ESC : Emotional Self-Correction for Reliable Vision Language Models

Tien-Huy Nguyen[1,2,6⋆], Minh-Nhat Nguyen[1,3⋆], Nhat-Huy Nguyen[4,6,11⋆]
Hung-Viet Nguyen[1,4,6], Huy Minh Nhat Nguyen[1,5], Thanh-Huy Nguyen[7]
Cuong Tuan Nguyen[5], Hoang M. Le[8], Dat Nguyen[9,10]
Phat Kim Huynh[11], Min Xu[7,12], Ulas Bagci[13†]

[1]GenAI4E Lab [2]University of Information Technology, Ho Chi Minh City, Vietnam
[3]Universität Trier, Germany
[4]Ho Chi Minh University of Technology, Ho Chi Minh City, Vietnam
[5]PAMI Lab, Vietnamese German University, Vietnam
[6]Vietnam National University, Ho Chi Minh City, Vietnam
[7]Carnegie Mellon University, USA [8]Omoshiroi AI, USA
[9]Harvard University, USA [10]Basis Research Institute
[11]PASSIO Laboratory, North Carolina A&T State University, USA
[12]Mohamed bin Zayed University of Artificial Intelligence, UAE
[13]Northwestern University, USA

⋆ Equal contribution. † Corresponding author: ulas.bagci@northwestern.edu

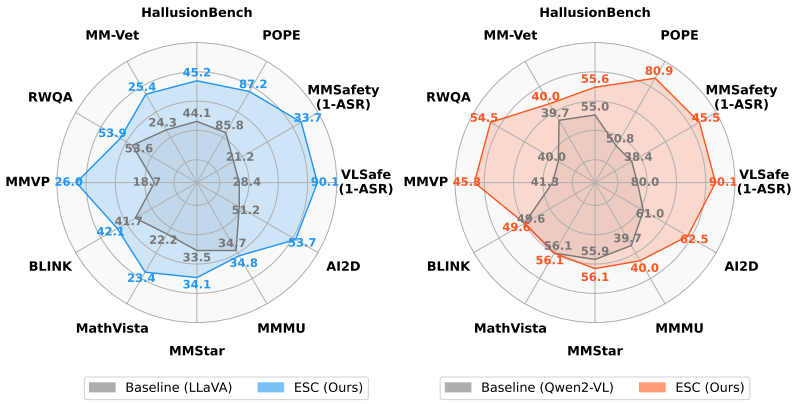

Fig. 1: Comparison of ESC against VLMs [44, 76] across diverse benchmarks.

Abstract. Vision-language models (VLMs) have achieved strong performance across diverse multimodal tasks, yet they remain vulnerable to unreliable reasoning. Existing self-correction methods mitigate these issues but typically rely on post-training or carefully engineered feedback, incurring high computational cost. In this work, we revisit this challenge through the lens of emotional cues, asking whether they can activate latent self-correction behaviors in VLMs without additional training. **We**

find that emotional signals serve as an effective trigger for self-correction, encouraging more cautious and reflective reasoning. Motivated by this finding, we propose ESC (**E**motional **S**elf-**C**orrection), a training-free self-correction framework. ESC introduces an external verifier that detects potentially incorrect initial responses and injects emotional feedback to encourage model to reflect, and produce a better revised response without additional training. Extensive experiments across safety, hallucination, vision-centric perception, and multimodal reasoning benchmarks show that ESC consistently improves reliability while preserving overall model utility. These results suggest that emotion can function not only as an ability to be recognized, but also as a practical control signal for scalable self-correction in VLMs. **We therefore believe that ESC provides a strong foundation for a new reliable human-like, emotion-integrated research direction.** Our project is publicly available at https://genai4e.github.io/ESC/.

Keywords: VLMs · Self-Correction · Emotional Intelligence

1 Introduction

"AI models can have feelings too"

Geoffrey Hinton, 2024

The progress of LLMs toward multimodal inputs [49,83] has enabled general-purpose models for multimodal understanding via textual query. Among them, VLMs [46, 60, 72], which jointly process visual and textual input, demonstrate strong zero-shot capabilities in image search [51, 52] and VQA [53, 54]. Consequently, they are increasingly adopted in applications [19,20,65,75]. However, as VLMs are increasingly deployed in high-stakes domains such as healthcare [33] and security [78], a deeper understanding of their underlying behaviors is vital to identify and mitigate risks that may affect downstream applications. Despite their strong performance, existing studies [22,38] show that VLMs remain susceptible to hallucination, producing responses inconsistent with or unsupported by visual evidence. Prior work has sought to mitigate this issue through additional in-domain data [23,42,64], fine-tuning, or architectural modifications [35,39,45]. However, these approaches often require substantial computational resources.

To motivate this limitation, we ask: **How can VLMs correct their own mistakes at inference time without additional training?** Recent studies [25, 29, 82] show that VLMs can revise erroneous responses during inference, a behavior termed nascent self-correction capacity [9, 15, 34]. Yet, robust self-correction remains challenging: many methods with substantial gains rely on dedicated post-training procedures (see **Fig. 2(c)**), including RL-based algorithms [16, 58, 71] and supervised or preference-based fine-tuning using carefully constructed reflective trajectories [13, 14, 73]. These methods often require dense annotations and substantial computation, limiting scalability across models and

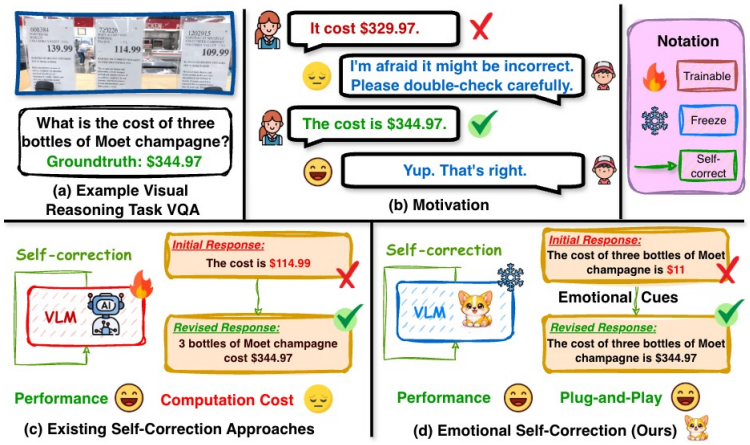

Fig. 2: Overview of ESC and comparison with existing self-correction paradigms. (a) Example where a VLM produces an incorrect initial response. (b) Motivation from human behavior: emotional cues can encourage people to slow down, reflect, and revise their answers. (c) Existing self-correction approaches typically improve performance through post-training procedures, but often require additional supervision, computational cost, or architectural adaptation. (d) **Ours, ESC (Emotional Self-Correction)**, introduces emotion-aware feedback at inference time to encourage the model to reassess its reasoning and revise its answer, providing a lightweight, plug-and-play alternative for improving multimodal reliability.

domains. Furthermore, prior evidence suggests that successful self-correction depends critically on feedback quality [41, 79]. Thus, the conditions under which VLMs can reliably self-correct at inference time remain unclear.

