Novelty Assessment Report

Paper: Multimodal Policy Internalization for Conversational Agents
PDF URL: https://openreview.net/pdf?id=fSE0rUngCX
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Modern conversational agents such as ChatGPT and Alexa+ have become indispensable in everyday life. To handle diverse business requirements and enable agentic capabilities, these LLM-based systems often rely on predefined policies, which specify instructions such as model metadata, response styles, and tool-using rules. These policies, typically implemented as in-context prompts, are becoming increasingly complex and lengthy, posing challenges for models in faithfully following them. Moreover, they impose a large fixed computational cost regardless of the input query. As multimodal conversational agents emerge, complex policies that govern multimodal tasks and even involve visual instructions are becoming increasingly necessary, yet they have been rarely studied in previous work. In particular, prior work on prompt compression has focused solely on reducing the length of task templates and demonstrations, which require limited reasoning compared to policies. Meanwhile, related work on policy alignment has been limited to internalizing text-only safety instructions. To bridge this gap, we introduce Multimodal Policy Internalization (MPI), a new task that aims to internalize reasoning-intensive multimodal policies into the parameters of a large multimodal model, enabling stronger policy-following behavior without requiring the policy to be included in-context during inference. MPI presents unique challenges from both data and algorithmic perspectives. We construct two new datasets that cover complex decision-making and tool-using tasks across both synthetic and real-world visual inputs. We investigate diverse internalization strategies and propose a novel three-stage training framework, TriMPI, which enables stronger guidance from the original policy during internalization. Specifically, we first introduce a continual pretraining stage before supervised finetuning, which directly injects policy knowledge into the model. We then propose PolicyRollout, a simple yet effective extension to GRPO-style RL algorithms, which enables more grounded exploration by augmenting the rollout space with policy-aware responses. We show significant improvements of TriMPI over strong baselines in end-to-end performance, generalization capability, and robustness to catastrophic forgetting. As the first work on multimodal policy internalization, we aim to build a strong foundation for future research by providing datasets, training recipes, and comprehensive evaluations.

Core Task Landscape

This paper addresses: **Internalizing Multimodal Policies into Large Multimodal Model Parameters**

A total of **15 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

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

- **Policy Internalization and Alignment Methods**
- **Policy Integration Architectures for Embodied Agents**
- **Unified Multimodal Policy Learning from Task Specifications**
- **Supporting Technologies for Multimodal Policy Systems**

Complete Taxonomy Tree

- Internalizing Multimodal Policies into Large Multimodal Model Parameters Survey Taxonomy
- Policy Internalization and Alignment Methods
  - Multimodal Policy Internalization via Multi-Stage Training ★ (1 papers)
  - [0] Multimodal Policy Internalization for Conversational Agents (Anon et al., 2026) View paper
  - Safety-Grounded Policy Alignment for Vision-Language Models (1 papers)
  - [4] MSR-Align: Policy-Grounded Multimodal Alignment for Safety-Aware Reasoning in Vision-Language Models (Xia Yinan, 2025) View paper
  - Task Vector-Based In-Context Policy Adaptation (1 papers)
  - [7] Multimodal task vectors enable many-shot multimodal in-context learning (Assaf Arbelle, 2024) View paper
- Policy Integration Architectures for Embodied Agents
  - Hierarchical Planning-Control Architectures with Multimodal Models
  - Goal-Conditioned Policy Architectures for Sequential Tasks (1 papers)
    - [5] Optimus-2: Multimodal minecraft agent with goal-observation-action conditioned policy (Li, 2025) View paper
  - Multimodal Perception-Planning Frameworks for Manipulation (1 papers)
    - [2] RoboMP2: A Robotic Multimodal Perception-Planning Framework with Multimodal Large Language Models (Lv Qi, 2024) View paper
  - 3D Feature Field Integration with Language Model Planners (1 papers)
    - [3] Integrating LMM Planners and 3D Skill Policies for Generalizable Manipulation (Li Yue-lei, 2025) View paper
  - Landmark-Centric Multimodal Navigation Frameworks (1 papers)
    - [14] LMNav: Landmark-Centric Multimodal Reasoning for Vision-And-Language Navigation in Continuous Environments (Kun Luo, 2025) View paper
  - Agile Locomotion-Navigation Integration for Quadruped Agents (1 papers)
    - [13] QuadrupedGPT: Towards a Versatile Quadruped Agent in Open-ended Worlds (Mei, 2024) View paper

