# Novelty Assessment Report

**Paper**: SophiaVL-R1: Reinforcing MLLMs Reasoning with Thinking Reward
**PDF URL**: https://openreview.net/pdf?id=0tzvmjMcXC
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2026-01-05

## Abstract

Recent advances have shown success in eliciting strong reasoning abilities in multimodal large language models (MLLMs) through rule-based reinforcement learning (RL) with outcome rewards. However, this paradigm typically lacks supervision over the thinking process leading to the final outcome. As a result, the model may learn sub-optimal reasoning strategies, which can hinder its generalization ability. In light of this, we propose SophiaVL-R1, as an attempt to add reward signals for the thinking process in this paradigm. To achieve this, we first train a thinking reward model that evaluates the quality of the entire thinking process. Given that the thinking reward may be unreliable for certain samples due to reward hacking, we propose the Trust-GRPO method, which assigns a trustworthiness weight to the thinking reward during training. This weight is computed based on the thinking reward comparison of responses leading to correct answers versus incorrect answers, helping to mitigate the impact of potentially unreliable thinking rewards. Moreover, we design an annealing training strategy that gradually reduces the thinking reward over time, allowing the model to rely more on the accurate rule-based outcome reward in later training stages. Experiments show that our SophiaVL-R1 surpasses a series of reasoning MLLMs on various benchmarks (\textit{e.g.}, MathVisita, MMMU), demonstrating strong reasoning and generalization capabilities. Notably, our SophiaVL-R1-7B even outperforms LLaVA-OneVision-72B on most benchmarks, despite the latter having 10 $\times$ more parameters. All code, models, and datasets will be made publicly available.



## Core Task Landscape

This paper addresses: **Reinforcing Multimodal Large Language Model Reasoning with Process-Level Supervision**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

### Taxonomy Overview

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

- **Reinforcement Learning Frameworks for Multimodal Reasoning**
- **Process Reward Models and Step-Level Supervision**
- **Chain-of-Thought and Structured Reasoning Paradigms**
- **Data Construction and Training Strategies**
- **Domain-Specific Multimodal Reasoning Applications**
- **Reasoning Enhancement through External Knowledge and Mechanisms**
- **Hallucination Mitigation and Reasoning Reliability**
- **Architectural Components and Efficiency**
- **Comprehensive Surveys and Taxonomies**

### Complete Taxonomy Tree

- Reinforcing Multimodal Large Language Model Reasoning with Process-Level Supervision Survey Taxonomy
- Reinforcement Learning Frameworks for Multimodal Reasoning
  - Policy Optimization Methods ★ (4 papers)
  - [0] SophiaVL-R1: Reinforcing MLLMs Reasoning with Thinking Reward (Anon et al., 2026) View paper
  - [1] Vision-r1: Incentivizing reasoning capability in multimodal large language models (Huang Wenxuan, 2025) View paper
  - [2] R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization (Zhang Jingyi, 2025) View paper
  - [3] Enhancing the reasoning ability of multimodal large language models via mixed preference optimization (Wang Wei-yun, 2024) View paper
  - Multi-Domain and Multi-Agent RL Frameworks (2 papers)
  - [36] Medmmv: A controllable multimodal multi-agent framework for reliable and verifiable clinical reasoning (Liu Hongjun, 2025) View paper
  - [39] MoDoMoDo: Multi-Domain Data Mixtures for Multimodal LLM Reinforcement Learning (Liang Yiqing, 2025) View paper
  - RL Paradigms and Theoretical Surveys (1 papers)
  - [11] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models (Guanghao Zhou, 2025) View paper
- Process Reward Models and Step-Level Supervision
  - Visual Process Reward Models (3 papers)
  - [10] Visualprm: An effective process reward model for multimodal reasoning (Wang Wei-yun, 2025) View paper
  - [20] Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned (Pala Tej Deep, 2025) View paper
  - [37] MM-PRM: Enhancing Multimodal Mathematical Reasoning with Scalable Step-Level Supervision (Du Ling-Xiao, 2025) View paper
  - Generative Process Reward Models (1 papers)
  - [28] Gm-prm: A generative multimodal process reward model for multimodal mathematical reasoning (Yan Yi-bo, 2025) View paper

