1

Theia: Distilling Diverse Vision Foundation Models for Robot Learning
Jinghuan Shang,Karl Schmeckpeper,Brandon B. May,Maria Vittoria Minniti,Tarik Kelestemur,David Watkins,Laura Herlant
https://openreview.net/forum?id=wwfJ0zSpfM
Vision-based robot policy learning, which maps visual inputs to actions, necessitates a holistic understanding of diverse visual tasks beyond singletask needs like classification or segmentation. Inspired by this, we introduce Theia, a vision foundation model for robot learning that distills multiple off-the-shelf vision foundation models trained on varied vision tasks. Theia's rich visual representations encode diverse visual knowledge, enhancing downstream robot learning. Extensive experiments demonstrate that Theia outperforms its teacher models and prior robot learning models using less training data and smaller model sizes. Additionally, we quantify the quality of pre-trained visual representations and hypothesize that higher entropy in feature norm distributions leads to improved robot learning performance.

2

Body Transformer: Leveraging Robot Embodiment for Policy Learning
Carmelo Sferrazza,Dun-Ming Huang,Fangchen Liu,Jongmin Lee,Pieter Abbeel
https://openreview.net/forum?id=Oce2215aJE
In recent years, the transformer architecture has become the de-facto standard for machine learning algorithms applied to natural language processing and computer vision. Despite notable evidence of successful deployment of this architecture in the context of robot learning, we claim that vanilla transformers do not fully exploit the structure of the robot learning problem. We propose Body Transformer (BoT), an architecture that exploits the robot embodiment by providing an inductive bias that guides the learning process. We represent the robot body as a graph of sensors and actuators, and rely on masked attention to pool information through the architecture. The resulting architecture outperforms the vanilla transformer, as well as the classical multilayer perceptron, with respect to task completion, scaling properties, and computational efficiency when representing either imitation or reinforcement learning policies.

3

Gameplay Filters: Robust Zero-Shot Safety through Adversarial Imagination
Duy Phuong Nguyen,Kai-Chieh Hsu,Wenhao Yu,Jie Tan,Jaime Fernández Fisac
https://openreview.net/forum?id=Ke5xrnBFAR
Despite the impressive recent advances in learning-based robot control, ensuring robustness to out-of-distribution conditions remains an open challenge. Safety filters can, in principle, keep arbitrary control policies from incurring catastrophic failures by overriding unsafe actions, but existing solutions for complex (e.g., legged) robot dynamics do not span the full motion envelope and instead rely on local, reduced-order models. These filters tend to overly restrict agility and can still fail when perturbed away from nominal conditions. This paper presents the gameplay filter, a new class of predictive safety filter that continually plays out hypothetical matches between its simulation-trained safety strategy and a virtual adversary co-trained to invoke worst-case events and sim-to-real error, and precludes actions that would cause failures down the line. We demonstrate the scalability and robustness of the approach with a first-of-its-kind full-order safety filter for (36-D) quadrupedal dynamics. Physical experiments on two different quadruped platforms demonstrate the superior zero-shot effectiveness of the gameplay filter under large

perturbations such as tugging and unmodeled terrain. Experiment videos and open-source software are available online: https://saferobotics.org/research/gameplay-filter

4

Mobile ALOHA: Learning Bimanual Mobile Manipulation using Low-Cost Whole-Body Teleoperation
Zipeng Fu,Tony Z. Zhao,Chelsea Finn
https://openreview.net/forum?id=FO6tePGRZj
Imitation learning from human demonstrations has shown impressive performance in robotics. However, most results focus on table-top manipulation, lacking the mobility and dexterity necessary for generally useful tasks. In this work, we develop a system for imitating mobile manipulation tasks that are bimanual and require whole-body control. We first present Mobile ALOHA, a low-cost and whole-body teleoperation system for data collection. It augments the ALOHA system with a mobile base, and a whole-body teleoperation interface. Using data collected with Mobile ALOHA, we then perform supervised behavior cloning and find that co-training with existing static ALOHA datasets boosts performance on mobile manipulation tasks. With 50 demonstrations for each task, co-training can increase success rates by up to 90%, allowing Mobile ALOHA to autonomously complete complex mobile manipulation tasks such as sauteing and serving a piece of shrimp, opening a two-door wall cabinet to store heavy cooking pots, calling and entering an elevator, and lightly rinsing a used pan using a kitchen faucet. We will open-source all the hardware and software implementations upon publication.

5

RP1M: A Large-Scale Motion Dataset for Piano Playing with Bi-Manual Dexterous Robot Hands
Yi Zhao,Le Chen,Jan Schneider,Quankai Gao,Juho Kannala,Bernhard Schölkopf,Joni Pajarinen,Dieter Büchler
https://openreview.net/forum?id=4Of4UWyBXE
Endowing robot hands with human-level dexterity is a long-lasting research objective. Bi-manual robot piano playing constitutes a task that combines challenges from dynamic tasks, such as generating fast while precise motions, with slower but contact-rich manipulation problems. Although reinforcement learning based approaches have shown promising results in single-task performance, these methods struggle in a multi-song setting. Our work aims to close this gap and, thereby, enable imitation learning approaches for robot piano playing at scale. To this end, we introduce the Robot Piano 1 Million (RP1M) dataset, containing bi-manual robot piano playing motion data of more than one million trajectories. We formulate finger placements as an optimal transport problem, thus, enabling automatic annotation of vast amounts of unlabeled songs. Benchmarking existing imitation learning approaches shows that such approaches reach state-of-the-art robot piano playing performance by leveraging RP1M.

6

Learning to Walk from Three Minutes of Real-World Data with Semi-structured Dynamics Models
Jacob Levy,Tyler Westenbroek,David Fridovich-Keil
https://openreview.net/forum?id=evCXwlCMli
Traditionally, model-based reinforcement learning (MBRL) methods exploit neural networks as flexible function approximators to represent $a\ priori$ unknown environment dynamics. However, training data are typically scarce in practice, and these black-box models often fail to generalize. Modeling architectures that leverage known physics can substantially reduce the complexity of system-identification, but break down in the face of complex phenomena such as contact. We

introduce a novel framework for learning semi-structured dynamics models for contact-rich systems which seamlessly integrates structured first principles modeling techniques with black-box auto-regressive models. Specifically, we develop an ensemble of probabilistic models to estimate external forces, conditioned on historical observations and actions, and integrate these predictions using known Lagrangian dynamics. With this semi-structured approach, we can make accurate long-horizon predictions with substantially less data than prior methods. We leverage this capability and propose Semi-Structured Reinforcement Learning (SSRL) a simple model-based learning framework which pushes the sample complexity boundary for real-world learning. We validate our approach on a real-world Unitree Go1 quadruped robot, learning dynamic gaits -- from scratch -- on both hard and soft surfaces with just a few minutes of real-world data. Video and code are available at: https://sites.google.com/utexas.edu/ssrl

7

Towards Open-World Grasping with Large Vision-Language Models
Georgios Tziafas,Hamidreza Kasaei
https://openreview.net/forum?id=QUzwHYJ9Hf
The ability to grasp objects in-the-wild from open-ended language instructions constitutes a fundamental challenge in robotics. An open-world grasping system should be able to combine high-level contextual with low-level physical-geometric reasoning in order to be applicable in arbitrary scenarios. Recent works exploit the web-scale knowledge inherent in large language models (LLMs) to plan and reason in robotic context, but rely on external vision and action models to ground such knowledge into the environment and parameterize actuation. This setup suffers from two major bottlenecks: a) the LLM's reasoning capacity is constrained by the quality of visual grounding, and b) LLMs do not contain low-level spatial understanding of the world, which is essential for grasping in contact-rich scenarios. In this work we demonstrate that modern vision-language models (VLMs) are capable of tackling such limitations, as they are implicitly grounded and can jointly reason about semantics and geometry. We propose \texttt{OWG}, an open-world grasping pipeline that combines VLMs with segmentation and grasp synthesis models to unlock grounded world understanding in three stages: open-ended referring segmentation, grounded grasp planning and grasp ranking via contact reasoning, all of which can be applied zero-shot via suitable visual prompting mechanisms. We conduct extensive evaluation in cluttered indoor scene datasets to showcase \texttt{OWG}'s robustness in grounding from open-ended language, as well as open-world robotic grasping experiments in both simulation and hardware that demonstrate superior performance compared to previous supervised and zero-shot LLM-based methods.

8

An Open-Source Soft Robotic Platform for Autonomous Aerial Manipulation in the Wild
Erik Bauer,Marc Blöchlinger,Pascal Strauch,Arman Raayatsanati,Cavelti Curdin,Robert K. Katzschmann
https://openreview.net/forum?id=SfaB20rjVo
Aerial manipulation combines the versatility and speed of flying platforms with the functional capabilities of mobile manipulation, which presents significant challenges due to the need for precise localization and control. Traditionally, researchers have relied on off-board perception systems, which are limited to expensive and impractical specially equipped indoor environments. In this work, we introduce a novel platform for autonomous aerial manipulation that exclusively utilizes onboard perception systems. Our platform can perform aerial manipulation in various indoor and outdoor environments without depending on external perception systems. Our experimental results demonstrate the platform's ability to autonomously grasp various objects in

diverse settings. This advancement significantly improves the scalability and practicality of aerial manipulation applications by eliminating the need for costly tracking solutions. To accelerate future research, we open source our modern ROS 2 software stack and custom hardware design, making our contributions accessible to the broader research community.

9

Safe Bayesian Optimization for the Control of High-Dimensional Embodied Systems
Yunyue Wei,Zeji Yi,Hongda Li,Saraswati Soedarmadji,Yanan Sui
https://openreview.net/forum?id=8PcRynpd1m
Learning to move is a primary goal for animals and robots, where ensuring safety is often important when optimizing control policies on the embodied systems. For complex tasks such as the control of human or humanoid control, the high-dimensional parameter space adds complexity to the safe optimization effort. Current safe exploration algorithms exhibit inefficiency and may even become infeasible with large high-dimensional input spaces. Furthermore, existing high-dimensional constrained optimization methods neglect safety in the search process. In this paper, we propose High-dimensional Safe Bayesian Optimization with local optimistic exploration (HdSafeBO), a novel approach designed to handle high-dimensional sampling problems under probabilistic safety constraints. We introduce a local optimistic strategy to efficiently and safely optimize the objective function, providing a probabilistic safety guarantee and a cumulative safety violation bound. Through the use of isometric embedding, HdSafeBO addresses problems ranging from a few hundred to several thousand dimensions while maintaining safety guarantees. To our knowledge, HdSafeBO is the first algorithm capable of optimizing the control of high-dimensional musculoskeletal systems with high safety probability. We also demonstrate the real-world applicability of HdSafeBO through its use in the safe online optimization of neural stimulation-induced human motion control.

10

LeLaN: Learning A Language-Conditioned Navigation Policy from In-the-Wild Video
Noriaki Hirose,Catherine Glossop,Ajay Sridhar,Oier Mees,Sergey Levine
https://openreview.net/forum?id=zIWu9Kmlqk
We present our method, LeLaN, which uses action-free egocentric data to learn robust language-conditioned object navigation. By leveraging the knowledge of large vision and language models and grounding this knowledge using pre-trained segmentation and depth estimation models, we can label in-the-wild data from a variety of indoor and outdoor environments with diverse instructions that capture a range of objects with varied granularity and noise in their descriptions. Leveraging this method to label over 50 hours of data collected in indoor and outdoor environments, including robot observations, YouTube video tours, and human-collected walking data allows us to train a policy that can outperform state-of-the-art methods on the zero-shot object navigation task in both success rate and precision.

11

Trajectory Improvement and Reward Learning from Comparative Language Feedback
Zhaojing Yang,Miru Jun,Jeremy Tien,Stuart Russell,Anca Dragan,Erdem Biyik
https://openreview.net/forum?id=1tCteNSbFH
Learning from human feedback has gained traction in fields like robotics and natural language processing in recent years. While prior works mostly rely on human feedback in the form of comparisons, language is a preferable modality that provides more informative insights into user

preferences. In this work, we aim to incorporate comparative language feedback to iteratively improve robot trajectories and to learn reward functions that encode human preferences. To achieve this goal, we learn a shared latent space that integrates trajectory data and language feedback, and subsequently leverage the learned latent space to improve trajectories and learn human preferences. To the best of our knowledge, we are the first to incorporate comparative language feedback into reward learning. Our simulation experiments demonstrate the effectiveness of the learned latent space and the success of our learning algorithms. We also conduct human subject studies that show our reward learning algorithm achieves a 23.9% higher subjective score on average and is 11.3% more time-efficient compared to preference-based reward learning, underscoring the superior performance of our method. Our website is at https://liralab.usc.edu/comparative-language-feedback/.

12

Policy Adaptation via Language Optimization: Decomposing Tasks for Few-Shot Imitation
Vivek Myers,Chunyuan Zheng,Oier Mees,Kuan Fang,Sergey Levine
https://openreview.net/forum?id=qUSa3F79am
Learned language-conditioned robot policies often struggle to effectively adapt to new real-world tasks even when pre-trained across a diverse set of instructions. We propose a novel approach for few-shot adaptation to unseen tasks that exploits the semantic understanding of task decomposition provided by vision-language models (VLMs). Our method, Policy Adaptation via Language Optimization (PALO), combines a handful of demonstrations of a task with proposed language decompositions sampled from a VLM to quickly enable rapid nonparametric adaptation, avoiding the need for a larger fine-tuning dataset. We evaluate PALO on extensive real-world experiments consisting of challenging unseen, long-horizon robot manipulation tasks. We find that PALO is able of consistently complete long-horizon, multi-tier tasks in the real world, outperforming state of the art pre-trained generalist policies, and methods that have access to the same demonstrations.

13

Learning Transparent Reward Models via Unsupervised Feature Selection
Daulet Baimukashev,Gokhan Alcan,Kevin Sebastian Luck,Ville Kyrki
https://openreview.net/forum?id=2sg4PY1W9d
In complex real-world tasks such as robotic manipulation and autonomous driving, collecting expert demonstrations is often more straightforward than specifying precise learning objectives and task descriptions. Learning from expert data can be achieved through behavioral cloning or by learning a reward function, i.e., inverse reinforcement learning. The latter allows for training with additional data outside the training distribution, guided by the inferred reward function. We propose a novel approach to construct compact and interpretable reward models from automatically selected state features. These inferred rewards have an explicit form and enable the learning of policies that closely match expert behavior by training standard reinforcement learning algorithms from scratch. We validate our method's performance in various robotic environments with continuous and high-dimensional state spaces.

14

VLM-Grounder: A VLM Agent for Zero-Shot 3D Visual Grounding
Runsen Xu,Zhiwei Huang,Tai Wang,Yilun Chen,Jiangmiao Pang,Dahua Lin
https://openreview.net/forum?id=IcOrwIXzMi
3D visual grounding is crucial for robots, requiring integration of natural language and 3D scene understanding. Traditional methods depend on supervised learning with 3D point clouds are limited by scarce datasets. Recently zero-shot methods leveraging LLMs have been proposed to address the data issue. While effective, these methods often miss detailed scene context, limiting their ability to handle complex queries. In this work, we present VLM-Grounder, a novel framework using vision-language models (VLMs) for zero-shot 3D visual grounding based solely on 2D images. VLM-Grounder dynamically stitches image sequences, employs a grounding and feedback scheme to find the target object, and uses a multi-view ensemble projection to accurately estimate 3D bounding boxes. Experiments on ScanRefer and Nr3D datasets show VLM-Grounder outperforms previous zero-shot methods, achieving 51.6% Acc@0.25 on ScanRefer and 48.0% Acc on Nr3D, without relying on 3D geometry or object priors.

15

MaIL: Improving Imitation Learning with Selective State Space Models
Xiaogang Jia,Qian Wang,Atalay Donat,Bowen Xing,Ge Li,Hongyi Zhou,Onur Celik,Denis Blessing,Rudolf Lioutikov,Gerhard Neumann
https://openreview.net/forum?id=IssXUYvVTg
This work introduces Mamba Imitation Learning (MaIL), a novel imitation learning (IL) architecture that offers a computationally efficient alternative to state-of-the-art (SoTA) Transformer policies. Transformer-based policies have achieved remarkable results due to their ability in handling human-recorded data with inherently non-Markovian behavior. However, their high performance comes with the drawback of large models that complicate effective training. While state space models (SSMs) have been known for their efficiency, they were not able to match the performance of Transformers. Mamba significantly improves the performance of SSMs and rivals against Transformers, positioning it as an appealing alternative for IL policies. MaIL leverages Mamba as a backbone and introduces a formalism that allows using Mamba in the encoder-decoder structure. This formalism makes it a versatile architecture that can be used as a standalone policy or as part of a more advanced architecture, such as a diffuser in the diffusion process. Extensive evaluations on the LIBERO IL benchmark and three real robot experiments show that MaIL: i) outperforms Transformers in all LIBERO tasks, ii) achieves good performance even with small datasets, iii) is able to effectively process multi-modal sensory inputs, iv) is more robust to input noise compared to Transformers.

16

Robotic Control via Embodied Chain-of-Thought Reasoning
Michał Zawalski,William Chen,Karl Pertsch,Oier Mees,Chelsea Finn,Sergey Levine
https://openreview.net/forum?id=S70MgnIA0v
A key limitation of learned robot control policies is their inability to generalize outside their training data. Recent works on vision-language-action models (VLAs) have shown that the use of large, internet pre-trained vision-language models as the backbone of learned robot policies can substantially improve their robustness and generalization ability. Yet, one of the most exciting capabilities of large vision-language models in other domains is their ability to reason iteratively through complex problems. Can that same capability be brought into robotics to allow policies to

improve performance by reasoning about a given task before acting? Naive use of "chain-of-thought" (CoT) style prompting is significantly less effective with standard VLAs because of the relatively simple training examples that are available to them. Additionally, purely semantic reasoning about sub-tasks, as is common in regular CoT, is insufficient for robot policies that need to ground their reasoning in sensory observations and the robot state. To this end, we introduce Embodied Chain-of-Thought Reasoning (ECoT) for VLAs, in which we train VLAs to perform multiple steps of reasoning about plans, sub-tasks, motions, and visually grounded features like object bounding boxes and end effector positions, before predicting the robot action. We design a scalable pipeline for generating synthetic training data for ECoT on large robot datasets. We demonstrate, that ECoT increases the absolute success rate of OpenVLA, the current strongest open-source VLA policy, by 28% across challenging generalization tasks, without any additional robot training data. Additionally, ECoT makes it easier for humans to interpret a policy's failures and correct its behavior using natural language.

17

Bootstrapping Reinforcement Learning with Imitation for Vision-Based Agile Flight
Jiaxu Xing,Angel Romero,Leonard Bauersfeld,Davide Scaramuzza
https://openreview.net/forum?id=bt0PX0e4rE
Learning visuomotor policies for agile quadrotor flight presents significant difficulties, primarily from inefficient policy exploration caused by high-dimensional visual inputs and the need for precise and low-latency control. To address these challenges, we propose a novel approach that combines the performance of Reinforcement Learning (RL) and the sample efficiency of Imitation Learning (IL) in the task of vision-based autonomous drone racing. While RL provides a framework for learning high-performance controllers through trial and error, it faces challenges with sample efficiency and computational demands due to the high dimensionality of visual inputs. Conversely, IL efficiently learns from visual expert demonstrations, but it remains limited by the expert's performance and state distribution. To overcome these limitations, our policy learning framework integrates the strengths of both approaches. Our framework contains three phases: training a teacher policy using RL with privileged state information, distilling it into a student policy via IL, and adaptive fine-tuning via RL. Testing in both simulated and real-world scenarios shows our approach can not only learn in scenarios where RL from scratch fails but also outperforms existing IL methods in both robustness and performance, successfully navigating a quadrotor through a race course using only visual information.

18

Autonomous Improvement of Instruction Following Skills via Foundation Models
Zhiyuan Zhou,Pranav Atreya,Abraham Lee,Homer Rich Walke,Oier Mees,Sergey Levine
https://openreview.net/forum?id=8Ar8b00GJC
Intelligent robots capable of improving from autonomously collected experience have the potential to transform robot learning: instead of collecting costly teleoperated demonstration data, large-scale deployment of fleets of robots can quickly collect larger quantities of autonomous data useful for training better robot policies. However, autonomous improvement requires solving two key problems: (i) fully automating a scalable data collection procedure that can collect diverse and semantically meaningful robot data and (ii) learning from non-optimal, autonomous data with no human annotations. To this end, we propose a novel approach that addresses these challenges, allowing instruction following policies to improve from autonomously collected data without human supervision. Our framework leverages vision-language models to collect and evaluate semantically meaningful experiences in new environments, and then utilizes a decomposition of

instruction following tasks into (semantic) language-conditioned image generation and (non-semantic) goal reaching, which makes it significantly more practical to improve from this autonomously collected data without any human annotations. We carry out extensive experiments in the real world to demonstrate the effectiveness of our approach, and find that in a suite of unseen environments, the robot policy can be improved significantly with autonomously collected data. We open-source the code for our semantic autonomous improvement pipeline, as well as our autonomous dataset of 25K trajectories collected across five tabletop environments: https://soar-autonomous-improvement.github.io

19

Learning Robotic Manipulation Policies from Point Clouds with Conditional Flow Matching
Eugenio Chisari,Nick Heppert,Max Argus,Tim Welschehold,Thomas Brox,Abhinav Valada
https://openreview.net/forum?id=vtEn8NJWlz
Learning from expert demonstrations is a popular approach to train robotic manipulation policies from limited data. However, imitation learning algorithms require a number of design choices ranging from the input modality, training objective, and 6-DoF end-effector pose representation. Diffusion-based methods have gained popularity as they allow to predict long horizon trajectories and handle multimodal action distributions. Recently, Conditional Flow Matching (CFM) (or Rectified Flow) has been proposed as a more flexible generalization of diffusion models. In this paper we investigate the application of CFM in the context of robotic policy learning, and specifically study the interplay with the other design choices required to build an imitation learning algorithm. We show that CFM gives the best performance when combined with point cloud input observations. Additionally, we study the feasibility of a CFM formulation on the SO(3) manifold and evaluate its suitability with a simplified example. We perform extensive experiments on RLBench which demonstrate that our proposed PointFlowMatch approach achieves a state-of-the-art average success rate of 67.8% over eight tasks, double the performance of the next best method.

20

Context-Aware Replanning with Pre-Explored Semantic Map for Object Navigation
Po-Chen Ko,Hung-Ting Su,CY Chen,Jia-Fong Yeh,Min Sun,Winston H. Hsu
https://openreview.net/forum?id=Dftu4r5jHe
Pre-explored Semantic Map, constructed through prior exploration using visual language models (VLMs), has proven effective as a foundational element for training-free robotic applications. However, existing approaches assume the map's accuracy and do not provide effective mechanisms for revising decisions based on incorrect maps. This work introduces Context-Aware Replanning (CARe),, which estimates map uncertainty through confidence scores and multi-view consistency, enabling the agent to revise erroneous decisions stemming from inaccurate maps without additional labels. We demonstrate the effectiveness of our proposed method using two modern map backbones, VLMaps and OpenMask3D, and show significant improvements in performance on object navigation tasks.

21

MBC: Multi-Brain Collaborative Control for Quadruped Robots
Hang Liu,Yi Cheng,Rankun Li,Xiaowen Hu,Linqi Ye,Houde Liu
https://openreview.net/forum?id=Lixj7WEGEy
In the field of locomotion task of quadruped robots, Blind Policy and Perceptive Policy each have their own advantages and limitations. The Blind Policy relies on preset sensor information and

algorithms, suitable for known and structured environments, but it lacks adaptability in complex or unknown environments. The Perceptive Policy uses visual sensors to obtain detailed environmental information, allowing it to adapt to complex terrains, but its effectiveness is limited under occluded conditions, especially when perception fails. Unlike the Blind Policy, the Perceptive Policy is not as robust under these conditions. To address these challenges, we propose a MBC:Multi-Brain collaborative system that incorporates the concepts of Multi-Agent Reinforcement Learning and introduces collaboration between the Blind Policy and the Perceptive Policy. By applying this multi-policy collaborative model to a quadruped robot, the robot can maintain stable locomotion even when the perceptual system is impaired or observational data is incomplete. Our simulations and real-world experiments demonstrate that this system significantly improves the robot's passability and robustness against perception failures in complex environments, validating the effectiveness of multi-policy collaboration in enhancing robotic motion performance.

22

Guided Reinforcement Learning for Robust Multi-Contact Loco-Manipulation
Jean Pierre Sleiman,Mayank Mittal,Marco Hutter
https://openreview.net/forum?id=9aZ4ehSTRc
Reinforcement learning (RL) has shown remarkable proficiency in developing robust control policies for contact-rich applications. However, it typically requires meticulous Markov Decision Process (MDP) designing tailored to each task and robotic platform. This work addresses this challenge by creating a systematic approach to behavior synthesis and control for multi-contact loco-manipulation. We define a task-independent MDP formulation to learn robust RL policies using a single demonstration (per task) generated from a fast model-based trajectory optimization method. Our framework is validated on diverse real-world tasks, such as navigating spring-loaded doors and manipulating heavy dishwashers. The learned behaviors can handle dynamic uncertainties and external disturbances, showcasing recovery maneuvers, such as re-grasping objects during execution. Finally, we successfully transfer the policies to a real robot, demonstrating the approach's practical viability.

23

Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation
Ria Doshi,Homer Rich Walke,Oier Mees,Sudeep Dasari,Sergey Levine
https://openreview.net/forum?id=AuJnXGq3AL
Modern machine learning systems rely on large datasets to attain broad generalization, and this often poses a challenge in robotic learning, where each robotic platform and task might have only a small dataset. By training a single policy across many different kinds of robots, a robotic learning method can leverage much broader and more diverse datasets, which in turn can lead to better generalization and robustness. However, training a single policy on multi-robot data is challenging because robots can have widely varying sensors, actuators, and control frequencies. We propose CrossFormer, a scalable and flexible transformer-based policy that can consume data from any embodiment. We train CrossFormer on the largest and most diverse dataset to date, 900K trajectories across 20 different robot embodiments. We demonstrate that the same network weights can control vastly different robots, including single and dual arm manipulation systems, wheeled robots, quadcopters, and quadrupeds. Unlike prior work, our model does not require manual alignment of the observation or action spaces. Extensive experiments in the real world show that our method matches the performance of specialist policies tailored for each

embodiment, while also significantly outperforming the prior state of the art in cross-embodiment learning.

24

OpenVLA: An Open-Source Vision-Language-Action Model
Moo Jin Kim,Karl Pertsch,Siddharth Karamcheti,Ted Xiao,Ashwin Balakrishna,Suraj Nair,Rafael Rafailov,Ethan P Foster,Pannag R Sanketi,Quan Vuong,Thomas Kollar,Benjamin Burchfiel,Russ Tedrake,Dorsa Sadigh,Sergey Levine,Percy Liang,Chelsea Finn
https://openreview.net/forum?id=ZMnD6QZAE6
Large policies pretrained on a combination of Internet-scale vision-language data and diverse robot demonstrations have the potential to change how we teach robots new skills: rather than training new behaviors from scratch, we can fine-tune such vision-language-action (VLA) models to obtain robust, generalizable policies for visuomotor control. Yet, widespread adoption of VLAs for robotics has been challenging as 1) existing VLAs are largely closed and inaccessible to the public, and 2) prior work fails to explore methods for efficiently fine-tuning VLAs for new tasks, a key component for adoption. Addressing these challenges, we introduce OpenVLA, a 7B-parameter open-source VLA trained on a diverse collection of 970k real-world robot demonstrations. OpenVLA builds on a Llama 2 language model combined with a visual encoder that fuses pretrained features from DINOv2 and SigLIP. As a product of the added data diversity and new model components, OpenVLA demonstrates strong results for generalist manipulation, outperforming closed models such as RT-2-X (55B) by 16.5% in absolute task success rate across 29 tasks and multiple robot embodiments, with 7x fewer parameters. We further show that we can effectively fine-tune OpenVLA for new settings, with especially strong generalization results in multi-task environments involving multiple objects and strong language grounding abilities, where we outperform expressive from-scratch imitation learning methods such as Diffusion Policy by 20.4%. We also explore compute efficiency; as a separate contribution, we show that OpenVLA can be fine-tuned on consumer GPUs via modern low-rank adaptation methods and served efficiently via quantization without a hit to downstream success rate. Finally, we release model checkpoints, fine-tuning notebooks, and our PyTorch codebase with built-in support for training VLAs at scale on Open X-Embodiment datasets.

25

Steering Your Generalists: Improving Robotic Foundation Models via Value Guidance
Mitsuhiko Nakamoto,Oier Mees,Aviral Kumar,Sergey Levine
https://openreview.net/forum?id=6FGlpzC9Po
Large, general-purpose robotic policies trained on diverse demonstration datasets have been shown to be remarkably effective both for controlling a variety of robots in a range of different scenes, and for acquiring broad repertoires of manipulation skills. However, the data that such policies are trained on is generally of mixed quality -- not only are human-collected demonstrations unlikely to perform the task perfectly, but the larger the dataset is, the harder it is to curate only the highest quality examples. It also remains unclear how optimal data from one embodiment is for training on another embodiment. In this paper, we present a general and broadly applicable approach that enhances the performance of such generalist robot policies at deployment time by re-ranking their actions according to a value function learned via offline RL. This approach, which we call Value-Guided Policy Steering (V-GPS), is compatible with a wide range of different generalist policies, without needing to fine-tune or even access the weights of the policy. We show that the same value function can improve the performance of five different state-of-the-art policies with different architectures, even though they were trained on distinct

datasets, attaining consistent performance improvement on multiple robotic platforms across a total of 12 tasks. Code and videos can be found at: https://nakamotoo.github.io/V-GPS

26

ViPER: Visibility-based Pursuit-Evasion via Reinforcement Learning
Yizhuo Wang,Yuhong Cao,Jimmy Chiun,Subhadeep Koley,Mandy Pham,Guillaume Adrien Sartoretti
https://openreview.net/forum?id=EPujQZWemk
In visibility-based pursuit-evasion tasks, a team of mobile pursuer robots with limited sensing capabilities is tasked with detecting all evaders in a multiply-connected planar environment, whose map may or may not be known to pursuers beforehand. This requires tight coordination among multiple agents to ensure that the omniscient and potentially arbitrarily fast evaders are guaranteed to be detected by the pursuers. Whereas existing methods typically rely on a relatively large team of agents to clear the environment, we propose ViPER, a neural solution that leverages a graph attention network to learn a coordinated yet distributed policy via multi-agent reinforcement learning (MARL). We experimentally demonstrate that ViPER significantly outperforms other state-of-the-art non-learning planners, showcasing its emergent coordinated behaviors and adaptability to more challenging scenarios and various team sizes, and finally deploy its learned policies on hardware in an aerial search task.

27

Adapting Humanoid Locomotion over Challenging Terrain via Two-Phase Training
Wenhao Cui,Shengtao Li,Huaxing Huang,Bangyu Qin,Tianchu Zhang,hanjinchao,Liang Zheng,Ziyang Tang,Chenxu Hu,NING Yan,Jiahao Chen,Zheyuan Jiang
https://openreview.net/forum?id=O0oK2bVist
Humanoid robots are a key focus in robotics, with their capacity to navigate tough terrains being essential for many uses. While strides have been made, creating adaptable locomotion for complex environments is still tough. Recent progress in learning-based systems offers hope for robust legged locomotion, but challenges persist, such as tracking accuracy at high speeds and on uneven ground, and joint oscillations in actual robots. This paper proposes a novel training framework to address these challenges by employing a two-phase training paradigm with reinforcement learning. The proposed framework is further enhanced through the integration of command curriculum learning, refining the precision and adaptability of our approach. Additionally, we adapt DreamWaQ to our humanoid locomotion system and improve it to mitigate joint oscillations. Finally, we achieve the sim-to-real transfer of our method. A series of empirical results demonstrate the superior performance of our proposed method compared to state-of-the-art methods.

28

Lifelong Autonomous Improvement of Navigation Foundation Models in the Wild
Kyle Stachowicz,Lydia Ignatova,Sergey Levine
https://openreview.net/forum?id=vBj5oC60Lk
Recent works have proposed a number of general-purpose robotic foundation models that can control a variety of robotic platforms to perform a range of different tasks, including in the domains of navigation and manipulation. However, such models are typically trained via imitation learning, which precludes the ability to improve autonomously through experience that the robot gathers on the job. In this work, our aim is to train general-purpose robotic foundation models in the domain of robotic navigation specifically with the aim of enabling autonomous self-

improvement. We show that a combination of pretraining with offline reinforcement learning and a complete system for continual autonomous operation leads to a robotic learning framework that not only starts off with broad and diverse capabilities, but can further improve and adapt those capabilities in the course of carrying out navigational tasks in a given deployment location. To our knowledge, our model LiReN is the first navigation robot foundation model that is capable of fine-tuning with autonomous online data in open-world settings.

29

Action Space Design in Reinforcement Learning for Robot Motor Skills
Julian Eßer,Gabriel B. Margolis,Oliver Urbann,Sören Kerner,Pulkit Agrawal
https://openreview.net/forum?id=GGuNkjQSrk
Practitioners often rely on intuition to select action spaces for learning. The choice can substantially impact final performance even when choosing among configuration-space representations such as joint position, velocity, and torque commands. We examine action space selection considering a wheeled-legged robot, a quadruped robot, and a simulated suite of locomotion, manipulation, and control tasks. We analyze the mechanisms by which action space can improve performance and conclude that the action space can influence learning performance substantially in a task-dependent way. Moreover, we find that much of the practical impact of action space selection on learning dynamics can be explained by improved policy initialization and behavior between timesteps.

30

DexGraspNet 2.0: Learning Generative Dexterous Grasping in Large-scale Synthetic Cluttered Scenes
Jialiang Zhang,Haoran Liu,Danshi Li,XinQiang Yu,Haoran Geng,Yufei Ding,Jiayi Chen,He Wang
https://openreview.net/forum?id=5W0iZR9J7h
Grasping in cluttered scenes remains highly challenging for dexterous hands due to the scarcity of data. To address this problem, we present a large-scale synthetic dataset, encompassing 1319 objects, 8270 scenes, and 426 million grasps. Beyond benchmarking, we also explore data-efficient learning strategies from grasping data. We reveal that the combination of a conditional generative model that focuses on local geometry and a grasp dataset that emphasizes complex scene variations is key to achieving effective generalization. Our proposed generative method outperforms all baselines in simulation experiments. Furthermore, it demonstrates zero-shot sim-to-real transfer through test-time depth restoration, attaining 91% real-world success rate, showcasing the robust potential of utilizing fully synthetic training data.

31

OKAMI: Teaching Humanoid Robots Manipulation Skills through Single Video Imitation
Jinhan Li,Yifeng Zhu,Yuqi Xie,Zhenyu Jiang,Mingyo Seo,Georgios Pavlakos,Yuke Zhu
https://openreview.net/forum?id=URj5TQTAXM
We study the problem of teaching humanoid robots manipulation skills by imitating from single video demonstrations. We introduce OKAMI, a method that generates a manipulation plan from a single RGB-D video and derives a policy for execution. At the heart of our approach is object-aware retargeting, which enables the humanoid robot to mimic the human motions in an RGB-D video while adjusting to different object locations during deployment. OKAMI uses open-world vision models to identify task-relevant objects and retarget the body motions and hand poses separately. Our experiments show that OKAMI achieves strong generalizations across varying

visual and spatial conditions, outperforming the state-of-the-art baseline on open-world imitation from observation. Furthermore, OKAMI rollout trajectories are leveraged to train closed-loop visuomotor policies, which achieve an average success rate of 79.2 without the need for labor-intensive teleoperation. More videos can be found on our website https://ut-austin-rpl.github.io/OKAMI/.

32

So You Think You Can Scale Up Autonomous Robot Data Collection?
Suvir Mirchandani,Suneel Belkhale,Joey Hejna,Evelyn Choi,Md Sazzad Islam,Dorsa Sadigh
https://openreview.net/forum?id=XrxLGzF0lJ
A long-standing goal in robot learning is to develop methods for robots to acquire new skills autonomously. While reinforcement learning (RL) comes with the promise of enabling autonomous data collection, it remains challenging to scale in the real-world partly due to the significant effort required for environment design and instrumentation, including the need for designing reset functions or accurate success detectors. On the other hand, imitation learning (IL) methods require little to no environment design effort, but instead require significant human supervision in the form of collected demonstrations. To address these shortcomings, recent works in autonomous IL start with an initial seed dataset of human demonstrations that an autonomous policy can bootstrap from. While autonomous IL approaches come with the promise of addressing the challenges of autonomous RL—environment design challenges—as well as the challenges of pure IL strategies—extensive human supervision—in this work, we posit that such techniques do not deliver on this promise and are still unable to scale up autonomous data collection in the real world. Through a series of targeted real-world experiments, we demonstrate that these approaches, when scaled up to realistic settings, face much of the same scaling challenges as prior attempts in RL in terms of environment design. Further, we perform a rigorous study of various autonomous IL methods across different data scales and 7 simulation and real-world tasks, and demonstrate that while autonomous data collection can modestly improve performance (on the order of 10%), simply collecting more human data often provides significantly more improvement. Our work suggests a negative result: that scaling up autonomous data collection for learning robot policies for real-world tasks is more challenging and impractical than what is suggested in prior work. We hope these insights about the core challenges of scaling up data collection help inform future efforts in autonomous learning.

33

D
3
RoMa: Disparity Diffusion-based Depth Sensing for Material-Agnostic Robotic Manipulation
Songlin Wei,Haoran Geng,Jiayi Chen,Congyue Deng,Cui Wenbo,Chengyang Zhao,Xiaomeng Fang,Leonidas Guibas,He Wang
https://openreview.net/forum?id=7E3JAys1xO
Depth sensing is an important problem for 3D vision-based robotics. Yet, a real-world active stereo or ToF depth camera often produces noisy and incomplete depth which bottlenecks robot performances. In this work, we propose D3RoMa, a learning-based depth estimation framework on stereo image pairs that predicts clean and accurate depth in diverse indoor scenes, even in the most challenging scenarios with translucent or specular surfaces where classical depth sensing completely fails. Key to our method is that we unify depth estimation and restoration into an image-to-image translation problem by predicting the disparity map with a denoising diffusion probabilistic model. At inference time, we further incorporated a left-right consistency constraint

as classifier guidance to the diffusion process. Our framework combines recently advanced learning-based approaches and geometric constraints from traditional stereo vision. For model training, we create a large scene-level synthetic dataset with diverse transparent and specular objects to compensate for existing tabletop datasets. The trained model can be directly applied to real-world in-the-wild scenes and achieve state-of-the-art performance in multiple public depth estimation benchmarks. Further experiments in both simulated and real environments show that accurate depth prediction significantly improves robotic manipulation in various scenarios.