- Simulation-Augmented Policy Learning with Generative Models (1 papers)
  - [1] MultiGen: Using Multimodal Generation in Simulation to Learn Multimodal Policies in Real (R Wang, 2025) View paper
- Unified Multimodal Policy Learning from Task Specifications
  - Cross-Modal Reasoning for Unified Policy Execution (1 papers)
  - [6] Mutex: Learning unified policies from multimodal task specifications (Shah, 2023) View paper
  - Unimodal-to-Multimodal Policy Transfer via Pretrained Models (1 papers)
  - [8] Robo-mutual: Robotic multimodal task specification via unimodal learning (Jianxiong Li, 2025) View paper
- Supporting Technologies for Multimodal Policy Systems
  - Controllable Multimodal Data Synthesis for Policy Training (1 papers)
  - [10] CtrlSynth: Controllable Image Text Synthesis for Data-Efficient Multimodal Learning (CAO Qingqing, 2024) View paper
  - Multimodal Retrieval-Augmented Generation for Enhanced Search (1 papers)
  - [9] CUE-M: Contextual Understanding and Enhanced Search with Multimodal Large Language Model (Go, 2024) View paper
  - Multimodal Data Fusion with Hybrid Learning and Reinforcement Optimization (1 papers)
  - [12] Design of an Improved Model for Multimodal Data Fusion Using XGBoost-LSTM-CNN and Proximal Policy Optimizations (Kamble, 2024) View paper
  - Domain-Specific Multimodal Policy Applications and Surveys (2 papers)
  - [11] Policy-driven, multimodal deep learning for predicting visual fields from the optic disc and OCT imaging (Yuka Kihara, 2022) View paper
  - [15] Safe Learning for Contact-Rich Robot Tasks: A Survey from Classical Learning-Based Methods to Safe Foundation Models (H Zhang, 2025) View paper

Narrative

Core task: Internalizing multimodal policies into large multimodal model parameters. This field addresses how to embed decision-making capabilities directly within the weights of large multimodal models, enabling them to act as embodied agents that perceive and respond to complex environments. The taxonomy organizes research into four main branches: Policy Internalization and Alignment Methods focus on training strategies that fuse behavioral policies with pretrained representations, often through multi-stage pipelines or alignment objectives; Policy Integration Architectures for Embodied Agents explore architectural designs that combine vision-language backbones with action prediction modules; Unified Multimodal Policy Learning from Task Specifications investigates how models can generalize across diverse tasks by conditioning on natural language or other modalities; and Supporting Technologies for Multimodal Policy Systems cover auxiliary techniques such as data synthesis, safety mechanisms, and evaluation frameworks. Representative works like MultiGen[1] and RoboMP2[2] illustrate how pretraining and fine-tuning stages can be orchestrated to internalize control policies, while LMM Planners Skills[3] and Optimus-2[5] demonstrate different ways to integrate planning and low-level skills within unified architectures.

