# FROM SURFACE TO DEPTH: TOWARDS COGNITIVE APPRAISAL REASONING IN MULTIMODAL EMOTION UNDERSTANDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

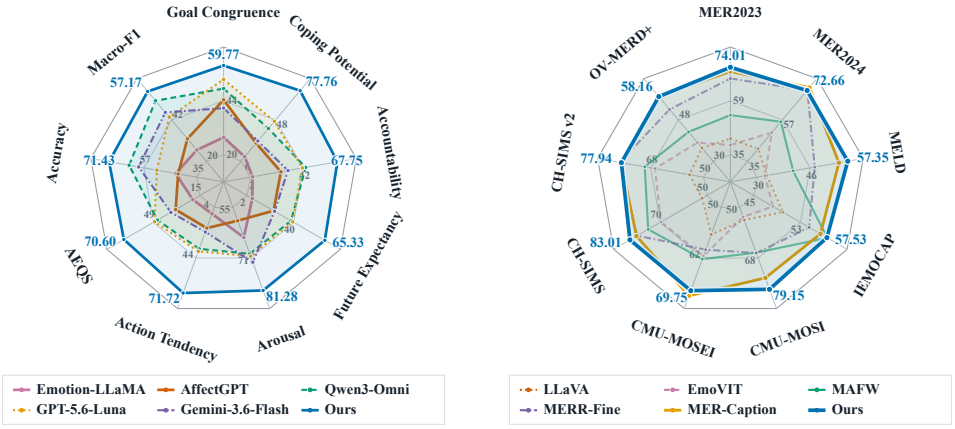
Recent multimodal large language models (MLLMs) increasingly incorporate explainable reasoning for emotion understanding. However, reasoning based mainly on observable affective cues can reduce emotion understanding to superficial cue-label associations, giving rise to the *Clever Hans effect*[1]. Such shortcuts become unreliable when affective cues are implicit, conflicting across modalities, linguistically misleading, or obscured by redundant details. In contrast, human emotions are shaped by how individuals interpret and evaluate surrounding events beyond observable cues. Inspired by appraisal theories of emotion, we formulate multimodal emotion understanding as a progression from perception to cognitive appraisal, and introduce a dataset, a model, and a benchmark to support this novel paradigm. ❶ **CogEmo-40K** is a large-scale instruction-tuning dataset constructed through a perception-to-appraisal pipeline to elicit evidence-grounded reasoning across six cognitive appraisal dimensions underlying emotion. ❷ **CogEmo-MoE** is a compact sparse MLLM that introduces interleaved MoE blocks for appraisal-specific adaptation, enabling effective appraisal reasoning at a substantially smaller scale than typical emotion MLLMs. ❸ **CogEmo-Bench** introduces an Appraisal Evidence Quality Score (AEQS) to assess cognitive-affective understanding across six complementary appraisal dimensions, addressing the limitation of conventional emotion metrics that evaluate what emotion is predicted but not why it arises. Extensive experiments show that our paradigm not only leads CogEmo-Bench (CogEmo-MoE: **70.60** vs. 56.30 AEQS for GPT-5.2; **71.43%** vs. 63.10% accuracy and **57.17%** vs. 52.19% M-F1 for Qwen3-Omni), but also exhibits strong cross-domain generalization. Our findings suggest that the perception-to-appraisal reasoning can move beyond surface-level cue-label associations toward more reliable multimodal emotion understanding and closer cognitive alignment between MLLMs and humans. Project Page: https://anonymous.4open.science/r/CogEmo-70C8/.

## 1 INTRODUCTION

Emotion understanding is not simply a matter of recognizing observable affective expressions. A prominent perspective in emotion research holds that emotional responses depend not only on what happens, but also on how individuals interpret and evaluate what happens to them (Ellsworth & Scherer, 2003; Moors et al., 2013). The same event may lead to different emotional responses depending on whether it supports or obstructs personal goals, whether the situation is perceived as manageable, who is considered responsible, and how future consequences are anticipated (Roseman, 1991; Yeo & Ong, 2024; Liang et al., 2026). From this perspective, observable behavior provides only part of the information needed for emotion understanding: perception reveals ***what is happening***, whereas cognitive appraisal concerns ***what the situation means to the individual*** (Ong et al., 2015; Skerry & Saxe, 2015; Teotia et al., 2026).

[1] **Clever Hans Effect:** A system exhibits seemingly intelligent behavior but actually relies on hidden cues, data biases, or superficial associations rather than genuinely understanding the mechanisms underlying the task.

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




(a) Comparison of Representative MLLMs on CogEmo-Bench (b) Cross-Domain Generalization Results on MERBench

Figure 1: Performance comparison of representative MLLMs on cognitive appraisal reasoning and cross-domain emotion understanding.

In recent years, multimodal large language models (MLLMs) have advanced emotion recognition toward more flexible and interpretable multimodal emotion understanding by generating natural-language explanations alongside emotion predictions (Lian et al., 2025a; Cheng et al., 2024b; Zhao et al., 2025; Zhang et al., 2026b). However, plausible explanations do not necessarily imply genuine reasoning about why an emotion arises (Luo et al., 2026). Models often still rely on statistical associations between observable facial, acoustic, or linguistic cues and emotion categories (Jiang et al., 2025; Yeo & Jaidka, 2025; Chen et al., 2026), while accurate perception alone does not ensure that perceptual evidence is explicitly organized into cognition-level appraisal reasoning (Ong et al., 2015; Houlihan et al., 2023; Yeo & Jaidka, 2025). Such ***cognitive misalignment*** becomes particularly problematic when multimodal evidence is ambiguous, conflicting, or context-dependent, and may contribute to inappropriate empathic responses (Wang et al., 2022) or unreliable emotion-informed decisions in mental-health settings (Zhang et al., 2023). Yet existing multimodal emotion datasets largely provide emotion labels or perception-level descriptions (Lian et al., 2026; Cheng et al., 2024b; Lian et al., 2025a), while conventional metrics mainly evaluate final predictions, leaving cognition-level reasoning insufficiently supervised and assessed.

From a psychological perspective, we model cognitive appraisal as a key intermediate process in emotion understanding, rather than treating emotion as a direct cue-to-label mapping. Specifically, we characterize appraisal reasoning through six complementary dimensions: goal congruence, coping potential, accountability, future expectancy, arousal, and action tendency (Roseman, 1991; Yeo & Ong, 2024), providing an interpretable structure for reasoning about why an emotion arises. However, existing emotion-oriented MLLMs provide limited support for cognition-specific specialization, and scaling model capacity alone is inefficient for targeted cognitive adaptation.

The main contributions of this paper are summarized as follows:

① We introduce a **perception-to-appraisal paradigm** for multimodal emotion understanding, shifting from direct cue-to-label prediction toward reasoning about how perceptual evidence is cognitively interpreted before emotion arises. To support this paradigm, we construct **CogEmo-40K**, a large-scale human-aligned instruction-tuning dataset for cognition-grounded emotion reasoning.

② We propose **CogEmo-MoE**, a lightweight sparse MLLM for cognitive appraisal reasoning with perception-preserving adaptation. Instead of scaling up or broadly fine-tuning the backbone, we preserve established multimodal perception and specialize only sparse experts for appraisal-specific adaptation, enabling strong appraisal reasoning within a compact model.

③ We develop **CogEmo-Bench**, a fine-grained benchmark with an Appraisal Evidence Quality Score (AEQS) for evaluating cognitive appraisal reasoning and emotion recognition. Across diverse open- and closed-source MLLMs, CogEmo-MoE improves AEQS over GPT-5.2 from $\mathbf{56.30} \rightarrow \mathbf{70.60}$ and accuracy over Qwen3-Omni from $\mathbf{63.10}\% \rightarrow \mathbf{71.43}\%$, revealing both the strengths and limitations of current models in cognitive emotion understanding.

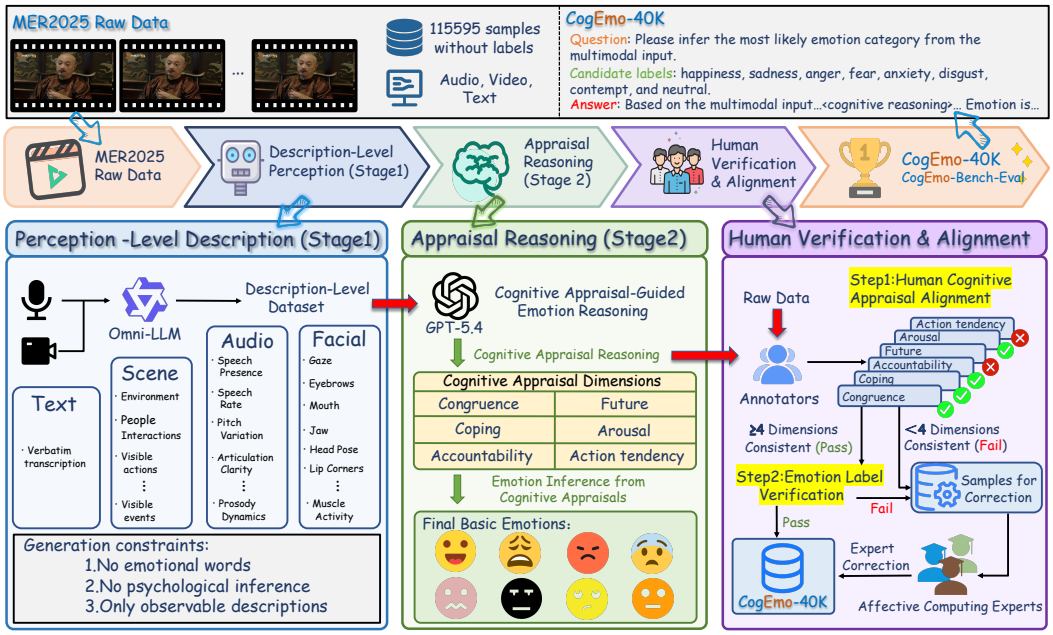



Figure 2: Illustration of CogEmo-40K Dataset Construction.

## 2 CogEmo-40K: Cognition-Grounded Multimodal Emotion Reasoning Dataset

### 2.1 Dataset Overview

Existing multimodal emotion datasets mainly provide emotion labels or perception-level descriptions, offering limited supervision for the cognitive appraisal processes underlying emotional responses. To address this gap, we construct **CogEmo-40K**, a cognition-grounded multimodal emotion reasoning dataset. Following the screening protocol of AffectGPT (Lian et al., 2025a), we annotate the same 115,595 MER2025 (Lian et al., 2025b) source samples used by MERCaption (Lian et al., 2025a) through a two-stage perception-to-appraisal pipeline, followed by human-guided alignment and refinement to obtain approximately 40K high-quality instances. A held-out subset forms **CogEmo-Bench-Eval**, providing over 6,000 human-verified dimension-level appraisal references across six appraisal dimensions. The overall construction pipeline is illustrated in Figure 2. Each instance explicitly follows a *multimodal observation* $\rightarrow$ *cognitive appraisal* $\rightarrow$ *emotion state* reasoning process, providing supervision not only for ***what emotion is expressed***, but also for ***how emotional responses emerge from interpreted situations***. Detailed dataset statistics are provided in Appendix B.4.

### 2.2 Two-Stage Perception-to-Appraisal Data Generation

As illustrated in Figure 2, we decompose annotation generation into *perception* and *appraisal* to reduce the entanglement between perceptual description and cognitive interpretation, while discouraging shortcut cue-label associations.

**Stage 1: Perception-Level Description.** The first stage converts multimodal inputs into emotion-agnostic descriptions of observable evidence across four sources: *Scene*, *Face*, *Speaker Audio*, and *Transcript*. Emotion labels and latent psychological interpretations are explicitly excluded to prevent appraisal leakage, yielding an evidence-grounded representation for subsequent cognitive reasoning.

**Stage 2: Cognition-Level Appraisal Reasoning.** The second stage interprets the perceptual evidence through six appraisal dimensions motivated by cognitive appraisal theory: *Goal Congruence*, *Coping*, *Accountability*, *Future Expectancy*, *Arousal*, and *Action Tendency*. Rather than directly inferring an emotion label from observable cues, the model is required to explain how the perceived

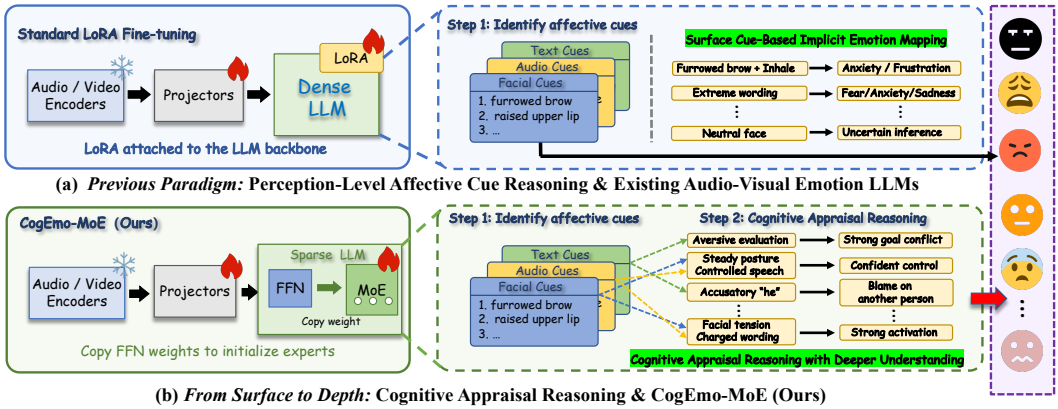



Figure 3: Comparison of surface cue-based emotion reasoning and our cognitive appraisal-based CogEmo-MoE framework.

situation is evaluated along these dimensions before deriving the emotional response. Each appraisal is grounded in supporting evidence from perception-level description, yielding an explicit reasoning trajectory from multimodal observation to cognitive interpretation and finally to emotion state.

