Self-Correction is More than Refinement: A Learning Framework for Visual and Language Reasoning Tasks

Jiayi He[♠*] **Hehai Lin**[◇*] **Qingyun Wang**[♡] **Yi R. (May) Fung**[◇] **Heng Ji**[♡]
[♡]University of Illinois at Urbana-Champaign [♠]Georgia Institute of Technology
[◇]The Hong Kong University of Science and Technology
`jhe478@gatech.edu` `hengji@illinois.edu`

Abstract

While Vision-Language Models (VLMs) have shown remarkable abilities, they invariably generate flawed responses. Self-correction that instructs models to refine their outputs presents a promising solution to this issue. Previous studies have mainly concentrated on Large Language Models (LLMs), while the self-correction abilities of VLMs, particularly concerning both visual and linguistic information, remain largely unexamined. This study investigates the self-correction capabilities of VLMs during both inference and fine-tuning stages. We introduce a Self-Correction Learning (SCL) approach that enables VLMs to learn from their self-generated self-correction data through Direct Preference Optimization (DPO) without relying on external feedback, facilitating self-improvement. Experimental results demonstrate that although VLMs struggle to self-correct effectively during iterative inference without additional fine-tuning and external feedback, they can enhance their performance and avoid previous mistakes through preference fine-tuning when their generated self-correction data are categorized into preferred and disfavored samples. This study emphasizes that self-correction is not merely a refinement process; rather, it should enhance models' reasoning ability through additional training, enabling them to generate high-quality responses directly without further refinement.[1]

1 Introduction

Large Language Models (LLMs) have shown exceptional versatility across numerous natural language processing domains (Sun et al., 2025; Chen et al., 2024b). Building upon the foundational capabilities of LLMs, Vision-Language Models (VLMs) (Liu et al., 2024b; Zhu et al., 2024) integrate visual recognition and language understanding by combining pre-trained LLMs and vision models through instruction fine-tuning, leading to significant advancements in multimodal tasks (Huang et al., 2024b; Peng et al., 2024).

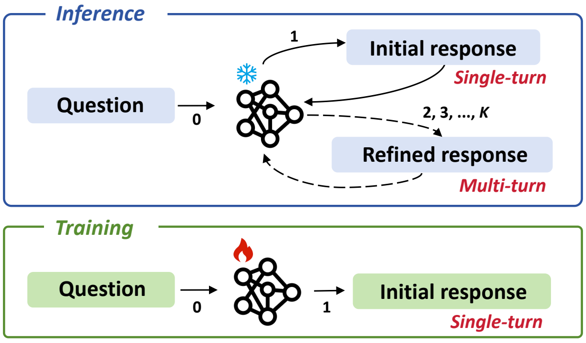


Figure 1: Comparison of inference-based and training-based self-correction. Inference-based methods aim to refine an initial response over K iterations, keeping model parameters fixed. Training-based methods focus on training the model to directly generate high-quality initial responses without iterative refinement.

Despite their robust vision-language understanding capabilities, VLMs are still prone to generating inaccurate information (Wu et al., 2024b; Fan et al., 2025; He et al., 2025; Zhang et al., 2025). Self-correction, an approach enabling models to identify and rectify mistakes in their outputs (Kamoi et al., 2024), is thus a promising method for enhancing the quality of VLM-generated responses. While previous studies have predominantly focused on self-correction in LLMs, the self-correction capabilities of VLMs remain underexplored. Given that VLMs integrate visual and linguistic information during reasoning, self-correction in VLMs presents additional challenges. This complexity arises from the inherent difficulty of accurately aligning and rectifying multimodal data, underscoring the critical need for a systematic investigation into VLM self-correction to advance their performance in vision-language reasoning tasks.

Existing self-correction strategies focus on the

---

[*] These two authors contribute equally to this work.
[1] Code is available at https://github.com/ivy3h/SCL.

Findings of the Association for Computational Linguistics: ACL 2025, pages 6405–6421
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

inference stage without parameter updates (Madaan et al., 2023; Shinn et al., 2023; Li et al., 2024; Liu et al., 2024a). These methods typically involve instructing models to revise their initially generated answers by leveraging self-correction prompts. While prior work has demonstrated the effectiveness of inference-time self-correction in improving LLM performance on reasoning tasks (Madaan et al., 2023), and such approaches offer advantages like no dedicated training cost and operational simplicity, recent studies have reported contradictory findings (Huang et al., 2024a; Xu et al., 2024). This controversy highlights two main shortcomings of inference-time self-correction: (1) *Unreliable performance*: The effectiveness of self-correction is highly sensitive to the formulation of self-correction prompts (Li et al., 2024). (2) *Limitations of models' reasoning abilities*: Without additional training to enhance their intrinsic reasoning capabilities, models often struggle to effectively self-correct when confronted with identical challenging tasks (Kamoi et al., 2024).

Besides these two challenges, a crucial distinction between existing self-correction methods during the inference stage and the more natural human self-correction process lies in their *correction goals*. As illustrated in Figure 1, the former approach focuses on better *refinement*, specifically enabling the model to correct its initial response through additional revisions (Madaan et al., 2023). Conversely, the latter approach emphasizes better *initial generation*, aiming to provide the correct answer on the first attempt without subsequent revisions (Zhang et al., 2024a,b; Tong et al., 2024). This discrepancy indicates that current inference-time self-correction methods offer only a temporary solution for rectifying mistakes. While a model can correct mistakes in its generated content through iterative self-correction, its underlying reasoning ability remains unchanged. Consequently, the model may continue to produce suboptimal answers when faced with identical questions in the future, leading to inefficient resource expenditure on repeated refinement. Therefore, we emphasize the **ultimate aim** of self-correction: ***not merely to fix initial mistakes but to fundamentally improve the model's capability to generate correct answers directly.***

In this paper, we investigate the self-correction capabilities of VLMs through two research questions (RQs): (1) **Inference-based self-correction mechanisms:** Can VLMs perform self-correction during inference without external feedback? (2) **Training-based self-correction mechanisms:** Can VLMs improve their performance by learning from their self-correction process and prevent similar future errors? Both RQs emphasize the concept of *self*, exploring the intrinsic abilities of VLMs to self-correct autonomously. For inference-based mechanisms, we design three distinct visual self-correction prompts. These prompts guide VLMs to critically examine their initial responses by scrutinizing input image details, understanding contextual cues, and comprehensively interpreting scenes. For training-based mechanisms, we propose **Self-Correction Learning (SCL)**, which utilizes Direct Preference Optimization (DPO) (Rafailov et al., 2023) to empower VLMs to self-improve by learning from their own generated self-correction preference data. The preference dataset, SELFCORSET, is constructed based on the intrinsic self-correction process during inference, where we designate the correct responses as preferred and the incorrect responses as disfavored.

We evaluate the intrinsic self-correction abilities of VLMs and the efficacy of SCL across multiple-choice questions (MCQ) benchmarks. Experimental results demonstrate that while VLMs initially struggle with intrinsic self-correction, they can benefit from learning from their self-correction samples. Specifically, VLMs fine-tuned with SCL are better able to avoid previous errors and exhibit superior performance compared to existing preference optimization methods for VLMs.

Our main contributions are as follows: **Firstly**, we define the ultimate objective of self-correction as not only rectifying initial errors but fundamentally enhancing a model's ability to generate accurate responses directly. **Secondly**, we systematically evaluate the self-correction abilities of VLMs during inference by developing three visual self-correction prompts. We further analyze the reliability of inference-based self-correction mechanisms. **Thirdly**, we introduce Self-Correction Learning (SCL), a novel approach that enables VLMs to self-improve via DPO by leveraging self-generated preference data. Our findings demonstrate the effectiveness of SCL and underscore the advantages of training-based self-correction mechanisms.

