Novelty Assessment Report

Paper: DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning
PDF URL: https://openreview.net/pdf?id=xUyMXkI958
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Large Vision-Language Models excel at multimodal understanding but struggle to deeply integrate visual information into their predominantly text-based reasoning processes, a key challenge in mirroring human cognition. To address this, we introduce DeepEyes, a model that learns to ``think with images'', trained end-to-end with reinforcement learning and without pre-collected reasoning data for supervised fine-tuning (SFT) as a cold-start. Notably, this ability emerges natively, leveraging the model's own grounding capability as an intrinsic function rather than relying on external specialized models or APIs. We enable this capability through active perception, where the model learns to strategically ground its reasoning in visual information, guided by a tailored data selection and reward strategy. DeepEyes achieves significant performance gains on general perception and reasoning benchmarks and also demonstrates improvement in grounding, hallucination, and mathematical reasoning tasks. Interestingly, we observe the distinct evolution of active perception from initial exploration to efficient and accurate exploitation, and diverse thinking patterns that closely mirror human visual reasoning processes. Code is available at \url{https://anonymous.4open.science/r/DeepEyes-97FE/}.

Core Task Landscape

This paper addresses: **integrating visual information into vision-language model reasoning processes**
A total of **50 papers** were analyzed and organized into a taxonomy with **29 categories**.

Taxonomy Overview

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

- **Structured Reasoning and Chain-of-Thought Approaches**
- **Spatial and Geometric Reasoning**
- **Visual Feature Representation and Integration**
- **Training Paradigms and Model Architectures**
- **Context and Demonstration Learning**
- **Domain-Specific Applications and Specialized Tasks**
- **Evaluation, Benchmarking, and Analysis**
- **Enhanced Capabilities and Auxiliary Mechanisms**
- **Cross-Domain and Multimodal Extensions**

Complete Taxonomy Tree

- integrating visual information into vision-language model reasoning processes Survey Taxonomy
- Structured Reasoning and Chain-of-Thought Approaches
  - Autonomous Multi-Stage Visual Reasoning ★ (2 papers)
  - [0] DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning (Anon et al., 2026) View paper
  - [1] LLaVA-CoT: Let Vision Language Models Reason Step-by-Step (Xu Guowei, 2024) View paper
  - Supervised Reasoning Decomposition with Visual Signals (2 papers)
  - [2] Self-rewarding vision-language model via reasoning decomposition (Li, 2025) View paper
  - [17] Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning (Tan Huajie, 2025) View paper
  - Visual Chain-of-Thought and Sketching (4 papers)
  - [16] Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models (Xingyu Fu, 2024) View paper
  - [27] Latent visual reasoning (Li, 2025) View paper
  - [28] CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models (Zhao Qingqing, 2025) View paper
  - [31] Simple o3: Towards Interleaved Vision-Language Reasoning (Wang Ye, 2025) View paper
  - Long-Chain Visual Reasoning (2 papers)
  - [35] Reasoning, scaling, generating with vision-language models (Wang, 2024) View paper
  - [43] Insight-v: Exploring long-chain visual reasoning with multimodal large language models (Dong Yu-hao, 2025) View paper
- Spatial and Geometric Reasoning
  - 3D Spatial Reasoning and Depth Integration (2 papers)
  - [4] SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities (Boyuan Chen, 2024) View paper
  - [40] SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model (Cheng, 2024) View paper
  - 2D Spatial Grounding and Region-Level Reasoning (2 papers)
  - [3] Spatialrgpt: Grounded spatial reasoning in vision-language models (An-Chieh Cheng, 2024) View paper
  - [24] Reinforcing Spatial Reasoning in Vision-Language Models with Interwoven Thinking and Visual Drawing (Wu, 2025) View paper
  - Geometric Problem Solving (1 papers)
  - [12] Integrating visual interpretation and linguistic reasoning for geometric problem solving (Z Guo, 2025) View paper

