Novelty Assessment Report

Paper: R1-Reward: Training Multimodal Reward Model Through Stable Reinforcement Learning

PDF URL: https://openreview.net/pdf?id=4Ewgw9M2xE

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

Multimodal Reward Models (MRMs) play a crucial role in enhancing the performance of Multimodal Large Language Models (MLLMs). While recent advancements have primarily focused on improving the model structure and training data of MRMs, there has been limited exploration into the effectiveness of long-term reasoning capabilities for reward modeling and how to activate these capabilities in MRMs. In this paper, we explore how Reinforcement Learning (RL) can be used to improve reward modeling. Specifically, we reformulate the reward modeling problem as a rule-based RL task. However, we observe that directly applying existing RL algorithms, such as Reinforce++, to reward modeling often leads to training instability or even collapse due to the inherent limitations of these algorithms. To address this issue, we propose the StableReinforce algorithm, which refines the training loss, advantage estimation strategy, and reward design of existing RL methods. These refinements result in more stable training dynamics and superior performance. To facilitate MRM training, we collect 200K preference data from diverse datasets. Our reward model, R1-Reward, trained using the StableReinforce algorithm on this dataset, significantly improves performance on multimodal reward modeling benchmarks. Compared to previous SOTA models, R1-Reward achieves a 8.4% improvement on the VL Reward-Bench and a 14.3% improvement on the Multimodal Reward Bench. Moreover, with more inference compute, R1-Reward's performance is further enhanced, highlighting the potential of RL algorithms in optimizing MRMs.

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: **Training Multimodal Reward Models Through Reinforcement Learning**

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:

- **Reward Model Architecture and Training Paradigms**
- **Reinforcement Learning Algorithms and Optimization Methods**
- **Application Domains and Task-Specific Implementations**
- **Evaluation and Benchmarking**
- **Safety and Security**

Complete Taxonomy Tree

- Training Multimodal Reward Models Through Reinforcement Learning Survey Taxonomy
- Reward Model Architecture and Training Paradigms
    - Unified and Multi-Task Reward Models (4 papers)
    - [2] Unified reward model for multimodal understanding and generation (Wang Yi-bin, 2025) View paper
    - [17] BaseReward: A Strong Baseline for Multimodal Reward Model (Zhang Yi Fan, 2025) View paper
    - [19] Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model (Zang, 2025) View paper
    - [24] Skywork-vl reward: An effective reward model for multimodal understanding and reasoning (Wang Xiao-kun, 2025) View paper
    - Chain-of-Thought and Reasoning-Enhanced Reward Models (2 papers)
    - [1] Unified multimodal chain-of-thought reward model through reinforcement fine-tuning (Wang Yi-bin, 2025) View paper
    - [18] Vrprm: Process reward modeling via visual reasoning (Chen Xinquan, 2025) View paper
    - Few-Shot and Activation-Based Reward Models (1 papers)
    - [13] Activation Reward Models for Few-Shot Model Alignment (Tianning Chai, 2025) View paper
    - Generative and Principle-Based Reward Models (1 papers)
    - [22] Generative RLHF-V: Learning Principles from Multi-modal Human Preference (Zhou JiaYi, 2025) View paper
- Reinforcement Learning Algorithms and Optimization Methods
    - Policy Optimization Methods ★ (6 papers)
    - [0] R1-Reward: Training Multimodal Reward Model Through Stable Reinforcement Learning (Anon et al., 2026) View paper
    - [5] Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning (Xing, 2025) View paper
    - [8] R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization (Zhang Jingyi, 2025) View paper
    - [12] Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning (Wang Peiyu, 2025) View paper
    - [15] Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models (Xu, 2025) View paper
    - [27] GLM-4.1 V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning (V. Team, 2025) View paper
    - Hybrid and Value-Based RL Approaches (1 papers)
    - [50] MORAL: A Multimodal Reinforcement Learning Framework for Decision Making in Autonomous Laboratories (Kumar Sathish A.P., 2025) View paper

