OVGER: EVIDENCE-GROUNDED OPEN-VOCABULARY GROUP EMOTION RECOGNITION VIA BLIND ANNOTATION AND PROGRESSIVE REWARD INJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Group Emotion Recognition (GER) infers collective affect from multi-person audio-visual recordings for social-interaction analysis, multimedia understanding, and human-centered computing. Most existing GER systems formulate the task as closed-set classification over coarse affect categories. Although effective for benchmark comparison, this discriminative interface provides limited fine-grained affect description and little explicit evidence for the prediction. Motivated by this gap, we formulate open-vocabulary GER as a benchmark-compatible extension beyond conventional closed-set prediction and instantiate it with **OVGER**. The model retains the coarse group-affect decision while adding more specific free-form emotion descriptors and an evidence-linked structured response. The central contribution is the supervision-and-reward interface that makes these outputs jointly trainable and auditable: **OV-VGAF** augments the Video-level Group AFfect (VGAF) benchmark with label-blind modality observations, structured supervision, and hidden evidence references for evaluating coarse labels together with open-vocabulary affect and evidence alignment. This design couples benchmark-compatible group decisions with fine-grained affect descriptions whose supporting multimodal cues can be inspected. Extensive experiments show competitive coarse-label performance relative to representative task-specific GER systems and clear gains over the evaluated zero-shot MLLMs on the same 766-clip validation split, reaching **74.67%** accuracy, **74.50%** macro-F1, and **85.51%** open-vocabulary emotion soft-F1.

1 INTRODUCTION

Group Emotion Recognition (GER) aims to infer the collective affect expressed by multiple people in an image or video, with applications in social-interaction analysis, multimedia understanding, and human-centered computing (Veltmeijer et al., 2023). Group affect is not simply the emotion of the most salient individual: interaction, shared attention, vocal behavior, and emotional contagion can jointly shape the state expressed by a group (Hatfield et al., 1993; Barsade, 2002). Existing GER benchmarks and methods therefore provide a useful basis for learning group-level affect from visual or audio-visual cues (Sharma et al., 2019; Dhall et al., 2020; Petrova et al., 2020; Sun et al., 2020; Liu et al., 2020; Evtodienko, 2021). Despite this progress, most systems ultimately return a small set of coarse affect classes. Such a discriminative output is easy to score, but it does not describe finer affective states or expose the observable multimodal cues associated with a prediction.

Multimodal large language models (MLLMs) make a richer interface possible because they can jointly interpret visual content and language, while omni-modal models can additionally consume audio (Xu et al., 2025; Wang et al., 2025; Bai et al., 2025). Reinforcement learning can further optimize structured free-form outputs with automatically computed rewards (Shao et al., 2024; Guo et al., 2025; Liu et al., 2025; Feng et al., 2025; Li et al., 2025). These capabilities expose a missing direction in GER: moving beyond a single coarse decision toward benchmark-compatible group prediction that also expresses fine-grained affect and grounds it in observable multimodal evidence. As Figure 1(a)–(c) illustrates, existing interfaces progress from closed-set prediction to open-ended description and task-oriented reasoning, but do not jointly expose the group-level decision, fine-grained affect, and

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.

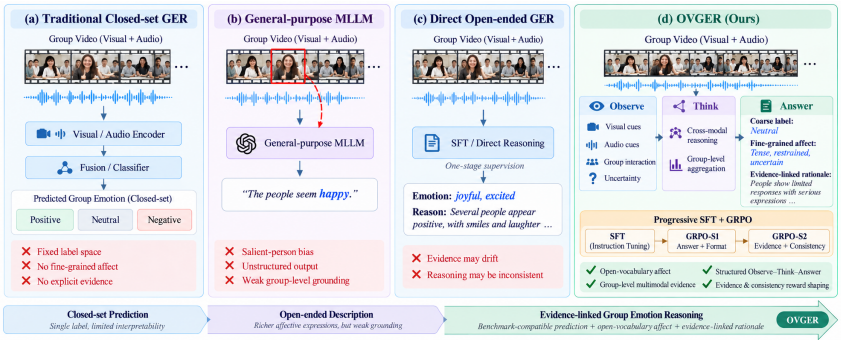


Figure 1: **Comparison of GER paradigms.** (a) Traditional closed-set GER maps a group clip to a predefined coarse label. (b) General-purpose MLLMs support open-ended descriptions, and (c) direct open-ended GER adds task-oriented reasoning. (d) OVGER goes beyond these interfaces by coupling group-level audio-visual observation and cross-modal reasoning with a benchmark-compatible coarse decision, fine-grained open-vocabulary affect, and evidence-linked observe-think-answer output.

checkable audio-visual evidence. Figure 1(d) depicts the target interface pursued here: group-level audio-visual observations support cross-modal reasoning, while the answer preserves the VGAF label and adds open-vocabulary affect whose evidence can be inspected.

However, directly applying open-ended generation to VGAF leaves three practical gaps. First, the original benchmark supplies class labels rather than evidence annotations, so structured supervision must be constructed without leaking the target label into the modality-observation stage. Second, fine-grained emotion descriptions are non-unique, which makes exact-string supervision and evaluation inappropriate. Third, generic Group Relative Policy Optimization (GRPO) rewards do not directly encode the GER-specific requirements of coarse-label correctness, structured output, multimodal evidence alignment, and consistency between evidence, interpretation, and answer.

To address these challenges, we propose **OVGER**, which connects the conventional benchmark label, open-vocabulary collective affect, and checkable multimodal evidence through a GER-specific supervision-and-reward interface. Figure 1(d) makes the distinction explicit: unlike Figure 1(a)–(c), OVGER preserves the coarse group decision while jointly exposing fine-grained affect, group-level multimodal observations, and evidence-linked reasoning. Figure 2 summarizes supervision construction and training, with annotation details in Appendix A. We first build **OV-VGAF**: *label-blind* observers describe audio and visual evidence without the VGAF target label or the three coarse label words, and label-conditioned fusion converts those observations into the visible observe-think-answer target and hidden evidence references. This separation keeps perception tied to observable cues. We then train Qwen2.5-Omni-7B with SFT and progressive GRPO: S1 stabilizes answer correctness and format, whereas S2 adds evidence-grounding and cross-section-consistency rewards.

On the fixed VGAF validation split, the final model reaches 74.67% accuracy, 74.50% macro-F1, and 85.51% open-vocabulary emotion soft-F1. It improves over the strongest evaluated zero-shot MLLM by 10.44 accuracy points and 12.33 macro-F1 points; published task-specific GER results remain contextual because their protocols differ. Under the same 200-update budget, S1→S1 reaches 72.45%, S2→S2 reaches 71.41%, and progressive S1→S2 reaches 74.67%, supporting reward ordering rather than extra updates as the source of the gain.

Our contributions are fourfold:

- We formulate open-vocabulary group emotion recognition (OVGER), extending benchmark-compatible coarse GER with fine-grained affect and evidence-linked structured output.
- We construct OV-VGAF with label-blind modality observations, structured targets, and hidden evidence references.
- We introduce progressive GRPO reward injection after SFT, stabilizing answer/format before evidence/consistency; equal-update controls isolate the ordering effect.
- Experiments show competitive coarse-label performance, clear gains over evaluated zero-shot MLLMs, and benefits from audio-visual input and progressive rewards.

2 Related Work

Closed-set group emotion recognition. The Video-level Group AFfect (VGAF) benchmark was introduced for video-level group affect and later used as the audio-video GER task in EmotiW 2020 (Sharma et al., 2019; Dhall et al., 2020). Representative methods improve the coarse group-affect decision through privacy-safe holistic visual modeling (Petrova et al., 2020), temporal and static fusion (Sun et al., 2020), hybrid audio-video networks (Liu et al., 2020), and cross-modal attention (Evtodienko, 2021). More recent work explores spatial-temporal graph modeling (Wang et al., 2024), multimodal spatiotemporal Transformers (Huang & Xu, 2026), hierarchical multimodal aggregation (Mouafo Mapikou & Cao, 2026), and kinematics-informed representations (Singh et al., 2026). These approaches establish strong task-specific modeling, but their primary prediction interface remains the benchmark's coarse label. By comparison, OVGER retains that label for direct GER evaluation while extending the output with fine-grained free-form affect descriptors and structured evidence-linked text.

