al. (2026); Shao et al. (2026); Li et al. (2026). However, generic rubrics lack the granularity required to capture the diverse failure modes in VEC. To bridge this gap, we propose **EmoCARE** , which leverages **C**anonical **A**ffective **R**ubric r**E**finement to enhance the quality of affective reasoning through instance-specific supervision. Concretely, MLLMs first propose a pool of candidate rubrics for each error type. We then align these rubrics with real model failures by collecting reasoning trajectories from multiple MLLMs, classifying their errors with an external judge model, and selecting the most representative rubrics. The affective discriminative rubrics are incorporated into RL as instance-specific corrective feedback. As shown in Figure 2, EmoCARE substantially reduces all four categories of reasoning mistakes, highlighting the effectiveness of fine-grained affective rubrics for reliable VEC. Our main contributions are as follows:

- We introduce **VECBench-Subjective**, a subjectivity-aware benchmark that extends conventional single-label annotations into validated candidate emotion sets.
- We propose **EmoCARE** , which introduces **C**anonical **A**ffective **R**ubric r**E**finement that converts fine-grained errors into instance-specific rubric supervision to train models for more reliable affective reasoning.
- Experiments demonstrate that EmoCARE consistently achieves leading performance under both conventional and subjectivity-aware VEC settings, validating the effectiveness of canonical affective rubric refinement.

2 Related Work

2.1 Visual Emotion Comprehension

The methodological landscape of VEC has progressively shifted from conventional feature-driven approaches Poria et al. (2015); Zhao et al. (2021) to data-driven foundation models Xie et al. (2024). Through large-scale multimodal pretraining, vision-language models (VLMs) learn transferable cross-modal representations that enable the unified modeling of heterogeneous visual and textual inputs Li et al. (2023). This capability makes VLMs particularly well suited for VEC, where emotion understanding often requires jointly interpreting visual affective cues and their semantic contexts. Building on such transferable multimodal knowledge, recent studies have explored adaptation strategies to enhance emotion perception in VEC Wu et al. (2025); Fang et al. (2025), while subsequent training-oriented efforts further optimize VLMs for task-specific affective reasoning Wu et al. (2026b); Gao et al. (2026). Nevertheless, their reliance on predefined reward rules restricts generalization to the diverse, instance-specific demands of affective reasoning.

2.2 Emotion Reasoning Optimization

Recent efforts have moved beyond conventional emotion classification to improve the reasoning behavior of multimodal affective models. In open-vocabulary multimodal emotion recognition Lian et al. (2025b;a), AffectGPT-R1 Lian et al. (2025c) formulates emotion-wheel-based evaluation as a reward signal and applies RL to better align model optimization with fine-grained affective metrics. For emotion reasoning under modality conflicts, CA-MER Han et al. (2025) introduces a conflict-aware benchmark, while MoSEAR mitigates the audio-dominance bias of existing MLLMs

through modality-specific experts and attention reallocation. Complementarily, EmoCaliber Wu et al. (2026a) incorporates a confidence dimension into affective prediction, enabling models to self-assess and calibrate their outputs under subjective or ambiguous scenarios.

Despite their effectiveness, these methods mainly optimize emotion reasoning at the model or task level. Such unified optimization signals are less capable of capturing instance-specific reasoning requirements, as different samples may involve distinct challenges such as modality inconsistency, subtle affective distinctions, or subjective uncertainty. To this end, we propose EmoCARE , which introduces canonical affective rubric refinement to derive instance-specific supervision from fine-grained reasoning errors. Results confirm that EmoCARE establishes robust performance on both conventional and subjectivity-aware datasets.

Dataset	Number of Labels				Samples
	1	2	3	≥4	
EmoSet-8	467	439	94	0	1000
FI-8	411	447	132	10	1000
WebEmo-7	579	370	51	0	1000
WebEmo-25	635	112	150	103	1000
Abstract-8	132	86	11	0	229
ArtPhoto-8	449	325	32	0	806
UnbiasedEmo-6	426	535	39	0	1000

Table 1: Statistics of the annotated subjective emotion labels.

3 VECBENCH-SUBJECTIVE CONSTRUCTION

To mitigate the intrinsic subjectivity of VEC, we construct **VECBench-Subjective**, extending conventional single-label annotations to validated subjective emotion sets. As illustrated in Figure 3, the annotation pipeline consists of two stages: model voting and human verification.

Model Voting. Starting from the image-only samples in VECBench Wu et al. (2026a), we employ multiple advanced MLLMs (*e.g.*, Qwen3.6-27B, InternVL3.5-38B, Gemma4-31B-it, GLM-4.5V) to independently predict the set of emotions each image may evoke. To ensure reliability, we adopt a hit-and-vote strategy: a prediction is retained only when its candidate set contains the original ground-truth label, filtering out models that fail to capture the reference emotion. Among the retained predictions, emotions receiving votes from at least two models are included in the preliminary subjective label set. This process preserves diverse affective responses while leveraging cross-model consensus.

Preference	**Average Count**	**Ratio (%)**
Previous Label	18.6	9.3
Subjective Label	175.8	87.9
Not Sure	5.6	2.8

Table 2: Human preference study on 200 VECBench-Subjective samples with five volunteers. The study compares previous annotations with our subjective annotations.

Human Verification. The generated label sets are subsequently verified by three annotators, who independently review each candidate set and remove or add labels according to their affective judgments. The final annotations are obtained from the verified results, ensuring that each sample captures a valid range of human-perceived emotional responses. Therefore, VECBench-Subjective provides a more comprehensive evaluation protocol for subjective emotion understanding and reduces the bias from ambiguous single-label supervision.

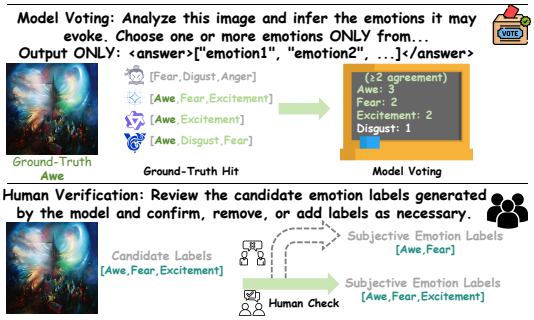

Figure 3: The subjective emotion annotation pipeline. Candidate labels are first generated through multi-MLLM voting and then refined via human verification to obtain the final subjective emotion annotations.

Annotation Analysis. Table 1 summarizes the distribution of subjective emotion annotations. Across all datasets, a substantial proportion of images are associated with multiple valid emotion labels. For instance, over half of the samples in FI-8 and UnbiasedEmo-6 receive at least two emotion labels, while WebEmo-25 exhibits an even higher degree of ambiguity, with 25.3% of images containing three or more labels. These observations confirm that emotional responses to visual content are inherently one-to-many, suggesting that fixed-label annotations cannot fully capture the diversity of human emotional perception.

