EMOR1: NEUTRAL-CALIBRATED REINFORCEMENT LEARNING FOR MULTIMODAL EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Emotion Recognition (MER) combines linguistic, acoustic, and visual cues. Reinforcement-learning-based multimodal large language models (MLLMs) improve coarse-grained emotion and polarity prediction, yet two reliability issues remain. Neutral samples may be forced into non-neutral labels (neutral under-prediction), while non-neutral samples may collapse into neutral (neutral over-prediction), distorting the boundary in opposite directions. Modality-aware rewards may also credit mentions of audio or visual cues that do not improve prediction. Aggregate evaluations obscure these failures when they do not isolate neutral performance or test modality contributions. We introduce the Multimodal Affective Reasoning-Integrated dataset (MARI), a 103K-sample corpus with step-wise reasoning chains and cross-model auditing, and EmoR1, a 7B model trained on MARI with NH-GRPO. NH-GRPO combines bidirectional neutral-boundary rewards with paired full-modal and text-only score comparisons. We also introduce NH-MERBench for evaluating exact labels, semantic clusters, and polarity under a strict neutral criterion. On seven public datasets, EmoR1 achieves competitive NH-MERBench performance and supports further study of neutral calibration and modality credit.

1 INTRODUCTION

Emotion-aware human–AI interaction, socially intelligent assistants, and embodied agents require affect inference from language, speech, and visual behavior in context (Huang et al., 2024; Lin et al., 2024b). MER integrates these cues, and recent multimodal large language models (MLLMs) extend it from closed-set classification to instruction following, open-vocabulary prediction, and free-form reasoning (Cheng et al., 2024; Lian et al., 2025b). Reinforcement learning (RL) enables direct optimization of emotion-oriented objectives (Lian et al., 2025c), supporting evidence-grounded judgments while making reward and metric design central to whether models learn the intended behavior.



Figure 1: Motivation for neutral-aware MER. (a) Across leading MLLM-based MER models, non-neutral scores stay competitive while Neutral-F1 drops sharply, revealing a calibration gap masked by aggregate metrics. NonNeutral-HScore is the mean hierarchical compatibility score over ground-truth non-neutral samples (formally defined in Section 3.4). Both scores are reported in percent. (b)–(c) On an MER2024 (Lian et al., 2024a) case, Emotion-LLaMA v2 (Peng et al., 2026) over-reads a factual utterance as anger, whereas EmoR1 preserves the neutral boundary through cross-modal evidence. Red/green mark incorrect/correct evidence and predictions, respectively.

Recent work improves robustness to incomplete observations, emotion-specific modeling, explainable reasoning, retrieval, and large-scale multimodal supervision (Wang et al., 2023; Huang et al., 2025;

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.

Lian et al., 2023b; Wang et al., 2026b). RL further aligns generative emotion prediction with evaluation metrics (Lian et al., 2025c; Wang et al., 2026a; Fang et al., 2026). Yet higher aggregate scores need not imply more reliable decisions. Figure 1 shows competitive non-neutral scores but much lower Neutral-F1 across the compared MLLMs, together with a factual utterance over-interpreted as anger. The score gap diagnoses a neutral-boundary failure without establishing its cause, while the case motivates testing whether unsupported surface cues contribute. MER reliability therefore depends on the final label, the decision boundary, and credit assigned to the supporting evidence. Misalignment can leave brittle shortcuts behind strong aggregate performance.

The first gap concerns *neutral*, a decision boundary between ambiguous and polarized affect that is fragile under imbalance. We call $y = \text{neutral}$ but $\hat{y} \neq \text{neutral}$ ***neutral under-prediction***, which assigns spurious polarity, and $y \neq \text{neutral}$ but $\hat{y} = \text{neutral}$ ***neutral over-prediction***, which erases valid affective distinctions. Most MLLM pipelines use one label-level objective for neutral and other classes instead of separate signals for these opposite errors (Cheng et al., 2024; Huang et al., 2025; Lian et al., 2025a). The second gap is *modality credit*. Emotion-oriented RL optimizes task-level prediction and reasoning (Lian et al., 2025c; Zhao et al., 2025), but rewards for mentioning audio-visual evidence can be exploited when decisions remain text-driven. This blurs multimodal reasoning with stylistic compliance. Existing benchmarks cover diverse settings but rarely isolate neutral-boundary quality, so distorted decisions and inflated modality claims can remain hidden by aggregate performance (Lian et al., 2024b; Hu et al., 2025; Zhang et al., 2025).

We address both gaps with EmoR1 and Neutral-aware Hierarchical Group Relative Policy Optimization (NH-GRPO), an MER-specific formulation that retains standard GRPO while combining bidirectional neutral-boundary rewards with paired full-modal and text-only score differences. MARI provides stepwise reasoning traces grounded in observable evidence, while NH-MERBench evaluates fine-grained labels, semantic clusters, and polarity under strict neutral criteria. Our contributions are:

- We treat neutral as a decision boundary and propose a symmetric calibration mechanism with bidirectional rewards, explicitly correcting boundary distortion.
- We make modality credit depend on a task-performance difference between paired full-modal and text-only scores, rather than on modality words alone.
- We release MARI (103K samples with structured reasoning) and NH-MERBench for neutral-aware, hierarchical evaluation.

2 Related Work

Multimodal Emotion Recognition. MER integrates noisy, incomplete, and weakly aligned language, speech, and visual evidence, motivating cross-modal and missing-modality methods (Wang et al., 2023). MLLMs extend closed-label classification to instruction-following affect understanding and visual instruction tuning (Cheng et al., 2024; Xie et al., 2024; Fang et al., 2025), followed by explainable reasoning, retrieval augmentation, and specialized architectures (Lian et al., 2023b; Huang et al., 2025; Wang et al., 2026b). Open-vocabulary work supports richer descriptions beyond fixed taxonomies (Lian et al., 2025b;a), while benchmarks now cover exact-label, sentiment-level, open-vocabulary, and emotional-intelligence evaluation (Lian et al., 2024b; Hu et al., 2025; Zhang et al., 2025). Despite these advances, most MLLM-based MER systems retain one undifferentiated label objective for neutral and other classes, while benchmarks rarely isolate neutral-boundary quality. Strong aggregate scores can therefore mask unstable calibration.

Reinforcement Learning for Multimodal Emotion Understanding. RL post-training is useful when target behavior depends on reasoning rather than token matching. GRPO samples multiple outputs per prompt and optimizes relative rewards, offering a practical PPO alternative for reasoning-oriented generation (Shao et al., 2024; Zhang & Zuo, 2025). GRPO has been applied to general multimodal reasoning (Rouditchenko et al., 2025) and emotion tasks (Lian et al., 2025c; Zhao et al., 2025), where task-level optimization improves prediction and reasoning over supervised instruction tuning. Generative MER must optimize both final labels and evidence use. Imperfect rewards, however, invite reward hacking and over-optimization (Skalse et al., 2022; Miao et al., 2024). Rewards based on audio or visual mentions are especially vulnerable because such references do not establish modality use. AffectGPT-R1 optimizes emotion-wheel rewards for open-vocabulary MER, R1-Omni uses verifiable rewards for omni-modal emotion recognition, and EMO-R3 emphasizes reflective visual-emotion reasoning (Lian et al., 2025c; Zhao et al., 2025; Fang et al., 2026). NH-GRPO instead retains

