

AffectOmni: RL-Verifiable People-Centric Grounded Affective Reasoning for Social and Art-Related Scenes

Yibo Wang, Rui Yang, Jisheng Dang, Bimei Wang, Yitao Wu, Pengfei Cao, Wencan Zhang, Hong Peng, Bin Hu, *Fellow, IEEE*, and Tat-Seng Chua

Abstract—Multimodal large language models (MLLMs) achieve strong performance on VQA and scene understanding, yet affective reasoning remains vulnerable to shortcut behavior. Models may predict correct answers while neglecting people-centric cues such as micro expressions and body language, which weakens traceability and external verification. Prior reinforcement learning approaches mainly reward context or logical coherence without explicitly enforcing attention to human evidence. In addition, LLM as a Judge scoring often suffers from score clustering, which reduces reward discriminability. We propose AffectOmni, a GRPO trained framework for verifiable affective reasoning. AffectOmni introduces People Focus and Temporal Order rewards to encourage people-centric evidence selection and temporally structured reasoning, and it adopts within-group comparative scoring to produce more stable and discriminative reward signals. For verification, a Thinking Summarizer converts free form rationales into executable evidence instructions, which are grounded into pixel level evidence regions via SAM3 to provide an externally auditable interface outside the training loop. Experiments on IntentBench, Daily Omni, and WorldSense show consistent improvements over open source 7B scale base-lines, including gains of 4.66% on emotion recognition and +14.29% on temporally sensitive tasks. Code is available at https://github.com/eliot127825-rgb/AffectOmni_nobody.

Index Terms—Affective Computing, Multimodal Large Language Models, Reinforcement Learning, Visual Grounding, Emotion Recognition, Social Scene Understanding.

I. INTRODUCTION

MULTIMODAL large language models (MLLMs) have made substantial progress [1]–[3], and reinforcement learning, especially GRPO [4], further strengthens their capacity for deep reasoning [5], [6]. This capability is widely expected to support high level social cognition tasks involving human emotion and intent [7]–[9]. However, Visionary-R1 [10] and HumanOmniV2 [11] report a pronounced shortcut problem in multimodal reasoning. Models often guess answers from global cues while failing to analyze fine-grained evidence. HumanOmniV2 introduces explicit context modeling and an LLM as a Judge reward [12], which partially alleviates limitations in global context understanding. Nevertheless, in affective computing [13], [14], we identify a more fundamental challenge. Even when a model predicts the correct emotion, its

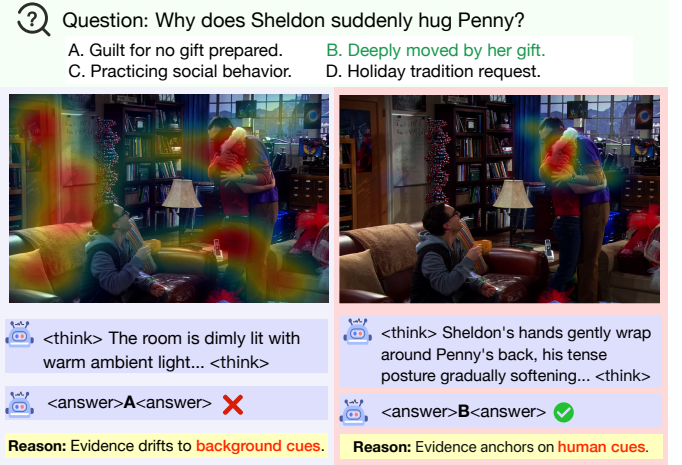


Fig. 1: **Shortcut reasoning vs. human-centric reasoning.** Left: without People Focus reward, evidence attribution drifts to background cues, leading to an incorrect answer. Right: with People Focus reward, evidence anchors on human cues such as hand placement and posture change, enabling evidence-grounded affective inference. The displayed frame is a representative snapshot from the full video, previous events provide context, while human cues provide discriminative evidence.

reasoning often overlooks people-centric cues such as micro expressions and body language and cannot be verified against external evidence [15], [16]. As shown in Fig. 1, current models [11] may rely on background cues rather than people-centric evidence, yielding an untrustworthy reasoning process and sometimes an incorrect answer. This remains a key barrier to deploying affective AI in real world settings [17], [18].

We distill the trustworthiness challenge in affective reasoning into three interrelated questions. (1) How can we design fine-grained reward signals that guide the model to learn a people-centric reasoning mode, rather than relying on coarse global cues. (2) How can we obtain reward signals with sufficient discriminative power during reinforcement learning to reliably separate evidence grounded reasoning from shortcut reasoning. (3) How can we transform long chain reasoning into an executable structured representation and establish a verification mechanism that links reasoning claims to visual evidence.

Existing approaches address these needs only partially. For (1), prior rewards such as global context completeness in Hu-

Yibo Wang, Rui Yang, Jisheng Dang, Bimei Wang, Pengfei Cao, Hong Peng, and Bin Hu are with Lanzhou University, Lanzhou, China.

Yitao Wu is with Hainan University, Haikou, China.

Wencan Zhang and Tat-Seng Chua are with the School of Computing, National University of Singapore, Singapore.

Corresponding authors: Bin Hu, Pengfei Cao, and Wencan Zhang.

manOmniV2 [11] mainly target holistic context understanding, but they do not explicitly model or encourage key affective cues, including micro expressions such as furrowed brows and teary eye corners, body language such as gestures and posture changes, and temporal dynamics that reflect emotion evolution over time [8], [19]. For (2), most LLM as a Judge mechanisms adopt an independent scoring strategy. Each candidate response is scored in isolation. However, LLMs can exhibit calibration drift and score clustering in absolute scoring settings [12], [20], causing reasoning paths of different quality to receive similar scores and yielding weakly discriminative rewards for policy optimization. For (3), current models often generate verbose chains of thought of roughly 800 to 1400 characters that contain rich details but remain loosely structured and low density, making them difficult to directly support downstream applications such as counseling assistance and keyframe extraction [21], [22]. Moreover, label only evaluation cannot test whether the model truly attends to the visual regions that are relevant to the emotion judgment [23], [24].

Our central thesis is that trustworthy affective reasoning depends not only on answer correctness, but more importantly on a reasoning process that is traceable, reward signals that are discriminative, and outcomes that are externally verifiable. Building on this view, we propose the AffectOmni framework. Our main contributions are as follows.

- We propose the first reasoning to evidence grounding paradigm for affective reasoning tasks. We compress model generated affective reasoning into executable, structured evidence instructions and interface them with SAM3 to produce pixel level evidence regions on video frames. This provides externally traceable and verifiable evidence anchors for the people, actions, and temporal cues referenced in reasoning.
- We propose a fine-grained reward mechanism for people-centric reasoning by introducing multi dimensional constraints, including People Focus and Temporal Order, into reinforcement learning. These constraints encourage the model to attend to micro expressions, body language, and human interactions, and to organize reasoning over time.
- During GRPO training, we introduce a within-group comparative scoring strategy that mitigates calibration drift and score clustering in LLM as a Judge evaluation, thereby providing a more stable and discriminative optimization signal.
- We achieve state of the art performance on multimodal benchmarks including IntentBench, Daily-Omni, and WorldSense, showing consistent improvements over open source 7B scale baselines, including **+4.66%** on emotion recognition and **+14.29%** on temporally sensitive tasks.

II. RELATED WORK

Multimodal Affect Understanding. Affective computing aims to enable machines to understand human affect, and the field has progressed from rule based classification to deep reasoning [7], [8]. Early studies focused on unimodal feature extraction and end to end recognition [13], followed by explorations of multimodal fusion strategies. Hazarika et

al. [25] propose modality invariant representation learning, and Hu et al. [26] build a unified framework for affect and emotion recognition. With the rise of MLLMs, affective computing has entered a new phase. Lian et al. [19] introduce an explainable multimodal affective reasoning benchmark. Zhou et al. [27] study an audio visual temporal alignment mechanism. Lian et al. [21] propose AffectGPT for open vocabulary affect understanding. Chain of thought prompting has also been explored for implicit affective reasoning [28], [29]. However, existing methods often exhibit shortcut learning. Models may bypass fine-grained cues such as micro expressions and body language and directly predict labels [10], [11]. We introduce a fine-grained reward design centered on people-centric cues to explicitly encourage attention to key affective evidence.