ID VER	EmoSet-8		FI-8		WebEmo-7		WebEmo-25		Average	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Open-Source MLLM										
Qwen2.5-VL-7B Bai et al. (2025b)	48.70	47.54	49.70	48.14	43.20	37.59	18.30	15.59	39.97	37.21
Qwen2.5-VL-72B Bai et al. (2025b)	53.30	52.42	53.60	50.90	46.70	40.03	17.60	18.36	42.08	40.42
Qwen3-VL-8B Bai et al. (2025a)	53.00	51.34	57.00	55.72	48.80	41.37	21.40	18.77	45.05	41.80
Qwen3-VL-32B Bai et al. (2025a)	45.90	48.05	46.80	52.30	46.80	42.20	23.30	19.24	40.70	40.44
InternVL3.5-8B Wang et al. (2025)	50.80	50.86	51.00	51.72	32.70	32.60	13.20	13.95	36.92	37.28
InternVL3.5-38B Wang et al. (2025)	51.90	52.28	45.80	47.29	43.40	39.32	24.60	19.89	41.25	39.69
GLM-4.6v-Flash Team et al. (2025)	50.70	51.26	49.40	50.80	45.70	39.39	19.10	16.76	41.22	39.55
Proprietary MLLM										
Qwen3.6-Plus	57.00	56.94	57.20	57.44	48.60	43.53	25.40	21.36	47.05	44.81
GPT-5 Singh et al. (2026)	56.90	55.40	54.90	55.36	42.70	33.84	20.10	17.02	43.65	40.40
Emotion-Oriented MLLM										
EmoVIT Xie et al. (2024)	58.90	62.32	58.60	59.23	41.20	35.67	15.80	15.85	43.62	43.26
Emotion-LLaMA Cheng et al. (2024)	18.70	13.26	23.50	18.97	18.70	20.11	10.30	10.04	17.80	15.60
EEmo-Logic Gao et al. (2026)	51.50	51.18	51.60	50.82	47.90	34.01	19.00	12.58	42.50	37.15
EmoCaliber Wu et al. (2026a)	<u>67.40</u>	<u>67.75</u>	<u>67.40</u>	<u>66.39</u>	<u>53.10</u>	<u>44.65</u>	**28.00**	<u>21.93</u>	<u>53.98</u>	<u>50.18</u>
EmoCARE (**Ours**)	**70.10**	**71.59**	**70.70**	**70.00**	**53.70**	**46.07**	<u>27.00</u>	**22.74**	**55.38**	**52.60**

Table 3: Comparison of advanced MLLMs on emotion prediction on ID VER tasks of VECBench. **Bold** and <u>underline</u> indicate the best and second-best results, respectively.

Furthermore, we conduct a human preference study on 200 randomly sampled images to assess annotation quality, with results presented in Table 2. Five volunteers independently compared the previous and subjective labels, with a "Not Sure" option for ambiguous cases, to validate the effectiveness of our VECBench-Subjective annotations (see Appendix B.2 for details). Subjective annotations are preferred in 87.9% of comparisons, significantly outperforming the original single-label annotations with only 9.3% preference. This confirms that VECBench-Subjective provides more comprehensive annotations that better align with human perceptual judgments.

4 METHOD

4.1 PROBLEM FORMULATION

Let x denote an input prompt and $\hat{y} \sim \pi_\theta(\cdot \mid x)$ a response sampled from the policy parameterized by θ. Since affective reasoning trajectories lack deterministic solutions and reliable automatic evaluation, we define a structured reward function based on instance-specific rubric criteria. Specifically, each prompt x is associated with a set of k rubric criteria $(w_j, c_j)_{j=1}^{k}$, where $w_j \in \mathbb{R}$ denotes the weight of the j-th criterion and $c_j : (x, \hat{y}) \mapsto 0, 1$ is a binary indicator that evaluates whether $\hat{y}$ satisfies the corresponding criterion.

4.2 CANONICAL AFFECTIVE RUBRIC REFINEMENT

Building on the identified error taxonomy, we observe that the four recurring failure modes reveal distinct deficiencies in model's affective reasoning. Existing reward signals, however, typically provide coarse correctness feedback and fail to capture fine-grained instance-specific reasoning flaws. To address this issue, we propose EmoCARE , which converts error-specific affective criteria into instance-level supervision through canonical affective rubric refinement, as illustrated in Figure 4.

Candidate rubrics construction. We first establish a hierarchical rubric library with the assistance of several advanced MLLMs, including Qwen3.6-27B, Gemma-31B-IT, InternVL3.5-38B, and GLM-4.5V. In this stage, each model acts as a *Rubric Designer*, formulating interpretable evaluation criteria for a specific affective reasoning error category. Given an input image, a reference reasoning chain, the ground-truth emotion label, and a predefined error type from our taxonomy, Rubric Designers are instructed to generate 7–8 atomic rubrics. Each rubric captures a minimal yet semantically meaningful criterion that characterizes a specific reasoning deficiency. Collectively, these rubrics form a candidate pool for subsequent refinement.

Diagnostic failure generation. To capture realistic and diverse reasoning failures, we leverage a diverse set of MLLMs, including Qwen3.6-27B, Gemma4-31B, InternVL3.5-8B and 38B, and LLaVA-v1.5-7B, as *Response Generators* to generate diverse diagnostic reasoning trajectories.

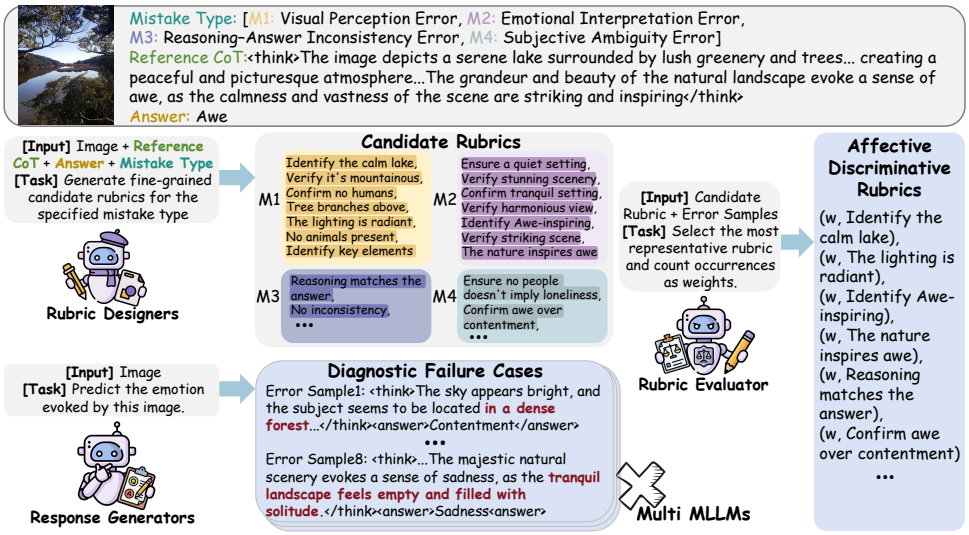

Figure 4: Overview of Canonical Affective Rubric rEfinement (CARE). CARE refines instance-specific, discriminative affective rubrics by aligning MLLM-generated candidate rubrics with empirically observed reasoning failures.

Specifically, under identical task inputs, each Response Generator independently produces multiple reasoning responses, yielding a diverse set of model-generated trajectories. These responses naturally expose heterogeneous failure patterns arising from different stages of affective reasoning, providing realistic diagnostic cases.

Rubric selection and weighting. Given the diagnostic failure cases, we employ GPT-5 as a dedicated *Rubric Evaluator* to analyze each reasoning trajectory. Rubric Evaluator first identifies the underlying error category of each trajectory and subsequently aligns the identified error with the corresponding candidate rubric pool. For each failure case, the evaluator selects one or two atomic rubrics that most precisely explain the observed reasoning deficiency. Aggregating these selections across the training corpus, we calculate the occurrence frequency of each rubric and assign an importance weight w_j. The resulting weighted rubrics are then consolidated into an affective discriminative rubric set, which serves as fine-grained reward guidance in the RL framework.

