Novelty Assessment Report

Paper: SimpleVLA-RL: Scaling VLA Training via Reinforcement Learning

PDF URL: https://openreview.net/pdf?id=TQhSodCM4r

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Vision-Language-Action (VLA) models have emerged as a powerful paradigm for robotic manipulation. Despite substantial progress enabled by large-scale pretraining and supervised fine-tuning (SFT), these models face two fundamental challenges: (i) the scarcity and high cost of large-scale robotic trajectories required for SFT scaling, and (ii) limited generalization to tasks under distribution shift. To overcome these limitations, we explore reinforcement learning (RL) as a pathway to scaling VLA training beyond limited datasets. Inspired by LLM breakthroughs where RL with outcome rewards enhances step-by-step reasoning, we ask: Can outcome-driven RL improve long-horizon step-by-step action planning of VLA? In this work, we introduce SimpleVLA-RL, an efficient RL framework tailored for VLA models. Building upon veRL, we introduce VLA-specific trajectory sampling, scalable parallelization, multi-environment rendering, and optimized loss computation. Applied to OpenVLA-OFT, SimpleVLA-RL achieves 99\% of SoTA performance on LIBERO and 80\% relative improvement on RoboTwin 1.0\&2.0, outperforming π_0 with our proposed exploration-enhancing strategies. SimpleVLA-RL reduces dependence on large-scale data, enables robust generalization, and remarkably surpasses SFT in real-world tasks. Moreover, we identify a novel phenomenon "pushcut'' during RL training, wherein the policy discovers unseen patterns beyond those seen in previous training process.

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: **Scaling Vision-Language-Action Model Training via Reinforcement Learning**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

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

- **Reinforcement Learning Algorithms for VLA Fine-Tuning**
- **Reward Design and World Model Integration**
- **Reasoning and Chain-of-Thought Integration**
- **Data Generation and Self-Improvement Strategies**
- **Application Domains and Task-Specific Adaptations**
- **Architectural Innovations and Training Frameworks**
- **Survey and Taxonomic Studies**

Complete Taxonomy Tree

- Scaling Vision-Language-Action Model Training via Reinforcement Learning Survey Taxonomy
- Reinforcement Learning Algorithms for VLA Fine-Tuning
  - Policy Gradient and Proximal Policy Optimization Methods (4 papers)
  - [5] Improving Vision-Language-Action Model with Online Reinforcement Learning (Yanjiang Guo, 2025) View paper
  - [11] TGRPO :Fine-tuning Vision-Language-Action Model via Trajectory-wise Group Relative Policy Optimization (Kong He, 2025) View paper
  - [32] STARE-VLA: Progressive Stage-Aware Reinforcement for Fine-Tuning Vision-Language-Action Models (Feng Xu, 2025) View paper
  - [49] VLA Model Post-Training via Action-Chunked PPO and Self Behavior Cloning (Wang Si-cheng, 2025) View paper
  - Flow-Based and Diffusion Policy Optimization (4 papers)
  - [14] Reinforcement Fine-Tuning of Flow-Matching Policies for Vision-Language-Action Models (Lyu Mingyang, 2025) View paper
  - [24] Balancing Signal and Variance: Adaptive Offline RL Post-Training for VLA Flow Models (Zhang HongYin, 2025) View paper
  - [31] $\pi_\texttt{RL}$: Online RL Fine-tuning for Flow-based Vision-Language-Action Models (Kang Chen, 2025) View paper
  - [37] $Ï□_\texttt{RL}$: Online RL Fine-tuning for Flow-based Vision-Language-Action Models (Chen Kang, 2025) View paper
  - Offline and Batch Reinforcement Learning (4 papers)
  - [2] Co-rft: Efficient fine-tuning of vision-language-action models through chunked offline reinforcement learning (Zhang Tianle, 2025) View paper
  - [43] DEAS: DEtached value learning with Action Sequence for Scalable Offline RL (Kim, 2025) View paper
  - [44] Scaling Vision-and-Language Navigation With Offline RL (Bundele, 2024) View paper
  - [47] Integrating Failures in Robot Skill Acquisition with Offline Action-Sequence Diffusion RL (Wang Hecheng, 2025) View paper
  - Online Reinforcement Learning and Interactive Training (3 papers)
  - [30] Interactive Post-Training for Vision-Language-Action Models (Tan, 2025) View paper
  - [35] RLinf-VLA: A Unified and Efficient Framework for VLA+RL Training (Zang Hong-zhi, 2025) View paper
  - [41] MindDrive: A Vision-Language-Action Model for Autonomous Driving via Online Reinforcement Learning (Haoyu Fu, 2025) View paper
  - Preference Optimization and Alignment Methods (1 papers)

