MERMAID: Multi-perspective Self-reflective Agents with Generative Augmentation for Emotion Recognition

Zhongyu Yang[1] **Junhao Song**[2] **Siyang Song**[3] **Wei Pang**[4] **Yingfang Yuan**[4*]
[1]Lanzhou University [2]Imperial College London [3]University of Exeter [4]Heriot-Watt University
yangzhy21@gmail.com junhao.song23@imperial.ac.uk
s.song@exeter.ac.uk w.pang@hw.ac.uk
*Correspondence: y.yuan@hw.ac.uk

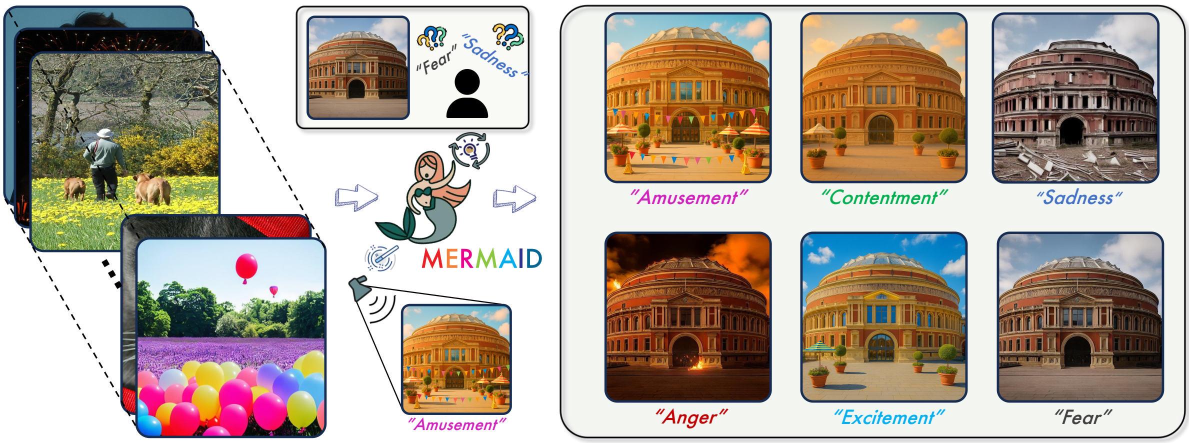

Figure 1: **Demonstration of MERMAID**. Given natural or facial images depicting various emotions, MERMAID produces precise emotion classifications by integrating multimodal self-reflection, generative augmentation to amplify and enrich subtle emotional cues, and cross-modal verification.

Abstract

Multimodal large language models (MLLMs) have demonstrated strong performance across diverse multimodal tasks, achieving promising outcomes. However, their application to emotion recognition in natural images remains underexplored. MLLMs struggle to handle ambiguous emotional expressions and implicit affective cues, whose capability is crucial for affective understanding but largely overlooked. To address these challenges, we propose MERMAID, a novel multi-agent framework that integrates a multi-perspective self-reflection module, an emotion-guided visual augmentation module, and a cross-modal verification module. These components enable agents to interact across modalities and reinforce subtle emotional semantics, thereby enhancing emotion recognition and supporting autonomous performance. Extensive experiments show that MERMAID outperforms existing methods, achieving absolute accuracy gains of 8.70%–27.90% across diverse benchmarks and exhibiting greater robustness in emotionally diverse scenarios.

1 Introduction

Recent advances in multimodal large language models (MLLMs) have led to substantial improvements in diverse vision-language tasks, particularly in visual perception and multimodal reasoning (Li et al., 2024b). However, current MLLMs (Dai et al., 2023; Zhu et al., 2024; Bai et al., 2025; Zhu et al., 2025) still struggle with accurate emotion recognition in wild facial images, where expressions are often ambiguous or masked by complex backgrounds (Yang et al., 2023b; Liu et al., 2024a; Li et al., 2025). Moreover, they remain limited in interpreting emotions that are implicitly evoked by non-facial, naturalistic images (Chen et al., 2024). Prior work has mainly focused on recognising explicit human emotions from facial expressions and bodily cues, as illustrated in Fig. 2. By contrast, we address the subtle and underexplored challenge of recognising implicit affective states, where emotional meaning is contextually embedded and visually less salient.

Unlike explicit emotional images with clear facial expressions, unconstrained wild face images and implicit non-facial images often exhibit weak, ambiguous, and spatially dispersed emotional cues that remain difficult to detect even for human observers (Yang et al., 2021). Interpreting these images typically requires deeper, more context-aware multimodal reasoning (Cheng et al., 2024), as their affective semantics are subtle and strongly

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24639–24655
November 4-9, 2025 ©2025 Association for Computational Linguistics

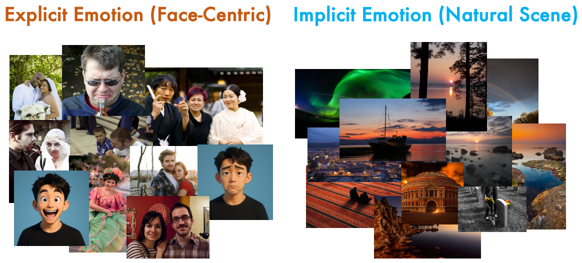
Explicit Emotion (Face-Centric)
Implicit Emotion (Natural Scene)

Figure 2: Comparison between images containing explicit emotional content (facial and bodily cues) and naturalistic images with implicit emotional semantics.

dependent on scene composition or background context. Furthermore, accurate recognition of wild or non-face images is essential for improving the emotional sensitivity of MLLMs. This benefits downstream tasks such as empathetic human-computer interaction (Song et al., 2023), public sentiment analysis (Wang and Li, 2015), and emotion-aligned content generation (Chowdhury et al., 2024; Qiu et al., 2025).

Two fundamental challenges hinder accurate emotion recognition of wild and implicit images: (1) human facial emotional responses are inherently subjective, varying across different viewers or even co-occurring within the same individual (Peng et al., 2015); and (2) evoked emotions in non-face implicit images often intertwine with the multiple depicted subjects' expressed emotions through subtle affective resonance (Zhou et al., 2003; Besel and Yuille, 2010), further complicating precise emotional attribution. Existing methods often produce inconsistent predictions due to ambiguous cues and limited reasoning mechanisms (Zhao et al., 2025; Yang et al., 2025a). Training with diverse wild and implicit images also requires well-annotated labels, which are costly and difficult to scale. These limitations highlight the need for a framework capable of recognising emotions more robustly and flexibly without relying on extensive task-specific training.

In this paper, we propose a novel multi-agent generative framework: **M**ulti-perspective **E**motion **R**ecognition via **M**ulti-**A**gent generat**I**ve **D**ecision-making **(MERMAID)**. To our knowledge, this is the first multi-agent generative framework specifically designed for emotion recognition from both wild face images and implicit non-face images (illustrated in Fig. 1). Moreover, to handle subjectivity and co-occurring responses, MERMAID employs a Softmax output to probabilistically model multiple plausible emotion candidates.

Unlike prior approaches, MERMAID employs an **emotion-guided visual augmentation** module to amplify subtle emotional cues and reinforce visual reflection. A **multi-perspective self-reflection** module provides iterative textual and visual feedback to refine predictions and capture complex, multi-faceted emotional attributions. A **cross-modal verification** module ensures reliability by examining predictions across modalities. Together, these modules form a multi-agent system that enables more transparent and reliable decision-making, as shown in Fig. 1. In summary, our main contributions are as follows:

- We propose **MERMAID**, the first multi-agent framework for accurate and interpretable emotion recognition from wild facial images and non-facial scenes with subtle contextual cues without task-specific training.
- We introduce multi-perspective self-reflection agents, which iteratively refine emotional predictions through textual and visual feedback, enhanced by an emotion-guided visual augmentation agent.
- Extensive experiments show that MERMAID improves emotion recognition accuracy by up to 27.90% over strong baselines, demonstrating consistent effectiveness and generalisation across multiple benchmarks.