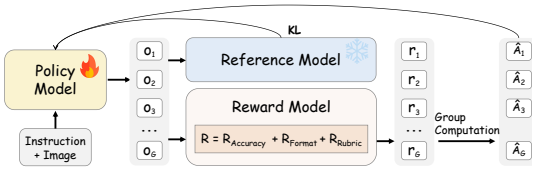

Figure 5: Illustration of the training pipeline of EmoCARE under the rubric-guided GRPO framework.

Different from conventional error-dependent supervision that requires explicit error prediction, the refined rubric set provides a unified and interpretable feedback mechanism that directly captures instance-level affective reasoning deficiencies. This enables RL to optimize not only answer correctness but also the underlying reasoning quality of MLLMs in VEC.

4.3 Rubric-Guided Training of EmoCARE

Initializing Emotion Reasoning via SFT. We first perform Supervised Fine-Tuning (SFT) Zhang et al. (2024) to initialize the policy model π_θ before RL. Given the SFT dataset $\mathcal{D}_{\text{SFT}}$, the model is trained to generate structured emotion reasoning responses. The training objective is the standard negative log-likelihood loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}}\left[\sum_{t=1}^{|y|}\log \pi_\theta(y_t|x, y_{<t})\right], \tag{1}$$

where x and y denote the input and target response, respectively.

RL with Canonical Affective Rubric as Rewards. Given a response $o \sim \pi_\theta(\cdot \mid x)$, we define a composite reward jointly supervising the final prediction, output format, and intermediate reasoning quality. The total reward comprises three complementary components:

Model	ID VSA						OOD VER							
	FI-2		WebEmo-2		Average		Abstract-8		Artphoto-8		UnbiasedEmo-6		Average	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Open-Source MLLM														
Qwen2.5-VL-7B Bai et al. (2025b)	83.20	81.17	71.20	72.08	77.20	77.62	23.58	24.17	41.19	38.66	85.60	81.42	50.12	48.08
Qwen2.5-VL-72B Bai et al. (2025b)	86.80	84.92	72.10	75.16	79.45	80.04	34.49	30.95	43.30	40.72	82.80	76.62	53.53	49.43
Qwen3-VL-8B Bai et al. (2025a)	87.40	85.33	75.60	75.53	81.50	80.43	20.58	22.17	32.10	31.14	72.10	71.32	41.59	41.54
Qwen3-VL-32B Bai et al. (2025a)	80.90	56.31	65.10	70.53	73.00	63.42	20.96	22.42	28.78	32.32	62.60	68.83	37.45	41.19
InternVL3.5-8B Wang et al. (2025)	76.50	77.91	57.00	62.55	66.75	70.23	23.58	26.79	38.70	40.89	67.60	66.36	43.29	44.68
InternVL3.5-38B Wang et al. (2025)	64.20	72.26	73.70	75.89	68.95	74.08	19.65	23.66	43.54	43.07	79.80	77.04	47.66	47.92
GLM-4.6v-Flash Team et al. (2025)	72.60	78.04	71.10	72.84	71.85	75.44	21.83	22.67	43.42	42.83	74.50	76.68	46.58	47.39
Proprietary MLLM														
Qwen3.6-Plus	87.90	86.82	75.10	75.98	81.50	81.40	33.18	30.44	**45.78**	44.81	**87.30**	**84.81**	**55.42**	53.35
GPT-5 Singh et al. (2026)	87.40	85.83	71.50	71.46	79.45	78.64	**38.86**	**35.22**	44.91	42.56	65.60	58.12	49.79	45.30
Emotion-Oriented MLLM														
EmoVIT Xie et al. (2024)	75.80	79.11	63.90	67.77	69.85	73.44	24.89	23.61	33.12	37.71	79.60	79.15	45.87	46.82
Emotion-LLaMA Cheng et al. (2024)	56.40	63.15	44.90	48.21	50.65	55.68	10.48	7.78	22.95	16.46	40.50	46.41	24.64	23.55
EEmo-Logic Gao et al. (2026)	84.70	85.44	74.80	77.08	79.75	81.26	31.44	25.30	39.08	37.57	59.20	50.75	43.24	37.87
EmoCaliber Wu et al. (2026a)	87.60	85.91	77.90	77.88	82.75	81.89	26.20	23.45	36.47	38.45	80.40	76.68	47.69	46.19
EmoCARE (**Ours**)	**89.30**	**87.73**	**80.70**	**80.68**	**85.00**	**84.21**	34.93	31.82	44.16	**44.83**	86.30	83.65	55.13	**53.43**

Table 4: Comparison of MLLMs on emotion prediction across ID VSA and OOD VER tasks of VECBench.

(i) Accuracy reward. We measure the correctness of the predicted emotion label and use it as the accuracy reward:

$$R_{\text{acc}}(x, o) = \begin{cases} 1, & \text{if } o \text{ matches the correct label,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

(ii) Format reward. Following Guo et al. (2025a), we require responses to follow the `<think></think><answer></answer>` template:

$$R_{\text{fmt}}(x, o) = \begin{cases} 1, & \text{if } o \text{ conforms to the format,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

(iii) Rubric reward. To provide fine-grained instance-specific supervision over the intermediate CoT, we leverage the canonical affective rubric refinement set $\{(w_j, c_j)\}_{j=1}^{k}$ associated with prompt x. The final normalized rubric reward is computed as a weighted sum of independently LLM-as-judge evaluated criteria:

$$R_{\text{rub}}(x, o) = \frac{\sum_{j=1}^{k} w_j \, c_j(x, o)}{\sum_{j=1}^{k} w_j}. \tag{4}$$

The overall reward integrates the three components:

$$R(x, o) = R_{\text{acc}} + R_{\text{fmt}} + \lambda_{\text{rub}} R_{\text{rub}}, \tag{5}$$

where λ_{rub} is a hyperparameter balancing the contribution of the rubric-based reward.

We optimize π_θ using Group Relative Policy Optimization (GRPO) Shao et al. (2024), which estimates advantages from a group of sampled responses, eliminating the need for a separate value network. As illustrated in Figure 5, for each prompt x (i.e., instruction and image), the policy π_θ samples a group of G responses $\{o_i\}_{i=1}^{G}$ and computes their rewards $\{r_i\}_{i=1}^{G}$ via Eq. equation 5. The group-normalized advantage for the i-th response is:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})}. \tag{6}$$

The policy is updated by maximizing the clipped surrogate objective with KL regularization:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{o_i\}_{i=1}^{G}} \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(\rho_i(\theta) \, \hat{A}_i, \; \text{clip}\big(\rho_i(\theta), \quad 1-\epsilon, \, 1+\epsilon\big) \, \hat{A}_i \right) - \beta \, \mathbb{D}_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}] \right], \tag{7}$$

where $\rho_i(\theta) = \dfrac{\pi_\theta(o_i \mid x)}{\pi_{\theta_{\text{old}}}(o_i \mid x)}$ is the importance sampling ratio, ϵ controls the clipping range, π_{ref} is the frozen reference model, and β weights the KL penalty preventing excessive policy drift.