    - Process Supervision Frameworks and Benchmarks (3 papers)
    - [12] Agentps: Agentic process supervision for multi-modal content quality assurance through multi-round qa (Sun Yu, 2024) View paper
    - [25] Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics (Luo Ruilin, 2025) View paper
    - [26] Mpbench: A comprehensive multimodal reasoning benchmark for process errors identification (Xu, 2025) View paper
    - General Process Reward Model Surveys (1 papers)
    - [47] A Survey of Process Reward Models: From Outcome Signals to Process Supervisions for Large Language Models (Zhu Jiachen, 2025) View paper
- Chain-of-Thought and Structured Reasoning Paradigms
    - Autonomous Multi-Stage Reasoning (4 papers)
    - [6] Insight-v: Exploring long-chain visual reasoning with multimodal large language models (Dong Yu-hao, 2025) View paper
    - [7] Llava-cot: Let vision language models reason step-by-step (Xu Guowei, 2025) View paper
    - [8] Corvid: Improving multimodal large language models towards chain-of-thought reasoning (Jiang JingJing, 2025) View paper
    - [29] Llamav-o1: Rethinking step-by-step visual reasoning in llms (Ahsan Noor, 2025) View paper
    - Prompting-Based Reasoning Strategies (3 papers)
    - [13] Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models (Zheng Ge, 2023) View paper
    - [18] Image-of-thought prompting for visual reasoning refinement in multimodal large language models (Zhou Qiji, 2024) View paper
    - [49] Advancing Multimodal Large Language Models: Optimizing Prompt Engineering Strategies for Enhanced Performance (Minjun Son, 2025) View paper
    - Visual Reasoning with Intermediate Representations (3 papers)
    - [5] Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing (Fang, 2025) View paper
    - [21] Enhancing visual reasoning with autonomous imagination in multimodal large language models (Jingming Liu, 2024) View paper
    - [43] Sifthinker: Spatially-aware image focus for visual reasoning (Chen Zhangquan, 2025) View paper
    - Step-Level Reasoning with Fine-Grained Rewards (2 papers)
    - [35] Moss-chatv: Reinforcement learning with process reasoning reward for video temporal reasoning (Tao Sicheng, 2025) View paper
    - [50] Unveiling Chain of Step Reasoning for Vision-Language Models with Fine-grained Rewards (Chen, 2025) View paper
- Data Construction and Training Strategies
    - Reasoning Data Generation and Augmentation (2 papers)
    - [33] Read and Think: An Efficient Step-wise Multimodal Language Model for Document Understanding and Reasoning (Zhang Jin-xu, 2024) View paper
    - [34] Uniapo: Unified multimodal automated prompt optimization (Zhu Qi-peng, 2025) View paper
    - Training Pipeline Optimization (4 papers)
    - [14] Mm1: methods, analysis and insights from multimodal llm pre-training (Brandon McKinzie, 2024) View paper
    - [19] MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training (McKinzie, 2024) View paper
    - [38] Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training (Jia, 2024) View paper
    - [48] MM1.5: Methods, Analysis & Insights from Multimodal LLM Fine-tuning (Zhang, 2024) View paper
- Domain-Specific Multimodal Reasoning Applications
    - Mathematical and Scientific Reasoning (2 papers)
    - [9] A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges (He Jianxiang, 2025) View paper
    - [23] Position: Multimodal large language models can significantly advance scientific reasoning (Yan Yi-bo, 2025) View paper
    - Spatial and Physical Reasoning (3 papers)
    - [4] Reasoning grasping via multimodal large language model (Jin Shiyu, 2024) View paper
    - [15] Enhancing Spatial Reasoning in Multimodal Large Language Models through Reasoning-based Segmentation (Ning Zhen-Hua, 2025) View paper
    - [22] Spatial-ormllm: Improve spatial relation understanding in the operating room with multimodal large language model (He PeiQi, 2025) View paper
    - Medical and Clinical Reasoning (3 papers)
    - [16] Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering (Dexuan Xu, 2024) View paper
    - [45] Deciphering Cognitive Distortions in Patient-Doctor Mental Health Conversations: A Multimodal LLM-Based Detection and Reasoning Framework (Gopendra Vikram Singh, 2024) View paper
    - [46] A Survey of LLM-based Agents in Medicine: How far are we from Baymax? (Wenxuan Wang, 2025) View paper
    - Multimodal Information Verification (1 papers)
    - [42] Debunk and Infer: Multimodal Fake News Detection via Diffusion-Generated Evidence and LLM Reasoning (Liu Yu-kun, 2025) View paper
- Reasoning Enhancement through External Knowledge and Mechanisms
    - Knowledge Graph Integration (1 papers)
    - [24] Multimodal reasoning with multimodal knowledge graph (Li Jing, 2024) View paper
    - Analogical and Comparative Reasoning (1 papers)
    - [17] Can multimodal large language model think analogically? (Diandian Guo, 2024) View paper
- Hallucination Mitigation and Reasoning Reliability
    - Visual Grounding and Attention Correction (2 papers)
    - [27] Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination (Zheng Hao-jie, 2024) View paper
    - [31] Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning (Chua, 2025) View paper
    - Reasoning Decoupling and Contrastive Methods (1 papers)
    - [32] Language Models Can See Better: Visual Contrastive Decoding For LLM Multimodal Reasoning (Pang Yuqi, 2025) View paper
- Architectural Components and Efficiency
    - Visual Projector and Token Optimization (1 papers)

- [41] TokenPacker: Efficient Visual Projector for Multimodal LLM (Wentong Li, 2024) View paper
- Internal Processing Analysis (1 papers)
- [40] How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding (Yu, 2025) View paper

• Comprehensive Surveys and Taxonomies (2 papers)
- [30] Multimodal chain-of-thought reasoning: A comprehensive survey (Wang Yao-ting, 2025) View paper
- [44] From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond (H Fei, 2024) View paper

### Narrative

Core task: reinforcing multimodal large language model reasoning with process-level supervision. The field has organized itself around several complementary directions. Reinforcement learning frameworks for multimodal reasoning explore policy optimization methods that align vision-language models with human preferences and reasoning objectives, often drawing on techniques like mixed preference optimization (Mixed Preference Optimization[3]) and step-level reward signals. Process reward models and step-level supervision form a dense branch focused on training verifiers that evaluate intermediate reasoning steps rather than only final answers, with works such as Vision Process Rewards[20] and MM-PRM[37] developing multimodal extensions of process-level feedback. Chain-of-thought and structured reasoning paradigms investigate how to elicit and represent explicit reasoning traces in vision-language settings, spanning methods like Llava-cot[7] and Insight-v[6]. Data construction and training strategies address the challenge of generating high-quality reasoning annotations at scale, while domain-specific applications target areas such as scientific reasoning, medical diagnosis, and robotic grasping. Additional branches cover reasoning enhancement through external knowledge, hallucination mitigation, architectural efficiency, and comprehensive surveys that synthesize emerging trends.

