ICLR2025 Workshop Proposal on Generative Models for Robot Learning (GenBot)

1 Workshop Summary

Next generation of robots should combine ideas from other fields such as computer vision, natural language processing, machine learning and many others, because the close-loop system is required to deal with complex tasks based on multimodal input in the complicated real environment. This workshop proposal focuses on generative models for robot learning, which lies in the important and fundamental field of AI and robotics. Learning-based methods in robotics have achieved high success rate and generalization ability in a wide variety of tasks such as manipulation, navigation, SLAM, scene reconstruction, proprioception, and physics modeling. However, robot learning faces several challenges including the expensive cost of data collection and weak transferability across different tasks and scenarios. Inspired by the significant progress in computer vision and natural language processing, efforts have been made to combine generative models with robot learning to address the above challenges such as synthesizing high-quality data, and incorporating generation frameworks into representation and policy learning. Besides, pre-trained large language models (LLMs), vision-language models (VLMs) and vision-language-action (VLA) models are adapted to various downstream tasks to fully leverage the rich commonsense knowledge. This progressive development enables robot learning frameworks to be applied in complex and diverse real-world tasks.

This workshop aims to enable interdisciplinary communication for researchers in the broader community, so that more attention can be drawn to this field. In this workshop, the state-of-the-art process and promising future directions will be discussed, which will inspire new ideas and fantastic applications in related fields. Our topics will include but not be limited to the following:

Robotics data generation. Acquiring synthesized robotics data including RGB-D images, action sequences of locomotion and manipulation, and IMU data which have low discrepancy with the real-world ones is necessary to enhance the capacity and generalization ability of robot learning frameworks. Under this topic, we will discuss: (i) How can we build simulators with diverse assets with rich interactive properties? And how can we accurately simulate physical consequences for diverse actions of robots? (ii) How can we accelerate the generation process for successful trajectory in the simulation environments? (iii) What are the challenges and possible solutions to alleviate the visual domain gap between the simulators and the real world?

Generative policy learning. Enabling generative models as powerful tools in robotic policy learning captures the distribution of complex multimodal data, which boosts the sample efficiency, transferability, and robustness of learned policy. Under this topic, we will discuss: (i) How can we design a generative visual representation learning framework that effectively embeds spatiotemporal information of the scene via self-supervision? (ii) How can we efficiently construct world model for scalable robot learning, and what information of the scene and the robot should be considered in order to acquire accurate feedback from the world model? (iii) How can we extend state-of-the-art generative models such as diffusion models in computer vision and auto-regressive models in natural language processing for policy generation?

Foundation model grounding. Levering rich commonsense in foundation models such as LLM, VLM, and VLA can significantly enhance the generalization ability across tasks and environments, while they should be grounded to realistic deployment scenes in order to output physically plausible actions. Under this topic, we will discuss: (i) What are the general criteria for designing prompts of LLMs for robot tasks? (ii) How can we build a scalable, efficient and generalizable representation of physical scenes to ground the action prediction of VLMs? (iii) How can we enhance the sample efficiency in VLA model training, and how can we efficiently adapt pre-trained VLA models to novel robot tasks?

On-device generative model deployment. Due to the extremely limited resources on robots, powerful pre-trained large generative models should be compressed to achieve fewer FLOPs, memory, storage and less energy consumption without obvious degradation of performance. Under this topic, we will discuss: (i) What is the complexity bottleneck in current pre-trained large generative models, and how can we distinguish and remove the redundant architectures? (ii) How can we dynamically keep the optimal accuracy-efficiency trade-off to adapt to the changing resource limit caused by battery level and utilization variance? (iii) How can we develop the compilation toolbox for pre-trained large generative models on robot-based computational platforms to achieve significant actual speedup and memory saving?