- Visual Feature Representation and Integration
  - Multi-Layer Visual Feature Fusion (2 papers)
  - [23] Exploring the Role of CLIP Global Visual Features in Multimodal Large Language Models (Zixing Bai, 2025) View paper
  - [44] Multi-Layer Visual Feature Fusion in Multimodal LLMs: Methods, Analysis, and Best Practices (Lin Junyan, 2025) View paper
  - Single-Layer Visual Encoding and Token Efficiency (3 papers)
  - [9] Matryoshka multimodal models (Cai, 2024) View paper
  - [41] Task-Oriented Feature Compression for Multimodal Understanding via Device-Edge Co-Inference (Yuan Cheng, 2025) View paper
  - [49] Evlm: An efficient vision-language model for visual understanding (Chen Kai-bing, 2024) View paper
  - Object-Centric and Region-Based Visual Representations (3 papers)
  - [8] GeoChat:Grounded Large Vision-Language Model for Remote Sensing (Kartik Kuckreja, 2023) View paper
  - [11] VinVL: Revisiting Visual Representations in Vision-Language Models (Pengchuan Zhang, 2021) View paper
  - [33] Incorporating Visual Experts to Resolve the Information Loss in Multimodal Large Language Models (Xin He, 2024) View paper
  - Visual Feature Space Analysis and Interpretation (2 papers)
  - [15] Exploring The Visual Feature Space for Multimodal Neural Decoding (Xia, 2025) View paper
  - [38] Towards Interpreting Visual Information Processing in Vision-Language Models (Clement Neo, 2024) View paper
- Training Paradigms and Model Architectures
  - Vision-Language Pre-Training Frameworks (2 papers)
  - [6] A survey of vision-language pre-trained models (Du, 2022) View paper
  - [13] BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (Li, 2022) View paper
  - Instruction Tuning and Alignment (3 papers)
  - [5] Instructblip: Towards general-purpose vision-language models with instruction tuning (Dai, 2023) View paper
  - [18] Otter: A multi-modal model with in-context instruction tuning (Bo Li, 2025) View paper
  - [21] Visual instruction tuning towards general-purpose multimodal large language model: A survey (Jiaxing Huang, 2025) View paper
  - Self-Improvement and Modality Alignment (1 papers)
  - [29] Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement (Xiyao Wang, 2024) View paper
  - Architectural Innovations for Multimodal Integration (2 papers)
  - [7] Incorporating Convolution Designs into Visual Transformers (Yuan Kun, 2021) View paper
  - [25] Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models (Kiros, 2022) View paper
- Context and Demonstration Learning
  - Visual In-Context Learning (1 papers)
  - [34] Visual In-Context Learning for Large Vision-Language Models (Zhou Yucheng, 2024) View paper
  - Long-Context Modeling (1 papers)
  - [10] InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output (Zhang Pan, 2024) View paper
- Domain-Specific Applications and Specialized Tasks
  - Embodied AI and Robotics (2 papers)
  - [14] Zero-shot Object Navigation with Vision-Language Models Reasoning (Congcong Wen, 2024) View paper
  - [19] PaLM-E: An Embodied Multimodal Language Model (Driess, 2023) View paper
  - Medical and Healthcare Applications (2 papers)
  - [22] Qilin-med-vl: Towards chinese large vision-language model for general healthcare (Liu Jun-ling, 2023) View paper
  - [46] Visionâ□□language foundation model for echocardiogram interpretation (M. Christensen, 2024) View paper
  - Segmentation and Grounding Tasks (1 papers)
  - [47] LISA: Reasoning Segmentation via Large Language Model (Xin Lai, 2023) View paper
  - Graph and Structural Reasoning (1 papers)
  - [48] GITA: Graph to Visual and Textual Integration for Vision-Language Graph Reasoning (Wei YanBin, 2024) View paper
  - Image Fusion and Synthesis (1 papers)
  - [37] Image Fusion via Vision-Language Model (Zhao Zixiang, 2024) View paper
- Evaluation, Benchmarking, and Analysis
  - Visual Reasoning Benchmarking (2 papers)
  - [42] Smart Vision-Language Reasoners (Roberts, 2024) View paper
  - [45] Zero-shot visual reasoning by vision-language models: Benchmarking and analysis (Jaiswal, 2024) View paper
  - Cognitive and Perceptual Analysis (2 papers)
  - [32] From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models (Zhou Chen-yue, 2025) View paper
  - [36] Visual cognition in multimodal large language models (Luca M. Schulze Buschoff, 2025) View paper
- Enhanced Capabilities and Auxiliary Mechanisms
  - High-Resolution and Multi-Scale Processing (1 papers)
  - [30] Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models (Li Yanwei, 2024) View paper
  - Image Tagging and Semantic Guidance (1 papers)
  - [39] Tag2text: Guiding vision-language model via image tagging (Huang XinYu, 2023) View paper
  - Collaborative and Distributed Inference (1 papers)
  - [50] Enhanced Reasoning via Multimodal LLMs and Collaborative Inference (Zhiquan Wen, 2025) View paper
- Cross-Domain and Multimodal Extensions
  - Audio-Visual Integration (1 papers)
  - [20] Integrating Audio-Visual Features For Multimodal Deepfake Detection (Sneha Muppalla, 2023) View paper
  - Neural Decoding and Brain-Vision-Language Integration (1 papers)
  - [26] Decoding visual neural representations by multimodal learning of brain-visual-linguistic features (Changde Du, 2023) View paper

