Benchmarking Dynamic Affective Reasoning: A Viewer-Centric Video Emotion Dataset

Zhiyan Zhang[1], Peipei Song[1]*, Jinpeng Hu[2], Jingyang Jia[1], Xun Yang[1], and Xiaojun Chang[1]

[1] University of Science and Technology of China, Hefei, China
zzyhang02@gmail.com, beta.songpp@gmail.com, jjygood@mail.ustc.edu.cn, xyang21@ustc.edu.cn, xjchang@ustc.edu.cn
[2] Hefei University of Technology, Hefei, China
135858hjp@gmail.com

Abstract. Video emotion analysis is typically framed as a static classification problem, treating each clip as an independent labeled unit. However, such a formulation overlooks a key psychological fact: emotions change as a result of cumulative reactions to consecutive causal events. To bridge this gap, we introduce **DAR** (Dynamic Affective Reasoning), the first large-scale benchmark for viewer-centric affect transitions and causal reasoning over consecutive video events. DAR contains 15,087 videos and 36,908 event-aligned affective segments annotated with 27 emotion categories. Unlike existing video-based emotion datasets, DAR presents a new viewer-centric perspective on fine-grained emotional expressions and transitions, and provides dense, temporally grounded, and causally explicit reasoning chains. Based on DAR, we formally define three challenging tasks: affective segmentation, fine-grained emotion classification, and affective reasoning. Complementing this benchmark, we propose **DAR-R1**, a two-stage framework that combines supervised fine-tuning with Group Relative Policy Optimization. Experiments across 10+ MLLMs show that DAR-R1 sets a new state-of-the-art for dynamic affective reasoning, in terms of both emotional localization and affective reasoning. Project page: https://github.com/Zhang-Zhiyan/DAR.

Keywords: Dynamic Affective Reasoning · Video Emotion Analysis · Large-scale Dataset · Reinforcement Learning

1 Introduction

Affective computing [31] aims to equip machines with the ability to perceive, interpret, and reason about human affect, enabling applications from human–computer interaction [13] to psychological counseling and psychotherapy [27, 43, 44, 48]. As AI systems become embedded in everyday products [45], robust emotion understanding is increasingly essential for natural, adaptive, and socially aware interaction [37]. Recent Multimodal Large Language Models (MLLMs) [1,

* Corresponding author.

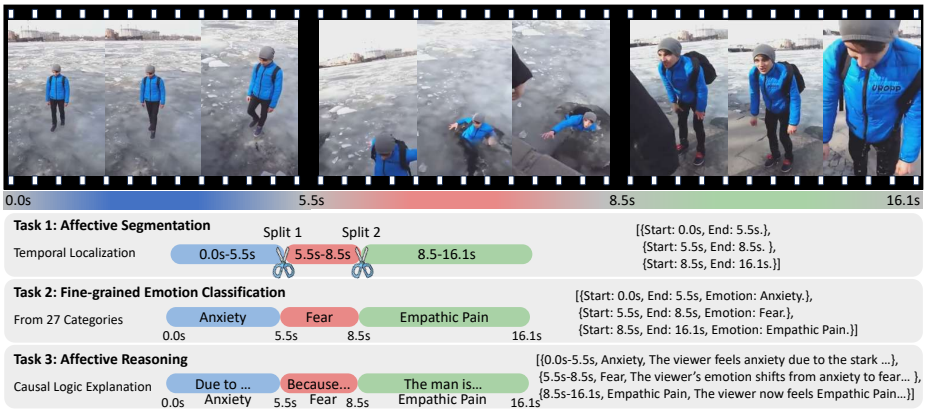


Fig. 1: Overview of the DAR Benchmark Tasks. We formulate three hierarchical tasks: (1) *Affective Segmentation* for locating temporal boundaries of emotion shifts; (2) *Fine-grained Emotion Classification* utilizing a 27-category viewer-centric taxonomy; and (3) *Affective Reasoning* for generating causal explanations grounded in visual evidence.

9, 46, 47] have further advanced video-based emotion understanding and recognition [10, 16]. Foundational benchmarks such as DFEW [19] and VCE [29] enabled systematic evaluation of video emotion understanding, while Social-IQ [50] extended evaluation to broader social and contextual cues. More recently, MER [10, 24] incorporated evidence-based reasoning into emotion recognition.

However, key aspects of affect remain underexplored in current benchmarks. Prior works either emphasize describing visual content [25, 28, 36, 54] or improve character-centric emotion recognition [23, 52], but rarely model how consecutive events drive moment-to-moment shifts in a viewer's emotional state. Furthermore, most datasets adopt coarse taxonomies following Ekman's theory [12], which are often insufficient to represent the diversity of real-world emotion reactions. As a result, video emotion analysis is still largely treated as assigning a single label to an entire clip [17, 24] or focusing primarily on character emotions [23], without explicit boundaries for emotion shifts or explanations of why they occur. These omissions hinder the deployment of emotionally intelligent agents, which are expected to respond continuously and causally to users. Therefore, we advocate reframing the video emotion analysis problem as fine-grained dynamic affective reasoning.

To support this paradigm shift, we draw inspiration from the *Affective Events Theory* (AET) [40], which views emotions as reactions to consecutive events that accumulate over time: a viewer's current affect is shaped by preceding stimuli. Guided by AET, we introduce **DAR**, a large-scale viewer-centric video emotion benchmark, designed to support multi-level emotion cognitive tasks. As illustrated in Figure 1, we define three challenging tasks based on DAR: (1) Affective Segmentation, locating temporal boundaries of emotional shifts; (2) Fine-grained Emotion Classification, identifying specific emotional states from 27 categories;

and (3) Affective Reasoning, explaining the causal logic behind each emotion. To scale annotation while preserving temporal precision, we adopt a coarse-to-fine data construction pipeline that couples high-level semantic boundary proposals with low-level visual cut refinement.

Extensive evaluations on 10+ mainstream MLLMs reveal that neither general-purpose models [28, 46] nor existing emotion-specialized models [23, 52] satisfy the requirements of dynamic affective reasoning. Generalist models often struggle with strict temporal schemas and boundary localization, whereas emotion-focused models typically lack the generative capacity to produce temporally grounded, viewer-centric explanations. To provide a strong baseline, we propose **DAR-R1**, a two-stage tuning pipeline consisting of cold-start supervised training for structural adaptation followed by GRPO-based reinforcement learning [14]. By explicitly rewarding boundary precision and reasoning quality, DAR-R1 better aligns MLLMs with fine-grained affect dynamics, yielding consistent gains over state-of-the-art baselines on DAR.