  - Reinforcement Learning from Human Feedback (4 papers)
  - [4] Tuning large multimodal models for videos using reinforcement learning from ai feedback (Ahn Daechul, 2024) View paper
  - [11] Improving multimodal interactive agents with reinforcement learning from human feedback (Abramson, 2022) View paper
  - [26] Direct preference optimization of video large multimodal models from language model reward (Zhang, 2025) View paper
  - [36] Safe RLHF-V: Safe Reinforcement Learning from Multi-modal Human Feedback (Ji, 2025) View paper
  - Self-Evolving and Automated Reward Systems (3 papers)
  - [31] EvoLMM: Self-Evolving Large Multimodal Models with Continuous Rewards (Omkar Thawakar, 2025) View paper
  - [33] Evolutionary reward design and optimization with multimodal large language models (Ali Narin, 2024) View paper
  - [43] C2-evo: Co-evolving multimodal data and model for self-improving reasoning (Chen Xiu-wei, 2025) View paper
  - Curriculum and Phased RL Training (1 papers)
  - [30] Infi-MMR: Curriculum-based Unlocking Multimodal Reasoning via Phased Reinforcement Learning in Multimodal Small Language Models (Liu, 2025) View paper
- Application Domains and Task-Specific Implementations
  - Vision-Language Understanding and Reasoning (7 papers)
  - [3] R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning (Zhao, 2025) View paper
  - [6] Fusing pre-trained language models with multimodal prompts through reinforcement learning (Young-Jae Yu, 2023) View paper
  - [10] Multimodal knowledge alignment with reinforcement learning (Yu, 2022) View paper
  - [20] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models (Guanghao Zhou, 2025) View paper
  - [23] Advancing Multimodal Reasoning Capabilities of Multimodal Large Language Models via Visual Perception Reward (T Xiao, 2025) View paper
  - [28] Reinforcement Fine-Tuning Powers Reasoning Capability of Multimodal Large Language Models (Sun Haoyuan, 2025) View paper
  - [39] Reinforcement Learning Tuning for VideoLLMs: Reward Design and Data Efficiency (Li, 2025) View paper
  - Visual Generation and Creative Applications (2 papers)
  - [40] Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation (XU Jiazheng, 2024) View paper
  - [49] Rewarddance: Reward scaling in visual generation (Wu Jie, 2025) View paper
  - Document Understanding and Structured Analysis (2 papers)
  - [14] SAM-R1: Leveraging SAM for Reward Feedback in Multimodal Segmentation via Reinforcement Learning (Huang, 2025) View paper
  - [34] Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding (Yu Wenwen, 2025) View paper
  - Dialogue and Interactive Systems (2 papers)
  - [45] Aligning Dialogue Agents with Global Feedback via Large Language Model Reward Decomposition (Lee Dong Won, 2025) View paper
  - [46] Global reward to local rewards: Multimodal-guided decomposition for improving dialogue agents (Dong Won Lee, 2024) View paper
  - Specialized Domain Applications
  - Emotion Recognition and Mental Health (1 papers)
    - [35] A reinforcement learning-based approach for promoting mental health using multimodal emotion recognition (Dumidu Kasun Rajakaruna, 2024) View paper
  - Time Series and Forecasting (2 papers)
    - [16] Text reinforcement for multimodal time series forecasting (Su Chen, 2025) View paper
    - [37] TimeMaster: Training Time-Series Multimodal LLMs to Reason via Reinforcement Learning (Zhang Junru, 2025) View paper
  - Robotics and Autonomous Systems (4 papers)
    - [38] RDDRL: a recurrent deduction deep reinforcement learning model for multimodal vision-robot navigation (Zhenyu Li, 2023) View paper
    - [44] An adaptive reinforcement learning-based multimodal data fusion framework for humanâ□□robot confrontation gaming (Wen Qi, 2023) View paper
    - [47] Multimodal fusion for autonomous navigation via deep reinforcement learning with sparse rewards and hindsight experience replay (Wendong Xiao, 2023) View paper
    - [48] Prefmmt: Modeling human preferences in preference-based reinforcement learning with multimodal transformers (Zhao Dezhong, 2024) View paper
  - Medical Imaging and Registration (2 papers)
    - [7] End-to-end multimodal image registration via reinforcement learning (Jing Hu, 2021) View paper
    - [32] Multimodal image registration with deep context reinforcement learning (Kai Ma, 2017) View paper
  - Translation and Cross-Modal Retrieval (2 papers)
    - [25] End-to-End Optimization for Multimodal Retrieval-Augmented Generation via Reward Backpropagation (Zhiyuan Fan, 2025) View paper
    - [29] Multimodal machine translation with reinforcement learning (Qian Xin, 2018) View paper
  - Optimization and Decision Making (2 papers)
    - [9] Rlemmo: Evolutionary multimodal optimization assisted by deep reinforcement learning (Hongqiao Lian, 2024) View paper
    - [42] Reinforcement learning-based multimodal model for the stock investment portfolio management task (Sha Du, 2024) View paper
- Evaluation and Benchmarking (1 papers)
  - [41] VideoRewardBench: Comprehensive Evaluation of Multimodal Reward Models for Video Understanding (Zhang Zhihong, 2025) View paper
- Safety and Security (1 papers)
  - [21] Secure tug-of-war (sectow): Iterative defense-attack training with reinforcement learning for multimodal model security (Muzhi Dai, 2025) View paper

Narrative

Core task: training multimodal reward models through reinforcement learning. The field has evolved into a structured landscape organized around five major branches. Reward Model Architecture and Training Paradigms explores how to design and train reward functions that can handle vision, language, and other modalities, with works like Unified Multimodal CoT Reward[1] and Unified Reward

