NEURIPS2025

Posters Report

Generated: January 06, 2026 at 05:22

Conference Highlights

Main Themes:

1. **Autonomous Systems**: Many papers focus on developing autonomous systems, including autonomous driving, robotics, and decision-making under uncertainty.
2. **Reinforcement Learning**: Several papers explore reinforcement learning techniques, including preference-based reinforcement learning, and applying reinforcement learning to complex tasks like autonomous driving.
3. **Deep Learning**: Deep learning models are used in various papers, including graph neural networks, vision-language models, and foundation models for scientific discovery.
4. **Simulation and Benchmarking**: Several papers focus on developing realistic simulators and benchmarks for autonomous systems, including driving world models and robotic manipulation simulators.
5. **Generalization and Transfer**: Some papers address the challenge of generalization and transfer in autonomous systems, including generalization in behavioral cloning and preference alignment.

Key Technologies:

1. **Graph Neural Networks (GNNs)**: GNNs are used in several papers, including enforcing convex constraints in GNNs and teaching LLMs to reason on graphs with reinforcement learning.
2. **Diffusion Models**: Diffusion models are used in several papers, including discrete diffusion models, and improving discrete diffusion with masking and adaptive chunking.
3. **Foundation Models**: Foundation models, such as GPT-4 and AlphaFold, are used in several papers, including understanding their role in scientific discovery and applying them to preference-based reinforcement learning.
4. **Reinforcement Learning (RL)**: RL is used in several papers, including preference-based reinforcement learning, and applying RL to complex tasks like autonomous driving.
5. **Deep Learning**: Deep learning models are used in various papers, including vision-language models, and foundation models for scientific discovery.

Emerging Trends:

1. **Autonomous Systems**: Autonomous systems are becoming increasingly important, with many papers focusing on developing autonomous driving, robotics, and decision-making under uncertainty.
2. **Reinforcement Learning**: Reinforcement learning is becoming a key technique for developing autonomous systems, with several papers exploring its applications.
3. **Deep Learning**: Deep learning models are being used in various papers, including graph neural networks, vision-language models, and foundation models for scientific discovery.
4. **Simulation and Benchmarking**: Realistic simulators and benchmarks are becoming increasingly important for autonomous systems, with several papers focusing on developing them.
5. **Generalization and Transfer**: Generalization and transfer are becoming increasingly important challenges in autonomous systems, with several papers addressing them.

Based on analysis of 70 papers

Statistics

Total Papers	70

Papers with PDFs	70 (100.0%)

All Papers (70)

[1] SutureBot: A Precision Framework & Benchmark For Autonomous End-to-End Suturing

Authors: Jesse Haworth, Joo-Tung Chen, Nigel J. W. ?, Massoud ?, Christina ?, Axel ?
PDF: https://arxiv.org/pdf/2510.20965

Overview

Problem Statement
The paper addresses the problem of autonomous end-to-end suturing, a long-horizon dexterous manipulation task that requires coordinated needle grasping, precise tissue penetration, and secure knot tying. This task is crucial in robotic surgery, but current approaches have limitations, such as lack of generalization and error recovery.

Key Contributions
The paper introduces SutureBot, an autonomous suturing benchmark on the da Vinci Research Kit (dVRK), which includes a high-fidelity dataset of 1,890 suturing demonstrations. The authors also propose a goal-conditioned framework that optimizes insertion-point precision, improving targeting accuracy by 59%-74% over a task-only baseline. This framework is evaluated on state-of-the-art vision-language-action (VLA) models, including ?0, GR00T N1, OpenVLA-OFT, and multitask ACT.

Methodology
The SutureBot framework consists of a high-level language policy, which selects the current task and generates the associated language condition, and a low-level policy, which produces precise, continuous control commands for the robot. The framework is trained on the high-fidelity dataset and evaluated on the dVRK. The authors also release a benchmark for dexterous imitation learning, which can be used to track progress in autonomous suturing.

Results
The goal-conditioned framework achieves significant improvements in targeting accuracy, with a 59%-74% improvement over the task-only baseline. The framework is also evaluated on state-of-the-art VLA models, which demonstrate robustness to disturbances and adaptability to unpredictable tissue interactions.

Significance
This work matters because it addresses a critical challenge in robotic surgery, autonomous end-to-end suturing. The SutureBot framework and benchmark provide a reproducible platform for evaluating and developing precision-focused, long-horizon dexterous manipulation policies. The high-fidelity dataset and goal-conditioned framework can be used to advance the field of robotic surgery and improve patient outcomes.

---

[2] HiMaCon: Discovering Hierarchical Manipulation Concepts from Unlabeled Multi-Modal Data

Authors: Ruihe Lui, Pai Zhou, Qian Lui, Li Sun, Jian Gen, Yibing Song, Yanchao Yang
PDF: https://arxiv.org/pdf/2510.11321v2.pdf

Overview

Problem Statement
The paper addresses the challenge of generalization in robotic manipulation, where policies trained on specific tasks and environments fail to adapt to novel scenarios. This is a critical issue in robotics, as it limits the deployment of robots in real-world settings. The authors propose that learning transferable manipulation concepts can help bridge this generalization gap.

Key Contributions

The main contributions of this work are:

* A self-supervised framework for learning hierarchical manipulation concepts from unlabeled multi-modal data.
* A dual-structure approach that combines cross-modal correlation learning and multi-horizon sub-goal organization.
* A policy enhancement approach that integrates learned concepts into policy learning through joint prediction.

Methodology

The proposed framework consists of two stages:
1. Concept encoder: processes multi-modal robot demonstrations to extract concept latents.
2. Concept refinement: uses two objectives to refine latents: Cross-Modal Correlation Network (C) and Multi-Horizon Future Predictor (F).
3. Policy enhancement: integrates learned concepts into policy learning through a backbone network with concept and action prediction heads.

Results

The authors demonstrate significant performance improvements in simulated benchmark tasks and real-world robot deployments. Specifically:

* Policies enhanced with manipulation concepts outperform conventional approaches in challenging scenarios.
* Learned concepts form interpretable clusters that resemble meaningful manipulation primitives.

Significance

This work advances the understanding of representation learning for manipulation and provides a practical approach to enhancing robotic performance in complex scenarios. The proposed framework can be applied to various robotic tasks and environments, and its impact could be significant in areas such as robotics, artificial intelligence, and human-robot interaction.

---

[3] Failure Prediction at Runtime for Generative Robot Policies

Authors: Ralf R"omer, Adrian K"obas, Luca W"orl, Angela P. Schoelig

PDF: https://arxiv.org/pdf/2510.09459v2.pdf

Overview

Problem Statement

The paper addresses the problem of predicting task failures in generative robot policies during runtime, without relying on labeled failure data. This is crucial in human-centered and safety-critical environments, where early failure prediction enables timely intervention or safe fallbacks. The challenge lies in the vast range of potential failure modes and the difficulty of generating examples of failures.

Key Contributions

The main contributions of this work are:

* FIPER (Failure Prediction at Runtime), a lightweight framework for predicting failures of generative robot policies that requires no examples of failures.
* Two key indicators of impending failure: out-of-distribution (OOD) observations detected via random network distillation (RND) and high uncertainty in generated actions measured by action-chunk entropy (ACE).
* Statistical calibration of uncertainty scores using conformal prediction on a small set of successful rollouts.

Methodology

FIPER uses RND to detect OOD observations in the policy's observation embedding space and ACE to measure uncertainty in generated actions. The uncertainty scores are calibrated using conformal prediction on a small set of successful rollouts. At runtime, FIPER raises an alarm when both scores exceed their respective thresholds.

Results

The paper evaluates FIPER across five simulation and real-world environments, demonstrating that it:

* Achieves higher accuracy and lower detection time compared to state-of-the-art baselines.

* Can distinguish actual failures from OOD situations more effectively.
* Is a promising step towards more interpretable and safer deployment of generative robot policies.

Significance
This work matters because it addresses a critical challenge in robotics: predicting task failures in generative robot policies. FIPER's ability to predict failures accurately and early can enable safer deployment of robots in human-centered environments, reducing the risk of accidents and improving overall safety.

---

[4] SVRPBench: A Realistic Benchmark for Stochastic Vehicle Routing Problems

Authors: Ahmed Hossan, Yahya Saadatshiraz, Martin Tukic, Salem LaHou, Zangir Kikasov
PDF: https://arxiv.org/pdf/2505.21887v2.pdf

Overview

Problem Statement
The paper addresses the Stochastic Vehicle Routing Problem (SVRP), which is a critical challenge in modern logistics and last-mile delivery. The problem involves planning routes for vehicles under uncertain and dynamic conditions, such as time-dependent congestion, random incidents, and diverse delivery preferences. The importance of this problem lies in its impact on urban logistics, where ignoring these factors can lead to overly optimistic performance assessments and misdirected algorithmic development.

Key Contributions
The main contributions of this work are:

* **Stochastic Realism**: The introduction of a novel benchmark suite, SVRPBench, which simulates realistic logistics scenarios with embedded uncertainty.
* **Constraint-Rich Instance Generation**: The development of a framework that supports multi-depot and multi-vehicle setups, strict capacity constraints, and diverse time window widths.
* **Diverse Baseline Evaluation**: The benchmarking of classical heuristics, metaheuristics, industrial solvers, and learning-based methods to highlight the effects of stochastic conditions on solution quality, feasibility, and robustness.

Methodology
The paper employs a range of techniques, including:

* **Time-Dependent Travel Time Modeling**: The use of Gaussian mixtures to model time-dependent congestion, log-normal distributions to model stochastic travel time delays, and Poisson processes to model accident-induced disruptions.
* **Customer Time Window Sampling**: The use of bimodal Gaussian mixtures to sample residential and commercial customer time windows.

Results
The benchmarking results show that state-of-the-art RL solvers degrade by over 20% under distributional shift, while classical and metaheuristic methods remain robust.

Significance
This work matters because it provides a realistic benchmark for the SVRP, which is essential for developing robust and deployable routing algorithms suited for real-world logistics. The release of the dataset and evaluation suite through a public repository will support reproducible research and foster future contributions.

---

[5] AutoDiscovery: Open-ended Scientific Discovery via Bayesian Surprise

Authors: Dhruv Agarwal, Roothsastiva Prasad Maumder, Reece Adamson, Megha Chakraborty, Satvika Reddy, Aditya Parasuram, Hanshith Surana, Bhavna Dalw, Mishra, Ashish Sabharwal, Peter Clark, Andrew McCallum
PDF: https://arxiv.org/pdf/2507.00310v2.pdf

Overview

Problem Statement

The paper addresses the problem of open-ended autonomous scientific discovery (ASD), where an AI system explores and generates hypotheses without human guidance. The goal is to accelerate scientific discovery by allowing the AI system to drive exploration by its own criteria, rather than relying on human-specified research questions.

Key Contributions

The main contributions of this work are:

* **AUTODISCOVERY**: a method for open-ended ASD that uses Bayesian surprise to drive scientific exploration.
* **Bayesian Surprise**: a measure of how data affects an observer's belief distributions, used to quantify the surprisal of a hypothesis.
* **Monte-Carlo Tree Search (MCTS)**: a procedure with progressive widening to balance exploration and exploitation of the vast hypothesis search space.

Methodology

The AUTODISCOVERY method uses an LLM model as the Bayesian observer to compute prior and posterior distributions about hypotheses. The MCTS procedure with progressive widening is used to sample hypotheses with high surprisal. The method is evaluated on 21 real-world datasets spanning domains such as biology, economics, finance, and behavioral science.

Results

The results show that AUTODISCOVERY substantially outperforms competitors by producing 5-29% more discoveries deemed surprising by the LLM. Human evaluation reveals that two-thirds of discoveries made by the system are surprising to domain experts as well.

Significance

This work matters because it addresses the key challenge of search in open-ended ASD. By using Bayesian surprise as a reward function, AUTODISCOVERY provides a principled mechanism to balance exploration and exploitation of the vast hypothesis search space. The results demonstrate the potential of AUTODISCOVERY to accelerate scientific discovery and make a significant impact on the field of ASD.

---

[6] Learning Robust Visuomotor Policies by Staying In-Distribution

Authors: Zhanyi Sun, Shuran Song
PDF: https://arxiv.org/pdf/2508.05941v1.pdf

Overview

Problem Statement

The paper addresses the problem of covariate shift in visuomotor policy learning, where small deviations from expert trajectories can quickly compound and cause task failures. This is a critical issue in robotics and autonomous systems, where precise imitation of expert behavior is essential for reliable manipulation.

Key Contributions

The main contributions of this work are:

* **Latent Policy Barrier (LPB)**: a framework for robust visuomotor policy learning that decouples precise expert imitation from out-of-distribution (OOD) correction.
* **Dynamics Model**: a learned model that predicts future latent states and optimizes them to stay within the expert distribution.

* **Policy Steering**: a latent-space steering approach that simultaneously achieves high task performance and robustness.

Methodology
The LPB framework consists of two complementary components:

* **Base Diffusion Policy**: trained exclusively on consistent, high-quality expert demonstrations.
* **Action-conditioned Visual Latent Dynamics Model**: trained on a mixed-quality dataset combining expert demonstrations and automatically generated rollout data.

Results
Experimental results across simulated and real-world manipulation tasks demonstrate that LPB:

* **Improves Sample Efficiency**: by decoupling expert imitation from OOD correction, enabling the policy to focus on learning from a small amount of high-quality human demonstrations.
* **Enhances Robustness**: through the use of a dynamics model trained on inexpensive, lower-quality policy rollout data.
* **Improves Performance**: in both simulated and real-world tasks, with significant improvements in robustness and sample efficiency.

Significance
This work matters because it provides a novel solution to the covariate shift problem in visuomotor policy learning, enabling more reliable and efficient manipulation in robotics and autonomous systems. The LPB framework is plug-and-play compatible with off-the-shelf pre-trained policies, improving their robustness without requiring policy retraining or fine-tuning.

---

[7] Optimal Estimation of the Best Mean in Multi-Armed Bandits

Authors: Takayuki Osogami, Jinya Hoda
PDF: https://openreview.net/pdf/cc07eb5b11b18db0faf89a07e94498a5c4f2bea5.pdf

Overview

Problem Statement
The paper addresses the problem of estimating the mean reward of the best arm in a multi-armed bandit (MAB) setting. This is a fundamental problem in decision-making under uncertainty, where an agent must choose the best action among multiple options with unknown rewards. The goal is to estimate the mean reward of the best arm with a target precision ? and confidence level 1 ? ?, while minimizing the number of samples.

Key Contributions
The paper makes several key contributions:

* Establishes an instance-dependent lower bound on the sample complexity, which is expressed as a non-convex optimization problem.
* Introduces a novel algorithm that achieves the asymptotic lower bound with matching constants in the leading term.
* Proposes a two-phase sampling strategy that adapts to the structure of each instance and achieves the corresponding sample complexity.

Methodology
The algorithm uses a martingale-based anytime confidence bound to jointly estimate the expected rewards of arms as a confidence ellipsoid. The key insight is that this ellipsoidal representation enables efficient testing of global hypotheses. The algorithm also conducts a detailed analysis to reduce the original non-convex characterizing optimization problem to a one-dimensional non-convex optimization over a narrow interval.

Results
The paper shows that the algorithm achieves asymptotically optimal sample complexity, with an expected number of

samples that matches the theoretical lower bound. The algorithm also guarantees an estimation error of at most ? with probability at least 1 ? ?.

Significance
This work matters because it provides a novel algorithm for best-mean estimation in MAB settings, which is a fundamental problem in decision-making under uncertainty. The algorithm's asymptotic optimality and adaptability to instance structure make it a significant contribution to the field. The practical relevance of the approach is also demonstrated through numerical experiments.

---

[8] Fast Monte Carlo Tree Diffusion: 100x Speedup via Parallel Sparse Planning

Authors: Jaeik Yoon, Hyeonseok Cho, Yoshua Bengio, Sungjin Ahn
PDF: https://arxiv.org/pdf/2506.09498v4.pdf

Overview

Problem Statement
The paper addresses the challenge of improving the efficiency of Monte Carlo Tree Diffusion (MCTD), a powerful approach for trajectory planning, while maintaining its strong planning capabilities. MCTD is a diffusion-based method that integrates tree search with diffusion-based planning, but its sequential nature and iterative denoising process incur significant computational overhead, limiting its practical effectiveness in challenging long-horizon tasks.

Key Contributions
The main contributions of this paper are the introduction of Fast-MCTD, a framework that improves the efficiency of MCTD through parallelization and rollout sparsification. Fast-MCTD integrates two key optimization techniques: Parallel MCTD (P-MCTD) and Sparse MCTD (S-MCTD). P-MCTD accelerates the tree search process by enabling parallel rollouts, deferring tree updates until multiple searches are completed, introducing redundancy-aware selection, and parallelizing both expansion and simulation steps. S-MCTD further improves efficiency by planning over coarsened trajectories, using diffusion models trained on these compressed representations.

Methodology
The paper proposes Fast-MCTD, which combines two techniques: Parallel MCTD (P-MCTD) and Sparse MCTD (S-MCTD). P-MCTD enables parallel rollouts via delayed tree updates and redundancy-aware selection, while S-MCTD reduces rollout length through trajectory coarsening. The paper also uses Diffusion Forcing, a token-level denoising control within trajectories, and Diffuser, a generative denoising process over full trajectories.

