Novelty Assessment Report

Paper: Through the Lens of Contrast: Self-Improving Visual Reasoning in VLMs
PDF URL: https://openreview.net/pdf?id=ZymCPON45y
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

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 fine-tuning. 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 Visual Contrastive Self-Taught Reasoner (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.

Core Task Landscape

This paper addresses: **Self-Improving Visual Reasoning in Vision Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Self-Improvement via Iterative Refinement and Feedback**
- **Self-Improvement via External Supervision or Synthetic Data Generation**
- **Modality Alignment and Perception Enhancement**
- **Task-Specific Self-Improving Applications**
- **Evaluation, Benchmarking, and Auxiliary Techniques**

Complete Taxonomy Tree

- Self-Improving Visual Reasoning in Vision Language Models Survey Taxonomy
- Self-Improvement via Iterative Refinement and Feedback
  - Self-Correction and Reflection Mechanisms (5 papers)
  - [2] Sherlock: Self-Correcting Reasoning in Vision-Language Models (Ding Yi, 2025) View paper
  - [30] REVISOR: Beyond Textual Reflection, Towards Multimodal Introspective Reasoning in Long-Form Video Understanding (Jiaze Li, 2025) View paper
  - [31] Self-correction is more than refinement: A learning framework for visual and language reasoning tasks (He, 2025) View paper
  - [33] Vision-Language Models Can Self-Improve Reasoning via Reflection (Xu, 2024) View paper
  - [38] Toward More Reliable Artificial Intelligence: Reducing Hallucinations in Vision-Language Models (Kassoum Sanogo, 2025) View paper
  - Actor-Critic and Multi-Agent Frameworks (4 papers)
  - [20] Enhancing safety of vision-language reasoning through model-to-model deliberation (Sungwoo Kim, 2025) View paper
  - [25] Critic-v: Vlm critics help catch vlm errors in multimodal reasoning (Di Zhang, 2025) View paper
  - [35] Vision-Language Model Dialog Games for Self-Improvement (Konyushkova, 2025) View paper
  - [36] Mmc: Iterative refinement of vlm reasoning via mcts-based multimodal critique (Liu, 2025) View paper
  - Iterative Reasoning and Chain-of-Thought Enhancement (3 papers)
  - [4] Measuring and Improving Chain-of-Thought Reasoning in Vision-Language Models (Chen Yangyi, 2023) View paper
  - [6] Insight-v: Exploring long-chain visual reasoning with multimodal large language models (Dong Yu-hao, 2025) View paper
  - [13] Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles (Deng Yihe, 2025) View paper
- Self-Improvement via External Supervision or Synthetic Data Generation
  - Reward-Based and Preference Optimization (5 papers)
  - [1] Calibrated self-rewarding vision language models (Zhaorun Chen, 2024) View paper
  - [15] Iterative Tool Usage Exploration for Multimodal Agents via Step-wise Preference Tuning (Li Pengxiang, 2025) View paper
  - [17] VisPlay: Self-Evolving Vision-Language Models from Images (Yicheng He, 2025) View paper
  - [43] Self-Improving VLM Judges Without Human Annotations (Inna Wanyin Lin, 2025) View paper
  - [44] Videosavi: Self-aligned video language models without human supervision (Kulkarni, 2024) View paper
  - Synthetic Data Generation and Self-Training (4 papers)
  - [21] Enhancing Large Vision Language Models with Self-Training on Image Comprehension (Kai-Wei Chang, 2024) View paper

