AURORA: ACTION UNIT-REINFORCED LATENT REASONING FOR MULTIMODAL EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary multimodal emotion recognition (OV-MER) aims to predict flexible emotion expressions from video, audio, and language signals. Despite recent advances, existing OV-MER methods remain predominantly text-supervised: models are optimized to generate rationales and emotion labels, while reinforcement learning rewards only the final textual outputs. Such output-level supervision provides limited guidance for preserving fine-grained facial evidence, causing subtle muscle activations, temporal expression dynamics, and co-occurring facial movements to be diluted during language generation. We propose AURORA, a latent reasoning framework that explicitly grounds OV-MER in facial Action Units (AUs). AURORA combines frame-wise AU prediction with a continuous latent reasoning stage supervised to reconstruct the projected visual tokens of an AU-selected facial frame. This reconstruction objective encourages the latent states to retain facial evidence before emotion prediction. A second stage uses standard Group-Relative Policy Optimization (GRPO) with emotion-wheel, output-format, and AU-consistency rewards for task-level alignment. Experiments on OV-MERD+ and MER-UniBench demonstrate that AURORA achieves state-of-the-art performance, reaching a mean score of 84.87 on MER-UniBench when built on Qwen3-VL-8B.

1 INTRODUCTION

Human emotion is conveyed through multimodal signals, including facial expressions, speech, and language (Poria et al., 2017; Lian et al., 2023; 2024a; Xie et al., 2024). Among them, subtle facial behaviors often provide critical affective evidence that may not be explicitly reflected in spoken content. Traditional multimodal emotion recognition (MER) formulates this task as closed-set classification over predefined emotion categories (Busso et al., 2008; Poria et al., 2019; Jiang et al., 2020), which struggles to capture the mixed, subtle, and context-dependent nature of real-world emotions.

Open-vocabulary multimodal emotion recognition (OV-MER) addresses the limitations of closed-set emotion classification by predicting free-form emotion words (Lian et al., 2024b; 2025). Current OV-MER frameworks, typically built upon Multimodal Large Language Models (MLLMs) (Peng et al., 2026; Lian et al., 2025), often rely on textual rationales to bridge raw sensory signals and emotion understanding. Although such rationales improve interpretability, language remains a lossy interface for subtle facial evidence. Fine-grained facial behaviors, including transient muscle activations and expression co-occurrences, are difficult to faithfully preserve once compressed into text, leading to over-reliance on linguistic priors rather than rigorous visual grounding.

Figure 1 illustrates this limitation with a representative example. The ground-truth emotion set of the video is *comforted, joyful, caring, pleased, happy*. However, existing emotion MLLMs such as AffectGPT (Lian et al., 2024b), Emotion-LLaMA (Cheng et al., 2024), and Emotion-LLaMAv2 (Peng et al., 2026) instead predict negative emotions dominated by *sadness*. Their predictions rely on coarse textual rationales, such as "somber expression", or hallucinated visual cues, despite the reassuring subtitle and smile-related facial evidence. This example highlights a key limitation of OV-MER: natural-language rationales are often too coarse to faithfully capture subtle facial configurations.

We therefore ask: *How can an OV-MER model preserve and leverage the granular facial evidence that eludes verbalization to facilitate precise open-vocabulary emotion reasoning?*

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.

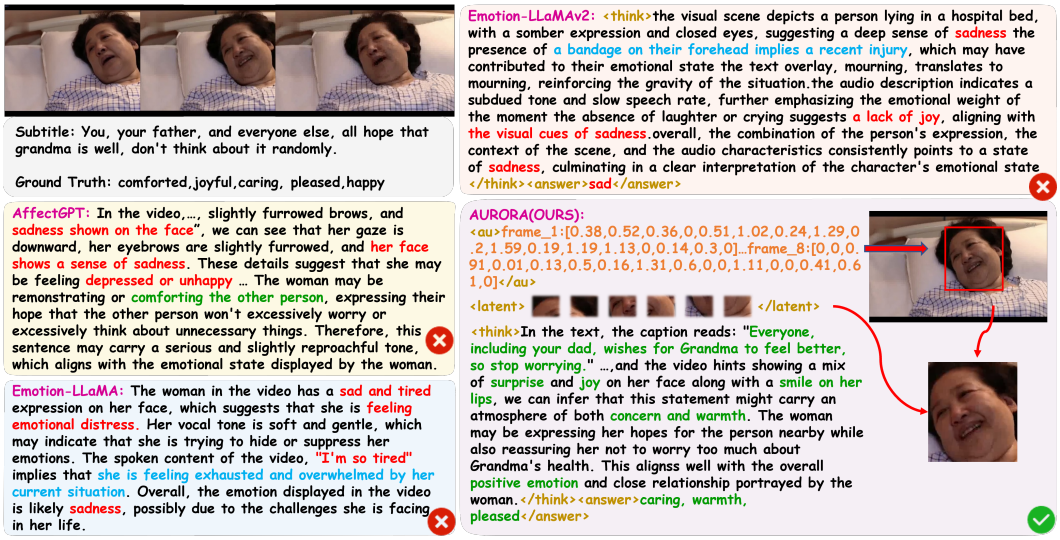


Figure 1: An OV-MER example comparing AURORA with existing emotion MLLMs. Red marks incorrect information, green marks correct evidence, and blue marks hallucinations.

We argue that facial Action Units (AUs), which taxonomize visible facial muscle activations, provide a subsymbolic yet grounded representation for affective behavior analysis (Ekman & Friesen, 1978; Cohn et al., 2007; Baltrušaitis et al., 2016). Unlike discrete textual rationales, frame-wise AU intensities capture the continuous dynamics of facial appearance, offering a more faithful bridge to high-level emotion concepts. We therefore propose utilizing AUs as fine-grained grounding signals for OV-MER. In a video context, these intensity streams provide dense supervision that closely tracks facial motion, effectively mitigating the linguistic bottleneck. Furthermore, their temporal profiles allow us to select expression-salient peaks.

Based on this observation, we propose **AURORA**, an Action Unit-Reinforced Latent Reasoning framework for OV-MER. AURORA first generates frame-wise AU traces, transforming subtle facial dynamics into structured intermediate supervision before emotion prediction. Beyond explicit AU perception, AURORA introduces a latent reasoning trajectory supervised to reconstruct the visual tokens of an AU-salient facial key frame, enabling the model to preserve fine-grained expression evidence that is difficult to verbalize. We then use standard GRPO with AU-aware task rewards for second-stage optimization of response generation.

Experiments on OV-MERD+ and MER-UniBench demonstrate the effectiveness of AURORA across modality settings and backbone scales.In the vision-audio-text setting, AURORA reaches 84.27% with Qwen2.5-VL-7B and further achieves 84.87% with the stronger Qwen3-VL-8B backbone. These results show that Action Unit-Reinforced Latent Reasoning is complementary to both audio cues and stronger multimodal backbones. Our main contributions are summarized as follows:

- We introduce a key-frame reconstruction mechanism for affective latent reasoning. Unlike existing methods that rely on sparse textual rationales, AURORA preserves holistic, subsymbolic facial configurations by supervising a latent reasoning trajectory to reconstruct the projected patch-token sequence of the AU-salient facial key frame at AU-defined expression peaks.

- We integrate AU-consistency supervision into standard GRPO for second-stage task alignment. Together with emotion-wheel and structural rewards, the AU-distance reward encourages generated AU traces and emotion predictions to remain aligned with observable facial behavior.

- Extensive evaluations on OV-MERD+ and MER-UniBench demonstrate the effectiveness of AURORA across modality settings and backbone scales. It improves Qwen2.5-VL-7B by **7.52 mean-score points** in the vision-text setting, reaches **84.27%** with audio on Qwen2.5-VL-7B, and further achieves **84.87%** with the stronger Qwen3-VL-8B backbone.

