Selen’s Interview Preparation

About me

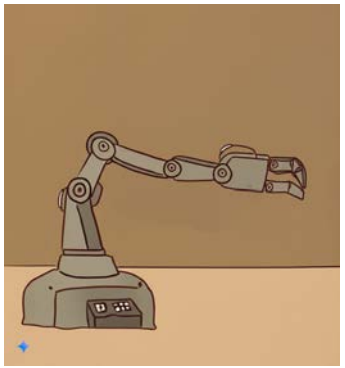
I am a third-year undergraduate(2022-2026) at Xidian University.

Researcher in robot learning currently, adversarial attacks previously.

Shanghai Jiao Tong University (SJTU)
July 2024 - Now
Research intern at *MVIG* Lab

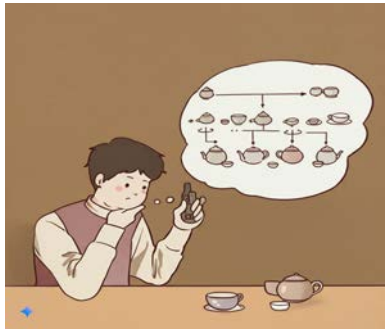
Xidian University (XDU)
September 2023 - July 2024
Research intern at *OMEGA* Lab

Robot Learning Pre



Motion Before Action

MBA

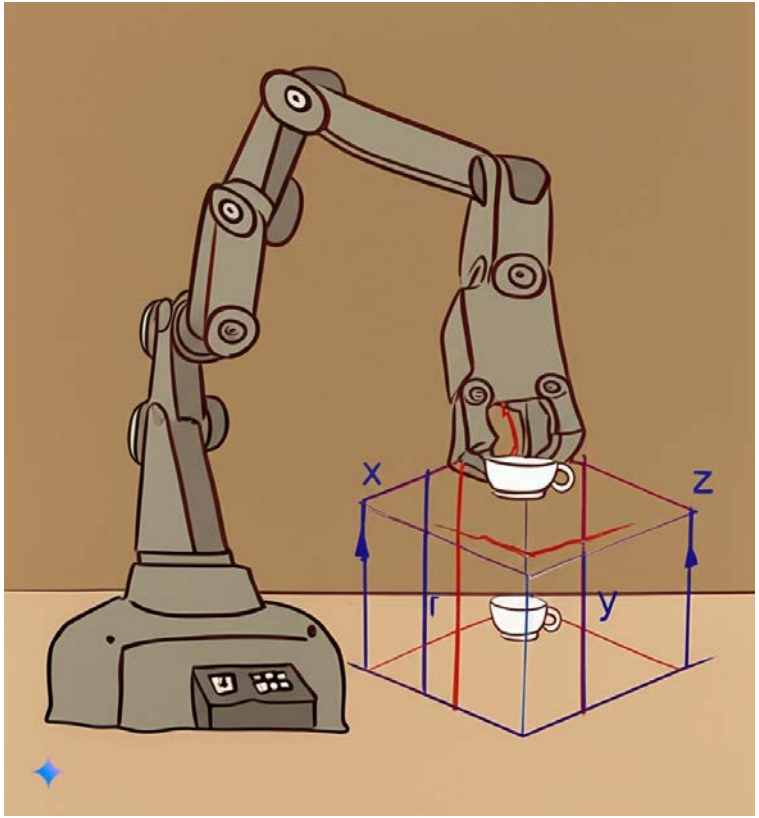


There is an interesting issue: Naive users can quickly master robot teleoperation, adapting to novel embodiments despite only having experience with human hands. Doesn't this stem from the fact that the ***human brain, through interaction, has developed extensive prior knowledge about objects*** *– particularly how to manipulate them? This knowledge, encompassing objects' shapes, affordances, and other semantic features, is largely embodiment-independent.*

