AFFECTREVEAL: EVENT-GROUNDED EMOTION RECOGNITION BEYOND VISUAL APPEARANCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual emotion recognition commonly assumes that all evidence required for prediction is contained in the observed image or video. Yet the same visible reaction can convey different emotions depending on events beyond the input: tears, for example, may indicate grief or joy. We formulate **Event-Grounded Emotion Recognition (EGER)**, where emotion recognition requires recovering the affect-determining event. We construct **EGER-Bench**, comprising 10,052 videos and 10,734 images across 11 emotions, two source domains, and four visual settings. A study with six annotators shows that event context raises human recognition accuracy from 33.96% to 72.08%, confirming that visual evidence alone is often insufficient. Semantic relevance alone does not solve EGER: a plausible event may imply the wrong emotion if its identity, focal-person role, relationship, or outcome is misinterpreted. We therefore propose **AffectReveal**, a tuning-free framework that first constructs and independently verifies evidence-grounded alternatives over these affect-critical factors. It then cross-checks the recovered event against face-masked in-media facts through bidirectional atomic evidence support, while retaining the original unmasked input for final prediction. Across three downstream models and four input settings, AffectReveal yields average UAR gains of 5.26–10.53 points. For three fine-tunable models, it also enables untuned models to outperform their fine-tuned visual-only counterparts in all 12 accuracy comparisons, without updating downstream parameters.

1 INTRODUCTION

Visual emotion recognition infers a person's emotional state from facial expressions, body movements, interactions, and scene context (Mollahosseini et al., 2017; Kosti et al., 2017; Lee et al., 2019). Despite substantial progress, it is still predominantly formulated as *closed-input recognition*, assuming that the observed image or video contains all evidence required for emotion prediction.

This assumption does not always hold. The same tears may arise from grief, relief, or victory, while the same smile may express happiness, embarrassment, or concealed distress. A visible reaction can therefore imply different emotions depending on what happened and how the focal person was involved (Xia & Ding, 2019; Poria et al., 2021). As illustrated in Figure 1, a crying athlete may appear sad from visual appearance alone, yet becomes understandable as happy once the underlying event—winning an Olympic final—is recovered.

Context-aware methods incorporate scenes, objects, actions, and social interactions (Kosti et al., 2017; 2019; Lee et al., 2019; Mittal et al., 2020), but still reason over evidence contained in the given media. They cannot resolve cases in which decisive information lies outside the observed sample, such as event identity, focal-person role, interpersonal relationship, or event outcome. We study this evidence gap through **Event-Grounded Emotion Recognition (EGER)**, where recognizing an emotion requires recovering the external event needed to interpret the observed reaction.

To enable systematic evaluation, we construct **EGER-Bench**, comprising 10,052 video clips and 10,734 images across 11 emotions, two source domains, and four visual input settings. The benchmark focuses on reactions whose short-form visual evidence is insufficient or potentially misleading. A controlled human study confirms this information gap: providing event context increases emotion-recognition accuracy from 33.96% to 72.08%.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

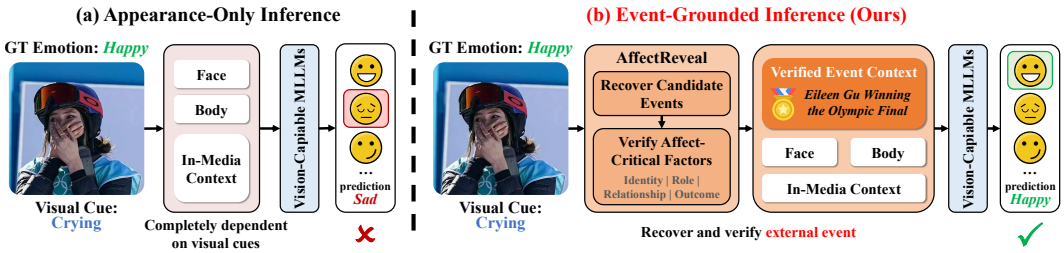

Figure 1: Appearance-based versus event-grounded emotion recognition. The visible reaction suggests sadness, whereas the recovered event context supports happiness. AffectReveal recovers and verifies the affect-determining event before prediction.

External retrieval is a natural starting point (Lewis et al., 2020; Yao et al., 2022; Asai et al., 2024), but introduces two task-specific challenges. First, a retrieved event may be topically plausible yet imply the wrong emotion if its identity, focal-person role, relationship, or outcome is misinterpreted, which we call *affect-critical event ambiguity*. Second, using a potentially misleading expression to validate the event retrieved from that expression creates *appearance-induced circularity*.

We propose **AffectReveal**, a tuning-free framework designed around these challenges. **Affect-Critical Event Recovery** constructs and independently verifies evidence-anchored alternatives over the four affect-critical factors before consolidating the external event context. **Expression-Disentangled Event Verification** then masks facial regions only in the verification branch and compares non-facial in-media facts with the recovered event through bidirectional atomic evidence support. The original unmasked input and the verified event evidence are jointly used for final prediction.

Across three downstream models and four input settings, AffectReveal improves both UAR and accuracy in all 16 direct comparisons, with average UAR gains of 5.26–10.53 points. An untuned model equipped with AffectReveal also outperforms its fine-tuned visual-only counterpart in all 12 corresponding accuracy comparisons. Component analysis further shows that non-facial context becomes useful when employed to verify the recovered event, rather than simply appended as additional text.

Our contributions are:

(1) We formulate **Event-Grounded Emotion Recognition (EGER)**, where the affect-determining event is not fully contained in the observed media and must be recovered to interpret the focal person's visual reaction;

(2) We construct **EGER-Bench**, which pairs visually insufficient or misleading short-form observations with source-grounded event annotations across two domains and four controlled input settings, enabling systematic evaluation of emotion recognition under missing event evidence;

(3) We propose **AffectReveal**, a tuning-free framework that verifies emotion-relevant event details and cross-checks the recovered event with non-facial visual evidence, improving multiple downstream models without parameter updates.

2 Related Work

Visual and Context-Aware Emotion Recognition. Visual emotion recognition has expanded from face-centered classification (Barsoum et al., 2016; Li et al., 2017; Jiang et al., 2020; Liu et al., 2022) to models incorporating body posture, scenes, objects, actions, and social interactions (Kosti et al., 2017; 2019; Lee et al., 2019; Hoang et al., 2021; Wu et al., 2022). Subsequent approaches further reduce dependence on facial appearance through masking, context deconfounding, and interaction modeling (Mittal et al., 2020; Yang et al., 2022; 2023; Li et al., 2026). However, their contextual evidence remains contained in the observed media. EGER instead addresses reactions whose affect-determining events are not fully observable in the input. Recent MLLMs and emotion-oriented variants provide strong downstream reasoners, but the EGER challenge concerns missing