Particularly active lines of work center on integrating reinforcement learning with process-level rewards to improve step-by-step reasoning quality in multimodal contexts. SophiaVL[0] sits within the policy optimization cluster, emphasizing how RL techniques can refine reasoning trajectories by leveraging fine-grained supervision signals at each reasoning step. This approach contrasts with outcome-based methods and aligns closely with Vision-r1[1] and R1-vl[2], which similarly apply policy gradient or actor-critic frameworks to vision-language reasoning. Compared to Mixed Preference Optimization[3], which blends multiple preference signals, SophiaVL[0] focuses more directly on process-level reward shaping to guide intermediate reasoning decisions. A key open question across these works is how to balance the cost of annotating step-level feedback with the gains in reasoning reliability, and whether learned process reward models can generalize across diverse multimodal tasks without extensive domain-specific tuning.

## Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. Vision-r1: Incentivizing reasoning capability in multimodal large language models

**Authors**: Huang Wenxuan, Jia Bohan, Wenxuan Huang, Zhai Zijie, Bohan Jia, et al. (19 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

DeepSeek-R1-Zero has successfully demonstrated the emergence of reasoning capabilities in LLMs purely through Reinforcement Learning (RL). Inspired by this breakthrough, we explore how RL can be utilized to enhance the reasoning capability of MLLMs. However, direct training with RL struggles to activate complex reasoning capabilities such as questioning and reflection in MLLMs, due to the absence of substantial high-quality multimodal reasoning data. To address this issue, we propose the reasoni...

#### Relationship Analysis

Both papers belong to the Policy Optimization Methods category, employing reinforcement learning techniques to enhance multimodal reasoning in MLLMs. They share overlapping approaches in using GRPO (Group Relative Policy Optimization) with outcome-based rewards to train reasoning models on multimodal tasks. The key difference is that SophiaVL-R1 introduces a thinking reward model with trustworthiness weighting to supervise the reasoning process holistically, while Vision-R1 focuses on cold-start initialization with high-quality CoT data and Progressive Thinking Suppression Training (PTST) to address overthinking problems during RL training.

---

### 2. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization

**Authors**: Zhang Jingyi, Huang Jiaxing, Liu Shunyu, Zhang Xikun, Lu, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

Recent studies generally enhance MLLMs' reasoning capabilities via supervised fine-tuning on high-quality chain-of-thought reasoning data, which often leads models to merely imitate successful reasoning paths without understanding what the wrong reasoning paths are. In this work, we aim to enhance the MLLMs' reasoning ability beyond passively imitating positive reasoning paths. To this end, we design Step-wise Group Relative Policy Optimization (StepGRPO), a new online reinforcement learning fra...

#### Relationship Analysis

Both papers belong to the Policy Optimization Methods category, employing reinforcement learning techniques to enhance multimodal large language model reasoning through process-level supervision. They overlap in using group relative policy optimization (GRPO) frameworks and addressing the challenge of reward signal quality during training. The key difference is that SophiaVL-R1 introduces a holistic thinking reward model with trustworthiness weighting to evaluate entire reasoning processes, while R1-VL proposes step-wise rewards (StepRAR and StepRVR) using rule-based mechanisms to provide dense, fine-grained supervision at each reasoning step without requiring additional reward models.

---

### 3. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization

**Authors**: Wang Wei-yun, Weiâ□□Yun Wang, Chen Zhe, Zhe Chen, Weiyun Wang, et al. (26 authors total) | **Year/Venue**: 2024 | **URL**: View paper

#### Abstract

Existing open-source multimodal large language models (MLLMs) generally follow a training process involving pre-training and supervised fine-tuning. However, these models suffer from distribution shifts, which limit their multimodal reasoning, particularly in the Chain-of-Thought (CoT) performance. To address this, we introduce a preference optimization (PO) process to enhance the multimodal reasoning capabilities of MLLMs. Specifically, (1) on the data side, we design an automated preference da...

#### Relationship Analysis

Both papers belong to the Policy Optimization Methods category, employing policy gradient techniques to enhance MLLM reasoning through reinforcement learning. They overlap in using rule-based outcome rewards and addressing reasoning quality, but differ fundamentally in their approach: the original paper (SophiaVL-R1) introduces a holistic thinking reward model with Trust-GRPO to assess entire reasoning processes and mitigate reward hacking, while the candidate paper focuses on mixed preference optimization

(MPO) combining DPO, BCO, and SFT losses with an automated preference data construction pipeline (DropoutNTP) to improve reasoning through preference learning. The original emphasizes process-level supervision via thinking rewards, whereas the candidate emphasizes preference-based optimization with diverse CoT methods.