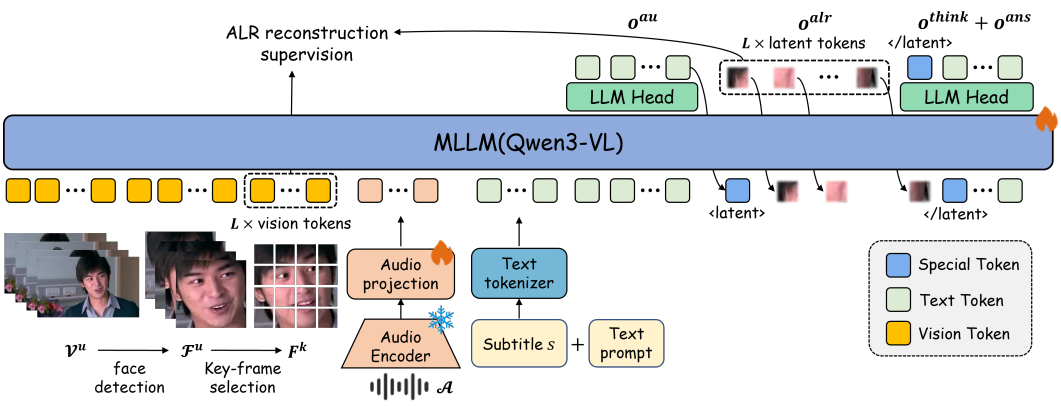


Figure 2: Overview of **AURORA**: multimodal encoding, AU-trace generation, latent reasoning, and textual rationale and emotion prediction. Facial patches illustrate the visual-token reconstruction targets.

2 Related Work

Multimodal and Open-vocabulary Emotion Understanding. Multimodal emotion recognition integrates facial, acoustic, and linguistic cues for sentiment and emotion prediction (Baltrušaitis et al., 2018; Zhao et al., 2021; Hu et al., 2024). Earlier systems emphasize closed-set multimodal fusion (Tsai et al., 2019; Hazarika et al., 2020; Shi & Huang, 2023). Recent models, including Emotion-LLaMA, AffectGPT, R1-Omni, and Emotion-LLaMAv2, generate emotion descriptions and rationales (Cheng et al., 2024; Lian et al., 2024b; Zhao et al., 2025; Peng et al., 2026). OV-MER supports variable-size emotion sets evaluated through emotion-wheel metrics, as in OV-MERD and AffectGPT-R1 (Lian et al., 2024b; 2025).

Latent Reasoning and Reinforcement Learning in MLLMs. Textual reasoning and reinforcement learning guide generation in modern MLLMs (Wei et al., 2022; Wang et al., 2022; Shao et al., 2024). Approaches such as Insight-V, Mirage, and Latent Visual Reasoning introduce intermediate visual computation beyond textual rationales (Dong et al., 2025; Yang et al., 2025; Li et al., 2025). These task-general methods typically align intermediate representations with visual evidence or helper abstractions.

Facial Action Units for Affective Grounding. Facial AUs describe muscle actions under FACS (Ekman & Friesen, 1978). OpenFace and subsequent AU-recognition systems support automatic analysis of facial activity (Baltrušaitis et al., 2016; Luo et al., 2022; Wang et al., 2024; Ge et al., 2024). Prior emotion systems typically use AUs as auxiliary predictors, preprocessing cues, or peak-frame heuristics.

3 Our Method: AURORA

Figure 2 illustrates the overall architecture of our proposed AURORA.

Task Formulation. Given a video clip, OV-MER aims to generate a flexible set of open-form emotion descriptors that capture nuanced affective states. We denote an input sample as $x = (\mathcal{V}, \mathcal{A}, s)$, where $\mathcal{V}$ is the visual stream, $\mathcal{A}$ is the audio stream, and s is the subtitle or speech transcript. The ground-truth emotion annotation is a set $y = \{e_1, e_2, \ldots, e_{|y|}\}$, where each e_i is an emotion word or phrase. The default response is $o = (o^{\text{au}}, o^{\text{alr}}, o^{\text{think}}, o^{\text{ans}})$, where o^{au} is a frame-wise AU trace enclosed by `<au>` and `</au>`, o^{alr} is an internal latent reasoning segment enclosed by `<latent>` and `</latent>`, o^{think} is a textual rationale enclosed by `<think>` and `</think>`, and o^{ans} is the final open-vocabulary emotion prediction enclosed by `<answer>` and `</answer>`. Rationale and answer generation are retained in all reported AURORA configurations and internal ablations. The final prediction is obtained by parsing o^{ans} into a set $\hat{y}$.

AU-Guided Facial Evidence Extraction. To capture fine-grained facial dynamics while maintaining computational efficiency, we uniformly sample M frames from the video. For each frame, we perform face detection and alignment to normalize pose and scale variations, yielding a standardized

sequence $\mathcal{F}^u = \{F_1, F_2, \ldots, F_M\}$. We employ OpenFace (Baltrušaitis et al., 2016) to estimate the AU intensity vector for each face frame. For each aligned face F_m, we extract a J-dimensional intensity vector $A_m = [a_{m,1}, a_{m,2}, \ldots, a_{m,J}] \in \mathbb{R}^J$, where $a_{m,j}$ quantifies the activation of the j-th AU. To identify an AU-salient moment for latent reasoning, we determine an expression-salient key frame based on the cumulative AU activation. We define the key frame index $m^\star$ by maximizing the aggregate intensity across all detected AUs: $m^\star = \arg\max_{m\in\{1,\ldots,M\}} \sum_{j=1}^{J} a_{m,j}$, and $F^k = F_{m^\star}$. This deterministic rule is a simple, reproducible key-frame selection heuristic that prioritizes strong aggregate AU activation. The selected frame supplies expression-focused input tokens and the latent reconstruction target, while the original and aligned face sequences retain contextual and temporal information. Table 2 supports its effectiveness under the evaluated settings. For subtle, mixed, or dynamically evolving emotions, relevant evidence may be distributed across frames rather than concentrated at the peak of aggregate AU activity.

Multimodal Encoding. AURORA adopts the foundational Encoder-Projector-LLM architecture common to modern MLLMs (Bai et al., 2025; Liu et al., 2023) to map heterogeneous inputs into a unified multimodal embedding space. We denote the uniformly sampled original video frames as $\mathcal{V}^u = \{V_1, \ldots, V_M\}$. Specifically, a visual encoder E_v and a projector P_v extract feature tokens from three distinct perspectives: $T^v = P_v(E_v(\mathcal{V}^u)), T^f = P_v(E_v(\mathcal{F}^u)), T^k = P_v(E_v(F^k))$. The visual encoder E_v and projector P_v remain frozen throughout SFT and GRPO. Here, T^v provides global context including scene and pose; T^f preserves local facial dynamics across the sampled sequence; and T^k encodes the AU-salient key frame used for affective latent reasoning. For the acoustic modality, we use HuBERT (Hsu et al., 2021) and an audio projector P_a to transform the input signal $\mathcal{A}$ into prosodic tokens $T^a = P_a(E_a(\mathcal{A}))$, providing paralinguistic cues complementary to the visual evidence. Finally, these multimodal tokens are integrated with the textual embeddings of the transcript s and task instruction q, processed via the language model's embedding layer E_t, through a fusion and alignment module $\Phi(\cdot)$ to form the emotion-aware context: $C = \Phi(T^v, T^f, T^k, T^a, E_t(s), E_t(q))$. The AU-salient tokens T^k are included in the input context during both training and inference. During SFT, they additionally provide the visual reconstruction target for ALR. In the vision-text setting, T^a is omitted.

Affective Latent Reasoning. The cornerstone of AURORA is AU-reinforced Affective Latent Reasoning (ALR), which establishes a dedicated internal visual workspace between AU perception and emotion reasoning.

For each sample, the AU-salient key frame F^k is resized to a fixed resolution and encoded by the visual encoder and projector:

$$K = P_v(E_v(F^k)) = [k_1, k_2, \ldots, k_L] \in \mathbb{R}^{L\times d}, \tag{1}$$

where $K = T^k$, d denotes the hidden dimension of the language model, and L is the number of projected patch tokens. The consistent resolution of key frames yields a fixed number of visual tokens, enabling a unified ALR length for stable optimization and efficient inference.