2 Related Work

Vision-Language Models and Preference Fine-Tuning. Vision-Language Models (VLMs), such as GPT-4o (OpenAI, 2024a), MiniGPT-4 (Zhu et al., 2024), and LLaVA-1.5 (Liu et al., 2024b), integrate the encoding of visual and textual data to solve various multimodal tasks such as image classification (Peng et al., 2024) and action recognition (Deng et al., 2024). Human preference alignment techniques have been widely applied to VLMs to train these models to generate content aligning with human intentions (Chen et al., 2024e). For instance, Silkie constructs a large-scale multimodal preference dataset annotated by GPT-4V, and distills these preferences into VLMs via Direct Preference Optimization (DPO) (Li et al., 2023b). Preference Optimization in VLLM with AI-Generated Dispreferences (POVID) utilizes preference fine-tuning to mitigate hallucinations (Zhou et al., 2024a). Calibrated Self-Rewarding (CSR) incorporates an iterative learning and rewarding paradigm into preference fine-tuning for modality alignment (Zhou et al., 2024b). Similarly, Inner Monologue Multi-Modal Optimization (IMMO) utilizes a combination of supervised learning and reinforcement learning approaches, performing an inner monologue to enhance model performance on complex vision-language tasks (Yang et al., 2024). While prior studies primarily achieve VLM alignment with human preferences through external feedback from humans or other language models, our work focuses on the self-improvement of VLMs through preference fine-tuning.

Intrinsic Self-Correction in Large Language Models. Self-correction in Large Language Models (LLMs) aims to guide these models in rectifying their flawed generated content, such as harmful outputs (Phute et al., 2024). Intrinsic self-correction refers to a self-correction paradigm during inference where a model revises its output solely by leveraging its inherent capabilities and the input context, without external feedback (Huang et al., 2024a; Kamoi et al., 2024; Liu et al., 2024a). This iterative, multi-turn self-correction process is distinct from single-turn test-time inference of GPT-o1 (OpenAI, 2024b). While recent research has demonstrated the effectiveness of intrinsic self-correction (Madaan et al., 2023; Shinn et al., 2023; Dhuliawala et al., 2024; Li et al., 2025), some studies indicate that LLMs face challenges with it. For instance, intrinsic self-correction may sometimes decrease output quality (Huang et al., 2024a) and potentially introduce bias (Xu et al., 2024). These conflicting results suggest that the self-correction ability of LLMs remains unreliable without external feedback. Previous work has primarily explored the intrinsic self-correction abilities of LLMs on unimodal tasks like arithmetic reasoning. Our study investigates the intrinsic self-correction abilities of VLMs on visual and language reasoning tasks.

Improvement in Language Models and Vision-Language Models. The enhancement of Language Models (LMs) can be broadly classified into self-improvement and external improvement. Self-improvement in LMs relies on their ability to learn from the data they generate, whereas external improvement involves leveraging external models or tools to enable LMs to learn from provided data. Regarding self-improvement, Huang et al. (2023); Wang et al. (2025b) show that LLMs can self-improve by learning from self-generated data selected via self-consistency (Wang et al., 2023). Zhang et al. (2024b); Wu et al. (2024a); Huang et al. (2025b) leverage the LLM's self-evaluation capability to generate training signals, such as for mitigating factual inaccuracies and enhancing response personalization. Kumar et al. (2024) introduce SCoRe, a multi-turn online reinforcement learning method that achieves self-correction improvements by training on self-generated data and utilizing a two-stage training process to prevent behavior collapse. Additionally, Wang et al. (2025a) propose Self-Improvement Modality Alignment (SIMA) which uses in-context self-critic to improve the modality alignment of VLMs. In SIMA, the model generates two one-turn responses for each question using greedy decoding and temperature sampling. It is then prompted to critique these responses as preferred or disfavored, thereby constructing a preference dataset. Distinguished from SIMA, our work constructs a preference dataset using pairs of initial and revised responses from VLMs, generated during intrinsic self-correction.

For external improvement, Tong et al. (2024) demonstrate that error data generated by strong LLMs can enhance the reasoning capabilities of weaker LLMs. Han et al. (2024) show that small LMs can enhance their self-correction capabilities through instruction fine-tuning. However, this method still requires the generation of self-modification responses. Our work emphasizes that the goal of self-correction is not only to correct mistakes iteratively, but rather to enhance the models' ability to produce correct answers directly. We explore whether VLMs can improve the quality of their responses without further refinement steps by leveraging both their successful and erroneous intrinsic self-correction data for self-improvement.

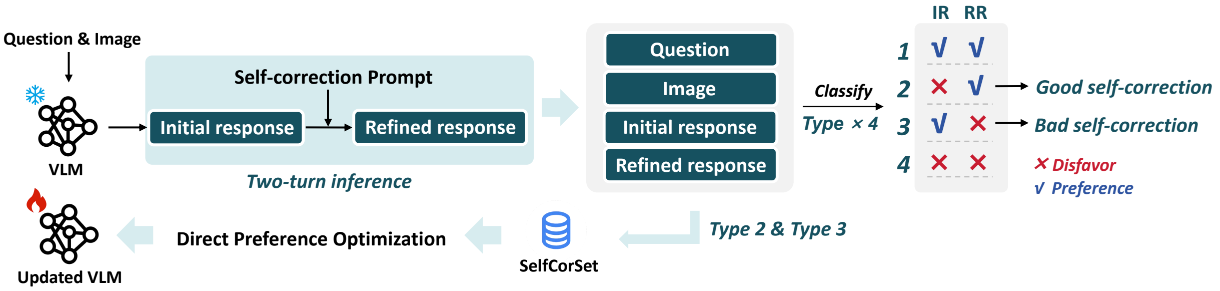


Figure 2: SCL begins with intrinsic self-correction applied to the VLM, generating four types of self-correction samples. Correct responses from Type 2 and incorrect responses from Type 3 samples are designated as preferences and disfavors, respectively, to construct the SELFCORSET preference dataset. The VLM then undergoes DPO on SELFCORSET for self-improvement.

3 Methodology

The framework of Self-Correction Learning (SCL) is illustrated in Figure 2, comprising three stages: inference, dataset construction, and preference fine-tuning. In the inference stage, we propose three visual self-correction prompts and investigate the intrinsic self-correction capabilities of VLMs to address **RQ1**. Subsequently, in the dataset construction and fine-tuning stage, we create SELFCORSET for each VLM based on its intrinsic self-correction and explore **RQ2** through DPO.

3.1 Inference: Intrinsic Self-Correction

Intrinsic self-correction comprises two distinct stages: initial answer generation and refined answer generation. During the initial answer generation stage, a **Standard Prompt** (**SP**) presents the complete question to the VLM, ensuring all requirements are included. For refined answer generation, the VLM engages in an iterative multi-turn process to enhance its initial responses. Due to computational resource constraints, we limit VLMs to a single refinement turn. We apply a critical prompt (Huang et al., 2024a) and develop three visual self-correction prompts to evaluate VLMs' intrinsic self-correction. The critical prompt directly guides models to detect issues in their initial responses. The visual self-correction prompts instruct models to identify problems by scrutinizing input image details, comprehending the portrayed context, and comprehensively interpreting entire scenes. Here are the prompts:

(1) **Critical Prompt** (**CP**): Review your previous answer and find problems with your answer. Based on the problems you found, improve your answer.

(2) **Comprehensive Detail Prompt** (**VP-1**): Review your previous answer and ensure that all relevant aspects of the image have been considered. Are there any elements or details that you missed? Based on your review, improve your answer.

