THROUGH THE LENS OF CONTRAST: SELF-IMPROVING VISUAL REASONING IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

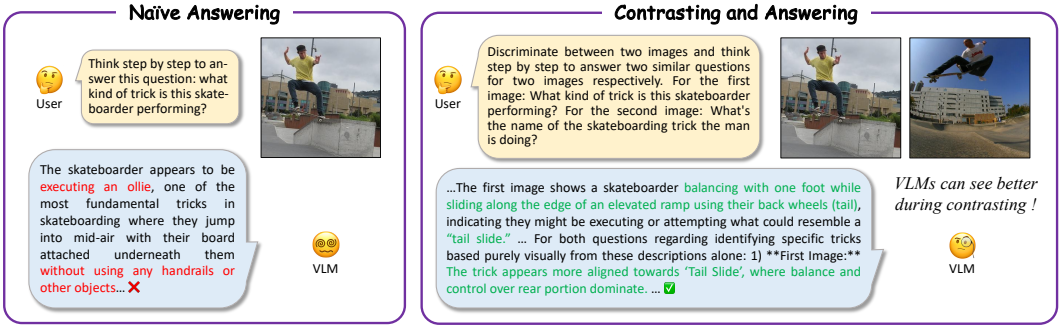
Reasoning has emerged as a key capability of large language models. In linguistic tasks, this capability can be enhanced by self-improving techniques that refine reasoning paths for subsequent finetuning. However, extending these language-based self-improving approaches to vision language models (VLMs) presents a unique challenge: visual hallucinations in reasoning paths cannot be effectively verified or rectified. Our solution starts with a key observation about visual contrast: when presented with a contrastive VQA pair, *i.e.*, two visually similar images with synonymous questions, VLMs identify relevant visual cues more precisely compared with when given a single VQA sample. Motivated by this observation, we propose **V**isual **C**ontrastive **S**elf-**Ta**ught **R**easoner (VC-STaR), a novel self-improving framework that leverages visual contrast to mitigate hallucinations in model-generated rationales. We collect a diverse suite of VQA datasets, curate contrastive pairs according to multi-modal similarity, and generate rationales using VC-STaR. Consequently, we obtain a new visual reasoning dataset, VisCoR-55K, which is then used to boost the reasoning capability of various VLMs through supervised finetuning. Extensive experiments show that VC-STaR not only outperforms existing self-improving approaches but also surpasses models finetuned on the SoTA visual reasoning datasets, demonstrating that the inherent contrastive ability of VLMs can bootstrap their own visual reasoning. The code, dataset and trained models will be released upon acceptance.

1 INTRODUCTION

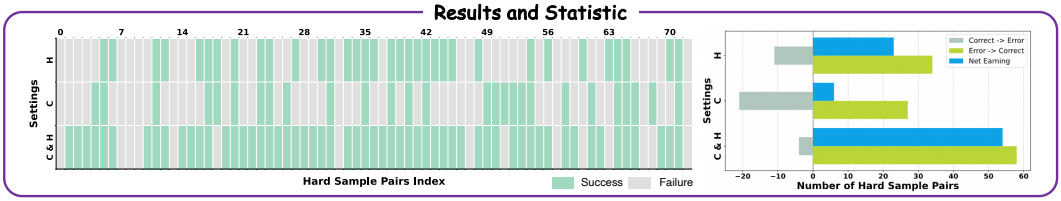
The scaling of large language models (LLM) has led to the emergence of reasoning capabilities (Wei et al., 2022a), making a transition from *System* 1 to *System* 2 (Kahneman, 2011) and enabling language models to tackle complex, multi-step problems (Wei et al., 2022b; Kojima et al., 2022). This emergent ability can be further enhanced by various techniques (Wang et al., 2023b; Li et al., 2023b; Hao et al., 2023; Gao et al., 2023; OpenAI, 2024b; Guo et al., 2025). Among them, self-improving approaches (Zelikman et al., 2022; Gulcehre et al., 2023; Madaan et al., 2023; Qu et al., 2024; Ma et al., 2025) form a prominent branch, mainly because they can be easily applied and extended without external reward models (Lu et al., 2024a), predefined step decomposition (Liu et al., 2025), or specially designed reasoning structures (Li et al., 2025).

However, it is infeasible to directly adapt such language-based self-improving methods to vision language models (VLMs) (Liu et al., 2023; Bai et al., 2025). Previous self-improving approaches focus on textual coherence and the quality of the final answer (Zelikman et al., 2022; Zhang et al., 2024a), while they are unable to verify or rectify the visual hallucinations that persist in current VLMs (Tong et al., 2024; Li et al., 2024). Even worse, they may get stuck in speculative reasoning that privileges textual priors over real visual evidences (Favero et al., 2024; Wu et al., 2025). We claim that the key problem for self-improving in VLMs is: *how to rectify visual hallucinations in VLMs' reasoning paths for high-quality visual rationale generation.*

Our solution is built upon an interesting observation: *VLMs can see better during contrasting.* As shown in Fig. 1a, the VLM generates a wrong rationale with visual hallucinations given a single visual question answering (VQA) sample. Instead, when presented with a contrastive VQA pair, *i.e.*, two similar images with synonymous questions (Setting C in sub-figure 1b), the model captures fine-grained visual evidence more accurately and rectifies the erroneous rationale. Statistics of this



(a) Visual hallucinations within the reasoning paths can mislead the model. Contrasting within a contrastive VQA pair, the VLM may rectify its own hallucinations.



(b) Results and statistics of rectifying hallucinatory outputs by three settings. H: with hint; C: via contrasting.

Figure 1: **Contrasting makes the VLM see better.** (a) Contrastive VQA pairs compels a more accurate response. (b) Compared with a previous self-improving method STaR (Zelikman et al., 2022) that enhances the quality of reasoning with hints (ground-truth answers), contrasting with hints can rectify more cases. The blocks along the x-axis mark initial VLM failures. The color of each block indicates the outcome of rectifying: green for success and gray for failure. Statistical analyses suggest that contrasting with hints produces minimal additional errors and rectify more erroneous cases. Tested VLM is Qwen2.5VL-7B (Bai et al., 2025).

phenomenon on a group of failure cases are shown in Fig. 1b. Compared with the hints-only (Setting H in sub-figure 1b) self-improving (provide the model with the ground-truth answers), the hints and contrasting (Setting C&H in sub-figure 1b) setting not only prevents the model from making new errors but also leads to the rectification of its original hallucinations.

Motivated by this, we propose a new self-improving framework, **V**isual **C**ontrastive **S**elf-**Ta**ught **R**easoner (VC-STaR). VC-STaR contains three steps: (1) *think step by step* and generate a coarse rationale; (2) *compare* visual queries in a contrastive VQA pair and provide a contrastive analysis; (3) *rethink* and refine the coarse rationale via an LLM based on the contrastive analysis. In order to guarantee the scalability of VC-STaR, we also propose a task-agnostic contrastive VQA pair curation framework, which can be readily adapted to various VQA tasks, *e.g.*, reasoning (Lu et al., 2021b), math (Gao et al., 2025), chart Liu et al. (2024a), and OCR (Yuan et al., 2022). Specifically, we curate the contrastive VQA pairs within individual datasets, based on the similarity of both images and questions. We utilize these contrastive VQA pairs to generate faithful rationales, resulting in a novel **Vis**ual **Co**ntrastive **R**easoning dataset (VisCoR-55K) as illustrated in Fig. 2. Finetuning with VisCoR-55K enhances VLMs' visual reasoning capability.

VC-STaR achieves prominent results on a wide range of challenging benchmarks, including MMVP (Tong et al., 2024), HallusionBench (Guan et al., 2024b), MathVista (Lu et al., 2024b), MathVision Wang et al. (2024), and MMStar (Chen et al., 2024d). On the one hand, VC-STaR outperforms existing self-improving baselines. On the other hand, it exhibits a clear advantage over models trained on recently proposed reasoning datasets. The experimental results validate that visual reasoning capability of VLMs can be bootstrapped through the lens of contrast.