Model[2] proposing architectures that unify cross-modal signals. Reinforcement Learning Algorithms and Optimization Methods focuses on policy optimization techniques and algorithmic innovations, including approaches such as R1-omni[3] and Skywork R1v2[12] that refine training dynamics. Application Domains and Task-Specific Implementations addresses concrete use cases ranging from video understanding (Tuning Video Models RLAIF[4]) to document reasoning (DocThinker[34]) and even medical image registration (Multimodal Image Registration RL[7]). Evaluation and Benchmarking provides standardized testbeds like VideoRewardBench[41] to measure progress, while Safety and Security tackles alignment and robustness concerns exemplified by Safe RLHF-V[36].

Within the policy optimization branch, a particularly active line of work centers on integrating reasoning traces and iterative refinement into multimodal reward learning. R1-Reward[0] sits squarely in this cluster, emphasizing policy-level optimization methods that leverage reinforcement signals to improve both reasoning quality and multimodal alignment. Nearby efforts such as EchoInk-R1[5] and R1-VL[8] share a similar focus on refining vision-language models through RL-driven reward shaping, though they differ in architectural choices and the granularity of feedback. A key trade-off across these methods involves balancing sample efficiency against the richness of multimodal supervision: some approaches rely on dense process-level rewards (VRPRM[18]), while others adopt sparser outcome-based signals (Mixed-R1[15]). Open questions remain around scaling these techniques to longer reasoning chains and ensuring that learned rewards generalize robustly across diverse visual and linguistic contexts.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning

Authors: Xing, Zhenghao, Hu Xiaowei, Zheng Xing, Fu, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Multimodal large language models (MLLMs) have advanced perception across text, vision, and audio, yet they often struggle with structured cross-modal reasoning, particularly when integrating audio and visual signals. We introduce EchoInk-R1, a reinforcement learning framework that enhances such reasoning in MLLMs. Built upon the Qwen2.5-Omni-7B foundation and optimized with Group Relative Policy Optimization (GRPO), EchoInk-R1 tackles multiple-choice question answering over synchronized audio-im...

Relationship Analysis

Both papers belong to the Policy Optimization Methods category, employing reinforcement learning algorithms (StableReinforce vs. GRPO) to train multimodal reward models or reasoning systems. They overlap in using RL policy gradient methods for multimodal tasks and addressing training stability challenges through refined loss functions and advantage estimation. However, R1-Reward focuses on training a multimodal reward model for preference ranking across diverse visual tasks using a novel StableReinforce algorithm with consistency rewards, while EchoInk-R1 specifically targets audio-visual reasoning in multiple-choice QA using GRPO on synchronized audio-image pairs, representing different application domains within the same methodological framework.

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

Authors: Zhang Jingyi, Huang Jiaxing, Jingyi Zhang, Jiaxing Huang, Liu Shunyu, et al. (15 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 frame...

Relationship Analysis

Both papers belong to the Policy Optimization Methods category, focusing on reinforcement learning approaches for training multimodal reward models and MLLMs. They overlap in using policy gradient methods (the original paper proposes StableReinforce while the candidate proposes StepGRPO) to improve multimodal reasoning through RL, and both address training stability issues in the RL process. The key difference is that the original paper focuses on training a reward model to evaluate response quality using consistency rewards and advantage filtering, while the candidate paper focuses on training the MLLM itself using step-wise reasoning rewards (StepRAR and StepRVR) to improve its reasoning capability directly.

3. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning

Authors: Wang Peiyu, Wei, Yichen, Chris, Peng Yi, et al. (28 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We present Skywork R1V2, a next-generation multimodal reasoning model and a major leap forward from its predecessor, Skywork R1V. At its core, R1V2 introduces a hybrid reinforcement learning paradigm that jointly leverages the Mixed Preference Optimization (MPO) and the Group Relative Policy Optimization (GRPO), which harmonizes reward-model guidance with rule-based strategies, thereby addressing the long-standing challenge of balancing sophisticated reasoning capabilities with broad generalizat...

Relationship Analysis

Both papers belong to the Policy Optimization Methods category, employing policy gradient-based RL algorithms (GRPO, PPO variants) to train multimodal reward models or reasoning systems. They overlap in using reinforcement learning for multimodal tasks, addressing training stability issues, and incorporating reward design strategies. However, R1-Reward focuses specifically on training a multimodal reward model through StableReinforce (a refined PPO variant with pre-CLIP and advantage filtering), while Skywork R1V2 trains a multimodal reasoning model using a hybrid approach combining MPO and GRPO with Selective Sample Buffer, targeting end-to-end reasoning rather than reward modeling.

4. Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models

Authors: Xu, Shilin, Li Yanwei, Shilin Xu, Yang Rui, et al. (26 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Recent works on large language models (LLMs) have successfully demonstrated the emergence of reasoning capabilities via reinforcement learning (RL). Although recent efforts leverage group relative policy optimization (GRPO) for MLLMs post-training, they constantly explore one specific aspect, such as grounding tasks, math problems, or chart analysis. There are no works that can leverage multi-source MLLM tasks for stable reinforcement learning. In this work, we present a unified perspective to s...

Relationship Analysis

Both papers belong to the Policy Optimization Methods category, employing reinforcement learning with policy gradient approaches (GRPO, PPO variants) to train multimodal reward models. They overlap in using RL frameworks to improve reward modeling through stable training algorithms and mixed reward functions for multimodal tasks. The key difference is that R1-Reward focuses on

StableReinforce algorithm refinements (pre-CLIP, advantage filtering) with consistency rewards for reasoning-result alignment, while Mixed-R1 emphasizes a unified mixed reward design (BMAS, IoU, chart rewards) across diverse task types with a Mixed-45K dataset for broader MLLM post-training.

5. GLM-4.1 V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning

Authors: V. Team, Hong Wenyi, Wenyi Hong, Yu, Wenmeng, et al. (168 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We present GLM-4.1V-Thinking and GLM-4.5V, a family of vision-language models (VLMs) designed to advance general-purpose multimodal understanding and reasoning. In this report, we share our key findings in the development of the reasoning-centric training framework. We first develop a capable vision foundation model with significant potential through large-scale pre-training, which arguably sets the upper bound for the final performance. We then propose Reinforcement Learning with Curriculum Sam...

Relationship Analysis

Both papers belong to the Policy Optimization Methods category, employing reinforcement learning with policy gradient approaches to train multimodal reward models or vision-language models. They overlap in using RL algorithms (R1-Reward uses StableReinforce, a refined PPO/Reinforce++ variant; GLM-4.1V-Thinking uses RLCS with curriculum sampling) to enhance multimodal reasoning through long-form chain-of-thought generation. The key difference is that R1-Reward focuses specifically on training reward models for preference ranking with novel stability improvements (pre-CLIP, advantage filtering, consistency rewards), while GLM-4.1V-Thinking develops a general-purpose vision-language model with curriculum-based RL across diverse multimodal tasks including STEM, video understanding, and GUI agents.

Contributions Analysis

Overall novelty summary. The paper proposes StableReinforce, a refinement of policy gradient methods for training multimodal reward models, and introduces R1-Reward trained on 200K preference pairs. It sits within the Policy Optimization Methods leaf, which contains six papers exploring policy gradient variants (GRPO, PPO, custom optimizers) for multimodal RL. This is a moderately populated research direction within the broader Reinforcement Learning Algorithms and Optimization Methods branch, indicating active but not overcrowded exploration of policy-based training for reward models.

The taxonomy reveals closely related work in neighboring leaves: Hybrid and Value-Based RL Approaches explores alternative optimization paradigms, while Reinforcement Learning from Human Feedback focuses on preference-based alignment (though the paper's rule-based formulation differs). The Chain-of-Thought and Reasoning-Enhanced Reward Models branch addresses explicit reasoning mechanisms, which connects to this work's emphasis on long-term reasoning capabilities. The sibling papers in Policy Optimization Methods share the core focus on policy gradients but vary in architectural choices and feedback granularity, suggesting this leaf represents a coherent but diverse research cluster.

Among 29 candidates examined, the StableReinforce algorithm contribution showed no clear refutation across 10 candidates, suggesting potential novelty in its specific refinements to loss, advantage estimation, and reward design. The rule-based RL reformulation similarly found no refuting work among 10 candidates. However, the R1-Reward-200K dataset contribution encountered one refutable candidate among nine examined, indicating some overlap in progressive difficulty training strategies for preference data. The limited search scope means these findings reflect top-K semantic matches rather than exhaustive coverage.

Based on the 29-candidate search, the algorithmic contributions appear more distinctive than the dataset contribution within the examined literature. The taxonomy context suggests the work occupies a moderately explored niche, with sibling papers pursuing related but distinct policy optimization approaches. A broader search beyond top-K semantic similarity might reveal additional relevant work, particularly in the reasoning-enhanced reward models direction or in general RL stability techniques adapted to multimodal settings.

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

Contribution 1: StableReinforce algorithm for stable reward model training

Description: The authors introduce StableReinforce, a novel reinforcement learning algorithm that addresses training instability in reward modeling through three key refinements: pre-clipping to prevent numerical overflow, advantage filtering using the 3-sigma rule to handle outliers, and a consistency reward mechanism that ensures alignment between reasoning processes and final outputs.