(3) **Contextual Understanding Prompt** (**VP-2**): Review your contextual understanding of the image. Have you correctly interpreted the overall context and purpose of the scene? Based on your review, improve your answer.

(4) **Comprehensive Scene Analysis Prompt** (**VP-3**): Review your answer and ensure that your understanding of the image is comprehensive and detailed. Are there any aspects of the scene that you have omitted or misinterpreted? Based on your review, improve your answer.

3.2 Data Construction: SELFCORSET

We construct the preference dataset SELFCORSET using intrinsic self-correction outputs generated from multiple-choice question (MCQ) samples. These MCQ samples are obtained from several multimodal MCQ datasets, including MMStar (Chen et al., 2024c), MMBench (Liu et al., 2024c), SEEDBench (Li et al., 2023a), ScienceQA (Lu et al., 2022), MMT-Bench (Ying et al., 2024), and MMEvalPro (Huang et al., 2025a). For MMEvalPro, all available samples are utilized. For the remaining datasets, 500 samples are randomly selected for evaluation, with the rest allocated for SELFCORSET construction. Intrinsic self-correction is conducted using the VP-1 prompt, identified as the best-performing self-correction prompt overall (Table 1).

The construction process for the SELFCORSET dataset, using a VLM denoted as M, is outlined as follows: Given an MCQ sample $s = \{Question, Image, Answer\}$, we first perform intrinsic self-correction with VLM M to obtain a self-correction sample $s' =$

Type 2

Question How many chairs are present in the image? A: Four B: Three C: Two D: Five

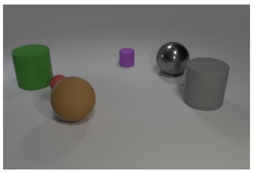
Initial response In the image, there are two blue chairs. So, the answer is ***C: Two***.

Self-correction prompt Review your previous answer and ensure that all relevant aspects of the image have been considered. Are there any elements or details that you missed? Based on your review, improve your answer.

Refined response After reviewing the image, I can confirm that there are indeed three chairs present: two blue ones and one white one. Therefore, the correct answer is ***B: Three***.

Correct answer: B

Type 3



Question The other object that is the same color as the large shiny thing is what shape?
A: cube B: sphere C: cylinder

Initial response The other object that is the same color as the large shiny thing is a cylinder. The reasoning behind this is that among the given options, only a cylinder has a similar shape to the large shiny object, which is also a cylinder. Therefore, the final answer is ***C: cylinder***.

Self-correction prompt Review your previous answer and ensure that all relevant aspects of the image have been considered. Are there any elements or details that you missed? Based on your review, improve your answer.

Refined response After reviewing the image, I can see that there are four objects in total: two green cubes, one purple cylinder, and one large shiny object. The question asks for the shape of the object that is the same color as the large shiny thing. Among the given options, only a sphere has a similar shape to the large shiny object, which is also a sphere. Therefore, the final answer is ***B: sphere***.

Correct answer: C

Figure 3: Two examples of intrinsic self-correction processes generated by InternLM-XComposer-2-7B.

$\{Question, Image, IR, RR\}$, where IR denotes the initial response and RR is the refined response. In this process, M generates IR when prompted with the initial question and image. Subsequently, RR is generated as a revision of IR after M receives the self-correction prompt. Based on the correctness of IR and RR, each self-correction sample s' is classified into four distinct types: Specifically, the corresponding correctness of IR and RR for the four types of samples are as follows: **Type 1** (correct⇒correct), **Type 2** (incorrect⇒correct), **Type 3** (correct⇒incorrect), and **Type 4** (incorrect⇒incorrect). The ⇒ symbol denotes the transition of correctness from the initial response (IR) to the refined response (RR). Type 2 samples represent successful self-corrections, as the model effectively revises an incorrect IR into a correct RR. Conversely, Type 3 samples indicate detrimental self-corrections, where the model erroneously alters a correct IR to an incorrect one. The SELFCORSET preference dataset is then constructed using Type 2 and Type 3 samples according to the following criterion: RR from Type 2 samples and IR from Type 3 samples are designated as preferred responses, as they represent correct answers. In contrast, IR from Type 2 samples and RR from Type 3 samples are classified as disfavored responses. Figure 3 presents two examples of Type 2 and Type 3 self-correction outputs generated by InternLM-XComposer-2-7B. In the Type 2 example, the model successfully revises an incorrect IR (C: Two) to a correct RR (B: Three) after reviewing the image. Conversely, the Type 3 example shows a detrimental self-correction, where the model incorrectly alters an initially correct IR (C: cylinder) to an incorrect RR (B: sphere).

We construct three distinct SELFCORSET, one for each of the three evaluated VLMs: LLaVA-V1.5-7B (Liu et al., 2024b), LLaVA-V1.5-13B (Liu et al., 2024b), and MiniCPM-Llama3-V2.5 (Yao et al., 2024). This construction emphasizes the uniqueness of *self*, as different VLMs possess distinct intrinsic self-correction behaviors and thus generate unique sets of self-correction samples. Each sample comprises a question, an image, a preferred response, and a disfavored response. Data samples are provided in Appendix A.

3.3 Training: Learn from Self-Correction

We fine-tune the VLM using Direct Preference Optimization (DPO) (Rafailov et al., 2023) on its dedicated preference dataset, SELFCORSET. We denote SELFCORSET as $\mathcal{D}_{sc} = \{(Q^{(i)}, I^{(i)}, R_c^{(i)}, R_r^{(i)})\}_{i=1}^{N}$, where $Q^{(i)}$ represents the input question, $I^{(i)}$ is the corresponding image, $R_c^{(i)}$ is the preferred response, and $R_r^{(i)}$ is the disfavored response. The DPO loss is defined as:

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(Q,I,R_c,R_r)} \left[\log \sigma f(\pi_\theta; \pi_{\text{ref}})\right] \quad (1)$$

$$f(\pi_\theta; \pi_{\text{ref}}) = \beta \left(\log \frac{\pi_\theta(R_c|Q,I)}{\pi_{\text{ref}}(R_c|Q,I)} - \log \frac{\pi_\theta(R_r|Q,I)}{\pi_{\text{ref}}(R_r|Q,I)}\right) \quad (2)$$

where σ represents the logistic sigmoid function, π_θ denotes the current VLM policy, π_{ref} denotes

		RealWorldQA	MMStar	MMBench	SEEDBench	ScienceQA	MMT-Bench	Rank
MiniCPM-Llama3-V2.5	SP	**61.70**	**50.40**	**79.00**	66.80	**75.40**	**49.00**	**1.17**
	+ CP	38.56	40.20	68.60	62.80	69.40	37.80	4.17
	+ VP-1	48.50	46.20	76.40	64.80	73.00	37.00	3.00
	+ VP-2	47.32	48.40	61.00	64.40	69.00	38.00	3.33
	+ VP-3	43.00	48.00	54.00	**69.80**	54.40	40.60	3.33
InternLM-XComposer-2-7B	SP	**60.13**	47.40	**76.80**	**69.00**	**78.20**	**48.80**	**1.17**
	+ CP	53.86	37.00	61.60	58.00	52.60	38.40	4.83
	+ VP-1	54.50	**48.00**	70.80	67.00	62.60	38.60	2.50
	+ VP-2	55.03	45.00	64.40	60.60	49.00	41.40	3.00
	+ VP-3	54.51	39.00	61.80	59.60	58.80	39.00	3.50
LLaVA-V1.5-7B	SP	**50.46**	**32.20**	**68.40**	**65.60**	**65.80**	**36.00**	**1.00**
	+ CP	36.60	24.00	54.00	36.20	56.80	32.00	2.83
	+ VP-1	43.01	22.80	57.20	42.40	58.20	29.00	2.33
	+ VP-2	17.78	18.60	45.40	29.00	45.80	12.00	5.00
	+ VP-3	36.21	20.40	54.00	37.00	54.80	28.40	3.67

Table 1: Quantitative comparisons (%) of MiniCPM-Llama3-V-2.5, InternLM-XComposer-2-7B, and LLaVA-V1.5-7B with intrinsic self-correction. Best results are **bolded**, and second-best are underlined. Rank indicates the overall performance ranking across the evaluated benchmarks (lower is better).

the reference policy, and β is a hyperparameter that controls the strength of the penalty against deviating from the reference policy. Both π_θ and π_{ref} are initialized with the same weight.

