Novelty Assessment Report

Paper: Masked Generative Policy for Robotic Control
PDF URL: https://openreview.net/pdf?id=KFu4p3pd11
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

We present Masked Generative Policy (MGP), a novel framework for visuomotor imitation learning. We represent actions as discrete tokens, and train a conditional masked transformer that generates tokens in parallel and then rapidly refines only low-confidence tokens. We further propose two new sampling paradigms: MGP-Short, which performs parallel masked generation with score-based refinement for Markovian tasks, and MGP-Long, which predicts full trajectories in a single pass and dynamically refines low-confidence action tokens based on new observations. With globally coherent prediction and robust adaptive execution capabilities, MGP-Long enables reliable control on complex and non-Markovian tasks that prior methods struggle with. Extensive evaluations on 150 robotic manipulation tasks spanning the Meta-World and LIBERO benchmarks show that MGP achieves both rapid inference and superior success rates compared to state-of-the-art diffusion and autoregressive policies. Specifically, MGP increases the average success rate by 9\% across 150 tasks while cutting per-sequence inference time by up to 35×. It further improves the average success rate by 60\% in dynamic and missing-observation environments, and solves two non-Markovian scenarios where other state-of-the-art methods fail. Further results and videos are available at: https://anonymous.4open.science/r/masked_generative_policy-8BC6.

Core Task Landscape

This paper addresses: **Visuomotor Imitation Learning for Robotic Manipulation**

A total of **50 papers** were analyzed and organized into a taxonomy with **33 categories**.

Taxonomy Overview

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

- **Demonstration Modality and Source**
- **Representation and Feature Learning**
- **Policy Architecture and Learning**
- **Generalization and Data Efficiency**
- **Task-Specific Manipulation Domains**
- **Execution and Deployment Optimization**
- **Benchmarking and Empirical Analysis**

Complete Taxonomy Tree

- Visuomotor Imitation Learning for Robotic Manipulation Survey Taxonomy
- Demonstration Modality and Source
  - Human Video Demonstrations
  - Third-Person Video Learning (4 papers)
    - [1] Watch and act: Learning robotic manipulation from visual demonstration (Shuo Yang, 2023) View paper
    - [12] Vision-based robot manipulation learning via human demonstrations (Jia Zhixin, 2020) View paper
    - [18] RoMaViD: Learning Robotic Manipulation from Video Demonstrations (Abhinav Upadhyay, 2024) View paper
    - [35] Learning by watching: A review of video-based learning approaches for robot manipulation (Chrisantus Eze, 2025) View paper
  - Eye-in-Hand Video Learning (1 papers)
    - [22] Giving robots a hand: Learning generalizable manipulation with eye-in-hand human video demonstrations (Kim, 2023) View paper
  - Bimanual Video Demonstrations (1 papers)
    - [9] You Only Teach Once: Learn One-Shot Bimanual Robotic Manipulation from Video Demonstrations (Zhou, 2025) View paper
  - Dexterous Manipulation from Video (1 papers)
    - [19] Vividex: Learning vision-based dexterous manipulation from human videos (Ze-Rui Chen, 2025) View paper
  - Robot Teleoperation and Direct Demonstration
  - Virtual Reality Teleoperation (1 papers)
    - [20] Deep imitation learning for complex manipulation tasks from virtual reality teleoperation (Tianhao Zhang, 2018) View paper
  - Kinesthetic and Direct Teaching (2 papers)
    - [17] Learning manipulation actions from human demonstrations (Tim Welschehold, 2016) View paper
    - [33] Teaching a robot manipulation skills through demonstration (Jeff Lieberman, 2004) View paper
  - Cross-Embodiment and Morphology Transfer (2 papers)
  - [2] Learning Robot Manipulation from Cross-Morphology Demonstration (Salhotra, 2023) View paper
  - [39] Mitigating the Human-Robot Domain Discrepancy in Visual Pre-training for Robotic Manipulation (Jiaming Zhou, 2024) View paper
  - Multimodal Demonstration Integration (2 papers)