### 2.3 Human-Guided Cognitive Alignment

Reliable cognition-grounded supervision requires consistency across the full *perception–appraisal–emotion* chain. We therefore conduct human-guided alignment at perceptual, cognitive, and emotional levels.

**Perception-Level Quality Assessment.** We randomly sample 1,000 generated descriptions and ask five researchers with backgrounds in affective computing to evaluate their *Accuracy*, *Completeness*, and *Objectivity* on a five-point scale, following Appendix B.3. The average score for each criterion exceeds 4.0, indicating that the generated descriptions provide reliable and largely emotion-agnostic perceptual evidence.

**Cognitive Appraisal and Emotion Alignment.** We recruit 300 trained volunteers to assess whether the generated cognitive appraisal reasoning is supported by the original multimodal evidence and aligned with plausible human interpretations. For each sample, volunteers compare the generated appraisals with human interpretations across the six cognitive appraisal dimensions, while also verifying whether the final emotion label is consistent with both the multimodal context and the inferred cognitive appraisal reasoning. Rather than directly discarding inconsistent, uncertain, or ambiguous cases, we adopt a human-in-the-loop refinement process to revise the cognitive reasoning and emotion labels based on human feedback, followed by further expert review. This process preserves challenging samples while improving the consistency between cognitive appraisals and emotion labels.

## 3 CogEmo-MoE: Model Design

### 3.1 Architecture Overview

Existing multimodal emotion models often rely on large backbones for affective understanding. We instead explore whether compact models can acquire effective cognitive appraisal through targeted specialization, and introduce **CogEmo-MoE**, a lightweight sparse model built upon MoE-LLaVA. It adopts `Qwen-1.8B` as the language backbone and `CLIP/CLAP` as the visual/audio encoders. As shown in Fig. 3, sparse experts initialized from the corresponding FFN weights are specialized for cognitive appraisal reasoning. Given video $V$, audio $A$, and instruction $I$, multimodal features are projected into the language embedding space and combined with the instruction embeddings:

$$Y = LM_{\text{MoE}}\left([P_v(f_{\text{CLIP}}(V)); P_a(f_{\text{CLAP}}(A)); E(I)]\right), \tag{1}$$

where $P_v(\cdot)$ and $P_a(\cdot)$ are modality-specific projectors, $E(I)$ is the instruction embedding, and $LM_{\text{MoE}}$ denotes the MoE-augmented language backbone for cognition-specific reasoning.

### 3.2 Sparse Cognitive Expert Adaptation

Cognitive appraisal reasoning builds on earlier multimodal perception. We therefore interleave sparse MoE modules across the Transformer backbone. Each MoE layer contains multiple experts and a lightweight router, which computes the selection probability for expert $e$ given hidden state $h$:

$$p(e|h) = \mathrm{Softmax}(W_r h), \tag{2}$$

where $W_r$ denotes the routing parameters. Only the top-$k$ experts are activated for each input:

$$\mathrm{MoE}(h) = \sum\nolimits_{i \in \mathrm{TopK}(p)} p_i E_i(h), \tag{3}$$

where $E_i(\cdot)$ denotes the $i$-th expert and $p_i$ is its routing weight.

### 3.3 Progressive Perception-to-Cognition Training

CogEmo-MoE follows a progressive three-stage training strategy. All stages are optimized with the same autoregressive objective, while differing in supervision granularity and trainable modules. Given the stage-specific dataset $\mathcal{D}_s$, the training objective is

$$\mathcal{L}_s = -\sum\nolimits_{(X,I,Y)\in\mathcal{D}_s} \sum\nolimits_{t=1}^{|Y|} \log p_{\theta_s}\left(y_t \mid y_{<t}, X, I\right), \qquad s \in \{1,2,3\}, \tag{4}$$

where $X$, $I$, and $Y$ denote the multimodal input, instruction, and stage-specific target, respectively, and $\theta_s$ are the trainable parameters. The training data used are summarized in Table 1. **MERR-Coarse** is a coarse-grained face–audio emotion dataset from Emotion-LLaMA (Cheng et al., 2024a), whereas our **CogEmo-Desc** comprises fine-grained perception-level descriptions generated during the automatic annotation process.

**Stage 1: Multimodal Alignment.** Using **MERR-Coarse**, we optimize only the modality-specific projectors, while keeping the pretrained encoders and language backbone frozen, to align visual and audio representations with the language embedding space.

**Stage 2: Perception Enhancement.** Starting from the aligned model, we introduce **CogEmo-Desc** as fine-grained description-level supervision to strengthen the model's perception of facial and acoustic details. During this stage, both the language backbone and modality-specific projectors are unfrozen and jointly optimized.

**Stage 3: Cognitive Reasoning Adaptation.** Finally, we use **CogEmo-40K** as cognition-level supervision to specialize the sparse experts for appraisal-based emotion reasoning on top of the multimodal perception acquired in the previous stages. During this stage, the language backbone is frozen and only the sparse expert modules are optimized, concentrating adaptation on cognition-specific reasoning.

Table 1: Training data used in the three stages of CogEmo-MoE.

| Usage | Source | #Samples |
|---|---|---|
| Stage I | MERR-Coarse | 28,618 |
| Stage II | CogEmo-Desc | 114,587 |
| Stage III | CogEmo-40K | 38,532 |

## 4 CogEmo-Bench: Cognition-Grounded Emotion Understanding Benchmark

### 4.1 CogEmo-Bench Overview

The introduced **CogEmo-Bench** jointly evaluates cognitive appraisal reasoning and conventional emotion recognition. An ***Appraisal Extractor*** normalizes free-form MLLM responses into the six appraisal dimensions defined in Sec. 2.2, without judging correctness or adding evidence. The extracted appraisals are compared with human-verified references using the **Appraisal Evidence Quality Score (AEQS)**, while final emotion predictions are evaluated with standard classification metrics.

### 4.2 Evaluation Metrics

Conventional emotion metrics evaluate final predictions but do not reveal whether the underlying cognitive evidence is well aligned with human interpretation. We therefore introduce a new metric to evaluate appraisal evidence across the six cognitive dimensions.

**Appraisal Evidence Quality Score (AEQS).** For sample $i$ and appraisal dimension $d$, let $\hat{e}_{i,d}$ and $e_{i,d}$ denote the extracted model evidence and human-verified reference, respectively. Using a fixed embedding encoder $\phi(\cdot)$, the evidence similarity is defined as

$$s_{i,d} = \cos(\phi(\hat{e}_{i,d}), \phi(e_{i,d})) \,. \tag{5}$$

If only one of $\hat{e}_{i,d}$ and $e_{i,d}$ is empty, we assign $s_{i,d} = 0$; if both are empty, the sample is excluded from the corresponding dimension. Let $\mathcal{S}_d$ denote the valid sample set for dimension $d$. The dimension-level score is

$$S_d = \frac{1}{|\mathcal{S}_d|} \sum\nolimits_{i \in \mathcal{S}_d} s_{i,d}. \tag{6}$$

AEQS is then computed by averaging across the six appraisal dimensions:

$$\text{AEQS} = \frac{\alpha}{D} \sum\nolimits_{d=1}^{D} S_d, \tag{7}$$

where $D$ denotes the number of appraisal dimensions and $\alpha$ is a scaling factor; we set $D = 6$ and $\alpha = 100$ in this work. Higher AEQS indicates stronger semantic alignment between model-generated appraisal evidence and human-verified cognitive references.

**Emotion Recognition.** We additionally report standard **Accuracy** and **Macro-F1** for final emotion prediction, with detailed computation provided in Appendix D.3.

## 5 Results and Observations

**Main results.** Detailed implementation information can be found in Appendix C. Table 2 compares representative MLLMs on both cognitive appraisal reasoning and emotion recognition. The results show substantial differences across models in AEQS and final emotion performance. Some models achieve competitive emotion classification despite relatively weak appraisal reasoning, while larger general-purpose and proprietary models generally perform better on both aspects. CogEmo-MoE achieves a favorable balance between appraisal reasoning and emotion recognition despite its compact model scale. Specifically, it improves AEQS over GPT-5.2 from $\mathbf{56.30} \rightarrow \mathbf{70.60}$, while outperforming Qwen3-Omni in accuracy from $\mathbf{63.10}\% \rightarrow \mathbf{71.43}\%$ and Macro-F1 from $\mathbf{52.19}\% \rightarrow \mathbf{57.17}\%$. Additional model-size comparisons are provided in Appendix C.2.

**Obs. 1: Emotion recognition does not necessarily imply cognitive understanding.** VideoChat-Flash (Li et al., 2025b) achieves strong emotion prediction despite limited explicit appraisal reasoning, while emotion-specialized models such as AffectGPT (Lian et al., 2025a) and Emotion-LLaMA (Cheng et al., 2024a) also show limited transfer to cognition-level understanding. These results indicate that emotion recognition and surface cue-to-label mapping cannot substitute for cognitive appraisal reasoning about why emotions arise. By contrast, cognition-oriented supervision enables our CogEmo-MoE to improve both appraisal reasoning and emotion discrimination within a compact model.

**Obs. 2: True emotion intelligence requires adaptive reasoning across different emotion spaces.** Strong performance within a fixed label space does not necessarily indicate robust emotion understanding. A capable emotion model should adapt its judgments to task-specific emotion taxonomies rather than rely on surface-level cue-to-label mappings. Although open-vocabulary approaches such as AffectGPT (Lian et al., 2025a) improve output flexibility, evaluation protocols that count a prediction as correct whenever the target emotion appears among multiple generated candidates may overestimate performance without establishing that the model selected the most appropriate emotion. In contrast, as shown in Fig. 1(b) and detailed in Appendix E.2, our results show that cognition-oriented reasoning transfers across different emotion spaces and consistently improves emotion recognition tasks, suggesting that cognitive appraisal provides a more generalizable basis for emotion understanding.

Table 2: Comparison of different MLLMs on **CogEmo-Bench**. A, V, and T denote audio, visual, and textual modalities, respectively. Cong., Cop., Account., Fut., Aro., and Act. denote Goal congruence, Coping potential, Accountability, Future expectancy, Arousal, and Action tendency.

| Model | A | V | T | Cong. | Cop. | Account. | Fut. | Aro. | Act. | AEQS | Acc. | M-F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Open-source MLLMs* | | | | | | | | | | | | |
| OneLLM (Han et al., 2024) | ✓ | × | ✓ | 31.47 | 8.02 | 10.10 | 9.59 | 30.90 | 6.81 | 16.15 | 30.16 | 26.03 |
| SECap (Xu et al., 2024) | ✓ | × | ✓ | 11.12 | 7.42 | 4.59 | 3.97 | 42.26 | 4.23 | 12.26 | 30.85 | 21.25 |
| Qwen2-Audio (Chu et al., 2024) | ✓ | × | ✓ | 39.27 | 18.16 | 22.26 | 24.85 | 30.55 | 15.60 | 25.12 | 43.55 | 32.68 |
| SALMONN (Tang et al., 2024) | ✓ | × | ✓ | 42.87 | 18.20 | 20.18 | 24.10 | 27.17 | 10.67 | 23.87 | 43.69 | 33.43 |
| Otter (Li et al., 2025a) | × | ✓ | ✓ | 26.09 | 3.83 | 8.90 | 3.89 | 15.65 | 3.58 | 10.32 | 36.61 | 24.69 |
| LLaMA-VID (Li et al., 2024) | × | ✓ | ✓ | 47.73 | 16.35 | 20.33 | 21.62 | 39.77 | 13.79 | 26.60 | 36.71 | 31.84 |
| VideoChat-Flash (Li et al., 2025b) | × | ✓ | ✓ | 6.82 | 1.22 | 1.26 | 1.75 | 4.32 | 0.78 | 2.69 | 55.36 | 39.11 |
| Chat-UniVi (Jin et al., 2024) | × | ✓ | ✓ | 49.96 | 22.80 | 26.66 | 27.92 | 46.63 | 19.73 | 32.28 | 34.03 | 29.86 |
| mPLUG-Owl3 (Ye et al., 2025) | × | ✓ | ✓ | 47.78 | 18.53 | 23.11 | 27.49 | 58.21 | 17.30 | 32.07 | 45.54 | 34.72 |
| Qwen3.8-Max (Team, 2026) | × | ✓ | ✓ | 51.30 | 44.37 | 47.12 | 42.56 | 67.71 | 43.94 | 49.50 | 57.04 | 46.29 |
| Kimi-K3 (Kimi Team, 2026) | × | ✓ | ✓ | 52.02 | 43.58 | 45.23 | 40.54 | 68.38 | 42.42 | 48.70 | 46.13 | 39.04 |
| PandaGPT (Su et al., 2023) | ✓ | ✓ | ✓ | 50.97 | 14.65 | 19.92 | 21.22 | 32.92 | 13.55 | 25.54 | 25.79 | 18.78 |
| Emotion-LLaMA (Cheng et al., 2024a) | ✓ | ✓ | ✓ | 27.74 | 10.42 | 3.71 | 6.24 | 64.49 | 9.04 | 20.27 | 43.75 | 25.90 |
| AffectGPT (Lian et al., 2025a) | ✓ | ✓ | ✓ | 44.46 | 24.26 | 25.05 | 23.58 | 59.20 | 20.39 | 32.83 | 42.76 | 31.90 |
| Qwen3-Omni (Xu et al., 2025) | ✓ | ✓ | ✓ | 49.50 | 39.21 | 43.74 | 37.79 | 69.69 | 36.76 | 46.12 | 63.10 | 52.19 |
| *Closed-source MLLMs* | | | | | | | | | | | | |
| GPT-5.2 (OpenAI, 2025) | × | ✓ | ✓ | 55.89 | 56.78 | 53.36 | 48.37 | 70.34 | 53.07 | 56.30 | 46.33 | 36.92 |
| GPT-5.6-Luna (OpenAI, 2026b) | × | ✓ | ✓ | 53.57 | 46.48 | 41.23 | 39.18 | 70.41 | 38.76 | 48.27 | 51.59 | 43.46 |
| Grok-4.3 (xAI, 2026) | × | ✓ | ✓ | 51.67 | 36.80 | 40.67 | 37.82 | 69.46 | 36.88 | 45.55 | 46.83 | 39.79 |
| Gemini-3.6-Flash (Google, 2026) | ✓ | ✓ | ✓ | 40.79 | 26.42 | 30.44 | 24.36 | 72.33 | 23.45 | 36.30 | 59.13 | 45.99 |
| *Our CogEmo Models* | | | | | | | | | | | | |
| CogEmo-MoE | ✓ | × | ✓ | 58.68 | 77.04 | 67.44 | 65.04 | 81.03 | **71.81** | 70.17 | 63.59 | 50.18 |
| CogEmo-MoE | × | ✓ | ✓ | 59.56 | 77.56 | 66.88 | 64.59 | 80.32 | 71.26 | 70.03 | 62.70 | 50.07 |
| CogEmo-MoE | ✓ | ✓ | ✓ | **59.77** | **77.76** | **67.75** | **65.33** | **81.28** | 71.72 | **70.60** | **71.43** | **57.17** |

