USC Institute for Creative Technologies

ICLR 2026

AVERE: Improving Audiovisual Emotion Reasoning via Preference Optimization

Ashutosh Chaubey, Jiacheng Pang, Maksim Siniukov, Mohammad Soleymani

Institute for Creative Technologies, University of Southern California

Project Page: avere-iclr.github.io

chaubeyG/EmoReAlM ihp-lab/AVERE chaubeyG/AVERE-7B

Introduction & Context

ICLR 2026

Social AI Agents

Emotion understanding is a cornerstone for fostering meaningful human-machine interactions. It is critical for domains such as:

- **Healthcare:** Enhancing therapeutic alliance.
- **Education:** Improving adaptive learning outcomes.

Interpretability Gap

ICLR 2026

- Emotion classification in domains such as healthcare needs to be interpretable.
- Traditional methods for emotion classification rely on black-box classification.
- **Multimodal (or "*Omni*") LLMs** generate *explanations (reasoning)* in addition to their predictions and hence are more interpretable.

Multimodal LLMs are prone to perception ("hallucinations") and reasoning errors.

Bottlenecks in MLLM Reasoning

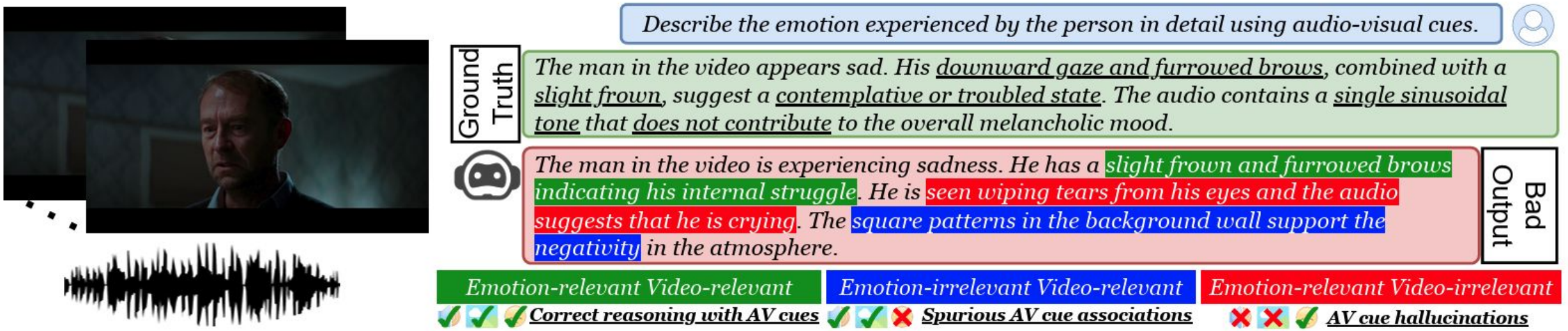

Figure 1: Existing MLLMs (i) include spurious associations between AV cues and emotions – *reasoning errors* (*blue highlight*) and (ii) hallucinate AV cues to explain emotions – *perception errors* (*red highlight*). AV: audiovisual.

