Beyond Plausibility: Diagnosing Rationalization in VLM Emotion Explanations

Anonymous ACL submission

Abstract

Vision-language models (VLMs) are increasingly expected to justify emotion predictions in natural language. We ask whether these explanations support the ground-truth affective label or rationalize the model's own prediction. We formulate **affective post-hoc rationalization** as an output-level diagnostic problem and introduce the **Emotion Explanation Faithfulness (EEF)** framework, which decomposes explanation quality into prediction-conditioned consistency, truth-conditioned consistency, evidence direction via the Rationalization Gap (RG), evidence quantity, and image-conditioned plausibility.

Across three VLMs, two visual emotion benchmarks, and three prompting strategies (7,815 triples), incorrect predictions favor the wrong label in **79–88%** of cases ($r = 0.62$–0.79, all $p < 0.001$). In this label-aware evaluation setting, RG detects errors more strongly (AUROC 0.852) than label-free evidence-quantity proxies such as EDL-inspired uncertainty (0.511) and softmax-style confidence (0.559). This suggests that rationalized errors are better diagnosed by prediction–truth support asymmetry than by evidence amount alone. The effect persists after emotion-label debiasing ($r_s = 0.495$, $p < 0.001$), across NLI scorer and hypothesis-template robustness checks, and across prompt subsets. A small-scale four-annotator sanity check further shows that RG aligns with human judgments of label-support direction ($\rho = 0.543$, $p < 10^{-8}$). Chain-of-thought prompting is not a reliable mitigation, revealing a decoupling among explanation consistency, output-format stability, and prediction reliability. These findings show that VLM emotion explanations often fail not by lacking plausible evidence, but by directing that evidence toward the wrong label.

1 Introduction

Explaining why an emotion is inferred from a visual scene is qualitatively different from classifying it. As VLMs are deployed in affectively sensitive contexts—mental health monitoring, social media analysis, human-robot interaction—explanation quality matters as much as accuracy. Yet a fundamental question has gone unasked: when a VLM generates an emotion explanation, does it provide evidence aligned with the ground-truth emotion, or construct a plausible-sounding justification aligned with its own prediction?

This distinction defines **affective post-hoc rationalization**: the tendency of VLM emotion explanations to support the model's predicted label more strongly than the ground truth, especially when the prediction is incorrect. We use rationalization as an *output-level diagnostic*: we do not claim to trace the model's internal causal computation, but measure whether the generated explanation supports the prediction over the ground truth. We distinguish this from hallucination (Li et al., 2023; Guan et al., 2024): a model may correctly describe visual elements yet rationalize the wrong emotion by misattributing their affective significance—a subtler failure mode not captured by existing metrics. Our goal is not to replace human judgment of explanation quality, but to isolate a narrower diagnostic question: whether the explanation's label support points toward the prediction or the ground truth. This matters because fluent explanations can appear plausible even when they justify an incorrect label.

Existing frameworks evaluate explanation quality by textual similarity to annotated descriptions (Lian et al., 2024; Cheng et al., 2024), rewarding resemblance to annotations rather than support for the true emotion. Evidential Deep Learning (Sensoy et al., 2018) motivates uncertainty signals based on the *amount* of evidence across categories. In our setting, we use an EDL-inspired transformation of all-label NLI support as a label-free evidence-quantity proxy. By contrast, RG is a label-aware evaluation-time diagnostic that com-

pares support for the prediction against support for the ground truth. Our experiments show that this prediction–truth support asymmetry is highly informative for diagnosing rationalized errors (AUROC = 0.852), whereas evidence-quantity proxies are much weaker in this diagnostic setting (EDL-u AUROC = 0.511).

Research questions. We organize our analysis around three questions:

> **RQ1.** Do VLM explanations for incorrect emotion predictions support the model's wrong label more than the ground truth?
>
> **RQ2.** In a label-aware evaluation-time setting, how informative is RG compared with label-free evidence-quantity proxies such as EDL-inspired uncertainty and softmax-style confidence?
>
> **RQ3.** Do prompting strategies improve prediction reliability, or merely alter explanation consistency and output-format stability?

We make four contributions:

1. We formulate **affective post-hoc rationalization** as a new output-level diagnostic problem for VLM emotion explanations.

2. We propose **EEF** as a diagnostic decomposition of emotion explanation quality, separating prediction support, truth support, evidence direction, evidence quantity, and image-conditioned plausibility.

3. We show that rationalization is **prediction-specific** rather than generic semantic drift: explanations over-support the model's own wrong label more than random non-truth labels. In a label-aware evaluation-time setting, RG separates incorrect from correct predictions more strongly than label-free evidence-quantity proxies (AUROC 0.852 vs. 0.511/0.559), and the pattern persists under debiasing, scorer/template robustness checks, prompt subsets, and a small-scale human directional-support sanity check ($\rho = 0.543$, $p < 10^{-8}$).

4. We show that **generic CoT prompting is not a reliable mitigation**: it can increase prediction consistency without improving correctness, and in some models destabilizes output format.

2 Related Work

Visual Emotion Analysis. VEA benchmarks span binary sentiment (Katsurai and Satoh, 2016), Ekman's six basic emotions (Peng et al., 2015), and Mikels' eight-category taxonomy (You et al., 2016; Yang et al., 2023). Methods range from attention-based classifiers (Yang et al., 2018; Zhao et al., 2019; Xu et al., 2022) to instruction-tuned VLMs (Xie et al., 2024). Recent visual emotion analysis work increasingly emphasizes expressive intention and affective context; our work asks whether VLM explanations actually capture such grounding.

Explainable Emotion Recognition. EMER (Lian et al., 2024) and Emotion-LLaMA (Cheng et al., 2024) evaluate explanations by clue overlap and label overlap with annotated descriptions, rewarding explanations that resemble human-written rationales rather than those that support the true emotion. Crucially, prior multimodal emotion models can capture correct visual features while still assigning them to the wrong affective interpretation: Cheng et al. (2024) report that models such as PandaGPT and Valley identify relevant visual cues yet fail to integrate cross-modal context, yielding incorrect emotion labels with plausible-sounding visual descriptions. Such overlap-based metrics reward this pattern—they cannot distinguish whether an explanation supports the ground-truth label or merely aligns with the model's prediction. High overlap can be achieved by label echo or generic clichés, the dominant rationalization patterns our error analysis identifies.

Emotion Confusion and VLM Failure Modes. Psychological studies show that emotion cues are not label-specific: fear is often confused with surprise, and disgust with anger, due to overlapping facial configurations (Duran and Atkinson, 2021). In computational settings, Bhattacharyya and Wang (2025) similarly document systematic confusions in VLM-based emotion recognition, showing that plausible surface-level cues can bias models toward unintended affective labels. However, these works do not examine whether the *natural-language explanations* accompanying incorrect predictions support the wrong predicted label over the ground truth. Prior work explains why emotion errors occur; we ask whether explanation support follows the wrong prediction rather than the ground truth.

Faithfulness in NLP and VLMs. Faithfulness concerns whether explanations reflect a model's

actual decision process rather than merely offering plausible post-hoc justifications (Jacovi and Goldberg, 2020; Lyu et al., 2024). Prior work examines this through counterfactual interventions (Ding et al., 2025), perturbation-based tests of visual grounding, and analyses of model-internal structures (Belinkov, 2021). These methods ask whether an explanation is causally tied to the visual input or the model's computation. Our setting is complementary: EEF/RG does not establish causal visual faithfulness, but provides a scalable output-level diagnostic of whether an explanation's *label support* is directed toward the model's prediction or the ground truth. VLM sycophancy studies similarly show that models can follow textual cues despite conflicting visual evidence (Li et al., 2024b); here, however, no external false label is given, and the rationalization target is the model's *own* emotion prediction.

