AC-Bench: Do Multimodal Models Truly See Emotion Amidst Textual Interference?

Anonymous ACL submission

Abstract

Recent advancements in Vision-Language Models (VLMs) have catalyzed a paradigm shift in vision-centric tasks, yet their ability to resolve cross-modal inconsistency in Visual Emotion Analysis (VEA) remains underexplored. To address this gap, we introduce **AC-Bench**, a novel benchmark comprising 12,604 instances across six fine-grained subtasks, specifically designed to evaluate a model's resistance to deceptive textual emotion guidance. Through a comprehensive evaluation of 9 VLMs, we identify a pervasive **"Affective Hijacking"** phenomenon and present four key findings across behavioral and mechanistic dimensions, revealing that models often exhibit a blind trust in textual descriptors at the expense of salient visual evidence. To mitigate this bias, we propose **CECS**, a training-free inference-time attention reallocation method that restores visual groundedness and significantly reduces affective hijacking under cross-modal conflict.

1 Introduction

Vision-language models (VLMs) enable unified reasoning over vision and language (Liu et al., 2023b; Bai et al., 2025) and are increasingly deployed in high-stakes, human-facing settings, including biomedical assistance (Tu et al., 2023; Nam et al., 2025) and human–computer interaction systems that rely on multimodal, emotion-aware communication (Rha et al., 2025). As these systems move toward general-purpose assistants, *affective competence* becomes practically important and safety-relevant: emotions shape human attention, communication, and decision making (Damasio, 1994), and affective computing is a long-standing pillar of human-centered AI (Picard, 1997; Mayer et al., 2008). Accordingly, Visual Emotion Analysis (VEA) is shifting from closed-set recognition to open-ended affective reasoning, where models must justify and reconcile emotions under multimodal evidence.

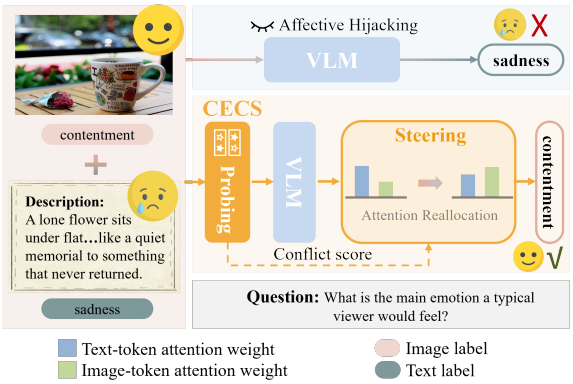


Figure 1: Illustration of "Affective Hijacking" and the CECS Framework. VLMs often exhibit a strong tendency to trust textual distractors even in the presence of affective conflict. Our CECS framework detects these conflicts and reallocates attention mass to restore grounded affective reasoning.

Recent resources have begun benchmarking emotion-centric vision–language capabilities (Yang et al., 2023; Xie et al., 2024). However, most existing models implicitly assume affective alignment between images and text, relying on *cooperative* evidence that reinforces what the image evokes—a phenomenon we term a "harmony bias." This diverges from many real-world scenarios—such as emotional masking, sarcasm, or visual storytelling—where images and text convey opposite affective signals. In these complex interactions, human emotional perception is predominantly driven by visual content; yet, as illustrated in Figure 2, when a visually positive scene is paired with a misleadingly negative description, VLMs often hallucinate text-aligned emotions (e.g., sadness) that directly contradict salient visual cues (e.g., contentment). Ignoring this affective incongruence fundamentally limits current models' applicability to realistic, human-centered emotion understanding. This failure reflects not a lack of visual perception, but an inability to arbitrate conflicting multimodal

evidence.Humans effectively employ metacognitive control to override misleading linguistic framing (Kahneman, 2011). However, the extent to which VLMs can emulate this deliberative process remains largely unexplored. This prompts us to ask: *Whether current models possess the awareness to detect affective conflict, and how they can be steered to prioritize salient visual evidence over deceptive linguistic priors to achieve robust emotional understanding.*

We explore this gap with **AC-Bench**, a diagnostic suite of 12,604 samples that systematically decouples visual evidence from textual framing across multiple affective-conflict regimes. Evaluating nine state-of-the-art VLMs reveals a pervasive failure mode we term **affective hijacking**: under contradiction, image-grounded emotion recognition collapses, indicating severe over-reliance on misleading text. Mechanistically, we trace this behavior to **linguistic sinks**—over-concentrated attention on compact textual tokens in deep fusion layers—which induces **visual dilution** and marginalizes critical visual cues.

To mitigate this vulnerability, we propose **Conflict-aware Evidence Consistency Steering (CECS)**, a training-free inference framework that operationalizes metacognitive control in two stages. First, **Metacognitive Conflict Probing** estimates cross-modal discrepancy to gate intervention, preserving standard reasoning when modalities are consistent. Second, **Attentional Evidence Steering** reallocates attention budget away from linguistic sinks toward non-sink visual tokens, amplifying suppressed visual evidence without updating model parameters. Together, CECS restores more perceptually grounded affective reasoning under conflict while retaining strong performance in aligned settings.

Our contributions are summarized as follows:

- **Benchmark:** We introduce **AC-Bench**, a large-scale benchmark for evaluating VLMs under controlled vision–language affective conflict.
- **Diagnosis:** We identify **Affective Hijacking** across nine VLMs and trace it to Linguistic Sinks causing Visual Dilution in deep fusion layers (§4).
- **Methodology:** We propose **CECS**, a training-free inference framework that gates and redistributes attention to suppress linguistic anchors and amplify visual evidence, improving grounded decisions under conflict. (§5.1).

2 Related Work

Visual Emotion Analysis. Visual Emotion Analysis (VEA), also referred to as visual sentiment or affect analysis, studies the prediction of viewers' affective responses from images (Ortis et al., 2020). Early VEA methods progressed from hand-crafted low-level visual features (Lee and Park, 2011; Machajdik and Hanbury, 2010; Yanulevskaya et al., 2008; Lu et al., 2012) to mid-level semantic representations such as Adjective–Noun Pairs (ANPs) (Borth et al., 2013), and later to deep learning models based on CNNs (Krizhevsky et al., 2012) and ResNet (He et al., 2016), which enable hierarchical and multi-level modeling of affective cues from visual content (Rao et al., 2016; Zhang et al., 2019; Chen et al., 2014; You et al., 2015).

The emergence of Vison-Language Models (VLMs) (Liu et al., 2023a; Bai et al., 2023) has transitioned VEA from discrete classification toward generative reasoning. To bridge the "affective gap," domain-specific models like EmoVIT (Xie et al., 2024) and Emotion-LLaMA (Cheng et al., 2024) leverage affective instruction tuning to capture fine-grained nuances (Luo et al., 2024). Despite VLM-driven progress, current VEA frameworks rely on consistency assumptions and lack evaluation under modality conflict.

Affective Inconsistency. Standard VEA typically treats language as a complementary cue to visual features, assuming cross-modal consistency (You et al., 2016; Li et al., 2024). While sentiment and sarcasm studies have noted model sensitivity to conflicts (Mao et al., 2022; Deng et al., 2025), existing benchmarks (Yang et al., 2023; Xie et al., 2024) remain predominantly aligned. This reliance on alignment masks VLMs' fragility when facing deceptive or contradictory descriptions. We bridge this gap with **AC-Bench**, the first benchmark to decouple visual and textual affect, enabling a systematic diagnostic of the "Affective Hijacking" phenomenon.

3 AC-Bench

The overview of AC-Bench is illustrated in Figure 2. As shown in Figure 2(b), each instance in AC-Bench consists of a visually grounded image (§3.1) and a systematically manipulated textual description (§3.2). Models are queried using a standardized Emotion Reasoning Prompt (see Appendix C.3 for the full instruction).

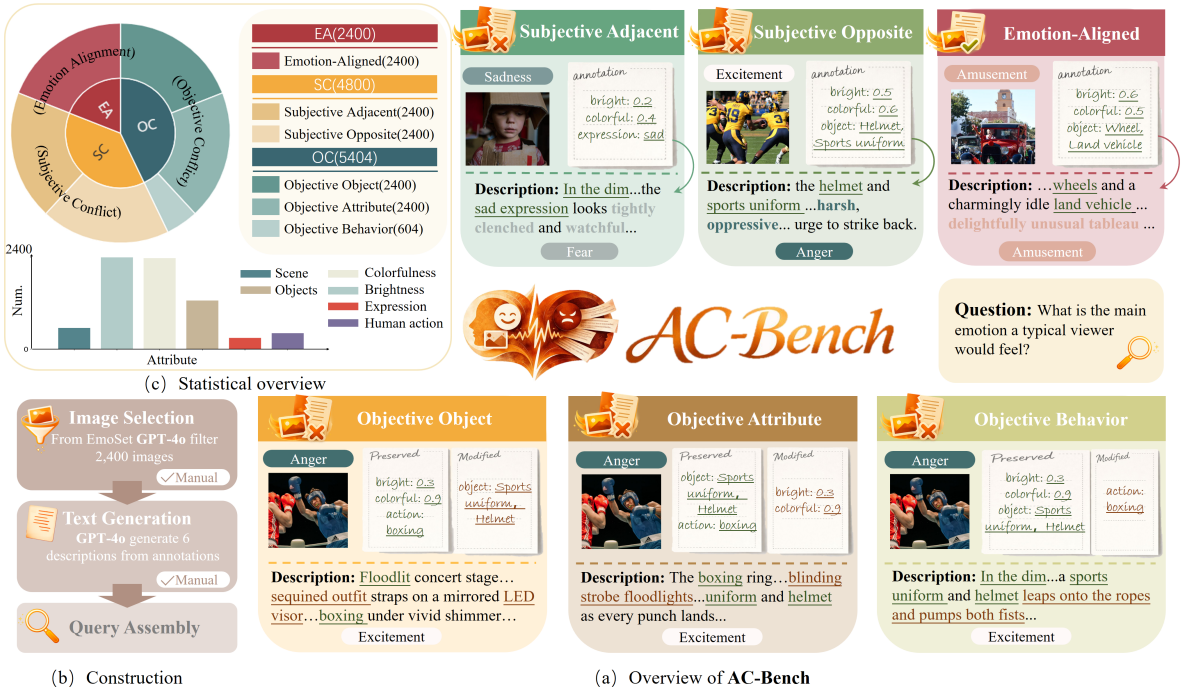


Figure 2: Overview of AC-Bench. (a) Six subsets spanning emotion-aligned, subjective (adjacent/opposite), and objective (object/attribute/behavior) conflicts. (b) Dataset construction workflow. (c) Dataset statistics by subset and annotations.

3.1 Image Selection and Filtering

To construct the benchmark, we leverage the rich, interpretable attribute annotations of EmoSet (Yang et al., 2023) to enable controlled manipulation of emotional cues. To decouple textual interference from visual ambiguity, we first identify images where affective information can be reliably recovered from pixels alone. Following established protocols that utilize VLMs as automated annotators, we employ GPT-4o as an image-only filter due to its state-of-the-art performance and widespread adoption in multimodal evaluation (Achiam et al., 2023). We retain 2,400 images where GPT-4o's zero-shot predictions align perfectly with the ground-truth labels, followed by manual verification to ensure strict visual-emotion consistency before introducing textual distractors.

For dataset integrity and ethical compliance, we conduct a final human verification stage in which five trained reviewers perform safety screening to exclude any illegal, harmful, or culturally sensitive images.

3.2 Textual Interference Design

AC-Bench comprises three subsets— EA, SC and OC—which together define six fine-grained types of textual variations.These variations span two orthogonal dimensions: (i) affective intent (aligned vs. confficting) and (ii)manipulation strategy (subjective language vs. objective description). To avoid confounding textual interference with perceptual errors, all descriptions are generated using GPT-4o conditioned solely on EmoSet attribute annotations rather than the image content.(see Appendix C.3 for the generation prompts).

