Novelty Assessment Report

Paper: Perception-Aware Policy Optimization for Multimodal Reasoning

PDF URL: https://openreview.net/pdf?id=izbBqTL8vb

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has proven to be a highly effective strategy for empowering Large Language Models (LLMs) with long chain-of-thought reasoning abilities. However, its design and optimizations remain tailored to purely textual domains, resulting in suboptimal performance when applied to multimodal reasoning tasks. In particular, we observe that a major source of error (67%) in current multimodal reasoning lies in the perception of visual inputs. To address this bottleneck, we propose PAPO, a novel policy gradient algorithm that encourages the model to generate visually grounded reasoning without external supervision. Specifically, we introduce the Implicit Perception Loss in the form of a KL divergence term, which maximizes the difference between two probability distributions over the same rollout sequence, conditioned on either the original or corrupted visual input. Notably, PAPO does not rely on any additional data annotation, reward models, or stronger teacher models, and can therefore be seamlessly integrated into mainstream RLVR algorithms such as GRPO and DAPO. To further enhance the training stability of PAPO, we introduce the Double Entropy Loss, which effectively regularizes the new KL objective without compromising performance. Despite its simplicity, PAPO yields significant overall improvements of 4.4%-17.5% on diverse multimodal benchmarks. The improvements are more pronounced, approaching 8.0%-19.1%, on tasks with high vision dependency. We also observe a substantial reduction of 30.5% in perception errors, indicating improved perceptual capabilities with PAPO. Overall, PAPO offers a new perspective on advancing multimodal RLVR via the optimization objective, moving beyond rollout or reward design and pointing toward deeper integration of perception and reasoning.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Improving Visual Perception in Multimodal Reasoning Through Policy Optimization**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

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

- **Reinforcement Learning Frameworks for Multimodal Reasoning**
- **Training Paradigms and Optimization Algorithms**
- **Domain-Specific Applications**
- **Specialized Reasoning Tasks**
- **Foundational Studies and Benchmarking**

Complete Taxonomy Tree

- Improving Visual Perception in Multimodal Reasoning Through Policy Optimization Survey Taxonomy
- Reinforcement Learning Frameworks for Multimodal Reasoning
  - General-Purpose Multimodal RL Frameworks
  - Cross-Modal Reasoning via Policy Optimization (4 papers)
    - [2] Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning (Xing, 2025) View paper
    - [5] R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning (Zhao, 2025) View paper
    - [11] Fusing pre-trained language models with multimodal prompts through reinforcement learning (Young-Jae Yu, 2023) View paper
    - [27] Multimodal knowledge alignment with reinforcement learning (Yu, 2022) View paper
  - Unified Multimodal Reasoning Systems (4 papers)
    - [8] Vision-r1: Incentivizing reasoning capability in multimodal large language models (Huang Wenxuan, 2025) View paper
    - [19] Unified multimodal chain-of-thought reward model through reinforcement fine-tuning (Wang Yi-bin, 2025) View paper
    - [38] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models (Guanghao Zhou, 2025) View paper
    - [46] WeThink: Toward General-purpose Vision-Language Reasoning via Reinforcement Learning (Yang, 2025) View paper
  - Perception-Centric RL Methods
  - Visual Grounding and Perception Optimization ★ (4 papers)
    - [0] Perception-Aware Policy Optimization for Multimodal Reasoning (Anon et al., 2026) View paper
    - [22] Advancing Multimodal Reasoning Capabilities of Multimodal Large Language Models via Visual Perception Reward (T Xiao, 2025) View paper
    - [23] Perception in Reflection (Wei Yana, 2025) View paper
    - [29] Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models (Chen Yan, 2025) View paper
  - Iterative Perception Refinement (3 papers)
    - [12] VRAG-RL: Empower Vision-Perception-Based RAG for Visually Rich Information Understanding via Iterative Reasoning with Reinforcement Learning (Wang Qiu-chen, 2025) View paper
    - [31] VideoChat-R1.5: Visual Test-Time Scaling to Reinforce Multimodal Reasoning by Iterative Perception (Yan, 2025) View paper