2 Related Works

Reasoning in Language. Dual-system theory (Kahneman, 2011) illustrates two systems in human cognition: a fast, intuitive *System* 1 and a slow, deliberate *System* 2 which is akin to emergent

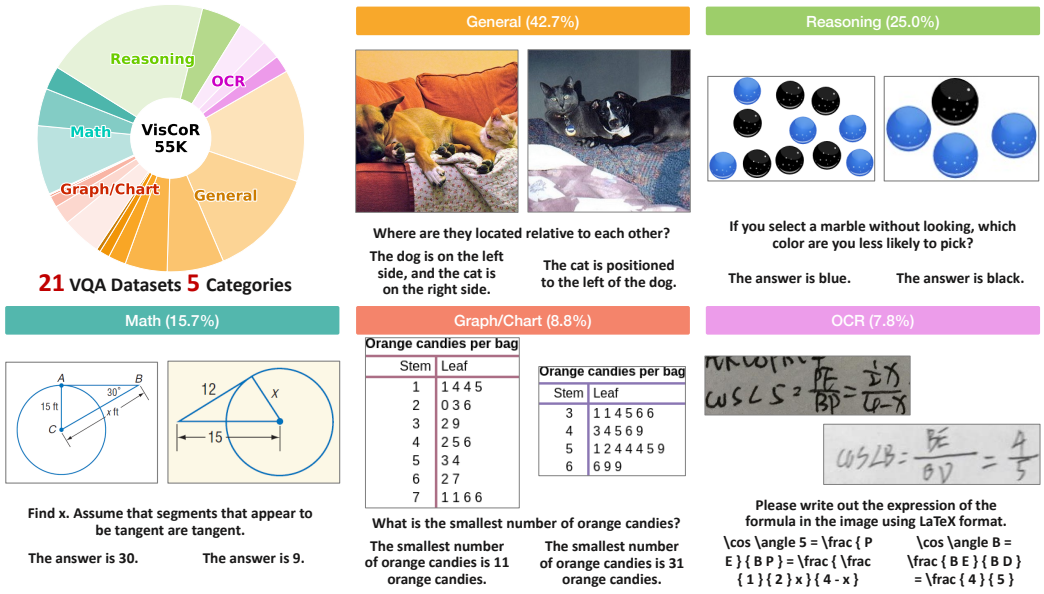


Figure 2: **VisCoR-55K.** We introduce the **Vis**ual **Co**ntrastive **R**easoning dataset (VisCoR-55K), a new collection of 55K high-quality visual reasoning samples. Spanning the domains of general VQA, reasoning, math, graph/chart, and OCR, each sample is created by leveraging a contrastive counterpart to generate a faithful rationale. Rationales are shown in the Sec. A.3.

reasoning capability of LLMs (Wei et al., 2022a). Consequently, reasoning enhancement (Wei et al., 2022b; Kojima et al., 2022) is considered a pathway to elevate LLMs' cognitive performance. One solution involves a reward model (Li et al., 2023b; Lu et al., 2024a), often coupled with Monte Carlo tree search (Hao et al., 2023; Zhang et al., 2024a), to discover optimal reasoning paths. However, this solution is constrained by the need of an auxiliary model and the requirement for reasoning step dividing (Liu et al., 2025). Another way employs macro reasoning actions (Gao et al., 2023; Khot et al., 2023; Yang et al., 2025a) to inject human prior knowledge, however, hand-crafted macro actions struggle to adapt to diverse reasoning scenarios. While reinforcement learning (Rafailov et al., 2023; Trung et al., 2024; Guo et al., 2025) has also attracted the attention, its success relies on the data format and the design of reward functions. Self-improving methods (Zhang et al., 2024a) offer a more scalable alternative, enabling LLMs to refine its own reasoning by constructing high-quality reasoning data (Wang et al., 2023b), utilizing ground-truth answers as hints (Zelikman et al., 2022), or leveraging internal feedback (Qu et al., 2024). With fewer external constraints, self-improving methods pave the way for more flexible and general language reasoners.

Reasoning in Vision. Human reasoning is stimulated not only by textual input but also by visually-related queries. Fostering the visual reasoning ability (Zhang et al., 2024c) for VLMs (Liu et al., 2023; Li et al., 2023a) is therefore a critical frontier topic. Early attempts often rely on external scaffolding like scene graphs (Mitra et al., 2024), macro actions (Xu et al., 2025; Dong et al., 2025), or bounding boxes highlighting key region in images (Shao et al., 2024). However, such approaches suffer from fundamental limitations: they are constrained by the data structure or tend to generate stereotyped reasoning paths. Despite these advances, the self-improving paradigm which has shown its effectiveness in text-only domain is underexplored for visual reasoning. The primary obstacle is the visual hallucinations embedded in reasoning paths cannot be easily rectified by existing text-centric self-improving frameworks (Zhang et al., 2024a; Zelikman et al., 2022; Qu et al., 2024). The proposed VC-STaR attempts to bridge this gap through the lens of contrast.

Power of Contrasting. Contrasting has shown effectiveness in a wide range of machine learning topics. By comparing different *views* (Tian et al., 2020), *e.g.*, data augmentations, of the same sample while distinguishing them from others (Wang & Isola, 2020), contrastive self-supervised learning methods (He et al., 2020; Grill et al., 2020; Radford et al., 2021; Liang et al., 2022; Pan et al., 2023) excel at learning potent feature representations. Explicitly cross-image contrasting is also studied under uni-modal setting (Pan et al., 2023; Ding et al., 2024; Chen et al., 2024a) and

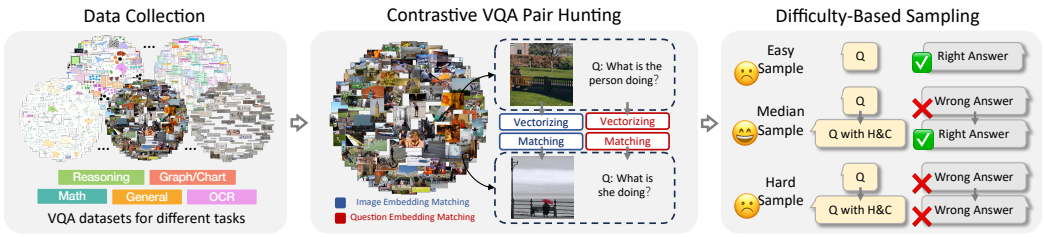


Figure 3: **Contrastive VQA pair curation pipeline.** To facilitate effective contrastive analysis, we curate corresponding challenging counterparts for VQA samples from a pool of diverse datasets. Each curated pair consists of two samples that share a synonymous question but feature distinct yet semantically similar images. Collected pairs are filtered by a difficulty-based sampling procedure.

multi-modal setting (Park et al., 2019; Kim et al., 2021; Yao et al., 2022; Dunlap et al., 2024). Based on these advancements, VLMs are endowed with robust capabilities for multi-image comprehension and comparison (Alayrac et al., 2022; Bai et al., 2025; Chameleon, 2025; Lin et al., 2025). Some prior works have leveraged contrasting to create better instruction-tuning data (Jiao et al., 2025; Ma et al., 2024). However, how contrasting can help visual reasoning remains an open question. We observe that VLMs' inherent comparative ability can be repurposed to actively suppress its own visual hallucinations, bootstrapping their visual reasoning capability. This discovery offers a new perspective about the power of contrasting in reasoning.

3 VISUAL CONTRASTIVE SELF-TAUGHT REASONER (VC-STAR)

Let θ be a VLM and $\mathcal{D} = \{(v_i, q_i, a_i)\}_{i=1}^{N}$ be a set of visual question answering (VQA). The VQA set consists N triplets, where v_i, q_i, and a_i represent the i-th image, question, and corresponding ground-truth answer, respectively. Following previous self-taught reasoners (Zelikman et al., 2022; Madaan et al., 2023), the original VQA dataset $\mathcal{D}$ can be enriched by generating a rationale r with θ for each triplet, which transforms $\mathcal{D}$ into a a visual reasoning dataset $\mathcal{R} = \{(v_i, q_i, a_i, r_i)\}_{i=1}^{M}$. However, as metioned in Sec. 1, rationale r_i may be contaminated by visual hallucinations. Motivated by the observation illustrated in Fig. 1, VC-STaR aims to refine rationale r_i into a more faithful one $\tilde{r_i}$ by contrasting the (v_i, q_i, a_i) with a contrastive VQA counterpart sample $(\hat{v_i}, \hat{q_i}, \hat{a_i})$ where q_i is synonymous with $\hat{q_i}$ and v_i shares similar context with $\hat{v_i}$. The contrastive VQA pairs $\mathcal{P} = \{\big((v_i, q_i, a_i), (\hat{v_i}, \hat{q_i}, \hat{a_i})\big)\}_{i=1}^{K}$ support the contrasting and rationale refining process. The contrastive VQA pairs are curated by searching $(\hat{v_i}, \hat{q_i}, \hat{a_i})$ for (v_i, q_i, a_i) within diverse data groups in $\mathcal{D}$ for different VQA tasks, ensuring the generalization of VC-STaR. The VC-STaR is designed to address two key challenges: (1) how to curate meaningful contrastive VQA pairs; (2) how to transfer the fine-grained discriminative ability from dual-image contrasting to refine the single-image reasoning. Sec. 3.1 elaborates on the pipeline for curating contrastive VQA pairs. Building upon this foundation, Sec. 3.2 introduces our *contrasting and rethinking* procedure which embeds the dual-image comparison into a new reasoning path, guided by an LLM, to produce a more faithful rationale. The refined rationales are then used to construct a new reasoning dataset $\tilde{\mathcal{R}} = \{(v_i, q_i, a_i, \tilde{r_i})\}_{i=1}^{L}$, which we name the **Vis**ual **Co**ntrastive **R**easoning dataset (VisCoR-55K). The VLM θ is updated to a new version $\tilde{\theta}$ with improved reasoning capability by finetuning on VisCoR-55K.