Model	ID				OOD			Mean
	EmoSet-8	FI-8	WebEmo-7	WebEmo-25	Abstract-8	Artphoto-8	UnbiasedEmo-6	
Open-Source MLLM								
Qwen3-VL-8B (Δ vs. VECBench)	69.50 (+16.50)	75.50 (+18.50)	62.50 (+13.70)	37.70 (+16.30)	43.72 (+23.14)	49.17 (+17.07)	82.70 (+10.60)	60.11 (+16.54)
InternVL3.5-8B (Δ vs. VECBench)	65.70 (+14.90)	68.00 (+17.00)	43.10 (+10.40)	22.90 (+9.70)	34.93 (+11.35)	51.98 (+13.28)	76.00 (+8.40)	51.80 (+12.15)
Proprietary MLLM								
Qwen3.6-Plus (Δ vs. VECBench)	72.80 (+15.80)	76.80 (+19.60)	62.30 (+13.70)	38.30 (+12.90)	**52.40** (+19.22)	**63.15** (+17.37)	93.00 (+5.70)	65.54 (+14.90)
GPT-5 (Δ vs. VECBench)	72.50 (+15.60)	74.10 (+19.20)	57.30 (+14.60)	34.10 (+14.00)	**52.40** (+13.54)	62.03 (+17.12)	77.90 (+12.30)	61.48 (+15.20)
Emotion-Oriented MLLM								
EEmo-Logic (Δ vs. VECBench)	71.70 (+20.20)	74.20 (+22.60)	60.90 (+13.00)	32.20 (+13.20)	45.03 (+13.59)	55.58 (+16.50)	83.40 (+24.20)	60.43 (+17.61)
EmoCaliber (Δ vs. VECBench)	78.90 (+11.50)	79.50 (+12.10)	63.30 (+10.20)	38.30 (+10.30)	35.80 (+9.60)	51.61 (+15.14)	90.70 (+10.30)	62.59 (+11.31)
Ours (Δ vs. VECBench)	**80.80** (+10.70)	**83.00** (+12.30)	**65.80** (+12.10)	**40.10** (+13.10)	46.28 (+11.35)	61.66 (+17.50)	**93.90** (+7.60)	**67.36** (+12.09)

Table 5: Performance comparison of different MLLMs on the VECBench-Subjective. The reported scores are hit score (%).

Model	EmoSet-8	FI-2	FI-8	WebEmo-2	WebEmo-7	WebEmo-25	Abstract-8	Artphoto-8	UnbiasedEmo-6	Mean
Qwen3-VL-8B	53.00	87.40	57.00	75.60	48.80	21.40	20.58	32.10	72.10	52.00
+SFT	65.80	87.60	67.00	78.20	50.20	25.40	29.49	43.67	80.80	58.68
+GRPO	54.60	88.40	57.50	78.70	50.30	23.00	31.87	44.91	85.90	57.24
+SFT+GRPO	64.30	86.20	66.10	76.10	49.20	24.10	30.14	44.23	79.50	57.76
+Full Rubric	69.70	86.90	67.40	77.00	50.40	24.10	28.38	**45.11**	83.40	59.15
+Part Rubric	69.20	88.90	68.90	79.60	51.20	26.50	27.21	43.65	84.60	59.97
EmoCARE	**70.10**	**89.30**	**70.70**	**80.70**	**53.70**	**27.00**	**34.94**	44.16	**86.30**	**61.88**

Table 6: Ablation study of different training strategies and rubric construction designs on VECBench. Performance is measured in terms of Accuracy (%).

5 Experimental Setup

Datasets. We evaluate on VECBench Wu et al. (2026a), which comprises six representative VEC datasets. The in-domain (ID) setting includes FI You et al. (2016), WebEmo Panda et al. (2018), and EmoSet Yang et al. (2023). FI is divided into FI-8 for visual emotion recognition (VER) and FI-2 for visual sentiment analysis (VSA), WebEmo into WebEmo-25, WebEmo-7, and WebEmo-2, and EmoSet into EmoSet-8. Each test subset contains 1,000 samples. The out-of-domain (OOD) datasets consist of Abstract-8 Machajdik & Hanbury (2010), ArtPhoto-8 Machajdik & Hanbury (2010), and UnbiasedEmo-6 Panda et al. (2018), containing 229, 806, and 1,000 test samples, respectively.

For training, we use the second-stage subset of VEC-CoT Wu et al. (2026a), denoted as VEC-CoT-2, which contains 43,033 instances and provides a large, disjoint corpus for optimization. The third-stage subset, VEC-CoT-3, is excluded from training and reserved for controlled preliminary analysis and protocol verification, thereby preventing data leakage and ensuring an independent basis for methodological assessment. VEC-CoT-2 is further split into SFT and RL subsets with a 9:1 ratio.

Implementation Details. We use Qwen3-VL-8B-Instruct as the backbone model. The model is trained with 5 epochs of SFT, followed by 1 epoch of GRPO optimization. Both training stages employ LoRA with rank $r = 16$, scaling factor $\alpha = 32$, and dropout rate of 0.1, together with BF16 mixed-precision training and a cosine learning rate scheduler. The learning rates for the SFT and GRPO stages are set to 6×10^{-5} and 5×10^{-5}, respectively. All experiments are conducted on 4 NVIDIA H20 GPUs. During GRPO training, we sample $G = 12$ responses for each input to estimate the policy optimization objective. For rubric-based LLM-as-Judge scoring in GRPO, we deploy Qwen3-32B Team (2025) with vLLM as the judge model to evaluate the generated responses.

Baselines. We compare our method with representative MLLMs on VECBench, covering three categories of baselines. The first category includes general-purpose MLLMs, such as Qwen2.5-VL-7B/72B Bai et al. (2025b), Qwen3-VL-8B/32B Bai et al. (2025a), InternVL3.5-8B/38B Wang et al. (2025), and GLM-4.6V-Flash Team et al. (2025). The second category consists of proprietary MLLMs, including Qwen3.6-Plus and GPT-5 Singh et al. (2026). The third category comprises emotion-oriented MLLMs, including EmoVIT Xie et al. (2024), Emotion-LLaMA Cheng et al. (2024), EmoCaliber Wu et al. (2026a), and EEmo-Logic Gao et al. (2026).

Comparison on VECBench. As shown in Tables 3 and 4, EmoCARE achieves state-of-the-art results on all ID tasks. For VER, it attains an average accuracy of 55.38% and F1 of 52.60%. For VSA, EmoCARE further surpasses EmoCaliber by 2.25% in accuracy and 2.32% in F1, reflecting consistent gains across both fine-grained and binary emotion classification. On OOD benchmarks, EmoCARE achieves an average accuracy of 55.13% and F1 of 53.43%, outperforming all emotion-oriented baselines by a substantial margin.

Comparison on VECBench-Subjective. Since VECBench-Subjective is designed to alleviate the subjective ambiguity introduced by single-label annotation, we further evaluate all models on this

Model	EmoSet-8	FI-8	WebEmo-7	WebEmo-25	Abstract-8	Artphoto-8	UnbiasedEmo-6	Mean
Qwen3-VL-8B	69.50	75.50	62.50	37.70	43.72	49.17	82.70	60.11
+SFT	78.60	80.70	63.90	36.70	41.54	59.68	92.10	65.46
+GRPO	71.40	75.80	63.30	38.40	42.03	60.04	91.10	64.03
+SFT+GRPO	77.40	78.80	62.10	36.70	**46.72**	60.04	91.70	64.78
+Full Rubric	78.90	77.60	64.10	37.90	43.72	60.17	90.40	64.68
+Part Rubric	79.20	80.90	63.20	38.50	45.41	**62.31**	92.60	66.02
EmoCARE	**80.80**	**83.00**	**65.80**	**40.10**	46.28	61.66	**93.90**	**67.36**

Table 7: Ablation study of different training strategies and rubric construction designs on VECBench-Subjective, with performance measured by Hit Score.

benchmark using a hit-based metric that accounts for multiple human-validated emotional interpretations. Specifically, a prediction is considered correct if it matches any emotion in the annotated subjective emotion set:

$$\text{HIT} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big[y_i \in \mathcal{S}_i\big], \tag{8}$$

where $\mathbb{I}[\cdot]$ is the indicator function, y_i is the prediction for the i-th sample, $\mathcal{S}_i$ is its subjective emotion set, and N denotes the number of samples.