Results
The experimental results show that Fast-MCTD achieves substantial speedups?up to 100× on some tasks?compared to standard MCTD, while maintaining comparable or superior planning performance. Fast-MCTD also outperforms Diffuser in inference speed on some tasks, despite Diffuser's lack of search and its substantially inferior performance.

Significance
This work matters because it addresses the critical challenge of improving the efficiency of MCTD, a powerful approach for trajectory planning. The proposed Fast-MCTD framework achieves substantial speedups while maintaining strong planning performance, demonstrating its practical effectiveness in challenging long-horizon tasks. This work has the potential to impact various applications, such as robotics, autonomous vehicles, and planning in complex environments.

---

[9] Heterogeneous Graph Transformers for Simultaneous Mobile Multi-Robot Task Allocation and Scheduling under Temporal Constraints

Authors: Baluhan Altundas, Shengkang Chen, Shivangi Deo, Mimwo Cho, Matthew Gombolay

PDF: https://openreview.net/pdf?id=k1fbdnwjCH

Overview

Problem Statement

The paper addresses the problem of Multi-Agent Task Allocation and Scheduling (MATAS) for heterogeneous mobile agents under temporal constraints. MATAS is a critical problem in various applications, including logistics, manufacturing, and disaster response, where efficient task allocation and scheduling are essential. The problem is NP-hard, and existing methods, including heuristics and optimization-based solvers, often fail to scale and overlook inter-task dependencies and agent heterogeneity.

Key Contributions

The paper proposes a novel Simultaneous Decision-Making model for Heterogeneous Multi-Agent Task Allocation and Scheduling (HM-MATAS), called TARGETNET. TARGETNET is built on a Residual Heterogeneous Graph Transformer with edge and node-level attention, which encodes agent capabilities, travel times, and temporospatial constraints into a rich graph representation. The model is trainable via reinforcement learning and can generalize effectively to larger scenarios.

Methodology

The paper uses a Graph Neural Network (GNN)-based approach, specifically a Heterogeneous Graph Transformer (HGT), to model the HM-MATAS problem. The HGT architecture leverages attention mechanisms to effectively model heterogeneous and pairwise dynamics. The model is trained on small-scale problems and then deployed on larger problems.

Results

The paper reports significant performance improvements, including:

* 164.10% more feasible tasks assigned given temporal constraints in 3.83% of the time compared to classical heuristics.
* 201.54% more feasible tasks assigned in 0.01% of the time compared to metaheuristics.
* 231.73% more feasible tasks assigned in 0.03% of the time compared to exact solvers.
* 20×-to-250× speedup from prior graph-based methods across scales.

Significance

The work is significant because it provides a scalable and efficient solution to the HM-MATAS problem, which is critical in various applications. The proposed TARGETNET model can generalize effectively to larger scenarios and outperforms state-of-the-art methods. The work has the potential to impact real-world multi-agent systems, such as logistics, manufacturing, and disaster response.

---

[10] Sequential Monte Carlo for Policy Optimization in Continuous POMDPs

Authors: Hany AbduLsamaD, Sahel Iqbal, Simo SarKka

PDF: https://arxiv.org/pdf/2505.16732v3.pdf

Overview

Problem Statement

The paper addresses the problem of optimal decision-making under partial observability in continuous partially observable Markov decision processes (POMDPs). POMDPs are a fundamental framework for modeling decision-making under uncertainty, but existing methods often rely on simplifications or approximations that limit their effectiveness. The problem is important because it is a long-standing challenge in decision theory, and solving it is crucial for building autonomous agents that can operate in real-world environments.

Key Contributions

The paper introduces a novel policy optimization framework for continuous POMDPs that explicitly addresses the challenge of balancing exploration and exploitation. The key contributions are:

* A non-Markovian Feynman-Kac model that captures the adaptive nature of decision-making in POMDPs and naturally incorporates the value of future observations.
* A nested sequential Monte Carlo (SMC) method that simulates adaptive decision-making by sampling from the optimal trajectory distribution of a POMDP.
* A policy optimization algorithm framed as maximum likelihood estimation within the Feynman-Kac model, resulting in a novel policy gradient method for POMDPs.

Methodology

The paper uses a combination of probabilistic inference and optimization techniques to develop a novel policy optimization algorithm. The methodology includes:

* A non-Markovian Feynman-Kac model that captures the adaptive nature of decision-making in POMDPs.
* A nested SMC method that simulates adaptive decision-making by sampling from the optimal trajectory distribution of a POMDP.
* Maximum likelihood estimation within the Feynman-Kac model to optimize policies.

Results

The paper demonstrates the effectiveness of the proposed algorithm on standard continuous POMDP benchmarks, where existing methods struggle to act under uncertainty. The results show that the proposed algorithm outperforms existing methods in terms of cumulative reward and decision-making efficiency.

Significance

The work matters because it provides a novel and effective solution to the problem of optimal decision-making under partial observability in continuous POMDPs. The proposed algorithm has the potential to improve the performance of autonomous agents in real-world environments, where uncertainty and partial observability are common. The work also contributes to the development of more effective decision-making algorithms for POMDPs, which is a fundamental problem in decision theory.

---

[11] Raw2Drive: Reinforcement Learning with Aligned World Models for End-to-End Autonomous Driving in CARLA v2

Authors: Zhenjie Yang, Xiaosong Jiat, Qifeng Li, Xue Yang, Maoqiang Yao, Junchi Yan

PDF: https://arxiv.org/pdf/2505.16394v2.pdf

Overview

Problem Statement

The paper addresses the problem of applying Reinforcement Learning (RL) to End-to-End Autonomous Driving (E2E-AD) with raw sensor inputs. Current methods rely on Imitation Learning (IL) or Model-free RL, which have limitations such as poor generalization and causal confusion. The goal is to develop a model-based RL approach that can effectively learn from raw sensor data.

Key Contributions

The main contributions of this work are:

* Proposing Raw2Drive, a dual-stream Model-based RL approach that leverages privileged information to train a world model and a neural planner.
* Introducing a Guidance Mechanism to align the raw sensor world model with the privileged world model during rollouts.
* Achieving state-of-the-art performance on CARLA Leaderboard 2.0, surpassing IL-based methods by a large margin.

Methodology

The training process consists of two stages:

1. Training a privileged world model and a paired neural planner using privileged information.
2. Jointly training a raw sensor world model and an end-to-end planner using raw video inputs.

The raw sensor world model is guided by the alignment with the privileged world model using the proposed Guidance Mechanism.

Results
Raw2Drive achieves state-of-the-art performance on CARLA Leaderboard 2.0, with a score of 94.5% on the CARLA v2 benchmark. This is a significant improvement over IL-based methods, which have saturated performance on CARLA Leaderboard 1.0.

Significance
This work matters because it addresses the limitations of current E2E-AD methods and provides a promising solution for applying RL to complex driving scenarios. The proposed approach can be applied to various autonomous driving tasks, and its performance improvement on CARLA Leaderboard 2.0 demonstrates its potential for real-world applications.

---

[12] SDTagNet: Leveraging Text-annotated Navigation Maps for Online HD Map Construction

Authors: Fabian Immele, Jan Hendrik Pauls, Richard Fenner, Frank Biedler, Jonas Merker, Christoph Stiller
PDF: https://arxiv.org/pdf/2506.08997v2.pdf

Overview

Problem Statement
The paper addresses the challenge of online High Definition (HD) map construction for autonomous vehicles. HD maps are essential for safe navigation, but their creation and maintenance are costly and time-consuming. Online HD map construction methods generate local HD maps from live sensor data, but they are limited by the short perception range of onboard sensors. The paper aims to improve the performance of online HD map construction by leveraging Standard Definition (SD) maps as prior information.

Key Contributions
The paper proposes SDTagNet, a novel online HD map construction method that fully utilizes the information of widely available SD maps, like OpenStreetMap. SDTagNet introduces two key innovations: (1) it incorporates not only polyline SD map data but also additional semantic information in the form of textual annotations, and (2) it uses a point-level SD map encoder together with orthogonal element identifiers to uniformly integrate all types of map elements.

Methodology
SDTagNet uses a BERT encoder to compute NLP tag embeddings and a graph transformer-like method to encode points, polylines, and element relations. The encoded information is then supplied to the base model via cross-attention. The paper also evaluates SDTagNet on Argoverse 2 and nuScenes datasets, showing that it outperforms existing SD map prior encoding modules by up to 20% and 35%, respectively.

Results
The paper reports that SDTagNet boosts map perception performance by up to +5.9 mAP (+45%) w.r.t. map construction without priors and up to +3.2 mAP (+20%) w.r.t. previous approaches that already use SD map priors.

Significance
This work matters because it addresses a critical challenge in autonomous driving and provides a novel solution that can improve the performance of online HD map construction. By leveraging SD maps as prior information, SDTagNet can enhance far-range detection accuracy and reduce the maintenance effort required for HD maps. The proposed method can have a significant impact on the scalability and reliability of autonomous driving systems.

---

[13] Scaffolding Dexterous Manipulation with Vision-Language Models

Authors: Vincent de Bakker, Joey Hejna, Tyler Lum, Onur Celik, Aleksandar Taranovic, Denis Blessing, Gerhard Neumann, Jeannette Bohg, Dorsa Sadigh

PDF: https://arxiv.org/pdf/2506.19212v2.pdf

Overview

Problem Statement

The paper addresses the challenge of training dexterous robotic hands for complex manipulation tasks. Dexterous hands are essential for tasks like using power-drills or twisting door knobs, but existing learning paradigms struggle to cope with their complexity. The scarcity of high-quality demonstrations and the need for task-specific reward functions hinder the development of generalist policies.

Key Contributions

The main contributions of this paper are:

* Using vision-language models (VLMs) to generate coarse motion plans ("scaffolds") for dexterous manipulation tasks.
* Demonstrating that these scaffolds can be used to guide exploration and provide high-level reward signals for reinforcement learning (RL).
* Introducing a framework that combines VLM-generated motion plans with residual RL to learn manipulation policies for dexterous robot hands.

Methodology

The proposed method uses an off-the-shelf VLM to generate 3D trajectories for hand and object motions from a natural language instruction and image. The VLM identifies relevant object keypoints and generates associated 3D trajectories. A low-level residual RL policy is then trained in simulation to track these trajectories and complete the desired task.

Results

The paper evaluates the method across 8 challenging dexterous manipulation tasks in simulation, achieving close performance to human teleoperation. The results demonstrate that the method can learn robust dexterous manipulation policies and transfer to real-world robotic hands without human demonstrations or handcrafted rewards.

Significance

This work matters because it provides a scalable and generalizable approach to training dexterous robotic hands. By leveraging VLMs to generate coarse motion plans, the method can overcome the challenges of demonstration collection and reward design. The impact of this work could be significant, enabling the development of more advanced robotic hands for a range of applications, from manufacturing to healthcare.

---

[14] FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving

Authors: Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, Xing Wei

PDF: https://arxiv.org/pdf/2505.17685v3.pdf

Overview

Problem Statement

The paper addresses the challenge of enabling Vision-Language-Action (VLA) models to reason about future scenarios in autonomous driving, while minimizing the loss of critical spatio-temporal relationships and fine-grained visual details. This is important because current VLA models rely on textual Chain-of-Thought (CoT) strategies, which can introduce a "modality gap" between perception and planning, leading to inaccurate trajectory planning and increased collisions.

Key Contributions

The main contributions of this paper are:

* Introducing FSDrive, a framework that empowers VLAs to "think visually" using a novel visual spatio-temporal

CoT.
* Proposing a pre-training paradigm that simultaneously preserves semantic understanding and activates visual generation capacity.
* Developing a progressive, easy-to-hard generation method to constrain physical laws and generate accurate future frames.

Methodology
The methodology involves:
* Using a VLA as a world model to generate a unified future frame that combines predicted background with explicit priors like future lane dividers and 3D object boxes.
* Representing temporal relationships through ordinary future frames, which intuitively characterize temporal progression and inherent laws of scene development.
* Pre-training a VLA with visual question answering (VQA) tasks for current scene comprehension and image generation capabilities.

Results
The results show that FSDrive:
* Improves trajectory accuracy and reduces collisions on nuScenes and NAVSIM datasets.
* Achieves competitive FID for video generation with a lightweight autoregressive model.
* Advances scene understanding on DriveLM dataset.

Significance
This work matters because it bridges the perception-planning gap in autonomous driving, enabling safer and more anticipatory decision-making. By unifying future scene representations and perception outputs in image format, FSDrive eliminates semantic gaps caused by cross-modal conversions, establishing an end-to-end visual reasoning pipeline.

---

[15] Uncertainty-aware preference alignment for diffusion policies

Authors: Runqing Miao, Sheng Xiu, Runyi Zhao, Wai Kin Victor Chan, Guiliang Liu
PDF: https://openreview.net/pdf/ca84b09982b76f75f6d05db21f8ff40216ce9fc0.pdf

Overview

Problem Statement
The paper addresses the challenge of aligning diffusion policies with human preferences in decision-making tasks. The problem is important because diffusion policies have shown promising performance, but their alignment with human preferences is often uncertain due to diverse and potentially conflicting preferences from different user groups.

Key Contributions
The main contributions of this work are:

* The introduction of Uncertainty-aware Preference Alignment for Diffusion Policies (Diff-UAPA), a novel algorithm that addresses the uncertainty in preference alignment.
* The use of a maximum posterior objective to align the diffusion policy with regret-based preferences under the guidance of an informative Beta prior.
* The development of an iterative preference alignment framework that adapts the diffusion policy to labels from different user groups.

Methodology
The methodology involves:

* Learning a Beta prior model to capture uncertainty arising from diverse human preferences.
* Parameterizing the Beta distribution with neural networks and training the model via variational inference.
* Integrating the learned Beta prior model into the alignment process with a regret-based preference model.

Results

The results show that Diff-UAPA achieves robust and reliable preference alignment in both simulated and real-world tasks. The paper reports extensive experiments on two simulated robotics environments and one real-world task, demonstrating the effectiveness of Diff-UAPA in handling uncertainty in preference data.

Significance
This work matters because it provides a novel solution to the challenge of aligning diffusion policies with human preferences. The use of an uncertainty-aware objective and the development of an iterative preference alignment framework make Diff-UAPA a robust and reliable approach for decision-making tasks. The impact of this work could be significant in applications such as robotics, autonomous vehicles, and decision-making systems.

---

[16] Deep Learning for Continuous-Time Stochastic Control with Jumps

Authors: Patrick Cheridito, Jean-Loup Dupret, Donatien Hainaut
PDF: https://arxiv.org/pdf/2505.15602v2.pdf

Overview

Problem Statement
The paper addresses the problem of solving finite-horizon continuous-time stochastic control problems with jumps. These problems are crucial in various fields, such as finance, engineering, and economics, where decision-makers need to optimize control processes under uncertainty. The importance of this problem lies in its ability to model complex dynamic systems, where the underlying stochastic processes are known.

Key Contributions
The main contributions of this paper are:

* A deep model-based approach for solving stochastic control problems with jumps, which leverages the Hamilton-Jacobi-Bellman (HJB) equation.
* Two neural networks are trained iteratively to approximate the value function and optimal control.
* The approach can handle high-dimensional problems with a combination of diffusive noise and random jumps with controlled intensities.

Methodology
The paper uses a deep learning approach, specifically two neural networks, to approximate the value function and optimal control. The training objectives are derived from the HJB equation, ensuring that the networks capture the underlying stochastic dynamics. The approach is meshfree, meaning it does not require discretization of the problem domain.

Results
Empirical evaluations on different problems demonstrate the accuracy and scalability of the approach. The results show that the approach can effectively handle complex, high-dimensional stochastic control tasks.

Significance
This work matters because it provides a novel approach for solving stochastic control problems with jumps, which are common in various fields. The approach can handle high-dimensional problems with complex dynamics, making it a significant contribution to the field of stochastic control. The code is available on GitHub, allowing for further development and application of the approach.

---

[17] Predictive Preference Learning from Human Interventions

Authors: Haoyuan Cai, Zhenghao Peng, Bolei Zhou

PDF: https://arxiv.org/pdf/2510.01545v2.pdf

Overview

Problem Statement

The paper addresses the challenge of effectively leveraging human demonstrations to teach and align autonomous agents in Reinforcement Learning (RL) and Imitation Learning (IL). Current methods require a large number of environment interactions, exploration can lead to dangerous states, and IL agents are susceptible to distributional shift.

Key Contributions

The main contributions are:
1. **Predictive Preference Learning from Human Interventions (PPL)**: a novel Interactive Imitation Learning algorithm that leverages trajectory prediction to inform human intervention and employs preference learning to deter the agent from returning to dangerous states.
2. **Efficient rollout-based trajectory prediction model**: forecasts the agent's future states, visualizing them in real-time for the user to proactively determine when an intervention is necessary.
3. **Theoretical analysis**: derives an upper bound on the performance gap of PPL, highlighting its efficacy in reducing distributional shifts while preserving preference data quality.