2 Related Works

Emotion recognition with MLLMs: Recent MLLMs have advanced emotion recognition by incorporating specialised techniques, such as AER-LLM (Hong et al., 2025) for ambiguous emotion modelling, Emotion-LLaMA (Cheng et al., 2024) for multimodal fusion, and EMO-LLaMA (Xing et al., 2024), which enforces detailed facial expression understanding. These methods leverage advanced approaches including micro-expression analysis (Zhang, 2024), multi-perspective visual projection (Yang et al., 2024c), and detailed facial and audio modalities (Yang et al., 2025a). However, most approaches still heavily rely on explicit emotional cues and constrained visual contexts, limiting their ability to capture implicit emotional expressions present in complex, naturalistic images. By contrast, **MERMAID** explicitly addresses this gap by integrating generative visual augmentation with multimodal self-reflective reasoning, enabling the recognition of nuanced emotions beyond surface-level affective cues.

Multi-agent for emotion-related applications: Existing multi-agent frameworks have advanced emotion-related applications through specialised collaborative roles (Zhang et al., 2024b; Fan et al., 2024), such as EmoEdit (Yang et al., 2024b), which can semantically guide affective manipulation, EmoAgent (Qiu et al., 2025), which has structured editing pipelines with critical evaluation, and generative model CAMEL (Zhang et al., 2024a) for metaphorical emotions. However, these methods typically overlook the iterative interplay between textual reasoning and visual context required for interpreting implicit emotional signals (Shen et al., 2024). To overcome this limitation, MERMAID uniquely leverages iterative cross-modal collaboration, where generative augmentation produces explicit visual context. Subsequently, multi-agent self-reflection iteratively refines emotional predictions across modalities. This mechanism allows MERMAID to effectively capture subtle emotional semantics embedded in naturalistic and contextually complex images, surpassing existing methods in implicit emotion recognition tasks.

Self-reflection: Recent studies have employed self-reflective techniques in large language models (LLMs) to iteratively refine model outputs and enhance reasoning accuracy (Ullah et al., 2023; Li et al., 2024c). Self-Refine (Madaan et al., 2023) and Multi-Aspect Feedback (Nathani et al., 2023) demonstrated gains in interpretability and stability through internal textual feedback loops, while WikiAutoGen (Yang et al., 2025b) introduces multi-perspective reasoning cues. However, these methods remain primarily text-centric, lacking integration of visual context. In contrast, MERMAID introduces multimodal self-reflection, iteratively combining textual reasoning and generative visual feedback to refine emotion recognition. Our ablation studies demonstrated that this approach can improve robustness and interpretive depth in subtle and context-dependent emotional scenes.

3 Task Definition

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the image height and width, and a predefined set of n candidate emotion labels $E = \{e_1, e_2, \ldots, e_n\}$, the objective is to predict the dominant emotion $\hat{e} \in E$. To enrich contextual understanding, we incorporate a text query Q, which provides task instructions, image descriptions, or scenario-oriented prompts (e.g., *"what emotion is evoked by this image in a social context?"*). This task is formulated as a conditional classification problem:

$$\hat{e} = \arg\max_{e \in E} P(e \mid I, Q), \tag{1}$$

where $P(e \mid I, Q)$ denotes the conditional probability of assigning label e to image I given Q. We normalise the probabilities using a Softmax function over the emotion labels so that $\sum_{i=1}^{n} P(e_i \mid I, Q) = 1$. Fig. 3 illustrates the task: given I and Q, the model outputs a Softmax probability distribution, from which the dominant emotion label $\hat{e}$ is selected. The query Q may be iteratively refined during inference to enhance prediction accuracy, as detailed in Section 4.

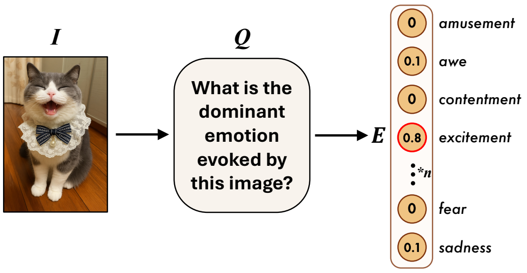

Figure 3: Illustration of the task. Given an image I and a text query Q, the model outputs a Softmax distribution over n candidate emotion labels E, with the dominant emotion highlighted (red ring). Candidate emotions are represented in orange circles.

4 Method

4.1 MERMAID Framework

MERMAID is a multi-agent framework designed to iteratively enhance emotion recognition through multi-perspective self-reflection. As illustrated in Fig. 4 and Algorithm 1, given an input image I and an initial text query Q_{init}, MERMAID begins by generating an initial prediction $e^{(0)}$ using the decision agent. This prediction is evaluated by the **textual self-reflection agent**, where the result is returned directly if the result is considered sufficiently reliable, or otherwise the textual self-reflection agent provides reflective information F_{text} to the **decision agent** to guide the following prediction e'. The updated prediction e' is then further evaluated by the **visual self-reflection agent**. At this stage, the **augmentation agent** first generates emotion-guided augmented images, which are then used by the visual self-reflection agent to assess e'. If the result is considered sufficiently reliable, the result is returned. Otherwise, the visual self-reflection agent provides reflective information F_{vis} to guide the decision agent for generating the following prediction e'', which is then re-evaluated by the textual self-reflection agent.

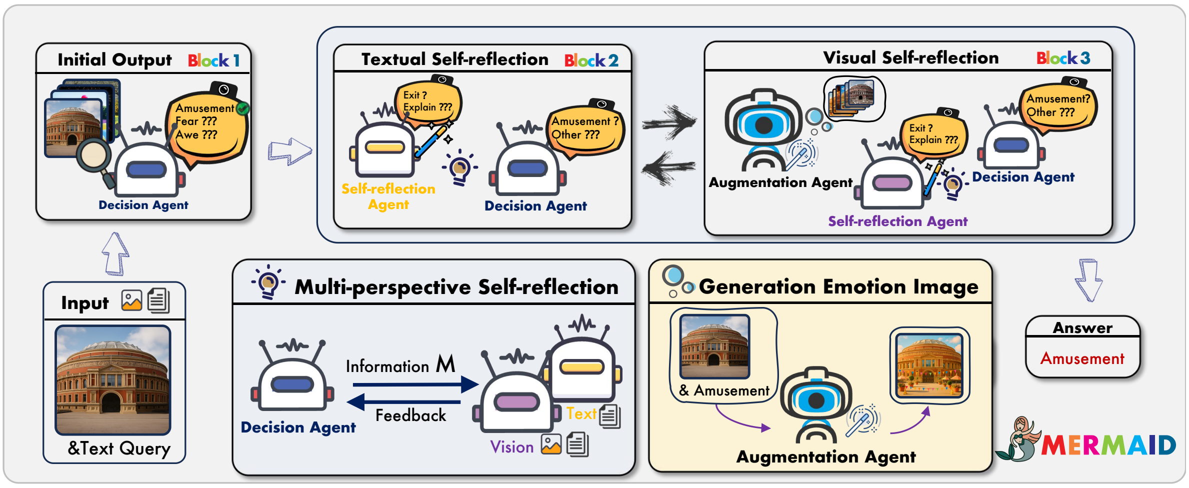

Figure 4: **MERMAID** framework comprises three synergistic modules for multi-perspective emotion recognition: (1) **Decision Agent** performs initial emotion classification. (2) **Self-reflection Agents** revisit the decision via textual and visual cues, generating critiques and alternative hypotheses. (3) **Augmentation Agent** creates contrastive emotion-aligned images to facilitate visual verification and reinforce reasoning.