    - [43] Agentic Jigsaw Interaction Learning for Enhancing Visual Perception and Reasoning in Vision-Language Models (Zeng Yu, 2025) View paper
  - Perception-Reasoning Decoupling (2 papers)
    - [13] Praxis-VLM: Vision-Grounded Decision Making via Text-Driven Reinforcement Learning (Hu Zhe, 2025) View paper
    - [17] Perceptual Decoupling for Scalable Multi-modal Reasoning via Reward-Optimized Captioning (Y Gou, 2025) View paper
  - Spatial and Temporal Reasoning Enhancement
  - Spatial Understanding via Self-Supervised RL (2 papers)
    - [3] Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning (Liu Yuhong, 2025) View paper
    - [35] PeRL: Permutation-Enhanced Reinforcement Learning for Interleaved Vision-Language Reasoning (Zhang, 2025) View paper
  - Video and Temporal Reasoning (3 papers)
    - [4] VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning (Li Xinhao, 2025) View paper
    - [14] Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning (Zhang Hao-ji, 2025) View paper
    - [37] VideoRFT: Incentivizing Video Reasoning Capability in MLLMs via Reinforced Fine-Tuning (Wang Qi, 2025) View paper
- Training Paradigms and Optimization Algorithms
  - SFT-RL Training Pipeline Analysis (2 papers)
  - [9] SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models (Tu, 2025) View paper
  - [15] Infi-MMR: Curriculum-based Unlocking Multimodal Reasoning via Phased Reinforcement Learning in Multimodal Small Language Models (Liu, 2025) View paper
  - Reward Design and Verification (3 papers)
  - [7] A vision-language-action-critic model for robotic real-world reinforcement learning (Zhai Shaopeng, 2025) View paper
  - [32] Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning of Vision Language Models (Tan Huajie, 2025) View paper
  - [40] GRPO-CARE: Consistency-Aware Reinforcement Learning for Multimodal Reasoning (Chen Yi, 2025) View paper
  - Efficiency and Scalability Optimization (3 papers)
  - [6] Visionthink: Smart and efficient vision language model via reinforcement learning (Yang, 2025) View paper
  - [26] Unleashing Perception-Time Scaling to Multimodal Reasoning Models (Li Yifan, 2025) View paper
  - [41] RLinf-VLA: A Unified and Efficient Framework for VLA+RL Training (Zang Hong-zhi, 2025) View paper
  - Visual Reasoning Paradigms (3 papers)
  - [20] Latent visual reasoning (Li, 2025) View paper
  - [34] Visual Planning: Let's Think Only with Images (Xu Yi, 2025) View paper
  - [47] Visual jigsaw post-training improves mllms (Wu, 2025) View paper
- Domain-Specific Applications
  - Medical and Healthcare Applications (3 papers)
  - [30] Med-R1: Reinforcement Learning for Generalizable Medical Reasoning in Vision-Language Models (Lai Yuxiang, 2025) View paper
  - [33] Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner (Zhang Wenchuan, 2025) View paper
  - [48] RL4Med-DDPO: Reinforcement Learning for Controlled Guidance Towards Diverse Medical Image Generation using Vision-Language Foundation Models (Parham Saremi, 2025) View paper
  - Document and Visual Information Understanding (2 papers)
  - [28] DocThinker: Explainable Multimodal Large Language Models with Rule-based Reinforcement Learning for Document Understanding (Yu Wenwen, 2025) View paper
  - [49] Geopqa: Bridging the visual perception gap in mllms for geometric reasoning (Guizhen Chen, 2025) View paper
  - Embodied AI and Robotics (4 papers)
  - [10] Embodied ai-enhanced vehicular networks: An integrated vision language models and reinforcement learning method (Ruichen Zhang, 2025) View paper
  - [18] Multi-Modal Attention Perception for Intelligent Vehicle Navigation Using Deep Reinforcement Learning (Zhenyu Li, 2025) View paper
  - [36] Towards multi-modal perception-based navigation: A deep reinforcement learning method (Xueqin Huang, 2021) View paper
  - [44] Vision-Based Mobile Robotics Obstacle Avoidance With Deep Reinforcement Learning (Wenzel, 2021) View paper
  - Quality Assessment and Evaluation (2 papers)
  - [1] Q-insight: Understanding image quality via visual reinforcement learning (Li Weiqi, 2025) View paper
  - [50] VQAThinker: Exploring Generalizable and Explainable Video Quality Assessment via Reinforcement Learning (Sun Wei, 2025) View paper
- Specialized Reasoning Tasks
  - Interactive and Cooperative Reasoning (1 papers)
  - [16] Learning cooperative visual dialog agents with deep reinforcement learning (Das, 2017) View paper
  - Tool-Augmented Reasoning (1 papers)
  - [39] Mindomni: Unleashing reasoning generation in vision language models with rgpo (Xiao, 2025) View paper
  - Multimodal Generation with Reasoning (1 papers)
  - [45] FaceCoT: A Benchmark Dataset for Face Anti-Spoofing with Chain-of-Thought Reasoning (Zhang Honglu, 2025) View paper
- Foundational Studies and Benchmarking
  - Pre-Training and Generalization Analysis (1 papers)
  - [25] Investigating Pre-Training Objectives for Generalization in Vision-Based Reinforcement Learning (Kimï¼□ Donghu, 2024) View paper
  - Interpretability and Explainability (1 papers)
  - [24] Rl-cam: Visual explanations for convolutional networks using reinforcement learning (Soumyendu Sarkar, 2023) View paper
  - Hierarchical and Multi-Level Reasoning (2 papers)
  - [21] Hierarchical adaptive value estimation for multi-modal visual reinforcement learning (Y Huang, 2023) View paper
  - [42] Sports-ACtrans Net: research on multimodal robotic sports action recognition driven via ST-GCN (Qi Lu, 2024) View paper