    - [14] A multi-modal framework for robots to learn manipulation tasks from human demonstrations (Congcong Yin, 2023) View paper
    - [21] Teaching robots with show and tell: Using foundation models to synthesize robot policies from language and visual demonstration (M Murray, 2024) View paper
- Representation and Feature Learning
    - Object-Centric Representations (3 papers)
    - [3] Viola: Imitation learning for vision-based manipulation with object proposal priors (Zhu Yi-feng, 2023) View paper
    - [10] Viola: Object-centric imitation learning for vision-based robot manipulation (Y Zhu, 2022) View paper
    - [34] S-Diffusion: Generalizing from Instance-level to Category-level Skills in Robot Manipulation (Q Yang, 2025) View paper
    - Correspondence-Based Representations (1 papers)
    - [30] CordViP: Correspondence-based Visuomotor Policy for Dexterous Manipulation in Real-World (Yinghua Fu, 2025) View paper
    - Latent Space and Generative Models (2 papers)
    - [6] Learning robotic manipulation through visual planning and acting (Angelina Wang, 2019) View paper
    - [25] Latent Diffusion Planning for Imitation Learning (Xie, 2025) View paper
    - Concept and Language Grounding (2 papers)
    - [5] Concept2robot: Learning manipulation concepts from instructions and human demonstrations (Lin Shao, 2021) View paper
    - [16] Learning Compositional Behaviors from Demonstration and Language (Liu Weiyu, 2025) View paper
- Policy Architecture and Learning
    - Transformer-Based Policies (1 papers)
    - [15] The Art of Imitation: Learning Long-Horizon Manipulation Tasks From Few Demonstrations (Jan Ole von Hartz, 2024) View paper
    - Diffusion Policies (1 papers)
    - [36] DemoSpeedup: Accelerating Visuomotor Policies via Entropy-Guided Demonstration Acceleration (Guo Lingxiao, 2025) View paper
    - Discrete Token Generation ★ (1 papers)
    - [0] Masked Generative Policy for Robotic Control (Anon et al., 2026) View paper
    - Hierarchical and Compositional Policies (1 papers)
    - [4] Learning multi-step manipulation tasks from a single human demonstration (Guo, 2023) View paper
    - Recurrent and Sequential Policies (1 papers)
    - [28] Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration (Rahmatizadeh, 2018) View paper
    - Hybrid Imitation and Reinforcement Learning (2 papers)
    - [43] Learning Multiple Robot Manipulation Tasks with Imperfect Demonstrations (Jiahua Dai, 2023) View paper
    - [48] Learning Deformable Object Manipulation From Expert Demonstrations (Gautam Salhotra, 2022) View paper
- Generalization and Data Efficiency
    - Few-Shot and One-Shot Learning (3 papers)
    - [7] You only demonstrate once: Category-level manipulation from single visual demonstration (Bowen Wen, 2022) View paper
    - [8] Coarse-to-Fine Imitation Learning: Robot Manipulation from a Single Demonstration (Johns, 2021) View paper
    - [49] Roboclip: One demonstration is enough to learn robot policies (Sontakke, 2023) View paper
    - Spatial Generalization (1 papers)
    - [24] R2RGEN: Real-to-Real 3D Data Generation for Spatially Generalized Manipulation (Xu, 2025) View paper
    - Data Augmentation and Synthesis (1 papers)
    - [31] Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning (Xue, 2025) View paper
    - Covariate Shift and Distribution Mismatch (2 papers)
    - [32] Data efficient behavior cloning for fine manipulation via continuity-based corrective labels (Abhay, 2024) View paper
    - [50] Decomposing the Generalization Gap in Imitation Learning for Visual Robotic Manipulation (Annie Xie, 2024) View paper
- Task-Specific Manipulation Domains
    - Deformable Object Manipulation (1 papers)
    - [29] TieBot: Learning to Knot a Tie from Visual Demonstration through a Real-to-Sim-to-Real Approach (Peng, 2024) View paper
    - Dexterous Manipulation (1 papers)
    - [26] Dexterous manipulation through imitation learning: A survey (An Shan, 2025) View paper
    - Assembly and Long-Horizon Tasks (3 papers)
    - [38] Robot learning from demonstration in robotic assembly: A survey (Zuyuan Zhu, 2018) View paper
    - [44] Robot Learning Assembly Tasks from Human Demonstrations (Zuyuan Zhu, 2020) View paper
    - [45] Teaching robots to perform construction tasks via learning from demonstration (Ciâ□□Jyun Liang, 2019) View paper
    - Loco-Manipulation (3 papers)
    - [27] Learning a Unified Policy for Position and Force Control in Legged Loco-Manipulation (Zhi, 2025) View paper
    - [40] DemoHLM: From One Demonstration to Generalizable Humanoid Loco-Manipulation (Fu Yuhui, 2025) View paper
    - [41] Learning Visual Quadrupedal Loco-Manipulation from Demonstrations (Zhengmao He, 2024) View paper
    - Contact-Rich and Force-Controlled Manipulation (1 papers)
    - [46] Diff-lfd: Contact-aware model-based learning from visual demonstration for robotic manipulation via differentiable physics-based simulation and rendering (X Zhu, 2023) View paper
- Execution and Deployment Optimization
    - Execution Speed Optimization (1 papers)
    - [47] SAIL: Faster-than-Demonstration Execution of Imitation Learning Policies (Arachchige, 2025) View paper
    - Constrained and Safe Execution (1 papers)
    - [42] Leto: Learning constrained visuomotor policy with differentiable trajectory optimization (Zhengtong Xu, 2024) View paper
    - Target Localization and Attention (1 papers)
    - [37] Improving Learning from Visual Demonstration Methods by Target Localization (Pasquale Foggia, 2024) View paper
- Benchmarking and Empirical Analysis
    - Comparative Benchmarking Studies (1 papers)
    - [23] What matters in learning from offline human demonstrations for robot manipulation (Mandlekar, 2021) View paper
    - Survey and Review Papers (1 papers)