34

Bimanual Dexterity for Complex Tasks
Kenneth Shaw,Yulong Li,Jiahui Yang,Mohan Kumar Srirama,Ray Liu,Haoyu Xiong,Russell Mendonca,Deepak Pathak
https://openreview.net/forum?id=55tYfHvanf
To train generalist robot policies, machine learning methods often require a substantial amount of expert human teleoperation data. An ideal robot for humans collecting data is one that closely mimics them: bimanual arms and dexterous hands. However, creating such a bimanual teleoperation system with over 50 DoF is a significant challenge. To address this, we introduce Bidex, an extremely dexterous, low-cost, low-latency and portable bimanual dexterous teleoperation system which relies on motion capture gloves and teacher arms. We compare Bidex to a Vision Pro teleoperation system and a SteamVR system and find Bidex to produce better quality data for more complex tasks at a faster rate. Additionally, we show Bidex operating a mobile bimanual robot for in the wild tasks. Please refer to https://bidex-teleop.github.io for video results and instructions to recreate Bidex. The robot hands (5k USD) and teleoperation system (7k USD) is readily reproducible and can be used on many robot arms including two xArms ($16k USD).

35

Learning Robotic Locomotion Affordances and Photorealistic Simulators from Human-Captured Data
Alejandro Escontrela,Justin Kerr,Kyle Stachowicz,Pieter Abbeel
https://openreview.net/forum?id=1TEZ1hiY5m
Learning reliable affordance models which satisfy human preferences is often hindered by a lack of high-quality training data. Similarly, learning visuomotor policies in simulation can be challenging due to the high cost of photo-realistic rendering. We present PAWS: a comprehensive robot learning framework that uses a novel portable data capture rig and processing pipeline to collect long-horizon trajectories that include camera poses, foot poses, terrain meshes, and 3D radiance fields. We also contribute PAWS-Data: an extensive dataset gathered with PAWS containing over 10 hours of indoor and outdoor trajectories spanning a variety of scenes. With PAWS-Data we leverage radiance fields' photo-realistic rendering to generate tens of thousands of viewpoint-augmented images, then produce pixel affordance labels by identifying semantically similar regions to those traversed by the user. On this data we finetune a navigation affordance model from a pretrained backbone, and perform detailed ablations. Additionally, We open source PAWS-Sim, a high-speed photo-realistic simulator which integrates PAWS-Data with IsaacSim, enabling research for visuomotor policy learning. We evaluate the utility of the affordance model on a quadrupedal robot, which plans through affordances to follow pathways and sidewalks, and avoid human collisions. Project resources are available on the website.

36

LiDARGrid: Self-supervised 3D Opacity Grid from LiDAR for Scene Forecasting
Chuanyu Pan,Aolin Xu
https://openreview.net/forum?id=MfuzopqVOX
Timely capturing the dense geometry of the surrounding scene with unlabeled LiDAR data is valuable but under-explored for mobile robotic applications. Its value lies in the huge amount of such unlabeled data, enabling self-supervised learning for various downstream tasks. Current dynamic 3D scene reconstruction approaches however heavily rely on data annotations to tackle the moving objects in the scene. In response, we present LiDARGrid, a 3D opacity grid representation instantly derived from LiDAR points, which captures the dense 3D scene and facilitates scene forecasting. Our method features a novel self-supervised neural volume densification procedure based on an autoencoder and differentiable volume rendering. Leveraging this representation, self-supervised scene forecasting can be performed. Our method is trained on NuScenes dataset for autonomous driving, and is evaluated by predicting future point clouds using the scene forecasting. It notably outperforms state-of-the-art methods in point cloud forecasting in all performance metrics. Beyond scene forecasting, our representation excels in supporting additional tasks such as moving region detection and depth completion, as shown by experiments.

37

ReMix: Optimizing Data Mixtures for Large Scale Imitation Learning
Joey Hejna,Chethan Anand Bhateja,Yichen Jiang,Karl Pertsch,Dorsa Sadigh
https://openreview.net/forum?id=flj88Tn3fc
Increasingly large robotics datasets are being collected to train larger foundation models in robotics. However, despite the fact that data selection has been of utmost importance to scaling in vision and natural language processing (NLP), little work in robotics has questioned what data such models should actually be trained on. In this work we investigate how to weigh different subsets or ``domains'' of robotics datasets during pre-training to maximize worst-case performance across all possible downstream domains using distributionally robust optimization (DRO). Unlike in NLP, we find that these methods are hard to apply out of the box due to varying action spaces and dynamics across robots. Our method, ReMix, employs early stopping and action normalization and discretization to counteract these issues. Through extensive experimentation on both the Bridge and OpenX datasets, we demonstrate that data curation can have an outsized impact on downstream performance. Specifically, domain weights learned by ReMix outperform uniform weights by over 40% on average and human-selected weights by over 20% on datasets used to train the RT-X models.

38

HiRT: Enhancing Robotic Control with Hierarchical Robot Transformers
Jianke Zhang,Yanjiang Guo,Xiaoyu Chen,Yen-Jen Wang,Yucheng Hu,Chengming Shi,Jianyu Chen
https://openreview.net/forum?id=wTKJge0PTq
Large Vision-Language-Action (VLA) models, leveraging powerful pre-trained Vision-Language Models (VLMs) backends, have shown promise in robotic control due to their impressive generalization ability. However, the success comes at a cost. Their reliance on VLM backends with billions of parameters leads to high computational costs and inference latency, limiting the testing scenarios to mainly quasi-static tasks and hindering performance in dynamic tasks requiring rapid interactions. To address these limitations, this paper proposes \textbf{HiRT}, a

\textbf{Hi}erarchical \textbf{R}obot \textbf{T}ransformer framework that enables flexible frequency and performance trade-off. HiRT keeps VLMs running at low frequencies to capture temporarily invariant features while enabling real-time interaction through a high-frequency vision-based policy guided by the slowly updated features. Experiment results in both simulation and real-world settings demonstrate significant improvements over baseline methods. Empirically, we achieve a 58% reduction in inference time delay while maintaining comparable success rates. Additionally, on novel dynamic manipulation benchmarks which are challenging for previous VLA models, HiRT improves the success rate from 48% to 75%.

39

Exploring Under Constraints with Model-Based Actor-Critic and Safety Filters
Ahmed Agha,Baris Kayalibay,Atanas Mirchev,Patrick van der Smagt,Justin Bayer
https://openreview.net/forum?id=s31IWg2kN5
Applying reinforcement learning (RL) to learn effective policies on physical robots without supervision remains challenging when it comes to tasks where safe exploration is critical. Constrained model-based RL (CMBRL) presents a promising approach to this problem. These methods are designed to learn constraint-adhering policies through constrained optimization approaches. Yet, such policies often fail to meet stringent safety requirements during learning and exploration. Our solution ``CASE'' aims to reduce the instances where constraints are breached during the learning phase. Specifically, CASE integrates techniques for optimizing constrained policies and employs planning-based safety filters as backup policies, effectively lowering constraint violations during learning and making it a more reliable option than other recent constrained model-based policy optimization methods.

40

Progressive Multi-Modal Fusion for Robust 3D Object Detection
Rohit Mohan,Daniele Cattaneo,Florian Drews,Abhinav Valada
https://openreview.net/forum?id=Qoy12gkH4C
Multi-sensor fusion is crucial for accurate 3D object detection in autonomous driving, with cameras and LiDAR being the most commonly used sensors. However, existing methods perform sensor fusion in a single view by projecting features from both modalities either in Bird's Eye View (BEV) or Perspective View (PV), thus sacrificing complementary information such as height or geometric proportions. To address this limitation, we propose ProFusion3D, a progressive fusion framework that combines features in both BEV and PV at both intermediate and object query levels. Our architecture hierarchically fuses local and global features, enhancing the robustness of 3D object detection. Additionally, we introduce a self-supervised mask modeling pre-training strategy to improve multi-modal representation learning and data efficiency through three novel objectives. Extensive experiments on nuScenes and Argoverse2 datasets conclusively demonstrate the efficacy of ProFusion3D. Moreover, ProFusion3D is robust to sensor failure, showing strong performance when only one modality is available.

41

SPIRE: Synergistic Planning, Imitation, and Reinforcement Learning for Long-Horizon Manipulation
Zihan Zhou,Animesh Garg,Dieter Fox,Caelan Reed Garrett,Ajay Mandlekar
https://openreview.net/forum?id=cvUXoou8iz
Robot learning has proven to be a general and effective technique for programming manipulators. Imitation learning is able to teach robots solely from human demonstrations but is bottlenecked by

the capabilities of the demonstrations. Reinforcement learning uses exploration to discover better behaviors; however, the space of possible improvements can be too large to start from scratch. And for both techniques, the learning difficulty increases proportional to the length of the manipulation task. Accounting for this, we propose SPIRE, a system that first uses Task and Motion Planning (TAMP) to decompose tasks into smaller learning subproblems and second combines imitation and reinforcement learning to maximize their strengths. We develop novel strategies to train learning agents when deployed in the context of a planning system. We evaluate SPIRE on a suite of long-horizon and contact-rich robot manipulation problems. We find that SPIRE outperforms prior approaches that integrate imitation learning, reinforcement learning, and planning by 35% to 50% in average task performance, is 6 times more data efficient in the number of human demonstrations needed to train proficient agents, and learns to complete tasks nearly twice as efficiently. View https://sites.google.com/view/spire-corl-2024 for more details.

42

I Can Tell What I am Doing: Toward Real-World Natural Language Grounding of Robot Experiences
Zihan Wang,Brian Liang,Varad Dhat,Zander Brumbaugh,Nick Walker,Ranjay Krishna,Maya Cakmak
https://openreview.net/forum?id=iZF0FRPgfq
Understanding robot behaviors and experiences through natural language is crucial for developing intelligent and transparent robotic systems. Recent advancement in large language models (LLMs) makes it possible to translate complex, multi-modal robotic experiences into coherent, human-readable narratives. However, grounding real-world robot experiences into natural language is challenging due to many reasons, such as multi-modal nature of data, differing sample rates, and data volume. We introduce RONAR, an LLM-based system that generates natural language narrations from robot experiences, aiding in behavior announcement, failure analysis, and human interaction to recover failure. Evaluated across various scenarios, RONAR outperforms state-of-the-art methods and improves failure recovery efficiency. Our contributions include a multi-modal framework for robot experience narration, a comprehensive real-robot dataset, and empirical evidence of RONAR's effectiveness in enhancing user experience in system transparency and failure analysis.

43

APRICOT: Active Preference Learning and Constraint-Aware Task Planning with LLMs
Huaxiaoyue Wang,Nathaniel Chin,Gonzalo Gonzalez-Pumariega,Xiangwan Sun,Neha Sunkara,Maximus Adrian Pace,Jeannette Bohg,Sanjiban Choudhury
https://openreview.net/forum?id=nQslM6f7dW
Home robots performing personalized tasks must adeptly balance user preferences with environmental affordances. We focus on organization tasks within constrained spaces, such as arranging items into a refrigerator, where preferences for placement collide with physical limitations. The robot must infer user preferences based on a small set of demonstrations, which is easier for users to provide than extensively defining all their requirements. While recent works use Large Language Models (LLMs) to learn preferences from user demonstrations, they encounter two fundamental challenges. First, there is inherent ambiguity in interpreting user actions, as multiple preferences can often explain a single observed behavior. Second, not all user preferences are practically feasible due to geometric constraints in the environment. To address these challenges, we introduce APRICOT, a novel approach that merges LLM-based Bayesian active preference learning with constraint-aware task planning. APRICOT refines its generated preferences by actively querying the user and dynamically adapts its plan to respect environmental constraints. We evaluate APRICOT on a dataset of diverse organization tasks and

demonstrate its effectiveness in real-world scenarios, showing significant improvements in both preference satisfaction and plan feasibility.

44

Leveraging Locality to Boost Sample Efficiency in Robotic Manipulation
Tong Zhang,Yingdong Hu,Jiacheng You,Yang Gao
https://openreview.net/forum?id=Qpjo8I8AFW
Given the high cost of collecting robotic data in the real world, sample efficiency is a consistently compelling pursuit in robotics. In this paper, we introduce SGRv2, an imitation learning framework that enhances sample efficiency through improved visual and action representations. Central to the design of SGRv2 is the incorporation of a critical inductive bias—$action\ locality$, which posits that robot's actions are predominantly influenced by the target object and its interactions with the local environment. Extensive experiments in both simulated and real-world settings demonstrate that action locality is essential for boosting sample efficiency. SGRv2 excels in RLBench tasks with keyframe control using merely 5 demonstrations and surpasses the RVT baseline in 23 of 26 tasks. Furthermore, when evaluated on ManiSkill2 and MimicGen using dense control, SGRv2's success rate is 2.54 times that of SGR. In real-world environments, with only eight demonstrations, SGRv2 can perform a variety of tasks at a markedly higher success rate compared to baseline models.

45

Accelerating Visual Sparse-Reward Learning with Latent Nearest-Demonstration-Guided Explorations
Ruihan Zhao,ufuk topcu,Sandeep P. Chinchali,Mariano Phielipp
https://openreview.net/forum?id=3NI5SxsJqf
Recent progress in deep reinforcement learning (RL) and computer vision enables artificial agents to solve complex tasks, including locomotion, manipulation, and video games from high-dimensional pixel observations. However, RL usually relies on domain-specific reward functions for sufficient learning signals, requiring expert knowledge. While vision-based agents could learn skills from only sparse rewards, exploration challenges arise. We present Latent Nearest-demonstration-guided Exploration (LaNE), a novel and efficient method to solve sparse-reward robot manipulation tasks from image observations and a few demonstrations. First, LaNE builds on the pre-trained DINOv2 feature extractor to learn an embedding space for forward prediction. Next, it rewards the agent for exploring near the demos, quantified by quadratic control costs in the embedding space. Finally, LaNE optimizes the policy for the augmented rewards with RL. Experiments demonstrate that our method achieves state-of-the-art sample efficiency in Robosuite simulation and enables under-an-hour RL training from scratch on a Franka Panda robot, using only a few demonstrations.

46

Dynamic 3D Gaussian Tracking for Graph-Based Neural Dynamics Modeling
Mingtong Zhang,Kaifeng Zhang,Yunzhu Li
https://openreview.net/forum?id=itKJ5uu1gW
Videos of robots interacting with objects encode rich information about the objects' dynamics. However, existing video prediction approaches typically do not explicitly account for the 3D information from videos, such as robot actions and objects' 3D states, limiting their use in real-world robotic applications. In this work, we introduce a framework to learn object dynamics

directly from multi-view RGB videos by explicitly considering the robot's action trajectories and their effects on scene dynamics. We utilize the 3D Gaussian representation of 3D Gaussian Splatting (3DGS) to train a particle-based dynamics model using Graph Neural Networks. This model operates on sparse control particles downsampled from the densely tracked 3D Gaussian reconstructions. By learning the neural dynamics model on offline robot interaction data, our method can predict object motions under varying initial configurations and unseen robot actions. The 3D transformations of Gaussians can be interpolated from the motions of control particles, enabling the rendering of predicted future object states and achieving action-conditioned video prediction. The dynamics model can also be applied to model-based planning frameworks for object manipulation tasks. We conduct experiments on various kinds of deformable materials, including ropes, clothes, and stuffed animals, demonstrating our framework's ability to model complex shapes and dynamics. Our project page is available at \url{https://gaussian-gbnd.github.io/}.

47

Surgical Robot Transformer (SRT): Imitation Learning for Surgical Tasks
Ji Woong Kim,Tony Z. Zhao,Samuel Schmidgall,Anton Deguet,Marin Kobilarov,Chelsea Finn,Axel Krieger
https://openreview.net/forum?id=fNBbEgcfwO
We explore whether surgical manipulation tasks can be learned on the da Vinci robot via imitation learning. However, the da Vinci system presents unique challenges which hinder straight-forward implementation of imitation learning. Notably, its forward kinematics is inconsistent due to imprecise joint measurements, and naively training a policy using such approximate kinematics data often leads to task failure. To overcome this limitation, we introduce a relative action formulation which enables successful policy training and deployment using its approximate kinematics data. A promising outcome of this approach is that the large repository of clinical data, which contains approximate kinematics, may be directly utilized for robot learning without further corrections. We demonstrate our findings through successful execution of three fundamental surgical tasks, including tissue manipulation, needle handling, and knot-tying.

48

Simple Masked Training Strategies Yield Control Policies That Are Robust to Sensor Failure
Skand Skand,Bikram Pandit,Chanho Kim,Li Fuxin,Stefan Lee
https://openreview.net/forum?id=AsbyZRdqPv
Sensor failure is common when robots are deployed in the real world, as sensors naturally wear out over time. Such failures can lead to catastrophic outcomes, including damage to the robot from unexpected robot behaviors such as falling during walking. Previous work has tried to address this problem by recovering missing sensor values from the history of states or by adapting learned control policies to handle corrupted sensors through fine-tuning during deployment. In this work, we propose training reinforcement learning (RL) policies that are robust to sensory failures. We use a multimodal encoder designed to account for these failures and a training strategy that randomly drops a subset of sensor modalities, similar to missing observations caused by failed sensors. We conduct evaluations across multiple tasks (bipedal locomotion and robotic manipulation) with varying robot embodiments in both simulation and the real world to demonstrate the effectiveness of our approach. Our results show that the proposed method produces robust RL policies that handle failures in both low-dimensional proprioceptive and high-dimensional visual modalities without a significant increase in training time or decrease

in sample efficiency, making it a promising solution for learning RL policies robust to sensory failures.

49

Scaling Robot Policy Learning via Zero-Shot Labeling with Foundation Models
Nils Blank,Moritz Reuss,Marcel Rühle,Ömer Erdinç Yağmurlu,Fabian Wenzel,Oier Mees,Rudolf Lioutikov
https://openreview.net/forum?id=EdVNB2kHv1
A central challenge towards developing robots that can relate human language to their perception and actions is the scarcity of natural language annotations in diverse robot datasets. Moreover, robot policies that follow natural language instructions are typically trained on either templated language or expensive human-labeled instructions, hindering their scalability. To this end, we introduce NILS: Natural language Instruction Labeling for Scalability. NILS automatically labels uncurated, long-horizon robot data at scale in a zero-shot manner without any human intervention. NILS combines pre-trained vision-language foundation models in a sophisticated, carefully considered manner in order to detect objects in a scene, detect object-centric changes, segment tasks from large datasets of unlabelled interaction data and ultimately label behavior datasets. Evaluations on BridgeV2 and a kitchen play dataset show that NILS is able to autonomously annotate diverse robot demonstrations of unlabeled and unstructured datasets, while alleviating several shortcomings of crowdsourced human annotations.

50

Vocal Sandbox: Continual Learning and Adaptation for Situated Human-Robot Collaboration
Jennifer Grannen,Siddharth Karamcheti,Suvir Mirchandani,Percy Liang,Dorsa Sadigh
https://openreview.net/forum?id=ypaYtV1CoG
We introduce Vocal Sandbox, a framework for enabling seamless human-robot collaboration in situated environments. Systems in our framework are characterized by their ability to adapt and continually learn at multiple levels of abstraction from diverse teaching modalities such as spoken dialogue, object keypoints, and kinesthetic demonstrations. To enable such adaptation, we design lightweight and interpretable learning algorithms that allow users to build an understanding and co-adapt to a robot's capabilities in real-time, as they teach new behaviors. For example, after demonstrating a new low-level skill for "tracking around" an object, users are provided with trajectory visualizations of the robot's intended motion when asked to track a new object. Similarly, users teach high-level planning behaviors through spoken dialogue, using pretrained language models to synthesize behaviors such as "packing an object away" as compositions of low-level skills -- concepts that can be reused and built upon. We evaluate Vocal Sandbox in two settings: collaborative gift bag assembly and LEGO stop-motion animation. In the first setting, we run systematic ablations and user studies with 8 non-expert participants, highlighting the impact of multi-level teaching. Across 23 hours of total robot interaction time, users teach 17 new high-level behaviors with an average of 16 novel low-level skills, requiring 22.1% less active supervision compared to baselines. Qualitatively, users strongly prefer Vocal Sandbox systems due to their ease of use (+31.2%), helpfulness (+13.0%), and overall performance (+18.2%). Finally, we pair an experienced system-user with a robot to film a stop-motion animation; over two hours of continuous collaboration, the user teaches progressively more complex motion skills to produce a 52 second (232 frame) movie. Videos & Supplementary Material: https://vocal-sandbox.github.io

51

What Makes Pre-Trained Visual Representations Successful for Robust Manipulation?
Kaylee Burns,Zach Witzel,Jubayer Ibn Hamid,Tianhe Yu,Chelsea Finn,Karol Hausman
https://openreview.net/forum?id=A1hpY5RNiH
Inspired by the success of transfer learning in computer vision, roboticists have investigated visual pre-training as a means to improve the learning efficiency and generalization ability of policies learned from pixels. To that end, past work has favored large object interaction datasets, such as first-person videos of humans completing diverse tasks, in pursuit of manipulation-relevant features. Although this approach improves the efficiency of policy learning, it remains unclear how reliable these representations are in the presence of distribution shifts that arise commonly in robotic applications. Surprisingly, we find that visual representations designed for control tasks do not necessarily generalize under subtle changes in lighting and scene texture or the introduction of distractor objects. To understand what properties do lead to robust representations, we compare the performance of 15 pre-trained vision models under different visual appearances. We find that emergent segmentation ability is a strong predictor of out-of-distribution generalization among ViT models. The rank order induced by this metric is more predictive than metrics that have previously guided generalization research within computer vision and machine learning, such as downstream ImageNet accuracy, in-domain accuracy, or shape-bias as evaluated by cue-conflict performance. We test this finding extensively on a suite of distribution shifts in ten tasks across two simulated manipulation environments. On the ALOHA setup, segmentation score predicts real-world performance after offline training with 50 demonstrations.

52

DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models
Xiaoyu Tian,Junru Gu,Bailin Li,Yicheng Liu,Yang Wang,Zhiyong Zhao,Kun Zhan,Peng Jia,XianPeng Lang,Hang Zhao
https://openreview.net/forum?id=928V4Umlys
A primary hurdle of autonomous driving in urban environments is understanding complex and long-tail scenarios, such as challenging road conditions and delicate human behaviors. We introduce DriveVLM, an autonomous driving system leveraging Vision-Language Models (VLMs) for enhanced scene understanding and planning capabilities. DriveVLM integrates a unique combination of reasoning modules for scene description, scene analysis, and hierarchical planning. Furthermore, recognizing the limitations of VLMs in spatial reasoning and heavy computational requirements, we propose DriveVLM-Dual, a hybrid system that synergizes the strengths of DriveVLM with the traditional autonomous driving pipeline. Experiments on both the nuScenes dataset and our SUP-AD dataset demonstrate the efficacy of DriveVLM and DriveVLM-Dual in handling complex and unpredictable driving conditions. Finally, we deploy the DriveVLM-Dual on a production vehicle, verifying it is effective in real-world autonomous driving environments.

53

Unpacking Failure Modes of Generative Policies: Runtime Monitoring of Consistency and Progress
Christopher Agia,Rohan Sinha,Jingyun Yang,Ziang Cao,Rika Antonova,Marco Pavone,Jeannette Bohg
https://openreview.net/forum?id=yqLFb0RnDW
Robot behavior policies trained via imitation learning are prone to failure under conditions that deviate from their training data. Thus, algorithms that monitor learned policies at test time and

provide early warnings of failure are necessary to facilitate scalable deployment. We propose Sentinel, a runtime monitoring framework that splits the detection of failures into two complementary categories: 1) Erratic failures, which we detect using statistical measures of temporal action consistency, and 2) task progression failures, where we use Vision Language Models (VLMs) to detect when the policy confidently and consistently takes actions that do not solve the task. Our approach has two key strengths. First, because learned policies exhibit diverse failure modes, combining complementary detectors leads to significantly higher accuracy at failure detection. Second, using a statistical temporal action consistency measure ensures that we quickly detect when multimodal, generative policies exhibit erratic behavior at negligible computational cost. In contrast, we only use VLMs to detect modes that are less time-sensitive. We demonstrate our approach in the context of diffusion policies trained on robotic mobile manipulation domains in both simulation and the real world. By unifying temporal consistency detection and VLM runtime monitoring, Sentinel detects 18% more failures than using either of the two detectors alone and significantly outperforms baselines, thus highlighting the importance of assigning specialized detectors to complementary categories of failure. Qualitative results are made available at sites.google.com/stanford.edu/sentinel.

54

ScissorBot: Learning Generalizable Scissor Skill for Paper Cutting via Simulation, Imitation, and Sim2Real

Jiangran Lyu,Yuxing Chen,Tao Du,Feng Zhu,Huiquan Liu,Yizhou Wang,He Wang

https://openreview.net/forum?id=PAtsxVz0ND

This paper tackles the challenging robotic task of generalizable paper cutting using scissors. In this task, scissors attached to a robot arm are driven to accurately cut curves drawn on the paper, which is hung with the top edge fixed. Due to the frequent paper-scissor contact and consequent fracture, the paper features continual deformation and changing topology, which is diffult for accurate modeling.To deal with such versatile scenarios, we propose ScissorBot, the first learning-based system for robotic paper cutting with scissors via simulation, imitation learning and sim2real. Given the lack of sufficient data for this task, we build PaperCutting-Sim, a paper simulator supporting interactive fracture coupling with scissors, enabling demonstration generation with a heuristic-based oracle policy. To ensure effective execution, we customize an action primitive sequence for imitation learning to constrain its action space, thus alleviating potential compounding errors. Finally, by integrating sim-to-real techniques to bridge the gap between simulation and reality, our policy can be effectively deployed on the real robot. Experimental results demonstrate that our method surpasses all baselines in both simulation and real-world benchmarks and achives performance comparable to human operation with a single hand under the same conditions.

55

Fleet Supervisor Allocation: A Submodular Maximization Approach

Oguzhan Akcin,Ahmet Ege Tanriverdi,Kaan Kale,Sandeep P. Chinchali

https://openreview.net/forum?id=9dsBQhoqVr

In real-world scenarios, the data collected by robots in diverse and unpredictable environments is crucial for enhancing their perception and decision-making models. This data is predominantly collected under human supervision, particularly through imitation learning (IL), where robots learn complex tasks by observing human supervisors. However, the deployment of multiple robots and supervisors to accelerate the learning process often leads to data redundancy and inefficiencies, especially as the scale of robot fleets increases. Moreover, the reliance on teleoperation for

supervision introduces additional challenges due to potential network connectivity issues. To address these issues in data collection, we introduce an Adaptive Submodular Allocation policy, ASA, designed for efficient human supervision allocation within multi-robot systems under uncertain connectivity conditions. Our approach reduces data redundancy by balancing the informativeness and diversity of data collection, and is capable of accommodating connectivity variances. We evaluate the effectiveness of ASA in simulations with 100 robots across four different environments and various network settings, including a real-world teleoperation scenario over a 5G network. We train and test our policy, ASA, and state-of-the-art policies utilizing NVIDIA's Isaac Gym. Our results show that ASA enhances the return on human effort by up to $3.37\times$, outperforming current baselines in all simulated scenarios and providing robustness against connectivity disruptions.

56

Learning Quadruped Locomotion Using Differentiable Simulation
Yunlong Song,Sang bae Kim,Davide Scaramuzza
https://openreview.net/forum?id=XopATjibyz
This work explores the potential of using differentiable simulation for learning robot control. Differentiable simulation promises fast convergence and stable training by computing low-variance first-order gradients using the robot model. Still, so far, its usage for legged robots is limited to simulation. The main challenge lies in the complex optimization landscape of robotic tasks due to discontinuous dynamics. This work proposes a new differentiable simulation framework to overcome these challenges. The key idea involves decoupling the complex whole-body simulation, which may exhibit discontinuities due to contact into two separate continuous domains. Subsequently, we align the robot state resulting from the simplified model with a more precise, non-differentiable simulator to maintain sufficient simulation accuracy. Our framework enables learning quadruped walking in simulation in minutes without parallelization. When augmented with GPU parallelization, our approach allows the quadruped robot to master diverse locomotion skills on challenging terrains in minutes. We demonstrate that differentiable simulation outperforms a reinforcement learning algorithm (PPO) by achieving significantly better sample efficiency while maintaining its effectiveness in handling large-scale environments. Our policy achieves robust locomotion performance in the real world zero-shot.

57

RoboKoop: Efficient Control Conditioned Representations from Visual Input in Robotics using Koopman Operator
Hemant Kumawat,Biswadeep Chakraborty,Saibal Mukhopadhyay
https://openreview.net/forum?id=NiA8hVdDS7
Developing agents that can perform complex control tasks from high-dimensional observations is a core ability of autonomous agents that requires underlying robust task control policies and adapting the underlying visual representations to the task. Most existing policies need a lot of training samples and treat this problem from the lens of two-stage learning with a controller learned on top of pre-trained vision models. We approach this problem from the lens of Koopman theory and learn visual representations from robotic agents conditioned on specific downstream tasks in the context of learning stabilizing control for the agent. We introduce a Contrastive Spectral Koopman Embedding network that allows us to learn efficient linearized visual representations from the agent's visual data in a high dimensional latent space and utilizes reinforcement learning to perform off-policy control on top of the extracted representations with a linear controller. Our method enhances stability and control in gradient dynamics over time,

significantly outperforming existing approaches by improving efficiency and accuracy in learning task policies over extended horizons.

58

Text2Interaction: Establishing Safe and Preferable Human-Robot Interaction
Jakob Thumm,Christopher Agia,Marco Pavone,Matthias Althoff
https://openreview.net/forum?id=s0VNSnPeoA
Adjusting robot behavior to human preferences can require intensive human feedback, preventing quick adaptation to new users and changing circumstances. Moreover, current approaches typically treat user preferences as a reward, which requires a manual balance between task success and user satisfaction. To integrate new user preferences in a zero-shot manner, our proposed Text2Interaction framework invokes large language models to generate a task plan, motion preferences as Python code, and parameters of a safety controller. By maximizing the combined probability of task completion and user satisfaction instead of a weighted sum of rewards, we can reliably find plans that fulfill both requirements. We find that 83% of users working with Text2Interaction agree that it integrates their preferences into the plan of the robot, and 94% prefer Text2Interaction over the baseline. Our ablation study shows that Text2Interaction aligns better with unseen preferences than other baselines while maintaining a high success rate. Real-world demonstrations and code are made available at sites.google.com/view/text2interaction.

59

VIRL: Self-Supervised Visual Graph Inverse Reinforcement Learning
Lei Huang,Weijia Cai,Zihan Zhu,Chen Feng,Helge Rhodin,Zhengbo Zou
https://openreview.net/forum?id=fDRO4NHEwZ
Learning dense reward functions from unlabeled videos for reinforcement learning exhibits scalability due to the vast diversity and quantity of video resources. Recent works use visual features or graph abstractions in videos to measure task progress as rewards, which either deteriorate in unseen domains or capture spatial information while overlooking visual details. We propose $\mathbf{V}$isual-Graph $\mathbf{I}$nverse $\mathbf{R}$einforcement $\mathbf{L}$earning (VIRL), a self-supervised method that synergizes low-level visual features and high-level graph abstractions from frames to graph representations for reward learning. VIRL utilizes a visual encoder that extracts object-wise features for graph nodes and a graph encoder that derives properties from graphs constructed from detected objects in each frame. The encoded representations are enforced to align videos temporally and reconstruct in-scene objects. The pretrained visual graph encoder is then utilized to construct a dense reward function for policy learning by measuring latent distances between current frames and the goal frame. Our empirical evaluation on the X-MAGICAL and Robot Visual Pusher benchmark demonstrates that VIRL effectively handles tasks necessitating both granular visual attention and broader global feature consideration, and exhibits robust generalization to $extrapolation$ tasks and domains not seen in demonstrations. Our policy for the robotic task also achieves the highest success rate in real-world robot experiments.

60

Task-Oriented Hierarchical Object Decomposition for Visuomotor Control
Jianing Qian,Yunshuang Li,Bernadette Bucher,Dinesh Jayaraman
https://openreview.net/forum?id=hV97HJm7Ag
Good pre-trained visual representations could enable robots to learn visuomotor policy efficiently. Still, existing representations take a one-size-fits-all-tasks approach that comes with two

important drawbacks: (1) Being completely task-agnostic, these representations cannot effectively ignore any task-irrelevant information in the scene, and (2) They often lack the representational capacity to handle unconstrained/complex real-world scenes. Instead, we propose to train a large combinatorial family of representations organized by scene entities: objects and object parts. This hierarchical object decomposition for task-oriented representations (HODOR) permits selectively assembling different representations specific to each task while scaling in representational capacity with the complexity of the scene and the task. In our experiments, we find that HODOR outperforms prior pre-trained representations, both scene vector representations and object-centric representations, for sample-efficient imitation learning across 5 simulated and 5 real-world manipulation tasks. We further find that the invariances captured in HODOR are inherited into downstream policies, which can robustly generalize to out-of-distribution test conditions, permitting zero-shot skill chaining. Appendix and videos: https://sites.google.com/view/ hodor-corl24

61

Learning to Look: Seeking Information for Decision Making via Policy Factorization
Shivin Dass,Jiaheng Hu,Ben Abbatematteo,Peter Stone,Roberto Martín-Martín
https://openreview.net/forum?id=B2X57y37kC
Many robot manipulation tasks require active or interactive exploration behavior in order to be performed successfully. Such tasks are ubiquitous in embodied domains, where agents must actively search for the information necessary for each stage of a task, e.g., moving the head of the robot to find information relevant to manipulation, or in multi-robot domains, where one scout robot may search for the information that another robot needs to make informed decisions. We identify these tasks with a new type of problem, factorized Contextual Markov Decision Processes, and propose DISaM, a dual-policy solution composed of an information-seeking policy that explores the environment to find the relevant contextual information and an information-receiving policy that exploits the context to achieve the manipulation goal. This factorization allows us to train both policies separately, using the information-receiving one to provide reward to train the information-seeking policy. At test time, the dual agent balances exploration and exploitation based on the uncertainty the manipulation policy has on what the next best action is. We demonstrate the capabilities of our dual policy solution in five manipulation tasks that require information-seeking behaviors, both in simulation and in the real-world, where DISaM significantly outperforms existing methods. More information at https://robin-lab.cs.utexas.edu/learning2look/.

62

SoftManiSim: A Fast Simulation Framework for Multi-Segment Continuum Manipulators Tailored for Robot Learning
Mohammadreza Kasaei,Hamidreza Kasaei,Mohsen Khadem
https://openreview.net/forum?id=ovjxugn9Q2
This paper introduces SoftManiSim, a novel simulation framework for multi-segment continuum manipulators. Existing continuum robot simulators often rely on simplifying assumptions, such as constant curvature bending or ignoring contact forces, to meet real-time simulation and training demands. To bridge this gap, we propose a robust and rapid mathematical model for continuum robots at the core of SoftManiSim, ensuring precise and adaptable simulations. The framework can integrate with various rigid-body robots, increasing its utility across different robotic platforms. SoftManiSim supports parallel operations for simultaneous simulations of multiple robots and generates synthetic data essential for training deep reinforcement learning models. This capability enhances the development and optimization of control strategies in dynamic

environments. Extensive simulations validate the framework's effectiveness, demonstrating its capabilities in handling complex robotic interactions and tasks. We also present real robot validation to showcase the simulator's practical applicability and accuracy in real-world settings. To our knowledge, SoftManiSim is the first open-source real-time simulator capable of modeling continuum robot behavior under dynamic point/distributed loading. It enables rapid deployment in reinforcement learning and machine learning applications. This simulation framework can be downloaded from https://github.com/MohammadKasaei/SoftManiSim.

63

InterACT: Inter-dependency Aware Action Chunking with Hierarchical Attention Transformers for Bimanual Manipulation
Andrew Choong-Won Lee,Ian Chuang,Ling-Yuan Chen,Iman Soltani
https://openreview.net/forum?id=lKGRPJFPCM
We present InterACT: Inter-dependency aware Action Chunking with Hierarchical Attention Transformers, a novel imitation learning framework for bimanual manipulation that integrates hierarchical attention to capture inter-dependencies between dual-arm joint states and visual inputs. InterACT consists of a Hierarchical Attention Encoder and a Multi-arm Decoder, both designed to enhance information aggregation and coordination. The encoder processes multi-modal inputs through segment-wise and cross-segment attention mechanisms, while the decoder leverages synchronization blocks to refine individual action predictions, providing the counterpart's prediction as context. Our experiments on a variety of simulated and real-world bimanual manipulation tasks demonstrate that InterACT significantly outperforms existing methods. Detailed ablation studies validate the contributions of key components of our work, including the impact of CLS tokens, cross-segment encoders, and synchronization blocks.

64

Continuously Improving Mobile Manipulation with Autonomous Real-World RL
Russell Mendonca,Emmanuel Panov,Bernadette Bucher,Jiuguang Wang,Deepak Pathak
https://openreview.net/forum?id=46SluHKoE9
We present a fully autonomous real-world RL framework for mobile manipulation that can learn policies without extensive instrumentation or human supervision. This is enabled by 1) task-relevant autonomy, which guides exploration towards object interactions and prevents stagnation near goal states, 2) efficient policy learning by leveraging basic task knowledge in behavior priors, and 3) formulating generic rewards that combine human-interpretable semantic information with low-level, fine-grained observations. We demonstrate that our approach allows Spot robots to continually improve their performance on a set of four challenging mobile manipulation tasks, obtaining an average success rate of 80% across tasks, a 3-4 times improvement over existing approaches. Videos can be found at https://continual-mobile-manip.github.io/.

65

EquiGraspFlow: SE(3)-Equivariant 6-DoF Grasp Pose Generative Flows
Byeongdo Lim,Jongmin Kim,Jihwan Kim,Yonghyeon Lee,Frank C. Park
https://openreview.net/forum?id=5lSkn5v4LK
Traditional methods for synthesizing 6-DoF grasp poses from 3D observations often rely on geometric heuristics, resulting in poor generalizability, limited grasp options, and higher failure rates. Recently, data-driven methods have been proposed that use generative models to learn the distribution of grasp poses and generate diverse candidate poses. The main drawback of these

methods is that they fail to achieve SE(3)-equivariance, meaning that the generated grasp poses do not transform correctly with object rotations and translations. In this paper, we propose \textit{EquiGraspFlow}, a flow-based SE(3)-equivariant 6-DoF grasp pose generative model that can learn complex conditional distributions on the SE(3) manifold while guaranteeing SE(3)-equivariance. Our model achieves the equivariance without relying on data augmentation, by using network architectures that guarantee the equivariance by construction. Extensive experiments show that \textit{EquiGraspFlow} accurately learns grasp pose distribution, achieves the SE(3)-equivariance, and significantly outperforms existing grasp pose generative models. Code is available at https://github.com/bdlim99/EquiGraspFlow.