To facilitate reasoning within a non-symbolic visual workspace, ALR is implemented as a fixed-length internal computation rather than a sequence of discrete tokens. The transition to this mode is triggered by a special token `<latent>`, shifting the decoder from standard linguistic mapping to a continuous latent reasoning state. Over a predefined trajectory of L steps, the model performs iterative state updates without projecting to the vocabulary space. Specifically, at each step $\ell \in \{1, ..., L\}$, the decoder generates a hidden state $h_\ell^{\text{alr}} \in \mathbb{R}^d$. Unlike standard decoding where states are mapped to text, h_ℓ^{alr} is directly fed forward as the input representation for the subsequent step. This mechanism enables the model to refine its internal representation of subtle affective evidence within the hidden manifold, eventually terminating the process with a `</latent>` token to resume textual output. Concretely, let h_0^{alr} be the final-layer hidden state at `<latent>`. At latent step ℓ, we replace the next input embedding with $h_{\ell-1}^{\text{alr}}$ and run the decoder with its existing causal attention and KV-cache to obtain h_ℓ^{alr}. Reserved placeholder IDs track these positions, while their vocabulary logits are ignored. After exactly L steps, the model inserts `</latent>` and resumes text decoding with the `<think>` rationale followed by the `<answer>` block. The cached keys and values of the latent positions remain available to subsequent rationale and answer tokens. The resulting latent trajectory is defined as:

$$H^{\text{alr}} = [h_1^{\text{alr}}, h_2^{\text{alr}}, \ldots, h_L^{\text{alr}}]. \tag{2}$$

Through this tripartite architecture, the AU block exposes frame-wise dynamics, the ALR trajectory preserves peak-frame semantics, and the final textual output translates this multi-grained evidence into nuanced emotion predictions.

4 Training and Optimization

AURORA is optimized in two stages: Supervised Fine-tuning (SFT) for AU-grounded latent representation learning, followed by standard Group-Relative Policy Optimization (GRPO) for task-level alignment.

4.1 Supervised Fine-tuning

SFT establishes the foundational mapping between physical facial behavior and affective semantics. The overall objective is defined as:

$$\mathcal{L}_{\mathrm{SFT}} = \mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{alr}} \mathcal{L}_{\mathrm{ALR}}. \tag{3}$$

Generative Grounding. We optimize all discrete textual components in the output sequence using the standard next-token prediction objective. The supervised tokens include the structured AU trace, format tags, textual rationale, and final emotion predictions:

$$\mathcal{L}_{\mathrm{gen}} = -\sum_{t \in \mathcal{T}} \log p_\theta \left(y_t \mid y_{<t}, C, H^{\mathrm{alr}}_{<t} \right), \tag{4}$$

where $\mathcal{T}$ denotes the supervised discrete-token positions and $H^{\mathrm{alr}}_{<t}$ denotes the latent states preceding position t. This latent prefix is empty during AU-trace generation and complete during rationale and answer generation. By serializing AU intensities into structured text, the model is encouraged to explicitly verbalize frame-wise facial muscle activations before affective reasoning. Conditioning rationale and answer generation on the latent trajectory H^{alr} makes the preceding latent facial representations available during textual reasoning and emotion prediction.

Latent Reconstruction. To preserve subtle facial evidence beyond discrete language space, we supervise the latent trajectory H^{alr} to reconstruct the projected visual semantics of the AU-salient key frame F^k.

The ALR positions carry continuous hidden states and are excluded from the discrete-token prediction loss. During SFT, we impose a patch-level reconstruction loss at these positions:

$$\mathcal{L}_{\mathrm{ALR}} = \frac{1}{L} \sum_{\ell=1}^{L} \left\| h^{\mathrm{alr}}_\ell - k_\ell \right\|_2^2, \tag{5}$$

where k_ℓ denotes the projected patch token of the key frame F^k. The frozen visual encoder and projector provide fixed reconstruction targets, so this loss updates the decoder without changing the target representations. This objective encourages the latent workspace to remain aligned with the projected patch tokens of the selected facial frame.

4.2 Reinforcement Learning via GRPO

For second-stage task alignment, we apply standard Group-Relative Policy Optimization (GRPO) (Shao et al., 2024) with rewards tailored to OV-MER. SFT learns the AU-grounded latent representation through generation and reconstruction supervision; GRPO then optimizes the resulting policy with task-level rewards. The reconstruction loss $\mathcal{L}_{\mathrm{ALR}}$ is used during SFT, while the RL stage uses the composite reward defined below and retains the same L-step latent computation.

Group Sampling with ALR. For each input sample, the old policy $\pi_{\theta_{\mathrm{old}}}$ generates a group of G complete response trajectories:

$$\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid x). \tag{6}$$

Each trajectory contains the AU trace, latent ALR segment, textual rationale, and final emotion prediction. During rollout, the model enters the continuous ALR mode after generating `<latent>`

and performs exactly L latent reasoning steps before generating the rationale and answer. These latent states are deterministic internal computations that condition the subsequent output-token probabilities. Rollouts are generated without gradient tracking. We retain the generated token sequence and the old-policy log-probabilities at sampled discrete-token positions for the policy update. The trajectory-level reward evaluates the generated AU trace, output format, and emotion answer.

Multi-faceted Reward Design. We define a composite reward:

$$R(o_i) = \alpha P(o_i^{\text{ans}}, y) R_{\text{EW}} + \beta R_{\text{fmt}} + \eta R_{\text{AU}}, \tag{7}$$

which jointly evaluates semantic correctness, structural validity, perceptual grounding, and prediction conciseness. **(1) Emotion Fidelity (R_{EW}).** Measures semantic agreement between the predicted emotion set and the ground-truth annotation using the official OV-MER emotion-wheel protocol. **(2) Structural Validity (R_{fmt}).** Verifies whether the response follows the required structured format, including valid output tags and properly formatted AU traces. **(3) Perceptual Accuracy (R_{AU}).** Encourages accurate AU perception by comparing generated AU intensities $\hat{a}_{m,j}$ against OpenFace annotations:

$$R_{\text{AU}} = 1 - \frac{1}{MJ} \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{|\hat{a}_{m,j} - a_{m,j}|}{B_j}. \tag{8}$$

(4) Conciseness (P). Let n_i be the number of nonempty comma-separated emotion labels parsed from o_i^{ans}, and let n_y be the number of ground-truth emotion labels. The counts are taken before emotion-wheel grouping:

$$P(o_i^{\text{ans}}, y) = \begin{cases} 1, & n_i \le n_y, \\ \dfrac{n_y}{n_i}, & n_i > n_y. \end{cases} \tag{9}$$

This factor scales the EW reward according to the final answer's label count, while AU-trace and rationale token lengths are excluded from the count.

Group-relative Advantage. Rewards are normalized within each sampled group:

$$\hat{A}_i = \frac{R(o_i) - \text{mean}(\{R(o_g)\}_{g=1}^{G})}{\text{std}(\{R(o_g)\}_{g=1}^{G}) + \epsilon_{\text{adv}}}. \tag{10}$$

Here, $\epsilon_{\text{adv}} > 0$ is a numerical stabilizer. The same response-level advantage $\hat{A}_i$ is assigned to each sampled discrete token in trajectory i.

