

SoVTR: Set-of-Vision-Text Rewards for Attribute-Grounded Emotion Recognition in VLLMs

Qixuan Zhang^{1*} Zhifeng Wang^{1*} Donghao Zhang² Jiyang Zheng³

Wenjia Niu⁴ Ramesh Sankaranarayana¹ Sabrina Caldwell¹

Tom Gedeon^{1,5} Yang Liu^{1†} Zhenyue Qin^{1†}

¹Australian National University ²Curiosity Pty Ltd

³University of Sydney ⁴Webumate Pty Ltd ⁵Curtin University

{qixuan.zhang, zhifeng.wang}@anu.edu.au

{ramesh.sankaranarayana, sabrina.caldwell}@anu.edu.au

info@curiosity.com.au, niuwenjia9064@hotmail.com

tom.gedeon@curtin.edu.au, {yang.liu1082, kf.zy.qin}@gmail.com

Project Page: <https://github.com/wangzhifengharrison/SoVTR>

Abstract

Vision Large Language Models (VLLMs) show promise for multi-modal understanding; yet, their application to emotion recognition remains limited by insufficient spatial and contextual awareness and by supervised fine-tuning that uniformly penalizes all reasoning tokens—suppressing valid alternatives. We present Set-of-Vision-Text Rewards (SoVTR), a unified, reward-based framework that guides VLLMs with interpretable and verifiable visual attributes, focusing optimization on the emotion-discriminative reasoning steps that truly determine correctness. SoVTR decomposes supervision into three rewards: (i) Face Reward capturing action units, gaze, and micro-expressions; (ii) Body Reward encoding pose and gesture cues; and (iii) Context Reward modeling person-object relations and scene activities. By combining BERTScore attribute matching with GRPO policy optimization, SoVTR produces attribute-grounded explanation traces while preserving non-critical reasoning diversity. Across DFEW, FERV39k, and ExpW, SoVTR consistently improves both recognition accuracy and grounding quality, delivering substantial gains over strong VLLM baselines.

1 Introduction

With the advancement of computer vision and natural language processing integration, Vision-Large Language Models (VLLMs) (Bai et al., 2025; Wang et al., 2025; Zhang et al., 2024) have revolutionized how machines interpret visual and textual data. While emotion recognition has garnered significant interest across multiple disciplines, it re-

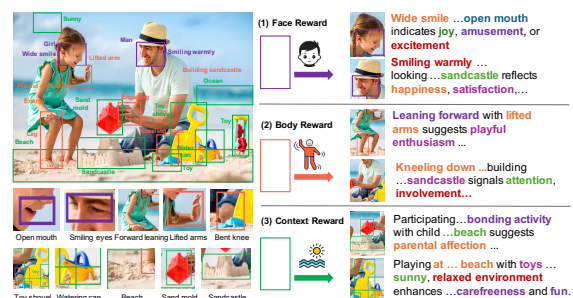


Figure 1: **Proposed SoVTR framework.** SoVTR integrates Face Reward (AUs, gaze, micro-expression), Body Reward (pose, gesture), and Context Reward (scene, activity recognition) for emotion prediction.

mains challenging (Yang et al., 2023a; Wang et al., 2024a; Cheng et al., 2024a; Wang et al., 2024b).

Prior work (Xenos et al., 2025; Zhang et al., 2023) has improved in-context emotion classification by training transformer-based models or vision-language models like CLIP to generate natural language descriptions of emotions in visual contexts. However, these methods suffer from two fundamental limitations. First, the effectiveness of SFT is highly dependent on data quality: they rely on generic scene-level descriptions from VLLMs and filter samples only by the correctness of the predicted label. Nevertheless, correctness at the label level does not guarantee that every reasoning step captures truly emotion-discriminative evidence—predictions. Correct emotion predictions may be supported by generic cues (e.g., overall scene mood or background clutter). Blindly imitating all tokens in the imperfect reasoning traces reduces optimization efficiency and weakens the model’s grounding in attribute-level emotion evidence. Moreover, only a small subset of reasoning tokens—those describing decisive emotion-

* Equal contribution

† Corresponding authors

which have demonstrated exceptional capabilities across diverse natural language processing tasks. Building on these successes, VLLMs incorporate large-scale video-text paired data to enable robust multimodal understanding. A representative example is LLaVA (Li et al., 2024), which adopts an end-to-end framework combining a vision encoder with a language model to support effective multimodal dialogue. More recent VLLMs (Ye et al., 2025b; Wu et al., 2024; Liu et al., 2024; Zhou et al., 2024) serve as strong foundations for vision-language reasoning, as they are architecturally optimized to integrate and align visual and textual modalities for enhanced multimodal inference.

2.2 Reinforcement Learning

Reinforcement Learning (RL) has become a key approach for enhancing the complex reasoning capabilities of Large Language Models (LLMs). Methods such as Reinforcement Learning from Human Feedback (RLHF), employed by OpenAI’s GPT models (Jaech et al., 2024), have significantly improved model alignment and reasoning quality. More recently, DeepSeek-R1 (Guo et al., 2025) introduced Group Relative Policy Optimization (GRPO), which replaces the traditional critic-based structure with verifiable, rule-based rewards—thereby simplifying training and improving reasoning performance. This RL-driven trend has extended into VLLMs (Ye et al., 2025a) to boost their visual reasoning capabilities, as evidenced by works such as Visual-RFT (Liu et al., 2025) and VLM-R1 (Shen et al., 2025). These studies show that applying GRPO yields substantial gains in visual grounding tasks over conventional supervised methods. Our approach aims to improve VLLM performance in fine-grained, interpretable emotion reasoning.

2.3 Emotion Recognition

Multimodal emotion recognition has progressed from unimodal facial or vocal classifiers toward datasets and models that jointly exploit visual, auditory, and textual cues (Chen et al., 2024; Wang et al., 2025; Zhang et al., 2026). Early benchmarks such as DFEW and FERV39k provide single-label, discrete emotion annotations collected under in-the-wild conditions, but offer limited contextual or explanatory supervision. To move beyond coarse classification, recent work fine-tunes large multimodal models directly on emotion-centric cor-

pora: Emotion-LLaMA (Cheng et al., 2024a) couples instruction tuning with multimodal cues for reasoning-based recognition, while related unified frameworks attempt to preserve general vision-language ability alongside emotion specialization by introducing dedicated facial-feature modules and expert-routing architectures. A persistent difficulty across this line of work is catastrophic forgetting: models fine-tuned on narrow emotion datasets often lose general-purpose reasoning capability, motivating architectural remedies such as mixture-of-experts compressors and multi-stage warm-up training. However, these approaches remain rooted in supervised fine-tuning, imitating every token of an imperfect reasoning trace regardless of whether it reflects emotion-discriminative evidence, and they typically lack spatial grounding to localize the facial, bodily, or contextual regions that justify a prediction. SoVTR instead departs from token-level imitation, using verifiable, attribute-conditioned rewards under reinforcement learning to selectively reinforce discriminative reasoning steps while preserving spatial grounding and reasoning diversity.

3 Methods

3.1 Problem Definition

Given an input image I , the goal is to detect all faces and extract a structured tuple for each face i (Eq. 1):

$$y_i = (\text{Position}_i, \text{Confidence}_i, \text{Label}_i \in \mathcal{E}) \quad (1)$$

where $\text{Position}_i = [x_1^{(i)}, y_1^{(i)}, x_2^{(i)}, y_2^{(i)}] \in \mathbb{Z}^4$ denotes the bounding box, Confidence_i is the model’s confidence score, and $\mathcal{E}$ is the emotion label set. A VLLM π_θ maps the image to an attribute-grounded explanation traces r and the answer set $\mathcal{Y} = \{y_1, \dots, y_N\}$ for N detected faces: $(r, \mathcal{Y}) = \pi_\theta(I)$.

3.2 SoVTR Overview

The SoVTR framework (Fig. 2) integrates structured visual and textual rewards into a reasoning-driven reinforcement learning pipeline for emotion recognition. Given a prompt and an image, the policy model π_θ generates multiple chain-of-thought (CoT) responses, which are evaluated by verifiable reward functions assessing face bounding boxes, facial expressions attributes, body language attributes, and contextual attributes. The highest-scoring response is selected based on these mul-

timodal rewards. We employ GRPO (Guo et al., 2025) to compute relative response quality within candidate groups and update π_θ , with KL divergence regularization against a pre-trained reference model to ensure training stability.