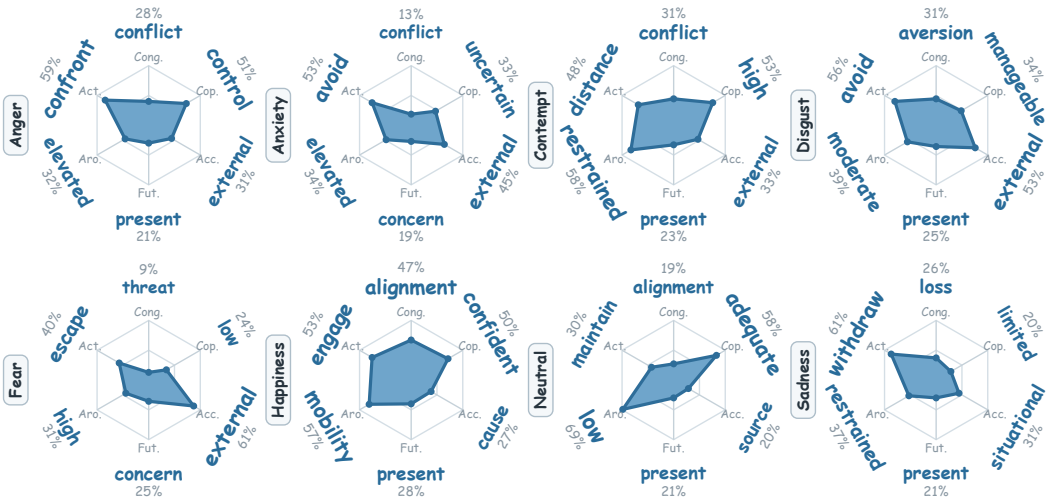
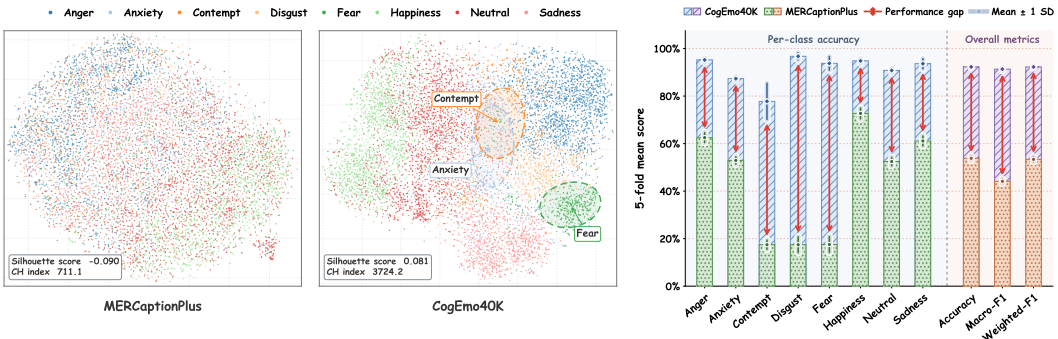



Figure 4: Emotion-Specific Cognitive Language Profiles in CogEmo-40K. Radar plots show six-dimensional appraisal patterns, with radial extent and label size indicating phrase prevalence.

**Cognitive Language Profiles.** We further analyze the structured outputs of the fixed Appraisal Extractor by aggregating recurring phrase-level patterns for each emotion and appraisal dimension. Figure 4 visualizes the resulting emotion-specific profiles across six cognitive dimensions, where radial extent and label size indicate the prevalence of the most frequent phrase within each emotion category. The surrounding terms summarize the corresponding appraisal semantics. These percentages reflect phrase prevalence rather than appraisal intensity. The original top-3 phrases and their occurrence rates are reported in Appendix B.4.3.

**Obs. 3: Emotion categories emerge from structured configurations of cognitive appraisals rather than isolated affective cues.** As shown in Figure 4, emotions exhibit distinct joint appraisal profiles even when individual dimensions overlap. For example, anger is characterized by conflict, maintained control, external attribution, elevated arousal, and confrontation, whereas sadness is associated with loss, limited coping, situational attribution, restrained arousal, and withdrawal. Likewise, fear emphasizes threat, low coping, high arousal, and escape, while anxiety is more associated with conflict, uncertain coping, elevated arousal, and avoidance. These patterns suggest that emotion differentiation is better characterized by configurations of cognitive appraisals than by isolated affective cues.




(a) Emotion-label structure in t-SNE space

(b) Emotion-label probing performance

Figure 5: Emotion Alignment of Cognition-Grounded Supervision. We compare CogEmo-40K and MERCaption+ on their intersected samples under the same eight-way emotion partition.

**Emotion Alignment of Cognition-Grounded Supervision.** On the samples shared by CogEmo-40K and MERCaption+ (Lian et al., 2025a), we compare how the two forms of supervision encode emotion-related information. As shown in Figure 5(a), cognition-grounded reasoning yields clearer emotion-label structure than open-ended descriptions. For Figure 5(b), we conduct five-fold cross-validation with the same `XLM-RoBERTa` classification framework for each supervision source. After removing explicit emotion-conclusion sentences, the classifier predicts eight emotion categories from each source's text. CogEmo-40K achieves stronger per-class and overall performance, suggesting that its appraisal reasoning encodes information more closely aligned with the final emotion labels. Detailed settings are provided in Appendix E.1.

**Effect of MoE Placement.** We investigate where the sparse expert layers should be introduced. We compare four placement strategies, including First-Half, Second-Half, Interval, and All. The Interval strategy achieves the best overall performance, raising accuracy from $\mathbf{64.29\%} \rightarrow \mathbf{71.43\%}$ over First-Half placement and supporting our design of inserting MoE layers at alternating Transformer blocks for more balanced cognition-specific adaptation.

**Dense vs. Sparse Experts.** We further compare the dense baseline with different sparse routing strategies. Simply increasing the number of experts with Top-1 routing does not improve performance. In contrast, Top-2 routing improves both emotion recognition and cognitive appraisal quality, including accuracy from $\mathbf{65.97\%} \rightarrow \mathbf{71.43\%}$ and Macro-F1 from $\mathbf{52.28\%} \rightarrow \mathbf{57.17\%}$ under the four-expert setting, suggesting that collaborative expert activation is more effective for cognition-level emotion reasoning.

Table 3: Effect of different MoE placement strategies on **CogEmo-Bench-Eval**.

| Architecture | AEQS | Acc. | M-F1 |
|---|---|---|---|
| First-Half | 68.76 | 64.29 | 50.80 |
| Second-Half | 70.10 | 68.45 | 54.57 |
| Interval | **70.60** | **71.43** | **57.17** |
| All | 69.76 | 70.04 | 55.49 |

Table 4: Comparison of dense and sparse expert configurations on **CogEmo-Bench-Eval**.

| Configuration | AEQS | Acc. | M-F1 |
|---|---|---|---|
| *Dense Model* | | | |
| Baseline | 70.48 | 67.26 | 53.27 |
| *Sparse MoE* | | | |
| 4E / Top-1 | 70.02 | 65.97 | 52.28 |
| 4E / Top-2 | **70.60** | **71.43** | **57.17** |
| 8E / Top-1 | 69.80 | 64.68 | 51.40 |
| 8E / Top-2 | 69.76 | 70.04 | 55.49 |

# 6 Related Work

## 6.1 From Multimodal Perception to Emotion Reasoning

MLLMs align visual, acoustic, and temporal evidence with language, providing a foundation for open-ended multimodal interpretation (Han et al., 2024; Li et al., 2024; Chu et al., 2024). In affective computing, earlier systems mapped fused multimodal representations to predefined emotion categories (Zhang et al., 2024; Zadeh et al., 2017; Tsai et al., 2019), while contextual models demonstrated the importance of scenes and conversational history beyond isolated expressions (Kosti et al., 2017; Lee et al., 2019; Poria et al., 2017). Emotion-oriented MLLMs further generate explanations alongside predictions (Cheng et al., 2024a; Lian et al., 2025a; Zhao et al., 2025), supported by explanation-rich datasets and unified benchmarks (Cheng et al., 2024a; Lian et al., 2025a; 2026; Zhang et al., 2026a). Nevertheless, current supervision and evaluation still emphasize final labels, observable cues, or textual plausibility. Because fluent rationales can be unstable and benchmark formats can admit shortcuts (Jiang et al., 2025; Salido et al., 2025), explanation generation alone does not establish that a model captures how an emotion arises.

## 6.2 Cognitive Appraisal as an Intermediate Representation

Appraisal theories locate emotion not in observable cues alone, but in how an individual evaluates an event with respect to goals, coping potential, responsibility, expected outcomes, action readiness and so on (Roseman, 1991; Frijda, 1987; Ellsworth & Scherer, 2003; Moors et al., 2013). Empirical and computational studies further link appraisal patterns and abstract event representations to emotion judgments (Yeo & Ong, 2024; Ong et al., 2015; Skerry & Saxe, 2015; Houlihan et al., 2023). Although appraisal has informed text-based emotion modeling and Theory-of-Mind evaluation (Hofmann et al., 2020; Troiano et al., 2023; Yeo & Jaidka, 2025), it has yet to be integrated systematically with multimodal perception and free-form reasoning. We operationalize appraisal as an explicit intermediate representation between observation and emotion: CogEmo-40K provides perception-to-appraisal supervision, CogEmo-MoE models this transition, and CogEmo-Bench evaluates it.

# 7 Conclusion

We introduce a perception-to-appraisal paradigm for multimodal emotion understanding. CogEmo-40K supplies cognition-grounded supervision, CogEmo-MoE applies sparse appraisal-specific adaptation, and CogEmo-Bench uses AEQS to evaluate appraisal evidence alongside emotion recognition. On CogEmo-Bench, CogEmo-MoE achieves 70.60 AEQS, 71.43% accuracy, and 57.17% Macro-F1. Analyses of appraisal language reveal distinct patterns across emotion categories, while cross-domain experiments suggest that cognition-grounded supervision can transfer beyond the training domain. These results support appraisal reasoning as an intermediate representation beyond surface-level cue-to-label emotion recognition.

## Acknowledgments

We gratefully acknowledge the psychology researchers and domain experts who provided valuable consultation on cognitive appraisal theory, annotation design, and human cognitive alignment throughout this project. We also thank the affective computing researchers, annotators, and volunteers who contributed to the verification and refinement of CogEmo-40K.

## AI use statement

Generative AI tools were used to improve the clarity, grammar, and readability of the manuscript. All AI-assisted edits were reviewed and revised by the authors, who take full responsibility for the final content.

## Reproducibility Statement

We provide detailed descriptions of the data construction, model training, and evaluation procedures to facilitate reproducibility. The complete two-stage data-generation prompts and human-verification protocols for CogEmo-40K are provided in Appendix B and Appendix B.3. Model architecture, three-stage training configurations, optimization settings, and training prompts for CogEmo-MoE are reported in Appendix C. For CogEmo-Bench, we provide the evaluation prompts, Appraisal Extractor, metric computation procedure, and model-specific evaluation settings in Appendix D. Additional experimental details, including the 5-fold cognition–emotion alignment analysis and cross-domain MERBench evaluation, are provided in Appendix E.1 and Appendix E.2, respectively. These details specify the data sources, model configurations, training hyperparameters, prompts, and evaluation protocols used throughout the experiments.

