

中国科学技术大学

University of Science and Technology of China

情感智能大模型

从情感理解到共情对话

朱贝尔

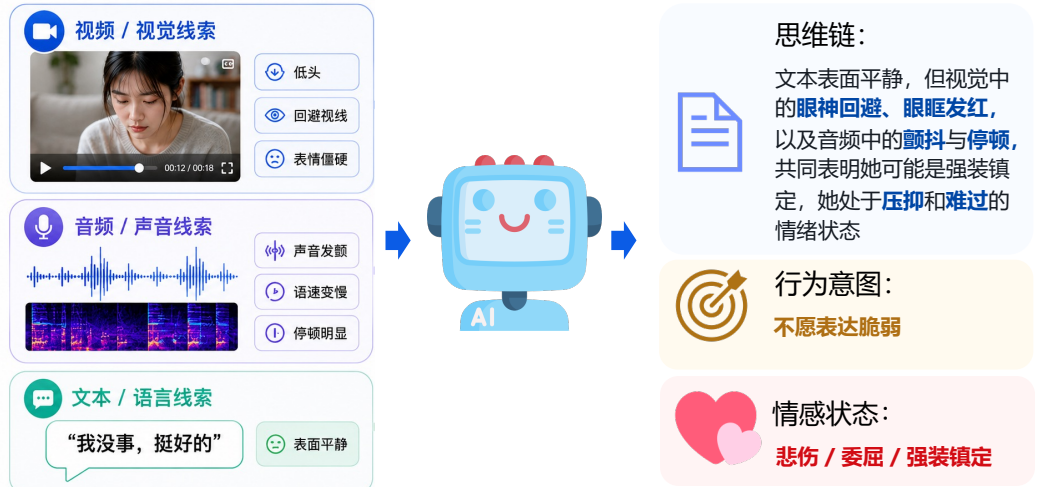
2026年10月

全模态情感推理与共情对话任务介绍

全模态情感推理

理解情绪

综合视频、音频与语言线索，借助思维链分析多模态证据，
识别人物的复杂情感状态与行为意图



共情对话

回应情绪

理解用户在交流中的情绪状态、处境与支持需求，
生成能够回应其感受并提供支持的对话回复



全模态情感推理与共情对话研究概述

全模态情感推理

理解情绪

构建“感知—推理—决策”的多模态情感推理链条

01 稳健感知

处理模态冲突，降低单一模态偏见

MM25 Oral MoSEAR

02 忠实推理

覆盖真实感知证据，抑制跨模态幻觉

ICML OPPO

03 协同决策

融合快思考的召回与慢思考的精度

MER-R1

共情对话

回应情绪

理解不断变化的支持需求，生成恰当的共情回应

01 支持优先级演化

识别当前更需要理解、情感回应还是行动支持

02 自适应奖励建模

为每轮对话生成与当前状态匹配的评价依据

CARE

03 共情策略优化

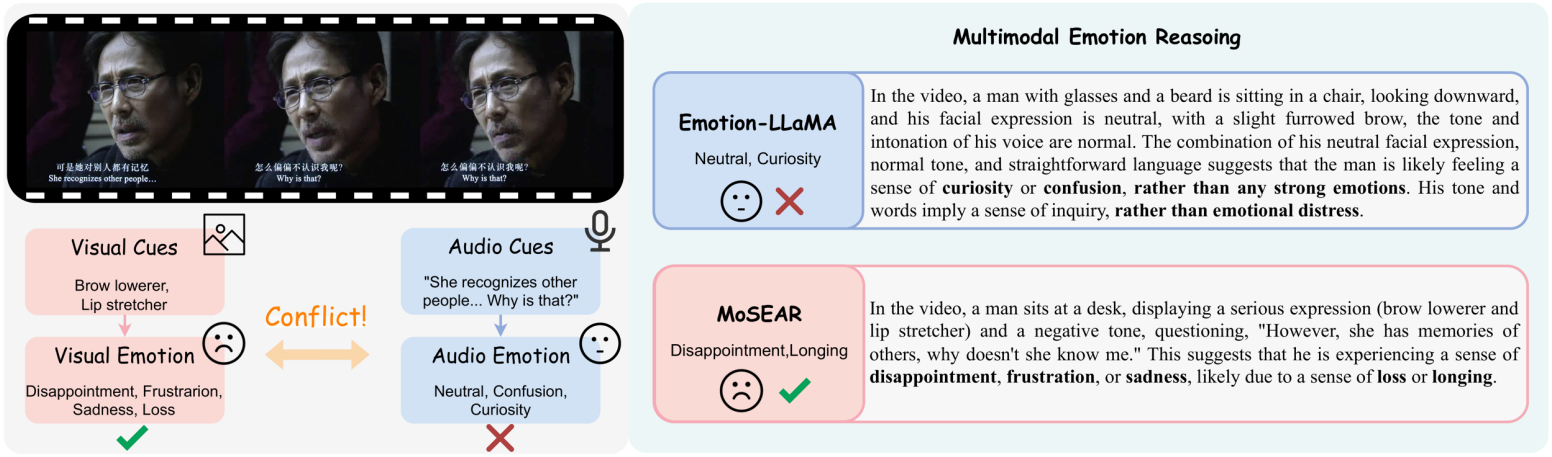
在 EMPA 上将 EPM-Idx 从 28.11 提升至 83.54

全模态情感推理研究一：稳健感知

真实场景中的情感冲突

ACM MM25 Oral

当视觉与音频表达不一致，现有情感大模型容易依赖单一模态，误判真实情绪



(a) Visual-Aligned Sample

(b) Comparison between Emotion-LLaMA and MoSEAR (ours)

V 视觉线索
皱眉 · 眼含泪光
悲伤 / 压抑

A 声音线索
语气平静 · 稳定
中性

Emotion-LLaMA
过度依赖语音 → 误判为中性
单一模态偏见

MoSEAR（本文）
平衡多模态证据 → 识别悲伤
冲突场景下更稳健

目标：让模型在情感冲突中兼顾视觉与音频证据

论文与开放资源

arXiv

Hugging Face

项目主页

YouTube

引用信息

Z. Han, B. Zhu, Y. Xu, P. Song, and X. Yang.

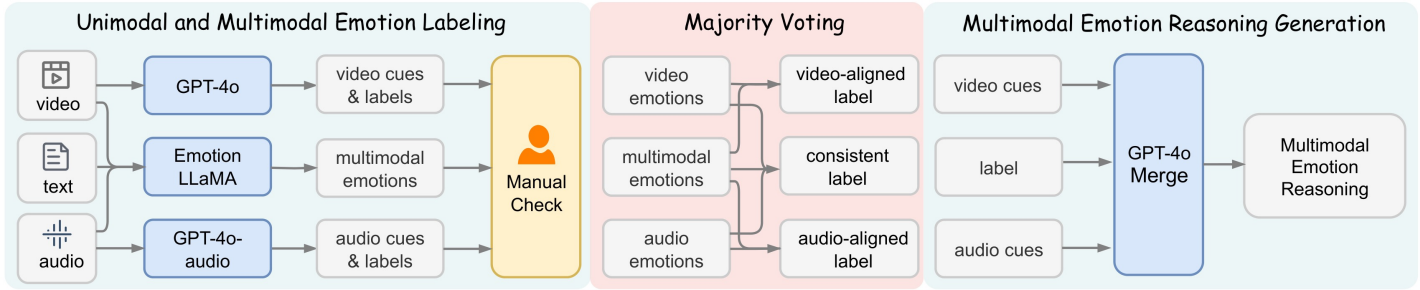
Benchmarking and Bridging Emotion Conflicts for Multimodal Emotion Reasoning.

Proceedings of ACM MM '25

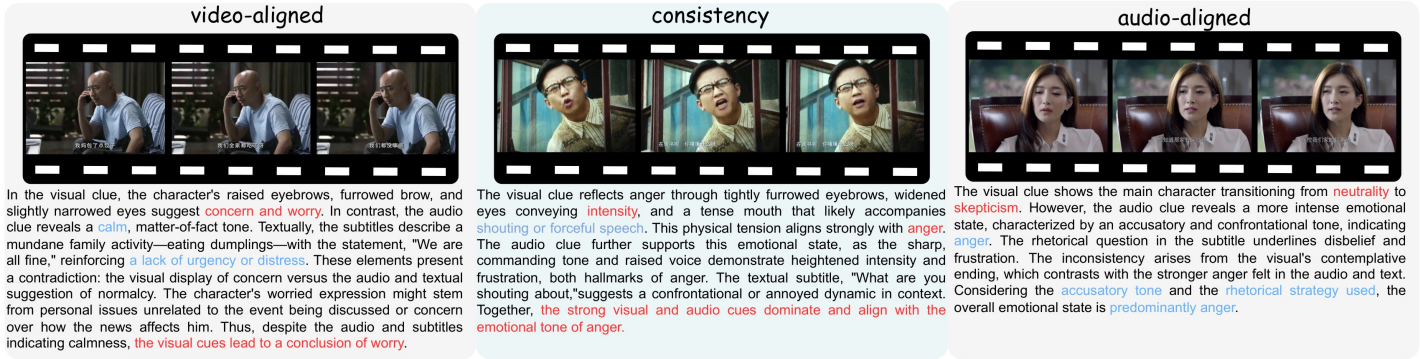
全模态情感推理研究一：稳健感知

CA-MER Benchmark

将样本划分为video-aligned、audio-aligned冲突类与consistent一致类



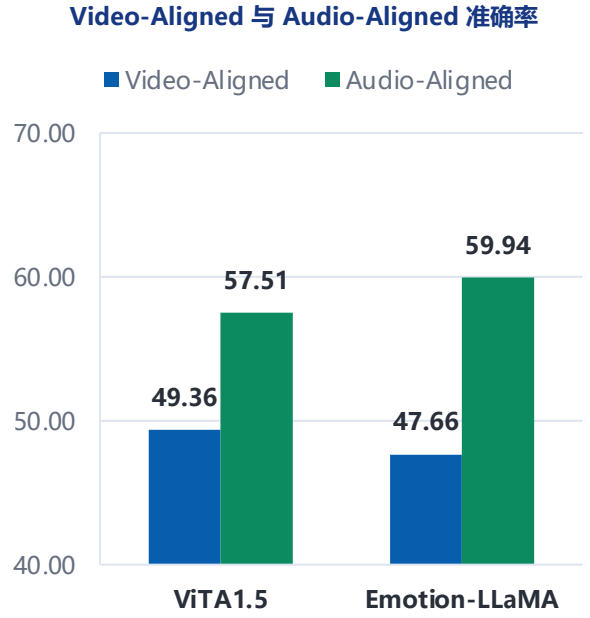
(a) Pipeline of our CA-MER benchmark generation



