Facial-R1: Aligning Reasoning and Recognition for Facial Emotion Analysis

Jiulong Wu[1,2*], **Yucheng Shen**[1*], **Lingyong Yan**[2], **Haixin Sun**[1], **Deguo Xia**[2], **Jizhou Huang**[2], **Min Cao**[1†]

[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Baidu Inc., Beijing, China
wjlwujiulong@gmail.com, mcao@suda.edu.cn

Abstract

Facial Emotion Analysis (FEA) extends traditional facial emotion recognition by incorporating explainable, fine-grained reasoning. The task integrates three subtasks—emotion recognition, facial Action Unit (AU) recognition, and AU-based emotion reasoning—to jointly model affective states. While recent approaches leverage Vision-Language Models (VLMs) and achieve promising results, they face two critical limitations: (1) hallucinated reasoning, where VLMs generate plausible but inaccurate explanations due to insufficient emotion-specific knowledge; and (2) misalignment between emotion reasoning and recognition, caused by fragmented connections between observed facial features and final labels. We propose Facial-R1, a three-stage alignment framework that effectively addresses both challenges with minimal supervision. First, we employ instruction fine-tuning to establish basic emotional reasoning capability. Second, we introduce reinforcement training guided by emotion and AU labels as reward signals, which explicitly aligns the generated reasoning process with the predicted emotion. Third, we design a data synthesis pipeline that iteratively leverages the prior stages to expand the training dataset, enabling scalable self-improvement of the model. Built upon this framework, we introduce FEA-20K, a benchmark dataset comprising 17,737 training and 1,688 test samples with fine-grained emotion analysis annotations. Extensive experiments across eight standard benchmarks demonstrate that Facial-R1 achieves state-of-the-art performance in FEA, with strong generalization and robust interpretability.

Code & Datasets — https://github.com/RobitsG/Facial-R1
Extended version — https://arxiv.org/abs/2511.10254

1 Introduction

Facial Emotion Analysis (FEA) (Chaubey et al. 2025; Lan et al. 2025; Li et al. 2024) is an evolved task derived from traditional Facial Emotion Recognition (FER) (Mao et al. 2025; Shu et al. 2022; Liu et al. 2024a). Compared to FER, which usually classifies the facial emotions without explanation, FEA extends the FER as an explainable and explicit emotion reasoning process, including three sub-tasks: facial action unit (AU) recognition, AU-based emotion reasoning, and facial emotion recognition. As shown in Figure 1, a complete FEA process simultaneously recognizes the final emotion (e.g., "disgust"), detects key facial action units like the brow lowerer (AU4), and generates a detailed reasoning process to explain its judgment. These tasks work together to enhance model interpretability and trustworthiness, and promote real-world applicability.

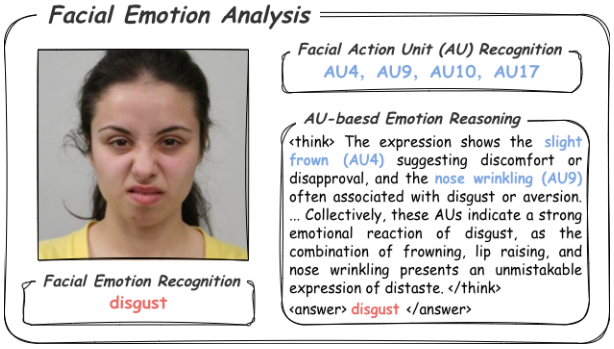

Figure 1: Illustration of facial emotion analysis task. Unlike traditional facial emotion recognition, which directly outputs a predicted emotion (e.g., *disgust*), facial emotion analysis decomposes the task into three interrelated sub-tasks: (1) **Facial Action Unit (AU) Recognition**, where local facial muscle movements (e.g., AU4: slight frown ...) are identified; (2) **AU-based Emotion Reasoning**, which generates natural language explanations linking the detected AUs to the predicted emotion; and (3) **Facial Emotion Recognition**, producing the final emotion label. Together, these results enable an explainable and interpretable emotion recognition, bridging the gap between low-level visual cues and high-level affective understanding.

Recent works (Li et al. 2024; Yuan, Zeng, and Shan 2023; Lan et al. 2025; Chaubey et al. 2025) leverage the powerful reasoning capabilities of Vision-Language Models (VLMs) (Liu et al. 2024b; Chen et al. 2024b; Bai et al. 2025) for the FEA task. Despite their promising performance, these approaches still present two key challenges: 1) **Hallucination in the reasoning process.** The reasoning process may generate plausible yet inaccurate emotion

*These authors contributed equally.
†Corresponding author.

interpretations that deviate from the actual input image or instruction. This issue arises due to the lack of emotion-specific prior knowledge in VLMs, leading to misinterpretation or omission of key facial features that are critical for accurate emotion recognition. 2) **Misalignment between reasoning and recognition.** Even when the model identifies relevant emotional cues during reasoning, the resulting analysis may conflict with the final emotion recognition output. This inconsistency stems from fragmented reasoning paths between observed facial features and emotion labels, where models fail to establish coherent causal relationships between visual evidence and emotional conclusions. Some approaches (Yuan, Zeng, and Shan 2023; Lan et al. 2025; Li et al. 2024) attempt to address these challenges by constructing fine-grained emotion analysis data for instruction fine-tuning. However, emotion reasoning is inherently complex and typically demands high-quality, large-scale data, which is often difficult to collect, limiting the interpretability and generalization of FEA methods. Additionally, overly strict instruction fine-tuning constrains the VLM's thinking, forcing it to follow predefined reasoning paths while ignoring potentially effective cues.