  - [19] Aligning Large Vision-Language Models by Deep Reinforcement Learning and Direct Preference Optimization (Nguyen, 2025) View paper
- Reward Design and World Model Integration
  - Process Reward and Dense Reward Models (2 papers)
  - [1] A vision-language-action-critic model for robotic real-world reinforcement learning (Zhai Shaopeng, 2025) View paper
  - [8] Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators (Li, 2025) View paper
  - World Model-Based Policy Learning (3 papers)
  - [4] Irl-vla: Training an vision-language-action policy via reward world model (Jiang An-qing, 2025) View paper
  - [33] WMPO: World Model-based Policy Optimization for Vision-Language-Action Models (Fangqi Zhu, 2025) View paper
  - [50] Reinforcing Action Policies by Prophesying (Jiahui Zhang, 2025) View paper
  - Verifiable Reward and Outcome-Based Learning (2 papers)
  - [21] ManipLVM-R1: Reinforcement Learning for Reasoning in Embodied Manipulation with Large Vision-Language Models (Song, 2025) View paper
  - [34] GTR: Guided Thought Reinforcement Prevents Thought Collapse in RL-based VLM Agent Training (Wei Tong, 2025) View paper
- Reasoning and Chain-of-Thought Integration
  - Chain-of-Thought Reasoning for Action Planning (4 papers)
  - [9] Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning (Hao Bai, 2024) View paper
  - [10] Thinkact: Vision-language-action reasoning via reinforced visual latent planning (Wu, 2025) View paper
  - [16] Vla-r1: Enhancing reasoning in vision-language-action models (Zhang Zeyu, 2025) View paper
  - [27] DeepThinkVLA: Enhancing Reasoning Capability of Vision-Language-Action Models (Cheng Yin, 2025) View paper
  - Reinforcement Fine-Tuning for Reasoning Enhancement (3 papers)
  - [7] GUI-R1 : A Generalist R1-Style Vision-Language Action Model For GUI Agents (Luo Run, 2025) View paper
  - [22] Agentic Jigsaw Interaction Learning for Enhancing Visual Perception and Reasoning in Vision-Language Models (Zeng Yu, 2025) View paper
  - [28] UIShift: Enhancing VLM-based GUI Agents through Self-supervised Reinforcement Learning (Longxi Gao, 2025) View paper
- Data Generation and Self-Improvement Strategies
  - RL-Driven Data Collection and Augmentation (1 papers)
  - [39] Discover, Learn, and Reinforce: Scaling Vision-Language-Action Pretraining with Diverse RL-Generated Trajectories (Rushuai Yang, 2025) View paper
  - Self-Improving and Residual Learning Frameworks (1 papers)
  - [25] Self-Improving Vision-Language-Action Models with Data Generation via Residual RL (Xiao Wenli, 2025) View paper
  - Exploration and Anti-Exploration Strategies (1 papers)
  - [26] Steering Vision-Language-Action Models as Anti-Exploration: A Test-Time Scaling Approach (Siyuan Yang, 2025) View paper
- Application Domains and Task-Specific Adaptations
  - Robotic Manipulation and Embodied Control ★ (6 papers)
  - [0] SimpleVLA-RL: Scaling VLA Training via Reinforcement Learning (Anon et al., 2026) View paper
  - [6] Large vlm-based vision-language-action models for robotic manipulation: A survey (Shao Rui, 2025) View paper
  - [15] SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Safe Reinforcement Learning (Borong Zhang, 2025) View paper
  - [17] SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Constrained Learning (Zhang, 2025) View paper
  - [23] MoRE: Unlocking Scalability in Reinforcement Learning for Quadruped Vision-Language-Action Models (Zhao Han, 2025) View paper
  - [42] MobileVLA-R1: Reinforcing Vision-Language-Action for Mobile Robots (Ting Huang, 2025) View paper
  - Autonomous Driving and Navigation (5 papers)
  - [12] Discrete Diffusion for Reflective Vision-Language-Action Models in Autonomous Driving (Li Pengxiang, 2025) View paper
  - [13] Reasoning-VLA: A Fast and General Vision-Language-Action Reasoning Model for Autonomous Driving (Dapeng Zhang, 2025) View paper
  - [20] Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail (Nvidia, 2025) View paper
  - [29] AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning (Zhou Zewei, 2025) View paper
  - [36] UrbanVLA: A Vision-Language-Action Model for Urban Micromobility (Li, 2025) View paper
  - Vision-Language Navigation and Spatial Reasoning (2 papers)
  - [18] Boosting Efficient Reinforcement Learning for Vision-and-Language Navigation With Open-Sourced LLM (Jiawei Wang, 2025) View paper
  - [48] NaVILA: Legged Robot Vision-Language-Action Model for Navigation (An-Chieh Cheng, 2024) View paper
  - GUI Agents and Digital Environment Interaction (1 papers)
  - [45] CLIP4MC: An RL-Friendly Vision-Language Model for Minecraft (Ziluo Ding, 2023) View paper
- Architectural Innovations and Training Frameworks
  - Model-Based Search and Planning Integration (1 papers)
  - [40] Improving Pre-Trained Vision-Language-Action Policies with Model-Based Search (Neary, 2025) View paper
  - Synthetic Environment Training and Sim-to-Real Transfer (1 papers)
  - [46] Enhancing Vision-Language Model Training with Reinforcement Learning in Synthetic Worlds for Real-World Success (Dereka, 2025) View paper
- Survey and Taxonomic Studies (2 papers)
  - [3] Pure vision language action (vla) models: A comprehensive survey (Zhang DaPeng, 2025) View paper
  - [38] A Survey on Reinforcement Learning of Vision-Language-Action Models for Robotic Manipulation (H Deng, 2025) View paper