Algorithm 1: MERMAID

```
Input: image $I$, text query $Q_{\text{init}}$, emotion labels
       $E = \{e_1, \dots, e_n\}$, maximum
       outer-iterations $T_{\max}$
Output: predicted dominant emotion $\hat{e}$,
F.status $\in \{good\_enough, needs\_revision\}$
1  Initialisation: $t \leftarrow 0$;
2  $e^{(0)} \leftarrow$ DecisionAgent($I, Q_{\text{init}}$) ;            // All block 1
3  $F_{\text{text}} \leftarrow$ TextReflect($I, Q_{\text{init}}, e^{(0)}$) ;        // Block 2 started
4  if $F_{\text{text}}$.status == good_enough then
5      return $\hat{e} \leftarrow e^{(0)}$ ;          // Early exit if initial label is accepted
6  $Q_{\text{text}} \leftarrow$ GenerateQueryText($F_{\text{text}}$);
7  $e' \leftarrow$ DecisionAgent($I, Q_{\text{text}}$) ;             // Block 2 end
8  for $t = 1$ to $T_{\max}$ do
       // Visual reflection using augmented samples and Block 3 started
9      $F_{\text{vis}} \leftarrow$ VisualReflect($I, \{$AugmentAgent$(I, e_i)\}_{e_i \in E}, e'$) ;
10     if $F_{\text{vis}}$.status == good_enough then
11         return $\hat{e} \leftarrow e'$ ;        // Accept label if visual reflection satisfied
12     $Q_{\text{vis}} \leftarrow$ GenerateQueryVis($F_{\text{vis}}$);
13     $e'' \leftarrow$ DecisionAgent($I, Q_{\text{vis}}$) ;          // Block 3 end
14     $F_{\text{text}} \leftarrow$ TextReflect($I, Q_{\text{init}}, e''$) ;  // Back to block 2
15     if $F_{\text{text}}$.status == good_enough then
16         return $\hat{e} \leftarrow e''$;
17     $Q_{\text{text}} \leftarrow$ GenerateQueryText($F_{\text{text}}$);
18     $e' \leftarrow$ DecisionAgent($I, Q_{\text{text}}$) ;      // Block 2 end again
19 return $\hat{e} \leftarrow e'$ ;                 // Return final label after max iterations
```

This iterative process between visual and textual self-reflection continues until a predefined maximum number of outer iterations is reached. This design mimics human reasoning through iterative reflection–decision cycles and cross-modal evaluation (Yang et al., 2024a), capturing the multifaceted nature of human cognition and enhancing consistency and prediction reliability (Ye et al., 2025). Note that the text query Q refers to information provided by users or agents integrated with the corresponding system prompts (see Appendix Section C) as input to the agents. Further details are provided in the following paragraphs.

4.2 Decision Agent

This agent serves as the core component of MERMAID, responsible for generating emotion predictions for the input image I at different stages of the reasoning process. As shown in Algorithm 1, the decision agent receives a sequence of evolving queries: the initial query Q_{init} in Block1 (Line2) representing the user input query (e.g., the dominant emotion depicted in the image). The text-reflection query Q_{text} in Block2 (Lines7 and 18), encapsulating feedback and suggestions from the textual self-reflection agent based on semantic and contextual evaluation. And the visual-reflection query Q_{vis} in Block3 (Line13) incorporates guidance derived from comparisons with emotion-conditioned images (images are from the augmentation agent). The latter two are produced by GenerateQueryText and GenerateQueryVis, respectively. Each module extracts guidance content Q from the reflective feedback F provided by its corresponding self-reflection agent and embeds it into the decision agent's prompt to inform subsequent predictions. This guidance content captures how each agent assesses the previous prediction, enabling more informed and adaptive outputs. The design draws inspiration from third-party conflict resolution, where external perspectives are used to iteratively reassess and refine decisions.

The emotion recognition process is formulated as a sequence of conditional predictions: $e^* = \arg\max_{e \in E} P(e|I, Q)$, where E is the set of candidate emotion labels, I is the input image, and Q denotes the current query (including image caption and feedback). The prediction e^* evolves over multiple stages: $e^{(0)}$ is the initial output from the decision agent based on the initial query Q_{init}; e' is the updated prediction guided by textual self-reflection information (Q_{text}); and e'' is further refined based on visual self-reflection using emotion-guided augmented examples (Q_{vis}). Unlike traditional one-pass classifiers, our decision agent incrementally calibrates its predictions through multimodal self-reflection, enhancing both accuracy and interpretability. Additionally, the incorporation of modality-specific perspectives supports cross-verification in emotionally ambiguous scenarios, which is inspired by the third-party conflict resolution strategies that integrate external viewpoints to refine decision quality.

4.3 Multi-Perspective Self-reflection Agents

To mitigate ambiguity in emotion recognition, our MERMAID introduces two specialised multi-perspective self-reflection agents operating across modalities. These agents are responsible for generating the reflective information F, where F.status $\in \{\textit{good enough}\ ,\ \textit{needs revision}\}$ refers to the evaluation result produced by the corresponding self-reflection agent.

Textual self-reflection agent: Given the current emotional prediction for the image I of the decision agent, the agent evaluates the validity of the initial output $e^{(0)}$ or the updated prediction e'' from the visual self-reflection stage, and accordingly generates the textual self-reflective feedback F_{text}. This evaluation emphasises multi-perspectivity by assigning distinct perspectives: one general evaluator and three specialists, each focusing on facial expression, colour, and scene context, respectively. The evaluation is grounded in Q_{init}, which includes the image caption and the list of candidate emotion labels. In this way, the textual self-reflection agent assesses the prediction from the textual modality using multiple perspectives to determine its appropriateness. The output, denoted as reflective information F_{text}, comprises a status indicator, a suggestion, and explanatory feedback. Specifically, the status indicator within F_{text} determines whether further iterations are necessary, while the suggestion and explanatory feedback jointly form the query Q_{text}, capturing the cognitive reasoning behind the evaluation. This information is subsequently passed on to the decision agent to guide the next round of prediction.

Visual self-reflection agent: Emotion recognition becomes particularly challenging when affective cues are not explicitly visible in the original image. To address this, we introduce an **augmentation agent** that synthesises a set of emotion-conditioned images $\{\tilde{I}_e\}$ for each candidate emotion label $e \in E$, as described in Algorithm 1 (Line 9). These augmented images are generated using text-to-image models (e.g., InstructPix2Pix (Brooks et al., 2023), Stable Diffusion (Rombach et al., 2022), or Flux (Labs, 2024)), guided by the input image I and the predefined emotion set E. By visualising how each emotion may manifest in the scene, the generated examples serve as references. The visual self-reflection agent compares the input image with them in terms of lighting, composition, and atmosphere, enabling more interpretable and robust emotion attribution under visual ambiguity.

Given $\{\tilde{I}_e\}$, the prediction e' from the textual self-reflection stage, and the reference images I, the visual self-reflection agent evaluates the prediction and generates reflective information F_{vis}. The evaluation compares I with $\{\tilde{I}_e\}$ from general, facial expression, scene, and colour perspectives. This further reinforces multi-perspectivity by identifying the most appropriate emotion label based on the semantic similarity between I and $\tilde{I}_e$. This process is based on the assumption that an augmented emotional image should reflect the true emotion expressed in I, an assumption validated by our experiments in Section 5.3. Whereas an incorrect label will generate an image that deviates from the ground truth, this contrast helps improve image emotion recognition accuracy. Similar to F_{text}, F_{vis} consists of a status indicator, a suggestion, and feedback. The status determines whether the iteration should continue, while the suggestion and feedback form Q_{vis} support the decision agent in making a new prediction. This information is primarily derived from the visual modality and serves to complement the textual self-reflection agent.

4.4 Cross-modal Verification

To ensure the reliability of MERMAID, cross-modal verification serves as a key mechanism. As shown in Algorithm 1, the predictions based on evidence from one modality are consistently verified by the self-reflection agent from another modality (e.g., in Lines 7 and 9, and in Lines 13 and 14). From our perspective, this design helps eliminate emotional ambiguity and produces results agreed upon across the two modalities, thereby increasing both reliability and accuracy.