4 Experiments

4.1 Experimental Settings

Test Models. For intrinsic self-correction evaluation, we conduct experiments on three open-source VLMs: MiniCPM-Llama3-V2.5 (Yao et al., 2024), InternLM-XComposer-2-7B (Dong et al., 2024), and LLaVA-V1.5-7B (Liu et al., 2024b). MiniCPM-Llama3-V2.5 is an advanced VLM with a total of 8B parameters. InternLM-XComposer-2-7B is designed for the comprehension and composition of free-form text-image pairs. LLaVA-V1.5-7B is a widely used VLM trained with visual instructions. For self-correction training evaluation, we conduct experiments on MiniCPM-Llama3-V2.5, InternLM-XComposer-2-7B, LLaVA-V1.5-7B, and LLaVA-V1.5-13B.

Evaluation Benchmarks. We conduct evaluations on eight multimodal multiple-choice question (MCQ) benchmarks: RealWorldQA (xAI, 2024), MMStar (Chen et al., 2024c), MMBench-en (Liu et al., 2024c), SEEDBench (Li et al., 2023a), ScienceQA (Lu et al., 2022), MMT-Bench (Ying et al., 2024), MMMU (Yue et al., 2024), and AI2D (Kembhavi et al., 2016). For intrinsic self-correction, we use the first six datasets with the following sample counts for evaluation: RealWorldQA (765), MMStar (500), MMBench (500), SEEDBench (500), ScienceQA (500), MMT-Bench (500). We further incorporate two benchmarks to evaluate fine-tuned models: MMMU (1050) and AI2D (3088). We adopt accuracy and average rank as the metrics.

Training Baselines. We compare SCL with three preference optimization methods: POVID (Zhou et al., 2024a), CSR (Zhou et al., 2024b), and SIMA (Wang et al., 2025a). POVID leverages GPT to enhance the quality of ground truth answers and employs DPO for training. CSR incorporates iterative learning and a reward-based paradigm into its preference fine-tuning process. SIMA deploys in-context self-critic to construct its preference dataset and utilizes DPO to enhance the comprehension capabilities of VLM. We also compare SCL with Supervised Fine-Tuning (SFT), which directly utilizes the preferred responses of SELFCORSET for fine-tuning.

Implementation Details. Intrinsic self-correction is conducted on a total of 26,981 samples. The resulting SELFCORSET dataset sizes for each VLM are as follows: MiniCPM-Llama3-V2.5 (1,853), InternLM-XComposer-2-7B (2,361), LLaVA-V1.5-7B (4,797), and LLaVA-V1.5-13B (738). For computational efficiency, we adopt LoRA(Hu et al., 2022), a widely recognized Parameter-Efficient Fine-Tuning technique, for training. We set the LoRA rank as 8. Training for 7/8B models is conducted on a single 4090 24GB GPU with 1.5 GPU hours per epoch. Training for 13B models is conducted on a single V100 32GB GPU with 1.5 GPU hours for three epochs.

	RealWorldQA	MMStar	MMBench	SEEDBench	ScienceQA	MMT-Bench	MMMU	AI2D	Rank
LLaVA-V1.5-7B	50.90	32.97	70.24	66.64	65.48	35.40	32.25	52.71	4.50
+POVID	51.50	33.68	**71.44**	65.52	64.60	35.04	33.96	53.59	3.25
+CSR	51.03	32.59	70.44	65.12	64.76	33.76	34.63	53.34	4.25
+SIMA	49.41	32.40	71.04	64.68	64.68	34.84	**35.14**	53.01	4.38
+SFT	52.21	35.59	70.72	67.12	66.48	36.36	32.95	52.73	3.00
+SCL(Ours)	**52.71**	**36.11**	71.00	**67.84**	**66.56**	**36.96**	33.62	**54.81**	**1.63**
LLaVA-V1.5-13B	54.85	35.25	74.40	**69.40**	71.36	40.08	34.88	55.95	2.75
+SFT	**56.59**	36.33	75.16	69.32	71.88	40.48	35.29	56.00	1.88
+SCL(Ours)	56.03	**37.60**	**75.40**	68.56	**72.16**	**41.16**	**35.87**	**58.92**	**1.38**
InternLM-XComposer-2-7B	60.45	47.70	76.60	68.90	78.30	49.20	56.48	80.27	3.00
+SFT	62.03	48.90	**77.80**	70.20	79.50	**50.20**	57.95	80.66	1.75
+SCL (Ours)	**62.23**	**49.60**	77.00	**70.60**	**79.90**	**50.20**	**58.20**	**81.11**	**1.13**
MiniCPM-Llama3-V2.5	61.70	50.40	79.00	66.80	75.40	49.00	45.24	77.56	3.00
+SFT	62.35	52.40	80.80	68.40	76.00	49.80	47.43	78.01	2.00
+SCL(Ours)	**63.53**	**53.00**	**81.40**	**69.20**	**76.40**	**50.40**	**47.52**	**78.72**	**1.00**

Table 2: Quantitative comparisons (%) of LLaVA-V1.5-7B, LLaVA-V1.5-13B, InternLM-XComposer-2-7B, and MiniCPM-Llama3-V2.5 with SCL and baseline methods. Best results are **bolded**, and second-best are underlined. Rank indicates the overall performance ranking across the evaluated benchmarks (lower is better).

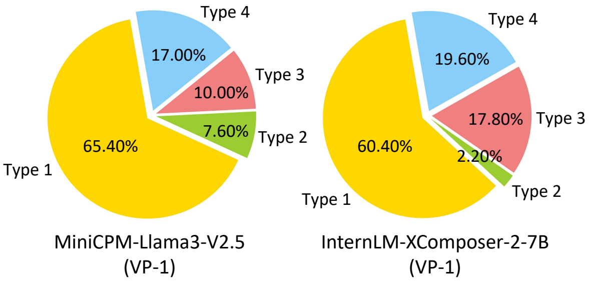


Figure 4: Distribution of self-correction examples of MiniCPM-Llama3-V2.5 and InternLM-XComposer-2-7B under VP-1 on ScienceQA.

4.2 Results and Analysis

VLMs struggle in intrinsic self-correction. Table 1 presents the results of VLMs regarding intrinsic self-correction. It can be observed that, in the majority of cases, VLMs achieve higher accuracy prior to intrinsic self-correction. For instance, MiniCPM-Llama3-V2.5's accuracy is 61.70% in its initial generation but drops to only 48.50% after intrinsic self-correction under the VP-1 prompt. These results consistently indicate that VLMs struggle to achieve effective performance improvement through intrinsic self-correction. Furthermore, the effectiveness of intrinsic self-correction varies significantly across different models, tasks, and self-correction prompts. While MiniCPM-Llama3-V2.5 improves its performance on SEEDBench with the VP-3 prompt (66.80% → 69.80%), it fails to self-correct on other benchmarks, highlighting the inherent instability of intrinsic self-correction.