Reinforcement Learning and Reward Design. Reinforcement learning has become a key approach for eliciting the reasoning capability of MLLMs. Huang et al. [5] and Yuan et al. [6] first apply GRPO to visual reasoning, and subsequent work extends it to video understanding [30], [31], omni modal reasoning [32], [33], and audio visual domains [34], [35]. To mitigate shortcut learning, recent studies design diverse reward functions. HumanOmniV2 [11] proposes context and logical consistency rewards. Chen et al. [36] introduce a step level reward, and Chen et al. [37] design a consistency reward. Other variants include format rewards [38], perceptual rewards [39], and hybrid reward strategies [40], [41]. Most of these methods follow the LLM as a Judge paradigm, where a language model assigns absolute scores to each generated output in isolation. However, calibration drift and score clustering can arise in absolute scoring, so reasoning paths of different quality may receive similar scores, reducing reward discriminability. We propose within-group comparative scoring, which evaluates multiple candidates jointly and enforces relative ranking to obtain more discriminative reward signals.

Visual Grounding and Reasoning Verification. Visual grounding aligns language descriptions with regions in images or videos [23], [24]. Reich et al. [15] propose faithful and plausible metrics for grounding, and Das et al. [16] analyze attention discrepancies between humans and deep networks. In medical VQA, studies validate the effectiveness of the localize then answer paradigm [42], [43]. For trustworthy MLLMs, grounding can mitigate hallucinations [44] and support multimodal fact checking [45], [46]. Yu et al. [47] and Sun et al. [48] use RLHF for behavior alignment, and Zhang et al. [49] employ reinforcement learning to improve multi image grounding. The SAM family [50], [51] enables pixel level video segmentation, providing a technical basis for reasoning verification. Prior work explores visual grounding and post-hoc localization, but it typically treats grounding and affective reasoning as separate components. We integrate a reasoning summarizer with SAM3 based pixel level grounding to provide an externally auditable evidence interface for post-hoc verification of affective reasoning.

III. METHOD

A. Problem Formulation and Framework Overview

Given a multimodal input $\mathbf{x}$ that includes a video frame sequence $\mathbf{v}$, an audio signal $\mathbf{a}$, and a textual question q , our

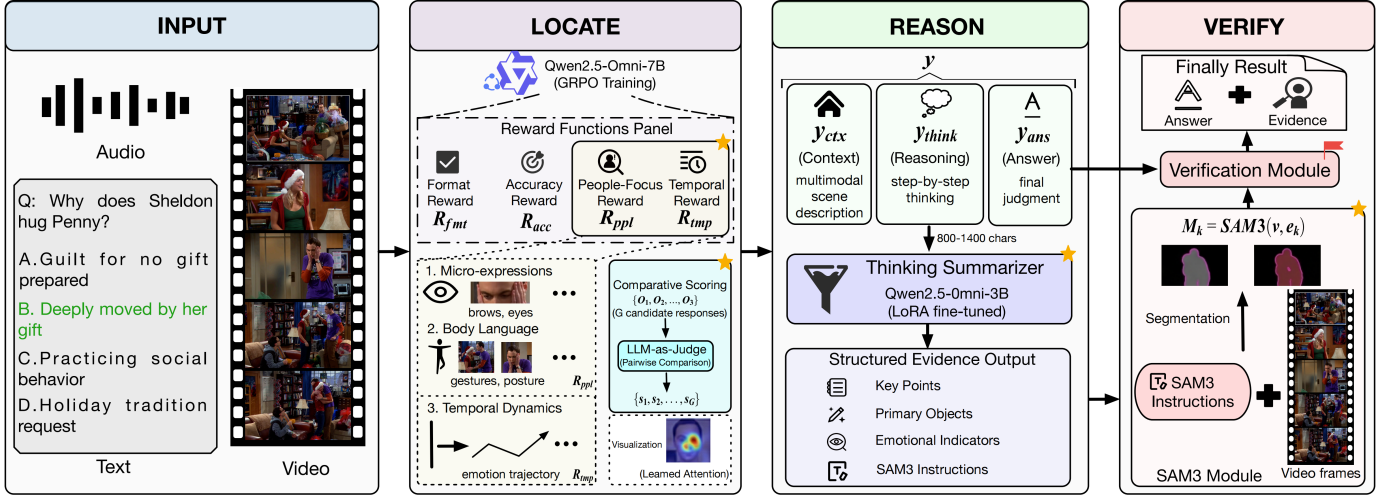


Fig. 2: **AffectOmni framework.** The pipeline consists of three stages. (1) **LOCATE** applies GRPO training with human-centric rewards (R_{ppl} , R_{tmp}) and comparative scoring to guide evidence-grounded reasoning. (2) **REASON** generates structured output and compresses the reasoning chain into a Minimal Evidence Package (MEP) via a Thinking Summarizer. (3) **VERIFY** grounds the MEP to pixel-level masks using SAM3, enabling post-hoc auditing of whether the evidence claims can be localized in the visual input. Stars indicate our key contributions. We further evaluate the verification module in Section V.

goal is to train a policy model π_θ to generate a structured reasoning output $o = \langle context, think, answer \rangle$. The context segment describes the multimodal context, the think segment provides step by step reasoning, and the answer segment gives the final decision.

We train the policy under the Group Relative Policy Optimization (GRPO) framework. For each input $\mathbf{x}$, the model samples G candidate responses $\{o_i\}_{i=1}^G$ and updates the policy using within-group relative advantages. The training objective is given in Eq. (1) and Eq. (2).

$$\mathcal{J}(\theta) = \mathbb{E} \left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} L_{i,t}(\theta) \right], \quad (1)$$

$$L_{i,t}(\theta) = \min \left(r_{i,t} \hat{A}_i, \text{clip}(r_{i,t}, 1 \pm \varepsilon) \hat{A}_i \right), \quad (2)$$

where $r_{i,t} = \pi_\theta(o_{i,t} | \mathbf{x}, o_{i,<t}) / \pi_{\theta_{old}}(o_{i,t} | \mathbf{x}, o_{i,<t})$ is the importance sampling ratio, and $\hat{A}_i = R_i - \text{mean}(\{R_j\}_{j=1}^G)$ denotes the advantage estimate with a within-group mean baseline. The clipping range is $(1 \pm \varepsilon)$.

The proposed AffectOmni framework is illustrated in Fig. 2. It consists of three complementary components, people-centric reward shaping for task aligned affective reasoning, within-group comparative scoring for more discriminative GRPO rewards, and a post-hoc auditing interface that grounds reasoning claims into visual evidence. The overall reward is defined as

$$R = R_{fmt} + R_{acc} + \lambda_p R_{ppl} + \lambda_t R_{tmp}, \quad (3)$$

where R_{fmt} and R_{acc} denote the format reward and the accuracy reward, R_{ppl} and R_{tmp} are our proposed People Focus and Temporal Order rewards, and λ_p , λ_t are weighting coefficients used to balance reward components.

B. People-Centric Fine-Grained Reward Shaping

Existing methods such as HumanOmniV2 adopt a context level reward that mainly evaluates global context completeness,

but they do not explicitly model key affective cues such as micro expressions and body language. To address this limitation, we design two complementary fine-grained reward functions. These rewards are task-aligned for affective reasoning rather than universal criteria across domains. For new tasks, the framework remains unchanged, while only the evaluation criteria in the judge prompts need to be replaced.

People Focus Reward. Let c and h denote the context text and the reasoning text generated by the model. We use an LLM as a Judge model to assess whether the reasoning attends to people-centric cues. As shown in Eq. (4),

$$R_{ppl} = f_{LLM}(c, h; \mathcal{P}_{ppl}) \in \{0, 1\}, \quad (4)$$

where $\mathcal{P}_{ppl}$ is an evaluation prompt that guides the judge to evaluate three dimensions. The first is facial expression description that covers micro expressions such as furrowed brows and teary eyes. The second is body motion analysis that covers non verbal signals such as gestures and posture changes. The third is interpersonal interaction modeling that captures affective exchange and responses between interlocutors.

Temporal Order Reward. Affective states evolve over time, so the model should capture temporal changes along the affective trajectory. We therefore introduce a Temporal Order reward

$$R_{tmp} = f_{LLM}(c, h; \mathcal{P}_{tmp}) \in \{0, 1\}, \quad (5)$$

where $\mathcal{P}_{tmp}$ instructs the judge to evaluate two aspects. The first is the proper use of temporal markers such as initially, then, and finally. The second is temporal coherence of the affective trajectory, such as a gradual transition from surprise to relief.

Implementation Details. To reduce API overhead, we adopt a joint evaluation scheme. We merge $\mathcal{P}_{ppl}$ and $\mathcal{P}_{tmp}$ into a single prompt $\mathcal{P}_{joint}$ and return two scores in one call. In addition, we apply a causal mask so that R_{ppl} and R_{tmp} act only on token positions in the context and think segments, which prevents reward leakage into the answer segment.