1.1 Broader Impact

The goal of this workshop is to foster interactions among interdisciplinary researchers from computer vision, robotics, machine learning, natural language processing for the general goal of better robot learning frameworks for perception, planning and control. We expect these engagements to foster the development of a future workforce that embraces interdisciplinary knowledge spanning multiple fields. We believe such a breadth of understanding will have a significant impact across various research domains, especially recent advancements in generative models for robot learning. Specifically, our workshop aims to 1) raise an awareness of sample efficiency in robot learning; 2) enable generative models to understand the physical laws in the real world; 3) provide insights of household and industrial practices of generative models and discuss future directions. We believe this workshop is beneficial to effectively address societal challenges especially for scenarios with general-purpose robots in house and factory environments. We will advertise our workshop, popularize the important findings, results and works through website updates, video publications, and online social networks including X, YouTube and Zhihu. We will provide Best Paper Awards with incentives for papers in high quality. The outcome of the workshop will be summarized into a white paper.

1.2 Relationship to Previous Workshops

There have been several related workshops in the past ICLR/NeurIPS/CoRL/CVPR/ICRA, while our workshop differs from them in the following aspects. 1) Most previous workshops of generative models (FMDM, NeurIPS2022; GenAI4DM, ICLR2024; GCV, CVPR2024; EDGE, CVPR2024) focused on the application in general tasks such as computer vision, natural language processing and decision making, none of them discussed the design of generative models for robots. The interest of the RoboNeRF Workshop in ICRA2024 was the generative learning of representation and policies with the NeRF-based models. On the contrary, we call for data generation, generative policy learning, foundation model grounding and on-device model deployment for all generation-based frameworks such as NeRF, diffusion, Gaussian splatting, and LLM/VLM/VLA. 2) Most workshops of robot learning (RoL, NeurIPS2019-2024; 3DVR, CVPR2023; VisPR, CVPR2023; Towards Generalist Robots, CoRL2023) considered the representation of perception in robot learning. Generative models were not involved, which could deal with challenging issues in robot learning such as insufficient realistic data and low task transferability. Therefore, we will discuss the solution of extending generative models for data synthesis and policy learning. 3) We will foster interdisciplinary communications with researchers working on robot learning (robotics, machine learning) and generative models (computer vision, natural language processing).

2 Invited Speakers

Jitendra Malik, UC Berkeley, he/him/his (confirmed) is an Arthur J. Chick Professor in the Department of Electrical Engineering and Computer Science at the University of California at Berkeley. He received the PhD degree in Computer Science from Stanford University in 1985 following which he joined UC Berkeley as a faculty member. He served as Chair of the Computer Science Division during 2002-2006, and of the Department of EECS during 2004-2006. Jitendra's group has worked on computer vision, computational modeling of biological vision, computer graphics, and machine learning. Several well-known concepts and algorithms arose in this work, such as anisotropic diffusion, normalized cuts, high dynamic range imaging, shape contexts, and R-CNN.

Sergey Levine, UC Berkeley, he/him/his (confirmed) is an Associate Professor in the Department of Electrical Engineering and Computer Science at the University of California at Berkeley. His work focuses on machine learning for decision making and control, with an emphasis on deep learning and reinforcement

learning algorithms. Applications of his work include autonomous robots and vehicles, as well as computer vision and graphics. His research includes developing algorithms for end-to-end training of deep neural network policies that combine perception and control, scalable algorithms for inverse reinforcement learning, deep reinforcement learning algorithms.

Shuran Song, Stanford, she/her/hers (confirmed) is an Assistant Professor of Electrical Engineering at Stanford University. Before joining Stanford, she was a faculty at Columbia University. Shuran received her Ph.D. in Computer Science at Princeton University, BEng. at HKUST. Her research interests lie at the intersection of computer vision and robotics. Song's research has been recognized through several awards, including the Best Paper Awards at RSS'22 and T-RO'20, Best System Paper Awards at CoRL'21, RSS'19, and finalists at RSS, ICRA, CVPR, and IROS. She is also a recipient of the NSF Career Award, Sloan Foundation fellowship as well as research awards from Microsoft, Toyota Research, Google, Amazon, and JP Morgan.