Multimodal emotion understanding and open-ended generation. General-purpose MLLMs provide strong image/video understanding, and omni-modal models additionally support joint audio-video inputs (Xu et al., 2025; Wang et al., 2025; Bai et al., 2025; GLM-V Team et al., 2025). Emotion-LLaMA and AffectGPT further show that multimodal emotion recognition can support natural-language reasoning and descriptive emotion understanding (Cheng et al., 2024; Lian et al., 2025). These developments show promising capabilities for open-ended affect description. OVGER targets a complementary group-level setting in which the original VGAF coarse label remains the primary GER target, while the open-vocabulary set provides more specific affect concepts and the observe/think fields expose externally checkable evidence and interpretation within the same benchmark-compatible output.

Reward-based multimodal post-training. GRPO was introduced in DeepSeekMath and later adopted by reasoning-oriented systems (Shao et al., 2024; Guo et al., 2025). Visual-RFT (Liu et al., 2025), Video-R1 (Feng et al., 2025), and VideoChat-R1 (Li et al., 2025) extend automatically scored reward-based post-training to image and video reasoning. For GER, the reward must simultaneously preserve the coarse label, permit semantically equivalent fine-grained emotion wording, follow a structured response contract, and remain aligned with multimodal evidence. OVGER therefore constructs task-specific annotation-side references and reward components for these requirements. Together, these components preserve coarse-label correctness while making free-form outputs structurally valid, evidence-aligned, and internally consistent.

3 Methodology

3.1 Task Formulation

Given a short group video v with its audio track, the model produces <observe> o </observe> <think> t </think> <answer> a </answer>. The observation o summarizes directly observable audio-visual group cues, t connects those cues to a group-affect interpretation, and a is a JSON object containing a coarse label $y \in \{\textit{positive}, \textit{neutral}, \textit{negative}\}$ and an open-vocabulary emotion set E. The coarse label preserves compatibility with standard VGAF evaluation, while E complements it with more specific free-form descriptors. We use *evidence-linked* to mean that observable claims in the visible response can be checked against annotation-side evidence references.

3.2 OV-VGAF Construction

OV-VGAF extends the EmotiW 2020 VGAF benchmark (Dhall et al., 2020), itself an expanded version of the video-level group-affect resource introduced by Sharma et al. (2019). Its construction separates label-blind modality observation from label-conditioned target construction: audio and visual observers first record temporally localized, directly observable cues without receiving the target class, and only the fusion stage receives the coarse supervision label. The resulting record contains a visible observe-think-answer target and a separate hidden reference record for automatic scoring and diagnostics. The left panel of Figure 2 summarizes this construction flow; observer roles, prompts, and quality-control details are provided in Appendices A and F.

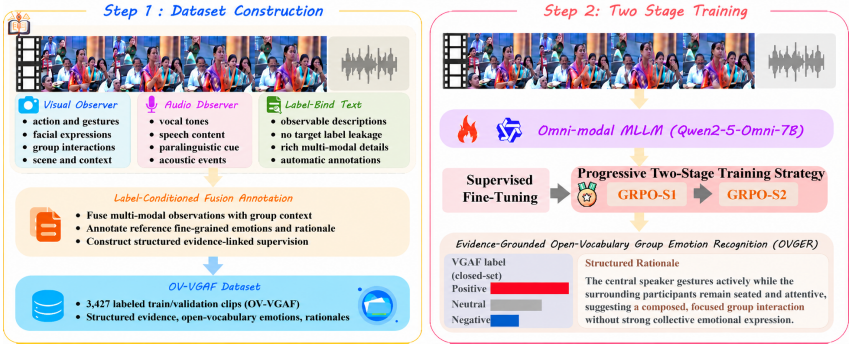


Figure 2: **OVGER overview.** The left panel constructs OV-VGAF from label-blind visual/audio observations and label-conditioned fusion; the right panel trains the omni-modal backbone with SFT and progressive GRPO-S1/GRPO-S2. The output preserves the VGAF coarse label while adding evidence-linked reasoning and open-vocabulary affect.

Label-blind observation and fusion. The audio stream records speech, prosody, loudness, rhythm, non-speech events, and uncertainty; the visual stream records actions, posture, facial/body dynamics, group interaction, scene context, temporal changes, and uncertainty. Neither observer receives the ground-truth VGAF label, and the words *positive*, *neutral*, and *negative* are prohibited in their outputs, keeping observer text tied to perceptual cues rather than the target class. The fusion annotator then receives both observation streams and the supervision label to construct the visible structured target together with reference fine-grained emotions, visual/audio/group/uncertainty evidence banks, temporal arcs, and forbidden claims. The hidden record remains separate from the visible training target.

Quality control and statistics. Rule-based checks reject label leakage and malformed records; label-inconsistent fusion outputs are retried, and an independent verifier audits grounding, label consistency, specificity, temporal consistency, and SFT usability before a final format check. These quality-control stages are applied automatically before a record is retained. OV-VGAF contains 2,661 training clips (802/923/936 positive/neutral/negative) and 766 validation clips (302/280/184), for 3,427 clips in total. Hidden records contain 14.58/14.01 evidence references and 4.48/4.43 forbidden-claim references per train/validation clip on average. Full class percentages, the corruption-based reward sanity check, and target-length statistics are reported in Appendices A and B.

3.3 Supervised Cold Start

To establish the observe-think-answer response contract before reinforcement learning, we apply LoRA-based supervised fine-tuning to the 2,661 structured training targets (Hu et al., 2022). For a serialized target $z = (o, t, a)$, multimodal input x, and prompt p,

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{j=1}^{|z|} \log \pi_\theta(z_j \mid x, p, z_{<j}). \tag{1}$$

Here θ denotes the trainable adapter parameters. We adapt the Qwen2.5-Omni thinker (Xu et al., 2025); the speech-generation talker is not needed for the textual GER output and remains frozen. Optimization details are summarized in Section 4.1 and Appendix B.

3.4 GRPO with Progressive Reward Injection

SFT teaches the output structure, but the open-vocabulary response admits multiple semantically valid realizations and contains properties that are easier to score than to specify with a single target string. We therefore use GRPO to optimize automatically computed GER-specific rewards after the supervised cold start (Shao et al., 2024). For each training clip, the behavior policy $\pi_{\theta_{\mathrm{old}}}$ samples G

response sequences s_i with scalar rewards r_i. Their group-relative advantages are

$$A_i = \frac{r_i - \text{mean}(r)}{\text{std}(r) + 10^{-8}}.$$

With token-level importance ratio $\rho_{i,t}(\theta) = \pi_\theta(s_{i,t} \mid x, s_{i,<t}) / \pi_{\theta_{\text{old}}}(s_{i,t} \mid x, s_{i,<t})$, optimization uses the standard clipped surrogate together with a KL penalty toward the frozen SFT reference policy π_{ref}:

$$\max_\theta \ \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|s_i|}\sum_t \min(\rho_{i,t}A_i, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)A_i) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right]. \tag{2}$$

Here ϵ is the clipping radius, β weights the KL penalty, and $\text{clip}(x, l, u) = \min(\max(x, l), u)$ bounds a scalar to $[l, u]$.

Reward overview. Four rewards target complementary properties: R_{ans} preserves the benchmark decision while allowing semantic variation in fine-grained emotion words, R_{format} enforces the response contract, R_{evidence} measures alignment with annotation-side multimodal evidence while penalizing forbidden claims, and $R_{\text{consistency}}$ checks agreement across observe, think, and answer. The stage-2 weights keep coarse-label correctness primary, assign the next-largest mass to evidence, use consistency as a coherence constraint, and reduce the format weight after SFT has established the structure. A recorded weight-sensitivity diagnostic is provided in Appendix C.

All semantic rewards use normalized BGE-M3 embeddings (Chen et al., 2024). For cosine similarity c, define

$$\phi(c) = \text{clip}\left(\frac{c - 0.40}{0.75 - 0.40}, 0, 1\right), \tag{3}$$

which maps similarities at or below 0.40 to zero, at or above 0.75 to one, and interpolates linearly in between.

