EMOTRUST: LEARNING WHICH MODALITY TO TRUST FOR MULTIMODAL EMOTION REASONING UNDER CROSS-MODAL CONFLICT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal emotion reasoning aims to infer human affect from visual, acoustic, and linguistic signals, but remains challenging when modalities convey conflicting emotional evidence. Existing methods improve perception and reasoning quality, yet they typically fuse multimodal cues without explicitly modeling which modality should be trusted more when cross-modal emotional conflict arises. To address this challenge, we propose EmoTrust, a framework for multimodal emotion reasoning that explicitly resolves such conflicts through adaptive modality preference learning. EmoTrust begins with multi-view supervised fine-tuning on full and modality-ablated inputs, enabling robust reasoning under varying modality availability. It then estimates cross-modal consistency through cue-grounded modality-level emotion support distributions. To resolve modality disagreement, EmoTrust further introduces conflict-conditioned policy optimization, which compares confidence under full and ablated inputs to estimate modality-specific contributions and reinforce reasoning grounded in the more informative modality. Experiments on several benchmarks show that EmoTrust consistently outperforms strong baselines, especially on low-consistency samples, while learning modality preferences that align well with human judgments.

1 INTRODUCTION

Multimodal emotion recognition aims to infer human affect from visual, acoustic, and linguistic signals in videos and conversations, and has become a central task in affective computing and human-centered interaction (Ma & Yarosh, 2021; Veltmeijer et al., 2023). With the rise of multimodal large language models (MLLMs), this task is evolving from fixed-label classification to open-vocabulary emotion reasoning with natural-language explanations (Xie et al., 2024; Cheng et al., 2024; Lee et al., 2024). This paradigm is appealing because it captures subtle, compound, and context-dependent affective states beyond predefined taxonomies, while making the prediction process more interpretable.

Recent work has advanced this direction through structured reasoning (Fang et al., 2026), salient cue modeling (Xu et al., 2026), and perception-oriented optimization Han et al. (2026). Yet a key challenge remains: when modalities imply different emotions, existing methods do not explicitly determine which modality should guide the final prediction. This issue is especially severe in naturalistic scenarios, where emotional evidence is sparse, asynchronous, and unevenly distributed across modalities (Tsai et al., 2019; Yang et al., 2026). For example, a speaker may use neutral words with a strained voice, smile while expressing disappointment, or suppress visible emotion despite strong internal arousal. In such cases, the main difficulty is not only perceiving each modality faithfully, but also deciding how much each should be trusted for the target emotion.

To address this challenge, we propose EmoTrust, a framework for multimodal emotion reasoning under cross-modal emotional conflict. EmoTrust first identifies potentially conflicting samples through cue-grounded modality-level emotion support, then estimates each modality's target-related contribution by comparing confidence under full and modality-ablated inputs, and finally performs contribution-weighted counterfactual policy optimization to reinforce reasoning grounded in the more informative modality. Experiments across sentiment analysis, basic emotion recognition, and

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.

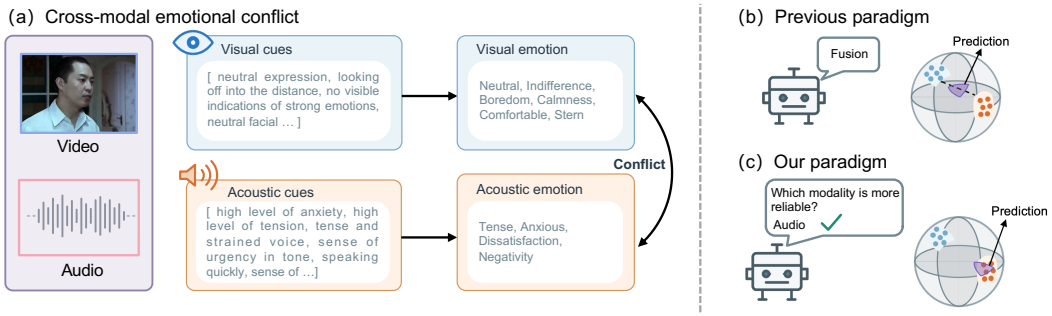


Figure 1: (a) Example of cross-modal emotional conflict, where visual and acoustic modalities convey divergent affective cues. (b) Previous paradigm, which directly predicts from fused multimodal representations. (c) Our paradigm, which estimates modality reliability under conflict and guides prediction with the more informative modality.

open-vocabulary emotion reasoning benchmarks show that EmoTrust consistently improves performance, especially on low-consistency samples, and learns modality preferences that align well with human judgments. Our contributions are summarized as follows:

1. We identify **cross-modal emotional conflict** as a fundamental challenge in multimodal emotion reasoning and introduce a **cue-grounded consistency estimation** method that maps visual and acoustic evidence into modality-level emotion distributions to explicitly quantify their agreement.

2. We propose **EmoTrust**, a two-stage training framework that combines multi-view supervised fine-tuning with conflict-conditioned policy optimization, enabling the model to learn **sample-dependent modality reliance** by comparing confidence under full and modality-ablated inputs.

3. We conduct extensive experiments and analyses across sentiment analysis, categorical emotion recognition, and open-vocabulary emotion reasoning benchmarks, showing that **EmoTrust consistently outperforms strong baselines**, yields the largest gains on low-consistency samples, and learns modality preferences that align well with **human judgments**.

2 Related Work

Multimodal emotion reasoning. Multimodal emotion recognition traditionally infers affect from visual, acoustic, and linguistic signals, mainly under closed-set settings with dedicated architectures and benchmark datasets (Zhang et al., 2024a; Veltmeijer et al., 2023). With the rise of MLLMs, this research is shifting toward open-vocabulary multimodal emotion reasoning (MER), where models generate both emotion predictions and natural-language explanations (Lian et al., 2025; 2024). Recent methods improve MER through instruction tuning (Cheng et al., 2024), structured reasoning and reflective learning (Fang et al., 2026), salient cue selection with step-by-step reasoning (Xu et al., 2026), and perception-oriented optimization (Han et al., 2026). However, they mainly improve overall reasoning quality and do not explicitly address which modality should be trusted when modalities convey inconsistent emotions. Our work fills this gap by modeling *sample-dependent modality reliance* under cross-modal emotional conflict.

Modality preference learning. A related line of work studies modality imbalance, asynchronous fusion, and unimodal bias in multimodal learning, showing that modalities often contribute unequally and may interfere during optimization (Tsai et al., 2019; Zhang et al., 2024b; Wei & Hu, 2024). Existing approaches mainly address this issue through fusion design, gradient modulation, or optimization strategies rather than through explicit reasoning. In affective computing, explicit treatment of cross-modal inconsistency remains relatively limited. A representative effort is MoSEAR (Han et al., 2025), which mitigates modality bias through expert-based modeling and attention reallocation. In contrast, EmoTrust models conflict in an emotion-semantic space and learns sample-level modality preference from counterfactual confidence changes under full and modality-ablated reasoning.

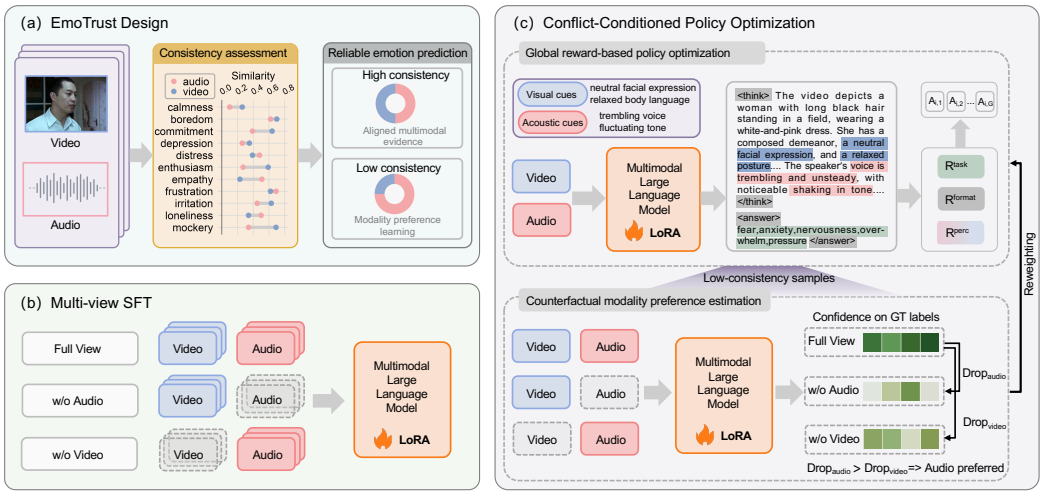


Figure 2: Overview of the EmoTrust framework. We propose a two-stage training paradigm. (1) Multi-view supervised fine-tuning (SFT) encourages the model to learn stable emotion representations across modalities. (2) Conflict-conditioned policy optimization estimates the relative contributions of visual and acoustic evidence under cross-modal conflict and leverages these estimates to guide modality-grounded reasoning.