Yilun Du, Harvard & Google, he/him/his (confirmed) is a Senior Research Scientist at Google Deepmind and an incoming Assistant Professor at the Department of Computer Science, Harvard University starting in Fall 2025. Yilun earned a PhD in Electrical Engineering and Computer Science from MIT, where he was advised by Professors Leslie Kaelbling, Tomas Lozano-Perez, and Joshua B. Tenenbaum. He previously completed his bachelor's degree at MIT and has been a Research Fellow at OpenAI, an intern, and a visiting researcher at FAIR and Google DeepMind. Additionally, he earned a gold medal at the International Biology Olympiad. Yilun's research focuses on generative models, decision making, robot learning, embodied agents, and their applications to scientific domains.

Qi Dou, CUHK, she/her/hers (confirmed) is an Assistant Professor with the Department of Computer Science & Engineering at the Chinese University of Hong Kong (CUHK). Her research interest lies in the interdisciplinary area of artificial intelligence and healthcare with expertise in medical image analysis and robotic surgery, with the mission to advance disease diagnosis and minimally invasive intervention via machine intelligence. She has won the IEEE ICRA Best Paper Award in Medical Robotics in 2021, Hong Kong Institute of Science Young Scientist Award in 2018, CUHK Engineering Faculty Outstanding Thesis Award in 2018, Best Paper Award of Medical Image Analysis-MICCAI in 2017, Best Paper Award of Medical Imaging and Augmented Reality in 2016, MICCAI Young Scientist Award Runner-up in 2016.

Xiaojuan Qi, HKU, she/her/hers (confirmed) is an Assistant Professor at the University of Hong Kong. Her research encompasses the broad areas of Computer Vision, Deep Learning, and Artificial Intelligence. Her objective is to equip machines with open-world capabilities for perceiving, understanding, and reconstructing the visual world, focusing on the following aspects: 1) 3D reconstruction, generation, and visual environment simulation; 2) Open-world, interactive and reliable visual understanding; 3) Efficient training and inference; 4) Applications on open-ended intelligent agents (e.g., autonomous driving, embodied agents). Besides, she is also interested in utilizing AI to advance scientific and medical research (AI for Science and Medicine).

Jiwen Lu, Tsinghua, he/him/his (confirmed) is a Professor at the Department of Automation, Tsinghua University. His current research interests include computer vision, pattern recognition, multimedia computing, and intelligent robotics. He was/is a member of multiple technical committees of the IEEE Circuits and Systems Society and the IEEE Signal Processing Society. He serves as the General Co-Chair for ICME 2022, the Program Co-Chair for ACCV 2026, ICME 2020, FG 2023, VCIP 2022, and AVSS 2021. He has served as the Co-Editor-of-Chief for Pattern Recognition Letters, a Senior Associate Editor for T-CSVT, an Associate Editor for the T-IP, T-MM, T-BIOM, Pattern Recognition, and Journal of Visual Communications and Image Representation.

Dieter Fox, UW & NVIDIA, he/him/his (tentative) is a Professor at the Department of Computer Science & Engineering at the University of Washington. He grew up in Bonn, Germany, and received his Ph.D. in 1998 from the Computer Science Department at the University of Bonn. He joined the UW faculty in the fall of 2000. He also leads the NVIDIA Robotics Research Lab in Seattle. His research interests are in robotics, artificial intelligence, and state estimation. He is the head of the UW Robotics and State Estimation Lab RSE-Lab and recently served as the academic PI of the Intel Science and Technology Center for Pervasive Computing ISTC-PC.

3 Diversity and Inclusion

Diversity of speakers and organizers: We have made extensive considerations on the diversity of invited speakers and the organizing committee in terms of race, gender, seniority, and background. To note, we have three organizers and three speakers being female. The speakers and organizers are from 14 different institutes in both industry and academia. The speakers and organizers are at different career stages, spanning from well-established professors to junior professors, and Ph.D. students. Our speakers come from different research fields including computer vision, robotics, machine learning, natural language processing, and graphics, and hold various perspectives on fundamental questions on the workshop topic for in-depth discussion.

Diversity of topics: Our workshop is devoted to fostering the discussion with diverse perspectives on generative models for robot learning. We have selected topics including data collection, representation learning, policy learning, model adaptation, and model deployment, which inclusively cover the breadth of the field. We also call for papers from fundamental theory to diverse applications, from sparking scientific ideas to practical engineering reports. We welcome contributors from different backgrounds which leads to richer discussions.