Evidential Deep Learning. EDL (Sensoy et al., 2018) models uncertainty through Dirichlet evidence, with later extensions to regression (Amini et al., 2020) and OOD detection (Bao et al., 2021). In our experiments, we use an EDL-inspired transformation of all-label NLI support as a quantity-oriented contrast: it summarizes the amount or concentration of support across labels, but does not compare support for the prediction against support for the ground truth. RG instead measures prediction–truth support direction, so we treat EDL-u as an evidence-quantity proxy rather than a competing visual-faithfulness metric.

3 The EEF Framework

3.1 Problem Formulation

Let $\mathcal{M}$ be a VLM, $\mathbf{x}$ an image, y the ground-truth emotion, $\hat{y}$ the predicted label, and e the generated explanation. **Affective post-hoc rationalization** is the tendency of e to be more semantically aligned with $\hat{y}$ than with y, especially when $\hat{y} \neq y$. The full pipeline is in Appendix B.

3.2 EEF Metrics

CCS_{pred} and $\text{CCS}_{\text{truth}}$. Using DeBERTa-v3-large (He et al., 2021):

$$\text{CCS}_{\text{pred}} = P_{\text{NLI}}(\text{ENTAIL} \mid e,\ h_{\hat{y}}) \quad (1)$$

$$\text{CCS}_{\text{truth}} = P_{\text{NLI}}(\text{ENTAIL} \mid e,\ h_{y}) \quad (2)$$

where h_y = "The primary emotion is y." Both are *text-only*: they assess semantic alignment, not visual grounding.

Signal	What it asks	Limitation
Softmax-style confidence	Is support concentrated on one label?	Cannot tell whether that label is correct
EDL-inspired uncertainty	Is there enough total evidence?	Misses wrong-directed evidence
Rationalization Gap	Does evidence favor prediction over truth?	Requires ground-truth label

Table 1: Conceptual distinction among confidence, evidential uncertainty, and RG. Softmax-style confidence and EDL-u measure evidence concentration or amount; RG measures evidence direction.

Rationalization Gap (RG).

$$\text{RG} = \text{CCS}_{\text{pred}} - \text{CCS}_{\text{truth}} \quad (3)$$

RG measures how much more the explanation supports the model's prediction than the ground-truth label. On incorrect predictions, RG > 0 indicates prediction-aligned rationalization. RG should therefore be read as a measure of *explanation-label support asymmetry*, not as direct evidence of the model's internal causal computation or full visual grounding.

All-label Support and EDL-inspired Uncertainty. We compute $s_k = P_{\text{NLI}}(\text{ENTAIL} \mid e, h_k)$ for all $k \in \mathcal{L}$. Softmax-style confidence: Conf $= \max_k \text{softmax}(s_k)$. EDL-inspired uncertainty ($c = 1$; robustness: Appendix I):

$$\alpha_k = c{\cdot}\max(s_k, 0){+}1, \quad S = \textstyle\sum_k \alpha_k, \quad u = K/S \quad (4)$$

This is an EDL-inspired transformation of all-label NLI support profiles, not a separately trained evidential model; we use it only as an evidence-quantity baseline against RG.

Image-conditioned Plausibility (PF). GPT-4o rates Q_1 (Visual Presence, 1–5) and Q_2 (Affective Relevance, 1–5) given image and explanation; PF $= (Q_1 + Q_2)/2$. Validated with image-conditioned, text-only, and mismatched-image controls (§4.7).

3.3 Datasets and Models

We evaluate on **FI** (You et al., 2016) (600 images, Mikels 8-class) and **Emotion6** (Peng et al., 2015) (300 images, Ekman 6-class), giving 900 unique images after deduplication. Models: LLaVA-1.6 (Liu et al., 2024a,b), LLaVA-OV (Li et al., 2024a), and Qwen2-VL (Wang et al., 2024), all 7B-class and run in float16 on NVIDIA RTX A6000. The

Component	Setting
Datasets	FI (600), Emotion6 (300)
Models	LLaVA-1.6, LLaVA-OV, Qwen2-VL
Prompts	Standard, CoT, Few-shot
Raw outputs	$900 \times 3 \times 3 = 8{,}100$
Final triples	7,815 after filtering
Main scorer	DeBERTa-v3-large NLI
Robustness checks	DeBERTa-v3-small, RoBERTa-base NLI, hypothesis templates
Primary comparison	RG vs. EDL-u vs. softmax-style confidence

Table 2: Experimental setup summary.

LLaVA-1.6/OV pair contrasts earlier vs. more recent LLaVA-style models; LLaVA-OV and Qwen2-VL share a Qwen-family backbone but differ in multimodal design, allowing us to examine whether rationalization persists across related architectures.

Three prompting strategies are applied: Standard, CoT, and Few-shot (Appendix A). After removing 285 outputs with exemplar contamination or grossly invalid label fields, we retain **7,815** generated triples (Appendix D).

Prediction label handling. Because VLMs generate free-form labels, we separate parsing from semantic scoring. Analyses involving RG exclude rows where CCS_{pred} or RG is undefined; we never impute undefined RG in the primary results. For all-label support analyses, we score each explanation against the canonical emotion labels in the dataset taxonomy. Controls requiring a canonical predicted label, such as the prediction-specific analysis in §4.5, are computed on the canonical-label subset only. This prevents open-vocabulary label variants from being conflated with rationalization scores. Prompt-wise parsing and empty-label behavior are reported separately in Appendix D.1.

4 Results

We report results in five steps:

RQ1: incorrect predictions systematically show positive RG.

RQ2: label-aware RG captures errors missed by label-free evidence-quantity proxies.

Validity: the effect is robust to debiasing, scorer and template variation, thresholding, prompt subsets, prediction-specific controls, prompt-format audits, and a small-scale human directional-support sanity check.

RQ3: prompting changes explanation consistency and output-format stability without reliable accuracy gains.

Model	CCS_p	CCS_t	RG	Acc.
LLaVA-1.6	0.527	0.215	$0.312_{\pm 0.450}$	0.209
LLaVA-OV	0.538	0.292	$0.246_{\pm 0.430}$	0.281
Qwen2-VL	0.557	0.326	$0.231_{\pm 0.435}$	0.250

Table 3: Mean EEF metrics and accuracy (all prompts pooled). Pairwise model differences in RG are significant after Bonferroni correction ($p < 0.001$). RG standard deviations are shown because RG is the signed gap used for rationalization analysis; CCS_p and CCS_t are also high-variance. We interpret rationalization magnitude primarily through error-conditioned means and confidence intervals in Table 4.

Model	RG_{err} [95% CI]	$\text{CCS}_p > \text{CCS}_t$	RG>0.8	r
LLaVA-1.6	$0.395_{[0.374, 0.416]}$	88.2%	36.5%	0.788
LLaVA-OV	$0.342_{[0.316, 0.367]}$	83.3%	20.6%	0.693
Qwen2-VL	$0.308_{[0.286, 0.329]}$	79.5%	30.9%	0.621

Table 4: Rationalization on incorrect predictions. RG_{err}: mean RG on errors; $\text{CCS}_p > \text{CCS}_t$: proportion favoring wrong label; RG>0.8: extreme rationalization ($n = 1{,}786$; threshold sensitivity: Appendix H); r: rank-biserial vs. correct predictions (all $p < 0.001$). Correct predictions: median RG ≈ 0.

Analysis: qualitative cases show plausible cues being directed toward the wrong label.

4.1 Model Overview

CCS_{pred} exceeds $\text{CCS}_{\text{truth}}$ by 0.231–0.312 across all models, before conditioning on prediction correctness. Absolute accuracies are modest because we evaluate zero-shot generative emotion predictions under fine-grained affective taxonomies; our goal is not to optimize emotion recognition accuracy, but to audit how models explain the predictions they actually make.