In summary, our contributions are fourfold:

1. ***New Task:*** We introduce a task of *Dynamic Affective Reasoning*, reframing video emotion analysis from static classification to temporally grounded segmentation and causal explanation.
2. ***Large-scale Dataset:*** We present DAR, the first large-scale, viewer-centric emotion dataset based on Affective Events Theory, with 15,087 videos, event-aligned segments, and verified causal reasoning chains.
3. ***Construction Method:*** We develop a scalable data construction pipeline that combines coarse-to-fine event segmentation, incremental differential captioning, and multi-dimensional quality verification to produce dense and reliable annotations.
4. ***Strong Baseline:*** We propose DAR-R1, a two-stage tuning approach that leverages reinforcement learning to improve temporal boundary localization and reasoning quality, establishing a strong baseline for this benchmark.

2 Related Work

2.1 Video Emotion Analysis

Traditional Video Emotion Analysis [35, 49] is often formulated as static emotion recognition on short clips or constrained scenarios [19], where a video segment is assigned a single categorical label [7] or emotions are inferred from affective facial cues [20]. While these benchmarks have driven progress in classification, they typically provide coarse temporal supervision and rarely evaluate whether a model can explain *why* an emotion arises or *when* it shifts. With the emergence of MLLMs, recent datasets and benchmarks have pushed VEA toward richer, language-based evaluation: AffectGPT [23] promotes descriptive emotion understanding at scale, and MME-Emotion [51] benchmarks both recognition and emotion-related reasoning. Despite these advances, fine-grained temporal localization and event-level causal attribution remain under-explored in standard VEA settings.

Table 1: Comparison of Video Emotion Datasets. Descriptive datasets allow for more diverse labels, facilitating the modeling of fine-grained dynamic emotions.

Dataset	Video	Data	Classes	Description	Reason	Timestamp	Dynamic Emotion
DFEW [19]	16,372	16,372	7	✗	✗	✗	✗
CREMA-D [7]	7,442	7,442	6	✗	✗	✗	✗
FERV39k [39]	38,935	38,935	7	✗	✗	✗	✗
MER2023 [24]	5,030	5,030	6	✗	✗	✗	✗
VCE [29]	61,046	61,046	27	✗	✗	✗	✗
MELD [32]	13,708	13,708	7	✗	✗	✓	✗
MERR [10]	28,618	28,618	113	✓	✗	✗	✗
MER-Caption+ [23]	31,327	31,327	1,972	✓	✗	✗	✗
DAR (Ours)	15,087	36,908	27	✓	✓	✓	✓

2.2 viewer-Centric Affective Reasoning

Beyond character-centric recognition, viewer-centric affect modeling studies emotions *induced in the observer* by visual stimuli. For images, ArtEmis [3] collects large-scale emotion attributions with free-form explanations for artworks, ArtEmis v2.0 [30] reduces emotional bias via contrastive data collection, and Affection [2] scales affective explanations to real-world images. For videos, VCE provides viewer-centric emotion annotations, highlighting the subjectivity of audience reactions, while StimuVAR [15] formulates video affective reasoning by predicting viewer emotions and generating rationales with stimuli-aware modeling. However, these resources often operate at the image level or provide video-level affect without dense, event-aligned boundaries, making it difficult to benchmark fine-grained affect transitions. In contrast, DAR provides event-aligned, temporally grounded annotations with explicit causal supervision for dynamic affective reasoning.

2.3 Multimodal Large Language Models

MLLMs connect visual encoders with LLM backbones and are adapted via instruction tuning [41], while video-oriented MLLMs extend the interface to temporal inputs [25, 28, 36]. Post-training methods, including reinforcement learning [34] and group-based optimization [14], have been explored to improve structured outputs and reasoning quality. Importantly, this paper does not focus on designing new policy optimization algorithms; instead, we emphasize crafting task-specific reward signals for fine-grained emotion recognition and dynamic affective reasoning.

3 The DAR Dataset

Table 1 summarizes representative video emotion datasets, most of which frame emotion recognition as static classification over short clips, overlooking the temporal dynamics and causal nature of affective experience. To bridge this gap, we introduce DAR, a large-scale, viewer-centric dataset designed for dynamic

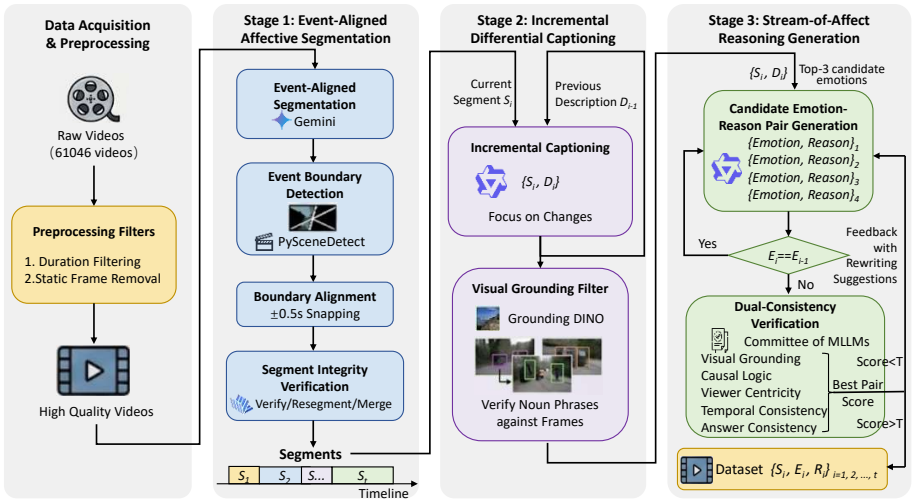


Fig. 2: Data Construction Pipeline of DAR. The process consists of three stages: (1) Coarse-to-fine event segmentation combining semantic proposals with visual cuts; (2) Incremental differential captioning to reduce redundancy; and (3) Stream-of-affect reasoning generation grounded in AET. A dual-consistency verification protocol ensures annotation quality.

affective reasoning. In this section, we first outline the theoretical foundation grounding our construction philosophy. We then detail our construction pipeline, quality assurance protocols, and dataset statistics.

