AICA-Bench: Holistically Examining the Capabilities of VLMs in Affective Image Content Analysis

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) have demonstrated strong capabilities in perception, yet holistic Affective Image Content Analysis (AICA)—which integrates perception, reasoning, and generation into a unified framework—remains underexplored. To address this, we introduce AICA-Bench, a comprehensive benchmark comprising three core tasks: Emotion Understanding (EU), Reasoning (ER), and Generation (EGCG). We evaluate 23 VLMs, revealing critical gaps: models struggle with intensity calibration and suffer from descriptive shallowness in open-ended tasks. To bridge these gaps, we propose Grounded Affective Tree (GAT) Prompting, a training-free framework that integrates visual scaffolding with hierarchical reasoning. Experiments show that GAT effectively corrects intensity errors and significantly enhances descriptive depth, establishing a robust baseline for future affective multimodal research.

1 Introduction

Recent years have witnessed rapid advances in Vision-Language Models (VLMs) (Liu et al., 2023a, 2024a,b; Wang et al., 2024; Bai et al., 2025; Chen et al., 2025; Zhu et al., 2025a; Yao et al., 2025), which integrate visual and textual modalities to perform a wide range of tasks, from image captioning and visual question answering to grounded reasoning. To assess these capabilities, a variety of benchmarks (Yue et al., 2024; Liu et al., 2024c; Fu et al., 2024; Wang et al., 2023) have been developed, focusing primarily on factual correctness, semantic grounding, visual reasoning, or multi-discipline understanding. Yet emotional intelligence remains an underexplored but essential aspect of evaluating VLMs and multimodal large language models (MLLMs).

In response, several recent benchmarks have started to explore the emotional capabilities of VLMs and MLLMs (Yang et al., 2024; Bhattacharyya and Wang, 2025; Lian et al., 2025; Gao et al., 2025; Song et al., 2024), as summarized in Table 1. Emo-Bench-M (Yang et al., 2024) is a benchmark designed to evaluate the emotional intelligence (EI) capabilities of MLLMs. EVE (Bhattacharyya and Wang, 2025) introduces an image-based benchmark for emotion recognition, evaluating seven open-source VLMs across classification, grounding, and explanation tasks. AffectGPT (Lian et al., 2025) proposes MER-UniBench, a benchmark designed for fine-grained multimodal emotion recognition, with a focus on video and dialogue understanding in MLLMs. EEmo-Bench (Gao et al., 2025) focuses on image-evoked emotion assessment. MOSABench (Song et al., 2024) introduces a benchmark for multi-object sentiment analysis in MLLMs.

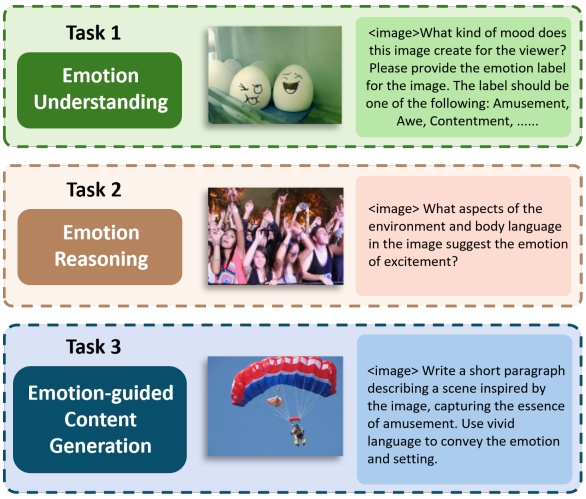


Figure 1: Illustration of the three affective tasks in the AICA-Bench benchmark.

Those benchmarks have made meaningful progress in evaluating the emotional capabilities of VLMs and MLLMs. However, they have yet to deeply investigate AICA, focusing on how emotions are interpreted, explained, and generated from

Benchmark	Model	Tasks	#Datasets	#AICA Datasets	#Instr.	#Models	Prompt
EVE (Bhattacharyya and Wang, 2025)	VLM	EU	5	5	8,009	7	B+CoT
AffectGPT (Lian et al., 2025)	MLLM	EU	9	3	-	17	B
EEmo-Bench (Gao et al., 2025)	MLLM	EU	1	1	6,773	19	B
EmoBench-M (Hu et al., 2025)	MLLM	EU	13	-	6,226	20	B
MOSABench (Song et al., 2024)	MLLM	EU	1	1	1,000	8	B
AICA-Bench (Ours)	**VLM**	**EU, ER, EGCG**	**9**	**9**	**18,124**	**23**	**B+CoT**

Table 1: Comparison of recent emotion benchmarks for VLMs and MLLMs. **EU**: Emotion Understanding, **ER**: Emotion Reasoning, **EGCG**: Emotion-guided Content Generation. #Instr. = total evaluation instructions; Prompt = **B**: Basic prompting, **CoT**: Chain-of-Thought prompting.

visual content. As shown in Table 1, most benchmarks include only a few image-based emotion datasets and focus primarily on basic *emotion understanding* tasks, typically framed as multiple-choice emotion classification. Yet prior work in affective computing and emotion psychology (Picard, 1997; Cambria, 2016) highlights that emotional intelligence involves not only recognizing affective cues, but also reasoning about emotional causes and producing contextually appropriate affective expressions.

The lack of comprehensive benchmarks for holistic Affective Image Content Analysis with VLMs —spanning understanding, reasoning, and generation—is a critical bottleneck in advancing affective intelligence. To bridge this gap, we propose **AICA-Bench** (Sec 3), a holistic benchmark evaluating VLMs across three complementary dimensions: **Emotion Understanding (EU)**, **Emotion Reasoning (ER)**, and **Emotion-Guided Content Generation (EGCG)**. AICA-Bench integrates 9 diverse datasets and 18,124 standardized instructions to assess not only cue recognition but also causal explanation and empathetic expression. Using this benchmark, we conduct a comprehensive evaluation of 23 open- and closed-source VLMs under zero-shot settings.

Beyond quantitative analysis (Sec 4), our detailed diagnostic error analysis uncovers two recurrent failure patterns(Sec 5): (1) *Emotion Intensity Hallucination*, where models frequently confuse high-arousal emotions (e.g., *Amusement*) with low-arousal ones (e.g., *Contentment*) due to weak visual grounding; and (2) *Descriptive Shallowness*, where generated responses suffer from generic, template-like content. In response, we introduce **Grounded Affective Tree (GAT)** Prompting (Sec 5). This training-free framework leverages visual scaffolding to steer models toward precise calibration and richer descriptive depth. Across the three tasks, GAT produces consistent improvements, with the EU Task increasing by 6.15 percentage points and the ER and EGCG Tasks increasing by 3.54 and 3.96 percentage points, respectively.

2 Related Work

Vision-Language Models (VLMs) have rapidly advanced from early pretraining frameworks such as Flamingo (Alayrac et al., 2022) and BLIP (Li et al., 2022) to modern instruction-following systems capable of handling diverse multimodal tasks. Recent open-source representatives include the LLaVA family (Liu et al., 2023a, 2024a,b), the Qwen2-VL and Qwen2.5-VL series (Wang et al., 2024; Bai et al., 2025), and InternVL2.5/3 (Chen et al., 2025; Zhu et al., 2025a), alongside compact models such as MiniCPM-V (Yao et al., 2025). Commercial systems such as GPT-4o and Gemini 2.5 further demonstrate the practical utility of this paradigm.

Most existing benchmarks primarily evaluate factual perception, VQA, captioning, and instruction following (Schwenk et al., 2022; Yin et al., 2024; Tu et al., 2024; Yue et al., 2024; Thapliyal et al., 2022; Liu et al., 2023b), leaving the affective dimension—how models perceive, interpret, and generate emotional content—largely understudied. Recent work has begun probing VLMs' affective abilities (Yang et al., 2024; Bhattacharyya and Wang, 2025; Lian et al., 2025; Gao et al., 2025; Song et al., 2024), but these studies typically rely on limited datasets and focus narrowly on emotion classification, without addressing affective reasoning or emotionally grounded generation.

Motivated by holistic evaluation efforts such as HELM (Liang et al., 2023a) and VHELM (Liang et al., 2023b), we introduce **AICA-Bench**, a holistic benchmark that systematically evaluates VLMs across the full spectrum of affective understanding, reasoning, and generation.

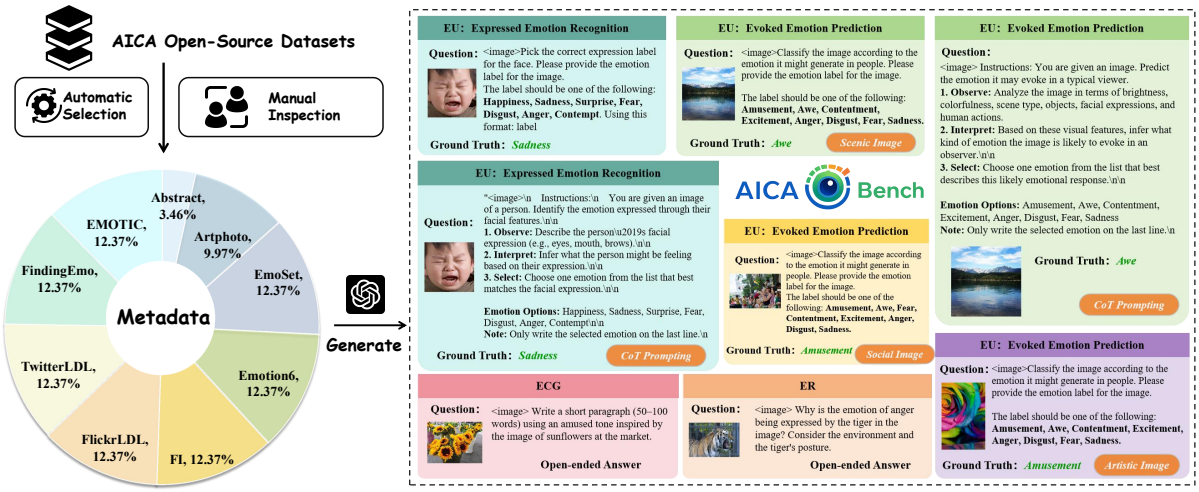


Figure 2: The instruction curation pipeline of the AICA-Bench benchmark. It consists of two stages: (1) image sample selection and filtering from multiple affective datasets, and (2) automatic instruction generation across three tasks—emotion understanding (EU), emotion reasoning (ER), and emotion-guided content generation (EGCG).

3 The AICA-Bench Benchmark

3.1 Evaluation Tasks Design

As discussed in the Introduction section, the existing affective benchmarks for VLMs focus primarily on *emotional understanding* tasks. However, emotional intelligence extends beyond perception, encompassing the ability to reason about the underlying causes of emotions and generate emotionally aligned content. To fill this gap, we propose the AICA-Bench benchmark, which evaluates VLMs on three tasks—emotional understanding(**EU**), emotion reasoning(**ER**), and emotion-guided content generation(**EGCG**)—reflecting different capability dimensions of VLMs in "understanding–reasoning–generation" for affective intelligence. (see Fig. 1 for task demonstrations).

1. **EU**: assesses a model's ability to identify emotions explicitly expressed in an image (*Expressed Emotion Recognition*) and infer the likely emotional response elicited in viewers (*Evoked Emotion Prediction*);

2. **ER**: measures a model's ability to explain why an image evokes a particular emotion, requiring causal reasoning grounded in visual context;

3. **EGCG**: evaluates a model's ability to produce emotionally congruent descriptions conditioned on an image and a target emotion.

3.2 Evaluation Dataset Construction

The dataset construction process in AICA-Bench comprises two stages to ensure both representativeness and quality of the evaluation prompts. Stage 1 curates high-quality affective image samples from existing public datasets, and Stage 2 leverages GPT-4o to automatically generate structured evaluation instructions for emotion understanding, reasoning, and generation tasks. The overall pipeline is illustrated in Figure 2.

Image Collection and Filtering We curate a total of 8,086 affective images as the visual foundation of our benchmark. These samples are drawn from a range of publicly available affective computing datasets and undergo a two-step filtering process. Candidate images are first selected using automated scripts, after which 20 trained annotators manually review them to remove any content that is not safe for work or emotionally ambiguous. Rejected samples are replaced to maintain both emotional clarity and dataset diversity. Specifically, we include 1,000 samples each from EmoSet (Yang et al., 2023), Emotion6 (Peng et al., 2015), FI (You et al., 2016), FlickrLDL (Yang et al., 2017), TwitterLDL (Yang et al., 2017), FindingEmo (Mertens et al., 2024), and EMOTIC (Kosti et al., 2020), along with 806 samples from Artphoto (Machajdik and Hanbury, 2010) and 280 from Abstract (Machajdik and Hanbury, 2010).

Evaluation Instruction Generation In the second stage, we automatically generate structured evaluation instructions for each pair of curated image-label using GPT-4o. This automated approach enables the consistent and scalable creation of task-specific prompts while ensuring alignment with the evaluation dimensions of the benchmark.

The instructions cover three task categories. **EU** includes two subtasks: *Expressed Emotion Recognition*, identifying the emotion explicitly displayed by individuals in the image, and *Evoked Emotion Prediction*, inferring the likely emotional response of viewers. Both subtasks use two prompt styles: Basic prompts, which directly query the emotion from a predefined category list, and CoT prompts, which guide the model to reason step by step over visual cues such as color and brightness, contextual elements like scene type and objects, and semantic indicators such as facial expressions before predicting the label. **ER** provides the image and its emotion label and asks the model to explain why the image evokes that emotion, testing causal reasoning grounded in visual context. **EGCG** similarly takes an image and target emotion but prompts the model to generate an emotionally congruent scene description, assessing its ability to produce affect-aligned content.