Methodology

PPL consists of two key designs:
1. **Trajectory prediction**: uses an efficient rollout-based model to forecast the agent's future states.
2. **Preference learning**: leverages the implicit preference signals contained in human interventions to inform predictions of future rollouts.

Results

Experiments on MetaDrive and Robosuite benchmarks show that PPL:
1. **Requires fewer expert monitoring efforts**: 2-3 times fewer than HG-DAgger and IWR.
2. **Achieves near-optimal policies**: with 10-20% fewer expert demonstrations than HG-DAgger and IWR.

Significance

PPL has the potential to significantly improve the efficiency and safety of Interactive Imitation Learning. By reducing the cognitive burden on human supervisors and leveraging preference learning, PPL can accelerate policy improvement and reduce the total number of expert interventions required.

---

[18] Blindfolded Experts Generalize Better: Insights from Robotic Manipulation and Video Games

Authors: Ev Zisselman, Mirco Mutti, Shelly Francis-Meretzki, Elisei Shafer, Aviv Tamar

PDF: https://arxiv.org/pdf/2510.24194v1

Overview

Problem Statement

The paper addresses the problem of generalization in behavioral cloning (BC), a technique for learning sequential decision-making from demonstrations. The key challenge is that BC typically requires abundant data to generalize to unseen tasks, which can be time-consuming and expensive to collect. The paper focuses on improving generalization to task variations by exploring the role of expert behavior in imitation learning.

Key Contributions

The main contributions of this work are:

* Introducing the concept of "blindfolded" experts, where the expert's observations are partially masked to induce more exploratory behavior.

* Theoretically proving an upper bound on the generalization error that scales with the amount of task information available to the demonstrator (I) and the number of demonstrated tasks (m).
* Empirically demonstrating the effectiveness of blindfolded experts on simulated games from the Procgen suite and a real-robot peg insertion task.

Methodology

The paper uses a combination of theoretical analysis and empirical experiments to evaluate the effectiveness of blindfolded experts. The methodology involves:

* Using recurrent neural networks (RNNs) or transformers to process a sequence of observations.
* Masking the expert's observations to induce exploratory behavior.
* Collecting demonstrations of a small selection of tasks and using them to train a policy.
* Evaluating the generalization performance on unseen tasks.

Results

The key findings are:

* Blindfolded experts generalize better to unseen tasks than fully-informed experts.
* The generalization error scales with I/m, where I measures the amount of task information available to the demonstrator.
* The blindfolded expert approach requires fewer demonstrations to achieve similar generalization performance.

Significance

This work matters because it provides a new and principled approach for collecting demonstrations, which can be time-consuming and expensive. By inducing exploratory behavior in experts, the blindfolded expert approach can improve generalization to unseen tasks, making it a valuable contribution to the field of imitation learning.

---

[19] DynaGuide: Steering Diffusion Policies with Active Dynamic Guidance

Authors: Maximilian Du, Shuran Song

PDF: https://arxiv.org/pdf/2506.13922

Overview

Problem Statement

The paper addresses the problem of steering large, complex policies in robots to match the needs of a specific scenario without retraining the policy, sampling excessively from it, or anticipating all possible steering during policy training. This is important because the deployment of complex policies in the real world requires adaptability to changing situations, and existing steering approaches rely on assumptions that limit their effectiveness.

Key Contributions

The main contributions of this work are:

* **DynaGuide**: a steering method for diffusion policies using guidance from an external dynamics model during the diffusion denoising process.
* **Separation of Steering Forces**: DynaGuide separates the steering forces from the base behavior policy, allowing for flexible steering structures, increased steering robustness, and plug-and-play modularity.
* **Dynamics Guidance**: DynaGuide uses a latent visual dynamics model to predict the final or far-future observation of a trajectory given the current observation and action, allowing for direct influence on the denoising process.

Methodology

The methodology involves:

* **Latent Visual Dynamics Model**: a model that predicts the final or far-future observation of a trajectory given the current observation and action.
* **Diffusion Denoising Process**: DynaGuide uses the dynamics model to guide the diffusion denoising process, allowing for direct influence on the action.
* **DinoV2 Embedder**: a model that projects visual observations into the expressive DinoV2 latent space.

Results
The results show that DynaGuide successfully guides the base policy up to 70% of the time and outperforms goal conditioning by 5.4x on lower quality objectives. It also successfully amplifies underrepresented behaviors over sampling methods and accommodates multiple positive and negative objectives.

Significance
This work matters because it provides a new approach to steering complex policies in robots, allowing for adaptability to changing situations without retraining the policy. The use of a dynamics model and diffusion denoising process enables direct influence on the action, making it a more effective and robust approach than existing methods. The code and collected data will be made publicly available, allowing for further research and development.

---

[20] Solving the Discretization Gap in Magic Gate Networks

Authors: Roger W. ..., ETH ...
PDF: https://arxiv.org/pdf/2506.07500

Overview

Problem Statement
The paper addresses the challenges of training Differentiable Logic Gate Networks (DLGNs), which are a type of neural network that uses logic gates to perform computations. The main issues are the discretization gap, where the trained model's performance degrades significantly when discretized for inference, and slow convergence, which makes training time-consuming.

Key Contributions
The paper proposes Gumbel Logic Gate Networks (Gumbel LGNs), which inject Gumbel noise into the gate selection process using the Gumbel-Softmax trick. This approach smooths the loss landscape, reducing the discretization gap and accelerating convergence. The authors also explore a training strategy that uses continuous relaxations only in the backward pass, while enforcing discrete gates in the forward pass.

Methodology
The paper uses a combination of theoretical analysis and empirical validation. The theoretical analysis shows that injecting Gumbel noise into DLGNs smooths their loss landscape by regularizing the Hessian's trace. The empirical validation involves designing and executing experiments on CIFAR-10 and CIFAR-100 datasets, using DLGNs and Gumbel LGNs architectures.

Results
The results show that Gumbel LGNs achieve faster convergence (4.5× faster in wall-clock time) and smaller discretization gaps (98% reduction) compared to baseline DLGNs. The training strategy that uses continuous relaxations only in the backward pass also reduces the discretization gap.

Significance
This work matters because it addresses the challenges of training DLGNs, which are a promising approach for efficient inference. The proposed Gumbel LGNs and training strategy can improve the performance and efficiency of DLGNs, making them more suitable for real-world applications. The results also demonstrate the scalability of neural architecture search (NAS) techniques to large parameter spaces.

---

[21] Enforcing Convex Constraints in Graph Neural Networks

Authors: Ahmed Rashwan, Keith Briggs, Chris Budd, Lisa Kreusel

PDF: https://arxiv.org/pdf/2510.11227v1.pdf

Overview

Problem Statement

The paper addresses the problem of enforcing complex, dynamic constraints in Graph Neural Networks (GNNs) for machine learning applications. This is important because many real-world applications require outputs that satisfy specific constraints, such as feasibility, safety, or interpretability. However, existing methods for enforcing constraints in GNNs are limited, particularly for input-dependent constraints.

Key Contributions

The paper makes several key contributions:
1. **Convergence Result**: The authors prove a convergence result for the Component-Averaged Dykstra (CAD) algorithm, which is used for solving the best-approximation problem.
2. **Surrogate Gradient**: They introduce a computationally efficient surrogate gradient for the CAD algorithm, tailored for ML applications.
3. **Sparse Vector Clipping**: They propose a refined variant of the vector clipping method, which leverages problem sparsity and is well-suited for use in GNN models.
4. **ProjNet Architecture**: They propose ProjNet, a GNN architecture for constraint satisfaction problems that integrates the CAD algorithm with sparse vector clipping.

Methodology

The authors use the CAD algorithm, which is an iterative scheme for solving the best-approximation problem. They also use a sparse variant of the vector clipping method to improve model expressiveness. The architecture is end-to-end differentiable and fully GPU-accelerated.

Results

The authors demonstrate the effectiveness of ProjNet on four classes of constrained optimization problems: linear programming, two classes of non-convex quadratic programs, and radio transmit power optimization. They show that ProjNet can handle large-scale inputs efficiently and provide significant speed-ups and scalability.

Significance

This work matters because it provides a scalable and efficient method for enforcing complex, dynamic constraints in GNNs. The authors' approach can be applied to a wide range of applications, including robotics, energy systems, and industrial processes. The use of a surrogate gradient and sparse vector clipping methods enables efficient computation and scalability, making this approach suitable for large-scale applications.

---

[22] Foundation Models for Scientific Discovery: From Paradigm Enhancement to Paradigm Transition

Authors: Fan Liu, Jindong Han, Tengfei Liu, Weijia Zhang, Zhe-Rui Yang

PDF: https://arxiv.org/pdf/2510.15280v1.pdf

Overview

Problem Statement

The paper addresses the problem of understanding the role of Foundation Models (FMs) in scientific discovery. FMs, such as GPT-4 and AlphaFold, have revolutionized various scientific fields, but their impact on the scientific paradigm is not yet fully understood. The authors argue that FMs are not just enhancing existing methodologies but are, in fact, catalyzing a transition towards a new scientific paradigm.

Key Contributions

The paper proposes a three-stage framework to describe the evolution of scientific paradigms empowered by FMs: (1) Meta-Scientific Integration, (2) Hybrid Human-AI Co-Creation, and (3) Autonomous Scientific Discovery. The authors

also review current applications and emerging capabilities of FMs across existing scientific paradigms, identifying risks and future directions for FM-enabled scientific discovery.

Methodology
The paper uses a qualitative approach, analyzing the impact of FMs on scientific discovery through a literature review and case studies. The authors draw on examples from various scientific fields, including protein folding, material design, and mathematical conjectures, to illustrate the potential of FMs in scientific discovery.

Results
The paper highlights the potential of FMs to accelerate scientific discovery, improve problem formulation, and enhance reasoning and abstraction capabilities. The authors also identify risks and challenges associated with the increasing reliance on FMs, including the potential for bias and the need for human oversight.

Significance
This work matters because it provides a framework for understanding the impact of FMs on scientific discovery. By highlighting the potential of FMs to catalyze a transition towards a new scientific paradigm, the authors provide a roadmap for future research and development in this area. The paper's findings have implications for the scientific community, policymakers, and industry leaders, who must consider the potential benefits and risks of FM-enabled scientific discovery.

---

[23] Improving Generative Behavior Cloning via Self-Guidance and Adaptive Chunking

Authors: Junhyuk So, Chiwoong Lee, Sinyoung Lee, Jungseul Ok, Eunhyeok Park
PDF: https://arxiv.org/pdf/2510.12392v1.pdf

Overview

Problem Statement
The paper addresses the limitations of Generative Behavior Cloning (GBC) methods, specifically the Diffusion Policy model, which suffers from stochasticity and delayed responses. These limitations can lead to task failures and poor performance in noisy or dynamic environments.

Key Contributions
The paper proposes two novel techniques to enhance the consistency and reactivity of diffusion policies:
1. **Self-Guidance**: incorporates negative score estimates from prior observations into the diffusion denoising process to guide the model toward more informed, high-fidelity action modes.
2. **Adaptive Chunking**: updates action chunks only when the benefits of increased reactivity outweigh the need for temporal consistency, striking a dynamic balance between responsiveness and stability.

Methodology
The paper uses a combination of self-guidance and adaptive chunking to improve the performance of the Diffusion Policy model. The self-guidance technique leverages information from past observations to guide the model's decision-making process, while adaptive chunking updates action chunks in response to changing environmental conditions.

Results
Extensive evaluations across simulated and real-world robotic environments demonstrate that the proposed approach outperforms the Vanilla Diffusion Policy by 23.25% and the state-of-the-art BID by 12.27%, while reducing computational cost by a factor of 16.

Significance
This work matters because it addresses the limitations of GBC methods and provides a more robust and efficient approach to decision-making in complex environments. The proposed techniques can be applied to a wide range of robotic tasks and can have a significant impact on the development of autonomous robots.

---

[24] Why Masking Diffusion Works: Condition on the Jump Schedule for Improved Discrete Diffusion

Authors: Alan N. Amir, Nate Gravert, Andrew G. Wilson
PDF: https://arxiv.org/pdf/2506.08316v2.pdf

Overview

Problem Statement

The paper addresses the problem of improving discrete diffusion models, which are used for conditional generation of discrete sequences. These models are crucial in applications such as biological sequence design, language data generation, and protein data modeling. However, the best-performing discrete diffusion model, masking diffusion, does not denoise gradually, which is counterintuitive.

Key Contributions

The paper proposes a new approach called schedule-conditioned diffusion (SCUD), which builds in the known distribution of jump times in the forward process and only learns where to jump to. This approach is shown to improve the performance of discrete diffusion models, including masking diffusion. The authors also demonstrate that SCUD can be applied to various discrete diffusion models, including structured and state-dependent models.

Methodology

The paper uses a combination of theoretical analysis and experimental evaluation. The authors derive a new evidence lower bound (ELBO) for discrete diffusion models, which is used to optimize the parameters of the model. They also propose a new algorithm for training discrete diffusion models, which is based on the SCUD approach. The authors evaluate their approach on several datasets, including CIFAR-10, UniRef50, and protein data.

Results

The paper reports significant improvements in the performance of discrete diffusion models using the SCUD approach. Specifically, the authors show that SCUD outperforms masking diffusion on several datasets, including CIFAR-10 and UniRef50. They also demonstrate that SCUD can be applied to structured and state-dependent models, achieving state-of-the-art results.

Significance

The work has significant implications for the field of discrete diffusion models. By improving the performance of these models, the authors enable more accurate and efficient generation of discrete sequences, which is crucial in applications such as biological sequence design, language data generation, and protein data modeling. The SCUD approach also provides a new framework for designing and optimizing discrete diffusion models, which can be applied to a wide range of applications.

---

[25] GI: Teaching LLMs to Reason on Graphs with Reinforcement Learning

Authors: Xiaojun Guo, Ang Li, Yifei Wang, Stefanie Jegelka, Yisen Wang
PDF: https://arxiv.org/pdf/2505.18499

Overview

Problem Statement

The paper addresses the limitation of Large Language Models (LLMs) in graph-related tasks, which hinders the development of truly general-purpose models. Despite their remarkable progress, LLMs struggle with graph reasoning, a critical capability for achieving general-purpose intelligence.

Key Contributions

The paper introduces G1, a simple yet effective approach that uses Reinforcement Learning (RL) to improve LLMs' graph reasoning abilities. The main contributions are:

* The first application of RL to improve LLMs on graph reasoning tasks.
* The introduction of Erd?s, the largest-scale and most comprehensive graph-theoretic dataset, comprising 50 distinct tasks of varying complexities.

* The demonstration of G1's ability to achieve substantial performance improvements on the Erd?s benchmark, with gains of up to 46% over baseline models.

Methodology
The paper uses RL to train LLMs on synthetic graph-theoretic tasks, leveraging the Erd?s dataset. The approach involves:

* Curating the Erd?s dataset, which includes 50 graph-theoretic tasks of varying difficulty levels.
* Using RL to train LLMs on the Erd?s dataset, with the goal of improving graph reasoning abilities.
* Employing chain-of-thought and supervised initialization to enhance training.

Results
The paper reports significant performance improvements, with G1-7B model achieving competitive performance with state-of-the-art reasoning models like OpenAI's o3-mini. The results also show strong zero-shot generalization on unseen graph tasks and domains.

Significance
This work matters because it demonstrates that RL can unlock latent graph understanding within general-purpose LLMs using synthetic data. This paves the way for more versatile AI systems capable of sophisticated reasoning across diverse data modalities. The Erd?s dataset and G1 approach provide a scalable and efficient path for building strong graph reasoners.

---

[26] ReSim: Reliable World Simulation for Autonomous Driving

Authors: Jiazi Yang, Kashyap Chitta, Shenyan Gao, Long Chen, Yugian Shao, Xiaosong Jia, Hongyong Li, Andreas Geiger, Xiangyu Yue, L. Chen
PDF: https://arxiv.org/pdf/2506.09981v1.pdf

Overview

Problem Statement
The paper addresses the challenge of reliably simulating future driving scenarios under a wide range of ego driving behaviors. Current driving world models, trained exclusively on real-world driving data, struggle to follow hazardous or non-expert behaviors, which are rare in such data. This limitation restricts their applicability to tasks such as policy evaluation.

Key Contributions
The main contributions of this work are:
1. **Heterogeneous data compilation**: Enriching real-world human demonstrations with diverse non-expert data collected from a driving simulator.
2. **ReSim**: A controllable world model that reliably simulates high-fidelity future outcomes by precisely executing diverse action inputs.
3. **Video2Reward**: A model that estimates a reward from ReSim's simulated future.

Methodology
The approach uses a scalable text-to-video generator with a multi-stage training pipeline to integrate visual and action conditions. The model applies an unbalanced noise sampling strategy and dynamics consistency loss to emphasize the learning of motion coherence.

Results
The results show:
1. **Improved visual fidelity**: Up to 44% higher visual fidelity compared to prior works.
2. **Improved controllability**: Over 50% improvement in controllability for both expert and non-expert actions.
3. **Planning and policy selection performance**: 2% and 25% improvement, respectively, on the NAVSIM benchmark.

Significance

This work matters because it addresses a critical challenge in autonomous driving: reliable simulation of diverse driving scenarios. The ReSim paradigm has the potential to improve the development of autonomous driving systems by providing a more realistic and controllable simulation environment.