3.1 Theoretical Foundation: Affective Events Theory

Our dataset construction is grounded in the *Affective Events Theory* (AET) proposed by Weiss and Cropanzano [40]. AET posits that human emotions are not static states, but rather dynamic reactions triggered by consecutive, precipitating events—such as sudden visual stimuli or narrative twists. Furthermore, these affective reactions are dynamic and cumulative, meaning current emotions are influenced by the history of preceding events. Based on this theory, we construct our dataset by segmenting videos based on semantic and visual event boundaries rather than arbitrary time intervals. This ensures that each segment represents a coherent *stimulus unit* that triggers a specific *affective reaction*, providing a solid psychological basis for dynamic reasoning.

3.2 Data Acquisition and Preprocessing

We source our raw videos from the Video Cognitive Empathy (VCE) dataset [29], which provides a rich taxonomy of 27 fine-grained emotion categories suitable for capturing nuanced viewer reactions.

Duration Filtering. We discard videos shorter than 1 second, as they lack sufficient temporal context to exhibit dynamic emotional changes.

Static Frame Removal. We compute the perceptual hash and optical flow between frames. Videos exhibiting static imagery, such as slideshows or single images, are removed as they do not possess the temporal motion cues essential for analyzing event-driven emotional shifts.

3.3 Construction Pipeline

Our construction pipeline, as illustrated in Figure 2, transforms raw videos into temporally grounded, causal affective reasoning chains through a multi-stage process.

Stage 1: Event-Aligned Affective Segmentation. To implement the AET-driven segmentation, we employ a coarse-to-fine strategy combining semantic understanding with visual boundary detection. First, we prompt Gemini-2.5-Pro [11] to propose high-level semantic event boundaries, dividing the video according to narrative shifts or affective turning points. Since LLM-predicted timestamps often suffer from drift, we utilize PySceneDetect [8] to detect precise hard cuts and fade transitions. We apply boundary snapping: semantic boundaries within a 0.5 s window of a detected visual cut are shifted to the corresponding cut frame. Finally, for segments that do not align with hard cuts, we employ InternVL3.5 [38] to verify segment integrity, recursively re-segmenting or merging those that fail to represent a complete event.

Stage 2: Incremental Differential Captioning. Conventional video captioning often yields redundant information when applied to streaming segments. Let S_i denote the i-th segment in a video, and let D_i denote its visual description. For the first segment, D_0 is set to an empty context. We introduce an *Incremental Differential Captioning* strategy in which Qwen3-VL [4] is conditioned on the visual tokens of S_i and the preceding description D_{i-1}. The model is explicitly instructed to focus on the changes in visual content, actions, and atmosphere. To ensure visual faithfulness, we use Grounding DINO [26] as a visual grounding filter to verify whether salient entities mentioned in each description can be grounded in the associated video frames.

Stage 3: Stream-of-Affect Reasoning Generation. This stage generates the core reasoning chain linking stimuli to emotions. We adopt a "Bottom-up" reasoning logic: *Visual Evidence* $\rightarrow$ *Contextual Appraisal* $\rightarrow$ *Emotion Trigger*. Let E_i denote the viewer-centric emotion label associated with segment S_i. For the first segment, E_0 is set to an empty context. We construct a structured prompt containing the historical context $\{S_{i-1}, D_{i-1}, E_{i-1}\}$, the current stimulus $\{S_i, D_i\}$, and the Top-3 candidate emotions drawn from VCE, and provide it to Qwen3-VL. The model generates four candidate `emotion`–`reason` pairs. Following AET principles, if the predicted emotion for segment S_i is identical to that of S_{i-1}, we merge the two segments into a coherent emotional phase and regenerate the rationale to reflect the sustained affective state.

3.4 Quality Assurance: Dual-Consistency Verification

To ensure the quality of our model-generated data, we implement a Dual-Consistency Verification Protocol. We employ InternVL3.5 and Qwen3-omni [42] as two independent MLLM judges. Instead of binary scoring, the generated `emotion` and `reason` pairs are evaluated across five distinct dimensions:

1. **Visual Grounding.** This metric verifies whether the visual cues cited in the reasoning, such as specific objects or actions, are physically present in the corresponding video frames.
2. **Causal Logic.** This dimension assesses whether the described event provides a sufficient and logical premise to trigger the predicted emotion.
3. **Viewer Centricity.** This criterion ensures the reasoning focuses on the viewer's affective reaction rather than merely describing the character's facial expressions or internal state.
4. **Temporal Consistency.** This metric checks if the reasoning coheres with the historical context of previous segments, which ensures a continuous narrative flow.
5. **Answer Consistency.** This verifies whether the final emotion label logically follows from the arguments presented in the reasoning text.

The committee assigns a comprehensive score based on these dimensions. Low-scoring samples are not simply discarded; they are fed back into the generation pipeline with specific feedback for rewriting. Only samples that pass a strict quality threshold after verification are included in the final dataset.

3.5 Dataset Statistics

We provide a comprehensive statistical analysis of DAR to highlight its scale, diversity, and complexity. The dataset contains 15,087 videos and 36,908 affective segments, with 13,646 videos in the training set and 1,441 videos in the test set.
Temporal Granularity and Dynamics. Unlike previous datasets that treat videos as static, DAR emphasizes fine-grained temporal dynamics. As shown in Figure 3(a), the average video duration is 14.3s, while the average segment duration is 5.8s, indicating that a typical video contains multiple distinct emotional phases. This is further corroborated by Figure 3(b), which illustrates the relationship between video length and segment count, reinforcing the idea that emotions can evolve quickly within brief video durations. Furthermore, the emotion transition matrix in Figure 3(c) reveals structured and non-random affect dynamics, with the most prominent transition being *Craving* $\rightarrow$ *Satisfaction* (0.47) and another frequent shift from *Anxiety* $\rightarrow$ *Fear* (0.30), indicating that our annotations capture coherent, event-driven emotional progressions rather than arbitrary fluctuations.
Emotion Distribution. Figure 3(d) presents the distribution of the 27 emotion categories. While the distribution follows a natural long-tail pattern typically observed in real-world data, the top-17 emotions account for approximately 80%