4.2 Incorrect Predictions Systematically Rationalize

On incorrect predictions, explanations favor the wrong label in **79–88%** of cases (Table 4). Extreme rationalization ($n = 1{,}786$) affects 20.6–36.5% of errors. Figure 1 shows the rationalization space: incorrect predictions concentrate above the RG $= 0$ diagonal, while correct predictions cluster near it.

This distinction matters because rationalization is not the same as ordinary classification failure. A model can be wrong without producing a prediction-aligned explanation; RG captures cases where the error is accompanied by an explanation that semantically supports the mistaken label. Appendix Figure 6 further shows that this separation

Metric	AUROC	AUPR	r_s	p
1−Softmax conf.	0.559	0.807	0.083	< 0.001
EDL-insp. uncert.	0.511	0.804	0.016	0.160
Ration. Gap (RG)	**0.852**	**0.958**	**0.527**	< 0.001
RG (debiased)	—	—	0.495	< 0.001

Table 5: Error detection (positive class = incorrect prediction). Conf. and EDL-u use all-label NLI profiles, not VLM logits. For confidence, we use $1 - \text{Conf}$ so that larger values indicate lower confidence and higher expected error. Because incorrect predictions are the majority class, random-ranking AUPR is approximately the positive-class rate (≈ 0.75), so $1 - \text{Conf}$ (0.807) and EDL-u (0.804) are near baseline, while RG reaches 0.958. EDL-inspired uncertainty has near-chance AUROC (0.511) and no significant monotonic association with error ($r_s = 0.016$, $p = 0.160$). Scorer and template robustness results: Appendix F.

Model	Mean RG		$\text{CCS}_p > \text{CCS}_t$ (incorr.)	
	Orig.	Debias.	Orig.	Debias.
LLaVA-1.6	0.312	0.281	88.2%	87.8%
LLaVA-OV	0.246	0.158	83.3%	81.3%
Qwen2-VL	0.231	0.176	79.5%	79.0%

Table 6: Original vs. debiased RG (emotion-label words removed before NLI scoring; Appendix E). Pattern fully preserved; debiased RG $r_s = 0.495$ with error, $p < 0.001$.

Concern	Control	Result
Label echo	Debiased RG	$r_s = 0.495$, $p < 0.001$
NLI scorer / template	DeBERTa-small, RoBERTa, templates	AUROC 0.772–0.862
Threshold choice	RG > 0.5–0.9	Ordering stable
Argmax dominance	All-label profile	Pred argmax 5–8%
Verbosity	Length vs. RG	$r_s = 0.012$–0.108
Prompt subsets + format audit	Subset AUROC; model parse audit	Std-only 0.842; LLaVA-OV CoT format collapse
Human alignment	4-ann. direction check	$\rho = 0.543$; 65.0% vs. 29.0%

Table 7: Validity controls summary. Each row addresses a distinct alternative explanation for the observed rationalization pattern.

is distributional rather than driven by a few outliers.

4.3 RG Provides a Strong Label-Aware Error Diagnostic

EDL-inspired uncertainty (AUROC 0.511) is near chance in this diagnostic: many rationalized explanations distribute NLI support broadly across categories, making them appear uncertain or evidentially balanced, while still showing a marginal directional bias toward the wrong label. For directional consistency with error detection, we report confidence as $1 - \text{Conf}$; 1−Softmax-style confidence (AUROC 0.559; Table 5) also provides only a weak label-free signal. By explicitly comparing prediction support against truth support, RG provides a strong label-aware error diagnostic (AUROC 0.852).

Why evidence quantity is insufficient here. Mean EDL-u is highly overlapping between correct and incorrect predictions: 0.927 ± 0.093 for correct vs. 0.918 ± 0.112 for incorrect. Thus, incorrect explanations are not simply evidence-poor. They can appear comparably evidence-supported while directing support toward the wrong label. All-label support confirms this: the predicted label is the argmax in only 5–8% of incorrect cases, below nominal chance (Appendix G). Rationalization therefore operates as marginal directional bias, which an evidence-quantity proxy is not designed to capture.

Confident rationalization. Figure 2 plots evidence quantity (u) vs. direction (RG). The top-left quadrant (low-u, high-RG) captures cases where explanations appear evidentially concentrated yet bias support toward the wrong label. This *confident rationalization* regime contains 23.9% of incorrect predictions ($n = 1{,}462$); EDL-u alone would not distinguish these cases in this evaluation setting.

4.4 Validity Controls Against Alternative Explanations

After removing emotion-label words (Table 6), the rationalization pattern is fully preserved. All-label support further shows rationalization operates as a marginal directional bias rather than argmax dominance (Appendix G). RG remains stronger than EDL-u and softmax-style confidence under an alternative DeBERTa scorer, a cross-family RoBERTa-base NLI scorer, and three hypothesis-template variants (Appendix F); absolute AUROC varies (0.772–0.862), but the directional conclusion is preserved.

To address whether low CoT accuracy inflates RG, we repeat error detection by prompt subset, excluding undefined RG rather than imputing it as zero (Table 8). RG remains strong under Standard-only (AUROC 0.842), Standard+Few-shot (0.843), and CoT-only (0.871), indicating that the finding is not driven solely by CoT collapse.

A prompt-wise format audit further shows that CoT failures have different mechanisms across models: LLaVA-1.6 and Qwen2-VL mostly produce parseable but unreliable outputs (RG NaN

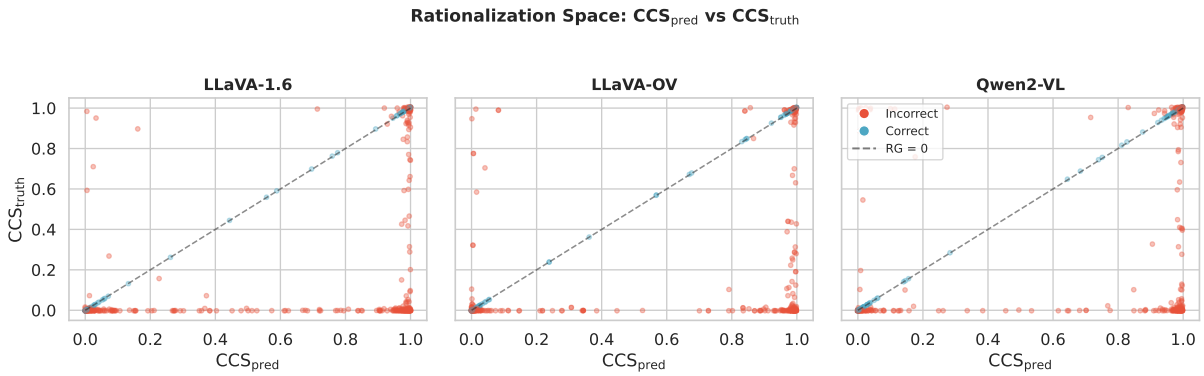

Figure 1: Rationalization space: CCS_{pred} vs. $\text{CCS}_{\text{truth}}$ per model. Both axes are text-only NLI scores. Incorrect predictions (red) concentrate above the RG $= 0$ diagonal; correct predictions (blue) cluster near it. Consistent spatial separation across all three models is the primary evidence for systematic affective post-hoc rationalization.