## Contributions Analysis

**Overall novelty summary.** The paper proposes SophiaVL-R1, which introduces a thinking reward model to evaluate the entire reasoning process in multimodal large language models, combined with a Trust-GRPO training method that weights process rewards by trustworthiness. This work resides in the 'Policy Optimization Methods' leaf under 'Reinforcement Learning Frameworks for Multimodal Reasoning', which contains four papers total. The leaf focuses on policy gradient and relative policy optimization techniques for MLLM training, explicitly excluding outcome-only reward models. This places the paper in a moderately populated research direction within a broader taxonomy of fifty papers across thirty-six topics, suggesting active but not overcrowded exploration of RL-based reasoning enhancement.

The taxonomy reveals several neighboring research directions that contextualize this work. The sibling leaf 'Multi-Domain and Multi-Agent RL Frameworks' addresses heterogeneous tasks and agent interactions, while the parent branch also includes 'RL Paradigms and Theoretical Surveys'. Adjacent branches cover 'Process Reward Models and Step-Level Supervision' (with visual PRMs and generative PRMs) and 'Chain-of-Thought and Structured Reasoning Paradigms' (including autonomous multi-stage reasoning). The paper bridges policy optimization methods with process-level supervision concepts, drawing on ideas from both the RL frameworks branch and the process reward models branch, though it sits formally within the former.

Among twenty-three candidates examined, the analysis found three refutable pairs for the 'SophiaVL-R1 multimodal reasoning model' contribution (ten candidates examined), while the 'thinking reward model' (ten candidates) and 'Trust-GRPO algorithm' (three candidates) showed no clear refutations. The limited search scope—top-K semantic matches plus citation expansion—means these statistics reflect a targeted rather than exhaustive review. The thinking reward model and Trust-GRPO components appear more distinctive within the examined literature, whereas the overall model architecture encounters some overlapping prior work among the candidates reviewed.

Based on the limited search of twenty-three candidates, the paper's process-level reward modeling and trustworthiness weighting mechanisms appear relatively novel, while the integrated model faces more substantial prior work. The taxonomy structure suggests this research direction—combining RL policy optimization with process supervision—remains an active area with room for methodological contributions. However, the analysis does not cover the full landscape of multimodal reasoning research, and a broader literature search might reveal additional related work in adjacent branches such as 'Step-Level Reasoning with Fine-Grained Rewards' or 'Generative Process Reward Models'.

---

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

### Contribution 1: Thinking reward model for holistic reasoning quality evaluation

**Description**: The authors introduce a thinking reward model trained on annotated reasoning responses that evaluates the entire thinking process holistically rather than step-by-step. This model assesses reasoning quality across dimensions such as logical soundness, consistency, and redundancy to help distinguish favorable from flawed reasoning patterns.

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. Rewardbench: Evaluating reward models for language modeling

**URL**: View paper

**Brief Assessment**

RewardBench[52] focuses on evaluating reward models across chat, reasoning, and safety domains using existing test sets, not on training holistic thinking reward models that assess reasoning quality across dimensions like logical soundness and redundancy.

---

#### 2. AceMath: Advancing Frontier Math Reasoning with Post-Training and Reward Modeling

**URL**: View paper

**Brief Assessment**

AceMath[57] focuses on developing math-specialized reward models for evaluating final solutions and identifying correct answers, not on holistic evaluation of intermediate reasoning processes across multiple quality dimensions.

---

#### 3. ToTRL: Unlock LLM Tree-of-Thoughts Reasoning Potential through Puzzles Solving

**URL**: View paper

**Brief Assessment**

ToTRL[59] focuses on tree-of-thoughts reasoning with rule-based rewards for puzzle solving, not on training a model-based reward model that evaluates holistic reasoning quality across multiple dimensions.

---

#### 4. Teaching large language models to reason with reinforcement learning

**URL**: View paper

**Brief Assessment**

Teaching Reasoning[58] focuses on sparse/dense outcome rewards and process reward models (PRMs) with step-wise evaluation, not holistic thinking process evaluation. The paper does not propose a reward model that evaluates entire reasoning processes holistically across dimensions like logical soundness, consistency, and redundancy as described in the original contribution.

---

#### 5. Multimodal RewardBench: Holistic Evaluation of Reward Models for Vision Language Models

**URL**: View paper

**Brief Assessment**

Multimodal RewardBench[56] focuses on benchmarking reward models across multiple domains (general correctness, preference, knowledge, reasoning, safety, VQA) rather than proposing a specific thinking reward model for holistic reasoning evaluation. The candidate is an evaluation benchmark, not a method for training reward models to assess reasoning processes.

---

#### 6. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models

**URL**: View paper

**Brief Assessment**

Reinforced MLLM Survey[11] discusses process reward mechanisms (PRM) that evaluate intermediate reasoning steps, but does not specifically describe a holistic thinking reward model trained on annotated responses that evaluates entire reasoning processes across dimensions like logical soundness, consistency, and redundancy as proposed in the original paper.

---

### 7. Improve Mathematical Reasoning in Language Models by Automated Process Supervision

**URL**: View paper

**Brief Assessment**

Automated Process Supervision[53] focuses on process reward models (PRMs) that provide step-wise supervision through Monte Carlo tree search, not holistic evaluation of entire reasoning processes across multiple quality dimensions.

---

### 8. Reasoning with language model is planning with world model

**URL**: View paper

**Brief Assessment**

Planning World Model[51] focuses on using LLMs as world models for planning-based reasoning with MCTS, not on training reward models to evaluate reasoning quality holistically. The candidate's reward design is task-specific and heuristic-based, fundamentally different from the original paper's trained thinking reward model approach.