As shown in Table 5, all evaluated MLLMs perform consistently better on VECBench-Subjective than on the original VECBench, with average hit-rate gains of 11.31%–17.61% across both ID and OOD datasets. This finding suggests that many errors on the original benchmark are attributable to annotation ambiguity rather than limited model capability. By incorporating diverse yet human-validated emotion sets, VECBench-Subjective reduces the bias of single-label evaluation and provides a more reliable benchmark for subjective visual emotion understanding. EmoCARE achieves the best overall result, reaching an average hit rate of 67.36%.

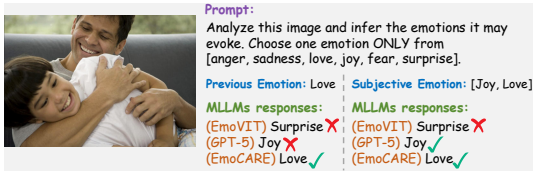

Figure 6: Qualitative examples illustrating the effect of VECBench-Subjective in alleviating subjective ambiguity.

Figure 6 provides consistent qualitative evidence: predictions judged incorrect under the original VECBench can become valid under subjective emotion sets. This observation highlights the limitation of deterministic annotations and further demonstrates the advantage of VECBench-Subjective in alleviating subjective ambiguity. More detailed qualitative case studies are presented in Appendix D, where we provide fine-grained analyses of EmoCARE's affective reasoning process and illustrate its advantages over competing models in resolving nuanced and subjective emotional cues.

Analysis of Rubric Construction. Tables 6 and 7 compare different training strategies and rubric construction schemes on VECBench and VECBench-Subjective. *Full Rubric* uses all handcrafted rubrics from the candidate construction stage, whereas *Part Rubric* retains only the rubrics corresponding to error types observed during RL. Compared with the base Qwen3-VL-8B model, SFT and GRPO generally improve performance, while their combination yields no consistent additional gains. Rubric-guided supervision further enhances results, with *Part Rubric* outperforming *Full Rubric* on both benchmarks, achieving 59.97% mean Accuracy on VECBench and 66.02% mean Hit Rate on VECBench-Subjective, respectively. EmoCARE achieves the best overall performance, reaching 61.88% mean Accuracy and 67.36% mean Hit Rate. These results indicate that removing redundant rubrics while preserving error-relevant feedback provides more precise optimization signals, thereby improving visual emotion comprehension.

6 Conclusion

In this work, we systematically analyzed the role of CoT reasoning in MLLMs for VEC and revealed its key limitations on this task. To alleviate emotion subjectivity, we introduced **VECBench-Subjective**, a benchmark with subjective emotion annotations. To improve reasoning quality, we proposed **EmoCARE** , which leverages **C**anonical **A**ffective **R**ubric r**E**finement to enhance affective reasoning. By advancing reliable reasoning for image-evoked emotion understanding, this work lays a solid foundation for developing empathic MLLMs.

AI USE STATEMENT

LLMs were used solely for language polishing of the manuscript.

ETHICS STATEMENT

This work studies VEC, where emotion annotations may reflect subjective, cultural, and demographic variations; therefore, VECBench-Subjective is intended to characterize such ambiguity rather than define a single universal emotional ground truth. All human annotations are used only for research purposes, and no personally identifiable information is collected or released beyond the visual data already contained in the evaluated benchmarks. We acknowledge that affective computing systems may inherit biases from data and annotators, and we encourage cautious use of our benchmark and method in real-world applications involving sensitive human emotions.

REPRODUCIBILITY STATEMENT

Our work is based on open-source models and datasets. In Section 5 and Appendix C, we provide detailed descriptions of the prompt templates, method implementation, and experimental setups.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems*, 2024.

Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. Catch your emotion: Sharpening emotion perception in multimodal large language models. In *International Conference on Machine Learning*, 2025.

Lancheng Gao, Ziheng Jia, Zixuan Xing, Wei Sun, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Eemo-logic: A unified dataset and multi-stage framework for comprehensive image-evoked emotion assessment. *International Conference on Machine Learning*, 2026.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *The Fourteenth International Conference on Learning Representations*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Haiyun Guo, ZhiYan Hou, Yandu Sun, Jinghan He, Yu Chen, Yuzhe Zhou, Yuheng Jia, Jinqiao Wang, and Tat-Seng Chua. Mllm-ctbench: A benchmark for continual instruction tuning with reasoning process diagnosis. *arXiv preprint arXiv:2508.08275*, 2025b.

Zhiyuan Han, Beier Zhu, Yanlong Xu, et al. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In *ACM International Conference on Multimedia*, 2025.

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

Sunzhu Li, Jiale Zhao, Huimin Ren, et al. RubricHub: A comprehensive and highly discriminative rubric dataset via automated coarse-to-fine generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, et al. Explainable multimodal emotion recognition. arXiv preprint arXiv:2306.15401, 2024a.

Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367, 2024b.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, et al. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *International Conference on Machine Learning*, 2025a.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, et al. OV-MER: Towards open-vocabulary multimodal emotion recognition. In *Forty-second International Conference on Machine Learning*, 2025b.

Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, et al. Affectgpt-r1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. arXiv preprint arXiv:2508.01318, 2025c.

Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM international conference on Multimedia*, 2010.

Takahiko Masuda, Phoebe C Ellsworth, Batja Mesquita, Janxin Leu, Shigehito Tanida, and Ellen Van de Veerdonk. Placing the face in context: cultural differences in the perception of facial emotion. *Journal of personality and social psychology*, 2008.

Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision*, 2018.

Soujanya Poria, Erik Cambria, Amir Hussain, and Guang-Bin Huang. Towards an intelligent framework for multimodal affective data analysis. *Neural Network*, 2015.

Md Rafi Ur Rashid, Vishnu Asutosh Dasu, Ye Wang, Gang Tan, and Shagufta Mehnaz. Chain-of-thought driven adversarial scenario extrapolation for robust language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, et al. DR tulu: Reinforcement learning with evolving rubrics for deep research. In *International Conference on Machine Learning*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv:2402.03300*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2026.

Qwen Team. Qwen3 technical report. 2025.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, et al. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Daiqing Wu, Dongbao Yang, Sicheng Zhao, Can Ma, and Yu Zhou. An empirical study on configuring in-context learning demonstrations for unleashing MLLMs' sentimental perception capability. In *International Conference on Machine Learning*, 2025.

Daiqing Wu, Dongbao Yang, Can Ma, and Yu Zhou. Emocaliber: Advancing reliable visual emotion comprehension via confidence verbalization and calibration. *Pattern Recognition*, 2026a.

Jiulong Wu, Yucheng Shen, Lingyong Yan, Haixin Sun, Deguo Xia, Jizhou Huang, and Min Cao. Facial-r1: Aligning reasoning and recognition for facial emotion analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026b.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 2024.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI conference on artificial intelligence*, 2016.

Guanghao Zhang, Tao Zhong, Yan Xia, Mushui Liu, Zhelun Yu, Haoyuan Li, Wanggui He, Dong She, Yi Wang, and Hao Jiang. Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026a.

Hongrui Zhang, Daiqing Wu, Yangyang Li, Kuien Liu, Yuhui Wang, Yu Zhou, and Sicheng Zhao. Multimodal emotion recognition with large language models. *arXiv preprint arXiv:2605.21239*, 2026b.

Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *Proceedings of the AAAI conference on artificial intelligence*, 2019.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. In *Proceedings of the European Conference on Computer Vision*, 2024.

Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. Affective image content analysis: A comprehensive survey. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

Sicheng Zhao, Xingxu Yao, Jufeng Yang, et al. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Sicheng Zhao, Hongrui Zhang, Daiqing Wu, et al. Emotionreasoner: Emotion-explanation-oriented reinforcement learning for explainable multimodal emotion recognition. *IEEE Transactions on Affective Computing*, 2026.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

A Additional Details of Preliminary Experiments

A.1 Construction of Reference CoT in Preliminary Experiments

The VEC-CoT dataset adopts a hierarchical reasoning structure that formalizes the transition from visual perception to affective categorization through five progressive steps: (1) identifying prominent visual elements in the image; (2) providing detailed descriptions of human subjects, when present; (3) describing contextual elements beyond the subjects; (4) analyzing the interactions among these visual elements; and (5) deriving the final emotional conclusion based on the preceding observations.

Our preliminary experiments are conducted based on this structured reasoning formulation. Specifically, in the Reference CoT setting, we aim to provide reliable intermediate reasoning supervision while preventing direct answer leakage. Therefore, we retain only the first four stages of the CoT, which focus on visual grounding, contextual understanding, and higher-level semantic interactions, while removing the final emotion conclusion step that explicitly maps observations to the target label. As illustrated in Figure 7, the highlighted red <analysis></analysis>segment corresponds to the removed answer-related reasoning component. Additionally, in emotion-related datasets involving humans, we also remove the <human></human>portions. By excluding this final affective inference stage, the model can leverage high-quality visual reasoning trajectories without accessing the ground-truth emotional answer, ensuring a fair evaluation of the effectiveness of reasoning supervision.

A.2 Details of Preliminary Experiments Results

To further examine the effect of different reasoning strategies on VEC, we report detailed results on six benchmark datasets in Table 8. Overall, the results are consistent with the findings presented in the main paper. Across different MLLMs, the *Reference CoT* setting achieves the best average performance, indicating that high-quality reasoning trajectories can provide effective guidance for emotion-oriented visual understanding. This trend is observed not only on 7–8B-scale models, such as Qwen2.5-VL-7B, Qwen3-VL-8B, and LLaVA-1.5-7B, but also on the larger InternVL3.5-38B model, suggesting that the observed gains are not merely attributable to the limited capacity of smaller models.

In contrast, *Self-generated CoT* consistently yields lower average performance than the direct *w/o CoT* setting across all evaluated models. This suggests that, although explicit reasoning has the potential to benefit VEC, current MLLMs are still prone to generating unreliable or misaligned affective reasoning chains when relying solely on their own intermediate rationales. Such low-quality reasoning may introduce additional noise and consequently degrade the final emotion prediction.

These results further support our hypothesis that the bottleneck of CoT-based VEC does not lie in the reasoning paradigm itself, but rather in the quality and reliability of the generated reasoning trajectories. When accurate and label-leakage-free reasoning traces are provided, CoT can substantially improve VEC performance; however, when the reasoning is self-generated without sufficient supervision, it may fail to capture the subjective and context-dependent nature of image-evoked emotions.



Prompt

Which emotion might this image evoke? Choose the most likely one from ['awe', 'amusement', 'contentment', 'excitement', 'anger', 'sadness', 'fear', 'disgust']. Think step by step. Respond in the format: <think>{your reasoning}</think><answer>{your final answer}</answer>.

<think><element>The image depicts a vibrant garden scene with blooming orange flowers, some wilted blooms, and a touch of purple flowers in the background.</element>
<human>There are no people present in the image.</human><context>The setting is an outdoor garden bathed in sunlight, creating a warm and lively atmosphere. The combination of fresh and fading flowers suggests the natural cycle of life.</context><interaction>The bright, healthy flowers contrast with the wilted ones, adding a sense of balance and harmony to the scene. The sunlight enhances the vivid colors, contributing to an overall pleasing visual experience.</interaction><analysis>The serene and natural beauty of the garden scene, with its mix of life - cycle stages and vibrant colors, conveys a sense of contentment.</analysis></think>
<answer>Contentment.</answer>

Figure 7: Example of the hierarchical reasoning structure in VEC-CoT. The red highlighted <analysis></analysis>segment indicates the answer-related reasoning component removed in the Reference CoT setting to prevent direct answer leakage.

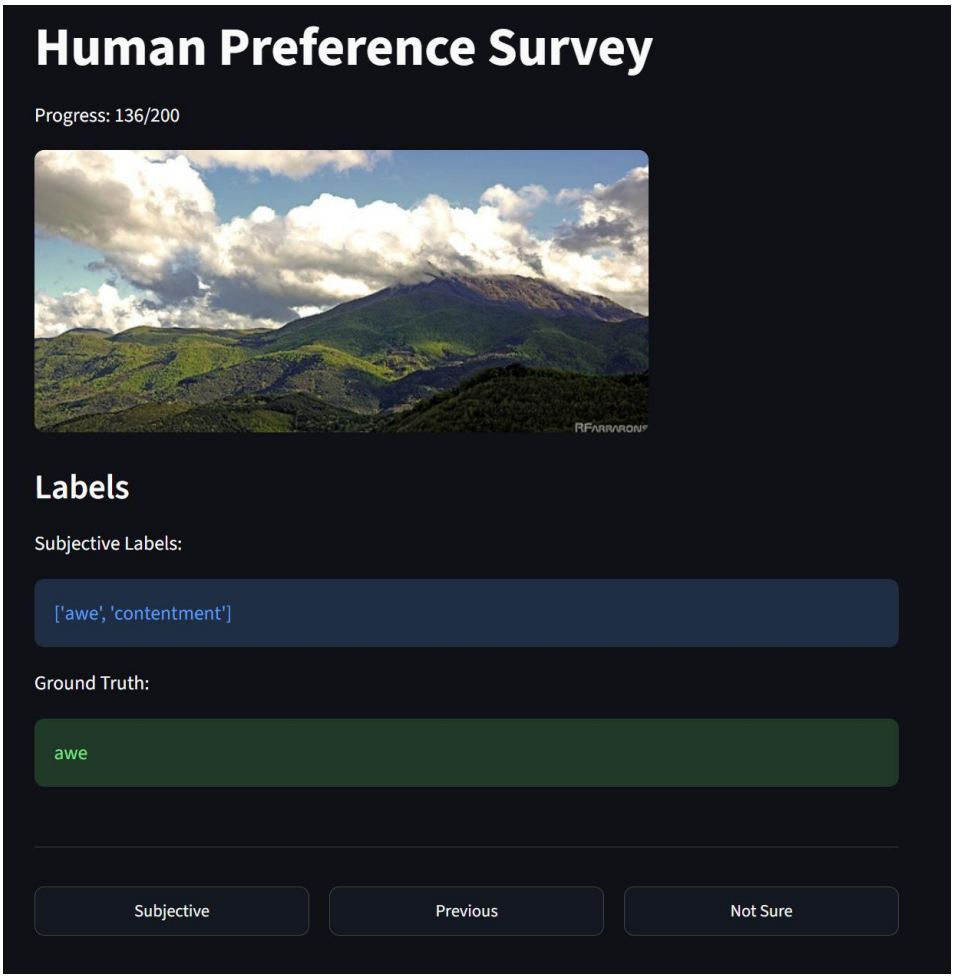

Figure 8: Interface of the human preference study for evaluating the quality of VECBench-Subjective annotations.