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. Avatar: Reinforcement learning to see, hear, and reason over video

URL: View paper

Brief Assessment

Avatar[69] focuses on multimodal video reasoning with off-policy RL and temporal advantage shaping, not on reward model training stability. The candidate addresses vanishing advantages in GRPO for video understanding tasks, while the original paper tackles training instability in reward modeling through pre-clipping, advantage filtering, and consistency rewards—fundamentally different problem domains and solutions.

2. Trust Region Reward Optimization and Proximal Inverse Reward Optimization Algorithm

URL: View paper

Brief Assessment

Trust Region Reward[75] focuses on inverse reinforcement learning (IRL) for reward function recovery from expert demonstrations, not on training reward models for evaluating multimodal responses. The technical domains and problem formulations are fundamentally different.

3. An improved deep reinforcement learning algorithm for path planning in unmanned driving

URL: View paper

Brief Assessment

Path Planning Deep RL[68] focuses on path planning for autonomous vehicles using deep reinforcement learning with reward function enhancements for navigation tasks. It does not address reward model training, multimodal learning, or the specific training instability challenges (pre-clipping, advantage filtering, consistency rewards) that StableReinforce targets for reward modeling in language models.

4. Hybrid group relative policy optimization: A multi-sample approach to enhancing policy optimization

URL: View paper

Brief Assessment

Hybrid Group RPO[71] focuses on policy optimization for general RL environments by balancing empirical multi-sample action evaluation with value function estimation. It does not address reward model training, training instability specific to reward modeling tasks, or the three key refinements (pre-clipping, advantage filtering via 3-sigma rule, consistency reward) that characterize StableReinforce.

5. Boosting policy learning in reinforcement learning via adaptive intrinsic reward regulation

URL: View paper

Brief Assessment

Adaptive Intrinsic Reward[70] focuses on regulating intrinsic reward coefficients in sparse reward environments to balance exploration-exploitation, not on stabilizing reward model training through loss refinement and advantage estimation techniques like pre-clipping and advantage filtering.

6. Quantile Advantage Estimation for Entropy-Safe Reasoning

URL: View paper

Brief Assessment

Quantile Advantage Estimation[76] addresses training instability in RL for LLM reasoning tasks through quantile-based advantage estimation, not reward model training. The candidate focuses on entropy regulation in policy optimization for mathematical reasoning, while the original paper targets multimodal reward modeling with consistency rewards and pre-clipping mechanisms. These are distinct application domains with different technical approaches.

7. Continuous reinforcement learning via advantage value difference reward shaping: A proximal policy optimization perspective

URL: View paper

Brief Assessment

Advantage Value Difference[67] focuses on exploration reward shaping in continuous RL environments, not on stabilizing reward model training through loss refinement and advantage estimation techniques like pre-clipping and advantage filtering.

8. Leftover lunch: Advantage-based offline reinforcement learning for language models

URL: View paper

Brief Assessment

Leftover Lunch[73] focuses on offline RL for language model alignment using advantage-based filtering of pre-existing data, not on stabilizing reward model training through loss refinement and advantage estimation as in the original paper's StableReinforce.

9. Enhancing stability and performance in mobile robot path planning with pmr-dueling dqn algorithm

URL: View paper

Brief Assessment

PMR-Dueling DQN[72] focuses on mobile robot path planning using dueling DQN architecture with prioritized experience replay and shaped rewards for navigation tasks. This is fundamentally different from StableReinforce, which addresses training instability in reward modeling for multimodal large language models through pre-clipping, advantage filtering, and consistency rewards. The candidate paper does not discuss reward model training, multimodal systems, or the specific stability challenges that StableReinforce addresses.

10. From Sparse to Dense: Toddler-inspired Reward Transition in Goal-Oriented Reinforcement Learning

URL: View paper

Brief Assessment

Toddler-Inspired Reward Transition[74] focuses on curriculum learning through sparse-to-dense reward transitions in goal-oriented RL tasks, not on reward model training stability or advantage estimation refinements for multimodal reward models.

Contribution 2: Reformulation of reward modeling as rule-based RL task

Description: The paper reformulates multimodal reward modeling as a reinforcement learning problem where the policy decides which of two answers is better, with rewards based on consistency with ground truth. This enables the application of RL techniques to activate long-term reasoning capabilities in reward 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. Decisionnce: Embodied multimodal representations via implicit preference learning

URL: View paper

Brief Assessment

DecisionNCE[53] focuses on multimodal representation learning through implicit preference learning in embodied AI contexts, not on reformulating reward modeling as an RL task for training reward models themselves.

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

URL: View paper

Brief Assessment

GraphFusion-HRL[55] focuses on hierarchical reinforcement learning for recommender systems with cross-modal reward signals, not on reformulating reward modeling for multimodal preference learning as a rule-based RL task.

3. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model

URL: View paper

Brief Assessment

InternLM-XComposer Reward[19] does not reformulate reward modeling as an RL task. Instead, it trains a discriminative reward model using standard pairwise preference loss, then applies this model within existing RL frameworks (PPO) for policy training—a conventional approach rather than reformulating reward modeling itself as RL.

4. Personalizing reinforcement learning from human feedback with variational preference learning

URL: View paper

Brief Assessment

Variational Preference Learning[51] focuses on modeling diverse user preferences through latent variable inference for personalized reward models, not on reformulating reward modeling as a rule-based RL task with long-term reasoning capabilities.

5. Tuning large multimodal models for videos using reinforcement learning from ai feedback

URL: View paper

Brief Assessment

Tuning Video Models RLAIF[4] applies RLAIF to video-text alignment tasks, not to reformulating reward modeling itself as an RL problem. The candidate focuses on using RL to align multimodal models with human preferences, whereas the original contribution reformulates the reward modeling task itself as a rule-based RL problem where the policy decides which answer is better.

6. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning

URL: View paper

Brief Assessment

Unified Multimodal CoT Reward[1] focuses on incorporating long chain-of-thought reasoning into reward models through GRPO-based reinforcement fine-tuning, rather than reformulating reward modeling as a rule-based RL task where the policy decides between two answers with rewards based on ground truth consistency.

7. Aligning Large Vision-Language Models by Deep Reinforcement Learning and Direct Preference Optimization

URL: View paper

Brief Assessment

Aligning Vision-Language Models[54] focuses on fine-tuning LVLMs using DRL and DPO for general alignment tasks, not specifically on reformulating reward modeling itself as an RL problem. The candidate discusses using rule-based rewards as one of several reward sources for LVLM fine-tuning, but does not present reward modeling as the primary RL task being solved.

8. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning

URL: View paper

Brief Assessment

Skywork R1v2[12] focuses on training multimodal reasoning models using hybrid RL (MPO+GRPO) with rule-based and model-based rewards, not on reformulating reward modeling itself as an RL task. The candidate applies RL to train vision-language models for reasoning, whereas the original reformulates the reward modeling problem as an RL task where the policy learns to judge preferences.

9. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning

URL: View paper

Brief Assessment

VL-Cogito[52] focuses on progressive curriculum RL for multimodal reasoning tasks (mathematics, science, logic), not on reformulating reward modeling itself as an RL problem. The candidate applies RL to improve reasoning capabilities, whereas the original reformulates the reward modeling task structure.

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

URL: View paper

Brief Assessment

VR-Thinker[56] focuses on video reward models with visual reasoning operations and thinking-with-image framework, not on reformulating reward modeling as an RL task. The candidate uses GRPO for training but does not reformulate the core reward modeling problem as an RL task where the policy decides preferences.

Contribution 3: R1-Reward-200K dataset with progressive difficulty training strategy

Description: The authors construct a 200K preference dataset from multiple sources and implement a progressive difficulty training strategy. They use GPT-4o to generate thinking processes for cold-start data and select challenging samples (requiring multiple attempts) for RL training, improving data efficiency.

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

1. Openprm: Building open-domain process-based reward models with preference trees

URL: View paper

Brief Assessment

OpenPRM[60] focuses on building process-based reward models for open-domain tasks using preference trees constructed from parallel sampling, not on progressive difficulty training strategies for multimodal reward models. The candidate addresses a different problem domain (open-domain instruction-following vs. multimodal reward modeling) with a different technical approach (preference trees with back-propagation vs. progressive difficulty based on GPT-4o sampling attempts).

2. 2D-Curri-DPO: Two-Dimensional Curriculum Learning for Direct Preference Optimization

URL: View paper

Brief Assessment

2D-Curri-DPO[61] focuses on direct preference optimization with a two-dimensional curriculum based on prompt complexity and pairwise distinguishability for language model alignment. The original paper constructs a multimodal reward model dataset with progressive difficulty based on GPT-4o sampling attempts for reward modeling tasks, which is a fundamentally different application domain and methodology.

3. Enhancing alignment using curriculum learning & ranked preferences

URL: View paper

Brief Assessment

Curriculum Ranked Preferences[58] focuses on ordering preference pairs for DPO training in text-based LLMs, not on multimodal reward model training or RL-based approaches for reward modeling.

4. De novo molecular design enabled by direct preference optimization and curriculum learning

URL: View paper

Brief Assessment

Molecular Design DPO[62] focuses on molecular generation using curriculum learning to progressively increase task difficulty in chemical space exploration, not on reward model training for multimodal tasks. The domains and applications are fundamentally different.

5. Curry-dpo: Enhancing alignment using curriculum learning & ranked preferences

URL: View paper

Brief Assessment

Curry-DPO[65] focuses on curriculum learning for DPO alignment using multiple preference pairs ranked by difficulty, not on reward model training or RL-based reward modeling which is the core of the original contribution.