Narrative

Core task: Improving visual perception in multimodal reasoning through policy optimization. This field addresses how reinforcement learning can enhance the way multimodal models extract and utilize visual information when solving complex reasoning problems. The taxonomy reveals several major branches: Reinforcement Learning Frameworks for Multimodal Reasoning explores perception-centric methods and visual grounding techniques, often emphasizing how agents learn to attend to relevant visual features through reward signals (e.g., Visual Perception Reward[22], Perception in Reflection[23]). Training Paradigms and Optimization Algorithms investigates algorithmic innovations such as group relative policy optimization and reward modeling strategies that guide perceptual improvements. Domain-Specific Applications targets concrete settings like medical imaging (Med-R1[30], Patho-R1[33]) and embodied navigation (Embodied AI Vehicular[10]), while Specialized Reasoning Tasks focuses on challenges such as spatial reasoning (Spatial-SSRL[3]) and document understanding (DocThinker[28]). Foundational Studies and Benchmarking provides evaluation frameworks and baseline comparisons across these diverse problem settings.

A particularly active line of work centers on integrating perception optimization directly into the policy learning loop, where models learn not only what reasoning steps to take but also which visual cues to prioritize. Perception-Aware Policy[0] exemplifies this approach by coupling perceptual attention mechanisms with policy gradients, situating itself within the Visual Grounding and Perception Optimization cluster alongside neighbors like Visual Perception Reward[22] and Perception in Reflection[23]. While Perception in Reflection[23] emphasizes iterative refinement of visual interpretations through self-critique, Perception-Aware Policy[0] more tightly integrates perceptual decisions into the forward reasoning trajectory. Contrasting approaches such as Perception before Reasoning[29] advocate for decoupling perception from downstream reasoning, raising open questions about whether joint optimization or modular pipelines better balance sample efficiency and interpretability. Across branches, trade-offs emerge between end-to-end learning (VideoChat-R1[4], Vision-r1[8]) and structured decomposition (Perceptual Decoupling[17]), reflecting ongoing debates about how best to scale visual understanding in policy-driven multimodal systems.

Related Works in Same Category

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

1. Advancing Multimodal Reasoning Capabilities of Multimodal Large Language Models via Visual Perception Reward

Authors: T Xiao, X Xu, Z Huang, H Gao, Q Liu | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ Existing methods [17, 35, 19] typically apply RLVR with two main components, including reward functions and the Group Relative Policy Optimization (GRPO) algorithm. â¦

Relationship Analysis

Both papers belong to the Visual Grounding and Perception Optimization category, focusing on improving visual perception in multimodal reasoning through reward-based policy optimization. They share overlapping approaches by addressing perception bottlenecks in multimodal LLMs using reinforcement learning frameworks like GRPO, with both introducing perception-aware reward signals to enhance visual grounding. The key difference is that the original paper (PAPO) introduces an implicit perception loss via KL divergence between original and corrupted visual inputs without external supervision, while the candidate paper explicitly incorporates visual perception reward models to assess and optimize perception quality.