Narrative

Core task: integrating visual information into vision-language model reasoning processes. The field has evolved into a rich landscape organized around several complementary directions. Structured Reasoning and Chain-of-Thought Approaches focus on making explicit the intermediate steps by which models combine visual and linguistic cues, often through multi-stage pipelines or self-reflective mechanisms. Spatial and Geometric Reasoning emphasizes understanding positional relationships and geometric properties, as seen in works like SpatialRGPT[3] and SpatialVLM[4]. Visual Feature Representation and Integration explores how to encode and fuse visual signals—ranging from early convolutional architectures to modern transformer-based embeddings—while Training Paradigms and Model Architectures address foundational questions of how to build and optimize these systems, exemplified by InstructBLIP[5] and related instruction-tuning methods. Meanwhile, Context and Demonstration Learning investigates few-shot and in-context strategies, Domain-Specific Applications target specialized tasks such as medical imaging or navigation, and Evaluation, Benchmarking, and Analysis provide the empirical grounding needed to compare approaches. Enhanced Capabilities and Auxiliary Mechanisms introduce tools like external memory or iterative refinement, and Cross-Domain and Multimodal Extensions push beyond vision-language pairs into audio-visual or embodied settings.

A particularly active line of work centers on autonomous multi-stage visual reasoning, where models iteratively refine their understanding by generating intermediate reasoning traces or self-critiques. DeepEyes[0] exemplifies this direction by orchestrating multiple reasoning steps that dynamically integrate visual evidence, closely aligning with LLaVA-CoT[1], which also structures chain-of-thought processes for vision-language tasks. These methods contrast with approaches that rely on fixed feature extractors or single-pass inference, trading computational cost for improved interpretability and accuracy on complex visual questions. Nearby efforts such as Self-rewarding VLM[2] explore self-improvement through reward-based learning, while works in spatial reasoning like SpatialRGPT[3] emphasize grounding in geometric relationships rather than general-purpose reasoning chains. DeepEyes[0] sits squarely within the autonomous multi-stage cluster, sharing with LLaVA-CoT[1] an emphasis on explicit intermediate steps, yet it distinguishes itself by deeper integration of visual cues at each reasoning stage, reflecting ongoing debates about how tightly vision and language should be coupled during inference.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step

Authors: Xu Guowei, Guowei Xu, Jin Peng, Peng Jin, Wu Ziang, et al. (15 authors total) | **Year/Venue**: 2024 • arXiv.org | **URL**: View paper

Abstract

Large language models have demonstrated substantial advancements in reasoning capabilities. However, current Vision-Language Models (VLMs) often struggle to perform systematic and structured reasoning, especially when handling complex visual question-answering tasks. In this work, we introduce LLaVA-CoT, a large VLM designed to conduct autonomous multistage reasoning. Unlike chain-of-thought prompting, LLaVA-CoT independently engages in sequential stages of summarization, visual interpretation, ...

Relationship Analysis

Both papers belong to the Autonomous Multi-Stage Visual Reasoning category, where models independently perform sequential reasoning stages without external prompting. They overlap in their approach to integrating visual information through multi-stage reasoning processes that combine visual perception with textual chain-of-thought. The key difference is that DeepEyes uses end-to-end reinforcement learning with active perception (dynamic image cropping/zooming) to interleave visual and textual reasoning, while LLaVA-CoT employs supervised fine-tuning on structured reasoning annotations with fixed sequential stages (summarization, visual interpretation, logical reasoning, conclusion) and introduces a test-time stage-wise retracing search method.

Contributions Analysis

Overall novelty summary. The paper introduces DeepEyes, a vision-language model trained end-to-end with reinforcement learning to perform multi-stage visual reasoning without supervised fine-tuning data. It resides in the 'Autonomous Multi-Stage Visual Reasoning' leaf of the taxonomy, which contains only two papers including the original work. This leaf sits within the broader 'Structured Reasoning and Chain-of-Thought Approaches' branch, indicating a relatively sparse but active research direction focused on models that independently decompose reasoning into sequential stages without external prompting or supervision.