6. Curriculum-RLAIF: Curriculum Alignment with Reinforcement Learning from AI Feedback

URL: View paper

Prior Art Analysis

Curriculum-RLAIF[59] demonstrates that progressive difficulty training strategies using preference datasets were established prior to the original paper's work. The candidate paper constructs a 200K preference dataset from multiple sources and implements a curriculum that progressively incorporates preference pairs of increasing difficulty for reward model training. Both papers use similar data construction approaches (sampling from multiple datasets to create ~200K samples) and employ progressive difficulty strategies, though with different specific implementations. The candidate's approach of gradually transitioning from easy contrastive pairs through bridging pairs to difficult random pairs parallels the original's strategy of using GPT-4o attempts as difficulty indicators and selecting challenging samples for RL training.

Evidence

Evidence 1 - **Rationale**: Both papers demonstrate that their progressive difficulty training strategies using ~200K preference datasets lead to improved performance, showing that the approach of using progressive difficulty with preference datasets for reward model training was established in prior work. - **Original**: we collect 200k preference data from diverse datasets. our reward model, r1-reward, trained using the stablereinforce algorithm on this dataset, significantly improves performance on multimodal reward modeling benchmarks. - **Candidate**: our experimental results on three widely used alignment datasets show that curriculum-rlaif substantially improves alignment performance over conventional rlaif methods that overlook data quality, and surpasses strong baselines by a large margin without incurring additional inference costs.

7. Curriculum direct preference optimization for diffusion and consistency models

URL: View paper

Brief Assessment

Curriculum DPO Diffusion[64] focuses on organizing image pairs by preference score differences for text-to-image generation tasks, not on constructing preference datasets with GPT-4o annotations or selecting challenging samples based on multiple attempts for reward model training in multimodal contexts.

8. Rovrm: A robust visual reward model optimized via auxiliary textual preference data

URL: View paper

Brief Assessment

RoVRM[63] focuses on three-phase progressive training to bridge modality gaps between textual and visual preference data, not on progressive difficulty selection based on GPT-4o sampling attempts. The candidate's approach addresses data scarcity through cross-modal transfer rather than difficulty-based curriculum learning.

9. Dast: Difficulty-adaptive slow-thinking for large reasoning models

URL: View paper

Brief Assessment

DAST[57] focuses on difficulty-adaptive reasoning length control during inference, not on constructing preference datasets or progressive difficulty training for reward models. The candidate addresses a different problem (overthinking mitigation) using a token length budget metric rather than preference data construction.

Appendix: Text Similarity Detection

Textual similarity detection checked 34 papers and found 1 similarity segment(s) across 1 paper(s).

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

1. Quantile Advantage Estimation for Entropy-Safe Reasoning

Detected in: Contribution: contribution_1

⚠ **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] R1-Reward: Training Multimodal Reward Model Through Stable Reinforcement Learning View paper
- [1] Unified multimodal chain-of-thought reward model through reinforcement fine-tuning View paper
- [2] Unified reward model for multimodal understanding and generation View paper
- [3] R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning View paper
- [4] Tuning large multimodal models for videos using reinforcement learning from ai feedback View paper
- [5] Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning View paper
- [6] Fusing pre-trained language models with multimodal prompts through reinforcement learning View paper
- [7] End-to-end multimodal image registration via reinforcement learning View paper
- [8] R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization View paper
- [9] Rlemmo: Evolutionary multimodal optimization assisted by deep reinforcement learning View paper
- [10] Multimodal knowledge alignment with reinforcement learning View paper
- [11] Improving multimodal interactive agents with reinforcement learning from human feedback View paper
- [12] Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning View paper