3.1 CONTRASTIVE VQA PAIR CURATION

To ensure the generalization of VC-STaR, the contrastive VQA pair curation pipeline should be flexible enough across a wide spectrum of VQA tasks. For better contrasting, each contrastive VQA pair $\big((v_i, q_i, a_i), (\hat{v_i}, \hat{q_i}, \hat{a_i})\big)$ should possess three key properties: (1) *q_i and $\hat{q_i}$ are synonymous.* This shared question acts as a semantic anchor, grounding the two images v_i and $\hat{v_i}$ at the same point in the semantic space. The images thus represent different manifestations of this anchor, providing a solid basis for contrasting; (2) *v_i and $\hat{v_i}$ are visually similar.* v_i and $\hat{v_i}$ should not be trivially distinct but exhibit visual similarity, creating a confusing contrasting. This visual proximity compels VLMs to engage in fine-grained contrasting to discriminate subtle differences; (3) *q_i is reasoning dependent.*

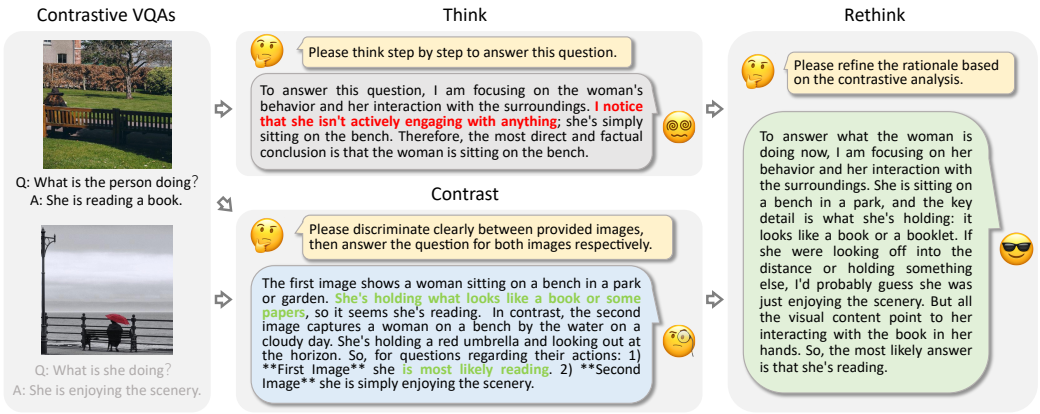


Figure 4: **Faithful rationale generation pipeline.** A contrastive analysis can be obtained based on the curated contrastive VQA pair. Leveraging the property of VLMs illustrated in Fig. 1, the contrastive analysis is then used to trigger a rethinking procedure, which refines the naive rationale into a more faithful one. This pipeline is designed to generate rationales for supervised finetuning.

q_i should be reasoning-provoking rather than one that can be solved by a straightforward answer. To achieve these requirements, as illustrated in Fig. 3, we propose a three-stage curation pipeline:

Data Collection. We collect 21 VQA datasets spanning five categories: reasoning (Zhang et al., 2019; Lu et al., 2021b), graph/chart (Kembhavi et al., 2016; Mathew et al., 2022; Masry et al., 2022; Tang et al., 2023; Lu et al., 2023; Liu et al., 2024a), math (Lu et al., 2021a; Cao & Xiao, 2022; Gao et al., 2025), general (Zhu et al., 2016; Johnson et al., 2017; Acharya et al., 2019; Schwenk et al., 2022; Wang et al., 2023a; Chen et al., 2024b), and OCR (ICDAR, 2019; Kiela et al., 2020; Yuan et al., 2022; Zhang et al., 2024b). This broad collection enriches the diversity of our curated pairs, which ensures the generalization ability of the finetuned model.

Contrastive VQA Pair Hunting. In order to compute the similarity of VQA pairs, we first represent the question q_i and the image v_i by high-dimensional embeddings, denoted as e_i^q and e_i^v respectively. We use GTE (Li et al., 2023c) text embeddings to represent the questions. In terms of image embedding, existing models fall into two types, *i.e.*, vision-language contrastive learning approaches (Radford et al., 2021; Tschannen et al., 2025) while vision-only self-supervised learning methods (Zhang et al., 2023; Oquab et al., 2024). The former ones mainly capture at global semantic information, while the later ones are good at instance discrimination. Neither of them are generic enough to adapt to the diverse domains. To tackle the dilemma, we build a versatile visual embedding model based on ID-based visual metric learning (Ypsilantis et al., 2024; An et al., 2023). Hunting for a counterpart $(\hat{v}_i, \hat{q}_i, \hat{a}_i)$ is then performed dataset-by-dataset. A sample (v_j, q_j, a_j) is recalled as a valid counterpart if it satisfies: $\gamma(e_i^v, e_j^v) < \phi_v$ and $\gamma(e_i^q, e_j^q) < \phi_q$, where $\gamma(\cdot, \cdot)$ is the cosine distance, and ϕ_v and ϕ_q are pre-defined thresholds for visual and question similarity, respectively. Any sample that fails to meet both conditions is dropped.

Difficulty-Based Data Sampling. For the goal of developing visual reasoning capability, q_i should be a difficult question requires reasoning rather than a straightforward one. We define the levels of difficulty based on the performance of VLM θ: (1) *easy samples* are with the simple q_i which can be correctly answered by θ without any auxiliary help; (2) *median samples* are with q_i which makes θ initially fails but succeeds when contrasting with $(\hat{v}_i, \hat{q}_i, \hat{a}_i)$ based on provided hint a_i (the $C\&H$ setting introduced in Fig. 1); (3) *hard samples* are the ones with q_i that cannot be correctly addressed by θ even with the help of contrasting. We only keep median-difficult contrastive VQA pairs for the rationale generating.

3.2 Contrasting and Rethinking

Rationales in the reasoning dataset $\mathcal{R} = \{(v_i, q_i, a_i, r_i)\}_{i=1}^M$ generated by the VLM θ itself include visual hallucinations. To achieve the goal:

$$\mathcal{R} = \{(v_i, q_i, a_i, r_i)\}_{i=1}^M \rightarrow \tilde{\mathcal{R}} = \{(v_i, q_i, a_i, \tilde{r}_i)\}_{i=1}^M , \tag{1}$$

where $\tilde{r_i}$ is the rectified rationale, we use the contrastive VQA counterpart $(\hat{v_i}, \hat{q_i}, \hat{a_i})$ to provoke a rethinking action to refine r_i into $\tilde{r_i}$. As illustrated in Fig. 4, this pipeline includes three steps:

Thinking step. Following the design of Zelikman et al. (2022) to provide the VLM θ with ground-truth answer a_i as hints, we prompt the VLM θ to generate the coarse rationale r_i for the target VQA sample (v_i, q_i, a_i) as follows:

$$r_i = f(v_i, q_i, a_i|\theta, \delta^t)\,, \tag{2}$$

where f is a inference process with a "thinking prompt" δ^t. Details of δ^t are in the Sec. A.2.

Contrasting step. Asking the VLM θ to compare the target VQA sample (v_i, q_i, a_i) with its contrastive counterpart $(\hat{v_i}, \hat{q_i}, \hat{a_i})$ results a contrastive analysis c_i which may provide more faithful visual information:

$$c_i = f(\big((v_i, q_i, a_i), (\hat{v_i}, \hat{q_i}, \hat{a_i})\big)|\theta, \delta^c)\,, \tag{3}$$

where the δ^c is the "contrasting prompt". When a_i has the same meaning with $\hat{a_i}$, δ^c requires summarizing the common patterns of v_i and $\hat{v_i}$; When a_i is different from $\hat{a_i}$, δ^c expects the analysis about the fine-grained differences between v_i and $\hat{v_i}$. Details of δ^c are in the Sec. A.2.