To further illustrate the correctness transition during intrinsic self-correction, Figure 4 displays the distribution of self-correction sample types for MiniCPM-Llama3-V2.5 and InternLM-XComposer-2-7B on ScienceQA under the VP-1 prompt. The proportion of Type 3 samples (correct⇒incorrect) exceeds that of Type 2 (incorrect⇒correct) for both models, indicating that while models can revise incorrect answers, they more frequently erroneously alter initially correct answers. Consistent with the findings of Huang et al. (2024a) regarding LLMs, which reports degraded performance following intrinsic self-correction, our results suggest that VLMs struggle to rectify their own answers reliably.

VLMs self-improve from their self-correction examples. Table 2 shows the performance of SCL in comparison to various preference optimization baselines. SCL achieves state-of-the-art performance for LLaVA-V1.5-7B on six benchmarks. While SCL exhibits suboptimal performance on MMMU, these results indicate that the preference data derived from self-correction samples effectively fine-tunes these models. All four evaluated VLMs achieve the best overall performance as indicated by the Rank metric. Despite using a relatively small fine-tuning dataset, our findings demonstrate that VLMs can benefit from both successful and erroneous self-correction samples. This suggests that VLMs possess a human-like behavior, i.e., effectively learning by internalizing the distinction between successful corrections and the insights gained from analyzing erroneous attempts. This ability enables them to efficiently enhance their comprehensive reasoning capabilities.

4.3 Case Study

Figure 5 presents the initial responses generated by LLaVA-V1.5-7B before and after preference fine-tuning. Prior to fine-tuning, the model incorrectly identified the sunrise as the primary object, misinterpreting it as the element that establishes the mood and atmosphere of the scene. Following fine-tuning with SCL, the model correctly recognizes that the question seeks the object occupying the most spatial area in the image and subsequently produces the correct answer, demonstrating the model's ability to avoid previous errors through preference fine-tuning.

5 Further Studies and Analysis

5.1 Reliability of Successful Intrinsic Self-Correction

We observe that some successful refinements, classified as Type 2 self-correction cases, stem from the model's incidental guessing of the correct answer after receiving a self-correction prompt, rather than from proper reasoning of the task. For instance, when presented with an image that does not contain a teapot, models might respond with "The teapot may exist behind the woman..." and then conclude that a teapot is present in the image after self-correction. This type of refinement reflects a degree of uncertainty. Moreover, the reasoning behind seemingly successful refinements may not be entirely accurate. For example, in the Type 2 case presented in Figure 3, the model correctly identifies the number of chairs after self-correction but fails to accurately determine their colors. These observations on the reliability of successful refinements collectively indicate that VLMs possess limited capabilities for truly accurate refinement. Further supporting data is provided in Appendix B.

Future research could explore the internal information flow that mediates the transition from initial to refined responses during intrinsic self-correction, for instance, by visualizing attention weights to improve both interpretability and reliability. More importantly, cases where the final answer is correct but the reasoning is flawed can negatively impact the self-correction mechanism. Future work should focus on designing more detailed evaluation criteria that account for the correctness of intermediate reasoning steps, such as by leveraging a strong model as an MLLM-as-a-judge (Chen et al., 2024a).

	Turn 0	Turn 1	Turn 2	Turn 3
RealWorldQA (VP-1)	**61.70**	48.50	39.22	42.61
MMStar (VP-2)	**50.40**	48.40	49.20	46.80
MMBench (VP-1)	**79.00**	76.40	75.60	72.60
SEEDBench (VP-3)	66.80	**69.80**	63.80	64.00

Table 3: Results (%) of MiniCPM-Llama3-V2.5 with multi-turn intrinsic self-correction.

5.2 Multi-Turn Intrinsic Self-Correction

Table 3 presents the results of multi-turn intrinsic self-correction for MiniCPM-Llama3-V2.5 on four benchmarks. Turn 0 represents the initial generation, while Turns 1 to 3 illustrate subsequent intrinsic self-correction iterations. Notably, the accuracy of the refined answers consistently degrades over multiple turns, ultimately falling below the initial accuracy after three correction iterations. This decline suggests that VLMs find it challenging to achieve effective intrinsic self-correction solely by increasing the number of correction iterations.

Despite this observed degradation, multi-turn intrinsic self-correction may generate a richer, more diverse set of self-correction samples. Each iterative attempt may produce different types of errors when responding to the same question, thereby providing varied erroneous data for preference fine-tuning. Therefore, future research could investigate whether VLMs can derive greater benefits from these additional, potentially more informative samples, even if the multi-turn accuracy diminishes.

5.3 Failure of VLM Intrinsic Self-Correction

We attribute the observed failures of intrinsic self-correction in VLMs to several potential factors: (1) **Limited ability to judge correctness** (Huang et al., 2024a): VLMs may lack the fundamental reasoning capabilities required to reliably detect and revise their own errors. The inherent multimodal nature of these models—integrating both visual and textual inputs—introduces additional complexity, making it difficult for them to synthesize information across modalities and accurately evaluate correctness. (2) **Susceptibility to prompt bias** (Li et al., 2024): As demonstrated in Table1, self-correction performance varies notably with different prompts, suggesting a sensitivity to prompt design. Existing self-correction prompts may not sufficiently capture the intricate interplay between visual and textual content, which may impair the models' ability to effectively revise their responses.

Question What is the primary subject of the image? A: The sunrise B: The grassy field C: The trees D: The fog

Before *A. The sunrise.* ☒

1. The image features a grassy field with a sunrise in the background. 2. The sunrise is the primary subject of the image, as it sets the mood and atmosphere of the scene. 3. The grassy field and the fog contribute to the overall aesthetic of the image, but the sunrise is the main focus.

After *B. The grassy field.* ☑

1. The primary subject of the image is the grassy field. 2. The image features a foggy sunrise, which adds a serene and atmospheric quality to the scene. 3. However, the focus is on the grassy field, which is the main subject of the photo. 4. The other elements, such as the trees and the sunrise, contribute to the overall composition and mood of the image, but they do not overshadow the grassy field as the primary subject.

Figure 5: LLaVA-V1.5-7B successfully answers the question after learning from its self-correction samples.

(3) **Lack of a grounding mechanism**: In contrast to self-correction methods that utilize external feedback or oracle labels (Huang et al., 2024a; Shinn et al., 2023), intrinsic self-correction operates without explicit supervision. This absence of grounding guidance limits the model's ability to converge on more accurate responses, resulting in ineffective or even detrimental revisions. Given these challenges, future work could explore the design of more diverse and task-specific self-correction prompts tailored for multimodal settings. Additionally, incorporating a strong teacher model to provide instructional feedback, or integrating oracle labels during the self-correction process, may help improve the reliability of revisions.

5.4 Effect of the Number of Training Samples

The limited proportion of Type 2 and Type 3 data, as illustrated in Figure 4, coupled with the tendency of more advanced models to generate fewer samples of these types, results in a relatively small sample size for SELFCORSET. To explore the influence of fine-tuning data quantity, we randomly divide SELFCORSET into five subsets, increasing in 20% increments from 0%. We evaluate the impact of these varying training set sizes on the performance of LLaVA-V1.5-7B on SEEDBench and AI2D, with the accuracy trend illustrated in Figure 6. Notably, even with smaller training datasets, the fine-tuned model exhibits significant performance gains. For instance, the model fine-tuned on the $p = 0.4$ subset achieves an accuracy of 67.80% on SEEDBench, reflecting a 2.2% improvement over the untrained model. These results indicate that SCL effectively improves the model's performance despite the relatively modest size of SELFCORSET. Moreover, as the number of training samples increases, the overall accuracy consistently improves, showcasing the potential of SCL with larger preference datasets. Therefore, future research could explore more straightforward and computationally efficient data augmentation methods to scale SELFCORSET and investigate SCL's performance under large-scale data conditions.