- [11] Visuomotor Policy Learning for Predictive Manipulation (Jayasimha, 2021) View paper
- Acceleration and Efficiency Analysis (1 papers)
- [13] Accelerating robot manipulation using demonstrations (Unknown, 2024) View paper

Narrative

Core task: visuomotor imitation learning for robotic manipulation. The field organizes around several complementary dimensions. Demonstration Modality and Source addresses how robots acquire training data—ranging from teleoperation and VR interfaces (Deep Imitation VR[20]) to cross-embodiment transfer (Cross-Morphology Demonstration[2]) and even single-demonstration methods (You Only Demonstrate Once[7]). Representation and Feature Learning explores how visual observations are encoded, including object-centric approaches (Viola Object-Centric[10]) and pre-trained vision models (Roboclip[49]). Policy Architecture and Learning encompasses the core algorithmic choices: end-to-end networks (Multi-Task End-to-End[28]), hierarchical planners (Visual Planning Acting[6]), diffusion-based policies (Diff-LfD[46]), and discrete token generation methods. Generalization and Data Efficiency investigates how policies transfer across tasks and environments, while Task-Specific Manipulation Domains targets challenges like assembly (Assembly from Demonstrations[44]) and deformable objects (Deformable Object Manipulation[48]). Execution and Deployment Optimization refines real-time performance, and Benchmarking and Empirical Analysis provides systematic evaluation frameworks.

Within Policy Architecture and Learning, a particularly active line explores discrete token generation as an alternative to continuous action prediction. Masked Generative Policy[0] exemplifies this direction by framing action sequences as masked token prediction, drawing inspiration from language modeling techniques. This contrasts with diffusion-based approaches (S-Diffusion[34], Latent Diffusion Planning[25]) that model action distributions through iterative denoising, and with hierarchical methods (Coarse-to-Fine Imitation[8]) that decompose policies into multiple stages. The discrete tokenization strategy offers potential advantages in sample efficiency and interpretability, positioning Masked Generative Policy[0] alongside recent efforts to leverage transformer architectures for sequential decision-making. Meanwhile, works like You Only Teach Once[9] and DemoHLM[40] emphasize learning from minimal or structured demonstrations, highlighting ongoing tensions between data requirements and generalization capability across the broader policy learning landscape.