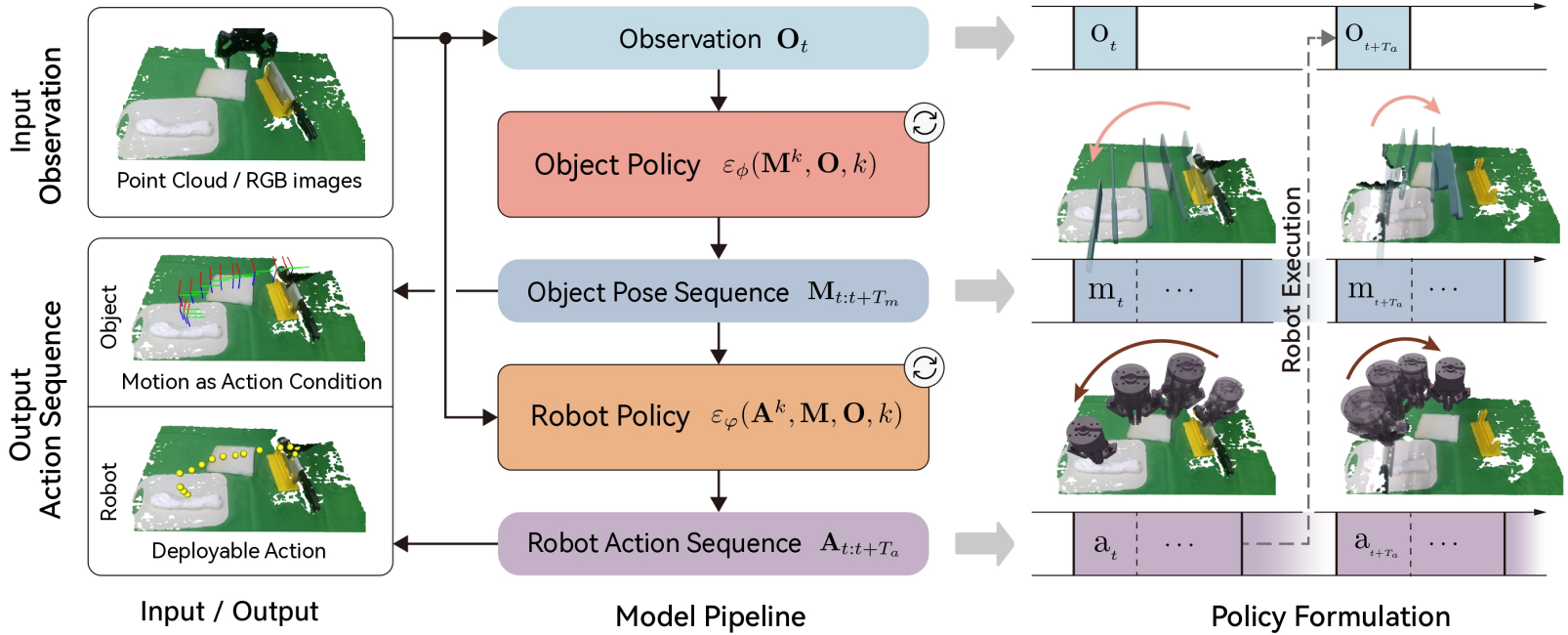
Select an aspect of the above knowledge: To bridge the gap between human motion and robotic action, ***using objects motion as a central medium representation.***

Motion Before Action

MBA is an initial exploration of this question, investigating whether object motion can predict robot actions. We leverage the mathematical similarity(6d pose representation) and spatial correlation between object poses and robot poses. From a deep learning perspective, **if diffusion models can predict robot pose sequences (i.e., actions), they should also be able to model object motion. Conversely, if current robot pose can condition a diffusion process, so too should object pose**.