C. Within-Group Comparative Scoring

The independent absolute scoring paradigm based on LLM as a Judge has inherent drawbacks. Large language models can exhibit calibration drift and score clustering in absolute scoring, causing reasoning paths of different quality to receive similar scores and reducing reward discriminability.

Comparative Scoring Mechanism. We propose a within-group comparative scoring strategy that presents the G candidates for the same prompt, $\{o_1, \dots, o_G\}$, jointly to the judge model and requires relative comparison rather than absolute scoring. Formally,

$$\{s_1, \dots, s_G\} = f_{\text{cmp}}(\{o_1, \dots, o_G\}; \mathcal{P}_{\text{cmp}}), \quad (6)$$

where $s_i \in [1, 10]$ is the relative score for the i th candidate. The prompt $\mathcal{P}_{\text{cmp}}$ instructs the judge to perform explicit comparison before assigning scores and to produce a differentiated ranking among candidates.

Advantage Computation. During training, we compute an overall reward R_i for each candidate as a weighted sum of multiple reward components

$$R_i = \sum_k \lambda_k r_i^{(k)}, \quad (7)$$

where $r_i^{(k)}$ denotes the k th reward component and λ_k is the corresponding weight. Comparative scoring is not a separate reward term; it is used to estimate judge-based components over the G candidates.

For the G candidates of the same prompt, we compute a within-group mean baseline and obtain the advantage

$$\hat{A}_i = R_i - \bar{R}_g, \quad \bar{R}_g = \frac{1}{G} \sum_{j=1}^G R_j, \quad (8)$$

and optionally apply within-group standardization $\hat{A}_i \leftarrow \hat{A}_i / (\sigma_g + \epsilon)$ to improve reward scale consistency and training stability.

Distributed Implementation. In multi GPU training, comparative scoring requires aggregating candidates across devices. We use a gather and broadcast scheme. Candidates from all GPUs are gathered to the rank 0 process, which calls the judge model to perform comparative scoring. The resulting scores are then broadcast to all ranks to resume backpropagation.

D. Reasoning to Evidence Grounding Framework

Long chain reasoning, often around 800 to 1400 characters, can contain rich analytic details, but it is typically low density and loosely structured, which limits its utility for downstream applications and external verification. We propose a reasoning to evidence grounding framework that converts free form reasoning into an executable structured representation and grounds the resulting evidence instructions through visual segmentation. The Thinking Summarizer and SAM3 are used only for post-hoc evidence execution and visualization-based evaluation. The entire verification stage is excluded from GRPO training and is not used as a reward signal. It serves as a decoupled auditing interface rather than an end-to-end optimization loop.

Thinking Summarizer. We define a mapping from the reasoning output o to a structured summary z as $g_\phi: o \mapsto z$, where z contains four fields, key points, primary objects, emotional indicators, and SAM3 instruction. We train g_ϕ with a two stage knowledge distillation pipeline. In the first stage, a large API based model produces high quality summaries to serve as training targets. In the second stage, we apply LoRA fine tuning to a lightweight base model with an autoregressive objective

$$\mathcal{L}_{\text{sum}} = - \sum_t \log p_\phi(z_t | o, z_{<t}). \quad (9)$$

SAM3 Based Segmentation Integration. The SAM3 instruction field is parsed into a target entity list $\{e_1, \dots, e_K\}$ and used to prompt SAM3 to produce pixel level segmentation over video frames

$$\mathbf{M}_k = \text{SAM3}(\mathbf{v}, e_k), \quad k = 1, \dots, K, \quad (10)$$

where $\mathbf{M}_k$ denotes the segmentation mask for the k th entity across frames.

Verification Framework. Given the SAM3 outputs, we verify whether the reasoning attends to the correct visual regions. For example, if the reasoning states that the man's expression shifts from surprise to relief, then the SAM3 instruction should include the corresponding person entity, and SAM3 should successfully segment the target. The segmentation success rate provides an external indicator of reasoning trustworthiness by establishing an externally auditable linkage between textual evidence claims and pixel level regions across frames. The full reasoning verification pipeline is illustrated in Fig. 3.

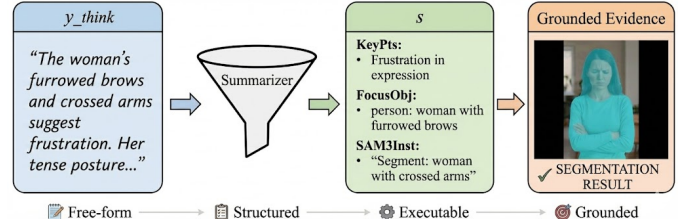


Fig. 3: Visualization of free-form reasoning transformed into structured evidence, i.e., key points, focus objects, and SAM3 instructions for SAM3 based pixel level grounding.

IV. EXPERIMENTS

In this section, we report experimental results and provide further analysis. Implementation details are provided in Appendix.

A. Experimental Setup

Training Data and Configuration. We use Qwen2.5-Omni-7B-Thinker as the base model. To obtain a stable and controllable policy initialization, we follow a standard multimodal reasoning alignment recipe for structured output initialization, enabling the model to consistently produce the three part output $\langle \text{context}, \text{think}, \text{answer} \rangle$. This stage serves to establish a strong baseline and control variables. We apply cold start SFT to stabilize long chain reasoning and format consistency, and then conduct two stage GRPO training to

suppress reasoning collapse and format drift. This stage does not include the proposed rewards R_{ppl} and R_{tmp} . The training corpus contains 24K video audio samples from Video-R1 [52], Social-IQ 2.0 [53] using the training split, and EMER [19]. During GRPO training, we sample $G=4$ candidate responses per instance, set $\varepsilon=0.2$, use a learning rate of 1×10^{-6} , and cap the maximum output length at 2048 tokens. We set reward weights to $\lambda_p=0.2$ and $\lambda_t=0.2$. All experiments run on $4 \times \text{A100 80GB}$ GPUs with DeepSpeed ZeRO Stage 2 for acceleration and memory optimization.

Evaluation Benchmarks. We evaluate on three multimodal benchmarks. IntentBench is our curated benchmark for human intent and emotion understanding. It contains 633 videos and 2,689 questions drawn from the Social-IQ 2.0, EMER, and MDPE subsets. Daily-Omni [27] is a daily scenario audio visual QA benchmark with 684 videos and 1,197 questions covering six task categories. WorldSense [54] is a world knowledge audio visual QA benchmark with 1,662 videos and 3,172 questions spanning eight domains.

Baselines. We compare against two groups of models. The first group includes proprietary models, GPT-4o [55], GPT-o1 (think) [56], Gemini-2.5-Pro (think) [57], Gemini 2.0 Flash, Gemini 2.0 Flash Lite, Gemini 1.5 Pro [58], and Claude 3.5 Sonnet [59]. The second group includes open source omni modal models, Qwen2.5-Omni [3], MiniCPM-o [60], Ola [61], HumanOmniV2 (7B) [11], Unified-IO-2 (8B) [62], VideoLLaMA2 (7B) [63], and VITA-1.5 (7B) [64].

B. Main Results

IntentBench Results. Table I reports results on IntentBench. AffectOmni reaches an average accuracy of 71.89, exceeding the strongest open source baseline HumanOmniV2 (7B) at 69.23. The improvements are moderate but structured, with larger gains on Emotion and How, which are more directly aligned with people-centric evidence. The largest gains occur on Emotion and How, with increases of 4.66 percentage points (pp) and 3.64 pp, the absolute difference between two accuracy percentages. We attribute these gains to the stronger reliance of these categories on people related fine-grained cues and temporally organized reasoning. Compared with prior models that can answer correctly while bypassing human evidence, AffectOmni more reliably forms people-centric evidence chains.

Daily-Omni Results. Table II reports results on Daily-Omni across six task dimensions. Compared with the strongest open source baseline HumanOmniV2 (7B), AffectOmni improves average accuracy from 58.47 to 61.90 and yields consistent gains across several representative dimensions. In particular, Context increases by 4.15 pp, Reason increases by 4.01 pp, and 60s tasks increase by 5.46 pp. On Infer, AffectOmni is slightly lower than HumanOmniV2, suggesting that this dimension relies more on general inference and knowledge transfer than on people-centric evidence constraints. We also note that the proprietary model Gemini 2.0 Flash reaches an average accuracy of 67.84 on this benchmark, indicating that performance under complex daily distributions remains influenced by model scale and general capability. Nevertheless, AffectOmni improves over open source 7B scale baselines

on several key dimensions, suggesting transfer beyond the diagnostic set, although the gain is not uniform across all dimensions.

