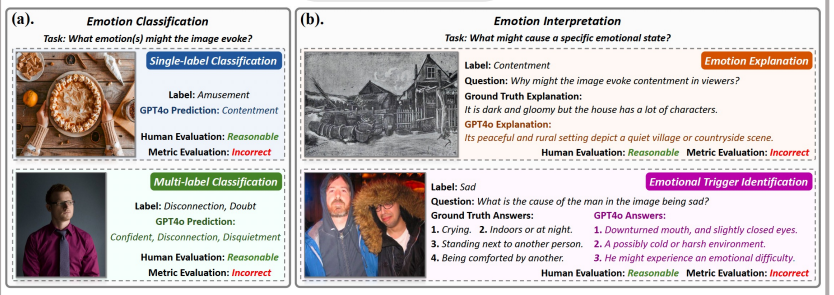
Customizing Visual Emotion Evaluation for MLLMs: An Open-Vocabulary, Multifaceted, and Scalable Approach

Daiqing Wu[1,4] · Dongbao Yang[1] · Sicheng Zhao[3] · Can Ma[1] · Yu Zhou[2]

[1] Institute of Information Engineering, CAS, [2] Nankai University,
[3] Tsinghua University, [4] University of Chinese Academy of Sciences

GitHub: WeChat:

MOTIVATION



Why Current Visual Emotion Evaluation Falls Short?

Oversight of Plausible Responses, Limited Emotional Taxonomies, Neglect of Contextual Factors, Labor-Intensive Annotations

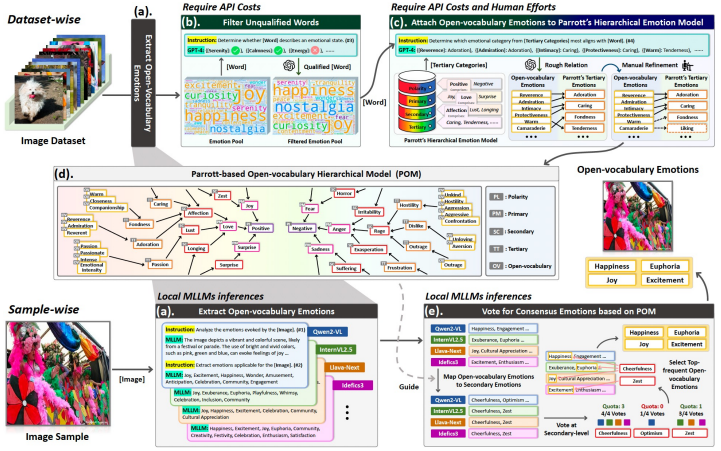
Core Idea: Emotion Statement Judgment (ESJ)

The ESJ task reformulates emotion evaluation by requiring MLLMs to validate emotion-centric statements for a given image, naturally supporting open-vocabulary emotions, inclusion of contextual factors, cost-efficient construction, and alleviation of response subjectivity.

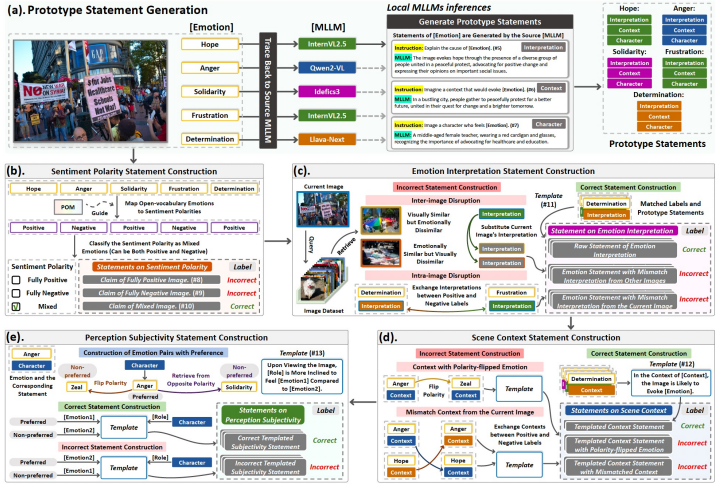
Sentiment Polarity: overall emotional tone.
Emotion Interpretation: affective reasoning capability.
Scene Context: dynamic influence of contextual factors.
Perception Subjectivity: personalized emotion prediction.

DATA PIPELINE

Automated Construction of 462k Corpus



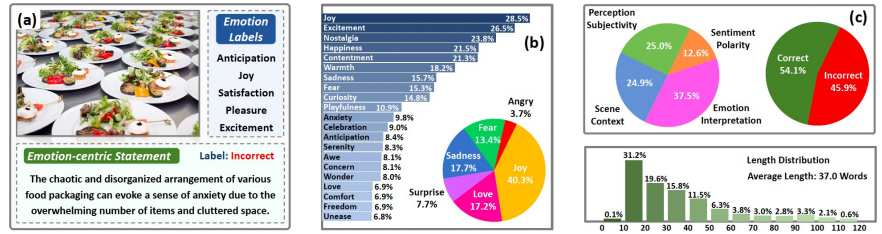
Step1: Open-Vocabulary Emotion Tagging



Step2: Emotional Statement Construction

HUMAN REFINEMENT

MVEI (Multifaceted Evaluation of Visual Emotion Intelligence): A Human-Refined Subset for Evaluation



5 human annotators refine 3,164 statements, reaching a **Fleiss's Kappa score of 0.61**. 3086 samples are finally retained into MVEI benchmark.

EVALUATION

Experiments on 18 Popular MLLMs, 25 Humans, and 5 Lightweight Model Adaptation Techniques

Finding 1: Asymmetric Strengths

MLLMs excel at emotion interpretation and context-based judgment, but struggle with perception subjectivity.

Finding 2: Persistent Human–AI Gap

Even GPT-4o trails human average by ~12.6%, revealing substantial room for improvement in emotional intelligence.

Finding 3: Adaptation Helps, But Unevenly

Fine-tuning dramatically boosts Sentiment Polarity yet barely improves Perception Subjectivity, suggesting the latter demands innovations beyond standard SFT.