A particularly active line of work examines trade-offs between end-to-end internalization and modular decomposition: some approaches embed all reasoning and control within a single model, while others retain separate planning or skill modules that interact with a central multimodal backbone. Another recurring theme is the tension between generalization across tasks and specialization for specific embodiments, with methods like MSR-Align[4] and Robo-mutual[8] exploring alignment strategies that balance broad pretraining with domain-specific adaptation. The original paper, Multimodal Policy Internalization[0], sits within the multi-stage training branch and emphasizes progressive internalization of policies through carefully designed training phases. Compared to nearby works such as LMM Planners Skills[3], which may retain explicit planning modules, and Optimus-2[5], which focuses on unified policy learning from task specifications, Multimodal Policy Internalization[0] appears to prioritize deeper integration of behavioral policies directly into model parameters, aiming for a more seamless fusion of perception, reasoning, and action generation within a single multimodal architecture.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on multi-stage training pipelines (continual pretraining, supervised finetuning, RL) that permanently modify model parameters to internalize multimodal policies. The sibling subtopics represent alternative approaches: one specializes in safety-specific alignment through fine-grained reasoning, while the other uses task vectors for in-context adaptation without parameter updates. These represent different points on the spectrum between permanent internalization and dynamic adaptation.

Similarities: - All three subtopics address how to incorporate policies into large multimodal models - All work with vision-language or multimodal inputs rather than text-only scenarios - All aim to improve model behavior alignment with desired policies or objectives

Differences: - Multi-Stage Training uses sequential training phases to permanently modify parameters; Safety-Grounded focuses on safety-specific reasoning alignment; Task Vector-Based uses in-context learning without parameter updates - Multi-Stage Training is general-purpose policy internalization; Safety-Grounded specializes in preventing harmful behaviors; Task Vector-Based enables rapid task adaptation - Multi-Stage Training requires full training pipeline; Safety-Grounded emphasizes fine-grained reasoning mechanisms; Task Vector-Based operates through many-shot prompting with derived vectors - Parameter permanence: Multi-Stage creates lasting changes through training; Task Vector-Based maintains base parameters and adapts contextually

Suggested Search Directions: - Hybrid approaches combining multi-stage training with task vector adaptation for flexible policy internalization - Safety-aware multi-stage training pipelines that integrate fine-grained reasoning into each training phase - Comparative studies on when parameter internalization versus in-context adaptation is more effective for policy learning

Sibling Subtopics

- **Safety-Grounded Policy Alignment for Vision-Language Models** (leaves: 1, papers: 1)
- Scope: Methods aligning models with safety policies through fine-grained reasoning over multimodal inputs to prevent harmful behaviors.
- Exclude: General policy internalization without safety focus belongs in Multi-Stage Training; task-specific alignment belongs elsewhere.
- **Task Vector-Based In-Context Policy Adaptation** (leaves: 1, papers: 1)
- Scope: Techniques using task vectors derived from multimodal examples to enable many-shot in-context learning without parameter updates.
- Exclude: Parameter-level internalization through training belongs in other internalization methods; architectural integration belongs in Policy Integration.

Contributions Analysis

Overall novelty summary. The paper introduces Multimodal Policy Internalization (MPI), a task aimed at embedding complex multimodal policies—including visual instructions and tool-using rules—directly into model parameters through multi-stage training. Within the taxonomy, it occupies the 'Multimodal Policy Internalization via Multi-Stage Training' leaf under 'Policy Internalization and

Alignment Methods'. Notably, this leaf contains only the original paper itself, with no sibling papers identified, suggesting this specific formulation of multi-stage multimodal policy internalization represents a relatively sparse research direction within the broader field of 15 surveyed papers.

The taxonomy reveals neighboring work in adjacent leaves: 'Safety-Grounded Policy Alignment for Vision-Language Models' focuses on safety-specific alignment rather than general policy internalization, while 'Task Vector-Based In-Context Policy Adaptation' explores parameter-free adaptation mechanisms. The broader 'Policy Integration Architectures' branch contains hierarchical planning-control systems that maintain modular separation between reasoning and execution, contrasting with the paper's emphasis on unified parameter-level internalization. The 'Unified Multimodal Policy Learning' branch addresses cross-modal reasoning but without the explicit multi-stage internalization strategy proposed here, highlighting how this work bridges internalization methods with unified policy execution.