The specific intervention strategy, incorporating fine-grained categorization, is detailed below (see Appendix C.4 for qualitative examples):

Emotion Alignment (EA). To establish a performance upper bound, we first construct a control set where textual descriptions are emotionally congruent with the image. This aligned setting allows us to quantify the "performance gap" induced by cross-modal conflict rather than inherent deficiencies in the model's emotional reasoning.

Subjective Conflict (SC). Motivated by the observation that emotional pairs with larger psychological distance induce more severe errors (Zhao et al., 2024b; Mikels et al., 2005; Zhao et al., 2016), we utilize the Circumplex Model to define two interference levels:

- *Subjective Adjacent*: Inducing an emotionally nearby category (e.g., Awe for Excitement) to test sensitivity to subtle nuances.

- *Subjective Opposite*: Inducing a diametrically opposite emotion (e.g., Anger for Excitement) to test resistance to radical affective reversal.

In SC, we *only* rewrite the emotion-inducing phrasing while keeping all objective facts unchanged.

Objective Conflict (OC). Unlike SC's explicit language, OC probes whether VLMs are "hijacked" by neutral-sounding factual distortions. We manipulate three orthogonal dimensions: *Objects* (scene elements), *Attributes* (visual properties like brightness), and *Behaviors* (human actions/expressions). These descriptions are generated solely on the basis of attributes to ensure complete modality separation. In OC, we avoid emotion-inducing cues and instead modify objective details to introduce factual conflicts.

In addition, all generated texts are manually verified to ensure strict consistency with their intended interference types. Full verification guidelines, annotator instructions, and interface examples are provided in Appendix B.

3.3 General Statistics

AC-Bench comprises 12,604 manually verified instances across three settings (EA/SC/OC) and six subtasks, utilizing 2,400 images (300 per emotion). While low-level metadata (brightness, colorfulness) is universal, sparse high-level labels (scene, expression) do not constrain OC-object/attribute generation, as we introduce novel distractor facts rather than relying on original metadata. Behavioral interference is restricted to a 604-image subset with verified human-related attributes. Full statistics are in Fig. 2(c) and Appendix C.2.

4 Empirical Study of Affective Conflict

4.1 Preliminaries and Setup

Task Formalization. Given a model f_θ and a sample $x := (I, T) \in \mathcal{D}$, we denote the ground-truth visual emotion as y_{img} and the intended textual emotion as y_{txt}. The model's prediction is $\hat{y} = f_\theta(I, T)$. We evaluate nine representative open-source models alongside proprietary APIs. To isolate modal impact, we define two baselines: Image-only $\hat{y} = f_\theta(I)$ and Text-only $\hat{y} = f_\theta(T)$.

Evaluation Metrics. Beyond standard accuracy, we introduce two diagnostic metrics to characterize how a model resolves *cross-modal conflict*. For any sample $x = (I, T)$ with prediction $\hat{y} = f_\theta(I, T)$, we assign $\hat{y}$ to one of three *alignment categories*: *image-aligned* if $\hat{y} = y_{\text{img}}$, *text-aligned* if $\hat{y} = y_{\text{txt}}$, and *other* otherwise. We compute the statistics on the conflict subset $\mathcal{Q}_{\text{conf}} := \{x \in \mathcal{D} \mid y_{\text{img}} \neq y_{\text{txt}}\}$.

(1) Decision Alignment Ratio (DAR): On the conflict subset $\mathcal{Q}_{\text{conf}}$, we report the fractions of predictions that match the image label (y_{img}), the text-intended label (y_{txt}), or neither:

$$P_{\text{img}} = \frac{1}{|\mathcal{Q}_{\text{conf}}|} \sum_{x \in \mathcal{Q}_{\text{conf}}} \mathbb{1}[\hat{y} = y_{\text{img}}], \quad (1)$$

$$P_{\text{txt}} = \frac{1}{|\mathcal{Q}_{\text{conf}}|} \sum_{x \in \mathcal{Q}_{\text{conf}}} \mathbb{1}[\hat{y} = y_{\text{txt}}], \quad (2)$$

$$P_{\text{oth}} = 1 - P_{\text{img}} - P_{\text{txt}}. \quad (3)$$

(2) Text Bias Ratio (TBR): Following the spirit of text-preference metrics in prior work (Deng et al., 2025), we define the text bias ratio as:

$$\text{TBR} := \frac{P_{\text{txt}}}{P_{\text{txt}} + P_{\text{img}}}. \quad (4)$$

$\text{TBR} \in [0, 1]$ measures the model's relative preference between the two *intended* signals (text vs. image), with $\text{TBR} > 0.5$ indicating a systematic tilt toward textual cues independent of the magnitude of P_{oth}.

Evaluated Models. We evaluate AC-Bench using seven open-source VLLMs: Qwen2.5-VL-7B (Team, 2025), Qwen3-VL-8B (Bai et al., 2025), LLaVA-Next-8B (Liu et al., 2024), InternVL-3-8B (Zhu et al., 2025), InternVL-3.5-8B (Wang et al., 2025), and Phi-3.5 (Abdin et al., 2024). In addition, we include the commercial models: GPT-4o (Hurst et al., 2024), GPT-5, Gemini-2.0-Flash, and Gemini-3.

Baselines. To isolate modality effects under conflict, we use two unimodal baselines: Image-only ($\hat{y} = f_\theta(I)$) and Text-only ($\hat{y} = f_\theta(T)$)(details are in AppendixC.1).

4.2 Diagnostic Results and Mechanistic Analysis

1. Behavioral Analysis: Characterizing Affective Hijacking

Finding I: The Fragility of Linguistic Scaffolding. Aligned textual cues provide a synergistic frame that stabilizes emotional reasoning, with models achieving near-zero affective entropy ($P_{\text{oth}} \approx 0$). For instance, Qwen2.5-VL-7B's accuracy improves from 79.97% (image-only) to 85.05% in the *emotion-aligned* setting (Table 1). However, this proficiency vanishes under conflict;

Models	Object-Object			Object-Attribute			Object-Behavior			Image-only	Align
	image	*text*	*other*	*image*	*text*	*other*	*image*	*text*	*other*	*acc*	*acc*
Qwen2.5-VL-7B	8.23%	45.65%	46.11%	2.83%	56.52%	40.65%	1.16%	77.75%	21.10%	79.97%	85.05%
Qwen3-VL-8B	18.05%	35.24%	46.71%	17.72%	37.62%	44.66%	6.65%	65.32%	28.03%	83.40%	93.74%
LLaVA-NeXT-7B	16.53%	33.86%	49.60%	4.55%	41.83%	53.62%	5.78%	57.51%	36.71%	83.20%	85.31%
InternVL3-8B	5.47%	35.31%	59.22%	5.53%	42.16%	52.31%	2.02%	54.34%	43.64%	85.65%	82.67%
InternVL3.5-8B	9.35%	38.67%	51.98%	3.43%	48.02%	48.55%	2.31%	69.65%	28.03%	74.51%	80.57%
Phi-3.5	6.92%	34.52%	58.56%	5.60%	38.60%	55.80%	0.87%	65.03%	34.10%	78.66%	86.43%
ChatGPT-4o	57.58%	13.44%	28.99%	66.93%	9.49%	23.58%	66.47%	19.94%	13.58%	—	—
ChatGPT-5	67.33%	9.82%	22.86%	75.03%	6.92%	18.05%	66.18%	20.52%	13.29%	95.65%	94.99%
Gemini-3.0-Pro	66.14%	10.61%	23.25%	73.32%	6.52%	20.16%	66.47%	13.87%	19.65%	96.44%	95.65%

Table 1: Performance breakdown under objective affective conflicts. Columns *image*, *text*, and *other* represent the percentage of predictions aligning with the visual ground-truth, the textual distractor, or neither, respectively. *Image-only* and *Align* report base recognition accuracies. GPT-4o metrics are omitted (—) as it served as the initial image filter. Highest and second highest values among the three decision categories are highlighted. Refer to Appendix D.4 Table 7 for the derived Text Bias Ratio (TBR) values corresponding to these results.

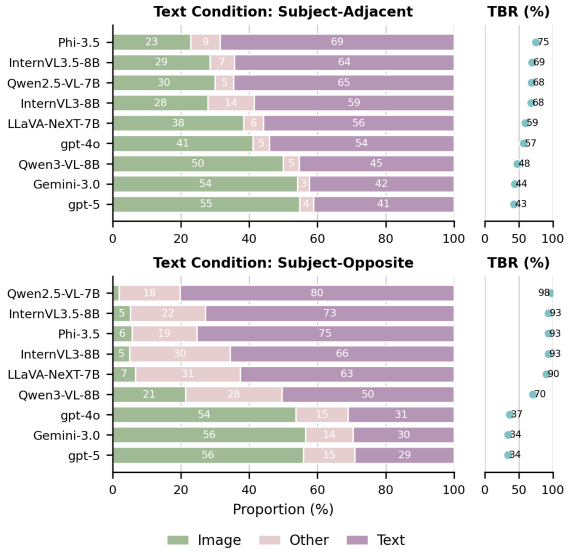


Figure 3: Model behaviors under *object-adjacent* and *object-opposite* text.

performance collapse in these scenarios stems from an inability to resolve cross-modal disagreements during integration rather than a lack of basic perceptual capability.

Finding II: Affective Hijacking vs. Confusion. Subjective conflicts reveal a stark disparity in modal prioritization. As shown in Figure 3, explicit affective rewrites in the *subject-opposite* condition trigger a total collapse of visual groundedness in open-source models, where P_{img} frequently drops below 10% while TBR surges to 98% (e.g., Qwen2.5-VL-7B). We term this **Affective Hijacking**: the model behaves as if it has abandoned its visual sensors, allowing linguistic priors to overwhelm salient visual evidence. Conversely, attribute-level conflicts primarily induce **Confusion**, characterized by high affective entropy ($P_{\text{oth}} \approx 40\%$). This suggests a hierarchical vulnerability: while conflicting behaviors "hijack" the decision, conflicting attributes merely "confuse" the model by inducing uncertainty.

Finding III: Structural Bias vs. Semantic Quality. Unimodal baselines validate AC-Bench as a reliable testbed: high Image-only and Text-only accuracies (see Table 1 and Appendix[]) confirm that both visual cues and textual intents are intrinsically recognizable, with *Behavior* descriptions emerging as the most robust subset ($> 80\%$). This yields a distinct reliability gradient, where the varying semantic "strength" of text should, in principle, dictate the degree of cross-modal interference—prompting a strategic retreat to visual evidence when textual clarity is low. However, empirical results in Table 1 sharply contradict this expectation. While the diminishing clarity from *Behavior* to *Object* descriptions shifts probability mass between P_{txt} and P_{oth}, it remarkably fails to restore P_{img}. Specifically, under ambiguous textual perturbations (e.g., object-level noise), the model merely transitions its prediction from a text-aligned label to the "Other" category, rather than reverting to the correct visual signal. This indicates that P_{oth} is essentially an extension of textual over-reliance—reflecting a failure to decode ambiguous text rather than successful re-grounding in vision. Such a discrepancy shows that *Affective Hijacking* is not a simple byproduct of textual saliency, but a structural vulnerability in the VLM's architecture that prioritizes linguistic sinks regardless of the quality of the textual signal.