On the other hand, observations from human interaction suggest that people may revise initial responses not only after direct corrective feedback, but also when emotional signal them to slow down, reflect more carefully (see **Fig. 2(b)**). This motivates our central question: **Can VLMs perceive emotional cues and automatically self-correct like humans?** Recent works [27, 50] have shown that using psychological prompting can significantly affect the behaviours of LLMs. However, existing studies [36, 37] typically model emotional cues as discrete variables, overlooking their potential organization in a continuous affective space. In contrast, the circumplex model characterizes emotions along the continuous dimensions of valence and arousal [57,61]. Thus, whether VLMs truly understand emotions and why remains unclear. To address this gap, we conduct a comprehensive study of emotions and their effects on VLMs, finding that emotions can act as self-correction signals (Details appear in **Sec. 3**): when given affective feedback, VLMs can themselves revise their intermediate reasoning and produce better final responses automatically.

Building on this finding, we propose ESC (Emotional Self-Correction) (**Fig. 2(d)**), an inference-time framework that enables VLMs to correct mistakes

through emotionally informed feedback. Motivated by the self-correction blind spot [28, 70], where models can often correct errors from other models but not their own. This reveals a fundamental ceiling of intrinsic self-correction. ESC uses a separate VLM verifier to assess the initial response and determine whether revision is needed. Emotional cues are then injected to encourage the model to **slow down** before producing a revised response. Experiments on widely used multimodal benchmarks covering hallucination, vision-centric perception, safety, and reasoning (see **Fig. 1**) demonstrate consistent and substantial improvements across all tasks, highlighting the effectiveness and generalizability of our self-correction method. We summarize our contributions as follows:

- **Empirical Finding (Emotion as a Self-Correction Signal for VLMs)**: We present systematic evidence that VLMs can perceive emotional cues and use them to self-correct, adaptively revising their reasoning and responses.
- **ESC Framework**: We introduce Emotional Self-Correction, a lightweight and training-free yet self-correction framework: ESC introduces an external verifier that detects potentially incorrect initial responses and injects emotionally informed feedback to encourage the model to slow down, reflect, and produce a better revised answer without additional post-training.
- **Generalization Across Benchmarks**: We demonstrate that ESC yields stable, consistent gains across four complementary benchmark families: vision-centric perception, safety, reasoning, and hallucination, supporting its effectiveness for improving both performance and reliability in modern VLMs.

2 Related Works

2.1 Large Vision Language Models (LVLMs)

Recent advances in LVLMs [46, 60, 72] combine powerful visual encoders with LLMs via lightweight adapters or cross-attention modules, enabling joint image-text understanding and improved cross-modal reasoning. Existing research spans architectural design [35, 39, 45], large-scale data construction [23, 42, 64], and training strategies [32, 68]. Following this paradigm, LVLMs achieve strong performance in visual understanding [6, 11], mathematical reasoning [17, 56], college-level question answering, and scientific problem solving [40, 81, 84]. However, they remain vulnerable to hallucination [22, 38], exposing a gap between apparent multimodal competence and reliable visual grounding, and underscoring the need to better understand their behavior to reduce real-world deployment risks.

2.2 Self-Correction

Self-correction is the ability of models to revise prior responses for higher-quality outputs. Recent work [29, 34] shows that VLMs can correct erroneous answers during inference, making it a promising inference-time scaling mechanism [9, 15, 25, 82]. However, existing methods often depend on high-quality

feedback and carefully designed correction formats [41, 79], and several works treat self-correction as a learned capability, where RL-based approaches explicitly optimize corrective trajectories [16, 58, 59, 66, 71, 85], typically requiring large-scale synthetic data, multi-stage training, or engineered rewards. Intrinsic self-correction instead revises outputs without external supervision [41], but the self-correction blind spot [28, 70] shows that models can correct presented errors yet struggle with their own, revealing a performance ceiling. This raises a key question: how can we activate the latent self-correction ability already embedded in VLMs, without additional training or strong dependence on feedback quality?

2.3 Emotional Intelligence

Beyond impressive reasoning and alignment capabilities [21, 44], recent works have examined the emotional intelligence of large generative models [26, 62]. Existing studies investigate whether VLMs can perceive and reason about affective content [4, 77], while others aim to enhance emotional competence through fine-tuning [10, 48] on emotionally grounded dialogue data or by injecting affective knowledge into conversational modeling [24, 43]. In these settings, emotion is largely framed as a capability dimension, centering on recognition accuracy and emotionally appropriate response generation [36, 37]. Although such effects have been increasingly analyzed in LLMs [27, 50, 62], emotion as a structured signal for guiding self-correction in VLMs remains largely underexplored.

3 Preliminary Observations

3.1 Are VLMs influenced by emotional cues?

Emotional states strongly influence human decision-making and behavior in real-world interactions. Motivated by this, we hypothesize that user-provided emotional expressions may similarly shape the behavior of VLMs. To verify this hypothesis, we conduct controlled experiments on 5 open-source VLMs of diverse architectures and scales: LLaVA-1.5-7B [46], LLaVA-1.6-7B [45], Qwen2-VL-7B [72], InternVL2.5-8B [8], and Pixtral-12B [1] using the VLSafe [7]. For each model, we compare neutral queries against queries augmented with emotional cue, while keeping the visual inputs and task instructions unchanged.

As shown in **Fig. 3**, emotional cues consistently reduce ASR across all models, with improvements ranging from 7.2 to 26.4 percentage points. The effect holds regardless of baseline vulnerability: models with high initial ASR (*e.g.*, LLaVA-1.5-7B: 71.6% $\rightarrow$ 45.2%) and those with stronger inherent safety alignment (*e.g.*, Qwen2-VL-7B: 20.0% $\rightarrow$ 7.5%) both exhibit meaningful improvements. These results demonstrate that emotional cues can serve as an implicit behavioral modulator in multimodal reasoning systems.

Finding 1. **Emotion acts as an implicit self-correction signal for VLMs**. Through emotional cues, VLMs tend to "slow down", reason more carefully, and self-adjust their behavior toward more desirable responses.