Narrative

Core task: Scaling vision-language-action model training via reinforcement learning. The field has organized itself around several complementary branches that address different facets of this challenge. Reinforcement Learning Algorithms for VLA Fine-Tuning explores policy optimization techniques such as PPO variants and flow-based methods, while Reward Design and World Model Integration examines how to construct effective learning signals and leverage predictive models for sample efficiency. Reasoning and Chain-of-Thought Integration investigates how to incorporate deliberative planning into action selection, and Data Generation and Self-

Improvement Strategies focuses on synthetic data creation and autonomous curriculum learning. Application Domains and Task-Specific Adaptations targets concrete settings like robotic manipulation, navigation, and GUI control, whereas Architectural Innovations and Training Frameworks addresses model design and scalable training infrastructure. Survey and Taxonomic Studies provide high-level perspectives on the rapidly evolving landscape, as seen in works like Pure VLA Survey[3] and VLA RL Survey[38].

A particularly active line of work centers on robotic manipulation and embodied control, where methods must balance generalization across diverse tasks with sample-efficient learning from limited real-world interactions. SimpleVLA-RL[0] sits within this branch, emphasizing straightforward RL-based fine-tuning for vision-language-action policies in manipulation settings. It shares thematic ground with VLA Online RL[5], which also explores online policy improvement, and contrasts with approaches like IRL-VLA[4] that leverage inverse reinforcement learning for reward specification. Nearby efforts such as SafeVLA[15] and SafeVLA Constrained[17] highlight the importance of safety constraints in physical systems, while MoRE[23] and MobileVLA-R1[42] address mixture-of-experts architectures and mobile manipulation respectively. The central tension across these works involves trading off exploration risk, computational cost, and the ability to generalize from pre-trained vision-language representations to precise low-level control, with SimpleVLA-RL[0] positioning itself as a practical entry point that prioritizes simplicity and scalability in the RL fine-tuning process.

Related Works in Same Category

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

1. Large vlm-based vision-language-action models for robotic manipulation: A survey

Authors: Shao Rui, Li Wei, Zhang Ren-shan, Liu, Zhiyang, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Robotic manipulation, a key frontier in robotics and embodied AI, requires precise motor control and multimodal understanding, yet traditional rule-based methods fail to scale or generalize in unstructured, novel environments. In recent years, Vision-Language-Action (VLA) models, built upon Large Vision-Language Models (VLMs) pretrained on vast image-text datasets, have emerged as a transformative paradigm. This survey provides the first systematic, taxonomy-oriented review of large VLM-based VL...

Relationship Analysis

Both papers belong to the same taxonomy category focusing on VLA RL approaches for robotic manipulation and embodied AI. They share overlapping areas in applying reinforcement learning to vision-language-action models for robotic tasks, addressing data scarcity and generalization challenges. However, the original paper (SimpleVLA-RL) presents a specific RL training framework with technical contributions like trajectory sampling and exploration strategies, while the candidate paper is a comprehensive survey that systematically reviews and taxonomizes the broader landscape of large VLM-based VLA models including but not limited to RL approaches.

2. SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Safe Reinforcement Learning

Authors: Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, et al. (7 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

N/A

Relationship Analysis

Both papers belong to the same taxonomy category focusing on VLA RL approaches for robotic manipulation and embodied AI tasks. They share the common goal of applying reinforcement learning to vision-language-action models for robotic control. However, SimpleVLA-RL focuses on scaling VLA training through outcome-driven RL to address data scarcity and improve long-horizon planning, while SafeVLA emphasizes safety alignment of VLA models through safe reinforcement learning techniques, addressing a distinct concern of ensuring safe robot behavior during RL training.

3. SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Constrained Learning

Authors: Zhang, Borong, Zhang Yuhao, Ji, Jiaming, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Vision-language-action models (VLAs) show potential as generalist robot policies. However, these models pose extreme safety challenges during real-world deployment, including the risk of harm to the environment, the robot itself, and humans. How can safety constraints be explicitly integrated into VLAs? We address this by exploring an integrated safety approach (ISA), systematically modeling safety requirements, then actively eliciting diverse unsafe behaviors, effectively constraining VLA polic...

Relationship Analysis

Both papers belong to the Robotic Manipulation and Embodied Control category, applying reinforcement learning to vision-language-action models for robotic tasks. While SimpleVLA-RL focuses on scaling VLA training through outcome-driven RL to improve performance and generalization under data scarcity, SafeVLA addresses a complementary dimension by integrating safety constraints into VLA policies via constrained RL (CMDP framework) to prevent harmful behaviors during deployment. The key difference is that SimpleVLA-RL optimizes for task success and data efficiency, whereas SafeVLA explicitly models and enforces safety requirements alongside performance objectives.