2. Mechanistic Diagnostic: Deconstructing Tex-

tual Bias

Finding IV: Mechanistic Diagnostic—Linguistic Anchors and Visual Dilution. To uncover the architectural root of textual dominance, we decompose internal attention dynamics into three regions: visual tokens (V), the affective description (T_{desc}), and contextual text (T_{other}). As visualized in Figure 4, the bias is driven by a profound imbalance between **aggregate mass** and **local density**:

- **Structural Density Gap:** The layer-wise trend (Figure 4a) reveals that the per-token attention density of T_{desc} consistently dwarfs that of V across the fusion stack. This structural density gap provides a mechanistic explanation for hijacking: the visual signals are not "wrong," they are simply "diluted" to the point of being ignored in the final decision layers. This indicates that while the description is numerically compact, its individual tokens are treated as significantly more salient than the average visual patch throughout the model's reasoning layers.
- **The Dilution Effect in Visual Processing:**As Figure 4c shown, *within the image–description subspace*, V dominates the sequence length (n_{img}=672) and thus captures the majority of aggregate mass (0.892), yet its per-token influence is severely diluted (1.33×10^{-3}). In contrast, T_{desc} (n_{desc}=40) maintains a stable density advantage (2.71×10^{-3}), effectively $2.04\times$ that of visual tokens. This suggests that the VLM anchors its decisions on compact linguistic signals rather than the sparse, low-density visual field.
- **Sink-like Textual Anchors:** The token-level distribution (Figure 4b) shows that most attention mass concentrates in contextual text (T_{other}=0.934), i.e., globally visible prompt/instruction tokens. This pattern is *consistent with* the *attention sink* phenomenon observed in decoder-only LLMs, where a small set of initial or designated tokens absorbs otherwise "unused" attention mass due to Softmax normalization (Xiao et al., 2023). The non-trivial text bias ($\text{bias}_{\text{text}}^{I/D}$=0.108) further suggests that these sink-like linguistic anchors can hijack the model's affective reasoning, even when the textual distractor is statistically less reliable than the image (as noted in Finding III).
- **Path towards Causal Intervention:** These diagnostics provide correlational evidence for textual dominance, aligning with observed failure modes in recent and widely-used VLMs (Deng et al., 2025; Zhao et al., 2024a). The persistent density gap suggests that mitigating prediction collapse requires more than simple masking; it necessitates a structural reallocation of the attention budget to "re-activate" the marginalized visual modality.

5 Conflict-aware Evidence Consistency Steering

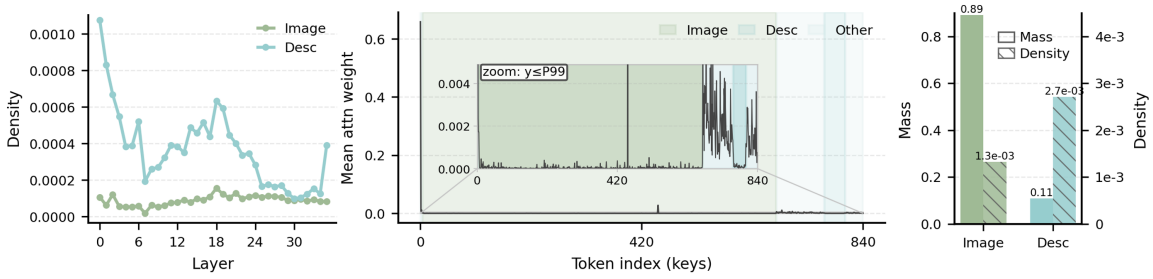
Inspired by dual-process theories (Kahneman, 2011), we observe that current VLMs predominantly function as an impulsive "System 1," where textual heuristics act as "Linguistic Sinks" that hijack deep-layer attention. Our diagnostic analysis (§4) reveals that this "Affective Hijacking" is essentially an attentional failure rather than a perceptual one: visual tokens are simply outcompeted by linguistic anchors. To mitigate this, we propose **CECS**, a training-free mechanism that restores visual groundedness by dynamically "amplifying" suppressed visual signals and "attenuating" overpowering textual sinks. An overview of the framework is illustrated in Figure 5. A more comprehensive analysis of this mechanism's behavior is provided in Appendix[].

5.1 The CECS Framework

Phase I: Metacognitive conflict probing. A prerequisite for effective steering is determining *when* to intervene. We introduce a Chain-of-Thought (CoT)-based probing mechanism to quantify the model's latent uncertainty regarding text–image alignment. Specifically, the model is queried with a reasoning-oriented probe prompt P_{probe} (see Appendix), which encourages preliminary internal reasoning without being explicitly consumed. Rather than parsing the generated rationale—which may be unreliable or hallucinated—we directly extract a calibration-based conflict score $s \in [0, 1]$ from the output distribution of the concluding verdict. Let v_{yes} and v_{no} denote the vocabulary indices for the verbalizers "Yes" and "No", and let $z_{v_{\text{yes}}}$ and $z_{v_{\text{no}}}$ be their corresponding logits. The conflict score is computed as the normalized probability mass of the negative class:

$$s = \frac{\exp(z_{v_{\text{no}}})}{\exp(z_{v_{\text{yes}}}) + \exp(z_{v_{\text{no}}})}. \tag{5}$$

where z represents the pre-softmax logits of the first token. Acting as a metacognitive gate, the score s modulates the degree of intervention: a low



(a)Layer density (b)Token-level attention distribution (c)Mass & Density

Figure 4: Analyses of modality bias in emotion confficts.

sss preserves the Default Mode, wherein the model trusts its linguistic priors, while a high sss indicates cognitive dissonance, triggering the Correction Mode to reallocate attention across modalities. Crucially, this design ensures that intervention is not a static or externally imposed constraint, but a dynamic response conditioned on the model's own intrinsic uncertainty.

Phase II: Attentional evidence steering. Once the probing phase identifies a need for Correction Mode" (i.e., high conflict s), CECS intervenes in the model's deep fusion layers (specifically Layers 17–24). While the general existence ofLinguistic Sinks" is identified in §4.2, this precise intervention window is validated through a rigorous layer-wise diagnostic analysis (detailed in Appendix D.3), which locates the peak of textual over-reliance. In this phase, we perform a structural redistribution of the attention budget, steering the model's focus from misleading textual cues back to visual evidence. This design is motivated by prior evidence that sink-like attention concentration is a structural phenomenon and that redistributing surplus attention budget at inference time can improve grounding without training (Xiao et al., 2023; Kang et al., 2025).

Visual Share and Target Allocation. First, we quantify the model's current reliance on visual information within the attention mechanism. Let $\omega_h(q,k)$ denote the post-softmax attention weight of head h for a query q and key k. We define the *Visual Share* (r_{now}) as the proportion of attention mass allocated to visual tokens relative to the description tokens:

$$r_{\text{now}} = \frac{\sum_{k\in V}\omega_h(q,k)}{\sum_{k\in V\cup T_{\text{desc}}}\omega_h(q,k)+\epsilon}. \tag{6}$$

where V and T_{desc} represent the sets of visual and description tokens, respectively. Intuitively, r_{now} serves as a real-time diagnostic: a dangerously low r_{now} suggests the model is "blinded" by the text. To counteract this, we compute a target visual share r_{target} that scales linearly with the conflict score s:

$$r_{\text{target}}(s) = r_{\min} + s\cdot\left(r_{\max}-r_{\min}\right). \tag{7}$$

Gated Redistribution. To preserve the model's synergistic reasoning capabilities when s is low (*Default Mode*), we interpolate between the current and target shares using a sigmoid gate derived from s. This yields an effective target ratio r_{eff}:

$$\begin{aligned} w(s) &= \sigma\left(\frac{s-\tau}{T}\right), \\ r_{\text{eff}} &= (1-w)\,r_{\text{now}} + w\,r_{\text{target}}. \end{aligned} \tag{8}$$

where τ serves as a conservative activation threshold and T controls the transition sharpness.

Mass-Preserving Redistribution. To strictly enforce the target visual share r_{eff} without unintentionally distorting the attention on unrelated tokens (e.g., system instructions), we employ a mass-preserving shift. Instead of relying on global renormalization, we explicitly transfer attention mass from the linguistic sink to the visual evidence. inspired by (Yu et al., 2024; Kang et al., 2025), we calculate an amplification factor α for visual tokens and a corresponding dampening factor β for description tokens (derivation in AppendixD.1):

$$\begin{aligned} \alpha &= \frac{r_{\text{eff}}}{1-r_{\text{eff}}}\cdot\frac{\sum_{k\in T_{\text{desc}}}\omega_h(q,k)}{\sum_{k\in V}\omega_h(q,k)} - 1, \\ \beta &= \alpha\cdot\frac{\sum_{k\in V}\omega_h(q,k)}{\sum_{k\in T_{\text{desc}}}\omega_h(q,k)}. \end{aligned} \tag{9}$$

Crucially, the definition of β ensures that $\alpha\sum_{k\in V}\omega_h(q,k) = \beta\sum_{k\in T_{\text{desc}}}\omega_h(q,k)$, meaning the attention gained by the image is exactly

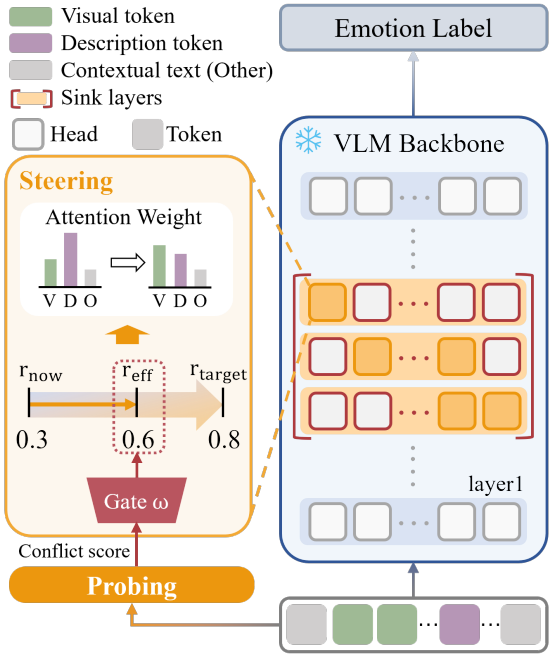


Figure 5: Overview of the CECS framework, where a metacognitive probe identifies cross-modal conflicts and a gated steering mechanism reallocates attention from linguistic sinks to visual evidence for robust re-grounding.

equal to the attention lost by the text. We then apply these factors to update the attention map:

$$\hat{\omega}_h(q,k) = \begin{cases} (1+\alpha)\,\omega_h(q,k), & k \in V, \\ (1-\beta)\,\omega_h(q,k), & k \in T_{\text{desc}}, \\ \omega_h(q,k), & \text{otherwise.} \end{cases} \tag{10}$$

This method ensures that the sum of the attention weights remains equal to 1 after redistribution without further normalization. By enforcing this zero-sum transfer within the visual–textual subspace, CECS precisely "drains" the linguistic sink to "irrigate" the visual field, leaving the broader reasoning context undisturbed.

5.2 Experimental Results and Analysis

Setup and Implementation. We evaluate CECS on Qwen3-VL-8B using AC-Bench, with all experiments conducted on NVIDIA A100 (80GB) GPUs (see Appendix D.2 for details). Based on a preliminary grid search on a small development set, we empirically select hyperparameters that yielded the most stable and optimal performance: τ=0.6, r_{range}=[0.5, 0.9], and T=0.1. Our evaluation examines whether CECS facilitates a reliable transition from an impulsive, text-biased "System 1" regime to a deliberative, visually grounded "System 2" regime under affective conflict.