Rethinking step. As demonstrated in Fig. 1, c_i is more trustworthy than r_i. Hence, we adopt a LLM ψ to transfer the information from c_i to a new reasoning path according to r_i:

$$\tilde{r_i} = f(r_i, c_i|\psi, \delta^r)\,, \tag{4}$$

where δ^r is the "rethinking prompt" which asks the LLM ψ to rectify the visual hallucinations in r_i according to the visual information from c_i. δ^r requires LLM ψ responding like directly answer the question q_i, details are in the Sec. A.2.

To ensure the quality of the $\tilde{\mathcal{R}}$, we finalize the visual reasoning dataset by employing a text-matching post-processing to filter out samples that contain incorrect reasoning patterns. The final visual reasoning dataset contains 55K VQA samples with corresponding rationales, *a.k.a.*, the VisCoR-55K.

4 Experiments

Section 4.1 details our experimental setup, including the supervised finetuning process and the benchmarks used to evaluate the effectiveness of the VC-STaR. In Section 4.2, we present a comprehensive performance comparison. As a self-improving method for visual reasoning, we benchmark VC-STaR against two primary groups: (1) other self-improving baselines adaptable to visual reasoning, and (2) models trained on off-the-shelf visual reasoning datasets. Finally, Section 4.3 provides in-depth ablation studies on designs of our method, including the contrastive VQA pair construction, the generalization on other base models, the difficulty sampling strategy, and the effect of the types of contrastive VQA counterpart.

4.1 Setup

Implementation Details. Using the LLaMA-factory framework (Zheng et al., 2024), we finetune the model for 3 epochs via full-parameter supervised finetuning (SFT), with the vision tower's parameters frozen. The SFT utilizes a learning rate of $1e$-5, a batch size of 256. The inference process of the finetuned model does not require such a contrastive pipeline illustrated in Fig. 4, and it follows the standard inference paradigm of VLMs. As for the curation of contrastive VQA pair, the question similarity threshold ϕ_q is set to 0.15 and the visual similarity threshold ϕ_v is set as 0.5 for datasets of general images. For the datasets including icon, geometry, chart or graph images, the visual similarity threshold ϕ_v is set as 0.3. The LLM ψ used in the rethinking step of our rationale generation pipeline is the open-sourced Qwen2.5-72B.

Evaluation Benchmarks. We employ 6 benchmarks designed to assess its robustness against hallucination, mathematical reasoning, and general abilities. The MMVP (Tong et al., 2024) and Hallusion (Guan et al., 2024a) benchmakrs focus on visual hallucination, and the MathVista (Lu et al., 2024b) and MathVision (Wang et al., 2024) benchmarks are about the mathematical reasoning. The MMStar (Chen et al., 2024c) is a highly curated benchmark, composed of purified samples from multiple benchmarks, *e.g.*, MMMU (Yue et al., 2024) and MMBench Liu et al. (2024b). The MME-RealWorld benchmark Zhang et al. (2025b) is a large-scale, human-annotated benchmark for difficult, real-world tasks. Therefore, MMStar and MME-RealWorld are suitable to evaluate the general perceptual and cognitive abilities under varied scenarios.

Table 1: Performance comparison with self-improving baselines and the models trained on off-the-shelf visual reasoning datasets on hallucination, math, and general benchmarks. We adopt the Qwen2.5VL-7B as our base model, and report its reasoning performance as a baseline. MME-RW is short for MME-RealWorld Zhang et al. (2025b); R1-OV is short for R1-Onevision (Yang et al., 2025b). Blue (red) numbers in parentheses represent performance gains (drops) relative to the baseline. The best performance is in **boldface**, and the second best is underlined.

Method \ Bench.	Hallucination		Math		General		Avg.
	MMVP	Hallusion	MathVista	MathVision	MMStar	MME-RW	
Base Model	70.0	53.1	68.4	24.0	61.8	55.9	55.5
VQA SFT	74.3(+4.3)	54.2(+1.1)	65.4(-3.0)	19.4(-4.6)	59.7(-2.1)	56.8(+0.9)	55.0(-0.5)
Self-Improving Approaches							
STaR(2022)	73.0(+3.0)	55.9(+2.8)	66.9(-1.5)	19.8(-4.2)	58.9(-2.9)	58.1(+2.2)	55.4(-0.1)
Verifier(2024a)	73.7(+3.7)	53.2(+0.1)	67.0(-1.4)	20.3(-3.7)	58.2(-3.6)	56.7(+0.8)	54.9(-0.6)
Feedback(2024)	75.0(+5.0)	53.4(+0.3)	68.8(+0.4)	22.1(-1.9)	63.2(+1.4)	56.0(+0.1)	56.4(+0.9)
Off-the-Shelf Visual Reasoning Datasets							
Virgo(2025)	68.0(-2.0)	47.2(-5.9)	63.5(-4.9)	21.5(-2.5)	59.7(-2.1)	29.4(-26.5)	48.2(-7.3)
LLaVA-CoT(2025)	71.7(+1.7)	50.3(-2.8)	68.4(+0.0)	24.4(+0.4)	63.1(+1.3)	**59.3(+3.4)**	56.2(+0.7)
R1-OV(2025b)	68.0(-2.0)	55.8(+2.7)	68.2(-0.2)	**25.4(+1.4)**	53.2(-8.6)	46.3(-9.6)	52.8(-2.7)
LPT(2025)	74.0(+4.0)	53.4(+0.3)	69.2(+0.8)	24.2(+0.2)	**64.3(+2.5)**	56.1(+0.2)	56.9(+1.4)
VC-STaR(Ours)	**75.7(+5.7)**	**56.3(+3.2)**	**69.7(+1.3)**	25.3(+1.3)	62.4(+0.6)	**59.3(+3.4)**	**58.1(+2.6)**

4.2 Main Results

Comparison with the base model. To evaluate the effectiveness of our approach, we employ Qwen2.5VL-7B as the base model and adopt the "think step by step" prompt to enable chain-of-thought reasoning. We compare our method against this baseline, with results summarized in Table 1. VC-STaR demonstrates consistent performance gains across diverse challenging benchmarks, achieving an average improvement of 2.4%. Notably, it yields substantial improvements of 5.7% and 3.2% on MMVP and the Hallusion Benchmark, respectively, validating its efficacy in mitigating hallucinations within the reasoning process. Our approach also shows its enhanced reasoning capabilities on mathematical benchmark, *i.e.*, MathVista and MathVision. Furthermore, the improvement on the MMStar and MME-RealWorld underscore the generalizability of the VC-STaR under varied challenging general-purpose scenarios.

For qualitative validation, Figure 5 provides visual comparisons that offer deeper insights. The visualizations reveal that our model excels at grounding its textual rationales in the corresponding visual evidence. This capability remains robust even when confronted with visually complex patterns, thereby effectively mitigating hallucinations.

Comparison with self-improving baselines. We reproduce three self-improving baselines and compare VC-STaR against them. Each baseline is applied to the Qwen2.5VL-7B base model and generates rationales on VisCoR-55K for finetuning, differing in their core improvement mechanism: (1) STaR (Zelikman et al., 2022): Leverages ground-truth answers to regenerate rationales for incorrect predictions. (2) Verifier (Lu et al., 2024a): Filters out visually hallucinated rationales via a self-verification step (Zhang et al., 2025a) to ensure visual grounding. (3) Feedback (Qu et al., 2024): Refines rationales based on self-generated feedback in a recursive manner. Table 1 reveals a critical trade-off: existing self-improving methods boost performance on hallucination benchmarks at the expense of math and general capabilities. Our approach mitigates this pattern and achieves robust, consistent performance gains.