To address these challenges, we introduce Facial-R1, a three-stage alignment training framework. Specifically, we first develop a minimal supervised fine-tuning (SFT), requiring only 300 high-quality emotion analysis samples generated by GPT-4o-mini (Hurst et al. 2024). The SFT stage effectively mitigates hallucinations by establishing basic emotion reasoning capability in VLMs. Second, we initiate reinforcement learning (RL) by designing two emotional factors—AUs (Ekman and Friesen 1978) and emotion label—as reward signals. The AU factor enables the model to ground its analysis in concrete facial features present in the image, enhancing the rationality of emotion reasoning. The emotion label factor enforces alignment between the reasoning process and the final emotion label, ensuring the accuracy of emotion recognition. Notably, our RL strategy only requires the model to consider two emotional facts without strictly limiting the thinking details, enhancing flexibility compared to SFT. In the third stage, we perform a data synthesis strategy that iteratively expands the training dataset using the model trained on the previous two stages, with automated filtering and manual inspection to ensure data quality. This stage bypasses the data collection bottleneck faced by previous methods (Li et al. 2024). Through multiple iterative training, we construct a large-scale emotion analysis dataset, FEA-20K, and achieve state-of-the-art in three emotion analysis tasks across eight diverse benchmarks. FEA-20K comprises 17,737 automatically constructed training samples and 1,688 manually verified test samples, divided into three tasks: facial emotion recognition, AU recognition, and AU-based emotion reasoning. Compared to existing methods that demand extensively labeled datasets (Lan et al. 2025; Li et al. 2024; Chaubey et al. 2025), our framework requires only minimal supervision—a small set of example data and weakly labeled emotional factors. Furthermore, our framework enables the model to flexibly reason about the potential connections between facial features and emotions through our innovative reward mechanism that leverages emotional factors during training.

In summary, unlike traditional methods that rely exclusively on manually labeled data—a key limitation that constrains their ability to solve hallucination and misalignment—Facial-R1 overcomes the limitation with three key innovations: (1) We introduce *FEA-20K*, a large-scale fine-grained emotion analysis dataset constructed with low initialization costs, effectively bypassing the data collection bottleneck that hinders the performance of previous approaches. (2) We propose *Facial-R1*, a three-stage reasoning training framework designed for FEA task. Our framework promotes flexible reasoning patterns that emerge naturally during training, rather than enforcing predetermined paths, enhancing adaptability and robustness. (3) Extensive experiments across eight diverse benchmarks demonstrate that Facial-R1 exhibits powerful generalization capabilities in various face-related tasks, comprehensively outperforming existing methods.

2 Related Work

Facial Emotion Analysis is an important research topic in the field of affective computing (Zhang et al. 2025). It includes three primary sub-tasks: facial emotion recognition, facial action units (AU) recognition, and AU-based emotion reasoning. Traditional methods (Ning et al. 2024; Liu et al. 2024a) relied on handcrafted features (e.g., SIFT (Lindeberg 2012)) and machine learning classifiers for sentiment analysis, but struggled with complex emotions and interpretability (Lian et al. 2023). Recent studies (Lan et al. 2025; Li et al. 2024; Cheng et al. 2024; Chaubey et al. 2025) shift towards unified models that generate explanatory reasoning alongside recognition, driven by critical applications in domains such as mental health monitoring.

Vision-Language Models have demonstrated impressive capabilities in multimodal reasoning tasks (Wang et al. 2024; Zhang et al. 2024; Sui et al. 2025; Chen et al. 2025a,b; Wu et al. 2025a). Recent research utilized the powerful reasoning capabilities of VLMs to enhance the interpretability of the emotion analysis process (Zhang et al. 2025). ExpBLIP (Yuan, Zeng, and Shan 2023) pioneered using VLMs for describing facial actions and emotional states through natural language. Face-LLaVA (Chaubey et al. 2025) proposed a facial-specific instruction dataset, enhancing facial analysis capabilities via visual encoder reconstruction. FABA (Li et al. 2024) and ExpLLM (Lan et al. 2025) implemented detailed facial affect analysis using Chain-of-Thought reasoning and LoRA fine-tuning strategies. Moreover, Reinforcement Learning from human feedback guides VLMs to generate more appropriate and interpretable outputs (Bai et al. 2022; Ouyang et al. 2022; Wu et al. 2025b). For example, Omni-Emotion (Yang et al. 2025) enhanced AU-based emotion reasoning through multi-stage instruction alignment and human feedback. However, existing VLM methods mostly rely on complex, costly manual annotations. In contrast, our proposed Facial-R1 employs verifiable reward RL (Shao et al. 2024) as its core algorithm, which requires only a small amount of weakly labeled data and can fully stimulate the reasoning capabili-