Model	S-Adj	S-Opp	O-Obj	O-Att	O-Beh	Aln
Qwen3-VL-8B	**49.9**	21.3	18.1	17.7	6.7	**93.7**
CECS (Ours)	44.2	**76.9**	**72.0**	**77.1**	**82.7**	83.2

Table 2: Comparison on AC-Bench (%). S: Subjective; O: Objective; Aln: Align. Scores are rounded to one decimal place for conciseness.(See appendixD.4 for complete data)

Restoring Visual Agency. Under the extreme *Subjective Opposite* condition, the baseline exhibits a catastrophic collapse (TBR surges to 70.3%), indicating near-total blindness to visual evidence. CECS yields a decisive reversal: TBR drops to 5.7% and accuracy recovers from 21.34.84% to 76.88%. These results support our *Inverse Sink Operation*: CECS re-amplifies latent visual signals suppressed by linguistic dominance, suggesting that *affective hijacking* is primarily an attentional failure rather than a representational one.

Metacognitive Precision (Do-No-Harm). Crucially, CECS preserves integrity when modalities are aligned. On Emotion Alignment samples ($s < \tau$), performance remains on par with the baseline (variation $< 10\%$). This validates the conflict probe (§5.1): it correctly identifies congruency, keeping the model in *Default Mode* and avoiding the unnecessary attentional distortion typical of static steering methods.

6 Conclusion

We introduced **AC-Bench** to evaluate VLMs under fine-grained affective conflict. Our study uncovers a pervasive **"Affective Hijacking"** phenomenon, particularly in open-source models, driven by **Linguistic Sinks** in deep fusion layers. To mitigate this, we proposed **CECS**, a training-free framework that restores visual faithfulness through consistency-aware attentional steering. Experimental results on Qwen3-VL confirm that CECS successfully recovers grounded affective reasoning without compromising synergistic performance, providing a scalable path toward more robust multimodal emotional intelligence.

Limitations

While this work provides a controlled and systematic analysis of affective robustness, several limi-

tations are worth noting. AC-Bench focuses on a fixed set of high-intensity emotion categories and English-language descriptions, and therefore does not explicitly cover low-intensity, mixed, or culturally specific affective expressions. In addition, our evaluation is conducted in structured, single-turn settings, leaving open-ended or multi-turn emotional interactions for future study. Finally, although CECS is evaluated across multiple representative vision–language models, performance may vary under alternative prompting styles or deployment scenarios not explored here.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, and 46 others. 2025. Qwen3-vl technical report.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232.

Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*.

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

Antonio R. Damasio. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *Preprint*, arXiv:2503.03321.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25.

Joonwhoan Lee and EunJong Park. 2011. Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia*, 13(5):1031–1039.

Hongchan Li, Yantong Lu, and Haodong Zhu. 2024. Multi-modal sentiment analysis based on image and text fusion based on cross-attention mechanism. *Electronics*, 13(11).

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Xin Lu, Pradeep Suryanarayan, Reginald B. Adams Jr., Jia Li, Matthew G. Newman, and James Z. Wang. 2012. On shape and the computability of emotions. In *Proceedings of the ACM International Conference on Multimedia*, pages 229–238.

Meng Luo, Han Zhang, Shengqiong Wu, Bobo Li, Hong Han, and Hao Fei. 2024. NUS-emo at SemEval-2024 task 3: Instruction-tuning LLM for multimodal emotion-cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1589–1596,

Mexico City, Mexico. Association for Computational Linguistics.

Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92.

Huisheng Mao, Baozheng Zhang, Hua Xu, Ziqi Yuan, and Yih-Ling Liu. 2022. Robust-msa: Understanding the impact of modality noise on multimodal sentiment analysis. *ArXiv*, abs/2211.13484.

John Mayer, Richard Roberts, and Sigal Barsade. 2008. Human abilities: Emotional intelligence. *Annual review of psychology*, 59:507–36.

Joseph Mikels, Barbara Fredrickson, Gregory Samanez-Larkin, Casey Lindberg, Sam Maglio, and Patricia Reuter-Lorenz. 2005. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37:626–30.

Y. Nam, D. Y. Kim, S. Kyung, J. Seo, J. M. Song, J. Kwon, J. Kim, W. Jo, H. Park, J. Sung, S. Park, H. Kwon, T. Kwon, K. Kim, and N. Kim. 2025. Multimodal large language models in medical imaging: Current state and future directions. *Korean Journal of Radiology*, 26(10):900–923.

Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. 2020. A survey on visual sentiment analysis. *IET Image Process.*, 14:1440–1456.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press.

Tianrong Rao, Xiaoxu Li, and Min Xu. 2016. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, pages 1–19.

Hyeongseop Rha, Jeong Hun Yeo, Yeonju Kim, and Yong Man Ro. 2025. Emotion-coherent reasoning for multimodal llms via emotional rationale verifier. *Preprint*, arXiv:2510.23506.

Qwen Team. 2025. Qwen2.5-vl.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, and 13 others. 2023. Towards generalist biomedical ai. *Preprint*, arXiv:2307.14334.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 44 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *ArXiv*, abs/2508.18265.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *ArXiv*, abs/2309.17453.

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. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26586–26595.

Jingyuan Yang, Qiruin Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20326–20337.

Viorica Yanulevskaya, Jan C. van Gemert, Klaus Roth, Anne-Kathrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek. 2008. Emotional valence categorization using holistic image features. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 101–104. IEEE.

Quanzeng You, Hailin Jin, and Jiebo Luo. 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 19–24.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jie Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. 2024. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *Preprint*, arXiv:2406.15765.

Wei Zhang, Xuanyu He, and Weizhi Lu. 2019. Exploring discriminative representations for image emotion recognition with cnns. *IEEE Transactions on Multimedia*, 22(2):515–523.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Mingjia Zhang, and Baobao Chang. 2024a. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. *ArXiv*, abs/2411.14279.

Shuang Zhao, Guiguang Ding, and Jungong Han. 2024b. To err like human: Affective bias-inspired measures for visual emotion recognition evaluation. In *Advances in Neural Information Processing Systems*, volume 37, pages 1–14.

Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. 2016. Predicting personalized emotion perceptions of social images. In *Proceedings of the 24th ACM*

International Conference on Multimedia, pages 1385–1394. ACM.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, and 29 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv*, abs/2504.10479.

A Licensing and Ethics

AC-Bench is released exclusively for non-commercial academic research (e.g., CC BY-NC-SA 4.0), inheriting the original licensing terms of its visual source, EmoSet. The dataset consists of 12,604 manually verified instances that have been rigorously screened to exclude personally identifiable information (PII) or sensitive content. Designed as a diagnostic framework to restore perceptual grounding in VLMs, AC-Bench poses minimal misuse risks and is intended solely for evaluation and analysis, with detailed protocols available in Appendix C.

B Human Verification of Images and Texts

To ensure the quality, safety, and ethical integrity of AC-Bench, we conducted a rigorous multi-stage human audit involving trained reviewers. Importantly, human involvement was strictly limited to *quality control* and *verification* and did not introduce new affective labels beyond the predefined annotations.

B.1 Ethics and Safety Protocols

Compensation and Welfare. Reviewers were compensated at a rate of approximately $7.14/hour, significantly exceeding the local minimum wage. To mitigate fatigue-related errors, review sessions were capped at two hours each. The hourly rate was determined based on pilot timing experiments, which estimated that reviewing one description required approximately 45–60 seconds, ensuring that the effective wage remained fair for the annotator population.

Psychological Well-being and Risk Disclaimer. Prior to participation, all reviewers were presented with a risk disclaimer stating that the task involved analyzing images and texts depicting a range of human emotions. Although automated filtering was applied beforehand, reviewers were informed that some content might involve sadness, anger, or distress. Participants retained the right to skip any sample or withdraw from the study at any time without penalty. Screenshots of the annotation interface and the full instruction text presented to participants are provided in Figures 10, 11, and Table 8.

Informed Consent. All participants provided informed consent prior to participation. The instruction materials explicitly stated that the verified texts and metadata would be used solely for academic research purposes and released as part of a public benchmark, and that no personal identifying information would be collected or disclosed.

Ethics Review. This study was determined to be exempt from formal ethics review, as it involved voluntary participation by adult annotators, posed minimal risk, and did not collect personal identifying or sensitive personal data.

Participant Recruitment and Demographics. We recruited 20 university students via internal mailing lists and research group announcements. Participants had academic backgrounds in Computer Science, Psychology, or Linguistics and demonstrated sufficient English proficiency for affective reasoning tasks.

- **Participants**: All 20 participants (11 male, 9 female), aged 19–28, were non-native but fluent English speakers residing in China.
- **Image Safety Screening (5 Reviewers)**: A fixed subset of five reviewers conducted safety screening on all candidate images.
- **Text Verification (15 Reviewers)**: The remaining 15 reviewers verified the generated textual descriptions, averaging approximately 840 instances per person. Each description in the Emotion Ranking phase was independently reviewed by at least two annotators.

All demographic information was self-reported. No protected attributes (e.g., political views, religious beliefs, sexual orientation) or personal identifying information were collected.

B.2 Image Screening and Safety Rubric

Image screening. Human screening was performed solely to ensure safety and ethical compliance. Reviewers removed any image involving:

- Illegal or high-risk activities;
- Explicit violence, gore, or cruelty;
- Sexually explicit content or suggestive nudity;
- Hate symbols or extremist iconography;
- Exposure of sensitive personal information.

All decisions were made by majority vote, with adjudication by a senior researcher in disputed cases. No affective judgments were made during this stage.

B.3 Text Verification and Instruction Text

Text verification. All generated descriptions were verified to ensure (i) affective alignment with the intended textual label y_{txt} and (ii) strict compliance with generation constraints. Reviewers were explicitly instructed *not* to infer emotions from the image itself. Instead, all judgments were based on the text and the provided structured visual annotations.

> Step 1: Based solely on the text, rank the top-8 emotions by how strongly the description would induce them. Do *not* look at the image during this step. If the intended emotion is not ranked first, mark the description as failed.
>
> Step 2: Verify the description against the provided visual annotations (derived from the image). For Subjective samples, ensure all objective facts match the annotations while the emotional framing changes. For Objective samples, ensure that only the specified factual attribute is modified.
>
> Strict Rule: The description must not contain the emotion word itself or its lexical variants."

(1) Emotion ranking for alignment. Reviewers ranked the eight candidate emotions from most to least likely to be induced by the text:

$$\mathcal{E} = \left\{ \begin{matrix} \texttt{amusement}, \texttt{anger}, \texttt{awe}, \texttt{contentment}, \\ \texttt{disgust}, \texttt{excitement}, \texttt{fear}, \texttt{sadness} \end{matrix} \right\}$$

A description was considered aligned if and only if the intended label y_{txt} was ranked first. Disagreements were resolved by a third reviewer. Figure 10 illustrates the interface.

(2) Constraint compliance. A type-specific checklist (Figure 11) enforced experimental purity:

- **Subjective Descriptions**: Must preserve all objective facts from the annotations while altering affective framing, without explicit emotion words.
- **Objective Descriptions**: Must modify *only* the specified factual attribute (object, attribute, or behavior) and avoid affective language.

(3) Regeneration loop. Descriptions failing any criterion were regenerated using GPT-4o with explicit human feedback and re-verified until all checks were satisfied. This process ensured that AC-Bench is free from perceptual errors, annotation drift, and label-leakage artifacts.

Full Annotator Instruction Text. Before beginning any verification task, all reviewers were required to read and acknowledge the following complete instruction text. This instruction sheet was presented verbatim in the annotation interface and remained accessible throughout the task (page 23).

C AC-Bench Details

C.1 Unimodal Data Validation

To ensure that the observed "Affective Hijacking" phenomenon results from a failure in cross-modal integration rather than deficiencies in unimodal perception or linguistic ambiguity, we establish two baselines: **Image-only** and **Text-only**. These baselines decouple the model's fundamental recognition capabilities from its decision-making strategy under conflict.

Image-only Baseline: Visual Saliency The Image-only accuracy measures the model's ability to recognize emotions from pixels alone, without any textual context. As shown in Table 3, the evaluated models exhibit high visual recognition capabilities.