4. MoRE: Unlocking Scalability in Reinforcement Learning for Quadruped Vision-Language-Action Models

Authors: Zhao Han, Song Wen-xuan, Wang, Donglin, Tong Xin-yang, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

N/A

Relationship Analysis

Both papers belong to the Robotic Manipulation and Embodied Control category, focusing on scaling VLA models through reinforcement learning for robotic tasks. They share the core approach of applying RL to vision-language-action models to improve manipulation performance beyond supervised fine-tuning. The key difference is that SimpleVLA-RL focuses on general manipulation tasks across multiple benchmarks (LIBERO, RoboTwin) with emphasis on data efficiency and outcome-driven rewards, while MoRE specifically targets quadruped locomotion and vision-language-action integration for legged robots, representing a distinct embodied AI subdomain within the broader robotic manipulation category.

5. MobileVLA-R1: Reinforcing Vision-Language-Action for Mobile Robots

Authors: Ting Huang, Dongjian Li, Rui Yang, Zeyu Zhang, Zida Yang, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Grounding natural-language instructions into continuous control for quadruped robots remains a fundamental challenge in vision language action. Existing methods struggle to bridge high-level semantic reasoning and low-level actuation, leading to unstable grounding and weak generalization in the real world. To address these issues, we present MobileVLA-R1, a unified vision-language-action framework that enables explicit reasoning and continuous control for quadruped robots. We construct MobileVLA...

Relationship Analysis

Both papers belong to the same taxonomy category of robotic manipulation and embodied control using VLA models with reinforcement learning. They share overlapping approaches in applying RL to VLA training, with both using outcome-based rewards and addressing data scarcity and generalization challenges. However, SimpleVLA-RL focuses on scaling VLA training through GRPO-based RL for general manipulation tasks across multiple benchmarks (LIBERO, RoboTwin), while MobileVLA-R1 specifically targets quadruped robot navigation and control by integrating Chain-of-Thought reasoning with continuous locomotion commands, emphasizing vision-language navigation tasks (R2R, RxR) and mobile robot deployment.

Contributions Analysis

Overall novelty summary. The paper introduces SimpleVLA-RL, an efficient reinforcement learning framework for vision-language-action models applied to robotic manipulation. It resides in the 'Robotic Manipulation and Embodied Control' leaf, which contains six papers including the original work. This leaf sits within the broader 'Application Domains and Task-Specific Adaptations' branch, indicating a moderately populated research direction focused on practical deployment. The taxonomy reveals that robotic manipulation is one of four application domains, suggesting this is an active but not overcrowded area compared to the algorithmic development branches.

The taxonomy structure shows that SimpleVLA-RL's leaf neighbors include autonomous driving, vision-language navigation, and GUI agents, each addressing distinct embodiment challenges. The paper's approach connects to algorithmic branches such as 'Policy Gradient and Proximal Policy Optimization Methods' and 'Online Reinforcement Learning and Interactive Training', which contain four and three papers respectively. The 'Reward Design and World Model Integration' branch, particularly 'Verifiable Reward and Outcome-Based Learning' with two papers, provides relevant context for the outcome-driven RL paradigm. This positioning suggests the work bridges application-specific concerns with established algorithmic foundations.

Among 27 candidates examined, the contribution-level analysis reveals mixed novelty signals. The core framework contribution (SimpleVLA-RL) examined 10 candidates with zero refutations, suggesting relative novelty in the specific engineering approach. Exploration-enhancing strategies examined 7 candidates, also with zero refutations, indicating potential originality in this aspect. However, the outcome-driven RL paradigm examined 10 candidates and found 1 refutable match, suggesting that using simple binary rewards for VLA training has precedent in the limited search scope. The statistics indicate a focused but not exhaustive literature review.

Based on the top-27 semantic matches examined, the work appears to offer engineering contributions in framework design and exploration strategies, while the outcome-driven RL concept shows overlap with prior work. The taxonomy placement in a moderately populated leaf suggests the paper addresses a recognized problem space with established context. The analysis does not cover the full breadth of robotics or RL literature, so definitive novelty claims require broader verification beyond this semantic search scope.

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

Contribution 1: SimpleVLA-RL: An efficient RL framework for VLA models

Description: The authors develop an end-to-end reinforcement learning framework specifically designed for Vision-Language-Action models. This framework extends veRL with VLA-specific components including interactive trajectory sampling, parallel multi-environment rendering, and optimized training-inference-rendering infrastructure to enable stable and sample-efficient online RL training for robotic manipulation.

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. Language-conditioned imitation learning for robot manipulation tasks

URL: View paper

Brief Assessment

Language-conditioned Imitation[53] focuses on imitation learning from demonstrations with language conditioning for manipulation tasks, not on reinforcement learning frameworks for VLA models. The candidate uses supervised learning to map language instructions and visual observations to control policies, whereas the original contribution develops an RL infrastructure with trajectory sampling, parallel rendering, and online training components.