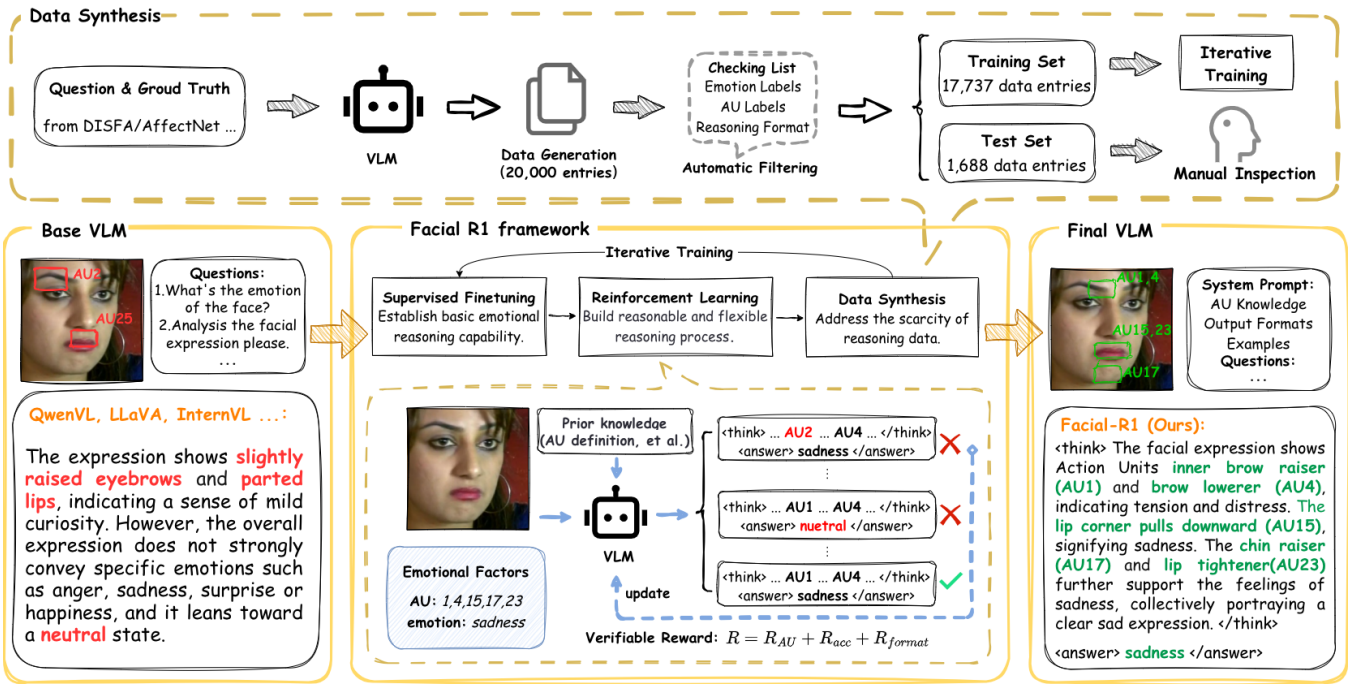

Figure 2: The Facial-R1 framework consists of three stages: (1) Supervised finetuning (SFT) mitigates hallucinations by establishing basic emotion reasoning capability; (2) Reinforcement Learning (RL) leverages verifiable emotional facts as reward signals to build reasonable and flexible reasoning process; (3) Data Synthesis iteratively leverages the prior two stages to expand the training dataset, enabling scalable self-improvement of the model.

ties of VLMs. We select Group Relative Policy Optimization (GRPO) (Shao et al. 2024; Peng et al. 2025) as our RL algorithm. It is a representative method of verifiable rewards RL, which generates multiple responses and computes advantages between each response for policy optimization.

3 Method

This section introduces our three-stage Facial-R1 framework, with the workflow shown in Figure 2. Section 3.1 introduces supervised finetuning as the first stage of our framework, where we mitigate hallucinations during the reasoning process by establishing basic emotion reasoning capability in VLM. Section 3.2 describes verifiable reward reinforcement learning as the core training methodology of Facial-R1, where we leverage verifiable emotional factors (AUs and emotion labels) as reward signals to build a reasonable and flexible emotion reasoning process. Section 3.3 elaborates how we address the data scarcity problem by synthesizing data and iteratively training the model to further enhance its performance and generalization.

3.1 Supervised Finetuning

VLMs inherently lack prior knowledge of facial emotion analysis, hindering their ability to comprehend the potential relationships between facial expressions and emotions, which can lead to reasoning hallucinations. To address this limitation, we first employ Supervised Fine-Tuning (SFT) with carefully designed instructions. Specifically, we utilize GPT-4o-mini (Hurst et al. 2024) to generate 300 high-quality instruction samples for fine-tuning the VLM. These instructions are crafted to incorporate essential emotional expertise, such as AU definitions, thereby equipping the VLM with the necessary domain-specific knowledge. After fine-tuning, the VLM demonstrates enhanced reasoning capabilities, effectively establishing a basic understanding of the relationship between facial expressions and emotions. The details for instructions can be found in Appendix C.

3.2 Reinforcement Learning

Although the VLM has acquired basic reasoning capabilities after the SFT stage, it still encounters challenges related to the interpretability of its reasoning process and potential misalignment with the final emotion labels. To address these issues, we introduce a Reinforcement Learning (RL) stage as the second stage of our approach. This stage is designed to refine the alignment between the VLM's outputs and key emotional factors. Specifically, we employ the GRPO algorithm (Shao et al. 2024) as our RL algorithm. GRPO operates by generating multiple responses and evaluating their relative advantages A_i by comparing them against each other using our designed verifiable reward. The relative advantages among these responses are calculated as:

$$A_i = \frac{R^i - \text{mean}\left(\{R^1, \ldots, R^G\}\right)}{\text{std}\left(\{R^1, \ldots, R^G\}\right)}, \tag{1}$$

where R^i is the reward for the i-th response, G is the number of responses in one step, mean calculates the arithmetic mean of all rewards, and std represents the standard deviation of the rewards. To guide the model toward more accurate and interpretable emotion reasoning, we design the reward function by integrating key factual cues in facial emotion analysis—specifically, AUs and emotion labels. Building upon the SFT-initialized model, we perform RL using a composite reward R (superscript i omitted for clarity in the following), which combines three core components: AU reward R_{AU}, emotion accuracy reward R_{acc}, and reasoning format reward R_{format}:

$$R = R_{AU} + R_{acc} + R_{format}. \tag{2}$$

AU Reward. AUs form the core physiological basis for FEA, as emotions are typically expressed through specific combinations of activated AUs. Consequently, we utilize AUs as the primary emotional factor to construct the reward signal R_{AU}. Training with R_{AU} encourages the VLM to ground its inferences in observable facial features rather than speculative interpretations, thereby improving the interpretability of emotion reasoning and mitigating hallucinations. In addition, to mitigate the inherent issue of reward sparsity in RL (Ibrahim et al. 2024), we adopt the $F1$ score as a metric for evaluating R_{AU}. It encourages the model to enhance the accuracy of AU recognition (i.e., minimizing false positives) and comprehensiveness (i.e., minimizing false negatives). The AU reward is modeled as follows:

$$R_{AU} = F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{3}$$

where $Precision$ and $Recall$ represent the proportions of correctly predicted AUs out of all predicted AUs and all true AUs, respectively.

Accuracy Reward. To address the misalignment between emotion reasoning and recognition, we propose the emotion accuracy reward R_{acc}. This reward ensures consistency between the final emotion label and AUs identified during the reasoning process by binarizing emotion labels. If the emotion derived through emotion reasoning is correct, the reward is 1; otherwise, it is 0. The accuracy reward is calculated as follows:

$$R_{acc} = \begin{cases} 1, & \text{if correct,} \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where "correct" denotes that the emotion label predicted by the model corresponds to the ground-truth annotation.

Format Reward. Following (Guo et al. 2025), we construct a format reward R_{format} to standardize the structure of emotion reasoning:

$$R_{format} = \begin{cases} 1, & \text{if valid,} \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where "valid" indicates that the model's output adheres to our specified format requirements, including the encapsulation of reasoning processes within "`<think></think>`" tags and the denotation of final emotion labels within "`<answer></answer>`" tags. This structured emotion reasoning protocol enhances both the interpretability of the emotion reasoning process and provides a systematic foundation for the subsequent data synthesis.

Overall, compared to the SFT stage which strictly regulates model outputs to establish foundational emotion reasoning capabilities, the RL stage emphasizes the extraction of key facial features and encourages flexible consideration of potential causal relationships between facial features and emotional states.

3.3 Data Synthesis

To address the scarcity of emotion reasoning data and enhance the VLM capability for facial emotion analysis, we develop an iterative data expansion stage.

Data Generation First, we construct the instruction x for emotion reasoning with a designed instruction template:

$$x = \mathrm{template}(q, gt) \tag{6}$$

where q denotes questions sourced from the FABA-Instruct dataset (Li et al. 2024) and gt represents ground-truth annotations encompassing AU and emotion labels, enabling VLM reasoning correctly. The instruction template is detailed in Appendix C. Secondly, we leverage the VLM trained in the prior two stages to synthesize high-quality emotion reasoning data y:

$$y = \mathrm{VLM}(v, x) \tag{7}$$

where v represents facial images curated from established emotion datasets as detailed in Section 4.1. Finally, we iteratively train the two prior stages to enable continuous data expansion, incorporating a data quality control to select high-quality samples for progressive model optimization.

Data Quality Control To mitigate noisy data—such as outputs with erroneous reasoning logic—we implement a rigorous two-stage filtering protocol. **(1) Automatic Filtering.** All generated samples undergo automated validation against three critical criteria: AU labels, emotion labels, and reasoning format. This process is formalized via the following checking function:

$$S(y) = \mathbb{I}[M_{AU}(y) \wedge M_{emotion}(y) \wedge M_{format}(y)], \tag{8}$$

where $M_{AU}(y)$ and $M_{emotion}(y)$ verify whether the predicted AUs and emotions align with their ground truth, respectively, and $M_{\text{format}}(y)$ evaluates whether the output adheres to the required textual structure. The indicator function $\mathbb{I}$ returns 1 if all conditions are satisfied simultaneously, and 0 otherwise. For a given input consisting of instruction x and image v, we employ a temperature-controlled sampling strategy to iteratively generate candidate responses y, repeating until a valid output is produced or the maximum retry limit is reached. **(2) Manual inspection**. After automatic filtering, we obtain the FEA-20K dataset, partitioned into a training set (17,737 samples) and a test set (1,688 samples) based on the source of data collection. For the test set, we perform strategic sampling and conduct a thorough manual evaluation by expert annotators, assessing action unit and emotion recognition accuracy, as well as the logical coherence and consistency of the generated reasoning process.

4 Experiments

4.1 Experimental Setups

Dataset Introduction. As summarized in Table 1, the FEA datasets typically cover three sub-tasks: emotion recognition, facial AU recognition, and emotion reasoning. To ensure data quality, we filtered out noisy data, such as missing annotations and images.