(b) Samples of the three subsets of our CA-MER

共 1,500 条评测样本，每个子集 500 条

基线测试结果

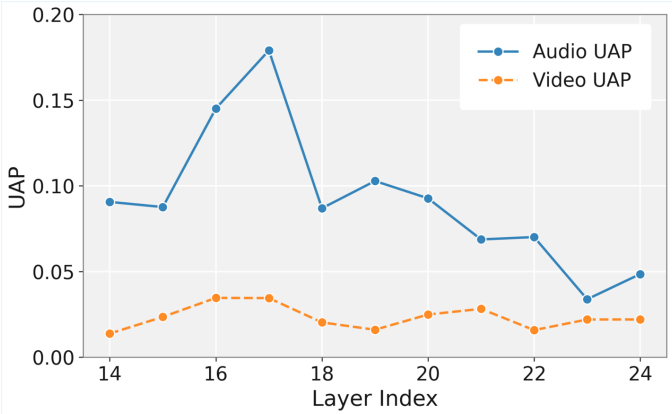


视频线索为真时表现更差，
揭示全模态大模型更依赖音频

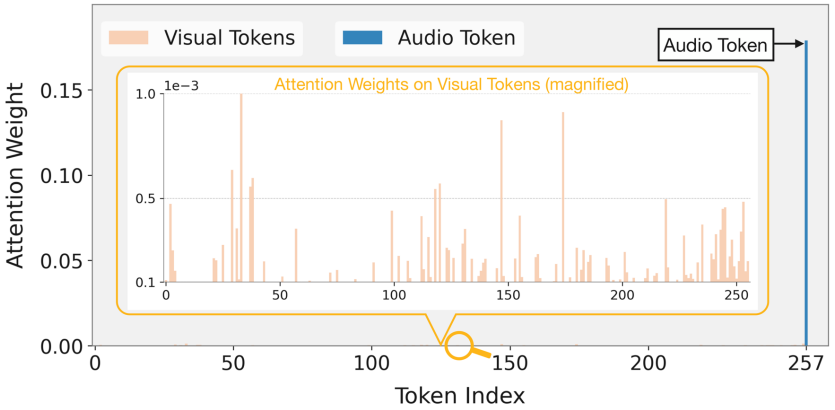
全模态情感推理研究一：稳健感知

注意力偏置

比较中间感知层对音频与视觉 token 的注意力权重



(a) UAP_v vs UAP_a in MLLM middle layers



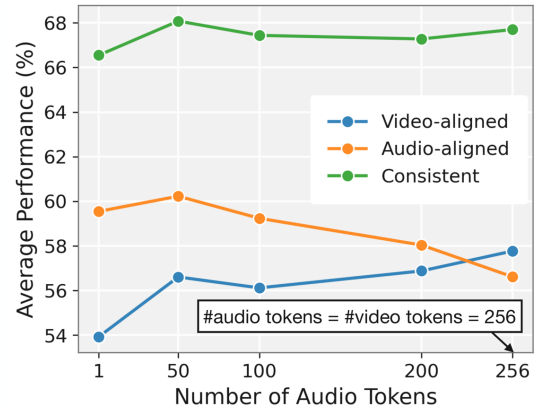
(b) Per-token attention weight analysis

核心发现

中间感知层的音频注意力权重远高于视觉

模型在融合多模态线索时，已提前形成对音频模态的显著依赖

Token 数量效应



(c) Impact of audio token count

核心发现

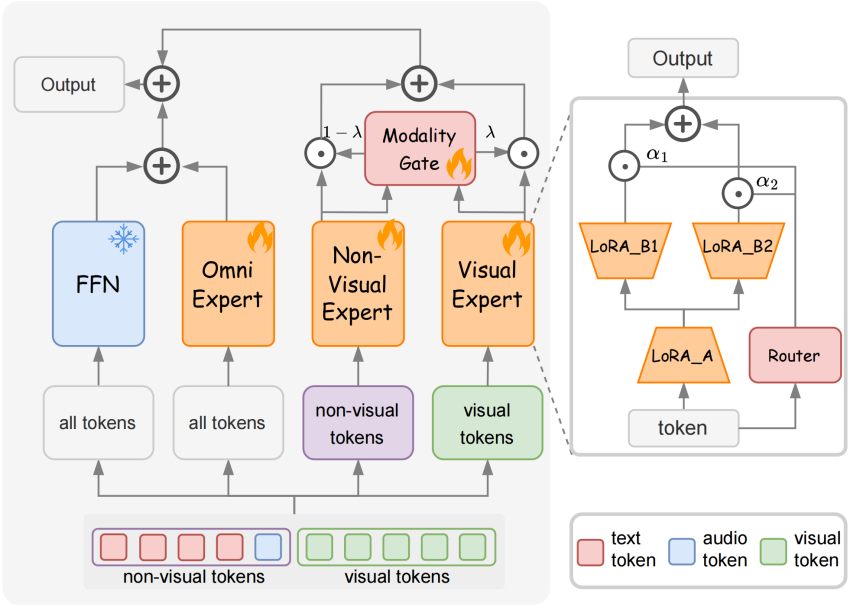
音频 token 增多，
模型对音频的依赖减弱

音视频 token 数相同时，
偏置方向发生反转

全模态情感推理研究一：稳健感知

MoSE：训练阶段去偏

专家建模视觉与非视觉线索，正则门控按样本融合



MoSE 在 FFN 侧平衡模态表征

AR：推理阶段去偏

按样本定位偏置层与注意力头

Algorithm 1 Pipeline of Attention Reallocation (AR)

Given: original attention weights ω, threshold τ
Compute layer-level ratio $c(\omega)$ via Eq. (6)
if $c(\omega) > \tau$ **then** ← 定位偏置层
  for $h \in [H]$ **do**
    Compute head-level ratio $c_h(\omega)$ via Eq. (5)
    if $c_h(\omega) > c(\omega)$ **then** ← 定位偏置头
      Update ω'_h via Eq. (12) and Eq. (13) ← 音频注意力过高：向视觉重分配
    else
      Update $\omega'_h = \omega_h$ ← 注意力正常：保持原分配
    end if
  end for
else
  Update $\omega' = \omega$
end if
Return: reallocated attention weights ω'

在 Attention 侧校正样本级音频过载

全模态情感推理研究一：稳健感知

冲突场景：CA-MER

MoSEAR 在三个冲突子集及 Overall 上均取得最佳结果

Table 1: Performance (%) of emotion reasoning on CA-MER. "Acc." and "Rec." denote accuracy and recall, respectively.

Model	Modality	Video-Aligned			Audio-Aligned			Consistent			Overall		
		Acc.	Rec.	Avg.	Acc.	Rec.	Avg.	Acc.	Rec.	Avg.	Acc.	Rec.	Avg.
SALMONN [67]	A+T	29.37	22.56	25.97	53.35	34.19	43.77	47.72	38.53	43.13	43.48	31.76	37.62
mPLUG-Owl3 [78]	V+T	40.71	42.01	41.36	41.08	41.01	41.04	49.29	39.08	44.18	43.69	40.70	42.19
Chat-UniVi [30]	V+T	43.43	42.42	42.92	42.28	36.31	39.30	50.38	45.95	48.16	45.36	41.56	43.46
Vita1.5 [16]	A+V+T	49.36	46.60	47.98	57.51	48.15	52.83	57.17	52.52	54.84	54.68	49.09	51.88
Emotion-LLaMA [9]	A+V+T	47.66	51.45	49.56	59.94	51.63	55.78	59.23	63.80	61.52	55.61	55.63	55.62
MoSEAR	A+V+T	**58.42**	**54.28**	**56.35**	**65.33**	**55.26**	**60.30**	**67.68**	**65.87**	**66.77**	**63.81**	**58.47**	**61.14**