Comparison with off-the-shelf visual reasoning datasets. We also evaluate VC-STaR against base model finetuned on four off-the-shelf visual reasoning datasets. These datasets represent diverse strategies for rationale generation. For instance, Virgo (Du et al., 2025) makes the VLM think slowly

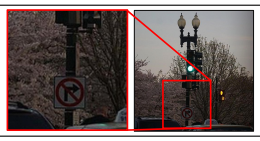
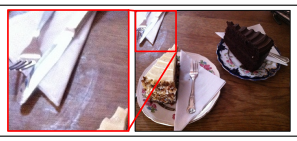
Question	 What is the meaning of the white sign?	 How many forks are visible in the image?
Qwen2.5VL-7B	The white sign in the picture has a red circle with a line through it, which is a universal symbol indicating "no parking." This sign is typically used to prohibit vehicles from parking in a specific area for a certain period of time. The exact duration or conditions under which parking is prohibited would be specified by additional text on the sign, but the symbol itself clearly indicates that parking is not allowed.	There are two forks visible in the image, one on each plate with the cake slices.
Qwen2.5VL-7B "think step by step"	The white sign in the image has a red circle with a line through it, which is a universal symbol indicating "no" or "prohibited." Specifically, this sign is a "No Parking" sign. It means that parking is not allowed in the area where the sign is posted. This type of sign is typically used to prevent vehicles from blocking traffic or causing congestion in certain areas.	To determine how many forks are visible in the image, let's follow these steps: 1. Identify the objects: Look closely at the image and identify all the objects that could be forks. 2. Count the forks: - There is one fork placed on the napkin next to the slice of cake on the left plate. - There is another fork partially visible on the right side of the image, resting on the plate with the chocolate cake. 3. Verify the count: Double-check to ensure no additional forks are hidden or partially obscured. After following these steps, we can conclude that there are two forks visible in the image.
Qwen2.5VL-7B + VC-STaR	<think>The white sign in the image contains a symbol that indicates "No Right Turn." This is evident because the sign features an arrow pointing to the right, which is universally recognized as a prohibition against turning right at that location. The absence of any conflicting symbols or additional text confirms this interpretation. Furthermore, the consistency in traffic signage conventions reinforces the conclusion that the sign's purpose is to prohibit right turns.</think><answer>The white sign means "No Right Turn." It prohibits vehicles from making a right turn at that particular location.</answer>	<think>The white sign in the image contains a symbol that indicates "No Right Turn." This is evident because the sign features an arrow pointing to the right, which is universally recognized as a prohibition against turning right at that location. The absence of any conflicting symbols or additional text confirms this interpretation. Furthermore, the consistency in traffic signage conventions reinforces the conclusion that the sign's purpose is to prohibit right turns.</think><answer>The white sign means "No Right Turn." It prohibits vehicles from making a right turn at that particular location.</answer>

Figure 5: **Qualitative Comparison with base model.** The second row shows the directly response from the base model, the third row shows the response when the base model is prompted to "think stey by step", the last row shows the model improved with our VC-STaR. We highlight the key visual evidences with red boxes for clarity of visualization. More results are in Sec. A.3.

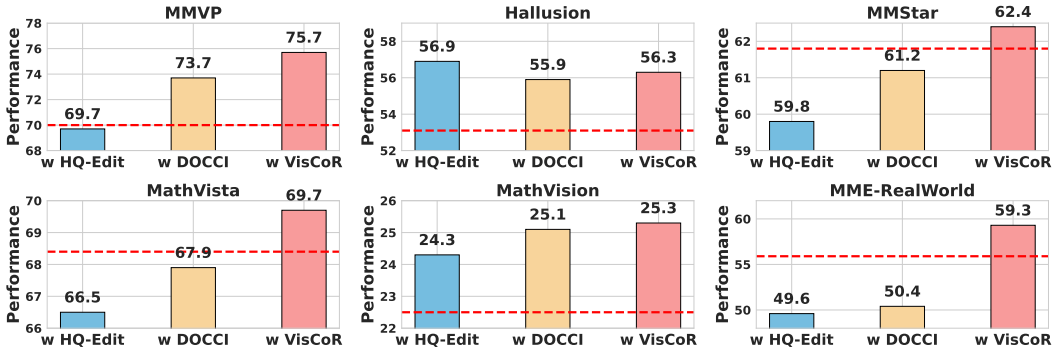


Figure 6: **Performance comparison with other contrastive VQA pair construction strategies.** Rationales in all settings are generated from the proposed VC-STaR. The red dashed line represents the base model (Qwen2.5VL-7B) performance.

with purely textual rationales. In contrast, LLaVA-CoT (Xu et al., 2025) leverages hand-crafted templates filled by the powerful GPT-4o (OpenAI, 2024a). Other approaches first convert visual information into text; R1-Onevision (Yang et al., 2025b) generates rationales from image captions using the DeepSeek-R1 model (Guo et al., 2025), while Long Perceptual Thought (LPT) (Liao et al., 2025) extends this by using dense captions (Onoe et al., 2024) and keywords like "wait" to elicit more detailed outputs from a similar LLM. In our experiments, we directly finetune the base model on each of these datasets. Based on the results shown in Table 1, we can draw the following conclusion: (a) Enhancing visual reasoning with purely textual rationales from Virgo is ineffective. This strongly indicates that visual modality matters. (b) The model trained on LLaVA-CoT suffers limited improvement, which demonstrates that the hand-crafted template struggle to generalize across diverse VQA tasks. (c) The models trained on datasets generated by DeepSeek-R1 based on captions achieves notable improvements. However, the performance gap between them and ours highlights the clear advantage of our visually-native approach over relying on textual captions.

4.3 Analysis

Can contrastive VQA pairs constructed in other ways? To answer this, we explore alternative strategies for curating contrastive VQA pairs. The first strategy is editing-based, utilizing the HQ-

Table 2: Evaluation of the effect of VC-STaR on other base models. Blue numbers in parentheses represent performance gains.

Model	VC-STaR	Hallusion	MathVision	MMStar
Qwen2.5VL 3B	✘	46.9	18.4	55.0
	✔	53.2(+6.3)	21.9(+3.5)	55.7(+0.7)
InternVL2.5 8B	✘	48.2	21.3	61.1
	✔	55.4(+7.2)	23.4(+2.1)	62.5(+1.4)

Table 3: Effect of the easy samples adding to VisCoR-55K. Red numbers in parentheses represent performance drops.

N_{easy}	Hallusion	MathVision	MMStar
0k	56.3	25.3	62.4
+20k	52.2(-4.1)	23.3(-2.0)	61.3(-1.1)
+40k	55.7(-0.6)	21.9(-3.4)	59.5(-2.9)

Table 4: Analysis about the effect of positive and negative contrastive VQA counterparts on GQA benchmark. We adopt the Qwen2.5VL-7B as our base model, and report its reasoning performance as a baseline. QR: query for relationships; QA: query for attributes; QG: query for global information; QC: query for category; CA: comparing of attribute; CC: choosing the object of one certain category; CAt: choosing the object of one certain attribute. Blue (red) numbers in parentheses represent performance gains (drops) relative to the baseline.

Setting	Pos.	Neg.	QR	QA	QG	QC	CA	CC	CAt	Total
Base Model	-	-	43.8	51.4	31.5	44.4	30.2	60.6	44.4	45.4
VC-STaR	✔	✘	48.3(+4.5)	52.8(+1.4)	46.8(+15.3)	57.1(+12.7)	42.9(+12.7)	60.1(-0.5)	44.4(+0.0)	50.6(+5.2)
VC-STaR	✘	✔	51.6(+7.8)	56.8(+5.4)	33.9(+2.4)	59.2(+14.8)	46.0(+15.8)	70.8(+10.2)	66.7(+22.3)	53.7(+8.3)
VC-STaR	✔	✔	53.5(+9.7)	57.3(+5.9)	46.3(+14.8)	55.6(+11.2)	36.5(+6.3)	72.7(+12.1)	55.6(+11.2)	54.7(+9.3)

Edit dataset (Hui et al., 2025). By prompting an LLM to create questions from editing instructions, we generate pairs where an original and an edited image yield different answers. The second strategy is caption-based, leveraging a dense caption dataset, *i.e.*DOCCI (Onoe et al., 2024). For this, we instruct an LLM to parse dense captions of visually similar images and generate a question that hinges on their subtle differences. For both strategies, we generate rationales for these newly created contrastive pairs using our proposed VC-STaR and finetune the Qwen2.5VL-7B. The results, presented in Fig. 6, lead to several observations: (a) VC-STaR is broadly effective, but performance is data-dependent. This is attributable to the biased data distribution of HQ-Edit and DOCCI, highlighting a key limitation of their curation scope. (b) VisCoR-55K includes contrastive pairs from a broader range of reasoning tasks, resulting in a more balanced performance.

Does VC-STaR generalize to other base models? We conduct experiments on Qwen2.5VL-3B and InternVL2.5-8B (Chen et al., 2025). Following the same self-improving procedure, we use VC-STaR to generate visual reasoning datasets from our VisCoR contrastive pairs, specifically for the two base models. We then finetune the Qwen2.5VL-3B and InternVL2.5-8B via the LLaMA-factory and SWIFT (Zhao et al., 2025), respectively. The results, presented in Table 2, demonstrate the model-agnostic effectiveness of our approach. These consistent and significant gains confirm that VC-STaR is a versatile and broadly applicable strategy for enhancing the visual reasoning ability.