- **Datasets with AU recognition** provide standard resources for facial action units. *DISFA* (Mavadati et al. 2013) and *BP4D* (Zhang et al. 2014) are all frame-annotated facial video datasets. We extract individual frames in each video and construct paired image-AU samples. *RAF-AU* (Yan et al. 2020; Li and Deng 2019) contributes in-the-wild facial images with AU annotations across diverse real-world conditions.
- **Datasets with emotion recognition** are annotated with categorical emotion labels for the standard FER task. *FER2013* (Goodfellow et al. 2013) is a widely used benchmark comprised of grayscale facial images categorized into seven emotion classes. *AffectNet* (Mollahosseini, Hasani, and Mahoor 2017) is a large-scale dataset with manually labeled images spanning eight emotion categories. *RAF-DB* (Li, Deng, and Du 2017) provides real-world facial emotion data collected under unconstrained conditions.
- **Datasets with emotion reasoning** include natural language descriptions in addition to AU and emotion labels, enabling comprehensive facial emotion analysis, including *FABA-Instruct* (Li et al. 2024) and our proposed *FEA-20K*. Compared to FABA-Instruct, FEA-20K features more diverse image sources and finer-grained AU annotations, making it a more rigorous benchmark for evaluating the reasoning capabilities of VLMs.

Evaluation Metrics. We categorize our evaluation metrics into three distinct tasks: AU recognition, emotion recognition, and AU-based emotion reasoning. For AU recognition, we employ the F1 score (Mavadati et al. 2013; Zhang et al. 2014), ensuring a balanced evaluation of precision and recall. Regarding emotion recognition, we adopt accuracy (Acc.) as the primary metric, following established protocols (Li, Deng, and Du 2017). For AU-based emotion reasoning, we employ task-specific metrics. Specifically, on the FABA-Instruct dataset, we utilize the SEGE metric proposed in FABA (Li et al. 2024), which aggregates the AU recognition F1 and the ROUGE-L score of textual descriptions. For the FEA-20K dataset, we report ROUGE-L to measure the textual similarity between generated and ground-truth reasonings. Additionally, to comprehensively evaluate the reliability of emotion reasoning, we propose employing GPT-4o-mini (Hurst et al. 2024) as an automated evaluator for measuring semantic similarity between generated and ground-truth reasonings, with scores ranging from 0 to 10.

Compared Methods. The compared Methods are divided into zero-shot and fine-tuned categories with distinct architectural approaches and training paradigms.

Dataset	AU	EL	ER	Images	Train	Test
DISFA	✓			87,192	52,392	27,654
BP4D	✓			146,847	100,813	46,034
RAF-AU	✓	✓		4,601	3,479	853
FER2013		✓		35,887	28,709	3,589
AffectNet		✓		303,330	287,618	3,493
RAF-DB		✓		29,672	12,271	3,068
FABA-Instruct	✓	✓	✓	14,379	6,060	314
FEA-20K (Ours)	✓	✓	✓	19,425	17,737	1,688

Table 1: Dataset summary. They are grouped according to the annotation availability of action unit (AU), emotion label (EL), and emotion reasoning (ER). The "Images" column indicates the number of filtered images. The "Train" column displays the total number of images for training. The "Test" column indicates the number of images used for evaluation.

- **Zero-shot methods** leverage general visual-language understanding capabilities acquired during pre-training for FEA without task-specific fine-tuning. *GPT-4o* (Hurst et al. 2024) and *GPT4o-mini* (Team 2024a) are state-of-the-art commercial multimodal methods with exceptional visual understanding capabilities. We also evaluate several open-source multimodal large language models, including *LLaVA-Next-7B* (Liu et al. 2024b), *InternVL-7B* (Chen et al. 2024b), and *Qwen2.5-VL-7B* (Bai et al. 2025). Though these methods demonstrate strong general vision-language capabilities, they lack domain-specific optimization for facial expression understanding.
- **Fine-tuned Methods** refer to models specifically optimized on emotion-related datasets. Traditional methods adopt end-to-end architectures that directly predict AU or emotion labels without providing interpretable reasoning, including *FMAE* (Ning et al. 2024), *Norface* (Liu et al. 2024a), *JÂA-Net* (Shao et al. 2020), *S2D* (Chen et al. 2024a), and *QCS* (Wang et al. 2025). Despite their impressive recognition performance, these methods typically lack intrinsic reasoning capabilities and are unable to generate coherent explanations for their predictions. In contrast, facial-specialized VLMs adapt general VLM to facial analysis tasks, enabling interpretable and language-based reasoning. *Exp-BLIP* (Yuan, Zeng, and Shan 2023) extends BLIP (Li et al. 2023) with expression-focused pre-training; *Face-LLaVA* (Chaubey et al. 2025) integrates facial attribute understanding into LLaVA (Liu et al. 2024b); *ExpLLM* (Lan et al. 2025) enables fine-grained expression description via multimodal reasoning; and *EmoLA* (Li et al. 2024) enhances facial analysis through language-augmented instruction tuning. We also evaluate several facial-specialized VLMs on FABA-Instruct dataset, including MiniGPT-4v2 (Chen et al. 2023a), mPLUG-Owl2 (Ye et al. 2024), and Shikra (Chen et al. 2023b).

Implementation Details. We build our proposed Facial-R1 upon Qwen2.5-VL-7B (Bai et al. 2025), a vision-language model that integrates a powerful vision encoder

Method	AU Recogniton (F1) ↑			
	DISFA	BP4D	RAF-AU	FABA-Instruct
Zero-shot				
GPT-4o	56.5	59.5	51.2	47.2
LLaVA-Next	23.7	21.4	18.8	24.1
InternVL	41.3	30.7	43.4	45.2
Qwen2.5-VL	22.1	16.9	20.7	26.3
Fine-tuned				
FMAE	70.1	67.1	63.2	61.9▲
Norface	67.0	**69.3**	-	-
$J\hat{A}A$-Net	56.0	60.0	-	-
Exp-BLIP	-	65.0	69.5	-
Face-LLaVA	72.9	65.8	-	-
EmoLA	65.1	64.2	44.6▲	56.3
Facial-R1 (Ours)	**73.1**	67.4	**70.2**	**68.3**

Table 2: Evaluation results of AU recognition. The best and second results are highlighted in **bold** and underlined respectively. ▲ denotes results reproduced from official code.