The taxonomy reveals several neighboring research directions that contextualize DeepEyes. The sibling leaf 'Supervised Reasoning Decomposition with Visual Signals' explores similar multi-stage reasoning but relies on explicit supervision or reward mechanisms for intermediate steps. Nearby branches address 'Visual Chain-of-Thought and Sketching' (four papers generating visual artifacts as reasoning steps) and 'Long-Chain Visual Reasoning' (two papers handling extended reasoning sequences). The 'Training Paradigms and Model Architectures' branch, particularly 'Self-Improvement and Modality Alignment', shares conceptual overlap with DeepEyes' RL-based approach but focuses on self-generated data rather than active perception mechanisms.

Among 29 candidates examined across three contributions, the analysis reveals limited prior work overlap. The core 'end-to-end RL-based iMCoT' contribution examined 9 candidates with no clear refutations, suggesting novelty in the training paradigm. The 'active perception mechanism' contribution examined 10 candidates, also without refutation, indicating the native grounding capability may be distinctive. However, the 'data selection and reward strategy' contribution found 1 refutable candidate among 10 examined, suggesting some overlap with existing reward-shaping techniques. The limited search scope (29 papers, not exhaustive) means these findings reflect top-K semantic matches rather than comprehensive field coverage.

Given the sparse taxonomy leaf (two papers total) and limited refutations across most contributions, DeepEyes appears to occupy a relatively novel position within autonomous multi-stage visual reasoning. The single refutable candidate for reward strategy suggests incremental refinement in that component, while the core RL-based training and active perception mechanisms show stronger novelty signals. However, the analysis is constrained by examining only 29 candidates from semantic search, leaving open the possibility of relevant work outside this scope, particularly in adjacent RL-for-VLM or grounding literature.

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

Contribution 1: DeepEyes model with end-to-end RL-based iMCoT

Description: The authors propose DeepEyes, a vision-language model that learns to integrate visual information into reasoning through end-to-end reinforcement learning. This approach eliminates the need for supervised fine-tuning with pre-collected reasoning data and enables interleaved multimodal chain-of-thought (iMCoT) reasoning.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens

URL: View paper

Brief Assessment

Machine Mental Imagery[57] focuses on latent visual tokens for mental imagery without explicit image generation, while DeepEyes uses active perception with actual image cropping and grounding coordinates. The candidate's approach does not challenge DeepEyes' novelty in end-to-end RL for interleaved multimodal chain-of-thought with visual grounding.

---

2. UniVG-R1: Reasoning Guided Universal Visual Grounding with Reinforcement Learning

URL: View paper

Brief Assessment

UniVG-R1[58] focuses on visual grounding tasks with rule-based RL for reasoning chains, not on general vision-language reasoning with interleaved multimodal chain-of-thought. The candidate addresses grounding-specific challenges rather than the broader 'thinking with images' paradigm.

---

3. Bridging Formal Language with Chain-of-Thought Reasoning to Geometry Problem Solving

URL: View paper

Brief Assessment

Formal Language Geometry[59] focuses on geometry problem solving by integrating chain-of-thought with formal language and symbolic solvers, not on general vision-language reasoning with interleaved multimodal chain-of-thought. The candidate addresses a specialized domain (geometry) with formal program generation, while the original paper targets general visual reasoning across diverse tasks through active perception mechanisms.

---

4. Reasoning-VLA: A Fast and General Vision-Language-Action Reasoning Model for Autonomous Driving

URL: View paper

Brief Assessment

Reasoning-VLA[55] focuses on autonomous driving with vision-language-action models for trajectory generation, not on general vision-language reasoning with interleaved multimodal chain-of-thought. The technical domains and objectives are fundamentally different.

---

5. PeRL: Permutation-Enhanced Reinforcement Learning for Interleaved Vision-Language Reasoning

URL: View paper

Brief Assessment

PeRL[56] focuses on multi-image positional reasoning with permutation-based exploration, while DeepEyes addresses single-image active perception through zoom-in operations for fine-grained visual reasoning.

---

6. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution

URL: View paper

Brief Assessment

Deer-vla[54] focuses on dynamic inference efficiency for robot execution through early-exit mechanisms in multimodal LLMs, not on end-to-end reinforcement learning for vision-language reasoning with interleaved multimodal chain-of-thought.