EmoReAIM Benchmark

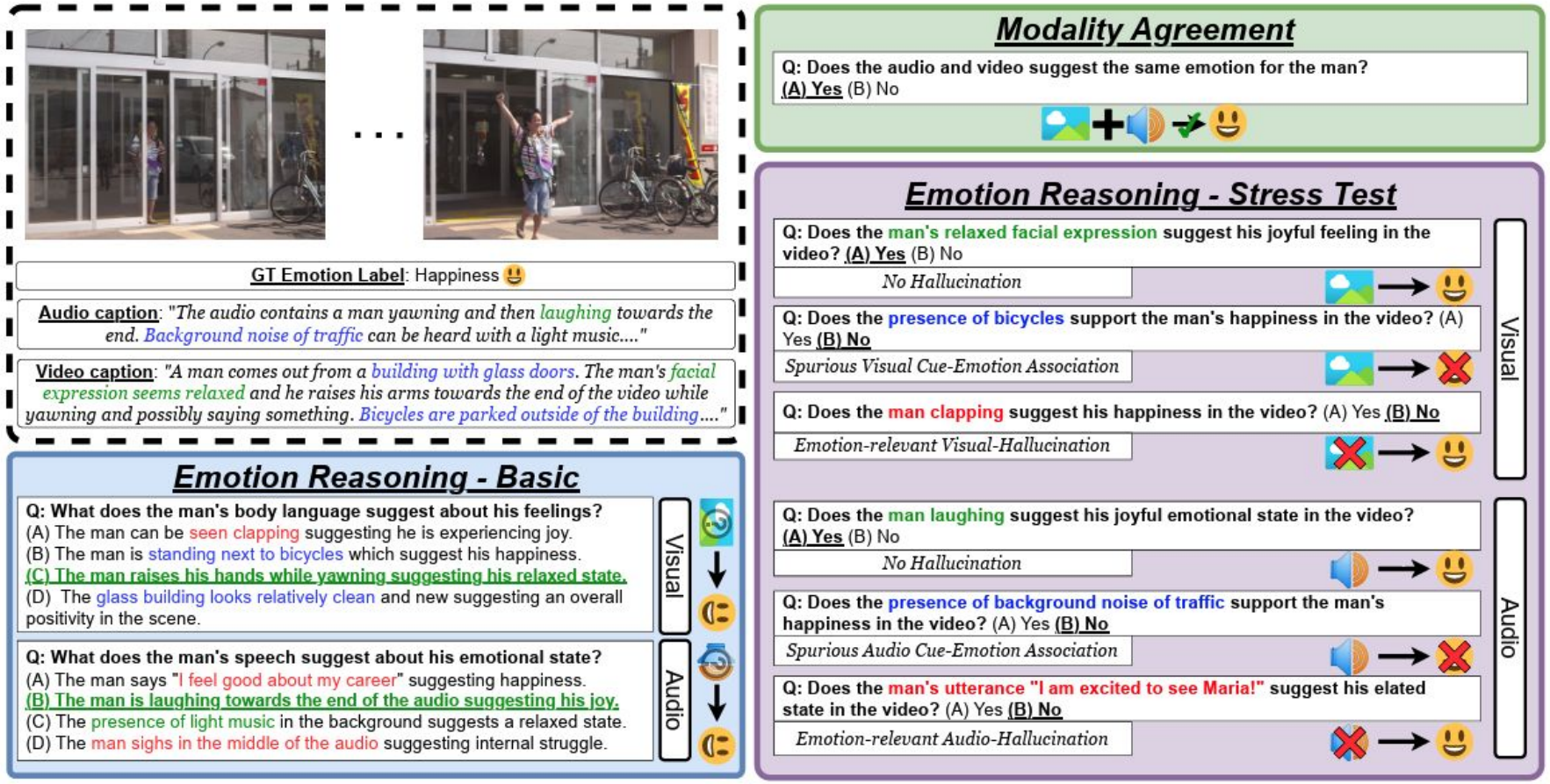

Figure 2: **EmoReAlM Tasks**. In addition to basic emotion reasoning, we include tasks for *Modality Agreement* and *Emotion Reasoning - Stress Test* to test spurious cue-emotion associations and cue hallucinations. Red text is a hallucinated cue, blue text is an emotion-irrelevant cue and green text is a cue relevant for emotion understanding. Correct choices are **underlined**.

Benchmark details

ICLR 2026

- Automatic annotation using existing videos from DFEW.
- Manually verified and filtered.
- All MCQs to ease evaluation.

Table 1: *EmoReAlM* Benchmark Statistics.

Task		# QA	# vid.	Rand. Acc.
Reasoning Basic	Audio	972	784	25%
	Visual	1024	883	25%
Modality Agreement		456	456	50%
Reas. Stress Test	Audio	820	655	50%
	Visual	728	593	50%
Total		**4000**	**2649**	

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, Jiateng Liu, "DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in-the-Wild,"in ACM Multimedia, 2020.