Answer reward. Let P and R be the predicted and reference open-vocabulary emotion sets. We compute bidirectional maximum semantic matching,

$$P_{\text{soft}} = \frac{1}{|P|}\sum_{p \in P}\max_{r \in R}\phi(c(p, r)), \tag{4}$$

$$R_{\text{soft}} = \frac{1}{|R|}\sum_{r \in R}\max_{p \in P}\phi(c(p, r)), \tag{5}$$

$$c_E = \frac{2P_{\text{soft}}R_{\text{soft}}}{P_{\text{soft}} + R_{\text{soft}}}, \tag{6}$$

with $c_E = 0$ if either set is empty. Here P_{soft} averages the closest reference match for each predicted emotion, R_{soft} performs the reverse coverage calculation, and c_E is their harmonic mean. The fine-emotion subreward and complete answer reward are

$$r_{\text{fine}} = \phi(c_E) + \big(1 - \phi(c_E)\big)0.3c_E, \tag{7}$$

$$R_{\text{ans}} = 0.8\,\mathbf{1}[\hat{y} = y] + 0.2r_{\text{fine}}, \tag{8}$$

where $\mathbf{1}[\hat{y} = y]$ equals one when the predicted coarse label matches the supervision label and zero otherwise.

Format reward. The format reward starts from one and subtracts penalties for missing sections, an illegal coarse label, extra explanation outside the contract, unexpected JSON keys, or an invalid emotion list:

$$\begin{aligned} R_{\text{format}} = \text{clip}\big(1 &- 0.20I_{o=\varnothing} - 0.15I_{t=\varnothing} - 0.30I_{\text{illegal } y} \\ &- 0.25I_{\text{extra explanation}} - 0.10I_{\text{other extra keys}} - 0.15I_{\text{invalid } E}, 0, 1\big). \end{aligned} \tag{9}$$

A 0.50 hard cap is applied if internal evidence identifiers or forbidden coarse-label wording leak into the observation section. The relative penalties follow structural severity: illegal labels and explanation outside the response contract receive the largest deductions, while missing sections, unexpected keys, and an invalid emotion list receive smaller deductions.

Evidence reward. Split the concatenated generated observe and think text into sentences g_j. For each reference evidence item $e_{k,m}$ in bank $k \in \{\text{visual}, \text{audio}, \text{group}, \text{uncertainty}\}$, compute its best generated-sentence match

$$u_{k,m} = \max_j \phi\big(c(e_{k,m}, g_j)\big). \tag{10}$$

The bank score e_k is the mean of the three largest available $u_{k,m}$ values. Thus "top-3" is taken over reference-level best matches, not over references for each generated sentence. Matching is not one-to-one: a generated sentence may be the best match for multiple references, and references are not removed after matching. The base score is

$$R^{\text{base}}_{\text{evidence}} = 0.35e_{\text{visual}} + 0.25e_{\text{audio}} + 0.25e_{\text{group}} + 0.15e_{\text{uncertainty}}. \tag{11}$$

Let m_{fc} be the maximum cosine similarity between the generated observe/think content and any annotation-side forbidden claim. The penalty is

$$\lambda_{\text{fc}} = \begin{cases} 0, & m_{\text{fc}} \le 0.75, \\ 0.60\dfrac{m_{\text{fc}} - 0.75}{0.10}, & 0.75 < m_{\text{fc}} \le 0.85, \\ 0.60, & m_{\text{fc}} > 0.85, \end{cases} \tag{12}$$

and $R_{\text{evidence}} = \text{clip}(R^{\text{base}}_{\text{evidence}} - \lambda_{\text{fc}}, 0, 1)$. The penalty therefore reduces the evidence score when generated text is too close to claims that the annotation record marks as unsupported.

Consistency reward. Let $h(q)$ denote a normalized BGE-M3 embedding and let a_y be the fixed polarity anchor for coarse label y. Define

$$\psi(q, y) = \text{clip}\left(\frac{\cos(h(q), a_y) - \max_{y' \neq y} \cos(h(q), a_{y'})}{0.2} + 0.5, 0, 1\right). \tag{13}$$

The four implemented subrewards are: (i) $r_{o\leftrightarrow t}$, the mean bidirectional sentence-level best-match similarity between observe and think after the semantic map, with 0.5 used when a side has no sentence; (ii) $r_{t\leftrightarrow a} = \psi(t, \hat{y})$, which tests whether the think paragraph supports the answer's coarse polarity; (iii) $r_{E\leftrightarrow y} = \psi(E, \hat{y})$, computed from the serialized predicted fine-emotion text and therefore measuring fine/coarse answer consistency rather than evidence-bank matching; and (iv) r_{group}, a lexical group-focus score,

$$r_{\text{group}} = \text{clip}\left(\min\left(1, \frac{n_{\text{group}}}{3}\right) - 0.25n_{\text{individual}}, 0, 1\right). \tag{14}$$

The complete consistency reward is

$$R_{\text{consistency}} = 0.30r_{o\leftrightarrow t} + 0.30r_{t\leftrightarrow a} + 0.25r_{E\leftrightarrow y} + 0.15r_{\text{group}}. \tag{15}$$

Progressive reward injection. GRPO-S1 concentrates on the benchmark answer and response contract; GRPO-S2 then shifts part of the reward mass toward evidence and consistency:

$$R_{\text{S1}} = 0.8R_{\text{ans}} + 0.2R_{\text{format}}, \tag{16}$$

$$R_{\text{S2}} = 0.5R_{\text{ans}} + 0.1R_{\text{format}} + 0.25R_{\text{evidence}} + 0.15R_{\text{consistency}}. \tag{17}$$

The schedule controls are reported in Section 4.5.

4 Experiments

4.1 Implementation Details

We train Qwen2.5-Omni-7B (Xu et al., 2025) on a single NVIDIA RTX PRO 6000 GPU, sample video frames at 1 fps with at most eight frames, and retain the original audio track. During SFT, we adapt the thinker for six epochs with rank-16 LoRA and seed 42. During GRPO, we use a separate rank-8 LoRA adapter with group size $G = 8$ for 100 optimizer updates in S1 followed by 100 in S2. Groups with reward standard deviation below 10^{-6} are skipped before backpropagation and do not advance the update counter, so checkpoint 200 represents 200 actual optimizer updates. Evaluation uses greedy decoding. Learning rates, optimizer settings, context/generation limits, software environment, and target lengths are in Appendix B.

4.2 Evaluation Protocol

Accuracy uses the 766 VGAF validation clips as a fixed denominator, with missing, unparsable, or illegal labels counted as incorrect; macro-F1 is the unweighted mean over positive, neutral, and negative. Strict format requires a parseable answer JSON with a legal coarse label. OVE soft-F1 uses bidirectional BGE-M3 semantic matching between predicted and reference fine-grained emotion sets. Evidence grounding, consistency, and forbidden-claim rate use the annotation-side reference system defined in Section 3.4; they are reported as reference-agreement diagnostics alongside the original-label metrics and qualitative analysis. Exact aggregation and parser rules are in Appendix D. MLLM baselines evaluated on the same 766-clip split include GLM-4.6V-Flash, Qwen3-VL-8B, InternVL3.5-8B, Qwen2.5-Omni-7B, Emotion-LLaMA, and AffectGPT (GLM-V Team et al., 2025; Bai et al., 2025; Wang et al., 2025; Xu et al., 2025; Cheng et al., 2024; Lian et al., 2025); published task-specific GER values are cited under their reported protocols. Full baseline configurations are in Appendix E.

4.3 Comparison with Representative Baselines

Table 1 separates published task-specific GER results from MLLMs evaluated on the fixed 766-clip VGAF validation split. Published rows provide contextual references under their reported protocols, while the zero-shot MLLM and OVGER rows use our common validation protocol.

Table 1: Comparison with representative VGAF systems and zero-shot MLLMs. Published GER values are contextual references under the evaluation setting shown in "Eval."; zero-shot MLLM and OVGER metrics use the fixed 766-clip validation split. "–" denotes unavailable metrics.

Category	Method	Size	Eval.	Acc (%)	Macro-F1 (%)	Format (%)
Published GER	Privacy-safe holistic GER (Petrova et al., 2020)	–	Val.	57.18	–	–
	Cross-modal attention (Evtodienko, 2021)	–	Val.	60.37	–	–
	Spatial-temporal GCN (Wang et al., 2024)	–	reported	70.23	–	–
	VIBE (Singh et al., 2026)	–	reported	70.17	–	–
	Spatio-temporal + static fusion (Sun et al., 2020)	–	Val.	71.93	–	–
	Hierarchical multimodal Transformer (Mouafo Mapikou & Cao, 2026)	–	5-fold	72.56	–	–
	Hybrid Networks (Liu et al., 2020)	–	Val.	74.28	–	–
Zero-shot MLLM	GLM-4.6V-Flash (GLM-V Team et al., 2025)	9B	Val.	51.31[†]	42.76[†]	9.14
	Qwen3-VL-8B (Bai et al., 2025)	8B	Val.	55.22	46.75	100
	InternVL3.5-8B (Wang et al., 2025)	8B	Val.	59.66	55.35	100
	Qwen2.5-Omni-7B (Xu et al., 2025)	7B	Val.	57.83	52.13	98.96
	Emotion-LLaMA (Cheng et al., 2024)	7B	Val.	62.27	59.48	100
	AffectGPT (Lian et al., 2025)	7B	Val.	64.23	62.17	100
OVGER	OVGER (vision-only)	7B	Val.	71.15	70.49	99.6
	OVGER (ours)	7B	Val.	**74.67**	**74.50**	**100**

[†] The recorded GLM run uses a lenient answer parser for accuracy/Macro-F1; strict-format compliance is shown separately.