66

Genetic Algorithm for Curriculum Design in Multi-Agent Reinforcement Learning
Yeeho Song,Jeff Schneider
https://openreview.net/forum?id=2CScZqkUPZ
As the deployment of autonomous agents in real-world scenarios grows, so does the interest in their application to competitive environments with other robots. Self-play in Reinforcement Learning (RL) enables agents to develop competitive strategies. However, the complexity arising from multi-agent interactions and the tendency for RL agents to disrupt competitors' training introduce instability and a risk of overfitting. While traditional methods depend on costly Nash equilibrium approximations or random exploration for training scenario optimization, this can be inefficient in large search spaces often prevalent in multi-agent problems. However, related works in single-agent setups show that genetic algorithms perform better in large scenario spaces. Therefore, we propose using genetic algorithms to adaptively adjust environment parameters and opponent policies in a multi-agent context to find and synthesize coherent scenarios efficiently. We also introduce GenOpt Agent—a genetically optimized, open-loop agent executing scheduled actions. The open-loop aspect of GenOpt prevents RL agents from winning through adversarial perturbations, thereby fostering generalizable strategies. Also, GenOpt is genetically optimized without expert supervision, negating the need for expensive expert supervision to have meaningful opponents at the start of training. Our empirical studies indicate that this method surpasses several established baselines in two-player competitive settings with continuous action spaces, validating its effectiveness and stability in training.

67

Learning Decentralized Multi-Biped Control for Payload Transport
Bikram Pandit,Ashutosh Gupta,Mohitvishnu S. Gadde,Addison Johnson,Aayam Kumar Shrestha,Helei Duan,Jeremy Dao,Alan Fern
https://openreview.net/forum?id=vhGkyWgctu
Payload transport over flat terrain via multi-wheel robot carriers is well-understood, highly effective, and configurable. In this paper, our goal is to provide similar effectiveness and configurability for transport over rough terrain that is more suitable for legs rather than wheels. For this purpose, we consider multi-biped robot carriers, where wheels are replaced by multiple bipedal robots attached to the carrier. Our main contribution is to design a decentralized controller for such systems that can be effectively applied to varying numbers and configurations of rigidly attached bipedal robots without retraining. We present a reinforcement learning approach for training the controller in simulation that supports transfer to the real world. Our experiments in simulation provide quantitative metrics showing the effectiveness of the approach over a wide variety of simulated transport scenarios. In addition, we demonstrate the controller in the real-

world for systems composed of two and three Cassie robots. To our knowledge, this is the first example of a scalable multi-biped payload transport system.

68

Monocular Event-Based Vision for Obstacle Avoidance with a Quadrotor
Anish Bhattacharya,Marco Cannici,Nishanth Rao,Yuezhan Tao,Vijay Kumar,Nikolai Matni,Davide Scaramuzza
https://openreview.net/forum?id=82bpTugrMt
We present the first static-obstacle avoidance method for quadrotors using just an onboard, monocular event camera. Quadrotors are capable of fast and agile flight in cluttered environments when piloted manually, but vision-based autonomous flight in unknown environments is difficult in part due to the sensor limitations of traditional onboard cameras. Event cameras, however, promise nearly zero motion blur and high dynamic range, but produce a very large volume of events under significant ego-motion and further lack a continuous-time sensor model in simulation, making direct sim-to-real transfer not possible. By leveraging depth prediction as a pretext task in our learning framework, we can pre-train a reactive obstacle avoidance events-to-control policy with approximated, simulated events and then fine-tune the perception component with limited events-and-depth real-world data to achieve obstacle avoidance in indoor and outdoor settings. We demonstrate this across two quadrotor-event camera platforms in multiple settings and find, contrary to traditional vision-based works, that low speeds (1m/s) make the task harder and more prone to collisions, while high speeds (5m/s) result in better event-based depth estimation and avoidance. We also find that success rates in outdoor scenes can be significantly higher than in certain indoor scenes.

69

T
2
SQNet: A Recognition Model for Manipulating Partially Observed Transparent Tableware Objects
Young Hun Kim,Seungyeon Kim,Yonghyeon Lee,Frank C. Park
https://openreview.net/forum?id=M0JtsLuhEE
Recognizing and manipulating transparent tableware from partial view RGB image observations is made challenging by the difficulty in obtaining reliable depth measurements of transparent objects. In this paper we present the Transparent Tableware SuperQuadric Network (T^2SQNet), a neural network model that leverages a family of newly extended deformable superquadrics to produce low-dimensional, instance-wise and accurate 3D geometric representations of transparent objects from partial views. As a byproduct and contribution of independent interest, we also present TablewareNet, a publicly available toolset of seven parametrized shapes based on our extended deformable superquadrics, that can be used to generate new datasets of tableware objects of diverse shapes and sizes. Experiments with T^2SQNet trained with TablewareNet show that T^2SQNet outperforms existing methods in recognizing transparent objects, in some cases by significant margins, and can be effectively used in robotic applications like decluttering and target retrieval.

70

Contrast Sets for Evaluating Language-Guided Robot Policies
Abrar Anwar,Rohan Gupta,Jesse Thomason
https://openreview.net/forum?id=dXSGw7Cy55
Robot evaluations in language-guided, real world settings are time-consuming and often sample only a small space of potential instructions across complex scenes. In this work, we introduce contrast sets for robotics as an approach to make small, but specific, perturbations to otherwise independent, identically distributed (i.i.d.) test instances. We investigate the relationship between experimenter effort to carry out an evaluation and the resulting estimated test performance as well as the insights that can be drawn from performance on perturbed instances. We use contrast sets to characterize policies at reduced experimenter effort in both a simulated manipulation task and a physical robot vision-and-language navigation task. We encourage the use of contrast set evaluations as a more informative alternative to small scale, i.i.d. demonstrations on physical robots, and as a scalable alternative to industry-scale real world evaluations.

71

DeliGrasp: Inferring Object Properties with LLMs for Adaptive Grasp Policies
William Xie,Maria Valentini,Jensen Lavering,Nikolaus Correll
https://openreview.net/forum?id=rY5T2aIjPZ
Large language models (LLMs) can provide rich physical descriptions of most worldly objects, allowing robots to achieve more informed and capable grasping. We leverage LLMs' common sense physical reasoning and code-writing abilities to infer an object's physical characteristics-mass m, friction coefficient μ, and spring constant k-from a semantic description, and then translate those characteristics into an executable adaptive grasp policy. Using a two-finger gripper with a built-in depth camera that can control its torque by limiting motor current, we demonstrate that LLM-parameterized but first-principles grasp policies outperform both traditional adaptive grasp policies and direct LLM-as-code policies on a custom benchmark of 12 delicate and deformable items including food, produce, toys, and other everyday items, spanning two orders of magnitude in mass and required pick-up force. We then improve property estimation and grasp performance on variable size objects with model finetuning on property-based comparisons and eliciting such comparisons via chain-of-thought prompting. We also demonstrate how compliance feedback from DeliGrasp policies can aid in downstream tasks such as measuring produce ripeness. Our code and videos are available at: https://deligrasp.github.io

72

Learning a Distributed Hierarchical Locomotion Controller for Embodied Cooperation
Chuye Hong,Kangyao Huang,Huaping Liu
https://openreview.net/forum?id=NCnplCf4wo
In this work, we propose a distributed hierarchical locomotion control strategy for whole-body cooperation and demonstrate the potential for migration into large numbers of agents. Our method utilizes a hierarchical structure to break down complex tasks into smaller, manageable sub-tasks. By incorporating spatiotemporal continuity features, we establish the sequential logic necessary for causal inference and cooperative behaviour in sequential tasks, thereby facilitating efficient and coordinated control strategies. Through training within this framework, we demonstrate enhanced adaptability and cooperation, leading to superior performance in task completion compared to the original methods. Moreover, we construct a set of environments as the benchmark for embodied cooperation.

73

Environment Curriculum Generation via Large Language Models
William Liang,Sam Wang,Hung-Ju Wang,Osbert Bastani,Dinesh Jayaraman,Yecheng Jason Ma
https://openreview.net/forum?id=F0rWEID2gb
Recent work has demonstrated that a promising strategy for teaching robots a wide range of complex skills is by training them on a curriculum of progressively more challenging environments. However, developing an effective curriculum of environment distributions currently requires significant expertise, which must be repeated for every new domain. Our key insight is that environments are often naturally represented as code. Thus, we probe whether effective environment curriculum design can be achieved and automated via code generation by large language models (LLM). In this paper, we introduce Eurekaverse, an unsupervised environment design algorithm that uses LLMs to sample progressively more challenging, diverse, and learnable environments for skill training. We validate Eurekaverse's effectiveness in the domain of quadrupedal parkour learning, in which a quadruped robot must traverse through a variety of obstacle courses. The automatic curriculum designed by Eurekaverse enables gradual learning of complex parkour skills in simulation and can successfully transfer to the real-world, outperforming manual training courses designed by humans.

74

SHADOW: Leveraging Segmentation Masks for Cross-Embodiment Policy Transfer
Marion Lepert,Ria Doshi,Jeannette Bohg
https://openreview.net/forum?id=MyyZZAPgpy
Data collection in robotics is spread across diverse hardware, and this variation will increase as new hardware is developed. Effective use of this growing body of data requires methods capable of learning from diverse robot embodiments. We consider the setting of training a policy using expert trajectories from a single robot arm (the source), and evaluating on a different robot arm for which no data was collected (the target). We present a data editing scheme termed Shadow, in which the robot during training and evaluation is replaced with a composite segmentation mask of the source and target robots. In this way, the input data distribution at train and test time match closely, enabling robust policy transfer to the new unseen robot while being far more data efficient than approaches that require co-training on large amounts of data from diverse embodiments. We demonstrate that an approach as simple as Shadow is effective both in simulation on varying tasks and robots, and on real robot hardware, where Shadow demonstrates over 2x improvement in success rate compared to the strongest baseline.

75

GenDP: 3D Semantic Fields for Category-Level Generalizable Diffusion Policy
Yixuan Wang,Guang Yin,Binghao Huang,Tarik Kelestemur,Jiuguang Wang,Yunzhu Li
https://openreview.net/forum?id=7wMlwhCvjS
Diffusion-based policies have shown remarkable capability in executing complex robotic manipulation tasks but lack explicit characterization of geometry and semantics, which often limits their ability to generalize to unseen objects and layouts. To enhance the generalization capabilities of Diffusion Policy, we introduce a novel framework that incorporates explicit spatial and semantic information via 3D semantic fields. We generate 3D descriptor fields from multi-view RGBD observations with large foundational vision models, then compare these descriptor fields against reference descriptors to obtain semantic fields. The proposed method explicitly considers geometry and semantics, enabling strong generalization capabilities in tasks requiring category-

level generalization, resolving geometric ambiguities, and attention to subtle geometric details. We evaluate our method across eight tasks involving articulated objects and instances with varying shapes and textures from multiple object categories. Our method demonstrates its effectiveness by increasing Diffusion Policy's average success rate on \textit{unseen} instances from 20% to 93%. Additionally, we provide a detailed analysis and visualization to interpret the sources of performance gain and explain how our method can generalize to novel instances. Project page: https://robopil.github.io/GenDP/

76

View-Invariant Policy Learning via Zero-Shot Novel View Synthesis
Stephen Tian,Blake Wulfe,Kyle Sargent,Katherine Liu,Sergey Zakharov,Vitor Campagnolo Guizilini,Jiajun Wu
https://openreview.net/forum?id=tqsQGrmVEu
Large-scale visuomotor policy learning is a promising approach toward developing generalizable manipulation systems. Yet, policies that can be deployed on diverse embodiments, environments, and observational modalities remain elusive. In this work, we investigate how knowledge from large-scale visual data of the world may be used to address one axis of variation for generalizable manipulation: observational viewpoint. Specifically, we study single-image novel view synthesis models, which learn 3D-aware scene-level priors by rendering images of the same scene from alternate camera viewpoints given a single input image. For practical application to diverse robotic data, these models must operate zero-shot, performing view synthesis on unseen tasks and environments. We empirically analyze view synthesis models within a simple data-augmentation scheme that we call View Synthesis Augmentation (VISTA) to understand their capabilities for learning viewpoint-invariant policies from single-viewpoint demonstration data. Upon evaluating the robustness of policies trained with our method to out-of-distribution camera viewpoints, we find that they outperform baselines in both simulated and real-world manipulation tasks.

77

Learning Differentiable Tensegrity Dynamics using Graph Neural Networks
Nelson Chen,Kun Wang,William R. Johson III,Rebecca Kramer-Bottiglio,Kostas Bekris,Mridul Aanjaneya
https://openreview.net/forum?id=5Awumz1VKU
Tensegrity robots are composed of rigid struts and flexible cables. They constitute an emerging class of hybrid rigid-soft robotic systems and are promising systems for a wide array of applications, ranging from locomotion to assembly. They are difficult to control and model accurately, however, due to their compliance and high number of degrees of freedom. To address this issue, prior work has introduced a differentiable physics engine designed for tensegrity robots based on first principles. In contrast, this work proposes the use of graph neural networks to model contact dynamics over a graph representation of tensegrity robots, which leverages their natural graph-like cable connec- tivity between end caps of rigid rods. This learned simulator can accurately model 3-bar and 6-bar tensegrity robot dynamics in simulation-to-simulation experiments where MuJoCo is used as the ground truth. It can also achieve higher accuracy than the previous differentiable engine for a real 3-bar tensegrity robot, for which the robot state is only partially observable. When compared against direct applications of recent mesh-based graph neural network simulators, the proposed approach is computationally more efficient, both for training and inference, while achieving higher accuracy. Code and data are available at https://github.com/nchen9191/tensegrity_gnn_simulator_public

78

Sparsh: Self-supervised touch representations for vision-based tactile sensing
Carolina Higuera,Akash Sharma,Chaithanya Krishna Bodduluri,Taosha Fan,Patrick Lancaster,Mrinal Kalakrishnan,Michael Kaess,Byron Boots,Mike Lambeta,Tingfan Wu,Mustafa Mukadam
https://openreview.net/forum?id=xYJn2e1uu8
In this work, we introduce general purpose touch representations for the increasingly accessible class of vision-based tactile sensors. Such sensors have led to many recent advances in robot manipulation as they markedly complement vision, yet solutions today often rely on task and sensor specific handcrafted perception models. Collecting real data at scale with task centric ground truth labels, like contact forces and slip, is a challenge further compounded by sensors of various form factor differing in aspects like lighting and gel markings. To tackle this, we turn to self-supervised learning (SSL) that has demonstrated remarkable performance in computer vision. We present Sparsh, a family of SSL models that can support various vision-based tactile sensors, alleviating the need for custom labels through pre-training on 460k+ tactile images with masking and self-distillation in pixel and latent spaces. We also build TacBench, to facilitate standardized benchmarking across sensors and models, comprising of six tasks ranging from comprehending tactile properties to enabling physical perception and manipulation planning. In evaluations, we find that SSL pre-training for touch representation outperforms task and sensor-specific end-to-end training by 95.1% on average over TacBench, and Sparsh (DINO) and Sparsh (IJEPA) are the most competitive, indicating the merits of learning in latent space for tactile images. Project page: https://sparsh-ssl.github.io

79

Harmon: Whole-Body Motion Generation of Humanoid Robots from Language Descriptions
Zhenyu Jiang,Yuqi Xie,Jinhan Li,Ye Yuan,Yifeng Zhu,Yuke Zhu
https://openreview.net/forum?id=UUZ4Yw3lt0
Humanoid robots, with their human-like embodiment, have the potential to integrate seamlessly into human environments. Critical to their coexistence and cooperation with humans is the ability to understand natural language communications and exhibit human-like behaviors. This work focuses on generating diverse whole-body motions for humanoid robots from language descriptions. We leverage human motion priors from extensive human motion datasets to initialize humanoid motions and employ the commonsense reasoning capabilities of Vision Language Models (VLMs) to edit and refine these motions. Our approach demonstrates the capability to produce natural, expressive, and text-aligned humanoid motions, validated through both simulated and real-world experiments. More videos can be found on our website https://ut-austin-rpl.github.io/Harmon/.

80

Toward General Object-level Mapping from Sparse Views with 3D Diffusion Priors
Ziwei Liao,Binbin Xu,Steven L. Waslander
https://openreview.net/forum?id=rEteJcq61j
Object-level mapping builds a 3D map of objects in a scene with detailed shapes and poses from multi-view sensor observations. Conventional methods struggle to build complete shapes and estimate accurate poses due to partial occlusions and sensor noise. They require dense observations to cover all objects, which is challenging to achieve in robotics trajectories. Recent work introduces generative shape priors for object-level mapping from sparse views, but is limited to single-category objects. In this work, we propose a General Object-level Mapping system,

GOM, which leverages a 3D diffusion model as shape prior with multi-category support and outputs Neural Radiance Fields (NeRFs) for both texture and geometry for all objects in a scene. GOM includes an effective formulation to guide a pre-trained diffusion model with extra nonlinear constraints from sensor measurements without finetuning. We also develop a probabilistic optimization formulation to fuse multi-view sensor observations and diffusion priors for joint 3D object pose and shape estimation. Our GOM system demonstrates superior multi-category mapping performance from sparse views, and achieves more accurate mapping results compared to state-of-the-art methods on the real-world benchmarks. We will release our code and model upon publication.

81

Scaling Safe Multi-Agent Control for Signal Temporal Logic Specifications
Joe Eappen,Zikang Xiong,Dipam Patel,Aniket Bera,Suresh Jagannathan
https://openreview.net/forum?id=N1K4B8N3n1
Existing methods for safe multi-agent control using logic specifications like Signal Temporal Logic (STL) often face scalability issues. This is because they rely either on single-agent perspectives or on Mixed Integer Linear Programming (MILP)-based planners, which are complex to optimize. These methods have proven to be computationally expensive and inefficient when dealing with a large number of agents. To address these limitations, we present a new scalable approach to multi-agent control in this setting. Our method treats the relationships between agents using a graph structure rather than in terms of a single-agent perspective. Moreover, it combines a multi-agent collision avoidance controller with a Graph Neural Network (GNN) based planner, models the system in a decentralized fashion, and trains on STL-based objectives to generate safe and efficient plans for multiple agents, thereby optimizing the satisfaction of complex temporal specifications while also facilitating multi-agent collision avoidance. Our experiments show that our approach significantly outperforms existing methods that use a state-of-the-art MILP-based planner in terms of scalability and performance.

82

Sparse Diffusion Policy: A Sparse, Reusable, and Flexible Policy for Robot Learning
Yixiao Wang,Yifei Zhang,Mingxiao Huo,Thomas Tian,Xiang Zhang,Yichen Xie,Chenfeng Xu,Pengliang Ji,Wei Zhan,Mingyu Ding,Masayoshi Tomizuka
https://openreview.net/forum?id=zeYaLS2tw5
The increasing complexity of tasks in robotics demands efficient strategies for multitask and continual learning. Traditional models typically rely on a universal policy for all tasks, facing challenges such as high computational costs and catastrophic forgetting when learning new tasks. To address these issues, we introduce a sparse, reusable, and flexible policy, Sparse Diffusion Policy (SDP). By adopting Mixture of Experts (MoE) within a transformer-based diffusion policy, SDP selectively activates experts and skills, enabling task-specific learning without retraining the entire model. It not only reduces the burden of active parameters but also facilitates the seamless integration and reuse of experts across various tasks. Extensive experiments on diverse tasks in both simulators and the real world show that SDP 1) excels in multitask scenarios with negligible increases in active parameters, 2) prevents forgetting in continual learning new tasks, and 3) enables efficient task transfer, offering a promising solution for advanced robotic applications. More demos and codes can be found on our https://anonymous.4open.science/w/sparse_diffusion_policy-24E7/.

83

What Matters in Range View 3D Object Detection
Benjamin Wilson,Nicholas Autio Mitchell,Jhony Kaesemodel Pontes,James Hays
https://openreview.net/forum?id=EifoVoIyd5
Lidar-based perception pipelines rely on 3D object detection models to interpret complex scenes. While multiple representations for lidar exist, the range view is enticing since it losslessly encodes the entire lidar sensor output. In this work, we achieve state-of-the-art amongst range view 3D object detection models without using multiple techniques proposed in past range view literature. We explore range view 3D object detection across two modern datasets with substantially different properties: Argoverse 2 and Waymo Open. Our investigation reveals key insights: (1) input feature dimensionality significantly influences the overall performance, (2) surprisingly, employing a classification loss grounded in 3D spatial proximity works as well or better compared to more elaborate IoU-based losses, and (3) addressing non-uniform lidar density via a straightforward range subsampling technique outperforms existing multi-resolution, range-conditioned networks. Our experiments reveal that techniques proposed in recent range view literature are not needed to achieve state-of-the-art performance. Combining the above findings, we establish a new state-of-the-art model for range view 3D object detection — improving AP by 2.2% on the Waymo Open dataset while maintaining a runtime of 10 Hz. We are the first to benchmark a range view model on the Argoverse 2 dataset and outperform strong voxel-based baselines. All models are multi-class and open-source. Code is available at https://github.com/benjaminrwilson/range-view-3d-detection.

84

MOSAIC: Modular Foundation Models for Assistive and Interactive Cooking
Huaxiaoyue Wang,Kushal Kedia,Juntao Ren,Rahma Abdullah,Atiksh Bhardwaj,Angela Chao,Kelly Y Chen,Nathaniel Chin,Prithwish Dan,Xinyi Fan,Gonzalo Gonzalez-Pumariega,Aditya Kompella,Maximus Adrian Pace,Yash Sharma,Xiangwan Sun,Neha Sunkara,Sanjiban Choudhury
https://openreview.net/forum?id=dUo6j3YURS
We present MOSAIC, a modular architecture for coordinating multiple robots to (a) interact with users using natural language and (b) manipulate an open vocabulary of everyday objects. At several levels, MOSAIC employs modularity: it leverages multiple large-scale pre-trained models for high-level tasks like language and image recognition, while using streamlined modules designed for low-level task-specific control. This decomposition allows us to reap the complementary benefits of foundation models and precise, more specialized models, enabling our system to scale to complex tasks that involve coordinating multiple robots and humans. First, we unit-test individual modules with 180 episodes of visuomotor picking, 60 episodes of human motion forecasting, and 46 online user evaluations of the task planner. We then extensively evaluate MOSAIC with 60 end-to-end trials. We discuss crucial design decisions, limitations of the current system, and open challenges in this domain

85

Gentle Manipulation of Tree Branches: A Contact-Aware Policy Learning Approach
Jay Jacob,Shizhe Cai,Paulo Vinicius Koerich Borges,Tirthankar Bandyopadhyay,Fabio Ramos
https://openreview.net/forum?id=zr2GPi3DSb
Learning to interact with deformable tree branches with minimal damage is challenging due to their intricate geometry and inscrutable dynamics. Furthermore, traditional vision-based modelling systems suffer from implicit occlusions in dense foliage, severely changing lighting conditions,

and limited field of view, in addition to having a significant computational burden preventing real-time deployment.In this work, we simulate a procedural forest with realistic, self-similar branching structures derived from a parametric L-system model, actuated with crude spring abstractions, mirroring real-world variations with domain randomisation over the morphological and dynamic attributes. We then train a novel Proprioceptive Contact-Aware Policy (PCAP) for a reach task using reinforcement learning, aided by a whole-arm contact detection classifier and reward engineering, without external vision, tactile, or torque sensing. The agent deploys novel strategies to evade and mitigate contact impact, favouring a reactive exploration of the task space. Finally, we demonstrate that the learned behavioural patterns can be transferred zero-shot from simulation to real, allowing the arm to navigate around real branches with unseen topology and variable occlusions while minimising the contact forces and expected ruptures.

86

Not All Errors Are Made Equal: A Regret Metric for Detecting System-level Trajectory Prediction Failures
Kensuke Nakamura,Thomas Tian,Andrea Bajcsy
https://openreview.net/forum?id=G0jqGG8Tta
Robot decision-making increasingly relies on data-driven human prediction models when operating around people. While these models are known to mispredict in out-of-distribution interactions, only a subset of prediction errors impact downstream robot performance.
We propose characterizing such ``system-level'' prediction failures via the mathematical notion of regret: high-regret interactions are precisely those in which mispredictions degraded closed-loop robot performance. We further introduce a probabilistic generalization of regret that calibrates failure detection across disparate deployment contexts and renders regret compatible with reward-based and reward-free (e.g., generative) planners.
In simulated autonomous driving interactions, we showcase that our system-level failure metric can automatically mine for closed-loop human-robot interactions that state-of-the-art generative human predictors and robot planners struggle with. We further find that the very presence of high-regret data during human predictor fine-tuning is highly predictive of robot re-deployment performance improvements. Furthermore, fine-tuning with the informative but significantly smaller high-regret data (23% of deployment data) is competitive with fine-tuning on the full deployment dataset, indicating a promising avenue for efficiently mitigating system-level human-robot interaction failures.

87

EquiBot: SIM(3)-Equivariant Diffusion Policy for Generalizable and Data Efficient Learning
Jingyun Yang,Ziang Cao,Congyue Deng,Rika Antonova,Shuran Song,Jeannette Bohg
https://openreview.net/forum?id=ueBmGhLOXP
Building effective imitation learning methods that enable robots to learn from limited data and still generalize across diverse real-world environments is a long-standing problem in robot learning. We propose EquiBot, a robust, data-efficient, and generalizable approach for robot manipulation task learning. Our approach combines SIM(3)-equivariant neural network architectures with diffusion models. This ensures that our learned policies are invariant to changes in scale, rotation, and translation, enhancing their applicability to unseen environments while retaining the benefits of diffusion-based policy learning such as multi-modality and robustness. We show on a suite of 6 simulation tasks that our proposed method reduces the data requirements and improves generalization to novel scenarios. In the real world, with 10 variations of 6 mobile manipulation

tasks, we show that our method can easily generalize to novel objects and scenes after learning from just 5 minutes of human demonstrations in each task.

88

ThinkGrasp: A Vision-Language System for Strategic Part Grasping in Clutter
Yaoyao Qian,Xupeng Zhu,Ondrej Biza,Shuo Jiang,Linfeng Zhao,Haojie Huang,Yu Qi,Robert Platt
https://openreview.net/forum?id=MsCbbIqHRA
Robotic grasping in cluttered environments remains a significant challenge due to occlusions and complex object arrangements. We have developed ThinkGrasp, a plug-and-play vision-language grasping system that makes use of GPT-4o's advanced contextual reasoning for grasping strategies. ThinkGrasp can effectively identify and generate grasp poses for target objects, even when they are heavily obstructed or nearly invisible, by using goal-oriented language to guide the removal of obstructing objects. This approach progressively uncovers the target object and ultimately grasps it with a few steps and a high success rate. In both simulated and real experiments, ThinkGrasp achieved a high success rate and significantly outperformed state-of-the-art methods in heavily cluttered environments or with diverse unseen objects, demonstrating strong generalization capabilities.

89

Perceive With Confidence: Statistical Safety Assurances for Navigation with Learning-Based Perception
Anushri Dixit,Zhiting Mei,Meghan Booker,Mariko Storey-Matsutani,Allen Z. Ren,Anirudha Majumdar
https://openreview.net/forum?id=cDXnnOhNrF
Rapid advances in perception have enabled large pre-trained models to be used out of the box for transforming high-dimensional, noisy, and partial observations of the world into rich occupancy representations. However, the reliability of these models and consequently their safe integration onto robots remains unknown when deployed in environments unseen during training. In this work, we address this challenge by rigorously quantifying the uncertainty of pre-trained perception systems for object detection via a novel calibration technique based on conformal prediction. Crucially, this procedure guarantees robustness to distribution shifts in states when perceptual outputs are used in conjunction with a planner. As a result, the calibrated perception system can be used in combination with any safe planner to provide an end-to-end statistical assurance on safety in unseen environments. We evaluate the resulting approach, Perceive with Confidence (PwC), with experiments in simulation and on hardware where a quadruped robot navigates through previously unseen indoor, static environments. These experiments validate the safety assurances for obstacle avoidance provided by PwC and demonstrate up to 40% improvements in empirical safety compared to baselines.

90

Enhancing Visual Domain Robustness in Behaviour Cloning via Saliency-Guided Augmentation
Zheyu Zhuang,RUIYU WANG,Nils Ingelhag,Ville Kyrki,Danica Kragic
https://openreview.net/forum?id=CskuWHDBAr
In vision-based behaviour cloning (BC), traditional image-level augmentation methods such as pixel shifting enhance in-domain performance but often struggle with visual domain shifts, including distractors, occlusion, and changes in lighting and backgrounds. Conversely, superimposition-based augmentation, proven effective in computer vision, improves model

generalisability by blending training images and out-of-domain images. Despite its potential, the applicability of these methods to vision-based BC remains unclear due to the unique challenges posed by BC demonstrations; specifically, preserving task-critical scene semantics, spatial-temporal relationships, and agent-target interactions is crucial. To address this, we introduce RoboSaGA, a context-aware approach that dynamically adjusts augmentation intensity per pixel based on input saliency derived from the policy. This method ensures aggressive augmentation within task-trivial areas without compromising task-critical information. Furthermore, RoboSaGA seamlessly integrates into existing network architectures without the need for structural changes or additional learning objectives. Our empirical evaluations across both simulated and real-world settings demonstrate that RoboSaGA not only maintains in-domain performance but significantly improves resilience to distractors and background variations.

91

Discovering Robotic Interaction Modes with Discrete Representation Learning
Liquan Wang,Ankit Goyal,Haoping Xu,Animesh Garg
https://openreview.net/forum?id=xcBH8Jhmbi
Abstract: Human actions manipulating articulated objects, such as opening and closing a drawer, can be categorized into multiple modalities we define as interaction modes. Traditional robot learning approaches lack discrete representations of these modes, which are crucial for empirical sampling and grounding. In this paper, we present ActAIM2, which learns a discrete representation of robot manipulation interaction modes in a purely unsupervised fashion, without the use of expert labels or simulator-based privileged information. Utilizing novel data collection methods involving simulator rollouts, ActAIM2 consists of an interaction mode selector and a low-level action predictor. The selector generates discrete representations of potential interaction modes with self-supervision, while the predictor outputs corresponding action trajectories. Our method is validated through its success rate in manipulating articulated objects and its robustness in sampling meaningful actions from the discrete representation. Extensive experiments demonstrate ActAIM2's effectiveness in enhancing manipulability and generalizability over baselines and ablation studies. For videos and additional results, see our website: https://actaim2.github.io/.

92

RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation
Yuxuan Kuang,Junjie Ye,Haoran Geng,Jiageng Mao,Congyue Deng,Leonidas Guibas,He Wang,Yue Wang
https://openreview.net/forum?id=8LPXeGhhbH
This work proposes a retrieve-and-transfer framework for zero-shot robotic manipulation, dubbed RAM, featuring generalizability across various objects, environments, and embodiments. Unlike existing approaches that learn manipulation from expensive in-domain demonstrations, RAM capitalizes on a retrieval-based affordance transfer paradigm to acquire versatile manipulation capabilities from abundant out-of-domain data. RAM first extracts unified affordance at scale from diverse sources of demonstrations including robotic data, human-object interaction (HOI) data, and custom data to construct a comprehensive affordance memory. Then given a language instruction, RAM hierarchically retrieves the most similar demonstration from the affordance memory and transfers such out-of-domain 2D affordance to in-domain 3D actionable affordance in a zero-shot and embodiment-agnostic manner. Extensive simulation and real-world evaluations demonstrate that our RAM consistently outperforms existing works in diverse daily tasks. Additionally, RAM shows significant potential for downstream applications such as automatic and

efficient data collection, one-shot visual imitation, and LLM/VLM-integrated long-horizon manipulation.

93

Bridging the Sim-to-Real Gap from the Information Bottleneck Perspective
Haoran He,Peilin Wu,Chenjia Bai,Hang Lai,Lingxiao Wang,Ling Pan,Xiaolin Hu,Weinan Zhang
https://openreview.net/forum?id=Bq4XOaU4sV
Reinforcement Learning (RL) has recently achieved remarkable success in robotic control. However, most works in RL operate in simulated environments where privileged knowledge (e.g., dynamics, surroundings, terrains) is readily available. Conversely, in real-world scenarios, robot agents usually rely solely on local states (e.g., proprioceptive feedback of robot joints) to select actions, leading to a significant sim-to-real gap. Existing methods address this gap by either gradually reducing the reliance on privileged knowledge or performing a two-stage policy imitation. However, we argue that these methods are limited in their ability to fully leverage the available privileged knowledge, resulting in suboptimal performance. In this paper, we formulate the sim-to-real gap as an information bottleneck problem and therefore propose a novel privileged knowledge distillation method called the Historical Information Bottleneck (HIB). In particular, HIB learns a privileged knowledge representation from historical trajectories by capturing the underlying changeable dynamic information. Theoretical analysis shows that the learned privileged knowledge representation helps reduce the value discrepancy between the oracle and learned policies. Empirical experiments on both simulated and real-world tasks demonstrate that HIB yields improved generalizability compared to previous methods.

94

3D Diffuser Actor: Policy Diffusion with 3D Scene Representations
Tsung-Wei Ke,Nikolaos Gkanatsios,Katerina Fragkiadaki
https://openreview.net/forum?id=gqCQxObVz2
Diffusion policies are conditional diffusion models that learn robot action distributions conditioned on the robot and environment state. They have recently shown to outperform both deterministic and alternative action distribution learning formulations. 3D robot policies use 3D scene feature representations aggregated from a single or multiple camera views using sensed depth. They have shown to generalize better than their 2D counterparts across camera viewpoints. We unify these two lines of work and present 3D Diffuser Actor , a neural policy equipped with a novel 3D denoising transformer that fuses information from the 3D visual scene, a language instruction and proprioception to predict the noise in noised 3D robot pose trajectories. 3D Diffuser Actor sets a new state-of-the-art on RLBench with an absolute performance gain of 18.1% over the current SOTA on a multi-view setup and an absolute gain of 13.1% on a single-view setup. On the CALVIN benchmark, it improves over the current SOTA by a 9% relative increase. It also learns to control a robot manipulator in the real world from a handful of demonstrations. Through thorough comparisons with the current SOTA policies and ablations of our model, we show 3D Diffuser Actor ’s design choices dramatically outperform 2D representations, regression and classification objectives, absolute attentions, and holistic non-tokenized 3D scene embeddings.

95

IMAGINATION POLICY: Using Generative Point Cloud Models for Learning Manipulation Policies
Haojie Huang,Karl Schmeckpeper,Dian Wang,Ondrej Biza,Yaoyao Qian,Haotian Liu,Mingxi Jia,Robert Platt,Robin Walters
https://openreview.net/forum?id=56IzghzjfZ
Humans can imagine goal states during planning and perform actions to match those goals. In this work, we propose IMAGINATION POLICY, a novel multi-task key-frame policy network for solving high-precision pick and place tasks. Instead of learning actions directly, IMAGINATION POLICY generates point clouds to imagine desired states which are then translated to actions using rigid action estimation. This transforms action inference into a local generative task. We leverage pick and place symmetries underlying the tasks in the generation process and achieve extremely high sample efficiency and generalizability to unseen configurations. Finally, we demonstrate state-of-the-art performance across various tasks on the RLbench benchmark compared with several strong baselines and validate our approach on a real robot.

96

UBSoft: A Simulation Platform for Robotic Skill Learning in Unbounded Soft Environments
Chunru Lin,Jugang Fan,Yian Wang,Zeyuan Yang,Zhehuan Chen,Lixing Fang,Tsun-Hsuan Wang,Zhou Xian,Chuang Gan
https://openreview.net/forum?id=7vzDBvviRO
It is desired to equip robots with the capability of interacting with various soft materials as they are ubiquitous in the real world. While physics simulations are one of the predominant methods for data collection and robot training, simulating soft materials presents considerable challenges. Specifically, it is significantly more costly than simulating rigid objects in terms of simulation speed and storage requirements. These limitations typically restrict the scope of studies on soft materials to small and bounded areas, thereby hindering the learning of skills in broader spaces. To address this issue, we introduce UBSoft, a new simulation platform designed to support unbounded soft environments for robot skill acquisition. Our platform utilizes spatially adaptive resolution scales, where simulation resolution dynamically adjusts based on proximity to active robotic agents. Our framework markedly reduces the demand for extensive storage space and computation costs required for large-scale scenarios involving soft materials. We also establish a set of benchmark tasks in our platform, including both locomotion and manipulation tasks, and conduct experiments to evaluate the efficacy of various reinforcement learning algorithms and trajectory optimization techniques, both gradient-based and sampling-based. Preliminary results indicate that sampling-based trajectory optimization generally achieves better results for obtaining one trajectory to solve the task. Additionally, we conduct experiments in real-world environments to demonstrate that advancements made in our UBSoft simulator could translate to improved robot interactions with large-scale soft material. More videos can be found at https://ubsoft24.github.io.

97

ALOHA Unleashed: A Simple Recipe for Robot Dexterity
Tony Z. Zhao,Jonathan Tompson,Danny Driess,Pete Florence,Seyed Kamyar Seyed Ghasemipour,Chelsea Finn,Ayzaan Wahid
https://openreview.net/forum?id=gvdXE7ikHI
Recent work has shown promising results for learning end-to-end robot policies using imitation learning. In this work we address the question of how far can we push imitation learning for

challenging dexterous manipulation tasks. We show that a simple recipe of large scale data collection on the ALOHA 2 platform, combined with expressive models such as Diffusion Policies, can be effective in learning challenging bimanual manipulation tasks involving deformable objects and complex contact rich dynamics. We demonstrate our recipe on 5 challenging real-world and 3 simulated tasks and demonstrate improved performance over state-of-the-art baselines.

98

D
3
Fields: Dynamic 3D Descriptor Fields for Zero-Shot Generalizable Rearrangement
Yixuan Wang,Mingtong Zhang,Zhuoran Li,Tarik Kelestemur,Katherine Rose Driggs-Campbell,Jiajun Wu,Li Fei-Fei,Yunzhu Li
https://openreview.net/forum?id=Uaaj4MaVIQ
Scene representation is a crucial design choice in robotic manipulation systems. An ideal representation is expected to be 3D, dynamic, and semantic to meet the demands of diverse manipulation tasks. However, previous works often lack all three properties simultaneously. In this work, we introduce D^3Fields---dynamic 3D descriptor fields. These fields are implicit 3D representations that take in 3D points and output semantic features and instance masks. They can also capture the dynamics of the underlying 3D environments. Specifically, we project arbitrary 3D points in the workspace onto multi-view 2D visual observations and interpolate features derived from visual foundational models. The resulting fused descriptor fields allow for flexible goal specifications using 2D images with varied contexts, styles, and instances. To evaluate the effectiveness of these descriptor fields, we apply our representation to rearrangement tasks in a zero-shot manner. Through extensive evaluation in real worlds and simulations, we demonstrate that D^3Fields are effective for zero-shot generalizable rearrangement tasks. We also compare D^3 Fields with state-of-the-art implicit 3D representations and show significant improvements in effectiveness and efficiency. Project page: https://robopil.github.io/d3fields/

99

KOROL: Learning Visualizable Object Feature with Koopman Operator Rollout for Manipulation
Hongyi Chen,ABULIKEMU ABUDUWEILI,Aviral Agrawal,Yunhai Han,Harish Ravichandar,Changliu Liu,Jeffrey Ichnowski
https://openreview.net/forum?id=A6ikGJRaKL
Learning dexterous manipulation skills presents significant challenges due to complex nonlinear dynamics that underlie the interactions between objects and multi-fingered hands. Koopman operators have emerged as a robust method for modeling such nonlinear dynamics within a linear framework. However, current methods rely on runtime access to ground-truth (GT) object states, making them unsuitable for vision-based practical applications. Unlike image-to-action policies that implicitly learn visual features for control, we use a dynamics model, specifically the Koopman operator, to learn visually interpretable object features critical for robotic manipulation within a scene. We construct a Koopman operator using object features predicted by a feature extractor and utilize it to auto-regressively advance system states. We train the feature extractor to embed scene information into object features, thereby enabling the accurate propagation of robot trajectories. We evaluate our approach on simulated and real-world robot tasks, with results showing that it outperformed the model-based imitation learning NDP by 1.08× and the image-to-action Diffusion Policy by 1.16×. The results suggest that our method maintains task success rates with learned features and extends applicability to real-world manipulation without GT object states. Project video and code are available at: https://github.com/hychen-naza/KOROL.