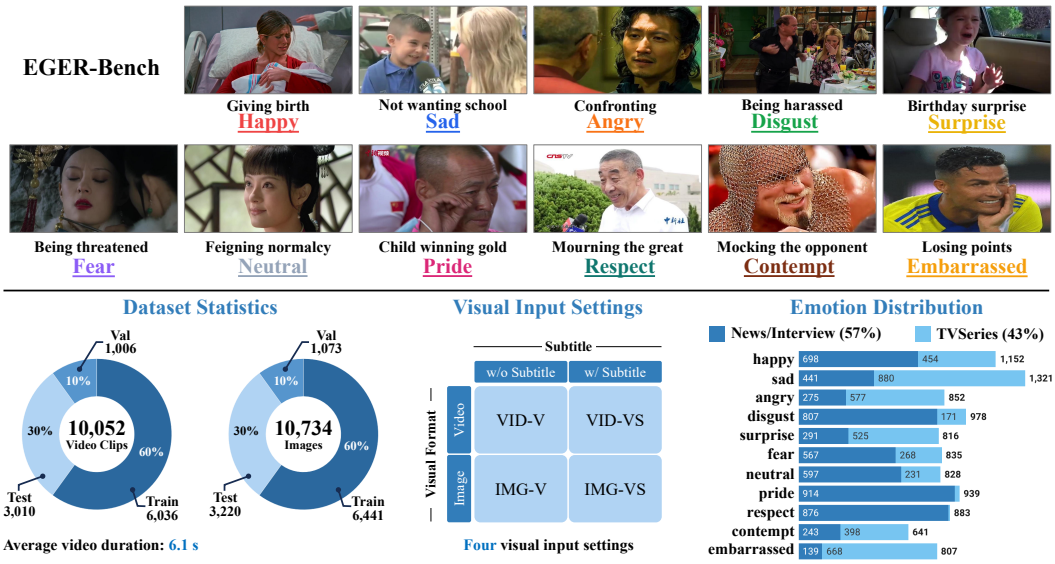

Figure 2: Overview of EGER-Bench, covering 11 emotions from news/interview and television-series sources. The lower panel summarizes dataset statistics, four visual settings, and class distributions across the two domains.

evidence rather than a particular model architecture (Liu et al., 2023; Bai et al., 2023; Chen et al., 2024; Cheng et al., 2024; Lian et al., 2025; Peng et al., 2026).

Emotion Causes and Event-Level Affect Understanding. Emotion-cause research relates affective states to the situations that elicit them (Gui et al., 2016; Xia & Ding, 2019; Rashkin et al., 2019; Zadeh et al., 2019). Existing methods typically identify causes from a provided conversation, document, or complete multimodal sequence (Poria et al., 2021; Wang et al., 2022; Li et al., 2025). In EGER, the relevant event is not given and must be recovered externally. The model must therefore establish not only which event occurred, but also which role, relationship, and outcome apply to the focal person.

Retrieval-Augmented Affect Reasoning. Retrieval-augmented and tool-using models acquire external evidence for knowledge-intensive reasoning (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2021; Yao et al., 2022; Schick et al., 2023; Qin et al., 2024), while iterative retrieval and critique improve factual support (Madaan et al., 2023; Asai et al., 2024; Gou et al., 2024). Retrieval-augmented multi-agent reasoning has also been explored for multimodal emotion recognition (Wang et al., 2026). AffectReveal targets a distinct affective failure mode: retrieved information can be topically relevant yet imply the wrong emotion. It therefore verifies evidence-grounded alternatives over affect-critical event details and cross-checks the recovered event against non-facial visual evidence.

3 Event-Grounded Emotion Recognition and EGER-Bench

3.1 Task Definition

Let x^m denote an image or video under visual setting $m \in \mathcal{M}$, and let $y \in \mathcal{Y}$ be the emotion of a designated focal person. EGER considers cases in which the emotion depends not only on the observed reaction but also on a latent affect-determining event z^*:

$$y = f(x^m, z^*), \qquad \hat{z} = \mathcal{R}(x^m; \mathcal{K}), \qquad \hat{y} = F(x^m, \hat{z}) \in \mathcal{Y}. \tag{1}$$

At inference time, z^* is unavailable. A method may instead query an external information space $\mathcal{K}$ to recover an event interpretation $\hat{z}$. Importantly, $\mathcal{K}$ is not sample-specific reference context supplied with the test input. Emotion prediction is the primary task; event recovery is evaluated separately through diagnostic analyses.

3.2 DATA COLLECTION

EGER-Bench draws from two complementary domains. **News and interview** sources provide real-world events documented by webpages and headlines, whereas **television series**, including *Friends* and *Empresses in the Palace*, provide richer relationships, intentions, and role-dependent reactions.

We use seed patterns such as "tears of joy," "putting on a brave face," and "forcing a smile" to discover reactions whose emotional meaning may differ from their appearance. For news and interview sources, we retrieve candidate videos and extract coherent, person-centered clips shorter than 10 seconds. For television series, we identify candidates from dialogue scripts, align them with episodes using subtitles, and extract clips centered on the focal person. These patterns are used only for discovery; inclusion and annotation are determined from the complete event context.

The resulting dataset contains 10,052 video clips with an average duration of 6.1 seconds. Sampling one representative frame per clip and adding 682 event-grounded news images yields 10,734 images, as shown in Figure 2.

3.3 EVENT-GROUNDED ANNOTATION

Using a dedicated interface, annotators inspect each restricted visual sample together with its source-grounded reference context: the original webpage and headline for news/interview samples, or expanded dialogue and scene context for television-series samples. They identify the focal person and retain samples for which the restricted input is insufficient or potentially misleading. Samples with low visual quality, no identifiable focal person, or inadequate event grounding are excluded.

Each retained sample is assigned one of 11 labels: *happy*, *sad*, *angry*, *disgust*, *fear*, *surprise*, *neutral*, *pride*, *embarrassed*, *contempt*, or *respect*. Labels are determined from the focal person's reaction and role in the complete event rather than facial appearance alone. The reference materials are retained for diagnostic evaluation but are not provided in the standard EGER setting.

Human validation. We conduct an independent study with six annotators on 480 randomly sampled instances. For each instance, different annotators evaluate the emotion under two evidence conditions: the restricted visual input alone, or the same input together with its reference event context. All annotators are blinded to the benchmark labels, and no annotator sees the same instance under both conditions. As shown in Table 1, event context raises accuracy from 33.96% to 72.08%, with corresponding gains in UAR and WAF. This substantial gap supports the premise that the restricted visual input alone does not reliably determine the emotion label.

Table 1: Human performance with and without reference event context on 480 instances.

Evidence Condition	UAR	Acc.	WAF
Visual Input	27.80	33.96	33.89
Visual + Event Context	**65.80**	**72.08**	**72.66**

3.4 VISUAL INPUT SETTINGS AND SPLITS

We define four settings, $\mathcal{M} = \{\text{VID-V}, \text{VID-VS}, \text{IMG-V}, \text{IMG-VS}\}$, where `VID`/`IMG` denotes video/image and `VS` retains visible subtitles or scene text, while `V` removes detected text regions. No audio or separately supplied transcript is used. PaddleOCR (Cui et al., 2025) detects text regions and ProPainter (Zhou et al., 2023) inpaints the resulting masks. Each text-retained and text-removed pair shares the same underlying frames, and AffectReveal and its downstream model always receive the same version.

The benchmark contains 57% news/interview and 43% television-series samples. We use an approximately 6:1:3 train/validation/test split, yielding 6,036/1,006/3,010 videos and 6,441/1,073/3,220 images. Representative frames inherit their source-video assignments to prevent frame-level overlap; the additional web images follow the same split ratio, with domain proportions maintained across subsets.