---

7. VLM-R: Region Recognition, Reasoning, and Refinement for Enhanced Multimodal Chain-of-Thought

URL: View paper

Brief Assessment

VLM-R[53] focuses on region-conditioned reinforcement policy optimization (r-grpo) for visual region localization in multimodal reasoning, while DeepEyes emphasizes end-to-end RL training without cold-start supervised fine-tuning. VLM-R[53] requires supervised fine-tuning on the VLIR dataset before applying reinforcement learning, representing a fundamentally different training paradigm.

---

8. Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning

URL: View paper

Brief Assessment

Point-rft[52] focuses on visually grounded reinforcement finetuning for visual document understanding (charts, plots) using point-level annotations, not on general vision-language models with interleaved multimodal chain-of-thought reasoning across diverse visual tasks. The technical approaches differ fundamentally: Point-rft[52] uses explicit point coordinates for grounding in documents, while the original paper's DeepEyes uses active perception with bounding box cropping for general multimodal reasoning.

---

9. LLM-I: LLMs are Naturally Interleaved Multimodal Creators

URL: View paper

Brief Assessment

LLM-I[51] focuses on interleaved image-text generation using tool orchestration (search, diffusion, code execution, editing), while DeepEyes addresses vision-language reasoning with active visual perception through image cropping. The technical approaches and problem domains differ fundamentally.

Contribution 2: Active perception mechanism with native grounding capability

Description: The authors introduce an active perception mechanism that encapsulates the model's native visual grounding capability as an internal tool. This allows the model to strategically ground its reasoning in visual information without depending on external specialized models or APIs.

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. Geoground: A unified large vision-language model for remote sensing visual grounding

URL: View paper

Brief Assessment

GeoGround[78] focuses on remote sensing visual grounding tasks (horizontal/oriented bounding boxes and segmentation masks) using a text-mask paradigm. It does not address the active perception mechanism for strategic visual reasoning during chain-of-thought processes that the original paper proposes.

---

2. Visual In-Context Learning for Large Vision-Language Models

URL: View paper

Brief Assessment

Visual In-Context Learning[34] focuses on in-context learning through visual demonstration retrieval and image summarization, not on active perception mechanisms or native grounding capabilities for strategic visual reasoning during chain-of-thought processes.

---

3. Learning Visual Grounding from Generative Vision and Language Model

URL: View paper

Brief Assessment

Learning Visual Grounding[75] focuses on scaling visual grounding data by prompting generative VLMs to generate object descriptions, not on developing an active perception mechanism for interleaved reasoning. The candidate addresses data annotation for grounding tasks, while the original paper introduces a dynamic reasoning framework where the model strategically decides when to ground during chain-of-thought reasoning.

---

4. Robot navigation using physically grounded vision-language models in outdoor environments

URL: View paper

Brief Assessment

Robot Navigation Outdoor[74] focuses on outdoor robot navigation using VLMs for terrain traversability estimation, not on developing native visual grounding capabilities within VLMs themselves. The candidate paper uses VLMs as external tools for classification tasks rather than developing internal grounding mechanisms.

---

5. Contrastive region guidance: Improving grounding in vision-language models without training

URL: View paper

Brief Assessment

Contrastive Region Guidance[70] focuses on training-free visual grounding through contrastive decoding between masked and unmasked images, not on an active perception mechanism that strategically grounds reasoning through learned tool-calling behavior.

---

6. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks

URL: View paper

Brief Assessment

VLM2Vec[76] focuses on converting vision-language models into universal embedding models for downstream tasks like retrieval and classification, not on active perception or visual grounding as an internal reasoning tool during chain-of-thought processes.

---

7. ViGoR: Improving Visual Grounding of Large Vision Language Models with Fine-Grained Reward Modeling

URL: View paper

Brief Assessment

ViGoR[73] focuses on improving visual grounding through fine-grained reward modeling applied to pre-trained models, addressing hallucination and grounding errors. It does not propose an active perception mechanism where the model autonomously decides to crop and zoom into image regions during reasoning, which is the core novelty of the original paper's contribution.

---

8. Direct visual grounding by directing attention of visual tokens

URL: View paper

Brief Assessment

Direct Visual Grounding[77] focuses on directing attention of visual tokens through KL divergence loss for visual grounding tasks, not on active perception mechanisms that strategically crop and zoom into image regions during reasoning. The candidate does not demonstrate prior work on encapsulating native grounding as an internal tool for active perception.