## Ethics Statement

CogEmo-40K is constructed from the existing MER2025 multimodal dataset rather than from newly collected audiovisual recordings. Human involvement in our study is limited to annotation, verification, and refinement of automatically generated perceptual descriptions, cognitive appraisal reasoning, and emotion labels. Because emotion and cognitive appraisal are inherently context-, culture-, and individual-dependent, the resulting annotations should be viewed as evidence-grounded interpretations rather than definitive descriptions of an individual's internal psychological state. We therefore caution against using model-generated appraisal reasoning for clinical diagnosis, psychological profiling, or other high-stakes decisions without appropriate human oversight. Potential biases inherited from the source data, including linguistic, cultural, and domain biases, are further discussed in the limitations section.

## References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do audio llms really listen, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5848–5877, 2026.

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

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-LLaMA: Multimodal Emotion Recognition and Reason-

ing with Instruction Tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024b.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Phoebe C Ellsworth and Klaus R Scherer. Appraisal processes in emotion. 2003.

Nico H Frijda. Emotion, cognitive structure, and action tendency. *Cognition and emotion*, 1(2): 115–143, 1987.

Google. Introducing gemini 3.6 flash, 3.5 flash-lite, and 3.5 flash cyber, July 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26574–26585. IEEE, 2024.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. Appraisal theories for emotion classification in text. In *Proceedings of the 28th international conference on computational linguistics*, pp. 125–138, 2020.

Sean Dae Houlihan, Max Kleiman-Weiner, Luke B Hewitt, Joshua B Tenenbaum, and Rebecca Saxe. Emotion prediction as computation over a generative theory of mind. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 381(2251):20220047, 2023.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13700–13710. IEEE, 2024.

Kimi Team. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1960–1968. IEEE, 2017.

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 10142–10151. IEEE, 2019.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7543–7557, 2025a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European conference on computer vision*, pp. 323–340. Springer, 2024.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 9610–9614, 2023.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 41–48, 2024.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025a.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, et al. Mer 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13837–13842, 2025b.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, et al. Ov-mer: Towards open-vocabulary multimodal emotion recognition. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025c.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Qiao Liang, Ying Shen, Yao Liu, Tiantian Chen, and Lin Zhang. Why do emotions change? appraisal-guided reasoning for emotion–cause triplet extraction in conversations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11750–11772, 2026.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, 2026.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pp. 247–258, 2022a.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 24–32, 2022b.

Meng Luo, Bobo Li, Shanqing Xu, Shize Zhang, Qiuchan Chen, Menglu Han, Wenhao Chen, Yanxiang Huang, Hao (Scofield) Fei, Mong-Li Lee, and Wynne Hsu. Unveiling the cognitive compass: Theory-of-mind–guided multimodal emotion reasoning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 85428–85498, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/89ab43b626b480347dbaccda7be8aedb-Paper-Conference.pdf.

Agnes Moors, Phoebe C Ellsworth, Klaus R Scherer, and Nico H Frijda. Appraisal theories of emotion: State of the art and future development. *Emotion review*, 5(2):119–124, 2013.

Desmond C Ong, Jamil Zaki, and Noah D Goodman. Affective cognition: Exploring lay theories of emotion. *Cognition*, 143:141–162, 2015.

OpenAI. Introducing gpt-5.2, December 2025. URL https://openai.com/index/introducing-gpt-5-2/.

OpenAI. Introducing gpt-5.4, mar 2026a. URL https://openai.com/index/introducing-gpt-5-4/.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition, July 2026b. URL https://openai.com/index/gpt-5-6/.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 873–883, 2017.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ira J Roseman. Appraisal determinants of discrete emotions. *Cognition & Emotion*, 5(3):161–200, 1991.

Eva Sánchez Salido, Julio Gonzalo, and Guillermo Marco. None of the others: a general technique to distinguish reasoning from memorization in multiple-choice llm evaluation benchmarks. *arXiv preprint arXiv:2502.12896*, 2025.

Amy E Skerry and Rebecca Saxe. Neural representations of emotion are organized around abstract event features. *Current biology*, 25(15):1945–1954, 2015.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the Era of Interactive Assistants!*, pp. 11–23, 2023.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024.

Qwen Team. Qwen3. 8: A new bar for coding and cowork. *Qwen, August*, 2026.

Jayant Teotia, Xiaowei Wang, Xulang Zhang, Rui Mao, and Erik Cambria. Evaluating multimodal large language model narrative interpretation through the lens of appraisal theory. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, volume 11, pp. 11417–11426, 2026.

Enrica Troiano, Laura Ana Maria Oberländer, and Roman Klinger. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72, 2023.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4634–4645, 2022.

xAI. Grok 4.3, 2026. URL https://docs.x.ai/developers/models/grok-4.3.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26586–26595. IEEE, 2024.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19323–19331, 2024.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 98891–98913, 2025.

Gerard C Yeo and Desmond C Ong. Associations between cognitive appraisals and emotions: A meta-analytic review. *Psychological Bulletin*, 150(12):1440, 2024.

Gerard Christopher Yeo and Kokil Jaidka. Beyond context to cognitive appraisal: Emotion reasoning as a theory of mind benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26517–26525, 2025.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3718–3727, 2020.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115/.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, YiFan Zhang, Renrui Zhang, et al. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 48767–48807, 2026a.

Hongrui Zhang, Daiqing Wu, Yangyang Li, Kuien Liu, Yuhui Wang, Yu Zhou, and Sicheng Zhao. Multimodal emotion recognition with large language models. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 8088–8096. International Joint Conferences on Artificial Intelligence Organization, 8 2026b. doi: 10.24963/ijcai.2026/897. URL https://doi.org/10.24963/ijcai.2026/897. Survey Track.

Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, 237:121692, 2024.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246, 2023.

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

# APPENDIX CONTENTS

**Appendix A.** Limitations and Social Impact . . . . . 16
**Appendix B.** Details of CogEmo-40K Dataset . . . . . 17
**Appendix C.** Implementation Details . . . . . 23
**Appendix D.** Details of CogEmo-Bench . . . . . 25
**Appendix E.** Additional Experimental Results . . . . . 28

## A LIMITATIONS AND SOCIAL IMPACT

While our work advances multimodal emotion understanding from surface-level affective cue recognition toward cognition-grounded appraisal reasoning, several directions remain open for future exploration:

- First, CogEmo-40K adopts six complementary appraisal dimensions inspired by established cognitive appraisal theories. These dimensions capture important aspects of emotion formation but are not intended to exhaust the full space of cognitive emotion processes. Future work may explore richer appraisal dimensions and alternative theoretical formulations.
- Second, CogEmo-40K is constructed from existing multimodal video data and therefore remains limited by its current domain, language, cultural, and emotion-space coverage. Extending cognition-grounded emotion understanding to more diverse multilingual, cross-cultural, and open-world settings is an important future direction.
- Third, CogEmo-Bench relies on a fixed Appraisal Extractor and semantic embedding model for normalizing and comparing free-form reasoning outputs. Although the same evaluation protocol is applied consistently to all models, future work may further investigate evaluator-independent cognition-level metrics.

Overall, cognition-grounded emotion understanding offers a promising direction for building more reliable and generalizable affective systems. However, model-generated appraisal interpretations should be treated as evidence-grounded reasoning rather than definitive assessments of an individual's internal psychological state, especially in high-stakes applications. We hope this work encourages further research toward more robust, flexible, and human-aligned emotional intelligence in MLLMs.

# B Details of CogEmo-40K Dataset

## B.1 Source Dataset: MER2025

The raw data for CogEmo-40K are derived from MER2025 (Lian et al., 2025b), the third edition of the MER challenge, which focuses on integrating large language models with multimodal emotion understanding. MER2025 includes a large-scale collection of multimodal samples without emotion annotations, providing video, audio, facial information, and subtitle text. Following the data screening protocol of AffectGPT (Lian et al., 2025a), we adopt the same MER2025 source samples used to construct MERCaption (Lian et al., 2025a) as the raw data for CogEmo-40K. Importantly, we do not use the automatically generated descriptions provided by MERCaption. Instead, we treat these source samples as unlabeled multimodal inputs and construct new cognition-grounded reasoning annotations through the two-stage perception-to-appraisal pipeline described below. The generated annotations are further refined through multi-round filtering and human-guided alignment (Section B.3), forming the final CogEmo-40K dataset.

## B.2 Two-Stage Data Generation Prompts

We adopt a two-stage prompt pipeline to generate CogEmo-40K: Stage 1 produces structured, emotion-free perceptual descriptions from the multimodal inputs; Stage 2 takes these perceptual descriptions as input and performs cognitive appraisal reasoning to derive cognition-grounded emotion interpretations. Stage 1 uses `Qwen3-Omni-30B-A3B-Instruct` (Xu et al., 2025), while Stage 2 uses `GPT-5.4` (OpenAI, 2026a). The complete prompts for both stages of data generation are provided below.

**Stage 1: Perception Generation**

**[Instruction]**
You are given a multimodal input (video + audio). Your task is to describe the scene, facial behavior, and audio signals in a structured and strictly objective way.
**[Goal]**

1. Provide a brief description of the overall scene.
2. Provide a HIGH-DENSITY, structured description of facial movements.
3. Provide a HIGH-DENSITY, structured description of audio features.

**[Strict Constraints]**

- DO NOT use any emotion words (e.g., happy, sad, angry, fear).
- Only describe observable visual or acoustic information.

**[Scene Description Rules]**

- Describe environment, people, and interactions
- Include visible actions only
- Keep it brief (1–2 sentences)
- Use neutral and factual language

**[Facial Description Rules]**
You MUST describe ALL facial feature categories below. For each category include: (1) state; (2) degree (slight / moderate / strong / stable); (3) dynamics (no change / gradual change / fluctuating).
Required categories:

- Gaze: direction, stability, shifts
- Eyes: openness, eyelid tension, blinking pattern, asymmetry
- Eyebrows: position, movement, contraction or relaxation, symmetry
- Mouth: opening degree, lip tension, corner direction, lip compression
- Jaw: tension level, clenching, movement
- Facial muscles: overall tension, rigidity, relaxation
- Left–Right asymmetry: differences in facial movement or tension
- Head pose: orientation, tilt, movement
- Temporal dynamics: changes over time (start → middle → end)
- Overall facial mobility: level of movement intensity (high / medium / low)

You MUST additionally describe fine-grained facial action components using observable anatomical or geometric descriptors, including but not limited to:

- Lip corners: upturned / downturned / stretched / retracted
- Cheek region: raised / tightened / relaxed / bulging

- Eye aperture: widened / narrowed / squinting / partially closed
- Eyelid behavior: lifted / lowered / tightened / compressed
- Nasolabial region: deepened / relaxed
- Teeth visibility: visible / partially visible / not visible
- Orbital tension: increased / reduced / stable
- Facial symmetry changes during motion: symmetric / asymmetric activation

Use only visible structural descriptions; do NOT label emotions or interpret meaning. These features must be embedded into Mouth / Eyes / Facial muscles sections.

**[Audio Description Rules]**

Describe audio features from the perspective of the ACTIVE SPEAKER (the person who is speaking or producing vocal sound in the video). For each category include: (1) state; (2) degree (slight / moderate / strong); (3) dynamics (no change / gradual change / fluctuating).

Required categories:

- Speech presence: continuous speech / intermittent speech / no speech
- Spoken content (verbatim transcription if audible): exact spoken words as heard, without interpretation, summarization, or paraphrasing
- Voice intensity: soft / moderate / loud / variable
- Speech rate: slow / normal / fast / changing
- Pitch variation: flat / slightly varied / highly varied
- Prosody dynamics: stable / gradual modulation / fluctuating
- Articulation clarity: clear / moderately clear / unclear

**[Writing Requirements]**

- Write in 1 unified paragraph for each modality (Scene / Facial / Audio)
- Preserve temporal order of changes
- Embed all attributes naturally into continuous flow
- Keep descriptions strictly visual or acoustic and measurable

**Stage 2: Cognitive Appraisal (CoT)**

You are an expert in multimodal emotion analysis based on cognitive appraisal theory. Your task is to infer the emotional state from scene, facial behavior, audio, and spoken content.

**Requirements:**

1. You MUST integrate evidence and appraisal inference together. Do NOT separate "evidence" and "appraisal" into different sections. Each appraisal dimension must be inferred together with its supporting evidence in the SAME sentence.
2. You MUST produce ONE single coherent paragraph. No bullet points, no section headers, and no line breaks.
3. For each appraisal dimension, you MUST follow this format: first briefly describe the supporting evidence, then explicitly state the appraisal result in parentheses.
4. You MUST include ALL six appraisal dimensions in the following order: Goal Congruence, Coping, Accountability, Future, Arousal, Action Tendency.
5. The paragraph MUST end with a brief summary of the overall appraisal pattern and a final emotion label selected from: happiness, sadness, anger, fear, anxiety, disgust, contempt, neutral.

### B.3 Human Verification and Quality Filtering

To ensure that CogEmo-40K provides reliable perceptual evidence and human-aligned cognitive appraisal labels, we conduct staged human quality control after automatic generation. Stage 1 audits description-level perception outputs; Stage 2 verifies cognition-level appraisal judgments and emotion labels. All annotators are trained with the standardized rubrics below and inspect each sample against the original video and audio.