3.2.1 IoU Reward in Face Detection

Given ground-truth boxes $\mathcal{G} = \{g_i\}_{i=1}^G$ and predicted boxes $\mathcal{P} = \{p_j\}_{j=1}^P$ parsed from the <answer> tag, we remove duplicates in $\mathcal{P}$ and greedily match each prediction to the unmatched ground-truth box with the highest IoU:

$$\text{IoU}(a, b) = \frac{\text{area}(a \cap b)}{\text{area}(a) + \text{area}(b) - \text{area}(a \cap b)}. \quad (2)$$

For IoU threshold τ (default 0.5), each prediction p_j is assigned $\widehat{\text{IoU}}_j = \widehat{\text{IoU}}_j$ if $\widehat{\text{IoU}}_j \geq \tau$, and 0 otherwise, where $\widehat{\text{IoU}}_j = \max_{i \in \text{unmatched}} \text{IoU}(p_j, g_i)$. The raw reward averages these scores normalized by $D = \max(G, P)$ to prevent inflation when $P < G$. The final reward is clipped with a small floor $\alpha = 0.05$ to prevent zero-reward collapse:

$$R_{iou} = \begin{cases} 1 - \alpha, & \text{if matched,} \\ \text{clip}(\frac{1}{D} \sum_{j=1}^P \widehat{\text{IoU}}_j, \alpha, 1 - \alpha), & \text{otherwise.} \end{cases} \quad (3)$$

This design rewards spatial accuracy, handles missing or extra face boxes.

3.2.2 Label Reward in Emotion Recognition

This reward measures the categorical correctness of predicted labels. Predictions are greedily matched to ground-truth boxes by maximum IoU with threshold τ . For each matched pair, we define a label-correctness indicator:

$$z_j = \begin{cases} 1, & \text{if } \hat{\ell}_j = \ell_{i^*(j)} \text{ and } \widehat{\text{IoU}}_j \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Each prediction contributes a reward $r_j^{\text{lab}} = (1 - \alpha) z_j + \alpha (1 - z_j)$, mapping correct and incorrect labels to $1 - \alpha$ and α , respectively. The final label reward averages over matches and normalizes by $D = \max(G, M)$ to prevent inflation when fewer predictions pass the IoU threshold:

$$R_{\text{lbl}} = \text{clip}\left(\frac{1}{D} \sum_{j=1}^M r_j^{\text{lab}}, \alpha, 1 - \alpha\right) + 10^{-4}, \quad (5)$$

where G is the number of ground-truth boxes and M is the number of valid matches.

Algorithm 1 Fine-Tuning with Verifiable Rewards for Vision-Language Face Detection & Emotion Recognition

Require: Pretrained policy π_θ ; dataset $\mathcal{D} = \{(I, p, \mathcal{Y}^*)\}$, $\mathcal{Y}^* = \{(b_i, \ell_i)\}$ with $b_i = [x_1, y_1, x_2, y_2]$; pixel domain $I = [W \times H \times 3]$; pixel bounds $p \in [I(p)_{\min}, I(p)_{\max}]$; τ_{iou} , GRPO settings, weights λ .

Ensure: Fine-tuned policy $\pi_{\theta'}$.

- 1: Initialize GRPO trainer with reward registry functions
- 2: **for** each training step **do**
- 3: Sample minibatch $\mathcal{B} \subset \mathcal{D}$.
- 4: **for all** $(I, \text{problem}, \mathcal{Y}^*) \in \mathcal{B}$ **do**
- 5: **Rollout:** Query π_θ with pixel constraints to produce <think> r </think> and <answer> A </answer>.
- 6: **Parse:**
- 7: $\hat{\mathcal{Y}} \leftarrow \text{ExtractBBboxes}(A)$;
- 8: $r \leftarrow \text{ExtractThink}(r)$;
- 9: // Removes duplicate boxes to avoid double counting
- 10: $\hat{\mathcal{Y}} \leftarrow \text{Deduplicate}(\hat{\mathcal{Y}})$.
- 11: **Match:**
- 12: // Sort by confidence, pair with best-IoU if $\text{IoU} > \tau$
- 13: $(\text{IoURes}, \text{LblRes}, \text{AttrRes}) = \text{MATCHANDSCORE}(\mathcal{Y}^*, \hat{\mathcal{Y}}, r)$.
- 14: **Rewards:**
- 15: $R_{iou} \leftarrow \text{IoUReward}(\text{IoURes}, |\mathcal{Y}^*|)$ (Eq. (3))
- 16: $R_{\text{lbl}} \leftarrow \text{LabelReward}(\text{LblRes}, |\mathcal{Y}^*|)$ (Eq. (5))
- 17: $R_{\text{conf}} \leftarrow \text{ConfidenceReward}(\text{IoURes}, \text{LblRes})$ (Eq. (8))
- 18: // Correct format for <think> and <answer>
- 19: $R_{\text{fmt}} \leftarrow \text{FormatReward}(r, A)$ (Eq. (9))
- 20: // BERTScore for face, body and context attributes
- 21: $R_{\text{attr}} \leftarrow \text{ReasoningAttrReward}(r, \hat{\mathcal{Y}})$ (Eq. (11))
- 22: **Aggregate:** $R \leftarrow R_{iou} + R_{\text{lbl}} + R_{\text{conf}} + R_{\text{fmt}} + R_{\text{attr}}$.
- 23: Store (prompt, answer, R) for GRPO.
- 24: **end for**
- 25: **Policy Update:** $\theta \leftarrow \text{GRPOUpdate}(\{(\text{prompt}, \text{answer}, R)\}_{\mathcal{B}})$.
- 26: **end for**
- 27: **return** π_θ

3.2.3 Confidence Reward

This reward encourages confidence consistent with both face localization quality and emotion label correctness. Let $\mathcal{M}$ denote the set of matched predictions with $|\mathcal{M}| = m$. For each match $i \in \mathcal{M}$, we have $\text{IoU } s_i \in [0, 1]$ with confidence c_i^{iou} , and label indicator $z_i \in \{0, 1\}$ with confidence c_i^{lbl} . We convert the label indicator to a value $v_i = (1 - \alpha) z_i + \alpha (1 - z_i) \in \{1 - \alpha, \alpha\}$, then apply piecewise shaping to penalize overconfident errors:

$$\tilde{c}_i^{\text{iou}} = \begin{cases} c_i^{\text{iou}}, & s_i > 0, \\ (1 - s_i) (1 - c_i^{\text{iou}}), & s_i = 0, \end{cases} \quad (6)$$

$$\tilde{c}_i^{\text{lbl}} = \begin{cases} c_i^{\text{lbl}}, & v_i = 1 - \alpha, \\ (1 - v_i) (1 - c_i^{\text{lbl}}), & v_i = \alpha. \end{cases} \quad (7)$$

The per-match score averages both shaped terms, and the final reward is the clipped mean over all matches:

$$R_{\text{conf}} = \text{clip}\left(\frac{1}{m} \sum_{i \in \mathcal{M}} \frac{1}{2} (\tilde{c}_i^{\text{iou}} + \tilde{c}_i^{\text{lbl}}), \alpha, 1 - \alpha\right). \quad (8)$$

This promotes high confidence when both localization and classification are correct, and suppresses overconfidence when either fails.

3.2.4 Format reward

Each response must contain `<think> . . . </think>` followed by `<answer> . . . </answer>`. Let $\zeta = 1$ if the response matches this format and 0 otherwise:

$$R_{\text{fmt}} = (1 - \alpha) \zeta + \alpha (1 - \zeta). \quad (9)$$

3.3 Reasoning Attributes Reward

We reward explanations that reference label-consistent and verifiable attributes. For each prediction j with label $\hat{y}_j$ and rationale t_j (from `<think> . . . </think>`), we compare t_j to three label-conditioned attribute sets: facial $\mathcal{A}_E(\hat{y}_j)$, body-language $\mathcal{A}_B(\hat{y}_j)$, and contextual $\mathcal{A}_C(\hat{y}_j)$. Using BERTScore (Zhang et al., 2020) $F1(t_j, a)$, we retain attributes exceeding threshold $\epsilon = 0.6$:

$$\mathcal{A}_\bullet^+(j) = \{a \in \mathcal{A}_\bullet(\hat{y}_j) : F1(t_j, a) > \epsilon\}. \quad (10)$$

The per-modality score x_j^X is the mean F1 over retained attributes in $\mathcal{A}_X^+(j)$ (zero if empty), for $X \in \{E, B, C\}$. We aggregate using hierarchical weights $\lambda_e, \lambda_b, \lambda_c$ that prioritize facial expressions while body language and contextual cues provide supplementary support:

$$r_j = \lambda_e e_j + \lambda_b b_j + \lambda_c c_j, \quad R_{\text{attr}} = \frac{1}{|G|} \sum_j r_j. \quad (11)$$

The total reward aggregates all components:

$$R = R_{\text{iou}} + R_{\text{lbl}} + R_{\text{conf}} + R_{\text{fmt}} + R_{\text{attr}}. \quad (12)$$

Each (prompt, answer, R) tuple updates the policy model via Algorithm 1, encouraging predictions that are accurate, localized, and grounded in multi-source visual evidence.