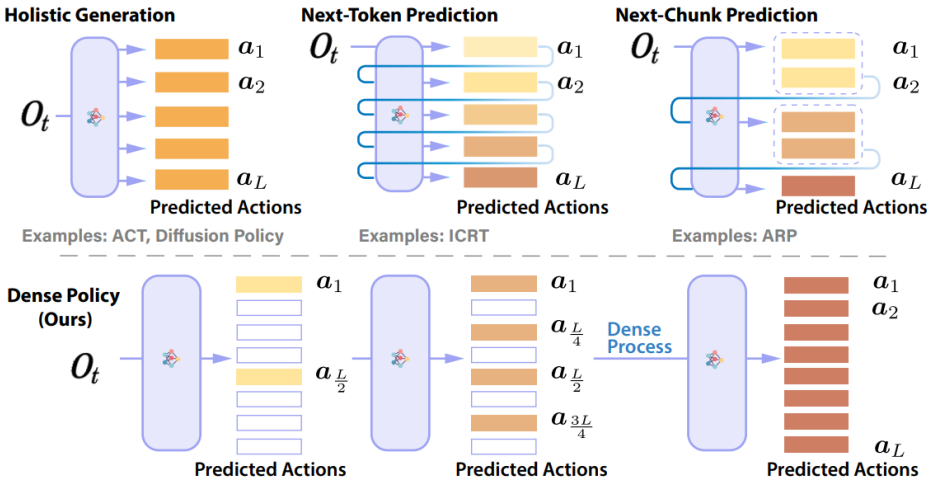


Motion Before Action

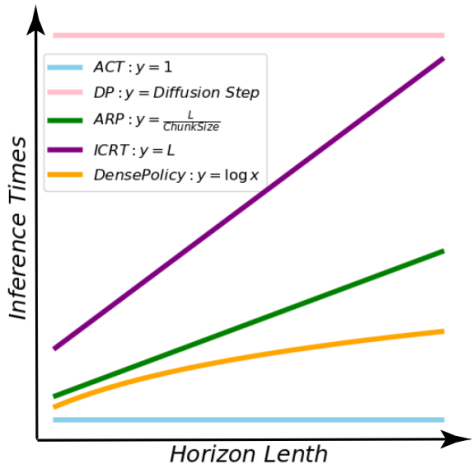


MBA is a cascaded diffusion framework. It predicts future object motion from observations, then uses this prediction to condition robot action generation. Leveraging pose consistency, MBA **avoids staged training unlike flow-based methods**. We jointly supervise object and robot poses, enabling end-to-end inference. Experiments also show MBA surpasses Flow-based methods in action generation, likely because it **avoids the 'vision-motion' gap** by directly modeling motion.