    - [23] Self-Training Large Language Models for Improved Visual Program Synthesis With Visual Reinforcement (Zaid Khan, 2024) View paper
    - [24] C2-Evo: Co-Evolving Multimodal Data and Model for Self-Improving Reasoning (Chen Xiu-wei, 2025) View paper
    - [48] VQA Training Sets are Self-play Environments for Generating Few-shot Pools (Mansoor Hassan, 2024) View paper
    - Knowledge Distillation and Model Compression (2 papers)
    - [10] Self-improving teacher cultivates better student: Distillation calibration for multimodal large language models (Xinwei Li, 2024) View paper
    - [29] SDRT: Enhance Vision-Language Models by Self-Distillation with Diverse Reasoning Traces (Wu, 2025) View paper
- Modality Alignment and Perception Enhancement
    - Visual-Language Modality Alignment (2 papers)
    - [11] Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement (Xiyao Wang, 2024) View paper
    - [18] Improving Generalization in Visual Reasoning via Self-Ensemble (Tien-Huy Nguyen, 2024) View paper
    - Active Perception and Visual Grounding (4 papers)
    - [3] Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing (Wu, 2025) View paper
    - [14] Viper: Empowering the self-evolution of visual perception abilities in vision-language model (Zhang Jun-tian, 2025) View paper
    - [22] Learning Active Perception via Self-Evolving Preference Optimization for GUI Grounding (Wang Wan-fu, 2025) View paper
    - [41] Cropper: Vision-Language Model for Image Cropping through In-Context Learning (Seung Hyun Lee, 2024) View paper
- Task-Specific Self-Improving Applications
    - Embodied Navigation and Tracking (6 papers)
    - [5] EvolveNav: Self-Improving Embodied Reasoning for LLM-Based Vision-Language Navigation (Lin, 2025) View paper
    - [7] TRACE: A Self-Improving Framework for Robot Behavior Forecasting with Vision-Language Models (Padrao, 2025) View paper
    - [8] Se-vln: A self-evolving vision-language navigation framework based on multimodal large language models (Dong Xiang-yu, 2025) View paper
    - [9] VLM Can Be a Good Assistant: Enhancing Embodied Visual Tracking with Self-Improving Vision-Language Models (Wu Kui, 2025) View paper
    - [12] VLM Can Be a Good Assistant: Enhancing Embodied Visual Tracking with Self-Improving Visual-Language Models (K Wu, 2025) View paper
    - [19] Iterative Vision-and-Language Navigation (Jacob Krantz, 2023) View paper
    - Visual Program Synthesis and Debugging (3 papers)
    - [37] Exovip: Step-by-step verification and exploration with exoskeleton modules for compositional visual reasoning (Wang Yu-xuan, 2024) View paper
    - [46] De-fine: De composing and re fin ing visual programs with auto-feedback (Minghe Gao, 2024) View paper
    - [47] Vdebugger: Harnessing execution feedback for debugging visual programs (Chang, 2024) View paper
    - Domain-Specific Reasoning Applications (4 papers)
    - [39] SERPENT-VLM : Self-Refining Radiology Report Generation Using Vision Language Models (Goyal, 2024) View paper
    - [42] Vurf: A general-purpose reasoning and self-refinement framework for video understanding (Mahmood Ahmad, 2024) View paper
    - [45] Video Visual Relation Detection via Iterative Inference (Xindi Shang, 2021) View paper
    - [50] Think Visually, Reason Textually: Vision-Language Synergy in ARC (Beichen Zhang, 2025) View paper
- Evaluation, Benchmarking, and Auxiliary Techniques
    - Prompt Optimization and In-Context Learning (2 papers)
    - [16] Advancing Multimodal In-Context Learning in Large Vision-Language Models with Task-aware Demonstrations (Yanshu, 2025) View paper
    - [34] Evolutionary Prompt Optimization Discovers Emergent Multimodal Reasoning Strategies in Vision-Language Models (Brown, 2025) View paper
    - Reasoning Evaluation and Verification (2 papers)
    - [27] Generative universal verifier as multimodal meta-reasoner (Zhang, 2025) View paper
    - [49] Probing Visual Language Priors in VLMs (Luo Tiange, 2025) View paper
    - Architectural and Efficiency Enhancements (4 papers)
    - [26] ARDN: Attention re-distribution network for visual question answering (Jinyang Yi, 2025) View paper
    - [28] Self-Bootstrapped Visual-Language Model for Knowledge Selection and Question Answering (Guo, 2024) View paper
    - [32] Enhancing advanced visual reasoning ability of large language models (Cai, 2024) View paper
    - [40] Perceiver-vl: Efficient vision-and-language modeling with iterative latent attention (Zineng Tang, 2023) View paper
    - Contrastive Learning and Visual Contrast ★ (1 papers)
    - [0] Through the Lens of Contrast: Self-Improving Visual Reasoning in VLMs (Anon et al., 2026) View paper

Narrative

Core task: self-improving visual reasoning in vision language models. The field organizes around several complementary strategies for enhancing VLM performance without extensive human annotation. One major branch focuses on iterative refinement and feedback mechanisms, where models learn to critique and revise their own outputs through self-play or internal verification loops (e.g., Calibrated Self-Rewarding[1], Self-Improving Teacher[10]). A second branch emphasizes external supervision or synthetic data generation, leveraging large-scale automated pipelines to produce training signals that guide model improvement (e.g., Self-Bootstrapped Knowledge[28], Self-Training Comprehension[21]). Modality alignment and perception enhancement address the core challenge of bridging vision and language representations, often through contrastive objectives or architectural innovations (e.g., Modality Alignment Enhancement[11], Perceiver-vl[40]). Task-specific applications demonstrate these principles in domains such as navigation, GUI interaction, and video understanding (e.g., EvolveNav[5], Active Perception GUI[22]). Finally, evaluation and auxiliary techniques provide the infrastructure for measuring progress and supporting self-improvement, including benchmarking frameworks, contrastive learning methods, and tool-augmented reasoning (e.g., Measuring Chain-of-Thought[4], Viper[14]).