4 Experiments

4.1 Models and Settings

We fine-tune Qwen2.5-VL-3B (Bai et al., 2025)¹ and Qwen2.5-VL-7B (Bai et al., 2025)² using SoVTR on the official DFEW training split. Training uses 4 H100 GPUs with DeepSpeed ZeRO-3 and generates 4 candidate responses per sample. For the Reasoning Attributes Reward, we

¹<https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct>

²<https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct>

compute BERTScore F1 using the distilbert-base-uncased backbone to keep reward computation lightweight during GRPO rollouts, where every candidate response in each sampling group requires scoring; we enable baseline rescaling to normalize raw cosine similarities onto a more interpretable range, omit IDF re-weighting. We evaluate on the DFEW test set and apply the trained model directly to FERV39k and ExpW test sets with zero-shot cross-dataset evaluation. Baselines include commercial models (MiMo-VL-7B-RL (Xiaomi, 2025)³, Qwen2.5-VL) and open-source models (VideoLLaMA2-7B (Cheng et al., 2024b)⁴, LLaVA-NeXT-Video-7B (Li et al., 2024)⁵, MiniCPM-V (Yao et al., 2025)⁶, and Video-CCAM-9B (Fei et al., 2024)⁷).

4.2 Quantitative Results

Our SoVTR-enhanced models demonstrate substantial improvements over baseline VLLMs on both DFEW and FERV39k (Table 1). On DFEW, Qwen2.5-VL-3B + SoVTR achieves 56.57% accuracy and 55.45% F1 (+14.37% and +14.61% over baseline), with consistent gains across all difficulty levels, even on Hard samples (+8.54% accuracy). Qwen2.5-VL-7B + SoVTR further attains 57.12% accuracy and 56.35% F1 (+13.59% and +14.21%). On FERV39k, Qwen2.5-VL-7B + SoVTR reaches 38.72% accuracy and 34.16% F1 (+6.50% and +9.14%), with gains pronounced on challenging samples, demonstrating SoVTR’s effectiveness in capturing subtle emotional cues in the wild. Table 2 compares emotion recognition performance on the ExpW dataset across baseline VLLMs, task-specific emotion methods, and our SoVTR-enhanced models. Among baselines, Qwen2.5-VL-7B achieves the highest performance (64.20% Acc, 63.61% F@1), while task-specific emotion methods such as FaRL reach 64.65% accuracy. Our SoVTR-enhanced models surpass all baselines, with Qwen2.5-VL-7B + SoVTR achieving 67.00% accuracy and 65.60% F@1, outperforming the strongest task-specific method FaRL by +2.35% and the base Qwen2.5-VL-7B by +2.80%. The consistent gains across both model scales confirm SoVTR’s effectiveness on static image-based emotion recognition.

³<https://huggingface.co/XiaomiMiMo/MiMo-VL-7B-RL>

⁴<https://github.com/DAMO-NLP-SG/VideoLLaMA2>

⁵<https://github.com/LLaVA-VL/LLaVA-NeXT>

⁶<https://huggingface.co/openbmb/MiniCPM-V>

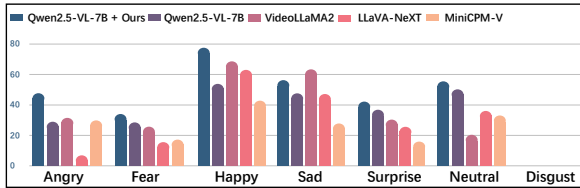
⁷<https://github.com/QQ-MM/Video-CCAM>

Table 1: **Performance comparison of VLLMs on DFEW and FERV39k datasets.** We evaluate VideoL-LaMA2 (Cheng et al., 2024b), LLaVA-NeXT (Li et al., 2024), MiniCPM-V (Yao et al., 2025), Video-CCAM (Fei et al., 2024), MiMo-VL-7B-RL (Xiaomi, 2025), Qwen2.5-VL (Bai et al., 2025), and our SoVTR-enhanced models. **Green** (Δ) indicates absolute improvements over baselines. **Red** and **Blue** denote best and second-best results.

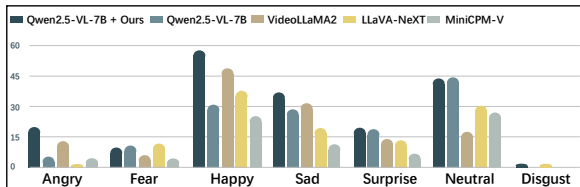
Method	Backbone	DFEW								FERV39k							
		Easy		Medium		Hard		Overall		Easy		Medium		Hard		Overall	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Baseline VLLMs																	
VideoLLaMA2	ViT-L	35.58	42.47	30.15	38.61	21.95	26.81	32.20	40.06	19.51	25.76	13.45	19.84	7.45	11.72	16.13	22.73
LLaVA-NeXT	SigLIP	31.93	34.96	29.26	33.21	31.71	35.19	30.39	33.95	15.12	18.54	18.34	22.58	9.17	13.69	16.33	20.07
MiniCPM-V	SigLIP	23.26	29.83	20.86	29.00	7.32	12.23	21.58	29.18	13.91	15.92	10.19	14.13	3.15	5.06	11.67	14.79
Video-CCAM	SigLIP	14.94	15.42	14.79	14.89	14.63	15.68	14.85	15.17	11.97	10.31	11.38	10.28	8.02	9.55	11.51	10.40
MiMo-VL-7B-RL	ViT	45.19	44.20	44.52	42.56	42.50	44.13	44.76	43.17	33.67	31.39	35.23	32.06	24.64	21.79	33.96	31.48
Qwen2.5-VL-3B	ViT	40.94	40.05	42.86	41.39	50.00	45.34	42.20	40.84	27.67	22.36	34.79	25.36	26.22	15.73	30.93	23.43
Qwen2.5-VL-7B	ViT	43.26	42.07	43.79	42.30	41.46	40.01	43.53	42.14	29.45	24.40	35.69	26.35	26.49	16.65	32.22	25.02
Our SoVTR-Enhanced Models																	
Qwen2.5-VL-3B + SoVTR	ViT	56.56	55.89	56.51	55.02	58.54	53.37	56.57	55.45	35.27	31.53	40.79	33.93	27.51	20.16	37.46	32.17
Δ		+15.62	+15.84	+13.65	+13.63	+8.54	+8.03	+14.37	+14.61	+7.60	+9.17	+6.00	+8.57	+1.29	+4.43	+6.53	+8.74
Qwen2.5-VL-7B + SoVTR	ViT	57.47	56.93	57.07	56.06	51.22	49.16	57.12	56.35	37.09	34.01	41.22	34.99	30.95	23.23	38.72	34.16
Δ		+14.21	+14.86	+13.28	+13.76	+9.76	+9.15	+13.59	+14.21	+7.64	+9.61	+5.53	+8.64	+4.46	+6.58	+6.50	+9.14

Table 2: **Comparison of emotion recognition performance across various VLLMs on the ExpW dataset.** Results shown for overall accuracy (Acc%) and F1-Score (F@1). **Green** (Δ) rows report absolute improvements over the corresponding Qwen2.5-VL baselines. **Red** and **blue** denote best and second-best results.

Methods	Backbone	Acc(%)	F@1
<i>Baseline VLLMs</i>			
VideoLLaMA2 (Cheng et al., 2024b)	ViT-L	35.20	42.50
LLaVA-NeXT (Li et al., 2024)	SigLIP	42.07	51.89
MiniCPM-V (Yao et al., 2025)	SigLIP	26.07	37.16
Video-CCAM (Fei et al., 2024)	SigLIP	24.40	28.04
MiMo-VL-7B-RL (Xiaomi, 2025)	ViT	46.33	54.60
Qwen2.5-VL-3B (Bai et al., 2025)	ViT	61.77	60.91
Qwen2.5-VL-7B (Bai et al., 2025)	ViT	64.20	63.61
<i>Emotion Methods</i>			
FaRL (Zheng et al., 2022)	ViT	64.65	—
Ada-DF (Liu et al., 2023)	ViT	59.59	—
HITL (Hasan et al., 2025)	ViT	54.00	40.60
<i>Our SoVTR-Enhanced Models</i>			
Qwen2.5-VL-3B + SoVTR	ViT	65.87	64.81
Δ		+4.10	+3.90
Qwen2.5-VL-7B + SoVTR	ViT	67.00	65.60
Δ		+2.80	+1.99



(a) DFEW (Jiang et al., 2020)



(b) FERV39k (Wang et al., 2022)

Figure 3: Performance comparison across seven emotion categories.