WorldSense Results. Table III reports domain wise performance on WorldSense. AffectOmni achieves an average accuracy of 48.80, exceeding the open source baseline HumanOmniV2 (7B) at 47.70 and slightly surpassing the proprietary baseline Gemini1.5 Pro. The largest domain wise gain appears in Music, with an improvement of 3.9 pp. Compared with Gemini1.5 Pro, AffectOmni remains behind on Film, Tech, and Perform., which rely more on external knowledge and cross domain semantics. At the same time, AffectOmni shows a clear advantage in Music, leading to a higher overall average. These results suggest that our gains primarily arise from improved reasoning structures that organize people related cues and temporal evidence, rather than from uniform improvements across knowledge intensive domains.

C. Diagnostic Analysis

Effectiveness of Reward Components. To assess how the People-Focus reward R_{ppl} and the Temporal-Order reward R_{tmp} mitigate shortcut reasoning and improve verifiability, we conduct an ablation on IntentBench that varies only reward configurations. Table IV shows that the Baseline with accuracy-only reward reaches 69.23 in Avg. Adding R_{ppl} increases Avg. to 69.78 and improves Emotion to 84.23, indicating that the People-Focus constraint more reliably encourages reliance on people-related cues. Adding R_{tmp} alone yields an Avg. of 69.62 and improves Emotion to 83.67, suggesting that explicit temporal constraints help organize reasoning over time-sensitive evidence.

Finally, the full model reaches 71.89 in Avg., improving by 2.66 pp over the baseline, and increases the When category by 14.29 pp, which confirms the importance of the complete reward design for temporally sensitive questions. We also observe a decrease on Deception. Since the bootstrap confidence interval includes zero, we do not interpret it as a statistically reliable degradation. Instead, it suggests a boundary of the current inductive bias, as deception and sarcasm often require modeling the mismatch between surface behavior and underlying intent. Lastly, although the Baseline, R_{ppl} , and R_{tmp} all obtain 57.14 on When, this does not imply saturation. These settings likely fall into the same granularity bin, whereas the full model moves beyond multiple granularity levels, further suggesting that joint rewards substantially improve temporal reasoning.

Comparative Scoring Analysis. We find that when the reward distribution becomes overly concentrated due to score clustering, or drifts across samples due to calibration drift, the advantage gaps between candidates are compressed. This reduces the effectiveness of GRPO [4] updates. To address this issue, we compare independent absolute scoring with comparative scoring that enforces ranking, focusing on reward discriminability and downstream performance. Because the two schemes use different raw score scales, we apply min max normalization to each set of scores and linearly rescale them to the common interval [1, 10]. This enables fair comparison of CV, defined as Std over Mean, within-group CV, and the shape of the score distribution.

TABLE I: **Comparison on IntentBench.** We report category-wise intent understanding and social reasoning performance. Models marked with $\dagger$ are proprietary. The last three rows are ablation variants of AffectOmni built incrementally on HumanOmniV2 by adding People-Focus (R_{ppl}), Temporal-Order (R_{tmp}), and both rewards jointly (Full Model).

Methods	LMM	Why	How	What	When	Who/Which	Other	Emotion	Deception	Avg
<i>Proprietary MLLMs</i>										
GPT-4o $\dagger$	–	61.46	55.69	60.00	35.71	76.00	63.31	60.99	59.00	59.98
GPT-o1 $\dagger$ (think)	–	68.19	65.82	66.04	57.14	76.00	68.83	67.26	59.50	66.69
Gemini-2.5-Pro $\dagger$ (think)	–	68.57	67.41	65.12	57.14	64.00	70.03	68.23	60.00	67.15
<i>Open-Source MLLMs</i>										
MiniCPM-o	8B	57.87	53.48	57.14	57.14	68.00	61.14	23.85	49.50	54.51
VITA-1.5	7B	53.15	49.36	51.66	71.42	64.00	61.14	53.20	59.50	54.17
Ola	7B	60.60	55.37	56.87	64.28	76.00	62.91	46.66	44.50	57.41
Qwen2.5-Omni	7B	62.60	63.44	63.53	57.14	76.00	69.03	59.74	63.50	64.20
HumanOmniV2	7B	66.05	64.08	68.75	57.14	76.00	74.56	80.78	66.50	69.23
<i>AffectOmni Ablation (built on HumanOmniV2)</i>										
+ People-Focus (R_{ppl})	7B	66.48	67.25	68.33	57.14	76.00	75.35	84.23	63.50	69.78
+ Temporal-Order (R_{tmp})	7B	64.33	68.20	68.96	57.14	80.00	74.16	83.67	60.50	69.62
Full Model (AffectOmni)	7B	66.91	67.72	70.21	71.43	76.00	74.95	85.44	62.50	71.89

TABLE II: Comparison with existing models on Daily-Omni. Models marked with $\dagger$ are proprietary. We report accuracy (%) on representative sub-dimensions; full results are provided in the supplemental material.

Method	Context	Infer.	Reason.	60s	Avg
Gemini 2.0 Flash $\dagger$	63.73	76.62	75.43	56.57	67.84
Gemini 2.0 Flash Lite $\dagger$	58.03	74.03	72.00	53.01	61.32
Unified-IO-2 (8B)	26.42	35.06	29.71	30.00	28.24
VideoLLaMA2 (7B)	35.75	40.91	34.29	31.82	35.17
Qwen2.5-Omni (7B)	38.86	57.79	61.71	38.36	47.45
Ola (7B)	39.89	61.03	66.28	48.72	49.87
MiniCPM-o (7B)	49.22	68.83	61.14	52.00	53.13
HumanOmniV2 (7B)	51.81	72.72	74.28	53.09	58.47
+ People-Focus (R_{ppl})	58.55	75.32	77.14	60.36	62.57
+ Temporal-Order (R_{tmp})	55.96	73.38	77.71	58.73	62.41
AffectOmni (Ours)	55.96	72.08	78.29	58.55	61.90

TABLE III: Comparison with existing models on WorldSense. Models marked with $\dagger$ are proprietary. We report accuracy (%) on representative domains; full results are provided in the supplemental material.

Method	Tech	Film	Perform.	Music	Avg
Claude 3.5 Sonnet $\dagger$	43.70	36.50	30.70	33.90	34.80
GPT-4o $\dagger$	48.00	43.50	41.90	42.70	42.60
Gemini 1.5 Pro $\dagger$	53.70	50.40	52.40	42.00	48.00
Unified-IO-2 XXL (7B)	27.10	23.70	25.50	27.30	25.90
VideoLLaMA2 (7B)	29.40	24.50	26.20	27.10	25.40
VITA-1.5 (7B)	38.20	39.80	41.20	39.90	36.90
Qwen2.5-Omni (7B)	47.80	43.80	48.30	47.30	45.40
HumanOmniV2 (7B)	49.60	47.50	48.40	43.30	47.70
+ People-Focus (R_{ppl})	50.90	50.40	46.50	46.70	47.60
+ Temporal-Order (R_{tmp})	51.30	48.50	48.80	46.40	48.60
AffectOmni (Ours)	51.50	49.30	50.40	47.20	48.80

Table V shows that comparative scoring substantially increases dispersion and within-group separability of reward signals relative to independent scoring. The overall CV increases from 0.153 to 0.325, and the within-group CV increases from

TABLE IV: We report accuracy (%) on IntentBench across key categories. All variants share the same base model and training setup, and differ only in the enabled reward terms, showing the marginal gain brought by each component.

Configuration	Emotion	Deception	When	Why	Avg
Baseline (Acc-only)	80.78	66.50	57.14	66.05	69.23
+ People-Focus (R_{ppl})	84.23	63.50	57.14	66.48	69.78
+ Temporal-Order (R_{tmp})	83.67	60.50	57.14	64.33	69.62
Full Model	85.44	62.50	71.43	66.91	71.89

0.049 to 0.138. These results indicate clearer relative differences among candidates within the same group, which yields a more informative and stable update signal for policy optimization. Consistent with these statistics, the unified scale visualization in Fig. 4 reveals pronounced clustering in the high score range under independent scoring. In contrast, comparative scoring reduces perfect score clustering, allocates more probability mass to the 9 to 10 and 8 to 9 ranges, and maintains non zero coverage in the low score tail. For example, the 1 to 2 bin has 7.31. This makes extreme calibration shifts caused by absolute scale drift less likely. Correspondingly, IntentBench average accuracy improves from 69.78 to 71.89, showing that stronger reward discriminability translates into downstream reasoning gains.

