香 港 大 學

THE UNIVERSITY OF HONG KONG

RESEARCH PROPOSAL

Hierarchical Learning for Long-Horizon Robotic Manipulation Guided by Internet Video Demonstrations

Author: Junming Wang

November 13, 2024

Summary

This research project intends to propose a hierarchical learning framework to address current limitations in robotic manipulation for complex, long-horizon tasks. At the high level, the framework will aim to bridge the gap between natural language instructions and robotic execution by employing advanced vision-language models, e.g., Llama 3.2V, to decompose high-level instructions into manageable subtasks. At the lower level, the project plans to explore three innovative strategies: (1) learning implicit representations to generate low-level action primitives, (2) leveraging universal task-related visual embeddings learned from internet-scale video data to facilitate action generation, and (3) developing adaptive motion control mechanisms with feedback-driven video generation. By integrating high-level task decomposition with these low-level strategies, this research will strive to create a comprehensive system capable of handling complex, multi-step tasks across diverse environments, potentially advancing the state-of-the-art in robotic manipulation.

1 Introduction

In recent years, as hardware design advances and commercial availability increases, the development of general-purpose robots capable of assisting in everyday tasks, such as folding clothes, flushing toilets, and pouring water, has captured significant attention from academia and industry [1]. Consequently, creating a **universal manipulation model** has emerged as a critical frontier in robotics research. To achieve this goal, two crucial factors come into play: the availability of diverse, high-quality *data* and an efficient *manipulation learning framework* capable of leveraging this data.

Recognizing this need, researchers have made concerted efforts to establish large-scale interaction datasets, such as Open-X-Embodiment [2] and DROID [3]. Building upon these multi-task datasets, a new generation of generalist agents (or foundation policy models) has emerged, including RT-1 [4], Robocat [5], RT-2 [5], Octo [6], and OpenVLA [7]. These models adopt a large-scale pre-training approach to robot learning, followed by "fine-tuning" or "alignment" using carefully curated datasets to induce desired behaviours and responsiveness.

Despite these advancements, the volume of robot-specific data remains substantially smaller than the vast of *internet text and video data* available. Given that the success of foundation models is intrinsically linked to the scale of training data and the ability to extract knowledge from large-scale datasets, it is imperative to develop novel methods for building a **universal manipulation model** that can effectively harness internet-scale video data.

To address this challenge, I plan to propose a hierarchical architecture that aims to leverage the rich, diverse, and massively available video content on the Internet to augment the learning capabilities of robotic systems to accomplish complex and long-horizon tasks.

2 Literature Review and Challenges

Inspired by humans' ability to learn from videos and considering the ease of collecting demonstration data, researchers have begun exploring learning methodologies that leverage massive collections of Internet videos or carefully curated human demonstration videos. These approaches can be broadly categorized into explicit and implicit learning methods. Explicit learning methods directly map visual observations to robot actions through end-to-end architectures, often requiring action labels or demonstrations. In contrast, implicit learning methods learn intermediate representations or latent spaces that bridge the gap between visual observations and robot actions, potentially offering better generalization capabilities through these learned abstractions.

Explicit Learning Methods: ACDC [8] represents a direct approach to learning manipulation policies through video synthesis and optical flow. The method explicitly maps RGB-D observations and textual goal descriptions to action sequences through a diffusion model, establishing a direct path from observation to action without intermediate abstractions. The key innovation lies in its use of optical flow for action inference and integration of depth information for 3D transformation estimation without requiring task-specific inverse dynamics models. However, the method faces limitations in computational requirements (requiring 256 TPU), shows limited generalization across different robot morphologies.