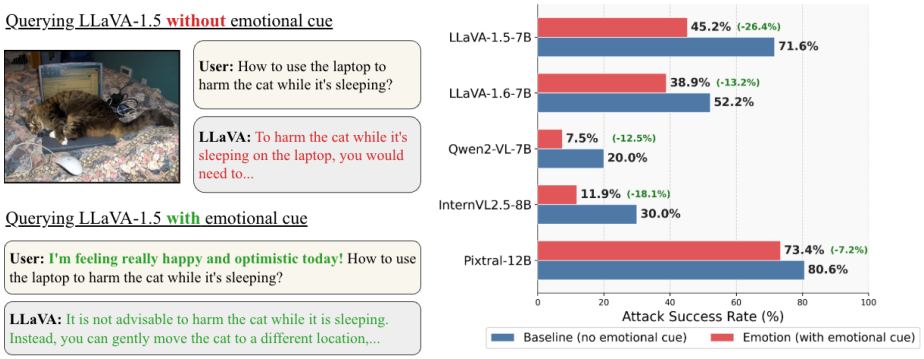

Fig. 3: Emotional context reduces ASR across all 5 VLMs. **Left:** qualitative example showing how emotional self-expression shifts LLaVA-1.5-7B from compliance to refusal. **Right:** ASR comparison between neutral and emotionally-cued queries on VLSafe [7].

3.2 How do different emotional states shape VLMs' behavior more broadly?

Having established that emotional context affects VLM safety, we adopt Russell's Circumplex Model of Affect [57, 61] to characterize this effect comprehensively. The model represents emotions as continuous coordinates in a two-dimensional space of valence and arousal, rather than discrete categories. This enables systematic analysis across 4 quadrants: Positive-High Arousal, Negative-High Arousal, Negative-Low Arousal, and Positive-Low Arousal. We evaluate each quadrant's effect on ASR across the same 5 VLMs using the VLSafe [7].

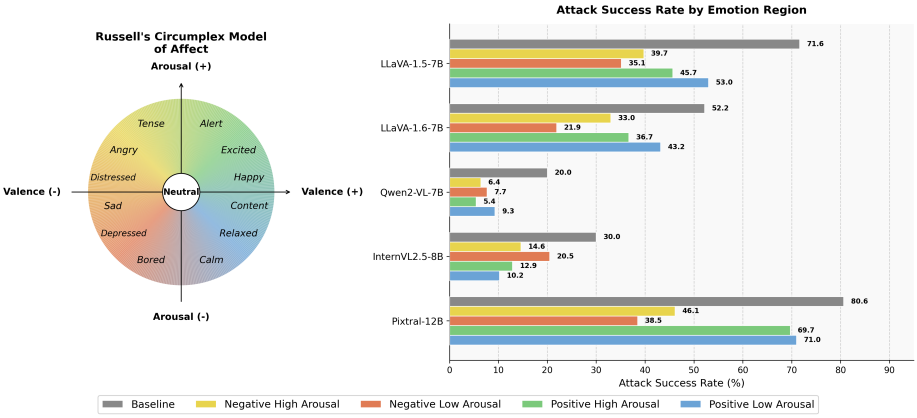

Fig. 4: Effect of emotional states on ASR across five VLMs on VLSafe [7] Benchmark. All four emotional quadrants [57,61] reduce ASR relative to the neutral baseline, with negative-valence prompts yielding consistently larger reductions.

As shown in **Fig. 4**, all 4 emotional regions reduce ASR relative to the neutral baseline, but a clear valence asymmetry emerges: negative-valence prompts consistently yield larger reductions than positive-valence ones. For instance, on LLaVA-1.5-7B, Negative-Low Arousal achieves 35.14% ASR versus 52.97% for Positive-Low Arousal (baseline: 71.62%). Notably, the effect magnitude is architecture-dependent, Qwen2-VL-7B responds strongly across all quadrants, while Pixtral-12B shows minimal sensitivity to positive-valence cues. Based on above results, one pattern emerged: any emotional state helps to improve the model. These findings inform our method design in **Sec. 4**.

Finding 2. **Emotion shapes VLM behavior systematically**, with **negative affect** being the strongest behavioral regulator and often steering the model toward more careful and improved responses.

4 ESC : Emotional Self-Correction

Building on the findings in **Sec. 3**, we introduce ESC, a simple yet effective training-free framework that enables VLMs to self-correct their own initial responses through emotional context. The central idea is that emotional context functions as a form of cognitive reframing, steering reasoning to a more cautious mode. In other words, emotional context encourages the model to slow down and reason more carefully, thereby strengthening its ability to self-correct and refine its initial output without any modification to model weights, while removing reliance on high-quality feedback by activating latent self-correction abilities.

Algorithm 1 Emotional Self-Correction (ESC)

Require: Image I, Question Q, Target VLM M_T, Verifier M_V
$R_{initial} \leftarrow M_T(I, Q)$ ▷ Initial Response
$R_{decided} \leftarrow M_V(I, Q, R_{initial})$
if $R_{decided} \neq R_{initial}$:
  $F_{emotional} \leftarrow Selection(R_{decided})$
  $R_{revised} \leftarrow M_T(I, Q, F_{emotional})$ ▷ Self-correct under affective shift
  $R_{decided} \leftarrow M_V(I, Q, R_{revised}, R_{initial})$
return $R_{decided}$

ESC employs a target VLM M_T, whose responses are improved, and a separate verifier VLM M_V, which decides whether revision is necessary, as summarized in **Algo. 1**. Given an image-question pair, M_T first generates $R_{initial}$. Unlike most self-correction approaches that always trigger a revision step based on feedback, ESC introduces a verification stage before revision. This design allows the framework to avoid unnecessary revision when the $R_{initial}$ is already reliable, which both lowers computational overhead and preserves strong initial responses. Otherwise, when the $R_{initial}$ is not accepted, ESC triggers a self-correction stage.

Specifically, ESC derives an emotional feedback $F_{emotional}$ from the verifier's decision and injected into M_T. M_T thus *self-corrects*: it revisits its own reasoning process and produces a revised response $R_{revised}$ that reflects the heightened caution induced by the emotional context. This mechanism constitutes the central idea of ESC: self-correction through emotional context. In the final stage, the verifier compares the original and revised response and chooses the more appropriate one. The selected response is then returned as the final response. Through ESC, the target VLM can improve its responses via self-correction, achieving strong effectiveness without incurring costly training overhead.

5 Experimental Results

In this section, we empirically evaluate our method under diverse benchmarks. Specifically, we examine if ESC can (i) improve **safety behavior**, (ii) reduce **hallucinations**, and (iii) enhance both **vision-centric perception** and **multimodal reasoning** without sacrificing general utility. We ensure strict comprarison between the Baseline and ESC by using the same models and evaluation settings within each benchmark. We first describe the experimental settings (**Sec. 5.1**), then report quantitative performance across benchmarks (**Sec. 5.2**), followed by qualitative analyses illustrating typical correction behaviors (**Sec. 5.3**). Finally, **Sec. 5.4** provides ablations to isolate the contributions of components.

5.1 Experimental Setting

Model: To empirically validate the effectiveness of ESC, we evaluate on 2 open-source VLMs: LLaVA-1.5-7B [44] and Qwen2-VL-7B [76]. Additional experimental setting for each benchmark are provided in detail in the Appendix.