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

Discrete Token Generation sits within a family of policy architecture approaches for visuomotor imitation learning, distinguished by its representation of actions as discrete tokens processed through masked or autoregressive generation. While siblings like Diffusion Policies and Transformer-Based Policies use continuous action spaces or direct transformer outputs, Discrete Token Generation quantizes the action space and applies sequence generation techniques. This approach shares the goal of learning manipulation policies from demonstrations but differs fundamentally in how actions are represented and generated.

Similarities: - All subtopics address policy architecture design for visuomotor imitation learning in robotic manipulation - Each approach learns from demonstration data to predict actions conditioned on visual observations - All methods aim to capture complex action distributions and temporal dependencies in manipulation tasks - Transformer-Based Policies and Discrete Token Generation both leverage attention mechanisms, though applied differently

Differences: - Discrete Token Generation uses quantized action spaces with discrete tokens, while Diffusion Policies, Transformer-Based Policies, and Recurrent policies typically operate in continuous action spaces - Generation mechanism varies: autoregressive/masked generation for discrete tokens vs. iterative denoising for diffusion vs. direct regression for transformers/recurrent networks - Hierarchical and Compositional Policies explicitly decompose tasks into skills, while Discrete Token Generation learns flat token sequences - Hybrid Imitation and Reinforcement Learning combines multiple learning paradigms, whereas Discrete Token Generation focuses purely on imitation with a specific representation choice - Recurrent architectures maintain hidden states across time, while Discrete Token Generation relies on attention over token sequences without persistent memory

Suggested Search Directions: - Investigate whether discrete tokenization provides better sample efficiency or generalization compared to continuous action representations - Explore hybrid approaches combining discrete token generation with hierarchical decomposition or reinforcement learning - Compare computational efficiency and inference speed between autoregressive token generation and diffusion-based action sampling - Examine how discrete vs. continuous action representations affect multi-modal action distribution modeling

Sibling Subtopics

- **Diffusion Policies** (leaves: 1, papers: 1)
- Scope: Methods employing diffusion models for generating action sequences or trajectories.
- Exclude: Excludes transformer or discrete token methods; see Transformer-Based Policies or Discrete Token Generation.
- **Hierarchical and Compositional Policies** (leaves: 1, papers: 1)
- Scope: Methods decomposing tasks into hierarchical skills or compositional action primitives.
- Exclude: Excludes flat, end-to-end policies; see other Policy Architecture subcategories.
- **Hybrid Imitation and Reinforcement Learning** (leaves: 1, papers: 2)
- Scope: Methods combining imitation learning with reinforcement learning for policy optimization.
- Exclude: Excludes pure imitation learning; see other Policy Architecture subcategories.
- **Recurrent and Sequential Policies** (leaves: 1, papers: 1)
- Scope: Methods using recurrent architectures for sequential decision-making in manipulation tasks.
- Exclude: Excludes transformer or diffusion architectures; see Transformer-Based Policies or Diffusion Policies.
- **Transformer-Based Policies** (leaves: 1, papers: 1)
- Scope: Methods using transformer architectures as the primary policy network for action prediction.
- Exclude: Excludes diffusion, autoregressive token-based, or other architectures; see Diffusion Policies or Discrete Token Generation.

Contributions Analysis

Overall novelty summary. The paper introduces Masked Generative Policy (MGP), which represents actions as discrete tokens and employs a conditional masked transformer for parallel generation with iterative refinement. According to the taxonomy, this work resides in the 'Discrete Token Generation' leaf under 'Policy Architecture and Learning'. Notably, this leaf contains only the original paper itself—no sibling papers are listed. This suggests the discrete token generation approach for visuomotor manipulation is a relatively sparse research direction within the taxonomy's 50-paper scope, contrasting with more populated areas like diffusion policies or transformer-based methods.