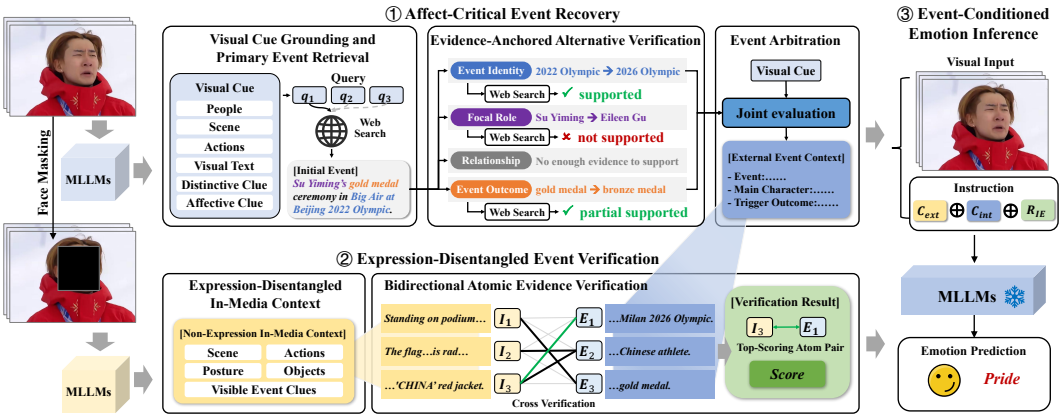

Figure 3: Overview of AffectReveal. The framework recovers external event context by retrieving a primary event and verifying evidence-anchored alternatives, then cross-checks the recovered event against non-facial in-media facts before emotion prediction.

4 AFFECTREVEAL

4.1 OVERVIEW

EGER presents two coupled challenges. First, limited visual evidence may support multiple event interpretations whose differences can change the implied emotion. We refer to this problem as *affect-critical event ambiguity*. Second, the expression that motivates event recovery may itself be misleading; using it again to validate the recovered event creates *appearance-induced circularity*. AffectReveal addresses these challenges through two corresponding stages, as illustrated in Figure 3.

First, **Affect-Critical Event Recovery** operationalizes event ambiguity through four recurrent factors: $\mathcal{D} = \{d_{\text{id}}, d_{\text{role}}, d_{\text{rel}}, d_{\text{out}}\}$, corresponding to event identity, focal-person role, interpersonal relationship, and event outcome. This factorization organizes event verification rather than providing an exhaustive definition of event context. The stage retrieves and verifies candidate interpretations to produce an external event context:

$$C_{\text{ext}} = \text{Recover}(x^m, \mathcal{K}; \mathcal{D}). \tag{2}$$

Second, **Expression-Disentangled Event Verification** masks facial regions and extracts non-facial in-media context:

$$C_{\text{int}} = \text{Extract}\left(\mathcal{M}_{\text{face}}(x^m)\right), \tag{3}$$

where $\mathcal{M}_{\text{face}}$ denotes face masking. It then evaluates the observable compatibility between the internal and external contexts:

$$R_{IE} = \text{Verify}(C_{\text{int}}, C_{\text{ext}}). \tag{4}$$

The original unmasked input, C_{ext}, C_{int}, and R_{IE} are jointly used for final emotion prediction.

4.2 AFFECT-CRITICAL EVENT RECOVERY

Visual Cue-Grounded Retrieval. To recover the external event context C_{ext}, we first extract structured clues from x^m, including people, scenes, actions, visible text, distinctive objects, and observable reactions. The strongest clues are combined into N ranked event-oriented queries, $Q = \{(q_i, p_i)\}_{i=1}^{N}$, where p_i denotes query priority. The queries and representative visual frames are used for web and image search. Retrieved evidence is checked against the visual clues and summarized as a primary event e_0, including its description, time, location, affect-critical details, and supporting sources.

Alternative Verification. A plausible e_0 may still contain an incorrect event identity, focal-person role, relationship, or outcome. For each uncertain factor $d_i \in \mathcal{D}$, we construct an alternative that changes only that factor: $e_i^{\text{alt}} = e_0[d_i \leftarrow \widetilde{d}_i]$. The alternative $\widetilde{d}_i$ must be grounded in primary retrieval evidence rather than freely generated. Each alternative is independently checked through

a factor-specific search and labeled as *supported*, *partially supported*, *not supported*, *not found*, or *conflicting*. Only supported and partially supported alternatives are retained.

Event Arbitration. The primary event, retained alternatives, visual clues, and retrieved evidence are jointly reconsidered to produce C_{ext}. It records the best-supported event identity, focal-person role, relationship, outcome, and sources without directly predicting an emotion label.

4.3 Expression-Disentangled Event Verification

Non-Facial In-Media Context. To verify whether C_{ext} is compatible with the observed sample without reusing the potentially misleading facial expression, we mask facial regions frame by frame and extract C_{int} from the masked input. C_{int} describes scenes, body actions, objects, interactions, and visible event clues. Masking is restricted to this verification branch; the original unmasked input remains available for final emotion prediction.

Bidirectional Atomic Evidence Verification. We decompose the internal and external contexts into atomic facts:

$$\mathcal{I} = \{i_j\}_{j=1}^{n}, \qquad \mathcal{E} = \{e_k\}_{k=1}^{m}, \tag{5}$$

where each statement contains one independently assessable fact. An evidence judge assigns every pair a support score $M_{jk} \in \{0, 0.5, 1.0\}$, denoting no, partial, or direct factual support.

Rather than averaging unrelated pairs, we retain the strongest counterpart for each fact and aggregate support in both directions:

$$S_{IE} = \frac{1}{2}\left(\frac{1}{n}\sum_{j=1}^{n}\max_{k} M_{jk} + \frac{1}{m}\sum_{k=1}^{m}\max_{j} M_{jk}\right). \tag{6}$$

The two terms measure how well the in-media facts are explained by the recovered event and how strongly the external event is grounded in observable evidence, respectively.

The strongest aligned pairs are retained as a textual explanation A_{IE}, giving the final verification result $R_{IE} = (S_{IE}, A_{IE})$. R_{IE} measures observable compatibility rather than requiring every external event fact to appear in the media.

4.4 Event-Conditioned Emotion Inference

The external context, in-media context, and their verification relation are serialized into a fixed-format evidence package and provided with the original unmasked input:

$$C_{\text{evi}} = C_{\text{ext}} \oplus C_{\text{int}} \oplus R_{IE}, \qquad \widehat{y} = F_\theta(x^m, C_{\text{evi}}), \qquad \widehat{y} \in \mathcal{Y} \tag{7}$$

Here, $\oplus$ denotes textual serialization. No parameter of the downstream model F_θ is updated.

5 Experiments

5.1 Experimental Setup

Benchmark and metrics. We evaluate on EGER-Bench under four visual input settings: video without visible text (VID-V), video with visible text (VID-VS), image without visible text (IMG-V), and image with visible text (IMG-VS). We report Accuracy (Acc.), Unweighted Average Recall (UAR), and Weighted Average F1-score (WAF). Given the imbalanced emotion distribution, we use UAR as the primary metric and report Acc. and WAF as complementary measures.

Downstream emotion models. We evaluate AffectReveal with two general-purpose models, Qwen3.5-Omni-Plus (Team, 2026) and Qwen2.5-Omni-7B(Xu et al., 2025), and two emotion-oriented models, Emotion-LLaMA (Cheng et al., 2024) and AffectGPT (Lian et al., 2025). External search is disabled for all downstream models. For the locally deployable models, we additionally report variants fine-tuned on the EGER-Bench training set, denoted by FT.

Table 2: Performance on EGER-Bench under four visual input settings. FT denotes task-specific fine-tuning, with gray rows indicating fine-tuned downstream models.