Model	Training	Inference	EmoSet-8	FI-2	FI-8	WebEmo-2	WebEmo-7	WebEmo-25	Mean
Qwen2.5-VL-7B	Base	w/o CoT	67.81	94.66	71.50	89.74	72.87	39.54	72.52
	Base	Self-generated CoT	61.54	91.63	65.44	86.66	70.21	33.82	68.22
	Base	Reference CoT	**71.93**	**96.79**	**73.89**	**93.37**	**79.88**	**41.20**	**76.18**
Qwen3-VL-8B	Base	w/o CoT	68.47	93.95	72.24	90.83	74.24	40.46	73.36
	Base	Self-generated CoT	67.11	92.77	68.56	90.37	73.47	39.29	71.93
	Base	Reference CoT	**71.31**	**98.04**	**78.30**	**95.93**	**74.95**	**41.98**	**76.75**
LLaVA-1.5-7B	Base	w/o CoT	64.73	87.36	70.40	80.40	49.83	**32.74**	64.24
	Base	Self-generated CoT	57.70	82.91	53.49	73.49	46.75	32.30	57.77
	Base	Reference CoT	**77.84**	**96.97**	**82.90**	**94.75**	**55.46**	31.76	**73.28**
InternVL3.5-38B	Base	w/o CoT	70.50	92.34	69.85	88.65	70.75	42.54	72.44
	Base	Self-generated CoT	69.25	89.14	64.15	84.92	68.72	39.45	69.27
	Base	Reference CoT	**74.65**	**97.82**	**75.22**	**92.65**	**75.67**	**43.12**	**76.62**

Table 8: Comparison of different reasoning strategies under various models. w/o CoT predicts emotion labels without explicit reasoning, while Self-generated CoT generates intermediate reasoning before prediction. Reference CoT provides a filtered reasoning chain with the final answer removed, serving as a high-quality supervision signal.

B Details of VECBench-Subjective Construction

B.1 More Comprehensive Experiments on VECBench-Subjective

In this appendix, we provide a more comprehensive evaluation on VECBench-Subjective, as shown in Table 9. We compare our method with a wide range of representative MLLMs, including open-source models, proprietary models, and emotion-oriented MLLMs. The evaluation covers both in-domain datasets and out-of-domain datasets, allowing us to better examine the robustness and generalization ability of different models in subjective visual emotion comprehension.

Overall, our method achieves the best average performance across all evaluated datasets, demonstrating its effectiveness on VECBench-Subjective. In particular, it obtains strong results on the in-domain benchmarks and also maintains competitive performance on out-of-domain datasets. These additional results further validate that our approach can better capture subjective emotional perceptions compared with existing general-purpose and emotion-oriented MLLMs.

B.2 Human Preference Study Details

To further validate the quality of VECBench-Subjective annotations, we conduct a human preference study where volunteers compare the original single-label annotations with our subjective annotations. The survey interface and annotation procedure are provided in Figure 8.

C Details of Canonical Affective Rubric Refinement

C.1 Prompt Templates for Rubric Refinement

We provide the prompt templates used in Canonical Affective Rubric Refinement for reproducibility. Figure 10 shows the prompt for generating instance-level candidate rubrics based on the question, reference answer, and error type. Figure 9 presents the prompt for selecting the most representative rubric(s) for reward supervision. Figure 11 shows the prompt used to classify reasoning failures according to our error taxonomy. Together, these prompts support a structured refinement process from rubric generation to rubric selection and error-type verification.

D Qualitative Results

To further illustrate the effectiveness of EmoCARE in enhancing affective reasoning, we present qualitative comparisons between our model and Qwen3-VL in Figure 12 and Figure 13. These examples demonstrate that our model produces more reliable and coherent reasoning trajectories by accurately grounding emotional predictions in visual evidence, contextual information, and affective cues. In contrast, Qwen3-VL tends to exhibit incomplete visual understanding or suboptimal affective reasoning, leading to less accurate emotional interpretations.

Rubrics Evaluator Selection Prompt

System Prompt: You are an expert in multimodal reasoning evaluation and reward function design.
User Prompt: Your task is to select the most appropriate candidate rubric(s) for evaluating the reasoning error of a given model prediction.
You are provided with:
1. Model reasoning: {model reasoning}
2. Identified Error Type in the Model Reasoning: {error type}
3. Reference Answer: {reference answer}
4. Candidate rubrics: {candidate rubrics}
The error type is considered correct and should NOT be re-evaluated. Your task is ONLY to identify the candidate rubric(s) that best characterize the observed reasoning error.
Selection Criteria:
- Select the rubric(s) that most directly explain the primary reasoning failure.
- The selected rubric(s) should be applicable as evaluation criteria for reward assignment, rather than merely describing this specific instance.
- Prefer rubrics that capture the fundamental error pattern instead of image-specific details.
- Do not select redundant rubrics with substantially overlapping meanings.
- Select one rubric whenever a single rubric sufficiently characterizes the error.
- Select two rubrics only when both describe complementary and indispensable aspects of the reasoning failure.
- Do not modify, rewrite, or summarize any rubric.
- Do not generate new rubrics.
Return ONLY a valid JSON array.
Example: ["Rubric A", "Rubric B"] OUTPUT:

Figure 9: Prompt template for the Rubric Evaluator to select the most representative rubric that characterize the identified reasoning failures for reward supervision.

Model	In Domain				Out of Domain			Mean
	EmoSet-8	**FI-8**	**WebEmo-7**	**WebEmo-25**	**Abstract-8**	**Artphoto-8**	**UnbiasedEmo-6**	
Open-Source MLLM								
Qwen2.5-VL-7B	67.40	70.50	57.10	28.60	48.03	51.98	90.05	59.09
Qwen2.5-VL-72B	71.30	74.00	61.30	30.80	51.09	60.17	91.00	62.81
Qwen3-VL-8B	69.50	75.50	62.50	37.70	43.72	49.17	82.70	60.11
Qwen3-VL-32B	63.30	72.40	58.50	37.70	48.47	56.07	88.80	60.75
InternVL3.5-8B	65.70	68.00	43.10	22.90	34.93	51.98	76.00	51.80
InternVL3.5-38B	68.70	64.20	56.10	36.80	34.06	58.31	87.80	58.00
GLM-4.6v-Flash	66.90	69.20	60.10	33.40	45.41	60.04	79.90	59.28
Proprietary MLLM								
Qwen3.6-Plus	72.80	76.80	62.30	38.30	**52.40**	**63.15**	<u>93.00</u>	<u>65.54</u>
GPT-5	72.50	74.10	57.30	34.10	**52.40**	<u>62.03</u>	77.90	61.48
Emotion-Oriented MLLM								
EmoVIT	72.80	70.60	53.10	23.90	35.37	44.66	85.70	55.16
Emotion-LLaMA	46.70	41.70	32.10	23.50	23.59	34.97	50.30	36.12
EEmo-Logic	71.70	74.20	60.90	32.20	45.03	55.58	83.40	60.43
EmoCaliber	<u>78.90</u>	<u>79.50</u>	<u>63.30</u>	<u>38.30</u>	35.80	51.61	90.70	62.59
EmoCARE (**Ours**)	**80.80**	**83.00**	**65.80**	**40.10**	<u>46.28</u>	61.66	**93.90**	**67.36**

Table 9: Performance comparison of different MLLMs on the Subjective-VECBench benchmark. The reported scores are hit score(%).