Method	Emotion Recognition (Acc.) ↑			
	FER2013	AffectNet	RAF-DB	FABA-Instruct
Zero-shot				
GPT-4o	61.1	48.9	62.7	64.2
LLaVA-Next	21.6	7.7	21.4	28.3
InternVL	56.3	43.6	56.3	62.7
Qwen2.5-VL	24.3	32.6	34.3	63.0
Fine-tuned				
FMAE	68.3▲	64.8	**93.4**	63.3▲
Norface	-	**68.6**	92.9	-
S2D	-	63.8	92.6	-
QCS	69.7▲	64.4	93.0	64.7▲
ExpLLM	59.8▲	62.9	91.0	63.7▲
EmoLA	59.7▲	39.7▲	92.1	64.5
Facial-R1 (Ours)	**69.8**	65.2	92.1	**67.8**

Table 3: Evaluation results of Emotion recognition.

with the Qwen2-7B large language model (Team 2024b). We employ full-parameter fine-tuning with 16-bit mixed precision to accelerate training and DeepSpeed ZeRO-3 (Rajbhandari et al. 2020) to reduce memory consumption. The optimizer uses a learning rate of 2×10^{-5} with cosine decay and a weight decay of 0.1 for regularization. All experiments are conducted on 8 NVIDIA A800-80GB GPUs. We use a batch size of 8, gradient accumulation over 4 steps, and 8 samples per group for GRPO-based reinforcement learning.

4.2 Main Results

Evaluation on AU Recognition Task. As summarized in Table 2, our proposed Facial-R1 model consistently outperforms existing methods across multiple AU recognition datasets. On the DISFA dataset, Facial-R1 achieves the highest F1 score of 73.1%, surpassing Face-LLaVA (72.9%) and other competitive methods. As specifically detailed in Table 5, Facial-R1 achieves the best performance on multiple action units such as AU6 (cheek raiser), and also ranks highly on other action units. For BP4D, our approach attains a competitive F1 score of 67.4%, ranking second only to Norface (69.3%). The similar performance advantages are observed on RAF-AU dataset and FABA-Instruct. Notably, compared to the zero-shot Qwen2.5-VL baseline (20.7%), Facial-R1 achieves an absolute improvement of 49.5%. It demonstrates that our Facial-R1 significantly enhances the ability of VLMs to recognize AUs, thereby improving the interpretability of emotion reasoning.

Method	FABA-Instruct	FEA-20K	
	REGE ↑	ROUGE-L ↑	Score ↑
Zero-shot			
GPT4o	79.8	32.3	5.80
LLaVA-Next	44.6	19.5	2.83
InternVL	60.7	26.7	5.41
Qwen2.5-VL	46.4	28.0	4.60
Fine-tuned			
MiniGPT4-v2	77.8	22.6	4.52
mPLUG-Owl2	82.0	24.5	5.30
Shikra	94.6	29.7	5.72
EmoLA	**96.2**	30.1	5.66
Facial-R1 (Ours)	95.5	**37.3**	**6.09**

Table 4: Evaluation results of emotion reasoning.

Evaluation on Emotion Recognition Task. As shown in Table 3, our proposed Facial-R1 achieves competitive results across multiple emotion recognition datasets, demonstrating strong generalization and effectiveness. On the widely used FER2013 benchmark, Facial-R1 attains a state-of-the-art accuracy of 69.8%. It further achieves 67.8% accuracy on FABA-Instruct, highlighting consistent performance across diverse evaluation settings. On RAF-DB, Facial-R1 achieves 92.1% accuracy, slightly below specialized models such as FMAE (93.4%). This marginal gap arises from the fact that such models employ end-to-end optimization tailored to specific datasets, often at the expense of interpretability. In contrast, Facial-R1 maintains strong recognition performance while providing transparent and robust explanatory mechanisms—offering a meaningful advancement over black-box classification approaches.

Evaluation on Emotion Reasoning Task. Beyond traditional AU and emotion recognition tasks, we further evaluate Facial-R1 on the more challenging emotion reasoning task—one that demands fine-grained understanding of facial cues and their semantic relationships with emotional states. As shown in Table 4, on the FABA-Instruct dataset, Facial-R1 achieves a REGE score of 95.5, demonstrating competitive performance. On our more comprehensive FEA-20K dataset, Facial-R1 shows clear superiority, attaining a ROUGE-L score of 37.3—substantially outperforming all methods, including EmoLA (30.1), and GPT-4o (32.3).

Similarly, in the GPT-aligned evaluation, Facial-R1 achieves the highest average score of 6.09. These results collectively underscore Facial-R1's strong capability in gener-

Method	AU1 ↑	AU2 ↑	AU4 ↑	AU6 ↑	AU9 ↑	AU12 ↑	AU25 ↑	AU26 ↑	Avg. ↑
Zero-shot									
GPT4o-mini *	29.2	30.2	56.5	41.6	49.3	24.4	58.2	53.6	42.9
LLaVA-Next	22.4	20.6	24.8	19.2	17.9	28.4	33.5	22.8	23.7
InternVL	38.6	36.7	42.4	33.2	32.5	49.1	56.4	42.3	41.3
Qwen2.5-VL	20.8	19.2	23.0	17.9	17.1	26.3	30.9	21.6	22.1
Fine-tuned									
$J\hat{A}A$-Net *	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
Norface	**76.4**	**66.1**	74.2	58.5	57.2	81.7	**97.6**	69.6	72.7
FMAE	62.7	59.5	67.3	55.6	61.8	77.9	95.0	**69.8**	68.7
EmoLA	50.5	56.9	**83.5**	55.2	43.1	80.1	91.6	60.0	65.1
Face-LLaVA *	63.6	62.3	79.0	73.3	71.0	83.2	90.2	60.6	72.9
Facial-R1 (Ours)	63.5	62.0	79.2	**73.7**	**71.4**	**83.5**	90.6	60.0	**73.1**

Table 5: Details analysis results on the 8 AUs of the DISFA dataset. * donates data source from Face-LLaVA.