Method	VID-V			VID-VS			IMG-V			IMG-VS		
	UAR	Acc.	WAF	UAR	Acc.	WAF	UAR	Acc.	WAF	UAR	Acc.	WAF
Emotion-LLaMA (Cheng et al., 2024)	20.42	20.90	15.69	21.36	21.99	16.05	18.87	19.72	15.36	18.38	19.63	15.11
Emotion-LLaMA + AffectReveal	**29.10**	**29.77**	**24.51**	**29.13**	**29.80**	**24.38**	**24.47**	**25.16**	**20.03**	**24.52**	**25.25**	**19.92**
Emotion-LLaMA (FT)	26.70	26.61	19.96	28.38	28.64	24.88	23.40	25.12	18.83	23.48	25.16	18.93
Emotion-LLaMA (FT) + AffectReveal	**31.17**	**31.69**	**26.96**	**32.08**	**32.69**	**28.39**	**28.14**	**28.91**	**24.35**	**28.73**	**29.60**	**25.41**
AffectGPT (Lian et al., 2025)	25.63	25.59	24.44	25.82	26.08	24.51	24.60	24.66	22.76	24.11	25.06	23.85
AffectGPT + AffectReveal	**29.48**	**30.00**	**28.02**	**29.93**	**30.46**	**28.51**	**30.70**	**30.25**	**28.70**	**31.07**	**30.87**	**29.44**
AffectGPT (FT)	25.66	26.06	24.68	25.42	26.02	24.28	25.39	25.28	23.68	26.02	25.68	24.53
AffectGPT (FT) + AffectReveal	**31.59**	**31.21**	**29.62**	**31.92**	**32.05**	**30.80**	**30.80**	**30.37**	**28.89**	**31.27**	**31.49**	**30.32**
Qwen2.5-Omni-7B (Xu et al., 2025)	17.11	17.94	11.13	21.85	22.52	17.02	16.66	18.04	13.77	18.95	20.37	16.99
Qwen2.5-Omni-7B + AffectReveal	**29.25**	**29.80**	**27.11**	**32.83**	**33.29**	**30.61**	**26.58**	**27.67**	**25.79**	**28.02**	**29.13**	**27.20**
Qwen2.5-Omni-7B (FT)	25.34	25.45	21.54	33.19	32.72	29.50	22.39	22.48	19.25	28.88	28.60	25.90
Qwen2.5-Omni-7B (FT) + AffectReveal	**36.55**	**36.08**	**35.44**	**36.48**	**36.78**	**33.54**	**33.13**	**32.55**	**32.07**	**35.94**	**35.31**	**34.86**

Implementation details. We instantiate Affect-Critical Event Recovery with Qwen3.5-Flash (Team, 2026). Primary retrieval uses one web-search call and one image-search call; alternative verification uses at most three additional web-search calls. Facial regions are detected frame by frame using MediaPipe Face Detection with the short-range BlazeFace model (Bazarevsky et al., 2019). In-media context extraction, atomic decomposition, and evidence scoring are implemented with Qwen2.5-Omni-7B. All supporting and downstream models remain frozen, and fixed-format prompts are used. Complete prompts and inference configurations are provided in Appendix E.

5.2 Main Results Across Models and Input Settings

We first evaluate whether AffectReveal is effective across downstream model families and visual input settings. For each model, we compare visual-only inference with the same model augmented by AffectReveal. For the three locally deployable models, we further repeat this comparison after task-specific fine-tuning on EGER-Bench, allowing us to examine whether inference-time event evidence remains useful after parameter adaptation. As shown in Table 2, AffectReveal improves UAR, accuracy, and WAF for every untuned downstream model under all four input settings. Averaged across the four settings, 7.05 for Emotion-LLaMA, 5.26 for AffectGPT, and 10.53 for Qwen2.5-Omni-7B. The improvements hold for both images and videos and remain substantial without visible text, indicating that AffectReveal is effective across visual formats and does not rely on textual cues being present.

AffectReveal also remains effective after task-specific fine-tuning. Across Emotion-LLaMA, AffectGPT, and Qwen2.5-Omni-7B, the untuned model augmented with AffectReveal outperforms its fine-tuned visual-only counterpart in all 12 accuracy comparisons. Applying AffectReveal to the fine-tuned models yields further gains across every model and input setting. This shows that task-specific parameter adaptation and instance-specific event evidence are complementary: fine-tuning improves how a model interprets the available input, whereas AffectReveal supplies event information that the input itself does not contain.

5.3 Diagnostic Value of External Event Context

To isolate the effect of missing event information, we conduct a diagnostic experiment with Qwen3.5-Omni-Plus under VID-V. We compare visual-only inference with two event-conditioned settings. **Reference Event Context** provides the source-grounded event description preserved during benchmark construction, whereas **AffectReveal** recovers event context automatically without access to this description. Explicit emotion labels and direct emotion-descriptive statements are removed from the reference context. Because it may still omit relevant role, relationship, or outcome details, it is treated as a diagnostic reference rather than a perfect oracle.

As shown in Table 3(a), reference event context raises UAR from 24.15% to 39.36%, confirming that missing event information constitutes a substantial bottleneck. AffectReveal reaches 34.76%

Table 3: Diagnostic and ablation results under VID-V. (a) Event-context diagnostic using Qwen3.5-Omni-Plus; (b) cumulative component analysis and (c) verification ablations using Qwen2.5-Omni-7B. UAR is the primary metric.

(a) Event Context

Evidence Setting	UAR	WAF
Visual Input Only	24.15	20.35
Visual + Reference Event	**39.36**	**39.02**
Visual + AffectReveal	34.76	33.32

(b) Evidence Components

Evidence Setting	UAR	WAF
Visual Input Only	17.11	11.13
+ External Event Context C_{ext}	27.84	25.73
+ In-Media Context C_{int}	27.75	25.56
+ AffectReveal R_{IE}	**29.25**	**27.11**

(c) Verification Designs

Variant	UAR	WAF
w/o Face Masking	28.89	24.37
w/o Atomic Support	28.95	26.74
Full AffectReveal	**29.25**	**27.11**

Table 4: Event recovery quality under VID-V. Generic Multi-Search and Affect-Critical Event Recovery use the same maximum retrieval budget.

Method	Affect-Critical Factor Correctness (%) ↑				Overall Event Recovery (%)		
	Event Identity	Focal Role	Relationship	Event Outcome	Match ↑	Partial ↑	Mismatch ↓
Primary Retrieval	16.35	36.67	23.33	12.25	13.25	21.55	65.20
Generic Multi-Search	15.44	38.90	25.53	14.32	12.96	25.92	61.12
Affect-Critical Event Recovery	**16.52**	**40.64**	**31.06**	**15.03**	**13.95**	**28.91**	**57.14**

UAR using automatically recovered evidence, closing 69.8% of the gap between the visual-only and reference-context conditions. Similar trends are observed for UAR and WAF. These results show that AffectReveal recovers a substantial portion of the task-relevant event information without privileged access to the benchmark reference context.

5.4 Does Affect-Critical Verification Improve Event Recovery?

Retrieving a relevant event does not necessarily recover the details needed for emotion interpretation. We therefore compare three strategies: **Primary Retrieval**, which uses only the initial search; **Generic Multi-Search**, which performs unconstrained query refinement; and **Affect-Critical Event Recovery**, which verifies evidence-grounded alternatives that differ in one affect-critical factor. Generic Multi-Search and our method use the same maximum retrieval budget, isolating the effect of how additional searches are organized.

We evaluate the recovered context against the source-grounded reference event at both event and factor levels. A recovery is labeled *Match* when the central event and its affect-critical details agree with the reference, *Partial* when the central event is recovered but some details remain incomplete or unresolved, and *Mismatch* when the event or an affect-determining detail conflicts with the reference. We additionally measure correctness for event identity, focal-person role, interpersonal relationship, and event outcome. The complete evaluation protocol and its agreement with human judgments are provided in Appendix C.