Comparison with task-specific GER systems. Published GER rows use heterogeneous protocols, so they provide context rather than a strict cross-paper ranking. OVGER's 74.67% accuracy lies in the range of the strongest values summarized here, including the reported 74.28% of Hybrid Networks (Liu et al., 2020). The key distinction is output capability: OVGER reaches this coarse-label range while also producing fine-grained open-vocabulary affect and evidence-linked text, rather than trading benchmark compatibility for a richer interface.

Comparison with evaluated MLLMs. Emotion-LLaMA and AffectGPT reach 62.27%/59.48% and 64.23%/62.17% accuracy/macro-F1, respectively. OVGER reaches 74.67%/74.50%, improving over the strongest evaluated zero-shot MLLM by 10.44 accuracy points and 12.33 macro-F1 points. The vision-only OVGER result (71.15%/70.49%) already exceeds these zero-shot baselines, while restoring audio adds a further 3.52 accuracy points and 4.01 macro-F1 points. Together, these comparisons indicate that task-specific structured supervision and acoustic context contribute complementary gains.

4.4 Effect of Training Stages

Table 2 shows complementary gains across training phases. SFT produces the largest initial improvement, raising accuracy from 57.83% to 69.32% and OVE soft-F1 from 75.18% to 83.65%, which shows the value of the structured supervision cold start. GRPO-S1 and GRPO-S2 then add another

5.35 accuracy points over SFT while OVE remains high and increases to 85.51%. Thus the reward stages refine the benchmark decision without collapsing the fine-grained affect vocabulary learned during SFT.

Table 2: Results along the training pipeline. "–" means that the diagnostic was not recorded for the same checkpoint as the reported Acc/Macro-F1. Lower is better for Forbidden.

Model	Acc (%)	Macro-F1 (%)	OVE (%)	Evidence (%)	Consistency (%)	Forbidden (%)
Base (Qwen2.5-Omni-7B)	57.83	52.13	75.18	80.51	82.66	55.61
SFT	69.32	68.97	83.65	88.82	87.00	4.44
+ GRPO-S1	72.58	72.34	85.32	–	–	–
+ GRPO-S2 (final)	**74.67**	**74.50**	**85.51**	85.74	87.58	18.02

4.5 Ablation Studies

Response structure. We retrain SFT from the same base with identical data and hyperparameters while varying which sections appear in the target (Table 3).

Table 3: Response-structure ablation at the SFT level with identical data and training hyperparameters.

Training target	observe	think	answer	Acc (%)	Macro-F1 (%)
Full three-part	✓	✓	✓	**69.32**	**68.97**
w/o think	✓	×	✓	67.75	66.87
w/o observe	×	✓	✓	69.19	67.36

Removing *think* costs 1.57 accuracy points and 2.10 macro-F1 points, whereas removing *observe* costs only 0.13 accuracy points but 1.61 macro-F1 points. The asymmetry suggests that *think* is the main cue-integration step, while *observe* provides a complementary evidence surface that supports class-balanced prediction and later grounding rewards; the full scaffold is strongest on both metrics.

Progressive reward schedule. Table 4 fixes the seed and 200 actual updates: S1→S1 uses answer/format rewards throughout, S2→S2 uses the full reward from the start, and S1→S2 is progressive. The progressive recipe reaches 74.67%/74.50%, versus 72.45%/72.22% and 71.41%/71.34% for the two controls. With budget held constant, the result isolates the benefit of first stabilizing answer/format behavior and then shaping evidence and consistency; skipped low-variance groups do not count as updates.

Table 4: GRPO schedule controls at the same cumulative step-200 optimizer-update budget. "–" denotes diagnostics not recorded for the corresponding checkpoint. Lower is better for Forbidden.

Recipe	1–100	101–200	Acc. (%)	F1 (%)	OVE (%)	Evid. (%)	Cons. (%)	Forbid. (%)
Progressive OVGER	R_{S1}	R_{S2}	**74.67**	**74.50**	**85.51**	**85.74**	87.58	18.02
S1-only continuation	R_{S1}	R_{S1}	72.45	72.22	–	–	–	–
Reward-matched one-shot	R_{S2}	R_{S2}	71.41	71.34	84.94	85.59	**87.78**	**17.49**

S2 reward components. Table 5 shows a clear division of labor. Removing R_{ans} causes the largest coarse-task drop (2.61 accuracy points); removing R_{evidence} lowers grounding from 85.74% to 82.94% and raises forbidden claims from 18.02% to 24.46%; removing $R_{\mathrm{consistency}}$ lowers its diagnostic from 87.58% to 85.12%. R_{format} has the smallest late-stage effect (0.39 accuracy points, 99.74% format), consistent with structure already being stabilized by SFT/S1. The components therefore protect different properties rather than duplicating one another. A separate archived sensitivity check involving an auxiliary temporal term not used in the final S2 objective is reported in Appendix C.

Input modality. Holding the checkpoint fixed, audio+vision improves over vision-only by 3.52 accuracy points and 4.01 macro-F1 points (Table 6). The weaker audio-only result shows that acoustic cues complement rather than replace visual evidence, helping resolve cases ambiguous from appearance and motion alone. The audio-only row uses its recorded lenient parser; strict-parser accuracy is in the caption.

Table 5: S2 reward-component ablation.

S2 setting	Acc. (%)	F1 (%)	OVE (%)	Evid. (%)	Cons. (%)	Forbid. (%)	Format (%)
Full	**74.67**	**74.50**	**85.51**	85.74	87.58	18.02	**100.00**
w/o Answer	72.06	71.89	84.41	**86.16**	**88.01**	**17.48**	100.00
w/o Format	74.28	74.09	85.47	85.78	87.52	18.21	99.74
w/o Evidence	73.50	73.28	85.19	82.94	86.87	24.46	100.00
w/o Consistency	73.89	73.69	85.30	85.18	85.12	19.37	100.00

Table 6: Modality ablation; audio-only Acc/F1 use the lenient parser (strict accuracy: 31.46%).

Input	Acc (%)	Macro-F1 (%)	Strict format (%)
Audio + vision (full)	**74.67**	**74.50**	**100**
Vision only	71.15	70.49	99.6
Audio only	57.83	57.25	50.5

4.6 Qualitative and Diagnostic Analysis

Group-level target and diagnostics. Individual-focused prompting changes 43.21% of coarse outputs and lowers group-label agreement from 74.67% to 58.36%, confirming the group/individual distinction. From the base model to OVGER, forbidden claims fall 55.61%→18.02%, while grounding rises 80.51%→85.74% and consistency 82.66%→87.58%, showing better-grounded, more coherent output.

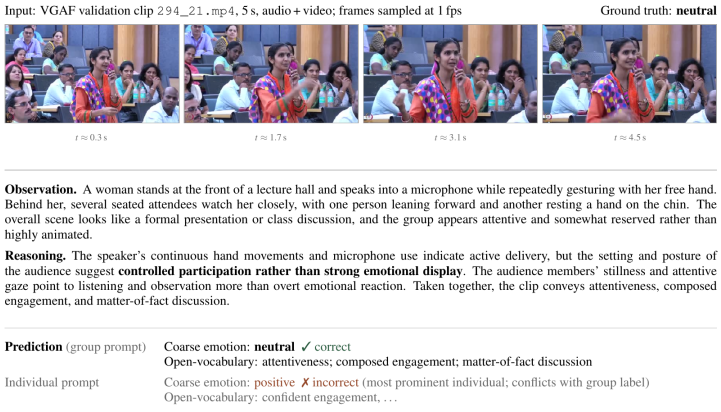


Figure 3: Group-level prediction; individual-focused prompting changes *neutral* to *positive*.