Within the evaluation and auxiliary techniques branch, contrastive learning and visual contrast methods have emerged as a small but important cluster. These approaches use contrastive objectives to sharpen visual discrimination and improve reasoning by highlighting differences between similar inputs or outputs. Contrast Lens[0] exemplifies this direction by introducing mechanisms that explicitly leverage visual contrasts to enhance interpretability and reasoning quality. This work shares thematic connections with broader evaluation efforts like Measuring Chain-of-Thought[4], which probes reasoning transparency, and with perception-focused methods such as Cropper[41], which refines visual attention. Compared to iterative refinement approaches like Calibrated Self-Rewarding[1] or task-

specific systems like Spatial Reasoning Drawing[3], Contrast Lens[0] emphasizes diagnostic and interpretive tools rather than end-to-end training loops, positioning itself as a complementary technique for understanding and improving how VLMs process visual information.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses specifically on contrastive learning methods that use visual contrast or contrastive pairs to improve visual reasoning and reduce hallucinations. The sibling subtopics cover complementary aspects of VLM self-improvement: architectural/efficiency innovations, prompt engineering without retraining, and evaluation/verification frameworks. Together, these represent different intervention points in the VLM reasoning pipeline—training methodology (original), model architecture, inference-time optimization, and quality assessment.

Similarities: - All subtopics aim to enhance visual reasoning capabilities in VLMs - All address the challenge of improving VLM performance through different mechanisms - All are concerned with the quality and reliability of VLM outputs - All exclude overlapping methods through their exclude_note boundaries (e.g., training vs. inference-time methods)

Differences: - Original leaf uses contrastive learning during training, while Prompt Optimization operates at inference time without model updates - Original leaf focuses on a specific training signal (visual contrast), while Architectural Enhancements addresses model design and computational efficiency - Reasoning Evaluation focuses on assessment and verification rather than improvement methods - Original leaf explicitly targets hallucination mitigation through contrastive pairs, while siblings address reasoning quality through different mechanisms (architecture, prompting, verification) - Original leaf requires model retraining with contrastive objectives, while Prompt Optimization explicitly excludes parameter updates

Suggested Search Directions: - Investigate whether contrastive learning methods can be combined with architectural innovations (e.g., contrastive attention mechanisms) - Explore how verification frameworks might provide feedback signals for contrastive learning objectives - Examine whether contrastive pairs can be generated or selected through prompt optimization techniques

Sibling Subtopics

- **Architectural and Efficiency Enhancements** (leaves: 1, papers: 4)
- Scope: Architectural innovations or efficiency improvements for VLM reasoning including attention mechanisms, retrieval augmentation, or computational optimization.
- Exclude: Excludes self-improvement training procedures; see Self-Improvement via External Supervision or Iterative Refinement.
- **Prompt Optimization and In-Context Learning** (leaves: 1, papers: 2)
- Scope: Techniques for optimizing prompts or demonstration selection to elicit better reasoning from VLMs without model retraining.
- Exclude: Excludes methods requiring model fine-tuning or parameter updates; see Self-Improvement via External Supervision.
- **Reasoning Evaluation and Verification** (leaves: 1, papers: 2)
- Scope: Frameworks for assessing VLM reasoning quality, detecting hallucinations, or verifying visual outcomes through meta-reasoning or universal verifiers.
- Exclude: Excludes training methods for improving reasoning; see Self-Improvement via Iterative Refinement and Feedback.

Contributions Analysis

Overall novelty summary. The paper introduces VC-STaR, a self-improving framework that uses visual contrast to reduce hallucinations in VLM-generated reasoning paths, and produces VisCoR-55K, a visual reasoning dataset for fine-tuning. According to the taxonomy, this work resides in the 'Contrastive Learning and Visual Contrast' leaf under 'Evaluation, Benchmarking, and Auxiliary Techniques'. Notably, this leaf contains only one paper (the original work itself), indicating a sparse research direction within the broader self-improving VLM landscape, which encompasses 50 papers across approximately 36 topics.

The taxonomy reveals that neighboring leaves focus on prompt optimization, reasoning evaluation, and architectural enhancements, while sibling branches address iterative refinement (e.g., self-correction mechanisms with 5 papers) and synthetic data generation (4 papers). The scope note for this leaf emphasizes 'leveraging visual contrast or contrastive pairs to enhance visual reasoning and mitigate hallucinations', explicitly excluding non-contrastive self-improvement methods. The taxonomy narrative mentions Contrast Lens as an exemplar, positioning contrastive approaches as diagnostic and interpretive tools complementary to end-to-end training loops found in denser branches like actor-critic frameworks or reward-based optimization.