Fig. 3 presents per-category analysis across

seven emotions. On DFEW, our method achieves strong performance on Happy (76.2%), Neutral (53.7%), and Angry (49.2%), significantly outperforming all baselines. On FERV39k, our approach attains the highest accuracy on six of seven categories. These consistent advantages demonstrate that SoVTR enhances VLLMs’ discriminative capability across diverse emotions, including challenging categories like Angry and Surprise.

4.3 Visual Prompting

Table. 3 presents a comparative analysis of different methods for emotion recognition across datasets categorized by varying levels of difficulty: Easy, Medium, and Hard. It details performance metrics such as Accuracy and Recall, comparing the effectiveness of different visual prompting strategies utilized by Qwen2.5-VL-7B. The baseline method (Qwen2.5-VL-7B), which uses plain text prompts, demonstrates moderate effectiveness, with an overall accuracy of 43.53% and a Recall of 42.14%. Methods such as ReCLIP, RedCircle, and SoVTR employ more complex combinations of visual prompts. SoVTR (Ours), incorporating facial cues, body language, and contextual information, achieves the highest overall accuracy and recall scores of 57.12% and 56.35%, respectively. This results suggests that the integration of multiple visual attributes cues, particularly those that enhance the recognition of facial features, significantly improves performance across all difficulty levels, especially in more challenging datasets. Fig. 4 compares visual prompting methods for emotion reasoning. RedCircle and ReCLIP focus on facial regions while masking the background. In contrast, our SoVTR approach retains facial, body, and contextual details, enabling richer and more holistic

Table 3: **Comparison of SOTA methods for emotion recognition across DFEW datasets (Jiang et al., 2020) with varying levels of difficulty—Easy, Medium, and Hard.** The types of visual prompts used by previous approaches are: **P**: Crop, **B**: Box, **R**: Blur Reverse, **C**: Circle, **F**: Face Attributes, **D**: Body Language Attributes, **T**: Contextual Attributes

SOTA methods	Visual Prompting	Easy		Medium		Hard		Total	
		Acc (%)	F@1	Acc (%)	F@1	Acc (%)	F@1	Acc (%)	F@1
Baseline (Bai et al., 2025)	Plain Text	43.26	42.07	43.79	42.30	41.46	40.01	43.53	42.14
RedCircle (Shtedritski et al., 2023)	P C R	45.27	42.22	43.25	39.72	46.34	37.51	44.13	40.79
ReCLIP (Sun et al., 2024)	P B R	47.55	45.30	46.24	42.64	43.90	41.17	46.73	43.74
SoVTR (Ours)	F D T	57.47	56.93	57.07	56.06	51.22	49.16	57.12	56.35

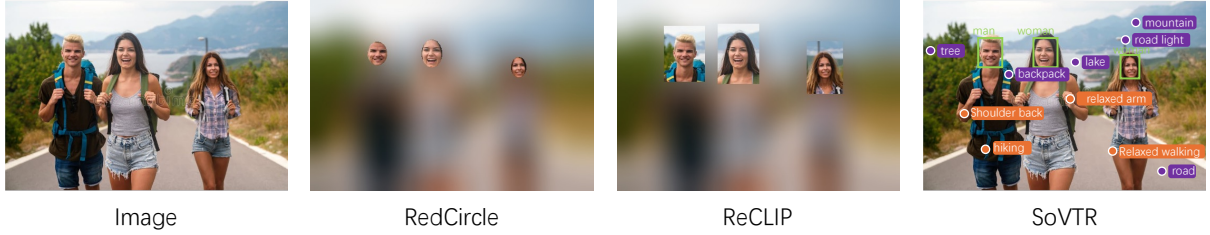


Figure 4: Visualization of the SOTA visual prompting approaches such as ReCLIP (Sun et al., 2024), RedCircle (Shtedritski et al., 2023) and our SoVTR prompts.

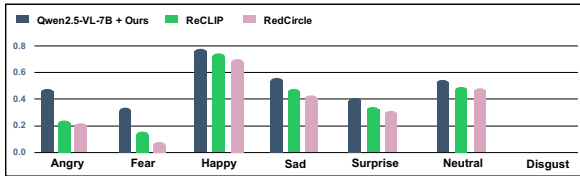


Figure 5: The bar chart illustrates the performance of SoVTR(Ours), RedCircle and ReCLIP across seven different emotional categories.

emotion understanding. Fig. 5 presents a comparative analysis of emotion recognition performance across seven emotion categories among three models: the proposed SoVTR-enhanced Qwen2.5-VL-7B, ReCLIP, and RedCircle. The SoVTR approach consistently outperforms the other two baselines in all emotion classes. This visualization highlights the robustness and superiority of the SoVTR visual rewards framework in fine-grained emotion recognition tasks.

4.4 Face Grounding

Table 4 compares reasoning-grounding performance (mIoU/gIoU) across VLLMs on DFEW, FERV39k, and ExpW. Baseline Qwen2.5-VL models achieve modest grounding scores, while MiMo-VL-7B-RL provides a somewhat stronger but still limited baseline (34.75%–45.96% mIoU). Applying SoVTR yields substantial, consistent gains across every dataset and model scale, lifting both 3B and 7B backbones to 77%–82% mIoU/gIoU — more than doubling grounding quality rela-

tive to their respective baselines (absolute improvements ranging from +40.45% to +63.54%). The gains are most pronounced for the 7B backbone, where SoVTR improves mIoU by up to +63.54% on FERV39k, indicating that the reward-based grounding supervision scales effectively with model capacity despite the 7B baseline’s comparatively weak starting point. Overall, the 7B+SoVTR configuration attains the strongest grounding performance on FERV39k and ExpW (82.65%/82.30% and 80.88%/80.13% mIoU/gIoU, respectively), while 3B+SoVTR remains competitive and even leads on DFEW (80.75%/80.25%), demonstrating that SoVTR’s grounding improvements are robust across both model scales and dataset domains.

Fig. 6 presents per-category reasoning-grounding performance on DFEW, FERV39k, and ExpW, respectively, comparing Qwen2.5-VL-3B + SoVTR (ours) with the Qwen2.5-VL-3B baseline and MiMo-VL-7B-RL across seven emotion classes (Angry, Fear, Happy, Sad, Surprise, Neutral, Disgust). Across all three datasets, our method achieves scores in the 78%–86% range for every emotion category, roughly doubling the grounding performance of both baselines on DFEW and FERV39k. Despite this baseline variability, our method’s advantage remains stable across categories within each dataset, with comparatively little fluctuation in its own scores (e.g., ranging narrowly from 78% to 84%

Table 4: Comparison of reasoning grounding across datasets (DFEW (Jiang et al., 2020), FERV39k (Wang et al., 2022), ExpW (Zhang et al., 2018)), reporting mIoU and gIoU for multiple VLLMs.

Dataset	Methods	#Param	mIoU	gIoU
DFEW	MiMo-VL-7B-RL (Xiaomi, 2025)	7B	39.35	38.18
	Qwen2.5-VL-3B	3B	38.45	37.65
	Qwen2.5-VL-3B + SoVTR (Ours)	3B	80.75 (+42.30)	80.25 (+42.60)
	Qwen2.5-VL-7B	7B	19.26	18.73
	Qwen2.5-VL-7B + SoVTR (Ours)	7B	77.71 (+58.45)	77.11 (+58.38)
FERV39k	MiMo-VL-7B-RL (Xiaomi, 2025)	7B	45.96	44.36
	Qwen2.5-VL-3B	3B	41.23	40.67
	Qwen2.5-VL-3B + SoVTR (Ours)	3B	81.68 (+40.45)	81.31 (+40.64)
	Qwen2.5-VL-7B	7B	19.11	18.89
	Qwen2.5-VL-7 + SoVTR (Ours)	7B	82.65 (+63.54)	82.30 (+63.41)
ExpW	MiMo-VL-7B-RL (Xiaomi, 2025)	7B	34.75	33.78
	Qwen2.5-VL-3B	3B	38.00	37.06
	Qwen2.5-VL-3B + SoVTR (Ours)	3B	80.30 (+42.30)	79.53 (+42.47)
	Qwen2.5-VL-7B	7B	20.23	19.58
	Qwen2.5-VL-7B + SoVTR (Ours)	7B	80.88 (+60.65)	80.13 (+60.55)

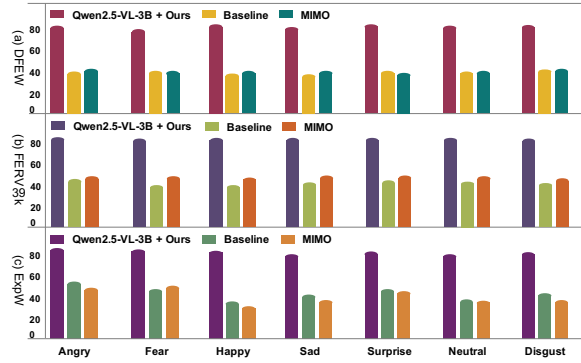


Figure 6: Per-category reasoning-grounding performance on three datasets. Our method attains consistently higher scores on all categories.

on DFEW) relative to the more pronounced category-wise variation exhibited by both baselines. This uniform, wide margin over diverse emotions and datasets underscores the robustness and cross-dataset generalization of SoVTR for fine-grained, emotion-aware grounding.