3 METHODOLOGY

3.1 OVERVIEW

Problem formulation. Given a multimodal input X^F, we denote its visual, acoustic, and textual streams as V, A, and T, respectively, where $X^F = \{V, A, T\}$. A multimodal emotion reasoning (MER) model π_θ generates a response $y = [r, \hat{Y}]$, where r is the intermediate reasoning trace and $\hat{Y}$ is the emotion prediction. A core challenge is that modalities may convey conflicting affective signals, making modality trust crucial for reliable prediction.

EmoTrust framework. We propose EMOTRUST (Figure 2), a framework for learning modality reliance in MER under cross-modal emotional conflict. First, multi-view supervised fine-tuning exposes the model to both full inputs and leave-one-modality-out views, improving reasoning under varying modality availability while discouraging hallucination from missing modalities. Next, a cue-to-emotion mapping converts fine-grained visual and acoustic cues into modality-level emotion distributions, whose similarity measures cross-modal emotional consistency. Finally, conflict-conditioned policy optimization learns modality preference for emotion prediction under conflict.

3.2 MULTI-VIEW SUPERVISED FINE-TUNING

We introduce multi-view SFT as a cold-start stage to enable the model to learn emotion expression patterns across modalities and to mitigate hallucinated evidence for missing or ambiguous modalities during emotion reasoning. For each SFT example, we construct three SFT views:

$$X_i^F = (V_i, A_i, T_i), \qquad X_i^{-A} = (V_i, \varnothing, T_i), \qquad X_i^{-V} = (\varnothing, A_i, T_i), \tag{1}$$

where X_i^{-A} and X_i^{-V} remove the acoustic and visual streams, respectively. Each view is paired with a supervised target consisting of a reasoning trace in the format `<think>...</think>`, followed by a prediction in the format `<answer>...</answer>`. All views share the same ground-truth emotion label, but their reasoning traces are view-specific and include only the evidence available in the corresponding input. Specifically, we use the complete annotated reasoning trace for the full view and construct each ablated-view target by removing the reasoning spans that rely on the omitted modality.

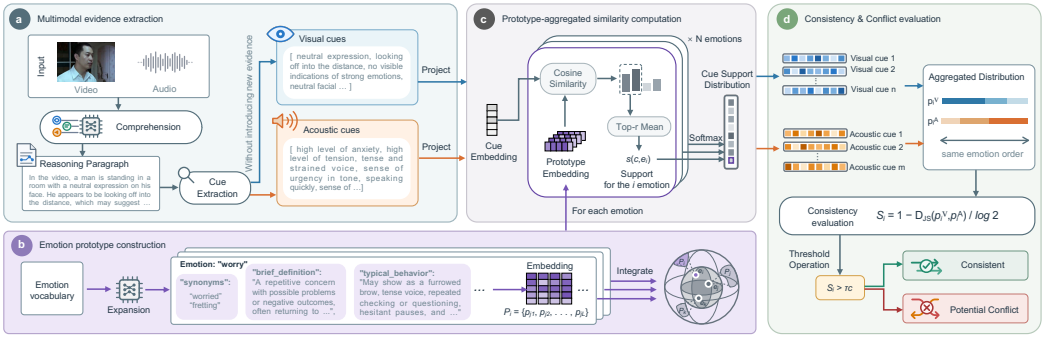


Figure 3: Cross-modal emotion consistency estimation. Explicit visual and acoustic cues are mapped to soft emotion-support distributions in a shared embedding space, aggregated into modality-level distributions, and compared with normalized Jensen-Shannon similarity to identify consistency and cross-modal conflict.

3.3 Cue-Grounded Cross-Modal Emotion Consistency Estimation

MER is not equally supported across modalities, as some emotions are conveyed more strongly by visual evidence while others rely more on acoustic cues. To make this heterogeneity explicit, we design a cue-to-emotion mapping scheme (Figure 3) that transforms fine-grained evidentiary statements from each modality into emotion-support distributions, from which we derive an inter-modal consistency measure.

Cue-level emotion support. Following Han et al. (2026), we use GPT-5.4 (Singh et al., 2025) to extract visual and acoustic cue sets $\mathcal{C}_i^V$ and $\mathcal{C}_i^A$ from each ground-truth reasoning paragraph, restricted to explicitly stated evidence. We then map each extracted cue to a soft support distribution over the global emotion vocabulary $\mathcal{E}$ (details are provided in Appendix A). To improve robustness to lexical variation in cue descriptions, we define a prototype set $P(e_j)$ for each $e_j \in \mathcal{E}$, containing its canonical label, synonyms, and short descriptions. For each extracted cue c, we compute its embedding $\mathbf{e}(c)$ using Qwen3-Embedding (Zhang et al., 2025), and encode all prototypes in the same embedding space. We then measure the compatibility between cue c and emotion e_j by a prototype-aggregated similarity score $s(c, e_j)$, defined as the mean cosine similarity between $\mathbf{e}(c)$ and the top-r most similar prototype embeddings in $P(e_j)$. Based on these similarities, the cue-level soft support distribution is computed as

$$q(e_j \mid c) = \frac{\exp\left(s(c, e_j)/\tau_e\right)}{\sum_{k=1}^{|\mathcal{E}|} \exp\left(s(c, e_k)/\tau_e\right)}, \tag{2}$$

where τ_e controls the sharpness of the distribution and is set to 0.02. This soft assignment allows each cue to contribute to multiple related emotions with different strengths, instead of enforcing a hard label assignment. Consequently, $q(e_j \mid c)$ provides a more fine-grained representation of emotion evidence.

Modality-level emotion distributions. We aggregate the cue-level supports within each modality to obtain

$$p_i^m(e_j) = \sum_{c \in \mathcal{C}_i^m} w_{i,c}^m q(e_j \mid c), \qquad \sum_{c \in \mathcal{C}_i^m} w_{i,c}^m = 1, \qquad m \in \{V, A\}. \tag{3}$$

Here, $w_{i,c}^m$ denotes the normalized reliability weight of cue c, which is set uniformly in this paper. Thus, p_i^V and p_i^A are modality-specific emotion-support distributions that summarize how the extracted cues in each modality distribute evidentiary support across emotions.

Consistency-aware conflict identification To measure agreement between the two modality-specific emotion-support distributions, we use the normalized Jensen-Shannon similarity:

$$S_i = 1 - \frac{\mathbb{D}_{JS}(p_i^V || p_i^A)}{\log 2}, \tag{4}$$

where $S_i \in [0, 1]$, with smaller values indicating greater discrepancy between the emotion interpretations supported by visual and acoustic cues. Based on this measure, a threshold τ_c is used to identify samples with potential cross-modal emotion conflict, defined by $S_i < \tau_c$.

3.4 Conflict-Conditioned Policy Optimization

Although standard GRPO is effective for LLM reasoning, it relies on trajectory-level advantage estimation and cannot determine which modality to trust when visual and acoustic cues conflict. We therefore propose conflict-conditioned policy optimization, which estimates their relative contributions under disagreement and guides modality-grounded reasoning accordingly. This encourages the policy to favor the modality providing stronger support for the target emotion, rather than treating conflicting modalities as equally reliable.

Reward Design. Given a full multimodal input X_i^F, we sample a rollout group of G complete responses, denoted by $y_{i,j}^F = [r_{i,j}^F, \hat{Y}_{i,j}^F]$. Here, $r_{i,j}^F$ is the intermediate reasoning trace and $\hat{Y}_{i,j}^F$ is the predicted emotion set. Each response receives three reward signals. The task reward $R_{i,j}^{\text{task}}$ measures how well $\hat{Y}_{i,j}^F$ matches the ground-truth label set Y_i under a normalized, emotion-structure-aware evaluation metric (Lian et al. (2025); see Appendix B). The format reward $R_{i,j}^{\text{format}} \in \{0, 1\}$ enforces structural validity, while the perception reward $R_{i,j}^{\text{perc}}$ measures how well the intermediate reasoning covers the ground-truth visual and acoustic evidence (Han et al., 2026). These three rewards are combined into a total reward, which is then normalized within the rollout group:

$$R_{i,j} = R_{i,j}^{\text{task}} + R_{i,j}^{\text{format}} + \lambda_{\text{perc}} R_{i,j}^{\text{perc}}, \tag{5}$$

$$A_{i,j} = \frac{R_{i,j} - \mu_i}{\sigma_i}. \tag{6}$$

Here $j = 1, \ldots, G$, while μ_i and σ_i denote the group mean and standard deviation of the scalar rewards. Every full-input response contributes to the policy update, preserving the general emotion-reasoning objective independently of its consistency status.