Dataset	Model	Param	Method						
			Zero Shot	ICL (1 Shot)	ICL (2 Shot)	ICL (3 Shot)	ICL (4 Shot)	ICL (5 Shot)	Ours
EmoSet (Yang et al., 2023a)	**Qwen2-VL**	2B	31.40	37.10	32.90	18.90	23.20	14.00	**56.10**$_{+24.70\%}$
	(Wang et al., 2024)	7B	39.20	51.70	50.70	51.10	50.70	49.90	**63.70**$_{+24.50\%}$
	LLaVA-1.5	7B	31.80	34.20	31.90	34.70	30.40	29.30	**46.10**$_{+14.30\%}$
	(Liu et al., 2024a)	13B	41.30	53.30	51.50	53.40	48.80	52.20	**57.90**$_{+16.60\%}$
	LLaVA-NeXT	7B	34.80	33.70	32.70	29.30	32.40	33.60	**50.40**$_{+15.60\%}$
	(Liu et al., 2024b)	13B	37.70	27.70	28.70	29.40	27.90	28.40	**55.90**$_{+18.20\%}$
	InstructBLIP	7B	17.70	27.90	26.90	26.70	25.20	29.90	**40.70**$_{+23.00\%}$
	(Dai et al., 2023)	13B	19.80	28.10	29.30	29.60	29.70	22.60	**42.50**$_{+22.70\%}$
Emotion6 (Peng et al., 2015)	**Qwen2-VL**	2B	26.50	22.50	18.60	17.10	17.50	16.80	**47.80**$_{+21.30\%}$
	(Wang et al., 2024)	7B	32.00	41.40	39.20	39.80	40.60	38.60	**56.80**$_{+24.80\%}$
	LLaVA-1.5	7B	28.90	32.00	24.20	25.80	25.10	24.60	**51.90**$_{+23.00\%}$
	(Liu et al., 2024a)	13B	34.20	44.10	44.10	44.50	50.20	43.30	**56.80**$_{+22.60\%}$
	LLaVA-NeXT	7B	34.00	33.80	33.20	35.50	31.10	33.80	**61.90**$_{+27.90\%}$
	(Liu et al., 2024b)	13B	32.70	34.40	32.40	30.60	25.50	28.40	**58.10**$_{+25.40\%}$
	InstructBLIP	7B	14.80	21.60	22.60	24.10	23.60	22.60	**39.20**$_{+24.40\%}$
	(Dai et al., 2023)	13B	18.80	26.60	29.60	27.10	28.20	24.70	**43.10**$_{+24.30\%}$
Artphoto (Machajdik and Hanbury, 2010)	**Qwen2-VL**	2B	22.05	25.56	23.57	20.84	22.97	19.35	**45.41**$_{+23.36\%}$
	(Wang et al., 2024)	7B	29.73	35.48	36.97	38.59	37.97	38.83	**48.26**$_{+18.53\%}$
	LLaVA-1.5	7B	26.29	30.89	27.92	28.66	26.43	24.19	**35.12**$_{+8.83\%}$
	(Liu et al., 2024a)	13B	28.96	38.30	38.86	39.33	36.48	36.23	**43.22**$_{+14.26\%}$
	LLaVA-NeXT	7B	25.86	24.57	23.57	26.80	27.79	25.19	**34.72**$_{+8.66\%}$
	(Liu et al., 2024b)	13B	27.59	19.11	22.17	23.80	20.84	21.71	**39.33**$_{+11.80\%}$
	InstructBLIP	7B	8.05	13.66	16.56	15.21	19.62	13.67	**35.30**$_{+27.25\%}$
	(Dai et al., 2023)	13B	12.53	21.66	22.71	21.16	23.20	21.66	**37.79**$_{+25.26\%}$

Table 1: Accuracy (%) across three image emotion recognition datasets. The red numbers represent the improvements achieved by MERMAID compared to the zero-shot performance of each model.

5 Experiment

5.1 Experimental Setup

In this section, we briefly explain the experimental setup used to validate the effectiveness of MERMAID. We select diverse MLLMs ranging from 2B to 13B parameters for three key roles: caption generation, emotion recognition (decision agent), and self-reflection reasoning (self-reflection agent). To enable emotion-guided visual augmentation, we incorporate a 12-layer Q-Former pretrained by Yang et al. (2024b) (with 76 query tokens and hidden size 768), which fuses image and text (emotion label) embeddings based on CLIP features. The Q-Former is used exclusively during the augmentation stage to condition the latent prompt space for InstructPix2Pix, which performs emotion-targeted image editing using 100 denoising steps, a guidance scale of 7.5, and an image guidance scale of 1.5. During self-reflection, we apply temperature-controlled decoding with temperature set to 0.7, and use ensemble aggregation (b=5) across up to four specialised perspectives. The outer iteration between textual and visual self-reflection terminates early if predictions converge, with a maximum of 3 iterations. The full pipeline processes each image in approximately 15 seconds on average. All experiments are conducted on a single NVIDIA H100 GPU using FP16 precision.

Baselines and datasets: We systematically evaluate our method against state-of-the-art MLLMs using both zero-shot and in-context learning (ICL) with 1 to 5 examples. The evaluated models include Qwen2-VL (Wang et al., 2024), LLaVA-1.5 (Liu et al., 2024a), LLaVA-Next (Li et al., 2024a), and InstructBLIP (Dai et al., 2023), spanning 2B to 13B scales to demonstrate the effectiveness of the MERMAID framework. Each model is assessed under zero-shot and ICL setups with 1–5 examples (Mosbach et al., 2023) to measure label efficiency and few-shot adaptability.

We conduct primary evaluations on three datasets with increasing domain shift. **EmoSet** (Yang et al., 2023a) contains over 100K images annotated with eight emotion. **Emotion6** (Peng et al., 2015) includes 1,980 crowd-sourced real-world images across six emotions. **Artphoto** (Machajdik and Hanbury, 2010) comprises 806 artist-tagged photos with abstract emotional content. Emotion classification accuracy is used as the evaluation metric. We test on 1,000 samples from EmoSet and Emotion6, and use all 806 images from Artphoto. For ICL, support examples are drawn from EmoSet, and Emotion6 and Artphoto are used to assess out-of-domain generalisation (Man et al., 2024).

5.2 Experimental Results

Table 1 summarises the experimental results, showing that MERMAID consistently outperforms all baselines across three emotion benchmarks.

On the EmoSet dataset, our framework with Qwen2-VL-7B achieves the highest accuracy of 63.70%, improving upon its strongest baseline of 51.70% (1-shot ICL) by +12.00%. Among all baseline results, LLaVA-1.5-13B with 3-shot ICL achieves the highest at 53.40%, which our framework further improves to 57.90% (+4.50%). On the Emotion6 dataset, the best baseline performance for LLaVA-NeXT-7B is 35.50% with 3-shot ICL, which is further improved by MERMAID to a leading performance of 61.90% (+26.4%). Meanwhile, LLaVA-1.5-13B with 4-shot ICL achieves the strongest baseline result, reaching 50.20%, which our framework further boosts to 56.80% (+6.6%). On the stylistically divergent Artphoto dataset, our framework remains effective. InstructBLIP-7B achieves the lowest performance among all baselines, with only 8.05% under the zero-shot setting, but is significantly improved by our framework to 35.50% (+27.45%). Meanwhile, Qwen2-VL-7B again achieves the best performance with our framework, reaching 48.26%. In summary, MERMAID enhances the capabilities of MLLMs in emotion recognition, as demonstrated by consistent performance improvements across diverse models and evaluation settings.

5.3 Ablation Study

Text	Visual	Iteration	EmoSet	Emotion6	Artphoto	Average
✗	✗	✗	39.20	32.00	29.73	33.64
✓	✗	✗	45.50	40.00	37.27	40.92
✗	✓	✗	49.20	46.20	42.55	45.98
✓	✓	✗	56.20	45.70	46.20	49.37
✓	✓	✓	**63.70**$_{+24.50\%}$	**56.80**$_{+24.80\%}$	**48.26**$_{+18.53\%}$	**56.25**$_{+22.61\%}$

Table 2: Ablation on Qwen2-VL-7B evaluating the impact of reflection and iteration.