2. Liv: Language-image representations and rewards for robotic control

URL: View paper

Brief Assessment

Liv[54] focuses on learning vision-language representations and reward functions from action-free human videos for robotic control, not on developing an RL training framework for VLA models. The candidate addresses representation learning and reward specification, while the original contribution is about infrastructure for online RL training of VLA models with specific components like parallel rendering and trajectory sampling.

3. Vlmpc: Vision-language model predictive control for robotic manipulation

URL: View paper

Brief Assessment

VLMPC[55] focuses on integrating vision-language models with model predictive control for robotic manipulation, not on developing an RL framework for VLA training. The candidate uses MPC with video prediction rather than reinforcement learning for policy optimization.

4. MAP-VLA: Memory-Augmented Prompting for Vision-Language-Action Model in Robotic Manipulation

URL: View paper

Brief Assessment

MAP-VLA[56] focuses on memory-augmented prompting via soft prompts and retrieval mechanisms for frozen VLA models, not on developing an end-to-end RL training framework with trajectory sampling, parallel rendering, and optimized training-inference infrastructure as in the original paper.

5. Thinkact: Vision-language-action reasoning via reinforced visual latent planning

URL: View paper

Brief Assessment

ThinkAct[10] focuses on dual-system reasoning with visual latent planning for embodied tasks, not on building an efficient RL training infrastructure for VLA models. The technical approaches differ fundamentally in architecture and objectives.

6. TinyVLA: Toward Fast, Data-Efficient Vision-Language-Action Models for Robotic Manipulation

URL: View paper

Brief Assessment

TinyVLA[51] focuses on building compact, fast VLA models through multimodal model initialization and diffusion policy decoders, without requiring pre-training. It does not present an RL framework for VLA training, which is the core contribution of the original paper.

7. ManipLVM-R1: Reinforcement Learning for Reasoning in Embodied Manipulation with Large Vision-Language Models

URL: View paper

Brief Assessment

ManipLVM-R1[21] focuses on rule-based reward design for affordance perception and trajectory matching in manipulation tasks, rather than developing a general RL training infrastructure with parallelization and trajectory sampling components like SimpleVLA-RL.

8. Recipe for Vision-Language-Action Models in Robotic Manipulation: A Survey

URL: View paper

Brief Assessment

VLA Recipe Survey[52] is a survey paper that reviews existing work in vision-language-action models. The provided context only contains brief mentions of reinforcement learning frameworks without detailed technical descriptions that could refute the novelty of SimpleVLA-RL's specific contributions (veRL extension, trajectory sampling, parallel rendering, training-inference-rendering infrastructure).

9. Vla-r1: Enhancing reasoning in vision-language-action models

URL: View paper

Brief Assessment

VLA-R1[16] focuses on enhancing reasoning capabilities in VLA models through chain-of-thought supervision and RLVR-based post-training with verifiable rewards (affordance, trajectory, format), rather than developing a general RL framework infrastructure for VLA training. The candidate addresses reasoning quality and execution accuracy, while the original contribution emphasizes scalable RL infrastructure with trajectory sampling, parallel rendering, and training-inference-rendering optimization.

10. A Survey on Reinforcement Learning of Vision-Language-Action Models for Robotic Manipulation

URL: View paper

Brief Assessment

VLA RL Survey[38] is a survey paper that reviews existing work on RL for VLA models, rather than proposing a novel framework. It does not present an implementation with specific technical components like trajectory sampling or parallel rendering infrastructure.

Contribution 2: Exploration-enhancing strategies for VLA RL training

Description: The authors introduce three key modifications to enhance exploration during RL training: dynamic sampling that excludes uniform-reward trajectory groups, higher clipping range in the GRPO objective (from [0.8, 1.2] to [0.8, 1.28]), and increased rollout temperature (from 1.0 to 1.6). These strategies collectively improve training stability and policy performance.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. DCPO: Dynamic Clipping Policy Optimization

URL: View paper

Brief Assessment

DCPO[67] focuses on token-level exploration in LLM reasoning tasks through dynamic clipping and advantage standardization, not on VLA models or robotic manipulation. The candidate addresses text generation optimization, while the original paper targets vision-language-action models with environment interaction, making direct comparison inappropriate.

2. From Uniform to Heterogeneous: Tailoring Policy Optimization to Every Token's Nature

URL: View paper

Brief Assessment

Heterogeneous Policy Optimization[69] focuses on token-level optimization strategies for LLM reasoning tasks, not vision-language-action models for robotic manipulation. The candidate addresses entropy-based token treatment in language generation, while the original contribution concerns dynamic sampling, clipping ranges, and temperature adjustments specifically for VLA trajectory rollout in embodied environments.