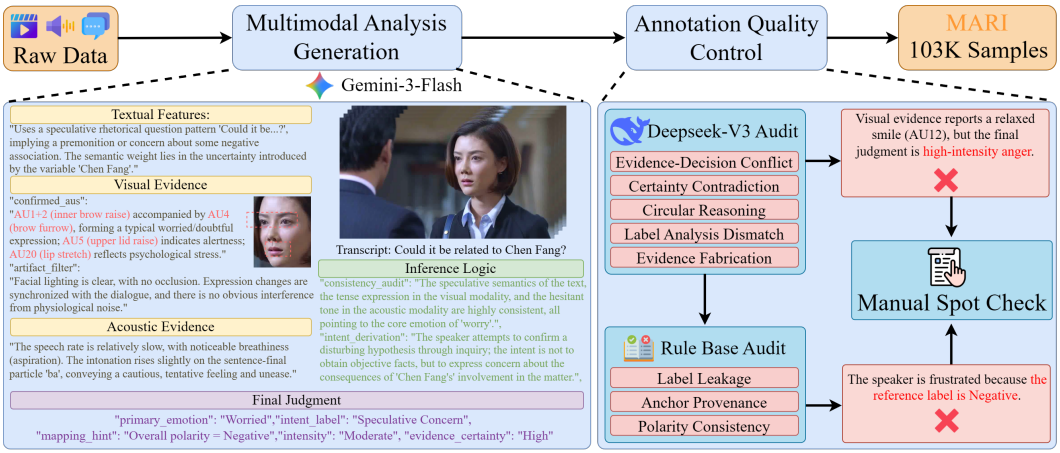

Figure 2: MARI data-construction pipeline. A multimodal teacher generates structured analyses from public video-audio-text records. DeepSeek-V3.2 audits the analyses, after which deterministic filters and stratified manual spot checks produce the 103K-sample MARI corpus.

the GRPO estimator but changes the MER reward by separating neutral-boundary error directions, preserving strict neutral handling in the hierarchy, and comparing full-modal with text-only scores. MARI provides structured evidence supervision before RL, and NH-MERBench reports neutral and non-neutral behavior under one label mapping. The contribution lies in the MER-specific task and reward formulation, not a new policy-gradient estimator.

3 METHOD

3.1 MULTIMODAL AFFECTIVE REASONING-INTEGRATED DATASET (MARI)

Existing MER datasets usually provide either categorical labels or natural-language explanations for emotion understanding (Lian et al., 2025a). Neither format explicitly separates observable evidence from cross-modal interpretation and the final decision. MARI instead provides a reusable record of evidence-grounded multimodal reasoning. The 103K records are derived from MOSI and MOSEI (Zadeh et al., 2016; 2018), MELD and MER2024 (Poria et al., 2019; Lian et al., 2024a), and SIMS_{v2}, MC-EIU, and EmotionTalk (Liu et al., 2022a; 2024; Sun et al., 2025). Together, these sources cover sentence-level sentiment videos, conversational MER, Chinese multimodal sentiment, emotion-intent interaction, and dialogue-centric emotion understanding. MARI adds structured evidence annotations while retaining each source's media unit, data split, and label semantics. The common schema applies only to the derived reasoning record. It does not replace the annotation unit or label ontology of the source dataset. Each record is built from one labeled audiovisual clip or utterance and its transcript, with the source label used only for final consistency checking. Sentence-level clips and conversational utterances can therefore share an output format without being treated as the same annotation task. Each annotation has three fields. `modality_extraction` records textual, acoustic, and visual observations, `inference_logic` reconciles the evidence, and `final_judgment` gives the emotion, intensity, evidence certainty, and key anchors. This structure trains the model to move from observations to a cross-modal decision instead of imitating a single unstructured explanation.

Figure 2 shows the construction process. The `gemini-3-flash-preview` teacher generates candidate analyses from prompts that request modality-specific evidence and stepwise reasoning. DeepSeek-V3.2 (DeepSeek-AI et al., 2025) audits the candidates independently for inconsistencies, unsupported or fabricated cues, contradictory certainty, and circular reasoning. Deterministic rules then remove failures such as explicit references to the source label. We also spot-check samples stratified by source, label type, and audit outcome, inspecting the transcript, audiovisual evidence, final label, and media path (Appendix B). Table 1 shows that MARI is the only compared resource jointly providing descriptive supervision, explicit stepwise reasoning, and multi-source coverage within one unified corpus.

3.2 EVIDENCE-GROUNDED SUPERVISED WARM START

Before RL, EmoR1 starts from an SFT-initialized Qwen2.5-Omni-7B checkpoint aligned to the MARI schema for evidence-grounded multimodal emotion generation. Given a multimodal sample

Table 1: Comparison of representative MER resources. Description denotes free-form descriptive supervision. Stepwise denotes whether a dataset explicitly supervises intermediate reasoning rather than only a single holistic explanation. Mix denotes multi-source aggregation.

Dataset	# Samples	Labels	Desc.	Stepwise	Mix	Annotation Manner
MOSEI (Zadeh et al., 2018)	23,453	✓	×	×	×	Human
MELD (Poria et al., 2019)	13,708	✓	×	×	×	Human
MER2023 (Lian et al., 2023a)	5,030	✓	×	×	×	Human
MAFW (Liu et al., 2022b)	10,045	✓	✓	×	×	Human
OV-MERD (Lian et al., 2025b)	332	✓	✓	×	×	Human-led+Model-assisted
MERR-Coarse (Cheng et al., 2024)	28,618	✓	✓	×	×	Model
MERR-Fine (Cheng et al., 2024)	4,487	✓	✓	×	×	Human-led+Model-assisted
MER-Caption (Lian et al., 2025a)	115,595	✓	✓	×	×	Model-led+Human-assisted
MER-Caption+ (Lian et al., 2025a)	31,327	✓	✓	×	×	Model-led+Human-assisted
EmotionTalk (Sun et al., 2025)	19,250	✓	✓	×	×	Human
MMEVerse (Peng et al., 2026)	129,128	✓	✓	×	✓	Multi-agent
MARI	103,945	✓	✓	✓	✓	Model-led+Human-assisted

$x = (v, a, t)$ with video, audio, and transcript, the model is trained to produce a structured response with a predicted label, text evidence, audio evidence, visual evidence, and a decision summary. Beyond format tuning, field-level evidence supervision grounds predictions in text, audio, and visual cues before reward optimization. This warm start aligns the base policy to a stable, evidence-grounded output schema and provides a reliable starting point for later RL optimization.

3.3 Reinforcement Learning with NH-GRPO

After the warm start, NH-GRPO optimizes the policy via the training pipeline in Figure 3. Unlike the broad SFT corpus, the RL subset is designed for reward learning rather than corpus coverage. We derive a high-quality reinforcement-learning subset from MARI by retaining samples whose labels, reasoning traces, and modality evidence remain consistent after Gemini regeneration, DeepSeek-V3.2 cross-audit (DeepSeek-AI et al., 2025), rule sanitization, and manual spot checks. Each retained sample must support stable paired comparison under one structured schema, allowing full-modal and text-only rollouts to be judged by the same parser and reward terms.

3.3.1 High-Quality RL Pool and Paired Rollouts

For each sample x, we generate a full-modal rollout o^{full} and a paired text-only rollout o^{text} under the same structured schema. Both paths must produce a predicted label, text evidence, audio evidence, visual evidence, and a decision summary, but under the text-only path the non-text fields are constrained to controlled abstention values rather than free-form content. Multimodal MER systems may still rely on a dominant modality instead of combining complementary evidence (Huang et al., 2026). Optimizing only one multimodal response does not reveal whether non-textual inputs improve the prediction. We therefore use the same output schema for full-modal and text-only rollouts so that their task scores can be compared directly. We then organize the retained samples into six error-aware and modality-aware buckets: Neutral-Suppression (NS), Neutral-Recovery (NR), Polarity-Reversal (PR), Fine-Grained-Refinement (FR), Modality-Gain (MG), and Stability-Anchor (SA). The buckets separate errors that otherwise receive the same aggregate reward. Class imbalance can disproportionately affect minority emotions (Meng et al., 2023), while neutral-boundary errors and same-polarity confusions require different corrections. If these cases are optimized under one undifferentiated reward stream, frequent error patterns can dominate policy updates and blur the intended decision boundary. The first four buckets target mirrored neutral-boundary errors, polarity flips, and same-polarity fine-grained confusions, the Modality-Gain bucket collects samples for which o^{full} achieves a measurable advantage over o^{text}, and the Stability-Anchor bucket preserves already-correct behavior. Because Modality-Gain is defined by paired performance rather than lexical mentions, modality credit is attached to counterfactual gain instead of stylistic references to audio or visual cues.