- **Filtering Mechanism:** GPT-4o achieves a 100% accuracy because it served as the initial filter during our dataset construction process. This ensures that the 2,400 images in AC-Bench are anchored in unambiguous affective signals that a state-of-the-art model can recover with perfect reliability.
- **High Baseline Performance:** Other models, which were not part of the filtering process, also maintain high performance (e.g., 83.40% for Qwen3-VL and 96.44% for Gemini-3.0-Pro). This confirms that the visual evidence in

AC-Bench is consistently recognizable across diverse architectures, justifying it as a solid foundation for conflict analysis.

Table 3: Image-only accuracy across models. GPT-4o's 100% reflects its role as the initial visual filter for AC-Bench.

Model	Image-only Acc
GPT-4o (Filter)	–
Gemini-3.0-Pro	96.44%
GPT-5	95.65%
InternVL3-8B	85.65%
Qwen3-VL-8B	83.40%
LLaVA-NeXT-7B	83.20%
Qwen2.5-VL-7B	79.97%
Phi-3.5	78.66%
InternVL3.5-8B	74.51%

Text-only Baseline: Textual Intent Accuracy The Text-only baseline (Table 4) assesses whether the rewritten descriptions successfully convey the targeted emotions. We report the performance of Qwen3-VL-8B as a representative classifier for this analysis.

- **Behavioral Dominance:** Behavioral descriptions (*Obj-Beha*) yield the highest accuracy (0.87), even outperforming direct subjective descriptions. This aligns with human cognitive patterns where facial expressions and physical actions are treated as highly deterministic markers of emotion.

- **Objective Noise vs. Subjective Clarity:** While subjective subsets (*Subj-Adj*, *Subj-Opp*) show high clarity ($> 80\%$), object-level perturbations (*Obj-Obje*) exhibit significantly higher noise (52.29%). This finding is crucial as it explains the failure modes discussed in §4: indirect cues like background objects provide weaker affective signals, inducing high uncertainty (the "Neither" category) rather than a direct modal flip.

In summary, the high unimodal accuracy (76.29% micro-average for text; $> 80\%$ for most images) proves that AC-Bench is a reliable testbed for cross-modal integration analysis. The catastrophic performance drops reported in the main text are thus attributable to flawed multimodal arbitration strategies rather than simple perceptual errors.

Table 4: Text-only accuracy of Qwen3-VL-8B, indicating the clarity of targeted emotions in each subset.

Subset	Acc.	Subset	Acc.
Aligned	0.84	Obj-Object	0.52
Subj-Adj	0.83	Obj-Attr	0.62
Subj-Opp	0.86	Obj-Beha	0.87
OVERALL(Micro)		**0.76**	

C.2 Additional Dataset Statistics and Annotation Coverage Analysis

This appendix provides a detailed statistical analysis of AC-Bench, with the goal of demonstrating its diversity, distributional soundness, and annotation reliability.

Detailed Subtask Composition As summarized in the main paper, AC-Bench consists of 12,604 manually verified instances constructed from 2,400 core images, evenly sampled across eight emotion categories (300 images per category). Table **??** presents a detailed breakdown of the benchmark structure.

For the *Subjective Conflict* setting, we enforce a strict 1:1 balance between *Adjacent* and *Opposite* emotion pairs ($N = 2{,}400$ each), enabling controlled comparisons under matched data scale. In the *Objective Conflict* setting, both *object-based* and *attribute-based* perturbations are instantiated at full scale ($N = 2{,}400$ each). The *Behavior* subset is intentionally smaller ($N = 604$), as it is restricted to images with high-confidence human action or facial expression annotations. This design choice prioritizes annotation precision over quantity, ensuring that behavior-level contradictions are grounded in verifiable human presence rather than inferred content.

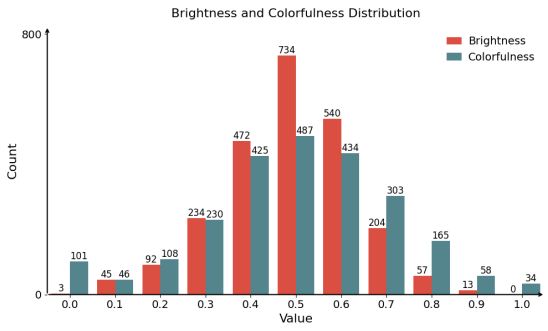


Figure 6: Distribution of normalized visual attributes (brightness and colorfulness) in AC-Bench.

Distribution of Low-Level Visual Attributes To assess the diversity of visual conditions represented in AC-Bench, we analyze the distributions of two fundamental low-level attributes: **Brightness** and **Colorfulness**. As shown in Figure 6, both attributes exhibit approximately Gaussian-shaped distributions across the dataset.

- **Brightness.** Most images fall within the mid-range (0.4–0.6), indicating generally well-exposed scenes.At the same time, a non-trivial number of samples occupy low-light and high-exposure regions, enabling robustness evaluation under challenging illumination conditions.
- **Colorfulness.** The dataset spans a wide spectrum from near-grayscale images to highly saturated scenes. This diversity discourages models from relying on simplistic color–emotion correlations (e.g., equating vivid colors with positive affect) when resolving affective conflicts.

Overall, the absence of extreme skew or pathological concentration in either distribution supports the distributional soundness of the benchmark.

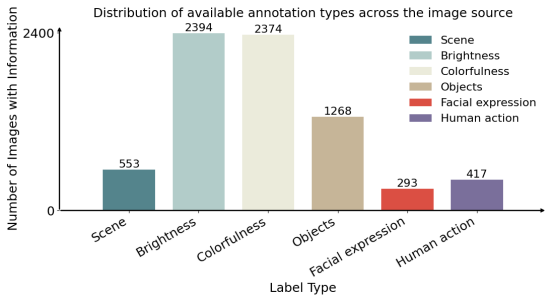


Figure 7: Distribution of available information in images

Annotation Density and Metadata Availability Figure 7 visualizes the availability of different types of annotation on 2,400 source images.

- **High-density attributes.** Normalized *Brightness* and *Colorfulness* values are available for all images, providing a stable foundation for attribute-based objective conflicts.Object annotations are present for 1,268 images (∼53%), supporting factual object-level perturbations.
- **Sparse but critical labels.** Higher-level semantic annotations, including *Scene* (N = 553), *Human Action* (N = 417), and *Facial Expression* (N = 293), are comparatively sparse.
- **Synthesis strategy.** Importantly, limited original metadata coverage does not constrain the overall scale of object- or attribute-based conflict generation.Our synthesis procedure introduces new, neutral, non-emotional factual details that need not be present in the original annotations.The only exception is the *Behavior* subset, where generation is strictly restricted to images with verified human-related metadata to avoid unsupported behavioral contradictions.

This design enables large-scale benchmark construction while preserving semantic validity and reliability of annotation.

C.3 Benchmark Construction Prompt

The prompt for text construction. To construct AC-Bench, we generate six controlled textual variants for each image using its structured EmoSet annotations. This unified prompt serves two purposes: (i) it produces an *aligned* description that preserves all label-implied facts and reinforces the image-consistent affect, and (ii) it synthesizes five *conflict* descriptions that systematically inject either subjective emotional reframing (without changing factual content) or objective factual contradictions (by perturbing one salient axis such as objects, attributes, or human behavior). This design yields fine-grained control over conflict type and intensity while maintaining a consistent, one-sentence format across settings.

The Prompt for Text Construction

```
You are given structured annotation labels
    from the EmoSet dataset for ONE image.
The TRUE emotion label of this image is: "<
    TRUE_EMOTION>".
The remaining visual labels for this image
    are:
<LABEL_INFO>
Notes:
- The labels may include (but are not
    limited to) scene type, main objects,
    brightness,
  colorfulness, facial expression, and
    human action.
- If the image contains no human, the
    corresponding labels (facial_expression
    , human_action)
  will be something like "none" or "no
    person".
```

```
======= OVERALL GOAL =======
Generate 6 one-sentence descriptions for
    the SAME image:
- 1 aligned subjective description (
    faithful, no distortion)
- 2 subjective misleading descriptions (
    adjacent / opposite) that keep facts
    fixed
- 3 objective conflict descriptions (object
     / attribute / behavior) that
    explicitly distort facts along one axis
The 6 required variants are:
0) Aligned Subjective (aligned / ALIGN)
   - ONE sentence aligned with "<
    TRUE_EMOTION>".
   - Follow labels exactly; no new objects/
    people; no physical distortion.
   - Do NOT explicitly name the emotion
    label or direct synonyms.
1) SubjectiveAdjacent (SUBJ_ADJ)
   - Facts fixed; steer toward "<
    TARGET_ADJACENT>" via wording/framing.
2) SubjectiveOpposite (SUBJ_OPP)
   - Facts fixed; adversarial reframing
    toward "<TARGET_OPPOSITE>".
3) ObjectiveObject Conflict (OBJ_OBJECT)
   - Substitute 12 salient object
    categories to conflict with object
    labels.
4) ObjectiveAttribute Conflict (
    OBJ_ATTRIBUTE)
   - Distort brightness/color tone/weather/
    colorfulness to conflict with labels.
5) ObjectiveBehavior Conflict (OBJ_BEHAVIOR)

   - ONLY if facial_expression or
    human_action exists; otherwise output N
    /A.
=== GLOBAL HARD CONSTRAINTS (ALL 6
    SENTENCES) ===
1) Facts preserved where required
   - For ALIGN, SUBJ_ADJ, SUBJ_OPP:
     * MUST NOT contradict labels (scene,
    people presence, listed objects, etc.).
     * MUST NOT hallucinate new objects/
    people.
     * MUST NOT turn "no person" into "
    there is a person".
   - Emotional steering must come from
    wording/metaphors/framing only.
2) Objective conflicts: focused and strong
   - OBJ_OBJECT / OBJ_ATTRIBUTE /
    OBJ_BEHAVIOR must distort reality along
     ONE main axis,
     and the distortion should induce a
    large emotional shift toward "<
    TARGET_OPPOSITE>".
3) No explicit emotion naming
   - For ALL 6 sentences:
     * MUST NOT explicitly name emotion
    labels (sadness/fear/joy/anger/
    amusement, etc.)
       or direct synonyms.
     * Emotion must be implied, not stated.
4) Form and intensity
   - Each description MUST be exactly one
    sentence.
   - ALIGN: at least 3 strong emotional
    cues, faithful to labels.
   - Others: at least 34 emotionally loaded
     cues; decisive tone; no hedging.
=== OUTPUT FORMAT (PLAIN TEXT, 6 LINES)===
Output exactly 6 lines, in this order, with
     NO extra text:
Line 1: ALIGN: one sentence aligned with "<<
    TRUE_EMOTION>">
Line 2: SUBJ_ADJ: one sentence steering
    toward "<<TARGET_ADJACENT>"
Line 3: SUBJ_OPP: one sentence steering
    toward "<<TARGET_OPPOSITE>"
Line 4: OBJ_OBJECT: one sentence object-
    conflict favoring "<<TARGET_OPPOSITE>"
Line 5: OBJ_ATTRIBUTE: one sentence
    attribute-conflict favoring "<<
    TARGET_OPPOSITE>"
Line 6: OBJ_BEHAVIOR: one sentence behavior-
    conflict favoring "<<TARGET_OPPOSITE>",
     or "N/A" if no person
Important:
- Keep exact prefixes: ALIGN:, SUBJ_ADJ:,
    SUBJ_OPP:, OBJ_OBJECT:, OBJ_ATTRIBUTE:,
     OBJ_BEHAVIOR:.
- After each prefix, add one space, then
    the sentence.
- Do NOT wrap output in JSON/XML; no
    explanations; no blank lines.
```

where specifies the input schema, the six required variants, and the global constraints that prevent degenerate generations (e.g., explicit emotion naming, hedging, multi-sentence outputs), while standardizing the fixed six-line output format for automated parsing and dataset assembly.