---

[27] SimWorld-Robotics: Synthesizing Photorealistic and Dynamic Urban Environments for Multimodal Robot Navigation and Collaboration

Authors: Yan Zhuang, Jiawei Ren, Xiaokang Ye, Jianzhi Shen, Rukuan Zhang, Tianai Yue, Muhammad Faayez, Kuhong He, Xian Zhang, Zigao Ma, Lianhui Qiu, Zhihui Hu, Tianmin Shu
PDF: https://arxiv.org/pdf/2512.10046v1.pdf

Overview

Problem Statement

The paper addresses the challenge of developing realistic and scalable embodied simulators for outdoor robotics tasks, particularly in large urban environments. Current simulators lack the necessary realism, customizability, scalability, and versatility to support the complex tasks required in real-world urban robotics.

Key Contributions

The main contributions of this work are:

* SimWorld-Robotics (SWR), a new embodied AI simulation platform for large-scale, photorealistic, and dynamic urban environments.
* Two novel benchmarks for robots in large, urban environments: SIMWORLD-MMNAV (multimodal instruction following) and SIMWORLD-MRS (multi-robot search).
* SimWorld-20K, a large-scale dataset for benchmarking multimodal robot navigation in photo-realistic urban environments.

Methodology

SWR is built on Unreal Engine 5 and offers diverse high-?delity building and object assets, multiple types of embodied agents with rich action spaces, a waypoint-based background traf?c system, and a comprehensive procedural city generation pipeline. The platform supports scalable data generation with fine-grained ground-truth annotations.

Results

The experimental results demonstrate that existing models, including state-of-the-art vision-language models (VLMs), fail to achieve meaningful success on the proposed benchmarks. The SimWorld-20K dataset contains 20K training steps sampled from 200 episodes, each averaging 500 m in length, across 100 procedurally generated city environments.

Significance

This work matters because it addresses the gap in current foundation models for challenging, realistic robot tasks in urban environments. The proposed platform and benchmarks have the potential to accelerate progress toward stronger embodied intelligence and improve the performance of robots in real-world urban environments.

---

[28] Chain-of-Action: Trajectory Autoregressive Modeling for Robotic Manipulation

Authors: Wenbo Zhang, Trannun Hu, Hanbo Zhang, Yanyuan Qiao
PDF: https://arxiv.org/pdf/2506.09990v1.pdf

Overview

Problem Statement

The paper addresses the problem of compounding errors in visuo-motor policies, which are models that enable robots to perform complex manipulation tasks from raw sensory observations. Compounding errors occur when the model's predictions accumulate and deviate from the intended goal, leading to misaligned behaviors during execution. This problem is important because it affects the reliability and efficiency of robotic manipulation tasks, which are critical in various industries such as manufacturing, healthcare, and logistics.

Key Contributions

The main contributions of this paper are:

* **Chain-of-Action (CoA)**: a novel visuo-motor policy paradigm that generates an entire trajectory by explicit backward reasoning with task-specific goals through an action-level Chain-of-Thought (CoT) process.
* **Continuous action representation**: a technique that preserves fine-grained structure and trajectory fidelity by representing actions as continuous values rather than discrete bins.
* **Dynamic stop**: a mechanism that enables the model to determine when to stop based on proximity to the goal, reducing over-generation and improving execution efficiency.
* **Reverse temporal ensemble**: a variant of ensemble strategies that ensembles multiple backward rollouts, mitigating temporal misalignment and reducing variance during closed-loop execution.

Methodology

The CoA paradigm is built upon Trajectory Autoregressive Modeling and is unified within a single autoregressive structure. The model consists of a Transformer decoder with a causal mask, which generates action tokens in reverse order, starting from a keyframe action that encodes the task-specific goal.

Results

The CoA paradigm achieves state-of-the-art performance across 60 RLBench tasks and 8 real-world manipulation tasks, outperforming ACT by 16% and Diffusion Policy by 23%. CoA also surpasses ACT by 15% in real-world robotic manipulation.

Significance

This work matters because it provides a novel and effective approach to visuo-motor policy learning, which can improve the reliability and efficiency of robotic manipulation tasks. The CoA paradigm can be applied to various industries and applications, such as manufacturing, healthcare, and logistics, where robotic manipulation is critical.

---

[29] PRIMT: Preference-based Reinforcement Learning with Multimodal Feedback and Trajectory Synthesis from Foundation Models

Authors: Ruiqi Wang, Dezhong Zhao, Ziqin Yuan, Tianyuan Shao, Guohua Chen, Dominic Kao, Sungeun Hong, Byung-Hee Min
PDF: https://arxiv.org/pdf/2509.15607v2.pdf

Overview

Problem Statement

The paper addresses the challenges of preference-based reinforcement learning (PbRL) in robotics, specifically the reliance on extensive human input and the difficulties in resolving query ambiguity and credit assignment during reward learning. These challenges limit the effectiveness of PbRL in teaching robots complex behaviors.

Key Contributions

The main contributions of this paper are:

* PRIMT, a PbRL framework that leverages foundation models (FMs) for multimodal synthetic feedback and trajectory synthesis.
* Hierarchical neuro-symbolic fusion strategy that combines the strengths of large language models (LLMs) and vision-language models (VLMs) for multimodal evaluation of robot behaviors.
* Bidirectional trajectory synthesis that mitigates early-stage query ambiguity through foresight generation and enhances credit assignment in reward learning via hindsight counterfactual augmentation with a causal auxiliary loss.

Methodology

The PRIMT framework consists of two core components:

* Multimodal feedback fusion: intra-modal fusion, inter-modal fusion using probabilistic soft logic (PSL), and hierarchical neuro-symbolic preference fusion.
* Bidirectional trajectory synthesis: foresight trajectory generation, hindsight trajectory augmentation, and causal auxiliary loss.

Results
The paper evaluates PRIMT on 2 locomotion and 6 manipulation tasks on various benchmarks, demonstrating superior performance over FM-based and scripted baselines.

Significance
This work matters because it addresses the challenges of PbRL in robotics, enabling more efficient and robust learning of complex behaviors. By leveraging FMs for multimodal synthetic feedback and trajectory synthesis, PRIMT has the potential to improve the scalability and effectiveness of PbRL in real-world applications.

---

[30] SonoGym: High Performance Simulation for Challenging Surgical Tasks with Robotic Ultrasound

Authors: Yunke Ao, Masoud Moghaddam, Mayank Mittal, Marsah Pralap, Lisong Wu, Federic Grudl, Fabio Caron, Andreas Kneel
PDF: https://arxiv.org/pdf/2507.01152v1.pdf

Overview

Problem Statement
The paper addresses the challenge of developing a high-performance simulation platform for robotic ultrasound tasks, particularly in orthopedic surgery. Robotic ultrasound has the potential to enhance imaging efficiency and improve surgical outcomes, but its use in complex tasks like anatomy reconstruction and surgery is limited due to the lack of realistic and efficient simulation environments.

Key Contributions
The main contributions of this work are:

* Developing SonoGym, a scalable simulation platform for complex robotic ultrasound tasks that enables parallel simulation across tens to hundreds of environments.
* Formulating ultrasound-guided navigation, reconstruction, and surgery as specialized Markov Decision Processes (MDPs) and adapting these models to partially observable MDPs (POMDPs), submodular MDPs, and state-wise constrained MDPs.
* Generating expert demonstration datasets within the simulator to enable training of recent IL agents, including the Action Chunking Transformer (ACT) and the Diffusion Policy (DP).

Methodology
The paper uses a combination of model-based and learning-based ultrasound simulation techniques, including:

* Physics-based simulation using CT-derived 3D models of the anatomy.
* Generative modeling approach using a 3D label map and CT scans from real patient datasets.
* Deep reinforcement learning (DRL) and imitation learning (IL) algorithms, including the Action Chunking Transformer (ACT) and the Diffusion Policy (DP).

Results
The results demonstrate successful policy learning across a range of scenarios, with the SonoGym platform enabling the training of high-performing DRL agents and IL agents. The paper also highlights the limitations of current methods in clinically relevant environments.

Significance
This work has significant implications for the development of robotic ultrasound systems for orthopedic surgery. By providing a high-performance simulation platform, SonoGym enables the training of more accurate and efficient DRL and IL agents, which can improve surgical outcomes and reduce the risk of complications. The platform can also facilitate research in robot learning approaches for challenging robotic surgery applications.

---

[31] Robo2VLM: Improving Visual Question Answering using Large-Scale Robot Manipulation Data

Authors: Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, Ken Goldberg
PDF: https://arxiv.org/pdf/2505.15517

Overview

Problem Statement

The paper addresses the challenge of improving the spatial reasoning capabilities of Vision-Language Models (VLMs) in robotic manipulation tasks. VLMs are trained on large-scale image-text corpora, but these datasets lack fine-grained spatial information, which is crucial for robots to identify long-tail objects, reason about spatial relationships, and plan physical interactions. The paper aims to bridge this gap by leveraging real-world robot data to enhance and evaluate VLMs.

Key Contributions

The paper presents Robo2VLM, a Visual Question Answering (VQA) dataset generation framework for VLMs from real-world robot data. Robo2VLM segments robot trajectories into distinct manipulation phases, selects representative frames, and generates questions whose answers are supported by synchronized proprioceptive and kinematic ground truth. The paper also curates Robo2VLM-1, a large-scale, in-the-wild VQA dataset with 684,710 questions covering 463 distinct scenes, 3,396 robotic manipulation tasks, and 149 manipulation skills.

Methodology

The paper uses a multiple-choice VQA dataset generation framework, which involves segmenting robot trajectories into manipulation phases, selecting representative frames, and generating questions based on synchronized proprioceptive and kinematic ground truth. The framework is applied to 176k diverse, real-world trajectories from the Open X-Embodiment (OXE) dataset, producing over 3 million VQA samples.

Results

The paper evaluates 14 model configurations with state-of-the-art open source models (LLaVA, Llama, and Qwen) and with different parameter sizes and prompting techniques. The results indicate that some VLMs can achieve near human performance in questions related to object reachability and interaction understanding. Finetuning LLaVA with Robo2VLM-1 improves most of the spatial and interaction capabilities with increasing training dataset size, with a maximum 50% accuracy gain in state reasoning and task understanding.

Significance

This work matters because it provides a new approach to improving the spatial reasoning capabilities of VLMs in robotic manipulation tasks. The Robo2VLM framework and dataset can be used to benchmark and improve VLMs, which can have a significant impact on the development of autonomous robots that can perform complex manipulation tasks. The paper's results demonstrate the potential of using real-world robot data to enhance VLMs, which can lead to more accurate and efficient robotic manipulation systems.

---

[32] Temporal Logic-Based Multi-Vehicle Backdoor Attacks Against Offline RL Agents in End-to-End Autonomous Driving

Authors: Xuan Chen, Shiwei Feng, Zikang Xiong, Shengwei An, Yunshu Ma, Lu Yan, Guanhong Tao, Wenbo Guo, Xiangyuan Zhang
PDF: https://arxiv.org/pdf/2509.16950v2.pdf

Overview

Problem Statement

The paper addresses the vulnerability of end-to-end autonomous driving (AD) systems to backdoor attacks, which can compromise their safety and reliability. Existing works focus on pixel-level triggers that are impractical to deploy in real-world scenarios. The authors aim to explore and understand the vulnerability of AD systems in simulation against backdoor attacks, which is a crucial capability for real-world deployment.

Key Contributions

The main contributions of this work are:

* A novel backdoor attack against end-to-end AD systems that leverage one or more other vehicles' trajectories as triggers.
* A temporal logic (TL)-based framework that can automatically generate sophisticated trajectories of different vehicles.
* A negative training strategy that generates patch trajectories to train the agent to behave normally under non-trigger scenarios.

Methodology

The authors propose a TL-based framework that can automatically generate sophisticated trajectories of different vehicles. The framework consists of two key components:

* Manually specifying the trigger trajectory with precise spatiotemporal coordination for multiple attacking vehicles is time-consuming and unrealistic. To address this, the authors propose a novel TL-based framework that can automatically generate sophisticated trajectories of different vehicles.
* Simply poisoning the ego car with trigger trajectories largely introduces false positives. To mitigate this, the authors develop a negative training strategy that generates patch trajectories, which are similar yet distinct from the original triggers.

Results

The authors conduct extensive evaluations on five offline RL agents using practical trigger designs, demonstrating the effectiveness and feasibility of their attack. They also examine the capability of existing defenses against their attack.

Significance

This work matters because it reveals an under-explored yet critical vulnerability to AD systems. The authors' approach shifts the focus from direct manipulation of the target vehicle input to exploiting the vehicle's contextual awareness algorithms. The use of temporal logic-based framework and negative training strategy makes the attack more practical and adaptable to deploy in the physical world.

---

[33] Neural Combinatorial Optimization for Time Dependent Traveling Salesman Problem

Authors: Ruixiao Yang, Chuchu Fan

PDF: https://openreview.net/pdf?id=UXTR6ZYV1x

Overview

Problem Statement

The Time-Dependent Traveling Salesman Problem (TDTSP) is a variant of the classic Traveling Salesman Problem (TSP) that incorporates time-varying edge weights, reflecting real-world scenarios where travel times fluctuate due to congestion patterns. This problem is crucial in logistics and transportation, as it can help optimize routes with time-dependent considerations, reducing environmental impact and improving customer service.

Key Contributions

The paper proposes a novel neural model that extends MatNet from static asymmetric TSP to time-dependent settings by using an adjacency tensor to capture temporal variations. The model addresses the unique challenge of asymmetry and triangle inequality violations that change dynamically over time. The key contributions are:

1. Empirical analysis of practical TDTSP data, identifying the limitations of the evaluation method in existing DRL work and proposing a new evaluation method.
2. An end-to-end neural network model that directly encodes the time-dependent adjacency tensors, effectively capturing the complicated spatiotemporal dynamics in TDTSP.
3. An effective inference process to enhance the solution quality based on the data distribution.

Methodology

The proposed model uses a neural network architecture that combines an adjacency tensor to capture temporal variations with a time-aware decoder. The model is trained using a REINFORCE algorithm and evaluated on real-world datasets from 12 cities.

Results

The results show that the proposed method achieves state-of-the-art average optimality gap on full instances and significant travel-time reduction on instances where time-aware routing saves time. The method is evaluated on real-world datasets, including datasets from 12 cities, and demonstrates strong support for learning spatiotemporal dependencies.

Significance

This work matters because it addresses a critical limitation in current evaluation practices for TDTSP, which rely solely on average travel time metrics. The proposed method provides a more effective way to evaluate time-dependent routing problems and can have a significant impact on logistics and transportation operations, reducing environmental impact and improving customer service.

---

[34] RoboCerebra: A Large-Scale Benchmark for Long-Horizon Robotic Manipulation Evaluation

Authors: Songhao Han, Boxiang Qiu, Yue Liao, Siyuan Huang, Chen Gao, Shuicheng Yan, Si Liu
PDF: https://arxiv.org/pdf/2506.06677v2.pdf

Overview

Problem Statement

The paper addresses the limitation of current robotic manipulation benchmarks in evaluating the System 2 capabilities of vision-language models (VLMs). These capabilities, including high-level reasoning and long-horizon planning, are underexplored due to the limited temporal scale and structural complexity of existing benchmarks.

Key Contributions

The main contributions of this work are:

* **RoboCerebra**: a novel benchmark designed to assess long-horizon planning and high-level reasoning in robotic manipulation.
* **Large-scale simulation dataset**: featuring extended task horizons and dynamically evolving environments.
* **Baseline framework**: integrating System 2?System 1 coordination for hierarchical policy execution.
* **Evaluation protocol**: tailored to isolate and measure System 2 performance.

Methodology

The dataset is constructed in simulation using a top-down pipeline, where GPT generates task instructions and decomposes them into subtask sequences. Human operators execute the subtasks in simulation, yielding high-quality trajectories with dynamic object variations. The baseline framework integrates a high-level VLM planner with a low-level vision-language-action (VLA) controller.

Results

The results show that RoboCerebra features significantly longer action sequences (approximately 6× those in existing benchmarks) and denser annotations. The evaluation protocol isolates and measures System 2 performance, advancing the development of more capable and generalizable robotic planners.

Significance

This work matters because it enables a comprehensive evaluation of System 2 capabilities in robotic manipulation. By providing a large-scale simulation dataset and a baseline framework, RoboCerebra can help unlock the full potential of VLMs in robotics, leading to more capable and generalizable robotic planners.

---

[35] Learning Parameterized Skills from Demonstrations

Authors: Vedant Gupta, Haotian Fu, Calvin Luo, Yiding Li
PDF: https://arxiv.org/pdf/2510.24095v1.pdf

Overview

Problem Statement

The paper addresses the problem of learning parameterized skills from expert demonstrations in multitask environments. This is important because traditional reinforcement learning methods often fail to leverage inherent behavioral patterns, leading to sample inefficiency. In contrast, humans can extract and reuse strong priors from past experiences, enabling effective generalization to novel tasks.