- [13] Activation Reward Models for Few-Shot Model Alignment View paper
- [14] SAM-R1: Leveraging SAM for Reward Feedback in Multimodal Segmentation via Reinforcement Learning View paper
- [15] Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models View paper
- [16] Text reinforcement for multimodal time series forecasting View paper
- [17] BaseReward: A Strong Baseline for Multimodal Reward Model View paper
- [18] Vrprm: Process reward modeling via visual reasoning View paper
- [19] Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model View paper
- [20] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models View paper
- [21] Secure tug-of-war (sectow): Iterative defense-attack training with reinforcement learning for multimodal model security View paper
- [22] Generative RLHF-V: Learning Principles from Multi-modal Human Preference View paper
- [23] Advancing Multimodal Reasoning Capabilities of Multimodal Large Language Models via Visual Perception Reward View paper
- [24] Skywork-vl reward: An effective reward model for multimodal understanding and reasoning View paper
- [25] End-to-End Optimization for Multimodal Retrieval-Augmented Generation via Reward Backpropagation View paper
- [26] Direct preference optimization of video large multimodal models from language model reward View paper
- [27] GLM-4.1 V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning View paper
- [28] Reinforcement Fine-Tuning Powers Reasoning Capability of Multimodal Large Language Models View paper
- [29] Multimodal machine translation with reinforcement learning View paper
- [30] Infi-MMR: Curriculum-based Unlocking Multimodal Reasoning via Phased Reinforcement Learning in Multimodal Small Language Models View paper
- [31] EvoLMM: Self-Evolving Large Multimodal Models with Continuous Rewards View paper
- [32] Multimodal image registration with deep context reinforcement learning View paper
- [33] Evolutionary reward design and optimization with multimodal large language models View paper
- [34] Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding View paper
- [35] A reinforcement learning-based approach for promoting mental health using multimodal emotion recognition View paper
- [36] Safe RLHF-V: Safe Reinforcement Learning from Multi-modal Human Feedback View paper
- [37] TimeMaster: Training Time-Series Multimodal LLMs to Reason via Reinforcement Learning View paper
- [38] RDDRL: a recurrent deduction deep reinforcement learning model for multimodal vision-robot navigation View paper
- [39] Reinforcement Learning Tuning for VideoLLMs: Reward Design and Data Efficiency View paper
- [40] Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation View paper
- [41] VideoRewardBench: Comprehensive Evaluation of Multimodal Reward Models for Video Understanding View paper
- [42] Reinforcement learning-based multimodal model for the stock investment portfolio management task View paper
- [43] C2-evo: Co-evolving multimodal data and model for self-improving reasoning View paper
- [44] An adaptive reinforcement learning-based multimodal data fusion framework for humanâ□□robot confrontation gaming View paper
- [45] Aligning Dialogue Agents with Global Feedback via Large Language Model Reward Decomposition View paper
- [46] Global reward to local rewards: Multimodal-guided decomposition for improving dialogue agents View paper
- [47] Multimodal fusion for autonomous navigation via deep reinforcement learning with sparse rewards and hindsight experience replay View paper
- [48] Prefmmt: Modeling human preferences in preference-based reinforcement learning with multimodal transformers View paper
- [49] Rewarddance: Reward scaling in visual generation View paper
- [50] MORAL: A Multimodal Reinforcement Learning Framework for Decision Making in Autonomous Laboratories View paper
- [51] Personalizing reinforcement learning from human feedback with variational preference learning View paper
- [52] Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning View paper
- [53] Decisionnce: Embodied multimodal representations via implicit preference learning View paper
- [54] Aligning Large Vision-Language Models by Deep Reinforcement Learning and Direct Preference Optimization View paper
- [55] GraphFusion-HRL: Multi-Modal Hierarchical Reinforcement Graph Learning for Context-Rich Recommender Systems View paper
- [56] Vr-thinker: Boosting video reward models through thinking-with-image reasoning View paper
- [57] Dast: Difficulty-adaptive slow-thinking for large reasoning models View paper
- [58] Enhancing alignment using curriculum learning & ranked preferences View paper
- [59] Curriculum-RLAIF: Curriculum Alignment with Reinforcement Learning from AI Feedback View paper
- [60] Openprm: Building open-domain process-based reward models with preference trees View paper
- [61] 2D-Curri-DPO: Two-Dimensional Curriculum Learning for Direct Preference Optimization View paper
- [62] De novo molecular design enabled by direct preference optimization and curriculum learning View paper
- [63] Rovrm: A robust visual reward model optimized via auxiliary textual preference data View paper
- [64] Curriculum direct preference optimization for diffusion and consistency models View paper
- [65] Curry-dpo: Enhancing alignment using curriculum learning & ranked preferences View paper
- [66] Weighted-reward preference optimization for implicit model fusion View paper
- [67] Continuous reinforcement learning via advantage value difference reward shaping: A proximal policy optimization perspective View paper
- [68] An improved deep reinforcement learning algorithm for path planning in unmanned driving View paper
- [69] Avatar: Reinforcement learning to see, hear, and reason over video View paper
- [70] Boosting policy learning in reinforcement learning via adaptive intrinsic reward regulation View paper
- [71] Hybrid group relative policy optimization: A multi-sample approach to enhancing policy optimization View paper
- [72] Enhancing stability and performance in mobile robot path planning with pmr-dueling dqn algorithm View paper
- [73] Leftover lunch: Advantage-based offline reinforcement learning for language models View paper
- [74] From Sparse to Dense: Toddler-inspired Reward Transition in Goal-Oriented Reinforcement Learning View paper
- [75] Trust Region Reward Optimization and Proximal Inverse Reward Optimization Algorithm View paper
- [76] Quantile Advantage Estimation for Entropy-Safe Reasoning View paper