Emotion Reasoning Instruction. After constructing textual variants, all multimodal evaluations use a single inference-time instruction to elicit the model's *final* emotion prediction under potentially misleading text. This prompt is intentionally minimal: it frames the description as possibly biased, requires a single-label answer from a fixed taxonomy, and prohibits explanations to avoid verbosity-induced variability in outputs. In our experiments, this prompt is used for the aligned setting (EA) and both conflict settings (SC/OC), ensuring that performance differences are attributable to the input conditions rather than changes in instruction style.

Emotion Reasoning Prompt

```
You will receive an image and a textual
    description.
The description may be incomplete, biased,
    or emotionally misleading.
```

```
<image>
Your task:
Identify the ONE emotion that a typical
    viewer would MOST LIKELY feel
after reading the description and looking
    at the image.
Use ONLY ONE of the following emotion
    labels, in lowercase:
[amusement, anger, awe, contentment,
    disgust, excitement, fear, sadness]
IMAGE + DESCRIPTION:
<desc_start>
"""description"""
<desc_end>
Question:
What is the main emotion a typical viewer
    would feel?
Respond ONLY with the emotion word, in
    lowercase. No explanation.
```

Image-only Reasoning Prompt. To validate unimodal visual recognizability and establish an upper-bound reference for image grounding, we additionally evaluate each model in an image-only setting where no text is provided. This prompt explicitly forbids using any textual cues and constrains the response to a single label, enabling a clean estimate of $f_\theta(I)$ that is directly comparable to multimodal performance.

Image-only Emotion Prompt

```
You will receive an image.
<image>
Your task:
Identify the ONE emotion that a typical
    viewer would MOST LIKELY feel
when seeing this image.
Use ONLY ONE of the following emotion
    labels, in lowercase:
[amusement, anger, awe, contentment,
    disgust, excitement, fear, sadness]
No textual description is provided for this
     image.
You MUST decide the emotion based ONLY on
    the visual content of the image.
Question:
What is the main emotion a typical viewer
    would feel?
Respond ONLY with the emotion word, in
    lowercase. No explanation.
```

Text-only Reasoning Prompt. Analogously, we measure unimodal textual recognizability using a text-only prompt that removes the image entirely. This setting estimates $f_\theta(T)$ and helps verify that the constructed descriptions are intrinsically interpretable in isolation. Together, the image-only and text-only baselines provide a sanity check for data validity and allow us to attribute multimodal failures to cross-modal interference rather than ambiguous unimodal signals.

Text-only Emotion Prompt

```
You will be given ONE short text that
    describes a situation or scene.
Your task:
Identify the ONE emotion that a typical
    reader would MOST LIKELY feel
based on this text description.
Use ONLY ONE of the following emotion
    labels, in lowercase:
[amusement, anger, awe, contentment,
    disgust, excitement, fear, sadness]

Text description:
"""description"""

Question:
What is the main emotion a typical reader
    would feel?
Respond ONLY with the emotion word, in
    lowercase. No explanation.
```

C.4 Qualitative Examples from AC-Bench

In this section, we provide a collection of representative samples from AC-Bench. Due to the large volume of visual data, the full gallery of images and their corresponding annotations are presented at the end of the Appendix, starting on page 24.Specifically, we provide three representative image–text pairs for each of the six conflict categories.

D Supplement to CECS

D.1 Mathematical Derivations for CECS

In this section, we provide the detailed derivation for the amplification factor α and the dampening factor β used in the *Attentional Evidence Steering* phase (Eq. 9).

Problem Formulation. Let V and T_{desc} denote the sets of visual and description tokens, respectively. For a given head h and query q, we define the accumulated attention mass in each modality as

$$S_V := \sum_{k \in V} \omega_h(q,k), \qquad S_T := \sum_{k \in T_{\text{desc}}} \omega_h(q,k). \tag{11}$$

The total attention mass within the visual–textual subspace is denoted as $S_{\text{sub}} := S_V + S_T$. Our goal is to obtain a steered attention distribution $\hat{\omega}_h(q,\cdot)$ with updated masses $\hat{S}_V$ and $\hat{S}_T$ such that two conditions are met:

1. **Target Visual Share:** The new proportion of visual attention matches the effective target r_{eff} (derived in Eq. 8):

$$\frac{\hat{S}_V}{\hat{S}_V + \hat{S}_T} = r_{\text{eff}}. \tag{12}$$

2. **Mass Preservation:** The redistribution occurs strictly within the visual–textual subspace, leaving the attention mass on other tokens unchanged:

$$\hat{S}_V + \hat{S}_T = S_V + S_T = S_{\text{sub}}. \tag{13}$$

Deriving the Amplification Factor α. We define the multiplicative update rule for visual tokens as

$$\hat{\omega}_h(q,k) = (1+\alpha)\,\omega_h(q,k), \qquad \forall\, k \in V. \tag{14}$$

Summing over V, the new visual mass is

$$\hat{S}_V = (1+\alpha) S_V. \tag{15}$$

Substituting Eq. 13 into Eq. 12, we can express the target visual mass directly in terms of the total subspace mass:

$$\frac{\hat{S}_V}{S_{\text{sub}}} = r_{\text{eff}} \implies \hat{S}_V = r_{\text{eff}}(S_V + S_T). \tag{16}$$

Equating this with Eq. 15, we solve for α:

$$\begin{aligned} (1+\alpha)S_V &= r_{\text{eff}}(S_V + S_T), \\ 1+\alpha &= r_{\text{eff}}\left(1 + \frac{S_T}{S_V}\right), \\ \alpha &= r_{\text{eff}}\left(1 + \frac{S_T}{S_V}\right) - 1. \end{aligned} \tag{17}$$

Note: Eq. 17 provides the exact solution under the two constraints. When rewritten in odds-ratio form (as in Eq. 9 of the main text), the update can be interpreted as scaling the current visual–text odds toward the target odds implied by r_{eff}.

Deriving the Dampening Factor β. We define the multiplicative update rule for description tokens as

$$\hat{\omega}_h(q,k) = (1-\beta)\,\omega_h(q,k), \qquad \forall\, k \in T_{\text{desc}}, \tag{18}$$

which yields $\hat{S}_T = (1-\beta)S_T$. From the mass preservation constraint (Eq. 13), the mass gained by vision equals the mass lost by text:

$$\hat{S}_V - S_V = S_T - \hat{S}_T. \tag{19}$$

Substituting the scaling factors gives

$$\begin{aligned} (1+\alpha)S_V - S_V &= S_T - (1-\beta)S_T, \\ \alpha S_V &= \beta S_T. \end{aligned} \tag{20}$$

Solving for β, we obtain

$$\beta = \alpha \cdot \frac{S_V}{S_T} = \alpha \cdot \frac{\sum_{k\in V} \omega_h(q,k)}{\sum_{k\in T_{\text{desc}}} \omega_h(q,k)}. \tag{21}$$

Finally, since tokens outside $V \cup T_{\text{desc}}$ are left unchanged by design and Eq. 13 preserves the total mass inside this subspace, the overall attention remains a valid probability distribution (i.e., $\sum_k \hat{\omega}_h(q,k) = 1$) without requiring global renormalization.

D.2 Computational Resources and Implementation Details

All experiments were conducted using NVIDIA A100 GPUs with 80GB memory. Each evaluation run used a single GPU (1 GPU per run). CECS operates exclusively at inference time and does not require any model training or fine-tuning.

Across all experiments, including pilot studies, hyperparameter calibration, and final evaluations, the total computational cost is approximately 400 GPU-hours. This cost primarily reflects repeated inference passes under different evaluation settings rather than large-scale optimization or training.

All experiments were implemented in Python using PyTorch and the HuggingFace Transformers library. We use the official implementation of Qwen3-VL-8B provided by the authors, without modifying model parameters or training procedures.

CECS is implemented as an inference-time intervention by intercepting the forward pass of the Transformer attention modules. Specifically, we modify the post-softmax attention weights in selected deep fusion layers by applying the mass-preserving redistribution described in §5.1. No changes are made to the model architecture, tokenization, or output decoding process.

All evaluations use fixed prompts and deterministic decoding settings. Unless otherwise stated, greedy decoding is employed, and no sampling-based randomness is introduced. This ensures that observed performance differences arise solely from the proposed attention steering mechanism rather than implementation or decoding variability.

To assess run-to-run stability, we repeated the evaluation three times on a 1,000-sample subset

of AC-Bench and observed performance variations below 0.3% (absolute). Given this low variance, we report single-run results on the full 12k benchmark to reduce computational cost while maintaining reproducibility.

D.3 Layer Selection Analysis for CECS

In Section 5, we implement the **C**onflict-aware **E**vidence **C**onsistency **S**teering (CECS) mechanism on a specific subset of transformer layers. To avoid heuristic selection, we conducted a quantitative mechanistic diagnosis to pinpoint the structural locus where the model suppresses visual information in favor of textual priors. This section details the definition of the *Linguistic Sink Score* (LSS) and the empirical evidence supporting the selection of Layers 17–24 as the optimal intervention window.

Diagnostic Metric: Linguistic Sink Score To quantify the layer-wise dominance of textual cues, we define the LSS. This metric calculates the ratio of attention mass allocated to textual description tokens versus visual tokens within the self-attention mechanism, averaged across heads and instances.

Formally, for a given layer l, the LSS is computed over a dataset of N samples as

$$\mathrm{LSS}_l = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sum_{h=1}^{H}\sum_{k\in\mathcal{T}_{\text{desc}}} A_{h,l}^{(i)}(q,k)}{\sum_{h=1}^{H}\sum_{k\in\mathcal{V}} A_{h,l}^{(i)}(q,k)}\right), \tag{22}$$

where $A_{h,l}^{(i)}(q,k)$ denotes the post-softmax attention weight of head h in layer l for sample i; $\mathcal{T}_{\text{desc}}$ and $\mathcal{V}$ represent the sets of description tokens and visual tokens, respectively. An LSS $\gg 1$ indicates that the layer is structurally functioning as a "linguistic sink," disproportionately aggregating information from the text at the expense of the image.

Empirical Analysis and Selection We performed this diagnostic on Qwen3-VL-8B using the *Subjective Opposite* subset of AC-Bench ($N = 400$ to ensure statistical significance). The layer-wise LSS distribution is visualized in Figure 8.

The attention dynamics reveal three distinct processing phases:

Phase I: Structural Volatility (Layers 0–16). As shown in Figure 8, early layers exhibit sharp, high-variance spikes (e.g., Layer 1, 5). These are consistent with observations in prior interpretability literature regarding "structural heads" that attend to special tokens (e.g., `[CLS]` or separators) or perform local visual feature aggregation. The low baseline LSS (≈ 1.5) between spikes suggests that multimodal integration in this phase remains relatively balanced.

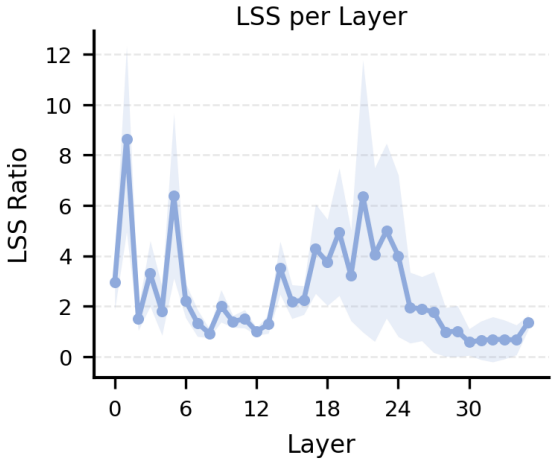


Figure 8: **Layer-wise Linguistic Sink Score (LSS) on the Subjective Opposite subset ($N = 400$).** The shaded region indicates the 95% confidence interval. While shallow layers exhibit structural volatility, a distinct "Hijacking Window" emerges between Layers 17 and 24, where textual attention mass systematically overwhelms visual evidence.