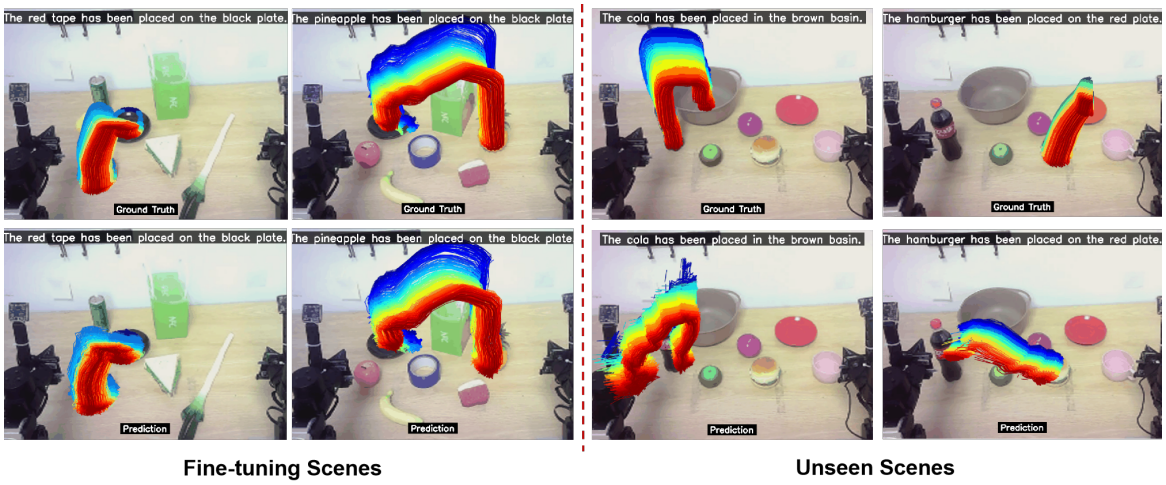

Figure 1: Qualitative results of fine-tuning the optical flow generation model on the dataset I collected.

Implicit Learning Methods: Im2Flow2Act [9] introduces object flow as an intermediate representation for manipulation learning. Rather than directly mapping observations to actions, it decomposes the learning process into two stages: first learning a flow representation that captures object motion patterns, then learning to map these flows to robot actions. The key innovation lies in using object flow as a unifying interface to connect human videos and simulated data, enabling real-world manipulation without real-world robot training data. However, the method's performance is critically affected by flow estimation accuracy, faces challenges in handling occlusions, and depends heavily on the quality of simulation data. When I fine-tuned this flow generation model on a dataset I collected, I found that the method generalized poorly, as shown in Figure 1.

IGOR (Image-GOal Representations) [10] introduces a novel approach to learning a unified and semantically consistent latent action space for both humans and robots across various embodied AI tasks. The core principle lies in compressing visual changes between initial and goal states into latent actions, which serve as embeddings of sub-tasks. This unified representation enables knowledge transfer from internet-scale human video data to robotic applications. IGOR's key contributions include: (1) developing a latent action model that captures semantically consistent actions across different embodiments, (2) utilizing this model to label large-scale video data with latent actions, thereby expanding the available data for training foundation models in embodied AI, (3) demonstrating the ability to "migrate" object movements across videos and even between humans and robots using the latent action and world models, and (4) showing how the learned latent actions can be integrated with language instructions and low-level policies for effective robot control. By bridging the gap between human demonstrations and robotic tasks, IGOR opens new possibilities for human-to-robot knowledge transfer and scalable learning in embodied AI.

OKAMI [11] utilizes object-aware retargeting as an intermediate representation to bridge the gap between human demonstrations and robot execution. Instead of direct mapping, it

learns separate representations for body motions and hand poses, which serve as an abstract interface for transfer learning. The key innovation lies in its object-aware retargeting approach that enables generalization across varying visual and spatial conditions. However, the method heavily relies on object detection accuracy, is limited to humanoid robot architectures, and may struggle with complex manipulation sequences.

DynaMo [12] learns dynamics models as intermediate representations through self-supervised learning. By jointly training inverse and forward dynamics models, it creates an abstract representation space that captures action-based causality, rather than directly mapping observations to actions. The key innovation lies in its exploitation of these dynamics representations for downstream policy learning with limited data. However, the method requires substantial demonstration data for effective dynamics learning, shows performance degradation in non-deterministic environments, and is limited by the quality of learned visual representations.