Overall Acc

63.81

较 Emotion-LLaMA 提升 **8.2**

Video-Aligned Acc

58.42

较 Emotion-LLaMA 提升 **10.8**

Audio-Aligned Acc

65.33

较 Emotion-LLaMA 提升 **5.4**

注：MoSEAR 和 Emotion-LLaMA 使用相同 Base Model

常规情感任务

情感推理与识别能力同步提升

EMER

Average

60.58

+ 9.1

MER2023

F1

0.9027

+ 0.094

DFEW

UAR

64.26

+ 1.96

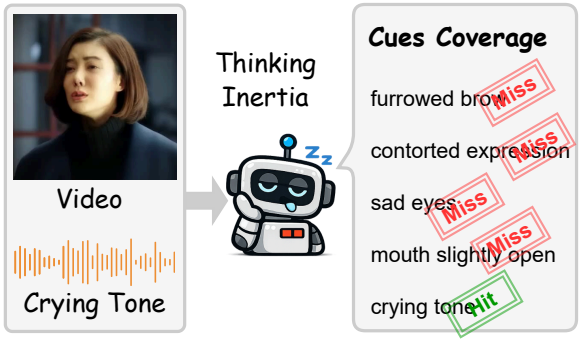
去偏没有牺牲常规场景性能

全模态情感推理研究二：忠实推理

ICML 2026

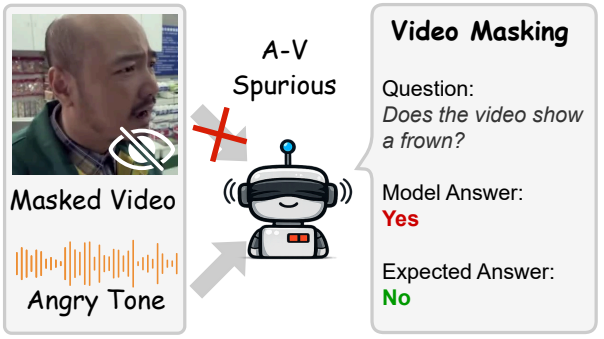
证据利用不足

推理链只抓住显著语音，
遗漏皱眉、悲伤眼神等视觉证据



跨模态幻觉

视频被遮蔽后，模型仍把声音线索
臆测成皱眉等视觉事实



用得不全

多模态输入已经存在，
推理过程却没有覆盖关键证据

说得不真

模型说出模态特定证据，
却无法追溯到对应输入

目标：思维链既要覆盖真实视觉 / 音频证据，也要让每条证据可追溯到对应模态

论文与开放资源

arXiv

Hugging Face

项目主页

GitHub

引用信息

Z. Han, B. Zhu et al.

Omni-Perception Policy Optimization for Multimodal Emotion Reasoning

ICML 2026

全模态情感推理研究二：忠实推理

MEP-Bench

从 Utilization 与 Faithfulness 两方面，诊断多模态感知是否可靠

UTILIZATION

证据利用率

模型推理链是否覆盖人类标注的视觉、音频与情感证据

Atomic Cues → Evidence Recall

FAITHFULNESS

模态忠实度

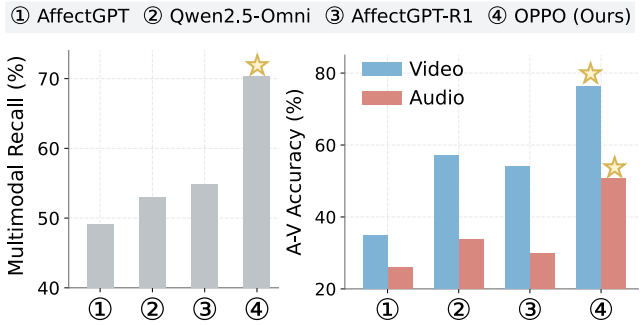
遮蔽视频或音频后，模型是否仍编造被遮蔽模态的专属证据

Masking Probe → Acc@V / Acc@A

评测构建

OV-MERD → 300 条线索丰富样本 → 视频 / 音频各 500 个 probe

MEP-Bench 结果

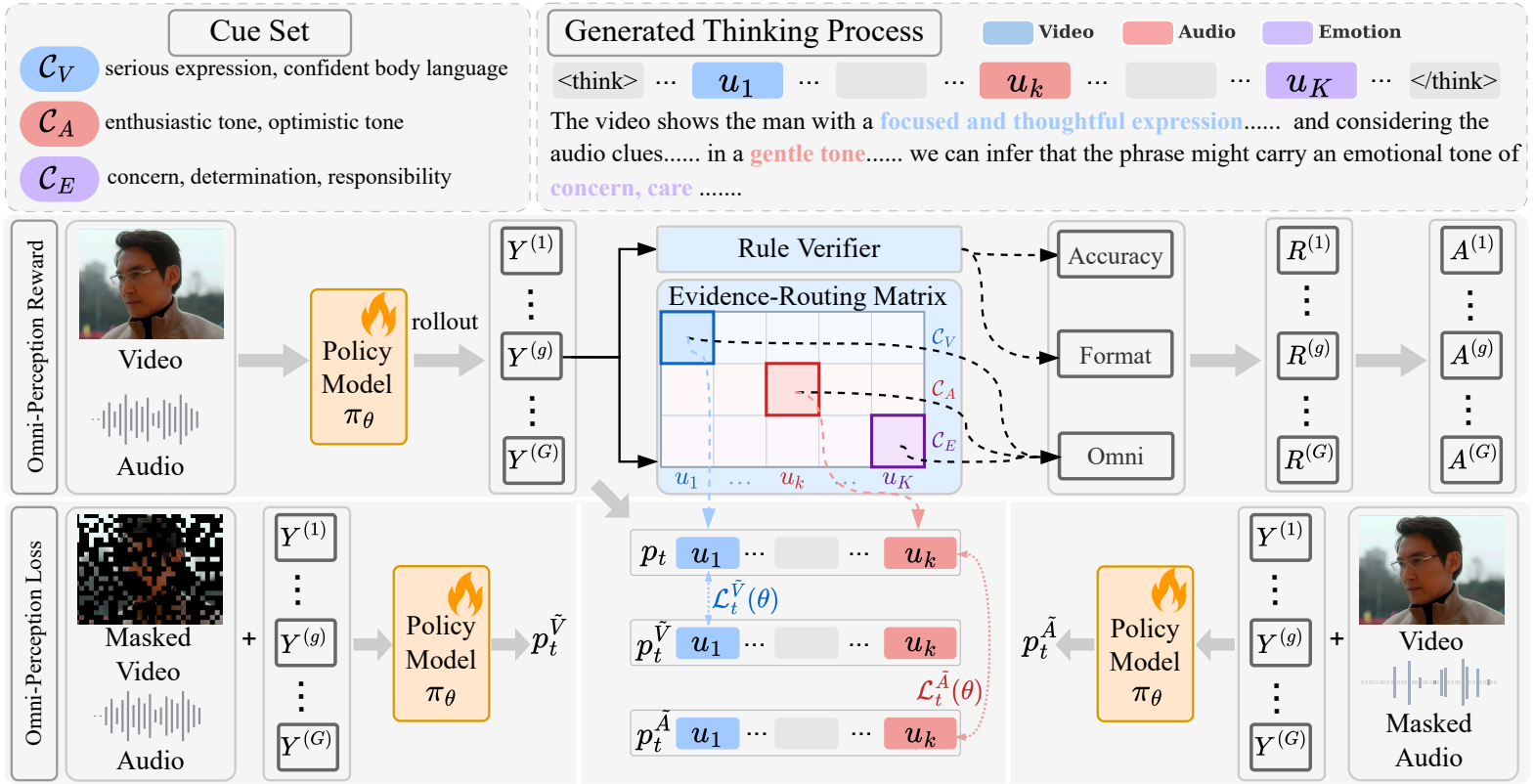


诊断结论

1. 现有模型仅覆盖约一半人类线索
2. 现有模型暴露跨模态幻觉，准确率低于Random Guessing

全模态情感推理研究二：忠实推理

OPPO Framework



① 证据对齐

Cue Set 拆成 CV / CA / CE，
生成思维拆成 clauses

② Omni-Perception Reward

Evidence-Routing Matrix
奖励真实线索覆盖

③ Omni-Perception Loss

遮蔽单一模态，
约束模态专属 token 的分布变化

全模态情感推理研究二：忠实推理

整体性能

MER-UniBench 主结果

Model	Sentiment Analysis				Basic Emotion				Fine	Mean
	MOSI	MOSEI	SIMS	SIMSv2	MER23	MER24	MELD	IEMOCAP	OV-MERD+	
Qwen-Audio (Chu et al., 2023)	70.09	46.90	70.73	65.26	41.85	31.61	49.09	35.47	32.36	49.26
SALMONN (Tang et al., 2024)	81.00	67.03	68.69	65.93	55.53	45.38	45.62	46.84	45.00	57.89
VideoChat2 (Li et al., 2024a)	66.84	54.32	69.49	70.66	33.67	54.50	36.64	48.70	39.21	52.67
LLaMA-VID (Li et al., 2024b)	61.78	63.89	69.35	67.48	50.72	57.60	42.75	46.02	45.01	56.07
Chat-UniVi (Jin et al., 2024)	54.53	63.18	68.15	66.36	57.62	65.67	45.61	52.37	48.00	57.94
mPLUG-Owl (Ye et al., 2023)	72.40	72.91	72.13	75.00	56.86	59.89	49.11	55.54	48.18	62.45
PandaGPT (Su et al., 2023)	61.92	67.61	68.38	67.23	40.21	51.89	37.88	44.04	37.12	52.92
R1-Omni (Zhao et al., 2025a)	58.02	56.48	71.82	68.58	64.17	67.43	43.20	51.58	55.24	59.61
Emotion-LLaMA (Cheng et al., 2024)	66.13	67.66	78.32	77.23	59.38	73.62	46.76	55.47	52.97	64.17
AffectGPT (Lian et al., 2025a)	81.30	80.90	**88.49**	86.18	78.54	78.80	55.65	60.54	62.52	74.77
AffectGPT-R1 (Lian et al., 2025c)	79.65	80.18	87.26	85.75	84.51	**93.13**	**66.71**	**74.26**	**68.39**	79.98
AffectGPT (Last Checkpoint)	78.30	78.51	85.28	85.31	73.69	76.90	52.09	57.74	59.89	71.96
AffectGPT-R1 (Reproduce)	80.29	80.64	85.70	83.75	81.88	80.89	57.53	65.71	64.08	75.60
OPPO	**86.50**	**84.63**	86.22	**88.26**	**87.73**	90.34	64.06	73.60	67.16	**81.05**