Representative cases. Figure 3 shows a neutral group despite an active speaker; the individual-focused prompt predicts positive. On `100_1.mp4`, OVGER uses group and temporal cues across the full observe-think-answer structure, whereas recorded baselines are shorter, observation-heavy, or format-incomplete. These cases expose collective evidence behind the decision; more are in Appendices H and I.

5 Conclusion

OVGER retains the VGAF group decision while adding fine-grained affect and evidence-linked text through label-blind observation, label-conditioned targets, and progressive SFT+GRPO. It reaches 74.67% accuracy, 74.50% macro-F1, and 85.51% OVE soft-F1; equal-budget and component ablations show distinct benefits from reward ordering and specialized evidence/consistency terms. Future work will extend to longer interactions and independent human evaluation.

Limitations

The study is limited to short, three-label VGAF clips, and the evidence, consistency, and open-vocabulary diagnostics are reference-based automatic measures rather than independent human judgments.

AI USE STATEMENT

AI models are used as methodological components of the OVGER framework. The annotation pipeline uses `gemini-3.1-flash-lite` as the audio observer, `qwen3.5-397b-a17b` as the visual observer, `gpt-5.4` as the fusion annotator, and `gpt-5.4-mini` as the independent verifier. The trainable OVGER backbone is Qwen2.5-Omni-7B, and BGE-M3 is used for semantic reward and evaluation matching. Generative AI assistance was also used for English-language editing and manuscript revision. All AI-assisted text and figures were reviewed by the authors, and the reported scientific claims, experimental values, citations, interpretations, and final submitted content remain the authors' responsibility.

ETHICS STATEMENT

Automatic emotion recognition carries substantial societal risks, including surveillance misuse, privacy intrusion, and cultural or contextual misinterpretation. Group-level inference introduces an additional aggregation risk: the behavior of a salient individual may be overgeneralized to the group, while minority or mixed affect within the scene may be obscured by a single collective label. OVGER exposes an evidence-linked rationale that can be inspected, but fluent explanatory text can itself create unwarranted confidence when the underlying prediction or cited evidence is wrong; inspectability should therefore not be equated with correctness or causal faithfulness. Label-blind observation prevents direct target-label injection into the modality-observation prompts, but it does not remove dataset bias, annotation bias, privacy concerns, or the possibility of harmful downstream use. Any deployment in consequential settings should require context-appropriate human review, explicit privacy safeguards, and compliance with applicable regulations.

REPRODUCIBILITY STATEMENT

The paper specifies the OV-VGAF annotation pipeline, model identifiers, core reward definitions and weights, the 6-epoch SFT schedule, progressive GRPO, input sampling, parsing rules, and evaluation metrics. Appendices A–I document the annotation roles and schemas, optimization settings, reward computation, fixed-denominator evaluation, baseline configurations, prompts, and qualitative cases. Upon acceptance, the release will include the OV-VGAF annotations and hidden reward references together with the corresponding schema, prompts, training configurations, and training/evaluation code needed to reproduce the reported pipeline, subject to the redistribution terms of the source VGAF videos.

REFERENCES

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Sigal G. Barsade. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, 47(4):644–675, 2002. doi: 10.2307/3094912.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi: 10.18653/v1/2024.findings-acl.137.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110805–110853, 2024. doi: 10.52202/079017-3518.

Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. EmotiW 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 784–789, 2020. doi: 10.1145/3382507.3417973.

Lev Evtodienko. Multimodal end-to-end group emotion recognition using cross-modal attention. *arXiv preprint arXiv:2111.05890*, 2021.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3310.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Elaine Hatfield, John T. Cacioppo, and Richard L. Rapson. Emotional contagion. *Current Directions in Psychological Science*, 2(3):96–100, 1993. doi: 10.1111/1467-8721.ep10770953.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Xiaohua Huang and Jinke Xu. Multimodal spatiotemporal semisupervised transformer network for video-based group-level emotion recognition. *IEEE Transactions on Computational Social Systems*, 13(2):1719–1733, 2026. doi: 10.1109/TCSS.2025.3622991.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014. PMLR, 2025. URL https://proceedings.mlr.press/v267/lian25a.html.

Chuanhe Liu, Wenqiang Jiang, Minghao Wang, and Tianhao Tang. Group level audio-video emotion recognition using hybrid networks. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 807–812, 2020. doi: 10.1145/3382507.3417968.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025. doi: 10.1109/ICCV51701.2025.00197.

Gaëlle Laetitia Mouafo Mapikou and Houwei Cao. Group emotion recognition using multimodal transformers with hierarchical aggregation. In *2026 International Conference on Computing, Networking and Communications (ICNC)*, pp. 955–960. IEEE, 2026. doi: 10.1109/ICNC68183.2026.11416990.

Anastasia Petrova, Dominique Vaufreydaz, and Philippe Dessus. Group-level emotion recognition using a unimodal privacy-safe non-individual approach. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 813–820, 2020. doi: 10.1145/3382507.3417969.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Garima Sharma, Shreya Ghosh, and Abhinav Dhall. Automatic group level affect and cohesion prediction in videos. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 161–167, 2019. doi: 10.1109/ACIIW.2019.8925231.

Abhishek Pratap Singh, Vaibhav Pratap Singh, Deepak Kumar, and Balasubramanian Raman. VIBE: Disentangling social dynamics via kinematics-informed variational inference for behavioral emotion. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Mo Sun, Jian Li, Hui Feng, Wei Gou, Haifeng Shen, Jian Tang, Yi Yang, and Jieping Ye. Multi-modal fusion using spatio-temporal and static features for group emotion recognition. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 835–840, 2020. doi: 10.1145/3382507.3417971.

Emmeke A. Veltmeijer, Charlotte Gerritsen, and Koen V. Hindriks. Automatic emotion recognition for groups: A review. *IEEE Transactions on Affective Computing*, 14(1):89–107, 2023. doi: 10.1109/TAFFC.2021.3065726.

Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xingzhi Wang, Tao Chen, and Dong Zhang. A spatial-temporal graph convolutional network for video-based group emotion recognition. In *Pattern Recognition – 27th International Conference, ICPR 2024*, volume 15313 of *Lecture Notes in Computer Science*, pp. 339–354. Springer, 2024. doi: 10.1007/978-3-031-78201-5_22.

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. *arXiv preprint arXiv:2503.20215*, 2025.

A Annotation and OV-VGAF Construction

Annotation roles. Table 7 summarizes the information available to each annotation component. The key separation is between label-blind modality observation and label-conditioned fusion. Here, label-blind means that the modality observer does not receive the coarse target label: neither modality observer receives the VGAF label, whereas the fusion stage receives the coarse label to construct the supervised target and hidden reward references.

Table 7: Annotation roles and information flow. "GT label" denotes the VGAF coarse supervision label.

Role	Model	Primary input	GT label	Output
Audio observer	`gemini-3.1-flash-lite`	raw audio	no	audio observations
Visual observer	`qwen3.5-397b-a17b`	original video	no	visual observations
Fusion annotator	`gpt-5.4`	both observations	yes	target + reward refs
Verifier	`gpt-5.4-mini`	fused record	–	quality-gate decision

Dataset statistics. OV-VGAF contains 2,661 training and 766 validation clips. The positive/neutral/negative counts are 802/923/936 for training and 302/280/184 for validation. The hidden reward references contain visual, audio, group-interaction, uncertainty, and temporal evidence. Counting all evidence types gives 14.58 items per training clip and 14.01 per validation clip on average; the four evidence types used by R_{evidence} average 12.15 and 11.74 items, respectively. The corresponding forbidden-claim banks contain 4.48 and 4.43 entries per clip on average. The retained annotation records span July 22–26, 2026; the model identifiers in Table 7 are the identifiers recorded for those annotation runs.

Table 8: OV-VGAF split statistics. Evidence and forbidden counts are per-clip averages over hidden annotation-side references.

Split	Clips	Positive	Neutral	Negative	Evidence refs.	Forbidden refs.
Train	2,661	802 (30.14%)	923 (34.69%)	936 (35.17%)	14.58	4.48
Validation	766	302 (39.43%)	280 (36.55%)	184 (24.02%)	14.01	4.43

Quality-control sequence. After modality observation, rule-based checks reject label-word leakage. Fusion then constructs the visible target and hidden references, retries label-inconsistent records, and is followed by an independent verifier stage and format checks. A corruption-based stress test replaces the correct answer label with an incorrect one while retaining the remaining annotation structure; the mean composite reward changes from 97.4% for the gold target to 55.0% for this corruption. This test is used as a pre-training sanity check that the reward reacts to a major answer-level error.