3. BAPO: Stabilizing Off-Policy Reinforcement Learning for LLMs via Balanced Policy Optimization with Adaptive Clipping

URL: View paper

Brief Assessment

BAPO[70] focuses on off-policy RL stabilization for LLMs through balanced optimization and adaptive clipping, not on VLA models or robotic manipulation. The candidate addresses entropy collapse and gradient imbalance in language model training, whereas the original paper's exploration strategies (dynamic sampling, clipping range adjustments, temperature tuning) are specifically designed for vision-language-action models in embodied tasks.

4. A dynamical clipping approach with task feedback for proximal policy optimization

URL: View paper

Brief Assessment

Dynamical Clipping PPO[73] focuses on dynamically adjusting PPO clipping bounds for general RL tasks using multi-armed bandit feedback, not specifically on VLA models or the combination of dynamic sampling, temperature adjustment, and clipping range modifications for vision-language-action training.

5. CE-GPPO: Coordinating Entropy via Gradient-Preserving Clipping Policy Optimization in Reinforcement Learning

URL: View paper

Brief Assessment

CE-GPPO[72] focuses on entropy regulation in policy gradient methods for LLMs through gradient-preserving clipping mechanisms, not on VLA-specific exploration strategies like dynamic sampling, clipping ranges, or rollout temperature for vision-language-action models.

6. Adaptive PPO With Multi-Armed Bandit Clipping and Meta-Control for Robust Power Grid Operation Under Adversarial Attacks

URL: View paper

Brief Assessment

Adaptive PPO Bandit[71] focuses on power grid control with multi-armed bandit clipping and meta-control mechanisms, not on vision-language-action models or robotic manipulation tasks. The technical domains and application contexts are fundamentally different.

7. Klear-reasoner: Advancing reasoning capability via gradient-preserving clipping policy optimization

URL: View paper

Brief Assessment

Klear-Reasoner[68] focuses on reasoning models for math/code tasks using GPPO and clipping mechanisms, not vision-language-action models for robotic manipulation. The exploration strategies (dynamic sampling, clipping range, temperature) target different domains and model architectures.

Contribution 3: Outcome-driven RL paradigm for VLA with simple binary rewards

Description: The authors apply an outcome-level reinforcement learning approach to VLA models using only sparse binary rewards (1 for task success, 0 for failure) rather than hand-crafted dense rewards. This paradigm, inspired by recent LLM breakthroughs, enables VLA models to improve long-horizon action planning through trial-and-error exploration without requiring task-specific reward engineering.

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. Stabilizing Long-term Multi-turn Reinforcement Learning with Gated Rewards

URL: View paper

Brief Assessment

Gated Rewards[59] focuses on software engineering tasks with multi-turn reasoning and verification-based reward shaping, not vision-language-action models for robotic manipulation. The technical domains and application contexts are fundamentally different.

2. Revisiting sparse rewards for goal-reaching reinforcement learning

URL: View paper

Brief Assessment

Sparse Goal-Reaching[62] focuses on goal-reaching tasks in general RL settings (e.g., robotic reaching, ball-in-cup) with constant negative rewards (-1 per step), not on vision-language-action models or long-horizon action planning. The candidate does not address VLA architectures or the integration of vision-language understanding with action generation.

3. Hierarchical reinforcement learning for handling sparse rewards in multi-goal navigation

URL: View paper

Brief Assessment

Hierarchical Multi-Goal[57] focuses on multi-goal navigation tasks in robotics using hierarchical RL with sparse binary rewards (1 for goal completion, 0 otherwise). However, it does not address vision-language-action models or long-horizon action planning in the VLA context that the original paper targets.

4. Exploring the Limit of Outcome Reward for Learning Mathematical Reasoning

URL: View paper

Brief Assessment

Outcome Reward Limit[61] focuses on mathematical reasoning tasks using binary outcome rewards for LLMs, not vision-language-action models for robotic manipulation. The technical domains and applications are fundamentally different.

5. Offline reinforcement learning with failure under sparse reward environments

URL: View paper

Brief Assessment

Offline Sparse Rewards[58] focuses on offline RL in sparse reward environments for general control tasks, not on vision-language-action models or outcome-driven paradigms for long-horizon manipulation planning.

6. Sqil: Imitation learning via reinforcement learning with sparse rewards

URL: View paper

Prior Art Analysis

SQIL[60] demonstrates that outcome-driven reinforcement learning with sparse binary rewards was already established for imitation learning tasks before the original paper. SQIL uses a constant reward of r=+1 for matching demonstrated actions in demonstrated states and r=0 for all other behavior, which is fundamentally the same sparse binary reward structure (1 for success, 0 for failure) claimed as novel in the original paper. The candidate paper explicitly states this approach enables long-horizon imitation without hand-crafted dense rewards, directly paralleling the original paper's claimed contribution of using 'only sparse binary rewards (1 for task success, 0 for failure) rather than hand-crafted dense rewards' for long-horizon action planning.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same problem (avoiding hand-crafted/learned reward functions) and propose the same solution (using simple constant rewards instead). - **Original**: traditional robotics rl requires hand-crafted reward functions for each task, limiting scalability and generalization to novel scenarios where rewards are undefined - **Candidate**: Since the true reward function for the task is unknown, these methods learn a reward function from the demonstrations, often using complex and brittle approximation techniques that involve adversarial training. we propose a simple alternative that still uses rl, but does not require learning a rewar...