4.5 Ablation Study

Table 5 investigates the effects of different visual attributes rewards on Qwen2.5-VL-3B. Relative to the baseline without attribute rewards (Acc = 42.20, F@1 = 40.84), adding Facial expression rewards yields large gains (51.23 Acc, 50.68 F@1). Incorporating facial expression and body language rewards improves further (52.85 Acc, 52.17 F@1). Our full SoVTR—which additionally injects context attributes—achieves the best results (56.57 Acc, 55.45 F@1), demonstrating consistent, step-wise improvements and highlighting the different contribution of facial expression, body language and contextual signals.

Table 6 compares three training paradigms — zero-shot inference, supervised fine-tuning (SFT), and our proposed SoVTR reinforcement learning framework — on two Qwen2.5-VL backbones. On

Table 5: **Ablation study for visual attributes rewards on Qwen2.5-VL-3B.** **Baseline:** represents the model’s performance without any additional attributes rewards. **Face:** indicates a visual reward that uses facial expression attributes. **Face+Body:** adding body language attributes. **SoVTR:** adding context attributes.

Visual Reward	#Param	Total	
		Acc (%)	F@1
Baseline (Bai et al., 2025)	3B	42.20	40.84
Face	3B	51.23	50.68
Face+Body	3B	52.85	52.17
SoVTR	3B	56.57	55.45

Table 6: Comparison of training paradigms on Qwen2.5-VL backbones. Δ denotes absolute improvement of SoVTR over Zero-shot.

Method	Training	Acc (%)	F1 (%)
Qwen2.5-VL-3B	Zero-shot	42.20	40.84
+ SFT	Supervised	47.47	47.22
+ SoVTR	RL	56.57	55.45
Δ	—	+14.37	+14.61
Qwen2.5-VL-7B	Zero-shot	43.53	42.14
+ SFT	Supervised	50.58	48.75
+ SoVTR	RL	57.12	56.35
Δ	—	+13.59	+14.21

the 3B model, SFT yields a moderate gain over the zero-shot baseline (+5.27 Acc, +6.38 F1), while SoVTR delivers a substantially larger improvement of +14.37 Acc and +14.61 F1, nearly tripling the gain from supervised training alone. The same trend holds on the 7B backbone: SFT improves over zero-shot by 7.05 Acc / 6.61 F1, whereas SoVTR achieves +13.59 Acc and +14.21 F1. These consistent double-digit gains across both model scales indicate that the improvement stems from the SoVTR-based optimization itself rather than from increased model capacity, and that the visual and textual reward signals in SoVTR provide a stronger learning signal for emotion recognition than supervised label fitting.

Table 7 reports an ablation study on impacts of the BERTScore filtering threshold ϵ and the emotion-attribute weight λ_e , with overall Accuracy and F@1 as metrics. Performance peaks at a moderate filter and mid-level weight: $\epsilon > 0.4$ with $\lambda_e = 0.6$ attains the best results (Acc 56.57, F@1 55.45). Raising ϵ to 0.6 yields similar but slightly lower scores, and the optimum shifts toward higher λ_e (0.8: Acc 55.78, F@1 54.88). At the strictest filter $\epsilon > 0.8$, performance improves with larger λ_e (0.8: Acc 56.33, F@1 54.49)

Table 7: Ablation study on Qwen2.5-VL-3B + SoVTR varying the BERTScore filtering threshold ϵ and the emotion-attribute weight λ_e .

Prompts	#Score	#Weight	Acc (%)	F@1
Qwen2.5-VL-3B (Bai et al., 2025)+ SoVTR	$\epsilon > 0.4$	0.2	54.80	53.81
		0.6	56.57	55.45
		0.8	54.71	53.73
Qwen2.5-VL-3B (Bai et al., 2025)+ SoVTR	$\epsilon > 0.6$	0.2	55.50	54.50
		0.6	55.55	54.56
		0.8	55.78	54.88
Qwen2.5-VL-3B (Bai et al., 2025)+ SoVTR	$\epsilon > 0.8$	0.2	55.45	54.27
		0.6	54.94	53.63
		0.8	56.33	54.49

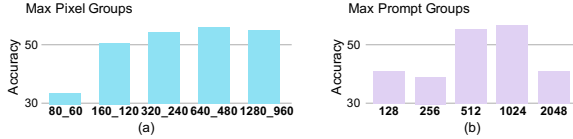


Figure 7: Ablation study of key hyperparameters. We evaluate the impact of (a) pixel size, (b) maximum prompt token sizes.

yet remains marginally below the $\epsilon > 0.4$ optimum. Overall, the study indicates that moderate evidence filtering and balanced emphasis on emotion attributes provide the most reliable gains, while overly small or large weights—and overly aggressive filtering—slightly degrade effectiveness.

Fig. 7 reports two ablations. (a) Input resolution: accuracy rises from low resolution (80×60) to mid/high resolutions, peaking around mid-high sizes, showing finer visual detail is key for reasoning. (b) Prompt length: varying the max tokens (128–2048) reveals a sweet spot at 512–1024; very short (128–256) or very long (2048) prompts reduce scores, implying too little or too much context hurts understanding. Overall, tuning resolution and prompt length is essential to balance efficiency and task accuracy.

4.6 Visualization

Fig. 8 illustrates qualitative results from the SoVTR-enhanced Qwen2.5-VL model across six emotion categories, demonstrating how the model uses structured visual and textual cues to support its predictions. Each sample includes an input image, an attribute-grounded explanation, and a predicted emotion with a confidence score. These examples showcase SoVTR’s ability to produce accurate, interpretable, and localized emotion reasoning in complex scenes.

5 Limitations

We acknowledge several limitations of this work. First, our evaluation focuses on the seven basic



Figure 8: Visualize the SoVTR-enhanced Qwen2.5-VL generated samples for different expressions: the top shows a target individual, with the ground-truth emotion label displayed in the top-left corner; the bottom shows the generated chain-of-thought reasoning and answer. Yellow text indicates location; green highlights body language; blue denotes facial expression; and purple captures contextual information.

emotions, which does not fully capture compound emotions (e.g., bittersweet, awe), or culturally-specific expressions that may follow different display conventions across cultures. Second, SoVTR is designed for structured emotion recognition with verifiable rewards. Its extension to open-ended affective tasks (e.g., empathetic dialogue, emotional storytelling), where valid responses are inherently ambiguous, remains underexplored.