Among 30 candidates examined, the VC-STaR framework and contrastive pair curation framework each showed no clear refutations across 10 candidates, suggesting these contributions occupy relatively unexplored methodological territory. However, the VisCoR-55K dataset contribution encountered 1 refutable candidate among 10 examined, indicating some overlap with existing visual reasoning datasets. The limited search scope (30 candidates total, not exhaustive) means these statistics reflect top-K semantic matches and citation expansion rather than comprehensive field coverage. The framework contributions appear more distinctive than the dataset contribution within this bounded search.

Given the sparse taxonomy leaf (1 paper) and the absence of sibling papers, the contrastive self-improvement angle appears underexplored relative to denser branches like self-correction (5 papers) or reward-based optimization (5 papers). The analysis covers top-30 semantic matches, so conclusions about novelty are provisional. The framework's emphasis on visual contrast as a hallucination mitigation strategy distinguishes it from iterative refinement or synthetic data generation approaches, though the dataset contribution shows more overlap with prior work within the examined scope.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Visual Contrastive Self-Taught Reasoner (VC-STaR) framework

Description: The authors introduce VC-STaR, a self-improving framework that uses contrastive VQA pairs (two visually similar images with synonymous questions) to help VLMs identify relevant visual cues more precisely and rectify visual hallucinations in reasoning paths. The framework includes three steps: generating a coarse rationale, performing contrastive analysis, and rethinking to refine the rationale.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Retrieve-then-compare mitigates visual hallucination in multi-modal large language models

URL: View paper

Brief Assessment

Retrieve-then-Compare[71] focuses on mitigating visual hallucinations through retrieval-based contrastive decoding at inference time, while VC-STaR is a self-improving training framework that generates reasoning datasets using contrastive VQA pairs for supervised finetuning. The candidate addresses hallucination mitigation through decoding strategies, not self-improving reasoning path generation.

---

2. Hallucination Augmented Contrastive Learning for Multimodal Large Language Model

URL: View paper

Brief Assessment

Hallucination Augmented[70] focuses on using contrastive learning with hallucinative text as hard negatives to align visual-textual representations and reduce hallucinations in MLLMs. The original paper's VC-STaR framework specifically addresses visual reasoning by using contrastive VQA pairs (two visually similar images with synonymous questions) to help VLMs identify relevant visual cues and rectify hallucinations in reasoning paths through a three-step process (coarse rationale generation, contrastive analysis, rethinking). These are distinct approaches: Hallucination Augmented[70] uses artificially generated hallucinative captions for representation alignment during pretraining, while VC-STaR leverages naturally occurring contrastive image pairs for reasoning path refinement.

---

3. Delve into Visual Contrastive Decoding for Hallucination Mitigation of Large Vision-Language Models

URL: View paper

Brief Assessment

Delve Contrastive[74] focuses on contrastive decoding methods for hallucination mitigation during inference (training-free), while VC-STaR is a self-improving framework that generates training data using contrastive VQA pairs for supervised finetuning. These are fundamentally different approaches to leveraging visual contrast.

---

4. Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models

URL: View paper

Brief Assessment

Self-Introspective Decoding[69] focuses on mitigating hallucinations during decoding by contrasting original and disturbed model outputs, rather than using contrastive VQA pairs for self-improving reasoning path generation. The candidate addresses inference-time hallucination reduction, not training-time rationale refinement through visual contrast.

---

5. Contrastive Learning Reduces Hallucination in Conversations

URL: View paper

Brief Assessment

Contrastive Conversations[72] focuses on reducing hallucinations in text-based conversational systems through contrastive learning on dialogue data, not on visual reasoning tasks with vision-language models. The candidate addresses knowledge-grounded dialogues using language models, while the original contribution specifically targets visual question answering with contrastive image pairs to improve VLM reasoning.

---

6. See different, think better: Visual variations mitigating hallucinations in lvlms

URL: View paper

Brief Assessment

See Different[76] focuses on mitigating hallucinations through visual variation images (controlled alterations of original images) rather than contrastive VQA pairs. The candidate does not use contrastive pairs of visually similar images with synonymous questions for self-improving reasoning paths.

---

7. Mitigating object hallucinations in large vision-language models through visual contrastive decoding

URL: View paper

Brief Assessment

Visual Contrastive Decoding[73] focuses on mitigating object hallucinations in vision-language models through contrastive decoding at inference time, not on self-improving reasoning frameworks with contrastive VQA pairs for training data generation.