2. Perception in Reflection

Authors: Wei Yana, Zhao Liang, Yana Wei, Liang Zhao, Yu, et al. (26 authors total) | **Year/Venue**: 2025 • International Conference on Machine Learning | **URL**: View paper

Abstract

We present a perception in reflection paradigm designed to transcend the limitations of current large vision-language models (LVLMs), which are expected yet often fail to achieve perfect perception initially. Specifically, we propose Reflective Perception (RePer), a dual-model reflection mechanism that systematically alternates between policy and critic models, enables iterative refinement of visual perception. This framework is powered by Reflective Perceptual Learning (RPL), which reinforces i...

Relationship Analysis

Both papers belong to the Visual Grounding and Perception Optimization category, focusing on improving visual perception accuracy in multimodal reasoning through optimization methods. While the original paper (PAPO) introduces a policy gradient algorithm using KL divergence between original and corrupted visual inputs to encourage visually grounded reasoning during reinforcement learning, the candidate paper (RePer) proposes a dual-model reflection mechanism that iteratively refines visual perception through alternating policy and critic models with reflective perceptual learning. The key difference is that PAPO integrates perception optimization directly into the RL objective function, whereas RePer employs a separate reflection paradigm with explicit critic feedback for iterative refinement.

3. Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models

Authors: Chen Yan, Li Long, Yan Chen, Xi, Teng, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Reinforcement learning (RL) has proven highly effective in eliciting the reasoning capabilities of large language models (LLMs). Inspired by this success, recent studies have explored applying similar techniques to vision-language models (VLMs), aiming to enhance their reasoning performance. However, directly transplanting RL methods from LLMs to VLMs is suboptimal, as the tasks faced by VLMs are inherently more complex. Specifically, VLMs must first accurately perceive and understand visual inp...

Relationship Analysis

Both papers belong to the Visual Grounding and Perception Optimization category, focusing on improving visual perception in multimodal reasoning through reinforcement learning with reward signals derived from perception quality. The original paper (PAPO) introduces an implicit perception loss via KL divergence between original and corrupted visual inputs to encourage visually grounded reasoning within a single-stage RL framework, while the candidate paper proposes a two-stage RL approach that explicitly separates perception enhancement (stage 1: coarse and fine-grained visual understanding) from reasoning improvement (stage 2), using dataset-level sampling to address different capabilities sequentially.

Contributions Analysis

Overall novelty summary. The paper introduces PAPO, a policy gradient algorithm that addresses visual perception errors in multimodal reasoning by incorporating an Implicit Perception Loss based on KL divergence between original and corrupted visual inputs. It resides in the 'Visual Grounding and Perception Optimization' leaf alongside three sibling papers (Visual Perception Reward, Perception in Reflection, and one other). This leaf sits within the broader 'Perception-Centric RL Methods' branch, which contains only three leaves total, suggesting a moderately populated but not overcrowded research direction focused specifically on perception as the primary optimization target.

The taxonomy reveals neighboring work in adjacent leaves: 'Iterative Perception Refinement' (three papers) explores feedback loops for progressive visual understanding, while 'Perception-Reasoning Decoupling' (two papers) advocates separating perception from reasoning. The parent branch 'Reinforcement Learning Frameworks for Multimodal Reasoning' also contains 'General-Purpose Multimodal RL Frameworks' with eight papers addressing cross-modal reasoning without perception-specific focus. PAPO's approach of jointly optimizing perception within the policy loop contrasts with decoupled architectures and differs from iterative refinement methods by embedding perceptual grounding directly into the gradient signal rather than through separate feedback stages.

Among 30 candidates examined across three contributions, none were identified as clearly refuting the work. The PAPO algorithm with Implicit Perception Loss examined 10 candidates with zero refutable matches, as did the Double Entropy Loss regularization (10 candidates, zero refutable) and the integration of perception into RLVR objectives (10 candidates, zero refutable). This suggests that within the limited search scope, the specific mechanism of using KL divergence over corrupted visual inputs to drive perceptual grounding appears distinct from examined prior work, though the search scale means potentially relevant papers outside the top-30 semantic matches may exist.

Based on the limited literature search covering 30 candidates, the work appears to occupy a recognizable but not densely populated niche within perception-centric multimodal RL. The taxonomy structure shows related directions exist (iterative refinement, decoupled architectures), but the specific technical approach of implicit perception loss through input corruption was not matched in the examined candidates. A more exhaustive search beyond top-K semantic similarity might reveal closer precedents, particularly in adjacent leaves or in domain-specific applications not fully captured here.