6 Conclusion

We proposed SoVTR, a unified reward-based framework that decomposes emotion supervision into Face, Body, and Context Rewards. By combining BERTScore attribute matching with GRPO policy optimization, SoVTR produces visually-grounded Chain-of-Thought predictions with bounding boxes and confidence scores while preserving reasoning diversity. On DFEW, FERV39k, and ExpW, SoVTR consistently surpasses strong VLLM baselines, with up to +14.37% accuracy gains and nearly doubled face-grounding IoU. These results demonstrate that structured attribute rewards enable robust, interpretable, and spatially-aware emotion reasoning in VLLMs.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. 2024. Finecliper: Multimodal fine-grained clip for dynamic facial expression recognition with adapters. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2301–2310.
- Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024a. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.
- Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024b. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *JMLR*, 24(240):1–113.
- Paul Ekman, Tim Dalgleish, and M Power. 1999. Basic emotions. *San Francisco, USA*.
- Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. 2024. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Mahedi Hasan, Naimul Islam Shuvo, Airin Akter, Fahim Shakil Tamim, Nazir Ahmed, Naeem Mia, and Shahinur Alam. 2025. Enhancing face-to-emotion recognition with vision transformer and human-in-the-loop approach. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, pages 1–6. IEEE.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 2020. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2881–2889.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.
- Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. 2024. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pages 1–18. Springer.
- Shu Liu, Yan Xu, Tongming Wan, and Xiaoyan Kui. 2023. A dual-branch adaptive distribution fusion framework for real-world facial expression recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.
- Emily B Prince, Katherine B Martin, Daniel S Messinger, and M Allen. 2015. Facial action coding system. *Environmental psychology & nonverbal behavior*, 1.
- Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.
- Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11987–11997.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.
- Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. 2022. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20922–20931.
- Zhifeng Wang, Kaihao Zhang, Wenhan Luo, and Ramesh Sankaranarayanan. 2024a. Htnet for micro-expression recognition. *Neurocomputing*, 602:128196.
- Zhifeng Wang, Kaihao Zhang, and Ramesh Sankaranarayanan. 2024b. Lrdif: Diffusion models for under-display camera emotion recognition. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 2048–2054. IEEE.
- Zhifeng Wang, Qixuan Zhang, Peter Zhang, Wenjia Niu, Kaihao Zhang, Ramesh Sankaranarayanan, Sabrina Caldwell, and Tom Gedeon. 2025. Visual and textual prompts in vlms for enhancing emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, and 1 others. 2024. Vision-llm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975.
- Alexandros Xenos, Niki M Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. 2025. Vllms provide better context for emotion understanding through common sense reasoning. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.
- LLM-Core-Team Xiaomi. 2025. **Mimo-vl technical report**. Preprint, arXiv:2506.03569.
- Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and 1 others. 2023a. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023b. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.
- Lingfeng Yang, Yuezhe Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2024. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, and 1 others. 2025. Efficient gpt-4v level multimodal large language model for deployment on edge devices. *Nature Communications*, 16(1):5509.
- Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2025a. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*.
- Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. 2025b. Voco-llama: Towards vision compression with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29836–29846.
- Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, YiFan Zhang, Renrui Zhang, and 1 others. 2026. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pages 48767–48807.
- Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, Sabrina Caldwell, Tom Gedeon, Yang Liu, and Zhenyue Qin. 2024. Visual prompting in llms for enhancing emotion recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4484–4499.
- Sitao Zhang, Yimu Pan, and James Z Wang. 2023. Learning emotion representations from verbal and nonverbal communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18993–19004.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, pages 1–18. Springer.
- Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569.
- Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. 2022. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709.
- Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2024. Segment everything everywhere all at once. *Proceedings of the Conference on Neural Information Processing Systems*, 36.

A1 Prompt Used to Construct the Dataset

Table A.1 presents the comprehensive prompt template employed for constructing a multi-task face detection and emotion recognition dataset. The prompt instructs the model to perform three interconnected tasks: (1) detect all human faces present in the input image and provide precise spatial localization through bounding box coordinates in the format $[x_1, y_1, x_2, y_2]$, where all values must be integers representing the top-left and bottom-right corners of each face; (2) assign a confidence score between 0 and 1 (reported to exactly two decimal places) to quantify face detection and emotion recognition certainty; and (3) classify each detected face into one of seven fundamental emotion categories—‘fear’, ‘angry’, ‘happy’, ‘sad’, ‘disgust’, ‘neutral’, or ‘surprise’—based on comprehensive analysis of facial expressions (eyebrow position, mouth shape, eye aperture), body language, and broader contextual details within the scene. The prompt explicitly handles edge cases by requiring the model to return “No Objects” when no faces are present, ensuring consistent data formatting across all scenarios.

The prompt emphasizes structured, machine-readable output through a two-stage response format: the model must first provide its reasoning process within `<think>`/`</think>` tags to ensure transparency and facilitate qualitative analysis of inference strategies, followed by the final predictions enclosed in `<answer>`/`</answer>` tags formatted as a JSON-compatible list structure: `[{'Position': [x1, y1, x2, y2], 'Confidence': number, 'Label': 'emotion'}, ...]`. This design concludes with an explicit directive to “strictly follow the format,” which is critical for maintaining dataset quality, enabling automated validation, and ensuring seamless integration with downstream data processing pipelines. By combining detection and classification tasks with mandatory reasoning documentation and rigid formatting requirements, this prompt design produces a rich, self-documenting dataset that minimizes annotation ambiguity, supports multiple research objectives, and scales effectively across large image collections while maintaining consistency throughout the dataset construction process. Throughout the training phase, we employ format-based rewards to ensure the model produces outputs conforming to the specified structure. The reasoning content within `<think>`/`</think>` tags facilitates model

Table A.1: Prompt used to construct the dataset.

Face Detection & Emotion Recognition Prompt: Detect the faces in the image, and provide the bounding boxes of the faces (integers) and a confidence score (between 0 and 1, with two decimal places). In addition, identify the emotion of the faces from ['fear', 'angry', 'happy', 'sad', 'disgust', 'neutral', 'surprise'] based on the image by analyzing facial expression, posture, and context. If no face is present in the image, return 'No Objects'. Output the thinking process in `<think>` `</think>` and the final answer in `<answer>` `</answer>` tags. The output answer format should be: `<think>` ... `</think>``<answer>``[{'Position': [x1, y1, x2, y2], 'Confidence': number, 'Label': ''}, ...]``</answer>`. Please strictly follow the format.

self-improvement during reinforcement learning by exposing the underlying decision-making logic in an interpretable manner. Concurrently, the predictions within `<answer>`/`</answer>` tags are assessed to generate task-oriented rewards that steer the model’s learning process toward improved face detection and emotion classification performance. This dual-component reward structure guarantees both process interpretability and prediction quality.

A1.1 Dataset details

We evaluate on three public emotion recognition datasets spanning easy to hard difficulty levels based on face image resolution: DFEW (Jiang et al., 2020) contains 16,372 movie clips with expert annotations under challenging in-the-wild conditions; FERV39k (Wang et al., 2022) which includes 39,000+ video clips labeled with seven basic emotions; and ExpW (Zhang et al., 2018) which comprises 91,793 in-the-wild face images with emotion annotations, providing greater scale and diversity.

A1.2 Vision and Text Prompts

After training, SoVTR assigns identity labels to individuals using bounding boxes, enabling person-level tracking across prompts. With each person tagged (e.g., Person 1, Person 2), the model analyzes emotions by jointly reasoning over facial expressions, body language, and contextual cues. Fig. A.1 demonstrates two prompting approaches on a family picnic scene: (1) plain text prompts for group-level emotion interpretation, and (2) vision-text prompts linking bounding boxes to queries for person-specific emotion recognition. This enables nuanced understanding of individual emotions grounded in multimodal reasoning.

A2 Construction and Validation of Emotion Reward Attributes.

The emotion reward attribute taxonomy was constructed through a systematic three-stage pipeline



Figure A.1: We use two types of prompt methods. **Left:** Plain text prompts are used to interpret group-level emotions, capturing general affective states such as joy, relaxation, and social engagement in an outdoor picnic scenario. **Right:** Combined vision-text prompts associate visual regions (e.g., bounding boxes) with corresponding textual descriptions, enabling the localization of individual emotional states (e.g., child's joy, father's engagement, mother's comfort). This allows for fine-grained emotion attribution at the person level.

grounded in both data-driven generation and expert validation. In the first stage, we employed Qwen2.5-VL to generate candidate facial expression, body language, and contextual attributes from single-frame images sampled exclusively from the DFEW training split, ensuring no test data leakage. The model was prompted to describe observable visual cues associated with each of the seven emotion categories, producing an initial pool of raw attributes per modality. In the second stage, two human annotators independently reviewed and refined the generated attributes by removing redundant or ambiguous entries, merging semantically overlapping descriptions, and ensuring label consistency—i.e., that each attribute is discriminative for its associated emotion category. Disagreements between annotators were resolved through discussion until consensus was reached. In the third stage, a third annotator with a professional background in psychology verified the correctness and completeness of the final attribute sets, cross-referencing them against established affective science frameworks including Ekman’s basic emotion theory (Ekman et al., 1999) and the Facial Action Coding System (FACS) (Prince et al., 2015). This expert review ensured that facial attributes align with recognized Action Unit (AU) configurations, that body language attributes reflect validated arousal and valence indicators, and that contextual attributes capture situational appraisal cues.

A2.1 Emotion Reward Attributes

To enable fine-grained and explainable emotion recognition, we propose a structured taxonomy of *Emotion Reward Attributes* that guide the model’s reasoning process. As illustrated in Fig. A.2 and Fig. A.3, the attributes are systematically organized into three complementary categories: (a) *Facial Expression Attributes*, (b) *Body Language Attributes*, and (c) *Contextual Attributes*. This design supports multi-modal affect recognition by incorporating visual, behavioral, and situational signals into the reinforcement learning framework.

Facial Expression Attributes. This category captures subtle visual cues that are strongly indicative of emotional states. Attributes such as *furrowed brows*, *wide open eyes*, *moist eyes*, *smile*, *tense facial muscles*, and *focused gaze* play a critical role in recognizing primary emotions like happiness, sadness, fear, and anger. These visual signals also include finer-grained expressions such as *calm*, *helplessness*, *despair*, and *confusion*, en-

abling nuanced emotion understanding. These features map naturally to the annotated emotion groups in Fig. A.3, where, for example, *smile*, *joy*, and *relaxed posture* are clustered under happiness, while *furrowed brows* and *heated argument* characterize anger.