3.3.2 NH-GRPO Objective

Let y and $\hat{y}$ denote the canonical ground-truth and predicted labels, and let b denote the bucket assignment of the current sample. For brevity, we use NS, NR, PR, FR, MG, and SA to denote the six

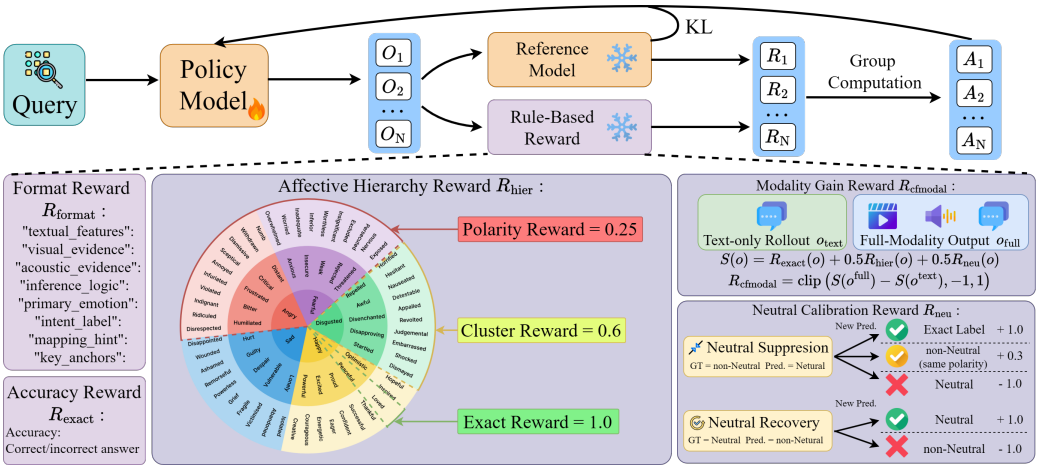

Figure 3: Training and reward pipeline of NH-GRPO. The SFT-aligned policy samples structured completions, scored by frozen reference and rule-based rewards. Group-normalized rewards produce GRPO advantages, updating the policy by relative quality within each multimodal input.

buckets defined above. Given a sample x and a full-modal rollout o^{full}, we define the total reward as

$$R = \lambda_{\text{exact}} R_{\text{exact}} + \lambda_{\text{hier}} R_{\text{hier}} + \lambda_{\text{neu}} R_{\text{neu}} + \lambda_{\text{cf}} R_{\text{cfmodal}} + \lambda_{\text{fmt}} R_{\text{format}}. \tag{1}$$

This response-level objective jointly evaluates exact correctness, hierarchical affective proximity, symmetric neutral calibration, counterfactual modality gain, and structural validity under one reward. We first preserve the primary recognition target with $R_{\text{exact}} = \mathbb{I}[\hat{y} = y]$.

To reflect semantically meaningful partial credit without collapsing neutral into the non-neutral hierarchy, we define a shared neutral-aware compatibility function

$$\mathcal{H}(y, \hat{y}) = \begin{cases} \mathbb{I}[\hat{y} = \text{neutral}], & y = \text{neutral}, \\ 1.0, & \hat{y} = y, \\ 0.6, & y \neq \text{neutral},\ \hat{y} \neq \text{neutral},\ C(\hat{y}) = C(y), \\ 0.25, & y \neq \text{neutral},\ \hat{y} \neq \text{neutral},\ P(\hat{y}) = P(y), \\ 0.0, & \text{otherwise}. \end{cases} \tag{2}$$

Following the emotion-wheel organization used in AffectGPT (Lian et al., 2025a), $C(\cdot)$ maps each non-neutral label to a cluster of related emotions. After projection to the dataset's canonical label space, $P(\cdot)$ maps the label to positive or negative polarity. Exact, cluster, and polarity matches receive 1.0, 0.6, and 0.25, respectively. Neutral is handled by its own strict branch, and the same mapping is used for training and evaluation. We then use $R_{\text{hier}} = \mathcal{H}(y, \hat{y})$.

To calibrate the two sides of the neutral boundary symmetrically, we define

$$R_{\text{neu}} = \begin{cases} -1.0, & b = \text{NS},\ \hat{y} = \text{neutral}, \\ 1.0, & b \in \{\text{NS}, \text{NR}\},\ \hat{y} = y, \\ 0.3, & b = \text{NS},\ 0 < \mathcal{H}(y, \hat{y}) < 1, \\ -1.0, & b = \text{NR},\ \hat{y} \neq y, \\ 0.0, & \text{otherwise}. \end{cases} \tag{3}$$

This term rewards restoration of the correct boundary in both directions rather than pushing the policy away from neutral with a one-sided heuristic. The value 0.3 gives limited credit when an NS sample moves into the correct non-neutral hierarchy but does not yet reach the exact label. To compare full-modal and text-only predictions, we compute the following base scores for the paired rollouts,

$$S_{\text{full}} = R_{\text{exact}}(o^{\text{full}}) + 0.5R_{\text{hier}}(o^{\text{full}}) + 0.5R_{\text{neu}}(o^{\text{full}}), \tag{4}$$

$$S_{\text{text}} = R_{\text{exact}}(o^{\text{text}}) + 0.5R_{\text{hier}}(o^{\text{text}}) + 0.5R_{\text{neu}}(o^{\text{text}}). \tag{5}$$

The R_{neu} term lets this comparison reflect changes at the neutral boundary as well as exact-label changes. We define the modality reward as

$$R_{\text{cfmodal}} = \text{clip}(S_{\text{full}} - S_{\text{text}}, -1, 1). \tag{6}$$

We apply this term only to the Modality-Gain bucket and explicitly flagged samples. It therefore rewards a score increase over the text-only prediction, not the use of audio or visual words. Finally, R_{format} checks that the response is parseable and contains all required fields.

We optimize these rewards with standard GRPO (Shao et al., 2024). In this paper, NH-GRPO refers to GRPO with the bucket-conditioned MER rewards defined above. The optimizer uses grouped token log-probabilities and parsed task outputs rather than hidden states specific to Qwen2.5-Omni. All reported experiments use Qwen2.5-Omni-7B. For each training sample x_i^{full}, the old policy samples a group of G full-modal completions,

$$o_{i,g}^{\text{full}} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_i^{\text{full}}), \qquad g = 1, \ldots, G. \tag{7}$$

Each completion is scored by the reward above, written as $R_{i,g} = R(o_{i,g}^{\text{full}})$. GRPO then converts these absolute rewards into within-group relative advantages:

$$\widehat{A}_{i,g} = \frac{R_{i,g} - \mu_i}{\sigma_i + \delta}, \qquad \mu_i = \frac{1}{G}\sum_{g=1}^{G} R_{i,g}, \tag{8}$$

$$\sigma_i^2 = \frac{1}{G}\sum_{g=1}^{G} \left(R_{i,g} - \mu_i\right)^2. \tag{9}$$

Here δ is a small numerical constant. Such normalization is essential for MER, where heterogeneous buckets produce incomparable raw reward ranges, so updates depend on within-prompt relative quality rather than absolute scores.

Let $h_{i,g,t}$ denote the token history before generating token $o_{i,g,t}$. The token-level policy ratio is

$$\rho_{i,g,t} = \frac{\pi_\theta(o_{i,g,t} \mid h_{i,g,t})}{\pi_{\theta_{\text{old}}}(o_{i,g,t} \mid h_{i,g,t})}. \tag{10}$$

We first define the clipped GRPO surrogate

$$\mathcal{L}_{i,g,t}^{\text{clip}}(\theta) = \min\Big(\rho_{i,g,t}\widehat{A}_{i,g}, \text{clip}(\rho_{i,g,t}, 1-\epsilon, 1+\epsilon)\widehat{A}_{i,g}\Big). \tag{11}$$

The resulting NH-GRPO objective is

$$\mathcal{J}_{\text{NH-GRPO}}(\theta) = \mathbb{E}_{i,g,t}\Big[\mathcal{L}_{i,g,t}^{\text{clip}}(\theta) - \beta D_{\text{KL}}(\pi_\theta(\cdot \mid h_{i,g,t}) \| \pi_{\text{ref}}(\cdot \mid h_{i,g,t}))\Big]. \tag{12}$$

Under this objective, non-textual modalities receive credit only when the full-modal score exceeds its text-only reference. Stability-Anchor samples retain already-correct cases in the training pool. Appendix E gives the full procedure.