Setting	DISFA	RAF-DB	FEA-20K	
	F1 ↑	Acc ↑	ROUGE-L ↑	Score ↑
Full model	**73.1**	**92.1**	**37.3**	**6.09**
w/o SFT	59.1	88.1	33.1	5.62
w/o RL	54.3	81.2	26.5	5.24
w/o Syn	62.4	84.5	34.8	5.78

Table 6: Ablation of three stages of Facial-R1. "w/o SFT" denotes removing the supervised finetuning stage, "w/o RL" denotes removing the reinforcement learning stage, "w/o Syn" denotes removing the data synthesis.

Setting	DISFA	RAF-DB	FEA-20K	
	F1 ↑	Acc ↑	ROUGE-L ↑	Score ↑
All Rewards	**60.4**	**85.7**	34.7	**6.01**
w/o R_{AU}	43.1	76.9	28.2	5.35
w/o R_{acc}	58.9	68.9	24.8	5.64
w/o R_{format}	60.0	71.6	**34.8**	5.86

Table 7: Ablation of different rewards in the RL stage. "w/o R_{AU} / R_{acc} / R_{format}" denotes removing the corresponding reward, respectively

ating coherent, accurate, and semantically rich emotion reasoning, effectively establishing principled links between facial features and underlying emotional states.

4.3 Ablation Studies

To rigorously evaluate the contribution of each component in our Facial-R1 framework and analyze the effectiveness of our reward design, we conduct comprehensive ablation experiments across three FEA tasks, examining both the impact of different training stages and reward configurations.

Ablation of Stages. Table 6 demonstrates the critical contribution of each training stage to Facial-R1's performance. The full model consistently outperforms all ablated variants across all evaluation metrics. Removing the SFT stage results in substantial performance degradation, with F1 score on DISFA dataset decreasing by 14.0% and accuracy on RAF-DB dropping by 4.0%. The RL stage is the most crucial component, as its removal leads to the most severe performance drops—18.8% in F1 score on DISFA and 10.9% in accuracy on the RAF-DB dataset. Additionally, removing the data synthesis stage causing 10.7% and 7.6% decreases in F1 on DISFA and accuracy on RAF-DB, respectively.

Ablation of Rewards. We validate the effectiveness of each reward component in the RL stage through limited-step inference experiments (600 steps) for computational efficiency, with results summarized in Table 7. The full configuration, incorporating all reward components, achieves the best overall performance across multiple tasks. The AU reward (R_{AU}) is critical for AU recognition, as its removal leads to a substantial 17.3% drop in F1 score on DISFA. The accuracy reward (R_{acc}) plays a key role in emotion recognition, with ablation causing a 16.8% decrease in accuracy on RAF-DB and a 9.9% drop in ROUGE-L score on FEA-20K. While the format reward (R_{format}) has a smaller impact on quantitative metrics, it enhances response structure and readability—evidenced by a 0.15% decline in human evaluation score when omitted.

5 Conclusion

This paper proposes Facial-R1, a three-stage alignment framework for Facial Emotion Analysis that effectively addresses the challenges of hallucination and misalignment between reasoning and recognition. The SFT stage employs instruction fine-tuning to reduce hallucinations in reasoning. The RL stage designs three verifiable rewards to enhance the explainability of emotion reasoning and the alignment with emotions. The third stage of data synthesis overcomes the data scarcity limitation through iterative data collection and training. Experiments across eight diverse benchmarks demonstrate Facial-R1's superior performance in all three FEA tasks. For future work, we plan to extend the powerful reasoning capabilities of VLMs to other face-related tasks, e.g., facial attribute editing, and explore more comprehensive facial understanding systems.

Ethics Statement

This study focuses on improving the reliability and trustworthiness of Facial Emotion Analysis by mitigating model hallucination. This study is based on publicly available and widely used data and models; therefore, our findings may inherit the biases and limitations present in these resources. To protect copyright, we only provide data annotations and not the face images. Those who need the images should download them from the original data source.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants 62476188, the Natural Science Foundation of the Jiangsu Higher Education Institutions of China, Key Laboratory of New Generation Artificial Intelligence Technology & Its Interdisciplinary Applications (Southeast University), Ministry of Education, China.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Chaubey, A.; et al. 2025. Face-LLaVA: Facial Expression and Attribute Understanding through Instruction Tuning. *arXiv preprint arXiv:2504.07198*.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Chen, K.; Ruan, D.; Dan, Y.; Wang, Y.; Yan, S.; Wu, X.; Zhang, Y.; Chen, Q.; Zhou, J.; He, L.; et al. 2025a. A Survey of Inductive Reasoning for Large Language Models. *arXiv preprint arXiv:2510.10182*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Chen, Q.; Qin, L.; Liu, J.; Peng, D.; Guan, J.; Wang, P.; Hu, M.; Zhou, Y.; Gao, T.; and Che, W. 2025b. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Chen, Y.; Li, J.; Shan, S.; Wang, M.; and Hong, R. 2024a. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Cheng, Z.; Cheng, Z.-Q.; He, J.-Y.; Wang, K.; Lin, Y.; Lian, Z.; Peng, X.; and Hauptmann, A. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37: 110805–110853.