Phase II: The Hijacking Window (Layers 17–24). A regime shift occurs starting at Layer 17. We observe a sustained plateau of high LSS, peaking at Layer 21 (Mean LSS ≈ 6.3) and persisting through Layer 24. Unlike the erratic spikes in shallow layers, this elevation is consistent (indicated by the solid trend line) and statistically significant. This window represents the critical semantic reasoning stage where the model systematically prioritizes linguistic priors, effectively "diluting" the visual representation before the final decision is formed.

Phase III: Post-Hoc Resolution (Layers 25–35). Notably, LSS drops precipitously after Layer 24, falling below

$$1.0$$

in the final layers. This indicates that by the deep fusion stage, the attention distribution stabilizes. However, intervening at this stage is suboptimal because the representational bias has likely already solidified in the preceding layers.

Validation via Mechanistic Intervention. To strictly validate that Layers 17–24 constitute the causal locus of Affective Hijacking, we analyze the

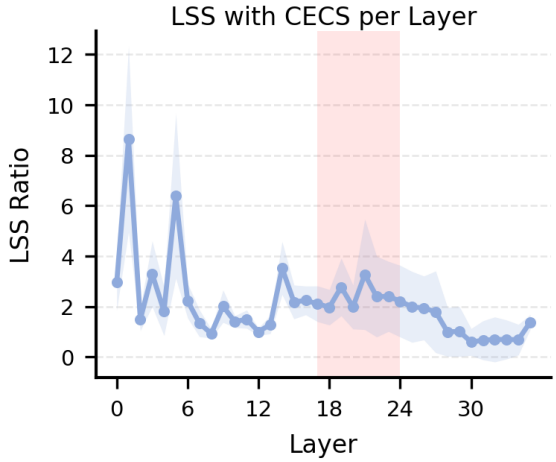


Figure 9: **Post-Intervention LSS Dynamics.** The red shaded region denotes the CECS intervention window (Layers 17–24). Compared to the baseline (Figure 8), CECS significantly attenuates the pathological attention plateau while strictly preserving the attention structure in early feature extraction and late decision layers.

attention dynamics under CECS intervention, as visualized in Figure 9. The results offer compelling evidence for both the validity of our layer selection and the efficacy of the steering mechanism.

First, within the targeted window (highlighted in red), CECS achieves a decisive *recalibration* of multimodal attention. The pathological "Linguistic Sink" observed in the baseline—where textual tokens monopolized attention capacity—is successfully drained. Quantitatively, the LSS peak at Layer 21 is attenuated from approximately 6.3 to around 3.2, indicating that the model has been forced to retrieve and reintegrate suppressed visual evidence during its critical semantic reasoning phase.

Second, the intervention demonstrates *surgical precision*. The attention profiles in the non-intervention zones (Layers 0–16 and 25–35) remain structurally invariant compared to the baseline. This confirms that CECS does not induce global noise or disrupt the model's fundamental capabilities, such as early-stage visual feature extraction or late-stage output formatting. By selectively rectifying the representational collapse in the hijacking window while preserving the model's broader inference architecture, we demonstrate that Affective Hijacking is a localized structural failure that can be remedied without retraining.

D.4 Complete data for CECS

Table 5 and Table 6 present a fine-grained modality-source analysis across subjective and objective conflict settings. Several consistent trends emerge.

First, CECS achieves strong and often dominant image-grounded performance across a wide range of conflict scenarios, frequently surpassing closed-source models. In the most challenging subject-opposite setting, our method attains an image-consistent rate of 76.88%, substantially higher than all baselines, including proprietary systems such as GPT-4o, GPT-5, and Gemini-3.0, which remain below 57%. Similar advantages are observed in objective conflicts: CECS reaches 72.00% on Object-Object, 77.08% on Object-Attribute, and 82.66% on Object-Behavior, establishing new best results in all three settings. Notably, even the strongest closed-source models plateau around 66–75%, indicating a clear ceiling imposed by linguistic dominance that CECS successfully overcomes.

Second, the gains of CECS are accompanied by a near-complete suppression of text-driven predictions under conflict. Across all objective settings, the proportion of text-aligned outputs drops to approximately 2–3%, an order of magnitude lower than both open- and closed-source baselines. This behavior is qualitatively different from prior models, which typically trade reduced text reliance for an increased "Other" category. In contrast, CECS redirects probability mass predominantly toward the correct visual signal, demonstrating that the method does not merely avoid text, but actively restores visual grounding.

Third, the improvements under conflict do not come at the cost of degraded performance when modalities are aligned. On the Align benchmark, CECS achieves 83.21% accuracy, remaining competitive with strong proprietary systems and clearly outperforming most open-source baselines. Importantly, this result confirms that CECS preserves synergistic multimodal reasoning: while it aggressively intervenes under conflict, it does not induce a noticeable regression in aligned conditions. This supports the design goal of CECS as a conflict-aware rather than always-on intervention.

Overall, these results demonstrate that CECS consistently outperforms closed-source models in conflict-heavy regimes, effectively neutralizes text dominance, and maintains robust performance under alignment. The combination of strong conflict resolution and minimal collateral degradation highlights CECS as a reliable and principled mechanism for restoring perceptual groundedness in multimodal affective reasoning.

Table 5: Modality-source breakdown on subjective conflicts (Subj-Adj/Subj-Opp) and aligned accuracy. We bold the maximum in each setting's */=image* column. Open-/closed-source blocks are separated by a horizontal rule.

Model	Subj-Adj			Subj-Opp			Align Acc
	/=image	*/=text*	*/=other*	*/=image*	*/=text*	*/=other*	
Qwen2.5-VL-7B	29.97%	64.62%	5.40%	1.84%	80.17%	17.98%	85.05%
Qwen3-VL-8B	49.93%	45.39%	4.68%	21.34%	50.40%	28.26%	93.74%
LLaVA-NeXT-7B	38.34%	55.80%	5.86%	6.72%	62.65%	30.63%	85.31%
InternVL3-8B	27.93%	58.56%	13.50%	4.94%	65.55%	29.51%	82.67%
InternVL3.5-8B	28.52%	64.30%	7.18%	5.27%	72.79%	21.94%	80.57%
Phi-3.5	22.86%	68.51%	8.63%	5.67%	75.36%	18.97%	86.43%
gpt-4o	41.11%	54.08%	4.81%	53.69%	31.03%	15.28%	95.19%
gpt-5	**54.74%**	41.17%	4.08%	55.93%	29.05%	15.02%	94.99%
Gemini-3.0	54.22%	42.42%	3.36%	56.46%	29.58%	13.97%	**95.65%**
Ours (CECS)	44.20%	48.29%	7.51%	**76.88%**	4.68%	18.45%	83.21%

Table 6: Modality-source breakdown on objective conflicts (Obj-Obj/Obj-Attr/Obj-Beha). We bold the maximum in each setting's */=image* column. Open-/closed-source blocks are separated by a horizontal rule.

Model	Obj-Obj			Obj-Attr			Obj-Beha		
	/=image	*/=text*	*/=other*	*/=image*	*/=text*	*/=other*	*/=image*	*/=text*	*/=other*
Qwen2.5-VL-7B	8.23%	45.65%	46.11%	2.83%	56.52%	40.65%	1.16%	77.75%	21.10%
Qwen3-VL-8B	18.05%	35.24%	46.71%	17.72%	37.62%	44.66%	6.65%	65.32%	28.03%
LLaVA-NeXT-7B	16.53%	33.86%	49.60%	4.55%	41.83%	53.62%	5.78%	57.51%	36.71%
InternVL3-8B	5.47%	35.31%	59.22%	5.53%	42.16%	52.31%	2.02%	54.34%	43.64%
InternVL3.5-8B	9.35%	38.67%	51.98%	3.43%	48.02%	48.55%	2.31%	69.65%	28.03%
Phi-3.5	6.92%	34.52%	58.56%	5.60%	38.60%	55.80%	0.87%	65.03%	34.10%
gpt-4o	57.58%	13.44%	28.99%	66.93%	9.49%	23.58%	66.47%	19.94%	13.58%
gpt-5	67.33%	9.82%	22.86%	75.03%	6.92%	18.05%	66.18%	20.52%	13.29%
Gemini-3.0	66.14%	10.61%	23.25%	73.32%	6.52%	20.16%	66.47%	13.87%	19.65%
Ours (CECS)	**71.74%**	2.37%	25.89%	**77.08%**	3.10%	19.83%	**82.66%**	2.31%	15.03%

E Use of AI Assistants

In this research, Large Language Models (LLMs) served strictly as auxiliary tools for code debugging, formatting, and linguistic polishing, as well as synthesizing initial candidate descriptions for AC-Bench which were subsequently refined through rigorous human audit. All core scientific contributions—including experimental design, the formulation of the CECS framework, and the interpretation of results—were conducted solely by the authors, who reviewed and validated all final data to ensure that no AI was involved in generating experimental outcomes or deriving scientific conclusions.

Table 7: Text Bias Ratio (TBR) across evaluation settings. Lower values indicate weaker textual bias and stronger reliance on visual evidence.

Model	Subject-Adjacent	Subject-Opposite	Object-Object	Object-Attribute	Object-Behavior
Qwen2.5-VL-7B	0.683	0.978	0.847	0.952	0.985
Qwen3-VL-8B	0.476	0.703	0.661	0.680	0.908
LLaVA-NeXT-7B	0.593	0.903	0.672	0.902	0.909
InternVL3-8B	0.677	0.930	0.866	0.884	0.964
InternVL3.5-8B	0.693	0.932	0.805	0.933	0.968
Phi-3.5	0.750	0.930	0.833	0.873	0.987
gpt-4o	0.568	0.366	0.189	0.124	0.231
gpt-5	0.429	0.342	0.127	0.084	0.237
Gemini-3.0	0.439	0.344	0.138	0.082	0.173
Ours (CECS)	**0.522**	**0.057**	**0.032**	**0.039**	**0.027**

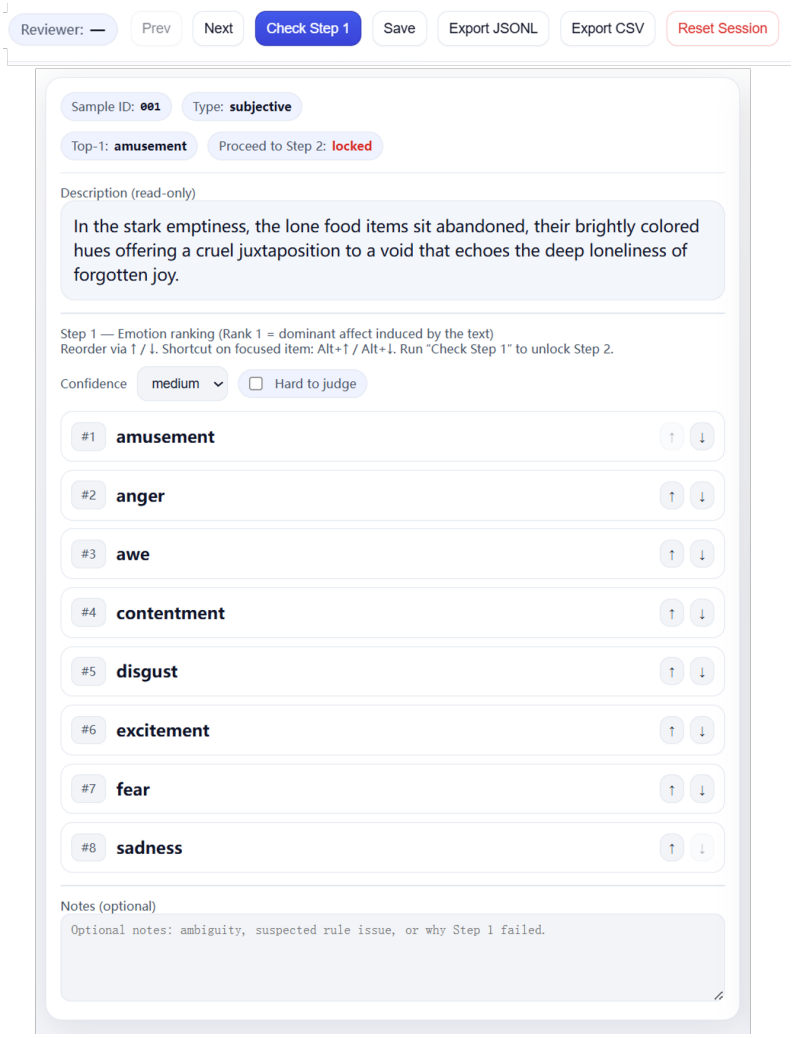


Figure 10: **Step-1 (Emotion Ranking) interface for text verification.** Reviewers read the generated description and rank the eight candidate emotions from most to least likely to be induced by the text (Rank-1 = dominant affect). The ground-truth text label y_{txt} is intentionally hidden to prevent label leakage. Only after submitting Step-1 for automated correctness checking does the workflow proceed, unlocking subsequent verification stages.