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

Contribution 1: PAPO algorithm with Implicit Perception Loss

Description: The authors introduce PAPO (Perception-Aware Policy Optimization), a new policy gradient algorithm that integrates an Implicit Perception Loss term into RLVR frameworks. This loss maximizes KL divergence between probability distributions conditioned on original versus corrupted visual inputs, encouraging visually grounded responses without requiring additional annotations, reward models, or teacher models.

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. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning

URL: View paper

Brief Assessment

VL-Rethinker[62] focuses on addressing vanishing advantages in GRPO through selective sample replay and forced rethinking mechanisms to encourage self-reflection. It does not propose modifications to the core policy gradient optimization objective through perception-aware loss terms like PAPO's implicit perception loss.

2. Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning of Vision Language Models

URL: View paper

Brief Assessment

Reason-RFT[32] focuses on a two-stage framework (SFT followed by GRPO) for visual reasoning across counting, perception, and spatial tasks. It does not propose a novel policy gradient algorithm with implicit perception loss like PAPO, nor does it address perception-aware optimization through KL divergence between original and corrupted visual inputs.

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

URL: View paper

Brief Assessment

PeRL[35] focuses on multi-image positional reasoning through permutation-based exploration and rollout filtering, not on perception-aware policy optimization with implicit perception loss for single-image visual grounding.

4. Latent chain-of-thought for visual reasoning

URL: View paper

Brief Assessment

Latent Chain-of-Thought[63] focuses on amortized variational inference for posterior reasoning with diversity-seeking RL, not on perception-aware policy optimization with KL divergence between original and corrupted visual inputs.

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

URL: View paper

Brief Assessment

R1-VL[64] focuses on step-wise reasoning rewards (StepRAR and StepRVR) for multimodal reasoning, not on perception-aware policy optimization with implicit perception loss via KL divergence between original and corrupted visual inputs.

6. Grounded Reinforcement Learning for Visual Reasoning

URL: View paper

Brief Assessment

Grounded Visual Reasoning[65] focuses on spatially grounded reasoning with coordinate-based visual attention and multi-turn zooming mechanisms for web navigation and GUI tasks, not on policy gradient algorithms with implicit perception loss for general multimodal reasoning.

7. Reason-rft: Reinforcement fine-tuning for visual reasoning

URL: View paper

Brief Assessment

Reason-RFT Visual[61] uses standard GRPO for reinforcement learning without introducing novel policy gradient algorithms or perception-specific loss terms. It focuses on a two-stage framework (SFT followed by GRPO) for visual reasoning tasks, not on modifying the optimization objective with perception-aware components.

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

URL: View paper

Brief Assessment

Vision-r1[8] focuses on constructing multimodal CoT datasets and using standard GRPO with a progressive thinking suppression strategy, not on developing novel policy gradient algorithms with perception-aware loss terms.

9. Latent visual reasoning

URL: View paper

Brief Assessment

Latent Visual Reasoning[20] focuses on autoregressive reasoning in visual embedding space through latent state generation, not policy gradient optimization with KL divergence-based perception loss. The candidate adapts GRPO for reinforcement learning on latent reasoning, which is architecturally distinct from PAPO's implicit perception loss mechanism.

10. Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models

URL: View paper

Brief Assessment

Perception before Reasoning[29] uses a two-stage RL framework with dataset-level sampling for perception and reasoning stages, not a policy gradient algorithm with implicit perception loss based on KL divergence between original and corrupted visual inputs.

Contribution 2: Double Entropy Loss regularization

Description: The authors propose a Double Entropy Loss regularization technique that stabilizes training by preventing model collapse during optimization of the unbounded KL divergence objective. This regularization encourages low entropy in both the original and corrupted policy distributions.

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. Greedification operators for policy optimization: Investigating forward and reverse kl divergences

URL: View paper

Brief Assessment

Greedification Operators[54] focuses on KL divergence choices for policy optimization in RL, not on entropy regularization for stabilizing training. The paper does not address model collapse prevention or the specific double entropy loss technique proposed in the original work.

2. Statistical analysis of Inverse Entropy-regularized Reinforcement Learning

URL: View paper

Brief Assessment

Inverse Entropy Analysis[58] focuses on entropy regularization in inverse reinforcement learning for reward recovery from expert demonstrations, not on stabilizing forward RL training with KL divergence objectives as in the original paper.