---

### 9. Unlocking Multimodal Mathematical Reasoning via Process Reward Model

**URL**: View paper

**Brief Assessment**

Multimodal Process Reward[55] focuses on process-level supervision with step-wise rewards and automated process labeling, while the original paper evaluates reasoning holistically across dimensions like logical soundness and redundancy without step-level constraints.

---

### 10. Unlocking the mysteries of OpenAI o1: A survey of the reasoning abilities of large language models

**URL**: View paper

**Brief Assessment**

OpenAI o1 Survey[54] does not discuss holistic reward models for reasoning quality evaluation. The survey focuses on reviewing reasoning techniques (tree-of-thoughts, MCTS, self-correction) and training methods (SFT, RL, DPO) but does not present a reward model that evaluates entire thinking processes holistically.

---

## Contribution 2: Trust-GRPO algorithm with trustworthiness weighting and annealing

**Description**: The authors propose Trust-GRPO, a training algorithm that assigns a trustworthiness weight to thinking rewards by comparing rewards of correct versus incorrect responses. It includes a time-based annealing strategy that gradually reduces thinking reward influence, allowing the model to rely more on accurate rule-based outcome rewards in later training stages.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. Beyond Correctness: Confidence-Aware Reward Modeling for Enhancing Large Language Model Reasoning

**URL**: View paper

**Brief Assessment**

Confidence-Aware Reward[61] focuses on penalizing low-confidence correct responses in STEM reasoning tasks, not on trustworthiness weighting of thinking rewards or annealing strategies in GRPO training. The candidate addresses confidence estimation for reward modeling, while the original contribution concerns dynamic weighting of process rewards during RL training.

---

### 2. From Hypothesis to Premises: LLM-based Backward Logical Reasoning with Selective Symbolic Translation

**URL**: View paper

**Brief Assessment**

Backward Logical Reasoning[62] focuses on hypothesis-driven backward reasoning with symbolic translation for logical inference tasks, not on reinforcement learning training algorithms with trustworthiness weighting or annealing strategies for reasoning models.

---

### 3. Confidence-Guided Stepwise Model Routing for Cost-Efficient Reasoning

**URL**: View paper

**Brief Assessment**

Stepwise Model Routing[60] focuses on routing between models at the step level using confidence scores for cost-efficient inference, not on training algorithms with trustworthiness weighting for reinforcement learning rewards.

---

## Contribution 3: SophiaVL-R1 multimodal reasoning model

**Description**: The authors develop SophiaVL-R1, a multimodal large language model that enhances reasoning by integrating model-generated thinking rewards with rule-based outcome rewards during reinforcement learning training. The model demonstrates strong reasoning and generalization capabilities across various benchmarks.

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. GraphFusion-HRL: Multi-Modal Hierarchical Reinforcement Graph Learning for Context-Rich Recommender Systems

**URL**: View paper

**Brief Assessment**

GraphFusion-HRL[64] focuses on hierarchical reinforcement learning for recommender systems with graph-based multi-modal fusion, not on multimodal reasoning models that combine thinking rewards with outcome rewards for chain-of-thought reasoning.

---

### 2. VLRMBench: A Comprehensive and Challenging Benchmark for Vision-Language Reward Models

**URL**: View paper

**Brief Assessment**

VLRMBench[70] is a benchmark for evaluating vision-language reward models, not a multimodal reasoning model that combines thinking rewards with outcome rewards during RL training.

---

### 3. Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned

**URL**: View paper

**Prior Art Analysis**

Vision Process Rewards[20] demonstrates that multimodal reasoning models can be enhanced by combining model-generated thinking rewards with outcome rewards during training, which directly challenges the novelty of SophiaVL-R1's core contribution. Both papers train reward models to evaluate reasoning quality and integrate these rewards with rule-based outcome rewards. Vision Process Rewards[20] explicitly trains process reward models (PRMs) that 'provide step-level supervision that improves the reliability of reasoning' and uses them alongside outcome evaluation, similar to SophiaVL-R1's approach of combining thinking rewards with outcome rewards. The candidate paper's hybrid reward framework and its application to multimodal reasoning predates or parallels the original's contribution.

**Evidence**

Evidence 1 - **Rationale**: Both papers train reward models to evaluate reasoning quality in multimodal contexts. The candidate explicitly describes PRMs providing supervision for reasoning reliability, which parallels the original's thinking reward model concept. - **Original**: we propose sophiavl-r1, as an attempt to add reward signals for the thinking process in this paradigm. to achieve this, we first train a thinking reward model that evaluates the quality of the entire thinking process. - **Candidate**: process reward models (prms) provide step-level supervision that improves the reliability of reasoning in large language models. while prms have been extensively studied in text-based domains, their extension to vision language models (vlms) remains limited.

Evidence 2 - **Rationale**: Both papers integrate model-generated process/thinking rewards with outcome evaluation. The candidate's VL-PRM assigns scores to reasoning steps, similar to the original's thinking reward model evaluating reasoning quality. - **Original**: we proposesophiavl-r1, an mllm that enhances reasoning by integrating model-generated thinking rewards with rule-based outcome rewards in rl training. - **Candidate**: our vl-prms are used as a critic model to judge the chain-of-thought (cot) step-by-step responses generated by the policy model. at inference time, the vision-language process reward model (vl-prm) receives an image-question pair (i, q)along with a partial reasoning trajectory {s1, . . . , si-1}, an...