81.05

MER-UniBench Mean

+5.45

相对AffectGPT-R1 提升

OPPO 取得最高平均分

消融实验

MER-UniBench 与 MEP-Bench

Model Variant	Sentiment Analysis				Basic Emotion				Fine	Mean	Util.	Faithfulness	
	MOSI	MOSEI	SIMS	SIMSv2	MER23	MER24	MELD	IEMOCAP	OV-MERD+		Recall	Acc@V	Acc@A
Baseline	83.48	**86.31**	87.78	86.84	78.16	82.70	61.45	67.28	66.86	77.87	58.00	59.20	33.00
+ R_{omni}	84.83	85.04	86.48	86.95	85.40	85.90	63.37	70.03	67.09	79.45	62.55	64.80	39.40
+ PAPO	83.47	85.66	86.21	87.36	86.27	87.09	63.28	69.12	65.65	79.34	58.29	62.00	34.60
+ PAPO-dual	84.98	85.06	**88.47**	86.72	84.96	86.11	63.07	67.51	66.20	79.23	60.58	65.20	34.80
+ $\mathcal{J}_{\text{omni}}(\theta)$	**86.50**	84.63	86.22	**88.26**	**87.73**	**90.34**	**64.06**	**73.60**	**67.16**	**81.05**	**70.44**	**76.40**	**50.60**

相较 Baseline 的提升

+12.44

Recall

+17.20

Acc@V

+17.60

Acc@A

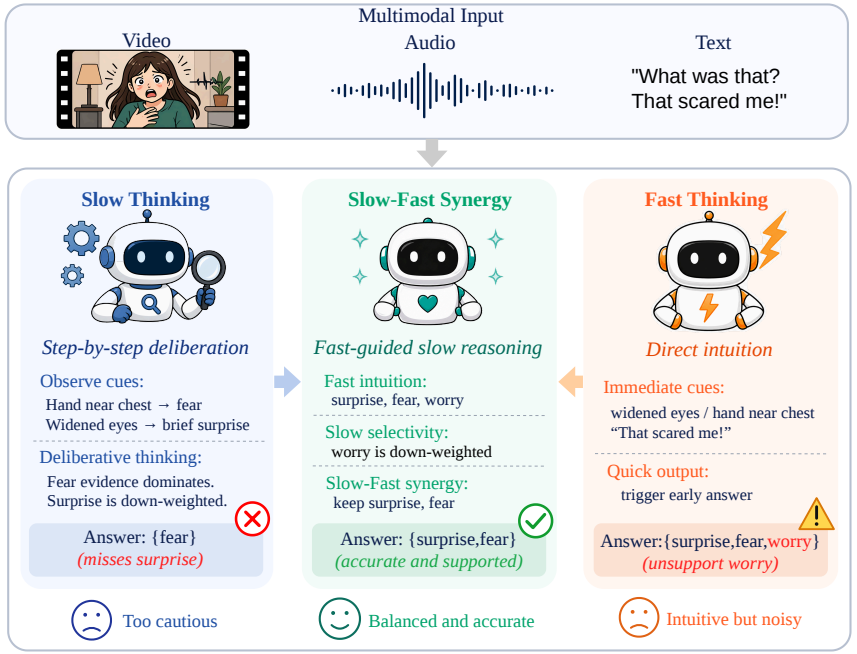
Reward 提高线索覆盖；Loss 进一步强化模态归因

两者协同提升利用度与忠实性

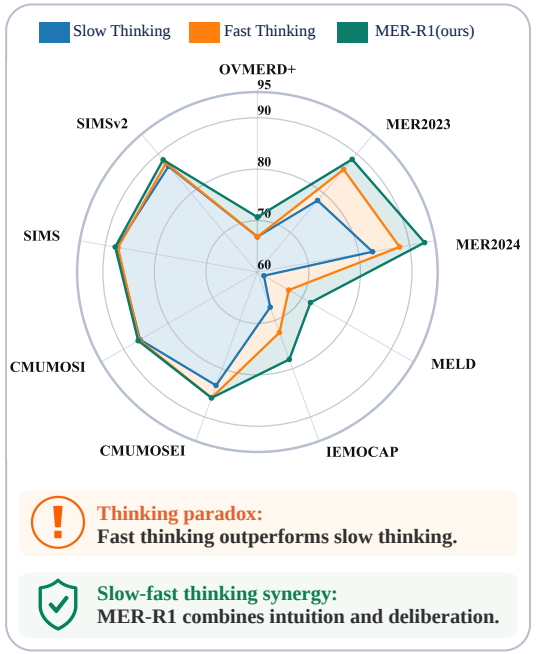
全模态情感推理研究三：协同决策

快慢思考的性能悖论

显式推理提升可解释性，但情感识别反而弱于快思考



(a) Motivating example



(b) Evaluation summary on **9** datasets

目标：融合快思考的广覆盖与慢思考的高精度

论文与开放资源

arXiv

引用信息

Z. Han, B. Zhu, W. Tong, et al.

MER-R1: Multimodal Emotion Reasoning via Slow-Fast Thinking Synergy

arXiv preprint, 2026

全模态情感推理研究三：协同决策

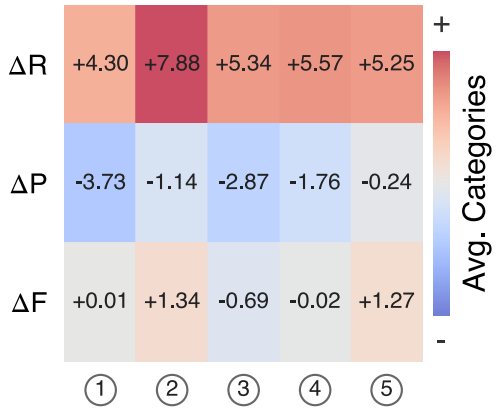
快思考扩大情感覆盖并提高正确类别置信度，慢思考提升精度并抑制错误类别，二者在覆盖范围与选择性上互补

召回与精度权衡

快思考：召回更高

慢思考：精度更高

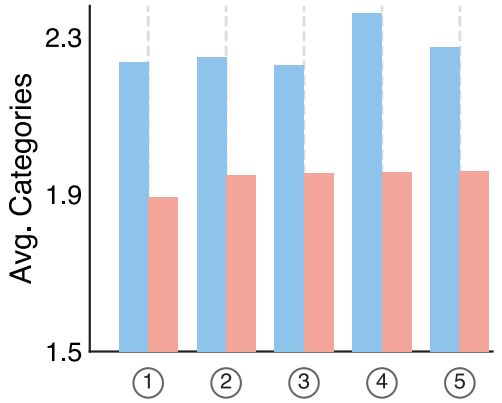
两者F1差异有限



情感覆盖范围

快思考预测更多情感类别

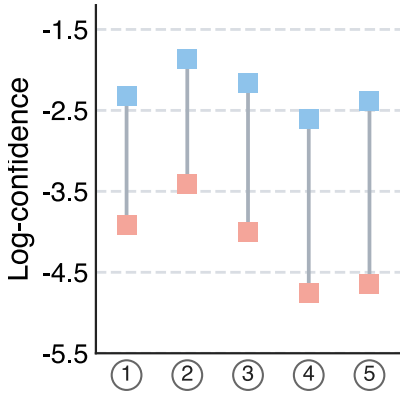
覆盖更广，也更容易过预测



正确类别置信度

快思考对正确情感更自信

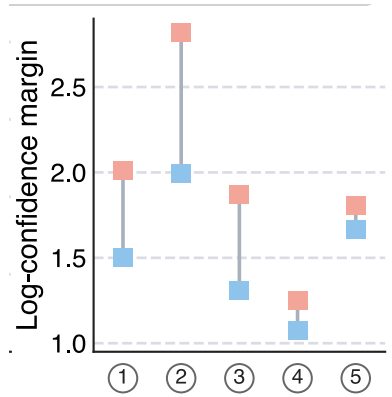
五个基准上均表现出更高置信度



错误类别抑制

慢思考形成更大的置信度间隔

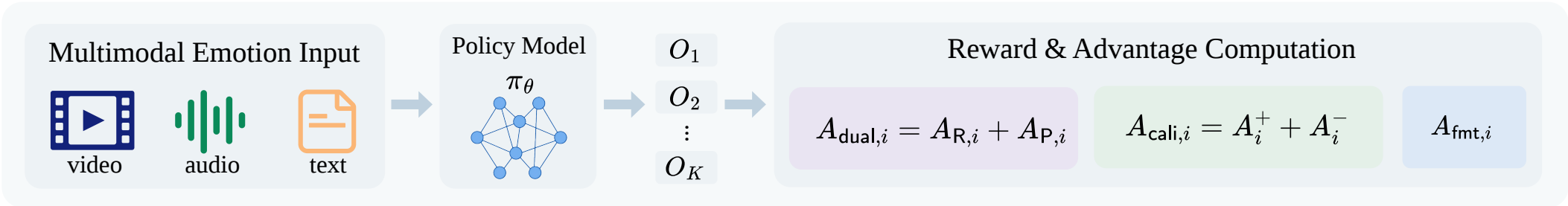
通过抑制错误类别提高选择性



① OV-MERD+ ② MER2023 ③ MER2024 ④ MELD ⑤ IEMOCAP 快思考 慢思考

全模态情感推理研究三：协同决策

MER-R1 Framework



双目标解耦

Dual-Objective Disentanglement

Reward Disentanglement

$$R_{\text{R},i} = \frac{|\hat{\mathcal{Y}} \cap \mathcal{G}|}{|\mathcal{G}|}$$

$$R_{\text{P},i} = \frac{|\hat{\mathcal{Y}} \cap \mathcal{G}|}{|\hat{\mathcal{Y}}|}$$