3. Cautious policy programming: exploiting KL regularization for monotonic policy improvement in reinforcement learning

URL: View paper

Brief Assessment

Cautious Policy Programming[56] uses double entropy loss to regularize KL divergence in a different context (monotonic policy improvement in general RL), not specifically for stabilizing unbounded KL divergence objectives in multimodal reasoning as proposed in the original paper.

4. Entropy Regularization for Scalable, Safe and Robust Reinforcement Learning

URL: View paper

Brief Assessment

Entropy Regularization Scalable[60] focuses on using KL divergence between consecutive policies for general RL stability, not specifically for preventing model collapse during unbounded KL divergence optimization in multimodal reasoning tasks.

5. Your Policy Regularizer is Secretly an Adversary

URL: View paper

Brief Assessment

Policy Regularizer Adversary[59] focuses on theoretical analysis of entropy regularization as implicit adversarial robustness in standard RL settings, not on stabilizing unbounded KL divergence objectives in multimodal RLVR training.

6. Enforcing kl regularization in general tsallis entropy reinforcement learning via advantage learning

URL: View paper

Brief Assessment

Tsallis Entropy KL[57] focuses on implicit KL regularization in Tsallis entropy RL through advantage learning, not on double entropy loss for stabilizing unbounded KL divergence objectives in multimodal policy optimization.

7. APO: Enhancing Reasoning Ability of MLLMs via Asymmetric Policy Optimization

URL: View paper

Brief Assessment

APO[53] focuses on difficulty-adaptive divergence shaping and suboptimal trajectory complexity regularization to address KL penalty issues and overthinking in MLLMs. It does not propose a double entropy loss regularization technique for stabilizing unbounded KL divergence objectives.

8. Relative entropy regularized sample-efficient reinforcement learning with continuous actions

URL: View paper

Brief Assessment

Relative Entropy Regularized[52] focuses on relative entropy regularization in actor-critic frameworks for continuous action spaces, not on preventing model collapse through double entropy loss in policy optimization with KL divergence objectives.

9. The entropy mechanism of reinforcement learning for reasoning language models

URL: View paper

Brief Assessment

Entropy Mechanism Reasoning[51] focuses on preventing entropy collapse through covariance-based token selection (clip-cov and kl-cov methods), while the original paper's Double Entropy Loss regularizes both original and corrupted policy distributions in a multimodal context to stabilize KL divergence objectives. These are distinct technical approaches addressing different aspects of entropy management.

10. Kl-entropy-regularized rl with a generative model is minimax optimal

URL: View paper

Brief Assessment

KL-Entropy Minimax[55] focuses on theoretical sample complexity analysis of KL divergence and entropy regularization in value iteration for RL with generative models, not on stabilizing training through double entropy loss in policy optimization for multimodal reasoning.

Contribution 3: Integration of perception into RLVR optimization objective

Description: The authors present a new perspective on multimodal RLVR by modifying the core optimization objective itself rather than only adjusting data, rollout, or reward components. This represents the first work to explore deeper integration of perception-aware supervision signals beyond reward-level modifications in multimodal 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. VideoChat-R1.5: Visual Test-Time Scaling to Reinforce Multimodal Reasoning by Iterative Perception

URL: View paper

Brief Assessment

VideoChat-R1.5[31] focuses on iterative perception during inference through test-time scaling, not on modifying the core RLVR optimization objective during training. The candidate uses reinforcement learning for spatio-temporal supervision in an iterative perception mechanism, which is architecturally different from integrating perception-aware supervision signals into the policy gradient objective itself.

2. Perceptual Decoupling for Scalable Multi-modal Reasoning via Reward-Optimized Captioning

URL: View paper

Brief Assessment

Perceptual Decoupling[17] focuses on separating perception from reasoning through reward-optimized captioning, rather than integrating perception-aware supervision into the core optimization objective itself. The candidate's approach is fundamentally different from PAPO's direct modification of the policy gradient objective.

3. VQAThinker: Exploring Generalizable and Explainable Video Quality Assessment via Reinforcement Learning

URL: View paper

Brief Assessment

VQAThinker[50] focuses on video quality assessment using reinforcement learning with domain-specific rewards (bell-shaped regression, pairwise ranking, temporal consistency). It does not address multimodal reasoning with perception-aware supervision signals integrated into the core optimization objective as described in the original paper.

4. Visuriddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning

URL: View paper

Brief Assessment

VisuRiddles[67] focuses on supervised fine-tuning with synthetic data containing fine-grained perceptual descriptions for abstract visual reasoning, not on modifying reinforcement learning optimization objectives for multimodal reasoning.

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

URL: View paper

Brief Assessment

Machine Mental Imagery[71] focuses on augmenting VLM decoding with latent visual tokens for mental imagery-based reasoning, not on modifying RLVR optimization objectives with perception-aware supervision signals. The candidate uses RL as a final enhancement stage after distillation and text-only supervision, rather than integrating perception directly into the core RL objective itself.

6. Learning only with images: Visual reinforcement learning with reasoning, rendering, and visual feedback

URL: View paper

Brief Assessment

Learning Only Images[70] focuses on image-to-code generation tasks (charts and web interfaces) using a reasoning-rendering-visual-feedback loop with GRPO, rather than modifying the core RLVR optimization objective for general multimodal reasoning as in the original paper.

7. Breaking the sft plateau: Multimodal structured reinforcement learning for chart-to-code generation

URL: View paper

Brief Assessment

Breaking SFT Plateau[66] focuses on chart-to-code generation with structured rewards at textual and visual levels, not on modifying the core RLVR optimization objective to integrate perception-aware supervision signals for general multimodal reasoning.

8. Advancing Multimodal Reasoning: From Optimized Cold Start to Staged Reinforcement Learning

URL: View paper

Brief Assessment

Optimized Cold Start[68] focuses on cold start initialization strategies and staged training (text-only followed by multimodal RL), not on modifying the core optimization objective to integrate perception-aware supervision signals as the original paper does.

9. Unlocking Multimodal Mathematical Reasoning via Process Reward Model

URL: View paper

Brief Assessment

Process Reward Model[69] focuses on process-level reward modeling for multimodal mathematical reasoning, not on modifying the core RLVR optimization objective to integrate perception-aware supervision signals as the original paper does.

10. Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models

URL: View paper

Brief Assessment

Perception before Reasoning[29] addresses perception through a two-stage training framework with dataset-level sampling rather than modifying the core optimization objective itself with perception-aware supervision signals.

Appendix: Text Similarity Detection

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

The following **2 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. Cautious policy programming: exploiting KL regularization for monotonic policy improvement in reinforcement learning

Detected in: Contribution: contribution_2

⚠ **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.

2. Breaking the sft plateau: Multimodal structured reinforcement learning for chart-to-code generation

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] Perception-Aware Policy Optimization for Multimodal Reasoning View paper
- [1] Q-insight: Understanding image quality via visual reinforcement learning View paper
- [2] Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning View paper
- [3] Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning View paper
- [4] VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning View paper
- [5] R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning View paper
- [6] Visionthink: Smart and efficient vision language model via reinforcement learning View paper
- [7] A vision-language-action-critic model for robotic real-world reinforcement learning View paper
- [8] Vision-r1: Incentivizing reasoning capability in multimodal large language models View paper
- [9] SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models View paper
- [10] Embodied ai-enhanced vehicular networks: An integrated vision language models and reinforcement learning method View paper
- [11] Fusing pre-trained language models with multimodal prompts through reinforcement learning View paper
- [12] VRAG-RL: Empower Vision-Perception-Based RAG for Visually Rich Information Understanding via Iterative Reasoning with Reinforcement Learning View paper
- [13] Praxis-VLM: Vision-Grounded Decision Making via Text-Driven Reinforcement Learning View paper
- [14] Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning View paper
- [15] Infi-MMR: Curriculum-based Unlocking Multimodal Reasoning via Phased Reinforcement Learning in Multimodal Small Language Models View paper
- [16] Learning cooperative visual dialog agents with deep reinforcement learning View paper
- [17] Perceptual Decoupling for Scalable Multi-modal Reasoning via Reward-Optimized Captioning View paper
- [18] Multi-Modal Attention Perception for Intelligent Vehicle Navigation Using Deep Reinforcement Learning View paper
- [19] Unified multimodal chain-of-thought reward model through reinforcement fine-tuning View paper
- [20] Latent visual reasoning View paper
- [21] Hierarchical adaptive value estimation for multi-modal visual reinforcement learning View paper
- [22] Advancing Multimodal Reasoning Capabilities of Multimodal Large Language Models via Visual Perception Reward View paper
- [23] Perception in Reflection View paper
- [24] Rl-cam: Visual explanations for convolutional networks using reinforcement learning View paper
- [25] Investigating Pre-Training Objectives for Generalization in Vision-Based Reinforcement Learning View paper
- [26] Unleashing Perception-Time Scaling to Multimodal Reasoning Models View paper
- [27] Multimodal knowledge alignment with reinforcement learning View paper
- [28] DocThinker: Explainable Multimodal Large Language Models with Rule-based Reinforcement Learning for Document Understanding View paper
- [29] Perception before reasoning: Two-stage reinforcement learning for visual reasoning in vision-language models View paper
- [30] Med-R1: Reinforcement Learning for Generalizable Medical Reasoning in Vision-Language Models View paper
- [31] VideoChat-R1.5: Visual Test-Time Scaling to Reinforce Multimodal Reasoning by Iterative Perception View paper
- [32] Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning of Vision Language Models View paper
- [33] Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner View paper
- [34] Visual Planning: Let's Think Only with Images View paper
- [35] PeRL: Permutation-Enhanced Reinforcement Learning for Interleaved Vision-Language Reasoning View paper
- [36] Towards multi-modal perception-based navigation: A deep reinforcement learning method View paper
- [37] VideoRFT: Incentivizing Video Reasoning Capability in MLLMs via Reinforced Fine-Tuning View paper
- [38] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models View paper
- [39] Mindomni: Unleashing reasoning generation in vision language models with rgpo View paper
- [40] GRPO-CARE: Consistency-Aware Reinforcement Learning for Multimodal Reasoning View paper
- [41] RLinf-VLA: A Unified and Efficient Framework for VLA+RL Training View paper