What is the effect of easy samples on visual reasoning? Starting with our VisCoR-55K datasets, we incrementally add easy samples of two batches with 20K each. As illustrated in Table 3, we observe that the inclusion of easy samples is harmful. Specifically, when the number of easy samples increases, performance decreases. Therefore, we do not use the easy samples to avoid the potential "overthink" for straightforward problems.

How the contrastive VQA pairs of different types contribute? A contrastive VQA pair can be categorized as "positive" if both samples yield the same answer, and "negative" if their answers differ. To investigate the respective contributions of these two types of counterparts to our method's performance, we conducted a controlled experiment on the GQA dataset (Hudson & Manning, 2019). The structured nature of GQA allows for the reliable curation of both positive and negative pairs via simple text matching. We applied VC-STaR to three distinct training sets: one generated from only positive contrastive pairs, one from only negative pairs, and a combined set including both. The results, detailed in Table 4, reveal a clear and significant trend. While both types of pairs are beneficial, negative counterparts are substantially more effective than positive ones, and their combination

yields the optimal total gain, highlighting their complementary roles. We attribute the superior efficacy of negative counterparts to their ability to induce stronger semantic contrast. Accordingly, our approach incorporates both positive and negative pairs without restriction to achieve optimal gain.

5 DISCUSSION

Rethinking VC-STaR from a cognitive perspective. Learning and reasoning are inherently comparative and contrastive processes. Humans rarely learn concepts in isolation. Instead, our human-beings refine our understanding by comparing examples, identifying distinguishing features, and reasoning through analogies and differences. The prototype theory concludes this cognitive behavior as that our human-beings identify new identities by comparing them with the prototype concept (Rosch, 1975). Besides, the structure-mapping theory says that analogical reasoning can recognize the relationships shared by two domains (Gentner, 1983). This mapping can be treated as a fine-grained contrasting process. In our work, the contrasting process provide an opportunity to learn visual concept from the prototype, and our rethinking strategy reinforce the structure-mapping by generating new reasoning paths via contrasting. Through this work, we hope to highlight the potential of porting such human-like cognitive behaviors to the domain of reasoning.

Computational cost of the VC-STaR. Our approach introduces a modest cost in data generation, while maintaining a fine-tuning cost comparable to existing self-improving frameworks like STaR. We train the 7B model using the LLaMAFactory framework once the VisCoR-55K dataset is generated. This process takes approximately 4 hours on a single node with $8\times$A800 GPUs. This cost is nearly identical to the standard fine-tuning procedure in approach like STaR. The additional cost stems from our data generation pipeline. The one-time contrastive VQA pair mining step processes approximately 7 samples per second, accelerated by our implementation featuring multi-processing concurrency and batch-wise computation. The rethinking step generates reasoning paths at about 1 sample per second, thanks to the advanced vLLM framework Kwon et al. (2023) for fast inference. We argue this is a reasonable and acceptable investment to create our reasoning dataset.

6 CONCLUSION

In this work, we demonstrate that visual hallucination, a critical bottleneck for VLM reasoning, can be effectively mitigated to achieve better visual reasoning ability through the lens of contrast. Our key finding is that VLMs can see better by contrast. This phenomenon motivates us to propose the VC-STaR. The VC-STaR refines hallucinatory reasoning paths through analysis over curated contrastive VQA pairs, which yields our high-quality VisCoR-55K. Finetuning on VisCoR-55K delivers a consistent performance gain across five benchmarks targeting hallucination, math, and general reasoning, significantly surpassing other self-improving baselines and models trained on state-off-the-art visual reasoning datasets.

ETHICS STATEMENT

We confirm that this research adheres to the ICLR Code of Ethics. The datasets utilized for this work are all publicly accessible and intended for academic research. No commercial or private data was used, ensuring that our study is free from associated privacy concerns. The authors take full responsibility for the ethical conduct of this research.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results and to foster future research, we commit to making our resources publicly available upon the acceptance of this paper. These resources will include: our newly curated VisCoR-55K dataset, and the final finetuned model weights. Crucially, our dataset release will not only contain the final rationales generated by VC-STaR but also the corresponding contrastive VQA counterparts used to produce them. We believe this unique data will be a valuable resource for the community. We hope that providing this data will enable researchers to explore the role of contrast in promoting visual reasoning and other cognitive abilities across a wider range of settings, such as reinforcement learning.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. *AAAI*, pp. 8076–8084, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, pp. 23716–23736, 2022.

Xiang An, Jiankang Deng, Kaicheng Yang, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Unicom: Universal and compact representation learning for image retrieval. In *ICLR*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, pp. 1511–1520, 2022.

Chameleon. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818.

Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatiotemporal state space model. *TGRS*, pp. 1–20, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, pp. 370–387, 2024b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, pp. 27056–27087, 2024c.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024d.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Lei Ding, Jing Zhang, Haitao Guo, Kai Zhang, Bing Liu, and Lorenzo Bruzzone. Joint spatio-temporal modeling for semantic change detection in remote sensing images. *TGRS*, 62:1–14, 2024.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *CVPR*, pp. 9062–9072, 2025.

Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm, 2025. URL https://arxiv.org/abs/2501.01904.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *CVPR*, pp. 24199–24208, 2024.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *CVPR*, pp. 14303–14312, 2024.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *ICLR*, 2025.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *ICML*, pp. 10764–10799, 2023.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2): 155–170, 1983.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, pp. 21271–21284, 2020.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pp. 14375–14385, 2024a.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pp. 14375–14385, 2024b.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling, 2023. URL https://arxiv.org/abs/2308.08998.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *EMNLP*, pp. 8154–8173, 2023.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Cihang Xie, and Yuyin Zhou. HQ-edit: A high-quality dataset for instruction-based image editing. In *ICLR*, 2025.

ICDAR. Overview - icdar 2019 robust reading challenge on scanned receipts ocr and information extraction, 2019. URL https://rrc.cvc.uab.es/?ch=13.

Qirui Jiao, Daoyuan Chen, Yilun Huang, Bolin Ding, Yaliang Li, and Ying Shen. Img-diff: Contrastive data synthesis for multimodal large language models. In *CVPR*, pp. 9296–9307, 2025.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pp. 2901–2910, 2017.

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pp. 235–251, 2016.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *ICLR*, 2023.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*, pp. 2611–2624, 2020.

Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyun a Park, and Gunhee Kim. Viewpoint-agnostic change captioning with cycle consistency. 2021 ieee. In *ICCV*, pp. 2075–2084, 2021.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *NeurIPS*, pp. 22199–22213, 2022.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pp. 611–626, 2023.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *NeurIPS*, pp. 17044–17068, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023a.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In *ACL*, pp. 5315–5333, 2023b.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023c. URL https://arxiv.org/abs/2308.03281.

Zhuoqun Li, Haiyang Yu, Xuanang Chen, Hongyu Lin, Yaojie Lu, Fei Huang, Xianpei Han, Yongbin Li, and Le Sun. DeepSolution: Boosting complex engineering solution design via tree-based exploration and bi-point thinking. In *ACL*, pp. 4380–4396, 2025.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, pp. 17612–17625, 2022.

Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuna. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. In *COLM*, 2025.

Wei Lin, Muhammad Jehanzeb Mirza, Sivan Doveh, Rogerio Feris, Raja Giryes, Sepp Hochreiter, and Leonid Karlinsky. Comparison visual instruction tuning. In *CVPR*, pp. 2973–2983, 2025.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *ICLR*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pp. 34892–34916, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, pp. 216–233, 2024b.

Yuliang Liu, Junjie Lu, Zhaoling Chen, Chaofeng Qu, Jason Klein Liu, Chonghan Liu, Zefan Cai, Yunhui Xia, Li Zhao, Jiang Bian, et al. Adaptivestep: Automatically dividing reasoning step through model confidence. *ICML*, 2025.

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yunlong Feng, and Zhijiang Guo. Autopsv: Automated process-supervised verifier. *NeurIPS*, pp. 79935–79962, 2024a.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *ACL*, 2021a.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *NeurIPS DB*, 2021b.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *ICLR*, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024b.

Ji Ma, Wei Suo, Peng Wang, and Yanning Zhang. C3 l: content correlated vision-language instruction tuning data generation via contrastive learning. In *IJCAI*, 2024.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. S^2R: Teaching LLMs to self-verify and self-correct via reinforcement learning. In *ACL*, pp. 22632–22654, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *NeurIPS*, pp. 46534–46594, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ACL findings*, pp. 2263–2279, 2022.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pp. 1697–1706, 2022.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *CVPR*, pp. 14420–14431, 2024.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of Connected and Contrasting Images. In *ECCV*, 2024.

OpenAI. Gpt-4o system card, 2024a. URL https://arxiv.org/abs/2410.21276.