As shown in Table 4, Primary Retrieval produces a fully or partially aligned event in 34.80% of the samples. Generic Multi-Search increases this rate to 38.88%, showing that additional retrieval itself is beneficial. Under the same retrieval budget, Affect-Critical Event Recovery further raises it to 42.86% and reduces the mismatch rate to 57.14%. The largest factor-level gain over Generic Multi-Search occurs for interpersonal relationship, increasing from 25.53% to 31.06%, with additional gains for event identity, focal-person role, and event outcome. These results show that organizing retrieval around affect-critical alternatives recovers emotion-relevant event details more reliably than unconstrained query refinement.

5.5 How Does Event Verification Contribute to Emotion Recognition?

The recovered external context may be incomplete or partially mismatched with the observed sample. We therefore examine whether in-media facts should be treated simply as additional context or used to assess the compatibility of the recovered event. We first incrementally add the external event

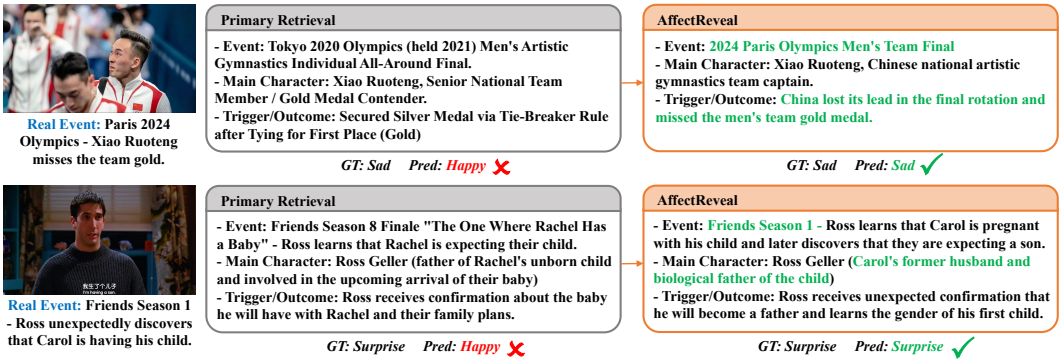

Figure 4: Qualitative comparisons between Primary Retrieval and AffectReveal. Although Primary Retrieval identifies the correct focal person, it retrieves a topically related but affectively incorrect event. AffectReveal corrects the event identity and outcome in the Olympic example, and the event identity and interpersonal relationship in the television-series example, leading to reference-aligned emotion predictions. Green text highlights the corrected event details.

context C_{ext}, the expression-disentangled in-media context C_{int}, and their explicit support relation R_{IE}. As shown in Table 3(b), external event context provides the dominant gain, raising UAR from 17.11% to 27.84%. Directly adding in-media context provides no further improvement, whereas explicitly modeling its support relation with the external event raises UAR to 29.25% and WAF to 27.11%. This indicates that in-media facts are more useful for assessing the recovered event than as another unconstrained textual description.

We next ablate the two mechanisms used to construct this support relation. The *w/o face masking* variant extracts C_{int} from the original input, allowing the facial expression being interpreted to also participate in event verification. The *w/o atomic support* variant replaces fact-level decomposition and alignment with a holistic comparison of the two contexts. As shown in Table 3(c), removing face masking reduces WAF by 2.74 points, supporting the concern that a potentially misleading expression can introduce circular evidence during verification. Replacing atomic support with holistic comparison causes smaller but consistent decreases across all metrics. Together, these results show that event verification is most effective when it uses expression-disentangled in-media facts and compares them with external evidence at the level of individual factual statements.

5.6 Qualitative Analysis

Figure 4 illustrates that recognizing the focal person or retrieving a related event is insufficient for EGER. In the Olympic example, Primary Retrieval identifies Xiao Ruoteng but associates him with the Tokyo individual all-around final and a silver-medal outcome, leading to a prediction of happiness. AffectReveal instead recovers the Paris team final and the missed opportunity for gold, supporting sadness. In the television example, Primary Retrieval identifies Ross but confuses Carol's pregnancy with a later storyline involving Rachel. Correcting the event identity and interpersonal relationship changes the prediction from happiness to surprise. These cases demonstrate that AffectReveal improves emotion recognition by resolving affect-critical event details rather than merely retrieving information about the visible person.

6 Conclusion

We introduced **Event-Grounded Emotion Recognition** and **EGER-Bench** for recognizing visual reactions whose affect-determining events are not fully observable. We also proposed **AffectReveal**, which verifies factorized event alternatives and cross-checks recovered context against expression-disentangled visual evidence. Across multiple models and visual settings, AffectReveal consistently improves recognition and complements task-specific fine-tuning without updating downstream parameters. These results highlight the importance of acquiring and verifying the events behind visually ambiguous reactions. Future work will extend EGER to dynamic and interactive settings where systems actively seek evidence and revise emotion interpretations as events unfold.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to assist with language polishing, text compression, and LaTeX formatting. The authors reviewed and revised all AI-assisted content and independently verified the technical claims, citations, experimental results, and numerical analyses reported in the paper. Generative AI was not used as an authoritative source for scientific claims or references, nor was it used to fabricate experimental data or results. The authors take full responsibility for the final content of the submission.

ETHICS STATEMENT

Emotion recognition from human behavior raises ethical concerns related to privacy, subjective interpretation, cultural variation, and potential misuse in surveillance or high-stakes decision-making. EGER-Bench is intended solely for research on event-grounded emotion understanding and should not be interpreted as providing objective psychological assessments of individuals. EGER-Bench is constructed from news/interview media and television-series content with source provenance retained for event grounding. For any release, we will respect applicable source licenses and copyright restrictions and provide documentation describing the intended research use and limitations of the benchmark. We do not advocate the use of AffectReveal or EGER-Bench for surveillance, clinical assessment, employment screening, law enforcement, or other high-stakes decisions about individuals.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Sections 3–5 describe the dataset construction, AffectReveal framework, and experimental settings. The appendix provides complete prompts, evaluation protocols, implementation details, and inference configurations used in our experiments.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pp. 279–283, 2016.

Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 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.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37: 110805–110853, 2024.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1639–1649, 2016.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Manh-Hung Hoang, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Context-aware emotion recognition based on visual relationship detection. *IEEE Access*, 9:90465–90474, 2021.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pp. 874–880, 2021.

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

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotic: Emotions in context dataset. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 2309–2317. IEEE, 2017.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42 (11):2755–2766, 2019.

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 10142–10151. IEEE, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(11):1–19, 2025.

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

Siyuan Li, Xinde Li, and Chuanfei Hu. Towards revealing trustworthy emotion states via human-object interaction-aware emotion recognition with vision-language model. *Pattern Recognition*, pp. 113342, 2026.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025.

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

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 24–32, 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14222–14231. IEEE, 2020.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE transactions on affective computing*, 10(1):18–31, 2017.

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

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, 2021.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5370–5381, 2019.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844, 2022.

Zeheng Wang, Zitong Yu, Yijie Zhu, Bo Zhao, Haochen Liang, Taorui Wang, Wei Xia, Jiayu Zhang, Zhishu Liu, Hui Ma, et al. Affectagent: Collaborative multi-agent reasoning for retrieval-augmented multimodal emotion recognition. *arXiv preprint arXiv:2604.12735*, 2026.