The taxonomy reveals neighboring approaches in adjacent leaves: 'Diffusion Policies' (1 paper), 'Transformer-Based Policies' (1 paper), and 'Hierarchical and Compositional Policies' (1 paper). The scope note for 'Discrete Token Generation' explicitly excludes continuous action diffusion and transformer policies, positioning MGP as an alternative to these paradigms. The broader 'Policy Architecture and Learning' branch also includes recurrent architectures and hybrid imitation-RL methods, indicating diverse algorithmic strategies. MGP's

masked generation mechanism appears to occupy a distinct niche between autoregressive token models and continuous diffusion approaches, though the taxonomy structure suggests limited prior exploration of this specific combination.

Among the three contributions analyzed, the literature search examined 10 candidates total. The core MGP framework (Contribution 1) had 4 candidates examined with 0 refutable, while MGP-Long with Adaptive Token Refinement (Contribution 3) examined 6 candidates with 0 refutable. MGP-Short (Contribution 2) had no candidates examined. The absence of refutable prior work across all contributions, combined with the limited search scope of 10 papers, suggests these specific mechanisms—parallel masked generation with score-based refinement and dynamic trajectory refinement—may not have direct precedents in the examined literature. However, this reflects the bounded search rather than exhaustive field coverage.

Based on the 10-candidate search and sparse taxonomy positioning, MGP appears to introduce a relatively unexplored combination of techniques within the surveyed literature. The lack of sibling papers in its taxonomy leaf and zero refutable candidates across contributions indicate potential novelty, though the limited search scope (10 papers from semantic search) means substantial related work may exist outside this analysis. The taxonomy's explicit exclusion of diffusion and transformer methods from the discrete token leaf further suggests MGP occupies a distinct methodological space, though comprehensive field coverage would require broader literature examination.

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

Contribution 1: Masked Generative Policy (MGP) framework for visuomotor imitation learning

Description: The authors introduce MGP, a new framework that represents robot actions as discrete tokens and uses a conditional masked transformer to generate these tokens in parallel with selective refinement of low-confidence tokens. This approach aims to overcome the inference bottlenecks of diffusion models and the sequential constraints of autoregressive models.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Enhancing Offline Reinforcement Learning with Decision Transformers: Evaluating Performance Across Simulated Robotic Control Tasks

URL: View paper

Brief Assessment

Decision Transformers Evaluation[60] focuses on evaluating decision transformers for offline RL in simulated robotic control tasks, not on masked generative transformers for visuomotor imitation learning with discrete action tokens and parallel generation.

---

2. Transformer-Based Sequence Modeling with Action Discretization for Robotic Grasping

URL: View paper

Brief Assessment

Action Discretization Grasping[59] focuses on discretizing actions by dimension for autoregressive modeling in robotic grasping tasks, whereas the original paper introduces a masked generative framework with parallel token generation and selective refinement. The candidate does not use masked transformers or parallel generation mechanisms that are central to MGP.

---

3. Sample-efficient Imitative Multi-token Decision Transformer for Real-world Driving

URL: View paper

Brief Assessment

Multi-token Decision Transformer[57] focuses on autonomous driving with real-world vehicle dynamics and physics-informed networks, not visuomotor robotic manipulation with masked transformers and parallel token generation.

---

4. Keypoint Action Tokens Enable In-Context Imitation Learning in Robotics

URL: View paper

Brief Assessment

Keypoint Action Tokens[58] focuses on in-context learning using pre-trained text transformers without training, operating on keypoint-based visual representations. MGP uses masked transformers trained specifically on robot action tokens with iterative refinement for visuomotor control, representing a fundamentally different technical approach.