Counterfactual confidence estimation and advantage reweighting. Modality preference is informative for low-consistency samples, where visual and acoustic evidence imply divergent emotion interpretations. However, counterfactual branch construction is expensive, as it requires additional reasoning generations. To further improve efficiency and reduce noise, we therefore restrict the counterfactual branches to high-quality positive rollouts from low-consistency samples, i.e., rollouts with positive group-relative advantage $A_{i,j} > 0$, task reward $R_{i,j}^{\text{task}} \geq \kappa$, and valid format reward $R_{i,j}^{\text{format}} = 1$. Let $h_{i,j} \in \{0, 1\}$ indicate whether rollout j of sample i satisfies these conditions. For each activated rollout, modality-ablated inputs X_i^{-A} and X_i^{-V} are used to generate the reasoning traces $r_{i,j}^{-A}$ and $r_{i,j}^{-V}$, respectively. For each setting $s \in \{F, -A, -V\}$, the setting-specific input and reasoning trace are evaluated against each ground-truth emotion label $e \in Y_i$, with token-level normalization within each label and averaging across labels:

$$C_{i,j}^s = \frac{1}{|Y_i|} \sum_{e \in Y_i} \frac{1}{N_e} \sum_{n=1}^{N_e} \log \pi_{\theta_{\text{old}}} \left(e_n \mid X_i^s, r_{i,j}^s, e_{<n} \right), \tag{7}$$

where N_e is the token length of emotion label e. All three settings are assessed against the same ground-truth answer, so differences in $C_{i,j}^s$ reflect the support provided by the corresponding input and reasoning trace. The confidence reduction after deleting modality m therefore provides a comparable proxy for its contribution:

$$g_{i,j}^{m} = [C_{i,j}^{F} - C_{i,j}^{-m}]_{+}. \tag{8}$$

Here, $[x]_+ = \max(x, 0)$. A larger $g_{i,j}^{m}$ indicates a greater loss of support for the correct emotion labels after removing modality m. These contribution scores are then normalized to produce modality preference weights:

$$w_{i,j}^{m} = \frac{\exp(g_{i,j}^{m}/\tau_g)}{\sum_{u\in\{V,A\}} \exp(g_{i,j}^{u}/\tau_g)}, \tag{9}$$

where τ_g is a temperature parameter that controls the sharpness of the distribution. These weights are used to modulate the full-input advantage as follows:

$$\widetilde{A}_{i,j}^{-A} = w_{i,j}^{V} A_{i,j}, \qquad \widetilde{A}_{i,j}^{-V} = w_{i,j}^{A} A_{i,j}. \tag{10}$$

Accordingly, when audio provides stronger support for the correct prediction, the audio-retaining branch receives a larger reweighted advantage, and vice versa for vision.

Joint policy objective. For every rollout, the full-input response is optimized with the group-relative advantage $A_{i,j}$. When $h_{i,j} = 1$, the two modality-ablated reasoning traces are additionally optimized with $\widetilde{A}_{i,j}^{-A}$ and $\widetilde{A}_{i,j}^{-V}$. For a generated token sequence u under branch s, the importance ratio at token t is the current-policy likelihood divided by the rollout-policy likelihood:

$$\rho_t^s(\theta; u) = \frac{\pi_\theta(u_t \mid X_i^s, u_{<t})}{\pi_{\theta_{\text{old}}}(u_t \mid X_i^s, u_{<t})}. \tag{11}$$

Let $\bar{\rho}_t^s(\theta; u) = \text{clip}(\rho_t^s(\theta; u), 1-\epsilon_p, 1+\epsilon_p)$ denote the clipped importance ratio. For a generated token sequence u under branch s, define $\mathcal{P}_{\text{clip}}^{s}(u; B)$ as the average of the token-level clipped terms $\min\left(\rho_t^s(\theta; u)B, \bar{\rho}_t^s(\theta; u)B\right)$ over all tokens in u. Therefore, the joint objective is

$$\begin{aligned}\mathcal{J}(\theta) = \mathbb{E}_{i\sim\mathcal{D},\,\mathcal{R}_i\sim\pi_{\theta_{\text{old}}}}\Bigg[\frac{1}{G}\sum_{j=1}^{G}\Bigg(\mathcal{P}_{\text{clip}}^{F}\left(y_{i,j}^{F}; A_{i,j}\right) + \lambda_{\text{cc}} h_{i,j}\Big[\mathcal{P}_{\text{clip}}^{-A}\left(r_{i,j}^{-A}; \widetilde{A}_{i,j}^{-A}\right) \\ + \mathcal{P}_{\text{clip}}^{-V}\left(r_{i,j}^{-V}; \widetilde{A}_{i,j}^{-V}\right)\Big]\Bigg)\Bigg] - \beta\mathbb{D}_{KL}(\pi_\theta||\pi_{ref}).\end{aligned} \tag{12}$$

Here, $\mathcal{R}_i$ denotes the full-input rollout group together with any modality-ablated reasoning branches generated from it, and $\mathcal{D}$ is the training-data distribution. The first term supplies the standard full-input GRPO update to every rollout. The two auxiliary terms are activated only when $h_{i,j} = 1$ and update the reasoning tokens generated under their respective ablated branches. And $\mathbb{D}_{KL}(\pi_\theta||\pi_{ref})$ regularizes the policy toward the reference policy.

4 Experiments

4.1 Experimental Settings

Datasets. EmoTrust is trained on MER-Caption (Lian et al., 2025), using 5,000 randomly sampled examples for SFT and the remainder for GRPO-style RL. We evaluate on four benchmarks spanning multimodal sentiment analysis, categorical emotion recognition, and open-vocabulary emotion understanding. MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018) assign each video clip a positive, neutral, or negative label. MELD (Poria et al., 2019) uses seven predefined emotion categories: anger, disgust, sadness, joy, neutral, surprise, and fear, whereas OV-MERD (Lian et al.,

Table 1: Performance comparison on MOSI, MOSEI, MELD, and OV-MERD. Best results are highlighted in bold. Under OV-MERD, EW denotes the score on the full benchmark, while EW_{human} denotes the score on $\text{OV-MERD}_{\text{human}}$.

Model	MOSI		MOSEI		MELD	OV-MERD	
	ACC	WAF	ACC	WAF	Hit	EW	EW_{human}
PandaGPT (Su et al., 2023)	78.81	78.92	78.34	78.60	50.38	52.48	51.92
LLaMA-VID (Li et al., 2024)	66.52	65.48	72.23	70.41	53.98	61.98	62.04
Emotion-LLaMA (Cheng et al., 2024)	69.21	69.25	59.71	59.66	46.02	47.16	49.02
Qwen2.5-Omni-3B (Xu et al., 2025)	80.03	79.88	76.25	73.35	52.42	65.03	64.69
Qwen2.5-Omni-7B (Xu et al., 2025)	80.79	80.45	78.22	77.75	54.23	66.48	65.21
R1-Omni (Zhao et al., 2025)	64.18	62.29	48.82	45.65	41.75	54.65	52.67
AffectGPT (Lian et al., 2025)	78.20	78.24	76.77	77.16	51.11	66.30	65.17
MoSEAR (Han et al., 2025)	80.42	80.16	80.45	79.58	61.22	68.41	66.43
EmoTrust	**85.68**	**84.26**	**83.19**	**83.16**	**62.36**	**71.13**	**69.74**

2024) uses free-form natural language annotations to capture diverse and fine-grained emotions. Further details are provided in Appendix C.1.

Implementation details. We use the open-source Qwen2.5-Omni-3B (Xu et al., 2025) as the base model, which provides strong general-purpose multimodal understanding. During the cold-start SFT stage, we fine-tune the base model with LoRA for two epochs using a learning rate of 2×10^{-5}, with LoRA rank and scaling factor $r = 16$ and $\alpha = 32$, respectively. During RL, we use a learning rate of 2×10^{-6} and train for one epoch with GRPO-style sampling, generating $G = 4$ responses per prompt. The batch size is fixed to 1, with two gradient-accumulation steps throughout training. We set the KL coefficient relative to the reference model to $\beta = 0.01$. For EmoTrust, we use $\lambda_{\text{perc}} = 0.5$, $\kappa = 0.5$, and $\lambda_{\text{cc}} = 0.2$, which yielded the best overall performance in our experiments. All experiments are conducted on 8 NVIDIA A800 GPUs, each with 80 GB of memory.

Human behavior analysis. To examine whether the modality preferences learned by EmoTrust are consistent with human judgments, we collected human modality-preference annotations on OV-MERD. Three annotators viewed each video clip and judged which modality provided stronger support for the emotional content, yielding one of three categorical outcomes for each sample: visual-dominant, acoustic-dominant, or tie. The inter-annotator agreement, measured by Krippendorff's alpha, is 0.678, indicating moderate agreement. We then aggregated the three annotators' judgments by majority vote to obtain a human preference label for each sample. Based on these annotations, we further derive $\text{OV-MERD}_{\text{human}}$, a subset of OV-MERD consisting of samples with low inter-modality consistency according to human judgments. Detailed procedures are provided in the Appendix D.

4.2 COMPARATIVE EXPERIMENTAL RESULTS

Table 1 reports results on multimodal sentiment analysis, categorical emotion recognition, and open-vocabulary emotion understanding (dataset-specific metrics are given in the Appendix C.2). EmoTrust achieves the best performance on all benchmarks, consistently outperforming general-purpose multimodal LLMs (e.g., PandaGPT, LLaMA-VID, Qwen2.5-Omni, and R1-Omni), affective-task-tuned LLMs (e.g., Emotion-LLaMA and AffectGPT), and the conflict-aware emotion analysis method MoSEAR. These results show that modality-preference learning provides a stronger basis for multimodal affective reasoning than generic multimodal scaling, task-specific tuning, or bias mitigation through fusion and reweighting alone, with particularly clear gains in settings involving ambiguous or conflicting cross-modal cues.