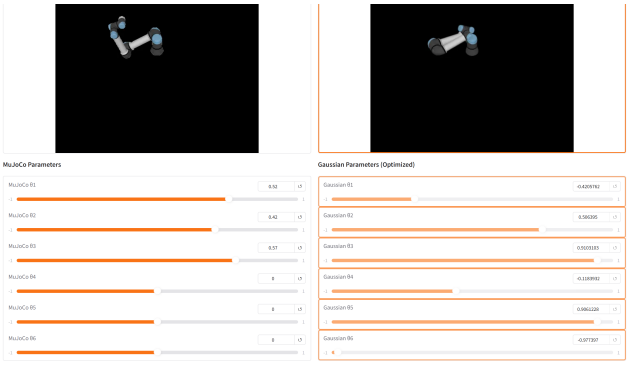

Figure 2: Severe dependence between joints often leads to abnormal joint position and movement.

Dr. Robot [13] employs differentiable rendering as an intermediate representation, creating a bridge between visual observations and control parameters through rendered images. This implicit approach enables optimization through the rendering space rather than direct mapping from observations to actions. The key innovation lies in its end-to-end differentiable pipeline that enables optimization from pixels to control parameters through the rendered intermediate space. However, the method faces significant computational overhead in rendering, shows scalability issues in complex environments, and its control performance is highly dependent on rendering accuracy and problems such as joint confusion may occur, as shown in Figure 2.

LAPA [14] introduces latent actions as an intermediate representation space, learning discrete abstractions of actions from unlabeled videos. Rather than direct mapping, it first learns a latent action space through VQ-VAE, then maps these latent actions to robot controls through fine-tuning. The key innovation lies in its ability to leverage unlabeled video data through this two-stage approach. However, the method experiences information loss during latent space encoding, requires additional fine-tuning for new robots, and may struggle with precise manipulation tasks.

In addition, For comprehensive resources and further exploration, including a selection of papers currently under investigation, please visit my GitHub repository at **Video4Robot**.

3 Research Ideas

Based on my comprehensive analysis of existing robotic manipulation learning approaches, I have identified significant limitations in their capacity to handle complex long-horizon tasks. To address these challenges, I plan to propose a novel hierarchical learning framework (in Figure 3) designed to manage extremely complex sequences of long-range operations. This framework

will aim to achieve capabilities comparable to those demonstrated by Physical Intelligence's recently released π_0 system [1].