Body Language Attributes. These attributes capture expressive cues from bodily posture and movement. Examples include *tense posture*, *leaning forward*, *relaxed demeanor*, *clenched fists*, *slumped posture*, and *expressive gestures*. Such attributes encode information about emotional intensity, engagement level, and behavioral disposition. For instance, *aggressive stance* and *defensive posture* are indicative of anger or anxiety, while *engaged in conversation* and *focused stance* reflect attentiveness or confidence. These body cues are particularly useful in scenarios where facial visibility is limited or occluded.

Contextual Attributes. Contextual attributes provide situational and environmental grounding that disambiguates emotional meaning in complex or ambiguous scenes. This includes both physical context (e.g., *indoor setting*, *dim lighting*, *formal attire*) and social-emotional context (e.g., *heated argument*, *pleasant conversation*, *emotional turmoil*, *unexpected news*). Such cues are essential for distinguishing between emotions with overlapping visual indicators—for example, both fear and surprise may involve *wide eyes* and *open mouth*, but the presence of a *threatening environment* or *shocking event* in the context refines the interpretation.

In summary, these three structured categories of Emotion Reward Attributes enable interpretable, multi-modal emotion reasoning by anchoring reinforcement learning outputs to human-verifiable, semantically grounded cues. This facilitates robust emotion understanding even in socially or visually complex scenarios, contributing to the broader goal of explainable emotion-aware AI systems.

A2.2 Reasoning Attributes Reward Algorithm

To operationalize the multi-modal Emotion Reward Attributes within our SoVTR framework, we introduce a structured reward computation process that leverages facial, bodily, and contextual evidence to guide model alignment. This mechanism ensures that generated outputs not only exhibit high spatial accuracy but also express emotionally grounded and interpretable justifications. The full algorithmic procedure is described in Algorithm A.1.

Algorithm A.1 Reasoning Attributes Reward (See Eq.(16)–Eq.(17))

Require: Completion C (with <think> & <answer>); ground-truth detections $G = \{(b_i^*, y_i^*)\}$

Ensure: $R_{\text{attr}} = \frac{1}{|G|} \sum_j r_j$

```

1: Hyperparams: IoU threshold  $\tau$ ; Bertscore similarity threshold  $\epsilon$ 
2: Parse predicted boxes/labels  $\hat{\mathcal{Y}} = \{(\hat{b}_j, \hat{y}_j)\}$  from <answer> in  $C$ 
3: Extract rationale text  $t$  from <think> in  $C$  (if per-face rationales exist, use  $t_j$ ; else set  $t_j \leftarrow t$ )
4: Greedy match predictions to ground truth by best IoU:
5:  $\mathcal{M} \leftarrow \{(i, j) \mid \text{IoU}(b_i^*, \hat{b}_j) \text{ is maximal for } j \text{ and } > \tau\}$ 
6: Initialize list  $\mathcal{R} \leftarrow []$ 
7: for all  $(i, j) \in \mathcal{M}$  do
8:    $t_j \leftarrow$  rationale for prediction  $j$  (default  $t$ )
9:   Obtain label-conditioned attribute sets  $\mathcal{A}_E(\hat{y}_j)$  (facial),  $\mathcal{A}_B(\hat{y}_j)$  (body),  $\mathcal{A}_C(\hat{y}_j)$  (context)
10:  for all  $X \in \{E, B, C\}$  do
11:    Compute BertScore F1 between  $t_j$  and every  $a \in \mathcal{A}_X(\hat{y}_j)$ 
12:    Retain attributes above threshold (Eq.(16)):
13:     $\mathcal{A}_X^+(j) \leftarrow \{a \in \mathcal{A}_X(\hat{y}_j) : \text{F1}(t_j, a) > \epsilon\}$ 
14:    Mean F1 over retained attributes (zero if none):
15:     $x_j^X \leftarrow \begin{cases} \frac{1}{|\mathcal{A}_X^+(j)|} \sum_{a \in \mathcal{A}_X^+(j)} \text{F1}(t_j, a), & |\mathcal{A}_X^+(j)| > 0, \\ 0, & \text{otherwise.} \end{cases}$ 
16:  end for
17:  Compute structured-evidence score (Eq.(17)):
18:   $r_j \leftarrow \lambda_e x_j^E + \lambda_b x_j^B + \lambda_c x_j^C$ 
19:  Append  $r_j$  to  $\mathcal{R}$ 
20: end for
21: return  $R_{\text{attr}} \leftarrow \frac{1}{|G|} \sum_{r \in \mathcal{R}} r$ 

```

Given a predicted completion C containing both rationale (<think>) and answer (<answer>) segments, the model first parses predicted detection boxes and associated labels $\hat{\mathcal{Y}} = \{(\hat{b}_j, \hat{y}_j)\}$. These are greedily matched to the set of ground-truth detections $G = \{(b_i^*, y_i^*)\}$ using Intersection-over-Union (IoU), retaining only matches with IoU above a threshold τ . For each matched prediction $(\hat{b}_j, \hat{y}_j)$, a rationale text t_j is extracted and associated with the prediction (defaulting to global t if instance-specific rationales are unavailable).

To evaluate the reasoning quality, the algorithm retrieves label-conditioned attribute sets for each prediction from the predefined taxonomy: $\mathcal{A}_E(\hat{y}_j)$ (facial), $\mathcal{A}_B(\hat{y}_j)$ (body), and $\mathcal{A}_C(\hat{y}_j)$ (context). For each modality $X \in \{E, B, C\}$, the algorithm computes the BertScore F1 similarity between the rationale t_j and each attribute $a \in \mathcal{A}_X(\hat{y}_j)$. Attributes with similarity above a BertScore threshold ϵ are retained, forming the verified set $\mathcal{A}_X^+(j)$. The attribute-based reward for each modality x_j^X is then computed as the mean F1 score over the retained attributes (or zero if none meet the threshold).

Finally, a structured evidence score r_j is calculated for each prediction as a weighted combination of modality-specific scores:

$$r_j \leftarrow \lambda_E x_j^E + \lambda_B x_j^B + \lambda_C x_j^C$$

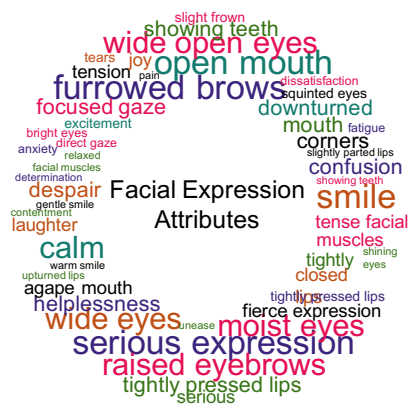
with weighting ratios as $\lambda_E, \lambda_B, \lambda_C$. The overall reasoning attributes reward R_{attr} is returned as the mean over all matched predictions:

$$R_{\text{attr}} = \frac{1}{|G|} \sum_j r_j$$

This reward design encourages the model to produce semantically meaningful rationales that align with fine-grained multimodal attributes. By grounding feedback in interpretable and verifiable cues, the algorithm enables robust training of emotionally aware systems within the SoVTR framework.

A2.3 Scene based emotion recognition

Figure A.4 qualitatively demonstrates how emotion inference is modulated by the interaction between facial, bodily, and contextual cues across diverse real-world situations. The examples illustrate that identical or similar facial expressions can yield distinct emotional interpretations depending on surrounding context and body language, whereas con-



(a) Facial expression attributes



(b) Body language attributes



(c) Context attributes

Figure A.2: Overview of Emotion Reward Attributes used for fine-grained emotion recognition and reasoning. The figure presents three categories of attributes that is used in reinforcement learning (RL) for aligning model outputs with desired behaviors through verifiable rewards. (a) Facial Expression Attributes include detailed facial cues. (b) Body Language Attributes capture postural and gestural. (c) Context Attributes provide environmental and situational cues. Together, these attributes form the foundation of multi-modal emotion reward modeling in SoVTR, enabling more robust and explainable emotion inference.

sistent contextual and bodily indicators tend to reinforce stable emotional judgments.

In scene Figure A.4 (a), Person 1 exhibits a partially open mouth and furrowed brows—facial features that, in isolation, may be ambiguous and could correspond to either fear or surprise. However, when coupled with a defensive running posture (arms bent and raised, body angled forward) and the chaotic, threatening background of smoke, debris, and fleeing individuals, these cues collectively indicate a high-arousal negative emotion, most plausibly *fear*. The convergence of tense posture, protective arm position, and hazardous environmental context strengthens this interpretation.