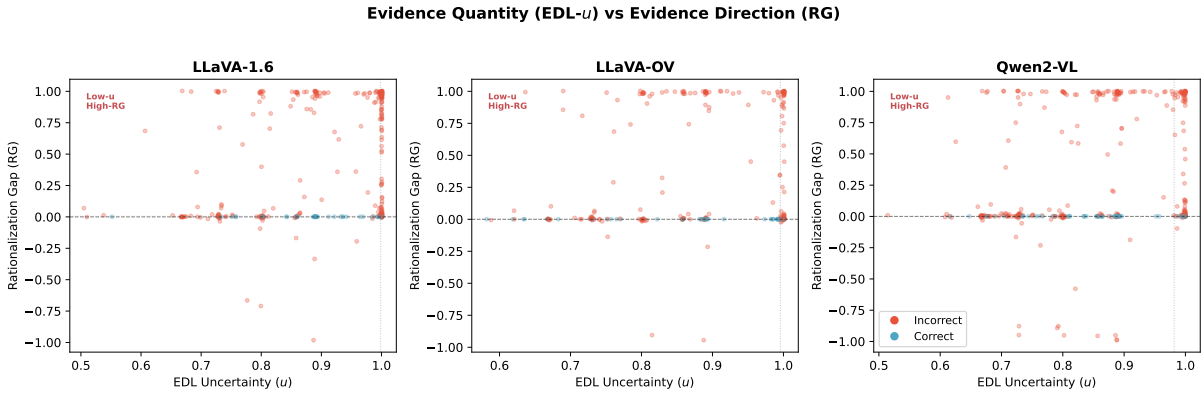

Figure 2: Evidence quantity (EDL-u) vs. evidence direction (RG) per model. Top-left quadrant: *confident rationalization* (23.9% of incorrect predictions, $n = 1{,}462$) — explanations that appear evidentially concentrated yet bias support toward the wrong label. EDL-u alone would not distinguish these cases in this evaluation setting. Correct predictions (blue) cluster near the center; incorrect (red) spread rightward (high RG) and toward the top-left quadrant.

rate: 2.8% and 0% respectively), whereas LLaVA-OV exhibits severe output-format collapse: 89.1% empty predictions, an additional 2.5% literal "[label]" placeholders, and 89.5% undefined RG under the metric parser (Appendix D.1). Together with the subset analysis, this audit indicates that the main rationalization pattern is not explained by CoT formatting artifacts alone: parseable CoT outputs in LLaVA-1.6 and Qwen2-VL can still be inaccurate and prediction-aligned.

We also conduct a small-scale human directional-support sanity check. Four anonymous external annotators judged 100 examples using only the explanation text and two randomized emotion labels. They did not see the image, model identity, ground truth, prediction, or RG score. Human pred-support rate correlates with RG ($\rho = 0.543$, $p = 5.3 \times 10^{-9}$; 95% CI $[0.392, 0.671]$). On high-RG incorrect examples, annotators selected the model-predicted label in 65.0% of judgments, compared with 29.0% on low-RG incorrect controls.

Subset	N	$\text{CCS}_p > \text{CCS}_t$	AUROC
All prompts	6,953	83.8%	0.852
Standard only	2,623	83.0%	0.842
Standard+Few-shot	5,095	82.9%	0.843
CoT only	1,858	86.0%	0.871

Table 8: Prompt-subset robustness. AUROC uses incorrect prediction as the positive class; rows with undefined RG are excluded rather than imputed. RG remains strong across all prompt subsets, confirming that the main rationalization pattern is not driven by CoT accuracy collapse. Detailed excluded-row composition is reported in Appendix D.1.

4.5 Rationalization Is Prediction-Specific

The RG result shows explanations support the predicted label more than the true label on incorrect

Model	n	$\mathrm{RG}_{\mathrm{pred}}$	$\mathrm{RG}_{\mathrm{rand}}$	Δ
Overall	809	+0.189	−0.131	+0.321
LLaVA-1.6	276	+0.201	−0.140	+0.341
LLaVA-OV	191	+0.239	−0.148	+0.386
Qwen2-VL	342	+0.152	−0.115	+0.268

Table 9: Prediction-specific rationalization control on incorrect predictions with canonical predicted labels (n = canonical-label subset per model; all Wilcoxon $p < 0.001$, one-sided). $\mathrm{RG}_{\mathrm{pred}}$: support gap between the predicted and true label. $\mathrm{RG}_{\mathrm{rand}}$: gap between mean non-truth/non-pred label support and truth. $\Delta = \mathrm{RG}_{\mathrm{pred}} - \mathrm{RG}_{\mathrm{rand}}$.

Model	Prompt	CCS_p	CCS_t	RG	Acc.
LLaVA-1.6	Standard	0.539	0.272	0.267	**0.325**
	CoT	0.619	0.176	**0.444**	0.087
	Few-shot	0.421	0.199	0.222	0.220
LLaVA-OV	Standard	0.665	0.382	0.283	**0.293**
	CoT	0.323	0.169	0.172	0.031
	Few-shot	0.426	0.212	0.214	0.268
Qwen2-VL	Standard	0.546	0.322	0.224	0.265
	CoT	0.659	0.358	0.301	0.207
	Few-shot	0.453	0.295	**0.158**	**0.283**

Table 10: EEF metrics and accuracy by model and prompting strategy.

predictions. However, this could reflect generic semantic drift rather than a specific bias toward the model's own prediction. We test this by comparing two gaps on incorrect predictions:

$$\mathrm{RG}_{\mathrm{pred}} = \mathrm{CCS}_{\mathrm{pred}} - \mathrm{CCS}_{\mathrm{truth}},$$
$$\mathrm{RG}_{\mathrm{rand}} = \bar{s}_{\mathrm{rand}} - \mathrm{CCS}_{\mathrm{truth}},$$

where $\bar{s}_{\mathrm{rand}}$ is the mean NLI support over all labels excluding the true and predicted labels. This control is computed on the canonical-label subset, where the model's predicted label belongs to the dataset taxonomy; open-vocabulary outputs that cannot be mapped to a canonical emotion are excluded only for this control. The canonical-label subset is relatively small ($n = 809$ among 5,262 incorrect predictions with defined RG) because many outputs—particularly CoT and few-shot responses—produce open-vocabulary labels (e.g., "serene", "satisfaction") or template-literal outputs that cannot be matched to a taxonomy emotion; these are excluded only here. The subset is not solely driven by the Standard prompt (Standard/CoT/Few-shot: 367/120/322 cases). A Standard-only restriction preserves the same sign pattern overall and within each model (Appendix D.2), reducing concern that the prediction-specific effect is an artifact of prompt composition.

As shown in Table 9, $\mathrm{RG}_{\mathrm{pred}}$ is positive for all models (+0.152 to +0.239; overall +0.189), whereas $\mathrm{RG}_{\mathrm{rand}}$ is consistently negative (−0.115 to −0.148; overall −0.131). The gap remains large across models ($\Delta = +0.268$ to +0.386; all Wilcoxon tests $p < 0.001$). Crucially, negative $\mathrm{RG}_{\mathrm{rand}}$ means that random wrong labels receive *less* NLI support than the true label on average. Thus, explanations do not merely drift away from the ground truth; they specifically over-support the model's own predicted label. Affective rationalization is **prediction-specific**, not generic semantic drift.

4.6 CoT Is Not a Reliable Mitigation

CoT affects models differently and in no case provides reliable accuracy improvement. For LLaVA-1.6, CoT raises $\mathrm{CCS}_{\mathrm{pred}}$ (+0.080) while accuracy drops sharply (0.325 → 0.087) and RG widens to 0.444: the model produces explanations that align more strongly with its own prediction while being more frequently wrong. For Qwen2-VL, CoT similarly raises $\mathrm{CCS}_{\mathrm{pred}}$ (+0.113) without accuracy gain. For LLaVA-OV, the failure mode is qualitatively different: CoT causes output-format collapse, with 89.1% empty predictions and an additional 2.5% literal "[label]" placeholders; 89.5% of CoT rows have undefined RG under the metric parser (Appendix D.1). The CoT accuracy of 0.031 for LLaVA-OV therefore reflects predominantly format non-compliance rather than reasoning failure.