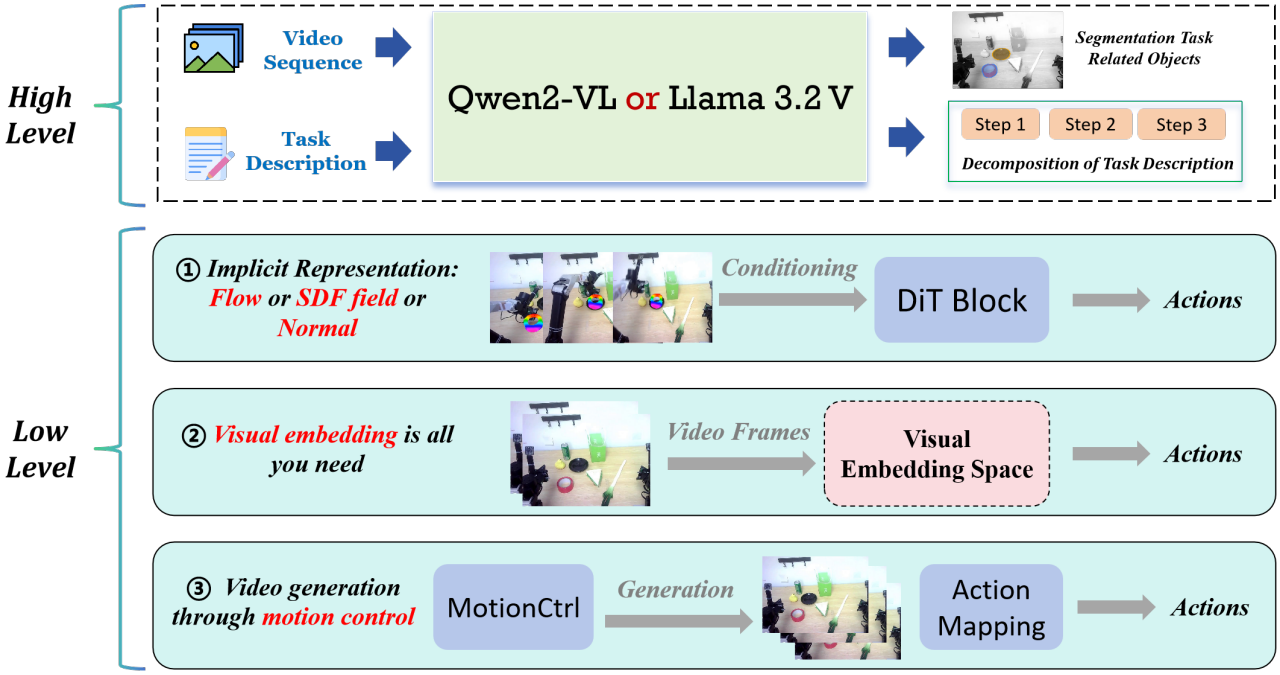

Figure 3: Hierarchical learning framework: (1) High-level task decomposition via visual-language models; (2) Low-level strategy generation using explicit and implicit representation learning.

High-level Task Decomposition: For complex instructions such as *"collect and wash dirty clothes, then fold and store them neatly"*, which involve cross-domain operations, I will argue that relying solely on implicit representations or end-to-end video generation becomes increasingly impractical as task complexity grows. My key insight will be to leverage large-scale vision-language models (e.g., Llama 3.2V, Qwen2-VL) to decompose such instructions into coherent subtasks. For instance, the clothing task can be decomposed into sequential subtasks: 1) activating the washing machine, 2) transferring clothes, 3) initiating the wash cycle, 4) retrieving clothes, and 5) executing folding operations. I will further explore how each subtask can be subdivided based on specific task attributes, allowing for granular control and optimization.

Low-level Execution Strategies: To efficiently execute these subtasks, drawing from my extensive research and iterative refinement of existing methodologies, I plan to propose three innovative approaches:

1. Learning Implicit Representations to Facilitate Action Generation

I will introduce a novel approach incorporating implicit representations as global conditions in diffusion models to generate actions. I plan to demonstrate that this method is particularly effective for subtasks involving operations in confined workspaces, such as the laundry area. For instance, when executing the subtask of "transferring clothes to the washing machine," I will utilize foundational models like optical flow or normal map generators to create initial operation trajectories between the laundry basket and the washing machine. These trajectories will then be integrated as implicit representations into conditional DDPM/DDIM/Flow models to generate specific, context-aware actions for grasping and moving clothes.

Furthermore, I will advocate for the integration of Signed Distance Fields (SDFs) into robotic operation tasks. While SDFs have been a common intermediate representation for aerial

and ground vehicle trajectory planning, I note that their application in robotic manipulation remains largely unexplored. I hypothesize that incorporating SDF information will be crucial for collision avoidance during complex manipulation processes [15], such as navigating around obstacles in the laundry room or avoiding contact with the washing machine door while loading clothes.