Ekman, P.; and Friesen, W. V. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, 117–124. Springer.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ibrahim, S.; Mostafa, M.; Jnadi, A.; Salloum, H.; and Osinenko, P. 2024. Comprehensive overview of reward engineering and shaping in advancing reinforcement learning applications. *IEEE Access*.

Lan, X.; Xue, J.; Qi, J.; Jiang, D.; Lu, K.; and Chua, T.-S. 2025. Expllm: Towards chain of thought for facial expression recognition. *IEEE Transactions on Multimedia*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, S.; and Deng, W. 2019. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127(6): 884–906.

Li, S.; Deng, W.; and Du, J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2852–2861.

Li, Y.; Dao, A.; Bao, W.; Tan, Z.; Chen, T.; Liu, H.; and Kong, Y. 2024. Facial affective behavior analysis with instruction tuning. In *European Conference on Computer Vision*, 165–186. Springer.

Lian, Z.; Sun, L.; Xu, M.; Sun, H.; Xu, K.; Wen, Z.; Chen, S.; Liu, B.; and Tao, J. 2023. Explainable multimodal emotion reasoning. *CoRR*.

Lindeberg, T. 2012. Scale invariant feature transform.

Liu, H.; An, R.; Zhang, Z.; Ma, B.; Zhang, W.; Song, Y.; Hu, Y.; Chen, W.; and Ding, Y. 2024a. Norface: Improving facial expression analysis by identity normalization. In *European Conference on Computer Vision*, 293–314. Springer.

Liu, S.; Cheng, H.; Liu, H.; Zhang, H.; Li, F.; Ren, T.; Zou, X.; Yang, J.; Su, H.; Zhu, J.; et al. 2024b. Llava-plus: Learning to use tools for creating multimodal agents. In *European Conference on Computer Vision*, 126–142. Springer.

Mao, J.; Xu, R.; Yin, X.; Chang, Y.; Nie, B.; Huang, A.; and Wang, Y. 2025. Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, 157: 110951.

Mavadati, S. M.; Mahoor, M. H.; Bartlett, K.; Trinh, P.; and Cohn, J. F. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2): 151–160.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1): 18–31.

Ning, M.; et al. 2024. Representation learning and identity adversarial training for facial behavior understanding. *arXiv preprint arXiv:2407.11243*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Peng, Y.; Zhang, G.; Zhang, M.; You, Z.; Liu, J.; Zhu, Q.; Yang, K.; Xu, X.; Geng, X.; and Yang, X. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. ZeRO: Memory optimizations Toward Training Trillion Parameter Models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16.

Shao, Z.; Liu, Z.; Cai, J.; and Ma, L. 2020. JÂA-Net: Joint Facial Action Unit Detection and Face Alignment Via Adaptive Attention. *International Journal of Computer Vision*, 129: 321 – 340.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shu, Y.; Gu, X.; Yang, G.-Z.; and Lo, B. 2022. Revisiting self-supervised contrastive learning for facial expression recognition. *arXiv preprint arXiv:2210.03853*.

Sui, Y.; Chuang, Y.-N.; Wang, G.; Zhang, J.; Zhang, T.; Yuan, J.; Liu, H.; Wen, A.; Zhong, S.; Chen, H.; et al. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

Team, O. 2024a. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. [Accessed 25-02-2025].

Team, Q. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wang, C.; Chen, L.; Wang, L.; Li, Z.; and Lv, X. 2025. QCS: Feature refining from quadruplet cross similarity for facial expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7563–7572.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wu, J.; Shi, Z.; Wang, S.; Huang, J.; Yin, D.; Yan, L.; Cao, M.; and Zhang, M. 2025a. Mitigating Hallucinations in Large Vision-Language Models via Entity-Centric Multimodal Preference Optimization. *arXiv preprint arXiv:2506.04039*.

Wu, Y.; Zhou, Y.; Ziheng, Z.; Peng, Y.; Ye, X.; Hu, X.; Zhu, W.; Qi, L.; Yang, M.-H.; and Yang, X. 2025b. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*.

Yan, W.-J.; Li, S.; Que, C.; Pei, J.; and Deng, W. 2020. Raf-au database: in-the-wild facial expressions with subjective emotion judgement and objective au annotations. In *Proceedings of the Asian conference on computer vision*.

Yang, Q.; Bai, D.; Peng, Y.-X.; and Wei, X. 2025. Omni-Emotion: Extending Video MLLM with Detailed Face and Audio Modeling for Multimodal Emotion Analysis. *arXiv preprint arXiv:2501.09502*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 13040–13051.

Yuan, Y.; Zeng, J.; and Shan, S. 2023. Describe Your Facial Expressions by Linking Image Encoders and Large Language Models. In *BMVC*, 377.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, X.; Yin, L.; Cohn, J. F.; Canavan, S.; Reale, M.; Horowitz, A.; Liu, P.; and Girard, J. M. 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10): 692–706.

Zhang, X.; Zhang, T.; Sun, L.; Zhao, J.; and Jin, Q. 2025. Exploring interpretability in deep learning for affective computing: a comprehensive review. *ACM Transactions on Multimedia Computing, Communications and Applications*.