100

Mobility VLA: Multimodal Instruction Navigation with Long-Context VLMs and Topological Graphs
Zhuo Xu,Hao-Tien Lewis Chiang,Zipeng Fu,Mithun George Jacob,Tingnan Zhang,Tsang-Wei Edward Lee,Wenhao Yu,Connor Schenck,David Rendleman,Dhruv Shah,Fei Xia,Jasmine Hsu,Jonathan Hoech,Pete Florence,Sean Kirmani,Sumeet Singh,Vikas Sindhwani,Carolina Parada,Chelsea Finn,Peng Xu,et al. (2 additional authors not shown)
https://openreview.net/forum?id=JScswMfEQ0
An elusive goal in navigation research is to build an intelligent agent that can understand multimodal instructions including natural language and image, and perform useful navigation. To achieve this, we study a widely useful category of navigation tasks we call Multimodal Instruction Navigation with demonstration Tours (MINT), in which the environment prior is provided through a previously recorded demonstration video. Recent advances in Vision Language Models (VLMs) have shown a promising path in achieving this goal as it demonstrates capabilities in perceiving and reasoning about multimodal inputs. However, VLMs are typically trained to predict textual output and it is an open research question about how to best utilize them in navigation. To solve MINT, we present Mobility VLA, a hierarchical Vision-Language-Action (VLA) navigation policy that combines the environment understanding and common sense reasoning power of long-context VLMs and a robust low-level navigation policy based on topological graphs. The high-level policy consists of a long-context VLM that takes the demonstration tour video and the multimodal user instruction as input to find the goal frame in the tour video. Next, a low-level policy uses the goal frame and an offline constructed topological graph to generate robot actions at every timestep. We evaluated Mobility VLA in a 836m^2 real world environment and show that Mobility VLA has a high end-to-end success rates on previously unsolved multimodal instructions such as ``Where should I return this?'' while holding a plastic bin.

101

Reasoning Grasping via Multimodal Large Language Model
Shiyu Jin,JINXUAN XU,Yutian Lei,Liangjun Zhang
https://openreview.net/forum?id=KPcX4jetMw
Despite significant progress in robotic systems for operation within human-centric environments, existing models still heavily rely on explicit human commands to identify and manipulate specific objects. This limits their effectiveness in environments where understanding and acting on implicit human intentions are crucial. In this study, we introduce a novel task: reasoning grasping, where robots need to generate grasp poses based on indirect verbal instructions or intentions. To accomplish this, we propose an end-to-end reasoning grasping model that integrates a multimodal Large Language Model (LLM) with a vision-based robotic grasping framework. In addition, we present the first reasoning grasping benchmark dataset generated from the GraspNet-1 billion, incorporating implicit instructions for object-level and part-level grasping, and this dataset will soon be available for public access. Our results show that directly integrating CLIP or LLaVA with the grasp detection model performs poorly on the challenging reasoning grasping tasks, while our proposed model demonstrates significantly enhanced performance both in the reasoning grasping benchmark and real-world experiments.

102

Learning Long-Horizon Action Dependencies in Sampling-Based Bilevel Planning
Bartłomiej Cieślar,Leslie Pack Kaelbling,Tomás Lozano-Pérez,Jorge Mendez-Mendez
https://openreview.net/forum?id=DsFQg0G4Xu
Autonomous robots will need the ability to make task and motion plans that involve long sequences of actions, e.g. to prepare a meal. One challenge is that the feasibility of actions late in the plan may depend on much earlier actions. This issue is exacerbated if these dependencies exist at a purely geometric level, making them difficult to express for a task planner. Backtracking is a common technique to resolve such geometric dependencies, but its time complexity limits its applicability to short-horizon dependencies. We propose an approach to account for these dependencies by learning a search heuristic for task and motion planning. We evaluate our approach on five quasi-static simulated domains and show a substantial improvement in success rate over the baselines.

103

OrbitGrasp: SE(3)-Equivariant Grasp Learning
Boce Hu,Xupeng Zhu,Dian Wang,Zihao Dong,Haojie Huang,Chenghao Wang,Robin Walters,Robert Platt
https://openreview.net/forum?id=clqzoCrulY
While grasp detection is an important part of any robotic manipulation pipeline, reliable and accurate grasp detection in $\mathrm{SE}(3)$ remains a research challenge. Many robotics applications in unstructured environments such as the home or warehouse would benefit a lot from better grasp performance. This paper proposes a novel framework for detecting $\mathrm{SE}(3)$ grasp poses based on point cloud input. Our main contribution is to propose an $\mathrm{SE}(3)$-equivariant model that maps each point in the cloud to a continuous grasp quality function over the 2-sphere S^2 using a spherical harmonic basis. Compared with reasoning about a finite set of samples, this formulation improves the accuracy and efficiency of our model when a large number of samples would otherwise be needed. In order to accomplish this, we propose a novel variation on EquiFormerV2 that leverages a UNet-style backbone to enlarge the number of points the model can handle. Our resulting method, which we name OrbitGrasp, significantly outperforms baselines in both simulation and physical experiments.

104

NOD-TAMP: Generalizable Long-Horizon Planning with Neural Object Descriptors
Shuo Cheng,Caelan Reed Garrett,Ajay Mandlekar,Danfei Xu
https://openreview.net/forum?id=rThtgkXuvZ
Solving complex manipulation tasks in household and factory settings remains challenging due to long-horizon reasoning, fine-grained interactions, and broad object and scene diversity. Learning skills from demonstrations can be an effective strategy, but such methods often have limited generalizability beyond training data and struggle to solve long-horizon tasks. To overcome this, we propose to synergistically combine two paradigms: Neural Object Descriptors (NODs) that produce generalizable object-centric features and Task and Motion Planning (TAMP) frameworks that chain short-horizon skills to solve multi-step tasks. We introduce NOD-TAMP, a TAMP-based framework that extracts short manipulation trajectories from a handful of human demonstrations, adapts these trajectories using NOD features, and composes them to solve broad long-horizon, contact-rich tasks. NOD-TAMP solves existing manipulation benchmarks with a handful of demonstrations and significantly outperforms prior NOD-based approaches on new tabletop

manipulation tasks that require diverse generalization. Finally, we deploy NOD-TAMP on a number of real-world tasks, including tool-use and high-precision insertion. For more details, please visit https://nodtamp.github.io/.

105

Cloth-Splatting: 3D Cloth State Estimation from RGB Supervision
Alberta Longhini,Marcel Büsching,Bardienus Pieter Duisterhof,Jens Lundell,Jeffrey Ichnowski,Mårten Björkman,Danica Kragic
https://openreview.net/forum?id=WmWbswjTsi
We introduce Cloth-Splatting, a method for estimating 3D states of cloth from RGB images through a prediction-update framework. Cloth-Splatting leverages an action-conditioned dynamics model for predicting future states and uses 3D Gaussian Splatting to update the predicted states. Our key insight is that coupling a 3D mesh-based representation with Gaussian Splatting allows us to define a differentiable map between the cloth's state space and the image space. This enables the use of gradient-based optimization techniques to refine inaccurate state estimates using only RGB supervision. Our experiments demonstrate that Cloth-Splatting not only improves state estimation accuracy over current baselines but also reduces convergence time by ~ 85 %.

106

Region-aware Grasp Framework with Normalized Grasp Space for Efficient 6-DoF Grasping
Siang Chen,Pengwei Xie,Wei Tang,Dingchang Hu,Yixiang Dai,Guijin Wang
https://openreview.net/forum?id=jPkOFAiOzf
A series of region-based methods succeed in extracting regional features and enhancing grasp detection quality. However, faced with a cluttered scene with potential collision, the definition of the grasp-relevant region stays inconsistent. In this paper, we propose Normalized Grasp Space (NGS) from a novel region-aware viewpoint, unifying the grasp representation within a normalized regional space and benefiting the generalizability of methods. Leveraging the NGS, we find that CNNs are underestimated for 3D feature extraction and 6-DoF grasp detection in clutter scenes and build a highly efficient Region-aware Normalized Grasp Network (RNGNet). Experiments on the public benchmark show that our method achieves significant >20 % performance gains while attaining a real-time inference speed of approximately 50 FPS. Real-world cluttered scene clearance experiments underscore the effectiveness of our method. Further, human-to-robot handover and dynamic object grasping experiments demonstrate the potential of our proposed method for closed-loop grasping in dynamic scenarios.

107

Neural Inverse Source Problem
Youngsun Wi,Jayjun Lee,Miquel Oller,Nima Fazeli
https://openreview.net/forum?id=BmvUg1FIWC
Reconstructing unknown external source functions is an important perception capability for a large range of robotics domains including manipulation, aerial, and underwater robotics. In this work, we propose a Physics-Informed Neural Network (PINN) based approach for solving the inverse source problems in robotics, jointly identifying unknown source functions and the complete state of a system given partial and noisy observations. Our approach demonstrates several advantages over prior works (Finite Element Methods (FEM) and data-driven approaches): it offers flexibility in integrating diverse constraints and boundary conditions; eliminates the need

for complex discretizations (e.g., meshing); easily accommodates gradients from real measurements; and does not limit performance based on the diversity and quality of training data. We validate our method across three simulation and real-world scenarios involving up to 4th order partial differential equations (PDEs), constraints such as Signorini and Dirichlet, and various regression losses including Chamfer distance and L2 norm.

108

Adaptive Language-Guided Abstraction from Contrastive Explanations
Andi Peng,Belinda Z. Li,Ilia Sucholutsky,Nishanth Kumar,Julie Shah,Jacob Andreas,Andreea Bobu
https://openreview.net/forum?id=OGjGtN6hoo
Many approaches to robot learning begin by inferring a reward function from a set of human demonstrations. To learn a good reward, it is necessary to determine which features of the environment are relevant before determining how these features should be used to compute reward. In particularly complex, high-dimensional environments, human demonstrators often struggle to fully specify their desired behavior from a small number of demonstrations. End-to-end reward learning methods (e.g., using deep networks or program synthesis techniques) often yield brittle reward functions that are sensitive to spurious state features. By contrast, humans can often generalizably learn from a small number of demonstrations by incorporating strong priors about what features of a demonstration are likely meaningful for a task of interest. How do we build robots that leverage this kind of background knowledge when learning from new demonstrations? This paper describes a method named ALGAE which alternates between using language models to iteratively identify human-meaningful features needed to explain demonstrated behavior, then standard inverse reinforcement learning techniques to assign weights to these features. Experiments across a variety of both simulated and real-world robot environments show that ALGAElearns generalizable reward functions defined on interpretable features using only small numbers of demonstrations. Importantly, ALGAE can recognize when features are missing, then extract and define those features without any human input -- making it possible to quickly and efficiently acquire rich representations of user behavior.

109

Dreaming to Assist: Learning to Align with Human Objectives for Shared Control in High-Speed Racing
Jonathan DeCastro,Andrew Silva,Deepak Gopinath,Emily Sumner,Thomas Matrai Balch,Laporsha Dees,Guy Rosman
https://openreview.net/forum?id=adf3pO9baG
Tight coordination is required for effective human-robot teams in domains involving fast dynamics and tactical decisions, such as multi-car racing. In such settings, robot teammates must react to cues of a human teammate's tactical objective to assist in a way that is consistent with the objective (e.g., navigating left or right around an obstacle). To address this challenge, we present Dream2Assist, a framework that combines a rich world model able to infer human objectives and value functions, and an assistive agent that provides appropriate expert assistance to a given human teammate. Our approach builds on a recurrent state space model to explicitly infer human intents, enabling the assistive agent to select actions that align with the human and enabling a fluid teaming interaction. We demonstrate our approach in a high-speed racing domain with a population of synthetic human drivers pursuing mutually exclusive objectives, such as "stay-behind" and "overtake". We show that the combined human-robot team, when blending its actions with those of the human, outperforms synthetic humans alone and several baseline assistance

strategies, and that intent-conditioning enables adherence to human preferences during task execution, leading to improved performance while satisfying the human's objective.

110

Meta-Control: Automatic Model-based Control Synthesis for Heterogeneous Robot Skills
Tianhao Wei,Liqian Ma,Rui Chen,Weiye Zhao,Changliu Liu
https://openreview.net/forum?id=cvVEkS5yij
The requirements for real-world manipulation tasks are diverse and often conflicting; some tasks require precise motion while others require force compliance; some tasks require avoidance of certain regions while others require convergence to certain states. Satisfying these varied requirements with a fixed state-action representation and control strategy is challenging, impeding the development of a universal robotic foundation model. In this work, we propose Meta-Control, the first LLM-enabled automatic control synthesis approach that creates customized state representations and control strategies tailored to specific tasks. Our core insight is that a meta-control system can be built to automate the thought process that human experts use to design control systems. Specifically, human experts heavily use a model-based, hierarchical (from abstract to concrete) thought model, then compose various dynamic models and controllers together to form a control system. Meta-Control mimics the thought model and harnesses LLM's extensive control knowledge with Socrates' "art of midwifery" to automate the thought process. Meta-Control stands out for its fully model-based nature, allowing rigorous analysis, generalizability, robustness, efficient parameter tuning, and reliable real-time execution.

111

Tag Map: A Text-Based Map for Spatial Reasoning and Navigation with Large Language Models
Mike Zhang,Kaixian Qu,Vaishakh Patil,Cesar Cadena,Marco Hutter
https://openreview.net/forum?id=eU5E0oTtpS
Large Language Models (LLM) have emerged as a tool for robots to generate task plans using common sense reasoning. For the LLM to generate actionable plans, scene context must be provided, often through a map. Recent works have shifted from explicit maps with fixed semantic classes to implicit open vocabulary maps based on queryable embeddings capable of representing any semantic class. However, embeddings cannot directly report the scene context as they are implicit, requiring further processing for LLM integration. To address this, we propose an explicit text-based map that can represent thousands of semantic classes while easily integrating with LLMs due to their text-based nature by building upon large-scale image recognition models. We study how entities in our map can be localized and show through evaluations that our text-based map localizations perform comparably to those from open vocabulary maps while using two to four orders of magnitude less memory. Real-robot experiments demonstrate the grounding of an LLM with the text-based map to solve user tasks.

112

Velociraptor: Leveraging Visual Foundation Models for Label-Free, Risk-Aware Off-Road Navigation
Samuel Triest,Matthew Sivaprakasam,Shubhra Aich,David Fan,Wenshan Wang,Sebastian Scherer
https://openreview.net/forum?id=AhEE5wrcLU
Traversability analysis in off-road regimes is a challenging task that requires understanding of multi-modal inputs such as camera and LiDAR. These measurements are often sparse, noisy, and difficult to interpret, particularly in the off-road setting. Existing systems are very engineering-

intensive, often requiring hand-tuning of traversability rules and manual annotation of semantic labels. Furthermore, existing methods for analyzing traversability risk and uncertainty are computationally expensive or not well-calibrated. We propose Velociraptor, a traversability analysis system that performs [veloci]ty-informed, [r]isk-[a]ware [p]erception and [t]raversability for [o]ff-[r]oad driving without any human annotations. We achieve this via the use of visual foundation models (VFMs) and geometric mapping to produce a rich visual-geometric representation of the robot's local environment. We then leverage this representation to produce costmaps, speedmaps, and uncertainty maps using state-of-the-art fully self-supervised techniques. Our approach enables intelligent high-speed off-road navigation with zero human annotation, and with about forty minutes of expert data, outperforms several geometric and semantic traversability baselines, both in offline and real-world robot trials across multiple challenging off-road sites.

113

DiffusionSeeder: Seeding Motion Optimization with Diffusion for Rapid Motion Planning
Huang Huang,Balakumar Sundaralingam,Arsalan Mousavian,Adithyavairavan Murali,Ken Goldberg,Dieter Fox
https://openreview.net/forum?id=B7Lf6xEv7l
Running optimization across many parallel seeds leveraging GPU compute [2] have relaxed the need for a good initialization, but this can fail if the problem is highly non-convex as all seeds could get stuck in local minima. One such setting is collision-free motion optimization for robot manipulation, where optimization converges quickly on easy problems but struggle in obstacle dense environments (e.g., a cluttered cabinet or table). In these situations, graph based planning algorithms are called to obtain seeds, resulting significant slowdowns. We propose DiffusionSeeder, a diffusion based approach that generates trajectories to seed motion optimization for rapid robot motion planning. DiffusionSeeder takes the initial depth image observation of the scene and generates high quality, multi-modal trajectories that are then fine-tuned with few iterations of motion optimization. We integrated DiffusionSeeder with cuRobo, a GPU-accelerated motion optimization method, to generate the seed trajectories which results in 12x speed up on average, and 36x speed up for more complicated problems, while achieving 10% higher success rate in partially observed simulation environments. Our results prove the effectiveness of using diverse solutions from learned diffusion model. Physical experiments on a Franka robot demonstrate the sim2real transfer of DiffusionSeeder to the real robot, with an average success rate of 86% and planning time of 26ms, increasing on cuRobo by 51% higher success rate and 2.5x speed up. The code and the model weights will be available after publication.

114

Learning Performance-oriented Control Barrier Functions Under Complex Safety Constraints and Limited Actuation
Lakshmideepakreddy Manda,Shaoru Chen,Mahyar Fazlyab
https://openreview.net/forum?id=8JLmTZsxGh
Control Barrier Functions (CBFs) offer an elegant framework for constraining nonlinear control system dynamics to an invariant subset of a pre-specified safe set. However, finding a CBF that simultaneously promotes performance by maximizing the resulting control invariant set while accommodating complex safety constraints, especially in high relative degree systems with actuation constraints, remains a significant challenge. In this work, we propose a novel self-supervised learning framework that holistically addresses these hurdles. Given a Boolean

composition of multiple state constraints defining the safe set, our approach begins by constructing a smooth function whose zero superlevel set provides an inner approximation of the safe set. This function is then used with a smooth neural network to parameterize the CBF candidate. Finally, we design a physics-informed training loss function based on a Hamilton-Jacobi Partial Differential Equation (PDE) to train the PINN-CBF and enlarge the volume of the induced control invariant set. We demonstrate the effectiveness of our approach on a 2D double integrator (DI) system and a 7D fixed-wing aircraft system (F16).

115

Learning Visuotactile Estimation and Control for Non-prehensile Manipulation under Occlusions
Juan Del Aguila Ferrandis,Joao Moura,Sethu Vijayakumar
https://openreview.net/forum?id=oSU7M7MK6B
Manipulation without grasping, known as non-prehensile manipulation, is essential for dexterous robots in contact-rich environments, but presents many challenges relating with underactuation, hybrid-dynamics, and frictional uncertainty. Additionally, object occlusions in a scenario of contact uncertainty and where the motion of the object evolves independently from the robot becomes a critical problem, which previous literature fails to address. We present a method for learning visuotactile state estimators and uncertainty-aware control policies for non-prehensile manipulation under occlusions, by leveraging diverse interaction data from privileged policies trained in simulation. We formulate the estimator within a Bayesian deep learning framework, to model its uncertainty, and then train uncertainty-aware control policies by incorporating the pre-learned estimator into the reinforcement learning (RL) loop, both of which lead to significantly improved estimator and policy performance. Therefore, unlike prior non-prehensile research that relies on complex external perception set-ups, our method successfully handles occlusions after sim-to-real transfer to robotic hardware with a simple onboard camera.

116

ANAVI: Audio Noise Awareness using Visual of Indoor environments for NAVIgation
Vidhi Jain,Rishi Veerapaneni,Yonatan Bisk
https://openreview.net/forum?id=IsZb0wT3Kw
We propose Audio Noise Awareness using Visuals of Indoors for NAVIgation for quieter robot path planning. While humans are naturally aware of the noise they make and its impact on those around them, robots currently lack this awareness. A key challenge in achieving audio awareness for robots is estimating how loud will the robot’s actions be at a listener’s location? Since sound depends upon the geometry and material composition of rooms, we train the robot to passively perceive loudness using visual observations of indoor environments. To this end, we generate data on how loud an `impulse' sounds at different listener locations in simulated homes, and train our Acoustic Noise Predictor (ANP). Next, we collect acoustic profiles corresponding to different actions for navigation. Unifying ANP with action acoustics, we demonstrate experiments with wheeled (Hello Robot Stretch) and legged (Unitree Go2) robots so that these robots adhere to the noise constraints of the environment. All simulated and real-world data, code and model checkpoints is released at https://anavi-corl24.github.io/.

117

Online Transfer and Adaptation of Tactile Skill: A Teleoperation Framework
Xiao Chen,Tianle Ni,Kübra Karacan,Hamid Sadeghian,Sami Haddadin
https://openreview.net/forum?id=6X3ybeVpDi
This paper presents a teleoperation framework designed for online learning and adaptation of tactile skills, which provides an intuitive interface without need for physical access to execution robot. The proposed tele-teaching approach utilizes periodical Dynamical Movement Primitives (DMP) and Recursive Least Square (RLS) for generating tactile skills. An autonomy allocation strategy, guided by the learning confidence and operator intention, ensures a smooth transition between human demonstration to autonomous robot operation. Our experimental results with two 7 Degree of Freedom (DoF) Franka Panda robot demonstrates that the tele-teaching framework facilitates online motion and force learning and adaptation within a few iterations.

118

SoloParkour: Constrained Reinforcement Learning for Visual Locomotion from Privileged Experience
Elliot Chane-Sane,Joseph Amigo,Thomas Flayols,Ludovic Righetti,Nicolas Mansard
https://openreview.net/forum?id=DSdAEsEGhE
Parkour poses a significant challenge for legged robots, requiring navigation through complex environments with agility and precision based on limited sensory inputs. In this work, we introduce a novel method for training end-to-end visual policies, from depth pixels to robot control commands, to achieve agile and safe quadruped locomotion. We formulate robot parkour as a constrained reinforcement learning (RL) problem designed to maximize the emergence of agile skills within the robot's physical limits while ensuring safety. We first train a policy without vision using privileged information about the robot's surroundings. We then generate experience from this privileged policy to warm-start a sample efficient off-policy RL algorithm from depth images. This allows the robot to adapt behaviors from this privileged experience to visual locomotion while circumventing the high computational costs of RL directly from pixels. We demonstrate the effectiveness of our method on a real Solo-12 robot, showcasing its capability to perform a variety of parkour skills such as walking, climbing, leaping, and crawling.

119

Handling Long-Term Safety and Uncertainty in Safe Reinforcement Learning
Jonas Günster,Puze Liu,Jan Peters,Davide Tateo
https://openreview.net/forum?id=97QXO0uBEO
Safety is one of the key issues preventing the deployment of reinforcement learning techniques in real-world robots. While most approaches in the Safe Reinforcement Learning area do not require prior knowledge of constraints and robot kinematics and rely solely on data, it is often difficult to deploy them in complex real-world settings. Instead, model-based approaches that incorporate prior knowledge of the constraints and dynamics into the learning framework have proven capable of deploying the learning algorithm directly on the real robot. Unfortunately, while an approximated model of the robot dynamics is often available, the safety constraints are task-specific and hard to obtain: they may be too complicated to encode analytically, too expensive to compute, or it may be difficult to envision a priori the long-term safety requirements. In this paper, we bridge this gap by extending the safe exploration method, ATACOM, with learnable constraints, with a particular focus on ensuring long-term safety and handling of uncertainty. Our approach is

competitive or superior to state-of-the-art methods in final performance while maintaining safer behavior during training.

120

HYPERmotion: Learning Hybrid Behavior Planning for Autonomous Loco-manipulation
Jin Wang,Rui Dai,Weijie Wang,Luca Rossini,Francesco Ruscelli,Nikos Tsagarakis
https://openreview.net/forum?id=ma7McOiCZY
Enabling robots to autonomously perform hybrid motions in diverse environments can be beneficial for long-horizon tasks such as material handling, household chores, and work assistance. This requires extensive exploitation of intrinsic motion capabilities, extraction of affordances from rich environmental information, and planning of physical interaction behaviors. Despite recent progress has demonstrated impressive humanoid whole-body control abilities, they struggle to achieve versatility and adaptability for new tasks. In this work, we propose HYPERmotion, a framework that learns, selects and plans behaviors based on tasks in different scenarios. We combine reinforcement learning with whole-body optimization to generate motion for 38 actuated joints and create a motion library to store the learned skills. We apply the planning and reasoning features of the large language models (LLMs) to complex loco-manipulation tasks, constructing a hierarchical task graph that comprises a series of primitive behaviors to bridge lower-level execution with higher-level planning. By leveraging the interaction of distilled spatial geometry and 2D observation with a visual language model (VLM) to ground knowledge into a robotic morphology selector to choose appropriate actions in single- or dual-arm, legged or wheeled locomotion. Experiments in simulation and real-world show that learned motions can efficiently adapt to new tasks, demonstrating high autonomy from free-text commands in unstructured scenes. Videos and website: hy-motion.github.io//

121

MimicTouch: Leveraging Multi-modal Human Tactile Demonstrations for Contact-rich Manipulation
Kelin Yu,Yunhai Han,Qixian Wang,Vaibhav Saxena,Danfei Xu,Ye Zhao
https://openreview.net/forum?id=7yMZAUkXa4
Tactile sensing is critical to fine-grained, contact-rich manipulation tasks, such as insertion and assembly. Prior research has shown the possibility of learning tactile-guided policy from teleoperated demonstration data. However, to provide the demonstration, human users often rely on visual feedback to control the robot. This creates a gap between the sensing modality used for controlling the robot (visual) and the modality of interest (tactile). To bridge this gap, we introduce "MimicTouch'', a novel framework for learning policies directly from demonstrations provided by human users with their hands. The key innovations are i) a human tactile data collection system which collects multi-modal tactile dataset for learning human's tactile-guided control strategy, ii) an imitation learning-based framework for learning human's tactile-guided control strategy through such data, and iii) an online residual RL framework to bridge the embodiment gap between the human hand and the robot gripper. Through comprehensive experiments, we highlight the efficacy of utilizing human's tactile-guided control strategy to resolve contact-rich manipulation tasks. The project website is at https://sites.google.com/view/MimicTouch.

122

ResPilot: Teleoperated Finger Gaiting via Gaussian Process Residual Learning
Patrick Naughton,Jinda Cui,Karankumar Patel,Soshi Iba
https://openreview.net/forum?id=B45HRM4Wb4
Dexterous robot hand teleoperation allows for long-range transfer of human manipulation expertise, and could simultaneously provide a way for humans to teach these skills to robots. However, current methods struggle to reproduce the functional workspace of the human hand, often limiting them to simple grasping tasks. We present a novel method for finger-gaited manipulation with multi-fingered robot hands. Our method provides the operator enhanced flexibility in making contacts by expanding the reachable workspace of the robot hand through residual Gaussian Process learning. We also assist the operator in maintaining stable contacts with the object by allowing them to constrain fingertips of the hand to move in concert. Extensive quantitative evaluations show that our method significantly increases the reachable workspace of the robot hand and enables the completion of novel dexterous finger gaiting tasks.

123

Rate-Informed Discovery via Bayesian Adaptive Multifidelity Sampling
Aman Sinha,Payam Nikdel,Supratik Paul,Shimon Whiteson
https://openreview.net/forum?id=bftFwjSJxk
Ensuring the safety of autonomous vehicles (AVs) requires both accurate estimation of their performance and efficient discovery of potential failure cases. This paper introduces Bayesian adaptive multifidelity sampling (BAMS), which leverages the power of adaptive Bayesian sampling to achieve efficient discovery while simultaneously estimating the rate of adverse events. BAMS prioritizes exploration of regions with potentially low performance, leading to the identification of novel and critical scenarios that traditional methods might miss. Using real-world AV data we demonstrate that BAMS discovers 10 times as many issues as Monte Carlo (MC) and importance sampling (IS) baselines, while at the same time generating rate estimates with variances 15 and 6 times narrower than MC and IS baselines respectively.

124

ClutterGen: A Cluttered Scene Generator for Robot Learning
Yinsen Jia,Boyuan Chen
https://openreview.net/forum?id=k0ogr4dnhG
We introduce ClutterGen, a physically compliant simulation scene generator capable of producing highly diverse, cluttered, and stable scenes for robot learning. Generating such scenes is challenging as each object must adhere to physical laws like gravity and collision. As the number of objects increases, finding valid poses becomes more difficult, necessitating significant human engineering effort, which limits the diversity of the scenes. To overcome these challenges, we propose a reinforcement learning method that can be trained with physics-based reward signals provided by the simulator. Our experiments demonstrate that ClutterGen can generate cluttered object layouts with up to ten objects on confined table surfaces. Additionally, our policy design explicitly encourages the diversity of the generated scenes for open-ended generation. Our real-world robot results show that ClutterGen can be directly used for clutter rearrangement and stable placement policy training.

125

One Policy to Run Them All: an End-to-end Learning Approach to Multi-Embodiment Locomotion
Nico Bohlinger,Grzegorz Czechmanowski,Maciej Piotr Krupka,Piotr Kicki,Krzysztof Walas,Jan Peters,Davide Tateo
https://openreview.net/forum?id=PbQOZntuXO
Deep Reinforcement Learning techniques are achieving state-of-the-art results in robust legged locomotion. While there exists a wide variety of legged platforms such as quadruped, humanoids, and hexapods, the field is still missing a single learning framework that can control all these different embodiments easily and effectively and possibly transfer, zero or few-shot, to unseen robot embodiments. To close this gap, we introduce URMA, the Unified Robot Morphology Architecture. Our framework brings the end-to-end Multi-Task Reinforcement Learning approach to the realm of legged robots, enabling the learned policy to control any type of robot morphology. The key idea of our method is to allow the network to learn an abstract locomotion controller that can be seamlessly shared between embodiments thanks to our morphology-agnostic encoders and decoders. This flexible architecture can be seen as a first step in building a foundation model for legged robot locomotion. Our experiments show that URMA can learn a locomotion policy on multiple embodiments that can be easily transferred to unseen robot platforms in simulation and the real world.

126

PianoMime: Learning a Generalist, Dexterous Piano Player from Internet Demonstrations
Cheng Qian,Julen Urain,Kevin Zakka,Jan Peters
https://openreview.net/forum?id=t0LkF9JnVb
In this work, we introduce PianoMime, a framework for training a piano-playing agent using internet demonstrations. The internet is a promising source of large-scale demonstrations for training our robot agents. In particular, for the case of piano-playing, Youtube is full of videos of professional pianists playing a wide myriad of songs. In our work, we leverage these demonstrations to learn a generalist piano-playing agent capable of playing any arbitrary song. Our framework is divided into three parts: a data preparation phase to extract the informative features from the Youtube videos, a policy learning phase to train song-specific expert policies from the demonstrations and a policy distillation phase to distil the policies into a single generalist agent. We explore different policy designs to represent the agent and evaluate the influence of the amount of training data on the generalization capability of the agent to novel songs not available in the dataset. We show that we are able to learn a policy with up to 57% F1 score on unseen songs.

127

Teaching Robots with Show and Tell: Using Foundation Models to Synthesize Robot Policies from Language and Visual Demonstration
Michael Murray,Abhishek Gupta,Maya Cakmak
https://openreview.net/forum?id=G8UcwxNAoD
We introduce a modular, neuro-symbolic framework for teaching robots new skills through language and visual demonstration. Our approach, ShowTell, composes a mixture of foundation models to synthesize robot manipulation programs that are easy to interpret and generalize across a wide range of tasks and environments. ShowTell is designed to handle complex demonstrations involving high level logic such as loops and conditionals while being intuitive and natural for end-users. We validate this approach through a series of real-world robot experiments, showing that

ShowTell out-performs a state-of-the-art baseline based on GPT4-V, on a variety of tasks, and that it is able to generalize to unseen environments and within category objects.

128

DiffuseLoco: Real-Time Legged Locomotion Control with Diffusion from Offline Datasets
Xiaoyu Huang,Yufeng Chi,Ruofeng Wang,Zhongyu Li,Xue Bin Peng,Sophia Shao,Borivoje Nikolic,Koushil Sreenath
https://openreview.net/forum?id=nVJm2RdPDu
Offline learning at scale has led to breakthroughs in computer vision, natural language processing, and robotic manipulation domains. However, scaling up learning for legged robot locomotion, especially with multiple skills in a single policy, presents significant challenges for prior online reinforcement learning (RL) methods. To address this challenge, we propose DiffuseLoco, a novel, scalable framework that leverages diffusion models to directly learn from offline multimodal datasets with a diverse set of locomotion skills. With design choices tailored for real-time control in dynamical systems, including receding horizon control and delayed inputs, DiffuseLoco is capable of reproducing multimodality in performing various locomotion skills, zero-shot transferred to real quadruped robots and deployed on edge computes. Through extensive real-world benchmarking, DiffuseLoco exhibits better stability and velocity tracking performance compared to prior RL and non-diffusion-based behavior cloning baselines. This work opens new possibilities for scaling up learning-based legged locomotion control through the scaling of large, expressive models and diverse offline datasets.

129

Tokenize the World into Object-level Knowledge to Address Long-tail Events in Autonomous Driving
Thomas Tian,Boyi Li,Xinshuo Weng,Yuxiao Chen,Edward Schmerling,Yue Wang,Boris Ivanovic,Marco Pavone
https://openreview.net/forum?id=M0Gv07MUMU
The autonomous driving industry is increasingly adopting end-to-end learning from sensory inputs to minimize human biases in system design. Traditional end-to-end driving models, however, suffer from long-tail events due to rare or unseen inputs within their training distributions. To address this, we propose TOKEN, a novel Multi-Modal Large Language Model (MM-LLM) that tokenizes the world into object-level knowledge, enabling better utilization of LLM's reasoning capabilities to enhance autonomous vehicle planning in long-tail scenarios. TOKEN effectively alleviates data scarcity and inefficient tokenization by producing condensed and semantically enriched representations of the scene. Our results demonstrate that TOKEN excels in grounding, reasoning, and planning capabilities, outperforming existing frameworks with a 27% reduction in trajectory L2 error and a 39% decrease in collision rates in long-tail scenarios. Additionally, our work highlights the importance of representation alignment and structured reasoning in sparking the common-sense reasoning capabilities of MM-LLMs for effective planning.

130

TidyBot++: An Open-Source Holonomic Mobile Manipulator for Robot Learning
Jimmy Wu,William Chong,Robert Holmberg,Aaditya Prasad,Yihuai Gao,Oussama Khatib,Shuran Song,Szymon Rusinkiewicz,Jeannette Bohg
https://openreview.net/forum?id=L4p6zTlj6k
Exploiting the promise of recent advances in imitation learning for mobile manipulation will require

the collection of large numbers of human-guided demonstrations. This paper proposes an open-source design for an inexpensive, robust, and flexible mobile manipulator that can support arbitrary arms, enabling a wide range of real-world household mobile manipulation tasks. Crucially, our design uses powered casters to enable the mobile base to be fully holonomic, able to control all planar degrees of freedom independently and simultaneously. This feature makes the base more maneuverable and simplifies many mobile manipulation tasks, eliminating the kinematic constraints that create complex and time-consuming motions in nonholonomic bases. We equip our robot with an intuitive mobile phone teleoperation interface to enable easy data acquisition for imitation learning. In our experiments, we use this interface to collect data and show that the resulting learned policies can successfully perform a variety of common household mobile manipulation tasks.

131

Jacta: A Versatile Planner for Learning Dexterous and Whole-body Manipulation
Jan Bruedigam,Ali Adeeb Abbas,Maks Sorokin,Kuan Fang,Brandon Hung,Maya Guru,Stefan Georg Sosnowski,Jiuguang Wang,Sandra Hirche,Simon Le Cleac'h
https://openreview.net/forum?id=vobaOY0qDl
Robotic manipulation is challenging due to discontinuous dynamics, as well as high-dimensional state and action spaces. Data-driven approaches that succeed in manipulation tasks require large amounts of data and expert demonstrations, typically from humans. Existing planners are restricted to specific systems and often depend on specialized algorithms for using demonstrations. Therefore, we introduce a flexible motion planner tailored to dexterous and whole-body manipulation tasks. Our planner creates readily usable demonstrations for reinforcement learning algorithms, eliminating the need for additional training pipeline complexities. With this approach, we can efficiently learn policies for complex manipulation tasks, where traditional reinforcement learning alone only makes little progress. Furthermore, we demonstrate that learned policies are transferable to real robotic systems for solving complex dexterous manipulation tasks.

132

Robust Manipulation Primitive Learning via Domain Contraction
Teng Xue,Amirreza Razmjoo,Suhan Shetty,Sylvain Calinon
https://openreview.net/forum?id=yNQu9zqx6X
Contact-rich manipulation plays an important role in everyday life, but uncertain parameters pose significant challenges to model-based planning and control. To address this issue, domain adaptation and domain randomization have been proposed to learn robust policies. However, they either lose the generalization ability to diverse instances or perform conservatively due to neglecting instance-specific information. In this paper, we propose a bi-level approach to learn robust manipulation primitives, including parameter-augmented policy learning using multiple models with tensor approximation, and parameter-conditioned policy retrieval through domain contraction. This approach unifies domain randomization and domain adaptation, providing optimal behaviors while keeping generalization ability. We validate the proposed method on three contact-rich manipulation primitives: hitting, pushing, and reorientation. The experimental results showcase the superior performance of our approach in generating robust policies for instances with diverse physical parameters.

133

In-Flight Attitude Control of a Quadruped using Deep Reinforcement Learning
Tarek El-Agroudi,Finn Gross Maurer,Jørgen Anker Olsen,Kostas Alexis
https://openreview.net/forum?id=67tTQeO4HQ
We present the development and real world demonstration of an in-flight attitude control law for a small low-cost quadruped with a five-bar-linkage leg design using only its legs as reaction masses. The control law is trained using deep reinforcement learning (DRL) and specifically through Proximal Policy Optimization (PPO) in the NVIDIA Omniverse Isaac Sim simulator with a GPU-accelerated DRL pipeline. To demonstrate the policy, a small quadruped is designed, constructed, and evaluated both on a rotating pole test setup and in free fall. During a free fall of 0.7 seconds, the quadruped follows commanded attitude steps of 45 degrees in all principal axes, and achieves an average base angular velocity of 110 degrees per second during large attitude reference steps.