CoT can either amplify rationalization or destabilize output format, depending on the model. **Explanation consistency, format stability, and prediction reliability should therefore be evaluated separately.**

4.7 Image-Conditioned Plausibility

GPT-4o judge controls show PF is image-sensitive but partially text-driven: mismatched-image PF (1.08) is far below image-conditioned PF (3.66, $p < 0.001$), while text-only PF (3.44) is nearly as high. Image-conditioned PF adds no significant accuracy variance beyond text-only ($r_{\mathrm{partial}} = 0.038$, $p = 0.64$). We treat PF as a plausibility-side check rather than the primary faithfulness metric; the human sanity check in Appendix P separately tests whether RG aligns with human judgments of label-support direction. Full PF details are in Appendix J.

Pattern	GT / Pred	Evidence excerpt	RG
High-RG — incorrect			
Label echo	anger / joy	"Open mouth, showing teeth, eye crinkle suggest a **joyful** expression."	0.998
Generic cue shortcut	disgust / joy	"Children are smiling, holding a toy, happy and playful mood."	0.998
Intention mismatch	excitement / awe	"Vast landscape evokes **awe**; grandeur overwhelms."	0.93
Low-RG — correct			
Grounded	amus. / amus.	"Crinkled eyes, open mouth: unguarded laughter."	0.02
Grounded	sad. / sad.	"Downcast eyes, hunched posture: inward withdrawal."	0.04

Table 11: Qualitative examples by error pattern. High-RG errors use plausible visual cues but direct them toward the predicted label rather than the ground-truth emotion. Low-RG correct cases provide contrast. Examples are lightly truncated, not rewritten.

4.8 Qualitative Analysis

Table 11 illustrates three high-RG error patterns: label echo, generic cue shortcut, and intention mismatch. All three produce plausible-sounding explanations, but the evidence points toward the predicted label rather than the ground truth.

Across these cases, high-RG failures are not merely perceptual absences: the model often describes plausible visual content, but maps its affective meaning toward the predicted label rather than the ground-truth emotion. This supports our central interpretation that rationalization is a misdirection of plausible evidence.

5 Discussion

Evidence direction vs. quantity. Affective rationalization is a label-support direction problem. In our evaluation-time setting, rationalized explanations can distribute NLI support broadly across emotion categories (EDL-u AUROC 0.511) while maintaining a marginal prediction-over-truth bias (RG AUROC 0.852). RG is deliberately simple: a scalable and interpretable label-aware diagnostic of prediction–truth support asymmetry, not a deployment-time uncertainty estimator or a causal visual-faithfulness test.

Implications for evaluation. Systems evaluated by clue overlap, label overlap, plausibility, or fluency may reward articulate rationalization of errors. Human or LLM-based ratings are useful, but can emphasize overall coherence rather than the direction of label support. RG isolates this narrower question by comparing support for the prediction against support for the ground truth. VLM emotion explanations should therefore be evaluated with two separate checks: whether the explanation is plausible, and whether its evidence points to the ground-truth label rather than the model's prediction. Appendix K shows RG also improves evaluation-time risk-coverage curves, though it requires ground-truth labels.

Implications for mitigation. Our prompting results should be read as a *negative mitigation result*: generic CoT does not reliably improve accuracy and can either increase prediction consistency without improving correctness (LLaVA-1.6, Qwen2-VL) or destabilize output format entirely (LLaVA-OV). This suggests that making explanations longer or more step-by-step is insufficient. A more principled use of RG is evaluation-time auditing: in validation settings with ground-truth labels, high-RG cases can be flagged for dataset, prompt, or model-level inspection. Deployment-time mitigation requires label-free proxies such as all-label support entropy, which we leave for future validation.

6 Conclusion

We introduced affective post-hoc rationalization as an output-level diagnostic for VLM emotion explanations and proposed EEF to separate prediction support, truth support, evidence direction, evidence quantity, and image-conditioned plausibility. Across three VLMs and two emotion benchmarks, incorrect explanations favor the model's wrong label in 79–88% of errors, and this bias persists after label-word debiasing, across NLI scorer and hypothesis-template robustness checks, and in a small-scale human directional-support sanity check. Prediction-specific controls show that explanations do not merely drift away from the ground truth; they over-support the model's own wrong label more than random non-truth labels on average. In this label-aware evaluation setting, RG captures this pattern more strongly than label-free confidence and evidential-quantity proxies (AUROC 0.852 vs 0.559/0.511), because it directly compares support for the model's prediction against support for the ground truth. Our results suggest that explanation-based evaluation in affective AI must audit not only whether explanations are fluent or evidence-rich, but *where* their evidence points.

Limitations

CCS and RG rely on NLI for semantic alignment; they do not directly verify image-explanation correspondence via intervention or masking, and should be understood as output-level diagnostics. We verify that RG remains stronger than EDL-u and softmax-style confidence under an alternative DeBERTa scorer (DeBERTa-small: AUROC 0.862), a cross-family RoBERTa-base NLI scorer (AUROC 0.772), and three hypothesis-template variants (AUROC 0.784–0.838). However, absolute AUROC varies across scorers and templates, so further robustness across additional semantic alignment models remains necessary. PF is image-conditioned but partially text-driven (§4.7). Evaluation covers 7B open-source VLMs on two English-language benchmarks; rationalization may differ for closed models and cross-lingual datasets. Emotion annotation is inherently subjective; boundary-case disagreement may inflate apparent rationalization near category boundaries. CoT results partly reflect prompt-format sensitivity, particularly for LLaVA-OV, where most CoT outputs are empty, template-literal, or yield undefined RG; they should not be interpreted solely as reasoning quality. Although our four-annotator sanity check shows that RG aligns with human directional-support judgments, it is not a substitute for a full human evaluation of visual explanation quality; inter-annotator agreement is modest, and larger and more diverse human studies remain necessary. Because RG requires ground-truth labels, it should be understood as an evaluation-time diagnostic rather than a deployment-time risk detector.

Ethical Considerations

This work studies failure modes of VLM-generated emotion explanations. Because emotion recognition is socially and culturally sensitive, our results should not be interpreted as supporting automated affective inference in high-stakes settings such as mental health assessment, hiring, education, or surveillance. We emphasize that RG is an evaluation-time diagnostic requiring ground-truth labels, not a deployment-time detector.

Our analysis primarily uses existing visual emotion benchmarks. We additionally conducted a small-scale anonymous text-only annotation sanity check, in which participants judged short explanation texts and emotion-label pairs. Annotators were recruited through an online university community and were compensated for a short annotation task. Participants were informed that their anonymous responses would be used for research evaluation of explanation-label support and that participation was voluntary. We did not collect names, student IDs, demographic attributes, health information, or personal affective states; any compensation-related communication was handled separately from annotation responses. The proposed metrics are intended for auditing and model evaluation, not for making claims about individual mental states. We plan to release code, prompts, and aggregate EEF scores to support reproducibility, while avoiding any use of the framework for individual-level affective profiling.

Artifact use and redistribution. We use existing visual emotion benchmarks and pretrained models only for offline research evaluation. We do not redistribute original benchmark images. Any released data will contain only anonymized derived outputs, aggregate EEF scores, prompts, and analysis scripts, subject to the licenses and terms of the original datasets and models. These derivative artifacts are intended for research auditing and model evaluation, not for deployment or individual-level affective profiling.

Use of AI assistants. We used GPT for LaTeX syntax checking and general English grammar polishing, and Claude for code debugging and terminal-command assistance. The research questions, experimental design, data analyses, numerical results, interpretations, and final paper claims were checked and approved by the authors.

Acknowledgments

Omitted for anonymous review.

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, pages 14927–14937.

Wentao Bao, Qi Yu, and Yu Kong. 2021. Evidential deep learning for open set action recognition. *arXiv preprint arXiv:2107.10161*.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *CoRR*, abs/2102.12452.

Sree Bhattacharyya and James Z. Wang. 2025. Evaluating vision-language models for emotion recognition. *Preprint*, arXiv:2502.05660.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Preprint*, arXiv:2406.11161.

