

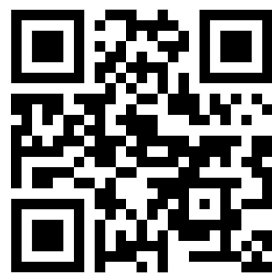


AVERE: Improving **A**udio**v**isual **E**motion **R**easoning via Preference Optimization

Ashutosh Chaubey, Jiacheng Pang, Maksim Siniukov, Mohammad Soleymani

Institute for Creative Technologies, University of Southern California

Project Page: [avere-iclr.github.io](https://github.com/chaubeyG/EmoReAlM)



Introduction & Context

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

- 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

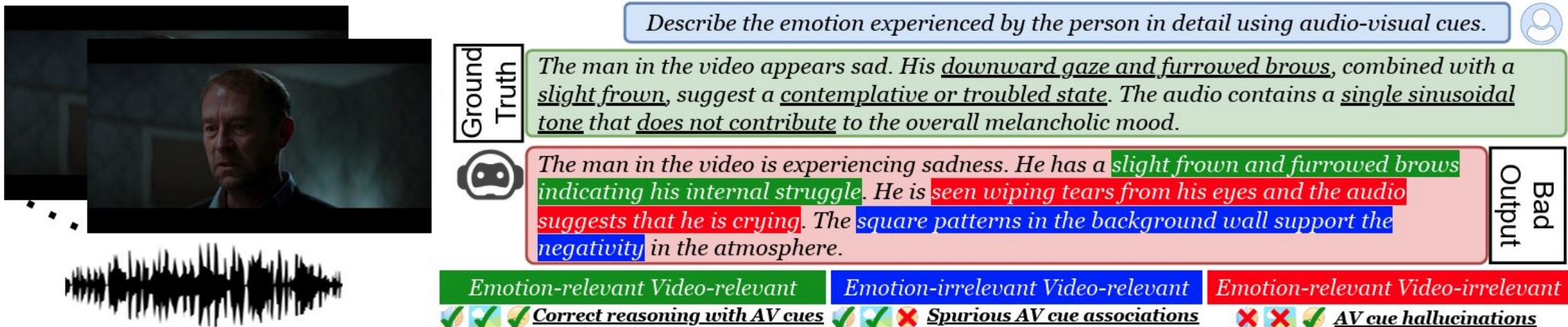



Figure 1: Existing MLLMs (i) include spurious associations between AV cues and emotions – *reasoning errors* (blue highlight) and (ii) hallucinate AV cues to explain emotions – *perception errors* (red highlight). AV: audiovisual.

EmoReAIM Benchmark



GT Emotion Label: Happiness 😊

Audio caption: "The audio contains a man yawning and then *laughing* towards the end. *Background noise of traffic* can be heard with a light music...."

Video caption: "A man comes out from a *building with glass doors*. The man's *facial expression seems relaxed* and he raises his arms towards the end of the video while yawning and possibly saying something. *Bicycles are parked outside of the building*...."

Modality Agreement

Q: Does the audio and video suggest the same emotion for the man?
(A) Yes (B) No

🌤️ + 🔊 → 🗑️ 😊

Emotion Reasoning - Basic

Q: What does the man's body language suggest about his feelings?
 (A) The man can be **seen clapping** suggesting he is experiencing joy.
 (B) The man is **standing next to bicycles** which suggest his happiness.
(C) The man raises his hands while yawning suggesting his relaxed state.
 (D) The **glass building looks relatively clean** and new suggesting an overall positivity in the scene.

Q: What does the man's speech suggest about his emotional state?
 (A) The man says "**I feel good about my career**" suggesting happiness.
(B) The man is laughing towards the end of the audio suggesting his joy.
 (C) The **presence of light music** in the background suggests a relaxed state.
 (D) The **man sighs in the middle of the audio** suggesting internal struggle.