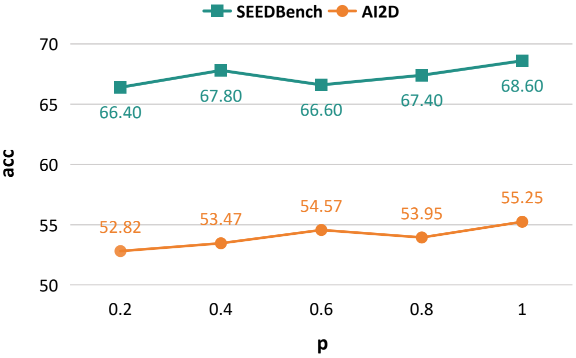


Figure 6: Results of LLaVA-V1.5-7B under different proportions (p) of SELFCORSET.

6 Conclusion

This work investigates the intrinsic self-correction capabilities and limitations of Vision-Language Models (VLMs) during the inference stage, and their potential for self-improvement through fine-tuning, across various multiple-choice question (MCQ) benchmarks. We propose Self-Correction Learning (SCL), a novel framework that employs Direct Preference Optimization (DPO) to enable VLMs to learn from their own self-correction responses. This approach aims to fundamentally enhance the models' ability to generate correct answers directly, rather than merely refining initial errors. Our experiments reveal that while VLMs face significant challenges and exhibit unreliable performance during intrinsic self-correction, they can effectively leverage their own self-correction samples (both successful and erroneous) to self-improve without requiring external feedback.

Limitations

Given the challenges of evaluating absolute correctness in open-ended multimodal tasks, this study primarily focuses on MCQ benchmarks to assess the accuracy of both initial and refined responses. However, this design inherently limits the exploration of more complex vision understanding tasks, such as visual question answering (Chen et al., 2025), video question answering (Chen et al., 2024d; Fu et al., 2024), and complex transportation system navigation (Li et al., 2023c). Future work could investigate fine-grained evaluation methods applicable across a wider range of multimodal tasks, as well as computational resource constraint-aware adaptive reasoning (Huang et al., 2025c).

The current construction of SelfCorSet is tailored to each VLM, leveraging the model's own preferred and disfavored responses to enable more targeted self-improvement. However, this design limits generalizability across different models. A potential direction for future work is to develop a unified version of SelfCorSet that can be applied across multiple VLMs. This would involve identifying common error patterns shared among VLMs and designing universal prompts that generalize well across different architectures.

References

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Jiao Chen, Jiayi He, Fangfang Chen, Zuohong Lv, Jianhua Tang, Weihua Li, Zuozhu Liu, Howard H. Yang, and Guangjie Han. 2024b. Towards general industrial intelligence: A survey of continual large models in industrial iot. *Preprint*, arXiv:2409.01207.

Jiao Chen, Ruyi Huang, Zuohong Lv, Jianhua Tang, and Weihua Li. 2025. Faultgpt: Industrial fault diagnosis question answering system by vision language models. *Preprint*, arXiv:2502.15481.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024c. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2024d. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pages 179–195. Springer.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024e. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14239–14250.

Zhichao Deng, Xiangtai Li, Xia Li, Yunhai Tong, Shen Zhao, and Mengyuan Liu. 2024. Vg4d: Vision-language model goes 4d video recognition. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5014–5020.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, and 4 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *Preprint*, arXiv:2401.16420.

Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R. Fung. 2025. Unveiling the lack of lvlm robustness to fundamental visual variations: Why and path forward. *Preprint*, arXiv:2504.16727.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Haixia Han, Jiaqing Liang, Jie Shi, Qianyu He, and Yanghua Xiao. 2024. Small language model can self-correct. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18162–18170.

Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R. Fung. 2025. Mmboundary: Advancing mllm knowledge boundary awareness through reasoning step confidence calibration. *Preprint*, arXiv:2505.23224.

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

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024a. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, Jingyang Yuan, Wei Ju, Luchen Liu, Tianyu Liu, Baobao Chang, and Ming Zhang. 2025a. MMEvalPro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4805–4822, Albuquerque, New Mexico. Association for Computational Linguistics.

Junsheng Huang, Zhitao He, Sandeep Polisetty, Qingyun Wang, and May Fung. 2025b. Mac-tuning: Llm multi-compositional problem reasoning with enhanced knowledge boundary awareness. *Preprint*, arXiv:2504.21773.

Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024b. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *Preprint*, arXiv:2403.12027.

Shijue Huang, Hongru Wang, Wanjun Zhong, Zhaochen Su, Jiazhan Feng, Bowen Cao, and Yi R. Fung. 2025c. Adactrl: Towards adaptive and controllable reasoning via difficulty-aware budgeting. *Preprint*, arXiv:2505.18822.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *Preprint*, arXiv:1603.07396.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. 2024. Training language models to self-correct via reinforcement learning. *Preprint*, arXiv:2409.12917.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Cheng Li, May Fung, Qingyun Wang, Chi Han, Manling Li, Jindong Wang, and Heng Ji. 2025. Mentalarena: Self-play training of language models for diagnosis and treatment of mental health disorders. *Preprint*, arXiv:2410.06845.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023b. Silkie: Preference distillation for large visual language models. *Preprint*, arXiv:2312.10665.

Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. 2024. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *Preprint*, arXiv:2402.12563.

Tianyi Li, Joshua Klavins, Te Xu, Niaz Mahmud Zafri, and Raphael Stern. 2023c. Understanding driver-pedestrian interactions to predict driver yielding: naturalistic open-source dataset collected in minnesota. *Preprint*, arXiv:2312.15113.

Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, Ruiyang Qin, Yiyu Shi, and Jinjun Xiong. 2024a. Large language models have intrinsic self-correction ability. *Preprint*, arXiv:2406.15673.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

OpenAI. 2024a. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2025-05-29.

OpenAI. 2024b. O1 system card. `https://cdn.openai.com/o1-system-card-20240917.pdf`. Accessed: 2025-05-29.

Wenshuo Peng, Kaipeng Zhang, Yue Yang, Hao Zhang, and Yu Qiao. 2024. Data adaptive traceback for vision-language foundation models in image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4506–4514.

Mansi Phute, Alec Helbling, Matthew Daniel Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. 2024. LLM self defense: By self examination, LLMs know they are being tricked. In *The Second Tiny Papers Track at ICLR 2024*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R. Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *Preprint*, arXiv:2402.11060.

Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024. Can LLMs learn from previous mistakes? investigating LLMs' errors to boost for reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3065–3080, Bangkok, Thailand. Association for Computational Linguistics.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, Furong Huang, and Cao Xiao. 2025a. Enhancing visual-language modality alignment in large vision language models via self-improvement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 268–282, Albuquerque, New Mexico. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh ICLR*.

Yumeng Wang, Zhiyuan Fan, Qingyun Wang, May Fung, and Heng Ji. 2025b. Calm: Unleashing the cross-lingual self-aligning ability of language model question answering. *Preprint*, arXiv:2501.18457.

Shujin Wu, May Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2024a. Aligning llms with individual preferences via interaction. *Preprint*, arXiv:2410.03642.