**Stage 1: Description-level quality assessment.** This stage evaluates the objective multimodal descriptions generated by `Qwen3-Omni-30B-A3B` (Xu et al., 2025), covering Scene, Facial, and Audio. Each sample is rated on a 1–5 scale along three criteria: **Accuracy** (whether the description matches the video and contains no hallucination), **Completeness** (whether Scene, Facial, and

Audio are covered), and **Objectivity** (whether the description stays within observable evidence without emotion or psychological inference). We randomly sample 1,000 descriptions and ask five researchers with backgrounds in affective computing to score each sample holistically under the rubric in Table 5, with priority given to Accuracy, then Completeness. Fluent wording alone does not justify a high score; annotators must strictly separate observable description from emotion or psychological inference.

Table 5: Stage 1 description-level human quality rubric (1–5).

| Score | Level | Criteria |
|---|---|---|
| 5 | Excellent | Fully accurate; covers Scene, Facial, and Audio; includes fine-grained cues (e.g., gaze, mouth dynamics, prosody); fully objective with no subjective inference; clear and unambiguous. Directly usable as high-quality data. |
| 4 | Good | Mostly accurate with only minor deviations; three modalities largely covered; slightly coarser or fewer details; remains largely objective. Usable with limited information loss. |
| 3 | Fair | Noticeable missing information (e.g., one modality absent); locally inaccurate or vague; occasional mild subjective phrasing. Marginally usable but unstable in quality. |
| 2 | Poor | Multiple inaccuracies; clear modality omissions; subjective inference or non-compliant description. Not recommended for use. |
| 1 | Bad | Severely incorrect or inconsistent with the video; obvious hallucination; lacks usable information. Should be discarded. |

In addition to the overall rubric, annotators apply the mandatory score caps in Table 6. Records are stored as `(id, score, notes)`, where `notes` are required whenever the score is $\leq 3$. When a sample cannot be judged reliably, annotators document the reason in `notes`.

Table 6: Mandatory score caps for Stage 1 description assessment.

| Cap | Triggering conditions |
|---|---|
| $\leq 3$ | Emotion words appear (e.g., angry / happy / nervous); subjective judgments occur (e.g., "looks angry"); any one of Scene / Facial / Audio is missing. |
| $\leq 2$ | Multiple modalities are missing; clear inconsistency with the video; description is severely vague or non-specific. |

Under this protocol, the average scores of Accuracy, Completeness, and Objectivity all exceed 4.0, indicating that Stage 1 descriptions provide reliable observable evidence for subsequent appraisal reasoning.

**Stage 2: Cognition-level human verification.** This stage verifies cognition-level reasoning across the six appraisal dimensions and the final emotion label. Given the substantial annotation cost, we randomly select approximately one-third of the 115,595 automatically generated cognition instances for human verification, without preferentially selecting samples based on their difficulty.

Before annotation, well-trained annotators receive standardized instructions and review the definitions of the six appraisal dimensions in Table 7. For each sample, annotators first inspect the original video and audio, considering scene context, facial behavior, body action, vocal cues, and spoken content. They then verify whether the generated reasoning for each appraisal dimension is supported by the multimodal evidence and consistent with the corresponding dimension definition. The final emotion label is additionally checked for consistency with both the multimodal evidence and the verified appraisal reasoning.

An instance is directly retained when its emotion label agrees with human judgment and at least four of the six appraisal dimensions are judged to be consistent. We adopt four dimensions as a practical majority-based threshold: it requires broad agreement across the appraisal structure while allowing limited disagreement that may arise from the inherent subjectivity and individual variation of cognitive appraisal.

Failed instances are not discarded: human annotations are used to correct emotion labels and appraisal reasoning, regenerate evidence-grounded reasoning explanations, and undergo additional review by five affective computing researchers. This correction-based filtering yields the final CogEmo-40K of approximately 40K high-quality instruction-tuning instances.

Table 7: Definitions of the six cognitive appraisal dimensions used in Stage 2 human verification.

| **Dimension** | **Definition** |
|---|---|
| Goal Congruence | Evaluates how the perceived situation relates to the subject's goals, needs, or expectations, including whether it supports, obstructs, or has little relevance to them. |
| Coping Potential | Reflects the subject's perceived ability to manage, control, adapt to, or respond effectively to the situation. |
| Accountability | Captures how responsibility or causal attribution is assigned, including whether the situation is attributed to the subject, another person, or no clearly identifiable agent. |
| Future Expectancy | Describes how the subject interprets the temporal development and expected consequences of the situation, including anticipation, continuation, or the absence of a salient future-oriented expectation. |
| Arousal | Reflects the perceived level of physiological or behavioral activation associated with the situation, as supported by observable multimodal cues such as vocal intensity, facial dynamics, or body movement. |
| Action Tendency | Captures the behavioral orientation implied by the appraisal, such as engaging with, avoiding, confronting, or otherwise responding to the situation. |

### B.4 Dataset Statistics and Comparison

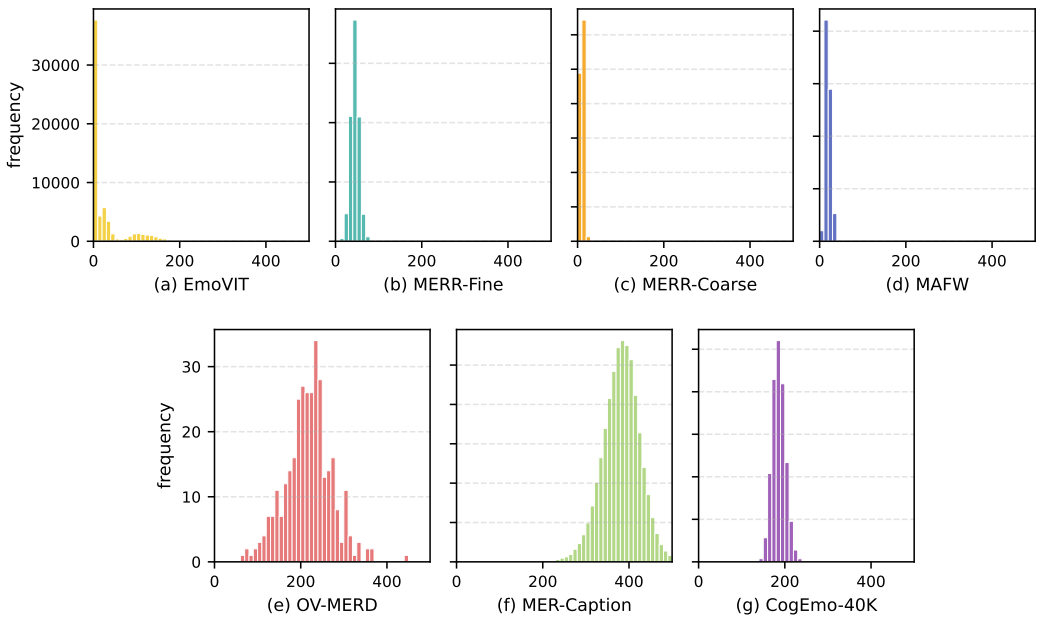



Figure 6: Distribution of description lengths (word count) across descriptive emotion datasets.

#### B.4.1 Text Length Comparison

We compare CogEmo-40K with several representative multimodal emotion datasets, including EmoVIT (Xie et al., 2024), MERR-Fine (Cheng et al., 2024a), MERR-Coarse (Cheng et al., 2024a), MAFW (Liu et al., 2022b), OV-MERD (Lian et al., 2025c), and MERCaption (Lian et al., 2025a), in terms of text length measured by word count. As shown in Figure 6, CogEmo-40K contains substantially longer reasoning annotations than MERR and MAFW, while remaining more concise

than MERCaption. This reflects the different supervision focus of CogEmo-40K, which emphasizes structured cognitive appraisal reasoning rather than purely descriptive emotion-related content.

### B.4.2 Emotion Distribution

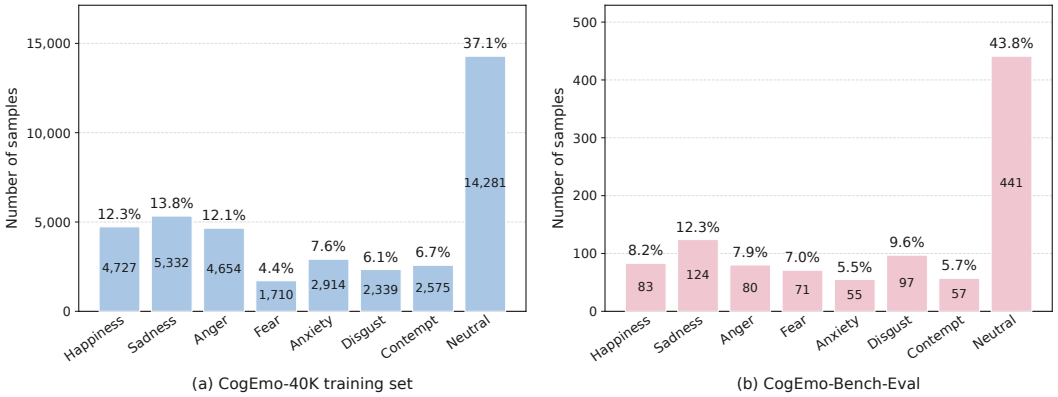



Figure 7: Emotion category distributions of CogEmo-40K and CogEmo-Bench-Eval. (a) CogEmo-40K training set ($N$=38,532). (b) CogEmo-Bench-Eval ($N$=1,008). Bars indicate the number of samples in each emotion category, with percentages annotated above.

As shown in Figure 7, both CogEmo-40K and CogEmo-Bench-Eval exhibit a distribution consistent with naturalistic emotion data, where *neutral* constitutes the dominant category. Beyond this overall trend, CogEmo-Bench-Eval assigns relatively larger proportions to *fear* and *disgust*, which are less frequent and often more challenging to distinguish. This design preserves the naturally imbalanced emotion distribution while providing stronger coverage of difficult low-frequency categories, making the benchmark more demanding for fine-grained cognition-grounded emotion reasoning.

### B.4.3 Cognitive Language Profiles

To further characterize the cognition-level supervision in CogEmo-40K, we analyze recurring language patterns across the six cognitive appraisal dimensions for each emotion category. Since the reasoning annotations are expressed in free-form natural language, we first apply the same fixed *Appraisal Extractor* used in CogEmo-Bench to normalize each annotation into the six predefined appraisal dimensions. The extractor only reorganizes explicitly expressed content and does not judge correctness or introduce additional evidence.

We then group the 38,532 reasoning annotations by emotion category and appraisal dimension, and report the three most frequent semantic phrases and their occurrence rates. Percentages indicate the proportion of samples within the corresponding emotion category containing each phrase. Table 8 provides the full phrase-level statistics for all eight emotion categories across the six appraisal dimensions.

These statistics reveal distinct appraisal tendencies across emotions. For example, anger frequently co-occurs with goal conflict and externally directed attribution, whereas fear is more strongly associated with perceived threat, limited coping capacity, and withdrawal-related tendencies. The compact cognitive profiles presented in the main paper are semantic summaries of these recurring phrase-level patterns rather than direct copies of the raw phrases reported here.

Table 8: Full Cognitive Language Profiles for Eight Emotion Categories.

**Part I: Goal Congruence, Coping, and Accountability**

| Emotion | Goal Congruence | Coping | Accountability |
|---|---|---|---|
| **Anger** | situation conflicts (28%) | able to (51%) | another person (31%) |
| | situation involves (6%) | clear articulation (43%) | the other (24%) |
| | unfavorable evaluation (3%) | feels able (42%) | source of (22%) |
| **Anxiety** | situation conflicts (13%) | clear articulation (33%) | source of (45%) |
| | situation involves (13%) | to manage (24%) | the source (43%) |
| | unfavorable evaluation (2%) | able to (20%) | attributed to (25%) |
| **Contempt** | situation conflicts (31%) | clear articulation (53%) | source of (33%) |
| | unfavorable evaluation (13%) | able to (51%) | the source (33%) |
| | situation involves (10%) | feels able (38%) | another person (31%) |
| **Disgust** | situation conflicts (31%) | clear articulation (34%) | the source (53%) |
| | situation involves (8%) | to manage (28%) | source of (53%) |
| | unfavorable evaluation (6%) | able to (25%) | another person (30%) |
| **Fear** | situation conflicts (9%) | perceived ability (24%) | source of (61%) |
| | situation involves (2%) | to manage (21%) | the source (58%) |
| | | low perceived (17%) | attributed to (30%) |
| **Happiness** | aligns with (47%) | able to (50%) | the cause (27%) |
| | situation supports (35%) | feels able (41%) | another person (15%) |
| | situation aligns (31%) | clear articulation (37%) | the other (15%) |
| **Neutral** | aligns with (19%) | clear articulation (58%) | source of (20%) |
| | situation aligns (14%) | able to (55%) | the source (19%) |
| | personal goals (7%) | to manage (41%) | another person (16%) |
| **Sadness** | situation conflicts (26%) | perceived ability (20%) | the cause (31%) |
| | situation involves (10%) | relaxed jaw (18%) | the source (22%) |
| | unfavorable evaluation (4%) | clear articulation (16%) | source of (21%) |