3.3 Evaluation Strategy

Our evaluation strategy is designed to accommodate the heterogeneous task formats in AICA-Bench, including multiple-choice emotion recognition (EU) and open-ended generation tasks (EGCG and ER).

EU: For the multiple-choice emotion understanding task, we adopt the weighted F1 score as the primary evaluation metric. The weighted F1 better reflects the performance in the unbalanced emotion categories, taking into account both the precision of each class and the recall while weighting each class proportionally to its frequency. These two sub-tasks are defined following the taxonomy in (Kosti et al., 2020), which distinguishes between *Expressed Emotion Recognition* and *Evoked Emotion Prediction*.

EGCG and ER: Both emotion-guided content generation and emotion reasoning involve open-ended outputs where traditional automatic metrics (e.g., BLEU (Papineni et al., 2002)) are insufficient. These conventional metrics emphasize lexical overlap or global semantic similarity but fail to capture critical affective dimensions.

To address these limitations, we propose an **AICA-Bench Scoring Model** based on QwenVL2.5-7B, fine-tuned with instruction-style supervision grounded in human-annotated evaluations. Instead of relying solely on synthetic LLM scores, we collect human ratings across multiple affective criteria and use them to guide the learning process. Our scoring system evaluates open-ended responses along three criteria: (1) **Emotion alignment**—the degree to which the response reflects the target emotion; (2) **Descriptiveness**—the richness and specificity of the generated content; and (3) **Causal soundness**, which we introduce specifically for the *emotion reasoning* task to assess whether the response provides a plausible explanation linking visual cues to the predicted emotion. These dimensions are inspired by prior work (Sedoc et al., 2017; Celikyilmaz et al., 2020) and extended to suit the multimodal and reasoning-oriented nature of our benchmark.

Model	ER			EGCG		
	MSE	MAE	Pearson	MSE	MAE	Pearson
Qwen2.5VL-7B	1.220	0.775	0.472	0.845	0.645	0.528
Gemini2.5-Pro	1.940	0.940	0.441	0.855	0.620	0.735
ChatGPT-4o	1.360	0.790	0.502	0.580	0.470	0.775
Ours	**0.400**	**0.295**	**0.880**	**0.260**	**0.240**	**0.900**

Table 2: Performance comparison on ER and EGCG tasks against human annotations.

We construct a dataset of 10,000 open-ended question–answer pairs via GPT-4o (5,000 for *emotion reasoning* and 5,000 for *emotion-guided content generation*), which is subsequently partitioned into training, validation, and test sets with a ratio of 8:1:1. Each response is independently annotated by five different annotators randomly selected from a pool of ten, using our proposed affective criteria on a 1 to 5 scale (interannotator agreement measured by Krippendorff's α is 0.78, indicating substantial agreement suitable for model training). We then fine-tune Qwen2.5VL-7B with the human-labeled data in a supervised manner, producing the AICA-Bench scoring model, which demonstrates superior alignment with human judgment compared to baseline and closed-source LLMs (Table 2). To facilitate comparison on the leaderboard, we normalize the raw 1–5 ratings (s) to a percentage score via $S_{\%} = \frac{s}{5} \times 100$ and report the macro-average across criteria.

4 Experiments

4.1 Experimental Setup

We benchmark a total of 23 VLMs, comprising 7 closed-source commercial models via APIs and 16 open-source models ranging from 2B to 16B parameters.

Closed-source models. We evaluate the following commercially hosted VLMs via API access:

Model	EU Basic	EU CoT	EU Avg.	ER Avg.	EG Avg.	**Overall Avg. (%)**
Closed-source Models						
Gemini-2.5-pro	66.97	67.57	67.27	79.08	74.13	**73.49**
Qwen-VL-max	64.07	65.98	65.02	77.75	75.93	**72.90**
ChatGPT-4o	64.44	65.42	64.93	77.81	75.73	**72.82**
Gemini-2.5-flash	68.05	69.32	68.68	76.55	68.19	**71.14**
ChatGPT-4o-mini	60.15	63.68	61.91	76.45	74.09	**70.81**
Qwen-VL-plus	60.04	67.81	63.92	72.39	66.86	**67.73**
Gemini-2.0-flash	67.16	68.98	68.07	71.05	63.93	**67.68**
Open-source Models (Size > 6B)						
Qwen2.5VL-7B (Bai et al., 2025)	56.43	57.25	56.84	74.50	66.00	**65.78**
Ovis2-16B (Lu et al., 2024)	54.38	54.70	54.54	68.24	71.56	**64.78**
Ovis2-8B (Lu et al., 2024)	53.63	52.73	53.18	68.89	70.81	**64.29**
InternVL3-14B (Zhu et al., 2025b)	52.91	52.04	52.47	68.27	66.50	**62.41**
InternVL3-8B (Zhu et al., 2025b)	52.18	52.98	52.58	67.21	67.27	**62.35**
InternVL2.5-8B (Chen et al., 2024)	51.89	51.03	51.46	66.48	68.86	**62.27**
MiniCPM-o-2.6 (Yao et al., 2025)	52.73	48.65	50.69	70.16	64.98	**61.94**
Qwen2VL-7B (Wang et al., 2024)	53.52	55.19	54.36	65.23	64.76	**61.45**
LLaVA-1.6-13B (Liu et al., 2024b)	36.78	46.82	41.80	73.57	64.51	**59.96**
LLaVA-1.6-7B (Liu et al., 2024b)	36.58	50.22	43.40	73.81	59.58	**58.93**
MiniCPM-V-2.6 (Yao et al., 2024)	43.70	47.25	45.48	65.77	63.00	**58.08**
LLaVA-onevision (Li et al., 2024)	54.02	53.25	53.64	63.78	54.18	**57.20**

Table 3: Main results on the AICA-Bench benchmark. **EU**: Emotion Understanding, **ER**: Emotion Reasoning, **EGCG**: Emotion-guided Content Generation.The complete leaderboard is provided in Appendix Table 8.

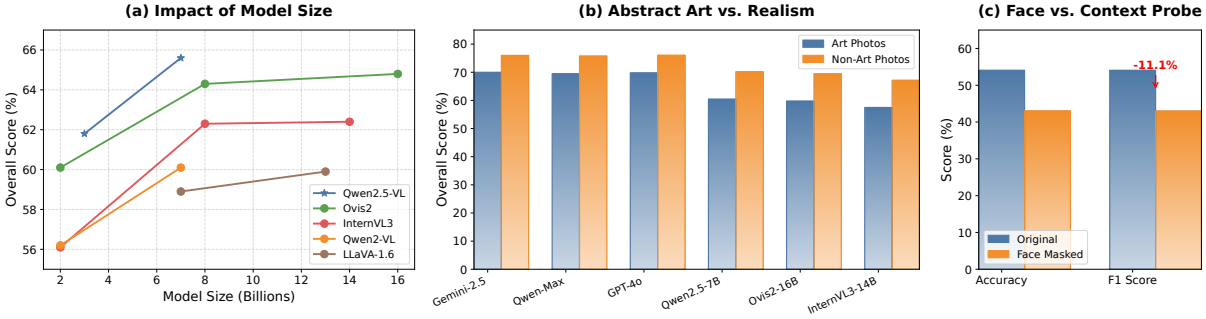


Figure 3: **(a)** Performance gains from model scaling diminish beyond 7B parameters. **(b)** Models consistently struggle with abstract art compared to realistic photos. **(c)** Masking facial cues causes an 11.1% performance drop, revealing the models' heavy reliance on visual shortcuts (faces) over holistic context.

Gemini-2.5-Pro, Gemini-2.5-Flash, Gemini-2.0-Flash, GPT-4o, GPT-4o-Mini, Qwen-VL-Max, Qwen-VL-Plus.

Open-source models. We include a diverse set of open-source models covering different architectures and sizes: Qwen series (Qwen2.5VL-7B-Instruct, Qwen2.5-VL-3B-Instruct (Bai et al., 2025), Qwen2-VL-7B-Instruct, Qwen2-VL-2B-Instruct (Wang et al., 2024)), InternVL series (InternVL3-14B-Instruct, InternVL3-8B-Instruct, InternVL3-2B-Instruct (Zhu et al., 2025b), InternVL2.5-8B (Chen et al., 2024)), Ovis series (Ovis2-16B, Ovis2-8B, Ovis2-2B (Lu et al., 2024)), LLaVA series (LLaVA-1.6-13B, LLaVA-1.6-7B, LLaVA-OneVision (Li et al., 2024)), and MiniCPM series (MiniCPM-V-2.6 (Yao et al., 2024), MiniCPM-O-2.6 (Yao et al., 2025)).

Computing environment. All open-source models are evaluated on a cloud computing platform using NVIDIA A100 GPUs. Closed-source models are accessed through their respective APIs under standard inference settings.

4.2 Main Results and Analysis

We present the comprehensive evaluation results in Table 3 and visualize key analytical dimensions in Figure 3. Our analysis highlights four critical observations regarding the current state of multimodal affective intelligence, with additional breakdowns and extended quantitative results provided in Appendix B.3.

The Understanding-Reasoning-Generation Gap. Although closed-source models still hold the lead, the gap between them and top-tier open-source mod-

els varies significantly across tasks. In reasoning-heavy tasks (ER), this gap has narrowed to just ~4.5%, but in visual perception (EU), it remains substantial at over 10%. This reflects a consistent **"Top-Heavy" pattern** across all evaluated models: scores for Reasoning and Generation are consistently 15-30% higher than Understanding scores. A clear example is LLaVA-1.6-13B, which achieves high reasoning performance (73.57%) comparable to state-of-the-art models, yet performs poorly in basic perception (41.80% in EU). This indicates that current models rely heavily on their language priors to infer emotions rather than truly perceiving them from visual cues. Furthermore, simply increasing model size offers limited benefits (Figure 3a); for instance, scaling parameters from 8B to 16B in the Ovis2 series yields negligible gains. This collectively suggests that the primary bottleneck is not the model size or textual capabilities, but the quality of fine-grained visual encoding.

Sensitivity to Visual Abstraction. We observe a consistent performance degradation when models process Abstract Art Photos compared to Realistic Non-Art Photos, as shown in Figure 3(b). Real-world photos typically contain clear objects and facial expressions that VLMs are well-trained to recognize. In contrast, art photos convey emotion through abstract features such as color theory and composition. The universal performance drop indicates that current VLMs lack effective affective transfer capability, struggling to generalize from concrete visual objects to abstract representations.

Over-reliance on High-Level Facial Cues. Inspired by psychological studies (Yang et al., 2023) which posit facial expressions as critical high-level emotion attributes, we investigate the extent to which current VLMs rely on this specific visual cue for EU. While facial expressions undoubtedly influence human emotional perception, a robust VLM should ideally integrate diverse visual information across levels. To rigorously test this, we conducted a controlled ablation study on a diagnostic subset of 500 images sampled from the context-rich EMOTIC and EmoSet datasets. For each image, facial regions were detected and masked with opaque bounding boxes using MediaPipe. Crucially, to ensure experimental validity, we manually verified the masked subset to explicitly retain only samples where the emotional intent remained visually inferable through remaining scene-level or action-level cues. The results, as illustrated in Figure 3(c), reveal a critical dependency: when high-level facial cues are occluded, the model's performance suffers a sharp decline, with the F1 score dropping by 11.1%. This significant degradation confirms that current VLMs heavily utilize facial expressions as a primary heuristic, struggling to effectively synthesize alternative contextual information when this visual shortcut is removed.

5 Diagnostic Analysis and Affective Steering

While Section 4 establishes the performance benchmarks, it does not reveal the underlying cognitive mechanisms behind model failures. In this section, we first conduct a diagnostic error analysis to pinpoint the limitations of current VLMs. Based on the identified bottlenecks—specifically the lack of visual grounding and the confusion in arousal estimation—we introduce a theoretically grounded framework, the Grounded Affective Tree (GAT) Prompting. This approach integrates visual scaffolding with hierarchical reasoning to steer models towards robust affective alignment and descriptive depth across all three tasks.

5.1 Diagnostic Error Analysis

EU Misclassification Patterns: Intensity vs. Polarity. A comprehensive evaluation of total errors reveals that misclassifications are disproportionately driven by **Intensity Errors** (72.25%), significantly overshadowing **Polarity Errors** (27.75%). This fundamental split indicates that while models generally succeed in distinguishing positive from negative affective states, they struggle severely with gauging arousal magnitude from visual cues. This trend is visually reinforced in Figure 4a, where the top ten misclassification patterns are dominated by red gradient bars representing within-polarity confusion, such as mislabeling *Amusement* as lower-arousal *Contentment* (10.38%) or higher-arousal *Excitement* (9.69%). Notably, the presence of a blue gradient bar signifies critical cross-valence failures, exemplified by the severe confusion of negative *Disgust* with positive *Awe* (3.62%). This demonstrates that despite the lower overall frequency of polarity errors in the top tier, models remain susceptible to catastrophic valence flips, particularly when faced with high-arousal visual features.