TABLE V: Comparison of scoring strategies. “Ind.” denotes independent scoring. “Comp.” denotes our comparative scoring. CV is the coefficient of variation, computed as standard deviation over mean.

Metric	Independent	Comparative
CV	0.153	0.325
Within-Group CV	0.049	0.138
IntentBench Avg (%)	69.78	71.89

Reasoning Verification Framework. A core challenge in

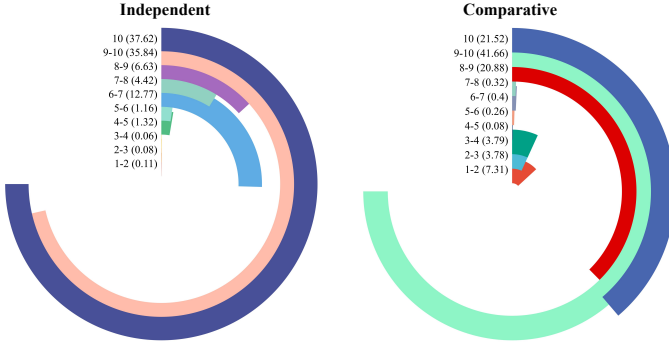


Fig. 4: **Reward Signal Discrimination Analysis.** Each arc represents a score interval (1–10 scale), with arc length proportional to the percentage of samples falling in that interval. Left (Independent): LLM as a Judge scores each candidate in isolation, causing severe high score clustering, over 73% of scores fall in the 9–10 range. Right (Comparative): within-group comparative scoring forces relative ranking among candidates, yielding a more balanced distribution with meaningful differentiation across the full score spectrum.

affective reasoning is to verify whether a model truly attends to the visual evidence it cites in its analysis. To address this challenge, we introduce a reasoning verification framework in AffectOmni that connects reasoning with visual localization. As shown in Fig. 5, given a video and the question “Is the man on the left surprised?”, AffectOmni generates a structured reasoning chain. It identifies that the man in a checkered shirt appears calm, with no visual cues of surprise such as widened eyes or a startle posture. The reasoning is then passed to a summarization module that extracts the focus entities in the video, for example the man in a checkered shirt with a calm and neutral expression, and converts them into SAM3 segmentation instructions. Finally, SAM3 produces pixel level masks that visualize the regions referenced by the reasoning. This provides a verifiable linkage between the language description and visual evidence in the video frames, enabling automatic consistency checks and human inspection of whether the claimed content is grounded in the attended regions.

Statistical Significance of Performance Gains. To verify that the observed gains reflect systematic effects of the proposed reward design rather than random variation, this study conducts one-sided paired bootstrap tests with 10,000 resamples on the full IntentBench test set ($n = 2,689$), with results reported in Table VI.

AffectOmni achieves statistically significant improvements in both categories directly targeted by R_{ppi} . The Emotion category yields the largest absolute gain at +4.66 pp ($p = 0.008$, 95% CI [+1.2, +8.3] pp), while How achieves the strongest significance at $p = 0.002$ with +3.64 pp ([+1.3, +5.9] pp). The lower p -value on How, despite its smaller absolute gain, is explained by its much larger sample size ($n = 632$ vs. $n = 133$), which provides greater statistical power. Both confidence intervals exclude zero, confirming that these gains arise from the reward mechanism rather than sampling noise.

Beyond the directly targeted categories, Why, What, and Other show positive but non-significant trends (+0.86, +1.46,

TABLE VI: One-sided paired bootstrap significance test (10,000 resamples) on IntentBench ($n = 2,689$), testing whether AffectOmni > HumanOmniV2 per category. Δ denotes the mean accuracy difference (AffectOmni – HumanOmniV2). Shaded rows are categories directly targeted by R_{ppi} . ** $p < 0.01$; * $p < 0.05$.

Category	n	Δ (pp)	95% CI (pp)	p
Emotion ($\leftarrow R_{\text{ppi}}$)	133	+4.66**	[+1.2, +8.3]	0.008
How ($\leftarrow R_{\text{ppi}}$)	632	+3.64**	[+1.3, +5.9]	0.002
Why	698	+0.86	[−1.4, +3.2]	0.484
What	480	+1.46	[−1.2, +4.2]	0.317
Other	507	+0.39	[−1.8, +2.6]	0.779
Deception	200	−4.00	[−9.0, +1.0]	1.000

+0.39 pp) with confidence intervals spanning zero, indicating partial but unreliable transfer of people-centric reasoning capabilities. The Deception category stands apart with a decline of −4.00 pp ($p = 1.000$, 95% CI [−9.0, +1.0] pp), the wide interval reflects high within-category variance, so the drop cannot be distinguished from random variation. Deception requires detecting mismatches between surface behavior and underlying intent, a pragmatic reasoning demand where stronger reliance on observable affective cues may reduce rather than increase sensitivity.

Cross-Domain Generalization on General Video QA. A natural concern is whether affective RL training causes catastrophic forgetting of general video understanding capabilities. To test this, this study evaluates AffectOmni on NExT-QA [65], a general-purpose video QA benchmark covering causal, temporal, and descriptive reasoning, without any domain-specific fine-tuning.

As shown in Table VII, AffectOmni achieves 80.52% overall accuracy on the multiple-choice test set, surpassing HumanOmniV2 (79.78%) and several general-purpose video models of comparable scale. Fine-grained results in Table VIII further reveal that the improvements are not uniform. AffectOmni outperforms HumanOmniV2 on the Temporal subset (79.44% vs. 76.50%) and the Descriptive subset (86.48% vs. 83.77%), while HumanOmniV2 retains a slight advantage on Causal (80.50% vs. 79.28%). The gains on Temporal and Descriptive are consistent with the design intent of this study: People-Focus and Temporal-Order rewards train transferable capabilities in entity anchoring, temporal organization, and evidence chain construction, which benefit reasoning tasks beyond affective inference. The modest Causal deficit suggests that causal inference relies more on general semantic priors not directly addressed by the reward design. Together, these results indicate that people-centric RL training does not degrade general video understanding and instead produces structured reasoning capabilities that transfer positively across task domains. Representative qualitative comparisons in Appendix F further illustrate the improvements in micro-expression analysis and temporal affective reasoning.

V. DISCUSSION

This section evaluates the Minimal Evidence Package (MEP) as an Executable Evidence Interface (EEI). We emphasize