---

8. Reflective instruction tuning: Mitigating hallucinations in large vision-language models

URL: View paper

Brief Assessment

Reflective Instruction[68] focuses on mitigating hallucinations through rationale learning (positive and negative rationales for correct/incorrect responses) rather than contrastive visual analysis. The candidate does not employ contrastive VQA pairs or visual contrast mechanisms central to VC-STaR.

---

9. ConVis: Contrastive Decoding with Hallucination Visualization for Mitigating Hallucinations in Multimodal Large Language Models

URL: View paper

Brief Assessment

ConVis[75] focuses on mitigating hallucinations in MLLMs through contrastive decoding using text-to-image models to visualize hallucinations in captions, not on self-improving visual reasoning through contrastive VQA pairs as in VC-STaR.

---

10. HSCL-RL: Mitigating Hallucinations in Multimodal Large Language Models

URL: View paper

Brief Assessment

HSCL-RL[77] focuses on hallucination mitigation through contrastive learning with false text as hard negatives and reinforcement learning, not on self-improving visual reasoning through contrastive VQA pairs as in VC-STaR.

Contribution 2: Task-agnostic contrastive VQA pair curation framework

Description: The authors develop a flexible pipeline for curating contrastive VQA pairs across diverse VQA tasks including reasoning, math, chart, and OCR. The pipeline involves data collection from 21 datasets, similarity-based pair hunting using image and question embeddings, and difficulty-based sampling to select median-difficulty samples suitable for reasoning enhancement.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Contrastive video question answering via video graph transformer

URL: View paper

Brief Assessment

Contrastive Video Question[66] focuses on video question answering with graph transformers and contrastive learning between video-text pairs, but does not describe a systematic pipeline for curating contrastive VQA pairs across diverse task categories (reasoning, math, chart, OCR) as the original paper does.

---

2. Ciem: Contrastive instruction evaluation method for better instruction tuning

URL: View paper

Brief Assessment

CIEM[63] focuses on generating factual/contrastive question-answer pairs from caption annotations to evaluate hallucination in VLMs, not on curating contrastive VQA pairs across diverse reasoning tasks (math, chart, OCR) for reasoning enhancement as in the original paper.

---

3. A Logic-based Approach to Contrastive Explainability for Neurosymbolic Visual Question Answering

URL: View paper

Brief Assessment

Logic-based Contrastive[67] focuses on contrastive explanations for neurosymbolic VQA using logical abduction, not on curating contrastive VQA pairs across diverse tasks. The candidate's contrastive approach is for explainability purposes rather than data curation.

---

4. Simple contrastive learning in a self-supervised manner for robust visual question answering

URL: View paper

Brief Assessment

Simple Contrastive[60] focuses on addressing linguistic bias in VQA through contrastive learning with same questions but different images, while the original paper develops a multi-task framework spanning reasoning, math, chart, and OCR tasks with similarity-based pair hunting and difficulty-based sampling for reasoning enhancement.

---

5. Design as Desired: Utilizing Visual Question Answering for Multimodal Pre-training

URL: View paper

Brief Assessment

Design as Desired[64] focuses on medical domain VQA pair generation from clinical reports for pre-training, not a general framework for curating contrastive VQA pairs across diverse reasoning tasks (math, chart, OCR, etc.) as in the original paper.

---

6. Overcoming language priors with self-contrastive learning for visual question answering

URL: View paper

Brief Assessment

The candidate paper (Overcoming Language Priors[65]) focuses on addressing language priors in VQA through self-contrastive learning, not on curating contrastive VQA pairs across diverse tasks. No full text context was provided for the candidate paper to enable detailed comparison.

---

7. Surgical-VQLA++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery

URL: View paper

Brief Assessment

Surgical-VQLA[61] focuses on surgical video question-localized answering with adversarial contrastive learning for calibration in robotic surgery contexts. This is a domain-specific application fundamentally different from the original paper's task-agnostic framework for curating contrastive VQA pairs across diverse reasoning tasks (math, chart, OCR, general reasoning).

---

8. Counterfactual samples synthesizing and training for robust visual question answering

URL: View paper

Brief Assessment

Counterfactual Samples[59] focuses on synthesizing counterfactual samples by masking critical objects/words within individual samples to reduce language biases, not on curating contrastive pairs across diverse VQA task categories for reasoning enhancement.

---

9. Language-guided Bias Generation Contrastive Strategy for Visual Question Answering

URL: View paper

Brief Assessment

Language-guided Bias[58] focuses on debiasing VQA models through contrastive learning to combat language guidance bias, not on curating contrastive VQA pairs across diverse reasoning tasks. The candidate's contrastive approach addresses bias mitigation rather than pair curation for reasoning enhancement.