134

Multi-Transmotion: Pre-trained Model for Human Motion Prediction
Yang Gao,Po-Chien Luan,Alexandre Alahi
https://openreview.net/forum?id=X3OfR3axX4
The ability of intelligent systems to predict human behaviors is essential, particularly in fields such as autonomous vehicle navigation and social robotics. However, the intricacies of human motion have precluded the development of a standardized dataset and model for human motion prediction, thereby hindering the establishment of pre-trained models. In this paper, we address these limitations by integrating multiple datasets, encompassing both trajectory and 3D pose keypoints, to further propose a pre-trained model for human motion prediction. We merge seven distinct datasets across varying modalities and standardize their formats. To facilitate multimodal pre-training, we introduce Multi-Transmotion, an innovative transformer-based model capable of cross-modality pre-training. Additionally, we devise a novel masking strategy to learn rich representations. Our methodology demonstrates competitive performance across various datasets on several downstream tasks, including trajectory prediction in the NBA and JTA datasets, as well as pose prediction in the AMASS and 3DPW datasets. The code will be made available upon publication.

135

Bridging the gap between Learning-to-plan, Motion Primitives and Safe Reinforcement Learning
Piotr Kicki,Davide Tateo,Puze Liu,Jonas Günster,Jan Peters,Krzysztof Walas
https://openreview.net/forum?id=ZdgaF8fOc0
Trajectory planning under kinodynamic constraints is fundamental for advanced robotics applications that require dexterous, reactive, and rapid skills in complex environments. These constraints, which may represent task, safety, or actuator limitations, are essential for ensuring the proper functioning of robotic platforms and preventing unexpected behaviors. Recent advances in kinodynamic planning demonstrate that learning-to-plan techniques can generate complex and reactive motions under intricate constraints. However, these techniques necessitate the analytical modeling of both the robot and the entire task, a limiting assumption when systems are extremely complex or when constructing accurate task models is prohibitive. This paper addresses this limitation by combining learning-to-plan methods with reinforcement learning, resulting in a novel integration of black-box learning of motion primitives and optimization. We evaluate our approach against state-of-the-art safe reinforcement learning methods, showing that

our technique, particularly when exploiting task structure, outperforms baseline methods in challenging scenarios such as planning to hit in robot air hockey. This work demonstrates the potential of our integrated approach to enhance the performance and safety of robots operating under complex kinodynamic constraints.

136

PointPatchRL - Masked Reconstruction Improves Reinforcement Learning on Point Clouds
Balazs Gyenes,Nikolai Franke,Philipp Becker,Gerhard Neumann
https://openreview.net/forum?id=3jNEz3kUSl
Perceiving the environment via cameras is crucial for Reinforcement Learning (RL) in robotics. While images are a convenient form of representation, they often complicate extracting important geometric details, especially with varying geometries or deformable objects. In contrast, point clouds naturally represent this geometry and easily integrate color and positional data from multiple camera views. However, while point-cloud processing with deep learning has seen many recent successes, RL on point clouds is under-researched, with only the simplest encoder architecture considered in the literature. We introduce PointPatchRL (PPRL), a method for RL on point clouds that builds on the common paradigm of dividing point clouds into overlapping patches, tokenizing them, and processing the tokens with transformers. PPRL provides significant improvements compared with other point-cloud processing architectures previously used for RL. We then complement PPRL with masked reconstruction for representation learning and show that our method outperforms strong model-free and model-based baselines on image observations in complex manipulation tasks containing deformable objects and variations in target object geometry.

137

Solving Offline Reinforcement Learning with Decision Tree Regression
Prajwal Koirala,Cody Fleming
https://openreview.net/forum?id=eTRncsYYdv
This study presents a novel approach to addressing offline reinforcement learning (RL) problems by reframing them as regression tasks that can be effectively solved using Decision Trees. Mainly, we introduce two distinct frameworks: return-conditioned and return-weighted decision tree policies (RCDTP and RWDTP), both of which achieve notable speed in agent training as well as inference, with training typically lasting less than a few minutes. Despite the simplification inherent in this reformulated approach to offline RL, our agents demonstrate performance that is at least on par with the established methods. We evaluate our methods on D4RL datasets for locomotion and manipulation, as well as other robotic tasks involving wheeled and flying robots. Additionally, we assess performance in delayed/sparse reward scenarios and highlight the explainability of these policies through action distribution and feature importance.

138

Learning to Open and Traverse Doors with a Legged Manipulator
Mike Zhang,Yuntao Ma,Takahiro Miki,Marco Hutter
https://openreview.net/forum?id=VoC3wF6fbh
Using doors is a longstanding challenge in robotics and is of significant practical interest in giving robots greater access to human-centric spaces. The task is challenging due to the need for online adaptation to varying door properties and precise control in manipulating the door panel and navigating through the confined doorway. To address this, we propose a learning-based controller

for a legged manipulator to open and traverse through doors. The controller is trained using a teacher-student approach in simulation to learn robust task behaviors as well as estimate crucial door properties during the interaction. Unlike previous works, our approach is a single control policy that can handle both push and pull doors through learned behaviour which infers the opening direction during deployment without prior knowledge. The policy was deployed on the ANYmal legged robot with an arm and achieved a success rate of 95.0% in repeated trials conducted in an experimental setting. Additional experiments validate the policy's effectiveness and robustness to various doors and disturbances. A video overview of the method and experiments is provided in the supplementary material.

139

TLDR: Unsupervised Goal-Conditioned RL via Temporal Distance-Aware Representations
Junik Bae,Kwanyoung Park,Youngwoon Lee
https://openreview.net/forum?id=deywgeWmL5
Unsupervised goal-conditioned reinforcement learning (GCRL) is a promising paradigm for developing diverse robotic skills without external supervision. However, existing unsupervised GCRL methods often struggle to cover a wide range of states in complex environments due to their limited exploration and sparse or noisy rewards for GCRL. To overcome these challenges, we propose a novel unsupervised GCRL method that leverages TemporaL Distance-aware Representations (TLDR). Based on temporal distance, TLDR selects faraway goals to initiate exploration and computes intrinsic exploration rewards and goal-reaching rewards. Specifically, our exploration policy seeks states with large temporal distances (i.e. covering a large state space), while the goal-conditioned policy learns to minimize the temporal distance to the goal (i.e. reaching the goal). Our results in six simulated locomotion environments demonstrate that TLDR significantly outperforms prior unsupervised GCRL methods in achieving a wide range of states.

140

SELFI: Autonomous Self-Improvement with RL for Vision-Based Navigation around People
Noriaki Hirose,Dhruv Shah,Kyle Stachowicz,Ajay Sridhar,Sergey Levine
https://openreview.net/forum?id=rRpmVq6yHv
Autonomous self-improving robots that interact and improve with experience are key to the real-world deployment of robotic systems. In this paper, we propose an online learning method, SELFI, that leverages online robot experience to rapidly fine-tune pre-trained control policies efficiently. SELFI applies online model-free reinforcement learning on top of offline model-based learning to bring out the best parts of both learning paradigms. Specifically, SELFI stabilizes the online learning process by incorporating the same model-based learning objective from offline pre-training into the Q-values learned with online model-free reinforcement learning. We evaluate SELFI in multiple real-world environments and report improvements in terms of collision avoidance, as well as more socially compliant behavior, measured by a human user study. SELFI enables us to quickly learn useful robotic behaviors with less human interventions such as pre-emptive behavior for the pedestrians, collision avoidance for small and transparent objects, and avoiding travel on uneven floor surfaces. We provide supplementary videos to demonstrate the performance of our fine-tuned policy.

141

KOI: Accelerating Online Imitation Learning via Hybrid Key-state Guidance
Jingxian Lu,Wenke Xia,Dong Wang,Zhigang Wang,Bin Zhao,Di Hu,Xuelong Li
https://openreview.net/forum?id=JZzaRY8m8r
Online Imitation Learning methods struggle with the gap between extensive online exploration space and limited expert trajectories, which hinder efficient exploration due to inaccurate task-aware reward estimation. Inspired by the findings from cognitive neuroscience that task decomposition could facilitate cognitive processing for efficient learning, we hypothesize that an agent could estimate precise task-aware imitation rewards for efficient online exploration by decomposing the target task into the objectives of "what to do" and the mechanisms of "how to do". In this work, we introduce the hybrid Key-state guided Online Imitation (KOI) learning approach, which leverages the integration of semantic and motion key states as guidance for task-aware reward estimation. Initially, we utilize the visual-language models to segment the expert trajectory into semantic key states, indicating the objectives of "what to do". Within the intervals between semantic key states, optical flow is employed to capture motion key states to understand the process of "how to do". By integrating a thorough grasp of both semantic and motion key states, we refine the trajectory-matching reward computation, encouraging task-aware exploration for efficient online imitation learning. Our experiment results prove that our method is more sample efficient than previous state-of-the-art approaches in the Meta-World and LIBERO environments. We also conduct real-world robotic manipulation experiments to validate the efficacy of our method, demonstrating the practical applicability of our KOI method.

142

Multi-Task Interactive Robot Fleet Learning with Visual World Models
Huihan Liu,Yu Zhang,Vaarij Betala,Evan Zhang,James Liu,Crystal Ding,Yuke Zhu
https://openreview.net/forum?id=DDIoRSh8ID
Recent advancements in large-scale multi-task robot learning offer the potential for deploying robot fleets in household and industrial settings, enabling them to perform diverse tasks across various environments. However, AI-enabled robots often face challenges with generalization and robustness when exposed to real-world variability and uncertainty. We introduce Sirius-Fleet, a multi-task interactive robot fleet learning framework to address these challenges. Sirius-Fleet monitors robot performance during deployment and involves humans to correct the robot's actions when necessary. We employ a visual world model to predict the outcomes of future actions and build anomaly predictors to predict whether they will likely result in anomalies. As the robot autonomy improves, the anomaly predictors automatically adapt their prediction criteria, leading to fewer requests for human intervention and gradually reducing human workload over time. Evaluations on large-scale benchmarks demonstrate Sirius-Fleet's effectiveness in improving multi-task policy performance and monitoring accuracy. We demonstrate Sirius-Fleet's performance in both RoboCasa in simulation and Mutex in the real world, two diverse, large-scale multi-task benchmarks. More information is available on the project website: https://ut-austin-rpl.github.io/sirius-fleet

143

Equivariant Diffusion Policy
Dian Wang,Stephen Hart,David Surovik,Tarik Kelestemur,Haojie Huang,Haibo Zhao,Mark Yeatman,Jiuguang Wang,Robin Walters,Robert Platt
https://openreview.net/forum?id=wD2kUVLT1g

Recent work has shown diffusion models are an effective approach to learning the multimodal distributions arising from demonstration data in behavior cloning. However, a drawback of this approach is the need to learn a denoising function, which is significantly more complex than learning an explicit policy. In this work, we propose Equivariant Diffusion Policy, a novel diffusion policy learning method that leverages domain symmetries to obtain better sample efficiency and generalization in the denoising function. We theoretically analyze the $\mathrm{SO}(2)$ symmetry of full 6-DoF control and characterize when a diffusion model is $\mathrm{SO}(2)$-equivariant. We furthermore evaluate the method empirically on a set of 12 simulation tasks in MimicGen, and show that it obtains a success rate that is, on average, 21.9% higher than the baseline Diffusion Policy. We also evaluate the method on a real-world system to show that effective policies can be learned with relatively few training samples, whereas the baseline Diffusion Policy cannot.

144

BiGym: A Demo-Driven Mobile Bi-Manual Manipulation Benchmark
Nikita Chernyadev,Nicholas Backshall,Xiao Ma,Yunfan Lu,Younggyo Seo,Stephen James
https://openreview.net/forum?id=EM0wndCeoD
We introduce BiGym, a new benchmark and learning environment for mobile bi-manual demo-driven robotic manipulation. BiGym features 40 diverse tasks set in home environments, ranging from simple target reaching to complex kitchen cleaning. To capture the real-world performance accurately, we provide human-collected demonstrations for each task, reflecting the diverse modalities found in real-world robot trajectories. BiGym supports a variety of observations, including proprioceptive data and visual inputs such as RGB, and depth from 3 camera views. To validate the usability of BiGym, we thoroughly benchmark the state-of-the-art imitation learning algorithms and demo-driven reinforcement learning algorithms within the environment and discuss the future opportunities.

145

Distribution Discrepancy and Feature Heterogeneity for Active 3D Object Detection
Huang-Yu Chen,Jia-Fong Yeh,Jiawei,Pin-Hsuan Peng,Winston H. Hsu
https://openreview.net/forum?id=6oESa4g05O
LiDAR-based 3D object detection is a critical technology for the development of autonomous driving and robotics. However, the high cost of data annotation limits its advancement. We propose a novel and effective active learning (AL) method called Distribution Discrepancy and Feature Heterogeneity (DDFH), which simultaneously considers geometric features and model embeddings, assessing information from both the instance-level and frame-level perspectives. Distribution Discrepancy evaluates the difference and novelty of instances within the unlabeled and labeled distributions, enabling the model to learn efficiently with limited data. Feature Heterogeneity ensures the heterogeneity of intra-frame instance features, maintaining feature diversity while avoiding redundant or similar instances, thus minimizing annotation costs. Finally, multiple indicators are efficiently aggregated using Quantile Transform, providing a unified measure of informativeness. Extensive experiments demonstrate that DDFH outperforms the current state-of-the-art (SOTA) methods on the KITTI and Waymo datasets, effectively reducing the bounding box annotation cost by 56.3% and showing robustness when working with both one-stage and two-stage models.

146

AnyRotate: Gravity-Invariant In-Hand Object Rotation with Sim-to-Real Touch
Max Yang,chenghua lu,Alex Church,Yijiong Lin,Christopher J. Ford,Haoran Li,Efi Psomopoulou,David A.W. Barton,Nathan F. Lepora
https://openreview.net/forum?id=8Yu0TNJNGK
Human hands are capable of in-hand manipulation in the presence of different hand motions. For a robot hand, harnessing rich tactile information to achieve this level of dexterity still remains a significant challenge. In this paper, we present AnyRotate, a system for gravity-invariant multi-axis in-hand object rotation using dense featured sim-to-real touch. We tackle this problem by training a dense tactile policy in simulation and present a sim-to-real method for rich tactile sensing to achieve zero-shot policy transfer. Our formulation allows the training of a unified policy to rotate unseen objects about arbitrary rotation axes in any hand direction. In our experiments, we highlight the benefit of capturing detailed contact information when handling objects of varying properties. Interestingly, we found rich multi-fingered tactile sensing can detect unstable grasps and provide a reactive behavior that improves the robustness of the policy.

147

RT-Sketch: Goal-Conditioned Imitation Learning from Hand-Drawn Sketches
Priya Sundaresan,Quan Vuong,Jiayuan Gu,Peng Xu,Ted Xiao,Sean Kirmani,Tianhe Yu,Michael Stark,Ajinkya Jain,Karol Hausman,Dorsa Sadigh,Jeannette Bohg,Stefan Schaal
https://openreview.net/forum?id=ty1cqzTtUv
Natural language and images are commonly used as goal representations in goal-conditioned imitation learning. However, language can be ambiguous and images can be over-specified. In this work, we study hand-drawn sketches as a modality for goal specification. Sketches can be easy to provide on the fly like language, but like images they can also help a downstream policy to be spatially-aware. By virtue of being minimal, sketches can further help disambiguate task-relevant from irrelevant objects. We present RT-Sketch, a goal-conditioned policy for manipulation that takes a hand-drawn sketch of the desired scene as input, and outputs actions. We train RT-Sketch on a dataset of trajectories paired with synthetically generated goal sketches. We evaluate this approach on six manipulation skills involving tabletop object rearrangements on an articulated countertop. Experimentally we find that RT-Sketch performs comparably to image or language-conditioned agents in straightforward settings, while achieving greater robustness when language goals are ambiguous or visual distractors are present. Additionally, we show that RT-Sketch handles sketches with varied levels of specificity, ranging from minimal line drawings to detailed, colored drawings. For supplementary material and videos, please visit http://rt-sketch.github.io.

148

Bi-Level Motion Imitation for Humanoid Robots
Wenshuai Zhao,Yi Zhao,Joni Pajarinen,Michael Muehlebach
https://openreview.net/forum?id=wH7Wv0nAm8
Imitation learning from human motion capture (MoCap) data provides a promising way to train humanoid robots. However, due to differences in morphology, such as varying degrees of joint freedom and force limits, exact replication of human behaviors may not be feasible for humanoid robots. Consequently, incorporating physically infeasible MoCap data in training datasets can adversely affect the performance of the robot policy. To address this issue, we propose a bi-level optimization-based imitation learning framework that alternates between optimizing both the robot policy and the target MoCap data. Specifically, we first develop a generative latent dynamics

model using a novel self-consistent auto-encoder, which learns sparse and structured motion representations while capturing desired motion patterns in the dataset. The dynamics model is then utilized to generate reference motions while the latent representation regularizes the bi-level motion imitation process. Simulations conducted with a realistic model of a humanoid robot demonstrate that our method enhances the robot policy by modifying reference motions to be physically consistent.

149

Avoid Everything: Model-Free Collision Avoidance with Expert-Guided Fine-Tuning
Adam Fishman,Aaron Walsman,Mohak Bhardwaj,Wentao Yuan,Balakumar Sundaralingam,Byron Boots,Dieter Fox
https://openreview.net/forum?id=gqFIybpsLX
The world is full of clutter. In order to operate effectively in uncontrolled, real world spaces, robots must navigate safely by executing tasks around obstacles while in proximity to hazards. Creating safe movement for robotic manipulators remains a long-standing challenge in robotics, particularly in environments with partial observability. In partially observed settings, classical techniques often fail. Learned end-to-end motion policies can infer correct solutions in these settings, but are as-yet unable to produce reliably safe movement when close to obstacles. In this work, we introduce Avoid Everything, a novel end-to-end system for generating collision-free motion toward a target, even targets close to obstacles. Avoid Everything consists of two parts: 1) Motion Policy Transformer (MπFormer), a transformer architecture for end-to-end joint space control from point clouds, trained on over 1,000,000 expert trajectories and 2) a fine-tuning procedure we call Refining on Optimized Policy Experts (ROPE), which uses optimization to provide demonstrations of safe behavior in challenging states. With these techniques, we are able to successfully solve over 63% of reaching problems that caused the previous state of the art method to fail, resulting in an overall success rate of over 91% in challenging manipulation settings.

150

Modeling the Real World with High-Density Visual Particle Dynamics
William F Whitney,Jake Varley,Deepali Jain,Krzysztof Marcin Choromanski,Sumeet Singh,Vikas Sindhwani
https://openreview.net/forum?id=pcPSGZFaCH
We present High-Density Visual Particle Dynamics (HD-VPD), a learned world model that can emulate the physical dynamics of real scenes by processing massive latent point clouds containing 100K+ particles. To enable efficiency at this scale, we introduce a novel family of Point Cloud Transformers (PCTs) called Interlacers leveraging intertwined linear-attention Performer layers and graph-based neighbour attention layers. We demonstrate the capabilities of HD-VPD by modeling the dynamics of high degree-of-freedom bi-manual robots with two RGB-D cameras. Compared to the previous graph neural network approach, our Interlacer dynamics is twice as fast with the same prediction quality, and can achieve higher quality using 4x as many particles. We illustrate how HD-VPD can evaluate motion plan quality with robotic box pushing and can grasping tasks. See videos and particle dynamics rendered by HD-VPD at https://sites.google.com/view/hd-vpd.

151

Provably Safe Online Multi-Agent Navigation in Unknown Environments
Zhan Gao,Guang Yang,Jasmine Bayrooti,Amanda Prorok
https://openreview.net/forum?id=0M7JiV1GFN
Control Barrier Functions (CBFs) provide safety guarantees for multi-agent navigation. However, traditional approaches require full knowledge of the environment (e.g., obstacle positions and shapes) to formulate CBFs and hence, are not applicable in unknown environments. This paper overcomes this issue by proposing an Online Exploration-based Control Lyapunov Barrier Function (OE-CLBF) controller. It estimates the unknown environment by learning its corresponding CBF with a Support Vector Machine (SVM) in an online manner, using local neighborhood information, and leverages the latter to generate actions for safe navigation. To reduce the computation incurred by the online SVM training, we use an Imitation Learning (IL) framework to predict the importance of neighboring agents with Graph Attention Networks (GATs), and train the SVM only with information received from neighbors of high `value'. The OE-CLBF allows for decentralized deployment, and importantly, provides provable safety guarantees that we derive in this paper. Experiments corroborate theoretical findings and demonstrate superior performance w.r.t. state-of-the-art baselines in a variety of unknown environments.

152

Goal-Reaching Policy Learning from Non-Expert Observations via Effective Subgoal Guidance
RenMing Huang,Shaochong Liu,Yunqiang Pei,Peng Wang,Guoqing Wang,Yang Yang,Heng Tao Shen
https://openreview.net/forum?id=kEZXeaMrkD
In this work, we address the challenging problem of long-horizon goal-reaching policy learning from non-expert, action-free observation data. Unlike fully labeled expert data, our data is more accessible and avoids the costly process of action labeling. Additionally, compared to online learning, which often involves aimless exploration, our data provides useful guidance for more efficient exploration. To achieve our goal, we propose a novel subgoal guidance learning strategy. The motivation behind this strategy is that long-horizon goals offer limited guidance for efficient exploration and accurate state transition. We develop a diffusion strategy-based high-level policy to generate reasonable subgoals as waypoints, preferring states that more easily lead to the final goal. Additionally, we learn state-goal value functions to encourage efficient subgoal reaching. These two components naturally integrate into the off-policy actor-critic framework, enabling efficient goal attainment through informative exploration. We evaluate our method on complex robotic navigation and manipulation tasks, demonstrating a significant performance advantage over existing methods. Our ablation study further shows that our method is robust to observation data with various corruptions.

153

OmniH2O: Universal and Dexterous Human-to-Humanoid Whole-Body Teleoperation and Learning
Tairan He,Zhengyi Luo,Xialin He,Wenli Xiao,Chong Zhang,Weinan Zhang,Kris M. Kitani,Changliu Liu,Guanya Shi
https://openreview.net/forum?id=oL1WEZQal8
We present OmniH2O (Omni Human-to-Humanoid), a learning-based system for whole-body humanoid teleoperation and autonomy. Using kinematic pose as a universal control interface, OmniH2O enables various ways for a human to control a full-sized humanoid with dexterous hands, including using real-time teleoperation through VR headset, verbal instruction, and RGB camera. OmniH2O also enables full autonomy by learning from teleoperated demonstrations or

integrating with frontier models such as GPT-4. OmniH2O demonstrates versatility and dexterity in various real-world whole-body tasks through teleoperation or autonomy, such as playing multiple sports, moving and manipulating objects, and interacting with humans. We develop an RL-based sim-to-real pipeline, which involves large-scale retargeting and augmentation of human motion datasets, learning a real-world deployable policy with sparse sensor input by imitating a privileged teacher policy, and reward designs to enhance robustness and stability. We release the first humanoid whole-body control dataset, OmniH2O-6, containing six everyday tasks, and demonstrate humanoid whole-body skill learning from teleoperated datasets. Videos at the anonymous website https://anonymous-omni-h2o.github.io/

154

Learning H-Infinity Locomotion Control
Junfeng Long,Wenye Yu,Quanyi Li,ZiRui Wang,Dahua Lin,Jiangmiao Pang
https://openreview.net/forum?id=uMZ2jnZUDX
Stable locomotion in precipitous environments is an essential task for quadruped robots, requiring the ability to resist various external disturbances. Recent neural policies enhance robustness against disturbances by learning to resist external forces sampled from a fixed distribution in the simulated environment. However, the force generation process doesn't consider the robot's current state, making it difficult to identify the most effective direction and magnitude that can push the robot to the most unstable but recoverable state. Thus, challenging cases in the buffer are insufficient to optimize robustness. In this paper, we propose to model the robust locomotion learning process as an adversarial interaction between the locomotion policy and a learnable disturbance that is conditioned on the robot state to generate appropriate external forces. To make the joint optimization stable, our novel H_∞ constraint mandates the bound of the ratio between the cost and the intensity of the external forces. We verify the robustness of our approach in both simulated environments and real-world deployment, on quadrupedal locomotion tasks and a more challenging task where the quadruped performs locomotion merely on hind legs. Training and deployment code will be made public.

155

DexCatch: Learning to Catch Arbitrary Objects with Dexterous Hands
Fengbo Lan,Shengjie Wang,Yunzhe Zhang,Haotian Xu,Oluwatosin OluwaPelumi Oseni,Ziye Zhang,Yang Gao,Tao Zhang
https://openreview.net/forum?id=VMqg1CeUQP
Achieving human-like dexterous manipulation remains a crucial area of research in robotics. Current research focuses on improving the success rate of pick-and-place tasks. Compared with pick-and-place, throwing-catching behavior has the potential to increase the speed of transporting objects to their destination. However, dynamic dexterous manipulation poses a major challenge for stable control due to a large number of dynamic contacts. In this paper, we propose a Learning-based framework for Throwing-Catching tasks using dexterous hands (LTC). Our method, LTC, achieves a 73% success rate across 45 scenarios (diverse hand poses and objects), and the learned policies demonstrate strong zero-shot transfer performance on unseen objects. Additionally, in tasks where the object in hand faces sideways, an extremely unstable scenario due to the lack of support from the palm, all baselines fail, while our method still achieves a success rate of over 60%.

156

Multi-agent Reinforcement Learning with Hybrid Action Space for Free Gait Motion Planning of Hexapod Robots
Huiqiao Fu,Kaiqiang Tang,Peng Li,Guizhou Deng,Chunlin Chen
https://openreview.net/forum?id=2AZfKk9tRI
Legged robots are able to overcome challenging terrains through diverse gaits formed by contact sequences. However, environments characterized by discrete footholds present significant challenges. In this paper, we tackle the problem of free gait motion planning for hexapod robots walking in randomly generated plum blossom pile environments. Specifically, we first address the complexity of multi-leg coordination in discrete environments by treating each leg of the hexapod robot as an individual agent. Then, we propose the Hybrid action space Multi-Agent Soft Actor Critic (Hybrid-MASAC) algorithm capable of handling both discrete and continuous actions. Finally, we present an integrated free gait motion planning method based on Hybrid-MASAC, streamlining gait, Center of Mass (COM), and foothold sequences planning into a single model. Comparative and ablation experiments in both of the simulated and real plum blossom pile environments demonstrate the feasibility and efficiency of our method.

157

JointMotion: Joint Self-Supervision for Joint Motion Prediction
Royden Wagner,Omer Sahin Tas,Marvin Klemp,Carlos Fernandez
https://openreview.net/forum?id=OznnnxPLiH
We present JointMotion, a self-supervised pre-training method for joint motion prediction in self-driving vehicles. Our method jointly optimizes a scene-level objective connecting motion and environments, and an instance-level objective to refine learned representations. Scene-level representations are learned via non-contrastive similarity learning of past motion sequences and environment context. At the instance level, we use masked autoencoding to refine multimodal polyline representations. We complement this with an adaptive pre-training decoder that enables JointMotion to generalize across different environment representations, fusion mechanisms, and dataset characteristics. Notably, our method reduces the joint final displacement error of Wayformer, HPTR, and Scene Transformer models by 3%, 8%, and 12%, respectively; and enables transfer learning between the Waymo Open Motion and the Argoverse 2 Motion Forecasting datasets.

158

Conformal Prediction for Semantically-Aware Autonomous Perception in Urban Environments
Achref Doula,Tobias Güdelhöfer,Max Mühlhäuser,Alejandro Sanchez Guinea
https://openreview.net/forum?id=aaY5fVFMVf
We introduce Knowledge-Refined Prediction Sets (KRPS), a novel approach that performs semantically-aware uncertainty quantification for multitask-based autonomous perception in urban environments. KRPS extends conformal prediction (CP) to ensure 2 properties not typically addressed by CP frameworks: semantic label consistency and true label coverage, across multiple perception tasks. We elucidate the capability of KRPS through high-level classification tasks crucial for semantically-aware autonomous perception in urban environments, including agent classification, agent location classification, and agent action classification. In a theoretical analysis, we introduce the concept of semantic label consistency among tasks and prove the semantic consistency and marginal coverage properties of the produced sets by KRPS. The results of our evaluation on the ROAD dataset and the Waymo/ROAD++ dataset show that KRPS

outperforms state-of-the-art CP methods in reducing uncertainty by up to 80% and increasing the semantic consistency by up to 30%, while maintaining the coverage guarantees.

159

Generative Factor Chaining: Coordinated Manipulation with Diffusion-based Factor Graph
Utkarsh Aashu Mishra,Yongxin Chen,Danfei Xu
https://openreview.net/forum?id=p6Wq6TjjHH
Learning to plan for multi-step, multi-manipulator tasks is notoriously difficult because of the large search space and the complex constraint satisfaction problems. We present Generative Factor Chaining (GFC), a composable generative model for planning. GFC represents a planning problem as a spatial-temporal factor graph, where nodes represent objects and robots in the scene, spatial factors capture the distributions of valid relationships among nodes, and temporal factors represent the distributions of skill transitions. Each factor is implemented as a modular diffusion model, which are composed during inference to generate feasible long-horizon plans through bi-directional message passing. We show that GFC can solve complex bimanual manipulation tasks and exhibits strong generalization to unseen planning tasks with novel combinations of objects and constraints. More details can be found at: https://sites.google.com/view/generative-factor-chaining

160

FlowRetrieval: Flow-Guided Data Retrieval for Few-Shot Imitation Learning
Li-Heng Lin,Yuchen Cui,Amber Xie,Tianyu Hua,Dorsa Sadigh
https://openreview.net/forum?id=FHnVRmeqxf
Imitation learning policies in robotics tend to require an extensive amount of demonstrations. It is critical to develop few-shot adaptation strategies that rely only on a small amount of task-specific human demonstrations. Prior works focus on learning general policies from large scale dataset with diverse behaviors. Recent research has shown that directly retrieving relevant past experiences to augment policy learning has great promise in few-shot settings. However, existing data retrieval methods fall under two extremes: they either rely on the existence of exact same behaviors with visually similar scenes in the prior data, which is impractical to assume; or they retrieve based on semantic similarity of high-level language descriptions of the task, which might not be that informative about the shared behaviors or motions across tasks. In this work, we investigate how we can leverage motion similarity in the vast amount of cross-task data to improve few-shot imitation learning of the target task. Our key insight is that motion-similar data carry rich information about the effects of actions and object interactions that can be leveraged during few-shot adaptation. We propose FlowRetrieval, an approach that leverages optical flow representations for both extracting similar motions to target tasks from prior data, and for guiding learning of a policy that can maximally benefit from such data. Our results show FlowRetrieval significantly outperforms prior methods across simulated and real-world domains, achieving on average 27% higher success rate than the best retrieval-based prior method. In the Pen-in-Cup task with a real Franka Emika robot, FlowRetrieval achieves 3.7x the performance of the baseline learning from all prior and target data.

161

Legolas: Deep Leg-Inertial Odometry
Justin Wasserman,Ananye Agarwal,Rishabh Jangir,Girish Chowdhary,Deepak Pathak,Abhinav Gupta
https://openreview.net/forum?id=VdyIhsh1jU
Estimating odometry, where an accumulating position and rotation is tracked, has critical applications in many areas of robotics as a form of state estimation such as in SLAM, navigation, and controls. During deployment of a legged robot, a vision system's tracking can easily get lost. Instead, using only the onboard leg and inertial sensor for odometry is a promising alternative. Previous methods in estimating leg-inertial odometry require analytical modeling or collecting high-quality real-world trajectories to train a model. Analytical modeling is specific to each robot, requires manual fine-tuning, and doesn't always capture real-world phenomena such as slippage. Previous work learning legged odometry still relies on collecting real-world data, this has been shown to not perform well out of distribution. In this work, we show that it is possible to estimate the odometry of a legged robot without any analytical modeling or real-world data collection. In this paper, we present Legolas, the first method that accurately estimates odometry in a purely data-driven fashion for quadruped robots. We deploy our method on two real-world quadruped robots in both indoor and outdoor environments. In the indoor scenes, our proposed method accomplishes a relative pose error that is 73% less than an analytical filtering-based approach and 87.5% less than a real-world behavioral cloning approach. More results are available at: learned-odom.github.io

162

InstructNav: Zero-shot System for Generic Instruction Navigation in Unexplored Environment
Yuxing Long,Wenzhe Cai,Hongcheng Wang,Guanqi Zhan,Hao Dong
https://openreview.net/forum?id=fCDOfpTCzZ
Enabling robots to navigate following diverse language instructions in unexplored environments is an attractive goal for human-robot interaction. However, this goal is challenging because different navigation tasks require different strategies. The scarcity of instruction navigation data hinders training an instruction navigation model with varied strategies. Therefore, previous methods are all constrained to one specific type of navigation instruction. In this work, we propose InstructNav, a generic instruction navigation system. InstructNav makes the first endeavor to handle various instruction navigation tasks without any navigation training or pre-built maps. To reach this goal, we introduce Dynamic Chain-of-Navigation (DCoN) to unify the planning process for different types of navigation instructions. Furthermore, we propose Multi-sourced Value Maps to model key elements in instruction navigation so that linguistic DCoN planning can be converted into robot actionable trajectories. With InstructNav, we complete the R2R-CE task in a zero-shot way for the first time and outperform many task-training methods. Besides, InstructNav also surpasses the previous SOTA method by 10.48% on the zero-shot Habitat ObjNav and by 86.34% on demand-driven navigation DDN. Real robot experiments on diverse indoor scenes further demonstrate our method's robustness in coping with the environment and instruction variations.

163

Humanoid Parkour Learning
Ziwen Zhuang,Shenzhe Yao,Hang Zhao
https://openreview.net/forum?id=fs7ia3FqUM
Parkour is a grand challenge for legged locomotion, even for quadruped robots, requiring active

perception and various maneuvers to overcome multiple challenging obstacles. Existing methods for humanoid locomotion either optimize a trajectory for a single parkour track or train a reinforcement learning policy only to walk with a significant amount of motion references. In this work, we propose a framework for learning an end-to-end vision-based whole-body-control parkour policy for humanoid robots that overcomes multiple parkour skills without any motion prior. Using the parkour policy, the humanoid robot can jump on a 0.42m platform, leap over hurdles, 0.8m gaps, and much more. It can also run at 1.8m/s in the wild and walk robustly on different terrains. We test our policy in indoor and outdoor environments to demonstrate that it can autonomously select parkour skills while following the rotation command of the joystick. We override the arm actions and show that this framework can easily transfer to humanoid mobile manipulation tasks. Videos can be found at https://humanoid4parkour.github.io

164

SkillMimicGen: Automated Demonstration Generation for Efficient Skill Learning and Deployment
Caelan Reed Garrett,Ajay Mandlekar,Bowen Wen,Dieter Fox
https://openreview.net/forum?id=YOFrRTDC6d
Imitation learning from human demonstrations is an effective paradigm for robot manipulation, but acquiring large datasets is costly and resource-intensive, especially for long-horizon tasks. To address this issue, we propose SkillGen, an automated system for generating demonstration datasets from a few human demos. SkillGen segments human demos into manipulation skills, adapts these skills to new contexts, and stitches them together through free-space transit and transfer motion. We also propose a Hybrid Skill Policy (HSP) framework for learning skill initiation, control, and termination components from SkillGen datasets, enabling skills to be sequenced using motion planning at test-time. We demonstrate that SkillGen greatly improves data generation and policy learning performance over a state-of-the-art data generation framework, resulting in the capability produce data for large scene variations, including clutter, and agents that are on average 24% more successful. We demonstrate the efficacy of SkillGen by generating over 24K demonstrations across 18 task variants in simulation from just 60 human demonstrations, and training proficient, often near-perfect, HSP agents. Finally, we apply SkillGen to 3 real-world manipulation tasks and demonstrate zero-shot sim-to-real transfer on a long-horizon assembly task. Videos, and more at https://skillgen.github.io.

165

Object-Centric Dexterous Manipulation from Human Motion Data
Yuanpei Chen,Chen Wang,Yaodong Yang,Karen Liu
https://openreview.net/forum?id=KAzku0Uyh1
Manipulating objects to achieve desired goal states is a basic but important skill for dexterous manipulation. Human hand motions demonstrate proficient manipulation capability, providing valuable data for training robots with multi-finger hands. Despite this potential, substantial challenges arise due to the embodiment gap between human and robot hands. In this work, we introduce a hierarchical policy learning framework that uses human hand motion data for training object-centric dexterous robot manipulation. At the core of our method is a high-level trajectory generative model, learned with a large-scale human hand motion capture dataset, to synthesize human-like wrist motions conditioned on the desired object goal states. Guided by the generated wrist motions, deep reinforcement learning is further used to train a low-level finger controller that is grounded in the robot's embodiment to physically interact with the object to achieve the goal. Through extensive evaluation across 10 household objects, our approach not only demonstrates superior performance but also showcases generalization capability to novel object geometries and

goal states. Furthermore, we transfer the learned policies from simulation to a real-world bimanual dexterous robot system, further demonstrating its applicability in real-world scenarios. Project website: https://cypypccpy.github.io/obj-dex.github.io/.

166

Contrastive Imitation Learning for Language-guided Multi-Task Robotic Manipulation
Teli Ma,Jiaming Zhou,Zifan Wang,Ronghe Qiu,Junwei Liang
https://openreview.net/forum?id=9HkElMlPbU
Developing robots capable of executing various manipulation tasks, guided by natural language instructions and visual observations of intricate real-world environments, remains a significant challenge in robotics. Such robot agents need to understand linguistic commands and distinguish between the requirements of different tasks. In this work, we present **Σ-agent**, an end-to-end imitation learning agent for multi-task robotic manipulation. **Σ-agent** incorporates contrastive Imitation Learning (contrastive IL) modules to strengthen vision-language and current-future representations. An effective and efficient multi-view querying Transformer (MVQ-Former) for aggregating representative semantic information is introduced. **Σ-agent** shows substantial improvement over state-of-the-art methods under diverse settings in 18 RLBench tasks, surpassing RVT by an average of 5.2% and 5.9% in 10 and 100 demonstration training, respectively. **Σ-agent** also achieves 62% success rate with a single policy in 5 real-world manipulation tasks. The code will be released upon acceptance.

167

VoxAct-B: Voxel-Based Acting and Stabilizing Policy for Bimanual Manipulation
I-Chun Arthur Liu,Sicheng He,Daniel Seita,Gaurav S. Sukhatme
https://openreview.net/forum?id=CPQW5kc0pe
Bimanual manipulation is critical to many robotics applications. In contrast to single-arm manipulation, bimanual manipulation tasks are challenging due to higher-dimensional action spaces. Prior works leverage large amounts of data and primitive actions to address this problem, but may suffer from sample inefficiency and limited generalization across various tasks. To this end, we propose VoxAct-B, a language-conditioned, voxel-based method that leverages Vision Language Models (VLMs) to prioritize key regions within the scene and reconstruct a voxel grid. We provide this voxel grid to our bimanual manipulation policy to learn acting and stabilizing actions. This approach enables more efficient policy learning from voxels and is generalizable to different tasks. In simulation, we show that VoxAct-B outperforms strong baselines on fine-grained bimanual manipulation tasks. Furthermore, we demonstrate VoxAct-B on real-world **Open Drawer** and **Open Jar** tasks using two UR5s. Code, data, and videos are available at https://voxact-b.github.io.