Policy Update. To evaluate current-policy token log-probabilities, we teacher-force the sampled discrete sequence and recompute the complete L-step ALR trajectory using the current parameters θ. Let $B_{\theta,i} = [b_{\theta,i,1}, \ldots, b_{\theta,i,L}]$ denote the resulting latent inputs, where $b_{\theta,i,\ell} = h^{\text{alr}}_{\theta,i,\ell-1}$ and $h^{\text{alr}}_{\theta,i,0}$ is the hidden state at `<latent>`. Each latent input is obtained from the preceding hidden state in the same differentiable forward computation. The current-policy trajectory and its attention keys and values are recomputed on every policy-update pass, with the computation graph preserved across all latent steps. We maximize the standard token-level clipped GRPO objective with KL regularization:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|\mathcal{D}_i|} \sum_{t \in \mathcal{D}_i} \left\{ \min\left(r_{i,t}(\theta)\hat{A}_i, \bar{r}_{i,t}(\theta)\hat{A}_i \right) - \beta_{\text{KL}} \widehat{D}^{\text{KL}}_{i,t} \right\} \right], \tag{11}$$

where $\bar{r}_{i,t}(\theta) = \text{clip}\big(r_{i,t}(\theta), 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}\big)$ and

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid x, o_{i,<t}, B_{\theta,i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid x, o_{i,<t}, B_{\theta_{\text{old}},i,<t})}. \tag{12}$$

Here, $\mathcal{D}_i$ contains the sampled discrete-token positions, including the AU trace, generated format tags, textual rationale, and emotion answer; forced latent placeholders and the forced `</latent>` marker are excluded. The discrete prefix $o_{i,<t}$ includes preceding format markers, and $B_{\theta,i,<t}$ contains only latent inputs preceding position t: it is empty during AU generation and complete during rationale and answer generation. The numerator and denominator share the same sampled discrete prefix, while each policy uses the latent trajectory determined by its own parameters. The denominator is retained from the old-policy rollout. The clipping threshold ϵ_{clip} is distinct from ϵ_{adv}.

Table 1: Main results on MER-UniBench under vision-text and audio-vision-text settings.

Method	Backbone	Sentiment				Basic Emotion				Fine-grained	Mean
		MOSI	MOSEI	SIMS	SIMS v2	MER2023	MER2024	MELD	IEMOCAP	OV-MERD+	
Input Modality: Video, Text											
AffectGPT	Qwen2.5-7B	82.39	81.57	87.20	86.29	74.58	75.29	57.63	62.19	61.65	74.31
AffectGPT-R1	Qwen2.5-7B	78.78	79.07	85.91	85.85	77.72	85.29	61.09	67.42	62.42	75.95
Baseline	Qwen2.5-VL-7B	78.01	80.23	83.96	84.10	76.71	77.73	56.59	58.74	58.85	72.77
AURORA(SFT)	Qwen2.5-VL-7B	84.57	84.46	88.50	87.06	81.72	81.41	64.57	64.22	62.37	77.65
AURORA(SFT+RL)	Qwen2.5-VL-7B	84.82	85.21	88.97	88.56	83.58	87.91	69.59	70.46	63.52	80.29
AURORA(SFT+RL)	**Qwen3-VL-8B**	**85.27**	**85.33**	**89.90**	**88.86**	**84.23**	**88.27**	**69.84**	**71.51**	**64.43**	**80.85**
Input Modality: Audio, Video, Text											
AffectGPT	Qwen2.5-7B	81.30	80.90	88.49	86.18	78.54	78.80	55.65	60.54	62.52	74.77
AffectGPT-R1	Qwen2.5-7B	79.65	80.18	87.26	85.75	84.51	93.13	66.71	74.26	68.39	79.98
Emotion-LLaMA	Llama2	66.13	67.66	78.32	77.23	59.38	73.62	46.76	55.47	52.97	64.17
Emotion-LLaMAv2 (Stage 1)	Llama2	86.28	87.69	87.72	86.02	77.28	86.90	51.32	84.05	62.90	78.91
Emotion-LLaMAv2 (Stage 2)	Llama2	86.96	86.15	86.35	87.42	76.72	86.90	48.61	83.60	64.00	78.52
AURORA(SFT)	Qwen2.5-VL-7B	86.82	85.51	88.17	86.89	82.85	83.52	59.73	68.65	63.00	78.35
AURORA(SFT+RL)	Qwen2.5-VL-7B	87.83	87.85	89.28	88.49	85.81	94.97	70.51	84.20	69.47	84.27
AURORA(SFT+RL)	**Qwen3-VL-8B**	**88.67**	**88.23**	**89.68**	**89.01**	**86.11**	**95.07**	**71.80**	**85.43**	**69.87**	**84.87**

The reference policy π_{ref} is a frozen copy of the SFT policy and uses its own deterministic latent computation when scoring the sampled discrete prefix. Let $\ell_{i,t}^{\theta}$ and $\ell_{i,t}^{\text{ref}}$ denote the current-policy and reference-policy log-probabilities of the sampled token under the same input and discrete prefix. Following GRPO (Shao et al., 2024), the token-level KL penalty is estimated as

$$\widehat{D}_{i,t}^{\text{KL}} = \exp\big(\ell_{i,t}^{\text{ref}} - \ell_{i,t}^{\theta}\big) - \big(\ell_{i,t}^{\text{ref}} - \ell_{i,t}^{\theta}\big) - 1. \tag{13}$$

The KL term regularizes the policy toward the SFT reference in the optimization objective. Its weight β_{KL} is separate from the format-reward weight β. Gradients from the sampled-token log-probabilities propagate through the current decoder, including the recomputed hidden states, their feedback into subsequent input embeddings, and the associated causal attention computations. The sampled token IDs, old-policy log-probabilities, and group-relative advantages are held fixed during optimization. Latent positions carry no independent action probabilities or direct policy-loss terms, but their computations receive gradients through the probabilities of subsequent sampled tokens. This allows the GRPO objective to jointly update latent computation and response generation toward task-level rewards.

5 Experiments

5.1 Experiment Setup

Datasets. Following AffectGPT-R1 (Lian et al., 2025), we use MER-Caption+ (Lian et al., 2024b) for SFT and MER2025-OV for GRPO, removing its overlap with OV-MERD+. MER-UniBench evaluates sentiment (WAF), basic emotion (HIT), and fine-grained OV-MER (emotion-wheel F_s). Sentiment and Basic each average four datasets, Fine-grained denotes OV-MERD+, and Mean averages all nine. Dataset statistics, implementation details, and metrics appear in Appendices A, B, and C.

Baselines. We compare AffectGPT (Lian et al., 2024b), AffectGPT-R1 (Lian et al., 2025), Emotion-LLaMA (Cheng et al., 2024), Emotion-LLaMAv2 (Peng et al., 2026), and the corresponding unmodified backbones. Internal ablations examine visual inputs, AU supervision, ALR, and reward design.

5.2 Main Results on MER-UniBench

Table 1 shows that AURORA achieves a mean score of 80.29 with Qwen2.5-VL-7B in the vision-text setting, improving the backbone by 7.52 points. GRPO further improves the SFT model from 77.65 to 80.29, with gains spanning all three task groups. Adding HuBERT audio features raises the mean score to 84.27, while Qwen3-VL-8B reaches 84.87 in the audio-video-text setting. These results support the value of AU-grounded latent reasoning across modality settings and backbone scales.

5.3 Ablation Study

Unless otherwise specified, the ablations use Qwen2.5-VL-7B with vision-text inputs.

Table 2: Key-frame selection under a fixed training and input configuration.

Selector	Sentiment	Basic	OV-MERD+	Mean
Random frame	84.23	73.05	61.20	76.70
Uniform middle frame	84.79	75.13	61.65	77.93
Max-Σ-AU	**86.89**	**77.89**	**63.52**	**80.29**

Table 3: AU reward and trace granularity ablations.

Setting	Sentiment	Basic	OV-MERD+	Mean
(a) AU reward after GRPO				
None	84.61	74.27	62.47	77.55
L1 reward	**86.89**	**77.89**	**63.52**	**80.29**
L2 reward	84.41	73.98	62.16	77.30
(b) AU trace granularity				
AU-sum (SFT)	83.37	68.40	60.44	74.17
Detailed (SFT)	86.15	72.98	62.37	77.65
AU-sum (SFT+RL)	83.87	72.99	61.98	76.60
Detailed (SFT+RL)	**86.89**	**77.89**	**63.52**	**80.29**

AU-Salient Key-Frame Selection. Table 2 varies the frame used as both decoder input and SFT reconstruction target, keeping the remaining inputs and training fixed. Max-Σ-AU achieves the highest mean score (80.29), exceeding random and uniform-middle selection by **3.59** and **2.36** points. This supports the key-frame heuristic under the evaluated settings.