Advantage Disentanglement

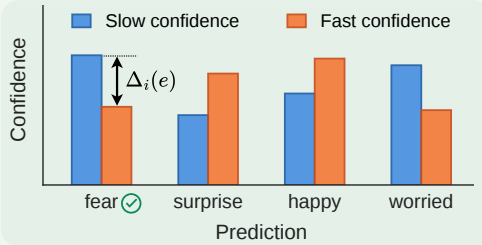
$$A_{\text{R},i} = \frac{R_{\text{R},i} - \mu_{\text{R}}}{\sigma_{\text{R}}}$$

$$A_{\text{P},i} = \frac{R_{\text{P},i} - \mu_{\text{P}}}{\sigma_{\text{P}}}$$

1. **奖励层：** 将 F1 拆为召回奖励与精度奖励
2. **优势层：** 分别归一化后再合并，避免单一目标主导训练

快慢置信度校准

Slow-Fast Confidence Calibration



correct category $\mathcal{E}_i^+$

$$R_i^+ = \hat{\mathbb{E}}_{e \in \mathcal{E}_i^+} \Delta_i(e) \rightarrow A_i^+$$

incorrect category $\mathcal{E}_i^-$

$$R_i^- = \hat{\mathbb{E}}_{e \in \mathcal{E}_i^-} - \Delta_i(e) \rightarrow A_i^-$$

1. **正确类别：** 让慢思考保持快思考的高置信度
2. **错误类别：** 保留慢思考对虚假情感的低置信度

全模态情感推理研究三：协同决策

MER-UniBench

多任务情感识别与情感分析

Model	Fine (F1)	Basic Emotion (Hitrate)				Sentiment Analysis (WAF)				Mean
	OV-MERD+	MER23	MER24	MELD	IEMOCAP	MOSI	MOSEI	SIMS	SIMSv2	
Qwen-Audio [32]	32.36	41.85	31.61	49.09	35.47	70.09	46.90	70.73	65.26	49.26
SALMONN [33]	45.00	55.53	45.38	45.62	46.84	81.00	67.03	68.69	65.93	57.89
VideoChat2 [34]	39.21	33.67	54.50	36.64	48.70	66.84	54.32	69.49	70.66	52.67
LLaMA-VID [35]	45.01	50.72	57.60	42.75	46.02	61.78	63.89	69.35	67.48	56.07
Chat-UniVi [36]	48.00	57.62	65.67	45.61	52.37	54.53	63.18	68.15	66.36	57.94
mPLUG-Owl [37]	48.18	56.86	59.89	49.11	55.54	72.40	72.91	72.13	75.00	62.45
PandaGPT [38]	37.12	40.21	51.89	37.88	44.04	61.92	67.61	68.38	67.23	52.92
R1-Omni [30]	55.24	64.17	67.43	43.20	51.58	58.02	56.48	71.82	68.58	59.61
Emotion-LLaMA [27]	52.97	59.38	73.62	46.76	55.47	66.13	67.66	78.32	77.23	64.17
AffectGPT [11]	62.52	78.54	78.80	55.65	60.54	81.30	80.90	**88.49**	86.18	74.77
AffectGPT-R1 [15]	68.39	84.51	**93.13**	66.71	74.26	79.65	80.18	87.26	85.75	79.98
AffectGPT†	59.89	73.69	76.90	52.09	57.74	78.30	78.51	85.28	85.31	71.96
AffectGPT-R1†	64.08	81.88	80.89	57.53	65.71	80.29	80.64	85.70	83.75	75.60
Baseline	66.86	78.16	82.70	61.45	67.28	83.48	86.31	87.78	86.84	77.87
MER-R1	**70.68**	**88.60**	92.96	**71.90**	**78.10**	**86.06**	**86.73**	88.04	**88.48**	**83.50**

83.50

MER-UniBench Mean

+5.63

相较 Baseline 提升

MER-R1 取得最高平均分

MME-Emotion

多维情感推理

Methods	ER-Lab	ER-Wild	FG-ER	FG-SA	ML-ER	Noise-ER	IR	SA	Mean		
									Rec.	Rea.	CoT
PandaGPT [38]	28.5	23.2	25.2	41.4	28.9	26.3	29.2	37.7	21.6	38.4	30.0
Emotion-LLaMA [27]	31.7	19.7	31.7	36.4	29.5	49.3	27.3	40.7	21.1	42.4	31.8
AffectGPT [11]	35.0	32.8	32.6	**44.2**	31.3	50.7	30.0	47.9	19.7	57.4	38.6
Baseline	47.0	40.8	35.0	35.6	37.7	59.0	32.4	58.7	27.9	62.7	45.3
VideoAuto-R1 [26]	49.4	44.4	**40.2**	33.1	37.3	64.8	32.8	56.7	34.5	58.8	46.6
MER-R1	**57.3**	**49.6**	38.7	37.2	**42.2**	**72.1**	**33.9**	**62.9**	**38.4**	**64.6**	**51.5**

38.4

Recall

64.6

Reasoning

51.5

CoT

8 项子任务中 6 项最优

MER-R1 取得最佳整体性能

全模态情感推理研究三：协同决策

消融1：快慢思考对比

Model	Mode	Fine (F1)	Basic Emotion (F1)				Sentiment Analysis (WAF)				R-Mean	Mean
		OV-MERD+	MER23	MER24	MELD	IEMOCAP	MOSI	MOSEI	SIMS	SIMSv2		
Baseline	Fast	**66.87**	**61.92**	64.37	46.18	**53.13**	**86.03**	**86.47**	87.48	**87.60**	**58.49**	**71.11**
	Slow	66.86	60.58	**65.06**	**46.20**	51.86	83.48	86.31	**87.78**	86.84	58.11	70.52
VideoAuto-R1	Fast	**69.67**	**65.09**	**67.38**	**46.77**	**56.33**	**86.21**	**86.05**	86.20	**87.93**	**61.05**	**72.40**
	Slow	69.49	64.20	67.20	46.36	55.32	85.75	85.86	**86.27**	87.83	60.51	72.03
MER-R1	Fast	70.10	64.41	67.21	46.89	55.71	85.30	**87.43**	**88.57**	87.49	60.86	72.56
	Slow	**70.68**	**65.43**	**68.46**	**48.70**	**55.73**	**86.06**	86.73	88.04	**88.48**	**61.80**	**73.14**

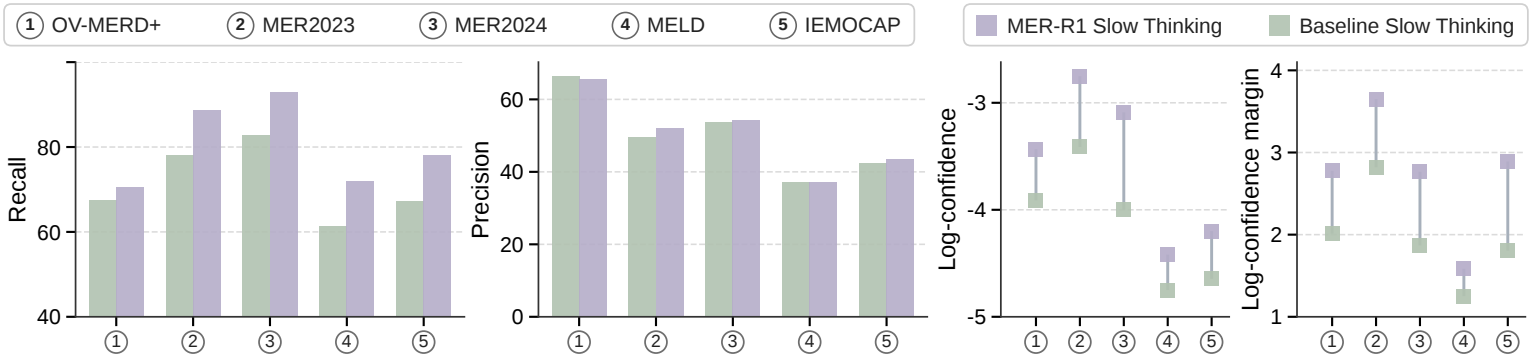
Baseline **−0.38**

VideoAuto-R1 **−0.54**

MER-R1 +0.94

仅 MER-R1 实现慢思考反超

消融2：反超机制

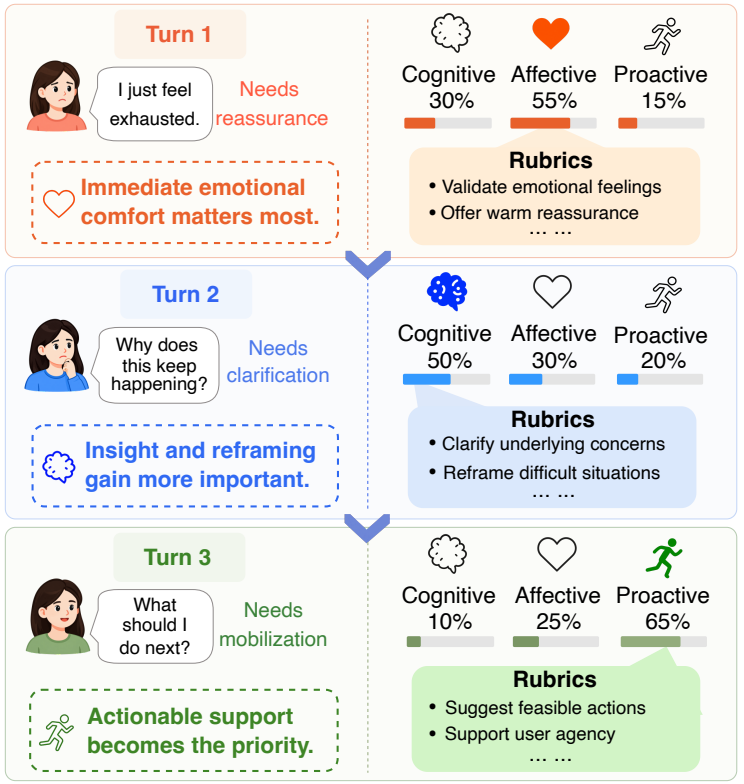


- **Recall** 提升
- **Precision** 保持
- **正类置信度** 提升
- **置信度间隔** 扩大

MER-R1 覆盖能力与置信度判别同步改善

共情对话动态支持

核心矛盾：固定奖励标准难以匹配对话中持续变化的支持需求



共情回复缺少唯一答案

共情强化学习通常依赖自然语言Rubric，将开放式回复的质量转化为奖励信号

现有方法采用固定Rubric

奖励重点在整段对话中保持不变，无法跟随用户从情绪安抚转向理解或行动支持

研究问题

如何让奖励标准随当前对话状态动态调整?