Descriptive Shallowness in Reasoning & Generation. For open-ended tasks, Figure 4(b)-(c) reveals a consistent dichotomy across both ER

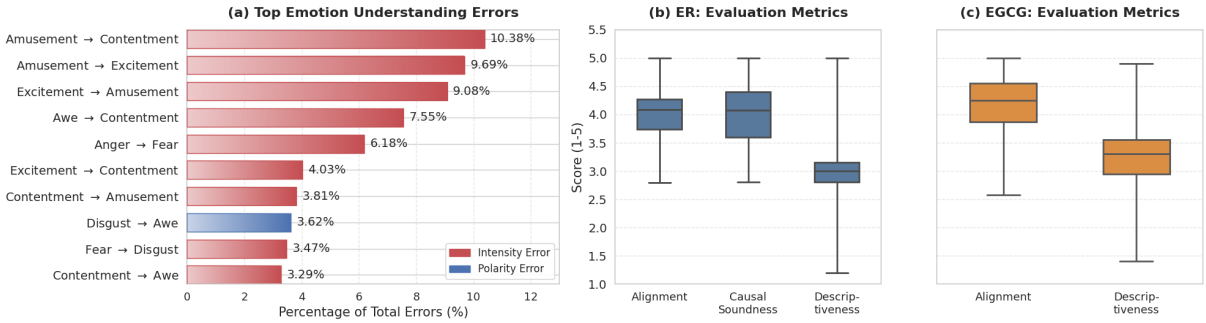


Figure 4: **(a) EU:** The dominance of intensity errors (Blue) over valence errors (Red) reveals an arousal bottleneck. **(b)-(c) ER & EGCG:** Across both tasks, models achieve high emotion alignment but consistently lack descriptive depth.

Figure 5: Representative failure cases of EU, ER, and EGCG tasks.

and EGCG: while models consistently achieve high scores in *Emotion Alignment* (Median ≈ 4.1), indicating effective topic adherence, their *Descriptiveness* lags significantly (Median ≈ 3.0). This performance gap highlights a "Safe Response Trap." Figure 5 qualitatively confirms this pattern, illustrating how models often resort to "template-filling"—producing generic, safe descriptions rather than grounding their reasoning in specific, subtle visual evidence like lighting, texture, or social interactions.

5.2 Steering via Grounded Affective Tree Prompting

To address the *Intensity* and *Polarity Errors* alongside the *Descriptive Shallowness* identified in Section 5.1, we introduce **Grounded Affective Tree (GAT) Prompting**.

Why Training-free-based Steering? First, *Efficiency*: simple instructions have been proven to steer latent capabilities effectively without altering weights (Wu et al., 2025). Second, *Scalability*: prompting facilitates test-time scaling, where structured strategies theoretically enhance reasoning performance (Liu et al., 2025). Third, *Universality*: unlike fine-tuning, prompting serves as a training-free, model-agnostic solution applicable to both open-source and black-box models. Training-free-based

5.2.1 GAT Prompting

Drawing inspiration from recent work on scaffolding coordinates for vision-language coordination (Lei et al., 2025) and the Tree-of-Thoughts (ToT) reasoning framework(Yao et al., 2023), our proposed GAT Prompting orchestrates a structured cognitive process.

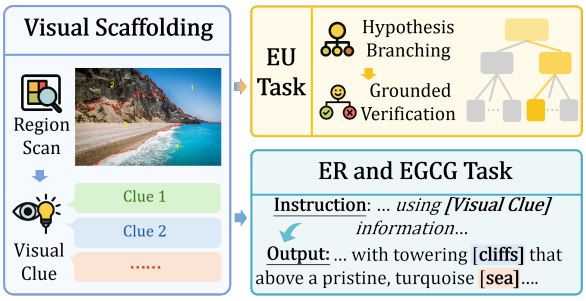


Figure 6: The GAT Prompting Framework.

Visual Scaffolding. The visual scaffolding is generated using an efficient graph-based image segmentation method(Chakrabarti, 2020), which creates large, contiguous regions that serve as explicit visual anchors in the prompt to guide the VLM's attention. (Please refer to Appendix C for illustrative examples). Based on these segmented anchors, the prompt instructs the VLM to execute a systematic "region scan." For each designated region, the model is required to identify and list objective visual elements, thereby extracting distinct visual

clues grounded in specific areas of the image.

AffectToT Reasoning (for EU). To address classification ambiguity in the EU task, we introduce AffectToT, a structured cognitive appraisal framework adapted from the ToT paradigm with a fixed search depth ($d = 3$) and breadth ($k = 3$). Leveraging the objective region descriptions from the initial scan (Step 1) as the foundation, the workflow proceeds to a *Hypothesis Branching* phase (Step 2), where the model generates $k = 3$ distinct competing emotion-intensity pairs, explicitly citing specific region IDs as grounding evidence. Subsequently, a *Grounded Verification* stage (Step 3) functions as a critic to evaluate the logical consistency between each proposed intensity and the visual facts. By identifying contradictions—such as a relaxed posture invalidating a high-arousal hypothesis—the system actively prunes invalid branches. This single-pass propose-and-verify mechanism effectively eliminates intensity hallucinations within a constrained token budget, ensuring the final decision is strictly grounded in verified visual evidence.

Grounded Generation (for ER & EGCG). In parallel, for open-ended generation tasks—ER and EGCG—the strategy pivots to directly leveraging the objective foundation established by visual scaffolding to enhance descriptive fidelity.

5.2.2 Effectiveness of GAT Prompting

Impact on Task-Level Performance. Table 4 demonstrates that GAT yields consistent performance improvements across EU, ER, and EGCG for all evaluated models. Specifically, the framework significantly boosts accuracy in EU while bridging the performance gap in higher-level ER and EGCG tasks, enabling smaller open-source models to approach the capabilities of proprietary baselines. These results confirm GAT as a robust enhancer that effectively grounds abstract emotional concepts in concrete visual evidence to benefit both perception and generation.

Improvements in Error Calibration and Descriptive Quality. Figure 7 illustrates the specific improvements in error handling and content quality. The left panel shows that GAT acts as an effective *intensity calibrator*, significantly suppressing confusion between high- and low-arousal emotions by correcting visual hallucinations. The right panel highlights a structural improvement in *descriptive depth* for ER and EGCG. Unlike baselines that tend to produce shallow responses, GAT shifts the score distribution upwards with condensed variance. This confirms that our method effectively eliminates generic outputs, ensuring that the generated content is descriptively rich and contextually grounded.

Table 4: Comprehensive performance comparison. We compare Baseline and GAT for EU, ER, and EGCG, where **green** values indicate improvements of GAT over the Baseline.

Model	EU		ER		EGCG	
	Basic	GAT	Basic	GAT	Basic	GAT
Gemini-2.5-Pro	66.97	**71.15**	79.08	**82.45**	74.13	**78.25**
ChatGPT-4o	64.44	**69.82**	77.81	**83.10**	75.73	**79.40**
Qwen2.5-VL-7B	56.43	**61.94**	74.50	**78.92**	66.00	**70.55**
Qwen2-VL-7B	53.52	**59.08**	65.23	**69.15**	64.76	**68.30**
Ovis2-8B	53.63	**57.41**	68.89	**71.50**	70.81	**73.20**
InternVL3-8B	52.18	**56.33**	67.21	**70.30**	67.27	**70.85**
LLaVA-1.6-7B	36.58	**51.05**	73.81	**75.88**	59.58	**63.45**

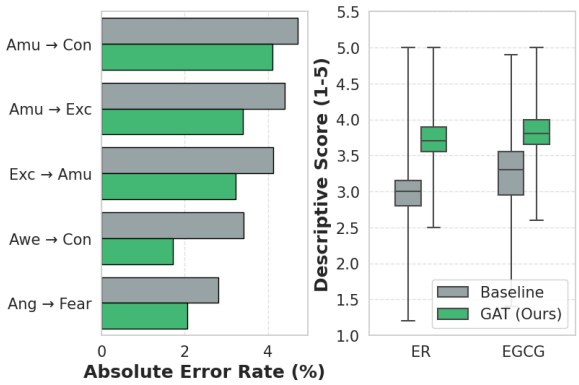


Figure 7: GAT corrects major intensity-confusion errors in EU (left) and improves descriptive depth for ER and EGCG (right). Abbreviations: Amu = Amusement, Con = Contentment, Exc = Excitement, Awe = Awe, Ang = Anger, Fear = Fear.

6 Conclusion

We introduced AICA-Bench, a holistic benchmark evaluating affective intelligence in vision-language models through perception, reasoning, and generation tasks. Our study reveals weaknesses in VLMs—particularly intensity miscalibration and shallow descriptive grounding—and demonstrates that GAT Prompting, a lightweight test-time strategy, substantially mitigates these issues across models. While establishing a foundation for multimodal affective understanding, broader challenges remain. We hope this advances research toward VLMs that understand, reason, and generate emotional content more accurately and naturally.

Limitations

While AICA-Bench provides a holistic and much-needed perspective on evaluating multimodal affective intelligence, it represents an initial step rather than a complete solution. However, several challenges remain intrinsic to this domain and are not fully addressed in our work. Emotional interpretation is inherently subjective and culturally situated, making it difficult to construct a benchmark that captures the full spectrum of human affect across demographics or social contexts. Our tasks operate on static images and English prompts, leaving broader settings—such as cross-cultural affective reasoning, temporal emotion dynamics, or multilingual understanding—outside our current scope. In addition, although our assessor offers consistent scoring across open-ended tasks, any automated judgment of emotional nuance inevitably carries modeling biases and may not fully approximate human interpretation. Finally, GAT Prompting focuses on test-time steering; its interaction with fine-tuning, memory-based reasoning, or personalized affect modeling remains unexplored. These limitations open important opportunities for future work toward developing more culturally aware, temporally grounded, and human-aligned affective VLMs.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Sree Bhattacharyya and James Z. Wang. 2025. Evaluating vision-language models for emotion recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1798–1820, Albuquerque, New Mexico. Association for Computational Linguistics.

Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. In *NeurIPS*.

Soumik Chakrabarti. 2020. Felzenszwalb segmentation: Python implementation of efficient graph-based image segmentation. https://github.com/soumik12345/felzenszwalb_segmentation. Accessed: 2025-12-10.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Lancheng Gao, Ziheng Jia, Yunhao Zeng, Wei Sun, Yiming Zhang, Wei Zhou, Guangtao Zhai, and Xiongkuo Min. 2025. Eemo-bench: A benchmark for multimodal large language models on image evoked emotion assessment. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 7064–7073, New York, NY, USA. Association for Computing Machinery.

He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. 2025. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *Preprint*, arXiv:2502.04424.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. 2017. Emotic: Emotions in context dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2309–2317.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. 2020. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2025. Scaffolding coordinates to

promote vision-language coordination in large multimodal models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2886–2903, Abu Dhabi, UAE. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C.H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. 2025. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Forty-second International Conference on Machine Learning*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023a. Holistic evaluation of language models. *Transactions on Machine Learning Research*. ArXiv:2211.09110.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023b. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. 2023b. Query-relevant images jailbreak large multi-modal models. *ArXiv*, abs/2311.17600.

Yexiang Liu, Zekun Li, Zhi Fang, Nan Xu, Ran He, and Tieniu Tan. 2025. Rethinking the role of prompting strategies in LLM test-time scaling: A perspective of probability theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27962–27994, Vienna, Austria. Association for Computational Linguistics.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model annbsp;all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *Preprint*, arXiv:2405.20797.

Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 83–92, New York, NY, USA. Association for Computing Machinery.

Laurent Mertens, Elahe Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. 2024. Findingemo: An image dataset for emotion recognition in the wild. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868.

Rosalind W. Picard. 1997. *Affective Computing*. The MIT Press.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark fornbsp;visual question answering using world knowledge. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, page 146–162, Berlin, Heidelberg. Springer-Verlag.

João Sedoc, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2017. Predicting emotional word ratings using distributional representations and signed clustering. In *Proceedings of the 15th Conference of the European*

Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 564–571, Valencia, Spain. Association for Computational Linguistics.

Shezheng Song, Chengxiang He, Shasha Li, Shan Zhao, Chengyu Wang, Tianwei Yan, Xiaopeng Li, Qian Wan, Jun Ma, Jie Yu, and Xiaoguang Mao. 2024. Mosabench: Multi-object sentiment analysis benchmark for evaluating multimodal large language models understanding of complex image. *Preprint*, arXiv:2412.00060.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. 2024. How many are in this image a safety evaluation benchmark for vision llms. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LI*, page 37–55, Berlin, Heidelberg. Springer-Verlag.

Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. 2023. What makes for good visual tokenizers for large language models? *Preprint*, arXiv:2305.12223.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20326–20337.

Jufeng Yang, Dongyu She, and Ming Sun. 2017. Joint image emotion classification and distribution learning via deep convolutional neural network. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3266–3272.

Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *Preprint*, arXiv:2406.16442.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Y. Yao, T. Yu, A. Zhang, J. Liu, H. Wang, and X. Chen. 2025. Efficient gpt-4v level multimodal large language model for deployment on edge devices. *Nature Communications*, 16:5509.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: the fine print and the benchmark. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 308–314. AAAI Press.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025b. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A More Details of AICA-Bench

A.1 Datasets

To construct a comprehensive and diverse affective image benchmark, we select 9 publicly available datasets spanning various sources, labeling models, and visual styles. These datasets include both social and artistic domains, and cover a range of emotion annotation schemes, such as Ekman's basic emotions, the Mikels model, and Plutchik's wheel. Table 5 summarizes the key statistics and features of each dataset used in AICA-Bench.