---

9. Navgpt: Explicit reasoning in vision-and-language navigation with large language models

URL: View paper

Brief Assessment

NavGPT[72] uses external visual foundation models (BLIP-2, Fast-RCNN) to translate visual observations into natural language for LLM processing, rather than leveraging a model's native visual grounding capability as an internal tool without external specialized models.

---

10. Cogvlm: Visual expert for pretrained language models

URL: View paper

Brief Assessment

CogVLM[71] focuses on visual grounding as a task output (predicting bounding boxes in response to queries), not as an internal active perception mechanism for iterative reasoning. The candidate does not demonstrate prior work on models that strategically invoke their own grounding capability during chain-of-thought reasoning.

Contribution 3: Data selection and reward strategy for active perception

Description: The authors design a data selection mechanism to choose training samples that encourage active perception behavior, along with a conditional reward strategy that assigns bonuses to trajectories successfully completing tasks through active perception. These components are crucial for optimizing the efficiency and accuracy of the model's visual reasoning.

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. 3D-R1: Enhancing Reasoning in 3D VLMs for Unified Scene Understanding

URL: View paper

Brief Assessment

3D-R1[65] focuses on 3D scene understanding with synthetic dataset construction and reward functions for detection accuracy, not on data selection mechanisms for active perception behavior in 2D visual reasoning tasks.

---

2. Active Vision for Embodied Agents Using Reinforcement Learning

URL: View paper

Brief Assessment

Active Vision Embodied[69] focuses on general embodied agent tasks using standard RL reward shaping techniques, not specifically on data selection mechanisms or conditional reward strategies tailored for active perception in visual reasoning contexts as described in the original paper.

---

3. Adaptive important region selection with reinforced hierarchical search for dense object detection

URL: View paper

Brief Assessment

Adaptive Region Selection[67] focuses on hierarchical object detection in images using reinforcement learning for region selection, not on training data selection or reward strategies for active perception in visual reasoning tasks as described in the original paper.

---

4. Training Vision-Language Process Reward Models for Test-Time Scaling in Multimodal Reasoning: Key Insights and Lessons Learned

URL: View paper

Brief Assessment

Vision-Language Process Rewards[64] focuses on process reward models (PRMs) for step-level supervision using MCTS-based data construction and perception-focused supervision for visual grounding errors. The original paper's contribution centers on data selection mechanisms to encourage active perception behavior and conditional reward bonuses for trajectories completing tasks through active perception in an end-to-end RL framework. These are fundamentally different approaches: one uses PRMs for step-level error detection, the other uses data curation and conditional rewards to incentivize active visual exploration.

---

5. Learning Active Perception via Self-Evolving Preference Optimization for GUI Grounding

URL: View paper

Brief Assessment

Active Perception GUI[66] focuses on GUI grounding tasks with coordinate prediction using Monte Carlo quality estimation and IoU-based evaluation, not on general visual reasoning with data selection mechanisms for encouraging active perception behavior across diverse tasks.

---

6. Learn 3D VQA Better with Active Selection and Reannotation

URL: View paper

Brief Assessment

Learn 3D VQA[63] focuses on active learning for data selection in 3D visual question answering tasks, not on reward strategies for active perception in visual reasoning. The candidate addresses annotation quality and training efficiency in 3D VQA, while the original paper designs mechanisms for incentivizing active perception behavior during multimodal reasoning.

---

7. Visrl: Intention-driven visual perception via reinforced reasoning

URL: View paper

Prior Art Analysis

VisRL[60] demonstrates prior work on data selection and reward strategies for visual perception tasks. Both papers employ data filtering mechanisms to select training samples based on difficulty and task suitability. The original paper uses a 'perception-utility filter' to retain samples solvable via active perception, while VisRL[60] applies filtering by setting win/lose thresholds to ensure questions have appropriate difficulty levels. Both papers also implement conditional reward strategies: the original assigns bonuses to trajectories completing tasks through active perception, while VisRL[60] uses outcome rewards combined with process rewards to guide the model's reasoning. These parallel approaches to data curation and reward design indicate that VisRL[60] established similar methodologies prior to the original paper.