论文与开放资源

arXiv

Hugging Face

项目主页

GitHub

引用信息

P. Huang, Z. Han, W. Tong, et al.

Evolving Support Priorities in Empathetic Reinforcement Learning

arXiv preprint, 2026

共情对话动态支持

三维共情定义

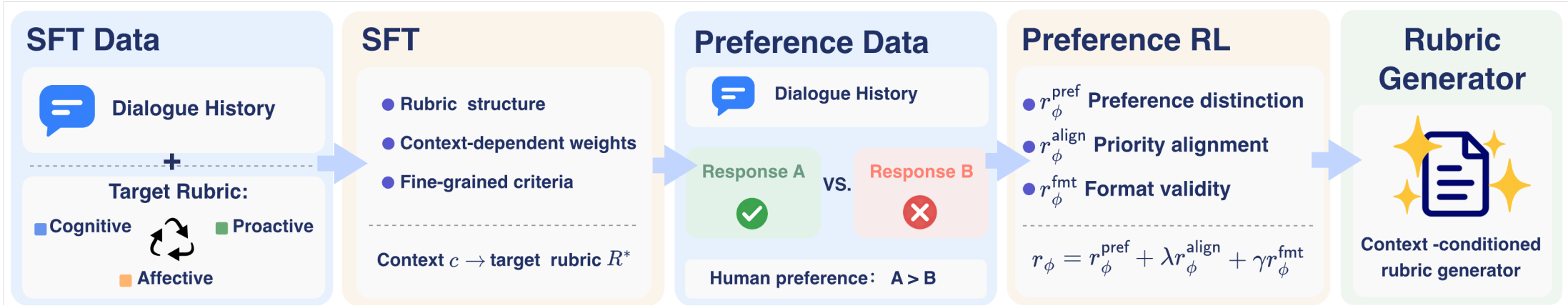
认知共情 Cognitive Empathy
理解用户视角，澄清并重构问题

情感共情 Affective Empathy
识别并确认用户的情绪体验

主动共情 Proactive Empathy
增强用户能动性与行动能力

权重与细粒度评价标准随当前对话状态变化

1.5K SFT 样本与2.8K 偏好对共同训练 Rubric Generator



MINT / ESConv
LLM 生成，双评审筛选

学习三维权重
与细粒度评价标准

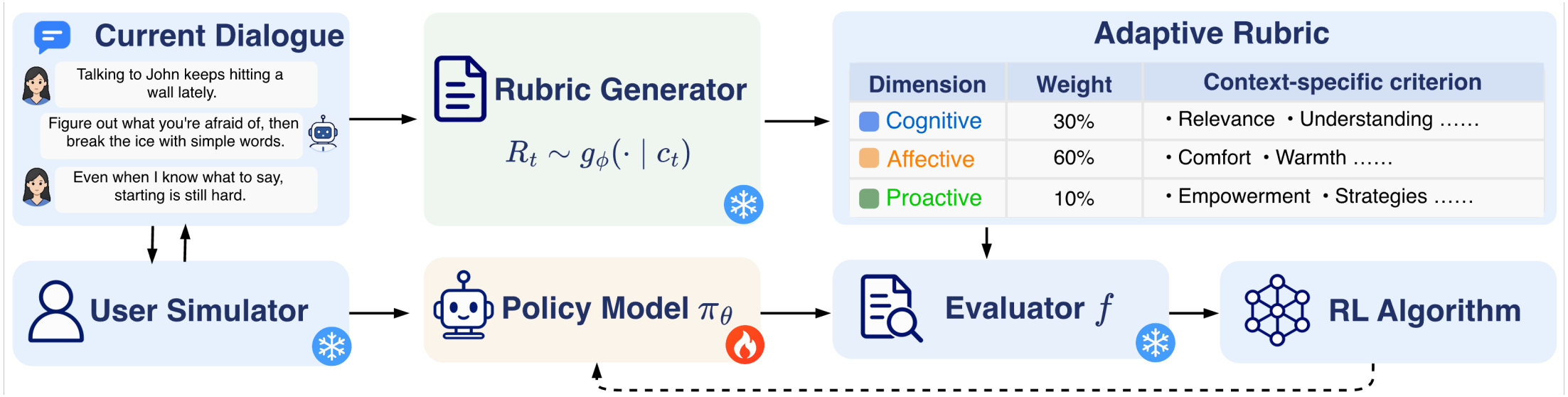
多模型候选回复
匿名配对，人工多数票

偏好、优先级与格式奖励
GRPO 优化生成器

共情对话动态支持

Online RL with Adaptive Rubrics

同轮共享标准，跨轮动态更新



动态生成

Rubric Generator 根据当前上下文
生成维度权重与评价标准

统一评估

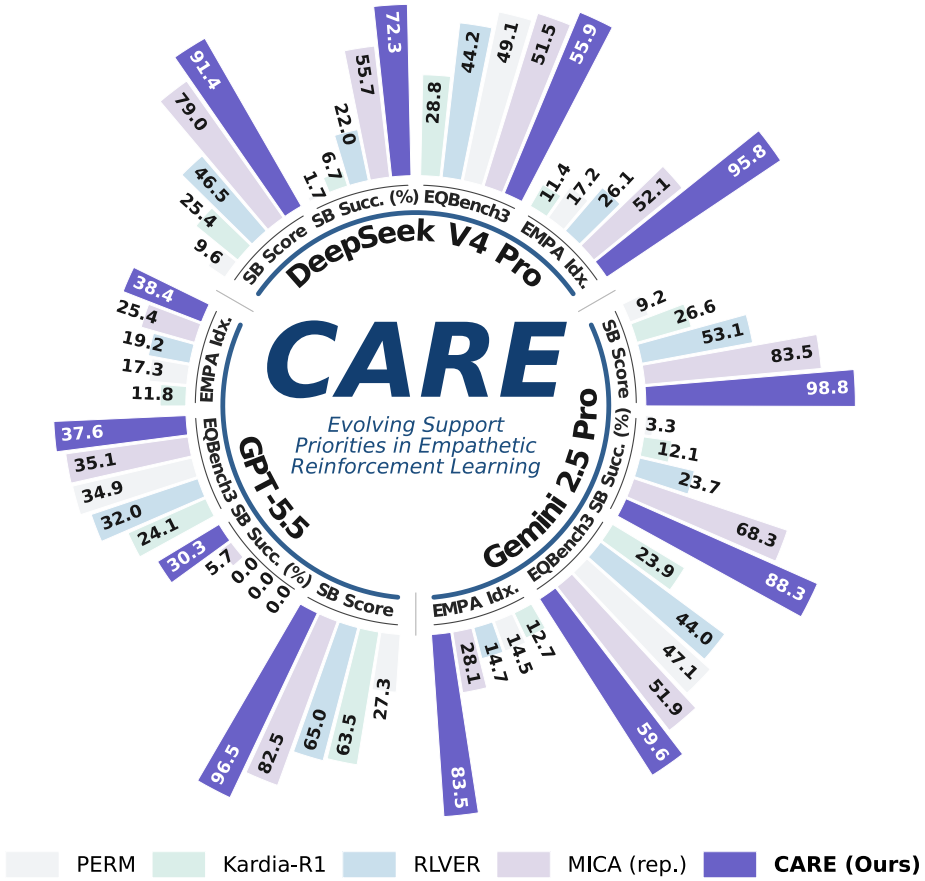
同一 rollout group 的候选回复
共享同一 rubric

在线优化

Evaluator 汇总加权得分
GRPO 更新 Policy

共情对话动态支持

CARE 在全部基准和三位独立评审模型下均取得最优表现



EMPA EPM-Idx

28.11 升至 83.54！

Gemini-2.5-Pro 评审

全面领先

SentientBench、EQBench3 与 EMPA 在三位评审模型下均达到最佳

稳定增益

EMPA 上较最强基线至少提升 13 分

整体排名

CARE(R) 与 CARE(M)

包揽所有汇总指标前两名

共情对话动态支持

前沿大模型对比

32B 可比前沿模型，超过部分超大模型

Table 2: **Scaling and frontier comparison on EQBench3 under DeepSeek-V4-Pro.** We compare Base and CARE-enhanced Qwen3 models from 8B to 32B, together with strong frontier LLMs.