4.3 ABLATION STUDY

Progressive validation of EmoTrust's full training pipeline. We perform an incremental ablation study from the base model to the full framework. As shown in Table 2, multi-view SFT consistently improves the base model across all benchmarks (2nd row vs 1st row), indicating that exposing the model to both full and modality-ablated views provides a stronger initialization for multimodal emo-

Table 2: Ablation study of EmoTrust on four benchmarks. EmoTrust# denotes the variant trained with standard SFT, i.e., without the modality-ablated views used in multi-view SFT. For MOSI and MOSEI, the reported metric is WAF.

Model	MOSI	MOSEI	MELD	OV-MERD$_{\text{human}}$
Base model	79.88	73.35	52.42	64.69
Multi-view SFT	81.96	79.58	59.44	65.37
Multi-view SFT + GRPO	83.14	81.68	61.32	66.68
EmoTrust#	83.23	81.94	60.86	67.42
EmoTrust	**84.26**	**83.16**	**62.36**	**69.74**

Table 3: Ablation study of modality-preference learning in EmoTrust. RevPref reverses the estimated visual and acoustic preference weights. UniPref uses uniform modality weights.

Model	MOSI	MOSEI	MELD	OV-MERD$_{\text{human}}$
w/o counterfactual branches	83.14	81.68	61.32	66.68
w/o quality gating	83.85	82.05	61.94	67.42
RevPref	82.03	80.75	59.78	64.39
UniPref	83.21	81.56	61.45	66.85
EmoTrust	**84.26**	**83.16**	**62.36**	**69.74**

tion reasoning. Adding GRPO with the total reward further yields consistent gains (3rd row vs 2nd row), showing the benefit of optimizing all samples with joint task, format, and perception supervision. Finally, introducing the counterfactual branches with modality-preference learning brings another clear improvement and leads to the best overall performance (5th row vs 3rd row). This gain is especially pronounced on samples with potential cross-modal conflicts (*i.e.,* OV-MERD$_{\text{human}}$), validating the effectiveness of conflict-conditioned policy optimization in teaching the model which modality to rely on under cross-modal inconsistency. We also observe that removing the modality-ablated views during SFT degrades the effectiveness of the subsequent counterfactual optimization (4th row vs 5th row), likely because the model is not sufficiently adapted to the ablated-input distribution encountered in later training.

Effectiveness of counterfactual modality-preference learning. As illustrated in Table 3, EmoTrust learns modality preferences on potentially conflicting samples through counterfactual comparisons. The results in Table 3 confirm the effectiveness of this design. Removing counterfactual branches consistently degrades performance across all benchmarks, suggesting that RL on full multimodal inputs alone is insufficient to capture modality-specific contributions under emotional conflict. Removing quality gating also hurts performance, indicating that counterfactual advantage reweighting is most effective when restricted to high-quality positive rollouts, while low-quality branches may introduce noisy supervision. We further examine two variants of the preference weighting mechanism. Reversing the estimated visual and acoustic preference weights causes the largest performance drop, showing that the learned preference signals capture meaningful modality contributions. Using uniform modality weights also underperforms EmoTrust, suggesting that equal treatment of the two modalities is suboptimal when their emotional evidence is asymmetric. Overall, these results demonstrate that high-quality counterfactual reweighting helps EmoTrust learn effective modality preferences, leading to more reliable emotion prediction under cross-modal conflict.

4.4 VISUALIZATION

Performance under different levels of cross-modal consistency. To better understand where EmoTrust is most effective, we compute an inter-modal consistency score S_{sim} for each OV-MERD sample and group the data into five bins accordingly (see Appendix D). As shown in Figure 4A, EmoTrust consistently outperforms all progressive variants across all bins, while its advantage is most pronounced on low-consistency samples (highlighted by the gray background), where the emotional evidence from different modalities is more likely to be conflicting or misleading. Figure 4B further highlights this pattern by reporting the gain of EmoTrust over the strongest competing variant in each bin: the improvement is largest in the low-consistency regime and gradually diminishes as S_{sim} increases. This trend suggests that EmoTrust improves the model's ability to resolve cross-

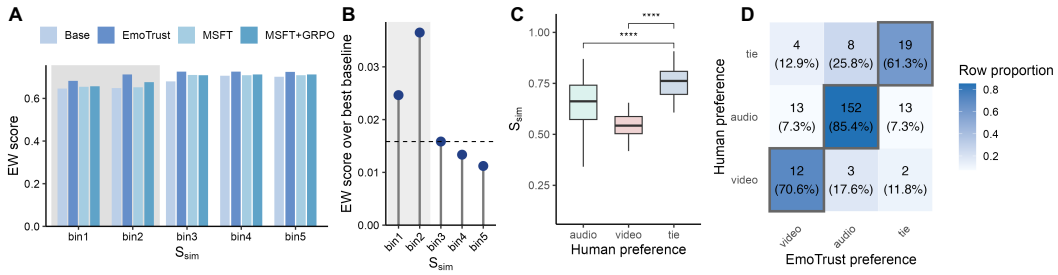


Figure 4: **Visualization of consistency and modality preference patterns in EmoTrust. A**. Performance comparison across five bins of inter-modal consistency score S_{sim}, showing that EmoTrust consistently outperforms progressive variants. **B**. Performance gain of EmoTrust over the strongest competing variant in each consistency bin. **C**. Distribution of S_{sim} under different human preference labels, indicating alignment between the estimated consistency and human perception. **D**. Agreement between model-inferred modality preference and human judgments.

modal disagreement by learning which modality should be trusted when multimodal signals are inconsistent.

Alignment between estimated consistency and human perception. We further examine whether the proposed cue-grounded cross-modal emotion consistency score S_{sim} aligns with human perception of cross-modal consistency. To this end, we group OV-MERD samples by human preference labels (*i.e.,* acoustic-dominant, visual-dominant, or tie) and compare the distribution of S_{sim} across groups. As shown in Figure 4C, *tie* samples exhibit noticeably higher S_{sim} values, whereas modality-dominant samples tend to have lower scores, indicating weaker perceived agreement between modalities. This difference is significant under the Kruskal–Wallis test ($H(2) = 32.67, p < 0.001$), and post-hoc pairwise comparisons further show that the *tie* group differs significantly from both acoustic-dominant and visual-dominant samples ($p < 0.001$). These results suggest that S_{sim} captures cross-modal consistency in a way that aligns well with human judgments, supporting the validity of our consistency estimation design.

Alignment between learned modality preference and human judgments. We further assess whether EmoTrust learns behaviorally meaningful modality preferences. Specifically, we compare model-side preference labels inferred from the learned modality weights, which take one of three values (visual-dominant, acoustic-dominant, or tie; see Appendix D.3), against human annotations on OV-MERD. As shown in Figure 4D, EmoTrust achieves strong agreement with human judgments, reaching 80.97% accuracy and 82.16% WAF. In particular, samples with a clearly perceived dominant modality are largely assigned to the same dominant side by EmoTrust, while *tie* cases are also identified with reasonable fidelity. These results suggest that the proposed counterfactual preference learning not only improves predictive performance, but also captures modality-reliance patterns that align well with human judgments under emotional conflict.

5 Conclusion

In this work, we introduced EmoTrust, a framework for multimodal emotion reasoning that explicitly addresses cross-modal emotional conflict by learning which modality should be trusted more. EmoTrust integrates multi-view supervised fine-tuning, cue-grounded cross-modal consistency estimation, and conflict-conditioned policy optimization to promote reasoning that is grounded in the more informative modality under disagreement. Extensive experiments across multiple benchmarks show that EmoTrust consistently outperforms strong baselines, with especially large gains on low-consistency samples where modality conflict is more pronounced. Furthermore, the learned modality preferences align closely with human judgments, suggesting that EmoTrust improves not only predictive performance but also the interpretability and behavioral plausibility of multimodal emotion reasoning.

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. *Advances in Neural Information Processing Systems*, 37: 110805–110853, 2024.

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 *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 745–755. IEEE, 2026.

Zhiyuan Han, Beier Zhu, Yanlong Xu, Peipei Song, and Xun Yang. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 5528–5537, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754856.

Zhiyuan Han, Beier Zhu, Wenwen Tong, Pengyang Shao, Peipei Song, Xinyi Wang, Jiangnan Chen, Lewei Lu, and Xun Yang. Omni-perception policy optimization for multimodal emotion reasoning, 2026. URL https://arxiv.org/abs/2606.25325.

Jaewook Lee, Yeajin Jang, Hongjin Kim, Woojin Lee, and Harksoo Kim. Analyzing key factors influencing emotion prediction performance of vllms in conversational contexts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5801–5816, 2024.

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, 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, 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. *arXiv preprint arXiv:2501.16566*, 2025.

Haiwei Ma and Svetlana Yarosh. A review of affective computing research based on function-component-representation framework. *IEEE Transactions on Affective Computing*, 14(2):1655–1674, 2021.

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.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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.

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 *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Emmeke A. Veltmeijer, Charlotte Gerritsen, and Koen V. Hindriks. Automatic emotion recognition for groups: A review. *IEEE Transactions on Affective Computing*, 14(1):89–107, 2023. doi: 10.1109/TAFFC.2021.3065726.

Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance, 2024.

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, 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, 2025.