Evidence 3 - **Rationale**: Both papers train reward models on annotated reasoning data to evaluate step-level correctness. The candidate's VL-PRM training approach directly parallels the original's thinking reward model training methodology. - **Original**: we introduce a thinking reward model trained on annotated reasoning responses collected from grpo training trajectories. this model evaluates intermediate reasoning quality based on criteria such as logical soundness, consistency across steps, and redundancy in the thinking process. - **Candidate**: vl-prms are trained on vl-prm300k to predict '+' or '-' as step-level judgments. training minimizes the cross-entropy loss over these label tokens for every step. the input to the vl-prm is{i, q, s0, p}. here, the promptpis the system prompt for vl-prm.

Evidence 4 - **Rationale**: Both papers demonstrate improved performance on MMMU and other multimodal reasoning benchmarks through their respective reward model approaches, showing similar evaluation strategies and performance claims. - **Original**: sophiavl-r1-7b consistently outperforms existing mllms on diverse benchmarks (e.g., mathvista, mmmu), highlighting its strong reasoning and generalization abilities. - **Candidate**: both qwen-vl-prm-3b and qwen-vl-prm-7b consistently enhance the performance of all evaluated policies under test-time scaling. on the expert-knowledge-oriented datasetmmmu, however, the observed gains are relatively modest, with improvements averaging around 2.5% across the three policies.

---

### 4. REVISOR: Beyond Textual Reflection, Towards Multimodal Introspective Reasoning in Long-Form Video Understanding

**URL**: View paper

**Brief Assessment**

REVISOR[66] focuses on long-form video understanding with visual segment reflection mechanisms, while SophiaVL-R1 addresses general multimodal reasoning by combining thinking rewards with outcome rewards during RL training. These are distinct technical approaches targeting different problem domains.

---

### 5. SAIL-RL: Guiding MLLMs in When and How to Think via Dual-Reward RL Tuning

**URL**: View paper

**Prior Art Analysis**

SAIL-RL[69] demonstrates that prior work exists on multimodal reasoning models that combine thinking rewards with outcome rewards. Both papers develop multimodal large language models that enhance reasoning by integrating model-generated thinking rewards with rule-based outcome rewards during reinforcement learning training. SAIL-RL[69] introduces a dual reward system with a 'thinking reward' that evaluates reasoning quality and a 'judging reward' for adaptive reasoning, while the original paper proposes SophiaVL-R1 with a thinking reward model and Trust-GRPO algorithm. The core innovation of combining thinking process rewards with outcome rewards in MLLM RL training is present in both works.

**Evidence**

Evidence 1 - **Rationale**: Both papers propose training multimodal reasoning models using thinking rewards that evaluate reasoning quality. SAIL-RL[69] explicitly describes a 'thinking reward' system that evaluates reasoning quality, demonstrating that this approach existed prior to or contemporaneously with the original paper's contribution. - **Original**: we propose sophiavl-r1, as an attempt to add reward signals for the thinking process in this paradigm. to achieve this, we first train a thinking reward model that evaluates the quality of the entire thinking process. - **Candidate**: sail-rl addresses these challenges with a dual reward system: the thinking reward, which evaluates reasoning quality through factual grounding, logical coherence, and answer consistency, and the judging reward, which adaptively determines whether deep reasoning or direct answering is appropriate.

Evidence 2 - **Rationale**: Both papers describe RL frameworks for enhancing MLLM reasoning capabilities. SAIL-RL[69] presents a reinforcement learning framework that combines thinking rewards with outcome supervision, showing that the core concept of using RL with thinking rewards for MLLMs was already established. - **Original**: we proposesophiavl-r1, an mllm that enhances reasoning by integrating model-generated thinking rewards with rule-based outcome rewards in rl training. - **Candidate**: we introduce sail-rl, a reinforcement learning (rl) post-training framework that enhances the reasoning capabilities of multimodal large language models (mllms) by teaching them when and how to think.

Evidence 3 - **Rationale**: Both papers evaluate reasoning quality using similar criteria. SAIL-RL[69]'s thinking reward evaluates 'logical coherence' and 'answer consistency,' which directly parallels the original paper's evaluation of 'logical soundness' and 'consistency across steps,' indicating prior work on this evaluation approach. - **Original**: we introduce a thinking reward model trained on annotated reasoning responses collected from grpo training trajectories. this model evaluates intermediate reasoning quality based on criteria such as logical soundness, consistency across steps, and redundancy in the thinking process. - **Candidate**: sail-rl addresses these challenges with a dual reward system: the thinking reward, which evaluates reasoning quality through factual grounding, logical coherence, and answer consistency

---

### 6. Semi-off-Policy Reinforcement Learning for Vision-Language Slow-thinking Reasoning

**URL**: View paper

**Brief Assessment**

Semi-off-Policy[68] focuses on semi-off-policy RL combining on-policy visual understanding with off-policy language reasoning, while the original paper develops a thinking reward model with trust-grpo for holistic reasoning quality assessment. These represent distinct technical approaches to multimodal reasoning enhancement.

---

### 7. Self-rewarding vision-language model via reasoning decomposition

**URL**: View paper

**Brief Assessment**

Self-Rewarding Vision[65] focuses on decomposing reasoning into visual perception and language reasoning stages with self-contained visual perceptions, whereas SophiaVL-R1 integrates thinking rewards (evaluating holistic reasoning quality) with outcome rewards. The candidate does not demonstrate prior work on combining model-generated thinking rewards with rule-based outcome rewards in the manner described by the original paper.