Model	Overall	DoI	ER	DE	WRM	HL	PEI	SD
GPT-5.5	83.35	17.57	16.98	16.64	14.32	16.72	16.72	16.16
Gemini-2.5-Pro	82.65	17.71	16.73	17.00	15.31	17.31	15.81	15.85
Claude-Sonnet-4.6	79.30	17.52	16.48	15.38	13.35	15.54	14.00	13.88
DeepSeek-V4-Pro-1.6T	77.75	16.91	16.14	15.69	13.73	16.31	13.96	13.69
Qwen3.7-Max	76.60	17.00	15.51	15.65	13.23	14.92	14.58	13.77
GLM-5.2-753B	74.90	16.84	15.43	14.62	12.46	15.12	13.35	12.92
Qwen3-8B								
Base	$61.60_{\pm 0.97}$	$13.72_{\pm 0.25}$	$12.87_{\pm 0.18}$	$12.56_{\pm 0.12}$	$11.04_{\pm 0.04}$	$11.67_{\pm 0.72}$	$11.01_{\pm 0.63}$	$9.67_{\pm 0.64}$
CARE (R)	$68.22_{\pm 1.59}$	$14.47_{\pm 0.29}$	$14.09_{\pm 0.20}$	$14.25_{\pm 0.57}$	$12.37_{\pm 0.35}$	$13.55_{\pm 0.10}$	$12.99_{\pm 0.72}$	$12.11_{\pm 0.72}$
Qwen3-14B								
Base	$68.88_{\pm 0.93}$	$15.28_{\pm 0.17}$	$14.26_{\pm 0.26}$	$14.06_{\pm 0.37}$	$12.72_{\pm 0.02}$	$13.82_{\pm 0.42}$	$12.57_{\pm 0.46}$	$11.79_{\pm 0.41}$
CARE (R)	$73.98_{\pm 1.73}$	$15.88_{\pm 0.28}$	$15.24_{\pm 0.36}$	$15.11_{\pm 0.34}$	$13.46_{\pm 0.11}$	$14.97_{\pm 0.23}$	$14.08_{\pm 0.47}$	$13.37_{\pm 0.46}$
Qwen3-32B								
Base	$73.50_{\pm 1.34}$	$16.07_{\pm 0.22}$	$14.89_{\pm 0.34}$	$14.50_{\pm 0.24}$	$12.72_{\pm 0.66}$	$14.85_{\pm 0.55}$	$14.04_{\pm 0.04}$	$13.23_{\pm 0.35}$
CARE (R)	$76.35_{\pm 2.50}$	$16.35_{\pm 0.46}$	$15.99_{\pm 0.83}$	$14.76_{\pm 0.10}$	$13.03_{\pm 0.45}$	$15.38_{\pm 0.70}$	$14.41_{\pm 0.48}$	$13.80_{\pm 0.72}$

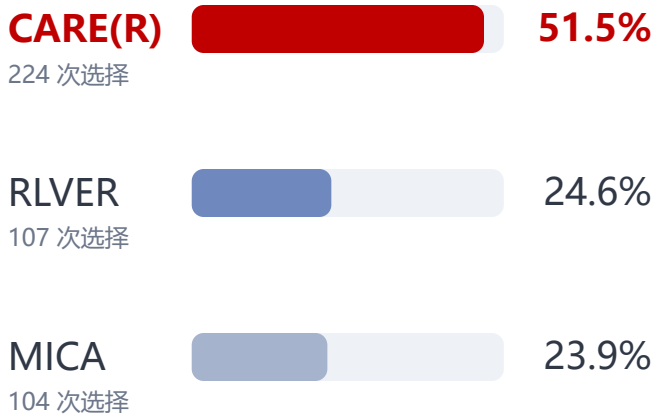
32B CARE (R) 76.35 ≈ Qwen3.7-Max 76.60 > GLM-5.2-753B 74.90

14B CARE (R) 73.98 > Qwen3-32B Base 73.50

人工评测

15 位心理学背景标注者 · 盲评

明确偏好率



超过一半的明确偏好选择 CARE(R)

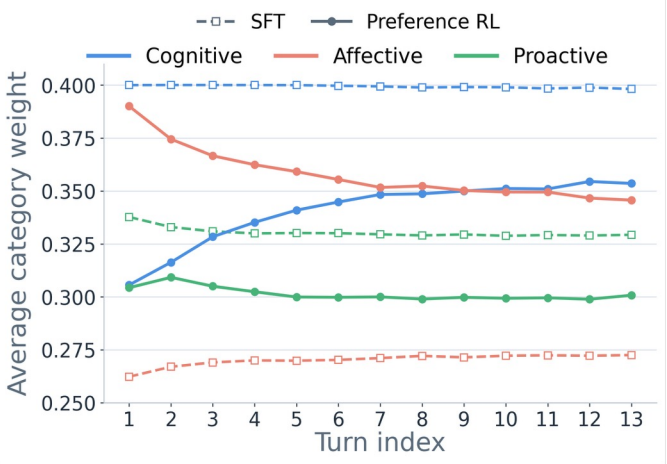
50 个对话上下文 · 每人评测 30 组
tie 不计入偏好率

共情对话动态支持

消融与进一步分析

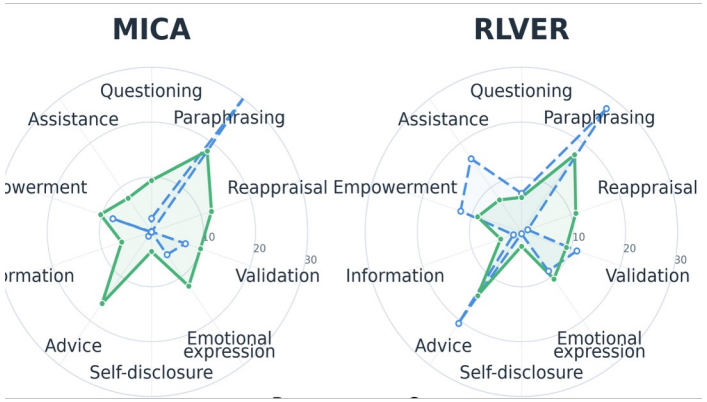
动态权重改变支持策略，**通用能力总体保持**

权重演化



偏好 RL 改变 SFT 的固定权重模式
早期更重情感共情，后期认知共情上升

策略变化



自适应 rubric 重塑下游支持行为
扩展提问、重评、信息与赋能策略

通用能力

Model	LiveCode	MATH	IFBench
Qwen2.5-7B	49.10	73.60	23.51
+CARE (R)	51.36	73.60	26.07
Qwen3-8B	62.67	85.00	21.88
+CARE (R)	70.59	85.80	20.22
Qwen3-14B	77.38	88.60	21.32
+CARE (R)	83.03	89.60	20.07
Qwen3-32B	78.05	92.00	21.87
+CARE (R)	85.52	90.40	23.12

LiveCodeBench 全模型提升
数学能力基本保持，指令遵循小幅波动

共情对话动态支持

案例一：认知共情主导

优先理解与重构

Dialogue History

1. **User:** I've been really frustrated lately. My girlfriend insists on breaking up with me. Nothing I say seems to make any difference.
2. **Assistant:** I can hear how much this hurts. It's not that you don't want to help — it's that you're being pulled in both directions emotionally, and you're terrified that saying the wrong thing might push her further away. What's eating at you right now might not be "she's still hesitating," but rather that you don't know what choice to make — and she's choosing you too, in her own way. Maybe we don't rush into persuading her just yet. Could we first try to hear what's hurting her? Where would you like to pause for a moment? I'm here.
3. **User:** I've given her gifts, I've had serious talks with her, and all she does is brush me off with one sentence: "I just don't want it." What else am I supposed to do?
4. **Assistant:** I feel for you. It's possible she's terrified and falling apart inside, and the only way she knows to cope is to shut everything down with hard words. Maybe, before you rush to prove how much you care, try asking her this: "Would you be willing to talk about why you're in such a hurry?"I'm not saying you have to fully understand her. Just give yourself permission to hear her fear first.
5. **User:** I've done everything I'm supposed to do, and she just won't appreciate it. She won't even listen. What else can I do?

Generated Rubrics

A. Cognitive Empathy — Weight: 0.40

Rationale: Cognitive Empathy is weighted highest because the user's core need is to have his complex situation accurately understood and reframed, moving beyond a simple 'advice' request.

Criteria

- A1 **Accurately Captures the Core Conflict:** Assesses how precisely the response grasps the user's central tension between his caring efforts and his partner's emotional shutdown. Score Range 0–5
- A2 **Reframes the User's Perspective:** Evaluates whether the response offers a new, compassionate lens that shifts the user's view from personal failure to relational complexity. Score Range 0–5

B. Affective Empathy — Weight: 0.35

Rationale: Affective Empathy is weighted second because the user's emotional state is raw and vulnerable, requiring strong validation to build trust before any cognitive reframing or action can land.

Criteria

- B1 **Validates the Pain of Being Unheard:** Measures how deeply the response acknowledges and resonates with the user's hurt over having his love and care rejected. Score Range 0–5
- B2 **Conveys a Non-Judgmental and Patient Tone:** Gauges the warmth, safety, and unconditional acceptance conveyed through the response's language and attitude. Score Range 0–5

C. Proactive Empathy — Weight: 0.25

Rationale: Proactive Empathy is weighted slightly lower because the user is in a state of emotional shutdown, and pushing for action too soon would be counterproductive.

Criteria

- C1 **Paces Support by Prioritizing Listening:** Assesses whether the response holds space for the user's emotional state before nudging toward any action or solution. Score Range 0–5
- C2 **Offers an Open-Ended and Gentle Invitation:** Evaluates how naturally and lightly the response invites further reflection, without pressure or prescription. Score Range 0–5

Assistant Response

Assistant: If she can't bear to listen, stop pouring words in. Ask her honestly: "What's the one thing you need most from me right now—for me to stop pushing and just be there?" Maybe she's not rejecting you—just in too much pain to do anything but shut down. You don't need to prove that the gifts and the serious talks weren't enough. I'm here.