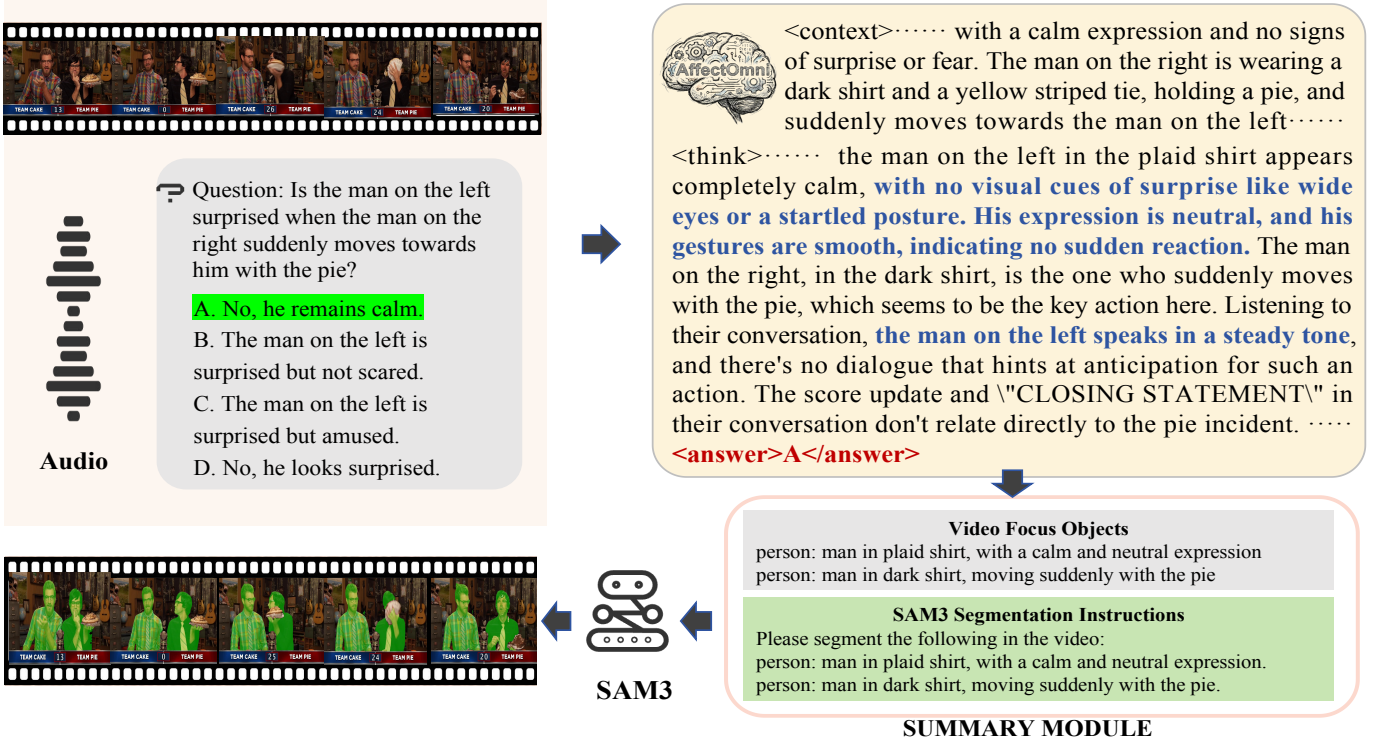


Fig. 5: **Reasoning to evidence grounding framework.** The framework connects structured reasoning chains to pixel-level grounding via SAM3: AffectOmni summarizes focus entities from the reasoning, converts them into segmentation instructions, and visualizes the exact regions to enable automatic consistency checks and human inspection.

TABLE VII: Results on NExT-QA multiple-choice test set. AffectOmni surpasses HumanOmniV2 without domain-specific fine-tuning, confirming that affective RL training does not cause catastrophic forgetting of general video understanding.

Method	Size	Overall Acc. (%)
LLaVA-Video [66]	7B	83.20
PLLaVA [67]	7B	81.00
Magma [68]	8B	80.90
AffectOmni (Ours)	7B	80.52
HumanOmniV2 [11]	7B	79.78
LLaVA-OneVision [69]	7B	79.40
mPLUG-Owl3 [70]	8B	78.60
VideoChat2 [71]	8B	63.20

TABLE VIII: Fine-grained NExT-QA results by question type. AffectOmni shows gains over HumanOmniV2 on Temporal and Descriptive subsets, consistent with the transferable reasoning capabilities trained by People-Focus and Temporal-Order rewards.

Method	Causal	Temporal	Descriptive
AffectOmni	79.28	79.44	86.48
HumanOmniV2	80.50	76.50	83.77

that its goal is not segmentation quality itself, but converting reasoning into an executable and checkable evidence carrier for post-hoc auditing. This abstraction is important in high stakes scenarios such as clinical decision support or affective companion systems. In such settings, a system should output not only conclusions but also evidence packages that are auditable and reviewable. This supports explanation and provides a safety fallback.

MEP encourages reasoning to produce locatable elements, for example the focus person, key actions, and temporal cues, and maps them into executable verification instructions. MEP interfaces with downstream verifiers through the EEI. The interface layer uses {FocusObj, TimeSpan, EvidenceType} as a minimal query form, and heterogeneous verifiers can be integrated with lightweight adaptation. This enables reuse of the auditing interface without modifying the upstream reasoning compression core. Our current implementation uses SAM3 pixel level localization as one instance. The same interface can extend to keyframe retrieval, object tracking, speaker segment localization, and textual evidence retrieval.

To validate these claims and quantify MEP as an EEI rather than a display component, we design three lightweight diagnostic experiments without retraining. These experiments probe evidence extractability, evidence to verification consistency, and falsifiability under external auditing.

A. Evidence Extractability

The first experiment tests whether the reasoning chain provides a stable basis of structured information. We randomly sample 100 think outputs from IntentBench and tally three evidence elements that are directly relevant to generating verification instructions. Visual evidence refers to people related observable cues such as expression, gaze, and posture. Temporal evidence refers to order, change, or stage descriptions over time. Behavioral evidence refers to executable actions or interaction cues such as gestures, approach and avoidance, and conversational interaction. To avoid missed detections and ambiguity from keyword matching, we use Qwen-Max as the

judge and apply a strict binary criterion. We also require a minimal supporting span and apply schema constrained parsing to restrict the output format and reduce drift from free form generation. This statistic is used to diagnose interface feasibility rather than as a final evaluation metric.

Table IX and Table X show that the occurrence rates of the three evidence types are 84%, 72%, and 89%, and that 93% of samples contain at least two evidence types. These results indicate that the model produces composable evidence primitives for most samples, which provides the information basis for the Summarizer to compress reasoning into an MEP.

TABLE IX: Occurrence rates of three types of structured evidence (Visual/Temporal/Behavioral) and coverage rate of ≥ 2 evidence types from 100 randomly sampled `think` outputs.

Evidence Type	Count	Rate
Visual Evidence	84	84%
Temporal Evidence	72	72%
Behavioral Evidence	89	89%
≥ 2 Types	93	93%

TABLE X: Evidence occurrence rates grouped by question type.

Type	Samples	Visual	Temporal	Behavioral
Other	28	75%	75%	89%
Why	27	81%	70%	81%
How	19	95%	74%	89%
What	14	86%	64%	93%
Deception	8	88%	62%	100%
Emotion	4	100%	100%	100%
Overall	100	84%	72%	89%

B. Evidence to Verification Consistency

The second experiment evaluates the operability of the EEI, namely whether the focus object description in `think` provides sufficient detail for successful localization by downstream visual verifiers. We count elements including person mentions, appearance attributes such as color and clothing, visual features such as expression and posture, and locatable actions. Based on these elements, we define two grounding criteria. Basic Grounding requires an explicitly mentioned and segmentable person. High Quality Grounding requires a person mention and at least one executable localization cue, either a visual feature or a locatable action.

Table XI shows that 98% of samples mention a person, 91% include visual feature descriptions, and 94% include locatable actions. Accordingly, 98% of samples satisfy Basic Grounding and 86% satisfy High Quality Grounding. By question type in Table XII, Basic Grounding ranges from 93% to 100% across categories, and High Quality Grounding ranges from 82% to 100%, which indicates consistent executability across question types. Appearance attributes appear in only 39% of cases, which is markedly lower than visual features and action cues. This suggests that current reasoning relies more on dynamic evidence such as expression, posture, and actions. Under occlusion or in crowded multi person scenes, grounding can become more

ambiguous without additional discriminative appearance cues or spatial references.

TABLE XI: Occurrence rates of grounding-relevant elements and two levels of grounding criteria from 100 randomly sampled reasoning chains.

Indicator	Count	Rate
Mentions Person	98	98%
Has Appearance (color/clothing)	39	39%
Has Visual Feature (expression/posture)	91	91%
Has Locatable Action	94	94%
Basic Grounding	98	98%
High-Quality Grounding	86	86%

TABLE XII: Grounding attainment rates grouped by question type.

Type	Samples	Basic	High-Quality
Other	28	96%	82%
Why	27	100%	85%
How	19	100%	89%
What	14	93%	86%
Deception	8	100%	88%
Emotion	4	100%	100%
Overall	100	98%	86%

C. External Auditing and Falsifiability

The third experiment evaluates falsifiability through external auditing. It tests whether the auditing interface provides an external signal that helps assess answer trustworthiness, rather than producing superficially plausible visualizations for arbitrary instructions. We randomly sample 100 items from IntentBench and ask annotators to view only the question and the SAM3 segmentation results. Annotators do not have access to the model textual reasoning or the ground truth label. They provide a binary relevance judgment, relevant or irrelevant, on whether the segmentation focuses on objects related to the question target. Relevant means that the segmented subject matches the asked for object and covers the key person.

Table XIII shows that when segmentation is judged relevant, conditional accuracy is 83.3%. When segmentation is judged irrelevant, conditional accuracy drops to 35.3%, a difference of 48.0 percentage points. The contingency table yields an overall agreement rate of 77%, computed as $(55 + 22)/100$, and the Phi coefficient is $\phi = 0.48$, which indicates a moderate

TABLE XIII: Annotators judge, based only on SAM3 segmentation, whether it is related to the question target, and we report the contingency table between this relevance judgment and answer correctness, together with conditional accuracy (n=100).