B Training and Reproducibility Details

SFT. The SFT adapter is trained in three consecutive two-epoch blocks (six epochs total) with learning rates 5×10^{-5}, 2.5×10^{-5}, and 1.25×10^{-5}. Each block uses LoRA rank 16, $\alpha = 32$, dropout 0.05, weight decay 0.01, cosine scheduling, zero warmup, per-device batch size 1, gradient accumulation 8 (effective batch size 8), maximum gradient norm 1.0, context length 4096, bf16, gradient checkpointing, and seed 42. LoRA targets the thinker attention and MLP projections (q/k/v/o and gate/up/down projections); the talker is not adapted. Video inputs use 1 fps with at most eight frames and the original audio track.

GRPO. The GRPO adapter uses rank 8, $\alpha = 16$, dropout 0.05, AdamW with weight decay 0.01, learning rate 5×10^{-7}, maximum gradient norm 0.5, bf16, and gradient checkpointing. The group size is $G = 8$, sampling temperature is 1.1, $\beta = 0.01$, $\epsilon = 0.2$, and generation is capped at 768 new tokens. Gradient accumulation is four prompts; because each prompt produces eight sampled completions, one optimizer update aggregates 32 completions. Groups with reward standard deviation below 10^{-6} are skipped before backpropagation and do not advance the accumulation counter. GRPO-S1 uses the answer/format reward mixture described in the main text; GRPO-S2 switches to the stage-2 mixture and continues from the S1 endpoint.

Reference and behavior policies. The reference policy π_{ref} is implemented by temporarily disabling the active GRPO LoRA adapter, yielding the frozen SFT-merged backbone. It remains fixed across the two GRPO stages. The behavior policy $\pi_{\theta_{\text{old}}}$ is the online policy at candidate generation time; the denominator of the PPO-style importance ratio uses the stored generation-time log-probability, while the KL penalty compares the current policy with π_{ref}.

Target lengths. Using the Qwen2.5-Omni-7B training tokenizer, supervised targets average 301.23 tokens on the 2,661 training clips and 284.36 tokens on validation. The observe/think/answer sections average 117.82/137.99/30.42 tokens in training and 109.27/129.71/30.38 in validation; the user-side task prompt is 55 tokens.

Training hardware and software environment. Training was conducted on a single NVIDIA RTX PRO 6000 GPU. The recorded training stack uses PyTorch/Transformers with PEFT and Accelerate, LLaMA-Factory for SFT, sentence-transformers for BGE-M3 encoding, and torch SDPA attention. GRPO is implemented by the project training loop rather than a TRL trainer.

C Limited Reward-Weight Sensitivity

A recorded single-parameter sweep from an earlier auxiliary-reward variant varies a temporal reward weight while holding its other coefficients fixed at answer=0.80, format=0.20, evidence=0.25, and consistency=0.10. This archived variant is not the final stage-2 objective in Equation (17); the sweep is retained only as a local sensitivity diagnostic.

Table 9: Temporal-weight sensitivity. Peak values are taken from the recorded checkpoints for each run.

Temporal weight	Peak checkpoint	Acc. (%)	Macro-F1 (%)
0.15	125	73.76	73.68
0.10	150/225	73.24	73.09
0.05	175	72.85	72.43

Across these three archived settings, peak accuracy decreases as the temporal weight is reduced, indicating that temporal information contributed useful signal in that variant. Because the sweep contains only three weights and does not use the final S2 objective, it is best interpreted as a local sensitivity check rather than a global hyperparameter search.

D Evaluation and Baseline Protocols

Canonical coarse-label evaluation. All reported coarse-label accuracies use the fixed 766-clip validation denominator. A missing `<answer>` section, unparsable JSON, or a `coarse_label` outside {positive, neutral, negative} is counted as an incorrect prediction. For macro-F1, such an invalid response contributes one false negative to its ground-truth class and no false positive to another class; macro-F1 is the arithmetic mean of the three classwise F1 values. Under this protocol the pipeline values are 443/766 (57.83%, 52.13% macro-F1) for the base model, 531/766 (69.32%, 68.97%) after SFT, 556/766 (72.58%, 72.34%) after GRPO-S1, and 572/766 (74.67%, 74.50%) for the final model.

Table 10: Canonical fixed-denominator validation results used throughout the paper. Invalid structured outputs are counted as incorrect.

Model	Correct / 766	Acc. (%)	Macro-F1 (%)	Invalid
Base	443 / 766	57.83	52.13	8
SFT	531 / 766	69.32	68.97	2
GRPO-S1	556 / 766	72.58	72.34	0
Final	572 / 766	74.67	74.50	0

Table 11: Class-wise validation metrics for the final GRPO-S2 checkpoint under the canonical strict 766-clip protocol.

Class	Precision (%)	Recall (%)	F1 (%)	Support
Positive	79.59	77.48	78.52	302
Neutral	76.60	64.29	69.90	280
Negative	66.67	85.87	75.06	184
Accuracy		74.67% (572/766)		
Macro-F1		74.50%		

Open-vocabulary emotion metric. Let P be the predicted emotion strings and R the fusion-annotation reference field `fine_emotions_ref`. For BGE-M3 cosine similarity $c(p, r)$, define

$$P_{\text{soft}} = \frac{1}{|P|} \sum_{p \in P} \max_{r \in R} \phi(c(p, r)), \tag{18}$$

$$R_{\text{soft}} = \frac{1}{|R|} \sum_{r \in R} \max_{p \in P} \phi(c(p, r)), \tag{19}$$

$$F_{1,\text{soft}} = \frac{2 P_{\text{soft}} R_{\text{soft}}}{P_{\text{soft}} + R_{\text{soft}}}. \tag{20}$$

This is bidirectional maximum semantic matching, not one-to-one Hungarian matching. If the prediction set is empty, precision and recall are zero; samples with an empty reference set are excluded from OVE aggregation. The same fixed BGE-M3 snapshot used by the rewards is used for OVE and the grounding metrics.

Automatic grounding metrics. The evidence-grounding diagnostic applies the same defined evidence-bank matching and forbidden-claim penalty used by R_{evidence}; it is therefore a reference-based diagnostic rather than an independent human grounding score. The consistency diagnostic applies the $R_{\text{consistency}}$ definition. The forbidden-claim rate counts clips for which the maximum cosine similarity between the concatenated observe/think text and any per-clip forbidden claim exceeds 0.75; the final value 18.02% corresponds to 138/766 clips. These metrics are reported together with original-label Acc/Macro-F1 and qualitative error analysis to make their scope explicit.

Prompt protocol. The Base/SFT evaluation prompt and the GRPO evaluation prompt specify the same group-emotion task, the same three tagged sections, and the same two answer fields, but are not word-for-word identical. The stricter Base/SFT form explicitly enumerates the three legal coarse labels and constrains the emotion list to 1–5 entries; the GRPO form is the compact training-time prompt and names the two required fields without those extra enumerations. In the GRPO-S1 and final evaluations, all 766 responses contain a legal coarse label. The prompt distinction is retained here for reproducibility.

E Zero-Shot Baseline Configuration

GLM-4.6V-Flash. The official GLM-4.6V release identifies the Flash variant as a nominal 9B visual-language model. The recorded evaluator gives 51.31% accuracy, 42.76% macro-F1, and 9.14% strict-format compliance for the GLM run. Because the GLM accuracy/Macro-F1 were produced with the run's lenient answer parser, Table 1 marks these two values with a dagger and reports strict-format compliance separately.

The final Qwen3-VL row in Table 1 uses the formatting-adjusted prompt. Relative to the original chat-template form, the adjustment only adds line breaks after the first two tagged sections and an explicit closing instruction; task semantics and answer fields are unchanged. The original form produced 1/766 strictly formatted responses (0.13%); the adjusted run reaches 100% strict format compliance and provides the reported 55.22% accuracy and 46.75% macro-F1.

Table 12: Recorded implementation details for the Qwen3-VL and InternVL zero-shot visual baselines.