OpenAI. Openai o1 system card, 2024b. URL https://arxiv.org/abs/2412.16720.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. ISSN 2835-8856.

Zhiyu Pan, Yinpeng Chen, Jiale Zhang, Hao Lu, Zhiguo Cao, and Weicai Zhong. Find beauty in the rare: Contrastive composition feature clustering for nontrivial cropping box regression. In *AAAI*, pp. 2011–2019, 2023.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *ICCV*, pp. 4624–4633, 2019.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. *NeurIPS*, pp. 55249–55285, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, pp. 53728–53741, 2023.

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pp. 146–162, 2022.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *NeurIPS*, pp. 8612–8642, 2024.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. In *ACL*, pp. 7268–7298, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *ECCV*, pp. 776–794, 2020.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pp. 9568–9578, 2024.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. ReFT: Reasoning with reinforced fine-tuning. In *ACL*, pp. 7601–7614, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning, 2023a. URL https://arxiv.org/abs/2311.07574.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-vision dataset. In *NeurIPS DB*, 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pp. 9929–9939, 2020.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *TMLR*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pp. 24824–24837, 2022b.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. Combating multimodal llm hallucination via bottom-up holistic reasoning. *AAAI*, pp. 8460–8468, 2025.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *ICCV*, 2025.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Minkai Xu, Joseph E Gonzalez, Bin Cui, and Shuicheng Yan. Supercorrect: Advancing small llm reasoning with thought template distillation and self-correction. *ICLR*, 2025a.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025b. URL https://arxiv.org/abs/2503.10615.

Linli Yao, Weiying Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *AAAI*, pp. 3108–3116, 2022.

Nikolaos-Antonios Ypsilantis, Kaifeng Chen, André Araujo, and Ondřej Chum. Udon: Universal dynamic online distillation for generic image representations. *NeurIPS*, pp. 86836–86859, 2024.

Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. Syntax-aware network for handwritten mathematical expression recognition. In *CVPR*, pp. 4553–4562, 2022.

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, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*, pp. 15476–15488, 2022.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *CVPR*, pp. 5317–5327, 2019.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. In *NeurIPS*, pp. 64735–64772, 2024a.

Fuxiang Zhang, Jiacheng Xu, Chaojie Wang, Ce Cui, Yang Liu, and Bo An. Incentivizing llms to self-verify their answers, 2025a. URL https://arxiv.org/abs/2506.01369.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *ICLR*, 2023.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding, 2024b. URL https://arxiv.org/abs/2306.17107.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *ICLR*, 2025b.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *TMLR*, 2024c.

Y. Zhao, J. Huang, J. Hu, X. Wang, Y. Mao, D. Zhang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *AAAI*, pp. 29733–29735, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *ACL*, pp. 400–410, 2024.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pp. 4995–5004, 2016.

A Appendix

A.1 Details about VisCoR-55K

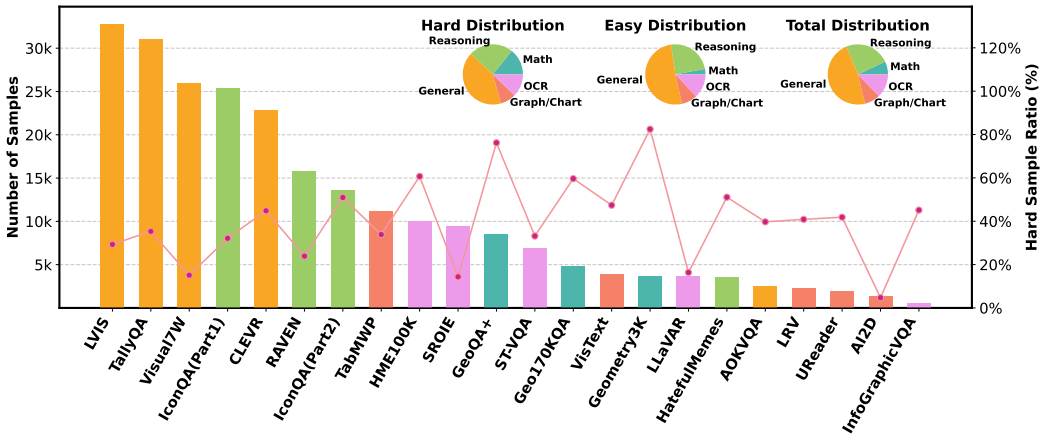


Figure 7: **Statistics of the contrastive VQA pair curation.** The bar chart (left y-axis) displays the total number of contrastive VQA pairs found in each dataset, with colors indicating the data category. The line graph (right y-axis) plots the ratio of hard samples identified within those pairs for each dataset. In the upper right, the pie charts provide a categorical breakdown of the sample distribution.

The construction of our VisCoR-55K dataset is a multi-stage process involving efficient pair curation, difficulty-based filtering, and quality-controlled rationale generation. The entire pipeline is designed to produce a high-quality, challenging visual reasoning dataset. Our curation process for contrastive VQA pairs begins with a dataset-by-dataset, divide-and-conquer strategy. To maintain computational tractability and avoid a costly $O(n^2)$ search complexity across the entire data pool, we implement a greedy, first-match-exit search algorithm: for each sample within a given source dataset, the search for a contrastive VQA counterpart terminates as soon as the first valid match is identified. This efficient approach allows us to scale the curation process effectively. Following this procedure, we initially curated a large pool of 240k raw contrastive VQA pairs. The distribution is visualized by the bar chart in Fig. 7, with the left y-axis indicating the number of samples.

This initial pool of 240k pairs then undergoes a rigorous filtering and refinement pipeline. First, we apply the difficulty-based sampling strategy (as detailed in Sec. 3.1) to select only the hard samples, which are most effective for enhancing the model's reasoning capabilities. The proportion of hard samples varies significantly across datasets, and is illustrated by the line graph in Fig. 7 (plotted against the right y-axis). This critical filtering step narrows our collection down to 86k challenging contrastive pairs. Subsequently, we leverage the contrasting and rethinking pipeline to generate a high-quality rationale for each of these 86k samples. As a final quality control measure, we employ a text-matching-based post-processing step to automatically filter out any rationales containing unexpected or erroneous reasoning patterns. This process culminates in our final VisCoR-55K dataset, a collection of high-fidelity visual reasoning samples ready for finetuning. The pie charts in Fig. 7 provide a categorical overview of the data composition throughout this pipeline.

A.2 Prompts for Thinking, Contrasting, and Rethinking

As introduced in Sec. 3.2, 3 steps lead to the final rationales. We design 3 prompts for the thinking, contrasting, and rethinking steps. The thinking prompt is:

> You are a helpful assistant to answer the question by thinking step by step.
> ### INPUT ###
> - Image: The image that serves as the basis for answering the question.
> - Question: The question pertains to the content of the image.
> - Answer: The correct answer for the question about the image.
> ### INSTRUCTION ###
> - You should analyze the question and decide to focus on which visual content.
> - You should parse the details of visual content based on the question.
> - You should conclude the visual evidence to answer the question.
> ### OUTPUT ###
> - The returned content MUST be in the natural flow.
> `<Image><Question><Answer>`

The contrasting prompt is:

> You are a helpful assistant to think step by step for discriminating between two images to answer two synonymous questions.
> ### INPUT ###
> - First Image: One image that serves as the basis for answering the question.
> - Second Image: The other image that serves as the basis for answering the question.
> - First Question: The question pertains to the content of the First Image.
> - Second Question: The question pertains to the content of the Second Image.
> - Answer: The correct answer for the question about the images.
> ### INSTRUCTION ###
> - When the correct answers for the two images are the same, you should summarize the common patterns in the visual content of the two images.
> - When the correct answers for the two images are different, you should identify the differences in visual content between two images.
> - Conclude the visual evidence to answer the questions respectively.
> ### OUTPUT ###
> - Return in the natural flow.
> `<FirstImage><SecondImage>`
> `<FirstQuestion><SecondQuestion><Answer>`

The "Answer" here is the concatenation of both samples. The rethinking prompt is:

You are a helpful assistant to rewrite the coarse rationale into a more correct and more logical one based on a contrastive analysis.
INPUT
- Question: The question to be answered based one given target image.
- Answer: The correct answer to answer the question.
- Coarse Rationale: The naive reasoning process answering the question.
- Contrastive Analysis: The reasoning process when comparing the first image with the second image for synonymous questions.
INSTRUCTION
- The contrastive analysis is more reliable than the coarse rationale.
- If the answers in the contrastive analysis are the same for the two images, the model should formulate a summary reasoning schema. This schema must summarize the key visual features and confirm that the provided visual evidence aligns with this schema to derive the conclusion.
- If the answers in the contrastive analysis are different for the two images, you can employ backward chaining hypothesizing the visual cues that would be present if the alternative answer were correct, and then highlighting the critical distinctions between this hypothetical scenario and the actual visual evidence.
OUTPUT
- The output MUST be in the format of '<think>the thinking content</think><answer>the answering content</answer>'.
- The content of thinking content MUST be between the special token of '<think>' and '</think>'
- The content of answering content MUST be between the special token of '<answer>' and '</answer>'.
<Question><Answer><CoarseRationale><ContrastiveAnalysis>

A.3 Additional Qualitative Results

Examples of rationales generated by VC-STaR in VisCoR-55K are illustrated in Fig. 8. After obtaining the model finetuned with VisCoR-55K on Qwen2.5VL-7B, we test it on some customized visual question answering cases and observe the interesting results shown in Fig. 9.