168

Structured Bayesian Meta-Learning for Data-Efficient Visual-Tactile Model Estimation
Shaoxiong Yao,Yifan Zhu,Kris Hauser
https://openreview.net/forum?id=TzqKmIhcwq
Estimating visual-tactile models of deformable objects is challenging because vision suffers from occlusion, while touch data is sparse and noisy. We propose a novel data-efficient method for dense heterogeneous model estimation by leveraging experience from diverse training objects. The method is based on Bayesian Meta-Learning (BML), which can mitigate overfitting high-capacity visual-tactile models by meta-learning an informed prior and naturally achieves few-shot

online estimation via posterior estimation. However, BML requires a shared parametric model across tasks but visual-tactile models for diverse objects have different parameter spaces. To address this issue, we introduce Structured Bayesian Meta-Learning (SBML) that incorporates heterogeneous physics models, enabling learning from training objects with varying appearances and geometries. SBML performs zero-shot vision-only prediction of deformable model parameters and few-shot adaptation after a handful of touches. Experiments show that in two classes of heterogeneous objects, namely plants and shoes, SBML outperforms existing approaches in force and torque prediction accuracy in zero- and few-shot settings.

169

OCCAM: Online Continuous Controller Adaptation with Meta-Learned Models
Hersh Sanghvi,Spencer Folk,Camillo Jose Taylor
https://openreview.net/forum?id=xeFKtSXPMd
Control tuning and adaptation present a significant challenge to the usage of robots in diverse environments. It is often nontrivial to find a single set of control parameters by hand that work well across the broad array of environments and conditions that a robot might encounter. Automated adaptation approaches must utilize prior knowledge about the system while adapting to significant domain shifts to find new control parameters quickly. In this work, we present a general framework for online controller adaptation that deals with these challenges. We combine meta-learning with Bayesian recursive estimation to learn prior predictive models of system performance that quickly adapt to online data, even when there is significant domain shift. These predictive models can be used as cost functions within efficient sampling-based optimization routines to find new control parameters online that maximize system performance. Our framework is powerful and flexible enough to adapt controllers for four diverse systems: a simulated race car, a simulated quadrupedal robot, and a simulated and physical quadrotor.

170

Multi-Strategy Deployment-Time Learning and Adaptation for Navigation under Uncertainty
Abhishek Paudel,Xuesu Xiao,Gregory J. Stein
https://openreview.net/forum?id=Isp19rFFV4
We present an approach for performant point-goal navigation in unfamiliar partially-mapped environments. When deployed, our robot runs multiple strategies for deployment-time learning and visual domain adaptation in parallel and quickly selects the best-performing among them. Choosing between policies as they are learned or adapted between navigation trials requires continually updating estimates of their performance as they evolve. Leveraging recent work in model-based learning-informed planning under uncertainty, we determine lower bounds on the would-be performance of newly-updated policies on old trials without needing to re-deploy them. This information constrains and accelerates bandit-like policy selection, affording quick selection of the best-performing strategy shortly after it would start to yield good performance. We validate the effectiveness of our approach in simulated maze-like environments, showing improved navigation cost and cumulative regret versus existing baselines.

171

Generalizing End-To-End Autonomous Driving In Real-World Environments Using Zero-Shot LLMs
Zeyu Dong,Yimin Zhu,Yansong Li,Kevin Mahon,Yu Sun
https://openreview.net/forum?id=s0vHSq5QEv
Traditional autonomous driving methods adopt modular design, decomposing tasks into sub-

tasks, including perception, prediction, planning, and control. In contrast, end-to-end autonomous driving directly outputs actions from raw sensor data, avoiding error accumulation. However, training an end-to-end model requires a comprehensive dataset. Without adequate data, the end-to-end model exhibits poor generalization capabilities. Recently, large language models (LLMs) have been applied to enhance the generalization property of end-to-end driving models. Most studies explore LLMs in an open-loop manner, where the output actions are compared to those of experts without direct activation in the real world. Other studies in closed-loop settings examine their results in simulated environments. In comparison, this paper proposes an efficient architecture that integrates multimodal LLMs into end-to-end real-world driving models in a closed-loop setting. The LLM periodically takes raw sensor data to generate high-level driving instructions. In our architecture, LLMs can effectively guide the end-to-end model, even at a slower rate than the raw sensor data, because updates aren't needed every time frame. This architecture relaxes the trade-off between the latency and inference quality of the LLM. It also allows us to choose a wide variety of LLMs to improve high-level driving instructions and minimize fine-tuning costs. Consequently, our architecture reduces the data collection requirements because the LLMs do not directly output actions, and we only need to train a simple imitation learning model to output actions. In our experiments, the training data for the end-to-end model in a real-world environment consists of only simple obstacle configurations with one traffic cone, while the test environment is more complex and contains different types of obstacles. Experiments show that the proposed architecture enhances the generalization capabilities of the end-to-end model even without fine-tuning the LLM.

172

Leveraging Mutual Information for Asymmetric Learning under Partial Observability
Hai Huu Nguyen,Long Dinh Van The,Christopher Amato,Robert Platt
https://openreview.net/forum?id=9jJP2J1oBP
Even though partial observability is prevalent in robotics, most reinforcement learning studies avoid it due to the difficulty of learning a policy that can efficiently memorize past events and seek information. Fortunately, in many cases, learning can be done in an asymmetric setting where states are available during training but not during execution. Prior studies often leverage the state to indirectly influence the training of a history-based actor (actor-critic methods) or a history-based critic (value-based methods). Instead, we propose using state-observation and state-history mutual information to improve the agent's architecture and ability to seek information and memorize efficiently through intrinsic rewards and an auxiliary task. Our method outperforms strong baselines through extensive experiments and achieves successful sim-to-real transfers to a real robot.

173

Verification of Neural Control Barrier Functions with Symbolic Derivative Bounds Propagation
Hanjiang Hu,Yujie Yang,Tianhao Wei,Changliu Liu
https://openreview.net/forum?id=jnubz7wB2w
Control barrier functions (CBFs) are important in safety-critical systems and robot control applications. Neural networks have been used to parameterize and synthesize CBFs with bounded control input for complex systems. However, it is still challenging to verify pre-trained neural networks CBFs (neural CBFs) in an efficient symbolic manner. To this end, we propose a new efficient verification framework for ReLU-based neural CBFs through symbolic derivative bound propagation by combining the linearly bounded nonlinear dynamic system and the gradient bounds of neural CBFs. Specifically, with Heaviside step function form for derivatives of activation

functions, we show that the symbolic bounds can be propagated through the inner product of neural CBF Jacobian and nonlinear system dynamics. Through extensive experiments on different robot dynamics, our results outperform the interval arithmetic-based baselines in verified rate and verification time along the CBF boundary, validating the effectiveness and efficiency of the proposed method with different model complexity. The code can be found at https://github.com/intelligent-control-lab/verify-neural-CBF.

174

PoliFormer: Scaling On-Policy RL with Transformers Results in Masterful Navigators
Kuo-Hao Zeng,Zichen Zhang,Kiana Ehsani,Rose Hendrix,Jordi Salvador,Alvaro Herrasti,Ross Girshick,Aniruddha Kembhavi,Luca Weihs
https://openreview.net/forum?id=KdVLK0Wo5z
We present PoliFormer (Policy Transformer), an RGB-only indoor navigation agent trained end-to-end with reinforcement learning at scale that generalizes to the real-world without adaptation despite being trained purely in simulation. PoliFormer uses a foundational vision transformer encoder with a causal transformer decoder enabling long-term memory and reasoning. It is trained for hundreds of millions of interactions across diverse environments, leveraging parallelized, multi-machine rollouts for efficient training with high throughput. PoliFormer is a masterful navigator, producing state-of-the-art results across two distinct embodiments, the LoCoBot and Stretch RE-1 robots, and four navigation benchmarks. It breaks through the plateaus of previous work, achieving an unprecedented 85.5% success rate in object goal navigation on the CHORES-S benchmark, a 28.5% absolute improvement. PoliFormer can also be trivially extended to a variety of downstream applications such as object tracking, multi-object navigation, and open-vocabulary navigation with no finetuning.

175

JA-TN: Pick-and-Place Towel Shaping from Crumpled States based on TransporterNet with Joint-Probability Action Inference
Halid Abdulrahim Kadi,Kasim Terzić
https://openreview.net/forum?id=SW8ntpJI0E
Towel manipulation is a crucial step towards more general cloth manipulation. However, folding a towel from an arbitrarily crumpled state and recovering from a failed folding step remain critical challenges in robotics. We propose joint-probability action inference JA-TN, as a way to improve TransporterNet's operational efficiency; to our knowledge, this is the first single data-driven policy to achieve various types of folding from most crumpled states. We present three benchmark domains with a set of shaping tasks and the corresponding oracle policies to facilitate the further development of the field. We also present a simulation-to-reality transfer procedure for vision-based deep learning controllers by processing and augmenting RGB and/or depth images. We also demonstrate JA-TN's ability to integrate with a real camera and a UR3e robot arm, showcasing the method's applicability to real-world tasks.

176

Flow as the Cross-domain Manipulation Interface
Mengda Xu,Zhenjia Xu,Yinghao Xu,Cheng Chi,Gordon Wetzstein,Manuela Veloso,Shuran Song
https://openreview.net/forum?id=cNI0ZkK1yC
We present Im2Flow2Act, a scalable learning framework that enables robots to acquire real-world manipulation skills without the need of real-world robot training data. The key idea behind

Im2Flow2Act is to use object flow as the manipulation interface, bridging domain gaps between different embodiments (i.e., human and robot) and training environments (i.e., real-world and simulated). Im2Flow2Act comprises two components: a flow generation network and a flow-conditioned policy. The flow generation network, trained on human demonstration videos, generates object flow from the initial scene image, conditioned on the task description. The flow-conditioned policy, trained on simulated robot play data, maps the generated object flow to robot actions to realize the desired object movements. By using flow as input, this policy can be directly deployed in the real world with a minimal sim-to-real gap. By leveraging real-world human videos and simulated robot play data, we bypass the challenges of teleoperating physical robots in the real world, resulting in a scalable system for diverse tasks. We demonstrate Im2Flow2Act's capabilities in a variety of real-world tasks, including the manipulation of rigid, articulated, and deformable objects.

177

Trust the PRoC3S: Solving Long-Horizon Robotics Problems with LLMs and Constraint Satisfaction
Aidan Curtis,Nishanth Kumar,Jing Cao,Tomás Lozano-Pérez,Leslie Pack Kaelbling
https://openreview.net/forum?id=r6ZhiVYriY
Recent developments in pretrained large language models (LLMs) applied to robotics have demonstrated their capacity for sequencing a set of discrete skills to achieve open-ended goals in simple robotic tasks. In this paper, we examine the topic of LLM planning for a set of continuously parameterized skills whose execution must avoid violations of a set of kinematic, geometric, and physical constraints. We prompt the LLM to output code for a function with open parameters, which, together with environmental constraints, can be viewed as a Continuous Constraint Satisfaction Problem (CCSP). This CCSP can be solved through sampling or optimization to find a skill sequence and continuous parameter settings that achieve the goal while avoiding constraint violations. Additionally, we consider cases where the LLM proposes unsatisfiable CCSPs, such as those that are kinematically infeasible, dynamically unstable, or lead to collisions, and re-prompt the LLM to form a new CCSP accordingly. Experiments across simulated and real-world domains demonstrate that our proposed strategy, \OursNoSpace, is capable of solving a wide range of complex manipulation tasks with realistic constraints much more efficiently and effectively than existing baselines.

178

Large Scale Mapping of Indoor Magnetic Field by Local and Sparse Gaussian Processes
Iad ABDUL-RAOUF,Vincent Gay-Bellile,Cyril JOLY,Steve Bourgeois,Alexis Paljic
https://openreview.net/forum?id=edP2dmingV
Magnetometer-based indoor navigation uses variations in the magnetic field to determine the robot's location. For that, a magnetic map of the environment has to be built beforehand from a collection of localized magnetic measurements. Existing solutions built on sparse Gaussian Process (GP) regression do not scale well to large environments, being either slow or resulting in discontinuous prediction. In this paper, we propose to model the magnetic field of large environments based on GP regression. We first modify a deterministic training conditional sparse GP by accounting for magnetic field physics to map small environments efficiently. We then scale the model on larger scenes by introducing a local expert aggregation framework. It splits the scene into subdomains, fits a local expert on each, and then aggregates expert predictions in a differentiable and probabilistic way. We evaluate our model on real and simulated data and show that we can smoothly map a three-story building in a few hundred milliseconds.

179

Gaitor: Learning a Unified Representation Across Gaits for Real-World Quadruped Locomotion
Alexander Luis Mitchell,Wolfgang Merkt,Aristotelis Papatheodorou,Ioannis Havoutis,Ingmar Posner
https://openreview.net/forum?id=ySI0tBYxpz
The current state-of-the-art in quadruped locomotion is able to produce a variety of complex motions. These methods either rely on switching between a discrete set of skills or learn a distribution across gaits using complex black-box models. Alternatively, we present Gaitor, which learns a disentangled and 2D representation across locomotion gaits. This learnt representation forms a planning space for closed-loop control delivering continuous gait transitions and perceptive terrain traversal. Gaitor's latent space is readily interpretable and we discover that during gait transitions, novel unseen gaits emerge. The latent space is disentangled with respect to footswing heights and lengths. This means that these gait characteristics can be varied independently in the 2D latent representation. Together with a simple terrain encoding and a learnt planner operating in the latent space, Gaitor can take motion commands including desired gait type and swing characteristics all while reacting to uneven terrain. We evaluate Gaitor in both simulation and the real world on the ANYmal C platform. To the best of our knowledge, this is the first work learning a unified and interpretable latent space for multiple gaits, resulting in continuous blending between different locomotion modes on a real quadruped robot. An overview of the methods and results in this paper is found at https://youtu.be/eVFQbRyilCA.

180

ManiWAV: Learning Robot Manipulation from In-the-Wild Audio-Visual Data
Zeyi Liu,Cheng Chi,Eric Cousineau,Naveen Kuppuswamy,Benjamin Burchfiel,Shuran Song
https://openreview.net/forum?id=wSWMsjuMTI
Audio signals provide rich information for the robot interaction and object properties through contact. These information can surprisingly ease the learning of contact-rich robot manipulation skills, especially when the visual information alone is ambiguous or incomplete. However, the usage of audio data in robot manipulation has been constrained to teleoperated demonstrations collected by either attaching a microphone to the robot or object, which significantly limits its usage in robot learning pipelines. In this work, we introduce ManiWAV: an 'ear-in-hand' data collection device to collect in-the-wild human demonstrations with synchronous audio and visual feedback, and a corresponding policy interface to learn robot manipulation policy directly from the demonstrations. We demonstrate the capabilities of our system through four contact-rich manipulation tasks that require either passively sensing the contact events and modes, or actively sensing the object surface materials and states. In addition, we show that our system can generalize to unseen in-the-wild environments, by learning from diverse in-the-wild human demonstrations. All data, code, and policy will be public.

181

UMI-on-Legs: Making Manipulation Policies Mobile with Manipulation-Centric Whole-body Controllers
Huy Ha,Yihuai Gao,Zipeng Fu,Jie Tan,Shuran Song
https://openreview.net/forum?id=3i7j8ZPnbm
We introduce UMI-on-Legs, a new framework that combines real-world and simulation data for quadruped manipulation systems. We scale task-centric data collection in the real world using a handheld gripper (UMI), providing a cheap way to demonstrate task-relevant manipulation skills without a robot. Simultaneously, we scale robot-centric data in simulation by training a whole-

body controller. The interface between these two policies are end-effector trajectories in the task-frame, which are inferred by the manipulation policy and passed to the whole-body controller for tracking. We evaluate UMI-on-Legs on prehensile, non-prehensile, and dynamic manipulation tasks, and report over 70% success rate for all tasks. Lastly, we also demonstrate the zero-shot cross-embodiment deployment of a pre-trained manipulation policy checkpoint from a prior work, originally intended for a fixed-base robot arm, on our quadruped system. We believe this framework provides a scalable path towards learning expressive manipulation skills on dynamic robot embodiments.

182

A3VLM: Actionable Articulation-Aware Vision Language Model
Siyuan Huang,Haonan Chang,Yuhan Liu,Yimeng Zhu,Hao Dong,Abdeslam Boularias,Peng Gao,Hongsheng Li
https://openreview.net/forum?id=IyhS75loxe
Vision Language Models (VLMs) for robotics have received significant attention in recent years. As a VLM can understand robot observations and perform complex visual reasoning, it is regarded as a potential universal solution for general robotics challenges such as manipulation and navigation. However, previous robotics VLMs such as RT-1, RT-2, and ManipLLM have focused on directly learning robot actions. Such approaches require collecting a significant amount of robot interaction data, which is extremely costly in the real world. Thus, we propose A3VLM, an object-centric, actionable, articulation-aware vision language model. A3VLM focuses on the articulation structure and action affordances of objects. Its representation is robot-agnostic and can be translated into robot actions using simple action primitives. Extensive experiments in both simulation benchmarks and real-world settings demonstrate the effectiveness and stability of A3VLM.

183

SLR: Learning Quadruped Locomotion without Privileged Information
Shiyi Chen,Zeyu Wan,Shiyang Yan,Chun Zhang,Weiyi Zhang,Qiang Li,Debing Zhang,Fasih Ud Din Farrukh
https://openreview.net/forum?id=RMkdcKK7jq
Traditional reinforcement learning control for quadruped robots often relies on privileged information, demanding meticulous selection and precise estimation, thereby imposing constraints on the development process. This work proposes a Self-learning Latent Representation (SLR) method, which achieves high-performance control policy learning without the need for privileged information. To enhance the credibility of our proposed method's evaluation, SLR is compared with open-source code repositories of state-of-the-art algorithms, retaining the original authors' configuration parameters. Across four repositories, SLR consistently outperforms the reference results. Ultimately, the trained policy and encoder empower the quadruped robot to navigate steps, climb stairs, ascend rocks, and traverse various challenging terrains.

184

Modeling Drivers' Situational Awareness from Eye Gaze for Driving Assistance
Abhijat Biswas,Pranay Gupta,Shreeya Khurana,David Held,Henny Admoni
https://openreview.net/forum?id=MwZJ96OkI3
Intelligent driving assistance can alert drivers to objects in their environment; however, such systems require a model of drivers' situational awareness (SA) (what aspects of the scene they

are already aware of) to avoid unnecessary alerts. Moreover, collecting the data to train such an SA model is challenging: being an internal human cognitive state, driver SA is difficult to measure, and non-verbal signals such as eye gaze are some of the only outward manifestations of it. Traditional methods to obtain SA labels rely on probes that result in sparse, intermittent SA labels unsuitable for modeling a dense, temporally correlated process via machine learning. We propose a novel interactive labeling protocol that captures dense, continuous SA labels and use it to collect an object-level SA dataset in a VR driving simulator. Our dataset comprises 20 unique drivers' SA labels, driving data, and gaze (over 320 minutes of driving) which will be made public. Additionally, we train an SA model from this data, formulating the object-level driver SA prediction problem as a semantic segmentation problem. Our formulation allows all objects in a scene at a timestep to be processed simultaneously, leveraging global scene context and local gaze-object relationships together. Our experiments show that this formulation leads to improved performance over common sense baselines and prior art on the SA prediction task.

185

FlowBotHD: History-Aware Diffuser Handling Ambiguities in Articulated Objects Manipulation
Yishu Li,Wen Hui Leng,Yiming Fang,Ben Eisner,David Held
https://openreview.net/forum?id=3ZAgXBRvla
We introduce a novel approach to manipulate articulated objects with ambiguities, such as opening a door, in which multi-modality and occlusions create ambiguities about the opening side and direction. Multi-modality occurs when the method to open a fully closed door (push, pull, slide) is uncertain, or the side from which it should be opened is uncertain. Occlusions further obscure the door’s shape from certain angles, creating further ambiguities during the occlusion. To tackle these challenges, we propose a history-aware diffusion network that models the multi-modal distribution of the articulated object and uses history to disambiguate actions and make stable predictions under occlusions. Experiments and analysis demonstrate the state-of-art performance of our method and specifically improvements in ambiguity-caused failure modes. Our project website is available at https://flowbothd.github.io/.

186

Uncertainty-Aware Decision Transformer for Stochastic Driving Environments
Zenan Li,Fan Nie,Qiao Sun,Fang Da,Hang Zhao
https://openreview.net/forum?id=LiwdXkMsDv
Offline Reinforcement Learning (RL) enables policy learning without active interactions, making it especially appealing for self-driving tasks. Recent successes of Transformers inspire casting offline RL as sequence modeling, which, however, fails in stochastic environments with incorrect assumptions that identical actions can consistently achieve the same goal. In this paper, we introduce an UNcertainty-awaRE deciSion Transformer (UNREST) for planning in stochastic driving environments without introducing additional transition or complex generative models. Specifically, UNREST estimates uncertainties by conditional mutual information between transitions and returns. Discovering 'uncertainty accumulation' and 'temporal locality' properties of driving environments, we replace the global returns in decision transformers with truncated returns less affected by environments to learn from actual outcomes of actions rather than environment transitions. We also dynamically evaluate uncertainty at inference for cautious planning. Extensive experiments demonstrate UNREST's superior performance in various driving scenarios and the power of our uncertainty estimation strategy.

187

EscIRL: Evolving Self-Contrastive IRL for Trajectory Prediction in Autonomous Driving
Siyue Wang,Zhaorun Chen,Zhuokai Zhao,Chaoli Mao,Yiyang Zhou,Jiayu He,Albert Sibo Hu
https://openreview.net/forum?id=1IzW0aniyg
While deep neural networks (DNN) and inverse reinforcement learning (IRL) have both been commonly used in autonomous driving to predict trajectories through learning from expert demonstrations, DNN-based methods suffer from data-scarcity, while IRL-based approaches often struggle with generalizability, making both hard to apply to new driving scenarios. To address these issues, we introduce EscIRL, a novel decoupled bi-level training framework that iteratively learns robust reward models from only a few mixed-scenario demonstrations. At the inner level, EscIRL introduces a self-contrastive IRL module that learns a spectrum of specialized reward functions by contrasting demonstrations across different scenarios. At the outer level, ESCIRL employs an evolving loop that iteratively refines the contrastive sets, ensuring global convergence. Experiments on two multi-scenario datasets, CitySim and INTERACTION, demonstrate the effectiveness of EscIRL, outperforming state-of-the-art DNN and IRL-based methods by 41.3% on average. Notably, we show that EscIRL achieves superior generalizability compared to DNN-based approaches while requiring only a small fraction of the data, effectively addressing data-scarcity constraints. All code and data are available at https://github.com/SiyueWang-CiDi/EscIRL.

188

Differentiable Discrete Elastic Rods for Real-Time Modeling of Deformable Linear Objects
Yizhou Chen,Yiting Zhang,Zachary Brei,Tiancheng Zhang,Yuzhen Chen,Julie Wu,Ram Vasudevan
https://openreview.net/forum?id=V5x0m6XDSV
This paper addresses the task of modeling Deformable Linear Objects (DLOs), such as ropes and cables, during dynamic motion over long time horizons. This task presents significant challenges due to the complex dynamics of DLOs. To address these challenges, this paper proposes differentiable Discrete Elastic Rods For deformable linear Objects with Real-time Modeling (DEFORM), a novel framework that combines a differentiable physics-based model with a learning framework to model DLOs accurately and in real-time. The performance of DEFORM is evaluated in an experimental setup involving two industrial robots and a variety of sensors. A comprehensive series of experiments demonstrate the efficacy of DEFORM in terms of accuracy, computational speed, and generalizability when compared to state-of-the-art alternatives. To further demonstrate the utility of DEFORM, this paper integrates it into a perception pipeline and illustrates its superior performance when compared to the state-of-the-art methods while tracking a DLO even in the presence of occlusions. Finally, this paper illustrates the superior performance of DEFORM when compared to state-of-the-art methods when it is applied to perform autonomous planning and control of DLOs.

189

Language-guided Manipulator Motion Planning with Bounded Task Space
Thies Oelerich,Christian Hartl-Nesic,Andreas Kugi
https://openreview.net/forum?id=yYujuPxjDK
Language-based robot control is a powerful and versatile method to control a robot manipulator where large language models (LLMs) are used to reason about the environment. However, the generated robot motions by these controllers often lack safety and performance, resulting in jerky movements. In this work, a novel modular framework for zero-shot motion planning for

manipulation tasks is developed. The modular components do not require any motion-planning-specific training. An LLM is combined with a vision model to create Python code that interacts with a novel path planner, which creates a piecewise linear reference path with bounds around the path that ensure safety. An optimization-based planner, the BoundMPC framework, is utilized to execute optimal, safe, and collision-free trajectories along the reference path. The effectiveness of the approach is shown on various everyday manipulation tasks in simulation and experiment, shown in the video at www.acin.tuwien.ac.at/42d2.

190

Task Success Prediction for Open-Vocabulary Manipulation Based on Multi-Level Aligned Representations

Miyu Goko,Motonari Kambara,Daichi Saito,Seitaro Otsuki,Komei Sugiura

https://openreview.net/forum?id=QtCtY8zl2T

In this study, we consider the problem of predicting task success for open-vocabulary manipulation by a manipulator, based on instruction sentences and egocentric images before and after manipulation. Conventional approaches, including multimodal large language models (MLLMs), often fail to appropriately understand detailed characteristics of objects and/or subtle changes in the position of objects. We propose Contrastive λ-Repformer, which predicts task success for table-top manipulation tasks by aligning images with instruction sentences. Our method integrates the following three key types of features into a multi-level aligned representation: features that preserve local image information; features aligned with natural language; and features structured through natural language. This allows the model to focus on important changes by looking at the differences in the representation between two images. We evaluate Contrastive λ-Repformer on a dataset based on a large-scale standard dataset, the RT-1 dataset, and on a physical robot platform. The results show that our approach outperformed existing approaches including MLLMs. Our best model achieved an improvement of 8.66 points in accuracy compared to the representative MLLM-based model.

191

Control with Patterns: A D-learning Method

Quan Quan,Kai-Yuan Cai,Chenyu Wang

https://openreview.net/forum?id=qoebyrnF36

Learning-based control policies are widely used in various tasks in the field of robotics and control. However, formal (Lyapunov) stability guarantees for learning-based controllers with nonlinear dynamical systems are challenging to obtain. We propose a novel control approach, namely Control with Patterns (CWP), to address the stability issue over data sets corresponding to nonlinear dynamical systems. For data sets of this kind, we introduce a new definition, namely exponential attraction on data sets, to describe nonlinear dynamical systems under consideration. The problem of exponential attraction on data sets is converted to a pattern classification one based on the data sets and parameterized Lyapunov functions. Furthermore, D-learning is proposed as a method for performing CWP without knowledge of the system dynamics. Finally, the effectiveness of CWP based on D-learning is demonstrated through simulations and real flight experiments. In these experiments, the position of the multicopter is stabilized using only real-time images as feedback, which can be considered as an Image-Based Visual Servoing (IBVS) problem.

192

ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation
Wenlong Huang,Chen Wang,Yunzhu Li,Ruohan Zhang,Li Fei-Fei
https://openreview.net/forum?id=9iG3SEbMnL
Representing robotic manipulation tasks as constraints that associate the robot and the environment is a promising way to encode desired robot behaviors. However, it remains unclear how to formulate the constraints such that they are 1) versatile to diverse tasks, 2) free of manual labeling, and 3) optimizable by off-the-shelf solvers to produce robot actions in real-time. In this work, we introduce Relational Keypoint Constraints (ReKep), a visually-grounded representation for constraints in robotic manipulation. Specifically, ReKep are expressed as Python functions mapping a set of 3D keypoints in the environment to a numerical cost. We demonstrate that by representing a manipulation task as a sequence of Relational Keypoint Constraints, we can employ a hierarchical optimization procedure to solve for robot actions (represented by a sequence of end-effector poses in SE(3)) with a perception-action loop at a real-time frequency. Furthermore, in order to circumvent the need for manual specification of ReKep for each new task, we devise an automated procedure that leverages large vision models and vision-language models to produce ReKep from free-form language instructions and RGB-D observation. We present system implementations on a mobile single-arm platform and a stationary dual-arm platform that can perform a large variety of manipulation tasks, featuring multi-stage, in-the-wild, bimanual, and reactive behaviors, all without task-specific data or environment models.

193

A Dual Approach to Imitation Learning from Observations with Offline Datasets
Harshit Sikchi,Caleb Chuck,Amy Zhang,Scott Niekum
https://openreview.net/forum?id=uHdVI3QMr6
Demonstrations are an effective alternative to task specification for learning agents in settings where designing a reward function is difficult. However, demonstrating expert behavior in the action space of the agent becomes unwieldy when robots have complex, unintuitive morphologies. We consider the practical setting where an agent has a dataset of prior interactions with the environment and is provided with observation-only expert demonstrations. Typical learning from observations approaches have required either learning an inverse dynamics model or a discriminator as intermediate steps of training. Errors in these intermediate one-step models compound during downstream policy learning or deployment. We overcome these limitations by directly learning a multi-step utility function that quantifies how each action impacts the agent's divergence from the expert's visitation distribution. Using the principle of duality, we derive DILO (Dual Imitation Learning from Observations), an algorithm that can leverage arbitrary suboptimal data to learn imitating policies without requiring expert actions. DILO reduces the learning from observations problem to that of simply learning an actor and a critic, bearing similar complexity to vanilla offline RL. This allows DILO to gracefully scale to high dimensional observations, and demonstrate improved performance across the board.

194

Physically Embodied Gaussian Splatting: A Visually Learnt and Physically Grounded 3D Representation for Robotics
Jad Abou-Chakra,Krishan Rana,Feras Dayoub,Niko Suenderhauf
https://openreview.net/forum?id=AEq0onGrN2
For robots to robustly understand and interact with the physical world, it is highly beneficial to

have a comprehensive representation -- modelling geometry, physics, and visual observations -- that informs perception, planning, and control algorithms. We propose a novel dual "Gaussian-Particle" representation that models the physical world while (i) enabling predictive simulation of future states and (ii) allowing online correction from visual observations in a dynamic world. Our representation comprises particles that capture the geometrical aspect of objects in the world and can be used alongside a particle-based physics system to anticipate physically plausible future states. Attached to these particles are 3D Gaussians that render images from any viewpoint through a splatting process thus capturing the visual state. By comparing the predicted and observed images, our approach generates "visual forces" that correct the particle positions while respecting known physical constraints. By integrating predictive physical modeling with continuous visually-derived corrections, our unified representation reasons about the present and future while synchronizing with reality. We validate our approach on 2D and 3D tracking tasks as well as photometric reconstruction quality. Videos are found at https://embodied-gaussians.github.io/

195

TaMMa: Target-driven Multi-subscene Mobile Manipulation
Jiawei Hou,Tianyu Wang,Tongying Pan,Shouyan Wang,Xiangyang Xue,Yanwei Fu
https://openreview.net/forum?id=EiqQEsOMZt
For everyday service robotics, the ability to navigate back and forth based on tasks in multi-subscene environments and perform delicate manipulations is crucial and highly practical. While existing robotics primarily focus on complex tasks within a single scene or simple tasks across scalable scenes individually, robots consisting of a mobile base with a robotic arm face the challenge of efficiently representing multiple subscenes, coordinating the collaboration between the mobile base and the robotic arm, and managing delicate tasks in scalable environments. To address this issue, we propose Target-driven Multi-subscene Mobile Manipulation (\textit{TaMMa}), which efficiently handles mobile base movement and fine-grained manipulation across subscenes. Specifically, we obtain a reliable 3D Gaussian initialization of the whole scene using a sparse 3D point cloud with encoded semantics. Through querying the coarse Gaussians, we acquire the approximate pose of the target, navigate the mobile base to approach it, and reduce the scope of precise target pose estimation to the corresponding subscene. Optimizing while moving, we employ diffusion-based depth completion to optimize fine-grained Gaussians and estimate the target's refined pose. For target-driven manipulation, we adopt Gaussians inpainting to obtain precise poses for the origin and destination of the operation in a \textit{think before you do it} manner, enabling fine-grained manipulation. We conduct various experiments on a real robotic to demonstrate our method in effectively and efficiently achieving precise operation tasks across multiple tabletop subscenes.

196

Learning Compositional Behaviors from Demonstration and Language
Weiyu Liu,Neil Nie,Ruohan Zhang,Jiayuan Mao,Jiajun Wu
https://openreview.net/forum?id=fR1rCXjCQX
We introduce Behavior from Language and Demonstration (BLADE), a framework for long-horizon robotic manipulation by integrating imitation learning and model-based planning. BLADE leverages language-annotated demonstrations, extracts abstract action knowledge from large language models (LLMs), and constructs a library of structured, high-level action representations. These representations include preconditions and effects grounded in visual perception for each high-level action, along with corresponding controllers implemented as neural network-based policies.

BLADE can recover such structured representations automatically, without manually labeled states or symbolic definitions. BLADE shows significant capabilities in generalizing to novel situations, including novel initial states, external state perturbations, and novel goals. We validate the effectiveness of our approach both in simulation and on real robots with a diverse set of objects with articulated parts, partial observability, and geometric constraints.

197

RoboEXP: Action-Conditioned Scene Graph via Interactive Exploration for Robotic Manipulation
Hanxiao Jiang,Binghao Huang,Ruihai Wu,Zhuoran Li,Shubham Garg,Hooshang Nayyeri,Shenlong Wang,Yunzhu Li
https://openreview.net/forum?id=UHxPZgK33I
We introduce the novel task of interactive scene exploration, wherein robots autonomously explore environments and produce an action-conditioned scene graph (ACSG) that captures the structure of the underlying environment. The ACSG accounts for both low-level information (geometry and semantics) and high-level information (action-conditioned relationships between different entities) in the scene. To this end, we present the Robotic Exploration (RoboEXP) system, which incorporates the Large Multimodal Model (LMM) and an explicit memory design to enhance our system's capabilities. The robot reasons about what and how to explore an object, accumulating new information through the interaction process and incrementally constructing the ACSG. Leveraging the constructed ACSG, we illustrate the effectiveness and efficiency of our RoboEXP system in facilitating a wide range of real-world manipulation tasks involving rigid, articulated objects, nested objects, and deformable objects. Project Page: https://jianghanxiao.github.io/roboexp-web/

198

Gaussian Splatting to Real World Flight Navigation Transfer with Liquid Networks
Alex Quach,Makram Chahine,Alexander Amini,Ramin Hasani,Daniela Rus
https://openreview.net/forum?id=ubq7Co6Cbv
Simulators are powerful tools for autonomous robot learning as they offer scalable data generation, flexible design, and optimization of trajectories. However, transferring behavior learned from simulation data into the real world proves to be difficult, usually mitigated with compute-heavy domain randomization methods or further model fine-tuning. We present a method to improve generalization and robustness to distribution shifts in sim-to-real visual quadrotor navigation tasks. To this end, we first build a simulator by integrating Gaussian Splatting with quadrotor flight dynamics, and then, train robust navigation policies using Liquid neural networks. In this way, we obtain a full-stack imitation learning protocol that combines advances in 3D Gaussian splatting radiance field rendering, crafty programming of expert demonstration training data, and the task understanding capabilities of Liquid networks. Through a series of quantitative flight tests, we demonstrate the robust transfer of navigation skills learned in a single simulation scene directly to the real world. We further show the ability to maintain performance beyond the training environment under drastic distribution and physical environment changes. Our learned Liquid policies, trained on single target maneuvers curated from a photorealistic simulated indoor flight only, generalize to multi-step hikes onboard a real hardware platform outdoors.

199

WoCoCo: Learning Whole-Body Humanoid Control with Sequential Contacts
Chong Zhang,Wenli Xiao,Tairan He,Guanya Shi
https://openreview.net/forum?id=Czs2xH9114
Humanoid activities involving sequential contacts are crucial for complex robotic interactions and operations in the real world and are traditionally solved by model-based motion planning, which is time-consuming and often relies on simplified dynamics models. Although model-free reinforcement learning (RL) has become a powerful tool for versatile and robust whole-body humanoid control, it still requires tedious task-specific tuning and state machine design and suffers from long-horizon exploration issues in tasks involving contact sequences. In this work, we propose WoCoCo (Whole-Body Control with Sequential Contacts), a unified framework to learn whole-body humanoid control with sequential contacts by naturally decomposing the tasks into separate contact stages. Such decomposition facilitates simple and general policy learning pipelines through task-agnostic reward and sim-to-real designs, requiring only one or two task-related terms to be specified for each task. We demonstrated that end-to-end RL-based controllers trained with WoCoCo enable four challenging whole-body humanoid tasks involving diverse contact sequences in the real world without any motion priors: 1) versatile parkour jumping, 2) box loco-manipulation, 3) dynamic clap-and-tap dancing, and 4) cliffside climbing. We further show that WoCoCo is a general framework beyond humanoid by applying it in 22-DoF dinosaur robot loco-manipulation tasks. Website: lecar-lab.github.io/wococo/.

200

TieBot: Learning to Knot a Tie from Visual Demonstration through a Real-to-Sim-to-Real Approach
Weikun Peng,Jun Lv,Yuwei Zeng,Haonan Chen,Siheng Zhao,Jichen Sun,Cewu Lu,Lin Shao
https://openreview.net/forum?id=Si2krRESZb
The tie-knotting task is highly challenging due to the tie's high deformation and long-horizon manipulation actions. This work presents TieBot, a Real-to-Sim-to-Real learning from visual demonstration system for the robots to learn to knot a tie. We introduce the Hierarchical Feature Matching approach to estimate a sequence of tie's meshes from the demonstration video. With these estimated meshes used as subgoals, we first learn a teacher policy using privileged information. Then, we learn a student policy with point cloud observation by imitating teacher policy. Lastly, our pipeline applies learned policy to real-world execution. We demonstrate the effectiveness of TieBot in simulation and the real world. In the real-world experiment, a dual-arm robot successfully knots a tie, achieving 50% success rate among 10 trials. Videos can be found on https://tiebots.github.io/.