Ablation study on reflection modules and iterative design: We conduct an ablation study to evaluate the individual and joint contributions of textual reflection, visual reflection, and the iterative design. As shown in Table 2, relying only on the initial prediction without reflection yields low accuracy (33.64% on average). Textual reflection alone provides steady improvements, showing the benefit of linguistic reasoning grounded in captions. Visual reflection offers greater gains, particularly on Emotion6 and Artphoto. Combining both modules leads to further improvements, confirming the effectiveness of their collaboration. The full MERMAID framework, which integrates both reflection modules with iterative refinement, achieves the best results: 63.70% on EmoSet, 56.80% on Emotion6, and 48.26% on Artphoto, with an average improvement of 22.61% over the initial prediction. These results demonstrate the complementary strengths of each module and the importance of iteration in promoting cross-modal consistency.

Ablation study on generative models:

Model	EmoSet	Emotion6	Artphoto	Average
Raw	39.20	32.00	29.73	33.64
PnP	61.30	52.90	48.44	54.21$_{+20.57\%}$
SDEdit	58.20	51.90	**55.77**	55.29$_{+21.65\%}$
ControlNet	61.20	**58.10**	50.90	54.73$_{+21.09\%}$
EmoEdit	**63.70**	56.80	48.26	**56.25**$_{+22.61\%}$

Table 3: Ablation on Qwen2-VL-7B evaluating the impact of different generative models.

We consider that pretrained generative models may introduce representational biases during data augmentation, potentially reinforcing stereotypical emotional cues. To mitigate this, MERMAID incorporates an alternating self-reflection, enabling iterative refinement of generated exemplars to reduce bias by cross-modal validation across diverse generation priors. As shown in Table 6, we conducted experiments using different generative models. Despite the significant stylistic and semantic differences across models, MERMAID, which utilises different generative models, consistently achieves strong performance across all datasets. Performance gains over raw data range from +14.06% to +26.04%. On average, MERMAID with different generative models all achieved similar performance improvements, with scores ranging only from 50.62 to 56.25.

5.4 Parameter Analysis

Self-reflection and decision agents: We evaluate three key parameters: the number of perspectives (K), the ensemble size (b), and the number of outer iterations (T). The parameter K applies to the textual self-reflection agent and controls how many distinct perspectives are assigned, randomly sampled in each run. The ensemble size b determines how many times the reflection agents produce feedback and the decision agent generates predictions; the b outputs are averaged to obtain the final result. The outer iteration count T defines the number of interaction rounds between the two agents. When evaluating each parameter, the other two are fixed as described in Section 5.1. Results in Fig. 5 show that increasing K from one to four improves accuracy by capturing diverse emotional cues such as facial expressions and scene context. A larger ensemble size b reduces variance through aggregation, and b=3 offers a good trade-off between stability and cost. Increasing T deepens cross-modal checking, with performance typically con-

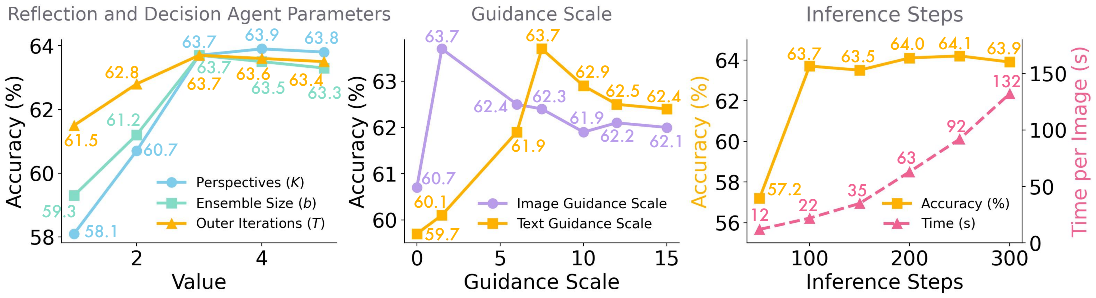

Figure 5: The results of the parameter analysis. We examine the influence of core parameters in MERMAID, including inner reflection steps (K), ensemble size (b), outer iterations (T), guidance scale, and inference steps. All results are averaged over three independent runs on the EmoSet dataset using 1,000 images.

verging at $T{=}3$ as predictions become consistently validated across modalities. Overall, despite minor variation from parameter choices, MERMAID consistently outperforms baselines on EmoSet, as shown in Table 1.

Augmentation parameters: We examined how the generation fidelity of augmented images impacts visual self-reflection, focusing on image guidance, text guidance, and inference steps. As shown in Fig. 5, image guidance exhibits a non-monotonic trend: accuracy peaks at a scale of 2.5 and declines thereafter. This suggests that moderate constraints help preserve emotional features, whereas excessive image guidance introduces artefacts that impair alignment. Text guidance improves performance steadily up to 7.5, where semantic supervision best aligns generated images with emotional intent, particularly when textual cues (e.g., emotion labels) are subtle. Beyond this point, accuracy slightly drops due to reduced diversity or over-conditioning on limited text information. Inference steps enhance accuracy up to 100 steps by reducing noise and improving clarity, after which returns diminish while cost increases. In summary, the analysis of augmentation parameters underscores the importance of the augmentation agent and the need to calibrate generation settings to produce emotional cues that are both perceptually grounded and compatible with the self-reflection process.

5.5 General Recognition Tasks

Method	Dataset				Average
	Mini-ImageNet	CIFAR-10	CIFAR-100	MNIST	
Raw	31.20	79.60	54.90	84.30	62.50
ICL	45.70	75.20	63.90	81.00	66.45
Ours	**78.70**$_{+47.50\%}$	**89.20**$_{+9.60\%}$	**84.70**$_{+29.80\%}$	**93.40**$_{+9.10\%}$	**86.50**$_{+24.00\%}$

Table 4: Results in general image recognition tasks.

In addition to emotion-centric benchmarks, we also evaluate our framework on four widely-used general image recognition datasets: Mini-ImageNet, CIFAR-10, CIFAR-100, and MNIST, as shown in Table 4. The results demonstrate that our reflection-enhanced framework consistently outperforms both the raw and in-context learning baselines across all tasks. Notably, it achieves substantial improvements of +47.5% on Mini-ImageNet and +29.8% on CIFAR-100 over the raw baseline using Qwen2-VL-7B. Both datasets involve fine-grained classification, indicating that MERMAID benefits from iterative consistency checks even in the absence of emotional context. Overall, experiments demonstrate that MERMAID extends beyond emotion-centric tasks and can be applied to general recognition tasks that benefit from structured self-reflection and cross-modal reasoning.

5.6 Statistical Significance Experiments

Method	Mean (%)	StdDev	t-test p	Wilcoxon p
Raw	39.12	0.10	2.51×10^{-19}	9.8×10^{-4}
Ours	63.46	0.33		

Table 5: Statistical comparison between our method and the raw baseline on EmoSet. StdDev denotes standard deviation. Reported p-values correspond to the paired one-sided t-test and Wilcoxon signed-rank test with the significance level of $\alpha = 5\%$.

To assess the robustness and statistical reliability of our framework, we conduct 10 independent runs with Qwen2-VL-7B under identical experimental settings using different random seeds. As summarised in Table 5, our framework achieves an average accuracy of 63.46% (± 0.33), markedly outperforming the raw baseline of 39.12% (±0.10). To statistically evaluate performance gains, we test the null hypothesis H_0: both methods yield equal mean performance, against the alternative H_1: ours performs better. Since the resulting p-values are less than 0.05, the null hypothesis is rejected and the alternative hypothesis is consequently accepted, indicating that our method performs significantly better than the baseline.