Qinfu Xu, Liyuan Pan, Yiwei Wei, Shaozu Yuan, Jiaqi Chen, and Tianyu Liu. Emothinker: Advancing visual-acoustic emotion analysis via structural token selection and chain-of-thought reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1672–1682, June 2026.

Kaixiang Yang, Yifan Luo, Zongyan Zhang, C. L. Philip Chen, and Tong Zhang. Multimodal affect perception with large language model enhancement network. *IEEE Transactions on Affective Computing*, 17(1):132–146, 2026. doi: 10.1109/TAFFC.2025.3607727.

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 (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

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 Syst. Appl.*, 237(PC), March 2024a. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.121692.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Yedi Zhang, Peter E. Latham, and Andrew Saxe. Understanding unimodal bias in multimodal deep linear networks, 2024b.

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

A Details of Cue-Grounded Cross-Modal Emotion Consistency Estimation

In this section, we describe the cue-to-emotion mapping procedure used in EmoTrust. As illustrated in Figure 3, the procedure consists of multimodal evidence extraction, emotion prototype construction, prototype-aggregated similarity computation, and consistency/conflict evaluation. Throughout this section, i indexes samples, $e_j \in \mathcal{E}$ denotes an emotion in the global vocabulary, and $m \in \{V, A\}$ denotes the visual or acoustic modality.

A.1 Multimodal Evidence Extraction

As illustrated in Figure 3(a), each video-audio input is paired with a ground-truth reasoning paragraph that summarizes the affective evidence exhibited by the character. From this paragraph, we use GPT-5.4 (Singh et al., 2025) to extract two modality-specific cue sets, C_i^V and C_i^A, corresponding to the visual and acoustic modalities, respectively. Visual cues capture explicitly described facial expressions, gaze, posture, and body movements, whereas acoustic cues capture explicitly described vocal properties, such as tone, tension, and speaking rate. This step only restructures evidence already present in the paragraph and does not introduce any additional observations.

Each cue is represented as a short, self-contained statement with its modality assignment preserved. Importantly, negation and uncertainty must also be retained. For example, "no visible indications of strong emotions" should not be rewritten as "strong emotions." As shown in Figure 3(a), "neutral expression" is treated as visual evidence, whereas "tense and strained voice" and "speaking quickly" are treated as acoustic evidence. Their potentially different affective implications are preserved for subsequent comparison rather than resolved at the extraction stage.

The extraction constraints are operationalized using the following prompt template.

Prompt template for modality-specific cue extraction

Given the following reasoning paragraph, extract the explicitly stated visual and acoustic cues concerning the main character. Use only information in the paragraph and do not introduce new evidence. Return short, self-contained cues. Preserve negation, uncertainty, and the original modality attribution. Visual cues concern appearance and visible behavior; acoustic cues concern vocal delivery and audible properties. Do not infer acoustic cues from visual descriptions or visual cues from acoustic descriptions. Do not convert subtitle meaning or a joint emotional conclusion into modality-specific evidence. Return a JSON object with two arrays, `visual_cues` and `acoustic_cues`. Use an empty array when no cue is explicitly available for a modality.

Reasoning paragraph: {`reasoning_paragraph`}

A.2 Emotion Prototype Construction

Global emotion vocabulary construction. We construct the global emotion vocabulary $\mathcal{E}$ using the training split of MER-Caption (Lian et al., 2025). Specifically, we first collect all emotion labels appearing in the open-form annotations. Because these labels exhibit a long-tailed distribution, many occur only rarely. To reduce sparsity and exclude categories with insufficient support, we discard labels that appear fewer than 100 times. We then merge synonymous labels into a shared canonical category; for example, glad, happy, and joyful are all mapped to happy. This normalization avoids treating semantically equivalent labels as distinct emotion categories. After frequency filtering and synonym consolidation, the resulting global vocabulary $\mathcal{E}$ contains 226 emotion terms.

Figure 3(b) expands each emotion $e_j \in \mathcal{E}$ into a set of textual prototypes $P(e_j)$. This set contains the canonical emotion label, synonyms, and short descriptions of its meaning and typical manifestations. The fields illustrated in the figure are `synonyms`, `brief_definition`, and `typical_behavior`. These complementary descriptions connect concrete observations to an emotion even when the cue and its canonical label have little lexical overlap. Prototype construction is performed at the vocabulary level, so that diverse textual realizations of the same emotion are mapped into a shared semantic space, which can then be used as a common reference for comparing both visual and acoustic cues across samples. For example, Figure 3(b) expands "worry" with

synonyms such as "worried" and "fretting," a definition involving repetitive concern about possible problems or negative outcomes, and descriptions of behaviors such as a furrowed brow, a tense voice, or hesitant pauses. These are candidate manifestations of an emotion, not additional observations about the current sample. The distinction ensures that enriching the prototype bank does not alter the evidence extracted in Section A.1.

The prototype expansion can be expressed with the following prompt template.

Prompt template for emotion prototype expansion

For the given emotion label, provide synonyms, a brief definition, and a short description of typical observable behavior, including visual or vocal manifestations where appropriate. Keep all descriptions specific to the given emotion. Describe behaviors as possible manifestations, not necessary conditions. Do not describe any particular video or invent sample-specific evidence. Return a JSON object with the fields `emotion`, `synonyms`, `brief_definition`, and `typical_behavior`.

Emotion label: {`emotion_label`}

We encode the resulting prototype texts with the same Qwen3-Embedding encoder used for cues.

A.3 Prototype-Aggregated Similarity Computation

Figure 3(c) maps each cue c to a soft support distribution over $\mathcal{E}$. Let $\mathbf{e}(c)$ be its Qwen3-Embedding representation and $\mathbf{e}(p)$ the representation of a prototype text $p \in P(e_j)$. Both are encoded in the same embedding space. Their cosine similarity is

$$\cos\big(\mathbf{e}(c), \mathbf{e}(p)\big) = \frac{\mathbf{e}(c)^\top \mathbf{e}(p)}{\|\mathbf{e}(c)\|_2 \|\mathbf{e}(p)\|_2}. \tag{13}$$

For a given emotion e_j, we first compute the cosine similarity between c and every prototype in $P(e_j)$, and rank these prototypes accordingly. We then retain the top-r most similar prototypes, where $r = 3$ in all experiments. Denoting this selected subset by $\mathrm{Top}_r(c, P(e_j))$, we define the support score from cue c to emotion e_j as

$$s(c, e_j) = \frac{1}{r} \sum_{p \in \mathrm{Top}_r(c, P(e_j))} \cos\big(\mathbf{e}(c), \mathbf{e}(p)\big). \tag{14}$$

This design allows each cue to match an emotion through its most relevant prototypes, while reducing the influence of less informative or noisy prototypes.

We then normalize the emotion support scores using a temperature-scaled softmax:

$$q(e_j \mid c) = \frac{\exp\big(s(c, e_j)/\tau_e\big)}{\sum_{k=1}^{|\mathcal{E}|} \exp\big(s(c, e_k)/\tau_e\big)}. \tag{15}$$

Here $q(e_j \mid c) \geq 0$ and $\sum_{j=1}^{|\mathcal{E}|} q(e_j \mid c) = 1$. A cue can support multiple related emotions with different strengths. For example, a tense voice may support more than one affective interpretation instead of being assigned a single hard label. Every cue uses the same emotion order, enabling direct aggregation and comparison in Figure 3(d).

A.4 Consistency and Conflict Evaluation

We first aggregate the cue-level support distributions separately for each modality, as illustrated in Figure 3(d):

$$p_i^m(e_j) = \sum_{c \in C_i^m} w_{i,c}^m q(e_j \mid c), \qquad \sum_{c \in C_i^m} w_{i,c}^m = 1, \qquad m \in \{V, A\}. \tag{16}$$

We use uniform weights $w_{i,c}^m = 1/|C_i^m|$ in this paper. Consequently, a modality with more cues does not receive greater total mass merely because of its cue count. The resulting p_i^V and p_i^A are probability distributions over the same ordered vocabulary.

We measure their discrepancy using Jensen-Shannon divergence:

$$D_{\mathrm{JS}}(p_i^V \| p_i^A) = \frac{1}{2} D_{\mathrm{KL}}\left(p_i^V \middle\| \frac{p_i^V + p_i^A}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(p_i^A \middle\| \frac{p_i^V + p_i^A}{2}\right), \tag{17}$$

$$D_{\mathrm{KL}}(p\|q) = \sum_{j=1}^{|\mathcal{E}|} p(e_j) \log \frac{p(e_j)}{q(e_j)}. \tag{18}$$

The normalized consistency score is therefore

$$S_i = 1 - \frac{D_{\mathrm{JS}}(p_i^V \| p_i^A)}{\log 2}, \qquad S_i \in [0, 1]. \tag{19}$$

Identical modality-level distributions yield $S_i = 1$, whereas increasingly different distributions yield lower consistency. Samples with $S_i > \tau_c$ are treated as consistent, and samples with $S_i < \tau_c$ are identified as exhibiting potential cross-modal emotion conflict. Accordingly, S_i is used to identify low-consistency samples for subsequent conflict-conditioned learning.