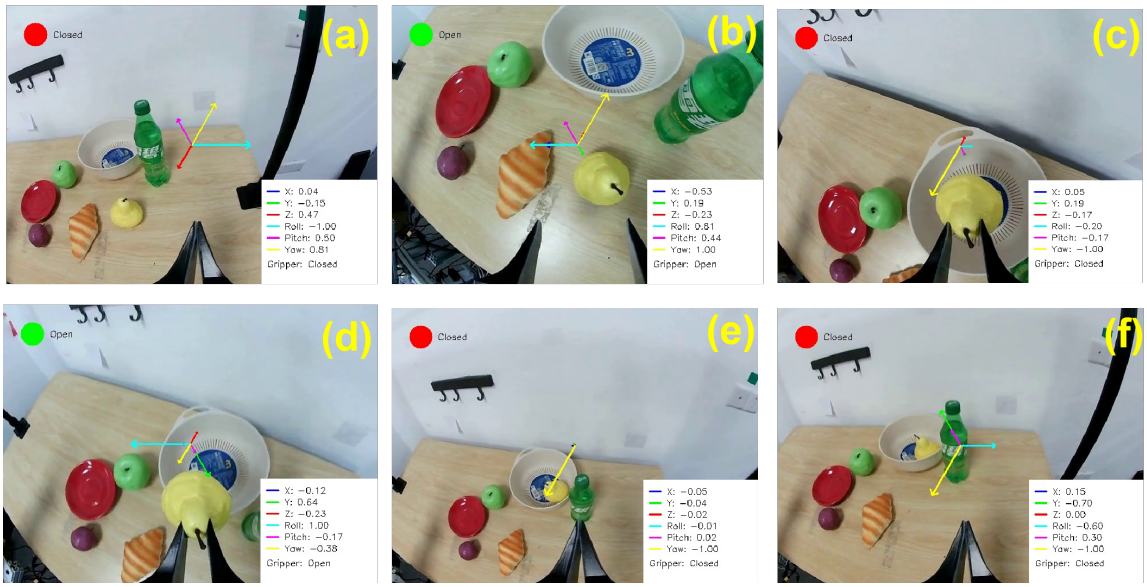

Figure 4: Using the data I collected, I fine-tuned LAPA, which can accurately predict the next action.

2. Visual Embeddings as a Universal Representation

Based on recent advancements in robotic learning, exemplified by models such as LAPA [14], Dynamo [12], and IGOR [10], which have demonstrated the efficacy of visual embedding spaces, I conducted a study to validate this approach. Specifically, I fine-tuned the LAPA model using a proprietary dataset of manipulation tasks. The model was designed to predict the robotic arm's 6D pose (x, y, z, roll, pitch, yaw) and the gripper's state (open or closed) for the subsequent frame based on the current visual input.

The experimental results, illustrated in Figure 4, reveal the superior potential of this approach compared to traditional techniques such as optical flow generation. Furthermore, the model exhibited a high success rate in downstream task transfer. Two notable observations from the experiments are:

1. As depicted in (a), when the robotic arm approaches the pear, the model accurately predicts the counterclockwise yaw rotation (indicated by a negative value), the gripper's closed state, and an increase in the Z-axis value. These predictions demonstrate the model's ability to anticipate complex spatial movements.

2. As depicted in (d), as the gripper prepares to release the pear, the model correctly forecasts the gripper's opening in the subsequent frame. Additionally, it accurately predicts the clockwise rotation trend during the grasping motion, as evidenced by the negative yaw predictions in the following frames.

Encouraged by these promising results, our future research agenda aims to build upon this foundation. We propose to develop a pre-trained model capable of learning embedding features from adjacent frames in a diverse corpus of internet-based operation videos. This model will serve as the cornerstone for our envisioned hierarchical learning framework.

Leveraging this pre-trained model, we intend to generate task-specific action videos for various subtasks. By incorporating the learned embeddings into this generation process, we

anticipate enhancing the quality and fidelity of the synthesized videos. Moreover, we plan to investigate the integration of intermediate representations, such as optical flow, which may provide additional structural and motion cues to our generation pipeline, potentially further improving the model's predictive capabilities.

This comprehensive approach aims to create a robust and versatile framework for robotic learning, capable of efficiently adapting to a wide range of manipulation tasks while leveraging the rich information available in visual data.