Among 30 candidates examined through semantic search, none were found to clearly refute any of the three main contributions: the MPI task formulation (10 candidates examined, 0 refutable), the ClevrPolicy and GTAPolicy datasets (10 candidates, 0 refutable), and the TriMPI training framework with PolicyRollout algorithm (10 candidates, 0 refutable). This suggests that within the limited search scope, the specific combination of multimodal policy internalization through multi-stage training with visual policy instructions appears relatively unexplored. However, the analysis is constrained by the top-30 semantic matches and does not constitute an exhaustive literature review.

Based on the limited search scope, the work appears to occupy a novel position by explicitly targeting multimodal policy internalization through parameter-level training, rather than relying on in-context prompting or modular architectures. The absence of sibling papers in its taxonomy leaf and the lack of refuting candidates among 30 examined works suggest potential novelty, though a broader literature search would be needed to confirm whether related approaches exist in adjacent research communities or under different terminology.

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

Contribution 1: Multimodal Policy Internalization (MPI) task

Description: The authors define a new task called Multimodal Policy Internalization (MPI), which aims to embed complex multimodal policies into model parameters so that models can generate policy-compliant responses without requiring the policy in-context during inference. This task extends prior work on text-only policy alignment to the multimodal domain.

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. Perception-aware policy optimization for multimodal reasoning

URL: View paper

Brief Assessment

Perception-Aware Policy[42] focuses on improving perception during multimodal reasoning through policy optimization (RLVR), not on internalizing policies into model parameters to remove in-context requirements during inference.

---

2. Think Then Embed: Generative Context Improves Multimodal Embedding

URL: View paper

Brief Assessment

Think Then Embed[43] focuses on universal multimodal embeddings for task-specific representations using chain-of-thought reasoning, not on internalizing complex policies into model parameters for conversational agents.

---

3. Understand, Think, and Answer: Advancing Visual Reasoning with Large Multimodal Models

URL: View paper

Brief Assessment

Understand Think Answer[38] focuses on visual reasoning through an 'understand-think-answer' process for compositional tasks, not on internalizing reasoning-intensive policies into model parameters as the original paper proposes.

---

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

URL: View paper

Brief Assessment

R1-VL[39] focuses on enhancing multimodal reasoning through step-wise reinforcement learning rewards for reasoning accuracy and validity, not on internalizing complex policies into model parameters to eliminate in-context policy requirements during inference.

---

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

URL: View paper

Brief Assessment

Reinforced MLLM Survey[40] focuses on RL-based reasoning methods for multimodal large language models in general, covering algorithmic designs and reward mechanisms. It does not address the specific task of internalizing complex multimodal policies into model parameters to enable policy-compliant responses without in-context policy prompts during inference.

---

6. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning

URL: View paper

Brief Assessment

Critic-V[45] focuses on improving VLM reasoning through external critic feedback for error detection and correction in multimodal reasoning tasks. It does not address policy internalization into model parameters, which is the core novelty of the original paper's MPI task.

---

7. Multimodal chain-of-thought reasoning: A comprehensive survey

URL: View paper

Brief Assessment

Multimodal Chain-of-Thought Survey[37] focuses on chain-of-thought reasoning methodologies across various modalities (image, video, audio, 3D) for tasks like VQA and embodied AI, not on internalizing complex policies into model parameters for conversational agents.

---

8. Perception, reason, think, and plan: A survey on large multimodal reasoning models

URL: View paper

Brief Assessment

The candidate survey focuses on multimodal reasoning models broadly, covering perception, understanding, and planning across modalities. It does not specifically address the task of internalizing reasoning-intensive policies into model parameters as defined in the original paper's MPI contribution.

---

9. Struct2D: A Perception-Guided Framework for Spatial Reasoning in Large Multimodal Models

URL: View paper

Brief Assessment

Struct2D[41] focuses on spatial reasoning in 3D environments using structured 2D representations (BEV images, metadata) for tasks like navigation and object grounding. It does not address policy internalization or embedding complex multimodal policies into model parameters.