Shujin Wu, Yi Fung, Sha Li, Yixin Wan, Kai-Wei Chang, and Heng Ji. 2024b. MACAROON: Training vision-language models to be your engaged partners. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7715–7731, Miami, Florida, USA. Association for Computational Linguistics.

xAI. 2024. Realworldqa. `https://x.ai/blog/grok-1.5v`. Accessed: 2025-05-29.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

Diji Yang, Kezhen Chen, Jinmeng Rao, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. 2024. Tackling vision language tasks through learning inner monologues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19350–19358.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, jiayi lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, and 3 others. 2024. MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Forty-first International Conference on Machine Learning*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R. Fung. 2025. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues. *Preprint*, arXiv:2502.12084.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965, Bangkok, Thailand. Association for Computational Linguistics.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024a. Aligning modalities in vision large language models via preference fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024b. Calibrated self-rewarding vision language models. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

A Data Examples of SELFCORSET

Our work introduces SELFCORSET, a novel dataset derived from the intrinsic self-correction processes of Vision-Language Models (VLMs) during inference. For each sample, the dataset captures both the VLM's initial and refined responses. Specifically, SELFCORSET comprises Type 2 (incorrect⇒correct) and Type 3 (correct⇒incorrect) self-correction instances, where the correct responses are designated as preferred and the incorrect ones as disfavored. In this section, we provide illustrative examples of the dataset content (Table 4 to 9).

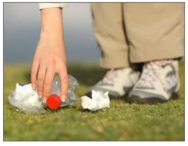
B Examples Supporting Further Studies

In our further studies, we argue that VLMs often exhibit limited capabilities for genuinely accurate refinement, and that even seemingly successful self-corrections can be unreliable due to flawed underlying reasoning. To substantiate this claim, we present two illustrative examples in Table 10.

In the first example, the model's refined response relies on the misleading assumption that a "Do not board" sign directly indicates a lack of crowding on the platform. However, platform crowding should primarily be assessed based on crowd density. Although the model ultimately selects the correct answer (B. no), this decision stems from an over-reliance on a single, potentially misleading cue rather than a comprehensive understanding of the entire visual scenario.

In the second example, the model confuses particle size with concentration during the refinement process. The model infers concentration levels from the size of the purple particles, entirely neglecting the actual number of particles present. While the model successfully corrects the error in its initial response, the self-correction process should have involved a proper reassessment of the quantity of purple particles in the solution, not their individual size.

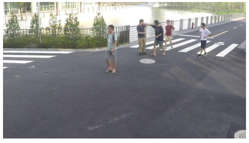
Q: What is the direction of this pull? A. toward her hand B. away from her hand

I: 

IR (disfavored): B. away from her hand. The person is pulling a bottle away from their hand, which indicates that they are moving the bottle in the opposite direction of their hand. This action suggests that they are trying to remove the bottle from the ground or from their hand, rather than placing it there.

RR (preferred): A. toward her hand. Upon reviewing the image, it is clear that the person is pulling a bottle toward their hand, rather than away from it. This action suggests that they are trying to pick up the bottle or place it on the ground. The person's hand is positioned above the bottle, indicating that they are in the process of lifting it.

Q: Detect and give a judgment of the anomalous object within the given image. A. normal B. anomaly

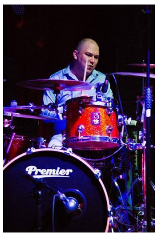
I: 

IR (disfavored): B. anomaly. The image shows a group of people walking on a sidewalk near a body of water. Among them, there is a person wearing a backpack, which is an unusual object in this context. The backpack is not a typical item one would expect to see while walking on a sidewalk, as it is usually associated with carrying personal belongings or items for outdoor activities. The presence of the backpack in this scene might indicate that the person is either on a trip or carrying something specific, but it is not a common sight in this setting.

RR (preferred): A. normal. After reviewing the image, I realized that the backpack is not an anomalous object in this context. It is a common item that people carry while walking on a sidewalk, especially if they are on a trip or carrying personal belongings. The presence of the backpack does not indicate that the person is carrying something specific or that they are on a trip. Therefore, the backpack is not an anomalous object in this scene.

Table 4: Type 2 samples of LLaVA-V1.5-7B.

Q: How would you describe the appearance of the man playing the drums? A. Young and long-haired B. Middle-aged with glasses C. Old and bald D. Young with a beard

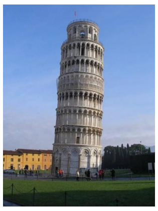
I: 

IR (preferred): C. Old and bald. The man playing the drums is described as an older man with a bald head. This suggests that he is not a young man, and his appearance is not typical of a young musician. The fact that he is playing drums indicates that he has some level of skill and experience in the field.

RR (disfavored): A. Young and long-haired. Upon reviewing the image, it is clear that the man playing the drums is not an older man with a bald head. Instead, he is described as a young man with long hair. This suggests that he is not a middle-aged man, and his appearance is more typical of a young musician. The fact that he is playing drums indicates that he has some level of skill and experience in the field

---

Q: What landmark is this and where is it? A. Mecca in Saudi Arabia B. Big Ben in London C. The Burj al Arab Hotel in Dubai D. Tower of Pisa, Italy

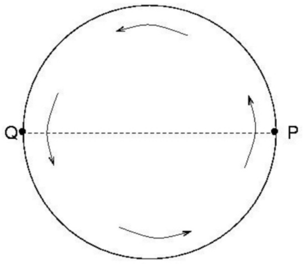
I: 

IR (preferred): D. Tower of Pisa, Italy. The image features a tall, white tower with a clock on its side, which is the iconic Tower of Pisa. The tower is located in Italy, and it is a famous landmark known for its leaning structure. The presence of a large group of people around the tower suggests that it is a popular tourist attraction. The image also shows a clock on the side of the tower, which is a distinctive feature of the structure.

RR (disfavored): A. Mecca in Saudi Arabia. I apologize for the confusion in my previous response. After reviewing the image, I realized that the Tower of Pisa is not the correct answer. The image features a large clock tower, which is a prominent landmark in Mecca, Saudi Arabia. The tower is known as the Abraj Al-Bait Clock Tower and is a significant attraction for pilgrims visiting the city. The image shows a large group of people around the tower, indicating its popularity as a tourist destination. The clock tower's distinctive design and its location in Mecca make it the most appropriate landmark in the image.

Table 5: Type 3 samples of LLaVA-V1.5-7B.

Q: How many arrows are there in total in the picture? A. 1 B. 2 C. 3 D. 4

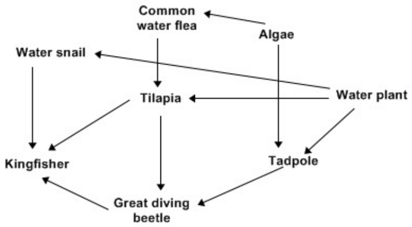


I:

IR (disfavored): There are a total of **3 arrows** in the picture. The first arrow is pointing upwards, the second arrow is pointing downwards, and the third arrow is pointing to the right.

RR (preferred): Upon reviewing my previous answer, I realized that I missed the fact that there are two arrows pointing in the same direction, which is upwards. Therefore, there are a total of **4 arrows** in the picture. The first arrow is pointing upwards, the second arrow is pointing upwards as well, the third arrow is pointing to the right, and the fourth arrow is pointing downwards.

Q: How many species are there in total? A. 5 B. 6 C. 7 D. 8 E. 9



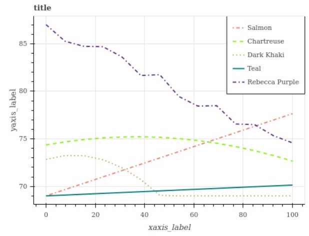
I:

IR (disfavored): There are a total of **7 species** in the diagram. The diagram shows the food chain of a water ecosystem, with each species being connected to the next one through arrows. The species include algae, water flea, tadpole, great diving beetle, and tapioca. The arrows indicate the flow of energy and nutrients from one species to another, as each species consumes the one before it.