	Seg. Relevant	Seg. Irrelevant	Total
Answer Correct	55	12	67
Answer Wrong	11	22	33
Conditional Acc.	83.3%	35.3%	67%

positive association between focusing on the key object and answering correctly. Among the 67 correct-answer cases, 12 have irrelevant segmentation, corresponding to a 17.9% false-negative rate. These false negatives reduce diagnostic coverage but do not affect training or final answer accuracy because the verifier is post-hoc and excluded from GRPO optimization. Such errors may come from summarization bias in evidence-query extraction, SAM3 grounding failure, or downstream relevance-judgment noise. Together, these results show that annotators do not need to read the reasoning text. By inspecting the post-hoc visual evidence alone, they can obtain an external signal for answer reliability, which supports the value of the MEP as an evidence carrier that is external, operational, and auditable. When segmentation decouples from the question target, accuracy declines substantially. This demonstrates falsifiability because evidence mismatch exposes potentially unreliable outputs in an observable manner, rather than always producing seemingly reasonable results. Representative failure cases in Appendix G further illustrate pragmatic reasoning failures in sarcasm interpretation and reasoning-selection disconnects.

VI. CONCLUSION

This paper proposes AffectOmni for verifiable affective reasoning, addressing the trustworthiness bottleneck in multimodal emotion and intent understanding where models produce correct answers while bypassing fine grained people-centric evidence. We introduce People-Focus and Temporal-Order rewards into GRPO training to incorporate people-centered evidence selection and temporal organization into the reinforcement learning objective, and propose within-group comparative scoring to mitigate score clustering and calibration drift in LLM as a Judge evaluation. A post-hoc reasoning-to-evidence auditing interface further compresses chain-of-thought outputs into executable evidence instructions grounded via SAM3, supporting external consistency and falsifiability diagnostics (Section V). Experiments on IntentBench, Daily-Omni, and WorldSense confirm consistent gains over open-source 7B scale baselines on people-centric and temporally sensitive tasks. Temporal interval localization remains coarse and error propagation in the verification module is unquantified, future work will address localization precision, auxiliary verification rewards, and reward generalizability.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62227807 and Grant No. U24B20186), in part by the Brain Science and Brain-like Intelligence Technology—National Science and Technology Major Project (No. 2021ZD0200600, No. 2021ZD0200408), and in part by the WQ & UCAS Research Academy Intelligent Computing Center (WRA-ICC) and the Supercomputing Center of Lanzhou University.

APPENDIX SUPPLEMENTARY MATERIAL

The supplementary material provides additional details and analyses supporting the main text. Appendix A describes the

full three-stage training pipeline and Thinking Summarizer configuration. Appendix B reports complete per-subcategory results on Daily-Omni and WorldSense. Appendix C analyzes how the group size G affects reward discriminability and verifies the absence of positional bias. Appendix D examines sensitivity to the reward weights λ_p and λ_t . Appendix E evaluates reasoning chain quality along five diagnostic dimensions using both automatic and human assessment. Appendices F and G present qualitative comparisons of success and failure cases, respectively, illustrating the behavioral patterns induced by the proposed rewards.

REFERENCES

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [2] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.
- [3] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang *et al.*, “Qwen2. 5-omni technical report,” *arXiv preprint arXiv:2503.20215*, 2025.
- [4] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv preprint arXiv:2402.03300*, 2024.
- [5] W. Huang, B. Jia, Z. Zhai, S. Cao, Z. Ye, F. Zhao, Z. Xu, Y. Hu, and S. Lin, “Vision-r1: Incentivizing reasoning capability in multimodal large language models,” *arXiv preprint arXiv:2503.06749*, 2025.
- [6] R. Yuan, C. Xiao, S. Leng, J. Wang, L. Li, W. Xu, H. P. Chan, D. Zhao, T. Xu, Z. Wei *et al.*, “VI-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning,” *arXiv preprint arXiv:2507.22607*, 2025.
- [7] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, “Affective computing and sentiment analysis,” in *A practical guide to sentiment analysis*. Springer, 2017, pp. 1–10.
- [8] S. Li and W. Deng, “Deep facial expression recognition: A survey,” *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.
- [9] H. Yang, Y. Zhao, Y. Wu, S. Wang, T. Zheng, H. Zhang, Z. Ma, W. Che, S. Wang, S. Wei *et al.*, “Large language models meet text-centric multimodal sentiment analysis: A survey,” *Science China Information Sciences*, vol. 68, no. 10, pp. 1–29, 2025.
- [10] J. Xia, Y. Zang, P. Gao, S. Li, and K. Zhou, “Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning,” *arXiv preprint arXiv:2505.14677*, 2025.
- [11] Q. Yang, S. Yao, W. Chen, S. Fu, D. Bai, J. Zhao, B. Sun, B. Yin, X. Wei, and J. Zhou, “Humanomniv2: From understanding to omnimodal reasoning with context,” *arXiv preprint arXiv:2506.21277*, 2025.
- [12] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, “Judging llm-as-a-judge with mt-bench and chatbot arena,” *Advances in neural information processing systems*, vol. 36, pp. 46 595–46 623, 2023.
- [13] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages,” *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.
- [14] G. Hu, Y. Xin, W. Lyu, H. Huang, C. Sun, Z. Zhu, L. Gui, R. Cai, E. Cambria, and H. Seifi, “Recent trends of multimodal affective computing: A survey from nlp perspective,” *arXiv preprint arXiv:2409.07388*, 2024.
- [15] D. Reich, F. Putze, and T. Schultz, “Measuring faithful and plausible visual grounding in VQA,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3129–3144. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.206/>
- [16] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, “Human attention in visual question answering: Do humans and deep networks look at the same regions?” *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, 2017.
- [17] Z. Dong, C. Chen, C. Liao, and X. M. Chen, “Integrating large language models and affective computing for human-machine symbiosis in intelligent driving,” *The Innovation*, vol. 6, no. 12, 2025.