3.4 NH-MERBENCH

NH-MERBench reports hierarchy-aware scores together with strict single-label metrics. The hierarchy-aware score reuses $\mathcal{H}$ during evaluation, while Overall Exact Acc, Macro-F1, and Neutral-F1 are computed from canonical single-label predictions without partial credit. For sample i with ground-truth label y_i, the resolver maps one model response to the dataset's canonical label space. It first normalizes explicit labels and dataset-specific affect fields. If this step does not identify a unique native label, a fixed closed-label judge selects one label from the allowed set without seeing y_i. The resulting candidate set $\widehat{\mathcal{Y}}_i$ represents one prediction in canonical form. Multiple strings are retained only when they are aliases of the same label. NH-MERBench then assigns a sample-level hierarchical score $h_i = \max_{\hat{y} \in \widehat{\mathcal{Y}}_i} \mathcal{H}(y_i, \hat{y})$, where the same compatibility function $\mathcal{H}$ is shared with the training objective. For consistency with the hierarchy above, NH-MERBench uses the same non-neutral cluster and polarity mappings defined in $\mathcal{H}$, while still treating neutral as an independent boundary condition. Let $\mathcal{I}_{\text{non}} = \{i \mid y_i \neq \text{neutral}\}$. The primary metric reported in the main result tables is Overall-HScore $= \frac{100}{N}\sum_{i=1}^{N} h_i$, which summarizes exact correctness, cluster-level proximity, polarity-level agreement, and strict neutral handling in a single number. Complementary diagnostics comprise Overall Exact Acc, Macro-F1, Neutral-F1, and NonNeutral-HScore $= \frac{100}{|\mathcal{I}_{\text{non}}|}\sum_{i \in \mathcal{I}_{\text{non}}} h_i$. We report both HScore metrics on a 0–100 scale by multiplying the mean compatibility by 100. Overall Exact Acc, Macro-F1, and Neutral-F1 use the same percentage scale. NonNeutral-HScore is computed only over samples whose ground-truth label is non-neutral. These metrics separate hierarchy-aware compatibility from exact recognition and report neutral and non-neutral behavior independently.

Table 2: Main results on NH-MERBench measured by Overall-HScore (%).

Method	MOSI	MOSEI	SIMS	$SIMS_{v2}$	MER2023	MER2024	MELD	AVG
Audio and Text Inputs								
SALMONN (Tang et al., 2024)	62.10	61.17	46.83	56.00	51.92	50.60	13.16	48.83
SECap (Xu et al., 2024)	40.96	28.80	43.98	39.26	33.60	40.84	16.08	34.79
Qwen2-Audio-7B-Instruct (Chu et al., 2024)	74.78	58.12	17.40	53.46	43.93	42.06	46.60	48.05
PandaGPT (Su et al., 2023)	71.78	63.28	49.30	53.25	27.01	40.72	14.29	45.66
AffectGPT (Lian et al., 2025a)	74.34	61.92	49.80	49.73	40.57	39.99	41.52	51.12
Emotion-LLaMA v2 (Peng et al., 2026)	66.76	64.12	50.40	53.34	49.20	48.63	41.72	53.45
EmoR1	**80.32**	**66.35**	**51.42**	**56.42**	**52.33**	**51.16**	**51.01**	**58.43**
Visual and Text Inputs								
OneLLM (Han et al., 2024)	80.17	65.12	64.93	61.57	46.33	52.54	22.30	56.14
Video-SALMONN-2 (Tang et al., 2025)	60.93	49.39	57.33	59.28	50.29	53.66	15.24	49.45
Chat-UniVi (Jin et al., 2024)	80.17	65.44	64.77	63.25	36.14	42.24	18.75	52.97
Video-LLaVA (Lin et al., 2024a)	65.11	58.08	54.05	62.25	51.68	51.93	19.70	51.83
PandaGPT (Su et al., 2023)	78.83	59.99	59.74	62.38	32.36	43.03	14.90	50.18
AffectGPT (Lian et al., 2025a)	75.36	60.94	61.33	65.05	44.06	49.05	41.27	56.72
Emotion-LLaMA v2 (Peng et al., 2026)	80.10	56.08	63.46	65.76	45.58	48.52	43.42	57.56
EmoR1	**81.34**	**65.64**	**66.79**	**66.10**	**52.16**	**54.17**	**51.26**	**62.49**
Audio, Visual, and Text Inputs								
HumanOmniV2-9B (Yang et al., 2025)	72.16	59.52	62.43	71.37	48.61	48.56	51.18	59.12
PandaGPT (Su et al., 2023)	78.13	60.87	58.42	62.48	38.83	49.71	20.39	52.69
Emotion-LLaMA (Cheng et al., 2024)	57.72	44.66	68.29	69.31	46.88	50.76	51.89	55.64
Light-MER (Zheng et al., 2026)	76.32	67.28	70.59	70.68	64.27	47.82	46.83	63.40
Emotion-LLaMA v2 (Peng et al., 2026)	82.51	68.47	69.96	**74.08**	69.22	74.26	52.06	70.08
AffectGPT (Lian et al., 2025a)	72.30	58.83	63.89	67.41	53.53	57.23	50.62	60.54
EmoR1	**82.65**	**69.52**	**71.77**	73.64	**69.95**	**74.51**	**52.89**	**70.70**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate all methods on MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), SIMS (Yu et al., 2020), $SIMS_{v2}$ (Liu et al., 2022a), MER2023 (Lian et al., 2023a), MER2024 (Lian et al., 2024a), and MELD (Poria et al., 2019) with official splits, reporting Overall-HScore as the primary metric and Overall Exact Acc, Macro-F1, Neutral-F1, and NonNeutral-HScore as diagnostics. The compared systems cover audio–text, video–text, and audio–video–text MLLMs, while EmoR1 uses Qwen2.5-Omni-7B, a 103K-sample MARI warm start, and a 6K high-quality RL subset. We feed video, audio, and transcripts through the native Qwen2.5-Omni interface, report the final SFT and RL checkpoints, and defer additional implementation details to Appendix C.

4.2 MAIN RESULTS ON NH-MERBENCH

Table 2 compares all systems under NH-MERBench across three input regimes. EmoR1 attains the highest average Overall-HScore in every regime, reaching 58.43 on audio–text, 62.49 on visual–text, and 70.70 on full audio–visual–text. Comparisons are made within each input regime because systems with different modality access are not directly comparable. EmoR1 also obtains its highest average with full audio–visual–text input. Table 3 isolates the pre-RL SFT data source's effect with backbone and training protocol fixed. MARI achieves the best average score, outperforming MER-Caption and MER-Caption+ by 4.05 and 4.17 points. Most of the difference comes from MOSEI, MER2024, and especially MELD, while MER-Caption remains higher on $SIMS_{v2}$. With the backbone and training protocol fixed, MARI gives the highest seven-dataset average in this SFT comparison.

4.3 NEUTRAL-AWARE AND SEMANTIC DIAGNOSTICS

Table 4 reports neutral and non-neutral performance separately. EmoR1 obtains the highest Neutral-F1 on all four representative benchmarks and improves NonNeutral-HScore on MOSEI, MER2024, and MELD by 1.32, 4.51, and 6.66 points over Emotion-LLaMA v2. The Neutral-F1 gains do not coincide with a general loss on non-neutral samples, since NonNeutral-HScore is preserved or improved on three of the four datasets. On $SIMS_{v2}$, NonNeutral-HScore is 1.46 points lower and Neutral-F1 is 1.56 points higher, indicating a localized trade-off.

Table 3: Training-data ablation for the supervised warm start on Qwen2.5-Omni-7B, evaluated by Overall-HScore (%) under NH-MERBench.

SFT Data	MOSI	MOSEI	SIMS	SIMS_{v2}	MER2023	MER2024	MELD	AVG
MER-Caption (Lian et al., 2025a)	71.87	64.05	**70.34**	**74.34**	63.31	65.20	35.40	63.50
MER-Caption+ (Lian et al., 2025a)	75.80	64.61	70.30	73.50	60.82	64.71	33.92	63.38
MARI (ours)	**79.88**	**67.18**	69.90	72.62	**65.08**	**68.61**	**49.61**	**67.55**

Table 4: Neutral-aware and semantic diagnostics (%). AVG is the mean across datasets.