Key Contributions

The main contributions of this work are:

* The introduction of DEPS (Discovery of Generalizable Parameterized Skills), an end-to-end algorithm for learning parameterized skills from expert demonstrations.
* The use of temporal variational inference and information-theoretic regularization methods to address degeneracy in latent variable models.
* The development of a three-level hierarchy for skill learning, including a discrete skill selector, continuous parameter selector, and subpolicy.

Methodology

DEPS uses a combination of temporal variational inference and information-theoretic regularization methods to learn parameterized skills. The algorithm consists of three levels:

1. Discrete skill selector: selects a skill from a library given the full environment observation.
2. Continuous parameter selector: outputs continuous parameters that modulate the chosen skill.
3. Subpolicy: produces the primitive action.

Results

The paper evaluates DEPS on two multitask environments: LIBERO and MetaWorld-v2. The results show significant quantitative performance improvements over prior work in low-data regimes. Specifically:

* DEPS achieves an average success rate of 92.1% on LIBERO, outperforming existing methods.
* DEPS achieves an average success rate of 85.6% on MetaWorld-v2, outperforming existing methods.

Significance

This work matters because it enables the learning of flexible and high-performing skills from expert demonstrations. The ability to generalize to novel tasks is crucial in many applications, including robotics, autonomous vehicles, and healthcare. The DEPS algorithm has the potential to improve sample efficiency and generalization in sequential decision-making problems.

---

[36] Learning Parameterized Skills from Demonstrations

Authors: Vedant Gupta, Haotian Fu, Calvin Luo, Yiding Jiang, George Konidaris
PDF: https://arxiv.org/pdf/2510.24095v1.pdf

Overview

Problem Statement

The paper addresses the problem of learning parameterized skills from expert demonstrations in multitask environments. This is important because traditional reinforcement learning methods often fail to leverage inherent behavioral patterns, leading to sample inefficiency. The goal is to discover modular and temporally extended skills that can be flexibly reused and composed, facilitating generalization to novel tasks.

Key Contributions

The main contributions of this work are:

* The introduction of DEPS (Discovery of Generalizable Parameterized Skills), an end-to-end algorithm for learning parameterized skills from expert demonstrations.
* The use of temporal variational inference and information-theoretic regularization methods to address degeneracy in latent variable models.
* The development of a three-level hierarchy for skill learning, including a discrete skill selector, continuous parameter selector, and subpolicy.

Methodology
DEPS uses a combination of temporal variational inference and information-theoretic regularization methods to learn parameterized skills. The algorithm consists of three levels:

1. Discrete skill selector: selects a skill from a library given the full environment observation.
2. Continuous parameter selector: outputs continuous parameters that modulate the chosen skill.
3. Subpolicy: produces the primitive action.

Results
The paper evaluates DEPS on two multitask environments: LIBERO and MetaWorld-v2. The results show significant quantitative performance improvements over prior work in low-data regimes. Specifically:

* DEPS achieves an average success rate of 92.1% on LIBERO, outperforming prior work by 10.4%.
* DEPS achieves an average success rate of 85.6% on MetaWorld-v2, outperforming prior work by 12.1%.

Significance
This work matters because it provides a scalable and generalizable approach to learning parameterized skills from expert demonstrations. The ability to learn flexible and high-performing skills has significant implications for robotics, autonomous systems, and other applications where multitask learning is essential.

---

[37] A Generalized Bisimulation Metric of State Similarity between Markov Decision Processes: From Theoretical Propositions to Applications

Authors: Zhenyu Tao, Wei Xu, Xiaohu You
PDF: https://arxiv.org/pdf/2509.18714v3.pdf

Overview

Problem Statement
The paper addresses the problem of computing state similarities between Markov Decision Processes (MDPs) in multiple-MDP scenarios. This is important because MDPs are a fundamental framework for modeling decision-making problems in Reinforcement Learning (RL), and state similarity is crucial for tasks like policy transfer, state aggregation, and sampling-based estimation.

Key Contributions
The main contributions of this paper are:

* A formal definition of the generalized bisimulation metric (GBSM) for computing state similarities between pairs of MDPs.
* Rigorous establishment of three fundamental properties of GBSM: symmetry, inter-MDP triangle inequality, and distance bound on identical state spaces.
* Theoretical analysis of policy transfer, state aggregation, and sampling-based estimation using GBSM, yielding explicit bounds for policy transfer performance, aggregation error, and estimation error.
* A closed-form sample complexity for estimation, improving upon existing asymptotic results.

Methodology
The paper uses a combination of theoretical analysis and numerical experiments to establish the properties and applications of GBSM. The methodology involves:

* Defining GBSM as a modified version of the bisimulation metric (BSM) for multiple-MDP scenarios.

* Establishing the three fundamental properties of GBSM using mathematical proofs.
* Applying GBSM to theoretical analyses of policy transfer, state aggregation, and sampling-based estimation.

Results
The key findings of this paper are:

* GBSM provides explicit bounds for policy transfer performance, aggregation error, and estimation error.
* The GBSM-derived bound is strictly tighter than the bound directly obtained from BSM.
* A closed-form sample complexity for estimation is established, improving upon existing asymptotic results.

Significance
This work matters because it provides a rigorous and efficient framework for computing state similarities between MDPs in multiple-MDP scenarios. The GBSM can be used to improve the performance of RL algorithms in tasks like policy transfer, state aggregation, and sampling-based estimation. The closed-form sample complexity for estimation can also be used to optimize the computational resources required for these tasks.

---

[38] Towards Robust Zero-Shot Reinforcement Learning

Authors: Kexin Zheng, Lauriane Teysier, Yinian Zheng, Xianyuan Zhan
PDF: https://arxiv.org/pdf/2510.15382v2.pdf

Overview

Problem Statement
The paper addresses the problem of zero-shot reinforcement learning (RL), which aims to learn a pre-trained generalist policy that can adapt to arbitrary new tasks in a zero-shot manner. This is important because traditional RL methods rely on human-provided reward functions and task-specific learning, limiting their adaptability to novel or multiple tasks.

Key Contributions
The main contributions of this paper are:

* **Behavior-REgularizEd Zero-shot RL with Expressivity enhancement (BREEZE)**: a novel framework that enhances offline learning stability and zero-shot generalization capability.
* **Behavioral regularization**: a reformulation of Forward-Backward representations (FB) that mitigates extrapolation errors while preserving representation fidelity.
* **Task-conditioned diffusion model**: a policy extraction method that enables high-quality multimodal action distributions in zero-shot RL settings.
* **Expressive attention-based architectures**: a representation modeling approach that captures complex relationships between environmental dynamics.

Methodology
The paper uses a combination of techniques, including:

* **Forward-Backward representations (FB)**: a rank-d approximation of the successor measure.
* **Behavioral regularization**: a reformulation of FB that mitigates extrapolation errors.
* **Task-conditioned diffusion model**: a policy extraction method.
* **Expressive attention-based architectures**: a representation modeling approach.

Results
The paper presents extensive experiments on the ExORL benchmark and the D4RL Kitchen dataset, demonstrating that BREEZE achieves the best or near-the-best performance while exhibiting superior robustness.

Significance
This work matters because it addresses the limitations of existing zero-shot RL methods, which often suffer from inconsistent and biased successor measures. BREEZE's innovations in behavioral regularization, task-conditioned diffusion models, and expressive attention-based architectures have the potential to improve the performance and

robustness of zero-shot RL agents.

[39] Skill-Driven Neurosymbolic State Abstractions

Authors: Alper Ahmedoglu, Steven James, Cameron Allen, Sam Loble, David Abel, George Konidaris
PDF: https://www.raillab.org/publication/ahmetoglu-2025-skill/ahmetoglu-2025-skill.pdf

Overview

Problem Statement

The paper addresses the problem of constructing state abstractions compatible with a given set of abstract actions to obtain a well-formed abstract Markov decision process (MDP). This is important because real-world tasks require decision-making at an abstract level, and current reinforcement learning (RL) methods focus on low-level action and perception, which is insufficient for complex tasks.

Key Contributions

The main contributions of this paper are:

* Deriving conditions under which abstract states represent distributions over ground states, making the resulting process Markov and approximately model-preserving.
* Developing algorithms for constructing state abstractions from data and planning with them.
* Generalizing these results to factored actions, which modify only some state variables in the ground MDP.
* Applying the resulting algorithms to visual chain and maze tasks, as well as a visual gridworld and Montezuma's Revenge.

Methodology

The paper uses a combination of theoretical analysis and algorithmic development. The authors derive conditions for abstract states to be Markov and approximately model-preserving using the Bellman equation. They then develop algorithms for constructing state abstractions from data and planning with them. The algorithms are applied to various tasks, including visual chain and maze tasks, as well as a visual gridworld and Montezuma's Revenge.

Results

The paper reports that the proposed algorithms can construct state abstractions that are Markov and approximately model-preserving. The results show that the abstract MDPs can be used for planning and learning, and that the factored action case can be handled using the proposed algorithms.

Significance

This work matters because it provides a principled and powerful framework for learning neurosymbolic abstract decision processes. The proposed algorithms can be used to construct state abstractions that are compatible with a given set of abstract actions, which is essential for complex tasks that require decision-making at an abstract level. The results have the potential to impact various fields, including robotics, computer vision, and artificial intelligence.

[40] Enhancing Safety in Reinforcement Learning with Human Feedback via Rectified Policy Optimization

Authors: Xiyue Peng, Hengguan Guo, Jiawei Zhang, Donggao Zou, Ziyu Shao, Honghao Wei, Xin Liu
PDF: https://arxiv.org/pdf/2410.19933v2.pdf

Overview

Problem Statement

The paper addresses the challenge of balancing helpfulness and safety in large language models (LLMs). The authors highlight that current approaches often decouple these two objectives, training separate preference models for helpfulness and safety, while framing safety as a constraint within a constrained Markov Decision Process (CMDP) framework. However, this approach has a potential issue, termed "safety compensation," where the constraints are satisfied on expectation, but individual prompts may trade off safety, resulting in some responses being overly restrictive while others remain unsafe.

Key Contributions

The main contributions of this paper are:

* The identification of "safety compensation" as a potential issue in current safety alignment approaches.
* The proposal of Rectified Policy Optimization (RePO), a new algorithm that replaces expected safety constraints with critical safety constraints imposed on every prompt.
* The use of rectified policy gradients to penalize strict safety violations, enhancing safety across nearly all prompts.

Methodology

The authors propose a new algorithm, RePO, which updates the policy with a rectified policy gradient by incorporating the critical safety metric as a penalty. This approach is based on a reinforcement learning algorithm and is designed to enhance safety across nearly all prompts without compromising helpfulness.

Results

The authors demonstrate the effectiveness of RePO on two large language models, Alpaca-7B and Llama3.2-3B, showing that it outperforms strong baseline methods and significantly enhances LLM safety alignment.

Significance

This work matters because it addresses a critical challenge in LLM safety alignment. By proposing a new algorithm that can guarantee safety for nearly all prompt-response pairs, the authors provide a more robust approach to safety alignment. This could have a significant impact on the development of safe and reliable LLMs, which are increasingly being used in a wide range of applications.

---

[41] Staggered Environment Resets Improve Massively Parallel On-Policy Reinforcement Learning

Authors: Sid Bhatnagar, Stone Tao, Hao Su

PDF: https://arxiv.org/pdf/2511.21011v1.pdf

Overview

Problem Statement

The paper addresses the problem of nonstationarity in massively parallel on-policy reinforcement learning (RL) environments. Specifically, it focuses on the issue of cyclical batch nonstationarity caused by synchronous full-episode resets combined with short rollouts in parallel RL settings. This problem is important because it can lead to destabilization of training and poor performance in RL algorithms.

Key Contributions

The main contributions of this paper are:

* Identifying and formulating the problem of cyclical batch nonstationarity in massively parallel on-policy RL.
* Proposing a simple yet effective technique called staggered resets, which ensures temporal diversity within training batches by desynchronizing the effective starting points of parallel environments across the task horizon.
* Characterizing the conditions under which this nonstationarity is most severe and staggered resets offer maximal benefit.
* Providing empirical evidence on challenging, high-dimensional robotics tasks, demonstrating that staggered resets significantly improve sample efficiency, wall-clock convergence speed, final policy performance, and scalability with increasing parallelism.

Methodology

The paper uses Proximal Policy Optimization (PPO) as the RL algorithm and employs a massively parallel simulation environment accelerated on GPUs. The staggered resets technique is introduced as a modification to the environment interaction protocol, which breaks the synchronicity of standard synchronous resets. The authors also use illustrative toy environments to characterize the conditions under which staggered resets are most effective.

Results

The paper reports significant improvements in sample efficiency, wall-clock convergence speed, final policy performance, and scalability with increasing parallelism using staggered resets. The authors also provide empirical evidence on challenging, high-dimensional robotics tasks, demonstrating the effectiveness of staggered resets.

Significance
This work matters because it addresses a critical issue in massively parallel on-policy RL and provides a simple yet effective solution to improve training stability and performance. The staggered resets technique can be applied to a wide range of RL algorithms and environments, making it a significant contribution to the field of RL.

---

[42] PARCO: Parallel AutoRegressive Models for Multi-Agent Combinatorial Optimization

Authors: Federico Bertoncini, Chuanbo Hu, Lauren Luttman, Jiwoo Son, Jinyoung Park
PDF: https://arxiv.org/pdf/2409.03811v3.pdf

Overview

Problem Statement
The paper addresses the problem of multi-agent combinatorial optimization (CO), which involves determining an optimal sequence of actions in discrete spaces with multiple agents. This class of problems is notoriously hard to solve due to their NP-hard nature and the necessity for effective agent coordination. Multi-agent CO problems arise in real-world applications such as coordinated vehicle routing, manufacturing, and last-mile delivery optimization.

Key Contributions
The paper proposes PARCO (Parallel AutoRegressive Combinatorial Optimization), a novel learning framework that addresses multi-agent CO problems effectively via parallel solution construction. PARCO integrates three key components: (1) transformer-based communication layers to enhance agent coordination, (2) a multiple pointer mechanism to reduce latency, and (3) priority-based conflict handlers to resolve decision conflicts.

Methodology
PARCO uses a parallel autoregressive framework for solution construction, which involves generating solutions step-by-step. The model leverages transformer-based communication layers to enable effective agent collaboration, a multiple pointer mechanism to reduce latency, and priority-based conflict handlers to resolve decision conflicts.

Results
The paper evaluates PARCO on multi-agent vehicle routing and scheduling problems, where it outperforms state-of-the-art learning methods in solution quality, generalization, and efficiency. Specifically, PARCO achieves a 23.1% improvement in solution quality and a 4.5x reduction in computational latency compared to the baseline method.

Significance
This work matters because it provides a novel solution to the challenging problem of multi-agent CO. PARCO's ability to generate high-quality solutions efficiently and effectively can have a significant impact on real-world applications such as logistics and supply chain management. The paper's contributions can also serve as a foundation for future research in multi-agent CO.

---

[43] HMARL-CBF -- Hierarchical Multi-Agent Reinforcement Learning with Control Barrier Functions for Safety-Critical Autonomous Systems

Authors: H.M. Sabbir Ahmad, Ehsan Sabouni, Alexander Wasikoff, Param Budhraj, Zijian Guo, Songyuan Zhang, Chuchu Fan, Christos Cassanaras, Wen-Chao Li
PDF: https://arxiv.org/pdf/2507.14850

Overview

Problem Statement
The paper addresses the problem of safe policy learning in multi-agent safety-critical autonomous systems. These systems require each agent to meet safety requirements at all times while cooperating with other agents to accomplish

the task. The problem is particularly challenging in partially observable settings, where learning a flat policy using existing Multi-Agent Reinforcement Learning (MARL) algorithms can suffer from scalability and high sample complexity.

Key Contributions
The paper proposes a novel Hierarchical Multi-Agent Reinforcement Learning (HMARL) approach based on Control Barrier Functions (CBFs) for safety-critical cooperative multi-agent partially observable systems. The approach decomposes the overall reinforcement learning problem into two levels: learning joint cooperative behavior at the higher level and learning safe individual behavior at the lower level. The proposed skill-based HMARL-CBF algorithm leverages skills to optimize cooperative behavior among agents while ensuring safety.

Methodology
The paper uses a hierarchical structure based on skills/options, where the high-level policy optimizes cooperative behavior among agents, and the low-level policy learns and executes safe skills. The approach is based on Control Barrier Functions (CBFs), which provide safety guarantees during both the training phase and real-world deployment. The paper also uses a Multi-Agent Semi-Markov Decision Process (MSMDP) model to formalize the problem.

Results
The paper validates the proposed approach on challenging conflicting environment scenarios where a large number of agents must each travel safely from its origin to its destination without colliding with other agents. Simulation results demonstrate superior performance and safety compliance, achieving near perfect (?95%) success/safety rate compared to existing benchmark methods.

Significance
This work matters because it addresses a critical problem in safety-critical autonomous systems, where safety is paramount. The proposed HMARL-CBF approach provides a scalable and safe solution for multi-agent reinforcement learning, which can be applied to various domains, including self-driving cars, UAVs, and swarm robotics. The approach's ability to guarantee safety during both training and deployment makes it a significant contribution to the field of autonomous systems.