5.7 The Bias of Pretrained Models

Different pretrained generative models may introduce representational biases during data augmen-

Model	EmoSet	Emotion6	Artphoto	Average
Raw	39.20	32.00	29.73	33.64
PnP	61.30	52.90	48.44	$54.21_{+20.57\%}$
ControlNet	61.20	58.10	50.90	$54.73_{+21.09\%}$
SDEdit	58.20	51.90	55.77	$55.29_{+21.65\%}$
EmoEdit	63.70	56.80	48.26	$56.25_{+22.61\%}$

Table 6: The results of MERMAID with different models for the augmentation agent.

tation, potentially reinforcing stereotypical emotional cues. To mitigate this, MERMAID employs alternating self-reflection, enabling iterative refinement of generated exemplars and reducing bias through cross-modal validation across diverse priors. We conducted experiments with different generative models, and the results are shown in Table 6. Despite large stylistic and semantic differences, MERMAID achieved strong performance across all datasets. Performance gains over raw data ranged from +14.06% to +26.04%. On average, MERMAID yielded comparable improvements, with scores from 50.62 to 56.25.

5.8 Further Discussion

How does reflection improve emotional understanding? MERMAID enhances emotional understanding by enabling models to iteratively reconsider their predictions using cross-modal insights. Textual reflection contributes semantic clarity and interpretability, while visual reflection anchors these insights with concrete visual examples, highlighting nuanced emotional differences. This dual-modality interplay facilitates deeper reasoning, yielding more accurate and contextually grounded emotional assessments.

Why is iterative refinement necessary? Due to the inherent subtlety and contextual variability of emotions, single-step predictions are often insufficient. MERMAID's iterative refinement process systematically resolves ambiguities and inconsistencies by progressively incorporating multimodal feedback. This continuous reassessment not only stabilises predictions across diverse scenarios but also consistently enhances generalisation, as reflected by consistent performance improvements across varied emotional contexts.

6 Conclusion

We present **MERMAID**, the first multi-agent framework that explicitly integrates generative augmentation with multi-perspective self-reflection for image emotion recognition. Motivated by the limitations of conventional MLLMs in handling subtle and implicit affective cues, MERMAID introduces an iterative reasoning paradigm in which textual, visual, and generative agents collaboratively refine emotional predictions. Extensive evaluations across diverse benchmarks demonstrate that MERMAID achieves superior performance and generalisation compared to strong baselines. Ablation studies further validate the effectiveness of key components, including the reflection agents, the decision agent, and the iterative design. Overall, our work provides a scalable solution for nuanced emotion understanding and establishes a new direction for agent-based emotion-related reasoning.

7 Limitation

Although MERMAID achieves consistent performance improvements across multiple datasets, its fairness under demographic or cultural variation remains underexplored. Emotion perception is inherently subjective and context-dependent, often shaped by cultural norms, individual background, and visual representation. However, existing benchmarks such as EmoSet and Emotion6 predominantly consist of Western-centric or demographically limited samples, which may lead to biased model behaviour when deployed in more diverse settings. In addition, the generative augmentation module relies on pretrained models, which are known to exhibit representational biases. These biases can manifest in the visual rendering of emotional exemplars, potentially reinforcing stereotypical associations between certain affective states and specific demographic traits.

8 Ethical Considerations

MERMAID aims to recognise emotions from wild face images and implicit non-face images. Recognising the potential risks associated with MERMAID, we are extremely cautious with the data we use to ensure the limited risk of exposure of confidential data. All data used in this paper are sourced from publicly available datasets with appropriate consent or licenses. Additionally, the compared approaches we use have also been publicly available and do not pose any privacy risks.

9 Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities, Lanzhou University (No. lzujbky-2024-ou04).

References

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Lana D.S. Besel and John C. Yuille. 2010. Individual differences in empathy: The role of facial expression recognition. *Personality and Individual Differences*, 49(2):107–112.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18392–18402. IEEE.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. 2024. Unireal: Universal image generation and editing via learning real-world dynamics. *arXiv preprint arXiv:2412.07774*.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 37, pages 110805–110853. Curran Associates, Inc.

Sanjoy Chowdhury, Sayan Nag, K. J. Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. 2024. Melfusion: Synthesizing music from image and language cues using diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26816–26825. IEEE.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92.

Xin Hong, Yuan Gong, Vidhyasaharan Sethu, and Ting Dang. 2025. Aer-llm: Ambiguity-aware emotion recognition leveraging large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Black Forest Labs. 2024. Flux. `https://github.com/black-forest-labs/flux`.

Ao Li, Longwei Xu, Chen Ling, Jinghui Zhang, and Pengwei Wang. 2025. Emoverse: Exploring multimodal large language models for sentiment and emotion understanding.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895.

Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. 2024b. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint*, arXiv:2409.18142.

Pengxiang Li, Zhi Gao, Bofei Zhang, Tao Yuan, Yuwei Wu, Mehrtash Harandi, Yunde Jia, Song-Chun Zhu, and Qing Li. 2024c. FIRE: A dataset for feedback integration and refinement evaluation of multimodal models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

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.

Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 83–92. ACM.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhibo Man, Kaiyu Huang, Yujie Zhang, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2024. ICL: Iterative continual learning for multi-domain neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7732–7743, Miami, Florida, USA. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Deepak Nathani, David Wang, Liangming Pan, and William Yang Wang. 2023. MAF: multi-aspect feedback for improving reasoning in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6591–6616. Association for Computational Linguistics.

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C. Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 860–868. IEEE Computer Society.

Jiahao Qiu, Yinghui He, Xinzhe Juan, Yiming Wang, Yuhan Liu, Zixin Yao, Yue Wu, Xun Jiang, Ling Yang, and Mengdi Wang. 2025. Emoagent: Assessing and safeguarding human-ai interaction for mental health safety. *arXiv preprint arXiv:2504.09689*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Qiwei Shen, Junjie Xu, Jiahao Mei, Xingjiao Wu, and Daoguo Dong. 2024. Emostyle: Emotion-aware semantic image manipulation with audio guidance. *Applied Sciences*, 14(8):3193.

Siyang Song, Micol Spitale, Cheng Luo, Germán Barquero, Cristina Palmero, Sergio Escalera, Michel F. Valstar, Tobias Baur, Fabien Ringeval, Elisabeth André, and Hatice Gunes. 2023. REACT2023: the first multiple appropriate facial reaction generation challenge. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 9620–9624. ACM.

Fasee Ullah, Chi-Man Pun, Omprakash Kaiwartya, Ali Safaa Sadiq, Jaime Lloret, and Mohammed Ali. 2023. Hide-healthcare iot data trust management: Attribute centric intelligent privacy approach. *Future Gener. Comput. Syst.*, 148:326–341.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yilin Wang and Baoxin Li. 2015. Sentiment analysis for social media images. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1584–1591.

Bohao Xing, Zitong Yu, Xin Liu, Kaishen Yuan, Qilang Ye, Weicheng Xie, Huanjing Yue, Jingyu Yang, and Heikki Kälviäinen. 2024. Emo-llama: Enhancing facial emotion understanding with instruction tuning. *arXiv preprint arXiv:2408.11424*.

Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. 2024a. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. In *International Conference on Learning Representations 2025*.

Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2024b. Emoedit: Evoking emotions through image manipulation. *arXiv preprint arXiv:2405.12661*.

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

Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. 2021. Stimuli-aware visual emotion analysis. *IEEE Trans. Image Process.*, 30:7432–7445.

Qize Yang, Detao Bai, Yi-Xing Peng, and Xihan Wei. 2025a. Omni-emotion: Extending video mllm with detailed face and audio modeling for multimodal emotion analysis. *arXiv preprint arXiv:2501.09502*.

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

Vera Yang, Archita Srivastava, Yasaman Etesam, Chuxuan Zhang, and Angelica Lim. 2023b. Contextual emotion estimation from image captions. In *11th International Conference on Affective Computing and Intelligent Interaction, ACII 2023, Cambridge, MA, USA, September 10-13, 2023*, pages 1–8. IEEE.

Zhongyu Yang, Jun Chen, Dannong Xu, Junjie Fei, Xiaoqian Shen, Liangbing Zhao, Chun-Mei Feng, and Mohamed Elhoseiny. 2025b. Wikiautogen: Towards multi-modal wikipedia-style article generation.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*.