Method	Metric	MOSEI	MER2024	MELD	SIMS_{v2}	AVG
HumanOmniV2-9B (Yang et al., 2025)	Neutral-F1	43.60	18.29	24.11	20.72	26.68
	NonNeutral-HScore	58.89	57.72	47.05	78.49	60.54
Emotion-LLaMA v2 (Peng et al., 2026)	Neutral-F1	47.54	55.31	68.30	34.86	51.50
	NonNeutral-HScore	78.45	75.15	46.32	**79.91**	69.96
AffectGPT (Lian et al., 2025a)	Neutral-F1	43.80	51.87	49.29	20.24	41.30
	NonNeutral-HScore	71.46	69.11	38.58	73.32	63.12
EmoR1	Neutral-F1	**54.88**	**58.81**	**72.08**	**36.42**	**55.55**
	NonNeutral-HScore	**79.77**	**79.66**	**52.98**	78.45	**72.72**

4.4 Ablation Studies

Table 5 evaluates the contribution of each reward term on MER2024.

The warm-start baseline without RL reaches 65.61 on MER2024. Each non-format reward improves on this baseline when tested alone. R_{neu} gives the largest increase at 3.42 points, followed by 2.85 points for R_{hier} and 0.89 points for R_{cfmodal}. Combining R_{neu} with R_{cfmodal} raises the score to 73.22, and adding R_{hier} gives a further 1.29-point increase to 74.51. Additional sensitivity analyses in Appendix F show that the default reward mixture gives the best four-dataset average and that weakening neutral calibration causes the largest drop. Table 6 isolates the three training mechanisms that operate outside the reward terms themselves. Removing the paired text-only rollout lowers Overall-HScore by 6.27 points on MOSEI and 5.11 points on SIMS_{v2}.

Table 6: Mechanism ablation (Overall-HScore, %).

Mechanism	MOSEI	MER2024	MELD	SIMS_{v2}
w/o text-only	63.25	72.80	51.63	68.53
w/o bucket	67.91	69.14	49.81	70.99
w/o neutral	66.10	67.50	47.17	61.14
Full	**69.52**	**74.51**	**52.89**	**73.64**

Removing bucketing reduces the score by 5.37 points on MER2024 and 3.08 points on MELD. Removing neutral calibration gives the largest reductions overall, including 12.50 points on SIMS_{v2}, 7.01 on MER2024, and 5.72 on MELD.

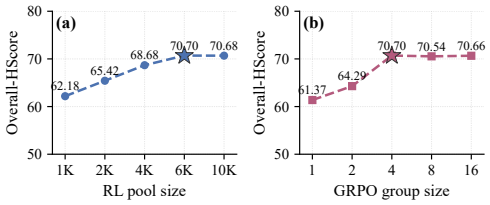

Figure 4: Sensitivity of Overall-HScore (%) to RL pool and GRPO group sizes. Stars mark the defaults.

Figure 4 reports sensitivity to the RL pool and GRPO group size. Increasing the pool from 1K to 6K raises average Overall-HScore from 62.18 to 70.70, while the 10K setting gives 70.68. For group size, the score rises from 61.37 at $G = 1$ to 70.70 at $G = 4$ and changes little at $G = 8$ or 16. We use 6K samples and $G = 4$ because the larger settings do not improve the reported average.

4.5 Case Study

Table 7 contrasts EmoR1 with two strong MLLM baselines on a representative case. AffectGPT maps the friendly, smiling exchange to neutral, whereas Emotion-LLaMA v2 predicts happy but mixes cor-

Table 5: Reward ablation on MER2024 measured by Overall-HScore (%).

Reward Functions					MER2024
R_{exact}	R_{neu}	R_{cfmodal}	R_{hier}	R_{format}	Overall-HScore (%)
Without RL					
×	×	×	×	×	65.61
Reduced rewards					
✓	✓	×	×	✓	69.03
✓	×	✓	×	✓	66.50
✓	×	×	✓	✓	68.46
Combined rewards					
✓	✓	✓	×	✓	73.22
✓	✓	✓	✓	✓	**74.51**

Table 7: An example of multimodal emotion reasoning comparing our EmoR1 with other MLLMs. Misleading evidence and incorrect predictions are marked in red, and correct evidence and predictions in green. AU denotes facial Action Units, the standard codes for observable facial muscle movements.

An Example of Multimodal Emotion Reasoning	
Frame	
Subtitle Ground truth	"Split it! Alright, I won't press you any further." happy
AffectGPT (Lian et al., 2025a)	**Pred: neutral.** "In the text, caption content: *'Alright, I won't press you any further.'* This phrase could be a response or conclusion from a young person to someone else. Given the audio clues described with an individual's sounding high-pitched and slightly shaky, along with a fast pace, we can infer that this tone might carry a sense of tension or anxiety. This sentence may reflect the emotional state of a young person who is feeling tense or anxious while responding or concluding a conversation."
Emotion-LLaMA v2 (Peng et al., 2026)	**Pred: happy.** `<think>` The person in the video is smiling and laughing, and their facial expressions are relaxed and carefree. They are also making gestures with their hands while speaking, such as waving and making a *"wow"* face. The vocal tone is light and playful, with a hint of sarcasm and teasing. `</think>` `<answer>`happy`</answer>`
EmoR1	**Pred: happy.** *Text:* The utterance contains a positive evaluative term meaning "split / share", together with consensus markers "alright" and "I won't press you", conveying satisfaction and friendliness after reaching agreement, with no aggressive or anxious modifiers. *Visual:* Pronounced AU12 (lip-corner raise) and AU6 (cheek raise) form a natural smile; gaze is soft and accompanied by a slight nod, indicating a relaxed and pleased state. *Audio:* Light prosody with an audible smile in the voice, moderate speech rate, and a slight pitch rise on the closing word — a relaxed, teasing, non-aggressive tone. *Logic:* Semantics (sharing / concession), visuals (smile), and acoustics (light prosody) are highly consistent, jointly pointing to a positive affective state. *Primary emotion:* Happiness / Satisfaction. *Intent:* Friendly Concession.

rect cues with unsupported sarcasm and no modality-level evidence. In this example, EmoR1 preserves the correct label and organizes its explanation by textual, visual, and acoustic evidence, making the basis of the decision easier to trace.

5 Conclusion

We presented EmoR1, which combines MARI supervision, neutral-aware hierarchical rewards, and paired modality-score shaping for MER. Across seven public benchmarks, EmoR1 records the highest average Overall-HScore in the audio–text, visual–text, and audio–visual–text settings. The ablations yield lower scores when neutral calibration, hierarchical compatibility, paired text-only rollouts, or error-aware bucketing is removed. Together, these results show that reliable MER needs training and evaluation protocols that make boundary decisions and modality credit explicit.

AI USE STATEMENT

In this work, we used Gemini to generate candidate annotations for MARI. We used DeepSeek-V3.2 to audit these candidate annotations and to perform closed-label resolution in NH-MERBench. We manually inspected sampled cases and take full responsibility for all AI-assisted outputs and for the final content of this paper.

REFERENCES

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 (NeurIPS)*, 2024.

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.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, et al. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025. doi: 10.48550/arXiv.2512.02556.

Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. Catch your emotion: Sharpening emotion perception in multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Yiyang Fang, Wenke Huang, Pei Fu, Yihao Yang, Kehua Su, Zhenbo Luo, Jian Luan, and Mang Ye. EMO-R3: Reflective reinforcement learning for emotional reasoning in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26584–26595, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

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

Dawei Huang, Qing Li, Chuan Yan, Zebang Cheng, Zihao Han, Yurong Huang, Xiang Li, Bin Li, Xiaohui Wang, Zheng Lian, Zhi-Qi Cheng, and Xiaojiang Peng. Emotion-Qwen: A unified framework for emotion and vision understanding. *arXiv preprint arXiv:2505.06685*, 2025.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. Apathetic or empathetic? evaluating LLMs' emotional alignments with humans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zhexian Huang, Bo Zhao, Hui Ma, Zhishu Liu, Jie Zhang, Ruixin Zhang, Shouhong Ding, and Zitong Yu. Complementarity-supervised spectral-band routing for multimodal emotion recognition. *arXiv preprint arXiv:2603.13340*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13700–13710, 2024.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. MER 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pp. 9610–9614, 2023a. doi: 10.1145/3581783.3612836.

Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, Lan Chen, Shan Liang, Ya Li, Jiangyan Yi, Bin Liu, and Jianhua Tao. Explainable multimodal emotion recognition. *arXiv preprint arXiv:2306.15401*, 2023b.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 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 (MRAC)*, 2024a.