- **EMOTIC(Kosti et al., 2017)**: Real-world images featuring individuals in diverse everyday contexts, annotated with Ekman's basic emotion categories. The dataset is designed to support *Expressed Emotion Recognition* based on visual cues within natural scenes.
- **EmoSet-118K(Yang et al., 2023)**: A subset of a large-scale dataset with human-annotated labels, designed primarily for *Evoked Emotion Prediction*. It provides rich multi-level emotion attributes, including low-level (e.g., color, brightness), mid-level (e.g., scene type, object category), and high-level (e.g., facial expression, human action) features, following the Mikels emotion model.
- **FindingEmo(Mertens et al., 2024)**: A dataset of 25,000 images tailored for *Expressed Emotion Recognition*, where each image is annotated with Plutchik emotion labels.
- **FI(You et al., 2016)**: A large-scale dataset from Flickr and Instagram, annotated using the Mikels model. Contains diverse everyday scenes and subtle affective cues in social media imagery.
- **Emotion6(Peng et al., 2015)**: Small-scale dataset labeled with six Ekman emotions plus neutral, offering clean and balanced data for *Evoked Emotion Prediction*.
- **FlickrLDL(Yang et al., 2017)**: Applies label distribution learning (LDL) on images from Flickr using the Mikels model, capturing the uncertainty of emotion perception through soft labels.
- **TwitterLDL(Yang et al., 2017)**: Similar to FlickrLDL but sourced from Twitter. Annotated via LDL under the Mikels model, reflecting noisy and ambiguous affective signals in social content.
- **Abstract(Machajdik and Hanbury, 2010)**: Consists of abstract paintings labeled with the Mikels model. Emotions are conveyed through shape, color, and form rather than representational content.
- **ArtPhoto(Machajdik and Hanbury, 2010)**: Aesthetic photographic images labeled with the Mikels model. Explores affect elicited through compositional and stylistic elements in artistic photos.

A.2 Instructions

To support diverse affective tasks, we construct four types of natural language instructions within the AICA-Bench benchmark: **EU Basic**, **EU CoT**, **ER**, and **EGCG**. In total, we generate **18,124** instructions covering a wide range of image-emotion combinations across all task types. Each instruction type corresponds to a specific task category. **EU Basic** and **EU CoT** are designed for Emotion Understanding. **ER** (Emotion Reasoning) instructions require the model to explain the cause of a given emotion in an image. **EGCG** (Emotion-guided Content Generation) prompts the model to generate emotionally aligned descriptions based on both visual input and a target emotion. Together, these instructions enable fine-grained evaluation of a model's ability to comprehend, reason about, and generate emotionally meaningful content.

Example Prefix of Instructions templates for Expressed Emotion Recognition in EU Basic

- "Identify the emotion displayed by the person in the image."
- "What emotion is the person in the picture showing?"
- "Determine the person's emotional expression in this photo."
- "Classify the emotion of the person visible in the image."
- "From the visual cues, what emotion is the person exhibiting?"
- "..."

Table 5: Datasets Used in the AICA Benchmark

Dataset	Year	Original Image Count	Images Used in AICA	Proportion in AICA	Type	Emotion Label Model
EMOTIC	2019	23,571	1,000	12.37%	social	Ekman
EmoSet-118K	2023	118,102	1,000	12.37%	social	Mikels
FindingEmo	2024	25,000	1,000	12.37%	social	Plutchik
FI	2016	23,308	1,000	12.37%	social	Mikels
Emotion6	2015	1,980	1,000	12.37%	social	Ekman
FlickrLDL	2017	10,700	1,000	12.37%	social	Mikels
TwitterLDL	2017	10,045	1,000	12.37%	social	Mikels
Abstract	2010	280	280	3.46%	abstract	Mikels
ArtPhoto	2010	806	806	9.97%	artistic	Mikels

Example Prefix of Instructions templates for Evoked Emotion Prediction in EU Basic

- "Identify the emotion this image is likely to evoke in a human viewer."
- "Based on the visual content, what emotion would a typical observer feel?"
- "Determine the emotional response this image is meant to trigger."
- "Classify the feeling this image might evoke in someone who sees it."
- "..."

For the **EU Basic** task, we construct the instruction set to directly prompt models to identify the most appropriate emotion from a predefined list. This task is divided into two subtypes: *Expressed Emotion Recognition* and *Evoked Emotion Prediction*. For each subtype, we use GPT-4o to generate 50 diverse question templates phrased in natural language, in order to enhance lexical variety and reduce model overfitting to fixed instruction patterns. These templates serve as natural language *prefixes*, generalizable across datasets and capturing variation in phrasing, tone, and focus.

To create a complete instruction for a given image instance, we concatenate a selected prefix with the emotion label space corresponding to the dataset's annotation schema (e.g., Ekman, Mikels, or Plutchik). Formally, we define a complete **EU Basic instruction** as:

$$\text{Instruction} = \text{Prefix}_i + \text{OptionList}(D_j) \quad (1)$$

where Prefix_i is the i-th natural language template, and $\text{OptionList}(D_j)$ denotes the set of emotion labels associated with dataset D_j. This dynamic construction ensures alignment between the instruction and the emotion taxonomy used in each dataset.

For the **EU CoT** setting, we design chain-of-thought style instructions to guide the model through a structured three-step reasoning process: *Observe*, *Interpret*, and *Select*. This format encourages the model to reflect on visual features before making a decision, aiming to improve both accuracy and interpretability.

We define two CoT templates corresponding to the two subtypes of Emotion Understanding:

- *Evoked Emotion Prediction*: The model is instructed to assess how an image may affect a typical viewer. It analyzes the scene based on brightness, colorfulness, scene type, and social context, then selects an emotion label from the provided list.

- *Expressed Emotion Recognition*: The model is asked to reason about the emotional state of a person depicted in the image by examining facial expressions, body posture, and environmental cues.

To illustrate how EU Basic instructions are instantiated for different subtypes, we present two representative examples. Figure 8 shows an *Evoked Emotion Prediction* prompt, where the model is asked to infer the likely emotional response of a typical viewer. In contrast, Figure 9 presents an *Expressed Emotion Recognition* prompt, which focuses on identifying the emotion explicitly displayed by the individual in the image.

EU: Evoked Emotion Prediction

<image> Which emotional reaction does this image most strongly suggest? Please provide the emotion label for the image. The label should be one of the following: Joy, Surprise, Anger, Disgust, Fear, Sadness. Using this format: label

Basic Prompt

<image> Instructions: You are given an image. Predict the emotion it may evoke in a typical viewer.
1. **Observe:** Analyze the image in terms of brightness, colorfulness, scene type, objects, facial expressions, and human actions.
2. **Interpret:** Based on these visual features, infer what kind of emotion the image is likely to evoke in an observer.
3. **Select:** Choose one emotion from the list that best describes this likely emotional response.
Emotion Options: Joy, Surprise, Anger, Disgust, Fear, Sadness
Note: Only write the selected emotion on the last line.

CoT Prompt

Figure 8: Example of an Evoked Emotion Prediction instruction in the EU Basic and CoT setting.

EU: Expressed Emotion Recognition

<image>From the visual cues, what emotion is the person exhibiting? Please provide the emotion label for the image. The label should be one of the following: Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning. Using this format: label

Basic Prompt

<image>Instructions: You are given an image of a person. Identify the emotion expressed through their facial features.
1. **Observe:** Describe the person's body posture, facial expression, and surrounding environment.
2. **Interpret:** Based on the visual context, infer how the person might feel.
3. **Select:** Choose one emotion from the list that best fits the person's emotional state.
Emotion Options: Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning
Note: Only write the selected emotion on the last line.\n

CoT Prompt

Figure 9: Example of an Expressed Emotion Recognition instruction in the EU Basic and CoT setting.

Prompt Used for ER and EGCG Instruction Generation

You are building a benchmark to evaluate a vision-language model's emotional reasoning and emotional content generation abilities.
You will receive an image and an emotion label.
Your task is to produce two distinct tasks: 1. Emotion Reasoning: A reasoning question that asks the model to explain **why** the given emotion is being expressed in the image. 2. Emotion-Guided Content Generation: Generate a writing instruction that asks the model to generate a short paragraph (50–100 words) using the target emotion tone, grounded in the image. Then provide an expressive answer.
Format your response like this:
Reasoning Question: ...
Reasoning Answer: ...
Generation Instruction: ...
Generation Answer: ...

To construct instructions for the **ER** and **EGCG** tasks, we leverage GPT-4o to generate natural language prompts along with corresponding sample answers. Each generation is conditioned on an image-emotion pair, and the model is asked to produce two distinct instruction types:

- *Emotion Reasoning*: A question that requires the model to explain why the given emotion is appropriate for the image.
- *Emotion-guided Content Generation*: A creative writing instruction that asks the model to describe the image using the tone or perspective of the given target emotion.

We use a single, structured system prompt to generate both instruction types in a consistent and controlled manner. A representative example of this prompt is shown in Figure A.2. To illustrate how

these instructions are applied in practice, Figure 10 presents an example based on a specific image, demonstrating how the model is guided to explain the given emotion and produce a corresponding emotionally aligned description.

A.3 Annotation for AICA-Bench

To construct the visual foundation of AICA-Bench, we relied on 20 trained annotators recruited from our institution and collaborating labs. The annotators were students (including undergraduate and M.S. students), typically 20–27 years old, and were required to be proficient in English to follow the annotation guidelines. They had prior experience in affective computing or related areas. Participation was voluntary, and annotators were informed that they could stop at any time without penalty. Each annotator received a flat honorarium of $20 for their contribution, which corresponds to a typical short annotation session in our locale and is comparable to the standard compensation level for student research assistants performing similar expert tasks. The 8,086 candidate images selected by automated scripts (Section 3.2) were divided into subsets and assigned to these annotators for manual review.

Before the main filtering phase, all annotators completed a brief training and calibration session. We provided written guidelines describing the goal of AICA-Bench (to build a clear and diverse affective image benchmark), the target emotion categories and valence–arousal space, and several worked examples of acceptable and unacceptable images. During training, annotators practiced on a small set of pilot images and received feedback to align their judgments. In the main annotation stage, we applied the two-step filtering process described in Section 3.2: images were first selected using automated scripts from a range of publicly available affective computing datasets, and then each candidate was manually reviewed by a trained annotator. Annotators were instructed to reject images that are unsafe for work, contain offensive or highly sensitive content, or are too emotionally ambiguous to support reliable labels; rejected samples were replaced to maintain both emotional clarity and dataset diversity.

The images used in AICA-Bench are drawn from existing publicly released affective image datasets that provide de-identified visual content under research-friendly licenses, and we do not collect any new personally identifying information about the individuals depicted. Our work focuses on secondary use and additional annotations over these public resources. Annotators were informed that their decisions would be used for research purposes and for releasing the benchmark to the community, and we store only anonymized annotator identifiers and binary keep/reject decisions. In line with the content policies of the source datasets, and consistent with our two-step filtering protocol, images that are NSFW, offensive, or otherwise inappropriate for an affective benchmark are removed prior to release, and annotators are encouraged to flag any remaining samples that may be sensitive or ambiguous so that they can be excluded from the final dataset.

A.4 Metrics

We use task-specific metrics to evaluate model performance across the AICA-Bench benchmark.

For the **EU** task, we report *Accuracy*, *Macro F1*, and *Weighted F1*. Among these, Weighted F1 is used as the primary metric, as it accounts for class imbalance across datasets.

For the open-ended generation tasks—**ER** and **EGCG**—we employ a fine-tuned scoring model to assess the quality of responses. For ER, we evaluate three dimensions: *Emotion Alignment*, which measures how well the response reflects the target emotion; *Descriptiveness*, which assesses the richness and specificity of the explanation; and *Causal Soundness*, which captures whether the reasoning provides a plausible link between the visual content and the stated emotion. For EGCG, we evaluate *Emotion Alignment* and *Descriptiveness* using the same criteria.

Table 6 summarizes the evaluation metrics used for each task.

A.5 VLM Models Details

To ensure a comprehensive evaluation, we include both closed-source and open-source vision-language models (VLMs). Within the open-source group, we select models at three parameter scales—2–3B, 7–8B, and 14–16B—covering a range of compute capacities and deployment settings.

Closed-source models. These models are accessed via API and represent strong general-purpose commercial systems. We include:

- **GPT-4o and GPT-4o-Mini**: The latest generation of OpenAI's VLMs with multimodal reasoning capabilities.

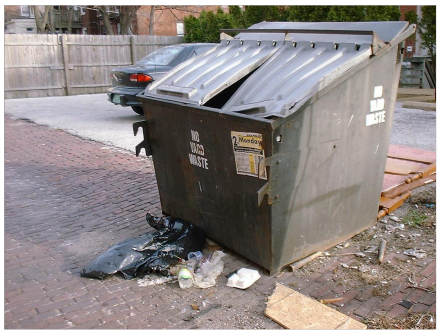
ER and ECG Prompt

<image> Why does the image of the ancient ruins at sunset evoke a feeling of awe?

ER Prompt

"<image> Generate a short paragraph that captures the feeling of awe inspired by the image of ancient ruins at sunset, using vivid and emotive language.

ECG Prompt



<image> Why might the emotion of disgust be expressed when viewing this image?

ER Prompt

"<image> Write a short paragraph (50–100 words) that captures the essence of disgust, inspired by the image.

ECG Prompt

<image> Why might the emotion of sadness be associated with the image of the dog lying down with a chain around its neck?

ER Prompt

“<image> Write a short paragraph (50-100 words) conveying a sense of sadness, inspired by the image of the dog lying down with a chain around its neck.

ECG Prompt

Figure 10: Examples of an ER and EGCG.

Table 6: Evaluation metrics used for each task in AICA-Bench

Task	Evaluation Metrics
Emotion Understanding (EU)	*Accuracy, Macro F1, Weighted F1*
Emotion Reasoning (ER)	*Emotion Alignment, Descriptiveness, Causal Soundness*
Emotion-guided Content Generation (EGCG)	*Emotion Alignment, Descriptiveness*

- **Gemini-2.5-Pro / Flash / 2.0-Flash**: Google's vision-language models optimized for both generation and perception tasks.
- **Qwen-VL-Max / Plus**: High-performing commercial VLMs developed by Alibaba, designed for instruction-following in multimodal contexts.