Sihao Ding, Santosh Vasa, and Aditi Ramadwar. 2025. Explanation-driven counterfactual testing for faithfulness in vision-language model explanations. *Preprint*, arXiv:2510.00047.

Nazire Duran and Anthony P. Atkinson. 2021. Foveal processing of emotion-informative facial features. *PLOS ONE*, 16(12):1–39.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *CoRR*, abs/2004.03685.

Marie Katsurai and Shin'ichi Satoh. 2016. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2837–2841.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Shuo Li, Tao Ji, Xiaoran Fan, Linsheng Lu, Leyi Yang, Yuming Yang, Zhiheng Xi, Rui Zheng, Yuran Wang, Xiaohui Zhao, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024b. Have the vlms lost confidence? a study of sycophancy in vlms. *Preprint*, arXiv:2410.11302.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, Lan Chen, Shan Liang, Ya Li, Jiangyan Yi, Bin Liu, and Jianhua Tao. 2024. Explainable multimodal emotion recognition. *Preprint*, arXiv:2306.15401.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`. Accessed: 2026-05-25.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Preprint*, arXiv:2209.11326.

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, pages 3179–3189.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. 2024. Emovit: Revolutionizing emotion insights with visual instruction tuning. *Preprint*, arXiv:2404.16670.

Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. 2022. MDAN: Multi-level dependent attention network for visual emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9479–9488.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. *Preprint*, arXiv:2307.07961.

Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. 2018. Weakly supervised coupled networks for visual sentiment analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7584–7592.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *Preprint*, arXiv:1605.02677.

Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. 2019. PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 192–201.

A Prompt Templates

Table 12 gives the full text of the three prompting strategies. For few-shot, responses reproducing exemplar keywords are excluded via keyword matching (see Appendix D for per-prompt counts).

B EEF Measurement Pipeline

Algorithm 1 EEF Measurement Pipeline

Require: Image $\mathbf{x}$, ground truth y, VLM $\mathcal{M}$, prompt π, NLI model, label set $\mathcal{L}$

$\hat{y},\ e \leftarrow \mathcal{M}(\mathbf{x}, \pi)$
$e_{\text{evi}} \leftarrow \textsc{ExtractEvidence}(e)$ ▷ Parse "Visual Evidence:" field; CoT uses Step-1 fallback
if $\textsc{HasExemplarMarkers}(e_{\text{evi}})$ **then**
  discard
end if
$e_{\text{clean}} \leftarrow \textsc{RemoveEmotionWords}(e_{\text{evi}})$
for $v \in \{e_{\text{evi}},\ e_{\text{clean}}\}$ **do**
  $\text{CCS}^{(v)}_{\text{pred}} \leftarrow P_{\text{NLI}}(\textsc{entail} \mid v,\ h_{\hat{y}})$
  $\text{CCS}^{(v)}_{\text{truth}} \leftarrow P_{\text{NLI}}(\textsc{entail} \mid v,\ h_{y})$
  $\text{RG}^{(v)} \leftarrow \text{CCS}^{(v)}_{\text{pred}} - \text{CCS}^{(v)}_{\text{truth}}$
end for
for $k \in \mathcal{L}$ **do**
  $s_k \leftarrow P_{\text{NLI}}(\textsc{entail} \mid e_{\text{evi}},\ h_k)$
end for
$\text{Conf} \leftarrow \max_k \textsc{Softmax}(s_k)$
$u \leftarrow K / \sum_k (c \cdot \max(s_k, 0) + 1)$
$Q_1, Q_2 \leftarrow \textsc{GPT4oJudge}(\mathbf{x},\ e_{\text{evi}},\ \hat{y})$
$\text{PF} \leftarrow (Q_1 + Q_2)/2$

C Statistical Testing Details

All tests are two-sided unless a directional hypothesis is explicitly stated.

- **Mann-Whitney U** with rank-biserial r: comparing RG between correct and incorrect predictions (one-sided: incorrect > correct). Bonferroni correction applied for model-wise comparisons.
- **Spearman** ρ: correlating EEF metrics with error indicator (0/1). Bootstrap 95% CI with $n = 2{,}000$ resamples.
- **AUROC / AUPR**: error (incorrect prediction) is the positive class.
- **Wilcoxon signed-rank**: for paired PF condition comparisons and prediction-specific control.
- **Partial correlation**: regressing out text-only PF from both image-conditioned PF and accuracy, then computing Spearman on residuals.
- **Bonferroni correction**: applied to all multiple comparisons (model-wise, prompt-wise, category-wise).

Effect size conventions: rank-biserial $|r| < 0.3$ = small, 0.3–0.5 = medium, > 0.5 = large. All analyses use `scipy.stats` and `sklearn.metrics`.

Compute budget. All VLM inference was run in float16 on NVIDIA RTX A6000 GPUs. The final reproducible GPU-based runs required approximately 6.9 GPU-hours ($\approx$ 0.3 GPU-days): 6.8 hours for VLM inference, 0.1 hours for main NLI scoring, and 0.1 hours for alternative scorer runs. End-to-end wall-clock experimentation took several days because generation, reruns, debugging, file I/O, and API-based PF judging were performed in multiple batches.

D Data Audit

Tables 13 and 14 report filtering by model and by prompting strategy, respectively. Few-shot accounts for the majority of exemplar contamination. The initial filtering step removes exemplar contamination and grossly invalid label fields. Prompt-specific output-format failures that lead to undefined RG are audited separately in Appendix D.1.

Model	Raw	Parsed	Contam.	Inv. label	Final
LLaVA-1.6	2,700	2,700	62	20	2,618
LLaVA-OV	2,700	2,700	47	13	2,640
Qwen2-VL	2,700	2,700	97	46	2,557
Total	8,100	8,100	206	79	7,815

Table 13: Data audit by model.

Strategy	Full Prompt Text

Standard	*Look at this image carefully. Answer strictly in this format:*
	`Emotion: [single emotion label]`
	`Visual Evidence: [specific visual cues you observed]`
	`Reasoning: [explanation of why those cues indicate this emotion]`

CoT	*Analyze this image step by step:* *Step 1 – Visual elements: list what you literally see.* *Step 2 – Intention: what does the subject seem to express?* *Step 3 – Emotion: given the above, what is the primary emotion?* *Answer:* `Emotion: [label] | Visual Evidence: [Step 1] | Reasoning: [Step 2→3]`

Few-shot	`Example 1: Emotion: sadness | Visual Evidence: hunched shoulders, grey sky | Reasoning: withdrawal posture signals depletion.`
	`Example 2: Emotion: awe | Visual Evidence: vast mountain, tiny figure, golden light | Reasoning: scale contrast triggers grandeur.`
	`Example 3: Emotion: disgust | Visual Evidence: furrowed brow, recoiling posture | Reasoning: tension confirms rejection.`
	Now analyze the following image in the exact same format.

Table 12: Full prompt templates. Standard: minimal structured request. CoT: step-by-step decomposition. Few-shot: three labeled exemplars.

Model	Standard	CoT	Few-shot
LLaVA-1.6	25	3	54
LLaVA-OV	2	6	52
Qwen2-VL	16	8	119

Table 14: Filtered outputs by prompt strategy (count out of 900). Counts include both exemplar contamination and invalid labels.