Evidence 2 - **Rationale**: SQIL[60] explicitly addresses long-horizon behavior through its reward structure, demonstrating that outcome-driven RL for long-horizon planning was already established. - **Original**: can outcome-driven rl improve long-horizon step-by-step action planning of vla? - **Candidate**: intuitively, adversarial methods encourage long-horizon imitation by providing the agent with (1) an incentive to imitate the demonstrated actions in demonstrated states, and (2) an incentive to take actions that lead it back to demonstrated states when it encounters new, out-ofdistribution states.

Evidence 3 - **Rationale**: Both papers use the same sparse outcome reward approach (binary rewards based on success) to enable step-by-step behavior learning. - **Original**: using only sparse outcome rewards, rl can significantly enhance models' ability to generate correct step-by-step reasoning chains - **Candidate**: the key idea is that, instead of using a learned reward function to provide a reward signal to the agent, we can simply give the agent a constant reward of r = +1 for matching the demonstrated action in a demonstrated state, and a constant reward of r= 0for all other behavior.

7. Multi-Goal Reinforcement Learning: Challenging Robotics Environments and Request for Research

URL: View paper

Brief Assessment

Multi-Goal Robotics[64] focuses on multi-goal RL for robotic manipulation with sparse binary rewards in traditional control tasks (pushing, sliding, pick & place), not on vision-language-action models or the outcome-driven paradigm for VLA training that the original paper proposes.

8. Tree Search for LLM Agent Reinforcement Learning

URL: View paper

Brief Assessment

Tree Search Agent[63] focuses on tree-based RL for LLM agents in QA tasks with step-wise process supervision, not vision-language-action models for robotic manipulation with binary outcome rewards.

9. Deep-reinforcement-learning-based autonomous UAV navigation with sparse rewards

URL: View paper

Brief Assessment

UAV Sparse Rewards[65] focuses on UAV navigation in airborne environments with sparse rewards, not vision-language-action models for robotic manipulation. The technical domains and model architectures are fundamentally different.

10. Seea-r1: Tree-structured reinforcement fine-tuning for self-evolving embodied agents

URL: View paper

Brief Assessment

SEEA-R1[66] focuses on embodied agents in interactive environments (e.g., household tasks in ALFWorld) using MCTS-based tree search with learned reward models, rather than robotic manipulation with VLA models. The technical approach and application domain differ substantially from the original paper's VLA manipulation framework.

Appendix: Text Similarity Detection

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

References

- [0] SimpleVLA-RL: Scaling VLA Training via Reinforcement Learning View paper
- [1] A vision-language-action-critic model for robotic real-world reinforcement learning View paper
- [2] Co-rft: Efficient fine-tuning of vision-language-action models through chunked offline reinforcement learning View paper
- [3] Pure vision language action (vla) models: A comprehensive survey View paper
- [4] Irl-vla: Training an vision-language-action policy via reward world model View paper
- [5] Improving Vision-Language-Action Model with Online Reinforcement Learning View paper
- [6] Large vlm-based vision-language-action models for robotic manipulation: A survey View paper
- [7] GUI-R1 : A Generalist R1-Style Vision-Language Action Model For GUI Agents View paper
- [8] Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators View paper
- [9] Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning View paper
- [10] Thinkact: Vision-language-action reasoning via reinforced visual latent planning View paper
- [11] TGRPO :Fine-tuning Vision-Language-Action Model via Trajectory-wise Group Relative Policy Optimization View paper
- [12] Discrete Diffusion for Reflective Vision-Language-Action Models in Autonomous Driving View paper
- [13] Reasoning-VLA: A Fast and General Vision-Language-Action Reasoning Model for Autonomous Driving View paper
- [14] Reinforcement Fine-Tuning of Flow-Matching Policies for Vision-Language-Action Models View paper
- [15] SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Safe Reinforcement Learning View paper
- [16] Vla-r1: Enhancing reasoning in vision-language-action models View paper
- [17] SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Constrained Learning View paper
- [18] Boosting Efficient Reinforcement Learning for Vision-and-Language Navigation With Open-Sourced LLM View paper
- [19] Aligning Large Vision-Language Models by Deep Reinforcement Learning and Direct Preference Optimization View paper
- [20] Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail View paper
- [21] ManipLVM-R1: Reinforcement Learning for Reasoning in Embodied Manipulation with Large Vision-Language Models View paper
- [22] Agentic Jigsaw Interaction Learning for Enhancing Visual Perception and Reasoning in Vision-Language Models View paper
- [23] MoRE: Unlocking Scalability in Reinforcement Learning for Quadruped Vision-Language-Action Models View paper
- [24] Balancing Signal and Variance: Adaptive Offline RL Post-Training for VLA Flow Models View paper
- [25] Self-Improving Vision-Language-Action Models with Data Generation via Residual RL View paper
- [26] Steering Vision-Language-Action Models as Anti-Exploration: A Test-Time Scaling Approach View paper
- [27] DeepThinkVLA: Enhancing Reasoning Capability of Vision-Language-Action Models View paper
- [28] UIShift: Enhancing VLM-based GUI Agents through Self-supervised Reinforcement Learning View paper
- [29] AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning View paper
- [30] Interactive Post-Training for Vision-Language-Action Models View paper
- [31] $\pi_\texttt{RL}$: Online RL Fine-tuning for Flow-based Vision-Language-Action Models View paper
- [32] STARE-VLA: Progressive Stage-Aware Reinforcement for Fine-Tuning Vision-Language-Action Models View paper
- [33] WMPO: World Model-based Policy Optimization for Vision-Language-Action Models View paper
- [34] GTR: Guided Thought Reinforcement Prevents Thought Collapse in RL-based VLM Agent Training View paper