RR (preferred): Upon reviewing my previous answer, I realized that I missed one species in the food chain. The correct answer is **8 species.** The diagram shows the food chain of a water ecosystem, with each species being connected to the next one through arrows. The species include algae, water flea, tadpole, great diving beetle, tapioca, and two other species that are not explicitly identified. The arrows indicate the flow of energy and nutrients from one species to another, as each species consumes the one before it.

Table 6: Type 2 samples of LLaVA-V1.5-13B.

Q: Where do Teal and Dark Khaki intersect on the x-axis? A. 40 B. 44 C. 60 D. 80

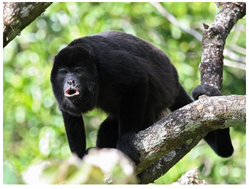


I:

IR (preferred): To find the intersection of Teal and Dark Khaki on the x-axis, we need to look at the x-axis values for both lines. Teal has a value of 40, while Dark Khaki has a value of 60. The intersection occurs at the point where these two values meet, which is at 44. Therefore, the answer is **B. 44.**

RR (disfavored): Upon reviewing my previous answer, I realized that I made an error in my interpretation of the x-axis values for Teal and Dark Khaki. Teal has a value of 40, while Dark Khaki has a value of 80. The intersection occurs at the point where these two values meet, which is at 60. Therefore, the correct answer is **C. 60.**

Q: What color is the Howler Monkey? A. Blue B. Green C. Brown D. Purple E. Black F. Dark brown G. White



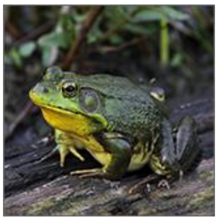
I:

IR (preferred): The Howler Monkey in the image is black. The monkey is sitting on a tree branch, and its color is clearly visible. The monkey's black fur stands out against the tree branch, making it easy to identify the color of the monkey. Therefore, the correct answer is **E. Black.**

RR (disfavored): Upon reviewing my previous answer, I noticed that I did not consider the possibility of the monkey's fur appearing dark brown in certain lighting conditions. Therefore, I would like to revise my answer to include the possibility of the monkey's fur appearing dark brown or brown. Therefore, the correct answer is **F. Dark brown.**

Table 7: Type 3 samples of LLaVA-V1.5-13B.

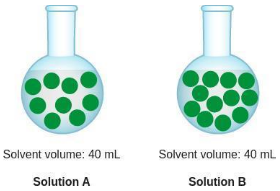
Q: Select the amphibian below. 0: brown tree frog 1: woodpecker

I: 

IR (disfavored): The image features a frog, which is an amphibian. The brown tree frog (option 0) is also an amphibian, but it is not depicted in the image; therefore, by elimination and based on the visual characteristics of the animal in the picture, the correct answer is **1 (woodpecker)**, as it does not match the description of the amphibian shown.

RR (preferred): Upon reviewing the context, it's clear that the image showcases a frog, which is an amphibian. The correct answer should be the amphibian present in the image. Therefore, the improved answer is **0 (brown tree frog)**, as it matches the description of the amphibian depicted in the image.

Q: Which solution has a higher concentration of green particles? A. neither; their concentrations are the same B. Solution B C. Solution A

I: 

IR (disfavored): The concentration of particles in a solution is determined by the volume of the solution and the number of particles present. In this case, both Solution A and Solution B have a solvent volume of 40 ml. Since there are more green particles in Solution A than in Solution B, we can conclude that Solution A has a higher concentration of green particles. Final answer: **C. Solution A**

RR (preferred): After reviewing the image, it is clear that there are more green particles in Solution B than in Solution A. Therefore, the concentration of green particles is higher in Solution B. Final answer: **B. Solution B**

Table 8: Type 2 samples of MiniCPM-Llama3-V-2.5.

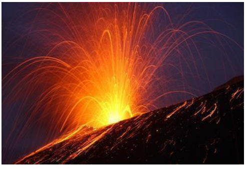
Q: Which term matches the picture? A. endotherms B. ectotherms

I:

IR (preferred): The picture you provided shows a horse, which is a mammal. Mammals are endotherms, meaning they can regulate their body temperature internally. The term that matches the picture is **A. endotherms**.

RR (disfavored): I have reviewed my previous answer and considered all relevant aspects of the image. Based on my review, I believe that the term ectotherms is more suitable to describe the horse in the picture because horses are cold-blooded animals that regulate their body temperature by external heat sources, rather than internally like endotherms (warm-blooded animals). Therefore, my final answer is **B. ectotherms**.

Q: Which term matches the picture? A. Hawaiian eruption B. Strombolian eruption

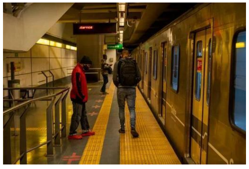
I: 

IR (preferred): A Strombolian eruption is characterized by moderate bursts of expanding gases that eject incandescent lava fragments, as shown in the image with fiery lava being thrown into the air. Hawaiian eruptions, on the other hand, are generally less explosive and produce more fluid lava flows. The term that matches the picture is **B. Strombolian eruption**.

RR (disfavored): I apologize for the mistake in my previous answer. Upon reviewing the image, it is clear that the term that matches the picture is **A. Hawaiian eruption**. This is because the image shows a volcanic eruption with lava flowing and a bright light at the top of the volcano, which are characteristics of a Hawaiian eruption. Strombolian eruptions typically involve explosive ejections of solid material and do not necessarily exhibit such visible lava flows or bright lights as seen in the image. Therefore, I have corrected my initial response to accurately reflect the information provided in the image.

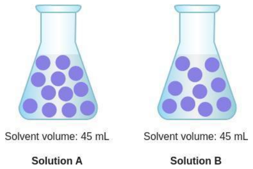
Table 9: Type 3 samples of MiniCPM-Llama3-V-2.5.

Q: is this place crowded? A. yes B. no

I: 

IR: The place is crowded. People are waiting for the train, and they are spread out along the platform. The final answer is **A**.
SC: Review your previous answer and ensure that all relevant aspects of the image have been considered. Are there any elements or details that you missed? Based on your review, improve your answer.
RR: I reviewed my previous answer and have identified a relevant detail that was missed. The sign Do not board on the platform indicates that the train may be not crowded, as it is likely the reason why people are standing near it. Based on this new information, I will revise my final answer to **B**.

Q: Which solution has a higher concentration of purple particles? A. Solution A B. Solution B C. neither; their concentrations are the same

I: 

IR: The concentration of particles in a solution is determined by the volume of the solution and the number of particles present. In this case, both Solution A and Solution B have the same volume (45 ml). Therefore, to determine which solution has a higher concentration of purple particles, we need to compare the number of particles in each solution. Unfortunately, the information provided does not include the number of particles in either solution. Without this information, it is impossible to determine which solution has a higher concentration of purple particles. So, the final answer is **C. neither; their concentrations are the same**.
SC: Review your previous answer and ensure that all relevant aspects of the image have been considered. Are there any elements or details that you missed? Based on your review, improve your answer.
RR: I apologize for the oversight in my previous response. Upon reviewing the image, I can now see that Solution A has more purple particles than Solution B. This is because Solution A has larger concentrated purple particles, while Solution B has smaller concentrated purple particles. Therefore, the final answer is **A. Solution A has a higher concentration of purple particles**.

Table 10: Successful refinements (Type 2 samples) with unreliable self-correction reasoning.