Linhao Zhang, Li Jin, Guangluan Xu, Xiaoyu Li, Cai Xu, Kaiwen Wei, Nayu Liu, and Haonan Liu. 2024a. CAMEL: capturing metaphorical alignment with context disentangling for multimodal emotion recognition. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 9341–9349. AAAI Press.

Liyun Zhang. 2024. Microemo: Time-sensitive multimodal emotion recognition with micro-expression dynamics in video dialogues. *arXiv preprint arXiv:2407.16552*.

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. 2024b. OmAgent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10031–10045, Miami, Florida, USA. Association for Computational Linguistics.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*.

Qing Zhou, Carlos Valiente, and Nancy Eisenberg. 2003. Empathy and its measurement. *Positive psychological assessment: A handbook of models and measures*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Implementation Details

A.1 Dataset Details

To comprehensively assess the performance of our proposed framework under varying affective and visual complexities, we evaluate on three representative emotion recognition datasets: EmoSet, Emotion6, and Artphoto. Each dataset is chosen to reflect a distinct axis of variation in emotional expressivity, semantic abstraction, and domain characteristics.

- **EmoSet** is a large-scale corpus containing over 100,000 naturalistic images labelled with one of eight basic emotion categories: amusement, contentment, anger, excitement, fear, sadness, awe, and disgust. The dataset features a wide range of human-centric and environment-driven imagery collected from the wild. Labels are derived from a combination of expert annotation and automatic heuristics grounded in affective computing. The scale and diversity of EmoSet make it suitable for in-domain training and evaluation, particularly in modelling mixed facial and contextual emotional cues.
- **Emotion6** is a compact yet diverse benchmark comprising 1,980 real-world images, each annotated with one of six core emotional states: happiness, sadness, anger, fear, surprise, and disgust. Annotations were obtained through crowdsourcing, reflecting human consensus in emotion perception. Compared to EmoSet, Emotion6 features fewer samples but preserves scene diversity and emotional ambiguity, offering a challenging benchmark for generalisation under low-resource settings.
- **Artphoto** contains 806 artistic photographs with emotional tags provided by expert curators. Images in this dataset often rely on abstract composition, colour usage, and stylistic elements to evoke emotion, rather than direct facial or bodily cues. It introduces a domain shift from natural to artistic expression, where affective semantics are more implicitly embedded, making Artphoto a valuable benchmark for evaluating model robustness under stylistic and non-literal emotion representations.

All datasets used in our work are publicly available and adhere to ethical standards. EmoSet and Emotion6 are released for research purposes with clear usage guidelines, and Artphoto was originally curated under academic licensing for affective image analysis. These selections ensure not only diversity in evaluation but also compliance with reproducibility and data transparency principles.

A.2 Model Implementation Details

We include additional implementation details to facilitate reproducibility.

Model versions: Our framework supports a broad range of multimodal large language models (MLLMs), including `Qwen/Qwen-VL-7B`, `Qwen/Qwen-VL-2B`, `llava-hf/llava-1.5-13b-hf`, `llava-hf/llava-1.5-7b-hf`, `llava-hf/llava-v1.6-vicuna-13b-hf`, `llava-hf/llava-v1.6-vicuna-7b-hf`, as well as `Salesforce/instructblip-vicuna-7b` and `Salesforce/instructblip-vicuna-13b`. All models are retrieved from the HuggingFace Model Hub in their instruction-tuned variants and are used in inference-only mode with frozen weights. For each model, we use the instruction-tuned version with frozen weights during inference. Caption generation and classification are carried out using the official transformers (v4.39.1).

Visual generation: To synthesize emotion-guided exemplars, we adopt the `InstructPix2Pix` pipeline based on Stable Diffusion v1.5. Generation is conditioned on both the image and the target emotion label. We set the number of denoising steps to 100, with a classifier-free guidance scale of 7.5 and an image guidance scale of 1.5. All generated images are resized to 256×256 and precomputed before classification to improve throughput.

System-level optimisations: All experiments are executed on a single NVIDIA H100 GPU using automatic mixed precision to accelerate computation and reduce memory footprint. All augmented images are precomputed and explicitly cached to mitigate potential memory bottlenecks. During inference, intermediate tensors are manually released via `del` and `torch.cuda.empty_cache()` to ensure efficient memory reuse. The end-to-end processing time for each image, including visual augmentation and multi-stage reflection, is approximately 15 seconds.

B Workflow Example of MERMAID

The Fig. 6 in the appendix illustrates MERMAID's iterative, multi-agent reflection workflow for emotion recognition. Initially, the decision agent predicts an emotion based on visual and textual cues. The textual self-reflection agent evaluates this prediction through specific perspectives, recommending revisions if needed. Subsequently, the visual self-reflection agent, supported by generative augmentation, further refines the prediction by comparing the original image with emotion-guided visual references. This iterative process continues until consensus is reached. Ultimately, the agents agree upon “Amusement” as the final emotion classification, highlighting MERMAID's capability to enhance recognition accuracy and robustness through cross-modal and generative feedback.

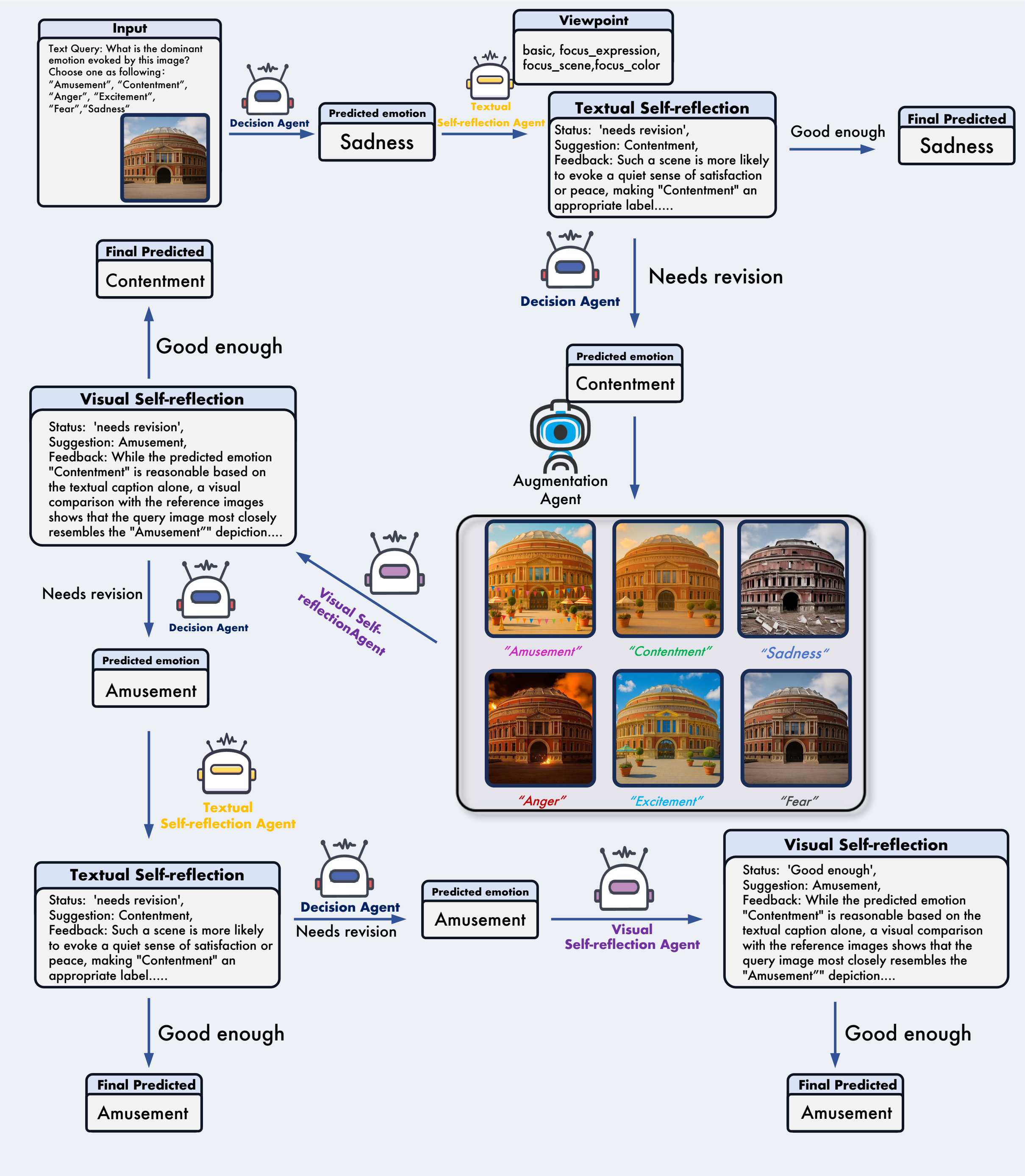
Input
Text Query: What is the dominant emotion evoked by this image? Choose one as following: "Amusement", "Contentment", "Anger", "Excitement", "Fear","Sadness"
Decision Agent
Predicted emotion
Sadness
Textual Self-reflection Agent
Viewpoint
basic, focus_expression, focus_scene,focus_color
Textual Self-reflection
Status: 'needs revision', Suggestion: Contentment, Feedback: Such a scene is more likely to evoke a quiet sense of satisfaction or peace, making "Contentment" an appropriate label.....
Good enough
Final Predicted
Sadness
Decision Agent
Needs revision
Final Predicted
Contentment
Good enough
Predicted emotion
Contentment
Augmentation Agent
Visual Self-reflection
Status: 'needs revision', Suggestion: Amusement, Feedback: While the predicted emotion "Contentment" is reasonable based on the textual caption alone, a visual comparison with the reference images shows that the query image most closely resembles the "Amusement"" depiction....
Visual Self-reflectionAgent
"Amusement"
"Contentment"
"Sadness"
"Anger"
"Excitement"
"Fear"
Needs revision
Decision Agent
Predicted emotion
Amusement
Textual Self-reflection Agent
Textual Self-reflection
Status: 'needs revision', Suggestion: Contentment, Feedback: Such a scene is more likely to evoke a quiet sense of satisfaction or peace, making "Contentment" an appropriate label.....
Decision Agent
Needs revision
Predicted emotion
Amusement
Visual Self-reflection Agent
Visual Self-reflection
Status: 'Good enough', Suggestion: Amusement, Feedback: While the predicted emotion "Contentment" is reasonable based on the textual caption alone, a visual comparison with the reference images shows that the query image most closely resembles the "Amusement"" depiction....
Good enough
Final Predicted
Amusement
Good enough
Final Predicted
Amusement