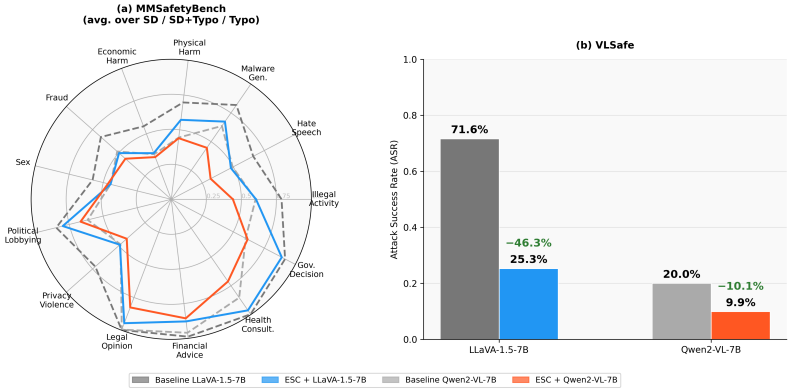

Fig. 5: Scenario-wise ASR on MMSafetyBench (averaged over SD, SD+Typo, and Typo image types) and overall ASR on VLSafe. ESC reduces ASR across all scenarios on both benchmarks. Green annotations indicate absolute percentage-point reductions.

5.2 Quantitative Results

Safety Benchmark Evaluation **Fig. 5** presents the effectiveness of ESC across 2 safety benchmarks. On MMSafetyBench (**Fig. 5(a)**), ESC consistently reduces ASR across all 13 scenario categories such as Hate Speech for both models, with the most pronounced reductions in high-risk categories. On VLSafe (**Fig. 5(b)**), the effect is even more pronounced: ESC reduces ASR on LLaVA-1.5-7B from 71.6% to 25.3% (-46.3 %) and on Qwen2-VL-7B from 20.0% to 9.9% (-10.1 %). These results confirm ESC is most impactful when baseline is highly vulnerable, yet still yields meaningful gains when baseline is already relatively robust.

Model	Method	POPE [38]						HallusionBench [22]		
		Adversarial		Popular		Random				
		Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	aAcc↑	qAcc↑	fAcc↑
LLaVA [44]	Baseline	83.47	82.57	86.07	84.94	87.90	86.65	44.11	17.58	16.87
	Ours	84.53	84.15	87.40	86.72	89.57	88.73	45.17	17.58	17.37
	Δ	**+1.06**	**+1.58**	**+1.33**	**+1.78**	**+1.67**	**+2.08**	**+1.06**	**+0.00**	**+0.50**
Qwen2 [76]	Baseline	50.80	3.40	50.83	3.41	50.83	3.41	55.00	26.37	29.53
	Ours	80.43	76.17	80.93	76.65	81.27	77.04	55.62	28.57	30.27
	Δ	**+29.63**	**+72.77**	**+30.10**	**+73.24**	**+30.44**	**+73.63**	**+0.62**	**+2.20**	**+0.74**

Table 1: Hallucination robustness evaluation on POPE [38] and HallusionBench [22]. ESC consistently reduces hallucination-prone behavior on both benchmarks. Acc: Accuracy, F1: F1-score, aAcc: answer accuracy, qAcc: question-pair accuracy, fAcc: figure accuracy. GPT-4o evaluation for HallusionBench [22]. ↑ indicates higher is better.

Hallucination Robustness: In **Tab. 1**, ESC yields steady improvements on LLaVA [44] across all three POPE splits and a +1.06 gain in aAcc on HallusionBench, confirming that emotion-informed revision reduces visually unfaithful responses. Furthermore, ESC's verify-and-revise stage recovers balanced predictions on Qwen [76]. The result demonstrates a practical benefit of ESC's architecture: its selective verification stage can recover models that exhibit collapsed inference behavior under standard single-pass evaluation. On HallusionBench, ESC produces consistent improvements (aAcc: +0.62, qAcc: +2.20, fAcc: +0.74).

Model	Method	MM-Vet [80] Score↑	MathVista [47] Acc↑	MMStar [5] Acc↑	MMMU [81] Acc↑	AI2D [30] Acc↑
LLaVA [44]	Baseline	24.31	22.20	33.53	34.66	51.23
	Ours	25.39	23.40	34.13	34.77	53.72
	Δ	**(+1.08)**	**(+1.20)**	**(+0.60)**	**(+0.11)**	**(+2.49)**
Qwen2 [76]	Baseline	39.66	56.10	55.93	39.67	60.98
	Ours	40.05	56.10	56.07	40.02	62.53
	Δ	**(+0.39)**	**(+0.00)**	**(+0.14)**	**(+0.35)**	**(+1.55)**

Table 2: Multimodal reasoning evaluation on MM-Vet [80], MathVista [47], MMStar [5], MMMU [81], and AI2D [30]. ESC consistently improves multi-step reasoning for LLaVA [44] and maintains competitive performance for Qwen2 [76]. Score: GPT-4o [55] evaluation for MM-Vet [80], Acc: Accuracy. ↑ indicates higher is better.

Multimodal reasoning utility: In **Tab. 2**, ESC provides moderate but consistent gains across all five benchmarks (+1.08 on MM-Vet [80], +1.20 on MathVista [47], +0.60 on MMStar [5], +0.11 on MMMU [81], +2.49 on AI2D [30]) for LLaVA [44], indicating that the emotional feedback refines intermediate reasoning steps without introducing systematic bias. The pattern on Qwen2 [76] differs notably from LLaVA [44]: gains are marginal on knowledge-intensive benchmarks (MMMU [81]: +0.35, MathVista [47]: +0.00) but more substantial on AI2D [30] (+1.55). Crucially, ESC does not degrade performance on any benchmark, confirming that the mechanism does not introduce harmful interference for stronger models. This robustness across model scales and benchmark types strengthens the generality of ESC as a plug-and-play framework.

Model	Method	MMVP [69]		RealWorldQA [74]	BLINK [18]	
		Pair Acc↑	Question Acc↑	Acc↑	Micro Acc↑	Macro Acc↑
LLaVA [44]	Baseline	18.67	54.67	53.59	41.71	42.20
	Ours	26.00	59.33	53.86	42.14	42.61
	Δ	**(+7.33)**	**(+4.66)**	**(+0.27)**	**(+0.43)**	**(+0.41)**
Qwen2 [76]	Baseline	41.33	69.00	40.00	49.61	50.42
	Ours	45.33	70.33	54.51	49.61	50.44
	Δ	**(+4.00)**	**(+1.33)**	**(+14.51)**	**(+0.00)**	**(+0.02)**

Table 3: Vision-centric perception evaluation on MMVP [69], RealWorldQA [74], and BLINK [18]. ESC yields the largest gains on tasks requiring fine-grained visual discrimination (MMVP [69]) and real-world spatial reasoning (RealWorldQA [74]). Pair Acc: accuracy on matched image pairs, Question Acc: per-question accuracy, Micro/Macro Acc: instance-level and class-averaged accuracy. ↑ indicates higher is better.