- [42] Sports-ACtrans Net: research on multimodal robotic sports action recognition driven via ST-GCN View paper
- [43] Agentic Jigsaw Interaction Learning for Enhancing Visual Perception and Reasoning in Vision-Language Models View paper
- [44] Vision-Based Mobile Robotics Obstacle Avoidance With Deep Reinforcement Learning View paper
- [45] FaceCoT: A Benchmark Dataset for Face Anti-Spoofing with Chain-of-Thought Reasoning View paper
- [46] WeThink: Toward General-purpose Vision-Language Reasoning via Reinforcement Learning View paper
- [47] Visual jigsaw post-training improves mllms View paper
- [48] RL4Med-DDPO: Reinforcement Learning for Controlled Guidance Towards Diverse Medical Image Generation using Vision-Language Foundation Models View paper
- [49] Geopqa: Bridging the visual perception gap in mllms for geometric reasoning View paper
- [50] VQAThinker: Exploring Generalizable and Explainable Video Quality Assessment via Reinforcement Learning View paper
- [51] The entropy mechanism of reinforcement learning for reasoning language models View paper
- [52] Relative entropy regularized sample-efficient reinforcement learning with continuous actions View paper
- [53] APO: Enhancing Reasoning Ability of MLLMs via Asymmetric Policy Optimization View paper
- [54] Greedification operators for policy optimization: Investigating forward and reverse kl divergences View paper
- [55] Kl-entropy-regularized rl with a generative model is minimax optimal View paper
- [56] Cautious policy programming: exploiting KL regularization for monotonic policy improvement in reinforcement learning View paper
- [57] Enforcing kl regularization in general tsallis entropy reinforcement learning via advantage learning View paper
- [58] Statistical analysis of Inverse Entropy-regularized Reinforcement Learning View paper
- [59] Your Policy Regularizer is Secretly an Adversary View paper
- [60] Entropy Regularization for Scalable, Safe and Robust Reinforcement Learning View paper
- [61] Reason-rft: Reinforcement fine-tuning for visual reasoning View paper
- [62] Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning View paper
- [63] Latent chain-of-thought for visual reasoning View paper
- [64] R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization View paper
- [65] Grounded Reinforcement Learning for Visual Reasoning View paper
- [66] Breaking the sft plateau: Multimodal structured reinforcement learning for chart-to-code generation View paper
- [67] Visuriddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning View paper
- [68] Advancing Multimodal Reasoning: From Optimized Cold Start to Staged Reinforcement Learning View paper
- [69] Unlocking Multimodal Mathematical Reasoning via Process Reward Model View paper
- [70] Learning only with images: Visual reinforcement learning with reasoning, rendering, and visual feedback View paper
- [71] Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens View paper