Open-source models. We include diverse open-source models from different families and architecture designs. All selected models support vision input and are instruction-tuned.

- **Qwen series** (2B, 3B, 7B):
  - *Qwen2-VL-2B / Qwen2.5-VL-3B / Qwen2.5-VL-7B*: Lightweight VLMs designed for practical deployment with strong performance in perception-based tasks.
- **InternVL series** (2B, 8B, 14B):
  - *InternVL3-2B / 8B / 14B, InternVL2.5-8B*: VLMs optimized for high-resolution visual reasoning, with a multi-scale fusion backbone.
- **Ovis series** (2B, 8B, 16B):
  - *Ovis2-2B / 8B / 16B*: Focused on vision-language alignment and long-form generation, designed for instruction-based multimodal understanding.

- **LLaVA series** (7B, 13B):
  - *LLaVA-1.6-7B / 13B, LLaVA-OneVision*: Popular open-source VLMs pretrained with image-text alignment and dialogue fine-tuning.
- **MiniCPM series**:
  - *MiniCPM-V / MiniCPM-O-2.6*: Compact VLMs designed for mobile and edge deployment, with surprisingly strong performance on vision tasks.

B Detail of AICA-Bench Scoring Model

B.1 Dataset Details

To effectively train our scoring model, we meticulously constructed a high-quality human-annotated dataset. This dataset serves as the foundation for fine-tuning our scoring model, enabling it to learn the nuances of human judgment regarding MLLM output.

Figure 11 illustrates the custom human annotation interface employed for collecting high-quality data for our scoring model. This platform facilitated expert evaluators in assessing model-generated responses (e.g., for emotion reasoning tasks) against predefined criteria, such as 'emotional_alignment', 'descriptiveness', and 'causal_soundness', ensuring a rigorous and consistent scoring process.

The creation of this dataset involved five domain experts independently evaluating 10K distinct data samples. Each sample comprised an image-based question, the MLLM's generated answer, and a set of predefined evaluation criteria. The experts provided scores for each criterion, reflecting the quality of the MLLM's response. This human-annotated data was then formatted into a specialized instruction-tuning dataset, specifically designed for training our scoring model.

Figure 12 provides illustrative examples of the instruction-tuning format used in our dataset. As depicted, each entry includes a "Question" prompt that instructs the expert evaluator on the task (e.g., Emotion Reasoning, Emotion-Guided Content Generation) and the specific criteria for assessment. An "Input Sample" section presents the original image-based prompt and the MLLM's generated "Output" response. Crucially, the "Evaluation Criteria" section lists the specific aspects to be rated on a 1-5 scale, such as "Emotional Alignment," "Descriptiveness," and "Causal Soundness." Finally, the "Ground Truth" section contains the actual scores provided by the human experts in a JSON format, serving as the target labels for our scoring model's training. This structured format ensures that our fine-tuning process directly aligns the model's output with expert human assessments of MLLM response quality.

B.2 Training Details

Our scoring model was fine-tuned from the Qwen2.5-VL-7B model using the LLaMA-Factory framework with Low-Rank Adaptation (LoRA). This approach was chosen for its efficiency in adapting large pre-trained models with reduced computational overhead. Training was performed on a single NVIDIA A100 80GB GPU, leveraging a high-quality human-annotated dataset. The detailed training configuration is summarized in Table 7. Key parameters, including an effective batch size of 8, a learning rate of 1.0e-4, a cosine learning rate schedule with a 0.1 warmup ratio, and 3 training epochs, were optimized to ensure stable convergence and prevent overfitting on our specialized dataset.

B.3 More Experiment Results

B.4 Comprehensive Model Comparison

Table 8 provides the full evaluation results for all closed-source and open-source models assessed in the AICA-Bench benchmark. While the main paper reports the aggregated findings, this appendix table offers a complete breakdown across five metrics: EU Basic, EU CoT, EU Average, ER Average, and EGCG Average, as well as the final overall average score.

B.5 Performance Radar Charts by Model Series

To complement the quantitative leaderboard presented in the main paper, Figure 13 visualizes the performance of all evaluated VLMs across seven core metrics using radar charts. Each subplot corresponds to a distinct model series—either open-source or closed-source—and includes all variants within that series. The plotted axes cover: EU Basic, EU CoT, ER Emotional Alignment (ER EA), ER Descriptiveness (ER Desc.), ER Causal Soundness (ER CS), EGCG Emotional Alignment (EGCG EA), and EGCG Descriptiveness (EGCG Desc.).

Radar chart comparisons reveal several consistent trends across model series. Closed-source models generally demonstrate strong and balanced performance across all affective tasks. In contrast,

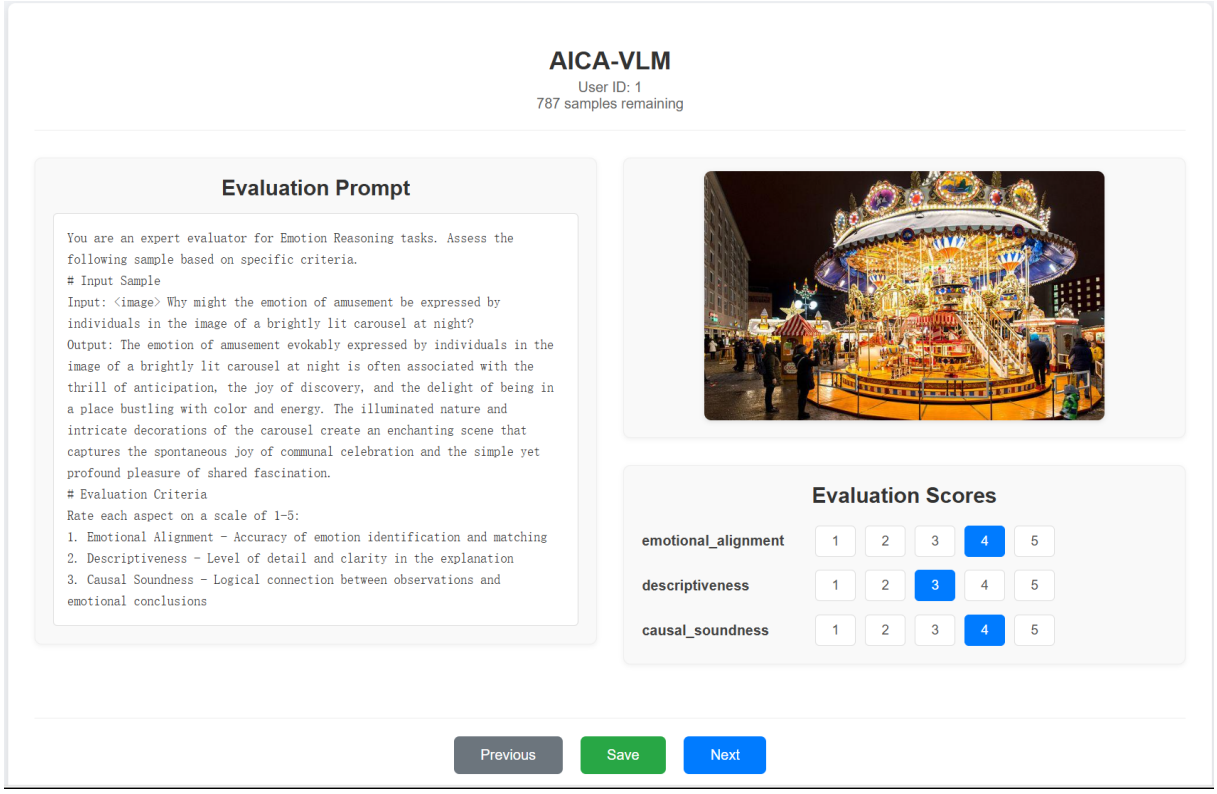


Figure 11: Human Annotation Interface

open-source models often achieve high scores in reasoning tasks (e.g., ER-EA, ER-CS) but underperform in emotion-guided generation, especially in descriptive quality. Additionally, within many model families, performance improvements do not scale uniformly with model size, highlighting the influence of optimization and alignment strategies.

B.6 Analysis of Model Size vs. Performance

To substantiate our observation that model size is not a reliable predictor of affective performance, we further analyze the scores of open-source models across four evaluation dimensions: *EU Avg.*, *ER Avg.*, *EGCG Avg.*, and *Overall Avg.*, plotted in Figure 14. Each curve represents a model series, and markers indicate performance at different parameter scales.

Several patterns reinforce our conclusion:

- **Qwen2.5VL-7B (65.78%)** achieves the highest overall average score among all open-source models, outperforming larger models such as **InternVL3-14B (62.41%)** and **LLaVA-1.6-13B (59.96%)**, underscoring that increased parameter size does not guarantee superior affective capability.

- **Qwen2.5VL-3B (61.81%)** outperforms multiple larger models such as **InternVL3-14B (62.41%)** and **LLaVA-1.6-13B (59.96%)**, and exhibits particularly strong performance on *ER Avg.* and *EGCG Avg.*.

- **Ovis2-8B (64.29%)** achieves higher average scores than the 14B and 13B models of other series, including InternVL3 and LLaVA, indicating that even mid-sized models can be highly competitive.

- Across series, performance fluctuations do not exhibit monotonic growth with scale: for example, within InternVL3, the smallest model (2B) performs worse overall (56.14%), but the 8B variant (62.35%) and the 14B (62.41%) are nearly tied, suggesting diminishing returns.

- In the EGCG dimension in particular, smaller models like **Qwen2VL-2B (58.64%)** and **MiniCPM-2.6 (58.08%)** are competitive with or even outperform some 7B–13B models.

Interestingly, when focusing specifically on the *EGCG Avg.* dimension, a more noticeable trend emerges: models with larger parameter sizes tend to perform better. For instance, **Ovis2-16B** achieve higher EGCG scores (71.56%) compared to their smaller counterparts like **Ovis2-2B (66.04%)**. This

Question
You are an expert evaluator for Emotion Reasoning tasks. Assess the following sample based on specific criteria.

Input Sample
Input: <image> Why might the emotion of amusement be expressed by individuals in the image of a brightly lit carousel at night?
Output: The emotion of amusement evokably expressed by individuals in the image of a brightly lit carousel at night is often associated with the thrill of anticipation, the joy of discovery, and the delight of being in a place bustling with color and energy. The illuminated nature and intricate decorations of the carousel create an enchanting scene that captures the spontaneous joy of communal celebration and the simple yet profound pleasure of shared fascination.

Evaluation Criteria
Rate each aspect on a scale of 1-5:
1. Emotional Alignment - Accuracy of emotion identification and matching
2. Descriptiveness - Level of detail and clarity in the explanation
3. Causal Soundness - Logical connection between observations and emotional conclusions

Response Format
Provide scores in JSON format only:
{
"emotional_alignment": <1-5>,
"descriptiveness": <1-5>,
"causal_soundness": <1-5>
}

Ground Truth
{
"emotional_alignment": 4,
"descriptiveness": 3,
"causal_soundness": 4
}

Question
You are an expert evaluator for Emotion-Guided Content Generation tasks. Assess the following sample based on specific criteria.

Input Sample
Input: <image> Write a short paragraph using a tone of fear, inspired by the image provided. The paragraph should describe the scene and evoke a sense of dread or anxiety.

Output: The image depicts a person dressed in a white lab coat, holding a beaker filled with a mysterious green liquid. The individual is wearing a mask, which adds to the eerie atmosphere. The setting appears to be a dark room, with the figure standing in the shadows. The overall tone of the image is one of fear and uncertainty, as the viewer is left to wonder about the intentions of the person in the lab coat.

Evaluation Criteria
Rate each aspect on a scale of 1-5:
1. Emotional Alignment(1-5) : Does the generated content match the target emotion?
2. Descriptiveness(1-5) : How detailed and vivid is the generated content?

Response Format
Provide scores in JSON format only:
{
"emotional_alignment": <score>,
"descriptiveness": <score>
}

Ground Truth
{
"emotional_alignment": 2,
"descriptiveness": 2
}

Figure 12: Examples of the Instruction-Tuning Dataset Format

suggests that model size might play a more significant role in generation tasks where maintaining coherence, style, and emotional grounding in generated content requires greater model capacity.

We speculate that this correlation may be due to the inherently more complex nature of text or image generation in affective settings, which benefits from the richer representation capabilities and longer context handling of larger models. Nevertheless, the trend is still not strictly linear, as some smaller models like **Qwen2.5VL-3B** and **Qwen2VL-7B** still demonstrate competitive performance, reinforcing the importance of model training and alignment strategies.

C GAT Prompting

This section outlines the concrete prompts used by the Grounded Affective Tree (GAT) framework across the three tasks in AICA-Bench. Each prompt is designed following the principles described in Section 5.2, combining region-level visual scaffolding with structured affective reasoning or grounded generation. In addition to the prompts, we also describe the implementation of visual scaffolding, which serves as a key component in guiding the model's understanding and generation of affective content. Furthermore, we provide results from human evaluation to validate the effectiveness of GAT, ensuring that the improvements are not only supported by automatic metrics but also by human judgments.

Visual Scaffolding Implementation. For image segmentation, we use the graph-based image segmentation method that divides the image into regions based on pixel similarities. The segmentation is controlled by three parameters: `scale`, `sigma`, and `min_size_ratio`. After the initial segmentation, we dynamically merge the smallest regions until the number of regions is reduced to a target range (2–4 regions) based on the adjacency of regions and their relative areas. The process was implemented in Python using the `skimage` library and is available on GitHub (Chakrabarti, 2020).