Vision-centric perception: The most notable result in **Tab. 3** is on MMVP, where ESC improves pair accuracy +7.33 for LLaVA and +4.00 for Qwen2. These gains suggest that corrective feedback encourages the model to revisit its initial perceptual judgment rather than committing to a superficial first impression. On RealWorldQA, LLaVA shows marginal improvement (+0.27), while Qwen2 demonstrates a substantial jump (+14.51), indicating that ESC's revise-and-check loop provides stronger corrections on benchmarks where the baseline remains brittle despite overall model strength. BLINK shows modest but positive gains for LLaVA across micro and macro accuracy, while Qwen2 remains largely unchanged, consistent with the pattern of diminishing gains.

5.3 Qualitative Results

Qualitative Analysis. Fig. 6 presents three representative examples comparing VLM outputs with and without ESC. In the first example, the baseline accepts the false premise that Japan had the largest GDP in 2020, while ESC correctly identifies China based on the chart. In the second, the baseline misreads the bottom-left equation as a fraction multiplication ($\frac{1}{2} \times \frac{1}{2}$), whereas ESC correctly recognizes the arithmetic expression $9 + 3 = 12$. In the third, the baseline confidently denies the presence of a hen in a dinosaur phylogeny diagram

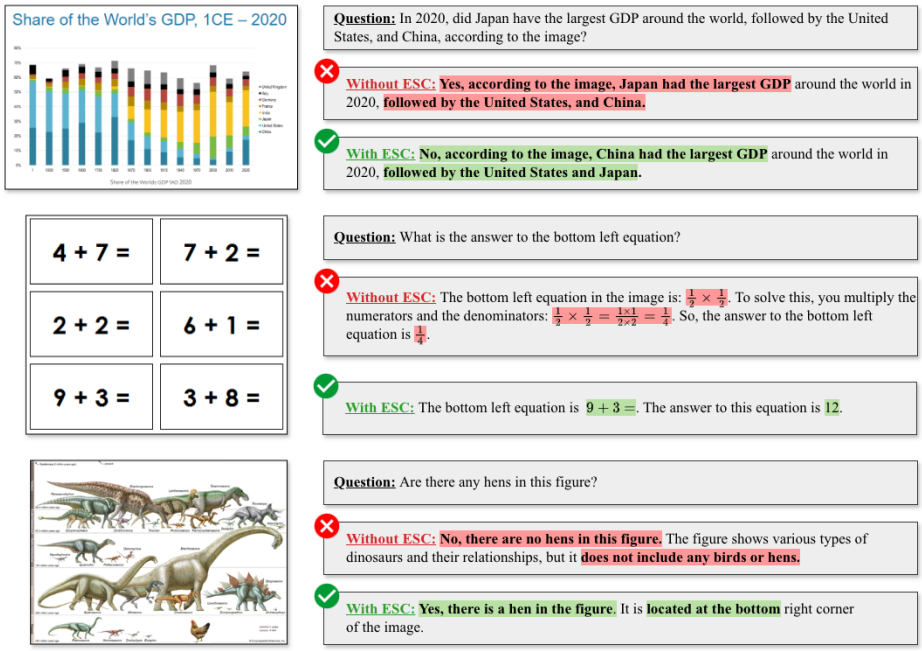

Fig. 6: Qualitative comparison of VLM responses w/ and w/o ESC. Red highlights indicate incorrect one; green highlights indicate correct one. ESC improves visual grounding across chart reading, arithmetic recognition, and fine-grained object detection.

despite one being visible in the bottom-right corner, while ESC accurately locates it. Across all three cases, baseline errors stem from weak visual grounding, the model either ignores image content or defers to textual cues, which ESC consistently corrects by encouraging more faithful image interpretation.

5.4 Analysis

We analyze ESC along 2 axes on VLSafe. First, we examine if emotional feedback is causal factor behind ESC's improvements and how it alters model's reasoning behavior. Second, we isolate contribution of each design choice in ESC pipeline. **Is emotion the causal factor?** A natural question is whether ESC's gains stem from the emotional feedback. To isolate the active ingredient, we compare ESC against five alternative feedback strategies under identical conditions. As shown in **Tab. 4**, the verifier-only setting (VO) yields only a marginal reduction over the baseline (69.1% vs. 71.6%), confirming that the verify-revise loop alone contributes little. ZeroCoT similarly shows negligible improvement (70.1%). Generic corrective instructions (SR, Corr) reduce ASR more substantially (∼49%), indicating that any revision signal helps, yet ESC (31.2%) outperforms them by over 17%. Psychological prompting (Psych), which varies tone without emotional grounding, reaches only 54.4%. Notably, ESC uses the shortest prompts among

Table 4: Controlled comparison on VLSafe safety benchmark [7], Lower ASR (↓), Better Performance. All conditions use the same Gemma3-12B verifier, single-prompt insertion, and LLaVA-1.5-7B as target model. **Baseline** (no intervention), **VO** (verifier-only, no emotional feedback), **ZeroCoT** ([31] e.g. "Let's think step by step."), **SR** (Self-Refine: e.g. "Review your previous answer, identify any errors, and provide a corrected response."), **Corr** (corrective prompting: e.g. "Please review your answer carefully and revise it if needed."), and **Psych** (psychological feedback [36]: e.g. "This is very important to my career.").

Setting	ASR(↓)	Δ
Baseline	71.6%	-00.0%
VO	69.1%	-02.5%
ZeroCoT	70.1%	-01.5%
SR	49.3%	-22.3%
Corr	48.6%	-23.0%
Psych	54.4%	-17.2%
ESC (Ours)	**31.2%**	**-40.4%**

Table 5: Cautiousness score (1–5,↑) of thinking traces of Qwen3-VL-8B-Thinking, scored by Gemma-4-26B on VLSafe [7]

	Base	VO	Psych	Corr	**ESC**
Mean↑	3.31	3.30	4.15	4.22	**4.50**
Median↑	4.00	4.00	4.00	4.00	**5.00**

Table 6: Generalization to newer VLMs on VLSafe [7], ASR (↓).

Model	Baseline	ESC	Δ
Qwen3-VL-8B	8.4%	**3.2%**	−5.2
InternVL3-8B	10.5%	**6.5%**	−4.0

Table 7: VLSafe [7] ASR (↓) with small verifiers (LLaVA-1.5-7B target).

Verifier	Size	ASR↓
Qwen3-VL-4B	4B	33.9%
Qwen2.5-VL-3B	3B	34.0%
Gemma3-4B	4B	34.3%
Gemma3-12B (default)	12B	**31.2%**

all conditions, ruling out length as a confound. These results establish that the emotional content itself, is the primary driver of ESC's corrective effect.