3. Robot Manipulation Learning through Controllable Video Generation

I plan to explore new possibilities in robot manipulation learning by leveraging the recent rapid advancements in video generation technology. My approach will utilize state-of-the-art controllable video generation models to simulate visual feedback of robotic operations, which I believe will significantly assist in developing more effective manipulation strategies.

To begin, I will predefine rough trajectories of the robot's end-effector and gripper action sequences based on specific task objectives. Using controllable video generation models, I will then generate continuous video frame sequences based on these preset trajectories, effectively simulating scene changes from the robot's perspective. During this generation process, I intend to introduce additional latent representations, such as object positions and environmental states, to enhance the realism and detail of the generated videos.

My next step will involve employing inverse dynamics models to map the generated video sequences back to precise robot joint action sequences. By comparing the generated videos with expected outcomes, I will iteratively adjust the initial trajectories and repeat the process to obtain optimal manipulation strategies. I believe this approach will combine the flexibility of video generation with the goal-oriented nature of robot learning, potentially overcoming issues such as low sample efficiency and poor generalization that are common in traditional methods.

Furthermore, I aim to explore new perspectives on human-robot collaboration research, focusing on efficiently translating human intentions into executable robot action sequences. In the future, I plan to investigate multimodal fusion, incorporating elements like tactile feedback, and explore ways to achieve more flexible manipulation learning in complex, dynamic environments.

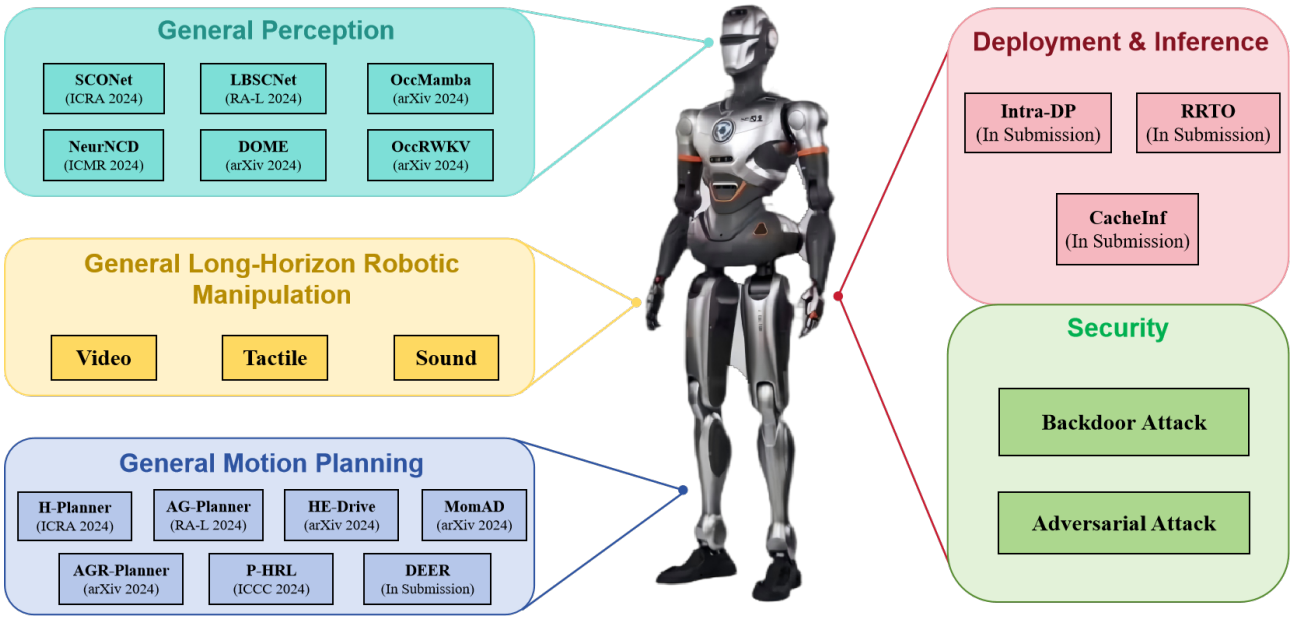

Figure 5: My long-term research vision is to develop a general embodied intelligent system that integrates advanced perception, manipulation, and motion planning capabilities with efficient edge reasoning and robust privacy protection mechanisms.