**Part II: Future, Arousal, and Action Tendency**

| Emotion | Future | Arousal | Action Tendency |
|---|---|---|---|
| **Anger** | the present (21%) | physiological activation (32%) | inclination to (59%) |
| | a present (17%) | facial tension (31%) | behavioral inclination (48%) |
| | distant possibility (13%) | point to (28%) | a behavioral (47%) |
| **Anxiety** | concern about (19%) | activation rather (34%) | inclination to (53%) |
| | points to (18%) | facial mobility (33%) | behavioral inclination (34%) |
| | the present (15%) | low facial (32%) | a behavioral (33%) |
| **Contempt** | the present (23%) | facial mobility (58%) | inclination to (48%) |
| | a present (21%) | low facial (55%) | behavioral inclination (27%) |
| | future threat (13%) | activation rather (54%) | a behavioral (27%) |
| **Disgust** | the present (25%) | point to (39%) | inclination to (56%) |
| | tied to (18%) | facial mobility (36%) | behavioral inclination (48%) |
| | a present (17%) | low facial (31%) | a behavioral (47%) |
| **Fear** | concern about (25%) | physiological activation (31%) | inclination to (40%) |
| | what may (21%) | point to (17%) | behavioral inclination (36%) |
| | about what (21%) | facial tension (16%) | a behavioral (33%) |
| **Happiness** | the present (28%) | facial mobility (57%) | inclination to (53%) |
| | points to (17%) | point to (41%) | inclination toward (43%) |
| | a present (12%) | low facial (41%) | behavioral inclination (39%) |
| **Neutral** | points to (21%) | low facial (69%) | inclination to (30%) |
| | the present (20%) | facial mobility (69%) | orientation toward (25%) |
| | a present (13%) | point to (46%) | toward the (24%) |
| **Sadness** | the present (21%) | activation rather (37%) | inclination to (61%) |
| | future threat (16%) | point to (35%) | behavioral inclination (47%) |
| | a present (14%) | facial mobility (29%) | a behavioral (46%) |

# C IMPLEMENTATION DETAILS

## C.1 DETAILS OF COGEMO-MOE MODEL

CogEmo-MoE extends the MoE-LLaVA (Lin et al., 2026) framework to support joint video, audio, and text modeling, while maintaining a compact model scale of approximately 3.1B parameters. It adopts a Qwen-1.8B (Bai et al., 2023) language model with CLIP ViT-L/14 (Radford et al., 2021) and CLAP (Elizalde et al., 2023) as the visual and audio encoders, respectively, and uses two-layer GELU MLP projectors for modality alignment. The model is trained progressively for multimodal alignment, perception enhancement, and cognitive appraisal specialization using DeepSpeed ZeRO-2 with bfloat16, and a maximum sequence length of 2048 tokens. All experiments are conducted on two NVIDIA A100 80GB GPUs. Detailed configurations are summarized in Table 9.

Table 9: Training configuration of CogEmo-MoE across three stages.

| **Configuration** | **Stage 1** | **Stage 2** | **Stage 3** |
|---|---|---|---|
| Trainable modules | Projectors | Full model | MoE FFN + router |
| MoE enabled | – | – | Sparse (4 experts, top-2) |
| Capacity factor | – | – | 1.5 |
| Router aux. loss $\lambda$ | – | – | 0.01 |
| Epochs | 4 | 1 | 6 |
| Learning rate | $1 \times 10^{-3}$ | $2 \times 10^{-5}$ | $1 \times 10^{-5}$ |
| Warmup ratio | 0.03 | 0.03 | 0.05 |
| LR scheduler | cosine | cosine | cosine |
| Batch size (per device) | 8 | 8 | 4 |
| Gradient accumulation | 1 | 2 | 1 |
| Max sequence length | 2048 | 2048 | 2048 |
| Precision | bf16 | bf16 | bf16 |
| DeepSpeed strategy | DeepSpeed ZeRO-2 | DeepSpeed ZeRO-2 | DeepSpeed ZeRO-2 |

### C.1.1 TRAINING PROMPTS FOR EACH STAGE

Each training stage uses a task-specific user instruction and multimodal inputs consisting of video frames and audio. The placeholders `<frames>` and `<audio>` indicate the positions where visual and acoustic features encoded by CLIP (Radford et al., 2021) and CLAP (Elizalde et al., 2023) are inserted, respectively. The supervision targets differ across stages: Stage 1 uses coarse multimodal descriptions from MERR-Coarse (Cheng et al., 2024a), Stage 2 uses perception-level descriptions from CogEmo-Desc, and Stage 3 uses cognition-level appraisal reasoning from CogEmo-40K.

**Stage 1: Multimodal Alignment.** The user instruction elicits a concise, free-form description of how the subject looks and sounds, providing a lightweight alignment signal for the multimodal projectors.

**Stage 1: Multimodal Alignment (user instruction)**

Here is the video: `<frames>`
Here is the audio: `<audio>`

Describe how the person looks and sounds in the video.

**Stage 2: Perception enhancement.** The instruction shifts toward structured, high-density perceptual description of the scene, facial behavior, and acoustic characteristics. The target response follows the emotion-free perception format produced by the dataset Stage 1 generation prompt.

**Stage 2: Perception Enhancement (user instruction)**

Here is the video: `<frames>`
Here is the audio: `<audio>`

Please briefly describe the video scene, and provide a detailed description of the main subject's facial details and the speaker's acoustic characteristics.

**Stage 3: Cognitive appraisal tuning.** The instruction asks the model to perform cognitive appraisal reasoning over all available modalities and output a final discrete emotion label. The target response follows the appraisal chain-of-thought format produced by the dataset Stage 2 generation prompt.

**Stage 3: Cognitive Appraisal Tuning (user instruction)**

Here is the video: `<frames>`
Here is the audio: `<audio>`

Please infer the most likely emotion category from multimodal cues by reasoning through six cognitive appraisal dimensions: Congruence, Coping, Accountability, Future, Arousal, and Action Tendency. The final emotion should be selected from happiness, sadness, anger, fear, anxiety, disgust, contempt, and neutral.

### C.2 Model Scale and Performance Comparison

Table 10 compares model scale and performance on CogEmo-Bench. While most multimodal emotion LLMs use backbones of at least 7B parameters, the 3.1B CogEmo-MoE achieves the best AEQS, accuracy, and Macro-F1, outperforming the strongest baselines by 14.30, 8.33, and 4.98 points, respectively. This demonstrates that its performance gains arise from effective cognitive modeling rather than increased model scale.

Table 10: Performance–Scale Comparison on CogEmo-Bench.

| Model | Parameters | AEQS | Acc. | Macro-F1 |
|---|---|---|---|---|
| *Open-source MLLMs* | | | | |
| OneLLM (Han et al., 2024) | 7B | 16.15 | 30.16 | 26.03 |
| SECap (Xu et al., 2024) | 7B | 12.26 | 30.85 | 21.25 |
| Qwen2-Audio (Chu et al., 2024) | 7B | 25.12 | 43.55 | 32.68 |
| SALMONN (Tang et al., 2024) | 7B | 23.87 | 43.69 | 33.43 |
| Otter (Li et al., 2025a) | 7B | 10.32 | 36.61 | 24.69 |
| LLaMA-VID (Li et al., 2024) | 7B | 26.60 | 36.71 | 31.84 |
| VideoChat-Flash (Li et al., 2025b) | 7B | 2.69 | 55.36 | 39.11 |
| Chat-UniVi (Jin et al., 2024) | 7B | 32.28 | 34.03 | 29.86 |
| mPLUG-Owl3 (Ye et al., 2025) | 7B | 32.07 | 45.54 | 34.72 |
| Qwen3.8-Max (Team, 2026) | 2.4T | 49.50 | 57.04 | 46.29 |
| Kimi-K3 (Kimi Team, 2026) | 2.8T (104B active) | 48.70 | 46.13 | 39.04 |
| PandaGPT (Su et al., 2023) | 7B | 25.54 | 25.79 | 18.78 |
| Emotion-LLaMA (Cheng et al., 2024a) | 7B | 20.27 | 43.75 | 25.90 |
| AffectGPT (Lian et al., 2025a) | 7B | 32.83 | 42.76 | 31.90 |
| Qwen3-Omni (Xu et al., 2025) | 30B (3B active) | 46.12 | 63.10 | 52.19 |
| *Closed-source MLLMs* | | | | |
| GPT-5.2 (OpenAI, 2025) | —— | 56.30 | 46.33 | 36.92 |
| GPT-5.6-Luna (OpenAI, 2026b) | —— | 48.27 | 51.59 | 43.46 |
| Grok-4.3 (xAI, 2026) | —— | 45.55 | 46.83 | 39.79 |
| Gemini-3.6-Flash (Google, 2026) | -. | 36.30 | 59.13 | 45.99 |
| *Our Model* | | | | |
| **CogEmo-MoE** | **3.1B** | **70.60** | **71.43** | **57.17** |

# D DETAILS OF COGEMO-BENCH

CogEmo-Bench is a held-out evaluation set providing over 6,000 human-verified dimension-level appraisal references across six cognitive appraisal dimensions, under the same eight-class emotion space as CogEmo-40K. It is used to assess both cognitive appraisal evidence quality and final emotion recognition under a unified protocol. Its label distribution is shown beside the training set in Figure 7. Consistent with that training panel, neutral remains dominant (43.8%), while the remaining seven categories range from 5.5% (anxiety) to 12.3% (sadness). Relative to CogEmo-40K, disgust and fear occupy a larger share of the evaluation split, which helps stress-test appraisal reasoning under less frequent negative emotions. The remainder of this section details the evaluated models, prompting templates, appraisal extractor, and metric computation.

## D.1 MODEL SETUP

We evaluate all models reported in Table 2 on the CogEmo-Bench test set. Each model is loaded from its official release and evaluated with the decoding configuration recommended by the original implementation. Table 11 summarizes the model category, supported modalities, and the corresponding evaluation-prompt template. For open-source models, we follow the default checkpoints and preprocessing pipelines released by the original authors. For closed-source models, we access the latest available API endpoints at the time of evaluation. CogEmo-MoE is evaluated under three modality settings (A+T, V+T, and A+V+T) to match the ablations in the main benchmark table. General-purpose MLLMs (Gen.) are prompted for cognitive appraisal reasoning with modality-specific templates; emotion-specialized MLLMs (Spec.) follow their original prompting protocols.

Table 11: **Evaluation setup for models on CogEmo-Bench.** Gen. and Spec. denote general-purpose MLLMs and emotion-specialized models, respectively. A, V, and T denote audio, visual, and textual modalities. Evaluation prompt templates are provided in Appendix D.2.

| Model | Type | A | V | T | Prompt Template |
|---|---|---|---|---|---|
| *Open-source MLLMs* | | | | | |
| OneLLM (Han et al., 2024) | Gen. | ✓ | × | ✓ | AT (Appendix D.2) |
| SECap (Xu et al., 2024) | Spec. | ✓ | × | ✓ | SECap (Appendix D.2) |
| Qwen2-Audio (Chu et al., 2024) | Gen. | ✓ | × | ✓ | AT (Appendix D.2) |
| SALMONN (Tang et al., 2024) | Gen. | ✓ | × | ✓ | AT (Appendix D.2) |
| Otter (Li et al., 2025a) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| LLaMA-VID (Li et al., 2024) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| VideoChat-Flash (Li et al., 2025b) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| Chat-UniVi (Jin et al., 2024) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| mPLUG-Owl3 (Ye et al., 2025) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| Qwen3.8-Max (Team, 2026) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| Kimi-K3 (Kimi Team, 2026) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| PandaGPT (Su et al., 2023) | Gen. | ✓ | ✓ | ✓ | AVT (Appendix D.2) |
| Emotion-LLaMA (Cheng et al., 2024a) | Spec. | ✓ | ✓ | ✓ | Emotion-LLaMA (Appendix D.2) |
| AffectGPT (Lian et al., 2025a) | Spec. | ✓ | ✓ | ✓ | AffectGPT (Appendix D.2) |
| Qwen3-Omni (Xu et al., 2025) | Gen. | ✓ | ✓ | ✓ | AVT (Appendix D.2) |
| *Closed-source MLLMs* | | | | | |
| GPT-5.2 (OpenAI, 2025) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| GPT-5.6-Luna (OpenAI, 2026b) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| Grok-4.3 (xAI, 2026) | Gen. | × | ✓ | ✓ | VT (Appendix D.2) |
| Gemini-3.6-Flash (Google, 2026) | Gen. | ✓ | ✓ | ✓ | AVT (Appendix D.2) |
| *Our CogEmo Models* | | | | | |
| CogEmo-MoE (A+T) | Spec. | ✓ | × | ✓ | AT (Appendix D.2) |
| CogEmo-MoE (V+T) | Spec. | × | ✓ | ✓ | VT (Appendix D.2) |
| CogEmo-MoE (A+V+T) | Spec. | ✓ | ✓ | ✓ | AVT (Appendix D.2) |

### D.2 Evaluation Prompt of Different Models

All models are evaluated on the same CogEmo-Bench instances using the input modalities supported by each model. Specifically, video frames, audio waveforms, and subtitle transcripts are provided whenever the corresponding modality is supported, while unsupported modalities are omitted. General-purpose MLLMs are prompted to perform cognitive appraisal reasoning and produce a final emotion prediction, whereas emotion-specialized models follow their original prompting protocols to preserve their intended inference behavior.
Accordingly, we use two types of evaluation prompts: model-specific prompts for emotion-specialized models and modality-specific appraisal prompts for general-purpose MLLMs, as detailed below.

**Emotion-specialized models.**

**AffectGPT (user instruction)**

Please infer the person's emotional state and provide your reasoning process.

**Emotion-LLaMA (user instruction)**

[reason] Please analyze all clues and reason out the emotional label.