Setting	Qwen3-VL-8B-Instruct	InternVL3.5-8B
Nominal size	8B	8B
HF revision	`0c351dd0...ff3b`	`853e3a79...dac9`
Frames	8, linear spacing incl. endpoints	8, segment centers
Audio input	no	no
Resolution	processor default	448×448
Decoding	greedy, 768 max new tokens	greedy, 768 max new tokens

F SANITIZED PROMPT TEMPLATES

This appendix records the task-relevant portions of the production prompts used to construct supervision and run the model. Provider keys, local paths, and API metadata are omitted; the task semantics are unchanged.

Label-blind audio observer. The observer receives only the clip audio and is instructed to report temporally localized acoustic evidence (speech activity, prosody, loudness, rhythm, non-speech events, and uncertainty) without inferring the VGAF label. The words *positive*, *neutral*, and *negative* are explicitly prohibited. Its output is a structured list of evidence items with timestamps and confidence.

Label-blind visual observer. The visual observer receives the original video and is instructed to describe only directly visible group-level cues—body posture, movement, interaction, scene context, temporal changes, and uncertainty—again without the three coarse label words. The prompt emphasizes observable evidence rather than inferred intent.

Fusion annotator. The fusion stage receives the two observer outputs and the coarse VGAF supervision label. It synthesizes (i) the visible `<observe>` evidence summary, (ii) a `<think>` reasoning passage that connects evidence to group affect, (iii) the JSON `<answer>` with `coarse_label` and `open_vocabulary_emotions`, and (iv) hidden `reward_reference` fields containing `fine_emotions_ref`, evidence banks, temporal arcs, and forbidden claims. The prompt requires the final coarse label to match the provided supervision label while grounding the explanation in the label-blind observations.

Verifier. The verifier audits format validity, evidence grounding, coarse-label consistency, fine-label grounding, open-vocabulary specificity, temporal consistency, hallucination risk, redundancy risk, and SFT usability. It also flags label leakage and over-assertion patterns. This stage is used as an annotation quality gate rather than as model input at training or inference time.

SFT task prompt.

> Analyze the group emotion in this video. First summarize observable audio-visual and temporal evidence, then reason about the group affect, and finally output the answer JSON. Use the format `<observe>...  </observe> <think>...  </think> <answer>...  </answer>`.

GRPO task prompt.

> Analyze the group emotion in this video. Output `<observe>...  </observe> <think>...  </think> <answer>{...} </answer>`. The answer JSON must include `coarse_label` and `open_vocabulary_emotions`.

Zero-shot visual-baseline prompt. Qwen3-VL and InternVL are instructed to produce the same three sections and the same two JSON fields, with the legal coarse labels explicitly enumerated and the open-vocabulary emotion list constrained to 1–5 entries. For Qwen3-VL, the final formatting-only variant places line breaks between the tagged sections and explicitly requires the response to end with `</answer>`; the task semantics and answer schema are unchanged.

G Representative Dataset Annotation Cases

This appendix records real annotation-side examples from OV-VGAF. They are drawn from existing validation annotations and are intended to show the supervision signal seen during SFT and the hidden references used by the reward computation.

VGAF-OVGER-D2 (validation split) **GT: positive** temporal type: dynamic *human-verified annotation*



frame 0/124 frame 41/124 frame 82/124 frame 124/124

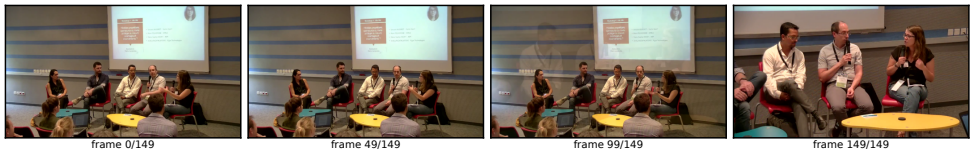
Reference fine-grained emotions
amusement, cheerful engagement, shared enthusiasm

<observe>
The audio is dominated by a high-pitched laugh delivered in quick, rhythmic bursts, with very little competing background sound. The laughter is strongest in the opening moments and then becomes less intense toward the end. Visually, a dense crowd stands around a central figure while many people hold smartphones aloft and point them inward. A person in white is escorted through the group by several men in black, and visible faces near the front show smiles and open mouths aimed toward the center of attention. [...]

<think>
The audio first suggests amusement because the dominant vocalization is unmistakable laughter rather than flat speech, shouting, or strained crying. Its rapid, staccato pattern adds lively expressiveness and social energy. The fact that this laughter begins strongly and then gradually tapers indicates an initially intense reaction that eases rather than turns tense or withdrawn. [...]

<answer>
{"coarse_label": "positive", "open_vocabulary_emotions": ["amusement", "cheerful engagement", "shared enthusiasm"]}

VGAF-OVGER-D3 (validation split) **GT: neutral** temporal type: stable *human-verified annotation*



frame 0/149 frame 49/149 frame 99/149 frame 149/149

Reference fine-grained emotions
attentiveness, composed engagement, matter-of-fact discussion

<observe>
The clip shows a formal panel discussion with one speaker delivering a continuous, brisk monologue at a steady intensity and with little variation in delivery. Visually, several panelists sit on stage in a semi-circle while one man keeps a microphone near his mouth and the others remain mostly still and attentive. Early in the clip the full panel and projection screen are visible, and as the scene progresses the camera dissolves into a tighter view of the panelists on the right. [...]

<think>
The audio suggests controlled engagement because the speaker maintains moderate vocal force throughout without noticeable shifts in loudness or tone. The rapid pace is steady rather than pressured, and the clean recording supports a structured speaking setting instead of a chaotic one. Visually, the microphone use and seated arrangement indicate an organized discussion format, while the panelists' restrained postures and limited movement point to attentive participation rather than strong outward display. [...]

<answer>
{"coarse_label": "neutral", "open_vocabulary_emotions": ["attentiveness", "composed engagement", "matter-of-fact discussion"]}

VGAF-OVGER-D4 (validation split) **GT: negative** temporal type: dynamic *human-verified annotation*



frame 0/124 frame 41/124 frame 82/124 frame 124/124

Reference fine-grained emotions
tension, confrontation, agitation

<observe>
The audio features a loud, assertive male command followed by a second male voice answering with a rapid, urgent pace, creating a forceful back-and-forth. Across the clip, the sound shifts from one dominant speaker to overlapping dialogue, keeping the exchange intense throughout. Visually, a tight cluster of men faces inward while a larger crowd watches around them. [...]

<think>
The loud, high-pitched command and the urgent reply indicate a strained verbal exchange rather than a calm discussion. When the audio shifts into overlapping dialogue, the interaction sounds more pressurized and contested. Visually, the close clustering of men, emphatic finger-pointing, and fixed face-to-face attention suggest confrontation within the group. In the middle portion, the raised phone and open-palm gesture add further activation at the center of the exchange. [...]

<answer>
{"coarse_label": "negative", "open_vocabulary_emotions": ["tension", "confrontation", "agitation"]}

Figure 4: **Additional dataset annotation examples.** Three validation examples are shown, one for each coarse class (positive, neutral, negative). Each case contains representative frames together with the ground-truth label, reference open-vocabulary emotions, and the fusion-generated, automatically verified `<observe>`/`<think>`/`<answer>` target.

Appendix Figure 4 provides one additional dataset example for each coarse class.

Positive example. **Coarse label:** positive. **Reference emotions:** animated enthusiasm, cheerful engagement, playful energy. **Evidence excerpt:** a three-person talk-show interaction with an animated speaker, shared attention, and a brief smile. **Forbidden claims:** fear or distress; hostile confrontation; sadness or withdrawal; boredom or disengagement.

Neutral example. **Coarse label:** neutral. **Reference emotions:** attentiveness, composed engagement, task-focused interaction. **Evidence excerpt:** two children remain focused on a toy stroller, one speaks continuously, and the scene remains stable and everyday. **Forbidden claims:** clear excitement; hostility; distress; sadness.

Negative example. **Coarse label:** negative. **Reference emotions:** confrontation, agitation, distress. **Evidence excerpt:** a nighttime crowd scene with aggressive motion around a person on the ground, together with explicit uncertainty from low light and motion blur. **Forbidden claims:** joy or celebration; calm neutral interaction; playful enjoyment; relaxed socializing.

H Representative Prediction Cases

The following excerpts are taken from existing validation predictions and annotation records; no additional inference is performed. They complement the aggregate results by showing how coarse labels, open-vocabulary emotions, and evidence-linked text appear in individual samples.

Appendix Figure 5 complements the textual excerpts below with rendered figure panels for multiple held-out predictions.