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. *arXiv preprint arXiv:2401.03429*, 2024b.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. 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 (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014, 2025a.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. OV-MER: Towards open-vocabulary multimodal emotion recognition. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37015–37050, 2025b.

Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, Xiaobai Li, and Bin He. AffectGPT-R1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2508.01318*, 2025c.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5971–5984, 2024a. doi: 10.18653/v1/2024.emnlp-main.342.

Wang Lin, Yueying Feng, Wenkang Han, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, and Jingyuan Chen. E^3: Exploring embodied emotion through a large-scale egocentric video dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W. Schuller, and Haizhou Li. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751*, 2024. doi: 10.48550/arXiv.2407.02751.

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 (ICMI)*, 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 (ACM MM)*, pp. 24–32, 2022b. doi: 10.1145/3503161.3548190.

Tao Meng, Yuntao Shou, Wei Ai, Nan Yin, and Keqin Li. Deep imbalanced learning for multimodal emotion recognition in conversations. *arXiv preprint arXiv:2312.06337*, 2023.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xiaojiang Peng, Jingyi Chen, Zebang Cheng, Bao Peng, Fengyi Wu, Yifei Dong, Shuyuan Tu, Qiyu Hu, Huiting Huang, Yuxiang Lin, Jun-Yan He, Kai Wang, Zheng Lian, and Zhi-Qi Cheng. Emotion-LLaMAv2 and MMEVerse: A new framework and benchmark for multimodal emotion understanding. *arXiv preprint arXiv:2601.16449*, 2026.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 527–536, 2019.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-R1: Do you really need audio to fine-tune your audio LLM? *arXiv preprint arXiv:2505.09439*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Haoqin Sun, Xuechen Wang, Jinghua Zhao, Shiwan Zhao, Jiaming Zhou, Hui Wang, Jiabei He, Aobo Kong, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. EmotionTalk: An interactive chinese multimodal emotion dataset with rich annotations. *arXiv preprint arXiv:2505.23018*, 2025. doi: 10.48550/arXiv.2505.23018.

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 (ICLR)*, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. Emotion-Thinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. In *International Conference on Learning Representations (ICLR)*, 2026a.

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zeheng Wang, Zitong Yu, Yijie Zhu, Bo Zhao, Haochen Liang, Taorui Wang, Wei Xia, Jiayu Zhang, Zhishu Liu, Hui Ma, Fei Ma, and Qi Tian. AffectAgent: Collaborative multi-agent reasoning for retrieval-augmented multimodal emotion recognition. *arXiv preprint arXiv:2604.12735*, 2026b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 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 (AAAI)*, 2024.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. HumanOmniV2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 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 (ACL)*, pp. 3718–3727, 2020.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2236–2246, 2018.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, Jieping Ye, and Pheng-Ann Heng. MME-Emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210*, 2025.

Jixiao Zhang and Chunsheng Zuo. GRPO-LEAD: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

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

Kaiwen Zheng, Junchen Fu, Wenhao Deng, Hu Han, Joemon M. Jose, and Xuri Ge. Do we really need multimodal emotion language models larger than 1B parameters?, 2026. URL `https://arxiv.org/abs/2607.12787`. Accepted by ACM MM 2026.

A Detailed Dataset Descriptions

This appendix describes the public datasets used to construct MARI and to evaluate NH-MERBench. MARI is built from seven sources, namely MOSI, MOSEI, SIMS_{v2}, MER2024, MELD, MC-EIU, and EmotionTalk. These datasets jointly cover English opinion videos, Chinese sentiment videos, short-video MER, multi-party dialogue, and emotion-intent interaction. NH-MERBench evaluates seven datasets, namely MOSI, MOSEI, SIMS, SIMS_{v2}, MER2023, MER2024, and MELD, always following the official train, validation, and test splits. Since the resources differ in label granularity and annotation ontology, we preserve native dataset annotations and evaluate each benchmark under its own label space.

Language and label handling. MOSI, MOSEI, and MELD provide English transcripts. SIMS_{v2}, MER2024, and EmotionTalk provide Chinese transcripts, while MC-EIU contains separately processed English and Mandarin subsets. We keep each transcript in its source language. A shared Chinese instruction template standardizes the reasoning-record format but does not replace dataset-specific labels. The `mapping_hint` field records only positive, negative, or neutral direction, and NH-MERBench maps predictions back to the native label space of each dataset. The original transcript, audiovisual evidence, and source labels therefore retain their dataset-specific meaning.

CMU-MOSI (Zadeh et al., 2016). CMU-MOSI contains 2,199 opinion-level video segments collected from YouTube movie reviews. Each segment is paired with audio, visual frames, and a transcript,

and is annotated with a continuous sentiment score in $[-3, +3]$. The dataset is widely used for sentence-level multimodal sentiment analysis, where the neutral boundary is often sensitive to small shifts in lexical or paralinguistic evidence.

CMU-MOSEI (Zadeh et al., 2018). CMU-MOSEI extends MOSI to 23,453 video segments from more than 1,000 speakers. It provides both sentiment intensity annotations and emotion labels, making it a broader testbed for multimodal affect recognition in unconstrained online videos. Its larger scale also exposes whether a method can maintain stable neutral and non-neutral decisions across diverse speakers and topics.

CH-SIMS (Yu et al., 2020). CH-SIMS is a Chinese single- and multi-modal sentiment dataset with 2,281 refined video segments in the wild. It provides fine-grained sentiment annotations for the multimodal sample and for individual modalities, which makes it useful for checking whether a model can reason over Chinese textual cues together with acoustic and visual evidence.

CH-SIMS v2 (Liu et al., 2022a). CH-SIMS v2 extends CH-SIMS by adding 2,121 refined supervised video segments and 10,161 unlabeled raw segments with richer acoustic and visual context. It follows the same multimodal sentiment analysis setting while increasing the diversity of acoustic and visual conditions. In NH-MERBench, it is reported separately from CH-SIMS because the two versions differ in scale and data construction.

MER2023 (Lian et al., 2023a). MER2023 is a multimodal emotion recognition challenge dataset built from short video clips with audio and textual transcripts. The original challenge covers multi-label learning, modality robustness, and semi-supervised learning, while our benchmark uses its discrete emotion labels for category-level MER evaluation. This setting is useful for compact clips where non-textual cues often disambiguate similar affective states.

MER2024 (Lian et al., 2024a). MER2024 extends the MER challenge setting toward semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. The original benchmark expands the data scale and introduces a track that allows flexible emotion labels beyond a fixed closed set. We use the dataset-specific evaluation labels in NH-MERBench, which keeps the benchmark aligned with the public protocol while still allowing MARI annotations to retain richer affective descriptions.

MELD (Poria et al., 2019). MELD is a multimodal multi-party dialogue dataset collected from *Friends*. Each utterance includes audio, visual context, and text, and is annotated with one of seven emotions, including anger, disgust, fear, joy, neutral, sadness, and surprise. Its dialogue structure makes it a useful complement to opinion-video datasets because emotion cues depend on conversational context and speaker interaction.

MC-EIU (Liu et al., 2024). MC-EIU focuses on joint emotion and intent understanding in multimodal conversations. It covers textual, acoustic, and visual modalities in English and Mandarin, with seven emotion categories and nine intent categories. The dataset is relevant to MARI because emotion explanations often require distinguishing expressed feeling from communicative purpose, so we keep these annotations in their native form rather than forcing them into a single global emotion taxonomy.

EmotionTalk (Sun et al., 2025). EmotionTalk is an interactive Chinese multimodal emotion dataset built from dyadic conversations. It contains 23.6 hours of speech and 19,250 utterances, with acoustic, visual, and textual modalities, seven utterance-level emotion categories, five sentiment labels, and speech-caption annotations. The dataset broadens the coverage of Chinese conversational affect and supports structured supervision that records both coarse affective polarity and fine-grained emotion descriptions.