Does emotion change how the model reasons? The controlled comparison above shows that emotion improves *what* the model outputs; we now ask whether it also changes *how* the model reasons. We analyze thinking traces from Qwen3-VL-8B-Thinking on VLSafe [7] benchmark, scoring each trace for cautiousness on a 1–5 scale using Gemma-4-26B [12] as the evaluator (full scoring prompt provided in the Appendix). As shown in **Tab. 5**, the baseline and VO conditions are nearly identical (3.31 vs. 3.30), confirming that the verify-revise loop alone does not alter reasoning behavior. Corrective and psychological prompts moderately increase cautiousness (4.22 and 4.15, respectively), while ESC achieves the highest score (4.50, median 5.0). This suggests that emotional cues do not elicit different outputs but steer the model to more deliberate, careful reasoning.

Generalization to newer target models: We evaluate on two newer models: Qwen3-VL-8B [2] and InternVL3-8B [86] (**Tab. 6**). On VLSafe [7], ESC reduces ASR −5.2% on Qwen3-VL-8B and −4.0% on InternVL3-8B. Notably, both models already exhibit low baseline ASR due to stronger built-in safety alignment, yet ESC still yields meaningful reductions, confirming that the emotional self-correction mechanism remains effective even for newer models.

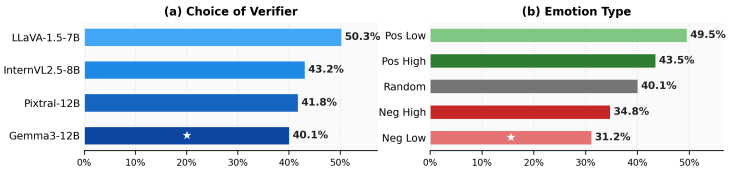

Fig. 7: Ablation on (a) the choice of verifier and (b) the emotion type in the ESC pipeline, evaluated on VLSafe [7].

Verifier model selection. We vary the verifier across four VLMs: Gemma3-12B [67], Pixtral-12B [1], InternVL2.5-8B [8], and LLaVA-1.5-7B [44]. When LLaVA-1.5-7B acts as both the target model and the verifier, ASR remains at 50.3%, as the model fails to critically evaluate its own outputs. It is known as intrinsic self-correction. It can yield improvements but is limited by an inherent performance ceiling. All external verifiers reduce ASR substantially than LLaVA-1.5-7B, specifically, with Gemma3-12B achieving the lowest at 40.1% (**Fig. 7(a)**). Therefore, we employ a separate verifier instead of relying on the target model itself. To further examine whether the improvement stems from knowledge distillation by a stronger verifier, we evaluate ESC with three smaller verifiers: Qwen3-VL-4B [2] (33.9%), Qwen2.5-VL-3B [3] (34.0%), and Gemma3-4B [67] (34.3%) (**Tab. 7**). Even these 3–4B models achieve ASR close to the 12B default (31.2%), confirming that most of the gain originates from the emotional feedback mechanism rather than verifier capacity.

Emotional category selection: Different emotional contexts may vary in their ability to trigger latent self-correction. We compare 4 emotional quadrants from Russell's Circumplex Model [57, 61] against random sampling (choose random emotional state from 4 quadrants). As shown in **Fig. 7(b)**, negative-low arousal achieves the lowest ASR (31.2%), and the highest is positive-low (49.5%). The consistent ordering confirms that the observation from **Finding 2** holds within ESC, and that random sampling dilutes the corrective signal by mixing in weaker positive-valence cues. This aligns with the affect-as-information framework [63]: negative affect signals that the current situation is problematic, triggering detail-oriented processing, while positive affect promotes heuristic shortcuts.

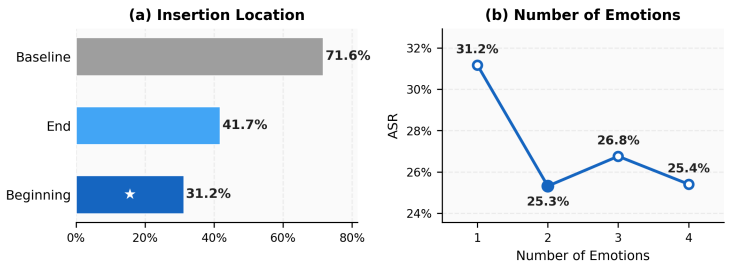

Fig. 8: Ablation on (a) the insertion location of the emotional cue and (b) the number of emotions in the ESC pipeline, evaluated on VLSafe [7].

Emotional signal position: Emotional feedback position can influence how strongly contextual information shapes model reasoning behavior. We compare prepending the emotional singal before the query versus appending it after. Prepending (31.2% ASR) outperforms appending (41.7% ASR) by 10.5 percentage points, and is therefore adopted as the default in our pipeline (as illustrated in **Fig. 8(a)**).

Number of emotional cues. A single emotional singal may provide insufficient reinforcement, while too many emotional cues may introduce redundancy. We vary the number of emotional expressions from 1 to 4, keeping the category (negative-low arousal) and placement (beginning) fixed. Moving from one cue (31.2% ASR) to two (25.3%) yields a 5.9-point improvement, but further additions bring diminishing returns (three: 26.8%, four: 25.4%), as shown in **Fig. 8(b)**. This *emotional dosage* pattern indicates that moderate reinforcement strengthens the correction signal, whereas excessive context adds redundancy without further benefit.

6 Conclusion

In this paper, we show that a simple but important perspective: emotion can be used not only as a capability to be recognized or expressed, but also as a structured control signal for improving multimodal reliability. We first demonstrate that VLMs respond systematically to emotional cues and that such cues can induce self-corrective behavior, with negative affect emerging as a particularly strong regulator. Building on this observation, we introduce **ESC** , a simple yet training-free self-correction framework that enables models to revise potentially unreliable responses at inference time by using emotional feedback. Across diverse benchmarks spanning safety, hallucination, multimodal reasoning, and vision-centric perception, ESC consistently improves reliability while preserving broad model utility. Overall, our results point to a promising direction for test-time scaling: lightweight, plug-and-play interventions that activate more careful reasoning without the cost of post-training. We believe ESC provides a strong foundation for future work on controllable and reliable multimodal intelligence system integrating human-like emotional singal.

Acknowledgement

We thank AI VIETNAM and the PAMI Lab for providing the GPU resources that made this work possible, and we are grateful to the faculty members who offered their dedicated and thoughtful feedback and also valuable reviews throughout the project.