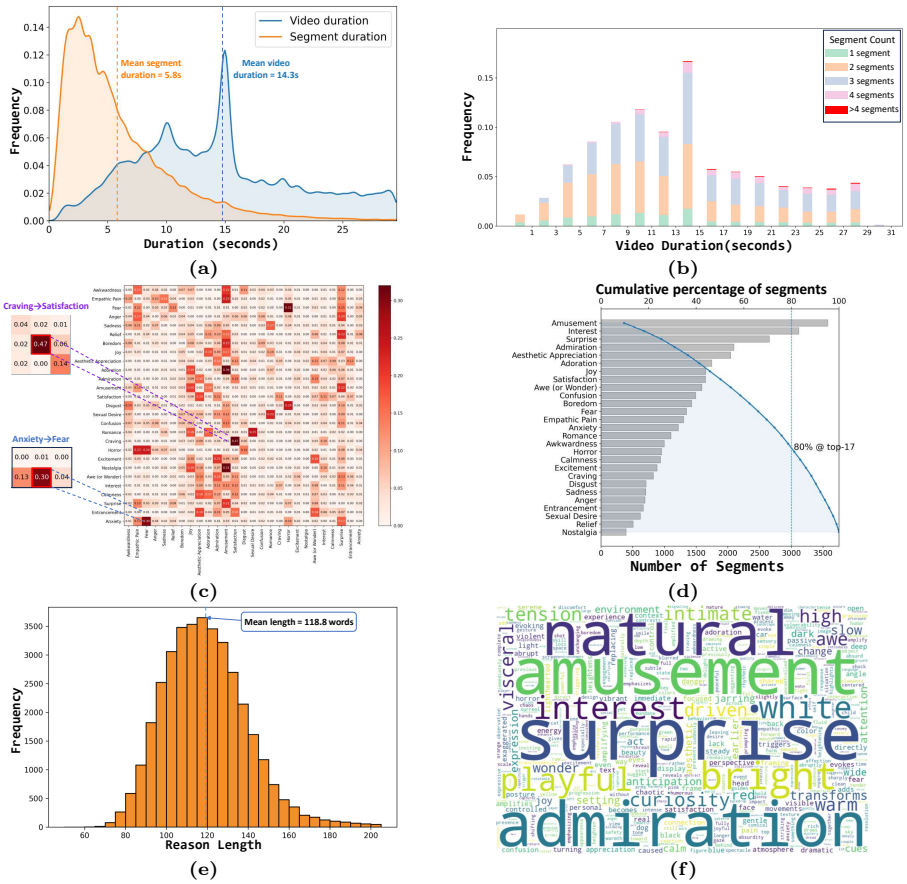


Fig. 3: Statistical Analysis of the DAR Dataset. (a) Video and segment duration distributions. (b) Segment counts by video duration. (c) Emotion transition matrix. (d) Distribution of 27 emotion categories. (e) Reasoning-text lengths. (f) High-frequency words in reasoning annotations.

of the occurrences. This indicates a balanced emotional coverage within the dataset, where emotions such as Amusement, Interest, and Surprise dominate, while complex emotions like Admiration and Aesthetic Appreciation are well-represented. This demonstrates that our dataset not only captures basic Ekman emotions but also includes more nuanced emotional and cognitive appraisals.

Reasoning Complexity. The average reasoning text length, as shown in Figure 3(e), is 118.8 words, which is significantly longer than standard video captions. This suggests that our annotations provide rich contextual information rather than merely summarizing the content. The word cloud in Figure 3(f) shows the high-frequency affect-related terms in the reasoning annotations after filtering generic nouns and connective words. The remaining terms cover emotion words, appraisal-related expressions, and visual descriptors.

4 Methodology

To address the challenges of dynamic affective reasoning—specifically, the need for precise temporal localization, accurate emotion classification, and grounded causal reasoning—we propose DAR-R1. Our framework adapts an MLLM to this task via a two-stage training paradigm: cold-start supervised fine-tuning (SFT) for structural adaptation and reinforcement learning (RL) for reasoning refinement.

4.1 Cold Start Training

The primary objective of this stage is to adapt the generalist MLLM to the complex output schema required for dynamic reasoning and to initialize a policy prior for the subsequent RL stage.

Given visual inputs $\mathcal{V}$ and task instructions $\mathcal{I}_{\text{task}}$, the model is trained to generate the ground-truth target sequence $\mathcal{S}_{gt} = (S_1^*, \ldots, S_N^*)$, where each segment $S_i^* = (t_{i,\text{start}}^*, t_{i,\text{end}}^*, e_i^*, r_i^*)$ comprises the start and end timestamps, emotion label, and causal rationale. We optimize the parameters θ by minimizing the standard negative log-likelihood over the ground-truth tokens.

4.2 Reinforcement Learning via GRPO

While SFT ensures format compliance, it often struggles to capture precise event boundaries or generate deeply grounded reasoning. To this end, we employ Group Relative Policy Optimization [14] (GRPO). Unlike Proximal Policy Optimization [34] (PPO) which requires a separate critic model, GRPO estimates the baseline from a group of outputs generated by the current policy itself, reducing computational overhead. For each input query q, we sample a group of G outputs $\{o_i\}_{i=1}^{G}$ from the behavior policy $\pi_{\theta_{\text{old}}}$. The objective function aims to maximize the advantage of high-reward outputs while constraining the policy shift via KL divergence:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^{G} \mathcal{G}_i(\theta) \right], \tag{1}$$

$$\mathcal{G}_i(\theta) = \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big(\rho_{i,t}(\theta)\, \hat{A}_i,\ \text{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\, \hat{A}_i \Big) - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}), \tag{2}$$

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \quad \hat{A}_i = \frac{R_i - \text{mean}(R_{group})}{\text{std}(R_{group})}. \tag{3}$$

We design a multi-dimensional reward function to guide the model across structure, localization, accuracy, and reasoning.