- [18] M. M. Amin, R. Mao, E. Cambria, and B. W. Schuller, "A wide evaluation of chatgpt on affective computing tasks," *IEEE Transactions on Affective Computing*, vol. 15, no. 4, pp. 2204–2212, 2024.
- [19] Z. Lian, H. Sun, L. Sun, H. Gu, Z. Wen, S. Zhang, S. Chen, M. Xu, K. Xu, K. Chen *et al.*, "Explainable multimodal emotion recognition," *arXiv preprint arXiv:2306.15401*, 2023.
- [20] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu *et al.*, "Large language models are not fair evaluators," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9440–9450.
- [21] Z. Lian, H. Chen, L. Chen, H. Sun, L. Sun, Y. Ren, Z. Cheng, B. Liu, R. Liu, X. Peng, J. Yi, and J. Tao, "AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: <https://openreview.net/forum?id=xmbdACIOxu>
- [22] M. Niu, Y. El-Tawil, A. Romana, and E. M. Provost, "Rethinking emotion annotations in the era of large language models," *IEEE Transactions on Affective Computing*, 2025.
- [23] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1307–1315.
- [24] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.
- [25] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.
- [26] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 7837–7851. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.534/>
- [27] Z. Zhou, R. Wang, and Z. Wu, "Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities," *arXiv preprint arXiv:2505.17862*, 2025.
- [28] W. Lai, H. Xie, G. Xu, and Q. Li, "Rvisa: reasoning and verification for implicit sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 16, no. 3, pp. 1760–1771, 2025.
- [29] L. Yang, X. Wang, X. Zhou, Z. Wu, and N. Tan, "Application of multiple chain-of-thought in contrastive reasoning for implicit sentiment analysis," *arXiv preprint arXiv:2503.07140*, 2025.
- [30] Q. Wang, Y. Yu, Y. Yuan, R. Mao, and T. Zhou, "VideoRFT: Incentivizing video reasoning capability in MLLMs via reinforced fine-tuning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: <https://openreview.net/forum?id=3pORFyKzh1>
- [31] X. Li, Z. Yan, D. Meng, L. Dong, X. Zeng, Y. He, Y. Wang, Y. Qiao, Y. Wang, and L. Wang, "Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning," *arXiv preprint arXiv:2504.06958*, 2025.
- [32] H. Zhong, M. Zhu, Z. Du, Z. Huang, C. Zhao, M. Liu, W. Wang, H. Chen, and C. Shen, "Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration," *arXiv preprint arXiv:2505.20256*, 2025.
- [33] M. Zhu, H. Zhong, C. Zhao, Z. Du, Z. Huang, M. Liu, H. Chen, C. Zou, J. Chen, M. Yang *et al.*, "Active-o3: Empowering multimodal large language models with active perception via grpo," *arXiv preprint arXiv:2505.21457*, 2025.
- [34] Z. Xing, X. Hu, C.-W. Fu, W. Wang, J. Dai, and P.-A. Heng, "Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning," *arXiv preprint arXiv:2505.04623*, 2025.
- [35] Q. Wang, X. Jiang, L. He, J. Wu, and N. Mesgarani, "Sightsound-r1: Cross-modal reasoning distillation from vision to audio language models," *arXiv preprint arXiv:2509.15661*, 2025.
- [36] H. Chen, X. Lou, X. Feng, K. Huang, and X. Wang, "Unveiling chain of step reasoning for vision-language models with fine-grained rewards," *arXiv preprint arXiv:2509.19003*, 2025.
- [37] Y. Chen, Y. Ge, R. Wang, Y. Ge, J. Cheng, Y. Shan, and X. Liu, "GRPO-CARE: Consistency-aware reinforcement learning for multimodal reasoning," 2026. [Online]. Available: <https://openreview.net/forum?id=XoUJk0aDCN>
- [38] J. Zhang, J. Huang, H. Yao, S. Liu, X. Zhang, S. Lu, and D. Tao, "R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization," *arXiv preprint arXiv:2503.12937*, 2025.
- [39] Z. Guo, M. Hong, and T. Jin, "Observe-r1: Unlocking reasoning abilities of mllms with dynamic progressive reinforcement learning," *arXiv preprint arXiv:2505.12432*, 2025.
- [40] P. Wang, Y. Wei, Y. Peng, X. Wang, W. Qiu, W. Shen, T. Xie, J. Pei, J. Zhang, Y. Hao *et al.*, "Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning," *arXiv preprint arXiv:2504.16656*, 2025.
- [41] M. Hong, Z. Guo, Y. Xia, Z. Wang, Z. Zhang, T. Jin, and Z. Zhao, "Apo: Enhancing reasoning ability of mllms via asymmetric policy optimization," *arXiv preprint arXiv:2506.21655*, 2025.
- [42] D. Nguyen, M. K. Ho, H. Ta, T. T. Nguyen, Q. Chen, K. Rav, Q. D. Dang, S. Ramchandre, S. L. Phung, Z. Liao *et al.*, "Localizing before answering: A benchmark for grounded medical visual question answering," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025, pp. 7670–7678.
- [43] C. Chen, S. Anjum, and D. Gurari, "Grounding answers for visual questions asked by visually impaired people," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19098–19107.
- [44] A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto, "Multi-modal hallucination control by visual information grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14303–14312.
- [45] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla *et al.*, "Factify 2: A multimodal fake news and satire news dataset," *arXiv preprint arXiv:2304.03897*, 2023.
- [46] R. F. Cekineli, P. Karagoz, and Ç. Çöltekin, "Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 4622–4633.
- [47] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun *et al.*, "Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13807–13816.
- [48] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L. Gui, Y.-X. Wang, Y. Yang *et al.*, "Aligning large multimodal models with factually augmented rlhf," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 13088–13110.
- [49] B. Zhang, H. Li, T. Zhang, C. Yan, J. Cai, and Y. Hao, "Improving the reasoning of multi-image grounding in mllms via reinforcement learning," *arXiv preprint arXiv:2507.00748*, 2025.
- [50] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.
- [51] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.
- [52] K. Feng, K. Gong, B. Li, Z. Guo, Y. Wang, T. Peng, B. Wang, and X. Yue, "Video-r1: Reinforcing video reasoning in mllms," *arXiv preprint arXiv:2503.21776*, 2025.
- [53] A. Wilf, L. Mathur, S. Mathew, C. Ko, Y. Kebe, P. P. Liang, and L.-P. Morency, "Social-iq 2.0 challenge: Benchmarking multimodal social understanding," <https://github.com/abwilf/Social-IQ-2.0-Challenge>, 2023.
- [54] J. Hong, S. Yan, J. Cai, X. Jiang, Y. Hu, and W. Xie, "Worldsense: Evaluating real-world omnimodal understanding for multimodal llms," *arXiv preprint arXiv:2502.04326*, 2025.
- [55] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.
- [56] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney *et al.*, "Openai o1 system card," *arXiv preprint arXiv:2412.16720*, 2024.
- [57] Google DeepMind, "Gemini 2.5 Pro Preview: even better coding performance," 2025, blog post. Available at: <https://deepmind.google/blog/gemini-25-pro-preview-even-better-coding-performance/> (accessed 2026-01-21).
- [58] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

- [59] Anthropic, “Introducing the next generation of claude,” 2024, accessed: 2024-10-22. [Online]. Available: <https://www.anthropic.com/news/claude-3-family>
- [60] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, “Minicpm-v: A gpt-4v level mllm on your phone,” *arXiv preprint arXiv:2408.01800*, 2024.
- [61] Z. Liu, Y. Dong, J. Wang, Z. Liu, W. Hu, J. Lu, and Y. Rao, “Ola: Pushing the frontiers of omni-modal language model,” *arXiv preprint arXiv:2502.04328*, 2025.
- [62] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi, “Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 439–26 455.
- [63] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao *et al.*, “Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms,” *arXiv preprint arXiv:2406.07476*, 2024.
- [64] C. Fu, H. Lin, X. Wang, Y.-F. Zhang, Y. Shen, X. Liu, H. Cao, Z. Long, H. Gao, K. Li *et al.*, “Vita-1.5: Towards gpt-4o level real-time vision and speech interaction,” *arXiv preprint arXiv:2501.01957*, 2025.
- [65] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, “Next-qa: Next phase of question-answering to explaining temporal actions,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9777–9786.
- [66] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, “Llava-video: Video instruction tuning with synthetic data,” *arXiv preprint arXiv:2410.02713*, 2024.
- [67] L. Xu, Y. Zhao, D. Zhou, Z. Lin, S. K. Ng, and J. Feng, “Pllava: Parameter-free llava extension from images to videos for video dense captioning,” *arXiv preprint arXiv:2404.16994*, 2024.
- [68] J. Yang, R. Tan, Q. Wu, R. Zheng, B. Peng, Y. Liang, Y. Gu, M. Cai, S. Ye, J. Jang *et al.*, “Magma: A foundation model for multimodal ai agents,” in *Proceedings of the computer vision and pattern recognition conference*, 2025, pp. 14 203–14 214.
- [69] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, “Llava-onevision: Easy visual task transfer,” *arXiv preprint arXiv:2408.03326*, 2024.
- [70] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou, “plug-owl3: Towards long image-sequence understanding in multi-modal large language models,” *arXiv preprint arXiv:2408.04840*, 2024.
- [71] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, “Mybench: A comprehensive multi-modal video understanding benchmark,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 195–22 206.

BIOGRAPHY SECTION



Yibo Wang is currently pursuing the M.S. degree at the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou, China. He received the B.S. degree from Dalian University of Technology, Dalian, China. His research interests include multimodal large language models, affective computing, and reinforcement learning.



Rui Yang is currently pursuing the Ph.D. degree with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University. His research interests include biometric authentication, affective computing, large language models (LLMs), and embodied intelligence.



including IEEE TIP/TNNLS/TITS/IJCAI/AAAI. He served as a reviewer at some major journals and conferences like IEEE TPAMI, ICML, NIPS, ICLR, IEEE TIP, CVPR, IJCAI, ACM MM, AAAI, IEEE TMM, IEEE TCSVT, ACM TOMM.



Yitao Wu received the B.S. degree in information system and information management from Hainan University. His research interests include large language models, vision-language-action models, and vision-language models.



in peer-reviewed journals, conferences, and book chapters. His research areas include bioinformation processing and ubiquitous affective computing.



of Technical Committee on Computational Psychophysiology, IEEE SMC. He has published over 300 papers in domestic and international academic journals and conferences. His research interests include pervasive computing, computational psychophysiology, data modeling, and artificial intelligence.



Tat-Seng Chua received the Ph.D. degree from the University of Leeds, U.K. He is the KITHCT chair professor with the School of Computing, National University of Singapore, where he was the acting and founding dean of the School from 1998 to 2000. He is the co-director of NExT, a joint center between NUS and Tsinghua University, to develop technologies for live social media search. He is the 2015 winner of the prestigious ACM SIGMM Award. He is the chair of Steering Committee of the ACM International Conference on Multimedia Retrieval (ICMR) and Multimedia Modeling (MMM) conference series. He is also the general co-chair of ACM Multimedia 2005, ACM CIVR (now ACM ICMR) 2005, ACM SIGIR 2008, and ACM Web Science 2015. He serves on the editorial boards of four international journals. He is the co-founder of two technology startups in Singapore and a Fellow of the Singapore Academy of Sciences, with 107,618 citations on Google Scholar.