B MARI Construction Details

MARI is a derived multimodal instruction corpus, not a new collection of raw media or a new benchmark split. Its ShareGPT records contain structured analyses and references to source media. The media files and native labels remain with the original datasets. Construction begins with public dataset metadata and local media files. A multimodal teacher generates structured affective analyses, an independent language model audits them, and deterministic rules remove invalid records. The retained records are then converted to a ShareGPT-compatible video-audio instruction format. The

Table 8: Source composition of the final MARI ShareGPT training file. Counts are measured after generation, audit filtering, deterministic sanitization, path conversion, and merge.

Source	# Samples	Split used for MARI	Construction note
MOSI	1,484	train+valid	English opinion videos with sentiment supervision.
MOSEI	17,840	train+valid	Larger English opinion-video corpus with diverse speakers.
SIMS_{v2}	3,344	train+valid	Chinese multimodal sentiment videos.
MER2024	4,919	labeled set	Short-video MER data with discrete emotion labels.
MELD	10,489	train+dev	Multi-party dialogue utterances; the test CSV is not read.
EmotionTalk	16,613	non-test speakers	Interactive Chinese dialogues, excluding speaker groups `G00003` and `G00015`.
MC-EIU	49,256	available train sources	English and Chinese emotion-intent conversation subsets, treated as one dataset family.
Total	103,945	–	Final merged MARI corpus used for supervised warm start.

resulting file contains 103,945 ShareGPT-format training samples. Table 8 summarizes the merged source composition.

Source filtering and generation. For MOSI, MOSEI, and SIMS_{v2}, we retain the train and validation splits and exclude the official test split. For MELD, we use train and development records. EmotionTalk excludes the README-defined test speaker groups `G00003` and `G00015`, MER2024 uses the labeled video set, and MC-EIU processes the English and Chinese subsets separately before merging them as one dataset family. For each eligible record, we verify that the referenced media is available, preserve a stable ordering, and send the video, audio, and transcript to `gemini-3-flash-preview` for candidate annotation. The generation prompt is designed to prevent label-conditioned rationalization while still allowing the source label to serve as a final consistency reference. The following excerpt reports the essential prompt instructions used for candidate generation, omitting only formatting boilerplate and dataset-specific path handling.

> *Act as an expert in multimodal behavior and intent analysis. Given the video, audio, and transcript, first extract textual sentiment cues, directly observable facial action-unit evidence, and acoustic prosody. Then reconcile the three modalities into an evidence-grounded affective judgment. Use any provided reference label only for final consistency checking, never as a premise in the reasoning chain. Do not invent missing transcript words, unobserved facial actions, unsupported acoustic cues, or external context. If evidence is weak, incomplete, or cross-modally conflicting, lower the evidence certainty rather than forcing a confident emotion. Return a JSON object with* `modality_extraction`*,* `inference_logic`*, and* `final_judgment`*.*

Candidate generation uses `gemini-3-flash-preview` with JSON-mode output and temperature 0.1. The same semantic template is used across sources; only the media, source-language transcript, and dataset-specific reference fields vary. Its design follows three fixed constraints: evidence is extracted before the final judgment, observations are separated from inference, and every response follows the same three-block schema. Cross-model auditing uses DeepSeek-V3.2 with temperature 0.1 and the fixed five-category conflict taxonomy below. EmoR1 benchmark inference uses a fixed system/user template with temperature 0. The NH-MERBench free-form label resolver uses `deepseek-chat` at temperature 0.1 with a fixed dataset-specific closed label list and receives the model response but not its ground-truth label. Candidate outputs must follow this fixed three-block schema. `modality_extraction` records textual, visual, and acoustic observations; `inference_logic` reconciles cross-modal evidence and applies affective decision rules; and `final_judgment` reports `primary_emotion`, `intent_label`, `mapping_hint`, intensity, evidence certainty, and key anchors.

Cross-model audit. We then audit the teacher-generated analyses with DeepSeek-V3.2 as an independent judge. The auditor receives the transcript, source label, and generated JSON analysis,

and returns a machine-readable audit record rather than free-form commentary. The core audit prompt is:

> *Act as a strict multimodal emotion-analysis auditor. Read the complete generated report before judging conflicts. Check five conflict types:* `modality_vs_conclusion`, `evidence_fabrication`, `certainty_contradiction`, `circular_reasoning`, *and* `label_analysis_mismatch`. *Report only conflicts that remain valid after reading the relevant fields. Prefer missing a subtle issue over falsely flagging a consistent sample. Output only JSON with* `has_conflict`, `conflicts`, `overall_quality`, *and* `summary`.

The audit separates internal self-consistency failures from external label-alignment mismatches. Internal failures cover contradictions between extracted modal evidence and the final emotion, fabricated anchors, inconsistent certainty, and circular use of the source label. The external mismatch category flags cases where the final affective direction or intent is incompatible with the source annotation.

Deterministic sanitization. Model-based auditing is followed by hard rules for failures that should not require another model judgment. The most important rule is label isolation: any generated analysis that explicitly mentions the reference label, a known label, or a shortcut such as "consistent with the label" is removed rather than repaired. We further enforce schema validity, anchor provenance, polarity consistency, certainty-intensity consistency, transcript hallucination guards, modality-conflict explanation guards, split isolation, media-path validity, and duplicate removal. When a deterministic rule fails, the record is removed rather than edited, preserving the independence of the retained supervision.

Merge format and spot checks. After filtering, retained analyses are converted to the Qwen2.5-Omni training format. Each final sample contains a `messages` list with a system role, a user role, and an assistant role, plus parallel `videos` and `audios` fields pointing to the same media file. The user message starts with `<video><audio>`, includes the transcript, and asks for multimodal emotion and intent analysis; the assistant message is the structured MARI JSON containing the three analysis blocks described above. Finally, stratified random spot checks are performed across source datasets, language groups, neutral/non-neutral label types, and automatic-audit outcomes. Reviewers inspect each sampled record together with its original audiovisual media, transcript, and generated analysis. They check transcript fidelity, observable modality anchors, cross-modal reconciliation, final-label consistency, and media-path validity. Spot checks are used to identify systematic errors missed by automatic filters, not to rewrite individual analyses.

C Implementation Details

EmoR1 is initialized from Qwen2.5-Omni-7B (Xu et al., 2025) and adapted with LoRA (Hu et al., 2022). We keep the native Qwen2.5-Omni video-audio-text processor and generate all MER predictions from the Thinker text branch. No handcrafted affective features are added.

Training follows a two-stage recipe. The SFT stage trains LoRA adapters on MARI for three epochs with rank $r = 16$, scaling $\alpha = 32$, dropout 0.05, no bias terms, AdamW with fused updates, cosine decay, warm-up ratio 0.05, and learning rate $3{\times}10^{-5}$; the SFT checkpoint is taken at step 9,747. The GRPO stage starts from the merged SFT checkpoint and optimizes NH-GRPO on the bucketed RL pool. Each prompt samples $G = 4$ full-modal completions, while the text-only counterfactual score is cached during RL-pool construction. The final GRPO checkpoint is taken at step 6,900 with learning rate $2{\times}10^{-5}$, clipping parameter $\epsilon = 0.2$, KL coefficient $\beta = 0.04$, and reward weights $(1.0, 0.35, 0.4, 0.25, 0.1)$ for exact, hierarchical, neutral, counterfactual-modality, and format rewards. All training is conducted on four H100 GPUs with 80GB memory each.

D Detailed Architecture Specifications

Table 9 gives the compact architecture and training hyperparameters used in all reported EmoR1 results.

Table 9: Architecture hyperparameters of EmoR1.