Figure 15 presents representative examples of the Visual Scaffolding used in GAT Prompting. Each image is decomposed into several coarse regions using graph-based segmentation, where each region is assigned a unique numerical ID (e.g., 1, 2, 3...). Subsequently, these IDs and contours are overlaid on the raw image before feeding it into the VLM. These annotated regions serve as explicit visual anchors for downstream prompting, providing a

Table 7: Scoring model training configuration

Parameter	Value
Base Model	Qwen2.5-VL-7B
Fine-tuning Framework	LLaMA-Factory
Fine-tuning Method	LoRA
GPU	1 × NVIDIA A100
Effective Batch Size	8 (per-device: 2, gradient accumulation: 4)
Learning Rate	1.0e-4
Learning Rate Scheduler Type	Cosine
Warmup Ratio	0.1
Number of Training Epochs	3

Model	EU Basic	EU CoT	EU Avg.	ER Avg.	EG Avg.	Overall Avg. (%)
Closed-source Models						
Gemini-2.5-pro	66.97	67.57	67.27	79.08	74.13	**73.49**
Qwen-VL-max	64.07	65.98	65.02	77.75	75.93	**72.90**
ChatGPT-4o	64.44	65.42	64.93	77.81	75.73	**72.82**
Gemini-2.5-flash	68.05	69.32	68.68	76.55	68.19	**71.14**
ChatGPT-4o-mini	60.15	63.68	61.91	76.45	74.09	**70.81**
Qwen-VL-plus	60.04	67.81	63.92	72.39	66.86	**67.73**
Gemini-2.0-flash	67.16	68.98	68.07	71.05	63.93	**67.68**
Open-source Models						
Qwen2.5VL-7B (Bai et al., 2025)	56.43	57.25	56.84	74.50	66.00	**65.78**
Ovis2-16B (Lu et al., 2024)	54.38	54.70	54.54	68.24	71.56	**64.78**
Ovis2-8B (Lu et al., 2024)	53.63	52.73	53.18	68.89	70.81	**64.29**
InternVL3-14B (Zhu et al., 2025b)	52.91	52.04	52.47	68.27	66.50	**62.41**
InternVL3-8B (Zhu et al., 2025b)	52.18	52.98	52.58	67.21	67.27	**62.35**
InternVL2.5-8B (Chen et al., 2024)	51.89	51.03	51.46	66.48	68.86	**62.27**
MiniCPM-o-2.6 (Yao et al., 2025)	52.73	48.65	50.69	70.16	64.98	**61.94**
Qwen2.5VL-3B (Bai et al., 2025)	54.58	53.67	54.13	69.68	61.63	**61.81**
Qwen2VL-7B (Wang et al., 2024)	53.52	55.19	54.36	65.23	64.76	**61.45**
Ovis2-2B (Lu et al., 2024)	49.32	46.45	47.88	66.38	66.04	**60.10**
LLaVA-1.6-13B (Liu et al., 2024b)	36.78	46.82	41.80	73.57	64.51	**59.96**
LLaVA-1.6-7B (Liu et al., 2024b)	36.58	50.22	43.40	73.81	59.58	**58.93**
Qwen2VL-2B (Wang et al., 2024)	51.85	49.38	50.62	67.33	57.97	**58.64**
MiniCPM-V-2.6 (Yao et al., 2024)	43.70	47.25	45.48	65.77	63.00	**58.08**
LLaVA-onevision (Li et al., 2024)	54.02	53.25	53.64	63.78	54.18	**57.20**
InternVL3-2B (Zhu et al., 2025b)	43.87	39.53	41.70	65.78	60.93	**56.14**

Table 8: Performance comparison on AICA-Bench tasks. EU: Emotion Understanding, ER: Emotion Reasoning, EGCG: Emotion-guided Content Generation. Scores are shown as percentages.

structured basis for guiding the model to identify objective visual cues before performing affective reasoning or generation.

Human evaluation. In addition to the automatic evaluation metrics reported in Table 4, we further conduct a human evaluation to validate the effectiveness of GAT from a human judgment perspective. Specifically, we randomly sample 200 instances and select two representative models, GPT-4o and Qwen2.5-VL-7B. For each instance, we collect both the Baseline and GAT-generated responses for the ER and EGCG tasks. These paired responses are then evaluated by 10 human annotators, who independently assign scores following the same affective criteria used in AICA-Bench, including emotion alignment and descriptiveness for both tasks, as well as causal soundness for ER. All scores are given on a 1–5 Likert scale. We report the mean ratings aggregated across annotators and samples in Table 9. The results show that GAT consistently receives higher human ratings across models and tasks, providing complementary evidence to the automatic evaluation results.

GAT Prompt for EU (Tree-of-Thoughts Strategy).

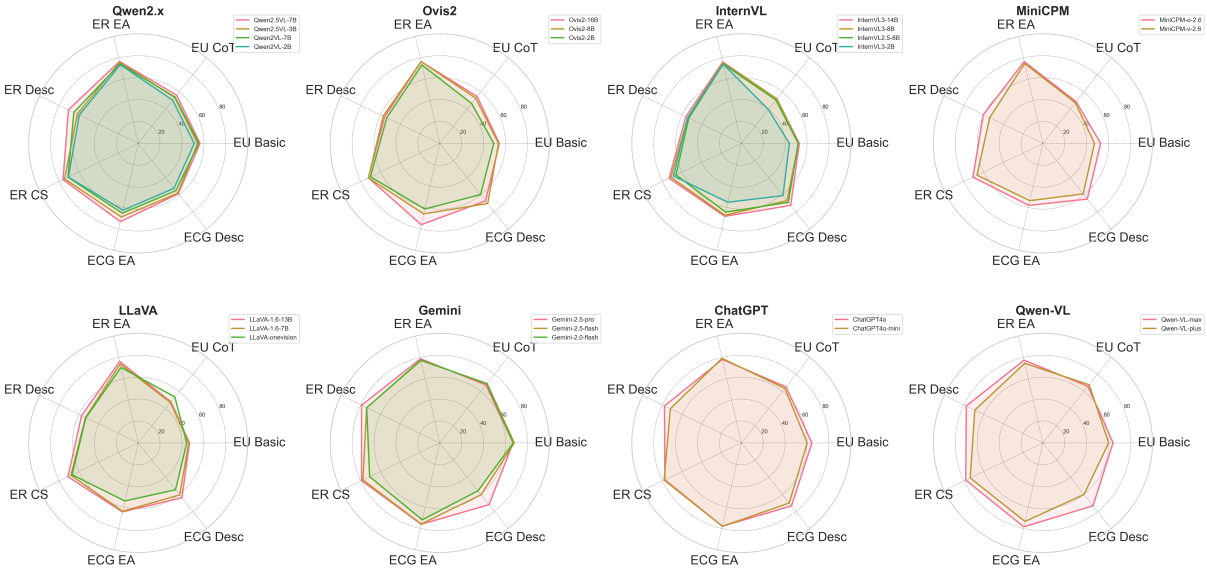


Figure 13: Radar charts comparing performance across tasks for eight major VLM model series (both open-source and closed-source)

Table 9: Human evaluation on ER and EGCG (200 images, 10 annotators). Values are mean ratings (no ±).

Model	**ER**								**EGCG**					
	Align		Desc		Causal		Overall		Align		Desc		Overall	
	Base	GAT	Base	GAT	Base	GAT	Base	GAT	Base	GAT	Base	GAT	Base	GAT
GPT-4o	4.12	4.36	3.05	3.61	3.18	3.74	3.45	**3.90**	4.20	4.48	3.22	3.79	3.71	**4.02**
Qwen2.5-VL-7B	3.88	4.14	2.94	3.47	3.02	3.56	3.28	**3.72**	3.95	4.21	3.10	3.62	3.53	**3.86**

Stage 1: Indexed Region Observation (Root Node)

Input: An image with numbered segmented regions.

Task: Systematically describe the visual content of each marked region based on its ID.

Instruction:
For each numbered region visible in the image, describe its key visual clues (shapes, textures, object parts, or interactions). Do NOT infer emotions yet. Focus on objective description.
Output Format:

- Region 1: [Description of visual elements in this specific area]
- Region 2: [Description of visual elements in this specific area]
- ...
- Region N: [Description...]

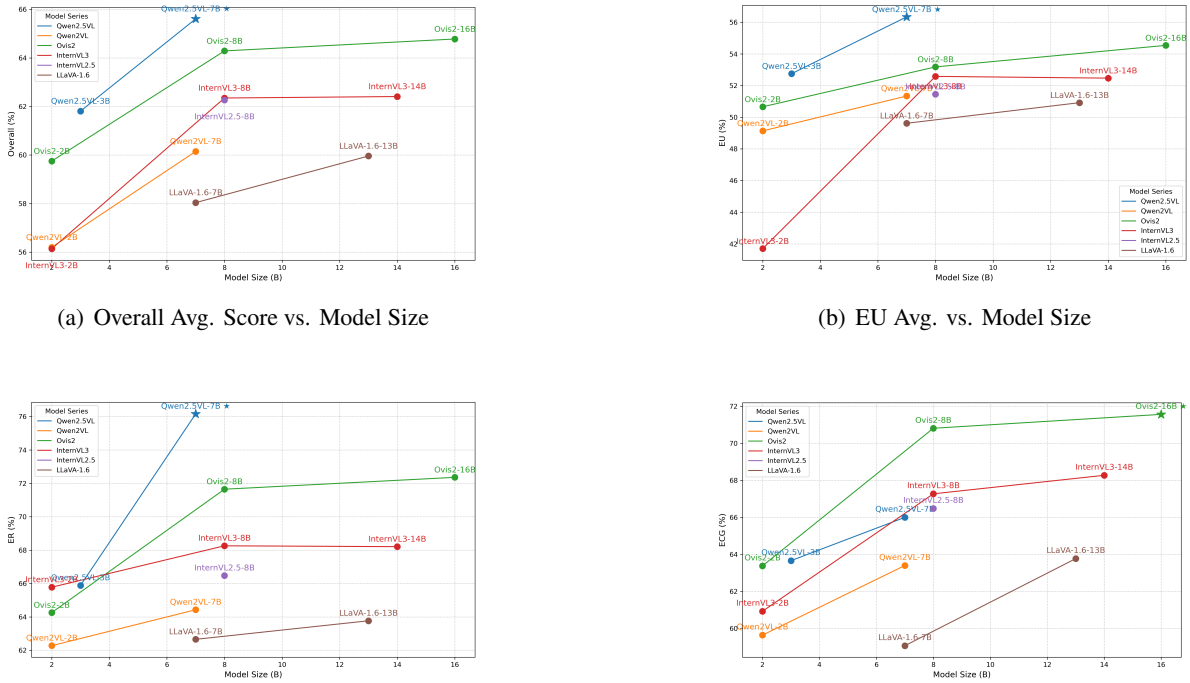
Stage 2: Candidate Generation (Branching)

Input: The descriptions of Regions 1 to N and the global polarity estimate.

Goal: Propose three distinct emotion candidates. You must cite specific Region IDs to justify your choice, specifically explaining the intensity.

Instruction:
Generate three hypotheses. For each, explain which Region number supports the emotion and its intensity level.

- Hypothesis A: [Emotion Label] (Intensity: Low/Medium/High)
  Evidence: Region [X] shows [detail] which suggests [level] arousal. Region [Y] shows [detail]...
- Hypothesis B: [Different Emotion Label] (Intensity: Different Level)
  Evidence: Region [Z] contradicts the previous hypothesis because it shows...
- Hypothesis C: ...



(a) Overall Avg. Score vs. Model Size

(b) EU Avg. vs. Model Size

(c) ER Avg. vs. Model Size

(d) EGCG Avg. vs. Model Size

Figure 14: Performance comparison across subtasks (EU, ER, EGCG) and overall average scores for open-source models. Each line represents a model series, with the best-performing model in each subtask marked by a star.

Stage 3: ID-Based Verification (Evaluation)

Role: You are a grounded verifier.
Input: A proposed Hypothesis (Emotion + Intensity) AND the Region Descriptions.

Instruction:
Verify if the contents of the cited Regions actually support the claimed intensity.

1. Fact Check: The hypothesis claims Region [X] shows evidence of High Intensity. Look at the description of Region [X]. Is this true? (e.g., If Region X is a "relaxed hand", it contradicts "High Excitement").

2. Completeness: Do other Regions (e.g., Region [Y]) suggest a different conclusion?

Verdict:

- Score (0.0 - 1.0): Probability that this hypothesis is correct based strictly on the regions.

- Critique: Mention specifically if a Region ID was misinterpreted.

GAT Prompt for ER.

Emotion Reasoning (ER)

You are given an image with segmented regions.

Step 1 – Extract visual clues.
Identify the important visual elements from the regions (people, actions, objects, background, lighting, color, and composition).

Step 2 – Explain the emotion.
Using only these grounded clues, explain which emotion the image conveys and why. Refer explicitly to the regions or elements that support your interpretation.

Provide a concise, evidence-based explanation grounded entirely in the visual content.

GAT Prompt for EGCG.

Emotion-Grounded Content Generation (EGCG)

You are given an image with segmented regions and a target emotion label.

Step 1 – Identify visual anchors.
From the segmented regions, list the key elements that should guide the emotion (e.g., characters, objects, environment, lighting, colors, and their relationships).

Step 2 – Generate an emotionally aligned description.
Write a short paragraph that is emotionally consistent with the target emotion and grounded in the listed visual elements. Use specific details rather than generic statements, and avoid adding objects or events that are not present in the image.