Shichao Wu, Lei Zhou, Zhengxi Hu, and Jingtai Liu. Hierarchical context-based emotion recognition with scene graphs. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3): 3725–3739, 2022.

Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1003–1012, 2019.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025.

Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. Emotion recognition for multiple context awareness. In *European conference on computer vision*, pp. 144–162. Springer, 2022.

Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. Context de-confounded emotion recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19005–19015. IEEE, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8799–8809. IEEE, 2019.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10477–10486, 2023.

APPENDIX CONTENTS

The appendix is structured as follows:

- Appendix A reports additional EGER-Bench results with Qwen3.5-Omni-Plus, an API-accessed closed-source downstream model.
- Appendix B presents a vision-only transfer evaluation on MER2023, examining whether recovered event evidence can also benefit conventional visual emotion recognition.
- Appendix C validates the automatic event recovery evaluation by comparing Qwen-based assessments with human annotations.
- Appendix D analyzes AffectReveal's performance across News/Interview and TV Series source domains.
- Appendix E provides the prompt templates and inference configuration used by AffectReveal's core components.

A ADDITIONAL RESULTS WITH QWEN3.5-OMNI-PLUS

In addition to the locally deployable downstream models reported in the main comparison, we evaluate AffectReveal with Qwen3.5-Omni-Plus, an API-accessed closed-source vision-capable MLLM. Since its model parameters are not available, we evaluate the untuned model only and do not include a task-specific fine-tuning variant. External search is disabled for the downstream model itself, following the same protocol as the main experiments; external evidence is introduced only through AffectReveal.

Table 5: Performance of Qwen3.5-Omni-Plus on EGER-Bench under four visual input settings.

Method	VID-V			VID-VS			IMG-V			IMG-VS		
	UAR	Acc.	WAF	UAR	Acc.	WAF	UAR	Acc.	WAF	UAR	Acc.	WAF
Qwen3.5-Omni-Plus (Team, 2026)	24.15	25.00	20.35	34.91	35.56	33.24	20.35	22.37	17.71	27.08	28.98	26.01
Qwen3.5-Omni-Plus + AffectReveal	**34.76**	**35.41**	**33.32**	**39.50**	**40.11**	**38.30**	**29.16**	**30.84**	**28.77**	**31.70**	**33.25**	**31.45**

As shown in Table 5, AffectReveal consistently improves Qwen3.5-Omni-Plus across all four visual input settings and all three evaluation metrics. The average UAR improvement across the four settings is 7.16 points. The largest UAR gain occurs under VID-V, increasing from 24.15 to 34.76, while IMG-V improves from 20.35 to 29.16. Improvements are also retained when visible text is available, with UAR increasing from 34.91 to 39.50 under VID-VS and from 27.08 to 31.70 under IMG-VS.

These results complement the locally deployable models reported in the main experiments and show that the benefit of AffectReveal is not restricted to downstream models whose parameters or training procedures are accessible.

B VISION-ONLY TRANSFER TO MER2023

As a secondary transfer analysis, we evaluate the same inference-time framework on the vision-only setting of MER2023. Table 6 reports the corresponding results. This experiment is not used to establish the EGER task because MER2023 does not explicitly require external event context; it instead tests whether recovered event evidence can help conventional visual emotion recognition when the observed media is limited.

C VALIDATION OF AUTOMATIC EVENT RECOVERY EVALUATION

To analyze the correspondence between automatic and human assessment of event recovery quality, we compare Qwen-based judgments with human annotations on a randomly sampled subset. Specifically, we randomly select 500 test samples under the VID-VS setting and ask six annotators

Table 6: Vision-only transfer results on MER2023.

Method	Acc.	UAR	WAF
Qwen3.5-Omni-Plus	78.63	**69.66**	79.20
Qwen3.5-Omni-Plus+AffectReveal	**81.51**	69.60	**81.61**
Emotion-LLaMA	42.81	32.71	36.60
Emotion-LLaMA+AffectReveal	**50.96**	**38.44**	**48.18**
AffectGPT	48.44	20.57	49.47
AffectGPT+AffectReveal	**53.84**	**23.69**	**56.73**
Qwen2.5-Omni-7B	36.69	37.64	36.86
Qwen2.5-Omni-7B+AffectReveal	**41.25**	**45.54**	**44.24**

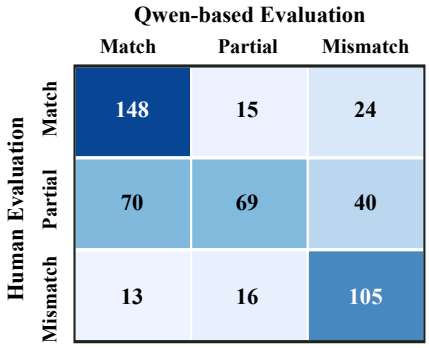

Figure 5: Comparison between Qwen-based and human assessments of event recovery quality. The confusion matrix is computed on 500 randomly sampled VID-VS test samples using three recovery levels: Match, Partial, and Mismatch.

to categorize each recovered event context into three levels: Match, Partial Match, and Mismatch. Figure 5 presents the confusion matrix between Qwen-based and human assessments. The two evaluations agree at 64.4%, with most samples concentrated along the diagonal. The remaining differences mainly occur between neighboring categories, especially for partially recovered events, reflecting the ambiguity in assessing event-level consistency.

D Performance Across Source Domains

EGER-Bench contains two source domains with different forms of event context. News and interview samples are typically associated with publicly documented events, whereas television-series samples depend more heavily on narrative context, character roles, and interpersonal relationships. We therefore examine whether AffectReveal improves upon initial event retrieval in both domains.

We partition the VID-V test set into News/Interview and TV Series subsets and evaluate Qwen2.5-Omni-7B separately on each. **Primary Retrieval** conditions the downstream model on the event interpretation obtained from the initial search, whereas **AffectReveal** applies the complete event recovery and cross-verification pipeline.

As shown in Table 7, AffectReveal improves all three metrics in both source domains. On News/Interview, it improves accuracy, UAR, and WAF over Primary Retrieval by 0.29, 0.58, and 0.76 points, respectively. The gains are larger on TV Series, reaching 2.08 points in accuracy, 1.60 in UAR, and 3.78 in WAF. Across all samples, AffectReveal improves accuracy from 28.74% to 29.80%, UAR from 28.09% to 29.25%, and WAF from 25.09% to 27.11%.

These results show that AffectReveal provides consistent benefits across both publicly documented and narrative-driven events. Its larger improvement on television-series samples is consistent with the motivation of resolving affect-critical ambiguity, as these samples more often require information

Table 7: Emotion recognition across source domains under VID-V using Qwen2.5-Omni-7B. News/Interview and TV Series account for 57% and 43% of the evaluation samples, respectively.

Method	New/Interview (57%)			TV Series (43%)			All		
	UAR	Acc.	WAF	UAR	Acc.	WAF	UAR	Acc.	WAF
Primary Retrieval	28.87	32.40	27.23	23.64	23.92	23.67	28.09	28.74	25.09
AffectReveal	**29.45**	**32.69**	**27.99**	**25.24**	**26.00**	**27.45**	**29.25**	**29.80**	**27.11**

about narrative outcomes, character roles, and interpersonal relationships beyond the focal visual input.

E Prompt Templates and Inference Configuration

This section lists the fixed prompts used by the core components of AffectReveal together with the corresponding inference settings. Following the presentation style of agentic systems, we represent sample-dependent inputs with brace-enclosed placeholders, while keeping the fixed instructions unchanged from the implementation.