Evidence

Evidence 1 - **Rationale**: Both papers describe filtering mechanisms to select appropriate training samples. The original filters for 'potential to encourage active perception behavior' while VisRL[60] filters based on win/lose thresholds to ensure data validity. - **Original**: we propose a data selection mechanism to choose training samples based on their potential to encourage active perception behavior. additionally, we design a reward strategy that assigns a conditional bonus to the trajectories that successfully complete their tasks through active perception. - **Candidate**: to ensure the validity of preference data pairs in the candidate set p, we apply filtering by setting win and lose thresholds, t b max and t b min for bounding boxes, t r max and t r min for responses

Evidence 2 - **Rationale**: Both papers implement data selection filters based on task suitability. The original uses a 'perception-utility filter' while VisRL[60] filters by difficulty level, both aiming to maximize training efficiency. - **Original**: the final stage applies aperception-utility filter, retaining only samples solvable via active perception with ground-truth regions, thereby maximizing informational gain and boosting initial rl sampling efficiency without an sft cold start. - **Candidate**: the intuition here is that we apply a filter to the questions in the dataset, selecting those with a difficulty level suitable for the current model. questions that are too difficult will prevent the model from generating meaningful answers, while questions that are too easy will result in the model...

Evidence 3 - **Rationale**: Both papers describe conditional reward strategies. The original grants bonuses when active perception leads to correct answers, while VisRL[60] uses outcome and process rewards to guide reasoning, demonstrating similar reward design principles. - **Original**: the total reward consists of three parts: an accuracy reward racc, a format reward rformat, and a conditional bonus rtool. accuracy measures whether the final answer is correct, while formatting penalizes poorly structured outputs. the conditional bonus is granted only when the answer is correct and... - **Candidate**: our method leverages the final task success or failure as the outcome reward, and the grades of intermediate steps as the process reward, guiding the model to gradually refine its reasoning process through reinforcement learning.

---

8. GM-PRM: A Generative Multimodal Process Reward Model for Multimodal Mathematical Reasoning

URL: View paper

Brief Assessment

GM-PRM[62] focuses on process reward models for mathematical reasoning verification and correction, not on data selection mechanisms or reward strategies for active perception in visual reasoning tasks.

---

9. Vr-thinker: Boosting video reward models through thinking-with-image reasoning

URL: View paper

Brief Assessment

VR-thinker[61] focuses on video reward models with frame selection tools for preference evaluation, not on general visual reasoning with active perception mechanisms for task completion as in the original paper.

---

10. SR-AIF: Solving Sparse-Reward Robotic Tasks From Pixels with Active Inference and World Models

URL: View paper

Brief Assessment

SR-AIF[68] focuses on robotic control tasks using active inference with contrastive learning for prior preference, not on data selection mechanisms for visual reasoning in vision-language models. The technical approaches and problem domains differ fundamentally.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Learning Active Perception via Self-Evolving Preference Optimization for GUI Grounding