---

[44] State-Covering Trajectory Stitching for Diffusion Planners

Authors: Kyoowoon Lee, Jaesik Choi
PDF: https://arxiv.org/pdf/2506.00895v3.pdf

Overview

Problem Statement
The paper addresses the challenge of improving the performance and generalization capabilities of diffusion planners in reinforcement learning (RL). Diffusion planners are a type of model that generates trajectories by transforming noise into trajectories that match a target distribution. However, their effectiveness is limited by the quality, diversity, and coverage of the offline training data. This restricts their ability to generalize to tasks outside their training distribution or longer planning horizons.

Key Contributions
The main contributions of this paper are:

* The introduction of State-Covering Trajectory Stitching (SCoTS), a reward-free trajectory augmentation framework that systematically extends trajectories to cover diverse, unexplored regions of the state space.
* A three-stage approach that includes learning a temporal distance-preserving latent representation, iterative stitching, and diffusion-based refinement.
* The demonstration of SCoTS's ability to significantly enhance the stitching capabilities and long-horizon generalization of diffusion planners.

Methodology
The SCoTS framework consists of three stages:

1. **Temporal Distance-Preserving Search**: A model is trained to encode states based on learned optimal temporal distances, facilitating efficient identification of viable trajectory segments.
2. **Iterative Stitching**: Trajectory segments are selected based on their progress along a learned direction in the latent space and their novelty relative to previously explored regions within the rollout.
3. **Diffusion-Based Refinement**: The resulting stitched trajectories are refined using a diffusion-based refinement procedure.

Results
Experiments on offline goal-conditioned benchmarks show that SCoTS significantly improves the performance and generalization capabilities of diffusion planners. Specifically:

* SCoTS improves the stitching capabilities of Hierarchical Diffuser (HD) by 25.6% on average.
* Augmented trajectories generated by SCoTS boost the performance of widely used offline goal-conditioned RL algorithms by 15.1% on average.

Significance
This work matters because it addresses a critical limitation of diffusion planners and provides a novel approach to improving their performance and generalization capabilities. The SCoTS framework has the potential to enable the development of more effective and robust RL agents that can generalize to a wider range of tasks and environments.

---

[45] Online Optimization for Offline Safe Reinforcement Learning

Authors: Yassine Chehade, Aryan Deswal, Alan Fern, Thanh Nguyen-Tang, Janardhan Rao Doppa
PDF: https://arxiv.org/pdf/2510.22027v1.pdf

Overview

Problem Statement
The paper addresses the problem of Offline Safe Reinforcement Learning (OSRL), where the goal is to learn a reward-maximizing policy from fixed data under a cumulative cost constraint. This is a critical problem in safety-critical domains such as healthcare and smart grid, where the decision-making agent needs to satisfy cost/safety constraints.

Key Contributions
The main contributions of this paper are:
* Formulating OSRL as a minimax optimization problem and providing an iterative framework based on no-regret algorithms with convergence guarantees.
* Developing a practical algorithm with convergence guarantee by avoiding the usage of off-policy evaluation and running offline RL algorithm to convergence inside each iteration of a multi-round method.
* Empirical evaluation of the practical algorithm and its variants on DSRL benchmark tasks to demonstrate its effectiveness over state-of-the-art methods.

Methodology
The proposed approach, called Online Optimization for Offline Safe RL (O3SRL), combines offline RL with online optimization algorithms. It uses a minimax optimization problem and solves it using an iterative approach that relies on two key components: an offline RL oracle and a no-regret algorithm. The practical algorithm uses K discrete values for the Lagrange variable and applies a multi-armed bandit algorithm.

Results
The experimental evaluation on multiple DSRL benchmark tasks shows that:
* Even the simplest version of the approach with two arms (K=2) is effective and results in state-of-the-art performance.
* The approach achieves excellent performance for the challenging setting of small cost constraint thresholds.
* The performance improves with higher arms (K > 2) but shows diminishing returns beyond K=5.

Significance
This work matters because it provides a novel and effective approach to solving OSRL problems, which are critical in

safety-critical domains. The proposed approach has convergence guarantees and can be combined with any offline RL algorithm, making it a robust and practical solution. The empirical evaluation demonstrates its effectiveness over state-of-the-art methods, making it a significant contribution to the field of reinforcement learning.

---

[46] Meta-Learning Objectives for Preference Optimization

Authors: Carlo Alfano, Silvia Sapora, Jakob N. Foester, Patrick Rebeschini, Yee Whye Teh
PDF: https://arxiv.org/pdf/2411.06568v3.pdf

Overview

Problem Statement

The paper addresses the problem of preference optimization (PO) in reinforcement learning, specifically in the context of Large Language Models (LLMs) alignment. PO is a paradigm that enables the alignment of machine learning systems to relative human preferences, without requiring access to absolute rewards. The problem is important because it allows for the fine-tuning of pre-trained LLMs to specific tasks and improves their safety and helpfulness.

Key Contributions

The main contributions of this work are:

* A systematic evaluation of eight existing PO algorithms on automatically generated preference datasets with varying levels of data quality, noise levels, and initial policy.
* The introduction of a novel family of offline PO algorithms using mirror descent, named Mirror Preference Optimization (MPO), which can be easily parameterized and explored via evolutionary strategies.
* The discovery of a PO algorithm within the MPO framework that largely outperforms all considered baselines in the MuJoCo benchmark.
* The demonstration that takeaways from the analysis on the MuJoCo setting can be successfully transferred onto LLM tasks.

Methodology

The paper uses a combination of techniques, including:

* MuJoCo environments and datasets for benchmarking PO algorithms.
* Mirror descent for PO algorithm design.
* Evolutionary strategies for searching and optimizing PO algorithms.
* A framework for finding PO algorithms that can be easily parametrized and explored.

Results

The key findings of this work are:

* Most existing PO algorithms struggle when dealing with noise and mixed-quality data.
* The discovered MPO algorithm outperforms all considered baselines in the MuJoCo benchmark.
* The TA-MPO algorithm demonstrates promising results in an LLM alignment task.

Significance

This work matters because it provides a comprehensive analysis of PO algorithms and introduces a novel family of offline PO algorithms that can be easily parameterized and explored. The results demonstrate the potential of PO algorithms in LLM alignment tasks, which is a critical area of research in natural language processing. The work has the potential to improve the safety and helpfulness of LLMs, which are increasingly being used in applications such as chatbots, virtual assistants, and language translation.

---

[47] Learning Human-Like RL Agents Through Trajectory Optimization with Action Quantization

Authors: Jian-Ting Gao, Tao-Cheng Chen, Peng-Chen Huang, Kuo-Hsiao Ho, Ji-Wen Huang, Tir-Long Wu, Chen Wu
PDF: https://arxiv.org/pdf/2511.15055v1.pdf

Overview

Problem Statement

The paper addresses the challenge of designing human-like reinforcement learning (RL) agents that not only succeed in tasks but also behave like humans. Current RL agents often exhibit unnatural behaviors, making them distinguishable from human players. This disparity raises concerns for both interpretability and trustworthiness.

Key Contributions

The main contributions of this paper are:

* Formulating human-likeness in RL as a trajectory optimization problem, where the objective is to produce trajectories that closely align with human behavior.
* Introducing Macro Action Quantization (MAQ), a human-likeness aware RL framework that distills human demonstrations into macro actions via Vector-Quantized Variational Autoencoder (VQVAE).
* Evaluating MAQ on the Adroit tasks in D4RL, a standard RL benchmark, and demonstrating its effectiveness in improving human-likeness.

Methodology

The paper uses the following techniques:

* Receding-horizon control to optimize over short action segments rather than entire trajectories.
* Vector-Quantized Variational Autoencoder (VQVAE) to distill human demonstrations into macro actions.
* Semi-Markov Decision Process (SMDP) to extend MDP by incorporating macro actions.

Results

The paper reports the following results:

* MAQ achieves higher success rates in task completion and substantially improves human-likeness.
* MAQ significantly increases trajectory similarity scores using Dynamic Time Warping (DTW) and Wasserstein distance metrics.
* In a human evaluation study, participants struggle to distinguish MAQ agents from humans.

Significance

This work matters because it provides a promising direction for future research in human-like RL studies. By effectively capturing human-like behavior, MAQ can improve the interpretability and trustworthiness of RL agents. The code is available at https://rlg.iis.sinica.edu.tw/papers/MAQ.

---

[48] Composite Flow Matching for Reinforcement Learning with Shifted-Dynamics Data

Authors: Lingkai Kong, Haichuan Wang, Tonghan Wang, Guojun Xiong, Milind Tambe
PDF: https://arxiv.org/pdf/2505.23062v3.pdf

Overview

Problem Statement

The paper addresses the problem of reinforcement learning (RL) with shifted-dynamics data, where the transition dynamics of the offline dataset differ from those of the online environment. This issue, known as shifted dynamics, can introduce severe mismatches that bias policy updates, destabilize the learning process, and ultimately degrade performance. The problem is important because it is common in real-world domains such as robotics, healthcare, and wildlife conservation, where access to online interactions is limited.

Key Contributions

The main contributions of this paper are:

* **Composite Flow Model**: The authors propose a composite flow model that estimates the dynamics gap by computing the Wasserstein distance between conditional transition distributions.
* **Principled Estimation**: The composite flow formulation reduces generalization error compared to standard flow matching by reusing structural knowledge embedded in the offline data.
* **Active Data Collection Strategy**: The authors propose an active data collection strategy that targets regions in the online environment where the dynamics gap relative to the offline data is high.

Methodology
The authors use a composite flow architecture, where the online flow is defined on top of the output distribution of a learned offline flow. They also use the Wasserstein distance to estimate the dynamics gap. The methodology is grounded in the theoretical connection between flow matching and optimal transport.

Results
The authors empirically validate their method on various RL benchmarks with shifted dynamics and demonstrate that COMPFLOW outperforms or matches state-of-the-art baselines across these tasks. The results show that COMPFLOW achieves lower generalization error and improves performance in regions with high dynamics gap.

Significance
This work matters because it addresses a critical challenge in RL with offline data and provides a principled approach to estimating the dynamics gap. The active data collection strategy can improve performance in regions with high dynamics gap, which is often underrepresented in the replay buffer. The impact of this work could be significant in real-world domains where access to online interactions is limited.

---

[49] ThinkBench: Dynamic Out-of-Distribution Evaluation for Robust LLM Reasoning

Authors: Shulin Huang, Liny Yang, Yan Song, Shuang Chen, Leyang Cui, Ziyu Wan, Qingzheng Zheng, Ying Wen, Kun Shao, Wenhan Zhang, Jun Wang, Yue Zhang
PDF: https://arxiv.org/pdf/2502.16268v1.pdf

Overview

Problem Statement
The paper addresses the problem of evaluating large language models (LLMs) accurately, particularly in reasoning tasks. The issue of data contamination and leakage of correct answers poses significant challenges in evaluating LLMs. The authors aim to develop a robust evaluation framework to assess the generalization abilities of LLMs, rather than their memorization capabilities.

Key Contributions
The main contributions of this work are:

* Introducing ThinkBench, a novel evaluation framework designed to evaluate LLMs' reasoning capability robustly.
* Proposing a dynamic data generation method for constructing out-of-distribution (OOD) datasets.
* Developing a unified evaluation framework for both reasoning models and non-reasoning models.
* Creating a high-quality OOD dataset of 2,912 samples, which is more challenging than the original datasets.

Methodology
The authors use a combination of techniques, including:

* Causal theory and semi-factual causality to design the OOD data generation method.
* Dynamic evaluation to test the generalization capabilities of LLMs.
* A diverse set of challenges, including math reasoning tasks and scientific questions, to evaluate the models.

Results
The key findings are:

* Most LLMs' performance is far from robust and faces a certain level of data leakage.
* The dynamically constructed OOD data construction reduces the impact of data contamination.
* The performance decay of 24.9% and 11.8% across all models on AIME-500 and AIME 2024, respectively, indicates the importance of mitigating data contamination.

Significance

This work matters because it provides a robust evaluation framework for LLMs, which is essential for developing and improving these models. The ThinkBench framework can help reduce data contamination and leakage, enabling more accurate evaluations of LLMs' reasoning capabilities. This has significant implications for the development of more generalizable and robust LLMs.

---

[50] Fully Autonomous Neuromorphic Navigation and Dynamic Obstacle Avoidance

Authors: Kai-Enn Chang, Peng-Eni Liang, Xin-Ming Yang, Jai-Hou Chen

PDF: https://openreview.net/pdf?id=11fe8wKkmk

Overview

Problem Statement

The paper addresses the challenge of enabling unmanned aerial vehicles (UAVs) to rely solely on onboard computation and sensing for real-time navigation and dynamic obstacle avoidance. This is crucial due to the limitations of external aids such as GPS and ground stations, which can be jammed or interfered with in various scenarios.

Key Contributions

The main contributions of this paper are:

* A fully neuromorphic framework for tiny UAVs to perform navigation and dynamic obstacle avoidance tasks using only onboard resources.
* A bio-inspired approach that enables accurate moving object detection and avoidance with a latency of 2.3 milliseconds.
* The creation of a monocular event-based pose correction dataset with over 50,234 paired and labeled event streams.

Methodology

The proposed framework uses a monocular event camera and an inertial measurement unit (IMU) to enable the UAV to navigate and avoid obstacles. The navigation module uses IMU data and an SCNN network to mitigate error drift. The obstacle avoidance module uses a bio-inspired algorithm to suppress events from static objects and directly output evasion maneuvers.

Results

The results show that the proposed framework achieves:

* A latency of 2.3 milliseconds for obstacle avoidance.
* Energy consumption reduced to 21% compared to traditional architectures.
* Superior performance compared to state-of-the-art dynamic obstacle avoidance approaches.

Significance

This work matters because it enables tiny UAVs to perform complex tasks autonomously, without relying on external aids. The proposed framework has the potential to improve the performance and safety of UAVs in various applications, such as surveillance, inspection, and search and rescue. The creation of a large-scale event-based pose correction dataset also contributes to the development of neuromorphic computer vision.

---

[51] ThinkAct: Vision-Language-Action Reasoning via Reinforced Visual Latent Planning

Authors: Chi-Pin Huang, Yue-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, Fu-En Yang
PDF: https://arxiv.org/pdf/2507.16815v2.pdf

Overview

Problem Statement

The paper addresses the challenge of enabling multimodal large language models (MLLMs) to reason before acting in physical environments, particularly in tasks requiring long-horizon planning and adaptation. This is important because current MLLMs excel in understanding multimodal inputs but struggle with multi-step planning and interacting with dynamic environments.

Key Contributions

The main contributions of this paper are:

* Proposing ThinkAct, a dual-system framework that connects structured reasoning with executable actions.
* Leveraging action-aligned rewards derived from visual goal completion and trajectory distribution matching to incentivize long-horizon planning.
* Advancing visual latent planning to steer downstream action execution by providing reasoning-enhanced trajectory guidance.

Methodology

ThinkAct uses a dual-system architecture that consists of:

* A multimodal LLM that generates embodied reasoning plans guided by reinforcing action-aligned visual rewards.
* A downstream action model that executes actions based on the compressed visual plan latent.

The framework uses reinforcement learning to incentivize reasoning behaviors and visual latent planning to steer action execution.

Results

The paper demonstrates that ThinkAct enables few-shot adaptation, long-horizon planning, and self-correction behaviors in complex embodied AI tasks. The results show that ThinkAct outperforms existing approaches on embodied reasoning and robot manipulation benchmarks, with a 25.1% improvement in few-shot adaptation.

Significance

This work matters because it enables MLLMs to reason before acting in physical environments, which is crucial for applications such as robotics and AR assistance. ThinkAct's ability to adapt to new environments and correct mistakes in real-time has significant implications for the development of more robust and efficient physical AI systems.

---

[52] ThinkAct: Vision-Language-Action Reasoning via Reinforced Visual Latent Planning

Authors: Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, Fu-En Yang
PDF: https://arxiv.org/pdf/2507.16815v2.pdf

Overview

Problem Statement

The paper addresses the challenge of enabling multimodal large language models (MLLMs) to reason before acting in physical environments. Current approaches to vision-language-action (VLA) tasks rely on end-to-end training, which hinders the ability to plan over multiple steps or adapt to complex task variations. This limitation is crucial, as it prevents the development of robust and flexible AI systems that can interact with dynamic environments.

Key Contributions

The main contributions of this paper are:

* The proposal of ThinkAct, a dual-system framework that connects structured reasoning with executable actions.
* The use of action-aligned rewards derived from visual goal completion and trajectory distribution matching to incentivize long-horizon planning.
* The advancement of visual latent planning to steer downstream action execution by providing

reasoning-enhanced trajectory guidance.

Methodology
ThinkAct leverages reinforcement learning to incentivize MLLMs to perform long-horizon planning. The framework consists of two main components: a multimodal LLM that generates embodied reasoning plans, and a downstream action model that executes actions based on the generated plans. The LLM is trained using a combination of fully supervised fine-tuning and reinforcement learning, while the action model is trained using a visual latent planning approach.

Results
The paper presents extensive experiments on embodied reasoning and robot manipulation benchmarks, demonstrating that ThinkAct enables few-shot adaptation, long-horizon planning, and self-correction behaviors in complex embodied AI tasks. The results show that ThinkAct outperforms existing approaches, such as OpenVLA, on tasks like picking up objects and placing them in compartments.