Inference Configuration. For Qwen3.5-Flash, thinking is disabled for the initial visual analysis within Visual Cue-Grounded Retrieval, enabled for the tool-using primary retrieval and Alternative Verification, and disabled for Event Arbitration and the final serialization of the recovered external event context C_{ext}. Tool selection is automatic during the retrieval calls. For Qwen2.5-Omni-7B, which is used for Non-Facial In-Media Context and Bidirectional Atomic Evidence Verification, we use greedy decoding with `do_sample=False` and `max_new_tokens=512`. Video input for Non-Facial In-Media Context is sampled at 1 FPS with audio disabled.

E.1 Affect-Critical Event Recovery

Visual Cue-Grounded Retrieval. The following is the complete video prompt used to extract retrieval-oriented visual cues and search queries in the reported visual-only setting.

```
VISUAL-ONLY MODE: Use visual information only. The supplied video has
    no audio track. Keep all audio fields empty and never use audio
    clues.

Analyze the video across multiple moments for an external-information
    retrieval agent that will later assist an emotion classification
    model.

Do not identify a final emotion label. Separate directly observed facts
     from hypotheses. Extract only clues useful for identifying the
    person, event, place, relationship, or outcome.

Return exactly one JSON object with this schema:
{
  "observed": {
    "people": ["brief visible description"],
    "scene": "brief scene description",
    "actions": ["important temporal action or interaction"],
    "visible_text": ["OCR/logo/name/number if readable"],
    "distinctive_clues": ["clothing, object, venue, organization,
    broadcast or event clue"],
    "affective_cues": ["visible tears, smile, restrained posture,
    cheering, embrace, etc."],
    "uncertainties": ["important unclear visual point"]
  },
  "queries": [
    {
      "query": "concise web-search query",
```

```
      "purpose": "identify_event|identify_person|verify_fact",
      "basis": "objective|hypothesis"
    }
  ],
  "visual_search_recommended": true
}

Rules:
1. Integrate the whole video rather than describing every frame.
2. Produce 2 queries by default and no more than 3.
3. Query 1 should be the most reliable objective query.
4. Always set visual_search_recommended=true.
5. Do not output a final emotion label, Markdown, or text outside JSON.
```

The structured visual analysis is then used for primary event retrieval and evidence-grounded alternative construction.

```
You are the primary retrieval module for a hidden-emotion understanding
     agent.
Your first task is to identify the most likely real-world event. Your
    second task is NOT to invent counterfactual stories, but to harvest
     a small set of evidence-anchored alternative event interpretations
     from the material you actually retrieved.

Stage-1 visual analysis:
{stage1_result_json}

Tool policy:
1. Perform ONE focused web_search using the strongest objective
    combination of
   visual, OCR, entity, number, place, and event clues.
2. Representative visual input is attached. Perform exactly ONE
    image_search for same/near-duplicate matching.
3. Use web_extractor only for up to 3 promising pages
   whose body is needed to verify identity, role, event outcome,
    relationship, or
   a concrete event title/description.
4. Prefer reputable/directly matching sources. Preserve contradictions.
5. Do not search generic emotion knowledge. Search the concrete event
    background.

STEP A | Build a PRIMARY EVENT ANCHOR.
Extract the stable event core from retrieved evidence: named people/
    teams/organization, place, date/time, competition/program/incident,
     score/result, quoted phrase, article headline, or exact/near-
    duplicate image match. Separate verified and uncertain slots.

STEP B | Harvest EVIDENCE-ANCHORED COUNTERFACTUAL CANDIDATES.
Generate 0 to 3 alternatives ONLY if the retrieved material itself
    supplies a concrete reason for that alternative. A valid candidate
    must:
- share the primary event core or be a directly confusable nearby real
    event;
- change exactly ONE uncertainty slot: subject_role, event_outcome,
    relationship, or event_identity;
- cite at least 1 concrete evidence_anchor_facts from the retrieved
    search result/snippet/extracted page/image match;
- remain compatible with Stage-1 directly observed visual facts;
- be specific enough to verify with a focused web query.

IMPORTANT:
- Do NOT create an alternative merely because it is semantically
    possible.
```

```
- Do NOT use a generic opposite such as win->lose, parent->coach, award
    ->elimination unless the retrieved material mentions or strongly
    implies that concrete alternative.
- Do NOT force a fixed number of candidates. Returning [] is correct
    when retrieval exposes no grounded alternative.
- Prefer a candidate mentioned by a second retrieved article/title/
    snippet over a candidate invented from world knowledge.

Return exactly one JSON object:
{
  "primary_query_used": "main text query used for web_search",
  "best_event": "current best identified event, or not reliably
    identified",
  "best_event_confidence": 0.0,
  "image_match_strength": "exact|near_duplicate|similar|none|not_used",
  "event_anchor": {
    "entities": ["person/team/organization explicitly supported"],
    "event_name": "event/program/match/incident name or uncertain",
    "time_place": "date/place if supported",
    "stable_facts": ["facts shared by plausible interpretations"],
    "uncertain_slots": ["subject_role|event_outcome|relationship|
    event_identity"]
  },
  "subject_identity_role": "current best role interpretation",
  "trigger_outcome": "current best result/trigger",
  "relationship_social_context": "current best relationship/social
    context",
  "verified_facts": ["facts directly supported by retrieved sources"],
  "candidate_facts": ["plausible but incompletely verified claims"],
  "contradictions": ["important mismatch or uncertainty"],
  "counterfactual_candidates": [
    {
      "id": "CF1",
      "dimension": "subject_role|event_outcome|relationship|
    event_identity",
      "priority": 1,
      "primary_event": "primary event interpretation",
      "alternative_event": "one concrete competing event interpretation
    ",
      "shared_anchor_facts": ["facts preserved from the primary event
    "],
      "evidence_anchor_facts": ["retrieved fact that motivated this
    alternative"],
      "evidence_grounding": "strong|medium|weak|none",
      "changed_slot_primary": "value under the primary interpretation",
      "changed_slot_alternative": "value under the alternative
    interpretation",
      "why_plausible": "why this is a real competing interpretation
    under the retrieved evidence",
      "why_emotion_critical": "why this one changed slot matters for
    later affect interpretation",
      "expected_discriminator": "specific fact that can distinguish the
     two",
      "verification_query": "focused open-web query for this concrete
    alternative"
    }
  ]
}

Return no more than 3 candidates. Never pad the list.
Do not output a final emotion label. Do not output Markdown or text
    outside JSON.
```

Alternative Verification. Each evidence-grounded alternative is independently checked with the following prompt.

```
You are an independent verification branch for one evidence-anchored
    competing event interpretation. The candidate already came from
    retrieved evidence; your job is to verify it, not to elaborate it
    into a new story.

Stage-1 multimodal analysis:
{stage1_result_json}

Primary retrieval result:
{primary_result_json}

Candidate to verify:
{candidate_json}

Perform exactly ONE focused web_search using plan.verification_query or
     a stricter equivalent. Do not repeat the broad primary query.
    image_search is not available.
Use web_extractor only for up to 2 pages when body text is needed.

Verification policy:
1. Stage-1 directly observed visual facts and plan.shared_anchor_facts
    are invariants. The candidate may change only plan.dimension.
2. First verify that plan.evidence_anchor_facts were not misread or
    taken out of context. A candidate collapses if its motivating
    evidence is itself incorrect.
3. Prefer facts with high specificity: exact name+role, score/result,
    relation, quoted statement, date/place, event title, official
    roster/report, or a news description tied to the same event.
4. Generic overlap (crying, hugging, interview, crowd, celebration) is
    weak evidence.
5. Explicitly assess media_consistency and evidence_specificity.
6. Do not invent a third interpretation. Only compare the supplied
    alternative with the primary interpretation.

Return exactly one JSON object:
{
  "counterfactual_id": "{candidate_id}",
  "dimension": "{candidate_dimension}",
  "counterfactual_query_used": "actual query",
  "primary_event": "primary interpretation",
  "alternative_event": "candidate interpretation",
  "counterfactual_status": "supported|partially_supported|not_supported
    |not_found|conflicting",
  "media_consistency": "high|medium|low|none",
  "evidence_specificity": "high|medium|low|none",
  "supporting_facts": ["specific facts supporting the candidate"],
  "refuting_facts": ["specific facts refuting it or supporting primary
    "],
  "discriminative_facts": ["facts that distinguish the two"],
  "effect_on_primary": "strengthens_primary|weakens_primary|ambiguous|
    unrelated",
  "confidence": 0.0,
  "summary": "concise evidence-based comparison"
}

Do not output a final emotion label. Do not output Markdown or text
    outside JSON.
```