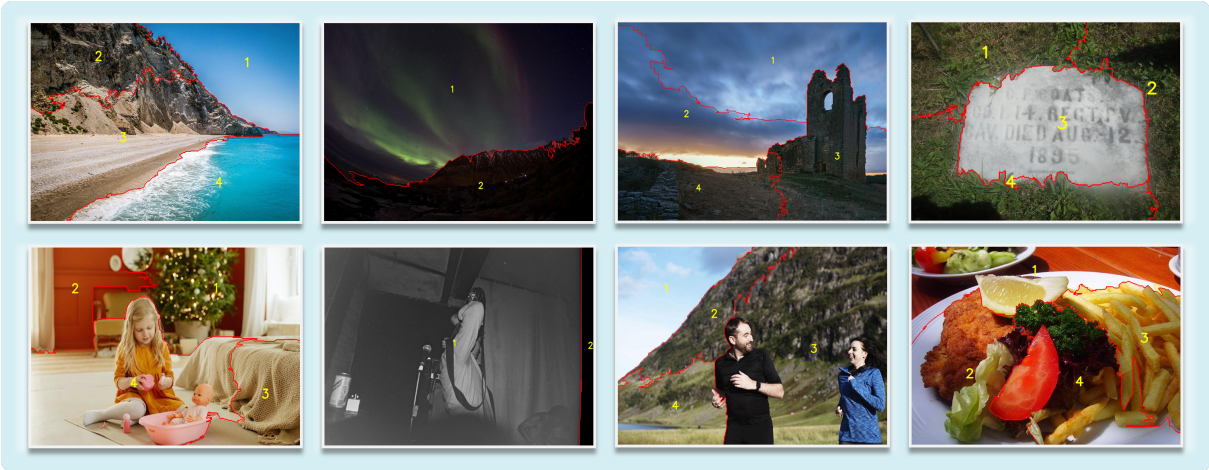


Figure 15: Visual Scaffolding examples used in GAT Prompting. The red contours highlight the segmented regions that serve as visual anchors for guiding the model's grounded reasoning.

D Experiments compute resources

All open-source models were evaluated using NVIDIA A100 GPUs on a cloud computing platform. Each evaluation task was executed on a single GPU, depending on model size. For most 7B models, each task (e.g., EU or ER) required approximately 1–2 hours to complete over the full benchmark set. Larger models (13B–16B) typically required 2–4 hours. Inference was run with float16 precision using the Hugging Face Transformers or official inference libraries where available.

Closed-source models were accessed via public APIs, including OpenAI, Google, and Alibaba endpoints. All requests were made under default inference settings without any batch acceleration.

E License and Intended Use

We will release the AICA-Bench artifacts under the CC BY 4.0 license for academic use. We follow the original terms of all source datasets used for secondary use and do not redistribute any restricted raw content; users should obtain any required original images from the respective sources under their terms. AICA-Bench is intended for research and benchmarking of affective image content analysis with vision-language models, and is not intended for high-stakes or sensitive applications or for identifying individuals.

F Show Case

F.1 Emotion Understanding Case

We present sample analyses of EU cases, including 5 correct and 8 error examples (see Figures 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 29) .

Figure 16: A sample correct case of EU-Basic.

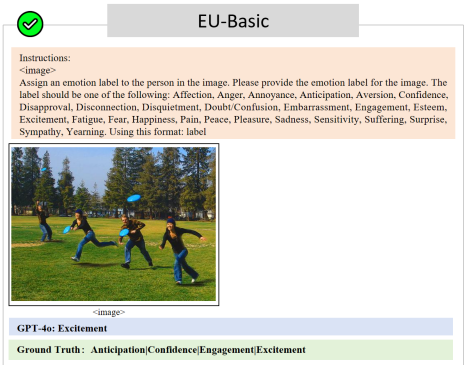


Figure 17: A sample correct case of EU-Basic.

EU-Basic

Instructions:
<image>
Which emotion is most likely to be felt when viewing this image? Please provide the emotion label for the image. The label should be one of the following: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sadness. Using this format: label

<image>

GPT-4o: Disgust

Error Category: Fine-grained category confusion

Error Reason: The model incorrectly labeled the emotion as Disgust instead of the correct Fear . The image depicts a menacing character (Red Skull from Marvel) with a stern, threatening expression. While both Disgust and Fear are negative emotions, they belong to distinct categories. **The error likely arose because the model conflated the character's intense, unpleasant appearance with disgust rather than recognizing the contextually appropriate fear response elicited by his villainous demeanor and aggressive posture.**

Ground Truth: fear

Figure 18: A sample error case of EU-Basic.

EU-Basic

Instructions:
<image>
Which emotion is most likely to be felt when viewing this image? Please provide the emotion label for the image. The label should be one of the following: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sadness. Using this format: label

<image>

GPT-4o: Contentment

Error Category: Valence-level misclassification

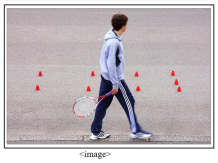
Error Reason: The model incorrectly labeled the emotion as Contentment (a positive/neutral valence) instead of the correct Sadness (a negative valence). The image shows a dog with a listless, vacant expression, drooping eyes, and a slack mouth, which typically signal sadness or melancholy. **However, the model may have focused on the dog's relaxed posture or the neutral background, misinterpreting these cues as contentment.** This error reflects a failure to distinguish between positive and negative emotional valences, likely due to limited contextual understanding of animal expressions or over-reliance on superficial features like body language without considering facial subtleties.

Ground Truth: sadness

Figure 19: A sample error case of EU-Basic.

EU-Basic

Instructions:
<image>
Can you tell what emotion this person is experiencing in the photo? Please provide the emotion label for the image. The label should be one of the following: Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning. Using this format: label



<image>

GPT-4o: Doubt/Confusion

Error Category: Valence-level misclassification

Error Reason: The model incorrectly labeled the emotion as Doubt/Confusion (a neutral/negative valence) instead of the correct Confidence (a positive valence). The image shows a person walking calmly with a tennis racket, surrounded by orange cones (likely for training drills). Their posture is relaxed and purposeful, suggesting self-assuredness in their activity. **However, the model may have fixated on the scattered cones or the monochrome tone of the image, interpreting these as ambiguous or disruptive cues that triggered a "confused" interpretation.** This error reflects a failure to distinguish between positive (confidence) and neutral/negative (doubt/confusion) emotional valences, likely due to overemphasis on contextual elements (e.g., cones) rather than the subject's composed demeanor.

Ground Truth: Confidence

Figure 20: A sample error case of EU-Basic.

EU-Basic

Instructions:
<image>
Given this image, classify the person's emotion as accurately as possible. Please provide the emotion label for the image. The label should be one of the following: Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning. Using this format: label

<image>

GPT-4o: Affection.

Error Category: Fine-grained category confusion

Error Reason: The model incorrectly labeled the emotion as Affection instead of the correct Engagement . While the scene depicts a warm, intimate interaction between two people in a cozy setting (shared blankets, candles, and close proximity), the key emotional cue is their mutual focus on each other during what appears to be a meaningful conversation. Affection implies romantic or familial warmth, whereas Engagement emphasizes active participation in shared activity or dialogue. **The model likely prioritized visual cues of physical closeness (e.g., entwined legs, proximity) over contextual indicators of collaborative interaction (e.g., eye contact, shared attention).** This reflects a failure to distinguish between closely related positive emotions based on subtle behavioral distinctions.

Ground Truth: Engagement

Figure 21: A sample error case of EU-Basic.

EU-CoT

Instructions:
<image>
You are given an image. Predict the emotion it may evoke in a typical viewer.
1. Observe: Analyze the image in terms of brightness, colorfulness, scene type, objects, facial expressions, and human actions.
2. Interpret: Based on these visual features, infer what kind of emotion the image is likely to evoke in an observer.
3. Select: Choose one emotion from the list that best describes this likely emotional response.
Emotion Options: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sadness
Note: Only write the selected emotion on the last line.

<image>

GPT-4o: Disgust

Ground Truth: disgust

Figure 22: A sample correct case of EU-CoT.

EU-CoT

Instructions:
<image>
You are given an image. Predict the emotion it may evoke in a typical viewer.
1. Observe: Analyze the image in terms of brightness, colorfulness, scene type, objects, facial expressions, and human actions.
2. Interpret: Based on these visual features, infer what kind of emotion the image is likely to evoke in an observer.
3. Select: Choose one emotion from the list that best describes this likely emotional response.
Emotion Options: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sadness
Note: Only write the selected emotion on the last line.

<image>

GPT-4o: Awe

Ground Truth: awe

Figure 23: A sample correct case of EU-CoT.

EU-CoT

Instructions:
<image>
You are given an image of a person. Identify the emotion expressed through their facial features.
1. Observe: Describe the person's body posture, facial expression, and surrounding environment.
2. Interpret: Based on the visual context, infer how the person might feel.
3. Select: Choose one emotion from the list that best fits the person's emotional state.
Emotion Options: Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning
Note: Only write the selected emotion on the last line.

<image>

GPT-4o: Engagement

Ground Truth: Confidence|Engagement

Figure 24: A sample correct case of EU-CoT.

EU-CoT

Instructions:
<image>
You are given an image. Predict the emotion it may evoke in a typical viewer.
1. Observe: Analyze the image in terms of brightness, colorfulness, scene type, objects, facial expressions, and human actions.
2. Interpret: Based on these visual features, infer what kind of emotion the image is likely to evoke in an observer.
3. Select: Choose one emotion from the list that best describes this likely emotional response.
Emotion Options: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sadness
Note: Only write the selected emotion on the last line.

<image>

GPT-4o: Fear

Error Category: Fine-grained category confusion

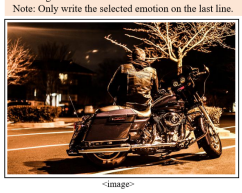
Error Reason: GPT-4o incorrectly classified the emotion as Fear instead of Anger , even though both emotions belong to the same valence (negative). **The error arises from a failure to distinguish between the specific emotional cues associated with "fear" (e.g., vulnerability, retreat) and "anger" (e.g., aggression, hostility) in the visual context.** The tiger's roaring and bared teeth strongly signal anger or aggression, not fear, leading to this fine-grained misclassification.

Ground Truth: anger

Figure 25: A sample error case of EU-CoT.

EU-CoT

Instructions:
<image>
You are given an image of a person. Identify the emotion expressed through their facial features.
1. Observe: Describe the person's body posture, facial expression, and surrounding environment.
2. Interpret: Based on the visual context, infer how the person might feel.
3. Select: Choose one emotion from the list that best fits the person's emotional state.
Emotion Options: Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning
Note: Only write the selected emotion on the last line.



<image>

GPT-4o: Contentment

Error Category: Valence-level misclassification

Error Reason: GPT-4o misclassified the emotion in the image as Contentment (positive valence) when it should have been Sadness (negative valence). **This error stems from the model's failure to accurately interpret key visual cues such as the nighttime setting, dim lighting, bare trees, and the lone motorcyclist riding away—all of which create a mood of solitude and melancholy.** Instead, the model focused on the act of riding a motorcycle, interpreting it as a symbol of freedom or adventure, which is typically associated with positive emotions. By overlooking the somber emotional context conveyed by the environment, the model made a valence-level misclassification, confusing a positive emotion with a negative one, rather than mistaking two emotions with similar valence.

Ground Truth: sadness

Figure 26: A sample error case of EU-CoT.

EU-CoT

Instructions:
<image>
You are given an image of a person. Identify the emotion expressed through their facial features.
1. Observe: Describe the person's body posture, facial expression, and surrounding environment.
2. Interpret: Based on the visual context, infer how the person might feel.
3. Select: Choose one emotion from the list that best fits the person's emotional state.
Emotion Options: Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning
Note: Only write the selected emotion on the last line.

<image>

GPT-4o: Peace

Error Category: Fine-grained category confusion

Error Reason: GPT-4o misclassified the emotion as Peace instead of Engagement, both of which fall within a similar positive or neutral valence, due to its inability to distinguish subtle visual cues differentiating relaxation from active focus. The man's closed eyes and posture leaning against the window suggest restfulness, supporting a peaceful interpretation; however, his hand on the steering wheel and slight forward lean also imply attentiveness or involvement, possibly indicating he is actively engaged in driving or observing. **GPT-4o prioritized the passive visual elements and overlooked how engagement can manifest through quiet concentration.** This represents a fine-grained category confusion, where the model confused two similarly valenced emotional states rather than misclassifying across opposing emotional valences.

Ground Truth: Engagement

Figure 27: A sample error case of EU-CoT.

EU-CoT

Instructions:
<image>
You are given an image of a person. Identify the emotion expressed through their facial features.
1. Observe: Describe the person's body posture, facial expression, and surrounding environment.
2. Interpret: Based on the visual context, infer how the person might feel.
3. Select: Choose one emotion from the list that best fits the person's emotional state.
Emotion Options: Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning
Note: Only write the selected emotion on the last line.

<image>

GPT-4o: Engagement

Error Category: Valence-level misclassification

Error Reason: GPT-4o misclassified the emotion as Engagement (positive/neutral valence) instead of Disconnection (negative valence), due to its failure to interpret subtle visual cues indicating emotional detachment. Although the person is walking—an action typically associated with active participation—their downward gaze, lack of interaction with the environment, and isolated posture suggest emotional withdrawal. The contrast between the busy street and the individual's disengaged demeanor further underscores their sense of disconnection. **However, GPT-4o focused on the physical movement as a sign of engagement, overlooking closed-off body language such as avoiding eye contact and minimal facial expression.** This represents a valence-level misclassification, as the model confused a positive or neutral emotion with a negative one, rather than mistaking two emotions of similar valence.

Ground Truth: Disconnection

Figure 28: A sample error case of EU-CoT.