AU Reward and Trace Granularity. Table 3(a) isolates the AU reward with the SFT checkpoint, data, visual inputs, ALR length, and other reward terms fixed. Normalized L1 yields a mean score of 80.29, compared with 77.55 without AU reward and 77.30 with L2, and is used by default. Panel (b) compares the 17-dimensional AU trace with one activation sum per frame, changing only the representation and corresponding reward. Detailed traces improve the mean score by 3.48 after SFT and 3.69 after SFT+RL, supporting per-AU supervision. Appendix H provides the full results and AU-sum reward definition.

ALR Length and Key-Frame Token Budget. After RL, Table 4 jointly varies latent steps, key-frame input tokens, and reconstruction-target granularity by resizing the key frame. The $L = 32$ configuration achieves the highest mean score (80.29) and is adopted by default. Latency is measured on H800 hardware with batch size 1, from the beginning to the end of model output generation, excluding preprocessing.

Table 4: Joint variation of ALR length and key-frame token budget.

Latent steps L	Sentiment	Basic	Fine-grained	Mean	Latency (s)
8	83.41	70.37	61.80	75.21	13.4512
16	85.68	75.79	62.05	78.66	13.7873
32	**86.89**	**77.89**	**63.52**	**80.29**	14.3789
48	85.59	75.37	62.62	78.50	14.6234
64	83.46	75.67	62.18	77.63	15.2551
72	83.47	74.38	61.89	77.03	15.9355
128	81.78	70.36	61.42	74.44	26.2109

Ablation on Visual Input Configurations. Table 5 varies decoder inputs during SFT+RL and inference, keeping AU supervision, the reconstruction target K, $L = 32$, and $\lambda_{\text{alr}} = 0.1$ fixed. Omitting F^k removes its input tokens while retaining K for SFT reconstruction. The full input set yields a mean score of 80.29 versus 77.32 with original and aligned face frames, supporting the key-frame input alongside temporal visual context.

Table 5: Visual input ablation with fixed ALR supervision.

Vision Inputs			**Performance**			
$\mathcal{V}^u$	$\mathcal{F}^u$	F^k	**Sentiment**	**Basic**	**Fine-grained**	**Mean**
✓	✗	✗	84.01	72.48	62.41	76.48
✗	✓	✗	84.33	72.57	58.13	76.19
✗	✗	✓	83.43	69.38	52.61	73.76
✓	✓	✗	83.75	74.83	61.57	77.32
✗	✓	✓	84.30	72.91	56.97	76.20
✓	✗	✓	82.10	72.81	61.51	75.68
✓	✓	✓	**86.89**	**77.89**	**63.52**	**80.29**

Table 6: Ablation on AU-trace generation and ALR after GRPO.

Response Components		Sentiment				Basic Emotion				Fine-grained	Mean
AU trace	ALR	MOSI	MOSEI	SIMS	SIMS v2	MER2023	MER2024	MELD	IEMOCAP	OV-MERD+	
✗	✗	76.56	78.65	81.88	83.00	77.23	83.66	60.25	66.07	60.29	74.18
✓	✗	82.36	83.77	85.92	86.83	79.54	85.70	63.42	68.00	61.75	77.48
✗	✓	83.62	83.53	87.77	86.19	78.12	86.14	64.92	68.20	61.08	77.73
✓	✓	**84.82**	**85.21**	**88.97**	**88.56**	**83.58**	**87.91**	**69.59**	**70.46**	**63.52**	**80.29**

Component Ablation. Table 6 retains rationale and answer generation in every variant, applying component-specific rewards and format constraints only to enabled components. Against the baseline mean of 74.18, AU traces and ALR alone yield 77.48 and 77.73; together they reach 80.29. Their complementary gains support combining explicit facial-action supervision with latent visual reconstruction.

5.4 Grounding Analysis

Table 7 reports answer-change rates relative to unmodified outputs under two settings. Panel (a) retains the original visual inputs and uses one evaluation subset for all three interventions. Panel (b) reports a separate mismatched-subtitle comparison.

Table 7: Answer sensitivity under latent and subtitle interventions.

System	Subtitle	Latent trajectory	Changed (%)
(a) Latent interventions and fixed-latent subtitle control			
AURORA	Original	Set to zero	56.3
AURORA	Original	From another video	51.8
AURORA	Replaced	Retained from original input	24.7
(b) Separate mismatched-subtitle setting			
w/o ALR	Replaced	None	47.6
AURORA	Replaced	Regenerated	**33.4**

Zeroing and donor substitution change 56.3% and 51.8% of answers, showing that latents participate in and influence prediction. The fixed-latent control (24.7%) conditions on the original video and subtitle, whose information can remain in the trajectory. In panel (b), AURORA regenerates latents after subtitle replacement (33.4%), while the model without ALR changes 47.6% of answers. Appendix F reports complementary OpenFace-based claim consistency and attention allocation analyses.

Additional Analyses. Appendices G, D, and E provide backbone comparisons, GRPO versus continued SFT under a matched budget, and qualitative facial-action analyses. Sensitivity to AU supervision is reported in Appendix H.2.

6 Limitations And Conclusion

Limitations. Our evaluation is constrained by the coverage of available OV-MER datasets and benchmarks. We have not conducted dedicated real-world tests under facial occlusion, image blur, or large head-pose variation. The Gaussian-noise and AU-dropout experiments quantify sensitivity to synthetic perturbations of AU estimates, which do not fully reproduce OpenFace extraction errors under these conditions.

Conclusion. We propose **AURORA**, an Action Unit-Reinforced Latent Reasoning framework for open-vocabulary multimodal emotion recognition. AURORA integrates structured AU perception with latent visual reasoning before generating a textual rationale and emotion prediction. Experiments on MER-UniBench, including OV-MERD+, demonstrate gains across modality settings and backbone scales. Component and key-frame ablations support the complementary roles of AU traces and latent visual preservation, while the compute-matched comparison demonstrates the benefit of second-stage task alignment using standard GRPO.

AI USE STATEMENT

In this work, we used generative AI tools to assist with coding and debugging during the implementation of our method. Additionally, we used generative AI tools to polish the manuscript for grammar, clarity, and readability. We also used GPT-5 to extract explicit facial-action claims from model-generated explanations, including AURORA's `<think>` rationales, for the OpenFace-based consistency analysis in Appendix F. The extracted claims were mapped to AUs using a fixed lexicon and evaluated against OpenFace reference estimates. We manually reviewed the implementation code for correctness and the polished text for accuracy and consistency with our intended meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–10. IEEE, 2016.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

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.

Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3): 203–221, 2007.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9062–9072, 2025.

Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

Xuri Ge, Junchen Fu, Fuhai Chen, Shan An, Nicu Sebe, and Joemon M Jose. Towards end-to-end explainable facial action unit recognition via vision-language joint learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8189–8198, 2024.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 1122–1131, 2020.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Guimin Hu, Yi Xin, Weimin Lyu, Haojian Huang, Chang Sun, Zhihong Zhu, Lin Gui, Ruichu Cai, Erik Cambria, and Hasti Seifi. Recent trends of multimodal affective computing: A survey from nlp perspective. *arXiv preprint arXiv:2409.07388*, 2024.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2881–2889, 2020.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 9610–9614, 2023.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 41–48, 2024a.

Zheng Lian, Haiyang Sun, Licai Sun, Jiangyan Yi, Bin Liu, and Jianhua Tao. Affectgpt: Dataset and framework for explainable multimodal emotion recognition. *arXiv preprint arXiv:2407.07653*, 2024b.

Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, Xiaobai Li, and Bin He. Affectgpt-r1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2508.01318*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022.

Xiaojiang Peng, Jingyi Chen, Zebang Cheng, Bao Peng, Fengyi Wu, Yifei Dong, Shuyuan Tu, Qiyu Hu, Huiting Huang, Yuxiang Lin, et al. Emotion-llamav2 and mmeverse: A new framework and benchmark for multimodal emotion understanding. *arXiv preprint arXiv:2601.16449*, 2026.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Tao Shi and Shao-Lun Huang. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14752–14766, 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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zihan Wang, Siyang Song, Cheng Luo, Songhe Deng, Weicheng Xie, and Linlin Shen. Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1270–1280, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

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

Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6): 59–73, 2021.