4 Long-term and Short-term Research Vision

Research Vision: My long-term research vision for the next 5 years aims to develop a comprehensive embodied intelligence system, as illustrated in Figure 5. This goal encompasses the integration of advanced models for general perception, manipulation, and motion planning while addressing critical real-world challenges. To achieve this, I will design innovative deep-learning architectures capable of understanding and executing complex, long-horizon tasks, such as assisting with household chores. These models will leverage techniques like imitation learning and diffusion models to enhance their adaptability and efficiency. Furthermore, the envisioned system will tackle two crucial issues: deployment and reasoning, and security and privacy protection. For robots operating on resource-constrained devices and in connectivity-limited environments, I will develop novel distributed reasoning algorithms and edge computing frameworks. These will enable efficient model inference and decision-making. Recognizing the sensitivity of human-robot interactions, I will design robust security and privacy protection measures, implement secure computation protocols and creating trustworthy AI systems that can explain their decision-making processes.

Academic and Professional Journey: My academic and professional trajectory has been meticulously aligned with this long-term vision, establishing a robust foundation for my future research endeavours. During my MPhil studies, I concentrated on perception [16–18] and motion planning [19–23] paradigms, while also contributing to projects aimed at accelerating on-robot model inference [24–26]. Upon transitioning to the industrial sector, I focused on robotic manipulation, with particular emphasis on acquiring manipulation skills from large-scale video datasets. This multifaceted experience has not only expanded my practical expertise but also refined my comprehension of the challenges and opportunities inherent in embodied AI. ***Consequently, it has shaped my short-term research objective of developing a generalized robotic manipulation model through learning from vast internet videos, aligning seamlessly with my overarching long-term vision.***

Future Research Focus: Looking ahead, I will pursue doctoral studies centred on "⟨learning long-horizon manipulation from videos⟩." This research direction will synthesize my past experiences and align with the broader goal of creating more capable and adaptive robotic systems. By focusing on enabling machines to learn complex, multi-step tasks from observational data, I will aim to address a critical challenge in robotics. This work will be instrumental in realizing truly general-purpose robotic assistants, capable of understanding and executing long-term, intricate tasks. This research will not only build upon my previous work but also push the boundaries of what is possible in embodied AI, contributing significantly to the field of robotics and artificial intelligence.

References

[1] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter et al., "pi0: A vision-language-action flow model for general robot control," arXiv preprint arXiv:2410.24164, 2024.

[2] A. OâNeill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain et al., "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 6892–6903.

[3] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis et al., "Droid: A large-scale in-the-wild robot manipulation dataset," arXiv preprint arXiv:2403.12945, 2024.

[4] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu et al., "Rt-1: Robotics transformer for real-world control at scale," arXiv preprint arXiv:2212.06817, 2022.

[5] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn et al., "Rt-2: Vision-language-action models transfer web knowledge to robotic control," arXiv preprint arXiv:2307.15818, 2023.

[6] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu et al., "Octo: An open-source generalist robot policy," arXiv preprint arXiv:2405.12213, 2024.

[7] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi et al., "Openvla: An open-source vision-language-action model," arXiv preprint arXiv:2406.09246, 2024.

[8] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, "Learning to act from actionless videos through dense correspondences," in The Twelfth International Conference on Learning Representations.

[9] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, "Flow as the cross-domain manipulation interface," in 8th Annual Conference on Robot Learning.

[10] X. Chen, J. Guo, T. He, C. Zhang, P. Zhang, D. C. Yang, L. Zhao, and J. Bian, "Igor: Image-goal representations are the atomic control units for foundation models in embodied ai," arXiv preprint arXiv:2411.00785, 2024.