---

10. HCCL: Hierarchical Counterfactual Contrastive Learning for Robust Visual Question Answering

URL: View paper

Brief Assessment

HCCL[62] focuses on automatically masking features in original VQA pairs to create counterfactual samples for bias mitigation, rather than curating contrastive pairs across diverse VQA task categories (reasoning, math, chart, OCR) from multiple datasets as described in the original paper's contribution.

Contribution 3: VisCoR-55K visual reasoning dataset

Description: The authors create VisCoR-55K, a new dataset containing 55K high-quality visual reasoning samples with faithful rationales generated using VC-STaR. The dataset spans five categories (general VQA, reasoning, math, graph/chart, and OCR) and is used to improve VLM reasoning capabilities through supervised finetuning.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Vlr-driver: Large vision-language-reasoning models for embodied autonomous driving

URL: View paper

Brief Assessment

VLR-Driver[55] focuses on autonomous driving with a VLR-Driver dataset containing scene descriptions, reasoning processes, and driving decisions for embodied AD. This is fundamentally different from VisCoR-55K, which is a general visual reasoning dataset spanning VQA, reasoning, math, graph/chart, and OCR tasks for improving VLM reasoning capabilities.

---

2. Insight-v: Exploring long-chain visual reasoning with multimodal large language models

URL: View paper

Brief Assessment

Insight-v[6] focuses on a different data generation approach and dataset structure. While both papers create visual reasoning datasets, Insight-v[6] generates long-chain reasoning data through a progressive strategy with multi-granularity assessment, whereas the original paper creates contrastive VQA pairs. The datasets serve different purposes and use distinct methodologies for rationale generation.

---

3. Understand, Think, and Answer: Advancing Visual Reasoning with Large Multimodal Models

URL: View paper

Brief Assessment

Understand Think Answer[54] presents a 334k visual reasoning dataset with a different construction methodology (semi-automatic expert-supervised annotation) and focuses on a unified 'understand-think-answer' mechanism, rather than contrastive visual reasoning pairs with rationales as in VisCoR-55K.

---

4. Is a picture worth a thousand words? delving into spatial reasoning for vision language models

URL: View paper

Brief Assessment

Picture Worth Spatial[53] focuses on spatial reasoning evaluation benchmarks (SpatialEval with tasks like spatial-map, maze-nav, spatial-grid) rather than creating visual reasoning datasets with rationales for finetuning VLMs. The candidate's datasets are designed for evaluation purposes, not for generating training rationales through contrastive methods.

---

5. Learn to explain: Multimodal reasoning via thought chains for science question answering

URL: View paper

Brief Assessment

Learn to Explain[56] focuses on a different dataset (Science QA with 21,208 examples) for multimodal science question answering with lectures and explanations, not the visual contrastive reasoning approach used in VisCoR-55K.

---

6. Fine-tuning large vision-language models as decision-making agents via reinforcement learning

URL: View paper

Brief Assessment

Decision-Making Agents[51] focuses on reinforcement learning for training vision-language models as decision-making agents in interactive environments, not on creating visual reasoning datasets with rationales for supervised finetuning.

---

7. Vision-Language Models Can Self-Improve Reasoning via Reflection

URL: View paper

Brief Assessment

Self-Improve Reflection[33] focuses on a self-training framework (R3V) for iteratively improving vision-language reasoning through reflection on CoT rationales, not on creating a visual reasoning dataset with contrastive pairs. The candidate does not describe dataset construction methodology comparable to VisCoR-55K's contrastive VQA pair curation approach.

---

8. Llava-cot: Let vision language models reason step-by-step

URL: View paper

Prior Art Analysis

Llava-cot[52] demonstrates that a similar visual reasoning dataset with structured rationales was created prior to the original paper's VisCoR-55K. Specifically, Llava-cot[52] constructed the llava-cot-100k dataset containing 99k image QA pairs with detailed reasoning processes including summary, caption, reasoning, and conclusion stages. This dataset was generated using GPT-4o to produce systematic reasoning annotations and was used for supervised finetuning to improve VLM reasoning capabilities. The dataset spans multiple VQA sources including general VQA, science-targeted VQA, and reasoning tasks, similar to VisCoR-55K's coverage of general VQA, reasoning, math, graph/chart, and OCR categories. Both datasets serve the same purpose: providing high-quality visual reasoning samples with faithful rationales for finetuning vision-language models.