A Dataset Statistics

Table 8: Dataset statistics for training and evaluation.

Dataset	Use	Task Type	#Samples	Label / Annotation Type
MER-Caption+	SFT training	OV + description	31,327	Open-vocabulary emotion labels and descriptive emotion rationales
MER2025-OV	RL training	OV	1,000	Open-vocabulary emotion labels
OV-MERD+	Evaluation	Fine-grained OV emotion recognition	532	Variable-size open-vocabulary emotion set
MER2023	Evaluation	Basic emotion recognition	411	One label among six basic emotions
MER2024	Evaluation	Basic emotion recognition	1,169	One label among six basic emotions
MELD	Evaluation	Basic emotion recognition	2,610	One label among seven emotions
IEMOCAP	Evaluation	Basic emotion recognition	1,241	One label among selected basic emotions
CMU-MOSI	Evaluation	Sentiment analysis	686	Sentiment intensity in $[-3, 3]$
CMU-MOSEI	Evaluation	Sentiment analysis	4,659	Sentiment intensity in $[-3, 3]$
CH-SIMS	Evaluation	Sentiment analysis	457	Sentiment intensity in $[-1, 1]$
CH-SIMS v2	Evaluation	Sentiment analysis	1,034	Sentiment intensity in $[-1, 1]$

Table 8 summarizes the datasets used for training and evaluation. OV denotes open-vocabulary emotion prediction with a variable-size emotion set. We use MER-Caption+ for cold-start supervised fine-tuning and MER2025-OV for reinforcement learning. For evaluation, OV-MERD+ measures fine-grained open-vocabulary emotion recognition, while MER-UniBench contains three task types: sentiment analysis, basic emotion recognition, and fine-grained open-vocabulary emotion recognition. These tasks use WAF, HIT, and emotion-wheel-based F_s, respectively.

B Implementation details

Unless otherwise specified, we uniformly sample $M = 8$ face frames from each video and estimate $J = 17$ AU intensity dimensions via OpenFace. Each intensity value is reported on a scale of $[0, 5]$, and normalized by $B_j = 5$. The default ALR length is $L = 32$, corresponding to the fixed number of projected patch tokens obtained after resizing the AU-salient key frame. All reported AURORA experiments and internal ablations enable the `<think>` rationale and `<answer>` prediction blocks.

For our primary vision-text experiments and ablation studies, we adopt Qwen2.5-VL-7B as the backbone model. For audio-video-text settings, we additionally incorporate HuBERT as the audio encoder. To evaluate scalability, we further instantiate AURORA with Qwen3-VL-4B and Qwen3-VL-8B, examining whether the proposed Action Unit-Reinforced Latent Reasoning framework generalizes consistently across different model scales. The LLaMA-2-Chat-7B visual adaptation is described in Appendix G.2; its training settings match the corresponding vision-text or audio-video-text experiments.

All training experiments are conducted on 8 NVIDIA H800 GPUs with 80GB memory using PyTorch. We use the AdamW optimizer for both SFT and RL stages, keeping the visual encoder and visual projector frozen throughout. During SFT, the learning rate is set to 1×10^{-5}, the batch size is 3, and the model is trained for 5 epochs. During RL, the learning rate is set to 5×10^{-6}, the batch size is 1, and the model is trained for 2 epochs. For GRPO training, we use a group size of $G = 4$. The SFT latent-reconstruction weight is $\lambda_{\text{alr}} = 0.1$; weight decay is 0.1, the warmup ratio is 0.03, and the learning-rate schedule is cosine. RL rollouts use a sampling temperature of 0.9. The reward weights are $\alpha = 0.7$, $\beta = 0.1$, and $\eta = 0.2$. The parameter-free P factor in Eq. 9 scales the EW component. GRPO uses symmetric clipping with $\epsilon_{\text{clip}} = 0.2$ and a KL coefficient $\beta_{\text{KL}} = 0.04$ against the frozen SFT reference policy. The KL coefficient follows DeepSeekMath (Shao et al., 2024). For the Qwen2.5-VL-7B vision-text configuration, SFT takes 16 hours and GRPO takes 12 hours on eight H800 GPUs, corresponding to 128 and 96 GPU-hours, respectively. The two-stage training total is 28 hours, or 224 GPU-hours. The latency measurements in Table 4 use NVIDIA H800 hardware and batch size 1. Timing starts when the model begins generating its output and ends when the response is complete; preprocessing is excluded.

C Evaluation Metrics

We follow the MER-UniBench evaluation protocol. Since MLLMs generate free-form emotion words rather than fixed classifier labels, the predicted response is first parsed into an emotion set. Different tasks are then evaluated with task-specific metrics.

C.1 Fine-grained Open-Vocabulary Emotion Recognition

For OV-MERD+, each sample may contain a variable number of ground-truth emotion words. Following the emotion-wheel protocol, we first normalize emotion words with a three-stage grouping function. Let $F_1(\cdot)$ map inflected words to their base forms, $F_2(\cdot)$ map synonyms to a unified emotion word, and $F_3^{w_k}(\cdot)$ map fine-grained words to their corresponding sector in the k-th emotion wheel w_k. The complete grouping function for the k-th wheel is

$$G_{w_k}(\cdot) = F_3^{w_k}\left(F_2\left(F_1(\cdot)\right)\right), \qquad k = 1, \ldots, K. \tag{14}$$

For a dataset with N samples, let Y_i denote the ground-truth emotion set of sample i, and let $\hat{Y}_i$ denote the predicted emotion set parsed from the model output. For the k-th emotion wheel, set-level precision and recall are computed as

$$\text{Precision}_s^k = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|G_{w_k}(Y_i) \cap G_{w_k}(\hat{Y}_i)\right|}{\left|G_{w_k}(\hat{Y}_i)\right|}, \tag{15}$$

and

$$\text{Recall}_s^k = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|G_{w_k}(Y_i) \cap G_{w_k}(\hat{Y}_i)\right|}{\left|G_{w_k}(Y_i)\right|}. \tag{16}$$

The set-level F-score for the k-th emotion wheel is

$$F_s^k = \frac{2 \cdot \text{Precision}_s^k \cdot \text{Recall}_s^k}{\text{Precision}_s^k + \text{Recall}_s^k}. \tag{17}$$

The final score is averaged over all K emotion wheels:

$$F_s = \frac{1}{K}\sum_{k=1}^{K} F_s^k. \tag{18}$$

We use F_s as the primary metric for OV-MERD+ because it jointly measures whether the predicted emotion words are accurate and whether they cover the ground-truth emotion set. Precision_s and Recall_s are reported as secondary metrics when needed.

C.2 Basic Emotion Recognition

For basic emotion recognition datasets, each sample has a single ground-truth basic emotion label y_i, while the MLLM may generate a variable-size emotion set $\hat{Y}_i$. Following MER-UniBench, we evaluate whether the generated emotion set covers the ground-truth label after emotion-wheel grouping:

$$\text{HIT} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[G_{w_k}(y_i) \in G_{w_k}(\hat{Y}_i)\right], \tag{19}$$

where $\mathbb{I}[\cdot]$ is the indicator function. A prediction is counted as correct if the grouped ground-truth basic emotion is included in the grouped predicted emotion set. We use HIT as the primary metric for MER2023, MER2024, MELD, and IEMOCAP.

C.3 Sentiment Analysis

For sentiment analysis datasets, the original annotations are continuous sentiment scores. Following MER-UniBench, scores smaller than zero are mapped to negative sentiment, scores larger than zero are mapped to positive sentiment, and neutral samples are handled according to the dataset protocol. The model output is first parsed into emotion words and then mapped to sentiment polarity. We report weighted average F-score (WAF) and accuracy (ACC):

$$\text{ACC} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\hat{s}_i = s_i\right], \tag{20}$$

where s_i and $\hat{s}_i$ denote the ground-truth and predicted sentiment labels. The WAF is computed as

$$\text{WAF} = \sum_{c \in \mathcal{C}} \frac{N_c}{N} \cdot F1_c, \tag{21}$$

where $\mathcal{C}$ is the sentiment label set, N_c is the number of samples in class c, and $F1_c$ is the class-wise F1 score. Due to label imbalance in sentiment datasets, WAF is used as the primary metric for CMU-MOSI, CMU-MOSEI, CH-SIMS, and CH-SIMS v2.