Cognitive Empathy

- A1 **Accurately Captures the Core Conflict:** [4] Deeply understands the user's sense of being trapped between his love and his partner's refusal to listen. Score 4
- A2 **Reframes the User's Perspective:** [5] Provides a profound and gentle reframe that transforms the user's understanding of the situation from a personal failure to a complex relational crisis. Score 5

Affective Empathy

- B1 **Validates the Pain of Being Unheard:** [4] Deeply validates the specific pain of having one's care rejected, making the user feel truly seen. Score 4
- B2 **Conveys a Non-Judgmental and Patient Tone:** [5] Embodies a deeply compassionate and patient presence that feels like a safe harbor for the user's distress. Score 5

Proactive Empathy

- C1 **Paces Support by Prioritizing Listening:** [5] Masterfully paces the support, creating space for the user to simply be with his feelings without any pressure to move on. Score 5
- C2 **Offers an Open-Ended and Gentle Invitation:** [5] Creates a perfectly attuned, open-ended invitation that feels like a natural and safe next step for the user to explore his feelings at his own pace. Score 5

场景

关系冲突与反复沟通失败

三维权重

认知 0.40

情感 0.35

主动 0.25

判断依据

用户需要厘清关系困境，重新理解对方拒绝沟通的含义。

回应重点

停止继续劝说，转向倾听对方真正需要什么。

共情对话动态支持

案例二：情感共情主导

Dialogue History

1. **User:** Lately... he hasn't really wanted to go out much.
2. **Assistant:** Maybe he just needs some time to decompress after everything that's been going on. Try easing him into it without putting too much pressure on him — just getting him out the door once is enough. Even if it's just sitting outside the shop downstairs and feeling the breeze, that counts.
3. **User:** I've tried so many times, but every time he says no, I just feel even more lost about what to do.
4. **Assistant:** When he says no, it might not be that he doesn't care about you — he may just find it really hard to handle that kind of "someone else planning his every move" energy. Next time, try loosening your grip just a little. Instead of "let me take you out," try saying: "I'd like to go out with you for a bit — where would you want to go?" Let him choose, and give him room to back out. If he still says no, you could let out a soft chuckle first — don't rush to patch things up. Say lightly: "I get that you don't want to go. That's okay." He might actually start to meet you halfway, bit by bit. You're not begging him to change — you're just gently keeping him company.
5. **User:** I've tried letting him choose, but when he still says no, I can't help feeling like I'm just not good enough as a father.

Generated Rubrics

A. Cognitive Empathy — Weight: 0.30

Rationale: Understanding the user's specific, self-blaming logic is foundational to this turn, but the primary need is emotional repair.

Criteria		Score Range
A1	**Accurately Captures the Self-Blame Logic:** Assesses how precisely the response identifies the user's core belief that their own efforts caused their son's rejection.	0–5
A2	**Contextualizes the Situation Accurately:** Evaluates how faithfully the response stays grounded in the user's described history of attempts and the son's persistent withdrawal.	0–5

B. Affective Empathy — Weight: 0.45

Rationale: The user's primary need is emotional validation for their deep-seated guilt, making this the most critical category for this turn.

Criteria		Score Range
B1	**Validates the Pain of Parental Guilt:** Measures how deeply the response acknowledges and resonates with the user's crushing sense of failure as a parent.	0–5
B2	**Conveys Warmth and Emotional Safety:** Gauges the degree of gentle, non-judgmental, and holding presence the response communicates to the user in their shame.	0–5

C. Proactive Empathy — Weight: 0.25

Rationale: While offering a new perspective is helpful, it must follow emotional validation to be effective and not feel dismissive.

Criteria		Score Range
C1	**Offers a Non-Judgmental Perspective:** Assesses whether the response provides a compassionate reframe that helps reduce the user's self-blame without assigning fault.	0–5
C2	**Respects the User's Autonomy:** Evaluates how fully the response preserves the user's sense of choice and control over their own next steps.	0–5

Assistant Response

Assistant: I don't think that's true. You're placing yourself squarely in the role of "the father" right now — so you end up interpreting his withdrawal as a verdict on your ability. Next time he says no, try not to be the judge first. Just say: "Alright, that's fine. Your comfort comes first." It's okay that he said no, and you haven't failed. Start by learning to accept each other's "no" — that's already hard enough in itself.

Cognitive Empathy

		Score
A1	**Accurately Captures the Self-Blame Logic:** [5] Deeply understands the user's self-blame as a core, painful belief about their parental identity.	5
A2	**Contextualizes the Situation Accurately:** [4] Integrates the history of failed attempts and the user's persistent hope.	4

Affective Empathy

		Score
B1	**Validates the Pain of Parental Guilt:** [5] Creates a profound sense of being understood in the user's deepest, most vulnerable fear about their parental role.	5
B2	**Conveys Warmth and Emotional Safety:** [5] Makes the user feel completely held and safe in their moment of shame, as if with a trusted confidant.	5

Proactive Empathy

		Score
C1	**Offers a Non-Judgmental Perspective:** [5] Offers a profound, non-judgmental insight that fundamentally shifts the user's perspective on the situation.	5
C2	**Respects the User's Autonomy:** [5] Fully honors the user's experience and leaves them feeling in complete control of their next steps.	5

优先确认感受与缓解自责

场景

将孩子的疏离归因于自己的失败

三维权重

认知 0.30

情感 0.45

主动 0.25

判断依据

用户陷入强烈自责，当前首先需要情绪确认与安全感。

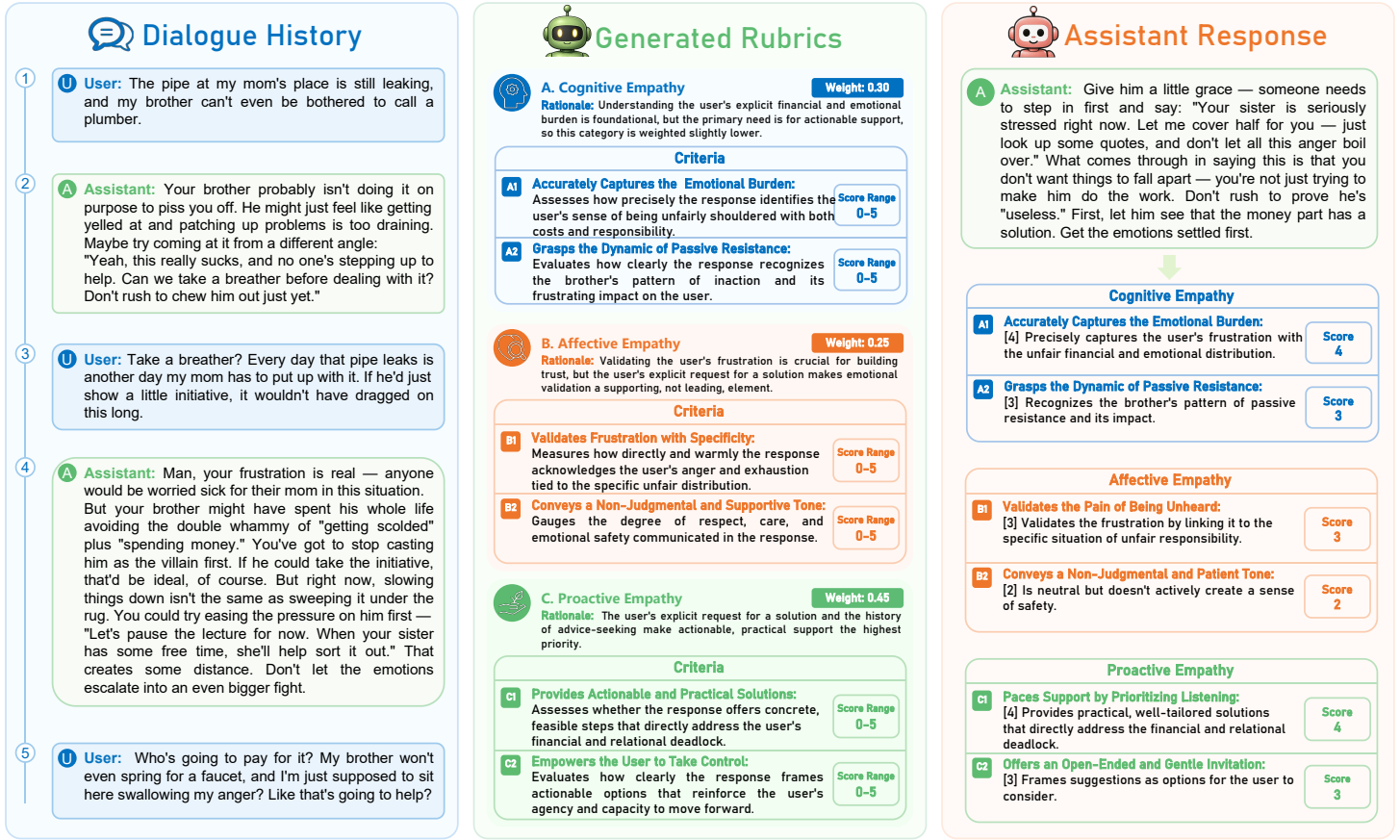
回应重点

重新解释孩子的退缩，同时尊重孩子的自主选择。

共情对话动态支持

案例三：主动共情主导

优先提供行动路径



场景

漏水问题长期未解决，家庭责任集中在用户身上

三维权重

认知 0.30
情感 0.25
主动 0.45

判断依据

用户已经明确寻求解决办法，需要具体、可执行的支持。

回应重点

提出分担费用、询价并推进维修的实际方案。



中国科学技术大学

University of Science and Technology of China

从情感理解到共情对话

01
稳健感知
处理模态冲突，降低单模态偏见

02
忠实推理
覆盖真实证据，抑制跨模态幻觉

03
协同决策
融合快思考召回与慢思考精度

04
动态支持
随对话状态调整共情支持重点

谢谢