Proposed Method: AVEm-DPO

Audiovisual Direct Preference Optimization

A framework aligning model responses with audiovisual inputs via explicit preference pairs and text-prior penalties.

AVEm-DPO

ICLR 2026

Propose two types of preference optimization

- Prompt-based Modality Preference – Only perform multimodal DPO based on the modalities that the prompt is related to.
- Emotion-based Response Preference – Perform vanilla DPO by creating two types of negative (rejected) responses tailored to *perception* and *reasoning* errors.

AVEm-DPO

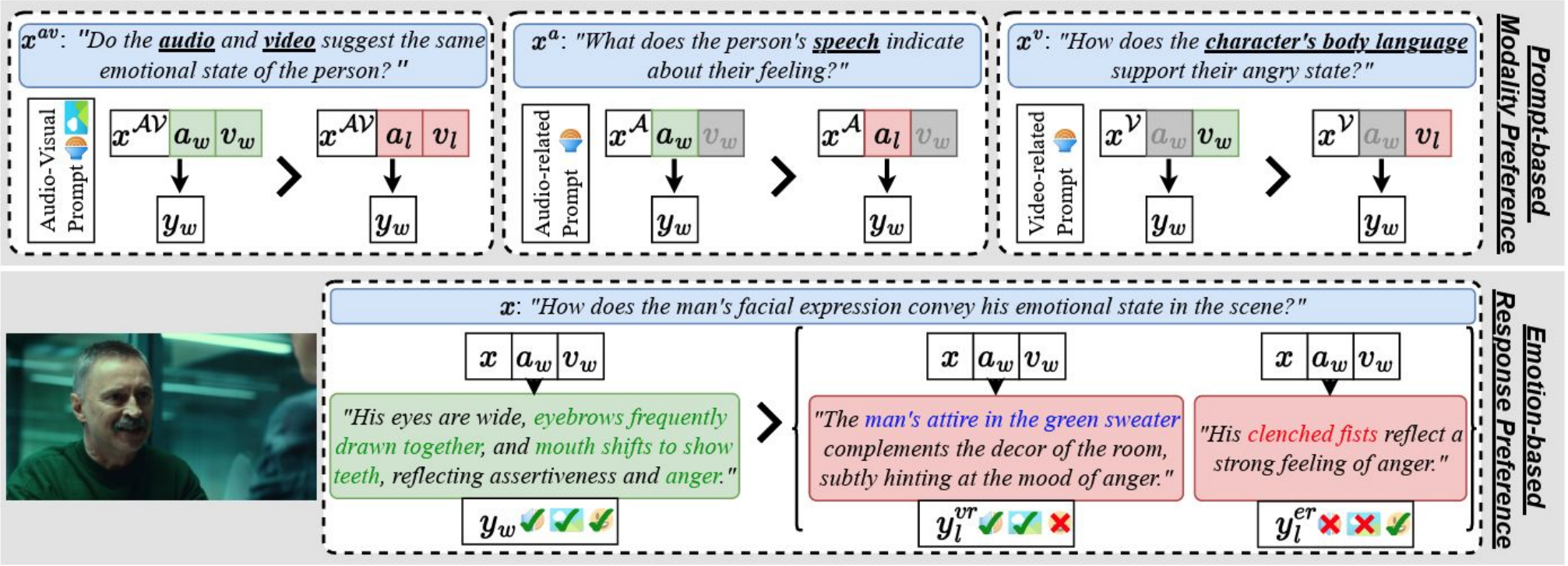

Figure 4: **Preference pairs in AVEm-DPO**. *(Top)* Fine-grained preference over modality input based on current prompt. *(Bottom)* Each chosen response y_w has two rejected responses – y_l^{vr} relevant to the video but with spurious emotion association and y_l^{er} irrelevant to the video (hallucinated) but related to the emotion.

Additional – Text Prior Debiasing (TPD)

ICLR 2026

The Bias Problem

LLMs rely on statistical correlations (e.g., "sad" implies "tears") regardless of visual evidence.