Event Arbitration. Reconsider the primary interpretation and retained alternatives together without further browsing.

```
Arbitrate ONE most likely event interpretation from a primary event and
     independently verified evidence-anchored alternatives. Do not
    browse and do not add facts.

Stage-1 multimodal analysis:
{stage1_result_json}
Primary event:
{primary_result_json}
Retained competing candidates:
{retained_candidates_json}
All branch results (including rejected ones for audit):
{counterfactual_results_json}
Tool counts:
{tool_counts_json}
Source records:
{source_records_json}

Arbitration policy:
1. Do not use majority voting.
2. Prefer the interpretation that jointly maximizes: media consistency,
     source specificity, direct event/role/outcome evidence, and
    independent support.
3. A candidate that needs changing more than its declared dimension is
    invalid.
4. Evidence compatible with both interpretations is non-discriminative
    and should not decide the winner.
5. If no interpretation is sufficiently discriminated, return uncertain
    /conflicting instead of forcing a winner.
6. Keep the final output focused on objective event, role, relationship
    , and outcome; do not assign an emotion label.

Return exactly one JSON object:
{
  "status": "confirmed|probable|uncertain|conflicting|not_found",
  "confidence": 0.0,
  "source_agreement": "high|medium|low|none",
  "image_match_strength": "exact|near_duplicate|similar|none|not_used",
  "causal_discrimination": "high|medium|low|none",
  "best_event": "single selected event or not reliably identified",
  "verified_facts": ["facts retained after comparison"],
  "main_character": "focal person's identity/role if supported",
  "event_trigger_outcome": "result/trigger if supported",
  "relationship_context": "relationship if supported",
  "counterfactual_test": "brief account of why alternatives were
    retained or rejected",
  "conflicts": ["remaining uncertainty"]
}
Do not output Markdown or text outside JSON.
```

The arbitration output is serialized into the compact external event context used by the subsequent verification stage.

```
Convert the materials below into a minimal JSON object for an emotion-
    understanding model.
Do not browse and do not add facts absent from the materials.

Current counterfactual output mode: arbitrate

Stage-1 media analysis:
{stage1_result_json}

Retrieval evidence note:
{retrieval_note}

Actual tool counts:
```

```
{tool_counts_json}

Actual source records:
{source_records_json}

ARBITRATION MODE:
- Use only the single BEST_MATCH selected by the arbitration note.
- Prefer a compact structure inside the context string: "Event: ...
    Main Character: ..."
- Include objective role/relationship and trigger/outcome; avoid
    unnecessary affective interpretation and never assign the final
    emotion label.

Return exactly this JSON schema:
{
  "status": "confirmed|probable|uncertain|conflicting|not_found",
  "confidence": 0.0,
  "source_agreement": "high|medium|low|none",
  "image_match_strength": "exact|near_duplicate|similar|none|not_used",
  "context": "one compact English string"
}

Context-writing rules:
1. context must never be empty, even when retrieval fails.
2. Write 45-110 English words when enough material exists.
3. Prioritize concrete event identity, main-character role,
    relationship, and outcome.
4. Never introduce an event/candidate absent from the retrieval note.
5. Do not mention URLs, tools, numerical scores, retrieval rounds, or
    source counts.
6. If identity/event is uncertain, say so explicitly rather than
    inventing a story.
7. Do not assign one final emotion-class label.
8. In multi_event mode, ambiguity is intentional: reliability refers to
     the candidate set, not to uniqueness of the true event.

Do not output Markdown or text outside JSON.
```

E.2 Expression-Disentangled Event Verification

All Qwen2.5-Omni-7B calls in this stage use the following system message.

```
You are a multimodal context analysis assistant.
```

Non-Facial In-Media Context. After facial regions are masked, the following video prompt extracts only non-facial visual information.

```
The input video has masked human faces.

Analyze ONLY non-facial visual information.

Describe:
- human actions
- body posture
- objects
- environment
- human interactions
- visible event clues

Do NOT infer emotion.
Do NOT use facial expressions.
Do NOT use audio.
Do NOT use subtitles.
```

```
Return JSON:

{
 "internal_context":"..."
}
```

Bidirectional Atomic Evidence Verification. The internal and external contexts are first decomposed into atomic factual evidence.

```
Decompose the following two contexts into atomic factual evidence.

Rules:
- One fact per evidence.
- No emotion inference.
- No additional information.
- Generate at most 4 atomic evidence items for Internal Context.
- Generate at most 4 atomic evidence items for External Context.
- If more than 4 facts are available, keep only the 4 most informative
    and event-relevant facts.
- Use IDs I1, I2, ... for Internal Evidence and E1, E2, ... for
    External Evidence.

Internal Context:
{internal}

External Context:
{external}

Return JSON:

{
 "internal_evidence":[
  {"id":"I1","text":"..."}
 ],
 "external_evidence":[
  {"id":"E1","text":"..."}
 ]
}
```

All internal–external evidence pairs are then scored in one ordered call.

```
Evaluate the support score for EVERY Internal-External evidence pair
    listed below.

Score meaning:
- 1.0: strongly supporting / describing the same factual clue
- 0.5: compatible or partially supporting
- 0.0: unrelated or conflicting

Use ONLY 0.0, 0.5, or 1.0.
Do NOT use intermediate values.

Pairs are listed in the EXACT output order:
{pair_lines}

There are exactly {n_pairs} pairs.

Output exactly ONE line in this format and nothing else:
SCORES: {score_placeholders}

Requirements:
- After "SCORES:" there MUST be exactly {n_pairs} numeric values
    separated by commas.
```

```
- The k-th value corresponds to the k-th pair above.
- Do not omit any pair.
- Do not output JSON, IDs, explanations, relations, markdown, or any
    other text.
```

E.3 Event-Conditioned Emotion Inference

For each downstream model, we retain the original emotion-classification instruction used by the corresponding baseline and only adapt its output label space to the 11 emotion categories of EGER-Bench. AffectReveal does not replace the baseline-specific classification instruction. Instead, the verified evidence package is appended to the original instruction as auxiliary context, while the original unmasked visual input remains unchanged.

The auxiliary evidence instruction is serialized as follows:

```
External Context:
{external}

Internal Context:
{internal}

Internal-External Evidence Alignment:
Alignment Score: {score}
Aligned Evidence: {aligned_evidence_pairs}
```

The resulting auxiliary instruction is concatenated with the original emotion-classification instruction of each downstream model before inference. No additional context-rewriting model is used.