---

10. Boosting Reasoning in Large Multimodal Models via Activation Replay

URL: View paper

Brief Assessment

Activation Replay[44] focuses on post-training analysis and test-time manipulation of activations in reinforcement learning with verifiable rewards (RLVR), not on internalizing reasoning-intensive policies into model parameters as a training objective.

Contribution 2: ClevrPolicy and GTAPolicy datasets

Description: The authors introduce two new datasets: ClevrPolicy, which focuses on reasoning-intensive decision-making with synthetic images and controllable policy complexity, and GTAPolicy, which targets tool-usage instructions with real-world images in a low-data regime. These datasets support training and evaluation of multimodal policy internalization methods.

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. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology

URL: View paper

Brief Assessment

Autonomous Oncology Agent[28] focuses on clinical decision-making in oncology using medical imaging and precision oncology tools, not on general decision-making or tool-usage datasets with visual inputs for training conversational agents.

---

2. Visualtoolagent (vista): A reinforcement learning framework for visual tool selection

URL: View paper

Brief Assessment

VisualToolAgent[31] focuses on reinforcement learning for visual tool selection in reasoning tasks, not on creating datasets for policy internalization with multimodal decision-making and tool-usage instructions.

---

3. VISION Datasets: A Benchmark for Vision-based InduStrial InspectiON

URL: View paper

Brief Assessment

VISION Datasets[27] focuses on industrial defect detection with segmentation masks for manufacturing inspection, not on decision-making or tool-usage tasks with multimodal policy internalization.

---

4. Ui-vision: A desktop-centric gui benchmark for visual perception and interaction

URL: View paper

Brief Assessment

UI-Vision[29] focuses on desktop GUI interaction benchmarks with element grounding, layout grounding, and action prediction tasks across real desktop applications. The candidate does not address decision-making datasets with controllable policy complexity or tool-usage instruction datasets, which are the core contributions of the original paper's ClevrPolicy and GTAPolicy datasets.

---

5. Openthinkimg: Learning to think with images via visual tool reinforcement learning

URL: View paper

Brief Assessment

OpenThinkImg[30] focuses on chart reasoning tasks and does not introduce datasets for decision-making or tool-usage tasks with visual inputs comparable to ClevrPolicy and GTAPolicy. The candidate's work addresses a different problem domain (chart analysis with vision tools) rather than policy internalization for conversational agents.

---

6. A Benchmarking Study of Vision-based Robotic Grasping Algorithms

URL: View paper

Brief Assessment

Robotic Grasping Benchmark[26] focuses on benchmarking vision-based robotic grasping algorithms with variations in physical conditions (lighting, cameras, grippers), not on datasets for reasoning-intensive decision-making or tool-usage tasks with multimodal policies.

---

7. Efficient and Accurate Pneumonia Detection Using a Novel Multi-Scale Transformer Approach

URL: View paper

Brief Assessment

Pneumonia Detection Transformer[35] focuses on medical image analysis for pneumonia detection using chest X-rays, not on decision-making or tool-usage tasks with visual inputs for conversational agents.

---

8. MedOrch: Medical Diagnosis with Tool-Augmented Reasoning Agents for Flexible Extensibility

URL: View paper

Brief Assessment

MedOrch[33] does not introduce datasets for decision-making or tool-using tasks with visual inputs. It focuses on medical diagnosis frameworks and does not present datasets comparable to ClevrPolicy or GTAPolicy.

---

9. Benchmarking vision, language, & action models on robotic learning tasks

URL: View paper

Brief Assessment

Vision Language Action Benchmark[32] focuses on evaluating vision-language-action models on robotic manipulation tasks using existing OpenX datasets, not on creating new datasets for decision-making and tool-using tasks with visual inputs.

---

10. Validation of computer vision-based ergonomic risk assessment tools for real manufacturing environments

URL: View paper

Brief Assessment

Ergonomic Risk Assessment[34] focuses on computer vision-based ergonomic risk assessment tools for manufacturing environments, not on datasets for decision-making and tool-using tasks with visual inputs. The candidate addresses workplace safety monitoring, while the original paper introduces datasets for training multimodal policy internalization methods in conversational agents.