Significance
This work matters because it enables the development of robust and flexible AI systems that can interact with dynamic environments. ThinkAct has the potential to unleash a wide range of physical AI applications, such as robotics and AR assistance, and can be applied to various tasks, including robot manipulation, navigation, and human-robot interaction.

---

[53] RLGF: Reinforcement Learning with Geometric Feedback for Autonomous Driving Video Generation

Authors: Tian-Yi Yan, Wencheng Han, Xia-Zhao, Kun-Zhao, Cheng-Zhong Xu, Jian-Bing Shen
PDF: https://arxiv.org/pdf/2509.16500v2.pdf

Overview

Problem Statement
The paper addresses the issue of geometric distortions in synthetic video generation for autonomous driving (AD) systems. Current state-of-the-art video generation models, despite their visual realism, suffer from subtle geometric flaws that limit their utility for downstream perception tasks. This problem is critical because it undermines the reliability of models trained or evaluated using such data, significantly constraining their applicability in essential use cases.

Key Contributions
The main contributions of this work are:

* **Reinforcement Learning with Geometric Feedback (RLGF)**: a novel framework that injects perception-model-driven geometric spatial constraints directly into the video generation process.
* **Latent-Space Windowing Optimization**: an efficient training strategy that applies rewards to noisy latent features within a randomly sampled sliding window of intermediate diffusion steps.
* **Hierarchical Geometric Reward (HGR)**: a multi-level feedback system designed to imbue generated videos with robust geometric fidelity and scene coherence.

Methodology
The paper uses a combination of techniques, including:

* **Diffusion-based video generation models**: such as DiVE, which are optimized via pixel-level supervision.
* **Pre-trained AD perception models**: as reward providers to ensure geometric fidelity.
* **GeoScores**: a metric suite that evaluates geometric fidelity by applying pre-trained perception models to both synthetic videos and their corresponding real-world counterparts.

Results
The paper reports significant performance improvements, including:

* **21% reduction in VP error** and **57% reduction in Depth error** using RLGF.
* **12.7% improvement in 3D object detection mAP** using RLGF.

* **Narrowing the gap to real-data performance** in 3D object detection.

Significance

This work matters because it addresses a critical issue in synthetic video generation for AD systems. By providing a plug-and-play solution for generating geometrically sound and reliable synthetic videos, RLGF has the potential to significantly improve the development and deployment of AD systems.

---

[54] Toward Artificial Palpation: Representation Learning of Touch on Soft Bodies

Authors: Zohar Rimon, Elisei Shafer, Tal Tepper, Efrat Shimron, Aviv Tamar
PDF: https://arxiv.org/pdf/2511.16596v1.pdf

Overview

Problem Statement

The paper addresses the problem of artificial palpation, which is the use of touch in medical examination. The authors aim to develop a system that can learn to interpret tactile measurements into corresponding mechanical structures, potentially leading to more accurate imaging than currently available. This is motivated by the fact that palpation is still an important tool in medical examination, particularly in breast cancer detection, where a large fraction of cases are discovered by palpation.

Key Contributions

The main contributions of this paper are:

* A proof of concept system for learning artificial breast palpation using self-supervised learning.
* A novel soft breast phantom with a modular component that can include lumps with various sizes and shapes.
* A neural representation learned by minimizing tactile force prediction error, which contains relevant information about the position and shape of the lump.
* Tactile images that are arguably easier to interpret than a map of forces, which can be used for change detection at a level comparable to humans.

Methodology

The authors use a combination of techniques, including:

* Self-supervised learning to learn a general, artificial, palpation representation.
* An encoder-decoder neural network to predict tactile measurements at given positions from a sequence of previous measurements.
* A robotic manipulator with a tactile sensor tip programmed to palpate the phantom.
* MRI scans of the objects as ground truth object models.

Results

The authors report that their learned representation contains relevant information about the position and shape of the lump, and yields tactile images that are arguably easier to interpret than a map of forces. They also report that their system can detect changes in the size of the lump at a level comparable to humans.

Significance

This work matters because it has the potential to improve tactile imaging and detection accuracy in medical examination. The authors' system can learn to interpret tactile measurements into corresponding mechanical structures, potentially leading to more accurate imaging than currently available. This could have a significant impact on medical diagnosis and treatment, particularly in breast cancer detection.

---

[55] GaussianFusion: Gaussian-Based Multi-Sensor Fusion for End-to-End Autonomous Driving

Authors: Shuai Lu, Quanmin Liang, Zeng Li, Boyang Li, Kai Huang

PDF: https://arxiv.org/pdf/2506.00034v2.pdf

Overview

Problem Statement

The paper addresses the problem of multi-sensor fusion in end-to-end (E2E) autonomous driving systems. E2E autonomous driving relies on deep learning to map sensor inputs to driving actions, but relying on a single sensor limits the system's ability to handle diverse and challenging driving scenarios. Multi-sensor fusion is essential to leverage complementary information from different sensors, enhancing perception reliability and providing richer input for learning robust driving policies.

Key Contributions

The paper proposes GaussianFusion, a Gaussian-based multi-sensor fusion framework for E2E autonomous driving. The key contributions are:

* Introducing Gaussian representations into multi-sensor fusion for E2E autonomous driving
* Proposing a dual-branch fusion pipeline tailored to the planning-centric task
* Designing a cascade planning head that iteratively refines trajectories through hierarchical Gaussian queries

Methodology

The proposed framework uses 2D Gaussians to represent the traffic scene, which are initialized uniformly across the driving scene. The Gaussians are progressively refined by integrating multi-modal features. A dual-branch fusion pipeline is designed, with one branch capturing local features for traffic scene reconstruction and the other aggregating global planning cues for motion planning. A cascade planning module refines anchor trajectories by querying the Gaussian representations.

Results

The paper evaluates GaussianFusion on the NAVSIM and Bench2Drive benchmarks. On NAVSIM, the approach achieves 92.0 PDMS (Planning Distance Metric Score) with the V2-99 backbone, surpassing current state-of-the-art methods. On Bench2Drive, the results consistently demonstrate the effectiveness of GaussianFusion.

Significance

GaussianFusion has the potential to improve the performance and robustness of E2E autonomous driving systems by leveraging Gaussian representations for multi-sensor fusion. The approach can be applied to various autonomous driving tasks, including planning, perception, and control. The use of Gaussian representations can also improve the interpretability and efficiency of multi-sensor fusion methods.

---

[56] Flow Matching-Based Autonomous Driving Planning with Advanced Interactive Behavior Modeling

Authors: Tianyi Tan, Yinian Zheng, Ruiming Liang, Zexu Wang, Kexin Zheng, Jinliang Zheng, Jianxiong Li, Xianyuan Zhan, Jingjing Liu

PDF: https://arxiv.org/pdf/2510.11083v1.pdf

Overview

Problem Statement

The paper addresses the challenge of modeling interactive driving behaviors in complex scenarios for autonomous driving planning. This is a critical problem because conventional rule-based approaches are limited by fundamental constraints, requiring substantial human engineering efforts and exhibiting poor generalization capability in dynamic environments. Learning-based methods aim to directly learn expert strategies from real-world driving data, but current approaches often fail to effectively capture intricate interdependencies among heterogeneous information.

Key Contributions

The main contributions of this paper are:

* Fine-grained trajectory tokenization, which decomposes the trajectory into overlapping segments to decrease the complexity of whole trajectory modeling.
* A sophisticatedly designed architecture that achieves efficient temporal and spatial fusion of planning and scene information.
* Flow matching with classifier-free guidance for multi-modal behavior generation, which dynamically reweights agent interactions during inference to maintain coherent response strategies.

Methodology

The proposed framework, Flow Planner, incorporates:

* Fine-grained trajectory tokenization to reduce complexity.
* A scale-adaptive attention mechanism for spatiotemporal fusion of scene and planning tokens.
* Flow matching loss with classifier-free guidance for multi-modal behavior generation.

Results

Experimental results on the nuPlan and interPlan datasets show that Flow Planner achieves state-of-the-art performance among learning-based planners, with significant improvements in interactive scenario understanding. Specifically, Flow Planner outperforms other methods on the nuPlan dataset, achieving a 10.5% improvement in planning accuracy.

Significance

This work matters because it addresses a critical challenge in autonomous driving planning, enabling more effective modeling of interactive driving behaviors in complex scenarios. The proposed framework, Flow Planner, has the potential to improve the safety and reliability of autonomous driving systems, and its impact could be significant in the development of autonomous vehicles.

---

[57] Temporal Representation Alignment: Successor Features Enable Emergent Compositionalities in Robot Instruction Following

Authors: Vivek Myers, Bill Chunyuan Zheng, Anca Dragan, Kuan Fang, Sergey Levine

PDF: https://arxiv.org/pdf/2502.05454v2.pdf

Overview

Problem Statement

The paper addresses the problem of compositional generalization in robot learning, specifically in the context of robotic manipulation tasks. Compositional generalization refers to the ability of an agent to generalize to new behaviors that are composed of known sub-behaviors. This is a crucial aspect of intelligent behavior, as it enables agents to adapt to new situations and tasks. However, current approaches to robot learning often struggle with compositional generalization, leading to limited performance in real-world settings.

Key Contributions

The main contributions of this paper are:

* The introduction of Temporal Representation Alignment (TRA), a method that learns to align state representations with tasks across time to enable compositional behavior.
* The use of a time-contrastive alignment loss to improve compositional performance.
* The demonstration of TRA's effectiveness in a real-world tabletop manipulation setting, with substantial improvements in compositional performance (>40% across 13 tasks in 4 evaluation scenes).

Methodology

The paper uses a combination of goal- and language-conditioned control, with a focus on temporal representation alignment. The TRA method involves learning a structured representation of the world through a time-contrastive alignment loss, which is added as an auxiliary loss to the policy learning objective. This approach enables the agent to learn to compose behaviors from known sub-behaviors.

Results
The paper presents results on a set of challenging multi-step manipulation tasks in the BridgeData setup and the OGBench simulation benchmark. TRA achieves substantial improvements in compositional performance, outperforming past imitation and reinforcement learning baselines.

Significance
This work matters because it provides a new approach to compositional generalization in robot learning, which is a crucial aspect of intelligent behavior. The TRA method has the potential to enable robots to adapt to new situations and tasks, and could have a significant impact on the development of autonomous robots and intelligent systems.

---

[58] Understanding while Exploring: Semantics-Driven Active Mapping

Authors: Lyan Chen, Huangyuan Zhan, Haorong Yin, Yi Xu, Philippos Mordohi
PDF: https://arxiv.org/pdf/2506.00225v2.pdf

Overview

Problem Statement
The paper addresses the problem of active semantic mapping, which is crucial for robotic autonomy in unknown environments. Effective exploration and understanding of both geometry and semantics are essential for robots to navigate and interact with their surroundings. Current approaches to semantic mapping are limited by their inability to determine the most informative path for the robot, leading to incomplete or suboptimal scene understanding.

Key Contributions
The main contributions of this paper are:
* The proposal of ActiveSGM, an active semantic mapping framework that predicts the informativeness of potential observations before execution.
* The use of a 3D Gaussian Splatting (3DGS) mapping backbone, which integrates semantic-aware mapping and planning for active reconstruction.
* A novel semantic exploration criterion that enhances semantic coverage and facilitates disambiguation across observations during exploration.
* A sparse semantic representation that retains the top-k most probable categories, reducing memory overhead without sacrificing semantic richness.

Methodology
The proposed method uses a 3DGS mapping backbone, which is combined with a semantic-aware mapping and planning module. The semantic exploration criterion is used to select the most informative views for the robot, and the sparse semantic representation is used to reduce memory overhead. The method also uses a pre-trained segmentation model to generate noisy semantic predictions, which are then refined progressively.

Results
The experiments on the Replica and Matterport3D datasets show that ActiveSGM outperforms state-of-the-art methods in terms of mapping completeness, accuracy, and robustness to noisy semantic data. Specifically, ActiveSGM achieves a 25% improvement in mapping completeness and a 15% improvement in accuracy compared to the baseline method.

Significance
This work matters because it provides a novel approach to active semantic mapping, which is essential for robotic autonomy in unknown environments. The proposed method can be used in a variety of applications, including autonomous navigation, object recognition, and scene understanding. The use of a 3DGS mapping backbone and a sparse semantic representation can also be applied to other areas of computer vision and robotics.

---

[59] Towards Reliable Code-as-Policies: A Neuro-Symbolic Framework for Embodied Task Planning

Authors: Sanghyun Ahn, Wonje Choi, Jinnyong Lee, Jinwoo Park, Honguk Woo

PDF: https://arxiv.org/pdf/2510.21302v1.pdf

Overview

Problem Statement

The paper addresses the challenge of generating reliable executable code for task planning in dynamic, partially observable environments. This is crucial for embodied agents, such as robots, that require accurate and complete code to perform tasks successfully. The current state-of-the-art approaches, like Code as Policies, suffer from limited environmental grounding, leading to suboptimal task success rates.

Key Contributions

The main contributions of this work are:

* The NESYRO framework, which incorporates explicit symbolic verification and interactive validation processes during code generation.
* A novel recursive mechanism that composes two phases: Neuro-symbolic Code Verification and Neuro-symbolic Code Validation.
* The framework's ability to ground generated code, resulting in improved task reliability and success rates in complex environments.

Methodology

The NESYRO framework uses a combination of symbolic verification and interactive validation to generate executable code. Symbolic verification statically checks code correctness using domain-specific symbolic tools, while interactive validation enables the agent to actively explore its environment to resolve ambiguities and acquire missing observations.

Results

Experimental results demonstrate that NESYRO improves task success rate by 46.2% over the state-of-the-art baseline, Code as Policies. The framework achieves over 86.8% executability of task-relevant actions in real-world settings.

Significance

This work matters because it addresses a critical challenge in embodied intelligence. The NESYRO framework has the potential to improve the reliability and success rates of task planning in dynamic, partially observable environments, enabling more robust and efficient robotic control. The framework's ability to ground generated code and its recursive mechanism make it a significant innovation in the field of neuro-symbolic robot task planning.

---

[60] InstructFlow: Adaptive Symbolic Constraint-Guided Code Generation for Long-Horizon Planning

Authors: Haotian Chi, Zeyu Feng, Yue-Ming Lyu, Cheng-En Lin, Yew-Son Ong, Ivory Tsang

PDF: Available (local only)

Overview

Problem Statement

The paper addresses the challenge of long-horizon planning in robotic manipulation tasks, where language model-based planners struggle with decomposing tasks, satisfying constraints, and recovering from failures. This is a critical problem because it affects the reliability and robustness of robotic systems, which require precise and adaptive planning to execute complex tasks.

Key Contributions

The main contributions of this work are:

* **InstructFlow framework**: a modular, multi-agent system that establishes a symbolic, feedback-driven flow of information for adaptive task planning and code generation.
* **Hierarchical instruction graph**: a structured representation of tasks that supports dynamic, feedback-driven updates based on induced constraints.
* **Symbolic constraint induction**: a mechanism that abstracts raw execution failures into interpretable symbolic predicates, enabling precise and generalizable plan and code refinement.

Methodology
The InstructFlow framework consists of three key components:

* **InstructFlow Planner**: constructs and traverses a hierarchical instruction graph that decomposes tasks into semantically meaningful subtasks.
* **Code Generator**: generates executable code snippets conditioned on the instruction graph.
* **Constraint Generator**: analyzes feedback and induces symbolic constraints, which are propagated back into the instruction graph to guide targeted code refinement.

Results
The paper presents comprehensive experiments on three benchmarks: Drawing, Arrange-block, and Arrange-YCB. The results show that InstructFlow significantly improves task success rates and robustness, especially in constraint-sensitive and long-horizon scenarios. Specifically:

* **Success rates**: InstructFlow achieves 92.5% success rate on Drawing, 85.2% on Arrange-block, and 81.4% on Arrange-YCB, outperforming strong LLM-based baselines.
* **Robustness**: InstructFlow demonstrates improved robustness in constraint-sensitive and long-horizon tasks, with a 25% reduction in failure rates compared to baselines.

Significance
This work matters because it addresses a critical challenge in robotic manipulation tasks and provides a robust and efficient solution. The InstructFlow framework has the potential to improve the reliability and adaptability of robotic systems, enabling them to execute complex tasks in dynamic environments.

---

[61] 3D Equivariant Visuomotor Policy Learning via Spherical Projection

Authors: Bocchini, Boce Hu, Dian Wang, David Klee, Heng Tian, Xupeng Zhu, Haajie Huang, Robert Platt, Robin Walters
PDF: https://arxiv.org/pdf/2505.16969v3.pdf

Overview

Problem Statement
This paper addresses the challenge of achieving SO(3)-equivariance in visuomotor policy learning using only monocular RGB inputs in eye-in-hand settings. This is important because existing equivariant models rely on point cloud data or multi-camera setups, which are not compatible with the common eye-in-hand configuration.

Key Contributions
The main contributions are:
* Introducing Image-to-Sphere Policy (ISP), the first SO(3)-equivariant policy learning framework that uses spherical projection from 2D RGB inputs to model 3D symmetries.
* Theoretically proving that ISP achieves global SO(3)-equivariance and local SO(2)-invariance, facilitating policy learning.
* Validating ISP through extensive experiments, achieving an average success rate improvement of 11.6% over twelve simulation tasks and 42.5% across four real-world tasks.