Accessibility: We will organize the workshop in a hybrid format, where we will organize synchronous and asynchronous activities for broader participation. We will use video conference apps, e.g. Zoom, to accommodate remote speakers and panelists, and livestream our workshop on video-sharing platforms, e.g. Youtube and Bilibili, for the virtual attendees, for those who cannot attend the workshop in person because of visa issues. Remote speakers and audiences can take and ask questions from our online tools. Our hosts will be in charge of the Q&A sessions to ensure smooth interaction for remote speakers, panelists, and the audience. To accommodate speakers and audiences in different time zones, we will invite speakers and authors of accepted papers to provide pre-recorded videos in advance for flexible participation of remote attendees. The onsite talks and panel discussions will also be recorded and uploaded on our website and social media.

In order to broaden the audience with enhanced accessibility, we plan to offer a registration fee grant to individuals from underrepresented groups and individuals who have difficulty in conference registration or traveling. Our organizers from industry affiliations are actively seeking sponsorship from leading companies such as Google, NVIDIA, ByteDance, or other startup companies.

4 Format and logistics

4.1 Schedule

Overview: We plan to hold a full-day workshop including both spotlight and poster presentations of peer-reviewed papers, talks from invited speakers, and a panel discussion. Eight 30-min talks (including 25-min presentation followed by 5-min Q&A session) will be given by invited renowned speakers from the fields of computer vision, robotics and machine learning. Three 10-min spotlight representations will be selected from accepted papers, and other accepted papers will have the poster representation in the 1-hour poster session. The 1-hour panel discussion with panelists from policymakers, researchers, and industry experts will warrant a broad and in-depth discussion of the recent progress in these themes and future directions. The tentative schedule is as follow:

Time	Event	Time	Event
9:00am - 9:15am	Welcome Speech	1:00pm - 2:00pm	Poster Session
9:15am - 9:45am	Invited Talk 1	2:00pm - 2:30pm	Invited Talk 5
9:45am - 10:15am	Invited Talk 2	2:30pm - 3:00pm	Invited Talk 6
10:15pm - 10:30pm	Coffee Break	3:00pm - 3:15pm	Coffee Break
10:30pm - 11:00pm	Invited Talk 3	3:15pm - 3:45pm	Invited Talk 7
11:00am - 11:30pm	Invited Talk 4	3:45pm - 4:15pm	Invited Talk 8
11:30pm - 12:00pm	Spotlight Session	4:15pm - 5:15pm	Panel Discussion
12:00pm - 1:00pm	Lunch Break	5:15pm - 5:30pm	Closing Remark

Expected size: As this workshop covers a wide spectrum of topics in AI, robotics, computer vision, and

natural language processing, we expect the GenBot workshop to have 300-400 audience from both academia and industry with different backgrounds.

Discussion opportunities: This workshop will provide discussion opportunities in the poster session and Q&A sessions in invited talks and panel discussions. For the poster session, all attendees can have a face-to-face discussion on the topics they are interested in. For the Q&A sessions in invited talks and panel discussions, despite the questions raised by the onsite audience, we will also take questions from our social media and video-sharing platforms for live streaming before and during the workshop.

Ethical considerations: Our workshop hosts discussions on generative models that exist the need for ethical considerations of 1) generated contents including images, videos, and text, and 2) learned robot behaviors in navigation, manipulation, and reasoning. For these considerations, we will offer general ethical conduct guidelines for the workshop including the talks, discussions, the spotlight, and poster representations. This contains the requirements of the materials and presentations by presenters on whether the contents contain any personally identifiable information, information that could be deduced which have not been consented to share, or information that potentially exacerbates bias against people of a certain gender, race, or sexuality.

Plans for sponsorship: We will seek sponsorship from leading companies including Google, NVIDIA, ByteDance, and other startup companies. The cash support will be used to set registration fee grants for those who are unable to register and be awarded the Best Paper Award. These are integral for inclusive discussion and foster high-quality submissions.