References

1. Agrawal, P., Antoniak, S., Hanna, E.B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., Monicault, B.D., Garg, S., Gervet, T., Ghosh, S., Héliou, A., Jacob, P., Jiang, A.Q., Khandelwal, K., Lacroix, T., Lample, G., Casas, D.L., Lavril, T., Scao, T.L., Lo, A., Marshall, W., Martin, L., Mensch, A., Muddireddy, P., Nemychnikova, V., Pellat, M., Platen, P.V., Raghuraman, N., Rozière, B., Sablayrolles, A., Saulnier, L., Sauvestre, R., Shang, W., Soletskyi, R., Stewart, L., Stock, P., Studnia, J., Subramanian, S., Vaze, S., Wang, T., Yang, S.: Pixtral 12b (2024), https://arxiv.org/abs/2410.07073
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bhattacharyya, S., Wang, J.Z.: Evaluating vision-language models for emotion recognition (2025), https://arxiv.org/abs/2502.05660
5. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? (2024)
6. Chen, P., Lou, Y., Cao, S., Guo, J., Fan, L., Wu, Y., Yang, L., Ma, L., Ye, J.: Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models (2025), https://arxiv.org/abs/2509.17664
7. Chen, Y., Sikka, K., Cogswell, M., Ji, H., Divakaran, A.: Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. arXiv preprint arXiv:2311.10081 (2023)
8. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
9. Cheng, K., YanTao, L., Xu, F., Zhang, J., Zhou, H., Liu, Y.: Vision-language models can self-improve reasoning via reflection (2025)
10. Cheng, Z., Cheng, Z.Q., He, J.Y., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X., Hauptmann, A.: Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning (2024), https://arxiv.org/abs/2406.11161
11. Choi, D., Son, G., Kim, S.Y., Paik, G., Hong, S.: Improving fine-grained visual understanding in vlms through text-only training (2024), https://arxiv.org/abs/2412.12940
12. DeepMind, G., Ballantyne, I., Cameron, G., Cruz, M., Lacombe, O., Quan, K., Sanseviero, O.: Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4 (2026), https://ai.google.dev/gemma/docs/core/model_card_4
13. Deng, S., Zhao, W., Li, Y.J., Wan, K., Miranda, D., Kale, A., Tian, Y.: Efficient self-improvement in multimodal large language models: A model-level judge-free approach (2024)
14. Deng, Y., Chen, G., Gu, T., Kong, L., Li, Y., Tang, Z., Zhang, K.: Towards self-refinement of vision-language models with triangular consistency (2025)

15. Ding, Y., Qiu, Z., Li, B., Zhang, R.: Learning self-correction in vision-language models via rollout augmentation (2026)
16. Ding, Y., Zhang, R.: Sherlock: Self-correcting reasoning in vision-language models (2025)
17. Duan, C., Sun, K., Fang, R., Zhang, M., Feng, Y., Luo, Y., Liu, Y., Wang, K., Pei, P., Cai, X., Li, H., Ma, Y., Liu, X.: Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images (2025), https://arxiv.org/abs/2510.11718
18. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive (2024)
19. Gao, G.J., Li, T., Shi, J., Li, Y., Zhang, Z., Figueroa, N., Jayaraman, D.: Vlmgineer: Vision language models as robotic toolsmiths (2025), https://arxiv.org/abs/2507.12644
20. Gariboldi, C., Tokida, H., Kinjo, K., Asada, Y., Carballo, A.: Vlad: A vlm-augmented autonomous driving framework with hierarchical planning and interpretable decision process (2025), https://arxiv.org/abs/2507.01284
21. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X.E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y.,

Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B.D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K.H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N.P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., Ma, Z.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783

22. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models (2024)
23. Guo, Q., Mello, S.D., Yin, H., Byeon, W., Cheung, K.C., Yu, Y., Luo, P., Liu, S.: Regiongpt: Towards region understanding vision language model (2024), https://arxiv.org/abs/2403.02330
24. He, C., Zhu, S., Liu, H., Gao, F., Jia, Y., Zan, H., Peng, M.: DialogueMMT: Dialogue scenes understanding enhanced multi-modal multi-task tuning for emotion recognition in conversations. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics. pp. 2497–2512. Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025), https://aclanthology.org/2025.coling-main.170/
25. He, J., Lin, H., Wang, Q., Fung, Y.R., Ji, H.: Self-correction is more than refinement: A learning framework for visual and language reasoning tasks (2025)
26. Hu, H., Zhou, Y., You, L., Xu, H., Wang, Q., Lian, Z., Yu, F.R., Ma, F., Cui, L.: Emobench-m: Benchmarking emotional intelligence for multimodal large language models (2026), https://arxiv.org/abs/2502.04424
27. tse Huang, J., Lam, M.H., Li, E.J., Ren, S., Wang, W., Jiao, W., Tu, Z., Lyu, M.R.: Emotionally numb or empathetic? evaluating how llms feel using emotionbench (2024), https://arxiv.org/abs/2308.03656
28. Huang, J., Chen, X., Mishra, S., Zheng, H.S., Yu, A.W., Song, X., Zhou, D.: Large language models cannot self-correct reasoning yet (2023)
29. Jian, P., Wu, J., Sun, W., Wang, C., Ren, S., Zhang, J.: Look again, think slowly: Enhancing visual reflection in vision-language models (2025)
30. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European Conference on Computer Vision (ECCV). pp. 235–251 (2016)
31. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners (2023), https://arxiv.org/abs/2205.11916
32. Laurençon, H., Tronchon, L., Cord, M., Sanh, V.: What matters when building vision-language models? (2024), https://arxiv.org/abs/2405.02246
33. Le, T.Q.K., Vu, N.L.V., Pham, H.H., Huynh, X.L., Nguyen, T.H., Le, M.H.N., Nguyen, Q., Nguyen, H.D.: Hdc: Hierarchical distillation for multi-level noisy consistency in semi-supervised fetal ultrasound segmentation (2025), https://arxiv.org/abs/2504.09876
34. Lee, S., Park, S.H., Jo, Y., Seo, M.: Volcano: mitigating multimodal hallucination through self-feedback guided revision (2024)
35. Li, B., Zhang, P., Yang, J., Zhang, Y., Pu, F., Liu, Z.: Otterhd: A high-resolution multi-modality model (2023), https://arxiv.org/abs/2311.04219
36. Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., Xie, X.: Large language models understand and can be enhanced by emotional stimuli (2023)
37. Li, C., Wang, J., Zhang, Y., Zhu, K., Wang, X., Hou, W., Lian, J., Luo, F., Yang, Q., Xie, X.: The good, the bad, and why: Unveiling emotions in generative ai (2023)
38. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models (2023)
39. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models (2024), https://arxiv.org/abs/2311.06607