**SECap (user instruction)**

Please describe the speaker's emotional state in one sentence.

SECap follows its original protocol of using a concise emotion-description prompt. The original implementation provides multiple Chinese prompt paraphrases with equivalent meaning; we use the single English sentence above for consistency across models.

**General-purpose MLLMs.** Depending on the supported modalities (Table 11), we use one of the following three prompt templates.

**A+V+T prompt (AVT)**

Given the multimodal input, analyze the person's emotional state and explain why the person may feel this way. Ground your reasoning in the available visual, audio, textual, and contextual evidence. Finally, identify the most likely emotion from: happiness, sadness, anger, fear, anxiety, disgust, contempt, neutral.

**A+T prompt (AT)**

Given the multimodal input, analyze the person's emotional state and explain why the person may feel this way. Ground your reasoning in the available audio, textual, and contextual evidence. Finally, identify the most likely emotion from: happiness, sadness, anger, fear, anxiety, disgust, contempt, neutral.

**V+T prompt (VT)**

Given the multimodal input, analyze the person's emotional state and explain why the person may feel this way. Ground your reasoning in the available visual, textual, and contextual evidence. Finally, identify the most likely emotion from: happiness, sadness, anger, fear, anxiety, disgust, contempt, neutral.

**Appraisal Extractor.** After each model generates a free-form response, we apply a fixed extraction prompt implemented with `Qwen3.5-9B` (Qwen Team, 2026) to obtain six appraisal-specific evidence units and the final emotion label. The extractor only reorganizes information explicitly present in the original response; it does not add new evidence or modify the model's reasoning. If a

dimension is absent in the response, its extracted evidence is left empty. All evaluated models share the same extractor and prompt to ensure a consistent evaluation protocol.

**Appraisal Extractor (user instruction)**

You are an appraisal extractor. Given a model-generated emotion reasoning response, extract the supporting evidence and appraisal judgment for each of the following dimensions if explicitly present: Goal Congruence, Coping, Accountability, Future, Arousal, and Action Tendency. Also extract the final emotion label. Do not add new evidence or modify the original reasoning. If a dimension is missing, leave it empty.

### D.3 Metric Computation Details

CogEmo-Bench jointly evaluates cognitive appraisal reasoning and final emotion recognition. The former is measured by the **Appraisal Evidence Quality Score (AEQS)**, while the latter is evaluated using **Accuracy** and **Macro-F1**. The formal definitions are introduced in Section 4.2; here we provide the corresponding implementation details.

**Evaluation set.** CogEmo-Bench-Eval is a human-verified multimodal evaluation set under the same eight-class emotion space as CogEmo-40K: {*happiness*, *sadness*, *anger*, *fear*, *anxiety*, *disgust*, *contempt*, *neutral*}. Each instance provides human-verified cognitive appraisal references across six dimensions together with a final emotion label.

**Response normalization.** For each instance, the evaluated model first generates a free-form response using the corresponding evaluation prompt described in Appendix D.2. The response is then processed by the same fixed *Appraisal Extractor* (Appendix D.2), which reorganizes the explicitly expressed content into six appraisal dimensions and a final emotion prediction. The extractor does not judge correctness or introduce additional evidence. Appraisal information not expressed in the original response is retained as empty.

**Appraisal Evidence Quality Score.** For each sample $i$ and appraisal dimension $d$, we compare the extracted model evidence $\hat{e}_{i,d}$ with the corresponding human-verified reference $e_{i,d}$. Both are encoded using the same fixed `Qwen3-Embedding` model, and their semantic similarity is computed by cosine similarity as defined in Eq. equation 5.
When only one side contains appraisal evidence, the similarity score is set to zero, penalizing either missing or unsupported evidence. When both sides are empty, the sample is excluded from the corresponding dimension because no evidence is available for semantic comparison. We first average the valid sample-level similarities within each appraisal dimension and then average the six dimension-level scores to obtain AEQS according to Eq. equation 7. For readability, the final AEQS is multiplied by 100 when reported.
**Emotion Recognition.** For final emotion prediction, we report **Accuracy** and **Macro-F1**. Accuracy measures the proportion of correctly predicted emotion labels, while Macro-F1 equally weights the class-wise F1 scores across all emotion categories:

$$\text{Acc.} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \qquad \text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2P_cR_c}{P_c + R_c}. \tag{8}$$

where $\hat{y}_i$ and $y_i$ denote the predicted and ground-truth emotion labels for sample $i$, respectively; $P_c$ and $R_c$ denote the precision and recall for class $c$; $N$ is the number of evaluation samples; and $C = 8$ is the number of emotion categories.

**Unified evaluation.** All evaluated models share the same Appraisal Extractor, embedding model, emotion label space, and metric computation procedure. This unified pipeline minimizes evaluation differences introduced by heterogeneous response formats and enables consistent comparison of both appraisal reasoning and emotion recognition across models.

# E ADDITIONAL EXPERIMENTAL RESULTS

## E.1 COGNITIVE REASONING–EMOTION ALIGNMENT

To examine whether the cognitive appraisal reasoning in CogEmo-40K is aligned with the corresponding final emotion, we conduct a 5-fold cross-validation experiment on the **intersection subset of CogEmo-40K and MERCaption+** (11,770 samples) (Lian et al., 2025a). Specifically, we use the cognitive reasoning annotations as representations and formulate emotion prediction as an 8-class classification task. This experiment is intended as a diagnostic analysis of the relationship between intermediate cognitive reasoning and the resulting emotion label, rather than an evaluation of a standalone text classifier. As shown in Table 12, CogEmo-40K reasoning representations exhibit strong emotion separability, suggesting that the generated appraisal chains encode emotion-relevant cognitive information that is closely aligned with the final emotion labels.

Table 12: 5-fold cross-validation results on the text-reason branch of the **CogEmo-40K & MER-Caption+ (Lian et al., 2025a) intersection subset**. Accuracy, Weighted-F1, and Macro-F1 are reported in %. The best average is shown in bold.

| Metric | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Avg±Std |
|---|---|---|---|---|---|---|
| Accuracy | 92.44 | 92.57 | 91.59 | 92.44 | 92.40 | **92.29±0.35** |
| Weighted-F1 | 92.43 | 92.57 | 91.50 | 92.48 | 92.31 | **92.26±0.39** |
| Macro-F1 | 91.44 | 91.99 | 90.27 | 91.72 | 91.23 | **91.33±0.59** |

**Training Details.** The 5-fold setup follows a fixed, pre-generated split (seed 42, 80%/20% per fold) on the intersection subset described above, ensuring full reproducibility. We list the key training configurations below:

1. **Backbone:** XLM-RoBERTa-base.
2. **Optimizer:** AdamW (HuggingFace Trainer default).
3. **Initial learning rate:** $2 \times 10^{-5}$ with a linear decay schedule and no warmup.
4. **Weight decay:** 0.01.
5. **Batch size:** 32 per device.
6. **Epochs:** 15.
7. **Maximum sequence length:** 512.
8. **Precision:** Full precision (FP32); no gradient accumulation.
9. **Hardware:** Single NVIDIA L40 (48GB).
10. **Data preprocessing:** A rule-based label-leakage stripping is applied to remove final explicit emotion-conclusion sentences (e.g., "therefore the emotion is anger"), forcing the model to rely on the actual reasoning content.

In addition to the classification results, we perform a feature-space analysis by extracting mean-pooled XLM-RoBERTa embeddings of the reasoning text, applying PCA, and computing the Silhouette score on 5,000 sampled instances. The resulting cluster separability further supports the quality of the cognitive reasoning annotations.

## E.2 CROSS-DOMAIN GENERALIZATION EXPERIMENTS

To evaluate the generalization ability of cognitive appraisal-guided training, we conduct a controlled comparison between our automatically generated cognition-grounded supervision and existing emotion description supervision. Specifically, we use the same 115,595 source samples as MERCaption (Lian et al., 2025a) and generate cognitive appraisal annotations under our perception-to-appraisal paradigm without human refinement. In contrast, MERCaption provides automatically generated free-form affective descriptions on the same source set. This setup controls for both data source and data scale, allowing us to more directly examine the effect of the supervision paradigm. All models are trained under the same AffectGPT (Lian et al., 2025a) framework and evaluated following the MERBench protocol (Lian et al., 2025a), covering nine emotion datasets. The corresponding zero-shot results are reported in Table 13 for Part I (MER2023–CMU-MOSI) and Table 14

for Part II (CMU-MOSEI–OV-MERD+), covering both General Instruction and Emotion Description tasks. A detailed description of MERBench is provided in Appendix E.3.

Table 13: Cross-domain generalization results on **General Instruction** and **Emotion Description** datasets (Part I: MER2023–CMU-MOSI). All models are trained on AffectGPT and evaluated on MERBench. Top: General Instruction models. Bottom: Emotion Description models. CogEmo-40K is highlighted in blue.

| Dataset | Filtering | MER2023 | MER2024 | MELD | IEMOCAP | CMU-MOSI |
|---|---|---|---|---|---|---|
| MiniGPT4 | × | 11.56 | 12.91 | 18.89 | 16.06 | 53.57 |
| MiniGPT4 | ✓ | 17.57 | 16.65 | 22.60 | 30.18 | 52.58 |
| VideoChat | × | 24.87 | 22.42 | 21.56 | 32.91 | 50.13 |
| VideoChat | ✓ | 27.70 | 24.73 | 27.66 | 39.46 | 45.45 |
| LLaVA | × | 42.21 | 41.54 | 32.97 | 49.96 | 54.48 |
| LLaVA | ✓ | 41.56 | 42.30 | 32.61 | 46.21 | 52.82 |
| WavCaps | × | 5.75 | 7.71 | 4.35 | 4.99 | 45.59 |
| WavCaps | ✓ | 23.72 | 26.97 | 23.39 | 27.30 | 54.67 |
| EmoVIT | – | 39.31 | 50.24 | 32.36 | 48.24 | 53.40 |
| MAFW | – | 52.67 | 55.99 | 40.85 | 57.60 | 66.11 |
| MERR-Coarse | – | 35.34 | 36.60 | 29.37 | 36.94 | 65.10 |
| MERR-Fine | – | 69.00 | 72.84 | 47.38 | 54.49 | 66.21 |
| MERCaption | – | 72.12 | **74.21** | 54.69 | 56.41 | 75.10 |
| **CogEmo-40K (Ours)** | – | **74.01** | 72.66 | **57.35** | **57.53** | **79.15** |

Table 14: Cross-domain generalization results on **General Instruction** and **Emotion Description** datasets (Part II: CMU-MOSEI–OV-MERD+). All models are trained on AffectGPT and evaluated on MERBench. Top: General Instruction models. Bottom: Emotion Description models. CogEmo-40K is highlighted in blue. Mean is computed across all nine datasets.

| Dataset | Filtering | CMU-MOSEI | CH-SIMS | CH-SIMS v2 | OV-MERD+ | Mean |
|---|---|---|---|---|---|---|
| MiniGPT4 | × | 45.98 | 57.66 | 55.16 | 13.86 | 31.74 |
| MiniGPT4 | ✓ | 56.50 | 52.36 | 51.19 | 20.16 | 35.53 |
| VideoChat | × | 56.17 | 50.07 | 51.71 | 24.56 | 37.16 |
| VideoChat | ✓ | 56.86 | 43.68 | 47.05 | 26.09 | 37.63 |
| LLaVA | × | 56.42 | 52.04 | 54.80 | 35.75 | 46.69 |
| LLaVA | ✓ | 57.72 | 52.78 | 53.44 | 36.96 | 46.27 |
| WavCaps | × | 22.76 | 53.04 | 45.68 | 4.95 | 21.65 |
| WavCaps | ✓ | 49.54 | 58.12 | 55.93 | 21.57 | 37.91 |
| EmoVIT | – | 61.53 | 69.72 | 66.53 | 38.09 | 51.05 |
| MAFW | – | 62.27 | 75.20 | 70.02 | 42.75 | 58.16 |
| MERR-Coarse | – | 63.27 | 75.12 | 73.76 | 33.18 | 49.85 |
| MERR-Fine | – | 60.03 | 79.90 | 78.54 | 52.56 | 64.55 |
| MERCaption | – | **70.97** | 80.21 | **78.06** | **58.42** | 68.91 |
| **CogEmo-40K (Ours)** | – | 69.75 | **83.01** | 77.94 | 58.16 | **69.95** |

**Discussion.** Compared with the strongest emotion description baseline, **MERCaption** (Lian et al., 2025a), **CogEmo-40K** improves performance on most benchmarks, including MER2023 (Lian et al., 2023), MER2024 (Lian et al., 2024), MELD (Poria et al., 2019), IEMOCAP (Busso et al., 2008), CMU-MOSI (Zadeh et al., 2017), CMU-MOSEI (Bagher Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), CH-SIMS v2 (Liu et al., 2022a)and OV-MERD+ (Lian et al., 2025c). Notably, CogEmo-40K improves MER2023 from **72.12** → **74.01** (**+1.89** points) and MELD from **54.69** → **57.35** (**+2.66** points), demonstrating that structured cognitive appraisal reasoning provides more transferable supervision than free-form affective descriptions. On the challenging CMU-MOSEI dataset, CogEmo-40K reaches **69.75** versus **70.97** for MERCaption, while on CH-SIMS v2 it trails by only **0.12** points. Overall, these results confirm that explicitly modeling the perception-to-appraisal pathway yields stronger cross-domain generalization.