4.2 Paper Submission

We welcome tiny papers up to 2 pages, short papers up to 4 pages, and long papers up to 8 pages (excluding references and supplementary materials). All papers can propose original research. Short and tiny papers can be technical reports to describe the applications of open-source frameworks and systems. Tiny papers can also propose problem statements and abstracts of possible solutions. The authors can select to make their submission archival or non-archival. Non-archival submissions allow dual submission if it is permitted by third parties. All accepted papers will be presented as posters. We will select 3 papers for short spotlight presentations and 2 papers for best paper awards from long papers with cash incentives from our sponsors.

Timeline: We will follow the suggested submission date for workshop contributions on Feb 3rd, 2025, and take three weeks to collect reviews, in order to catch the mandatory accepted paper notification deadline of March 5th, 2025. The timeline is listed as follows.

Feb 3rd, 2025: workshop submission deadline

Feb 24th, 2025: acceptance notification

Review process: All the submissions will be processed by the OpenReview system. For archival papers, we will ensure each submission will be reviewed by at least three reviewers, and each reviewer is assigned no more than three papers. The reviewing process will be double-blind. We will send reviewer invitations to researchers in the related fields and maintain a pool of emergency reviewers to ensure the author notification deadline.

5 Organizers

Our organizing committee has extensive experience in organizing and contributing to workshops in AI and/or robotics workshops, where most organizers have recently served as organizers or invited speakers including NeurIPS, ICLR, CVPR, ECCV, CoRL, and ICRA. Specifically, the organizers have adequate experience in coordinating with the invited speakers, setting up the paper review process, advertising for the workshop, etc. For example: Ziwei Liu has co-organized over 10 workshops at top AI conferences. Zhenyu Jiang led the organization team of workshops on 3D Vision and Robotics in CVPR2023. Congyue Deng also has created websites for her previous workshops (https://equivision.github.io, https://geo-lme.github.io/). The organizers are also experienced with coordinating in-person and online setups for the talks.

Ziwei Wang, NTU, he/him/his is currently an Assistant Professor at the School of Electrical and Electronic Engineering, Nanyang Technological University. Before joining NTU, he was a postdoc fellow at Robotics

Institute, Carnegie Mellon University. He received the Ph.D and the B.S degrees from the Department of Automation, Tsinghua University in 2023 and the Department of Physics, Tsinghua University in 2018 respectively. His research focuses on general robotic manipulation and efficient robotic foundation models. He has published a series of related works in TPAMI, RAL, ICRA, IROS, NeurIPS, CVPR, ICCV and ECCV. He served as an invited speaker of ECCV'24 tutorial on Recent Advances in Video Content Understanding and Generation (VENUE).

Congyue Deng, Stanford, she/her/hers is a fifth-year Ph.D. student in computer science at Stanford University, the Geometric Computing Group. Her research interests include 3D computer vision and geometric deep learning. She is particularly interested in designing feature representations for visual understanding with the awareness of symmetries and geometric relations. She has co-organized the ECCV'24 workshop on Geometry in the Large Model Era (GeoLME) and the CVPR'24 workshop on Equivariant Vision: From Theory to Practice (EquiVision).

Changliu Liu, CMU, she/her/hers is an Assistant Professor in the Robotics Institute, School of Computer Science, Carnegie Mellon University (CMU), where she leads the Intelligent Control Lab. Her research interests lie in the design and verification of human-centered intelligent systems with applications to manufacturing and transportation and on various robot embodiments, including robot arms, mobile robots, legged robots, and humanoid robots. Dr. Liu holds senior membership in IEEE, and membership in ASME and AAAI. She published the book "Designing robot behavior in human-robot interactions" with CRC Press in 2019. She is the founder of the International Neural Network Verification Competition launched in 2020. She has organized more than ten workshops in top conferences: NeurIPS, IROS, CAV, ACC, etc.