Component	Configuration
Backbone (Qwen2.5-Omni-7B)	
Hidden dimension d	3584
Text decoder layers	28
Audio / vision encoders	32 / 32 layers
Output branch	Thinker text generation
LoRA Adaptation	
LoRA rank r / scaling α	16 / 32
LoRA dropout	0.05
LoRA target modules	Thinker attention and MLP projections
Trainable parameters	40.37M (0.37%)
Training	
SFT	3 epochs, lr 3×10^{-5}, step 9,747
GRPO	1 epoch, lr 2×10^{-5}, step 6,900
Precision and memory strategy	bf16, gradient checkpointing, DeepSpeed ZeRO-2
Rollout sampling	$G = 4$, temperature 0.7, top-$p = 0.9$
Length limits	sequence 12288, completion 2048
Reward weights	1.0 / 0.35 / 0.4 / 0.25 / 0.1

Algorithm 1 EmoR1 Training with MARI and NH-GRPO

Require: MARI corpus $\mathcal{D}_{\text{sft}}$; RL pool $\mathcal{D}_{\text{rl}}$ with $(x, y, b, S_{\text{text}})$; base policy π_0
Ensure: Final policy $\pi_{\theta^\star}$
Initialize LoRA adapters on the Thinker branch of π_0.
for each SFT epoch **do**
  for each supervised sample $(x, z) \in \mathcal{D}_{\text{sft}}$ **do**
    Encode $x = (v, a, t)$ with the native Qwen2.5-Omni processor.
    Update LoRA parameters by next-token cross entropy on target z.
  end for
end for
Merge the SFT adapters and initialize the GRPO policy.
for each GRPO update **do**
  Sample $(x, y, b, S_{\text{text}})$ from $\mathcal{D}_{\text{rl}}$, $b \in \{\text{NS}, \text{NR}, \text{PR}, \text{FR}, \text{MG}, \text{SA}\}$.
  Draw $G = 4$ full-modal completions $\{o_g^{\text{full}}\}_{g=1}^{G}$ from $\pi_{\theta_{\text{old}}}$.
  for $g = 1$ to G **do**
    Parse o_g^{full} into emotion fields and format-validity flags.
    Compute R_{exact}, R_{hier}, R_{neu}, and R_{format}.
    $S_{\text{full}} \leftarrow R_{\text{exact}} + 0.5R_{\text{hier}} + 0.5R_{\text{neu}}$.
    if $b = \text{MG}$ or the sample is modality-sensitive **then**
      $R_{\text{cfmodal}} \leftarrow \text{clip}(S_{\text{full}} - S_{\text{text}}, -1, 1)$.
    else
      $R_{\text{cfmodal}} \leftarrow 0$.
    end if
    $R_g \leftarrow R_{\text{exact}} + 0.35R_{\text{hier}} + 0.4R_{\text{neu}} + 0.25R_{\text{cfmodal}} + 0.1R_{\text{format}}$.
  end for
  Normalize $\{R_g\}_{g=1}^{G}$ to obtain advantages $\{\widehat{A}_g\}_{g=1}^{G}$.
  Update π_θ with the clipped GRPO surrogate and KL regularization.
end for
return final GRPO checkpoint $\pi_{\theta^\star}$.

E ALGORITHM

Algorithm 1 gives the implementation-level training procedure. The algorithm keeps the paired-rollout design of Section 3, but follows the local Stage-2 implementation by storing the text-only counterfactual score in $\mathcal{D}_{\text{rl}}$ and sampling only full-modal completions during online GRPO updates.

Table 10: Reward-weight sensitivity measured by Overall-HScore (%) on four representative benchmarks. Each row changes one reward coefficient while keeping the others fixed. Δ Avg. is the change in percentage points from the default.

Setting	λ_{hier}	λ_{neu}	λ_{cf}	MOSEI	MER2024	MELD	SIMS_{v2}	Avg.	Δ Avg.
Default	0.35	0.40	0.25	69.52	**74.51**	52.89	73.64	**67.64**	–
Weak hierarchical	0.20	0.40	0.25	68.97	73.20	53.42	72.45	67.01	-0.63
Strong hierarchical	0.50	0.40	0.25	69.05	74.36	51.50	**73.81**	67.18	-0.46
Weak neutral	0.35	0.20	0.25	68.30	74.62	49.85	73.18	66.49	-1.15
Strong neutral	0.35	0.60	0.25	**70.18**	73.40	**53.62**	72.20	67.35	-0.29
Weak counterfactual	0.35	0.40	0.10	68.96	73.30	52.40	72.72	66.85	-0.79
Strong counterfactual	0.35	0.40	0.40	69.10	74.38	51.95	73.21	67.16	-0.48

F Additional Experimental Details and Results

This section reports sensitivity to the reward weights, RL pool size, and GRPO group size around the default configuration.

Table 10 tests whether the reward mixture is robust to coefficient changes. The default configuration obtains the best average score, while each one-dimensional perturbation lowers the four-dataset average by 0.29–1.15 points. The largest drop appears when λ_{neu} is weakened, mainly due to the decrease on MELD, showing that neutral calibration must be strong enough to shape the decision boundary in conversational data. Increasing λ_{neu} improves MOSEI and MELD but hurts MER2024 and SIMS_{v2}, indicating that over-emphasizing neutral can trade off against dataset-specific non-neutral distinctions. Changing λ_{hier} and λ_{cf} gives smaller average losses, but the best per-dataset values occur at different weights, so the default mixture is used as a cross-dataset compromise rather than a setting tuned to a single benchmark.

Figure 4 shows that the gains come from enough targeted RL data and enough within-prompt response diversity, not from unbounded scaling. Increasing the RL pool from 1K to 6K raises the average Overall-HScore from 62.18 to 70.70, but expanding it to 10K gives 70.68, suggesting that the bucketed 6K pool already covers the main correctable error modes. For GRPO, group size 1 and 2 underperform sharply, whereas group size 4 reaches 70.70 and group sizes 8 and 16 remain essentially flat. We therefore use 6K samples and group size 4 as the default because they achieve the best accuracy-cost trade-off among the tested settings.

Table 11: An example of multimodal emotion reasoning comparing our EmoR1 with other MLLMs. Misleading evidence and incorrect predictions are marked in red, and correct evidence and predictions in green. AU denotes facial Action Units, the standard codes for observable facial muscle movements.

A Neutral Example of Multimodal Emotion Reasoning	
Frame	
Subtitle Ground truth	"How much is it? Five hundred yuan." neutral
AffectGPT (Lian et al., 2025a)	**Pred: angry.** "In the text, the caption reads: *"How much is it? Five hundred yuan."* This sentence might be a question or reaction from a little girl to someone beside her. Based on the audio clues where the individual speaks in a high pitch and excited tone, as well as the video clues showing the little girl's curious and surprised expression, we can infer that the sentence might carry a tone of curiosity or surprise. Therefore, this sentence could be the little girl asking or reacting to the price of something, which is consistent with her overall sense of curiosity and surprise."
Emotion-LLaMA v2 (Peng et al., 2026)	**Pred: surprise.** `<think>` The person in the video appears to be expressing surprise or disbelief, as they raise their eyebrows and furrow their brows upon hearing the amount of money mentioned. Their mouth is slightly open, and they utter the phrase in a slightly incredulous tone. `</think> <answer>`surprise`</answer>`
EmoR1	**Pred: neutral.** *Text:* The text contains an interrogative sentence asking about the price, "How much is it?", and a confirmation of the price, "Five hundred yuan." Semantically, it is a purely factual inquiry and feedback, without any subjective modifiers such as *too expensive* or *cheap*. *Visual:* AU12 (slight lip-corner raise) shows a polite social smile; AU1/2 (slight eyebrow raise) accompanies the inquiry action, showing attention and curiosity. The gaze is directed at the interlocutor, the expression is natural, and there are no exaggerated emotional fluctuations. *Audio:* The tone is steady and the speech rate is moderate. The first sentence has a standard interrogative intonation, and the second sentence has the confirming intonation of a factual statement, with no obvious stress emphasis, high-pitched tone, or laughter; it appears as the natural state of everyday conversation. *Logic:* The textual, visual, and acoustic modalities are highly consistent, all pointing to a calm and objective interaction state. No conflict is found between the semantic tendency and physiological signals. *Primary emotion:* Neutral. *Intent:* Information Seeking.

Table 11 shows a representative neutral example where the baselines over-interpret routine information seeking as high-arousal affect. AffectGPT copies the price-question caption, but then introduces unsupported cues such as a "little girl," a "reaction," and a "high pitch and excited tone," before assigning an angry label that is inconsistent with its own curiosity/surprise rationale. Emotion-LLaMA v2 predicts surprise by treating a normal price question, slight eyebrow movement, and ordinary response contour as disbelief. EmoR1 instead preserves the neutral label because its modality-separated evidence identifies factual price inquiry, natural AU12/AU1/2 movements, steady prosody, and high cross-modal consistency without semantic–physiological conflict. This example complements the neutral-aware diagnostics in Section 4: improved neutral calibration is not only a score-level effect, but also appears in the model's ability to avoid turning routine conversational cues into surprise, anger, or disbelief.