Scene Figure A.4 (b) presents a contrasting situation, in which Person 1 displays a broad smile, slightly squinted eyes, and relaxed facial muscles—strong indicators of positive affect. The accompanying body language, including a forward-leaning but casual posture and relaxed shoulders, suggests comfort and social engagement. The picnic setting, featuring food, greenery, and companionship, provides additional contextual evidence of leisure and enjoyment. The integration of these multimodal cues leads to a confident inference of *happiness*.

Scene Figure A.4 (c) further reinforces this pattern of contextual alignment. Here, Person 1 shows a similar smiling expression with raised cheeks and visible teeth, but within a family-oriented barbecue scene. The body language—open arm gestures, upright and engaged stance—combined with a bright, relaxed environment near a pool, supports an interpretation of social warmth and joyful interaction. Despite differing activities between scenes (b) and (c), the parallel presence of friendly social context, relaxed demeanor, and affiliative gestures leads to the same emotional inference of *happiness*.

Overall, these qualitative examples highlight the importance of integrating multi-source evidence across facial expression, body language, and environmental context. By jointly reasoning over these modalities, the model can resolve ambiguities arising from isolated facial cues and achieve more robust and explainable emotion recognition across complex real-world scenes.

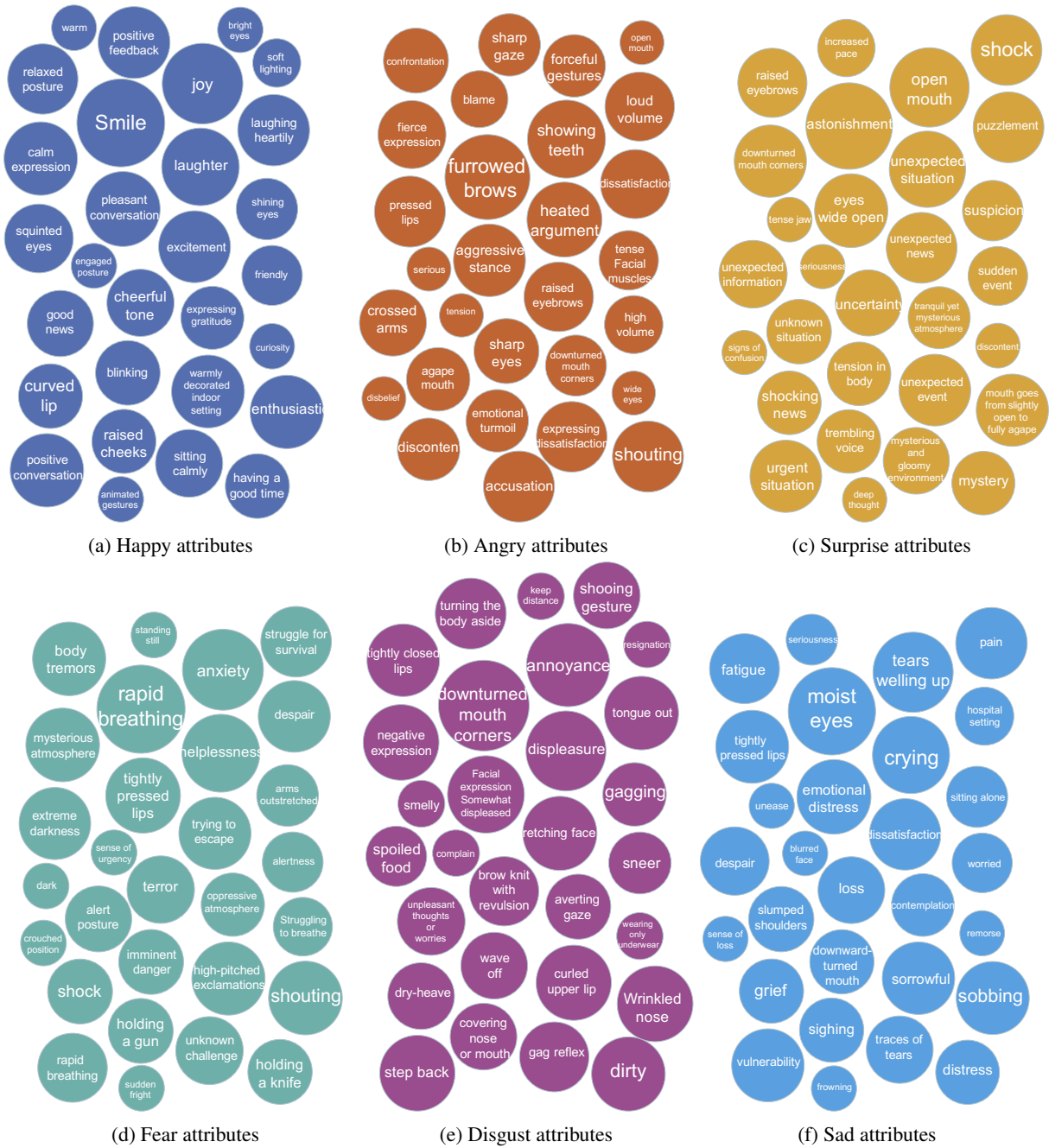


Figure A.3: Taxonomy of Open Emotion Reward Attributes by Emotion Categories. This figure illustrates the fine-grained taxonomy of Emotion Reward Attributes used for multi-modal emotion recognition and reasoning in the proposed reinforcement learning (RL)-based systems. The attributes are grouped according to six core emotion categories: (a) Happy, (b) Angry, (c) Surprise, (d) Fear, (e) Disgust, and (f) Sad. Each emotion cluster includes verifiable and label-consistent cues derived from facial expressions, body language, and contextual features. For instance, the Happy category includes attributes such as ‘smile’, ‘positive feedback’, and ‘cheerful tone’, while Angry encompasses ‘furrowed brows’, ‘forceful gestures’, and ‘heated argument’. These curated attributes form the backbone of the structured attribute reward model, providing interpretable and discriminative signals for aligning model rationales with human-understandable emotion concepts. By leveraging this categorization, the SoVTR framework promotes robust and explainable emotion inference via multi-source evidence integration.



What is the emotion of Person 1?

(a) 

Emotion Analysis for Person 1:
Facial Expression
Eyes wide open, brows slightly furrowed, and mouth tense — these are strong indicators of fear or distress. The child's expression suggests acute emotional arousal, likely due to immediate danger or panic.
Body Language
The child is running with arms bent and hands raised near the head, which reflects a defensive or protective reaction. The body is angled forward in a hurried motion, suggesting an attempt to escape a threatening situation. Overall posture is tense and urgent — classic signs of flight response under fear.
Context
The background shows a burning building, destroyed cars, smoke, rubble, and people running in different directions — a chaotic war or disaster scene. This traumatic environment heavily reinforces the likelihood that the emotion being expressed is fear, possibly accompanied by shock or panic.
Answer: Fear
The convergence of intense facial expression, defensive and urgent body movement, and a dangerous environment makes fear the most plausible emotion for Person 1.

(b) 

Emotion Analysis for Person 1:
Facial Expression
Person 1 is smiling broadly with visible teeth and slightly squinted eyes — strong indicators of genuine happiness or enjoyment. Eyebrows are relaxed, no tension visible in the face, further suggesting a positive affect.
Body Language
Sitting cross-legged and leaning slightly toward the other person, indicating engagement and openness. Arm extended toward the food, showing comfort and involvement in the shared activity. Relaxed shoulders and casual posture reinforce a state of ease.
Context
Outdoors in a large, green park with a picnic setup (food, drinks, basket), implying a leisurely and pleasant social setting. Interacting with another person in a friendly, informal atmosphere.
Answer: Happiness
The combination of a genuine smile, open and relaxed body posture, and a friendly picnic context strongly indicates that Person 1 is experiencing happiness and social enjoyment.

(c) 

Emotion Analysis for Person 1:
Facial Expression
Person 1 is smiling with visible teeth, slightly squinted eyes, and raised cheeks — classic indicators of genuine happiness and positive engagement. The expression shows relaxed facial muscles, suggesting comfort and enjoyment rather than forced emotion.
Body Language
He is standing in a relaxed posture, slightly leaning forward while holding tongs to cook on the grill. His open arm gestures and casual stance indicate that he is actively engaged in the activity, while remaining at ease.
Context
The setting is a leisurely outdoor barbecue with food, family, and a sunny environment next to a pool. The presence of smiling companions (including a child and another adult) reinforces a joyful and relaxed social context.
Answer: Happiness
Combining the cheerful facial expression, open and engaged body language, and relaxed family-oriented context, Person 1 is most likely experiencing a state of happiness, with signs of social bonding and enjoyment.

Figure A.4: Qualitative Emotion Inference for Person 1 Across Diverse Situational Contexts. This figure presents three case studies illustrating the reasoning process for determining the emotional state of Person 1 based on facial expressions, body language, and environmental context. Each scenario demonstrates multi-source visual evidence integration to support emotion classification