Structural Constraints Reward (R_{struct}). This component ensures the output adheres to the strict JSON schema, maintains temporal validity, and prevents runaway generation. It aggregates three sub-rewards: format compliance (r_{fmt}), minimum segment duration (r_{dur}), and total length penalty (r_{len}). Let o denote the generated output text, and $\mathcal{S}_{pred} = (\hat{S}_1, \ldots, \hat{S}_N)$ be the set of parsed segments, where each segment $\hat{S}_i$ contains a start time $\hat{t}_{i,\text{start}}$ and end time $\hat{t}_{i,\text{end}}$. The structural reward is defined as:

$$R_{struct} = \lambda_1 \cdot \mathbb{I}(o \in \mathcal{J}) + \lambda_2 \cdot \prod_{\hat{S}_i \in \mathcal{S}_{pred}} \mathbb{I}(\hat{t}_{i,\text{end}} - \hat{t}_{i,\text{start}} \geq \tau_{min}) + \lambda_3 \cdot \psi_{len}(|o|), \quad (4)$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $\mathcal{J}$ represents the set of valid JSON strings that conform to the required schema (e.g., valid keys, continuous timestamps). We set $\lambda_1 = 0.5$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.3$. The second term penalizes unrealistically short segments ($\tau_{min} = 0.4s$) to avoid fragmentation. The third term $\psi_{len}(|o|)$ applies a soft penalty on the total token count $|o|$ to discourage excessively long or runaway generations, defined as $\psi_{len}(n) = \exp(-\max(0, n - L_{soft})/\tau_{len})$ if $n < L_{hard}$, and 0 otherwise. We use $L_{soft} = 1200$, $L_{hard} = 2400$, and $\tau_{len} = 400$.

Segment Count Reward (R_{count}). To prevent over-segmentation or under-segmentation, we encourage the model to generate a number of segments consistent with the ground truth. Let $N_{pred} = |\mathcal{S}_{pred}|$ and $N_{gt} = |\mathcal{S}_{gt}|$, respectively. The reward is modeled as an exponential decay function based on the count difference:

$$R_{count} = \exp\left(-\frac{|N_{pred} - N_{gt}|}{\tau_{cnt}}\right), \quad (5)$$

where $\tau_{\text{cnt}} = 1.5$ controls the sensitivity of the penalty. This formulation provides a dense reward signal, incentivizing predictions that closely approximate the ground-truth event density.

Temporal Segmentation Reward (R_{seg}). To improve boundary precision, we employ an Intersection over Union (IoU) matching mechanism combined with a boundary distance decay. For each predicted segment $\hat{S}_i \in \mathcal{S}_{pred}$, we identify the best-matching ground-truth segment $S_i^* \in \mathcal{S}_{gt}$ that maximizes the temporal IoU. The segment-level reward is then computed as:

$$r_{seg}(\hat{S}_i) = \alpha \cdot \text{IoU}(\hat{S}_i, S_i^*) + (1 - \alpha) \cdot \exp\left(-\frac{|\hat{t}_{i,\text{start}} - t^*_{i,\text{start}}| + |\hat{t}_{i,\text{end}} - t^*_{i,\text{end}}|}{\tau_{temp}}\right). \quad (6)$$

The final R_{seg} is the average of r_{seg} over all predicted segments. The first term incentivizes overlap, while the second term, governed by the boundary decay temperature $\tau_{\text{temp}} = 0.5$, encourages the start and end timestamps to snap precisely to visual event boundaries. We use $\alpha = 0.5$ to balance temporal overlap and boundary-distance accuracy.

Emotion Accuracy Reward (R_{emo}). This reward evaluates the correctness of the predicted emotion labels. We align predicted segments to ground truth based on maximum temporal overlap. Let $\hat{e}_i$ and e_i^* be the emotion labels for the

matched segments $\hat{S}_i$ and S_i^*. The reward is binary, conditioned on both label matching and sufficient temporal overlap:

$$R_{emo} = \frac{1}{N_{pred}} \sum_{\hat{S}_i \in \mathcal{S}_{\text{pred}}} \mathbb{I}(\hat{e}_i = e_i^*) \cdot \mathbb{I}(\text{IoU}(\hat{S}_i, S_i^*) > 0.5). \tag{7}$$

This ensures that the model is rewarded only when it correctly identifies the emotion for the correct temporal event.

Reasoning Quality Reward (R_{reason}). To ensure the generated rationale r_i is informative and non-repetitive, we define R_{reason} based on length and uniqueness constraints. Let $W(r_i)$ be the word count of the reasoning text for segment S_i. We define an exponential length score $\phi(l) = \exp(-|l - \mu_{len}|/\sigma_{len})$. The uniqueness is measured using the Jaccard similarity between adjacent reasoning texts r_i and r_{i-1}. The reward is computed as:

$$R_{reason} = \frac{1}{N_{pred}} \sum_{i=1}^{N_{pred}} \phi(W(\hat{r}_i)) \cdot (1 - \mathbb{I}(\text{Jaccard}(\hat{r}_i, \hat{r}_{i-1}) > \tau_{dup})), \tag{8}$$

where $\mu_{len} = 120$ encourages detailed explanations, $\sigma_{len} = 40$ controls the length-score decay, and $\tau_{dup} = 0.75$ serves as a threshold to penalize repetitive hallucination loops across consecutive segments.

The total reward R_{total} is a weighted sum of five components:

$$R_{total} = w_1 R_{struct} + w_2 R_{count} + w_3 R_{seg} + w_4 R_{emo} + w_5 R_{reason}, \tag{9}$$

where w_k represents the weight for each reward component.

5 Experiments

In this section, we validate the effectiveness of our proposed DAR-R1 framework on the DAR dataset. We first describe the experimental setup, including implementation details and evaluation metrics. We then present a comprehensive comparison against state-of-the-art Multimodal Large Language Models. Finally, we conduct ablation studies to analyze the impact of our two-stage training paradigm and provide a qualitative analysis.

5.1 Experimental Setup

Implementation Details. We utilize Qwen2.5-VL-3B as our backbone model, keeping the vision encoder frozen while fine-tuning the LLM and aligner throughout both stages. We adopt a two-phase training scheme on 4×H100 GPUs: in the cold-start stage, we run supervised fine-tuning for 0.5 epochs using AdamW with a learning rate of 1×10^{-5}; in the reinforcement learning stage, we perform GRPO training for 1 epoch with a learning rate of 2×10^{-6}. The overall reward is a weighted sum of structural constraints, segment count reward, temporal

segmentation, emotion accuracy, and reasoning quality with weights 0.10, 0.25, 0.25, 0.25, and 0.15, respectively.