EU-Basic

Instructions:
<image>
What feeling does this image most likely convey to a person? Please provide the emotion label for the image. The label should be one of the following: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sadness. Using this format: label

<image>

GPT-4o: Awe

Error Category: Fine-grained category confusion

Error Reason: The image depicts beautiful purple flowers, which typically evoke feelings of peace and satisfaction rather than "Awe." Therefore, the correct emotional label should be "Contentment," not "Awe."

Ground Truth: contentment

EU-CoT

Instructions:
<image>
You are given an image. Predict the emotion it may evoke in a typical viewer.
1. Observe: Analyze the image in terms of brightness, colorfulness, scene type, objects, facial expressions, and human actions.
2. Interpret: Based on these visual features, infer what kind of emotion the image is likely to evoke in an observer.
3. Select: Choose one emotion from the list that best describes this likely emotional response. Emotion Options: Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sadness
Note: Only write the selected emotion on the last line.

<image>

GPT-4o: Contentment

Ground Truth: contentment

Figure 29: Basic Prompting vs. CoT Prompting Case Study

F.2 Emotion Reasoning Case

The following presents our sample analysis of Emotion Reasoning (ER) cases, including representative correct and error examples. Refer to Figures 30,31,32, and 33 for detailed illustrations.

In each figure, we manually annotate the *True Answer*, which represents the ground-truth reasoning outcome based on the emotion context of the image. This information was not provided by the model, but serves as a human-verified reference to facilitate comparison. For the bad cases, we further conduct manual analysis to identify typical reasoning failures, such as incorrect emotion attribution, misunderstanding of causal context, or mismatches between textual and visual cues.

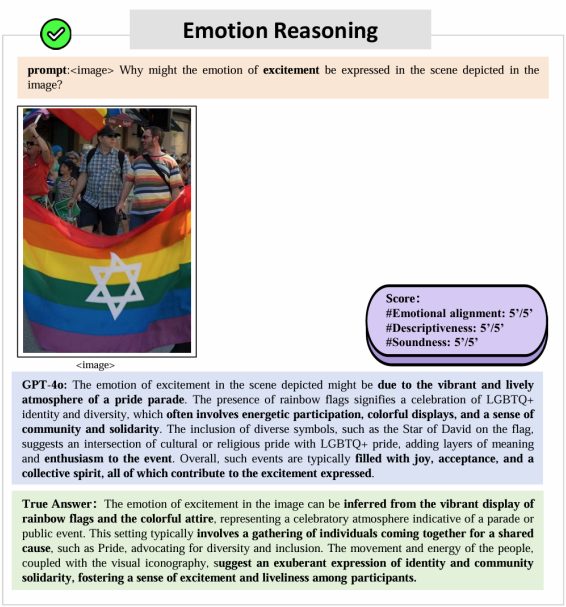


Figure 31: A sample correct case of EGCG

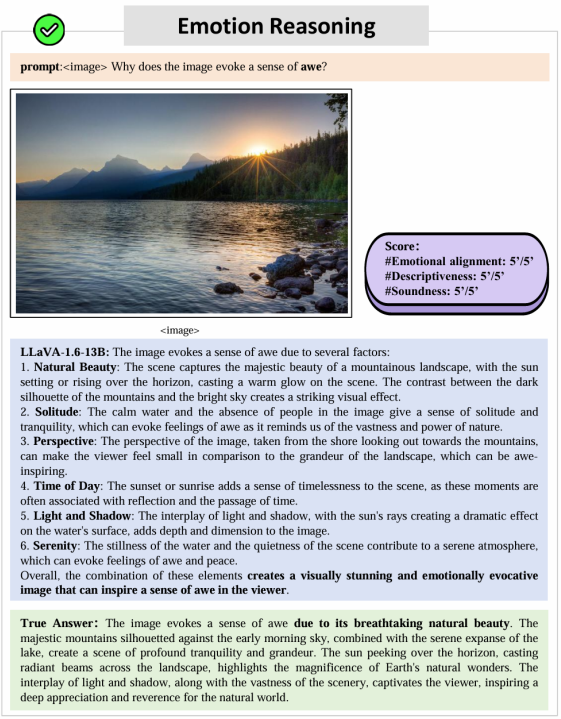


Figure 30: A sample correct case of EGCG

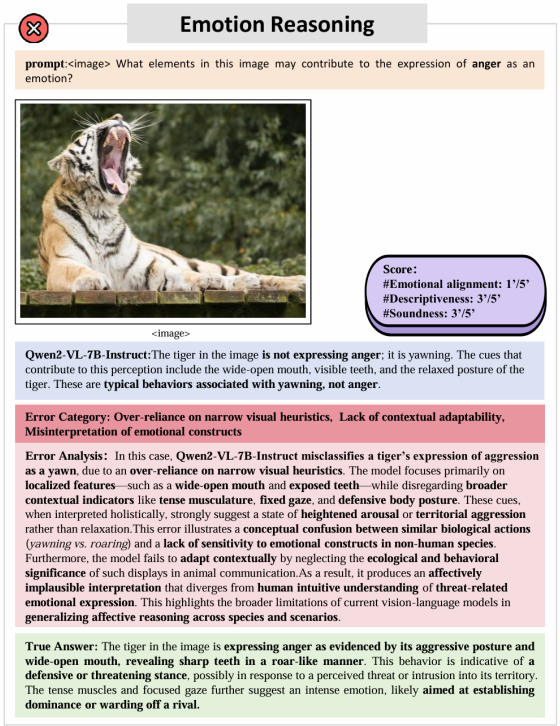


Figure 32: A sample correct case of EGCG

Emotion Reasoning

prompt:<image> Why is the emotion of **awe** being expressed in the image of a couple standing under a stone archway?

<image>

Score:
#Emotional alignment: 2'/5'
#Descriptiveness: 2'/5'
#Soundness: 2'/5'

Ovis2-8B: The emotion of awe is not explicitly expressed in the image. The couple appears to be posing for a wedding photo, smiling and looking content, which suggests happiness and joy rather than awe.

Error Category: Over-reliance on narrow visual heuristics, Lack of contextual adaptability, Misinterpretation of emotional constructs

Error Analysis: The model's failure to identify the emotion of awe stems from an over-reliance on facial expressions while neglecting the architectural and symbolic elements of the scene. **By focusing solely on the couple's smiles, the model overlooks the grandeur of the stone archway and the ritualistic solemnity associated with the wedding setting**. Moreover, it conflates emotional expression with emotional elicitation, **ignoring how visual aesthetics and cultural context can evoke awe in the observer**. This case highlights the need for more context-aware, multimodal reasoning in emotion attribution systems.

True Answer: The emotion of awe in the image is evoked by the grandeur and beauty of the wedding setting, particularly the striking stone archway that frames the couple. The archway's architectural elegance and historical significance add to the solemnity and majesty of the occasion, creating a sense of wonder. Furthermore, the sight of the bride in her flowing white gown holding a bouquet of pink flowers enhances this feeling of awe, as it symbolizes the beginning of a new chapter filled with love and commitment. This combination of visual elements contributes to the powerful emotional experience captured in the image.

Figure 33: A sample correct case of EGCG

ECG

prompt: <image> Write a short paragraph using a **disgusted** tone, inspired by the image of the insect caught in a spider's web. Focus on expressing the visceral reaction one might have upon viewing this scene.

<image>

Score:
#Emotional alignment: 5'/5'
#Descriptiveness: 5'/5'

GPT-4o: There it hangs, a grotesque spectacle of nature's cruel artistry—a hapless insect ensnared in the sticky lattice of a spider's web, its vibrant green now a sickening hue against the ominous strands that bind it. The sheer horror of its predicament is enough to churn one's stomach; limbs twitching futilely, each jerky movement a testament to its desperate struggle against the inevitable. The web, glistening treacherously in the light, seems to mock its plight, whispering the sinister promise of a slow, inexorable demise. This tableau of primal brutality is an unsettling reminder of the raw, unfeeling mechanics of survival, leaving an indelible mark of revulsion that lingers long after the image fades.

True Answer : The sight of the insect, grotesquely ensnared in the web, evokes a visceral shudder. The sticky threads cling to its spindly legs, and its bulbous eyes seem to bulge with helplessness. It writhes pathetically, trapped in an intricate lattice of silk, a prisoner in this macabre spectacle. The scene is unsettling, a grim reminder of nature's dark side, where beauty and horror intertwine. The insect's segmented body, with its sickly green hue, appears almost alien, stirring an instinctive repulsion deep within. It's a haunting tableau that leaves one squirming, eager to avert their gaze from such a repugnant display.

Figure 35: A sample correct case of EGCG

F.3 Emotion-guided Content Generation Case

We also include representative cases from the EGCG task to illustrate model performance (Figures 34, 35, 36, 37). Each case displays a manually annotated *True Answer*—the target content aligned with the intended emotion. For generation errors, we briefly analyze where the output fails to reflect the emotional intent.

ECG

prompt:<image>Write a short paragraph **in an excited tone** about a camping adventure inspired by this image.

<image>

Score:
#Descriptiveness: 5'/5'
#Soundness: 5'/5'

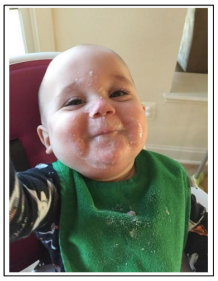
GPT-4o: The sun is shining brightly, casting its golden rays over a lively campsite filled with colorful tents and the promise of adventure! Imagine stepping into this vibrant scene, where every day is an exciting blend of exploration and laughter. The air is buzzing with the joyful chatter of fellow campers, all eager to share stories and create unforgettable memories. Picture yourself donning a playful, tropical shirt and a wide-brimmed hat, ready to embrace the day with a refreshing drink in hand. The majestic mountains in the distance beckon you to explore, while the cozy tents provide the perfect spot to relax after a day of discovery. This camping escapade is a thrilling escape, where the vivid hues of nature and the camaraderie of friends create an exhilarating experience like no other!

True Answer : The flames in the image, along with the individual whose body appears to be on fire and fleeing, suggest the occurrence of a fire-related incident. The inherent danger associated with fire serves as a potent emotional stimulus, primarily evoking a sense of fear in the observer.

Figure 34: A sample correct case of EGCG

ECG

prompt:<image> Write a short paragraph describing a scene using the tone of **amusement**, based on the image of the baby with the stained bib.



<image>

Score:
#Emotional alignment: 3'/5'
#Descriptiveness: 2'/5'

LLaVA-onevision: The baby with the stained bib looks like they've been up to some mischief, perhaps even trying to make a cake that's just a bit too crumbly.

Error Category: Emotion Overgeneralization, Lack of Visual Grounding

Error Analysis: The model's output reflects **emotion overgeneralization** by using vague phrasing ("up to some mischief") that **lacks the vivid, playful detail expected in an amused tone**. At the same time, it demonstrates **a lack of visual grounding**, **failing to reference key elements in the image such as the baby's expressive face, the stained green bib, or the joyful mess**. The response gestures at amusement but **lacks both emotional richness and image-specific detail**, resulting in a shallow and generic description. This case shows the model's difficulty in integrating affective tone with visual context, a critical challenge in multimodal emotion-guided generation.

True Answer : The dining room was filled with giggles and tiny squeals of delight as the baby, ensconced in their high chair, embarked on a culinary adventure. Pureed carrots and mashed peas adorned the green bib like badges of honor, each splatter a testament to the joyous chaos of discovery. Tiny hands waved enthusiastically, sending little blobs of food skyward, much to the amusement of the family gathered around. This mealtime spectacle, complete with impromptu food art, was a source of endless laughter and cherished memories, as the baby reveled in the simple joy of making a delightful mess.

Figure 36: A sample error case of EGCG

ECG

prompt:<image> Generate a short paragraph using a **disgusted** tone, describing the caterpillar in the image as if you are witnessing it firsthand.

<image>

Score:
#Emotional alignment: 1'/5'
#Descriptiveness: 4'/5'

Qwen2-VL-2B-Instruct: The caterpillar, with its vibrant yellow and black stripes, is a sight to behold. It's crawling along the slender branches of a plant, its movements slow and deliberate. The caterpillar's body is sleek and elongated, with its head and tail ending in a pair of antennae that seem to be scanning the environment for any potential threats. The caterpillar's eyes are large and round, and its mouthparts are clearly visible, ready to feed on the plant's nectar. The caterpillar's journey through the plant's leaves and branches is a testament to its adaptability and resilience, as it navigates through the intricate network of plant life.

Error Category: Emotional Misalignment, Emotion Overgeneralization

Error Analysis: This case illustrates a clear **emotional misalignment**, as the model, when prompted to generate a disgusted description of a caterpillar, instead produces a neutral or even admiring response, referring to the insect as "a sight to behold" and praising its "adaptability and resilience." Such phrasing directly contradicts the intended affective tone of disgust. In addition, the output suffers from **emotion overgeneralization**, relying on generic, emotionally flat language rather than vivid or figurative expressions that would typically convey aversion. While some visual elements are mentioned, they are not framed affectively, revealing a **lack of integration between visual grounding and emotional expression**. This example highlights a key limitation in current models—their frequent inability to **coherently align emotional tone, semantic content, and visual context** in emotion-guided generation tasks

True Answer : As I gaze upon the caterpillar clinging to the delicate branch, a wave of revulsion washes over me. Its bulbous, segmented body, adorned with unsightly protrusions, creeps along with a grotesque slowness that makes my skin crawl. The garish mix of yellow and green is almost nauseating, reminiscent of the colors found in a murky swamp. Each movement seems deliberate, as if intended to unsettle the observer with its unnatural, writhing form. I can't help but feel an intense urge to recoil, wishing for the immediate disappearance of this unsettling creature from my sight..

Figure 37: A sample error case of EGCG