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

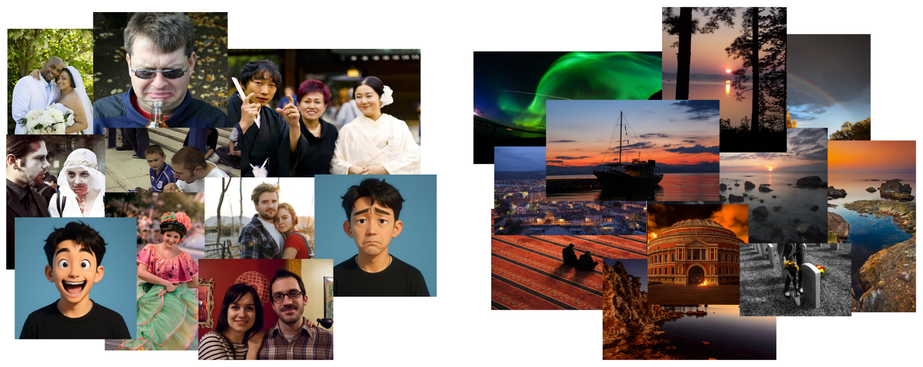
Zhongyu Yang[1] Junhao Song[2] Siyang Song[3] Wei Pang[4] Yingfang Yuan[4]
[1]Lanzhou University [2]Imperial College London [3]University of Exeter [4]Heriot-Watt University

30th ANNIVERSARY EMNLP 2025
Suzhou, China | 中国苏州
November 4-9 | 11月4日–9日

## Introduction

- ✘ Current MLLMs struggle with accurate emotion recognition in wild facial images.
- ✘ Moreover, MLLMs remain limited in interpreting emotions that are implicitly evoked by non-facial, naturalistic images.
- ✘ 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.

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

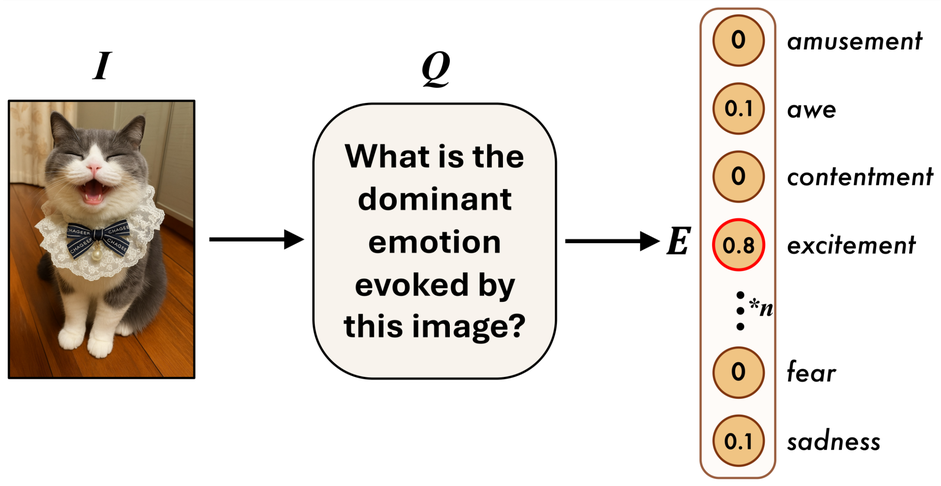


Project Website | WeChat

## Contribution

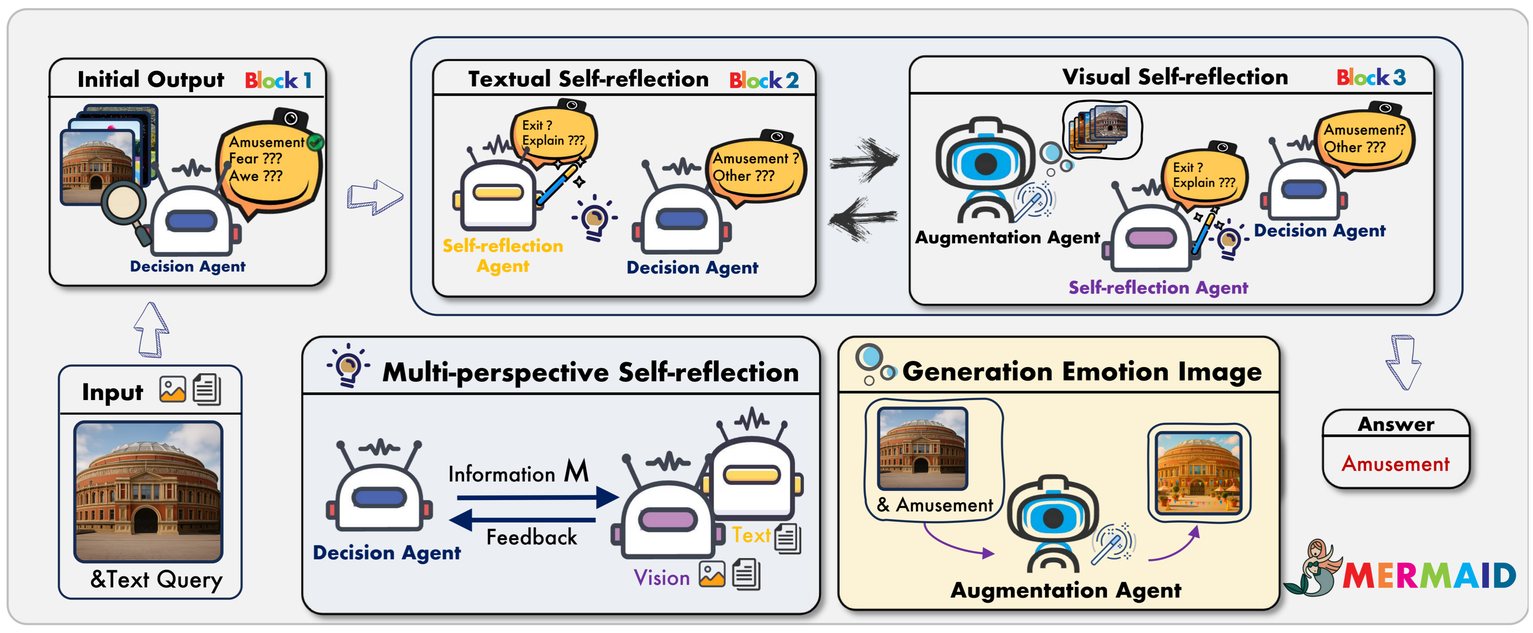
✓ We introduce **MERMAID**, a multi-agent framework that reflects, augments, and verifies emotions across modalities.