C.4 Overall Mean Score

For MER-UniBench, we report the dataset-wise mean score:

$$\text{Mean} = \frac{1}{D} \sum_{d=1}^{D} m_d, \tag{22}$$

where D is the number of evaluation datasets and m_d is the primary metric of dataset d. Specifically, m_d is WAF for sentiment analysis datasets, HIT for basic emotion recognition datasets, and F_s for OV-MERD+.

D Task Alignment under a Matched Training Budget

To evaluate the benefit of task-level alignment at a fixed additional training budget, we compare GRPO with continued SFT initialized from the same converged Qwen2.5-VL-7B AURORA-SFT checkpoint in the vision-text setting. Both receive an additional 12 hours on eight H800 GPUs, equivalent to 96 GPU-hours. Continued SFT reuses the original MER-Caption+ training data with the same supervised objective and training hyperparameters as the first SFT stage. Its AU-trace, rationale, and answer targets, together with key-frame reconstruction supervision, are unchanged from the original SFT examples. GRPO uses the MER2025-OV training split and the composite reward in Section 4.2. As shown in Table 9, GRPO reaches 80.29, exceeding continued SFT by **2.47** points under the same compute budget. Both stages retain the inference architecture and the $L = 32$ latent schedule.

Table 9: Continued SFT versus GRPO under a matched training budget.

Additional optimization	Training data	GPU-hours	Mean	Gain over SFT
None	—	0	77.65	—
Continued SFT	MER-Caption+	96	77.82	+0.17
GRPO with AU-aware reward	MER2025-OV	96	**80.29**	**+2.64**

E Qualitative Facial-Action Analysis

We examine representative OV-MERD+ examples to connect emotion predictions with generated facial-action traces. We distinguish the OpenFace reference intensity $a_{m,j}$, used in key-frame selection and supervision, from AURORA's generated intensity $\hat{a}_{m,j}$. The facial actions in Table 10 make the numeric trace interpretable. Emotion labels are inferred jointly from facial actions and the other input modalities. Case IDs in Table 11 are the numeric suffixes of the dataset sample identifiers.

Table 10: Facial-action names for the AUs discussed in the case analysis.

AU	Facial action	AU	Facial action
AU01	Inner brow raiser	AU02	Outer brow raiser
AU04	Brow lowerer	AU06	Cheek raiser
AU10	Upper lip raiser	AU12	Lip-corner puller
AU25	Lips part	AU26	Jaw drop

Table 11: Predictions on three OV-MERD+ examples.

Case	Ground truth	Emotion-LLaMA	Emotion-LLaMAv2	AffectGPT	AURORA
338	anxiety, nervousness, worry	relaxed, content	resignation, frustration	anxiety	worry, anxiety
3238	disappointed, angry, displeased	calm, professional	neutral, composed	negative	angry, dissatisfied, disappointed, unhappy
1371	concern, nervous, worry	sadness, disappointment	fear, surprise	concern, anxiety	worry, anxiety

Transient facial evidence. For `sample_00000338`, the generated trace exhibits a brief activation peak involving AU01, AU04, AU10, AU25, and AU26 at the OpenFace-selected key-frame position, followed by weaker activation. AURORA predicts worry and anxiety, consistent with the annotated emotion set. This case illustrates how the selected facial moment and the frame-wise AU trace represent complementary spatial and temporal evidence.

Sustained facial action. For `sample_00003238`, AURORA generates AU10 intensities of 1.22, 1.20, 1.20, 1.02, 1.30, 1.30, 1.10, and 1.10 across the eight frames. The trace represents sustained upper-lip movement, accompanying the prediction of angry, dissatisfied, disappointed, and unhappy. AU10 describes the facial action; its affective interpretation also depends on the clip's linguistic and contextual evidence.

Weak facial activation. For `sample_00001371`, the generated AU trace is nearly flat over most of the clip, with a late increase in AU01, AU02, and AU26. The predicted worry and anxiety are consistent with the annotation, illustrating a correct prediction in a clip with weak generated AU activation.

Cross-modal disagreement. For `sample_00000499`, the annotated emotions are relaxed, caring, pleasant, friendly, and happy, while AURORA predicts concern and worry. The generated trace includes AU06 and AU12, with AU12 active throughout the clip, and the subtitle contains concern-related language.

F Quantitative Visual Grounding Analysis

We examine facial-claim consistency and attention allocation in the default rationale-enabled AURORA configuration. The latent and subtitle interventions are presented in Section 5.4.

F.1 OpenFace-Based Facial-Claim Inconsistency

For each output on OV-MERD+, we use GPT-5 to extract explicit facial-action claims, such as *furrowed brow*, *smile*, and *wide eyes*, from the generated explanation, using the default `<think>` block for AURORA. A fixed phrase-to-AU lexicon maps the extracted claims to AUs, which are compared against the OpenFace AU trace from the corresponding video. A claim is marked unsupported when none of its associated AUs reaches an intensity of 0.5 in any of the $M = 8$ sampled frames. The same extraction procedure, AU mapping, and threshold are applied to all models. We report the total number of unsupported claims divided by the total number of evaluable, AU-mappable claims across outputs. Outputs with no evaluable facial claim contribute to neither count. This diagnostic measures agreement with OpenFace, which also supplies AURORA's AU supervision.

Table 12 reports an inconsistency rate of 14.6% for AURORA, compared with 30.8–37.2% for the other models, indicating closer agreement between its generated facial descriptions and OpenFace reference estimates. The same ordering holds across AU activation thresholds from 0.3 to 1.0.

Table 12: OpenFace-based facial-claim inconsistency on OV-MERD+.

Model	Claims unsupported by OpenFace (%)
Emotion-LLaMA	35.6
Emotion-LLaMAv2	30.8
AffectGPT	37.2
AURORA	**14.6**

F.2 Attention Allocation

We compare AURORA with the separately trained AU-trace model without ALR from Table 6. The two systems obtain mean scores of 80.29 and 77.48, respectively, and otherwise use the same backbone, data, visual inputs, reward design, and training schedule. Both models generate a `<think>` rationale followed by an `<answer>` prediction.

We average decoder attention over all heads in the final quarter of layers. We collect attention at generation steps that predict tokens in the `<answer>...</answer>` span, including its opening and closing tags. At each step, we sum attention weights over the visible key-token positions belonging to each source. The first $M = 8$ image spans form the face-frame group, and the separate AU-selected image span forms the key-frame group; video, subtitle, generated AU-trace, and latent positions form the other reported groups. We average these source-level attention masses over the selected generation steps within each clip and normalize over all context groups, including unlisted prompt, rationale, previously generated answer, and special-token positions. Both models are evaluated on the same valid sample set, retaining clips with successfully segmented outputs and at least one answer-token prediction step for each model. The resulting shares are averaged equally across this common set; the subtitle column uses its subset containing subtitle tokens for both models.

Table 13: Decoder attention shares during answer generation.

System	Face frames	Key frame	Latent states	Subtitle	Video frames	AU trace
w/o ALR	0.101	0.087	–	0.336	0.150	0.119
AURORA	**0.112**	**0.103**	**0.092**	**0.224**	0.141	**0.130**
Difference	+0.011	+0.016	+0.092	−0.112	−0.009	+0.011

As shown in Table 13, AURORA assigns less attention to the subtitle while assigning more attention to face frames, the AU-salient key frame, AU traces, and latent states. These results describe a change in attention allocation during answer generation.