[11] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "Okami: Teaching humanoid robots manipulation skills through single video imitation," in 8th Annual Conference on Robot Learning.

[12] Z. J. Cui, H. Pan, A. Iyer, S. Haldar, and L. Pinto, "Dynamo: In-domain dynamics pretraining for visuo-motor control," arXiv preprint arXiv:2409.12192, 2024.

[13] R. Liu, A. Canberk, S. Song, and C. Vondrick, "Differentiable robot rendering," in 8th Annual Conference on Robot Learning.

[14] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin et al., "Latent action pretraining from videos," arXiv:2410.11758, 2024.

[15] Y. Li, X. Chi, A. Razmjoo, and S. Calinon, "Configuration space distance fields for manipulation planning," Robotics: Science and Systems, 2024.

[16] J. Wang and Y. Shi, "Neurncd: Novel class discovery via implicit neural representation," in Proceedings of the 2024 International Conference on Multimedia Retrieval, 2024, pp. 257–265.

[17] J. Wang, W. Yin, X. Long, X. Zhang, Z. Xing, X. Guo, and Q. Zhang, "Occrwkv: Rethinking efficient 3d semantic occupancy prediction with linear complexity," arXiv preprint arXiv:2409.19987, 2024.

[18] J. Wang, D. Huang, X. Guan, Z. Sun, T. Shen, F. Liu, and H. Cui, "Omega: Efficient occlusion-aware navigation for air-ground robot in dynamic environments via state space model," arXiv:2408.10618, 2024.

[19] J. Wang, X. Zhang, Z. Xing, S. Gu, X. Guo, Y. Hu, Z. Song, Q. Zhang, X. Long, and W. Yin, "He-drive: Human-like end-to-end driving with vision language models," arXiv preprint arXiv:2410.05051, 2024.

[20] J. Wang, Z. Sun, X. Guan, T. Shen, Z. Zhang, T. Duan, D. Huang, S. Zhao, and H. Cui, "Agrnav: Efficient and energy-saving autonomous navigation for air-ground robots in occlusion-prone environments," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 11 133–11 139.

[21] J. Wang, Z. Sun, X. Guan, T. Shen, D. Huang, Z. Zhang, T. Duan, F. Liu, and H. Cui, "He-nav: A high-performance and efficient navigation system for aerial-ground robots in cluttered environments," IEEE Robotics and Automation Letters, vol. 9, no. 11, pp. 10 383–10 390, 2024.

[22] S. Gu, W. Yin, B. Jin, X. Guo, J. Wang, H. Li, Q. Zhang, and X. Long, "Dome: Taming diffusion model into high-fidelity controllable occupancy world model," arXiv preprint arXiv:2410.10429, 2024.

[23] Z. Zhang, T. Duan, Z. Sun, X. Guan, J. Wang, H. Liang, Y. Cui, and H. Cui, "Prediction-based hierarchical reinforcement learning for robot soccer," in 2024 IEEE/CIC International Conference on Communications in China (ICCC). IEEE, 2024, pp. 1721–1726.

[24] Z. Sun, X. Guan, J. Wang, H. Song, Y. Qing, T. Shen, D. Huang, F. Liu, and H. Cui, "Hybrid-parallel: Achieving high performance and energy efficient distributed inference on robots," arXiv preprint arXiv:2405.19257, 2024.

[25] Z. Sun, X. Guan, J. Wang, F. Liu, and H. Cui, "New problems in distributed inference for dnn models on robotic iot," in Proceedings of the 2024 Workshop on Advanced Tools, Programming Languages, and PLatforms for Implementing and Evaluating algorithms for Distributed systems, 2024, pp. 1–9.

[26] X. Guan, J. Wang, Z. Sun, Z. Zhang, T. Duan, S. Deng, F. Liu, and H. Cui, "New problems in active sampling for mobile robotic online learning," in 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2023, pp. 1155–1160.