B Details of Reward Functions

B.1 Details of Task Reward

Consistent with AffectGPT (Lian et al., 2025), we define the task reward as an Emotion Wheel (EW)-based score computed between the predicted emotion set and the ground-truth label set. Since the model outputs open-vocabulary free-form emotion words, we do not use exact string matching. Instead, both the prediction and the reference labels are first normalized through a multi-stage mapping procedure, and the reward is then computed as a set-level F1 score under multiple emotion wheels. Formally, let $\hat{Y}_{i,j}^F$ denote the predicted emotion set for the j-th rollout of sample i, and let Y_i denote the corresponding ground-truth emotion set. To reduce surface-form variation, each emotion word is normalized in three stages.

First, we apply morphological normalization, denoted by $F_{l1}(\cdot)$, which maps inflected or derived forms to their base emotion form. Second, we apply synonym normalization, denoted by $F_{l2}(\cdot)$, which maps semantically equivalent emotion words to a shared canonical form. Third, for each emotion wheel w_k, we apply an emotion-wheel grouping function $F_{l3}^{w_k}(\cdot)$, which maps fine-grained outer-ring emotions to their corresponding inner-ring base emotions.

Combining the above three steps, the final mapping of an emotion word x under the k-th emotion wheel is defined as

$$G_{w_k}(x) = F_{l3}^{w_k}\big(F_{l2}(F_{l1}(x))\big), \quad k \in [1, K]. \tag{20}$$

For an emotion set S, the mapped set under wheel w_k is

$$G_{w_k}(S) = \{\, G_{w_k}(x) \mid x \in S \,\}, \tag{21}$$

where duplicate items are removed after mapping.

After normalization, we compute a set-level precision, recall, and F1 score for each emotion wheel. Specifically, under wheel w_k, precision and recall are defined as

$$P_k^{(i,j)} = \frac{\left|G_{w_k}(\hat{Y}_{i,j}^F) \cap G_{w_k}(Y_i)\right|}{\left|G_{w_k}(\hat{Y}_{i,j}^F)\right|}, \tag{22}$$

$$R_k^{(i,j)} = \frac{\left|G_{w_k}(\hat{Y}_{i,j}^F) \cap G_{w_k}(Y_i)\right|}{|G_{w_k}(Y_i)|}. \tag{23}$$

The corresponding F1 score is

$$F_{s,k}^{(i,j)} = \frac{2P_k^{(i,j)} R_k^{(i,j)}}{P_k^{(i,j)} + R_k^{(i,j)}}. \tag{24}$$

Finally, the EW score is obtained by averaging the F1 scores across all emotion wheels:

$$EW_{i,j} = \frac{1}{K} \sum_{k=1}^{K} F_{s,k}^{(i,j)}. \tag{25}$$

In our implementation, we follow Lian et al. (2025) and use $K = 5$ emotion wheels. The task reward is then defined as

$$R_{i,j}^{\text{task}} = EW_{i,j}. \tag{26}$$

This reward provides a normalized measure of prediction quality that is robust to lexical variation and better reflects the structural similarity between the predicted and ground-truth emotion sets than exact-match evaluation.

B.2 Details of format Reward

The format reward is a binary reward used to enforce structural validity of the model output. Specifically, it checks whether the response contains both required fields with valid XML-style tags, namely `<think>...</think>` and `<answer>...</answer>`.

Formally, the format reward is defined as

$$R_{i,j}^{\text{format}} = \begin{cases} 1, & \text{if the output contains both required fields with valid tags,} \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

B.3 Details of perception Reward

Following (Han et al., 2026), we use a perception reward to measure whether the generated reasoning covers the ground-truth multimodal evidence. Specifically, for each sample i, we pre-extract two sets of perceptual cues from the ground-truth reasoning, $\mathcal{C}_{V,i}$ and $\mathcal{C}_{A,i}$, corresponding to visual and audio cues, respectively; see Appendix A.1. Given the reasoning trace generated in the j-th full-input rollout, we segment it into a set of clauses, denoted by $\mathcal{U}_{i,j} = \{u_{i,j,1}, \ldots, u_{i,j,K}\}$. For each cue c, we compute its maximum cosine similarity with all generated clauses using the same Qwen3-Embedding model $\mathbf{e}(\cdot)$:

$$M_{i,j}(c) = \max_{k} \cos\big(\mathbf{e}(c), \mathbf{e}(u_{i,j,k})\big). \tag{28}$$

We then convert the similarity score into a soft cue-hit score using a threshold δ:

$$h_{i,j}(c) = \frac{\max(M_{i,j}(c) - \delta, 0)}{1 - \delta}, \tag{29}$$

where we set $\delta = 0.5$ in our implementation. This transformation maps similarities below the threshold to zero and linearly rescales similarities above the threshold to the range $[0, 1]$.

Based on the cue-hit scores, we compute the coverage rewards for the visual and audio modalities:

$$R_{V,i,j}^{\text{cov}} = \frac{1}{|\mathcal{C}_{V,i}|} \sum_{c \in \mathcal{C}_{V,i}} h_{i,j}(c), \tag{30}$$

$$R_{A,i,j}^{\text{cov}} = \frac{1}{|\mathcal{C}_{A,i}|} \sum_{c \in \mathcal{C}_{A,i}} h_{i,j}(c). \tag{31}$$

The perception reward is computed as follows:

$$R_{i,j}^{\text{perc}} = \frac{1}{|\mathcal{M}_i|} \sum_{m \in \mathcal{M}_i} R_{m,i,j}^{\text{cov}}, \tag{32}$$

where $\mathcal{M}_i \subseteq \{V, A\}$.

C Datasets and Evaluation Protocols

C.1 Datasets

CMU-MOSI and CMU-MOSEI are standard benchmarks for multimodal sentiment analysis, where each sample consists of a short opinion-oriented video segment paired with language, acoustic, and visual signals. CMU-MOSI contains 2,199 annotated video clips, which are split into 1,284/229/686 samples for training, validation, and testing, respectively. CMU-MOSEI is a larger-scale benchmark containing 22,856 annotated video clips collected from YouTube movie reviews, with the standard split of 16,326/1,871/4,659 samples for training, validation, and testing, respectively. For both datasets, each sample is annotated with a sentiment score in the range $[-3, 3]$. Following common practice, we convert these annotations into three sentiment polarities: negative for scores smaller than 0, neutral for scores equal to 0, and positive for scores greater than 0. We evaluate our method on the standard test set of each dataset.

MELD is a benchmark for multimodal conversational emotion recognition, constructed by extending the EmotionLines dataset with aligned audio and visual modalities in addition to text. It contains more than 1,400 dialogues and approximately 13,000 utterances collected from the *Friends* TV series, involving multiple speakers in conversational settings. Each utterance is annotated with one of seven emotion categories: *anger, disgust, sadness, joy, neutral, surprise, and fear*. We follow the standard split with 9,989/1,109/2,610 utterances for training, validation, and testing, respectively. We evaluate our method on the standard test set of this dataset.

OV-MERD is a benchmark for open-vocabulary multimodal emotion understanding, designed to move beyond fixed emotion taxonomies by using free-form natural language annotations. The dataset contains 332 short video clips with aligned audio and subtitle modalities, enabling evaluation of fine-grained emotion understanding under multimodal inputs. Unlike closed-set benchmarks, OV-MERD characterizes emotions with unconstrained descriptors that capture nuanced and culturally diverse affective states. This makes it particularly suitable for evaluating a model's ability to reason about complex emotions in open-vocabulary settings.

C.2 Dataset-specific Evaluation Protocols

Our evaluation follows prior work (Lian et al., 2025). Given a multimodal input, the model first generates a response, from which an emotion set is then extracted and scored using the task-specific metric. Table 1 compares all models on multimodal sentiment analysis, categorical emotion recognition, and open-vocabulary emotion understanding. To ensure a fair comparison across models with different output styles, we adopt a unified evaluation pipeline before applying task-specific scoring rules.

First, each model is evaluated in a generation setting. Given the video, audio, and available transcript of an input sample, the model produces a free-form description of the main character's emotional state. The response may include one or more emotion labels together with supporting explanations. For all models and all datasets, we use the same response-generation prompt:

Prompt for response generation

Please infer the emotional state of the character.

Second, the generated response is converted into a standardized emotion-set representation using a shared language model, implemented with Qwen3.8. This step removes differences in response format and allows all models to be compared in a common label space. Let $\hat{Y}_i$ denote the extracted prediction for sample i, and let Y_i denote the corresponding reference emotion set. Multiple predicted emotions are allowed, and an empty list indicates that no identifiable emotion is extracted. The extraction prompt is:

Prompt for emotion extraction

Please assume the role of an expert in the field of emotions. We provide clues that may be related to the emotions of the characters. Based on the provided clues, please identify the emotional states of the main character. Please separate different emotional categories with commas and output only the clearly identifiable emotional categories in a list format. If none are identified, please output an empty list.

After this shared preprocessing, different benchmark tasks apply different postprocessing and scoring rules. For sentiment analysis, the extracted emotion set is further converted into sentiment polarity. For categorical emotion recognition, the extracted set is compared against the reference basic-emotion annotation. For open-vocabulary emotion understanding, the predicted and reference emotion sets are compared directly using the emotion-wheel-based metric defined in Appendix B.1.