H.1 Case A: Positive Group Interaction

GT / prediction: positive / positive. **Reference emotions:** cheerful engagement, amusement, camaraderie. The annotation evidence highlights a talk-show setting in which the active speaker uses frequent hand gestures, the other participants maintain attention, and one participant briefly smiles.

> <observe> A seated group remains oriented toward an animated speaker; frequent hand gestures and a brief smile provide visible signs of engaged interaction, while the camera gradually focuses on the speaker. </observe>
> <think> Shared orientation, expressive delivery, and the brief smile jointly support a socially engaged rather than withdrawn or conflictual interpretation. </think>
> <answer> {"coarse_label": "positive", "open_vocabulary_emotions": ["cheerful engagement", "amusement", "camaraderie"]} </answer>

H.2 Case B: Neutral, Task-Focused Interaction

GT / prediction: neutral / neutral. **Reference emotions:** attentiveness, composed engagement, task-focused interaction. Two children remain focused on a shared toy stroller; one speaks continuously while the scene remains stable and controlled.

> <observe> The two children remain close to the stroller and sustain attention toward one another and the shared object, with no strongly valenced event dominating the clip. </observe>
> <think> Continued social focus is present, but the interaction remains controlled and matter-of-fact rather than excited or distressed. </think>
> <answer> {"coarse_label": "neutral", "open_vocabulary_emotions": ["composed engagement", "attentiveness", "matter-of-fact interaction"]} </answer>

H.3 Case C: Negative Interaction under Visual Uncertainty

GT / prediction: negative / negative. **Reference emotions:** confrontation, agitation, distress. The evidence bank records a nighttime group scene, a person on the ground, rapid overlapping movement, and low-light/motion-blur uncertainty.

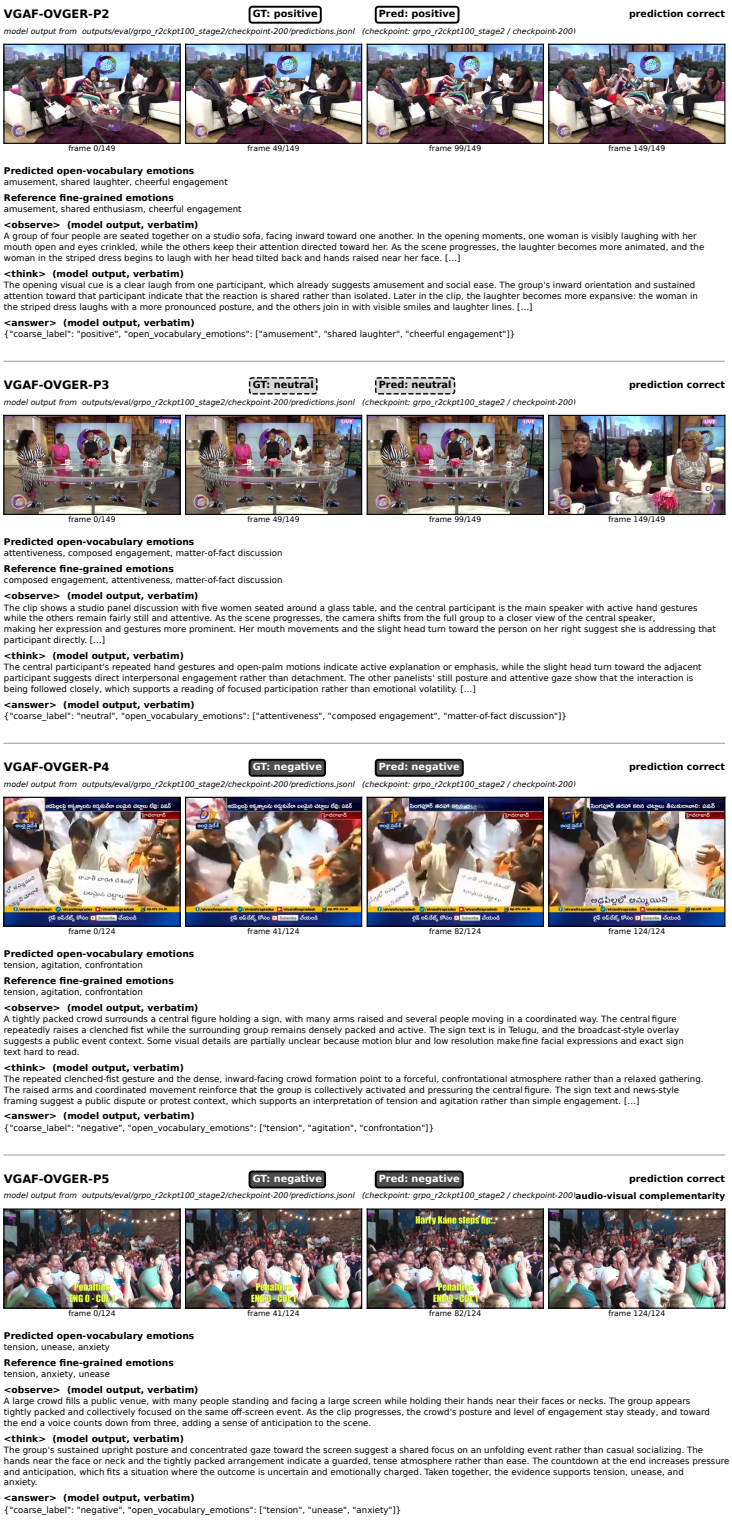


Figure 5: **Additional prediction examples.** Three class-balanced held-out examples and one extra audio–visual-complementarity example are shown with frames, ground-truth labels, model predictions, predicted and reference open-vocabulary emotions, and the model's structured outputs. All examples are taken verbatim from the recorded validation predictions.

<observe> A group gathers near a storefront at night; despite low light and camera shake, visible aggressive physical actions occur around a person on the ground. </observe>
<think> The sustained physical aggression and crowd behavior provide direct negative evidence, while the uncertainty statement limits claims about fine facial details or exact contact. </think>
<answer> {"coarse_label": "negative", "open_vocabulary_emotions": ["tension", "hostility", "agitation"]} </answer>

H.4 Case D: Audio–Visual Complementarity

GT / prediction: negative / negative. **Reference emotions:** tension, confrontation, agitation. In this clip the recorded audio temporal arc is stable, whereas the visual/affective arc moves from a dense physical standoff to a more ordered but still tense public accusation. The final model outputs the same three open-vocabulary reference concepts.

<observe> A tightly packed group in a formal chamber shows inward-leaning postures and raised arms; the scene later shifts to a speaker at a podium while surrounding participants remain watchful. </observe>
<think> The visual transition from crowded confrontation to a focused accusation supplies the dominant temporal signal, complementing the comparatively stable audio arc. </think>

I Additional Failure and Baseline Cases

The final validation predictions contain 194 coarse-label errors. Table 13 summarizes four representative examples spanning socially ambiguous behavior, positive over-interpretation of an uncertain neutral scene, confusion between concentration and unease, and underestimation of positive affect intensity.

Table 13: Representative failure cases from the final validation predictions.

Clip	GT → Pred.	Predicted open-vocabulary emotions	Error pattern
104_4.mp4	negative → positive	attentiveness; animated engagement; warm interaction	Vocal/interaction intensity is interpreted as socially warm rather than conflictual.
104_3.mp4	neutral → positive	playful engagement; cheerful involvement; camaraderie	The observation acknowledges ambiguity, but the interpretation resolves it too strongly toward positive affect.
125_5.mp4	neutral → negative	attentiveness; concentration; unease	Concentration and task focus are over-interpreted as strain or unease.
181_5.mp4	positive → neutral	attentiveness; composed engagement; matter-of-fact interaction	Positive affect intensity is understated with neutral descriptors.

Across all 194 errors, the recorded direction counts are 57 neutral→negative, 46 positive→neutral, 43 neutral→positive, 22 positive→negative, 17 negative→positive, and 9 negative→neutral.

For 100_1.mp4, recorded outputs are available for OVGER, Qwen3-VL, InternVL, and GLM-4.6V-Flash. OVGER supplies the complete observe-think-answer structure with group-interaction and temporal cues; InternVL supplies the three sections but with shorter reasoning and more bystander-focused description; Qwen3-VL's recorded output is observation-heavy; and GLM-4.6V-Flash begins with <think> rather than the requested <observe> section. This same-clip comparison is qualitative and does not introduce additional inference.