Model	Prompt	n	Acc.	Mean RG	RG NaN%	Canonical%
LLaVA-1.6	Standard	875	0.325	0.267	3.9%	48.6%
	CoT	897	0.087	0.444	2.8%	12.8%
	Few-shot	846	0.220	0.222	0.0%	33.6%
LLaVA-OV	Standard	898	0.293	0.283	0.0%	41.1%
	CoT	894	0.031	0.172	**89.5%**	**4.6%**
	Few-shot	848	0.268	0.214	0.4%	35.3%
Qwen2-VL	Standard	884	0.265	0.224	0.0%	40.0%
	CoT	892	0.207	0.301	0.0%	28.6%
	Few-shot	781	0.283	0.158	0.0%	47.8%

Table 15: Prompt-wise output-format audit on retained triples. RG NaN%: fraction of rows with undefined CCS_{pred} or RG. Canonical%: fraction of predicted labels belonging to the dataset taxonomy. LLaVA-OV CoT shows an 89.5% RG NaN rate, measuring rows with undefined CCS_{pred} or RG. Separately, 89.1% of its CoT predictions are empty, and an additional 2.5% contain the literal "[label]" placeholder. Thus, the CoT failure mainly reflects output-format collapse rather than a pure reasoning failure. LLaVA-1.6 and Qwen2-VL CoT format compliance is near-complete ($\leq$3%).

D.1 Prompt-wise Output-Format Audit

Table 15 reports parsing rates for each model–prompt combination on the retained 7,815 triples. RG NaN rate measures the fraction of rows with undefined RG (excluded from AUROC analyses rather than imputed), and canonical rate measures the fraction whose predicted label belongs to the dataset taxonomy.

D.2 Prediction-Specific Control under Prompt Restriction

Subset	n	$\mathrm{RG}_{\mathrm{pred}}$	$\mathrm{RG}_{\mathrm{rand}}$	Δ
All canonical	809	+0.189	−0.131	+0.321
Standard only	367	+0.204	−0.146	+0.350

Table 16: Prediction-specific control under prompt restriction. The Standard-only subset preserves the same sign pattern ($\mathrm{RG}_{\mathrm{pred}} > 0$, $\mathrm{RG}_{\mathrm{rand}} < 0$; Wilcoxon $p < 0.001$), reducing concern that the canonical-label analysis is driven by prompt composition.

E Label-Word Debiasing

Emotion	Removed words (partial)
amusement	amused, amusing, fun, funny, playful, lively
anger	angry, rage, furious, outrage, hostile, mad
awe	awestruck, wonder, astonish, majestic, grand
contentment	content, serene, peaceful, calm, tranquil
disgust	disgusted, revolting, repuls*, aversion
excitement	excited, thrill, energetic, enthusiastic
fear	fearful, scared, terrify, dread, horror
sadness	sad, melancholy, grief, sorrow, gloomy
joy	joyful, happy, happiness, delight, cheerful
surprise	surprised, shock, startl*, unexpected

Table 17: Emotion synonym sets for debiased RG (* = prefix match). Full sets will be released with the anonymized supplementary material.

F Semantic Scorer and Hypothesis Template Robustness

We verify that RG's advantage over EDL-u and softmax-style confidence is not specific to the DeBERTa-v3-large scorer or to a particular hypothesis template.

Alternative scorers. Table 18 reports RG error detection under DeBERTa-v3-small (same cross-encoder family, smaller size) and RoBERTa-base NLI (a different model family). Absolute AUROC varies across scorers, but all configurations substantially outperform EDL-u (0.511) and softmax-style confidence (0.559), preserving the directional conclusion.

Metric / Scorer	AUROC	Spearman r_s
RG, DeBERTa-v3-large (main)	0.852	0.527
RG, DeBERTa-v3-small	0.862	0.543
RG, RoBERTa-base NLI	0.772	0.408
1−Softmax-style confidence	0.559	0.083
EDL-inspired uncertainty	0.511	0.016

Table 18: RG error detection under alternative NLI scorers (positive class = incorrect prediction). All RG variants substantially outperform EDL-u and softmax-style confidence.

Hypothesis template robustness. The main analysis uses "The primary emotion is {label}." Table 19 shows results under three semantically related hypothesis formulations. The T1 row independently recomputes NLI scores under the main hypothesis wording; small differences from cached main scores do not affect the conclusion. All templates preserve the conclusion that evidence direction is more informative than evidence-quantity baselines.

Template	AUROC	r_s	Wrong-label%
T1: Primary emotion	0.838	0.507	82.4%
T2: Explanation supports label	0.784	0.428	74.7%
T3: Affective interpretation	0.818	0.478	79.7%
EDL-u reference	0.511	0.016	—

Table 19: RG error detection under three hypothesis-template variants ($n = 6{,}953$; rows with undefined RG excluded). T1: "The primary emotion is {label}."; T2: "This explanation supports {label}."; T3: "The affective interpretation is {label}." Wrong-label% is the fraction of incorrect predictions with RG > 0. All templates substantially outperform EDL-u, confirming that the direction-based advantage is not specific to a single hypothesis formulation.

G All-label Support Profile

Model	Argmax = Truth	Argmax = Pred	Other
LLaVA-1.6	23.4%	8.2%	68.4%
LLaVA-OV	27.5%	5.2%	67.2%
Qwen2-VL	23.6%	8.3%	68.1%

Table 20: All-label support on incorrect predictions. Nominal chance: 12.5% (FI, 8-class) and 16.7% (Emotion6, 6-class). Argmax = Pred (5–8%) is below nominal chance for both taxonomies, while Argmax = Truth (23–28%) is above both. Rationalization operates as a marginal directional bias (RG > 0), not as dominant argmax-level label echo.

H Extreme Rationalization Threshold Sensitivity

Threshold	% errors flagged
RG > 0.5	22.4–40.4%
RG > 0.6	21.8–39.3%
RG > 0.7	21.4–38.0%
RG > 0.8	20.6–36.5%
RG > 0.9	19.1–33.9%

Table 21: Extreme rationalization threshold sensitivity. The proportion of errors flagged decreases gradually, while the model ordering is identical across all thresholds: LLaVA-1.6 $>$ Qwen2-VL $>$ LLaVA-OV. This confirms that the choice of RG > 0.8 is not arbitrary.

I EDL Scaling Robustness

Scaling c	AUROC	r_s vs. error
$c = 1$	0.511	0.016
$c = 5$	0.510	0.015
$c = 10$	0.509	0.014

Table 22: EDL-u near-chance performance across scaling constants, confirming the quantity-direction distinction is robust to hyperparameter choice.

J Image-Conditioned Plausibility: Controls and Rubric

Condition	Mean PF	Wilcoxon vs. Image-cond.
Image-conditioned	3.662	—
Text-only	3.442	$p = 0.019$
Mismatched-image	1.075	$p < 0.001$

Table 23: GPT-4o judge validity controls ($n = 154$ paired samples).

Judge rubric. *You are given an IMAGE, an emotion label, and a visual explanation. Rate:* ***Q1 (Visual Presence, 1–5)****: Are the mentioned visual cues visible in the image?* ***Q2 (Affective Relevance, 1–5)****: Do those cues plausibly support the stated emotion? JSON only: {"Q1": <1-5>, "Q2": <1-5>}*

Text-only: only Q2 ("If present, would they support the emotion?").

K Evaluation-time Risk-Coverage Analysis

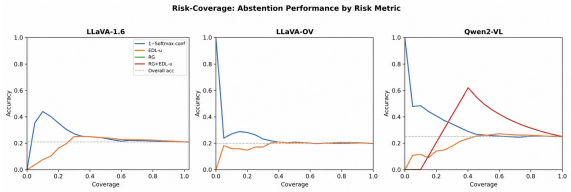

Figure 3: Risk-coverage curves (evaluation-time; ground-truth labels required). RG achieves the steepest accuracy gain as coverage decreases. EDL-u provides near-zero benefit, consistent with AUROC 0.511. RG is not a deployment-time signal without ground-truth labels; label-free proxies such as all-label support entropy require separate validation.

L Dataset and Emotion-Category Analyses

DS	Model	CCS_p	CCS_t	RG	Acc.
FI	LLaVA-1.6	0.587	0.276	0.311	0.217
	LLaVA-OV	0.597	0.357	0.240	0.285
	Qwen2-VL	0.628	0.419	0.210	0.238
E6	LLaVA-1.6	0.412	0.098	0.315	0.193
	LLaVA-OV	0.421	0.164	0.258	0.274
	Qwen2-VL	0.417	0.145	0.273	0.275

Table 24: EEF metrics by dataset. RG model ordering preserved across both. Emotion6 shows lower CCS_{truth}, which may reflect differences in dataset construction, image style, or label semantics rather than taxonomy granularity alone.