Sentiment analysis. For multimodal sentiment analysis, the extracted emotions are further mapped to one of three sentiment polarities: positive, negative, or neutral. This conversion is performed using the following prompt:

Prompt for sentiment prediction

Please act as an expert in the field of emotions. We provide a few words to describe the emotions of a character. Please choose the most likely sentiment from the given candidates: [positive, negative, neutral].

Although the original sentiment annotations are first converted into three polarities, the final evaluation follows the standard binary protocol commonly used for MOSI and MOSEI. Specifically, samples with neutral ground-truth sentiment (i.e., sentiment score equal to 0) are excluded from evaluation, and the remaining samples are labeled according to the sign of their annotated sentiment score. We therefore assign the reference binary label $b_i = 1$ to positive samples and $b_i = 0$ to negative samples. On the prediction side, the polarity-conversion stage may still output one of three classes, but for binary evaluation we map *positive* to $\hat{b}_i = 1$ and *negative* to $\hat{b}_i = 0$. Let $\mathcal{I}_{\text{sent}}$ denote the resulting evaluation set. Accuracy is then computed as

$$\text{ACC} = \frac{1}{|\mathcal{I}_{\text{sent}}|} \sum_{i \in \mathcal{I}_{\text{sent}}} \mathbf{1}[b_i = \hat{b}_i]. \tag{33}$$

We also report weighted F1 (WAF), which accounts for the relative frequency of each reference class:

$$\text{WAF} = \sum_{b \in \{0,1\}} \frac{n_b}{|\mathcal{I}_{\text{sent}}|} \frac{2\text{TP}_b}{2\text{TP}_b + \text{FP}_b + \text{FN}_b}. \tag{34}$$

Here n_b is the number of reference samples in class b, and TP_b, FP_b, and FN_b are its true-positive, false-positive, and false-negative counts. WAF is the primary metric for sentiment analysis, while ACC is reported as a complementary measure. In Table 1, these metrics are reported for MOSI and MOSEI.

Categorical emotion recognition. For MELD, each sample is annotated with a single basic-emotion label, which we represent as a singleton set $Y_i = \{y_i\}$ for consistency with the set-based evaluation framework. We reuse the same emotion normalization and wheel-specific mapping $G_{w_k}(\cdot)$ defined for the EW score, but replace set-level F1 with a coverage-based hit criterion. The final score is the average hit rate across the K emotion wheels:

$$\text{HIT} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbf{1}\Big[G_{w_k}(Y_i) \cap G_{w_k}(\hat{Y}_i) \neq \varnothing \Big] \right), \tag{35}$$

where $\mathcal{I}$ denotes the evaluation set and $\mathbf{1}[\cdot]$ is the indicator function. Intuitively, a prediction is counted as correct under wheel w_k if the mapped predicted set overlaps with the mapped reference set. Unlike set-level F1, HIT only measures whether the reference emotion is covered by the prediction and does not penalize additional predicted emotions once a valid match is found. In Table 1, we report HIT for MELD.

Table 4: Pairwise agreement between annotators for modality-preference labeling on OV-MERD.

Rater A	Rater B	Agreement Rate	Cohen's κ	p-value
Annotator 1	Annotator 2	0.899	0.749	< 0.001
Annotator 1	Annotator 3	0.833	0.597	< 0.001
Annotator 2	Annotator 3	0.877	0.689	< 0.001

Open-vocabulary emotion understanding. For open-vocabulary emotion understanding, evaluation is performed directly on the predicted and reference emotion sets. To reduce mismatches caused by lexical variation, both sets undergo word-form normalization, synonym mapping, and grouping into the core categories of the emotion wheel. The final score is computed using the emotion-wheel metric EW defined in Appendix B.1. In Table 1, we report both the score on the full benchmark (EW) and the score on OV-MERD$_{\text{human}}$ (EW$_{\text{human}}$).

D Human and Model Modality Preference Analysis

D.1 Human Modality Preference Annotation

To assess whether the modality preferences learned by EmoTrust align with human judgments, we collected human modality-preference annotations on OV-MERD. Three annotators independently reviewed each video clip together with its original audio track. For every sample, annotators were presented with its ground-truth emotion labels one by one and were asked to judge, for each label, how much the visual and acoustic modalities contributed to the expression of that emotion. To minimize potential bias introduced by a fixed presentation order, the emotion labels associated with each sample were randomly permuted separately for each annotator. Specifically, for each emotion label, the annotator was required to allocate a total of 100 points between the two modalities, yielding a visual contribution score v_i and an acoustic contribution score a_i, subject to the constraint that $v_i + a_i = 100$. This forced-allocation protocol was used to explicitly capture the annotator's relative belief about which modality provided stronger support for the given emotional content. In addition, the annotator assigned a confidence score from 1 to 5 to each judgment, where a higher score indicates greater confidence in the assigned modality contributions.

For a sample with multiple ground-truth emotion labels, we first computed a contribution difference for each label as $d_i = v_i - a_i$. We then aggregated these label-level differences into a sample-level modality preference score using confidence-weighted averaging, $D = \frac{\sum_i c_i d_i}{\sum_i c_i}$, where c_i denotes the confidence score assigned to the i-th emotion label. This score summarizes, for a given annotator, the overall relative contribution of the visual and acoustic modalities to the emotional content of the sample, while assigning greater influence to judgments made with higher confidence. Based on the sample-level score D, we assigned a categorical modality-preference label to each annotator–sample pair using a margin threshold m: *visual-dominant* if $D > m$, *acoustic-dominant* if $D < -m$, and *tie* otherwise. The margin defines a small indifference region around parity, preventing minor visual–acoustic differences from being over-interpreted as clear modality dominance. In our experiments, we set $m = 5$, and found the main conclusions to be stable under nearby choices of m.

We then assessed inter-annotator consistency over these categorical labels. In the rare case where the three annotators gave completely conflicting labels—*i.e.*, one *visual-dominant*, one *acoustic-dominant*, and one *tie*—no majority label could be established, and the sample was excluded from the human-preference analysis. This occurred for only one sample in the dataset. For the remaining samples, we aggregated the three annotations by majority vote to obtain the final human modality-preference label.

To quantify annotation reliability, we computed Krippendorff's alpha over the three annotators, obtaining a value of 0.678, which indicates moderate agreement. We further report pairwise agreement statistics in Table 4, including percent agreement and Cohen's κ for each annotator pair. These results show that annotator agreement is consistently and significantly above chance level.

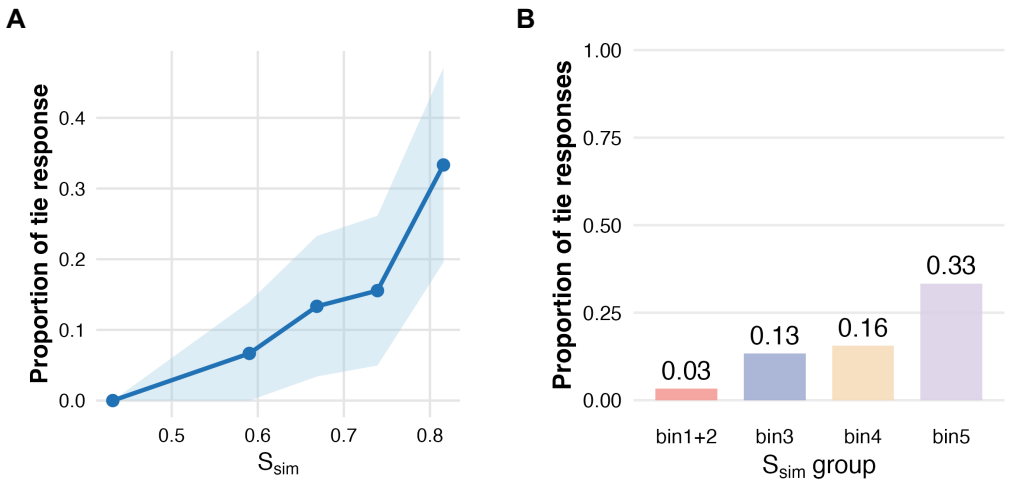


Figure 5: Human *tie* judgments as a function of estimated visual-acoustic consistency S_{sim}. **A.** Proportion of samples whose final human modality-preference label is *tie* across five equal-sized bins of S_{sim}, ordered from low to high estimated visual-acoustic consistency. Shaded areas indicate confidence intervals. **B.** Bin 1 and Bin 2 merged as the low-S_{sim} reference group. Relative to this reference group, Bin 3, Bin 4, and Bin 5 show significantly higher probabilities of *tie* judgments.

D.2 Construction of OV-MERD$_{\text{HUMAN}}$

We further construct OV-MERD$_{\text{human}}$, a human-validated subset of OV-MERD consisting of samples with low inter-modality consistency. The construction is based on the estimated cross-modal emotional consistency score S_{sim}, which measures how well the emotional signals conveyed by the visual and acoustic modalities agree with each other. A larger S_{sim} indicates stronger visual-acoustic consistency, whereas a smaller value suggests that the two modalities provide less aligned emotional evidence.