Emotion Taxonomy. To capture fine-grained viewer reactions, our dataset utilizes the VCE taxonomy consisting of 27 emotion categories, including *Admiration, Amusement, Anger, Anxiety, Awe, Boredom, Calmness, Confusion, Craving, Disgust, Empathic Pain, Entrancement, Excitement, Fear, Horror, Interest, Joy, Nostalgia, Relief, Romance, Sadness, Satisfaction, Sexual Desire, Surprise, Awkwardness, Adoration, Aesthetic Appreciation.*

Evaluation Metrics. Evaluating dynamic affective reasoning requires assessing both temporal localization and semantic correctness. We adopt a holistic suite of metrics:

- **Segment Count Accuracy:** We evaluate whether the number of predicted segments correctly reflects the emotional transitions in the video. This metric ensures that the predicted number of emotional shifts aligns with the ground truth, addressing the accuracy of emotional boundary detection.
- **Segmentation Accuracy:** We calculate the Mean Temporal Intersection over Union (mIoU) between the predicted segments and ground-truth segments to evaluate temporal localization accuracy.
- **Emotion Accuracy:** A prediction is considered correct only if the predicted emotion category matches the ground truth and the segment has a temporal IoU ≥ 0.5 with the corresponding ground truth segment.
- **Reasoning Quality:** To evaluate rationale quality, we employ GPT-4o [18] as a judge. The judge rates each generated rationale on a 0–5 scale, focusing on visual grounding, causal logic, viewer centricity, temporal consistency, and answer consistency.

5.2 Main Results

Table 2 compares our models with representative state-of-the-art MLLMs on the DAR test set. The results highlight both the difficulty of Dynamic Affective Reasoning and the effectiveness of our two-stage training paradigm.

Limitations of Existing Models. Existing models struggle significantly with this benchmark, reflecting the combined challenges of strict temporal localization and causally grounded, viewer-centric explanation. Emotion-focused MLLMs like Videmo and AffectGPT perform poorly due to their static training paradigms, lacking the granularity for temporal localization. Large-scale generalist models such as Qwen3-VL-8B demonstrate stronger inherent capabilities but still fall short in temporal precision and reasoning coherence without domain-specific adaptation.

Incorporating the DAR dataset via cold-start training yields a substantial performance improvement. Compared to the base model Qwen-2.5-VL-3B, DAR-SFT achieves gains across all evaluation metrics, with particularly large gains in segment-count accuracy and emotion accuracy by 12.3% and 10.8%, respectively. This confirms that our dataset effectively bridges the gap between general multimodal understanding and dynamic affective alignment. Building upon this,

Table 2: Quantitative Comparison on the DAR Test Set. We compare our method with leading proprietary and open-source MLLMs. SC-Acc (%) denotes segment count accuracy, mIoU (%) denotes temporal segmentation overlap, Emo-Acc (%) denotes emotion prediction accuracy, and GPT-Score (0–5) denotes reasoning quality, as evaluated by GPT across five dimensions: visual grounding (VG), causal logic (CL), viewer centricity (VC), temporal consistency (TC), and answer consistency (AC), along with the average score (Avg). "DAR-SFT" is after SFT and "DAR-R1" is the final model. Best results are in **bold**.

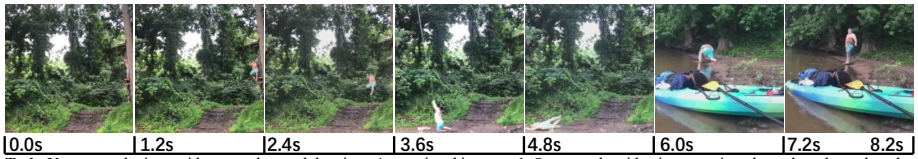
Models	SC-Acc	mIoU	Emo-Acc	GPT-Score					
				VG	CL	VC	TC	AC	Avg
Emotion Focused Multimodal Large Language Models									
Videmo [52]	7.9	9.2	4.3	0.7	0.6	0.5	0.4	0.3	0.5
EmotionLlama [10]	21.3	10.8	2.1	1.0	0.9	0.8	0.7	0.6	0.8
AffectGPT [23]	24.3	11.9	5.8	1.0	1.4	1.3	0.7	1.1	1.1
General Multimodal Large Language Models									
Video-ChatGPT [28]	18.4	20.9	6.4	2.4	1.2	0.8	1.6	1.3	1.5
PandaGPT [36]	21.3	32.7	7.1	1.9	2.4	1.2	2.2	1.5	1.8
Video-LLaVA [25]	14.6	32.6	7.0	2.1	1.9	1.0	1.8	1.7	1.7
InternVL-3.5-2B [38]	21.2	33.6	8.2	1.8	2.0	2.3	2.1	1.0	1.8
InternVL-3.5-8B [38]	26.0	39.2	11.5	2.2	2.5	2.5	2.3	1.6	2.2
Qwen2.5-VL-3B [5]	25.4	41.7	15.0	2.4	2.3	2.2	2.5	2.0	2.3
Qwen2.5-VL-7B [5]	31.9	45.2	14.9	2.8	2.7	2.6	2.6	2.4	2.6
Qwen3-VL-4B [4]	29.2	48.2	17.9	2.8	2.9	2.6	2.8	2.4	2.7
Qwen3-VL-8B [4]	26.2	43.1	17.0	2.6	3.0	2.4	2.6	2.2	2.6
DAR-SFT	37.7	47.6	25.8	3.1	3.0	2.9	2.8	3.0	3.0
DAR-R1	**41.5**	**52.3**	**28.6**	**3.1**	**3.8**	**3.2**	**3.2**	**3.1**	**3.3**

the RL stage drives further gains. DAR-R1 consistently outperforms DAR-SFT across all metrics, achieving 52.3% mIoU and a causal-logic score of 3.8, and establishes a new state of the art on our benchmark. This trajectory indicates that while SFT establishes structural compliance, the specific reward modeling in RL is crucial for refining boundary precision and enforcing logical consistency in affective reasoning.

Human Evaluation. To complement automatic metrics and validate alignment with human perception, we recruited five experts to evaluate 100 randomly sampled videos. As summarized in Table 3, existing emotion-focused models like AffectGPT struggle with the multi-task complexity. Generalist MLLMs exhibit better foundational reasoning but lack domain specificity. In contrast,

Table 3: Human Evaluation of Reasoning Quality. Average expert ratings across the five predefined dimensions.