Rubric Designers Generation Prompt

System Prompt: You are an expert evaluator for Multimodal Emotion Recognition (MER).
User Prompt: Your task is to generate an instance-level rubric list for evaluating the reasoning process of a multimodal large language model.
You will be given:
1. Question: the original multimodal emotion recognition question.
2. Reference Answer: a high-quality reference reasoning and emotion label.
3. Error Type: the predefined error category that specifies the primary failure mode to be evaluated and determines the aspects that the generated rubrics should focus on.
Your goal is to generate 7-8 evaluation criteria that are: 1. Closely grounded in the Reference Answer.
2. Specific to this particular instance rather than generic emotion recognition criteria.
3. Sensitive to the given Error Type.
4. Short, atomic, and independently assessable.
5. Written as declarative checklist items.
6. Focus on key visual evidence, emotional cues, reasoning steps, or label consistency depending on the Error Type.
Error Type Definitions
Rubric Generation Rules
1. Extract important visual entities, emotional cues, events, and relations from the Reference Answer.
2. Generate rubrics centered around these instance-specific elements.
3. Avoid generic criteria such as:
- ”The model understands the image.”
- ”The model predicts the correct emotion.”
4. Prefer criteria like:
GOOD:
- ”Correctly identifies that the woman is wiping tears while smiling.”
- ”Recognizes the supportive interaction between the two individuals.”
- ”Determines whether the visible tears indicate sadness, relief, or joy.”
- ”Checks whether the final emotion label is consistent with the described emotional state.”
BAD:
- ”Understands facial expression.”
- ”Identifies emotion correctly.”
- ”Reasons logically.”
5. Each criterion should evaluate ONE aspect only.
———————————————————
Output format: Return ONLY a Python list of strings.
Question: {question}
Reference Answer: {reference answer}
Error Type: {error type}

Figure 10: Prompt template for the Rubric Designers to generate instance-level evaluation rubrics conditioned on the reference reasoning and predefined error categories.

Rubrics Evaluator Judge the Failure Sample Error Type Prompt

System Prompt: You are an expert evaluator of reasoning chains (CoT, Chain-of-Thought).
Your task is to compare a candidate CoT with a Correct CoT and classify the type of error, if any, according to the following categories:
- M1: The reasoning process contains incorrect visual interpretations or hallucinated visual elements, such as misidentifying objects, scenes, attributes, or other observable details in the image. The error primarily originates from inaccurate visual understanding.
- M2: The reasoning process contains incorrect emotional or affective inferences, including misjudging emotions, intentions, attitudes, or sentiments, while the visual interpretation itself is largely correct. The error primarily arises from flawed affective reasoning rather than visual perception.
- M3: The reasoning process is largely correct and leads to a plausible emotional interpretation; however, the final answer in <answer>...</answer>is incorrect or inconsistent with the reasoning process. This category captures cases where the error stems from the mismatch between reasoning and the final prediction.
- M4: The reasoning process is coherent and the final answer in <answer>...</answer>is also reasonable; however, the predicted emotion does not exactly match the reference label due to the inherent subjectivity and ambiguity of emotion perception. In such cases, the predicted label represents a plausible alternative emotional interpretation rather than a clear reasoning failure.
- C: The reasoning process is correct, and the final answer in <answer>...</answer>is correct.
Instructions:
1. Carefully read both the candidate CoT and the Correct CoT.
2. Identify if the candidate CoT has an error.
3. If there is an error, classify it strictly as M1, M2, M3, M4 or C based on the definitions above.
4. Provide only the label M1, M2, M3, M4 or C, and no other text.
Format your output exactly as follows: Type: M1/M2/M3/M4/C
Correct CoT: {reference cot}
Candidate CoT: {candidate cot}
Type:

Figure 11: Prompt template for the Rubric Evaluator to identify reasoning failure types based on the proposed error taxonomy.



Prompt
Which emotion might this image evoke? Choose the most likely one from ['awe', 'amusement', 'contentment', 'excitement', 'anger', 'sadness', 'fear', 'disgust'].
Think step by step. Respond in the format:
<think>your reasoning</think><answer>your final answer</answer>.

Answer: Anger

Qwen3-VL-8B-Instruct
<think>The image shows a male lion with its mouth wide open, revealing its teeth and tongue. This is a typical display of a roar, which lions use for various purposes such as asserting dominance, communicating over long distances, or intimidating rivals. While a roar can be associated with aggression or anger, in this context, the lion's posture doesn't appear overtly threatening — its eyes are somewhat narrowed but not glaring, and the surrounding foliage suggests a calm, natural setting. "Fear" is possible, but less likely given the lack of overt aggression. "Amusement" or "contentment" don't fit the context. "Excitement" might be a secondary reaction, but "awe" is the most direct and emotionally resonant response to such a display of primal power. Therefore, the most likely emotion is "awe".</think><answer>awe</answer>

Subjective Ambiguity Error

EmoCARE
<think><element>The image depicts a lion with its mouth wide - open, showing its teeth, set against a backdrop of green foliage.</element><human>There are no people present in the image to observe facial expressions or body postures. The focus is entirely on the lion.</human><context>The lion's display is a natural behavior often associated with dominance or aggression in the animal kingdom. The setting, a natural habitat with dense foliage, emphasizes the wild and untamed nature of the scene.</context><interaction>The lion's open - mouth display stands out against the foliage, highlighting its assertiveness. The overall composition gives a sense of power and intensity, which are key elements in conveying the emotion at play.</interaction><analysis>Given the lion's aggressive - looking display and the overall sense of intensity, the emotion conveyed in the image is anger.</analysis></think><answer>Anger.</answer>

Correct

Figure 12: Qualitative results of affective reasoning. The examples demonstrate the effectiveness of our method in generating coherent reasoning trajectories and reliable emotion predictions.

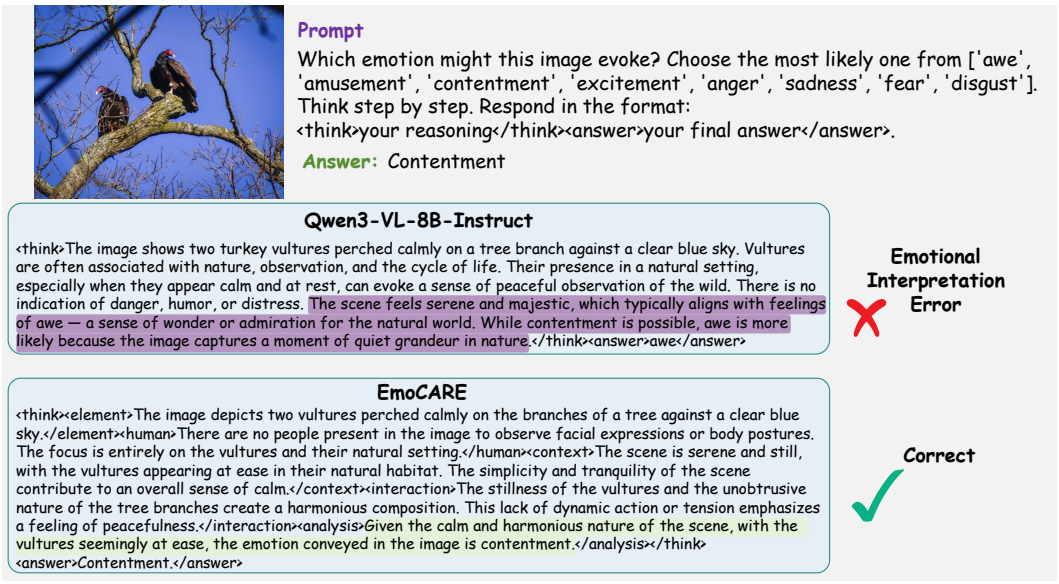

Figure 13: Qualitative results of affective reasoning. The examples demonstrate the effectiveness of our method in generating coherent reasoning trajectories and reliable emotion predictions.