Contribution 2: MGP-Short sampling paradigm for Markovian tasks

Description: The authors develop MGP-Short, a sampling method that performs parallel masked token generation with score-based refinement specifically designed for Markovian manipulation tasks. This method achieves rapid inference while maintaining high success rates on standard benchmarks.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: MGP-Long sampling paradigm with Adaptive Token Refinement for non-Markovian tasks

Description: The authors propose MGP-Long, which predicts complete action trajectories in one pass and then dynamically refines low-confidence tokens using new observations through an Adaptive Token Refinement strategy. This enables globally coherent predictions and robust execution for complex, long-horizon, and non-Markovian manipulation tasks.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. MVP: Memory-enhanced Vision-Language-Action Policy with Feedback Learning

URL: View paper

Brief Assessment

MVP Feedback Learning[56] focuses on memory-enhanced vision-language-action policies using episodic memory and SO(3) trajectory perturbation for feedback learning in robotic manipulation. The original paper's MGP-Long addresses non-Markovian tasks through masked generative transformers with adaptive token refinement for action sequences, which is a fundamentally different technical approach from MVP's memory-based policy learning with visual-language integration.

---

2. Adaptive Progressive Transformer-Based Trajectory Prediction Under Fine-Grained Trajectory-Scene Interaction Constraint

URL: View paper

Brief Assessment

Adaptive Progressive Transformer[55] focuses on trajectory prediction in autonomous driving scenarios with scene interaction constraints, not robotic manipulation with action token refinement for non-Markovian tasks.

---

3. Augmented transformer with adaptive graph for temporal action proposal generation

URL: View paper

Brief Assessment

Adaptive Graph Transformer[53] focuses on temporal action proposal generation in videos using transformers and graph networks for capturing temporal context. This is fundamentally different from MGP-Long's robotic manipulation trajectory prediction with dynamic token refinement for non-Markovian control tasks.

---

4. Multilevel semantic and adaptive actionness learning for weakly supervised temporal action localization.

URL: View paper

Brief Assessment

Adaptive Actionness Learning[54] focuses on weakly supervised temporal action localization in videos, which is a video understanding task. This is fundamentally different from MGP-Long's robotic manipulation and trajectory prediction domain.

---

5. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation

URL: View paper

Brief Assessment

MemoryVLA[51] addresses temporal dependencies through a memory bank mechanism for VLA models in robotic manipulation, while the original paper proposes MGP-Long for trajectory prediction with token refinement in visuomotor control. The candidate focuses on vision-language-action models with hippocampus-inspired memory systems, whereas the original introduces masked generative transformers for action token generation. These represent distinct technical approaches to handling temporal information in robotics.

---

6. Smart: Scalable multi-agent real-time motion generation via next-token prediction

URL: View paper

Brief Assessment

Smart[52] focuses on multi-agent motion generation for autonomous driving simulation using next-token prediction with decoder-only transformers. The original paper addresses robotic manipulation with adaptive token refinement for action sequences, which is a fundamentally different domain and task structure.