---

### 8. Vision-r1: Incentivizing reasoning capability in multimodal large language models

**URL**: View paper

**Prior Art Analysis**

Vision-r1[1] demonstrates that prior work exists in developing multimodal reasoning models that combine thinking rewards with outcome rewards through reinforcement learning. Both papers develop multimodal large language models that enhance reasoning by integrating model-generated thinking/process rewards with rule-based outcome rewards during RL training. Vision-r1[1] constructs a thinking reward model through cold-start initialization and applies progressive thinking suppression training (PTST) with GRPO, while the original paper develops a thinking reward model trained on annotated reasoning responses and introduces trust-GRPO with trustworthiness weighting. The core contribution of combining thinking rewards with outcome rewards in multimodal RL training was demonstrated by Vision-r1[1], challenging the novelty claim.

**Evidence**

Evidence 1 - **Rationale**: Both papers propose multimodal reasoning models that integrate thinking/process evaluation with RL training, demonstrating prior work in this approach. - **Original**: we propose sophiavl-r1, as an attempt to add reward signals for the thinking process in this paradigm. to achieve this, we first train a thinking reward model that evaluates the quality of the entire thinking process. - **Candidate**: we propose vision-r1, a reasoning mllm that integrates cold-start initialization with rl training. first, we construct a high-quality multimodal cot dataset without requiring manual annotations.

Evidence 2 - **Rationale**: Vision-r1[1] demonstrates the integration of thinking process evaluation (through cold-start initialization and PTST) with outcome rewards (hard formatting result rewards) in GRPO training, showing prior work combining these reward types. - **Original**: we proposesophiavl-r1, an mllm that enhances reasoning by integrating model-generated thinking rewards with rule-based outcome rewards in rl training. - **Candidate**: we implement group relative policy optimization (grpo) with hard formatting result rewards for the model's self-learning.

Evidence 3 - **Rationale**: Both papers identify the same limitation of outcome-only rewards and propose solutions involving thinking process evaluation, demonstrating that Vision-r1[1] addresses the same problem space. - **Original**: The key of these methods is to utilize a rule-based function that yields accurate outcome reward signals for rl training (guo et al., 2025; leng et al., 2025; deng et al., 2025b). however, solely relying on the outcome reward usually fails to ensure the quality of the thinking process - **Candidate**: however, direct training with rl struggles to activate complex reasoning capabilities such as questioning and reflection in mllms, due to the absence of substantial high-quality multimodal reasoning data. to address this issue, we propose the reasoning mllm, visionr1, to improve multimodal reasoning...

Evidence 4 - **Rationale**: Both papers develop methods to evaluate thinking process quality in multimodal reasoning, with Vision-r1[1] using cold-start initialization with complex CoT data to guide thinking processes, demonstrating prior work in this area. - **Original**: we introduce a thinking reward model trained on annotated reasoning responses collected from grpo training trajectories. this model evaluates intermediate reasoning quality based on criteria such as logical soundness, consistency across steps, and redundancy in the thinking process. - **Candidate**: we first construct a high-quality multimodal cot dataset without requiring manual annotations. specifically, we leverage an existing mllm to generate 'pseudo-cot' reasoning text from multimodal image-text pairs.

Evidence 5 - **Rationale**: Vision-r1[1] demonstrates a method to guide reasoning quality through PTST during RL training, showing prior work in evaluating and improving thinking processes in multimodal RL frameworks. - **Original**: in summary, our contributions are as follows: • we propose a thinking reward model that evaluates reasoning quality from various dimensions at a holistic level, enabling the model to distinguish between sound and flawed reasoning processes during rule-based rl training. - **Candidate**: by applying ptst, we compress the model's thought length in the early training stages to guide correct reasoning and gradually relax these constraints in later stages. as illustrated in fig. 1, this progressive strategy enables vision-r1 to generate more complex cot and significantly enhances its re...

---

### 9. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning

**URL**: View paper

**Brief Assessment**

Audio-Thinker[67] focuses on large audio language models (LALMs) for audio question answering tasks, not multimodal visual-language reasoning. The candidate addresses audio-specific reasoning challenges rather than the visual reasoning domain of the original paper.

---

### 10. Progressive Multimodal Reasoning via Active Retrieval

**URL**: View paper

**Brief Assessment**

Progressive Retrieval[63] focuses on active retrieval with MCTS for multimodal reasoning, while SophiaVL-R1 combines thinking rewards with outcome rewards in RL training. These represent distinct technical approaches to enhancing multimodal reasoning.