Contribution 3: TriMPI training framework with PolicyRollout algorithm

Description: The authors propose TriMPI, a three-stage training framework consisting of visually-masked continual pretraining, chain-of-thought supervised finetuning, and reinforcement learning with PolicyRollout. PolicyRollout is a novel extension to GRPO-style RL algorithms that augments the rollout space with policy-aware responses to enable more grounded exploration during training.

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. Automated unit test generation via chain of thought prompt and reinforcement learning from coverage feedback

URL: View paper

Brief Assessment

Unit Test Generation[17] focuses on automated unit test generation for code using chain-of-thought prompts and reinforcement learning from coverage feedback (PPO optimization). The original paper addresses multimodal policy internalization for conversational agents with a three-stage framework (VM-CPT, CoT-SFT, RL with PolicyRollout). These are fundamentally different application domains and technical approaches.

---

2. Bridging formal language with chain-of-thought reasoning to geometry problem solving

URL: View paper

Brief Assessment

Geometry Problem Solving[24] focuses on geometry problem solving with formal language and chain-of-thought reasoning, not on multimodal policy internalization for conversational agents. The training framework and application domain are fundamentally different.

---

3. VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks

URL: View paper

Brief Assessment

VAPO[19] focuses on value-based reinforcement learning for mathematical reasoning tasks (AIME 2024), not on policy internalization for conversational agents. The frameworks address fundamentally different problems with different objectives.

---

4. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning

URL: View paper

Brief Assessment

High-Entropy Tokens[22] focuses on token-level entropy analysis and selective gradient updates in RLVR for mathematical reasoning, not on multimodal policy internalization with chain-of-thought finetuning and policy-aware rollout augmentation.

---

5. ARES: Alternating reinforcement learning and supervised fine-tuning for enhanced multi-modal chain-of-thought reasoning through diverse AI feedback

URL: View paper

Brief Assessment

ARES[20] focuses on alternating RL and SFT for multi-modal chain-of-thought reasoning using sentence-level AI feedback, not on policy internalization for conversational agents. The technical approaches and problem domains are fundamentally different.

---

6. S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models

URL: View paper

Brief Assessment

S-GRPO[25] focuses on reducing overthinking in reasoning models through early exit mechanisms during chain-of-thought generation, not on policy internalization for conversational agents. The frameworks address fundamentally different problems.

---

7. Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning

URL: View paper

Brief Assessment

RL of Thoughts[23] focuses on inference-time RL for general reasoning tasks using a navigator to select logic blocks, not on multimodal policy internalization with visually-masked continual pretraining and chain-of-thought supervised finetuning as in TriMPI.

---

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

URL: View paper

Brief Assessment

Unified Multimodal Reward[21] focuses on reward model training with chain-of-thought reasoning for multimodal preference learning, not policy internalization for conversational agents. The technical approaches differ fundamentally in objectives and methodology.

---

9. Demystifying long chain-of-thought reasoning in llms

URL: View paper

Brief Assessment

Demystifying Chain-of-Thought[16] focuses on scaling inference compute through long chain-of-thought reasoning in general LLMs using reinforcement learning methods like PPO. The original paper addresses multimodal policy internalization for conversational agents, which is a fundamentally different task involving internalizing complex multimodal policies into model parameters without requiring them at inference time.

---

10. AdaCoT: Pareto-Optimal Adaptive Chain-of-Thought Triggering via Reinforcement Learning

URL: View paper

Brief Assessment

AdaCoT[18] focuses on adaptive chain-of-thought triggering using PPO-based RL for computational efficiency, not on multimodal policy internalization with visually-masked continual pretraining and policy-aware rollout augmentation.