### E.3 MERBench Evaluation Protocol

We adopt MERBench (Lian et al., 2025a) to evaluate the cross-domain generalization of different supervision paradigms. As summarized in Table 15, MERBench provides a unified evaluation protocol covering nine representative multimodal emotion datasets and three complementary affective tasks: fine-grained emotion recognition, basic emotion recognition, and sentiment analysis.

Table 15: **Statistics of the nine datasets in MERBench.** MERBench covers fine-grained emotion recognition, basic emotion recognition, and sentiment analysis.

| Task | Dataset | # Samples | # Classes |
|---|---|---|---|
| Fine-grained | OV-MERD+ | 532 | Unified |
| Basic | MER2023 | 411 | 6 |
| | MER2024 | 1,169 | 6 |
| | IEMOCAP | 1,241 | 4 |
| | MELD | 2,610 | 7 |
| Sentiment | CMU-MOSI | 686 | 3 |
| | CMU-MOSEI | 4,659 | 3 |
| | CH-SIMS | 457 | 3 |
| | CH-SIMS v2 | 1,034 | 3 |

**Fine-grained emotion recognition.** This setting evaluates open-vocabulary emotion understanding using **OV-MERD+** (Lian et al., 2025c), where emotion predictions are not restricted to a fixed closed-set taxonomy. Compared with conventional categorical emotion datasets, OV-MERD+ contains more diverse and fine-grained affective expressions, making it suitable for evaluating whether learned representations can generalize to open-ended emotion semantics.

**Basic emotion recognition.** This setting includes **MER2023** (Lian et al., 2023), **MER2024** (Lian et al., 2024), **IEMOCAP** (Busso et al., 2008), and **MELD** (Poria et al., 2019). MER2023 and MER2024 are large-scale Chinese multimodal emotion datasets collected primarily from movies and television programs. IEMOCAP contains acted dyadic interactions and is evaluated under the commonly used four-class setting, while MELD consists of multi-party conversational clips from television dialogues and contains seven basic emotion categories.

**Sentiment analysis.** This setting includes **CMU-MOSI** (Zadeh et al., 2017), **CMU-MOSEI** (Bagher Zadeh et al., 2018), **CH-SIMS** (Yu et al., 2020), and **CH-SIMS v2** (Liu et al., 2022a). CMU-MOSI (Zadeh et al., 2017) and CMU-MOSEI contain multimodal opinion videos with sentiment annotations, with CMU-MOSEI providing a substantially larger-scale extension. CH-SIMS and CH-SIMS v2 consist of Chinese multimodal sentiment samples and provide complementary evaluation in a different language and cultural context.

Following the original MERBench protocol, task-specific metrics are used for the three settings. Fine-grained emotion recognition adopts a set-based evaluation after normalizing semantically equivalent emotion expressions; basic emotion recognition evaluates whether the predicted emotion matches the ground-truth basic category; and sentiment analysis reports class-imbalance-aware classification performance after mapping continuous sentiment annotations into discrete polarity categories. Although MERBench is built on conventional emotion recognition and sentiment datasets, the evaluated models produce free-form emotion analyses that are mapped to benchmark label spaces for scoring. Our perception-to-appraisal supervision further structures these analyses through cognitive dimensions, allowing us to test whether cognition-grounded emotion understanding transfers to standard downstream benchmarks.

All compared models in our cross-domain experiments follow the same MERBench preprocessing, prompting, and metric computation protocol to ensure consistent evaluation.

### E.4 Case Study: Qualitative Comparison of Cognitive Appraisal Reasoning

To complement the quantitative results in Table 2, we present a qualitative case study comparing the reasoning trajectories of our CogEmo-MoE with the best-performing baseline (GPT-5.2). We selected three samples from CogEmo-Bench that highlight typical failure modes of existing MLLMs: (1) emotion confusion between contempt and anger, (2) surface-level textual misleading, and (3)

over-interpretation of neutral contexts. For each sample, we show the ground-truth cognitive appraisal reasoning, the output of our model, and the output of GPT-5.2.

#### E.4.1 CASE 1: DISTINGUISHING CONTEMPT FROM ANGER



(a) Frame 1 (b) Frame 2 (c) Frame 3

Figure 8: Selected video frames for Case 1 (Contempt vs. Anger). The frames show the controlled downward gaze, slightly downturned mouth corners, and stable head pose characteristic of contempt, as opposed to high-arousal anger cues.

**Ground Truth Emotion:** Contempt
**Context:** A man is criticized as "the weakest link" in a military confrontation.

**Ground-Truth Cognitive Appraisal:**

> "The dismissive verbal content calling the others merely 'three hundred defeated soldiers,' together with the downturned lip corners and steady off-camera address, suggests a devaluing, unfavorable evaluation of the target; meanwhile, his clear articulation, moderate voice intensity, stable gaze, and relaxed jaw imply he feels able to handle the situation and speak from a position of control; furthermore, because the criticism is directed outward at the off-camera group through the insulting label 'defeated soldiers,' the cause of concern is assigned to others rather than himself. Overall, the pattern is one of negative evaluation, confident control, blame toward others, present-focused engagement, and restrained activation, therefore the most likely emotion is contempt."

**Our Model (CogEmo-MoE) Output:**

> "The man's steady slight downward gaze, mildly downturned mouth corners, and the critical verbal content ('you are the weakest link') convey an unfavorable evaluation of the target; meanwhile, his controlled posture, clear articulation, and stable low-mobility face with moderate vocal intensity suggest he remains composed enough to address the issue directly rather than being overwhelmed; furthermore, because the utterance explicitly points to 'you' as the source of the problem and his facial expression stays directed outward rather than self-reproachful, responsibility is assigned externally. Overall, the pattern is one of controlled negative evaluation with external blame, present-focused criticism, low arousal, and direct engagement, therefore the most likely emotion is contempt."

**Prediction:** Contempt ✓

**Baseline Model (GPT-5.2) Output:**

> "Across the sampled frames, the person appears tense and increasingly confrontational. His brows are drawn and his gaze is fixed off to the side as if reacting to someone nearby; his mouth stays slightly open as though speaking defensively. The final frame shows a sharp escalation: eyes widened, jaw set, and teeth visible in what looks like a raised voice or shouted remark—classic cues of agitation and aggression... Together, the facial tension, sudden wide-eyed glare, and the insulting/critical text point to an angry response rather than sadness or fear."

**Prediction:** Anger × (Ground Truth: Contempt)

**Analysis:** The baseline model interprets the raised voice and confrontational facial cues as anger, focusing on observable affective signals of agitation. In contrast, our model explicitly reasons about *Accountability* ("responsibility is assigned externally") and *Coping* ("remains composed enough to address the issue directly"), which are the key appraisal dimensions distinguishing contempt (characterized by moral superiority and external blame with low arousal) from anger (characterized by high arousal and goal obstruction). This demonstrates that cognitive appraisal reasoning enables our model to differentiate emotionally similar but cognitively distinct negative states.

### E.4.2 CASE 2: OVERCOMING SURFACE-LEVEL TEXTUAL MISLEADING



(a) Frame 1 (b) Frame 2 (c) Frame 3

Figure 9: Selected video frames for Case 2 (Fear vs. Sadness). The frames illustrate the tightened jaw, narrowed eyes, and tense mouth while recounting being comforted, indicating a lingering threat appraisal rather than past-oriented sadness.

**Ground Truth Emotion:** Fear
**Context:** A woman testifies in court, recounting that someone comforted her.

**Ground-Truth Cognitive Appraisal:**

> "Her tightened jaw, narrowed eyes, raised yet unstable brows, and downturned but tense mouth while recounting that she felt okay because another person was there in a formal hearing-like setting suggest an unpleasant and threatening evaluation of the situation; meanwhile, although she continues speaking clearly and remains physically composed, the growing facial tension and reliance in her words on another person's reassurance imply only limited personal control over the threatening situation; furthermore, because her statement frames safety as depending on someone else's presence and comfort rather than her own action, the source of regulation is centered outside herself... Overall, the pattern is of negatively evaluated threat, limited control, externally centered responsibility, anticipation of danger, high activation, and a withdrawal-oriented response, therefore the emotion is fear."

**Our Model (CogEmo-MoE) Output:**

> "The speaker's statement that someone comforted her, together with the otherwise subdued and serious interaction, points to a negatively toned situation rather than a pleasant one; meanwhile, the relaxed face, steady gaze, moderate voice, and lack of forceful movement suggest the person is not actively exerting strong control over the situation; furthermore, the spoken content centers on another person's comforting behavior as the relevant cause, which places causation outside the speaker... Overall, the pattern is of a negative, externally caused, ongoing situation with limited coping, low arousal, and withdrawal, therefore the emotion is fear."

**Prediction:** Fear ✓

**Baseline Model (GPT-5.2) Output:**

> "The woman appears emotionally distressed and vulnerable. Her furrowed brows, raised inner eyebrows, tense eyes, pursed/downturned lips, and lowered gaze suggest worry and lingering sadness rather than happiness or neutrality. The subtitles indicate that Mu Zi comforted her and that she felt things would be okay because Mu Zi was there, implying she had previously been upset or anxious and is now only partially reassured. Her expression remains strained while speaking, suggesting emotional pain and concern beneath the relief... Most likely emotion: sadness."

**Prediction:** Sadness × (Ground Truth: Fear)

**Analysis:** The baseline model appears to rely on the textual cues "comforted" and "felt better" and predicts sadness, interpreting the situation as grief or loss. Our model instead describes an ongoing situation and limited coping, which are more consistent with the fear label in this example. The case illustrates how appraisal-based descriptions can help distinguish interpretations when the textual cues are ambiguous.

### E.4.3 CASE 3: CORRECTLY IDENTIFYING NEUTRAL AFFECT



(a) Frame 1 (b) Frame 2 (c) Frame 3

Figure 10: Selected video frames for Case 3 (Neutral vs. Anxiety). The woman's relaxed facial expression, stable gaze, and slight smile indicate low arousal and adequate coping, supporting a neutral appraisal rather than anxiety.

**Ground Truth Emotion:** Neutral
**Context:** A woman politely apologizes and explains that someone has offended a noble.

**Ground-Truth Cognitive Appraisal:**

> "The woman's polite wording asking for an audience, along with her relaxed face, slight mouth upturn, and steady moderate speech toward the man, suggests the interaction is not strongly adverse and may even be socially formal; her clear articulation, stable head pose, direct gaze, and controlled vocal delivery imply she feels able to manage the situation and speak deliberately; meanwhile, because she explicitly says she is sorry and that Wang Shang has offended Mi Yuan, the source of the issue is framed as lying with external parties rather than herself... Overall, the pattern is mildly positive, controlled, externally attributed, present-focused, low-arousal, and socially engaging, therefore the most likely emotion is neutral."

**Our Model (CogEmo-MoE) Output:**

> "The woman's polite wording asking for an audience, along with her relaxed face, slight mouth upturn, and steady moderate speech toward the man, suggests the interaction is not strongly adverse and may even be socially formal; her clear articulation, stable head pose, direct gaze, and controlled vocal delivery imply she feels able to manage the situation... Overall, the pattern is mildly positive, controlled, externally attributed, present-focused, low-arousal, and socially engaging, therefore the most likely emotion is neutral."

**Prediction:** Neutral ✓

**Baseline Model (GPT-5.2) Output:**

> "The woman appears emotionally distressed and concerned. Her raised, knitted eyebrows, slightly widened eyes, tense mouth, and pleading gaze toward the man suggest worry about the interpersonal conflict. The text indicates that Wang Shang has offended Mi Yuan, and she apologizes ('Sorry to bother you'), implying she may feel anxious about the consequences... Most likely emotion: anxiety."

**Prediction:** Anxiety × (Ground Truth: Neutral)

**Analysis:** The baseline model over-interprets the apologetic and formal social context as emotional distress, projecting anxiety onto a neutral interaction. Our model correctly recognizes that the woman's *Coping* ("able to manage the situation") and *Arousal* (low activation, "relaxed face") do not support an anxiety interpretation. This case illustrates how appraisal reasoning may help the model avoid assigning anxiety to a neutral conversational exchange.

### E.4.4 SUMMARY OF CASE STUDY

Across these three representative cases, we observe consistent patterns in how our model outperforms the baseline:

Table 16: Appraisal dimensions correctly identified by CogEmo-MoE in each case study.

| **Dimension** | **Case 1** (Contempt vs. Anger) | **Case 2** (Fear vs. Sadness) | **Case 3** (Neutral vs. Anxiety) |
|---|---|---|---|
| Goal Congruence | Negative evaluation | Negative threat | Not strongly adverse |
| Coping Potential | High (controlled) | Low (limited) | High (able) |
| Accountability | External blame | External cause | External parties |
| Future Expectancy | Present-focused | Anticipatory threat | Present-focused |
| Arousal | Low (restrained) | High (activation) | Low (relaxed) |
| Action Tendency | Confrontational | Withdrawal | Socially engaging |

These qualitative examples illustrate that cognitive appraisal reasoning provides critical discriminative information that is not accessible through surface-level affective cue matching alone. By explicitly modeling how individuals evaluate situations along dimensions such as accountability, coping potential, and future expectancy, our model achieves more reliable emotion understanding, particularly in cases involving subtle emotional distinctions, misleading textual content, or contextually appropriate neutral affect.

We acknowledge that these three cases are qualitatively selected for illustrative purposes; the statistical superiority of our model is rigorously validated in the main experiments (Table 2).