201

One Model to Drift Them All: Physics-Informed Conditional Diffusion Model for Driving at the Limits
Franck Djeumou,Thomas Jonathan Lew,NAN DING,Michael Thompson,Makoto Suminaka,Marcus Greiff,John Subosits
https://openreview.net/forum?id=0gDbaEtVrd
Enabling autonomous vehicles to reliably operate at the limits of handling— where tire forces are saturated — would improve their safety, particularly in scenarios like emergency obstacle avoidance or adverse weather conditions. However, unlocking this capability is challenging due to the task's dynamic nature and the high sensitivity to uncertain multimodal properties of the road, vehicle, and their dynamic interactions. Motivated by these challenges, we propose a framework

to learn a conditional diffusion model for high-performance vehicle control using an unlabelled multimodal trajectory dataset. We design the diffusion model to capture the distribution of parameters of a physics-informed data-driven dynamics model. By conditioning the generation process on online measurements, we integrate the diffusion model into a real-time model predictive control framework for driving at the limits, and show that it can adapt on the fly to a given vehicle and environment. Extensive experiments on a Toyota Supra and a Lexus LC 500 show that a single diffusion model enables reliable autonomous drifting on both vehicles when operating with different tires in varying road conditions. The model matches the performance of task-specific expert models while outperforming them in generalization to unseen conditions, paving the way towards a general, reliable method for autonomous driving at the limits of handling.

202

RobotKeyframing: Learning Locomotion with High-Level Objectives via Mixture of Dense and Sparse Rewards
Fatemeh Zargarbashi,Jin Cheng,Dongho Kang,Robert Sumner,Stelian Coros
https://openreview.net/forum?id=wcbrhPnOei
This paper presents a novel learning-based control framework that uses keyframing to incorporate high-level objectives in natural locomotion for legged robots. These high-level objectives are specified as a variable number of partial or complete pose targets that are spaced arbitrarily in time. Our proposed framework utilizes a multi-critic reinforcement learning algorithm to effectively handle the mixture of dense and sparse rewards. Additionally, it employs a transformer-based encoder to accommodate a variable number of input targets, each associated with specific time-to-arrivals. Throughout simulation and hardware experiments, we demonstrate that our framework can effectively satisfy the target keyframe sequence at the required times. In the experiments, the multi-critic method significantly reduces the effort of hyperparameter tuning compared to the standard single-critic alternative. Moreover, the proposed transformer-based architecture enables robots to anticipate future goals, which results in quantitative improvements in their ability to reach their targets.

203

FREA: Feasibility-Guided Generation of Safety-Critical Scenarios with Reasonable Adversariality
Keyu Chen,Yuheng Lei,Hao Cheng,Haoran Wu,Wenchao Sun,Sifa Zheng
https://openreview.net/forum?id=3bcujpPikC
Generating safety-critical scenarios, which are essential yet difficult to collect at scale, offers an effective method to evaluate the robustness of autonomous vehicles (AVs). Existing methods focus on optimizing adversariality while preserving the naturalness of scenarios, aiming to achieve a balance through data-driven approaches. However, without an appropriate upper bound for adversariality, the scenarios might exhibit excessive adversariality, potentially leading to unavoidable collisions. In this paper, we introduce FREA, a novel safety-critical scenarios generation method that incorporates the Largest Feasible Region (LFR) of AV as guidance to ensure the reasonableness of the adversarial scenarios. Concretely, FREA initially pre-calculates the LFR of AV from offline datasets. Subsequently, it learns a reasonable adversarial policy that controls critical background vehicles (CBVs) in the scene to generate adversarial yet AV-feasible scenarios by maximizing a novel feasibility-dependent objective function. Extensive experiments illustrate that FREA can effectively generate safety-critical scenarios, yielding considerable near-miss events while ensuring AV's feasibility. Generalization analysis also confirms the robustness of FREA in AV testing across various surrogate AV methods and traffic environments.

204

Non-rigid Relative Placement through 3D Dense Diffusion
Eric Cai,Octavian Donca,Ben Eisner,David Held
https://openreview.net/forum?id=rvKWXxIvj0
The task of "relative placement" is to predict the placement of one object in relation to another, e.g. placing a mug on a mug rack. Recent methods for relative placement have made tremendous progress towards data-efficient learning for robot manipulation; using explicit object-centric geometric reasoning, these approaches enable generalization to unseen task variations from a small number of demonstrations. State-of-the-art works in this area, however, have yet to represent deformable transformations, despite the ubiquity of non-rigid bodies in real world settings. As a first step towards bridging this gap, we propose "cross-displacement" - an extension of the principles of relative placement to geometric relationships between deformable objects - and present a novel vision-based method to learn cross-displacement for a non-rigid task through dense diffusion. To this end, we demonstrate our method's ability to generalize to unseen object instances, out-of-distribution scene configurations, and multimodal goals on a highly deformable cloth-hanging task beyond the scope of prior works.

205

General Flow as Foundation Affordance for Scalable Robot Learning
Chengbo Yuan,Chuan Wen,Tong Zhang,Yang Gao
https://openreview.net/forum?id=nmEt0ci8hi
We address the challenge of acquiring real-world manipulation skills with a scalable framework. We hold the belief that identifying an appropriate prediction target capable of leveraging large-scale datasets is crucial for achieving efficient and universal learning. Therefore, we propose to utilize 3D flow, which represents the future trajectories of 3D points on objects of interest, as an ideal prediction target.

206

CoViS-Net: A Cooperative Visual Spatial Foundation Model for Multi-Robot Applications
Jan Blumenkamp,Steven Morad,Jennifer Gielis,Amanda Prorok
https://openreview.net/forum?id=KULBk5q24a
Autonomous robot operation in unstructured environments is often underpinned by spatial understanding through vision. Systems composed of multiple concurrently operating robots additionally require access to frequent, accurate and reliable pose estimates. Classical vision-based methods to regress relative pose are commonly computationally expensive (precluding real-time applications), and often lack data-derived priors for resolving ambiguities. In this work, we propose CoViS-Net, a cooperative, multi-robot visual spatial foundation model that learns spatial priors from data, enabling pose estimation as well as general spatial comprehension. Our model is fully decentralized, platform-agnostic, executable in real-time using onboard compute, and does not require existing networking infrastructure. CoViS-Net provides relative pose estimates and a local bird's-eye-view (BEV) representation, even without camera overlap between robots, and can predict BEV representations of unseen regions. We demonstrate its use in a multi-robot formation control task across various real-world settings. We provide supplementary material online and will open source our trained model in due course. https://sites.google.com/view/covis-net

207

Splat-MOVER: Multi-Stage, Open-Vocabulary Robotic Manipulation via Editable Gaussian Splatting
Olaolu Shorinwa,Johnathan Tucker,Aliyah Smith,Aiden Swann,Timothy Chen,Roya Firoozi,Monroe David Kennedy,Mac Schwager
https://openreview.net/forum?id=8XFT1PatHy
We present Splat-MOVER, a modular robotics stack for open-vocabulary robotic manipulation, which leverages the editability of Gaussian Splatting (GSplat) scene representations to enable multi-stage manipulation tasks. Splat-MOVER consists of: (i) ASK-Splat, a GSplat representation that distills semantic and grasp affordance features into the 3D scene. ASK-Splat enables geometric, semantic, and affordance understanding of 3D scenes, which is critical for many robotics tasks; (ii) SEE-Splat, a real-time scene-editing module using 3D semantic masking and infilling to visualize the motions of objects that result from robot interactions in the real-world. SEE-Splat creates a "digital twin" of the evolving environment throughout the manipulation task; and (iii) Grasp-Splat, a grasp generation module that uses ASK-Splat and SEE-Splat to propose affordance-aligned candidate grasps for open-world objects. ASK-Splat is trained in real-time from RGB images in a brief scanning phase prior to operation, while SEE-Splat and Grasp-Splat run in real-time during operation. We demonstrate the superior performance of Splat-MOVER in hardware experiments on a Kinova robot compared to two recent baselines in four single-stage, open-vocabulary manipulation tasks. In addition, we demonstrate Splat-MOVER in four multi-stage manipulation tasks, using the edited scene to reflect changes due to prior manipulation stages, which is not possible with existing baselines. Video demonstrations and the code for the project are available at https://splatmover.github.io.

208

Theia: Distilling Diverse Vision Foundation Models for Robot Learning
Jinghuan Shang,Karl Schmeckpeper,Brandon B. May,Maria Vittoria Minniti,Tarik Kelestemur,David Watkins,Laura Herlant
https://openreview.net/forum?id=ylZHvlwUcI
Vision-based robot policy learning, which maps visual inputs to actions, necessitates a holistic understanding of diverse visual tasks beyond single-task needs like classification or segmentation. Inspired by this, we introduce Theia, a vision foundation model for robot learning that distills multiple off-the-shelf vision foundation models trained on varied vision tasks. Theia's rich visual representations encode diverse visual knowledge, enhancing downstream robot learning. Extensive experiments demonstrate that Theia outperforms its teacher models and prior robot learning models using less training data and smaller model sizes. Additionally, we quantify the quality of pre-trained visual representations and hypothesize that higher entropy in feature norm distributions leads to improved robot learning performance. Code, models, and demo are available at https://theia.theaiinstitute.com.

209

3D-ViTac: Learning Fine-Grained Manipulation with Visuo-Tactile Sensing
Binghao Huang,Yixuan Wang,Xinyi Yang,Yiyue Luo,Yunzhu Li
https://openreview.net/forum?id=bk28WlkqZn
Tactile and visual perception are both crucial for humans to perform fine-grained interactions with their environment. Developing similar multi-modal sensing capabilities for robots can significantly enhance and expand their manipulation skills. This paper introduces 3D-ViTac, a multi-modal sensing and learning system designed for dexterous bimanual manipulation. Our system features

tactile sensors equipped with dense sensing units, each covering an area of $3mm^2$. These sensors are low-cost and flexible, providing detailed and extensive coverage of physical contacts, effectively complementing visual information. To integrate tactile and visual data, we fuse them into a unified 3D representation space that preserves their 3D structures and spatial relationships. The multi-modal representation can then be coupled with diffusion policies for imitation learning. Through concrete hardware experiments, we demonstrate that even low-cost robots can perform precise manipulations and significantly outperform vision-only policies, particularly in safe interactions with fragile items and executing long-horizon tasks involving in-hand manipulation. Our project page is available at https://binghao-huang.github.io/3D-ViTac/.

210

Adaptive Diffusion Terrain Generator for Autonomous Uneven Terrain Navigation
Youwei Yu,Junhong Xu,Lantao Liu
https://openreview.net/forum?id=xYIeTh2QhS
Model-free reinforcement learning has emerged as a powerful method for developing robust robot control policies capable of navigating through complex and unstructured terrains. The effectiveness of these methods hinges on two essential elements: (1) the use of massively parallel physics simulations to expedite policy training, and (2) the deployment of an environment generator tasked with crafting terrains that are sufficiently challenging yet attainable, thereby facilitating continuous policy improvement. Existing methods of environment generation often rely on heuristics constrained by a set of parameters, limiting the diversity and realism. In this work, we introduce the Adaptive Diffusion Terrain Generator (ADTG), a novel method that leverages Denoising Diffusion Probabilistic Models (DDPMs) to dynamically expand an existing training environment by adding more diverse and complex terrains tailored to the current policy. Unlike conventional methods, ADTG adapts the terrain complexity and variety based on the evolving capabilities of the current policy. This is achieved through two primary mechanisms: First, by blending terrains from the initial dataset within their latent spaces using performance-informed weights, ADTG creates terrains that suitably challenge the policy. Secondly, by manipulating the initial noise in the diffusion process, ADTG seamlessly shifts between creating similar terrains for fine-tuning the current policy and entirely novel ones for expanding training diversity. Our experiments show that the policy trained by ADTG outperforms both procedural generated and natural environments, along with popular navigation methods.

211

A Planar-Symmetric SO(3) Representation for Learning Grasp Detection
Tianyi Ko,Takuya Ikeda,Hiroya Sato,Koichi Nishiwaki
https://openreview.net/forum?id=LmOF7UAOZ7
Planar-symmetric hands, such as parallel grippers, are widely adopted in both research and industrial fields. Their symmetry, however, introduces ambiguity and discontinuity in the SO(3) representation, which hinders both the training and inference of neural network-based grasp detectors. We propose a novel SO(3) representation that can parametrize a pair of planar-symmetric poses with a single parameter set by leveraging the 2D Bingham distribution. We also detail a grasp detector based on our representation, which provides a more consistent rotation output. An intensive evaluation with multiple grippers and objects in both the simulation and the real world quantitatively shows our approach's contribution.

212

Sim-to-Real Transfer via 3D Feature Fields for Vision-and-Language Navigation
Zihan Wang,Xiangyang Li,Jiahao Yang,Yeqi Liu,Shuqiang Jiang
https://openreview.net/forum?id=VFs1vbQnYN
Vision-and-language navigation (VLN) enables the agent to navigate to a remote location in 3D environments following the natural language instruction. In this field, the agent is usually trained and evaluated in the navigation simulators, lacking effective approaches for sim-to-real transfer. The VLN agents with only a monocular camera exhibit extremely limited performance, while the mainstream VLN models trained with panoramic observation, perform better but are difficult to deploy on most monocular robots. For this case, we propose a sim-to-real transfer approach to endow the monocular robots with panoramic traversability perception and panoramic semantic understanding, thus smoothly transferring the high-performance panoramic VLN models to the common monocular robots. In this work, the semantic traversable map is proposed to predict agent-centric navigable waypoints, and the novel view representations of these navigable waypoints are predicted through the 3D feature fields. These methods broaden the limited field of view of the monocular robots and significantly improve navigation performance in the real world. Our VLN system outperforms previous SOTA monocular VLN methods in R2R-CE and RxR-CE benchmarks within the simulation environments and is also validated in real-world environments, providing a practical and high-performance solution for real-world VLN.

213

Lessons from Learning to Spin “Pens”
Jun Wang,Ying Yuan,Haichuan Che,Haozhi Qi,Yi Ma,Jitendra Malik,Xiaolong Wang
https://openreview.net/forum?id=SFJz5iLvur
In-hand manipulation of pen-like objects is a most basic and important skill in our daily lives, as many tools such as hammers and screwdrivers are similarly shaped. However, current learning-based methods struggle with this task due to a lack of high-quality demonstrations and the significant gap between simulation and the real world. In this work, we push the boundaries of learning-based in-hand manipulation systems by demonstrating the capability to spin pen-like objects. We use reinforcement learning to train a policy and generate a high-fidelity trajectory dataset in simulation. This serves two purposes: 1) pre-training a sensorimotor policy in simulation; 2) conducting open-loop trajectory replay in the real world. We then fine-tune the sensorimotor policy using these real-world trajectories to adapt to the real world. With less than 50 trajectories, our policy learns to rotate more than ten pen-like objects with different physical properties for multiple revolutions. We present a comprehensive analysis of our design choices and share the lessons learned during development. Videos are shown on https://corl-2024-dexpen.github.io/.

214

Play to the Score: Stage-Guided Dynamic Multi-Sensory Fusion for Robotic Manipulation
Ruoxuan Feng,Di Hu,Wenke Ma,Xuelong Li
https://openreview.net/forum?id=N5IS6DzBmL
Humans possess a remarkable talent for flexibly alternating to different senses when interacting with the environment. Picture a chef skillfully gauging the timing of ingredient additions and controlling the heat according to the colors, sounds, and aromas, seamlessly navigating through every stage of the complex cooking process. This ability is founded upon a thorough comprehension of task stages, as achieving the sub-goal within each stage can necessitate the

utilization of different senses. In order to endow robots with similar ability, we incorporate the task stages divided by sub-goals into the imitation learning process to accordingly guide dynamic multi-sensory fusion. We propose MS-Bot, a stage-guided dynamic multi-sensory fusion method with coarse-to-fine stage understanding, which dynamically adjusts the priority of modalities based on the fine-grained state within the predicted current stage. We train a robot system equipped with visual, auditory, and tactile sensors to accomplish challenging robotic manipulation tasks: pouring and peg insertion with keyway. Experimental results indicate that our approach enables more effective and explainable dynamic fusion, aligning more closely with the human fusion process than existing methods.

215

Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks
Jialiang Zhao,Yuxiang Ma,Lirui Wang,Edward Adelson
https://openreview.net/forum?id=KXsropnmNI
This paper presents T3: Transferable Tactile Transformers, a framework for tactile representation learning that scales across multi-sensors and multi-tasks.T3 is designed to overcome the contemporary issue that camera-based tactile sensing is extremely heterogeneous, i.e. sensors are built into different form factors, and existing datasets were collected for disparate tasks. T3 captures the shared latent information across different sensor-task pairings by constructing a shared trunk transformer with sensor-specific encoders and task-specific decoders. The pre-training of T3utilizes a novel Foundation Tactile (FoTa) dataset, which is aggregated from several open-sourced datasets and it contains over 3 million data points gathered from 13 sensors and 11 tasks. FoTa is the largest and most diverse dataset in tactile sensing to date and it is made publicly available in a unified format. Across various sensors and tasks, experiments show that T3 pre-trained with FoTa achieved zero-shot transferability in certain sensor-task pairings, can be further fine-tuned with small amounts of domain-specific data, and its performance scales with bigger network sizes. T3 is also effective as a tactile encoder for long horizon contact-rich manipulation. Results from sub-millimeter multi-pin electronics insertion tasks show that T3 achieved a task success rate 25% higher than that of policies trained with tactile encoders trained from scratch, or 53% higher than without tactile sensing. Data, code, and model checkpoints are open-sourced at https://t3.alanz.info.

216

Learning Granular Media Avalanche Behavior for Indirectly Manipulating Obstacles on a Granular Slope
Haodi Hu,Feifei Qian,Daniel Seita
https://openreview.net/forum?id=Qz2N4lWBk3
Legged robot locomotion on sand slopes is challenging due to the complex dynamics of granular media and how the lack of solid surfaces can hinder locomotion. A promising strategy, inspired by ghost crabs and other organisms in nature, is to strategically interact with rocks, debris, and other obstacles to facilitate movement. To provide legged robots with this ability, we present a novel approach that leverages avalanche dynamics to indirectly manipulate objects on a granular slope. We use a Vision Transformer (ViT) to process image representations of granular dynamics and robot excavation actions. The ViT predicts object movement, which we use to determine which leg excavation action to execute. We collect training data from 100 real physical trials and, at test time, deploy our trained model in novel settings. Experimental results suggest that our model can accurately predict object movements and achieve a success rate ≥ 80% in a variety of manipulation tasks with up to four obstacles, and can also generalize to objects with different

physics properties. To our knowledge, this is the first paper to leverage granular media avalanche dynamics to indirectly manipulate objects on granular slopes. Supplementary material is available at https://sites.google.com/view/grain-corl2024/home.

217

EXTRACT: Efficient Policy Learning by Extracting Transferable Robot Skills from Offline Data
Jesse Zhang,Minho Heo,Zuxin Liu,Erdem Biyik,Joseph J Lim,Yao Liu,Rasool Fakoor
https://openreview.net/forum?id=uEbJXWobif
Most reinforcement learning (RL) methods focus on learning optimal policies over low-level action spaces. While these methods can perform well in their training environments, they lack the flexibility to transfer to new tasks. Instead, RL agents that can act over useful, temporally extended skills rather than low-level actions can learn new tasks more easily. Prior work in skill-based RL either requires expert supervision to define useful skills, which is hard to scale, or learns a skill-space from offline data with heuristics that limit the adaptability of the skills, making them difficult to transfer during downstream RL. Our approach, EXTRACT, instead utilizes pre-trained vision language models to extract a discrete set of semantically meaningful skills from offline data, each of which is parameterized by continuous arguments, without human supervision. This skill parameterization allows robots to learn new tasks by only needing to learn when to select a specific skill and how to modify its arguments for the specific task. We demonstrate through experiments in sparse-reward, image-based, robot manipulation environments that EXTRACT can more quickly learn new tasks than prior works, with major gains in sample efficiency and performance over prior skill-based RL.

218

OPEN TEACH: A Versatile Teleoperation System for Robotic Manipulation
Aadhithya Iyer,Zhuoran Peng,Yinlong Dai,Irmak Guzey,Siddhant Haldar,Soumith Chintala,Lerrel Pinto
https://openreview.net/forum?id=cvAIaS6V2I
Open-sourced, user-friendly tools form the bedrock of scientific advancement across disciplines. The widespread adoption of data-driven learning has led to remarkable progress in multi-fingered dexterity, bimanual manipulation, and applications ranging from logistics to home robotics. However, existing data collection platforms are often proprietary, costly, or tailored to specific robotic morphologies. We present OPEN TEACH, a new teleoperation system leveraging VR headsets to immerse users in mixed reality for intuitive robot control. built on the affordable Meta Quest 3, which costs $500, OPEN TEACH enables real-time control of various robots, including multi-fingered hands, bimanual arms, and mobile manipulators, through an easy-to-use app. Using natural hand gestures and movements, users can manipulate robots at up to 90Hz with smooth visual feedback and interface widgets offering closeup environment views. We demonstrate the versatility of OPEN TEACH across 38 tasks on different robots. A comprehensive user study indicates significant improvement in teleoperation capability over the AnyTeleop framework. Further experiments exhibit that the collected data is compatible with policy learning on 10 dexterous and contact-rich manipulation tasks. Currently supporting Franka, xArm, Jaco, Allegro, and Hello Stretch platforms, OPEN TEACH is fully open-sourced to promote broader adoption. Videos are available at https://anon-open-teach.github.io/.

219

TOP-Nav: Legged Navigation Integrating Terrain, Obstacle and Proprioception Estimation
Junli Ren,Yikai Liu,Yingru Dai,Junfeng Long,Guijin Wang
https://openreview.net/forum?id=O05tIQt2d5
Legged navigation is typically examined within open-world, off-road, and challenging environments. In these scenarios, estimating external disturbances requires a complex synthesis of multi-modal information. This underlines a major limitation in existing works that primarily focus on avoiding obstacles. In this work, we propose TOP-Nav, a novel legged navigation framework that integrates a comprehensive path planner with Terrain awareness, Obstacle avoidance and close-loop Proprioception. TOP-Nav underscores the synergies between vision and proprioception in both path and motion planning. Within the path planner, we present a terrain estimator that enables the robot to select waypoints on terrains with higher traversability while effectively avoiding obstacles. In the motion planning level, we construct a proprioception advisor from the learning-based locomotion controller to provide motion evaluations for the path planner. Based on the close-loop motion feedback, we offer online corrections for the vision-based terrain and obstacle estimations. Consequently, TOP-Nav achieves open-world navigation that the robot can handle terrains or disturbances beyond the distribution of prior knowledge and overcomes constraints imposed by visual conditions. Building upon extensive experiments conducted in both simulation and real-world environments, TOP-Nav demonstrates superior performance in open-world navigation compared to existing methods.

220

Promptable Closed-loop Traffic Simulation
Shuhan Tan,Boris Ivanovic,Yuxiao Chen,Boyi Li,Xinshuo Weng,Yulong Cao,Philipp Kraehenbuehl,Marco Pavone
https://openreview.net/forum?id=5iXG6EgByK
Simulation stands as a cornerstone for safe and efficient autonomous driving development. At its core a simulation system ought to produce realistic, reactive, and controllable traffic patterns. In this paper, we propose ProSim, a multimodal promptable closed-loop traffic simulation framework. ProSim allows the user to give a complex set of numerical, categorical or textual prompts to instruct each agent's behavior and intention. ProSim then rolls out a traffic scenario in a closed-loop manner, modeling each agent's interaction with other traffic participants. Our experiments show that ProSim achieves high prompt controllability given different user prompts, while reaching competitive performance on the Waymo Sim Agents Challenge when no prompt is given. To support research on promptable traffic simulation, we create ProSim-Instruct-520k, a multimodal prompt-scenario paired driving dataset with over 10M text prompts for over 520k real-world driving scenarios. We will release data, benchmark, and labeling tools of ProSim-Instruct-520k upon publication.

221

Reinforcement Learning with Foundation Priors: Let Embodied Agent Efficiently Learn on Its Own
Weirui Ye,Yunsheng Zhang,Haoyang Weng,Xianfan Gu,Shengjie Wang,Tong Zhang,Mengchen Wang,Pieter Abbeel,Yang Gao
https://openreview.net/forum?id=dsxmR6IYIg
Reinforcement learning (RL) is a promising approach for solving robotic manipulation tasks. However, it is challenging to apply the RL algorithms directly in the real world. For one thing, RL is data-intensive and typically requires millions of interactions with environments, which are

impractical in real scenarios. For another, it is necessary to make heavy engineering efforts to design reward functions manually. To address these issues, we leverage foundation models in this paper. We propose Reinforcement Learning with Foundation Priors (RLFP) to utilize guidance and feedback from policy, value, and success-reward foundation models. Within this framework, we introduce the Foundation-guided Actor-Critic (FAC) algorithm, which enables embodied agents to explore more efficiently with automatic reward functions. The benefits of our framework are threefold: (1) \textit{sample efficient}; (2) \textit{minimal and effective reward engineering}; (3) \textit{agnostic to foundation model forms and robust to noisy priors}. Our method achieves remarkable performances in various manipulation tasks on both real robots and in simulation. Across 5 dexterous tasks with real robots, FAC achieves an average success rate of 86% after one hour of real-time learning. Across 8 tasks in the simulated Meta-world, FAC achieves 100% success rates in 7/8 tasks under less than 100k frames (about 1-hour training), outperforming baseline methods with manual-designed rewards in 1M frames. We believe the RLFP framework can enable future robots to explore and learn autonomously in the physical world for more tasks.

222

Twisting Lids Off with Two Hands
Toru Lin,Zhao-Heng Yin,Haozhi Qi,Pieter Abbeel,Jitendra Malik
https://openreview.net/forum?id=3wBqoPfoeJ
Manipulating objects with two multi-fingered hands has been a long-standing challenge in robotics, due to the contact-rich nature of many manipulation tasks and the complexity inherent in coordinating a high-dimensional bimanual system. In this work, we share novel insights into physical modeling, real-time perception, and reward design that enable policies trained in simulation using deep reinforcement learning (RL) to be effectively and efficiently transferred to the real world. Specifically, we consider the problem of twisting lids of various bottle-like objects with two hands, demonstrating policies with generalization capabilities across a diverse set of unseen objects as well as dynamic and dexterous behaviors. To the best of our knowledge, this is the first sim-to-real RL system that enables such capabilities on bimanual multi-fingered hands.

223

RoboPoint: A Vision-Language Model for Spatial Affordance Prediction in Robotics
Wentao Yuan,Jiafei Duan,Valts Blukis,Wilbert Pumacay,Ranjay Krishna,Adithyavairavan Murali,Arsalan Mousavian,Dieter Fox
https://openreview.net/forum?id=GVX6jpZOhU
From rearranging objects on a table to putting groceries into shelves, robots must plan precise action points to perform tasks accurately and reliably. In spite of the recent adoption of vision language models (VLMs) to control robot behavior, VLMs struggle to precisely articulate robot actions using language. We introduce an automatic synthetic data generation pipeline that instruction-tunes VLMs to robotic domains and needs. Using the pipeline, we train RoboPoint, a VLM that predicts image keypoint affordances given language instructions. Compared to alternative approaches, our method requires no real-world data collection or human demonstration, making it much more scalable to diverse environments and viewpoints. In addition, RoboPoint is a general model that enables several downstream applications such as robot navigation, manipulation, and augmented reality (AR) assistance. Our experiments demonstrate that RoboPoint outperforms state-of-the-art VLMs (GPT-4o) and visual prompting techniques (PIVOT) by 21.8% in the accuracy of predicting spatial affordance and by 30.5% in the success rate of downstream tasks. Anonymous project page: https://robopoint.github.io.

224

Robot See Robot Do: Imitating Articulated Object Manipulation with Monocular 4D Reconstruction
Justin Kerr,Chung Min Kim,Mingxuan Wu,Brent Yi,Qianqian Wang,Ken Goldberg,Angjoo Kanazawa
https://openreview.net/forum?id=2LLu3gavF1
Humans can learn to manipulate new objects by simply watching others; providing robots with the ability to learn from such demonstrations would enable a natural interface specifying new behaviors. This work develops Robot See Robot Do (RSRD), a method for imitating articulated object manipulation from a single monocular RGB human demonstration given a single static multi- view object scan. We first propose 4D Differentiable Part Models (4D-DPM), a method for recovering 3D part motion from a monocular video with differentiable rendering. This analysis-by-synthesis approach uses part-centric feature fields in an iterative optimization which enables the use of geometric regularizers to re- cover 3D motions from only a single video. Given this 4D reconstruction, the robot replicates object trajectories by planning bimanual arm motions that induce the demonstrated object part motion. By representing demonstrations as part- centric trajectories, RSRD focuses on replicating the demonstration's intended behavior while considering the robot's own morphological limits, rather than at- tempting to reproduce the hand's motion. We evaluate 4D-DPM's 3D tracking accuracy on ground truth annotated 3D part trajectories and RSRD's physical ex- ecution performance on 9 objects across 10 trials each on a bimanual YuMi robot. Each phase of RSRD achieves an average of 87% success rate, for a total end- to-end success rate of 60% across 90 trials. Notably, this is accomplished using only feature fields distilled from large pretrained vision models — without any task-specific training, fine-tuning, dataset collection, or annotation. Project page: https://robot-see-robot-do.github.io

225

Continuous Control with Coarse-to-fine Reinforcement Learning
Younggyo Seo,Jafar Uruç,Stephen James
https://openreview.net/forum?id=WjDR48cL3O
Despite recent advances in improving the sample-efficiency of reinforcement learning (RL) algorithms, designing an RL algorithm that can be practically deployed in real-world environments remains a challenge. In this paper, we present Coarse-to-fine Reinforcement Learning (CRL), a framework that trains RL agents to zoom-into a continuous action space in a coarse-to-fine manner, enabling the use of stable, sample-efficient value-based RL algorithms for fine-grained continuous control tasks. Our key idea is to train agents that output actions by iterating the procedure of (i) discretizing the continuous action space into multiple intervals and (ii) selecting the interval with the highest Q-value to further discretize at the next level. We then introduce a concrete, value-based algorithm within the CRL framework called Coarse-to-fine Q-Network (CQN). Our experiments demonstrate that CQN significantly outperforms RL and behavior cloning baselines on 20 sparsely-rewarded RLBench manipulation tasks with a modest number of environment interactions and expert demonstrations. We also show that CQN robustly learns to solve real-world manipulation tasks within a few minutes of online training.

226

Generative Image as Action Models
Mohit Shridhar,Yat Long Lo,Stephen James
https://openreview.net/forum?id=cocHfT7CEs
Image-generation diffusion models have been fine-tuned to unlock new capabilities such as image-editing and novel view synthesis. Can we similarly unlock image-generation models for

visuomotor control? We present GENIMA, a behavior-cloning agent that fine-tunes Stable Diffusion to “draw joint-actions” as targets on RGB images. These images are fed into a controller that maps the visual targets into a sequence of joint-positions. We study GENIMA on 25 RLBench and 9 real-world manipulation tasks. We find that, by lifting actions into image-space, internet pre-trained diffusion models can generate policies that outperform state- of-the-art visuomotor approaches, especially in robustness to scene perturbations and generalizing to novel objects. Our method is also competitive with 3D agents, despite lacking priors such as depth, keypoints, or motion-planners.

227

Event3DGS: Event-Based 3D Gaussian Splatting for High-Speed Robot Egomotion
Tianyi Xiong,Jiayi Wu,Botao He,Cornelia Fermuller,Yiannis Aloimonos,Heng Huang,Christopher Metzler
https://openreview.net/forum?id=EyEE7547vy
By combining differentiable rendering with explicit point-based scene representations, 3D Gaussian Splatting (3DGS) has demonstrated breakthrough 3D reconstruction capabilities. However, to date 3DGS has had limited impact on robotics, where high-speed egomotion is pervasive: Egomotion introduces motion blur and leads to artifacts in existing frame-based 3DGS reconstruction methods. To address this challenge, we introduce Event3DGS, an event-based 3DGS framework. By exploiting the exceptional temporal resolution of event cameras, Event3GDS can reconstruct high-fidelity 3D structure and appearance under high-speed egomotion. Extensive experiments on multiple synthetic and real-world datasets demonstrate the superiority of Event3DGS compared with existing event-based dense 3D scene reconstruction frameworks; Event3DGS substantially improves reconstruction quality (+3dB) while reducing computational costs by 95%. Our framework also allows one to incorporate a few motion-blurred frame-based measurements into the reconstruction process to further improve appearance fidelity without loss of structural accuracy.

228

RiEMann: Near Real-Time SE(3)-Equivariant Robot Manipulation without Point Cloud Segmentation
Chongkai Gao,Zhengrong Xue,Shuying Deng,Tianhai Liang,Siqi Yang,Lin Shao,Huazhe Xu
https://openreview.net/forum?id=eJHy0AF5TO
We present RiEMann, an end-to-end near Real-time SE(3)-Equivariant Robot Manipulation imitation learning framework from scene point cloud input. Compared to previous methods that rely on descriptor field matching, RiEMann directly predicts the target actions for manipulation without any object segmentation. RiEMann can efficiently train the visuomotor policy from scratch with 5 to 10 demonstrations for a manipulation task, generalizes to unseen SE(3) transformations and instances of target objects, resists visual interference of distracting objects, and follows the near real-time pose change of the target object. The scalable SE(3)-equivariant action space of RiEMann supports both pick-and-place tasks and articulated object manipulation tasks. In simulation and real-world 6-DOF robot manipulation experiments, we test RiEMann on 5 categories of manipulation tasks with a total of 25 variants and show that RiEMann outperforms baselines in both task success rates and SE(3) geodesic distance errors (reduced by 68.6%), and achieves 5.4 frames per second (fps) network inference speed.

229

Implicit Grasp Diffusion: Bridging the Gap between Dense Prediction and Sampling-based Grasping
Pinhao Song,Pengteng Li,Renaud Detry
https://openreview.net/forum?id=VUhlMfEekm
There are two dominant approaches in modern robot grasp planning: dense prediction and sampling-based methods. Dense prediction calculates viable grasps across the robot's view but is limited to predicting one grasp per voxel. Sampling-based methods, on the other hand, encode multi-modal grasp distributions, allowing for different grasp approaches at a point. However, these methods rely on a global latent representation, which struggles to represent the entire field of view, resulting in coarse grasps. To address this, we introduce \emph{Implicit Grasp Diffusion} (IGD), which combines the strengths of both methods by using implicit neural representations to extract detailed local features and sampling grasps from diffusion models conditioned on these features. Evaluations on clutter removal tasks in both simulated and real-world environments show that IGD delivers high accuracy, noise resilience, and multi-modal grasp pose capabilities.

230

Shelf-Supervised Cross-Modal Pre-Training for 3D Object Detection
Mehar Khurana,Neehar Peri,James Hays,Deva Ramanan
https://openreview.net/forum?id=eeoX7tCoK2
State-of-the-art 3D object detectors are often trained on massive labeled datasets. However, annotating 3D bounding boxes remains prohibitively expensive and time-consuming, particularly for LiDAR. Instead, recent works demonstrate that self-supervised pre-training with unlabeled data can improve detection accuracy with limited labels. Contemporary methods adapt best-practices for self-supervised learning from the image domain to point clouds (such as contrastive learning). However, publicly available 3D datasets are considerably smaller and less diverse than those used for image-based self-supervised learning, limiting their effectiveness. We do note, however, that such data is naturally collected in a multimodal fashion, often paired with images. Rather than pre-training with only self-supervised objectives, we argue that it is better to bootstrap point cloud representations using image-based foundation models trained on internet-scale image data. Specifically, we propose a shelf-supervised approach (e.g. supervised with off-the-shelf image foundation models) for generating zero-shot 3D bounding boxes from paired RGB and LiDAR data. Pre-training 3D detectors with such pseudo-labels yields significantly better semi-supervised detection accuracy than prior self-supervised pretext tasks. Importantly, we show that image-based shelf-supervision is helpful for training LiDAR-only and multi-modal (RGB + LiDAR) detectors. We demonstrate the effectiveness of our approach on nuScenes and WOD, significantly improving over prior work in limited data settings.

231

Automated Creation of Digital Cousins for Robust Policy Learning
Tianyuan Dai,Josiah Wong,Yunfan Jiang,Chen Wang,Cem Gokmen,Ruohan Zhang,Jiajun Wu,Li Fei-Fei
https://openreview.net/forum?id=7c5rAY8oU3
Training robot policies in the real world can be unsafe, costly, and difficult to scale. Simulation serves as an inexpensive and potentially limitless source of training data, but suffers from the semantics and physics disparity between simulated and real-world environments. These discrepancies can be minimized by training in digital twins, which serve as virtual replicas of a real

scene but are expensive to generate and cannot produce cross-domain generalization. To address these limitations, we propose the concept of digital cousins, a virtual asset or scene that, unlike a digital twin, does not explicitly model a real-world counterpart but still exhibits similar geometric and semantic affordances. As a result, digital cousins simultaneously reduce the cost of generating an analogous virtual environment while also facilitating better robustness during sim-to-real domain transfer by providing a distribution of similar training scenes. Leveraging digital cousins, we introduce a novel method for their automated creation, and propose a fully automated real-to-sim-to-real pipeline for generating fully interactive scenes and training robot policies that can be deployed zero-shot in the original scene. We find that digital cousin scenes that preserve geometric and semantic affordances can be produced automatically, and can be used to train policies that outperform policies trained on digital twins, achieving 90% vs. 25% success rates under zero-shot sim-to-real transfer. Additional details are available at https://digital-cousins.github.io/.

232

RoVi-Aug: Robot and Viewpoint Augmentation for Cross-Embodiment Robot Learning
Lawrence Yunliang Chen,Chenfeng Xu,Karthik Dharmarajan,Richard Cheng,Kurt Keutzer,Masayoshi Tomizuka,Quan Vuong,Ken Goldberg
https://openreview.net/forum?id=ctzBccpolr
Scaling up robot learning requires large and diverse datasets, and how to efficiently reuse collected data and transfer policies to new embodiments remains an open question. Emerging research such as the Open-X Embodiment (OXE) project has shown promise in leveraging skills by combining datasets including different robots. However, imbalances in the distribution of robot types and camera angles in many datasets make policies prone to overfit. To mitigate this issue, we propose RoVi-Aug, which leverages state-of-the-art image-to-image generative models to augment robot data by synthesizing demonstrations with different robots and camera views. Through extensive physical experiments, we show that, by training on robot- and viewpoint-augmented data, RoVi-Aug can zero-shot deploy on an unseen robot with significantly different camera angles. Compared to test-time adaptation algorithms such as Mirage, RoVi-Aug requires no extra processing at test time, does not assume known camera angles, and allows policy fine-tuning. Moreover, by co-training on both the original and augmented robot datasets, RoVi-Aug can learn multi-robot and multi-task policies, enabling more efficient transfer between robots and skills and improving success rates by up to 30%.

233

Dynamics-Guided Diffusion Model for Sensor-less Robot Manipulator Design
Xiaomeng Xu,Huy Ha,Shuran Song
https://openreview.net/forum?id=AzP6kSEffm
We present Dynamics-Guided Diffusion Model (DGDM), a data-driven framework for generating task-specific manipulator designs without task-specific training. Given object shapes and task specifications, DGDM generates sensor-less manipulator designs that can blindly manipulate objects towards desired motions and poses using an open-loop parallel motion. This framework 1) flexibly represents manipulation tasks as interaction profiles, 2) represents the design space using a geometric diffusion model, and 3) efficiently searches this design space using the gradients provided by a dynamics network trained without any task information. We evaluate DGDM on various manipulation tasks ranging from shifting/rotating objects to converging objects to a specific pose. Our generated designs outperform optimization-based and unguided diffusion baselines relatively by 31.5% and 45.3% on average success rate. With the ability to generate a

new design within 0.8s, DGDM facilitates rapid design iteration and enhances the adoption of data-driven approaches for robot mechanism design. Qualitative results are best viewed on our project website https://dgdmcorl.github.io.