We sort all OV-MERD samples by their S_{sim} values and divide them into five equal-sized bins, ordered from low to high consistency. For each bin, we compute the proportion of samples whose final human modality-preference label is *tie*. The resulting trend is shown in Fig. 5A. We observe that the probability of a *tie* judgment generally increases with S_{sim}, which is consistent with the interpretation that larger S_{sim} corresponds to stronger agreement between the two modalities in human perception. More specifically, the first two bins exhibit similarly low proportions of *tie* responses, indicating that samples in these bins are consistently perceived as having relatively weak visual-acoustic agreement. In contrast, starting from the third bin, the proportion of *tie* judgments increases noticeably as S_{sim} becomes larger. To assess whether this increasing trend was statistically reliable, we fit a logistic regression model with *tie* judgment as the binary outcome and the S_{sim}-based group assignment as the predictor. Bin 1 and Bin 2 were merged into a single low-S_{sim} reference group, which was then contrasted with Bin 3, Bin 4, and Bin 5 (see Fig. 5B). Relative to the merged Bin 1-2 group, Bin 3 showed a significantly higher probability of *tie* judgments ($z = 2.056$, $p = 0.0397$), Bin 4 also showed a significantly higher probability ($z = 2.353$, $p = 0.0186$), and Bin 5 exhibited the strongest increase ($z = 4.027$, $p < 0.001$).

Based on the above observations, we define samples in Bin 1 and Bin 2 as *low-consistency samples*, and collect them to form OV-MERD$_{\text{human}}$. In other words, OV-MERD$_{\text{human}}$ is a human-validated subset of OV-MERD whose members are characterized by relatively low estimated visual-acoustic emotional consistency and are also less likely to be judged by annotators as modality-balanced.

D.3 Model-Side Modality Preference Inference

To examine whether the modality reliance learned by EmoTrust is aligned with human judgments, we derive model-side modality-preference labels from the preference weights introduced in the main

Method	κ	z	Acc. (%)	Macro F1 (%)	WAF (%)
EmoTrust	0.526	10.5	80.97	66.60	82.16
MSFT+GRPO	0.316	6.28	72.56	51.17	74.30
MSFT	0.212	4.00	73.00	41.40	72.25
Base	0.154	3.06	67.25	39.84	68.68

Table 5: Agreement between model-inferred modality preferences and human judgments on OV-MERD for EmoTrust and its progressive variants. Higher is better for all metrics.

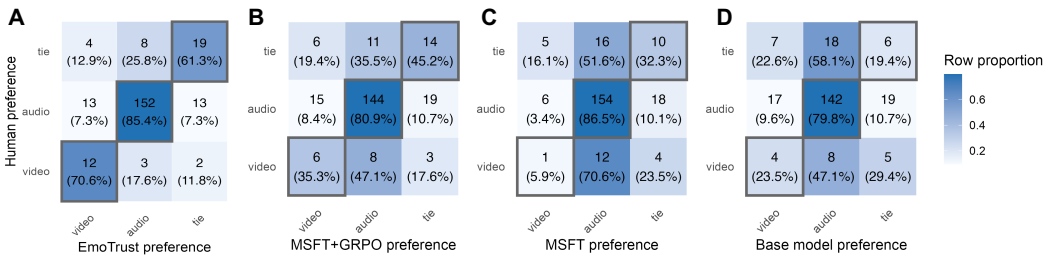


Figure 6: Confusion matrices for human-model modality-preference agreement on OV-MERD. (A) EmoTrust. (B) MSFT+GRPO. (C) MSFT. (D) Base model. Rows denote human preference labels, and columns denote model-inferred preference labels.

methodology. As described in Section 3, EmoTrust estimates the relative contribution of the visual and acoustic modalities through the counterfactual confidence reductions induced by modality deletion. Concretely, for an activated rollout, the contribution scores g^V and g^A are converted into normalized modality-preference weights w^V and w^A, where $w^V + w^A = 1$. Intuitively, a larger w^V indicates that removing visual evidence causes a greater loss of support for the correct emotion labels, implying stronger model reliance on the visual modality; w^A is interpreted analogously for the acoustic modality.

For the purpose of comparison with human annotations, we convert these continuous weights into the same three-way categorical space used in the human-preference analysis, namely *visual-dominant*, *acoustic-dominant*, and *tie*. Specifically, we define the model-side preference score as the weight difference $D_{\text{model}} = w^V - w^A$. A sample is labeled as *visual-dominant* if $D_{\text{model}} > m_{\text{model}}$, *acoustic-dominant* if $D_{\text{model}} < -m_{\text{model}}$, and *tie* otherwise. Since the human annotations were obtained under a forced-allocation protocol with $v_i + a_i = 100$ and a margin threshold of $m = 5$, we set the model-side threshold proportionally to $m_{\text{model}} = 0.05$, so that the indifference region occupies the same relative width under the normalization $w^V + w^A = 1$.

D.4 Comparison with Progressive Variants on Human-Judgment Alignment

To complement the main-text results on alignment with human modality judgments, we further compare EmoTrust with its progressive variants on OV-MERD. Specifically, we evaluate the agreement between model-side modality-preference labels and human annotations under the same three-way categorization used in the main analysis, namely *visual-dominant*, *acoustic-dominant*, and *tie*. The compared systems include the full EmoTrust model, the GRPO variant, the SFT variant, and the base Qwen model.

Table 5 summarizes the overall agreement results. EmoTrust achieves the best performance on all reported metrics, substantially outperforming the progressive variants and the base model. In particular, EmoTrust reaches 80.97% accuracy, 66.60% macro F1 and 82.16% WAF, while also obtaining the highest inter-rater agreement score ($\kappa = 0.526$). These results provide additional evidence that the proposed preference-learning mechanism yields modality-reliance patterns that are more consistent with human judgments than those produced by simpler training variants.

To further examine how these models differ, Figure 6 reveals the error patterns of different models at the class level. EmoTrust shows the strongest diagonal concentration overall, indicating better



Label: impatient, excited, angry, dissatisfied

MSFT+GRPO

<think>In the video, a man is lying on a bed with a pillow under his head, and he appears to be in a state of distress or discomfort. His facial expression is tense, and his body language suggests that he may be experiencing anxiety or fear. The environment around him is not clearly visible, but the fact that he is lying on a bed indicates that he may be in a private or secluded space. In the audio, as an expert in the field of emotions, I can hear that the individual's voice is shaking and trembling. This suggests that the individual may be experiencing strong emotions such as fear, anxiety, or sadness. In the text, the subtitle content reads: "Isn't it 38 degrees? It's not a fever yet. I'm going to make people laugh." This sentence may be the man's way of expressing his dissatisfaction or frustration with his current state, possibly due to his physical discomfort or a stressful situation. </think>
<answer>anxiety, fear, sadness</answer>

EmoTrust *acoustic-dominant*

<think>In the video, a man is lying on a bed with a pillow under his head, and he appears to be in a state of distress or discomfort. His facial expression is tense, and his body language suggests that he may be experiencing pain. Upon analyzing the acoustic information in the audio, I can hear that the individual's voice is trembling slightly with a hint of impatience. This suggests that the individual may be experiencing strong emotions such as anxiety, nervousness or irritation. Additionally, the individual's speech rate is fast, which may indicate that they are feeling anxious or stressed. In the text, the caption reads: "Isn't it 38 degrees? It's not a fever yet. I'm going to make people laugh." This sentence might be a man's reaction to someone's comment or situation. Therefore, this sentence could be the man's way of expressing his dissatisfaction or frustration with the other person's comment or situation, aligning with his overall negative emotional state.</think>
<answer>anxiety, nervousness, irritation, frustration, dissatisfaction</answer>

Figure 7: Case study between MSFT+GRPO and EmoTrust on the OV-MERD$_{\text{human}}$ dataset.

agreement with human judgments across the three preference categories. These confusion patterns are consistent with the aggregate results in Table 5 and further highlight the benefit of EmoTrust in learning modality preferences that better match human perception.

E Case Study

We further present a representative case from OV-MERD$_{\text{human}}$ to provide an intuitive illustration of EmoTrust's advantage. As shown in Figure 7, the ground-truth labels are *impatient, excited, angry, dissatisfied*, while the visual stream mainly conveys physical discomfort and distress, which can easily bias the model toward internalizing emotions such as *anxiety* or *fear*. As a result, the Multi-view SFT + GRPO baseline is misled by the visual evidence and produces a prediction centered on *anxiety, nervousness, irritation, frustration*, failing to capture the more interaction-oriented emotions expressed in the utterance. In contrast, EmoTrust identifies this sample as *acoustic-dominant* and places greater reliance on the speaker's vocal cues, such as trembling delivery and impatient tone, together with the subtitle content. Consequently, its reasoning is better aligned with the communicative intent of the speaker and yields a prediction that more accurately reflects the annotated emotional state, especially the negative reactive emotions such as *dissatisfaction* and *impatience*.

F AI use statement

In this work, we used generative AI tools only to aid or polish the writing of the manuscript, such as improving readability, grammar, and phrasing. We did not use generative AI tools to generate synthetic datasets, develop theoretical models, formulate or prove mathematical claims. All technical content, experimental findings, and conclusions were produced and verified by the authors. We reviewed all AI-assisted edits and take full responsibility for the final content of this work.