G Backbone Comparisons

G.1 Qwen3-VL Backbone Scaling

Table 14 evaluates AURORA across different backbone sizes. For each Qwen3-VL backbone, we compare the standard SFT/RL baseline with the corresponding model augmented by our AU-reinforced reasoning framework under the same training and evaluation protocol. AURORA improves the mean score by 2.95 and 3.12 points on the Qwen3-VL-4B and Qwen3-VL-8B backbones, respectively. These within-backbone gains show that fine-grained AU supervision and latent reconstruction complement increased model capacity.

Table 14: Backbone comparisons on Qwen3-VL-4B and Qwen3-VL-8B.

Backbone Setting		**Sentiment**				**Basic Emotion**				**Fine-grained**	**Mean**
Backbone	Method	MOSI	MOSEI	SIMS	SIMS v2	MER2023	MER2024	MELD	IEMOCAP	OV-MERD+	
Qwen3-VL-4B	Standard SFT/RL	76.09	78.71	80.54	81.87	73.86	82.84	60.99	63.17	58.57	72.96
Qwen3-VL-4B	+AURORA	77.68	81.53	82.73	83.63	75.93	87.24	67.17	68.01	59.23	75.91
Qwen3-VL-8B	Standard SFT/RL	80.99	83.32	87.81	88.29	80.74	86.08	62.70	67.34	62.26	77.73
Qwen3-VL-8B	+AURORA	**85.27**	**85.33**	**89.90**	**88.86**	**84.23**	**88.27**	**69.84**	**71.51**	**64.43**	**80.85**

G.2 LLaMA-2 Backbone Comparison

We use LLaMA-2-Chat-7B with the visual adaptation configuration of Emotion-LLaMA (Cheng et al., 2024): EVA extracts global visual features, MAE encodes local facial appearance, and VideoMAE captures temporal facial dynamics. Linear projection layers map the visual features into the decoder's 4096-dimensional embedding space, where they are combined with subtitle and instruction embeddings. AURORA additionally retains the AU-salient key-frame patch-token reconstruction objective defined in Section 3.

Table 15 compares standard SFT/RL with AURORA using the same LLaMA-2-Chat-7B backbone and visual adaptation. The Δ column reports the vision-text gain over standard SFT/RL. The vision-text variants use $M = 8$ sampled frames, and the remaining training settings follow the primary experiments in Appendix B. The rationale and answer blocks are enabled in all variants. AURORA improves the vision-text OV-MERD+ score from 60.83 to 62.91, a gain of 2.08 points. Adding audio raises AURORA's score to 64.79, a further gain of 1.88 points. The audio branch uses the same HuBERT encoder and audio-projector setup as the main audio-video-text experiments. These results support the value of AU supervision and latent reconstruction on the evaluated LLaMA-2 backbone, with complementary gains from audio.

Table 15: LLaMA-2-Chat-7B results on OV-MERD+ across input modalities.

Method	Inputs	OV-MERD+	Δ
Standard SFT/RL	Video, text	60.83	—
+AURORA	Video, text	**62.91**	**+2.08**
+AURORA	Audio, video, text	**64.79**	—

H Additional Ablation Details

H.1 AU Reward and Trace Granularity

Tables 16 and 17 give the per-dataset results summarized in Table 3, using Qwen2.5-VL-7B with vision-text inputs.

In the default setting, AURORA generates a frame-wise 17-dimensional AU intensity trace, preserving the activation pattern of individual facial action units. As a compact alternative, we construct an AU-sum variant in which each frame is represented by a single scalar $s_m = \sum_{j=1}^{J} a_{m,j}$. The corresponding reward is defined as:

$$R_{\text{sum}}(o_i) = 1 - \frac{1}{M} \sum_{m=1}^{M} \frac{|\hat{s}_m - s_m|}{\sum_{j=1}^{J} B_j}, \tag{23}$$

where $\hat{s}_m$ denotes the generated AU-sum for frame m, and $B_j = 5$ is the maximum intensity of the j-th AU. Except for the AU representation and corresponding reward formulation, all experimental settings remain identical under the Qwen2.5-VL-7B vision-text setting.

H.2 Sensitivity to AU Supervision Quality

We evaluate AURORA under additive Gaussian noise and random AU-dimension dropout to quantify its sensitivity to automatic AU estimates. For the intensity $a_{m,j} \in [0, 5]$ of AU j in frame m, Gaussian corruption is defined as

$$a^c_{m,j} = \min(5, \max(0, a_{m,j} + \epsilon_{m,j})), \qquad \epsilon_{m,j} \sim \mathcal{N}(0, \sigma^2). \tag{24}$$

For AU-dimension dropout, we sample $z_j \sim \text{Bernoulli}(1-p)$ and set $a^c_{m,j} = z_j a_{m,j}$, using the same mask for all eight frames within a clip. During training, the corrupted intensities replace the reference values in AU-trace supervision, key-frame selection, and R_{AU}. The selected frame, $F^k_c = F_{m^\star_c}$ with $m^\star_c = \arg\max_m \sum_j a^c_{m,j}$, supplies both input tokens and the latent-reconstruction target. During inference, corrupted AU estimates are used for key-frame selection.

Table 16: Ablation on AU reward design during GRPO.

	Sentiment				Basic Emotion				Fine-grained	Mean
AU reward	MOSI	MOSEI	SIMS	SIMS v2	MER2023	MER2024	MELD	IEMOCAP	OV-MERD+	
None	80.45	82.97	87.85	87.16	78.92	84.21	68.47	65.46	62.47	77.55
L1 reward	**84.82**	**85.21**	**88.97**	**88.56**	**83.58**	**87.91**	**69.59**	**70.46**	**63.52**	**80.29**
L2 reward	82.12	81.92	86.70	86.89	78.89	81.98	67.81	67.23	62.16	77.30

Table 17: AU trace granularity under SFT and SFT+RL.

AU Supervision Setting		Sentiment				Basic Emotion				Fine-grained	Mean
AU format	Training stage	MOSI	MOSEI	SIMS	SIMS v2	MER2023	MER2024	MELD	IEMOCAP	OV-MERD+	
AU-sum trace	SFT	82.43	81.56	85.09	84.41	78.79	74.36	58.13	62.30	60.44	74.17
Detailed AU trace	SFT	84.57	84.46	88.50	87.06	81.72	81.41	64.57	64.22	62.37	77.65
AU-sum trace	SFT+RL	81.92	81.40	86.28	85.89	80.29	83.29	62.66	65.71	61.98	76.60
Detailed AU trace	SFT+RL	**84.82**	**85.21**	**88.97**	**88.56**	**83.58**	**87.91**	**69.59**	**70.46**	**63.52**	**80.29**

For each corruption level, we independently train the complete SFT+RL pipeline once from the original Qwen2.5-VL-7B checkpoint and apply the same corruption level at test time. Training and test corruptions are sampled independently. The original video frames and aligned face sequences remain available under every condition. These experiments measure performance under matched training/test AU corruption.

Table 18: Synthetic AU perturbations under matched training and test conditions.

AU condition	Sentiment	Basic	OV-MERD+	Mean	Δ Mean
Clean	86.89	77.89	63.52	**80.29**	—
Gaussian $\sigma = 0.25$	86.24	77.19	63.02	79.64	-0.65
Gaussian $\sigma = 0.5$	86.14	76.68	62.65	79.33	-0.96
Gaussian $\sigma = 1.0$	85.72	75.12	62.10	78.38	-1.91
Gaussian $\sigma = 2.0$	84.87	74.34	61.45	77.59	-2.70
Dropout $p = 10\%$	86.16	77.04	63.01	79.53	-0.76
Dropout $p = 30\%$	85.71	75.95	62.63	78.81	-1.48
Dropout $p = 50\%$	83.30	73.57	61.12	76.51	-3.78
Backbone reference	81.58	67.44	58.85	72.77	—

The mean score remains 79.64 at $\sigma = 0.25$ and 79.53 under 10% dropout. At $\sigma = 2.0$ and 50% dropout, the mean scores are 77.59 and 76.51, respectively. The Δ column measures changes from clean AURORA; the backbone row provides context. These results characterize sensitivity under matched training/test AU corruption.