Models	VG	CL	VC	TC	AC	Avg
AffectGPT	0.6	1.8	0.2	0.1	1.0	0.7
Qwen2.5-VL-3B	2.5	2.3	2.5	2.4	2.8	2.5
Qwen3-VL-4B	3.1	3.5	2.7	2.8	3.0	3.0
DAR-SFT	3.4	3.5	4.0	3.1	3.8	3.6
DAR-R1	**4.2**	**4.5**	**4.2**	**3.8**	**4.4**	**4.2**



Task: You are analyzing a video to understand the viewer's emotional journey. 1. Segment the video into emotion phases based on when the viewer's dominant emotion changes. 2. Predict the viewer's emotion for each segment. 3. Reason why the viewer feels that emotion. {candidate_emotions} {Sample_format}

Emotion-Llama:
[{0.0s-10.0s, **Happy**, the viewer is happy because they are walking on a path in the woods, surrounded by nature, and enjoying the peaceful atmosphere.},
{10.0s-20.0s, **Sad**, the viewer is sad because they are walking alone and feeling lonely.},
{20.0s-30.0s, **Fear**, the viewer is fearful because they are walking through a dark forest with no clear path ahead.}]

Qwen2.5-VL-3B:
[{0.0s-2.2s, **Excitement**, The person is seen jumping off a log into a river, indicating a sense of thrill and excitement.},
{2.2s-4.2s, **Relief**, The person is seen lying on the ground after jumping, suggesting a sense of relief after the jump.},
{4.2s-8.2s, **Aesthetic Appreciation**, The lush green forest and the muddy river create a visually pleasing and serene environment, evoking a sense of appreciation for nature.}]

DAR-R1:
[{0.0s-1.5s, **Interest**, The viewer feels interest because the scene is rich with visual detail and potential narrative intrigue: the dense, overgrown forest creates a sense of mystery and hidden adventure, while the person's presence on the tree branch adds a human element that invites curiosity about their purpose or destination. The natural, overcast lighting and the lush greenery, combined with the slightly blurred background, draw the eye to the foreground, heightening focus on the person and the tree, which feels like a moment of quiet tension or anticipation.},
{1.5s-2.5s, **Surprise**, The viewer feels surprise because the previously still, mysterious scene is suddenly disrupted by someone grabbing a rope and swinging down from the tree branch. The visual shift from a calm, focused composition to a dynamic, uncontrolled arc—especially the person's body sweeping through the air—creates a jarring contrast that feels unexpected and almost violent. This transition from interest to surprise is triggered by the abrupt change in action: the person's sudden, unscripted swing breaks the quiet tension of the forest, replacing curiosity with a visceral reaction to the unexpected event.},
{2.5s-8.2s, **Empathic Pain**, The viewer feels Empathic Pain because the sudden, violent impact of the jump is now followed by a clear, unambiguous physical consequence: the person lands on the ground with a visible, forceful impact that causes them to fall backward, their body contorting in pain. The visual cues—falling forward, arms flailing, and a visible, uncontrolled fall—evoke a visceral sense of physical distress, making the viewer instinctively feel the person's discomfort. This emotional shift from Surprise to Empathic Pain is triggered by the transition from a sudden, unexplained jump to a visible, painful consequence—the person's body language and the immediate aftermath of the impact create a direct, relatable sense of suffering, which the viewer can't help but feel.}]

Fig. 4: Qualitative Comparison on DAR. We visualize the temporal alignment, predicted emotion labels, and reasoning text. Green highlights indicate correct reasoning, while red indicates hallucinations or errors. DAR-R1 demonstrates superior grounding capabilities compared to baselines.

our proposed methods demonstrate superior performance. The high scores in Causal Logic and Viewer Centricity particularly verify that our RL paradigm effectively aligns the model with the nuanced, viewer-centric logic required for high-quality affective reasoning. Importantly, we observe strong agreement between GPT-based judging and expert preferences, suggesting that our automatic GPT evaluation provides a reliable proxy for human assessment on this benchmark.

5.3 Qualitative Analysis

To provide concrete insights into the capabilities of DAR-R1, we present a qualitative comparison in Figure 4. As shown in Figure 4, emotion-focused MLLM fails on this benchmark. It often produces temporally invalid outputs (e.g., timestamps exceeding the video duration) and relies on generic, template-like rationales that ignore salient high-arousal events. This behavior suggests that static or coarse-grained supervision is insufficient for temporally grounded affective

reasoning in videos. The generalist model, Qwen2.5-VL-3B, correctly perceives the "jumping" action but fails in affective alignment. It misinterprets the crash scene—where the man hits the mud—as "Relief" or "Aesthetic Appreciation" of the lush forest. It focuses on the background scenery rather than the salient event (the accident) that drives the viewer's emotion. In contrast, our DAR-R1 model demonstrates precise temporal grounding and causal logic. It successfully identifies the narrative turning point. It segments the initial buildup as "Interest" and "Surprise", and crucially, it aligns the onset of "Empathic Pain" (2.5s) with the moment the swing loses control. Unlike Qwen2.5-VL, DAR-R1 explicitly captures the visual evidence of distress, referencing specific physical cues as highlighted in green. This causal link between the visual impact and the viewer's visceral reaction confirms that our RL training effectively aligns the model with fine-grained human emotional dynamics.

6 Conclusion

In this paper, we presented a comprehensive framework for transitioning Video Emotion Analysis from static classification to Dynamic Affective Reasoning. Grounded in Affective Events Theory, we introduced DAR, the first large-scale dataset featuring event-aligned segmentation and causal reasoning chains. To benchmark this task, we proposed DAR-R1, a novel training pipeline that leverages Reinforcement Learning to align MLLMs with fine-grained emotional dynamics. Extensive experiments demonstrate that our method establishes a new state-of-the-art, significantly outperforming existing models in temporal localization, emotion recognition, and reasoning quality. We hope this work serves as a foundation for advancing affective computing by enabling models to predict *when* emotions shift, *what* they are, and *why* they arise from temporally grounded evidence.