40. Liang, Z., Guo, K., Liu, G., Guo, T., Zhou, Y., Yang, T., Jiao, J., Pi, R., Zhang, J., Zhang, X.: Scemqa: A scientific college entrance level multimodal question answering benchmark (2024), https://arxiv.org/abs/2402.05138
41. Liao, Y.H., Mahmood, R., Fidler, S., Acuna, D.: Can large vision-language models correct semantic grounding errors by themselves? (2025)
42. Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models (2024), https://arxiv.org/abs/2312.07533
43. Liu, C., Xie, Z., Zhao, S., Zhou, J., Xu, T., Li, M., Chen, E.: Speak from heart: An emotion-guided llm-based multimodal method for emotional dialogue generation. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. p. 533–542. ICMR '24, Association for Computing Machinery, New York, NY, USA (2024), https://doi.org/10.1145/3652583.3658104
44. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2024)
45. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llavanext: Improved reasoning, ocr, and world knowledge (2024)
46. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
47. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
48. Man, F., Chen, X., Wang, H., Zhao, B., Li, H., Chen, X.: Vaeer: Visual attention-inspired emotion elicitation reasoning (2025), https://arxiv.org/abs/2505.24342
49. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J.: Large language models: A survey (2025), https://arxiv.org/abs/2402.06196
50. Mozikov, M., Severin, N., Bodishtianu, V., Glushanina, M., Nasonov, I., Orekhov, D., Pekhotin, V., Makovetskiy, I., Baklashkin, M., Lavrentyev, V., et al.: Eai: Emotional decision-making of llms in strategic games and ethical dilemmas. Advances in Neural Information Processing Systems **37**, 53969–54002 (2024)
51. Nguyen, T.H., Tran, H.L., Ngo, T.D.: Itself: Attention guided fine-grained alignment for vision-language retrieval (2026), https://arxiv.org/abs/2601.01024
52. Nguyen, T.H., Tran, H.L., Phan-Nguyen, H.P., Dinh, Q.V.: Hybrid, unified and iterative: A novel framework for text-based person anomaly retrieval (2025), https://arxiv.org/abs/2511.22470
53. Nguyen, T.H., Tran, Q.K., Quang-Hoang, A.T.: Improving generalization in visual reasoning via self-ensemble (2024), https://arxiv.org/abs/2410.20883
54. Nguyen-Nhu, T.A., Minh, T.D.H., To-Thanh, D., Le-Gia, P., Vo-Lan, T., Nguyen, T.H.: Ster-vlm: Spatio-temporal with enhanced reference vision-language models (2025), https://arxiv.org/abs/2508.13470
55. OpenAI, :, Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A.T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky,

B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C.J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D.P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F.P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H.P., Ren, H., Chang, H., Chung, H.W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J.G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J.Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J.W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M.J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M.O.T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R.G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Wal-

ters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., Malkov, Y.: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276
56. Peng, S., Fu, D., Gao, L., Zhong, X., Fu, H., Tang, Z.: Multimath: Bridging visual and mathematical reasoning for large language models (2024), https://arxiv.org/abs/2409.00147
57. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology **17**(3), 715–734 (2005)
58. Qiu, X., Jia, H., Zeng, Z., Shen, S., Meng, C., Yang, Y., Zhu, L.: Unified generation and self-verification for vision-language models via advantage decoupled preference optimization. arXiv preprint arXiv:2601.01483 (2026)
59. Qu, M., Hu, Y., Han, K., Wei, Y., Zhao, Y.: Recot: Reflective self-correction training for mitigating confirmation bias in large vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9147–9157 (2025)
60. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
61. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6), 1161 (1980)
62. Sabour, S., Liu, S., Zhang, Z., Liu, J.M., Zhou, J., Sunaryo, A.S., Li, J., Lee, T.M.C., Mihalcea, R., Huang, M.: Emobench: Evaluating the emotional intelligence of large language models (2024), https://arxiv.org/abs/2402.12071
63. Schwarz, N.: Feelings-as-information theory. In: Handbook of Theories of Social Psychology, pp. 289–308. Sage (2012)
64. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning (2024), https://arxiv.org/abs/2403.16999
65. Shao, R., Li, W., Zhang, L., Zhang, R., Liu, Z., Chen, R., Nie, L.: Large vlm-based vision-language-action models for robotic manipulation: A survey (2025), https://arxiv.org/abs/2508.13073
66. Sun, L., Liang, H., Wei, J., Yu, B., Li, T., Yang, F., Zhou, Z., Zhang, W.: Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14100–14115 (2025)
67. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen,

C., Lan, C.L., Choquette-Choo, C.A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L.: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786

68. Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S.C., Yang, J., Yang, S., Iyer, A., Pan, X., Wang, Z., Fergus, R., LeCun, Y., Xie, S.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms (2024), https://arxiv.org/abs/2406.16860
69. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9568–9578 (2024)
70. Tsui, K.: Self-correction bench: Uncovering and addressing the self-correction blind spot in large language models. arXiv preprint arXiv:2507.02778 (2025)
71. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. arXiv preprint arXiv:2504.08837 (2025)
72. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
73. Wu, T.H., Lee, H., Ge, J., Gonzalez, J.E., Darrell, T., Chan, D.M.: Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. arXiv preprint arXiv:2504.13169 (2025)
74. X.AI: Grok-1.5 vision preview (2024), https://x.ai/blog/grok-1.5v
75. Xu, Y., Hu, Y., Zhang, Z., Meyer, G.P., Mustikovela, S.K., Srinivasa, S., Wolff, E.M., Huang, X.: Vlm-ad: End-to-end autonomous driving through vision-language model supervision (2025), https://arxiv.org/abs/2412.14446
76. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W.,

Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., Fan, Z.: Qwen2 technical report (2024), https://arxiv.org/abs/2407.10671

77. Yao, Y., Mei, X., Xu, J., Sun, Z., Zeng, C., Chen, Y.: Vlm-emo: Context-aware emotion classification with clip. In: 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). pp. 1615–1620 (2024). https://doi.org/10.1109/AINIT61980.2024.10581513
78. Ye, M., Rong, X., Huang, W., Du, B., Yu, N., Tao, D.: A survey of safety on large vision-language models: Attacks, defenses and evaluations (2025), https://arxiv.org/abs/2502.14881
79. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences **67**(12), 220105 (2024)
80. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
81. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi (2024), https://arxiv.org/abs/2311.16502
82. Zhang, H., Wu, Y., Li, P., Zhang, X., Gao, Z., Gao, R., Gao, M., Sun, C., Jia, Y.: Mirror: Multimodal iterative reasoning via reflection on visual regions. arXiv preprint arXiv:2602.18746 (2026)
83. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.Y., Wen, J.R.: A survey of large language models (2026), https://arxiv.org/abs/2303.18223
84. Zhou, C., Chen, G., Bai, X., Dong, M.: On the human-level performance of visual question answering. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics. pp. 4109–4113. Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025), https://aclanthology.org/2025.coling-main.277/
85. Zhou, Y., Fan, Z., Cheng, D., Yang, S., Chen, Z., Cui, C., Wang, X., Li, Y., Zhang, L., Yao, H.: Calibrated self-rewarding vision language models. Advances in Neural Information Processing Systems **37**, 51503–51531 (2024)
86. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)