Appendix: Text Similarity Detection

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

References

- [0] Masked Generative Policy for Robotic Control View paper
- [1] Watch and act: Learning robotic manipulation from visual demonstration View paper
- [2] Learning Robot Manipulation from Cross-Morphology Demonstration View paper
- [3] Viola: Imitation learning for vision-based manipulation with object proposal priors View paper
- [4] Learning multi-step manipulation tasks from a single human demonstration View paper
- [5] Concept2robot: Learning manipulation concepts from instructions and human demonstrations View paper
- [6] Learning robotic manipulation through visual planning and acting View paper
- [7] You only demonstrate once: Category-level manipulation from single visual demonstration View paper
- [8] Coarse-to-Fine Imitation Learning: Robot Manipulation from a Single Demonstration View paper
- [9] You Only Teach Once: Learn One-Shot Bimanual Robotic Manipulation from Video Demonstrations View paper
- [10] Viola: Object-centric imitation learning for vision-based robot manipulation View paper
- [11] Visuomotor Policy Learning for Predictive Manipulation View paper
- [12] Vision-based robot manipulation learning via human demonstrations View paper
- [13] Accelerating robot manipulation using demonstrations View paper
- [14] A multi-modal framework for robots to learn manipulation tasks from human demonstrations View paper
- [15] The Art of Imitation: Learning Long-Horizon Manipulation Tasks From Few Demonstrations View paper
- [16] Learning Compositional Behaviors from Demonstration and Language View paper
- [17] Learning manipulation actions from human demonstrations View paper
- [18] RoMaViD: Learning Robotic Manipulation from Video Demonstrations View paper
- [19] Vividex: Learning vision-based dexterous manipulation from human videos View paper
- [20] Deep imitation learning for complex manipulation tasks from virtual reality teleoperation View paper
- [21] Teaching robots with show and tell: Using foundation models to synthesize robot policies from language and visual demonstration View paper
- [22] Giving robots a hand: Learning generalizable manipulation with eye-in-hand human video demonstrations View paper
- [23] What matters in learning from offline human demonstrations for robot manipulation View paper
- [24] R2RGEN: Real-to-Real 3D Data Generation for Spatially Generalized Manipulation View paper
- [25] Latent Diffusion Planning for Imitation Learning View paper
- [26] Dexterous manipulation through imitation learning: A survey View paper
- [27] Learning a Unified Policy for Position and Force Control in Legged Loco-Manipulation View paper
- [28] Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration View paper
- [29] TieBot: Learning to Knot a Tie from Visual Demonstration through a Real-to-Sim-to-Real Approach View paper
- [30] CordViP: Correspondence-based Visuomotor Policy for Dexterous Manipulation in Real-World View paper
- [31] Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning View paper
- [32] Data efficient behavior cloning for fine manipulation via continuity-based corrective labels View paper
- [33] Teaching a robot manipulation skills through demonstration View paper
- [34] S-Diffusion: Generalizing from Instance-level to Category-level Skills in Robot Manipulation View paper
- [35] Learning by watching: A review of video-based learning approaches for robot manipulation View paper
- [36] DemoSpeedup: Accelerating Visuomotor Policies via Entropy-Guided Demonstration Acceleration View paper
- [37] Improving Learning from Visual Demonstration Methods by Target Localization View paper

- [38] Robot learning from demonstration in robotic assembly: A survey View paper
- [39] Mitigating the Human-Robot Domain Discrepancy in Visual Pre-training for Robotic Manipulation View paper
- [40] DemoHLM: From One Demonstration to Generalizable Humanoid Loco-Manipulation View paper
- [41] Learning Visual Quadrupedal Loco-Manipulation from Demonstrations View paper
- [42] Leto: Learning constrained visuomotor policy with differentiable trajectory optimization View paper
- [43] Learning Multiple Robot Manipulation Tasks with Imperfect Demonstrations View paper
- [44] Robot Learning Assembly Tasks from Human Demonstrations View paper
- [45] Teaching robots to perform construction tasks via learning from demonstration View paper
- [46] Diff-lfd: Contact-aware model-based learning from visual demonstration for robotic manipulation via differentiable physics-based simulation and rendering View paper
- [47] SAIL: Faster-than-Demonstration Execution of Imitation Learning Policies View paper
- [48] Learning Deformable Object Manipulation From Expert Demonstrations View paper
- [49] Roboclip: One demonstration is enough to learn robot policies View paper
- [50] Decomposing the Generalization Gap in Imitation Learning for Visual Robotic Manipulation View paper
- [51] Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation View paper
- [52] Smart: Scalable multi-agent real-time motion generation via next-token prediction View paper
- [53] Augmented transformer with adaptive graph for temporal action proposal generation View paper
- [54] Multilevel semantic and adaptive actionness learning for weakly supervised temporal action localization. View paper
- [55] Adaptive Progressive Transformer-Based Trajectory Prediction Under Fine-Grained Trajectory-Scene Interaction Constraint View paper
- [56] MVP: Memory-enhanced Vision-Language-Action Policy with Feedback Learning View paper
- [57] Sample-efficient Imitative Multi-token Decision Transformer for Real-world Driving View paper
- [58] Keypoint Action Tokens Enable In-Context Imitation Learning in Robotics View paper
- [59] Transformer-Based Sequence Modeling with Action Discretization for Robotic Grasping View paper
- [60] Enhancing Offline Reinforcement Learning with Decision Transformers: Evaluating Performance Across Simulated Robotic Control Tasks View paper