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning View paper
- [1] LLaVA-CoT: Let Vision Language Models Reason Step-by-Step View paper
- [2] Self-rewarding vision-language model via reasoning decomposition View paper
- [3] Spatialrgpt: Grounded spatial reasoning in vision-language models View paper
- [4] SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities View paper
- [5] Instructblip: Towards general-purpose vision-language models with instruction tuning View paper
- [6] A survey of vision-language pre-trained models View paper
- [7] Incorporating Convolution Designs into Visual Transformers View paper
- [8] GeoChat:Grounded Large Vision-Language Model for Remote Sensing View paper
- [9] Matryoshka multimodal models View paper
- [10] InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output View paper
- [11] VinVL: Revisiting Visual Representations in Vision-Language Models View paper
- [12] Integrating visual interpretation and linguistic reasoning for geometric problem solving View paper
- [13] BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation View paper
- [14] Zero-shot Object Navigation with Vision-Language Models Reasoning View paper
- [15] Exploring The Visual Feature Space for Multimodal Neural Decoding View paper
- [16] Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models View paper
- [17] Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning View paper
- [18] Otter: A multi-modal model with in-context instruction tuning View paper
- [19] PaLM-E: An Embodied Multimodal Language Model View paper
- [20] Integrating Audio-Visual Features For Multimodal Deepfake Detection View paper
- [21] Visual instruction tuning towards general-purpose multimodal large language model: A survey View paper
- [22] Qilin-med-vl: Towards chinese large vision-language model for general healthcare View paper
- [23] Exploring the Role of CLIP Global Visual Features in Multimodal Large Language Models View paper
- [24] Reinforcing Spatial Reasoning in Vision-Language Models with Interwoven Thinking and Visual Drawing View paper
- [25] Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models View paper
- [26] Decoding visual neural representations by multimodal learning of brain-visual-linguistic features View paper
- [27] Latent visual reasoning View paper
- [28] CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models View paper
- [29] Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement View paper
- [30] Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models View paper
- [31] Simple o3: Towards Interleaved Vision-Language Reasoning View paper
- [32] From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models View paper
- [33] Incorporating Visual Experts to Resolve the Information Loss in Multimodal Large Language Models View paper
- [34] Visual In-Context Learning for Large Vision-Language Models View paper
- [35] Reasoning, scaling, generating with vision-language models View paper
- [36] Visual cognition in multimodal large language models View paper
- [37] Image Fusion via Vision-Language Model View paper
- [38] Towards Interpreting Visual Information Processing in Vision-Language Models View paper
- [39] Tag2text: Guiding vision-language model via image tagging View paper
- [40] SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model View paper
- [41] Task-Oriented Feature Compression for Multimodal Understanding via Device-Edge Co-Inference View paper
- [42] Smart Vision-Language Reasoners View paper
- [43] Insight-v: Exploring long-chain visual reasoning with multimodal large language models View paper
- [44] Multi-Layer Visual Feature Fusion in Multimodal LLMs: Methods, Analysis, and Best Practices View paper
- [45] Zero-shot visual reasoning by vision-language models: Benchmarking and analysis View paper
- [46] Visionâ□□language foundation model for echocardiogram interpretation View paper
- [47] LISA: Reasoning Segmentation via Large Language Model View paper
- [48] GITA: Graph to Visual and Textual Integration for Vision-Language Graph Reasoning View paper
- [49] Evlm: An efficient vision-language model for visual understanding View paper
- [50] Enhanced Reasoning via Multimodal LLMs and Collaborative Inference View paper
- [51] LLM-I: LLMs are Naturally Interleaved Multimodal Creators View paper
- [52] Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning View paper
- [53] VLM-R: Region Recognition, Reasoning, and Refinement for Enhanced Multimodal Chain-of-Thought View paper
- [54] Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution View paper
- [55] Reasoning-VLA: A Fast and General Vision-Language-Action Reasoning Model for Autonomous Driving View paper
- [56] PeRL: Permutation-Enhanced Reinforcement Learning for Interleaved Vision-Language Reasoning View paper
- [57] Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens View paper
- [58] UniVG-R1: Reasoning Guided Universal Visual Grounding with Reinforcement Learning View paper

- [59] Bridging Formal Language with Chain-of-Thought Reasoning to Geometry Problem Solving View paper
- [60] Visrl: Intention-driven visual perception via reinforced reasoning View paper
- [61] Vr-thinker: Boosting video reward models through thinking-with-image reasoning View paper
- [62] GM-PRM: A Generative Multimodal Process Reward Model for Multimodal Mathematical Reasoning View paper
- [63] Learn 3D VQA Better with Active Selection and Reannotation View paper
- [64] Training Vision-Language Process Reward Models for Test-Time Scaling in Multimodal Reasoning: Key Insights and Lessons Learned View paper
- [65] 3D-R1: Enhancing Reasoning in 3D VLMs for Unified Scene Understanding View paper
- [66] Learning Active Perception via Self-Evolving Preference Optimization for GUI Grounding View paper
- [67] Adaptive important region selection with reinforced hierarchical search for dense object detection View paper
- [68] SR-AIF: Solving Sparse-Reward Robotic Tasks From Pixels with Active Inference and World Models View paper
- [69] Active Vision for Embodied Agents Using Reinforcement Learning View paper
- [70] Contrastive region guidance: Improving grounding in vision-language models without training View paper
- [71] Cogvlm: Visual expert for pretrained language models View paper
- [72] Navgpt: Explicit reasoning in vision-and-language navigation with large language models View paper
- [73] ViGoR: Improving Visual Grounding of Large Vision Language Models with Fine-Grained Reward Modeling View paper
- [74] Robot navigation using physically grounded vision-language models in outdoor environments View paper
- [75] Learning Visual Grounding from Generative Vision and Language Model View paper
- [76] VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks View paper
- [77] Direct visual grounding by directing attention of visual tokens View paper
- [78] Geoground: A unified large vision-language model for remote sensing visual grounding View paper