Evidence

Evidence 1 - **Rationale**: Both papers create visual reasoning datasets with structured rationales generated by GPT-4o for supervised finetuning of VLMs. Llava-cot[52]'s dataset (99k samples) predates and serves the same purpose as VisCoR-55K (55k samples). - **Original**: we obtain a new visual reasoning dataset, viscor-55k, which is then used to boost the reasoning capability of various vlms through supervised finetuning. - **Candidate**: we compile a new dataset, integrating samples from several widely used vqa datasets, resulting in a total of 99k image qa pairs (each pair may include one or multiple rounds of questioning). As shown in figure 3, since no multimodal model currently exists that can directly produce systematic, struct...

Evidence 2 - **Rationale**: Both datasets collect from diverse VQA sources spanning multiple categories. Llava-cot[52] includes general VQA, science-targeted VQA, and reasoning tasks, similar to VisCoR-55K's coverage of general VQA, reasoning, math, graph/chart, and OCR. - **Original**: 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 - **Candidate**: we include data from both general-purpose vqa datasets and science-targeted vqa datasets specified blow: general vqa datasets. we include several generalpurpose vqa datasets with distinct focuses. sharegpt4v [9] provides multi-turn question-answering data from gpt4v [59] interactions. chartqa [41] fo...

Evidence 3 - **Rationale**: Both datasets provide structured, multi-stage reasoning annotations. Llava-cot[52] explicitly describes their four-stage rationale structure (summary, caption, reasoning, conclusion) used for training, similar to the structured rationales in VisCoR-55K. - **Original**: rationales are shown in the sec. a.3. - **Candidate**: our annotated training data includes four distinct stages: summary, caption, reasoning, and conclusion, enabling the model to systematically address questions in a multi-stage manner. each stage serves a unique purpose in the reasoning process. • summary: a brief outline in which the model summarize...

---

9. Measuring and Improving Chain-of-Thought Reasoning in Vision-Language Models

URL: View paper

Brief Assessment

Measuring Chain-of-Thought[4] focuses on the CURE benchmark (1,622 samples) for evaluating reasoning consistency through chain-of-thought subquestions, not on creating a large-scale visual reasoning dataset with rationales for finetuning like VisCoR-55K (55K samples spanning five categories).