The Solution

We introduce a regularization term that penalizes the model if it generates a response that is highly probable given *only* the text prompt, effectively discounting text priors.

$$r(a, v, x, y) = \beta \log \frac{\pi_\theta(y \mid a, v, x)}{\pi_{\text{ref}}(y \mid a, v, x)} + \beta \log Z(a, v, x) - \gamma_{\text{TPD}} \log \pi_{\text{text}}(y \mid x)$$

Quantitative Results

ICLR 2026

Table 3: Performance comparison of different methods on the proposed *EmoReAlM* Benchmark.

Model	Reas. Basic		Modality Agree.	Reas. - Stress	
	Audio	Visual		Audio	Visual
	Acc.	Acc.	F1	F1	F1
VideoLLaMA2	63.1	66.8	52.5	53.2	58.4
OLA	63.2	60.4	42.7	56.6	54.8
VITA-1.5	63.1	84.3	30.2	52.8	56.3
Qwen 2.5 Omni	76.8	89.2	33.3	55.0	56.8
Our base	69.2	85.3	34.6	50.3	59.9
+ Naive-DPO	71.3	85.9	41.6	54.8	65.9
+ Vista-DPO†	72.4	87.8	52.1	73.6	86.7
+ AVEm-DPO	**77.9**	**92.5**	**60.0**	**80.9**	**94.6**
Emot.-LLaMA*	64.8	84.9	33.1	46.7	63.2
+ Naive-DPO	67.2	85.7	42.8	52.6	67.6
+ Vista-DPO†	69.0	86.9	40.9	68.6	87.3
+ AVEm-DPO	76.5	89.9	56.8	75.4	91.7

Reasoning improves classification

ICLR 2026

Table 2: Zero-shot performance comparison of different methods on existing audiovisual emotion recognition benchmarks. Mod. are the modalities input to the model with the prompt. A: Audio, V:Video, T: Text Subtitles. ‡: evaluation without text subtitle input.

Model	Mod.	DFEW		RAVDESS		MER2023	EMER			
		UAR	WAR	UAR	WAR	F1	Clue	Label	Spurious	Halluc.
VideoLLaMA 2	A,V	43.65	48.66	41.81	31.62	50.79	3.82	3.80	4.25	4.23
OLA	A,V	38.17	41.73	27.45	22.11	55.82	3.80	3.33	3.93	4.22
VITA-1.5	A,V	39.31	42.56	50.67	46.88	66.94	4.77	4.72	5.16	5.70
Qwen-2.5 Omni	A,V	46.94	54.34	32.88	28.05	79.72	5.85	6.78	6.39	6.21
EmotionLLaMA	A,V,T	45.59	59.37	28.20	29.24	90.36	6.03	6.99	5.89	5.26
EmotionLLaMA‡	A,V	42.72	54.06	30.36	30.45	89.05	2.76	2.78	3.44	2.36
MoSEAR	A,V,T	44.48	56.60	-	-	90.27	-	-	-	-
Our base	A,V	56.78	60.14	53.59	53.01	89.19	5.63	6.45	5.41	5.19
+ Naive-DPO		55.67	59.90	53.63	52.94	88.59	5.81	6.30	5.96	5.48
+ Vista-DPO†		56.42	62.33	56.94	53.64	90.06	6.08	6.89	6.58	6.07
+ AVEm-DPO		**58.54**	**64.24**	**58.66**	**55.48**	**92.18**	**6.37**	**7.08**	**7.09**	**6.75**
EmotionLLaMA⋆	A,V	54.89	58.26	52.59	48.12	90.01	5.78	6.21	5.36	5.23
+ Naive-DPO		54.97	58.12	52.69	49.01	89.35	5.89	6.35	5.89	5.62
+ Vista-DPO†		56.28	61.58	56.42	50.96	91.19	6.05	6.56	6.85	6.31
+ AVEm-DPO		57.06	62.12	56.21	51.03	91.68	6.02	6.99	7.02	6.62

ICLR 2026

Questions?

achaubey@usc.edu

Project Page: avere-iclr.github.io

chaubeyG/EmoReAlM ihp-lab/AVERE chaubeyG/AVERE-7B