---

## Appendix: Text Similarity Detection

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

## References


- [0] SophiaVL-R1: Reinforcing MLLMs Reasoning with Thinking Reward View paper
- [1] Vision-r1: Incentivizing reasoning capability in multimodal large language models View paper
- [2] R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization View paper
- [3] Enhancing the reasoning ability of multimodal large language models via mixed preference optimization View paper
- [4] Reasoning grasping via multimodal large language model View paper
- [5] Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing View paper
- [6] Insight-v: Exploring long-chain visual reasoning with multimodal large language models View paper
- [7] Llava-cot: Let vision language models reason step-by-step View paper
- [8] Corvid: Improving multimodal large language models towards chain-of-thought reasoning View paper
- [9] A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges View paper
- [10] Visualprm: An effective process reward model for multimodal reasoning View paper

- [11] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models View paper
- [12] Agentps: Agentic process supervision for multi-modal content quality assurance through multi-round qa View paper
- [13] Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models View paper
- [14] Mm1: methods, analysis and insights from multimodal llm pre-training View paper
- [15] Enhancing Spatial Reasoning in Multimodal Large Language Models through Reasoning-based Segmentation View paper
- [16] Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering View paper
- [17] Can multimodal large language model think analogically? View paper
- [18] Image-of-thought prompting for visual reasoning refinement in multimodal large language models View paper
- [19] MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training View paper
- [20] Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned View paper
- [21] Enhancing visual reasoning with autonomous imagination in multimodal large language models View paper
- [22] Spatial-ormllm: Improve spatial relation understanding in the operating room with multimodal large language model View paper
- [23] Position: Multimodal large language models can significantly advance scientific reasoning View paper
- [24] Multimodal reasoning with multimodal knowledge graph View paper
- [25] Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics View paper
- [26] Mpbench: A comprehensive multimodal reasoning benchmark for process errors identification View paper
- [27] Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination View paper
- [28] Gm-prm: A generative multimodal process reward model for multimodal mathematical reasoning View paper
- [29] Llamav-o1: Rethinking step-by-step visual reasoning in llms View paper
- [30] Multimodal chain-of-thought reasoning: A comprehensive survey View paper
- [31] Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning View paper
- [32] Language Models Can See Better: Visual Contrastive Decoding For LLM Multimodal Reasoning View paper
- [33] Read and Think: An Efficient Step-wise Multimodal Language Model for Document Understanding and Reasoning View paper
- [34] Uniapo: Unified multimodal automated prompt optimization View paper
- [35] Moss-chatv: Reinforcement learning with process reasoning reward for video temporal reasoning View paper
- [36] Medmmv: A controllable multimodal multi-agent framework for reliable and verifiable clinical reasoning View paper
- [37] MM-PRM: Enhancing Multimodal Mathematical Reasoning with Scalable Step-Level Supervision View paper
- [38] Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training View paper
- [39] MoDoMoDo: Multi-Domain Data Mixtures for Multimodal LLM Reinforcement Learning View paper
- [40] How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding View paper
- [41] TokenPacker: Efficient Visual Projector for Multimodal LLM View paper
- [42] Debunk and Infer: Multimodal Fake News Detection via Diffusion-Generated Evidence and LLM Reasoning View paper
- [43] Sifthinker: Spatially-aware image focus for visual reasoning View paper
- [44] From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond View paper
- [45] Deciphering Cognitive Distortions in Patient-Doctor Mental Health Conversations: A Multimodal LLM-Based Detection and Reasoning Framework View paper
- [46] A Survey of LLM-based Agents in Medicine: How far are we from Baymax? View paper
- [47] A Survey of Process Reward Models: From Outcome Signals to Process Supervisions for Large Language Models View paper
- [48] MM1.5: Methods, Analysis & Insights from Multimodal LLM Fine-tuning View paper
- [49] Advancing Multimodal Large Language Models: Optimizing Prompt Engineering Strategies for Enhanced Performance View paper
- [50] Unveiling Chain of Step Reasoning for Vision-Language Models with Fine-grained Rewards View paper
- [51] Reasoning with language model is planning with world model View paper
- [52] Rewardbench: Evaluating reward models for language modeling View paper
- [53] Improve Mathematical Reasoning in Language Models by Automated Process Supervision View paper
- [54] Unlocking the mysteries of OpenAI o1: A survey of the reasoning abilities of large language models View paper
- [55] Unlocking Multimodal Mathematical Reasoning via Process Reward Model View paper
- [56] Multimodal RewardBench: Holistic Evaluation of Reward Models for Vision Language Models View paper
- [57] AceMath: Advancing Frontier Math Reasoning with Post-Training and Reward Modeling View paper
- [58] Teaching large language models to reason with reinforcement learning View paper
- [59] ToTRL: Unlock LLM Tree-of-Thoughts Reasoning Potential through Puzzles Solving View paper
- [60] Confidence-Guided Stepwise Model Routing for Cost-Efficient Reasoning View paper
- [61] Beyond Correctness: Confidence-Aware Reward Modeling for Enhancing Large Language Model Reasoning View paper
- [62] From Hypothesis to Premises: LLM-based Backward Logical Reasoning with Selective Symbolic Translation View paper
- [63] Progressive Multimodal Reasoning via Active Retrieval View paper
- [64] GraphFusion-HRL: Multi-Modal Hierarchical Reinforcement Graph Learning for Context-Rich Recommender Systems View paper
- [65] Self-rewarding vision-language model via reasoning decomposition View paper
- [66] REVISOR: Beyond Textual Reflection, Towards Multimodal Introspective Reasoning in Long-Form Video Understanding View paper
- [67] Audio-thinker: Guiding audio language model when and how to think via reinforcement learning View paper
- [68] Semi-off-Policy Reinforcement Learning for Vision-Language Slow-thinking Reasoning View paper
- [69] SAIL-RL: Guiding MLLMs in When and How to Think via Dual-Reward RL Tuning View paper
- [70] VLRMBench: A Comprehensive and Challenging Benchmark for Vision-Language Reward Models View paper