Acknowledgements

This work was supported by the National Science and Technology Major Project (Grant No. 2025ZD0215301), the National Natural Science Foundation of China (Grant Nos. 62402471, U22A2094, and U25A20530), the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project (Grant No. 2025CSJZN01600), the New Generation Artificial Intelligence National Science and Technology Major Project (Grant No. 2025ZD0123100), and the Key Science and Technology Project of Anhui Province (Grant No. 202523j08050001). This research was also supported by the advanced computing resources provided by the Supercomputing Center of the University of Science and Technology of China.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Achlioptas, P., Ovsjanikov, M., Guibas, L.J., Tulyakov, S.: Affection: Learning affective explanations for real-world visual data. In: CVPR. pp. 6641–6651 (2023)
3. Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L.J.: Artemis: Affective language for visual art. In: CVPR. pp. 11569–11579 (2021)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
6. Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., Wermter, S.: The omg-emotion behavior dataset. arXiv preprint arXiv:1803.05434 (2018)
7. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing **5**(4), 377–390 (2014)
8. Castellano, B.: PySceneDetect: Video scene cut and transition detection. `https://github.com/Breakthrough/PySceneDetect` (2025)
9. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
10. Cheng, Z., Cheng, Z.Q., He, J.Y., Wang, K., Lin, Y., Lian, Z., Peng, X., Hauptmann, A.: Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. Advances in Neural Information Processing Systems **37**, 110805–110853 (2024)
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
12. Ekman, P.: An argument for basic emotions. Cognition emotion **6**(3-4), 169–200 (1992)
13. Filippini, C., Perpetuini, D., Cardone, D., Chiarelli, A.M., Merla, A.: Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review. Applied Sciences **10**(8), 2924 (2020)
14. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
15. Guo, Y., Siddiqui, F., Zhao, Y., Chellappa, R., Lo, S.Y.: Stimuvar: Spatiotemporal stimuli-aware video affective reasoning with multimodal large language models. International Journal of Computer Vision **133**(9), 6456–6472 (2025)
16. Han, Z., Zhu, B., Xu, Y., Song, P., Yang, X.: Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 5528–5537 (2025)
17. Hu, J., Shi, H., Dai, C., Li, Z., Song, P., Wang, M.: Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark. In:

Proceedings of the 33rd ACM International Conference on Multimedia. pp. 5814–5823 (2025)
18. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
19. Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J.: Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2881–2889 (2020)
20. Kollias, D., Zafeiriou, S.: Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770 (2018)
21. Kossaifi, J., Tzimiropoulos, G., Todorovic, S., Pantic, M.: Afew-va database for valence and arousal estimation in-the-wild. Image and Vision Computing **65**, 23–36 (2017)
22. Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B., et al.: Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. IEEE transactions on pattern analysis and machine intelligence **43**(3), 1022–1040 (2019)
23. Lian, Z., Chen, H., Chen, L., Sun, H., Sun, L., Ren, Y., Cheng, Z., Liu, B., Liu, R., Peng, X., et al.: Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. International Conference on Machine Learning (2025)
24. Lian, Z., Sun, H., Sun, L., Chen, K., Xu, M., Wang, K., Xu, K., He, Y., Li, Y., Zhao, J., et al.: Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 9610–9614 (2023)
25. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
26. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
27. Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., Huang, M.: Towards emotional support dialog systems. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers). pp. 3469–3483 (2021)
28. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
29. Mazeika, M., Tang, E., Zou, A., Basart, S., Chan, J.S., Song, D., Forsyth, D., Steinhardt, J., Hendrycks, D.: How would the viewer feel? estimating wellbeing from video scenarios. Advances in Neural Information Processing Systems **35**, 18571–18585 (2022)
30. Mohamed, Y., Khan, F.F., Haydarov, K., Elhoseiny, M.: It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In: CVPR. pp. 21231–21240 (2022)

31. Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective multimodal human-computer interaction. In: Proceedings of the 13th annual ACM international conference on Multimedia. pp. 669–676 (2005)
32. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Conference of the Association for Computational Linguistics. pp. 527–536 (2019)
33. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). pp. 1–8. IEEE (2013)
34. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
35. Song, P., Guo, D., Yang, X., Tang, S., Wang, M.: Emotional video captioning with vision-based emotion interpretation network. IEEE Transactions on Image Processing **33**, 1122–1135 (2024)
36. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. In: Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants! pp. 11–23 (2023)
37. Wang, S., Yuan, F., Wang, K., Yang, X., Zhang, X., Wang, M.: Dual-state personalized knowledge tracing with emotional incorporation. IEEE Transactions on Knowledge and Data Engineering (2025)
38. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
39. Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., Ge, W., Zhang, W.: 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. pp. 20922–20931 (2022)
40. Weiss, H.M., Cropanzano, R.: Affective events theory. Research in organizational behavior **18**(1), 1–74 (1996)
41. Xie, H., Peng, C.J., Tseng, Y.W., Chen, H.J., Hsu, C.F., Shuai, H.H., Cheng, W.H.: Emovit: Revolutionizing emotion insights with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26596–26605 (2024)
42. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025)
43. Xu, Y., Hu, J., Zhao, Z., Duan, Z., Sun, X., Yang, X.: Multiagentesc: A llm-based multi-agent collaboration framework for emotional support conversation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 4665–4681 (2025)
44. Xu, Y., Zhao, Z., Sun, X., Yang, X.: Prompt learning with multiperspective cues for emotional support conversation systems. IEEE Transactions on Computational Social Systems (2025)
45. Yadegaridehkordi, E., Noor, N.F.B.M., Ayub, M.N.B., Affal, H.B., Hussin, N.B.: Affective computing in education: A systematic review and future research. Computers & education **142**, 103649 (2019)
46. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)

47. Yang, X., Pan, H., Wang, X., Cao, Y., Li, J., Wang, M.: Fine: Fine-grained neuron-level model editing for reliable and safe llms. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
48. Yannakakis, G.N.: Enhancing health care via affective computing (2018)
49. Ye, C., Chen, W., Song, P., Liu, X., Zhang, L., Mao, Z.: Multi-round mutual emotion-cause pair extraction for emotion-attributed video captioning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3320–3329 (2025)
50. Zadeh, A., Chan, M., Liang, P.P., Tong, E., Morency, L.P.: Social-iq: A question answering benchmark for artificial social intelligence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8807–8817 (2019)
51. Zhang, F., Cheng, Z., Deng, C., Li, H., Lian, Z., Chen, Q., Liu, H., Wang, W., Zhang, Y.F., Zhang, R., et al.: Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. International Conference on Learning Representations (2025)
52. Zhang, Z., Wang, W., Zhu, Y., Qin, W., Wan, P., Zhang, D., Yang, J.: Videmo: Affective-tree reasoning for emotion-centric video foundation models. Advances in Neural Information Processing Systems (2025)
53. Zhang, Z., Yang, J.: Temporal sentiment localization: Listen and look in untrimmed videos. In: ACM MM (2022)
54. Zhang, Z., Song, P., Hu, J., Chen, W., Ni, L., Yang, X.: Stimuli-aware emotion adaptor for enhancing llm in affective explanation captioning. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10662–10666. IEEE (2026)