Zhenyu Jiang, UT Austin, he/him/his is a fifth-year Ph.D. student in computer science at the University of Texas at Austin, the Robot Perception and Learning Lab. His research lies at the intersection of robotics and computer vision. He is interested in building the digital twin of the real world in simulation and learning real-world robot policies with digital twins. His work has been recognized as the best student paper finalists at RSS 2022. He has co-organized the CVPR 2023 Workshop on 3D Vision and Robotics.

Haoran Geng, UC Berkeley, he/him/his is currently a PhD student at the Berkeley AI Research (BAIR). Previously, he was a visiting scholar at Stanford University through the UGVR program. He received his Bachelor's degree with honors from Turing Class, Peking University. His research interests include robotics, reinforcement learning, and 3D computer vision.

Huazhe Xu, Tsinghua, he/him/his is currently an Assistant Professor at IIIS, Tsinghua University. Before joining Tsinghua, he was a postdoc fellow at Stanford University. He received a Ph.D degree from UC Berkeley and a B.S degree from Tsinghua University. His research focuses on generalizable manipulation, reinforcement learning, and imitation learning. His representative works have been covered by media outlets such as MIT Tech Review and Stanford HAI.

Yansong Tang, Tsinghua, he/him/his is currently a tenure-track Assistant Professor of Tsinghua Shenzhen International Graduate School, Tsinghua University. His current research focuses on multimodal human activity understanding and generation. He has published a series of related works in CVPR, ECCV, NeurIPS, and TPAMI. He was the primary organizer of CVPR'24 workshop on New Trends in Multimodal Human Action Perception, Understanding and Generation (Mango) and ECCV'24 tutorial on Recent Advances in Video Content Understanding and Generation (VENUE).

Philip H.S. Torr, Oxford, he/him/his is a Professor at the University of Oxford. He received his PhD degree from the University of Oxford. After working for another three years at Oxford, he worked for six years for Microsoft Research, first in Redmond, then in Cambridge, founding the vision side of the Machine Learning and Perception Group. He has won awards from top vision conferences, including ICCV, CVPR, ECCV, NeurIPS, and BMVC. He served as general co-chair for CVPR 2019, and co-organizer of multiple workshops in CVPR, ICCV, and ECCV.

Ziwei Liu, NTU, he/him/his is currently an Assistant Professor at Nanyang Technological University, Singapore. His research revolves around computer vision, machine learning and computer graphics. He has published extensively on top-tier conferences and journals in relevant fields, including CVPR, ICCV, ECCV, NeurIPS, ICLR, SIGGRAPH, TPAMI, TOG and Nature - Machine Intelligence. He is the recipient of PAMI Mark Everingham Prize, MIT TR Innovators under 35 Asia Pacific, ICBS Frontiers of Science Award, CVPR Best Paper Award Candidate and Asian Young Scientist Fellowship. He serves as an Area Chair of CVPR, ICCV, ECCV, NeurIPS and ICLR, as well as an Associate Editor of IJCV.

Angelique Taylor, Cornell Tech, she/her/hers is an Assistant Professor in the Information Science Department at Cornell University. Before joining Cornell, she was a Visiting Research Scientist at Meta Reality Labs. She received a PhD in Computer Science from the University of California San Diego in 2021, a B.S. in Electrical and Computer Engineering from the University of Missouri-Columbia in 2015, and a A.S. in Engineering Science from Saint Louis Community College at Florissant Valley. Her research focuses on enabling robots, augmented and virtual reality systems to assist people in real-world settings. She has published related articles in ICRA, HRI, CSCW, and AAAI. She served on the Program Committee for HRI and as an Associate Editor for ICRA.

Yuke Zhu, UT Austin & NVIDIA, he/him/his is an Assistant Professor in the Computer Science Department of UT-Austin, where he directs the Robot Perception and Learning (RPL) Lab. He also co-leads the Generalist Embodied Agent Research (GEAR) lab at NVIDIA Research, which builds foundation models for embodied agents in virtual and physical worlds, particularly for humanoid robots. He focuses on developing intelligent algorithms for generalist robots and embodied agents to reason about and interact with the real world. His research spans robotics, computer vision, and machine learning. He received his Master's and Ph.D. degrees from Stanford University.