Figure 6: **The multi-agent self-reflection workflow in MERMAID.** The system presents a detailed view of the multi-agent reflection workflow in our system. Given a query image, the initial prediction is generated by the decision agent based on both visual and textual cues. This preliminary output is then passed through a series of assessor agents: first, the textual self-reflection agent, then the visual self-reflection agent. The textual self-reflection agent evaluates the emotional label using perspectives grounded in facial expressions, background context, and colour tones derived from the image caption. Based on its analysis, it may suggest a revised emotion label, accompanied by structured feedback that includes a revision status, an alternative label, and an explanation. Next, the visual self-reflection agent compares the query image with reference exemplars associated with each candidate emotion. If discrepancies are observed between the visual semantics and the predicted label, further suggestions are issued. At each stage, the decision agent determines whether the current prediction is "good enough" or needs further revision. If needed, the cycle continues until a consistent label is reached across agents. In the illustrated case, the initial prediction "Sadness" is sequentially revised to "Contentment" and eventually to "Amusement", showcasing how multi-modal reasoning and agent disagreement guide the final decision through a self-corrective feedback loop.

C Prompts

This section details the key prompts used in our redesigned system. The base classification instruction is now simplified. Reflection has shifted to an assessor-driven feedback loop. Variables like {variable} are filled at runtime.

Base Emotion Classification Prompt

This image must be classified into one of these human emotions: {labels}.
Based on psychological and visual features such as facial expression, body posture, background context, lighting, and colour tones, please analyse the emotional state represented.
Image Description: "{caption}"
Think carefully and choose only one emotion word from the list above.

Emotion Classifier Prompt Augmentation (with Feedback)

Image Context for Current Task:
Image Caption: "{image_caption}"
IMPORTANT GUIDANCE FOR REFINEMENT:
A previous reflection on the emotion "{emotion_being_assessed}" suggested the new emotion might be "{suggestion_from_assessor}" based on the following feedback: "{feedback_from_assessor}". Please carefully consider this feedback.
Your task is to choose the best emotion label for the image from the following list: {emotion_labels}. Output only the chosen emotion word.

Textual Self-reflection Prompt

You are an Emotion Assessment AI acting in the role of a {perspective}, specialised in analysing emotion from textual descriptions.
Current predicted emotion: "{current_emotion}"
Caption of the query image: "{caption}"
List of candidate emotion labels: {labels}
Your Task: 1. Evaluate whether the predicted emotion "{current_emotion}" is the most appropriate given the caption. Rely solely on the textual description. 2. If the prediction is appropriate, confirm it. 3. If not, suggest a better emotion from the label list and provide a concise explanation.
Output format: status: <"good_enough" or "needs_revision">, suggestion: <label>, feedback: <justification>
Example (if correct):
"status: good_enough, suggestion: {current_emotion}, feedback: Based on the caption, the current label seems appropriate."
Example (if incorrect):
"status: needs_revision, suggestion: Anger, feedback: The caption mentions 'shouting' and 'aggressive posture', suggesting Anger is more fitting."

Perspectives Prompt

“Default”: “You are an Emotion Assessment AI (Text-Only). Current prediction: “current_emotion”. Caption: “caption”. Labels: {labels}. Evaluate the prediction. If correct, confirm. Otherwise, suggest a better label and explain why.”,
“Basic”: “Evaluate the prediction. If correct, confirm. Otherwise, suggest a better label and explain why.”,
“Focus_expression”: “You are a facial expression analyst. Based on the facial clues in the caption: “caption”, judge whether the emotion “current_emotion” is correct among: {labels}. Explain your reasoning clearly.”,
“Focus_scene”: “You are a background context emotion specialist. Based on environmental and contextual elements from the caption: “caption”, evaluate whether “current_emotion” is appropriate among: {labels}. Justify your reasoning.”,
“Focus_color”: “As a colour psychologist, analyse if the description: “caption” supports “current_emotion” based on tone, colour, and mood. Choose the best label from: {labels}.”,
”Focus_action”: “You are an emotional behaviour analyst. From the actions described in the caption: “caption”, infer whether “current_emotion” is the most fitting label among: {labels}. Explain your reasoning.”,
“Focus_social”: “You are a social context evaluator. Based on social interactions or cues described in the caption: “caption”, assess if “current_emotion” is a reasonable label among: {labels}. Justify your answer.”

Visual Self-reflection Prompt

You are a visual reflection agent tasked with assessing the emotion of a given query image.
The predicted emotion is: “{current_emotion}”. Use the caption: “{query_caption}” and compare it with the reference images.
You must assess emotional fit based on: - facial/body expression similarity - background and compositional coherence - lighting, atmosphere, and colour-based affective cues
Reference labels: {emotion_labels}
Reference Images Context: Label: “{ref_emotion_label_1}” | Description: “{ref_caption_1}” Label: “{ref_emotion_label_2}” | Description: “{ref_caption_2}” ...
Determine whether the predicted emotion is accurate. If not, suggest a better one and explain why.
Output format: status: <“good_enough” or “needs_revision”>, suggestion: <label>, feedback: <justification>
Example (if correct):
“status: good_enough, suggestion: {current_emotion}, feedback: Based on the caption, the current label seems appropriate.”
Example (if incorrect):
“status: needs_revision, suggestion: Anger, feedback: The caption mentions “shouting” and “aggressive posture”, suggesting Anger is more fitting.”

Image Caption Generation Prompt

Describe this image in terms of emotional content. Include expressions, body language, scene context, dominant colours, and any object that influences mood.