Methodology
The ISP framework uses spherical projection to transform 2D RGB features onto a sphere, and then rotates the resulting spherical signal to compensate for camera motion. This yields a stable, SO(3)-equivariant representation that is well-suited for downstream equivariant architectures.

Results
The experiments demonstrate that ISP consistently outperforms strong baselines in terms of both performance and sample efficiency. The average success rate improvement is 11.6% over twelve simulation tasks and 42.5% across four real-world tasks.

Significance
This work matters because it provides a novel framework for achieving SO(3)-equivariance in visuomotor policy learning using only monocular RGB inputs. This has the potential to improve data efficiency and generalization in robotic manipulation tasks, and can serve as a modular, plug-and-play component that generalizes seamlessly to richer sensing setups.

---

[62] DexFlyWheel: A Scalable and Self-Improving Data Generation Framework for Dextrous Manipulation

Authors: Kefei Zhu, Fengshuo Bai, Yuanhao Xiang, Yishu Cai, Xinglin Chen, Ruchong Li, Xingtao Wang, Hao Dong, Yaoong Yang, Xiaopeng Fan, Yuanpei Chen

PDF: Available (local only)

Overview

Problem Statement
The paper addresses the problem of generating diverse and high-quality datasets for dexterous manipulation tasks, which is crucial for advancing robot capabilities in real-world applications. The scarcity of such datasets is a significant bottleneck in robotics research, and existing methods either rely on human teleoperation or generate data with limited diversity.

Key Contributions
The main contributions of this paper are:

* The proposal of DexFlyWheel, a scalable and self-improving data generation framework for dexterous manipulation.
* The combination of Imitation Learning (IL) and residual Reinforcement Learning (RL) to generate human-like and diverse data.
* The design of a data flywheel mechanism that iteratively expands data diversity, enhances policy generalization, and evolves into a robust data generation agent.

Methodology
DexFlyWheel employs a self-improving cycle that integrates IL, residual RL, policy rollouts, and data augmentation. The framework starts with efficient seed demonstrations, which are then expanded through iterative cycles. Each cycle follows a closed-loop pipeline that extracts human-like behaviors from demonstrations, enhances policy generalization, and generates diverse trajectories.

Results
The experimental results demonstrate the effectiveness of DexFlyWheel on four dexterous manipulation tasks. The framework generates over 2,000 diverse demonstrations across 500+ scenarios, outperforming baseline methods. Policies trained on the generated data achieve an average success rate of 81.9% on challenging test sets and transfer to a real-world dual-arm robot system with a 78.3% success rate on the dual-arm lift task.

Significance
This work matters because it addresses the scarcity of dexterous manipulation data and provides a scalable and self-improving framework for generating diverse and high-quality datasets. The proposed framework has the potential to significantly improve the performance of policies in dexterous manipulation tasks and enable the development of more advanced robot capabilities.

---

[63] Knowledge insulating VLAs: Train fast, run fast, generalize better

Authors: Danny Driess, Jost Tobias Springenberg, Brian Ichtter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z. Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, Sergey Levine

PDF: Available (local only)

Overview

Problem Statement

The paper addresses the challenge of adapting large language models (LLMs) and vision-language models (VLMs) to real-world control, specifically for robotic manipulation tasks. The goal is to leverage the semantic knowledge distilled from web-scale pre-training to improve the performance of vision-language-action (VLA) models in controlling robots. However, the constraints of real-time control and the large number of parameters in VLMs pose significant obstacles.

Key Contributions

The main contributions of this work are:

* A novel approach called "knowledge insulation" that separates the training of the VLM backbone from the action expert, allowing for faster and more stable training.
* A technique for insulating the VLM backbone during VLA training, which mitigates the issue of degraded knowledge transfer.
* An extensive analysis of various design choices and their impact on performance and knowledge transfer.

Methodology

The paper proposes a training recipe that fine-tunes the VLM backbone with discretized actions while adapting an action expert to produce continuous actions. The key idea is to use next-token prediction to learn good representations for robotic control, while the action expert is trained with flow-matching or diffusion. The approach is illustrated in Figure 1.

Results

The experimental evaluation provides an extensive analysis of the various modeling choices in continuous-action VLAs. The results show that knowledge insulation improves training speed and knowledge transfer, and enables co-training on general vision-language data.

Significance

This work matters because it addresses a critical challenge in adapting LLMs and VLMs to real-world control. The proposed approach has the potential to improve the performance of VLA models in controlling robots, and could have a significant impact on the field of robotics and AI. The results demonstrate the importance of knowledge insulation in preserving the semantic knowledge contained in pre-trained VLMs.

---

[64] OSVI-WM: One-Shot Visual Imitation for Unseen Tasks using World-Model-Based Trajectory Synthesis

Authors: Ratin Gadeon Grewal, Prashant Krishna, Yann LeCun, Farnah Khorram

PDF: https://arxiv.org/pdf/2505.20425v2

Overview

Problem Statement

The paper addresses the problem of one-shot visual imitation (OSVI) learning, where a robotic agent must derive a policy from a single expert demonstration video without additional training. This is a challenging task, as existing methods often rely on strong assumptions that the training and testing tasks are nearly identical, and struggle to generalize to unseen tasks with different semantic or structural requirements.

Key Contributions

The main contributions of this work are:

* An efficient end-to-end imitation learning architecture trained solely on in-domain data, without requiring large-scale pretraining.

* A novel world-model-guided trajectory generation module tailored for OSVI on unseen tasks.
* Robustness enhancement at test time by using a waypoint controller with re-planning.

Methodology
The proposed method, OSVI-WM, uses a world-model-guided trajectory generation module to predict a sequence of latent states and actions from an expert demonstration video and the agent's initial observation. The predicted latent trajectory is then decoded into physical waypoints that guide the agent's execution. The method uses a combination of neural networks, including a world model, action model, and waypoint controller, to achieve this.

Results
The paper presents extensive experiments on two simulated benchmarks and three real-world robotic platforms, demonstrating that OSVI-WM outperforms existing methods on unseen tasks. Specifically, OSVI-WM achieves over 30% improvement in some cases, and consistently outperforms prior approaches.

Significance
This work matters because it addresses a critical limitation of existing OSVI methods, which struggle to generalize to unseen tasks. By using a world-model-guided trajectory generation module, OSVI-WM can reason about future states and plan into the future, making it a more robust and effective approach. The code is available at https://github.com/raktimgg/osvi-wm.

---

[65] VideoVLA: Video Generators Can Be Generalizable Robot Manipulators

Authors: Yichao Shen, Fangyuan Wei, Zhiyong Du, Taobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, Banning Guo
PDF: https://arxiv.org/pdf/2512.06963v1.pdf

Overview

Problem Statement
The paper addresses the problem of generalization in robot manipulation, which is essential for deploying robots in open-world environments and advancing toward artificial general intelligence. Despite recent advances in Vision-Language-Action (VLA) models, their ability to generalize to novel tasks, objects, and settings remains limited.

Key Contributions
The main contributions of this work are:

* Proposing a simple yet effective approach called VideoVLA, which transforms a Video Diffusion Transformer into a Video-Action Diffusion Transformer by adding actions as a new output modality.
* Using pre-trained video generation models as the backbone for VLA models, which enables the transfer of knowledge from general-purpose models to robotic manipulation tasks.
* Demonstrating strong generalization capabilities, including imitating other embodiments' skills and handling novel objects.

Methodology
The proposed approach, VideoVLA, uses a multi-modal Diffusion Transformer architecture, which jointly models video, language, and action modalities. The model operates by taking the language tokens and the latent of the current visual observation as conditions, and jointly predicts the future actions and generates the corresponding future visual contents.

Results
The experiments show a strong correlation between the predicted actions and the generated video clips, with a correlation coefficient of 0.85. The model also demonstrates strong generalization capabilities, including imitating other embodiments' skills and handling novel objects.

Significance
This work matters because it explores a paradigm shift in robot learning by using pre-trained video generation models as the backbone for VLA models. The proposed approach, VideoVLA, has the potential to unlock generalization capabilities in manipulation systems, enabling robots to handle unseen tasks, manipulate novel objects, and operate in

unfamiliar environments.

[66] Continual Optimization with Symmetry Teleportation for Multi-Task Learning

Authors: Zhipeng Zhou, Zigao Meng, Pengcheng Wu, Pelin Zhu, Chunyan Ma
PDF: https://arxiv.org/pdf/2503.04046v1.pdf

Overview

Problem Statement

The paper addresses the problem of multi-task learning (MTL), where a single model is trained to perform multiple tasks simultaneously. MTL is important because it enables efficient learning of multiple tasks using a single model, reducing computational and storage demands.

Key Contributions

The main contributions of this paper are:

* A novel approach to MTL called Continual Optimization with Symmetry Teleportation (COST), which seeks an alternative loss-equivalent point on the loss landscape to reduce conflict.
* A practical teleportation method using low-rank adapters (LoRA) to facilitate symmetry teleportation.
* A historical trajectory reuse strategy to continually benefit from advanced optimizers.

Methodology

The paper proposes a new approach to MTL optimization, which involves:

* Using LoRA to facilitate symmetry teleportation and reduce conflict.
* Designing convergent, loss-invariant objectives to promote progress.
* Introducing a historical trajectory reuse strategy to continually benefit from advanced optimizers.

Results

The paper presents extensive experiments on multiple mainstream datasets, including:

* Achieving state-of-the-art performance on several MTL benchmarks.
* Enhancing the performance of existing MTL methods by up to 10% on average.
* Demonstrating the effectiveness of COST in reducing conflict and improving convergence.

Significance

This work matters because it provides a novel approach to MTL optimization, which can improve the performance of MTL models. The proposed method, COST, is a plug-and-play solution that can enhance existing MTL methods, making it a significant contribution to the field of MTL.

[67] Improved Regret Bounds for Gaussian Process Upper Confidence Bound in Bayesian Optimization

Authors: Shogo Iwazaki
PDF: https://arxiv.org/pdf/2506.01393v3.pdf

Overview

Problem Statement

The paper addresses the Bayesian optimization problem, where a learner seeks to minimize regret under a function drawn from a known Gaussian process (GP). This problem is important because it has applications in various fields, such as robotics, finance, and healthcare, where optimizing a complex function is crucial.

Key Contributions

The main contributions of this paper are:

* Improving the regret upper bound for the Gaussian process upper confidence bound (GP-UCB) algorithm from e^O(?T) to e^O(?T) with high probability under a Mat´ern kernel with a certain degree of smoothness.

* Establishing an O(?Tln2T) cumulative regret of GP-UCB for a squared exponential kernel, improving the existing O(?Tln(d+2)T) upper bound.

Methodology

The paper uses a combination of techniques, including:

* Leveraging algorithm-dependent behavior and sample path properties of the GP to refine information gain bounds.
* Decomposing cumulative regret into lenient regret-based terms and analyzing them using techniques from [8, 28].
* Using the GP-UCB algorithm, which combines the posterior distribution of GP with the optimism principle.

Results

The key findings are:

* GP-UCB achieves e^O(?T) regret with high probability under a Mat´ern kernel with a certain degree of smoothness.
* GP-UCB achieves O(?Tln2T) cumulative regret for a squared exponential kernel.

Significance

This work matters because it improves the theoretical understanding of the performance of GP-UCB, a widely used algorithm in Bayesian optimization. The results have implications for the design of more efficient algorithms and the development of new applications in fields such as robotics, finance, and healthcare.

---

[68] Self-Generated In-Context Examples Improve LLM Agents for Sequential Decision-Making Tasks

Authors: Vishnu Sarukkai, Zhiqiang Xie, Kayvon Fatahalian

PDF: https://arxiv.org/pdf/2505.00234v3.pdf

Overview

Problem Statement

The paper addresses the problem of improving Large Language Model (LLM) agents for sequential decision-making tasks. These tasks require agents to produce a series of actions over time based on observations of the environment, making it challenging to improve agent performance using traditional knowledge engineering methods.

Key Contributions

The main contributions of this work are:

* Developing a method for LLM agents to autonomously improve by learning from their own successful experiences without human intervention.
* Introducing a technique for constructing and refining a database of self-generated trajectories that serve as in-context examples for future tasks.
* Proposing two database construction enhancements: database-level curation and exemplar-level curation.

Methodology

The paper uses a ReAct-style agent architecture that employs recent best practices for in-context retrieval. The agent operates through a three-phase approach (planning, reasoning, and acting) and incorporates an initial planning step to generate a high-level plan for the entire task. The method also uses a database of self-generated trajectories to serve as in-context examples for future tasks.

Results

The results show that even naive accumulation of successful trajectories yields substantial performance gains across three diverse benchmarks: ALFWorld (73% to 89%), Wordcraft (55% to 64%), and InterCode-SQL (75% to 79%). The enhanced method achieves 93% success on ALFWorld, surpassing approaches that use more powerful LLMs and hand-crafted components.

Significance

This work matters because it demonstrates that LLM agents can autonomously improve through experience, offering a scalable alternative to labor-intensive knowledge engineering. The results highlight the practical value of trajectory bootstrapping as a dimension for scaling test-time compute.

---

[69] Causal Spatio-Temporal Prediction: An Effective and Efficient Multi-Modal Approach

Authors: Yuting Huang, Ziquan Fang, Zhihao Zeng, Lu Chen, Yunjun Gao
PDF: https://arxiv.org/pdf/2505.17637v2.pdf

Overview

Problem Statement

The paper addresses the problem of spatio-temporal prediction, which is crucial in various applications such as intelligent transportation, weather forecasting, and urban planning. Accurate predictions can improve decision-making and optimize resource allocation. However, current methods face challenges in integrating multi-modal data, mitigating confounding factors, and reducing computational complexity.

Key Contributions

The main contributions of this work are:

* **Unified Multi-Modal Spatio-Temporal Fusion**: The authors propose a framework that integrates various modalities (environmental images, event-related text, and spatio-temporal time-series data) through cross-modal attention and adaptive gating mechanisms.
* **Dual-Branch based Causal Disentanglement**: The authors introduce a dual-branch causal inference design that focuses on learning spatio-temporal patterns and models additional modalities to reduce confounding bias.
* **Efficient and Hybrid Model Design**: The authors incorporate GCN and Mamba for efficient spatio-temporal encoding, reducing computational complexity and accelerating model inference.

Methodology

The proposed framework, E2-CSTP, leverages cross-modal attention and gating mechanisms to integrate multi-modal data. The dual-branch design focuses on spatio-temporal patterns and models additional modalities to reduce confounding bias. The efficient and hybrid model design incorporates GCN and Mamba for accelerated spatio-temporal encoding.

Results

The authors conduct extensive experiments on 4 real-world datasets and achieve significant improvements:

* Up to 9.66% improvements in accuracy
* 17.37%?56.11% reductions in computational overhead

Significance

This work matters because it addresses the challenges in multi-modal spatio-temporal prediction and provides a unified framework for integrating heterogeneous data. The proposed framework can improve decision-making and optimize resource allocation in various applications. The efficient and hybrid model design can accelerate model inference, making it more practical for large-scale datasets.

---

[70] Causal Spatio-Temporal Prediction: An Effective and Efficient Multi-Modal Approach

Authors: Yuting Huang, Ziquan Fang, Zhihao Zeng, Lu Chen, Yunjun Gao
PDF: https://arxiv.org/pdf/2505.17637v2.pdf

Overview

Problem Statement

The paper addresses the problem of spatio-temporal prediction, which is crucial in various applications such as intelligent transportation, weather forecasting, and urban planning. The goal is to improve prediction accuracy by effectively integrating multi-modal data, mitigating confounding factors, and reducing computational complexity.

Key Contributions
The main contributions of this work are:

* **Unified Multi-Modal Spatio-Temporal Fusion**: The authors propose a framework that jointly models heterogeneous features from various modalities (environmental images, event-related text, and spatio-temporal time-series data) using cross-modal attention and adaptive gating mechanisms.
* **Dual-Branch based Causal Disentanglement**: The authors introduce a dual-branch causal inference design that separates the main branch for spatio-temporal pattern learning from the auxiliary branch for modeling additional modalities and reducing confounding bias.
* **Efficient and Hybrid Model Design**: The authors incorporate GCN and Mamba for efficient spatio-temporal encoding, reducing computational complexity and accelerating model inference.

Methodology
The proposed framework, E2-CSTP, leverages cross-modal attention and gating mechanisms to integrate multi-modal data. The dual-branch design separates the main branch for spatio-temporal pattern learning from the auxiliary branch for modeling additional modalities and reducing confounding bias. The efficient and hybrid model design incorporates GCN and Mamba for spatio-temporal encoding.

Results
The authors conduct extensive experiments on 4 real-world datasets, achieving up to 9.66% improvements in accuracy and 17.37%?56.11% reductions in computational overhead compared to 9 state-of-the-art methods.

Significance
This work matters because it addresses the challenges of multi-modal spatio-temporal prediction, improving prediction accuracy and reducing computational complexity. The proposed framework, E2-CSTP, has the potential to impact various applications, including intelligent transportation, weather forecasting, and urban planning.