Cross-dataset results.

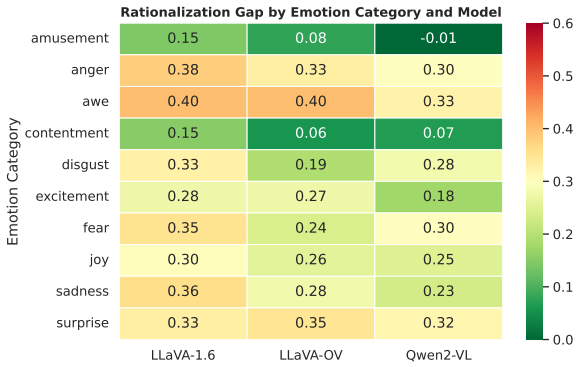

Figure 4: Mean RG by emotion category. *Awe*: highest RG (0.326–0.400); one interpretation is that awe requires mapping surface grandeur to an abstract affective response, though we treat this as a qualitative hypothesis. *Amusement/contentment*: lowest. Pattern consistent across FI and Emotion6.

Emotion category analysis.

M Evidence Length Analysis

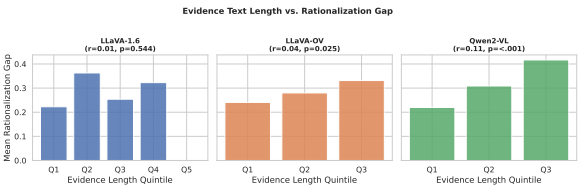

Figure 5: Evidence length quintile vs. RG. Spearman r: 0.012 ($p = 0.54$), 0.037 ($p = 0.02$), 0.108 ($p < 0.001$) for LLaVA-1.6, LLaVA-OV, Qwen2-VL. Small effect sizes suggest rationalization is not primarily explained by response verbosity.

N Qualitative Examples

Examples are lightly truncated for length but not rewritten; raw outputs will be released with the anonymized supplementary material.

Model	GT / Pred	Evidence and pattern	RG
High-RG — incorrect			
LLaVA-1.6	content / fear	"Dark shadows and a looming figure suggest fear." [generic cliche]	0.97
Qwen2-VL	excite / awe	"The vast landscape evokes awe; grandeur overwhelms." [wrong attribution]	0.93
LLaVA-OV	disgust / anger	"Clenched fists and confrontation express anger." [label echo]	0.89
Low-RG — correct			
Qwen2-VL	amus. / amus.	"Crinkled eyes, open mouth: unguarded laughter."	0.02
LLaVA-OV	sad. / sad.	"Downcast eyes, hunched posture: inward withdrawal."	0.04

Table 25: Additional high-RG and low-RG examples. Examples are lightly truncated, not rewritten.

O Additional Figures

These figures provide diagnostic checks that support the main claims while remaining supplementary to the core argument:

Figure 6 supports RQ1 by showing that rationalization is distributional: incorrect predictions form a right-skewed RG distribution, while correct predictions concentrate near RG $= 0$.

Figure 7 complements Table 10 by visualizing how prompting changes CCS_p, CCS_t, and RG jointly.

Figure 8 explains why PF is auxiliary: the judge is image-sensitive, but partially text-driven.

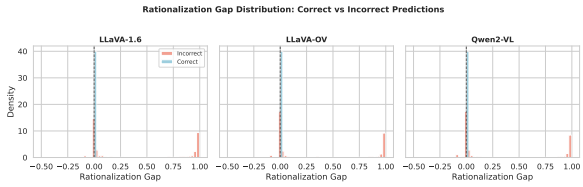

Figure 6: RG density for correct vs. incorrect predictions. Incorrect predictions show a right-skewed distribution with a heavy tail at high RG, while correct predictions concentrate near RG $= 0$.

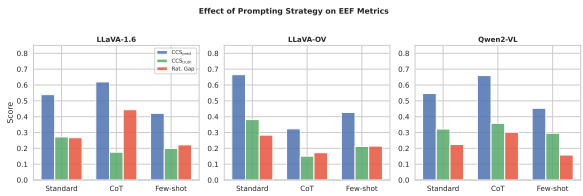

Figure 7: Prompt strategy effect on CCS_p, CCS_t, and RG per model. CoT can raise CCS_p without proportional CCS_t gains, widening RG most clearly for LLaVA-1.6 (RG: 0.267→0.444, Acc: 0.325→0.087).

Example annotation item
Explanation: "The person's wide open mouth, showing teeth, and crinkling around the eyes suggest a joyful expression." **Question:** Which emotion is better supported by this explanation? A. anger B. joy C. Both equally D. Neither / unclear

Table 26: Example item from the human directional-support sanity check. Annotators saw only the explanation text and two randomized labels.

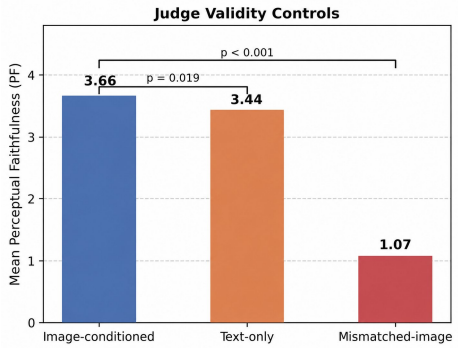

Figure 8: GPT-4o judge validity controls. The large gap between image-conditioned and mismatched-image PF confirms image sensitivity, while the smaller gap between image-conditioned and text-only PF indicates partial text-drivenness.

P Human Directional-Support Sanity Check

We conduct a small-scale human sanity check to test whether RG aligns with human judgments of label-support direction. Four annotators were recruited through an online university community. Annotators were not authors of the paper and were blinded to model identity, ground-truth label, model prediction, correctness, and RG score.

Task. Each annotator judged 100 items. For each item, annotators saw only an explanation text and two emotion labels with randomized order. They were asked which label was better supported by the explanation. The task was not to judge the image-level emotion or overall explanation quality, but to evaluate a narrower construct: direction of label support.

Response options. Annotators selected one of four options: (A) Label A is better supported, (B) Label B is better supported, (C) Both equally, or (D) Neither / unclear.

Item construction. The 100 examples consist of 50 high-RG incorrect examples and 50 low-RG incorrect controls, balanced across the three evaluated VLMs where possible. The order of the two labels was randomized for each item, and annotators were not shown whether either label was the model prediction or the ground truth.

Results. At the item level, RG correlates with the human pred-support rate (Spearman $\rho = 0.543$, $p = 5.3 \times 10^{-9}$; bootstrap 95% CI $[0.392, 0.671]$). On high-RG incorrect examples, annotators selected the model-predicted label in 65.0% of judgments (130/200; binomial $p = 1.3 \times 10^{-5}$), compared with 29.0% on low-RG incorrect controls (58/200). The difference is significant by Fisher's exact test (OR $= 4.55$, $p = 3.2 \times 10^{-13}$). This provides a small-scale sanity check that RG captures the direction of label support perceived by human readers.

Agreement and feedback. Inter-annotator agreement is modest (Fleiss' $\kappa = 0.231$ overall, falling in the "fair" range under the conventional scale of Landis and Koch (1977), consistent with the ambiguity of fine-grained emotion labels and short explanation texts. Annotators reported ambiguity between semantically adjacent labels such as joy/amusement, difficulty judging scene-level descriptions, and cases where the explanation suggested affective states outside the provided label set. We therefore treat this annotation as a directional-support sanity check rather than a full human evaluation of visual explanation quality.