---

10. Empowering vision-language models for reasoning ability through large language models

URL: View paper

Brief Assessment

Empowering Reasoning[57] focuses on a learning-free framework (TREE) that transfers LLM reasoning to VLMs without training or finetuning, rather than creating a visual reasoning dataset with rationales for supervised finetuning like VisCoR-55K.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Through the Lens of Contrast: Self-Improving Visual Reasoning in VLMs View paper
- [1] Calibrated self-rewarding vision language models View paper
- [2] Sherlock: Self-Correcting Reasoning in Vision-Language Models View paper
- [3] Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing View paper
- [4] Measuring and Improving Chain-of-Thought Reasoning in Vision-Language Models View paper
- [5] EvolveNav: Self-Improving Embodied Reasoning for LLM-Based Vision-Language Navigation View paper
- [6] Insight-v: Exploring long-chain visual reasoning with multimodal large language models View paper
- [7] TRACE: A Self-Improving Framework for Robot Behavior Forecasting with Vision-Language Models View paper
- [8] Se-vln: A self-evolving vision-language navigation framework based on multimodal large language models View paper
- [9] VLM Can Be a Good Assistant: Enhancing Embodied Visual Tracking with Self-Improving Vision-Language Models View paper
- [10] Self-improving teacher cultivates better student: Distillation calibration for multimodal large language models View paper
- [11] Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement View paper
- [12] VLM Can Be a Good Assistant: Enhancing Embodied Visual Tracking with Self-Improving Visual-Language Models View paper
- [13] Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles View paper
- [14] Viper: Empowering the self-evolution of visual perception abilities in vision-language model View paper
- [15] Iterative Tool Usage Exploration for Multimodal Agents via Step-wise Preference Tuning View paper
- [16] Advancing Multimodal In-Context Learning in Large Vision-Language Models with Task-aware Demonstrations View paper
- [17] VisPlay: Self-Evolving Vision-Language Models from Images View paper
- [18] Improving Generalization in Visual Reasoning via Self-Ensemble View paper
- [19] Iterative Vision-and-Language Navigation View paper
- [20] Enhancing safety of vision-language reasoning through model-to-model deliberation View paper
- [21] Enhancing Large Vision Language Models with Self-Training on Image Comprehension View paper
- [22] Learning Active Perception via Self-Evolving Preference Optimization for GUI Grounding View paper
- [23] Self-Training Large Language Models for Improved Visual Program Synthesis With Visual Reinforcement View paper
- [24] C2-Evo: Co-Evolving Multimodal Data and Model for Self-Improving Reasoning View paper
- [25] Critic-v: Vlm critics help catch vlm errors in multimodal reasoning View paper
- [26] ARDN: Attention re-distribution network for visual question answering View paper
- [27] Generative universal verifier as multimodal meta-reasoner View paper
- [28] Self-Bootstrapped Visual-Language Model for Knowledge Selection and Question Answering View paper
- [29] SDRT: Enhance Vision-Language Models by Self-Distillation with Diverse Reasoning Traces View paper
- [30] REVISOR: Beyond Textual Reflection, Towards Multimodal Introspective Reasoning in Long-Form Video Understanding View paper
- [31] Self-correction is more than refinement: A learning framework for visual and language reasoning tasks View paper
- [32] Enhancing advanced visual reasoning ability of large language models View paper
- [33] Vision-Language Models Can Self-Improve Reasoning via Reflection View paper
- [34] Evolutionary Prompt Optimization Discovers Emergent Multimodal Reasoning Strategies in Vision-Language Models View paper
- [35] Vision-Language Model Dialog Games for Self-Improvement View paper
- [36] Mmc: Iterative refinement of vlm reasoning via mcts-based multimodal critique View paper
- [37] Exovip: Step-by-step verification and exploration with exoskeleton modules for compositional visual reasoning View paper
- [38] Toward More Reliable Artificial Intelligence: Reducing Hallucinations in Vision-Language Models View paper
- [39] SERPENT-VLM : Self-Refining Radiology Report Generation Using Vision Language Models View paper
- [40] Perceiver-vl: Efficient vision-and-language modeling with iterative latent attention View paper
- [41] Cropper: Vision-Language Model for Image Cropping through In-Context Learning View paper
- [42] Vurf: A general-purpose reasoning and self-refinement framework for video understanding View paper
- [43] Self-Improving VLM Judges Without Human Annotations View paper
- [44] Videosavi: Self-aligned video language models without human supervision View paper
- [45] Video Visual Relation Detection via Iterative Inference View paper
- [46] De-fine: De composing and re fin ing visual programs with auto-feedback View paper
- [47] Vdebugger: Harnessing execution feedback for debugging visual programs View paper
- [48] VQA Training Sets are Self-play Environments for Generating Few-shot Pools View paper
- [49] Probing Visual Language Priors in VLMs View paper
- [50] Think Visually, Reason Textually: Vision-Language Synergy in ARC View paper
- [51] Fine-tuning large vision-language models as decision-making agents via reinforcement learning View paper
- [52] Llava-cot: Let vision language models reason step-by-step View paper
- [53] Is a picture worth a thousand words? delving into spatial reasoning for vision language models View paper
- [54] Understand, Think, and Answer: Advancing Visual Reasoning with Large Multimodal Models View paper
- [55] Vlr-driver: Large vision-language-reasoning models for embodied autonomous driving View paper
- [56] Learn to explain: Multimodal reasoning via thought chains for science question answering View paper

- [57] Empowering vision-language models for reasoning ability through large language models View paper
- [58] Language-guided Bias Generation Contrastive Strategy for Visual Question Answering View paper
- [59] Counterfactual samples synthesizing and training for robust visual question answering View paper
- [60] Simple contrastive learning in a self-supervised manner for robust visual question answering View paper
- [61] Surgical-VQLA++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery View paper
- [62] HCCL: Hierarchical Counterfactual Contrastive Learning for Robust Visual Question Answering View paper
- [63] Ciem: Contrastive instruction evaluation method for better instruction tuning View paper
- [64] Design as Desired: Utilizing Visual Question Answering for Multimodal Pre-training View paper
- [65] Overcoming language priors with self-contrastive learning for visual question answering View paper
- [66] Contrastive video question answering via video graph transformer View paper
- [67] A Logic-based Approach to Contrastive Explainability for Neurosymbolic Visual Question Answering View paper
- [68] Reflective instruction tuning: Mitigating hallucinations in large vision-language models View paper
- [69] Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models View paper
- [70] Hallucination Augmented Contrastive Learning for Multimodal Large Language Model View paper
- [71] Retrieve-then-compare mitigates visual hallucination in multi-modal large language models View paper
- [72] Contrastive Learning Reduces Hallucination in Conversations View paper
- [73] Mitigating object hallucinations in large vision-language models through visual contrastive decoding View paper
- [74] Delve into Visual Contrastive Decoding for Hallucination Mitigation of Large Vision-Language Models View paper
- [75] ConVis: Contrastive Decoding with Hallucination Visualization for Mitigating Hallucinations in Multimodal Large Language Models View paper
- [76] See different, think better: Visual variations mitigating hallucinations in lvlms View paper
- [77] HSCL-RL: Mitigating Hallucinations in Multimodal Large Language Models View paper