Appendix: Text Similarity Detection

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

References

- [0] Multimodal Policy Internalization for Conversational Agents View paper
- [1] MultiGen: Using Multimodal Generation in Simulation to Learn Multimodal Policies in Real View paper
- [2] RoboMP2: A Robotic Multimodal Perception-Planning Framework with Multimodal Large Language Models View paper
- [3] Integrating LMM Planners and 3D Skill Policies for Generalizable Manipulation View paper
- [4] MSR-Align: Policy-Grounded Multimodal Alignment for Safety-Aware Reasoning in Vision-Language Models View paper
- [5] Optimus-2: Multimodal minecraft agent with goal-observation-action conditioned policy View paper
- [6] Mutex: Learning unified policies from multimodal task specifications View paper
- [7] Multimodal task vectors enable many-shot multimodal in-context learning View paper
- [8] Robo-mutual: Robotic multimodal task specification via unimodal learning View paper
- [9] CUE-M: Contextual Understanding and Enhanced Search with Multimodal Large Language Model View paper
- [10] CtrlSynth: Controllable Image Text Synthesis for Data-Efficient Multimodal Learning View paper
- [11] Policy-driven, multimodal deep learning for predicting visual fields from the optic disc and OCT imaging View paper
- [12] Design of an Improved Model for Multimodal Data Fusion Using XGBoost-LSTM-CNN and Proximal Policy Optimizations View paper
- [13] QuadrupedGPT: Towards a Versatile Quadruped Agent in Open-ended Worlds View paper
- [14] LMNav: Landmark-Centric Multimodal Reasoning for Vision-And-Language Navigation in Continuous Environments View paper
- [15] Safe Learning for Contact-Rich Robot Tasks: A Survey from Classical Learning-Based Methods to Safe Foundation Models View paper
- [16] Demystifying long chain-of-thought reasoning in llms View paper
- [17] Automated unit test generation via chain of thought prompt and reinforcement learning from coverage feedback View paper
- [18] AdaCoT: Pareto-Optimal Adaptive Chain-of-Thought Triggering via Reinforcement Learning View paper
- [19] VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks View paper
- [20] ARES: Alternating reinforcement learning and supervised fine-tuning for enhanced multi-modal chain-of-thought reasoning through diverse AI feedback View paper
- [21] Unified multimodal chain-of-thought reward model through reinforcement fine-tuning View paper
- [22] Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning View paper
- [23] Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning View paper
- [24] Bridging formal language with chain-of-thought reasoning to geometry problem solving View paper
- [25] S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models View paper
- [26] A Benchmarking Study of Vision-based Robotic Grasping Algorithms View paper
- [27] VISION Datasets: A Benchmark for Vision-based InduStrial InspectiON View paper
- [28] Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology View paper
- [29] Ui-vision: A desktop-centric gui benchmark for visual perception and interaction View paper
- [30] Openthinkimg: Learning to think with images via visual tool reinforcement learning View paper
- [31] Visualtoolagent (vista): A reinforcement learning framework for visual tool selection View paper
- [32] Benchmarking vision, language, & action models on robotic learning tasks View paper
- [33] MedOrch: Medical Diagnosis with Tool-Augmented Reasoning Agents for Flexible Extensibility View paper
- [34] Validation of computer vision-based ergonomic risk assessment tools for real manufacturing environments View paper
- [35] Efficient and Accurate Pneumonia Detection Using a Novel Multi-Scale Transformer Approach View paper
- [36] Perception, reason, think, and plan: A survey on large multimodal reasoning models View paper
- [37] Multimodal chain-of-thought reasoning: A comprehensive survey View paper
- [38] Understand, Think, and Answer: Advancing Visual Reasoning with Large Multimodal Models View paper
- [39] R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization View paper
- [40] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models View paper
- [41] Struct2D: A Perception-Guided Framework for Spatial Reasoning in Large Multimodal Models View paper
- [42] Perception-aware policy optimization for multimodal reasoning View paper
- [43] Think Then Embed: Generative Context Improves Multimodal Embedding View paper
- [44] Boosting Reasoning in Large Multimodal Models via Activation Replay View paper
- [45] Critic-v: Vlm critics help catch vlm errors in multimodal reasoning View paper