## Task Definition




Given an input image and a text query, MERMAID identifies the dominant emotion from a predefined set of emotion categories. It performs conditional classification by estimating the probability of each emotion and selecting the most likely one based on multimodal cues.

## Method




Overview: MERMAID integrates decision, reflection, and generation agents to iteratively refine emotion recognition through multimodal self-reflection.
Pipeline Highlights:

- ✓ Decision Agent: predicts and updates emotion labels.
- ✓ Textual Reflection: critiques predictions from semantic and contextual views.
- ✓ Visual Reflection & Augmentation: generates emotion-guided images for cross-modal verification.
- ✓ Cross-Modal Iteration: refines predictions until textual–visual consensus is achieved.

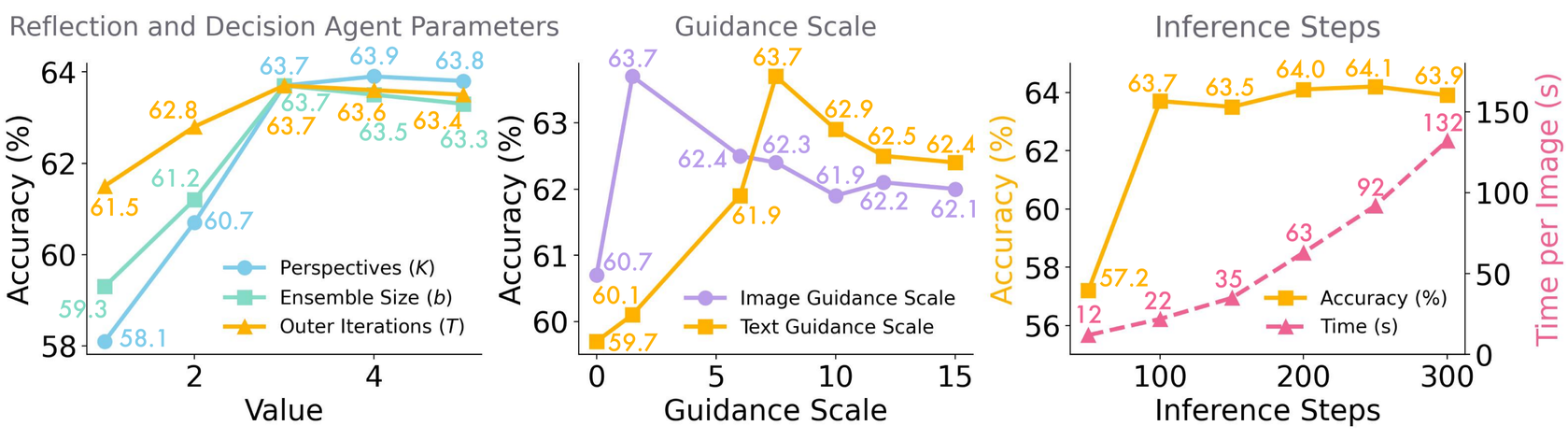
## Experiment

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

| 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**$_{+22.61\%}$ |

| 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**$_{+24.00\%}$ |

| 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**$_{+22.61\%}$ |




**Main Finding:** MERMAID improves accuracy by up to 27.9% over strong baselines, showing robust generalization and interpretable emotion reasoning across diverse scenes..

**Highlights:**

- ✓ **Consistent Gains:** On EmoSet, Emotion6, and ArtPhoto, MERMAID improves accuracy by 8.7%–27.9% across Qwen2-VL, LLaVA, and InstructBLIP models.
- ✓ **Generalizability:** Extends beyond emotion tasks: on Mini-ImageNet, CIFAR, and MNIST, accuracy rises from 62.5% to 86.5% (+24%), confirming broad recognition benefits.
- ✓ **Parameter Insights:** Increasing reflection depth improves performance, with optimal guidance (≈7.5) and 100 steps balancing cost, while joint text–visual reflection achieves the best accuracy.