Visual

↓

Audio

Emotion Reasoning - Stress Test

Q: Does the **man's relaxed facial expression** suggest his joyful feeling in the video? (A) Yes (B) No
 No Hallucination 🌤️ → 😊

Q: Does the **presence of bicycles** support the man's happiness in the video? (A) Yes (B) No
 Spurious Visual Cue-Emotion Association 🌤️ → ❌

Q: Does the **man clapping** suggest his happiness in the video? (A) Yes (B) No
 Emotion-relevant Visual-Hallucination ❌ → 😊

Q: Does the **man laughing** suggest his joyful emotional state in the video?
(A) Yes (B) No
 No Hallucination 🔊 → 😊

Q: Does the **presence of background noise of traffic** support the man's happiness in the video? (A) Yes (B) No
 Spurious Audio Cue-Emotion Association 🔊 → ❌

Q: Does the **man's utterance "I am excited to see Maria!"** suggest his elated state in the video? (A) Yes (B) No
 Emotion-relevant Audio-Hallucination ❌ → 😊

Visual

Audio

Figure 2: **EmoReAIM 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

- 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	

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

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

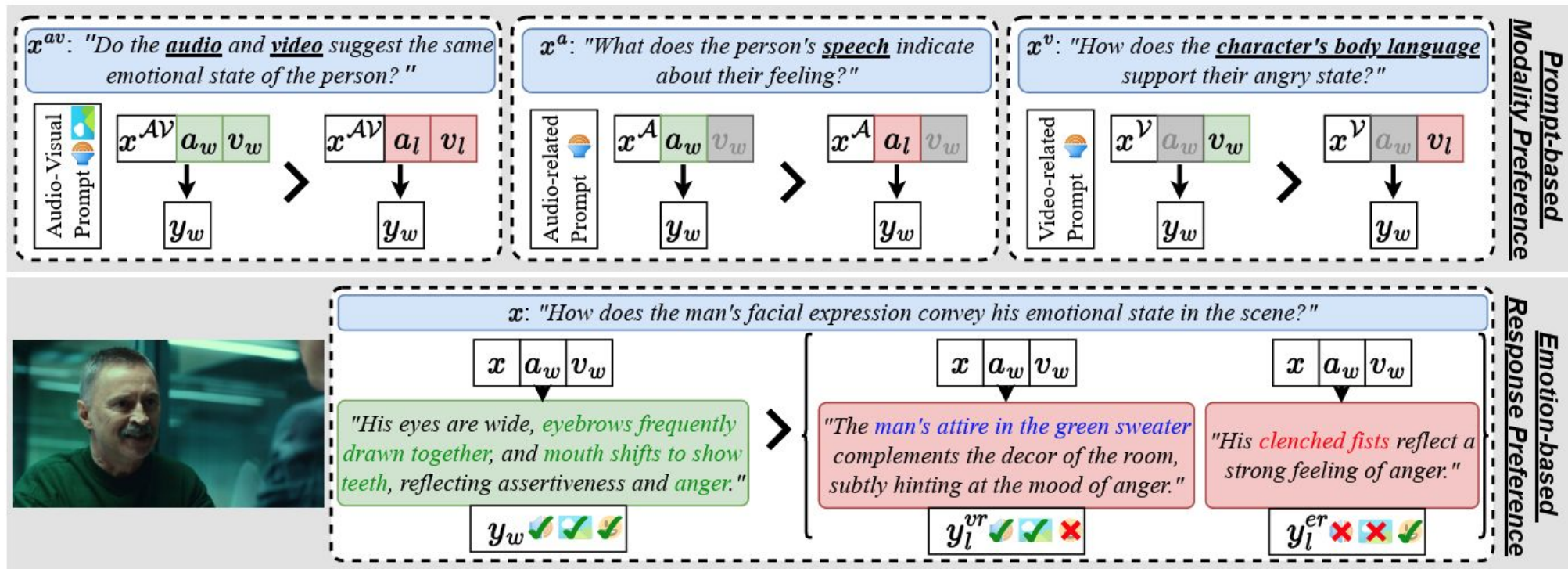


Figure 4: **Preference pairs in AVEm-DPO.** (Top) Fine-grained preference over modality input based on current prompt. (Bottom) Each chosen response y_w has two rejected responses – y_l^{vr} relevant to the video but with spurious emotion association and y_l^{er} irrelevant to the video (hallucinated)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

Additional – Text Prior Debiasing (TPD)

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.

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

Quantitative Results

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

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

Questions?

achaubey@usc.edu

Project Page: [averre-iclr.github.io](https://github.com/averre-iclr)



 [chaubeyG/EmoReAlM](https://github.com/chaubeyG/EmoReAlM)  [ihp-lab/AVERE](https://github.com/ihp-lab/AVERE)  [chaubeyG/AVERE-7B](https://github.com/chaubeyG/AVERE-7B)