234

Learning Robot Soccer from Egocentric Vision with Deep Reinforcement Learning
Dhruva Tirumala,Markus Wulfmeier,Ben Moran,Sandy Huang,Jan Humplik,Guy Lever,Tuomas Haarnoja,Leonard Hasenclever,Arunkumar Byravan,Nathan Batchelor,Neil sreendra,Kushal Patel,Marlon Gwira,Francesco Nori,Martin Riedmiller,Nicolas Heess
https://openreview.net/forum?id=fC0wWeXsVm
We apply multi-agent deep reinforcement learning (RL) to train end-to-end robot soccer policies with fully onboard computation and sensing via egocentric RGB vision. This setting reflects many challenges of real-world robotics, including active perception, agile full-body control, and long-horizon planning in a dynamic, partially-observable, multi-agent domain. We rely on large-scale, simulation-based data generation to obtain complex behaviors from egocentric vision which can be successfully transferred to physical robots using low-cost sensors. To achieve adequate visual realism, our simulation combines rigid-body physics with learned, realistic rendering via multiple Neural Radiance Fields (NeRFs). We combine teacher-based multi-agent RL and cross-experiment data reuse to enable the discovery of sophisticated soccer strategies. We analyze active-perception behaviors including object tracking and ball seeking that emerge when simply optimizing perception-agnostic soccer play. The agents display equivalent levels of performance and agility as policies with access to privileged, ground-truth state. To our knowledge, this paper constitutes a first demonstration of end-to-end training for multi-agent robot soccer, mapping raw pixel observations to joint-level actions that can be deployed in the real world.

235

SonicSense: Object Perception from In-Hand Acoustic Vibration
Jiaxun Liu,Boyuan Chen
https://openreview.net/forum?id=CpXiqz6qf4
We introduce SonicSense, a holistic design of hardware and software to enable rich robot object perception through in-hand acoustic vibration sensing. While previous studies have shown promising results with acoustic sensing for object perception, current solutions are constrained to a handful of objects with simple geometries and homogeneous materials, single-finger sensing, and mixing training and testing on the same objects. SonicSense enables container inventory status differentiation, heterogeneous material prediction, 3D shape reconstruction, and object re-identification from a diverse set of 83 real-world objects. Our system employs a simple but effective heuristic exploration policy to interact with the objects as well as end-to-end learning-based algorithms to fuse vibration signals to infer object properties. Our framework underscores the significance of in-hand acoustic vibration sensing in advancing robot tactile perception.

236

Autonomous Interactive Correction MLLM for Robust Robotic Manipulation
Chuyan Xiong,Chengyu Shen,Xiaoqi Li,Kaichen Zhou,Jiaming Liu,Ruiping Wang,Hao Dong
https://openreview.net/forum?id=S8jQtafbT3
The ability to reflect on and correct failures is crucial for robotic systems to interact stably with real-life objects. Observing the generalization and reasoning capabilities of Multimodal Large Language Models (MLLMs), previous approaches have aimed to utilize these models to enhance

robotic systems accordingly. However, these methods typically focus on high-level planning corrections using an additional MLLM, with limited utilization of failed samples to correct low-level contact poses which is particularly prone to occur during articulated object manipulation. To address this gap, we propose an Autonomous Interactive Correction (AIC) MLLM, which makes use of previous low-level interaction experiences to correct SE(3) pose predictions for articulated object. Specifically, AIC MLLM is initially fine-tuned to acquire both pose prediction and feedback prompt comprehension abilities. We design two types of prompt instructions for interactions with objects: 1) visual masks to highlight unmovable parts for position correction, and 2) textual descriptions to indicate potential directions for rotation correction. During inference, a Feedback Information Extraction module is introduced to recognize the failure cause, allowing AIC MLLM to adaptively correct the pose prediction using the corresponding prompts. To further enhance manipulation stability, we devise a Test Time Adaptation strategy that enables AIC MLLM to better adapt to the current scene configuration. Finally, extensive experiments are conducted in both simulated and real-world environments to evaluate the proposed method. The results demonstrate that our AIC MLLM can efficiently correct failure samples by leveraging interaction experience prompts.

237

FetchBench: A Simulation Benchmark for Robot Fetching
Beining Han,Meenal Parakh,Derek Geng,Jack A Defay,Gan Luyang,Jia Deng
https://openreview.net/forum?id=U5RPcnFhkq
Fetching, which includes approaching, grasping, and retrieving, is a critical challenge for robot manipulation tasks. Existing methods primarily focus on table-top scenarios, which do not adequately capture the complexities of environments where both grasping and planning are essential. To address this gap, we propose a new benchmark FetchBench, featuring diverse procedural scenes that integrate both grasping and motion planning challenges. Additionally, FetchBench includes a data generation pipeline that collects successful fetch trajectories for use in imitation learning methods. We implement multiple baselines from the traditional sense-plan-act pipeline to end-to-end behavior models. Our empirical analysis reveals that these methods achieve a maximum success rate of only 20%, indicating substantial room for improvement. Additionally, we identify key bottlenecks within the sense-plan-act pipeline and make recommendations based on the systematic analysis.

238

GenSim2: Scaling Robot Data Generation with Multi-modal and Reasoning LLMs
Pu Hua,Minghuan Liu,Annabella Macaluso,Yunfeng Lin,Weinan Zhang,Huazhe Xu,Lirui Wang
https://openreview.net/forum?id=5u9l6U61S7
Robotic simulation today remains challenging to scale up due to the human efforts required to create diverse simulation tasks and scenes. Simulation-trained policies also face scalability issues as many sim-to-real methods focus on a single task. To address these challenges, this work proposes GenSim2, a scalable framework that leverages coding LLMs with multi-modal and reasoning capabilities for complex and realistic simulation task creation, including long-horizon tasks with articulated objects. To automatically generate demonstration data for these tasks at scale, we propose planning and RL solvers that generalize within object categories. The pipeline can generate data for up to 100 articulated tasks with 200 objects and reduce the required human efforts. To utilize such data, we propose an effective multi-task language-conditioned policy architecture, dubbed proprioceptive point-cloud transformer (PPT), that learns from the generated demonstrations and exhibits strong sim-to-real zero-shot transfer. Combining the

proposed pipeline and the policy architecture, we show a promising usage of GenSim2 that the generated data can be used for zero-shot transfer or co-train with real-world collected data, which enhances the policy performance by 20% compared with training exclusively on limited real data.

239

MILES: Making Imitation Learning Easy with Self-Supervision
Georgios Papagiannis,Edward Johns
https://openreview.net/forum?id=y8XkuQIrvI
Data collection in imitation learning often requires significant, laborious human supervision, such as numerous demonstrations, and/or frequent environment resets for methods that incorporate reinforcement learning. In this work, we propose an alternative approach, MILES: a fully autonomous, self-supervised data collection paradigm, and we show that this enables efficient policy learning from just a single demonstration and a single environment reset. MILES autonomously learns a policy for returning to and then following the single demonstration, whilst being self-guided during data collection, eliminating the need for additional human interventions. We evaluated MILES across several realworld tasks, including tasks that require precise contact-rich manipulation such as locking a lock with a key. We found that, under the constraints of a single demonstration and no repeated environment resetting, MILES significantly outperforms state-of-the-art alternatives like imitation learning methods that leverage reinforcement learning. Videos of our experiments and code can be found on our webpage: www.robot-learning.uk/miles.

240

Differentiable Robot Rendering
Ruoshi Liu,Alper Canberk,Shuran Song,Carl Vondrick
https://openreview.net/forum?id=It0Yf8Wh5O
Vision foundation models trained on massive amounts of visual data have shown unprecedented reasoning and planning skills in open-world settings. A key challenge in applying them to robotic tasks is the modality gap between visual data and action data. We introduce differentiable robot rendering, a method allowing the visual appearance of a robot body to be directly differentiable with respect to its control parameters. Our model integrates a kinematics-aware deformable model and Gaussians Splatting and is compatible with any robot form factors and degrees of freedom. We demonstrate its capability and usage in applications including reconstruction of robot poses from images and controlling robots through vision language models. Quantitative and qualitative results show that our differentiable rendering model provides effective gradients for robotic control directly from pixels, setting the foundation for the future applications of vision foundation models in robotics.

241

Scaling Manipulation Learning with Visual Kinematic Chain Prediction
Xinyu Zhang,Yuhan Liu,Haonan Chang,Abdeslam Boularias
https://openreview.net/forum?id=Yw5QGNBkEN
Learning general-purpose models from diverse datasets has achieved great success in machine learning. In robotics, however, existing methods in multi-task learning are typically constrained to a single robot and workspace, while recent work such as RT-X requires a non-trivial action normalization procedure to manually bridge the gap between different action spaces in diverse environments. In this paper, we propose the visual kinematics chain as a precise and universal

representation of quasi-static actions for robot learning over diverse environments, which requires no manual adjustment since the visual kinematic chains can be automatically obtained from the robot's model and camera parameters. We propose the Visual Kinematics Transformer (VKT), a convolution-free architecture that supports an arbitrary number of camera viewpoints, and that is trained with a single objective of forecasting kinematic structures through optimal point-set matching. We demonstrate the superior performance of VKT over BC transformers as a general agent on Calvin, RLBench, ALOHA, Open-X, and real robot manipulation tasks. Video demonstrations and source code can be found at https://mlzxy.github.io/visual-kinetic-chain.

242

Detect Everything with Few Examples
Xinyu Zhang,Yuhan Liu,Yuting Wang,Abdeslam Boularias
https://openreview.net/forum?id=HlxRd529nG
Few-shot object detection aims at detecting novel categories given only a few example images. It is a basic skill for a robot to perform tasks in open environments. Recent methods focus on finetuning strategies, with complicated procedures that prohibit a wider application. In this paper, we introduce DE-ViT, a few-shot object detector without the need for finetuning. DE-ViT's novel architecture is based on a new region-propagation mechanism for localization. The propagated region masks are transformed into bounding boxes through a learnable spatial integral layer. Instead of training prototype classifiers, we propose to use prototypes to project ViT features into a subspace that is robust to overfitting on base classes. We evaluate DE-ViT on few-shot, and one-shot object detection benchmarks with Pascal VOC, COCO, and LVIS. DE-ViT establishes new state-of-the-art results on all benchmarks. Notably, for COCO, DE-ViT surpasses the few-shot SoTA by 15 mAP on 10-shot and 7.2 mAP on 30-shot and one-shot SoTA by 2.8 AP50. For LVIS, DE-ViT outperforms few-shot SoTA by 17 box APr. Further, we evaluate DE-ViT with a real robot by building a pick-and-place system for sorting novel objects based on example images. The videos of our robot demonstrations, the source code and the models of DE-ViT can be found at https://mlzxy.github.io/devit.

243

Hint-AD: Holistically Aligned Interpretability in End-to-End Autonomous Driving
Kairui Ding,Boyuan Chen,Yuchen Su,Huan-ang Gao,Bu Jin,Chonghao Sima,Xiaohui Li,Wuqiang Zhang,Paul Barsch,Hongyang Li,Hao Zhao
https://openreview.net/forum?id=KcW31O0PtL
End-to-end architectures in autonomous driving (AD) face a significant challenge in interpretability, impeding human-AI trust. Human-friendly natural language has been explored for tasks such as driving explanation and 3D captioning. However, previous works primarily focused on the paradigm of declarative interpretability, where the natural language interpretations are not grounded in the intermediate outputs of AD systems, making the interpretations only declarative. In contrast, aligned interpretability establishes a connection between language and the intermediate outputs of AD systems. Here we introduce Hint-AD, an integrated AD-language system that generates language aligned with the holistic perception-prediction-planning outputs of the AD model. By incorporating the intermediate outputs and a holistic token mixer sub-network for effective feature adaptation, Hint-AD achieves desirable accuracy, achieving state-of-the-art results in driving language tasks including driving explanation, 3D dense captioning, and command prediction. To facilitate further study on driving explanation task on nuScenes, we also introduce a human-labeled dataset, Nu-X. Codes, dataset, and models are publicly available at https://anonymous.4open.science/r/Hint-AD-1385/.

244

Manipulate-Anything: Automating Real-World Robots using Vision-Language Models
Jiafei Duan,Wentao Yuan,Wilbert Pumacay,Yi Ru Wang,Kiana Ehsani,Dieter Fox,Ranjay Krishna
https://openreview.net/forum?id=2SYFDG4WRA
Large-scale endeavors like RT-1 and widespread community efforts such as Open-X-Embodiment have contributed to growing the scale of robot demonstration data. However, there is still an opportunity to improve the quality, quantity, and diversity of robot demonstration data. Although vision-language models have been shown to automatically generate demonstration data, their utility has been limited to environments with privileged state information, they require hand-designed skills, and are limited to interactions with few object instances. We propose Manipulate-Anything, a scalable automated generation method for real-world robotic manipulation. Unlike prior work, our method can operate in real-world environments without any privileged state information, hand-designed skills, and can manipulate any static object. We evaluate our method using two setups. First, Manipulate-Anything successfully generates trajectories for all 5 real-world and 12 simulation tasks, significantly outperforming existing methods like VoxPoser. Second, Manipulate-Anything's demonstrations can train more robust behavior cloning policies than training with human demonstrations, or from data generated by VoxPoser and Code-As-Policies. We believe Manipulate-Anything can be the scalable method for both generating data for robotics and solving novel tasks in a zero-shot setting. Anonymous project page: manipulate-anything.github.io.

245

Pre-emptive Action Revision by Environmental Feedback for Embodied Instruction Following Agents
Jinyeon Kim,Cheolhong Min,Byeonghwi Kim,Jonghyun Choi
https://openreview.net/forum?id=cq2uB30uBM
When we, humans, perform a task, we consider changes in environments such as objects' arrangement due to interactions with objects and other reasons; e.g., when we find a mug to clean, if it is already clean, we skip cleaning it. But even the state-of-the-art embodied agents often ignore changed environments when performing a task, leading to failure to complete the task, executing unnecessary actions, or fixing the mistake after it was made. Here, we propose Pre-emptive Action Revision by Environmental feeDback (PRED) that allows an embodied agent to revise their action in response to the perceived environmental status before it makes mistakes. We empirically validate PRED and observe that it outperforms the prior art on two challenging benchmarks in the virtual environment, TEACh and ALFRED, by noticeable margins in most metrics, including unseen success rates, with shorter execution time, implying an efficiently behaved agent. Furthermore, we demonstrate the effectiveness of the proposed method with real robot experiments.

246

Let Occ Flow: Self-Supervised 3D Occupancy Flow Prediction
Yili Liu,Linzhan Mou,Xuan Yu,Chenrui Han,Sitong Mao,Rong Xiong,Yue Wang
https://openreview.net/forum?id=WLOTZHmmO6
Accurate perception of the dynamic environment is a fundamental task for autonomous driving and robot systems. This paper introduces Let Occ Flow, the first self-supervised work for joint 3D occupancy and occupancy flow prediction using only camera inputs, eliminating the need for 3D annotations. Utilizing TPV for unified scene representation and deformable attention layers for

feature aggregation, our approach incorporates a novel attention-based temporal fusion module to capture dynamic object dependencies, followed by a 3D refine module for fine-gained volumetric representation. Besides, our method extends differentiable rendering to 3D volumetric flow fields, leveraging zero-shot 2D segmentation and optical flow cues for dynamic decomposition and motion optimization. Extensive experiments on nuScenes and KITTI datasets demonstrate the competitive performance of our approach over prior state-of-the-art methods.

247

Q-SLAM: Quadric Representations for Monocular SLAM
Chensheng Peng,Chenfeng Xu,Yue Wang,Mingyu Ding,Heng Yang,Masayoshi Tomizuka,Kurt Keutzer,Marco Pavone,Wei Zhan
https://openreview.net/forum?id=k4Nnxqcwt8
In this paper, we reimagine volumetric representations through the lens of quadrics. We posit that rigid scene components can be effectively decomposed into quadric surfaces. Leveraging this assumption, we reshape the volumetric representations with million of cubes by several quadric planes, which results in more accurate and efficient modeling of 3D scenes in SLAM contexts. First, we use the quadric assumption to rectify noisy depth estimations from RGB inputs. This step significantly improves depth estimation accuracy, and allows us to efficiently sample ray points around quadric planes instead of the entire volume space in previous NeRF-SLAM systems. Second, we introduce a novel quadric-decomposed transformer to aggregate information across quadrics. The quadric semantics are not only explicitly used for depth correction and scene decomposition, but also serve as an implicit supervision signal for the mapping network. Through rigorous experimental evaluation, our method exhibits superior performance over other approaches relying on estimated depth, and achieves comparable accuracy to methods utilizing ground truth depth on both synthetic and real-world datasets.

248

TRANSIC: Sim-to-Real Policy Transfer by Learning from Online Correction
Yunfan Jiang,Chen Wang,Ruohan Zhang,Jiajun Wu,Li Fei-Fei
https://openreview.net/forum?id=lpjPft4RQT
Learning in simulation and transferring the learned policy to the real world has the potential to enable generalist robots. The key challenge of this approach is to address simulation-to-reality (sim-to-real) gaps. Previous methods often require domain-specific knowledge a priori. We argue that a straightforward way to obtain such knowledge is by asking humans to observe and assist robot policy execution in the real world. The robots can then learn from humans to close various sim-to-real gaps. We propose TRANSIC, a data-driven approach to enable successful sim-to-real transfer based on a human-in-the-loop framework. TRANSIC allows humans to augment simulation policies to overcome various unmodeled sim-to-real gaps holistically through intervention and online correction. Residual policies can be learned from human corrections and integrated with simulation policies for autonomous execution. We show that our approach can achieve successful sim-to-real transfer in complex and contact-rich manipulation tasks such as furniture assembly. Through synergistic integration of policies learned in simulation and from humans, TRANSIC is effective as a holistic approach to addressing various, often coexisting sim-to-real gaps. It displays attractive properties such as scaling with human effort. Videos and code are available at https://transic-robot.github.io/.

249

Learning Visual Parkour from Generated Images
Alan Yu,Ge Yang,Ran Choi,Yajvan Ravan,John Leonard,Phillip Isola
https://openreview.net/forum?id=cGswIOxHcN
Fast and accurate physics simulation is an essential component of robot learning, where robots can explore failure scenarios that are difficult to produce in the real world and learn from unlimited on-policy data. Yet, it remains challenging to incorporate RGB-color perception into the sim-to-real pipeline that matches the real world in its richness and realism. In this work, we train a robot dog in simulation for visual parkour. We propose a way to use generative models to synthesize diverse and physically accurate image sequences of the scene from the robot's ego-centric perspective. We present demonstrations of zero-shot transfer to the RGB-only observations of the real world on a robot equipped with a low-cost, off-the-shelf color camera.

250

Dreamitate: Real-World Visuomotor Policy Learning via Video Generation
Junbang Liang,Ruoshi Liu,Ege Ozguroglu,Sruthi Sudhakar,Achal Dave,Pavel Tokmakov,Shuran Song,Carl Vondrick
https://openreview.net/forum?id=InT87E5sr4
A key challenge in manipulation is learning a policy that can robustly generalize to diverse visual environments. A promising mechanism for learning robust policies is to leverage video generative models, which are pretrained on large-scale datasets of internet videos. In this paper, we propose a visuomotor policy learning framework that fine-tunes a video diffusion model on human demonstrations of a given task. At test time, we generate an example of an execution of the task conditioned on images of a novel scene, and use this synthesized execution directly to control the robot. Our key insight is that using common tools allows us to effortlessly bridge the embodiment gap between the human hand and the robot manipulator. We evaluate our approach on 4 tasks of increasing complexity and demonstrate that capitalizing on internet-scale generative models allows the learned policy to achieve a significantly higher degree of generalization than existing behavior cloning approaches.

251

Visual Manipulation with Legs
Xialin He,Chengjing Yuan,Wenxuan Zhou,Ruihan Yang,David Held,Xiaolong Wang
https://openreview.net/forum?id=E4K3yLQQ7s
Animals have the ability to use their arms and legs for both locomotion and manipulation. We envision quadruped robots to have the same versatility. This work presents a system that empowers a quadruped robot to perform object interactions with its legs, drawing inspiration from non-prehensile manipulation techniques. The proposed system has two main components: a visual manipulation policy module and a loco-manipulator module. The visual manipulation policy module decides how the leg should interact with the object, trained with reinforcement learning (RL) with point cloud observations and object-centric actions. The loco-manipulator controller controls the leg movements and body pose adjustments, implemented based on impedance control and Model Predictive Control (MPC). Besides manipulating objects with a single leg, the proposed system can also select from left or right legs based on the critic maps and move the object to distant goals through robot base adjustment. In the experiments, we evaluate the proposed system with the object pose alignment tasks both in simulation and in the real world, demonstrating object manipulation skills with legs more versatile than previous work.

252

Neural Attention Field: Emerging Point Relevance in 3D Scenes for One-Shot Dexterous Grasping
Qianxu Wang,Congyue Deng,Tyler Ga Wei Lum,Yuanpei Chen,Yaodong Yang,Jeannette Bohg,Yixin Zhu,Leonidas Guibas
https://openreview.net/forum?id=AGG1zIrrMw
One-shot transfer of dexterous grasps to novel scenes with object and context variations has been a challenging problem. While distilled feature fields from large vision models have enabled semantic correspondences across 3D scenes, their features are point-based and restricted to object surfaces, limiting their capability of modeling complex semantic feature distributions for hand-object interactions. In this work, we propose the neural attention field for representing semantic-aware dense feature fields in the 3D space by modeling inter-point relevance instead of individual point features. Core to it is a transformer decoder that computes the cross-attention between any 3D query point with all the scene points, and provides the query point feature with an attention-based aggregation. We further propose a self-supervised framework for training the transformer decoder from only a few 3D pointclouds without hand demonstrations. Post-training, the attention field can be applied to novel scenes for semantics-aware dexterous grasping from one-shot demonstration. Experiments show that our method provides better optimization landscapes by encouraging the end-effector to focus on task-relevant scene regions, resulting in significant improvements in success rates on real robots compared with the feature-field-based methods.

253

Open-TeleVision: Teleoperation with Immersive Active Visual Feedback
Xuxin Cheng,Jialong Li,Shiqi Yang,Ge Yang,Xiaolong Wang
https://openreview.net/forum?id=Yce2jeILGt
Teleoperation serves as a powerful method for collecting on-robot data essential for robot learning from demonstrations. The intuitiveness and ease of use of the teleoperation system are crucial for ensuring high-quality, diverse, and scalable data. To achieve this, we propose an immersive teleoperation system **Open-TeleVision** that allows operators to actively perceive the robot's surroundings in a stereoscopic manner. Additionally, the system mirrors the operator's arm and hand movements on the robot, creating an immersive experience as if the operator's mind is transmitted to a robot embodiment. We validate the effectiveness of our system by collecting data and training imitation learning policies on four long-horizon, precise tasks (can sorting, can insertion, folding, and unloading) for 2 different humanoid robots and deploy them in the real world. The entire system will be open-sourced.

254

LLARVA: Vision-Action Instruction Tuning Enhances Robot Learning
Dantong Niu,Yuvan Sharma,Giscard Biamby,Jerome Quenum,Yutong Bai,Baifeng Shi,Trevor Darrell,Roei Herzig
https://openreview.net/forum?id=Q2lGXMZCv8
In recent years, instruction-tuned Large Multimodal Models (LMMs) have been successful at several tasks, including image captioning and visual question answering; yet leveraging these models remains an open question for robotics. Prior LMMs for robotics applications have been extensively trained on language and action data, but their ability to generalize in different settings has often been less than desired. To address this, we introduce LLARVA, a model trained with a novel instruction tuning method that leverages structured prompts to unify a range of robotic

learning tasks, scenarios, and environments. Additionally, we show that predicting intermediate 2-D representations, which we refer to as visual traces, can help further align vision and action spaces for robot learning. We generate 8.5M image-visual trace pairs from the Open X-Embodiment dataset in order to pre-train our model, and we evaluate on 12 different tasks in the RLBench simulator as well as a physical Franka Emika Panda 7-DoF robot. Our experiments yield strong performance, demonstrating that LLARVA — using 2-D and language representations — performs well compared to several contemporary baselines, and can generalize across various robot environments and configurations.

255

Real-to-Sim Grasp: Rethinking the Gap between Simulation and Real World in Grasp Detection
Jia-Feng Cai,Zibo Chen,Xiao-Ming Wu,Jian-Jian Jiang,Yi-Lin Wei,Wei-Shi Zheng
https://openreview.net/forum?id=uJBMZ6S02T
For 6-DoF grasp detection, simulated data is expandable to train more powerful model, but it faces the challenge of the large gap between simulation and real world. Previous works bridge this gap with a sim-to-real way. However, this way explicitly or implicitly forces the simulated data to adapt to the noisy real data when training grasp detectors, where the positional drift and structural distortion within the camera noise will harm the grasp learning. In this work, we propose a Real-to-Sim framework for 6-DoF Grasp detection, named R2SGrasp, with the key insight of bridging this gap in a real-to-sim way, which directly bypasses the camera noise in grasp detector training through an inference-time real-to-sim adaption. To achieve this real-to-sim adaptation, our R2SGrasp designs the Real-to-Sim Data Repairer (R2SRepairer) to mitigate the camera noise of real depth maps in data-level, and the Real-to-Sim Feature Enhancer (R2SEnhancer) to enhance real features with precise simulated geometric primitives in feature-level. To endow our framework with the generalization ability, we construct a large-scale simulated dataset cost-efficiently to train our grasp detector, which includes 64,000 RGB-D images with 14.4 million grasp annotations. Sufficient experiments show that R2SGrasp is powerful and our real-to-sim perspective is effective. The real-world experiments further show great generalization ability of R2SGrasp. Project page is available on https://isee-laboratory.github.io/R2SGrasp.

256

CtRL-Sim: Reactive and Controllable Driving Agents with Offline Reinforcement Learning
Luke Rowe,Roger Girgis,Anthony Gosselin,Bruno Carrez,Florian Golemo,Felix Heide,Liam Paull,Christopher Pal
https://openreview.net/forum?id=MfIUKzihC8
Evaluating autonomous vehicle stacks (AVs) in simulation typically involves replaying driving logs from real-world recorded traffic. However, agents replayed from offline data are not reactive and hard to intuitively control. Existing approaches address these challenges by proposing methods that rely on heuristics or generative models of real-world data but these approaches either lack realism or necessitate costly iterative sampling procedures to control the generated behaviours. In this work, we take an alternative approach and propose CtRL-Sim, a method that leverages return-conditioned offline reinforcement learning to efficiently generate reactive and controllable traffic agents. Specifically, we process real-world driving data through a physics-enhanced Nocturne simulator to generate a diverse offline reinforcement learning dataset, annotated with various reward terms. We then train a return-conditioned multi-agent behaviour model that allows for fine-grained manipulation of agent behaviours by modifying the desired returns for the various reward components. This capability enables the generation of a wide range of driving behaviours beyond the scope of the initial dataset, including adversarial behaviours. We demonstrate that CtRL-Sim

can generate diverse and realistic safety-critical scenarios while providing fine-grained control over agent behaviours.

257

GraspSplats: Efficient Manipulation with 3D Feature Splatting
Mazeyu Ji,Ri-Zhao Qiu,Xueyan Zou,Xiaolong Wang
https://openreview.net/forum?id=pPhTsonbXq
The ability for robots to perform efficient and zero-shot grasping of object parts is crucial for practical applications and is becoming prevalent with recent advances in Vision-Language Models (VLMs). To bridge the 2D-to-3D gap for representations to support such a capability, existing methods rely on neural fields (NeRFs) via differentiable rendering or point-based projection methods. However, we demonstrate that NeRFs are inappropriate for scene changes due to its implicitness and point-based methods are inaccurate for part localization without rendering-based optimization. To amend these issues, we propose GraspSplats. Using depth supervision and a novel reference feature computation method, GraspSplats can generate high-quality scene representations under 60 seconds. We further validate the advantages of Gaussian-based representation by showing that the explicit and optimized geometry in GraspSplats is sufficient to natively support (1) real-time grasp sampling and (2) dynamic and articulated object manipulation with point trackers. With extensive experiments on a Franka robot, we demonstrate that GraspSplats significantly outperforms existing methods under diverse task settings. In particular, GraspSplats outperforms NeRF-based methods like F3RM and LERF-TOGO, and 2D detection methods. The code will be released.

258

Get a Grip: Multi-Finger Grasp Evaluation at Scale Enables Robust Sim-to-Real Transfer
Tyler Ga Wei Lum,Albert H. Li,Preston Culbertson,Krishnan Srinivasan,Aaron Ames,Mac Schwager,Jeannette Bohg
https://openreview.net/forum?id=1jc2zA5Z6J
This work explores conditions under which multi-finger grasping algorithms can attain robust sim-to-real transfer. While numerous large datasets facilitate learning generative models for multi-finger grasping at scale, reliable real-world dexterous grasping remains challenging, with most methods degrading when deployed on hardware. An alternate strategy is to use discriminative grasp evaluation models for grasp selection and refinement, conditioned on real-world sensor measurements. This paradigm has produced state-of-the-art results for vision-based parallel-jaw grasping, but remains unproven in the multi-finger setting. In this work, we find that existing datasets and methods have been insufficient for training discriminitive models for multi-finger grasping. To train grasp evaluators at scale, datasets must provide on the order of millions of grasps, including both positive and negative examples, with corresponding visual data resembling measurements at inference time. To that end, we release a new, open-source dataset of 3.5M grasps on 4.3K objects annotated with RGB images, point clouds, and trained NeRFs. Leveraging this dataset, we train vision-based grasp evaluators that outperform both analytic and generative modeling-based baselines on extensive simulated and real-world trials across a diverse range of objects. We show via numerous ablations that the key factor for performance is indeed the evaluator, and that its quality degrades as the dataset shrinks, demonstrating the importance of our new dataset. Project website at: https://sites.google.com/view/get-a-grip-dataset.

259

DextrAH-G: Pixels-to-Action Dexterous Arm-Hand Grasping with Geometric Fabrics
Tyler Ga Wei Lum,Martin Matak,Viktor Makoviychuk,Ankur Handa,Arthur Allshire,Tucker Hermans,Nathan D. Ratliff,Karl Van Wyk
https://openreview.net/forum?id=S2Jwb0i7HN
A pivotal challenge in robotics is achieving fast, safe, and robust dexterous grasping across a diverse range of objects, an important goal within industrial applications. However, existing methods often have very limited speed, dexterity, and generality, along with limited or no hardware safety guarantees. In this work, we introduce DextrAH-G, a depth-based dexterous grasping policy trained entirely in simulation that combines reinforcement learning, geometric fabrics, and teacher-student distillation. We address key challenges in joint arm-hand policy learning, such as high-dimensional observation and action spaces, the sim2real gap, collision avoidance, and hardware constraints. DextrAH-G enables a 23 motor arm-hand robot to safely and continuously grasp and transport a large variety of objects at high speed using multi-modal inputs including depth images, allowing generalization across object geometry. Videos at https://sites.google.com/view/dextrah-g.

260

Generalized Animal Imitator: Agile Locomotion with Versatile Motion Prior
Ruihan Yang,Zhuoqun Chen,Jianhan Ma,Chongyi Zheng,Yiyu Chen,Quan Nguyen,Xiaolong Wang
https://openreview.net/forum?id=9XV3dBqcfe
The agility of animals, particularly in complex activities such as running, turning, jumping, and backflipping, stands as an exemplar for robotic system design. Transferring this suite of behaviors to legged robotic systems introduces essential inquiries: How can a robot be trained to learn multiple locomotion behaviors simultaneously? How can the robot execute these tasks with a smooth transition? How to integrate these skills for wide-range applications? This paper introduces the Versatile Instructable Motion prior (VIM) – a Reinforcement Learning framework designed to incorporate a range of agile locomotion tasks suitable for advanced robotic applications. Our framework enables legged robots to learn diverse agile low-level skills by imitating animal motions and manually designed motions. Our Functionality reward guides the robot's ability to adopt varied skills, and our Stylization reward ensures that robot motions align with reference motions. Our evaluations of the VIM framework span both simulation environments and real-world deployment. To the best of our knowledge, this is the first work that allows a robot to concurrently learn diverse agile locomotion skills using a single learning-based controller in the real world.

261

ACE: A Cross-platform and visual-Exoskeletons System for Low-Cost Dexterous Teleoperation
Shiqi Yang,Minghuan Liu,Yuzhe Qin,Runyu Ding,Jialong Li,Xuxin Cheng,Ruihan Yang,Sha Yi,Xiaolong Wang
https://openreview.net/forum?id=7ddT4eklmQ
Bimanual robotic manipulation with dexterous hands has a large potential workability and a wide workspace as it follows the most natural human workflow. Learning from human demonstrations has proven highly effective for learning a dexterous manipulation policy. To collect such data, teleoperation serves as a straightforward and efficient way to do so. However, a cost-effective and easy-to-use teleoperation system is lacking for anthropomorphic robot hands. To fill the deficiency, we developed \our, a cross-platform visual-exoskeleton system for low-cost dexterous

teleoperation. Our system employs a hand-facing camera to capture 3D hand poses and an exoskeleton mounted on a base that can be easily carried on users' backs. ACE captures both the hand root end-effector and hand pose in real-time and enables cross-platform operations. We evaluate the key system parameters compared with previous teleoperation systems and show clear advantages of \our. We then showcase the desktop and mobile versions of our system on six different robot platforms (including humanoid-hands, arm-hands, arm-gripper, and quadruped-gripper systems), and demonstrate the effectiveness of learning three difficult real-world tasks through the collected demonstration on two of them.

262

Visual Whole-Body Control for Legged Loco-Manipulation
Minghuan Liu,Zixuan Chen,Xuxin Cheng,Yandong Ji,Ri-Zhao Qiu,Ruihan Yang,Xiaolong Wang
https://openreview.net/forum?id=cT2N3p1AcE
We study the problem of mobile manipulation using legged robots equipped with an arm, namely legged loco-manipulation. The robot legs, while usually utilized for mobility, offer an opportunity to amplify the manipulation capabilities by conducting whole-body control. That is, the robot can control the legs and the arm at the same time to extend its workspace. We propose a framework that can conduct the whole-body control autonomously with visual observations. Our approach, namely Visual Whole-Body Control (VBC), is composed of a low-level policy using all degrees of freedom to track the body velocities along with the end-effector position, and a high-level policy proposing the velocities and end-effector position based on visual inputs. We train both levels of policies in simulation and perform Sim2Real transfer for real robot deployment. We perform extensive experiments and show significant improvements over baselines in picking up diverse objects in different configurations (heights, locations, orientations) and environments.

263

Evaluating Real-World Robot Manipulation Policies in Simulation
Xuanlin Li,Kyle Hsu,Jiayuan Gu,Oier Mees,Karl Pertsch,Homer Rich Walke,Chuyuan Fu,Ishikaa Lunawat,Isabel Sieh,Sean Kirmani,Sergey Levine,Jiajun Wu,Chelsea Finn,Hao Su,Quan Vuong,Ted Xiao
https://openreview.net/forum?id=LZh48DTg71
The field of robotics has made significant advances towards generalist robot manipulation policies. However, real-world evaluation of such policies is not scalable and faces reproducibility challenges, issues that are likely to worsen as policies broaden the spectrum of tasks they can perform. In this work, we demonstrate that simulation-based evaluation can be a scalable, reproducible, and reliable proxy for real-world evaluation. We identify control and visual disparities between real and simulated environments as key challenges for reliable simulated evaluation and propose approaches for mitigating these gaps without needing to painstakingly craft full-fidelity digital twins. We then employ these techniques to create SIMPLER, a collection of simulated environments for policy evaluation on common real robot manipulation setups. Through over 1500 paired sim-and-real evaluations of manipulation policies across two embodiments and eight task families, we demonstrate strong correlation between policy performance in SIMPLER environments and that in the real world. Beyond aggregated trends, we find that SIMPLER evaluations effectively reflect the real-world behaviors of individual policies, such as sensitivity to various distribution shifts. We are committed to open-sourcing all SIMPLER environments along with our workflow for creating new environments to facilitate research on general-purpose manipulation policies and simulated evaluation frameworks. Website: https://simpler-env.github.io/

264

Learning to Manipulate Anywhere: A Visual Generalizable Framework For Reinforcement Learning
Zhecheng Yuan,Tianming Wei,Shuiqi Cheng,Gu Zhang,Yuanpei Chen,Huazhe Xu
https://openreview.net/forum?id=jart4nhCQr
Can we endow visuomotor robots with generalization capabilities to operate in diverse open-world scenarios? In this paper, we propose Maniwhere, a generalizable framework tailored for visual reinforcement learning, enabling the trained robot policies to generalize across a combination of multiple visual disturbance types. Specifically, we introduce a multi-view representation learning approach fused with Spatial Transformer Network (STN) module to capture shared semantic information and correspondences among different viewpoints. In addition, we employ a curriculum-based randomization and augmentation approach to stabilize the RL training process and strengthen the visual generalization ability. To exhibit the effectiveness of Maniwhere, we meticulously design 8 tasks encompassing articulate objects, bi-manual, and dexterous hand manipulation tasks, demonstrating Maniwhere's strong visual generalization and sim2real transfer abilities across 3 hardware platforms. Our experiments show that Maniwhere significantly outperforms existing state-of-the-art methods. Videos are provided at https://maniwhere.github.io.

265

HumanPlus: Humanoid Shadowing and Imitation from Humans
Zipeng Fu,Qingqing Zhao,Qi Wu,Gordon Wetzstein,Chelsea Finn
https://openreview.net/forum?id=WnSl42M9Z4
One of the key arguments for building robots that have similar form factors to human beings is that we can leverage the massive human data for training.Yet, doing so has remained challenging in practice due to the complexities in humanoid perception and control, lingering physical gaps between humanoids and humans in morphologies and actuation, and lack of a data pipeline for humanoids to learn autonomous skills from egocentric vision. In this paper, we introduce a full-stack system for humanoids to learn motion and autonomous skills from human data. We first train a low-level policy in simulation via reinforcement learning using existing 40-hour human motion datasets. This policy transfers to the real world and allows humanoid robots to follow human body and hand motion in real time using only a RGB camera, i.e. shadowing. Through shadowing, human operators can teleoperate humanoids to collect whole-body data for learning different tasks in the real world. Using the data collected, we then perform supervised behavior cloning to train skill policies using egocentric vision, allowing humanoids to complete different tasks autonomously by imitating human skills. We demonstrate the system on our customized 33-DoF 180cm humanoid, autonomously completing tasks such as wearing a shoe to stand up and walk, folding a sweatshirt, rearranging objects, typing, and greeting another robot with 60-100% success rates using up to 40 demonstrations.