A.4 Statement on the Use of Large Models

During the preparation of this paper, we utilized the large language model to assist with grammar checking and language polishing. The outputs of the model were used solely as a reference to improve the clarity and readability of our writing. All final content presented in this paper was written and verified by the human authors, who take full responsibility for its accuracy and integrity. The core intellectual contributions of this work, *e.g.*, initial motivation, experimental design, and execution of the experiments, were developed exclusively through discussions among the authors, with no involvement from any large language model.

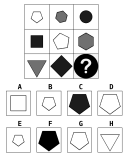
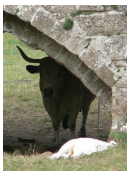
Question	Rationales
 Choose the figure that would logically complete the sequence.	<think>To determine the figure that logically completes the sequence, let's carefully analyze the target image: 1. **Shape Analysis**: - The shapes in the sequence are pentagons, and they remain consistent in form across all positions. 2. **Size Progression**: - Observing the figures, there is a clear trend of increasing size as the sequence progresses from left to right and top to bottom. Each subsequent shape grows larger than the previous one. 3. **Color/Pattern Observation**: - There is no apparent change in color or shading patterns influencing the placement of the figures. The consistency in color suggests that size is the primary factor driving the sequence. 4. **Conclusion Based on Size Increase**: - Given the established pattern of gradual size expansion, the next figure in the sequence must continue this trend by being larger than the preceding shapes while maintaining the same form. The details in the image confirm that the progression relies solely on incremental growth in size without introducing additional complexities such as color changes or new shapes. Therefore, the figure that aligns perfectly with this principle is option D.</think><answer>D</answer>
 How many buses are there?	<think>The image shows two red London United buses parked side by side, with their front ends visible. Therefore, based on this observation: There are **two** buses shown in the picture.</think><answer>2.</answer>
 How many horns are visible?	<think>The image shows an animal positioned in such a way that only one horn is clearly visible, as the rest of its head is partially obscured by an archway. This indicates that while there may be another horn present, it is not visible from this angle or due to the obstruction. The visual evidence supports the conclusion that only one horn is identifiable in the given view.</think><answer>1</answer>
 In the given diagram, triangle ABC is inscribed within circle O, and D is the midpoint of side BC. If OD is extended to intersect circle O at point E, and EC is then connected, what is the measure of angle A if angle OEC is 65°? Choices: A: 50° B: 55° C: 60° D: 65°	<think>To determine the measure of angle A, we start by analyzing the given information and leveraging geometric properties. Triangle ABC is inscribed in circle O, with D as the midpoint of side BC. Since D is the midpoint, OD is perpendicular to BC, and OD extended intersects the circle at point E. Connecting EC, we are given that angle OEC is 65°. From the property of isosceles triangles, since OE and OC are radii of the circle, triangle OEC is isosceles, and thus angles OEC and OCE are equal. Therefore, angle EOC can be calculated as: \[\text{Angle EOC} = 180^\circ - 2 \times 65^\circ = 50^\circ. \] Next, because D is the midpoint of BC and OD is perpendicular to BC, it follows that CE equals BE. This symmetry implies that angle EOB is also 50°. Consequently, the central angle BOC, which subtends arc BC, is twice angle EOB: \[\text{Angle BOC} = 2 \times 50^\circ = 100^\circ. \] Finally, angle A, being an inscribed angle subtending the same arc BC, is half of the central angle BOC: \[\text{Angle A} = \frac{1}{2} \times 100^\circ = 50^\circ. \] This reasoning aligns with the specific details provided in the problem and confirms the measure of angle A.</think> <answer>A: 50°</answer>

Figure 8: **Examples of rationales in VisCoR-55K.**

Question	Qwen2.5VL-7B	Qwen2.5VL-7B "think step by step"	Qwen2.5VL-7B+VC-STaR
 What makes the people so shocked?	The image shows a group of people in what appears to be a sports stadium, reacting with shock and surprise. The person in the center is holding a baseball bat, which could suggest that something unexpected or dramatic has occurred on the field, such as a foul ball hitting someone or a player making an extraordinary play. The expressions and body language of the spectators indicate a sudden and intense reaction to an event that has just taken place. Without additional context, it's not possible to determine the exact cause of their shock, but it is clear that whatever happened was surprising and possibly dangerous.	To determine what makes the people in the image so shocked, let's analyze the context and details: 1. Setting: The image appears to be taken at a baseball game, as indicated by the presence of a baseball bat and the crowd's attire, which includes team colors and logos. 2. Action: One person is holding a baseball bat, which suggests that something unexpected or dramatic has occurred on the field. This could be a play that resulted in a controversial or surprising outcome, such as a home run, a foul ball, or an unusual play. 3. Reactions: The people in the image are displaying strong reactions—some are covering their faces with their hands, while others are looking towards the field with expressions of shock or surprise. This indicates that whatever happened was unexpected and likely had a significant impact on the game. 4. Contextual Clues: The fact that the bat is being held up suggests that it might have been used in a way that surprised the spectators, possibly due to a foul ball or a hit that went into the stands, leading to a potential injury or disruption. Given these observations, the people are likely shocked because of a surprising event on the field, such as a foul ball hitting someone in the stands, a controversial play, or an unexpected turn of events during the game. The exact cause would depend on the specific circumstances of the game, but the general reaction points to a moment of high drama or surprise.	<think>The image depicts a scene at what appears to be a baseball game, with spectators showing expressions of surprise or shock. The key detail is the presence of a bat in mid-air, which suggests an unexpected event such as a foul ball or a player's swing. This aligns with the idea that something sudden and surprising has occurred during the game, causing the audience's reaction. While other scenarios might involve different causes for shock—such as accidents or unrelated incidents—the visual evidence here points directly to the bat as the source of their astonishment.</think><answer>The people are shocked because a bat has been thrown into the crowd, likely during a moment of excitement or surprise at the baseball game.</answer>
 What color is the car on the parking spot 33?	The car parked in the parking spot labeled "33" is blue.	The car on the parking spot labeled "33" is blue.	<think>The question asks about the color of the car in parking spot 33. The analysis shows that there is no car occupying parking spot 33 in the image, as it appears empty. Since there is no vehicle present in this specific spot, it is impossible to determine any color associated with it. This conclusion aligns with the observation that the absence of a car directly answers the question.</think> <answer> There is no car parked in parking spot 33, so I cannot determine its color.</answer>
 What answer will the girl with a single braid write on the board?	The girl with a single braid is working on the problem "11 − 2". The answer to this problem is 9.	To determine what answer the girl with a single braid will written on the board, we need to solve the mathematical expression she is working on: The expression on the board for the girl with a single braid is : 7×2. Let's break it down step by step: 1. Identify the operation: the operation here is multiplication. 2. Multiply the numbers: 7 multiplied by 2 is 14. So, the girl with a single braid will write the number 14 on the board.	<think>To answer the question "What answer will the girl with a single braid write on the board?", let's break down the analysis step by step: 1. Identify the Girl with Single Braid: - The girl with a single braid is located on the far left side of the image. 2. Position on the Board: - She is positioned directly below the equation (3 \times 3 =). 3. Action with Chalk: - The girl is holding a piece of chalk, indicating that she is ready to write. 4. Conclusion from Visual Evidence: The equation (3 (times 3) equals 9. The girl is positioned directly below this equation, suggesting she will write the result. Based on these observations, it is clear that the girl with a single braid will write the number 9 on the board as the solution to the equation (3 \times 3 =). Therefore, the answer is: The girl with a single braid will write the number 9 on the board.</think><answer> The girl with a single braid will write the number 9 on the board.</answer>

Figure 9: **Additional qualitative comparison.**