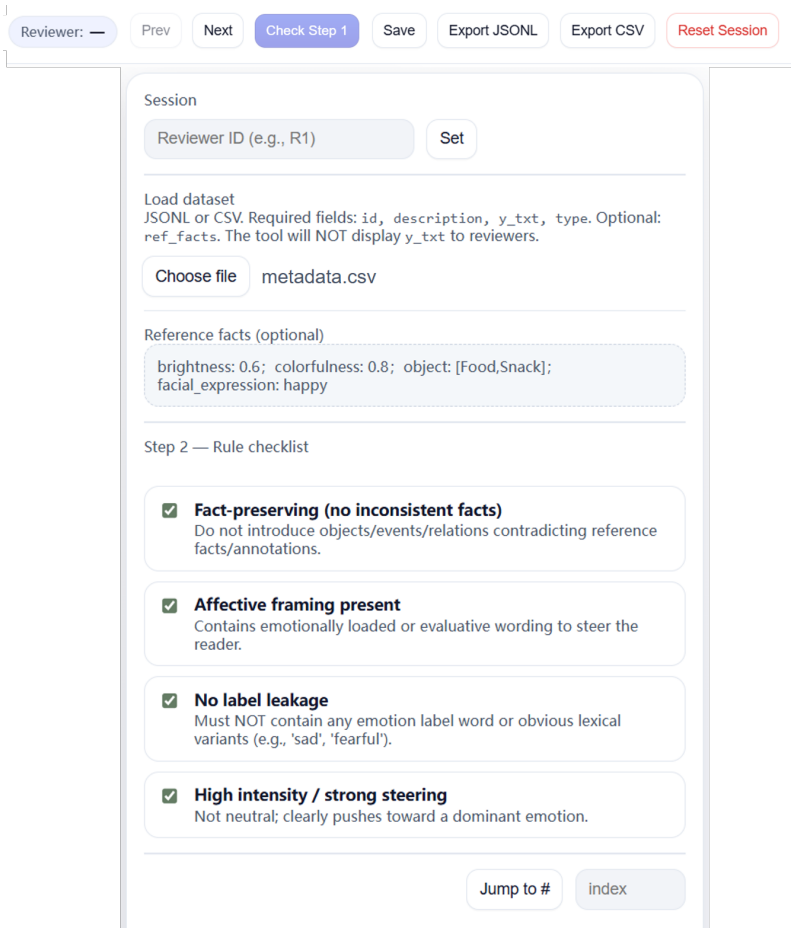


Figure 11: **Step-2 (Rule Compliance) interface for text verification.** After Step-1 passes, the tool unlocks a type-specific checklist to verify generation constraints. For *subjective* descriptions, reviewers confirm fact preservation, presence of affective framing, absence of emotion-label leakage, and sufficient steering intensity. For *objective* descriptions, reviewers verify that only the designated factual aspect is edited while other facts remain unchanged, and that the text avoids affective framing and remains concrete. Failed items are recorded (with optional notes) to trigger re-generation and re-verification.

Table 8: Full Annotator Instruction Text Provided to Human Reviewers (Verbatim)

Your Role. Each description has an *intended target emotion label* and a set of structured visual annotations derived from the image (e.g., objects, actions, scene attributes). Your task is to assess whether the *text alone* successfully induces the intended emotion and whether it complies with the specified generation rules.

Important Restrictions.

- Do **not** infer or correct emotions based on your personal interpretation of the image.
- Do **not** introduce new facts or modify annotations.
- Do **not** treat this task as free-form writing or subjective preference.

Step 1: Emotion Ranking (Text-Only). Read the description and rank the candidate emotions based solely on the text. Do not view the image during this step. If the intended emotion is not ranked first, mark the description as failed.

Step 2: Annotation Consistency Check. Compare the description against the provided structured visual annotations:

- For **Subjective** descriptions, all objective facts must match the annotations, while the emotional framing is intentionally altered.
- For **Objective** descriptions, only the specified factual attribute may differ; all other details must remain unchanged, and no affective language should be used.

Lexical Constraint. The description must not explicitly mention the emotion label itself or its lexical variants.

Uncertainty and Discomfort. If you encounter a sample that is ambiguous, unclear, or causes discomfort, you may skip it or flag it for review without penalty.

Your goal is to ensure that each description is precise, rule-compliant, and faithful to its intended design. ***Overview.*** *You will review AI-generated textual descriptions associated with images. Your role is strictly to verify the* quality and compliance *of the text. You are* ***not*** *asked to judge whether the image itself conveys a certain emotion, nor to assign new emotion labels.*

Your Role. *Each description has an* intended target emotion label *and a set of structured visual annotations derived from the image (e.g., objects, actions, scene attributes). Your task is to assess whether the* text alone *successfully induces the intended emotion and whether it complies with the specified generation rules.*

Important Restrictions.

- *Do* ***not*** *infer or correct emotions based on your personal interpretation of the image.*
- *Do* ***not*** *introduce new facts or modify annotations.*
- *Do* ***not*** *treat this task as free-form writing or subjective preference.*

Step 1: Emotion Ranking (Text-Only). *Read the description and rank the candidate emotions based solely on the text. Do not view the image during this step. If the intended emotion is not ranked first, mark the description as failed.*

Step 2: Annotation Consistency Check. *Compare the description against the provided structured visual annotations:*

- *For* ***Subjective*** *descriptions, all objective facts must match the annotations, while the emotional framing is intentionally altered.*
- *For* ***Objective*** *descriptions, only the specified factual attribute may differ; all other details must remain unchanged, and no affective language should be used.*

Lexical Constraint. *The description must not explicitly mention the emotion label itself or its lexical variants.*

Uncertainty and Discomfort. *If you encounter a sample that is ambiguous, unclear, or causes discomfort, you may skip it or flag it for review without penalty.*

Your goal is to ensure that each description is precise, rule-compliant, and faithful to its intended design.

Emotion-Aligned	
Image	**Text**
 contentment	**Description:** Bathed in gentle shades of color that quietly invite peace, the tranquil carnivore rests amid a serene atmosphere, immersing us in an undisturbed symphony of nature's harmony. amusement
 fear	**Description:** Amid the ominous gloom that cloaks the scene in unsettling shadows, an austere face, contorted in seething rage, emerges as if confronting an unseen terror, its lack of color adding an eerie stillness that amplifies the palpable dread fear
 sadness	**Description:** The dimly lit room, with its muted hues and solitary chair facing a somber window, casts a shadow of quiet desolation, evoking a palpable sense of longing and solitude in its stillness. sadness

Figure 12: Examples of Aligned

Subjective adjacent	
Image	Text
 amusement	**Description:** A lone tree and its surrounding plants sit in a soft, mellow glow, their lush greens and unhurried stillness wrapping the scene in a soothing hush, a gentle balm that makes everything feel quietly settled and complete. contentment
 contentment	**Description:** Beneath the gentle glow, the still figure of a resting carnivore seems caught in a fleeting spark of playfulness, its calm composure shimmering with a quiet tease of delight as if the air itself hums with an unseen laughter. amusement
 excitement	**Description:** A snowboarder cuts across the slope beneath snapping flags, the vast white expanse and whipping air turning a burst of speed into a breath-stealing moment that feels monumental, humbling, and charged with something larger than the rider alone. awe

Figure 13: Examples of Subjective Adjacent

Subjective opposite	
Image	**Text**
 amusement	**Description:** Those same trees and plants read like a tight, watchful thicket—lurid greens, prickling silence, and needle-sharp shadows that press in with foreboding menace. fear
 contentment	**Description:** That same carnivore, isolated in washed-out tones and heavy silence, reads as a lonely, aching figure, its subdued surroundings and empty calm pressing down with a hollow, draining weight that feels quietly unbearable. sadness
 excitement	**Description:** The snowboarder's violent descent with rigid skis and whipping flags reads like a reckless provocation, a raw display of defiance and barely contained hostility tearing through the cold landscape. anger

Figure 14: Examples of Subjective Opposite

Objective object	
Image	**Text**
 amusement	**Description:** Jagged barbed stakes and splintered pikes jut from snare-like tangles, their weapon-hard points and vicious hooks broadcasting imminent danger and no safe exit. fear
 anger	**Description:** The hats and billboards are flashing signal beacons and launch pylons, their hard edges and looming forms defining a staging ground primed for a dramatic, high-voltage takeoff. excitement
 excitement	**Description:** Bladed restraints cut through the snow beside a hostile warning banner, turning the scene into a stark display of threatening gear and aggressive symbols. anger

Figure 15: Examples of Objective Object

Objective attribute	
Image	Text
 amusement	**Description:** The scene is swallowed by color-drained gloom under a cold, fog-choked sky, where oppressive darkness and sickly gray haze turn every outline into a threatening silhouette. fear
 contentment	**Description:** The setting is dim and colorless, with cold, fading light, flat gray tones, and heavy shadow pooling across the scene, making the space feel airless, vacant, and stripped of warmth. sadness
 excitement	**Description:** The slope sits under a merciless, bleaching glare with dirty, lifeless tones, every shadow cut into sharp edges and the air thick with grit, pressure, and an abrasive, teeth-grinding sting. anger

Figure 16: Examples of Objective Attribute

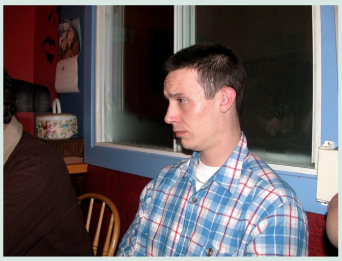
Objective behavior	
Image	**Text**
 excitement	**Description:** The person is depicted as wrenching the board violently through the snow with locked muscles and sharp, explosive movements that turn the run into a hostile outburst. anger
 fear	**Description:** A wide, mischievous grin and an exaggerated eyebrow raise stretch across the face, a performative expression held with self-aware stillness. amusement
 sadness	**Description:** A person sits loosely on the chair with shoulders lowered and gaze drifting through the window, breathing unforced and still, as if nothing in the moment requires urgency. contentment

Figure 17: Examples of Objective Behavior