- [35] RLinf-VLA: A Unified and Efficient Framework for VLA+RL Training View paper
- [36] UrbanVLA: A Vision-Language-Action Model for Urban Micromobility View paper
- [37] $Ï□_\texttt{RL}$: Online RL Fine-tuning for Flow-based Vision-Language-Action Models View paper
- [38] A Survey on Reinforcement Learning of Vision-Language-Action Models for Robotic Manipulation View paper
- [39] Discover, Learn, and Reinforce: Scaling Vision-Language-Action Pretraining with Diverse RL-Generated Trajectories View paper
- [40] Improving Pre-Trained Vision-Language-Action Policies with Model-Based Search View paper
- [41] MindDrive: A Vision-Language-Action Model for Autonomous Driving via Online Reinforcement Learning View paper
- [42] MobileVLA-R1: Reinforcing Vision-Language-Action for Mobile Robots View paper
- [43] DEAS: DEtached value learning with Action Sequence for Scalable Offline RL View paper
- [44] Scaling Vision-and-Language Navigation With Offline RL View paper
- [45] CLIP4MC: An RL-Friendly Vision-Language Model for Minecraft View paper
- [46] Enhancing Vision-Language Model Training with Reinforcement Learning in Synthetic Worlds for Real-World Success View paper
- [47] Integrating Failures in Robot Skill Acquisition with Offline Action-Sequence Diffusion RL View paper
- [48] NaVILA: Legged Robot Vision-Language-Action Model for Navigation View paper
- [49] VLA Model Post-Training via Action-Chunked PPO and Self Behavior Cloning View paper
- [50] Reinforcing Action Policies by Prophesying View paper
- [51] TinyVLA: Toward Fast, Data-Efficient Vision-Language-Action Models for Robotic Manipulation View paper
- [52] Recipe for Vision-Language-Action Models in Robotic Manipulation: A Survey View paper
- [53] Language-conditioned imitation learning for robot manipulation tasks View paper
- [54] Liv: Language-image representations and rewards for robotic control View paper
- [55] Vlmpc: Vision-language model predictive control for robotic manipulation View paper
- [56] MAP-VLA: Memory-Augmented Prompting for Vision-Language-Action Model in Robotic Manipulation View paper
- [57] Hierarchical reinforcement learning for handling sparse rewards in multi-goal navigation View paper
- [58] Offline reinforcement learning with failure under sparse reward environments View paper
- [59] Stabilizing Long-term Multi-turn Reinforcement Learning with Gated Rewards View paper
- [60] Sqil: Imitation learning via reinforcement learning with sparse rewards View paper
- [61] Exploring the Limit of Outcome Reward for Learning Mathematical Reasoning View paper
- [62] Revisiting sparse rewards for goal-reaching reinforcement learning View paper
- [63] Tree Search for LLM Agent Reinforcement Learning View paper
- [64] Multi-Goal Reinforcement Learning: Challenging Robotics Environments and Request for Research View paper
- [65] Deep-reinforcement-learning-based autonomous UAV navigation with sparse rewards View paper
- [66] Seea-r1: Tree-structured reinforcement fine-tuning for self-evolving embodied agents View paper
- [67] DCPO: Dynamic Clipping Policy Optimization View paper
- [68] Klear-reasoner: Advancing reasoning capability via gradient-preserving clipping policy optimization View paper
- [69] From Uniform to Heterogeneous: Tailoring Policy Optimization to Every Token's Nature View paper
- [70] BAPO: Stabilizing Off-Policy Reinforcement Learning for LLMs via Balanced Policy Optimization with Adaptive Clipping View paper
- [71] Adaptive PPO With Multi-Armed Bandit Clipping and Meta-Control for Robust Power Grid Operation Under Adversarial Attacks View paper
- [72] CE-GPPO: Coordinating Entropy via Gradient-Preserving Clipping Policy Optimization in Reinforcement Learning View paper
- [73] A dynamical clipping approach with task feedback for proximal policy optimization View paper