Dense Policy



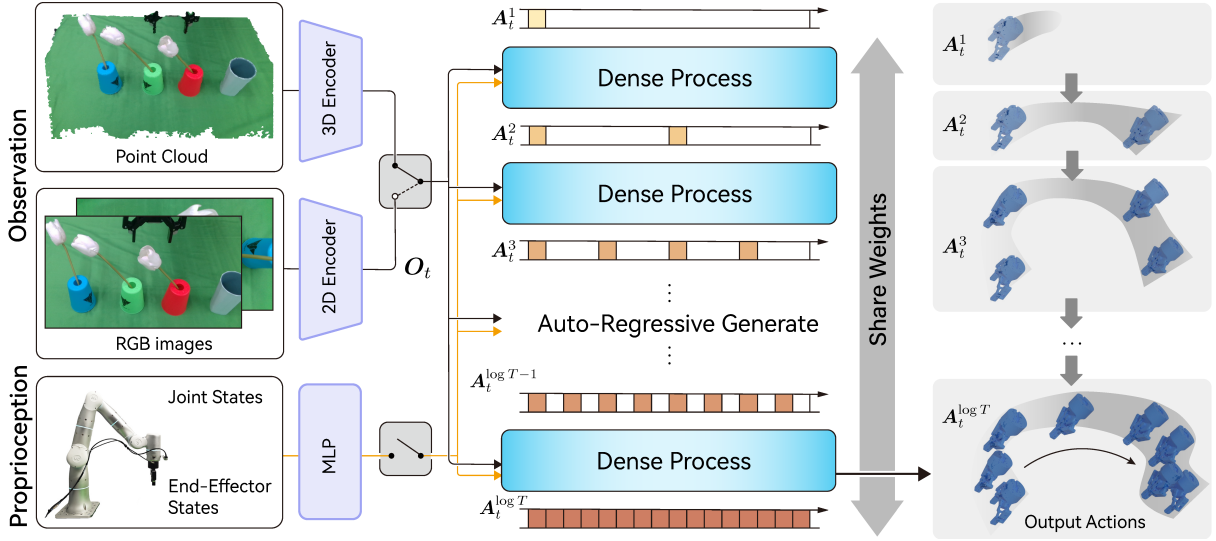
a) Different Generation Paradigms



b) Inference Time Complexity

In follow-up experiments after submitting MBA, we observed that slight camera jitter significantly degraded the performance of diffusion policies (DP). **This is because DP relies on modeling the joint distribution of predicted sequences, making it highly sensitive to distributional shifts.** While autoregressive generation is generally more robust, its next-token prediction paradigm is inefficient for robotics. Furthermore, it struggles to capture bidirectional temporal dependencies in action sequences.

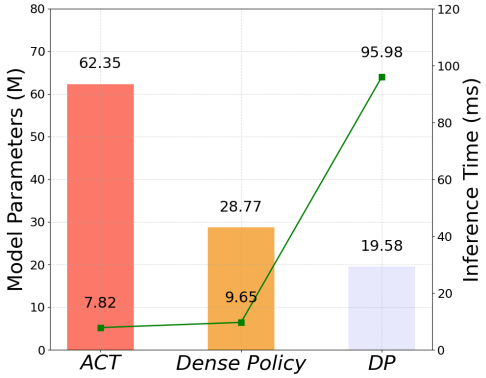
Dense Policy

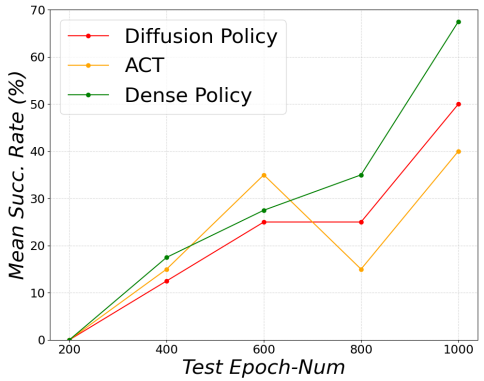


Thus we propose a bi-directional, autoregressive approach. **First, we generate a sparse set of keyframe actions from observations. Then, we iteratively refine this into a fine-grained trajectory**, with each refinement conditioned on the preceding sparse action priors. This distributes the inference complexity across a **coarse-to-fine refinement process**, promoting robustness. Furthermore, the refinement process completes in **logarithmic time**. Our experiments demonstrate that Dense Policy outperforms previous methods by a significant margin.

Dense Policy

Thus we propose a bi-directional, autoregressive approach. First, we generate a sparse set of keyframe actions from observations. Then, we iteratively refine this into a fine-grained trajectory, with each refinement conditioned on the preceding sparse action priors. This distributes the inference complexity across a coarse-to-fine refinement process, promoting robustness. Furthermore, the refinement process completes in logarithmic time. Our experiments demonstrate that Dense Policy outperforms previous methods by a significant margin.

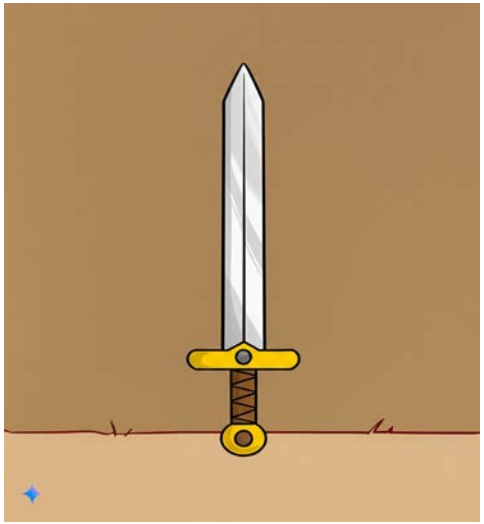




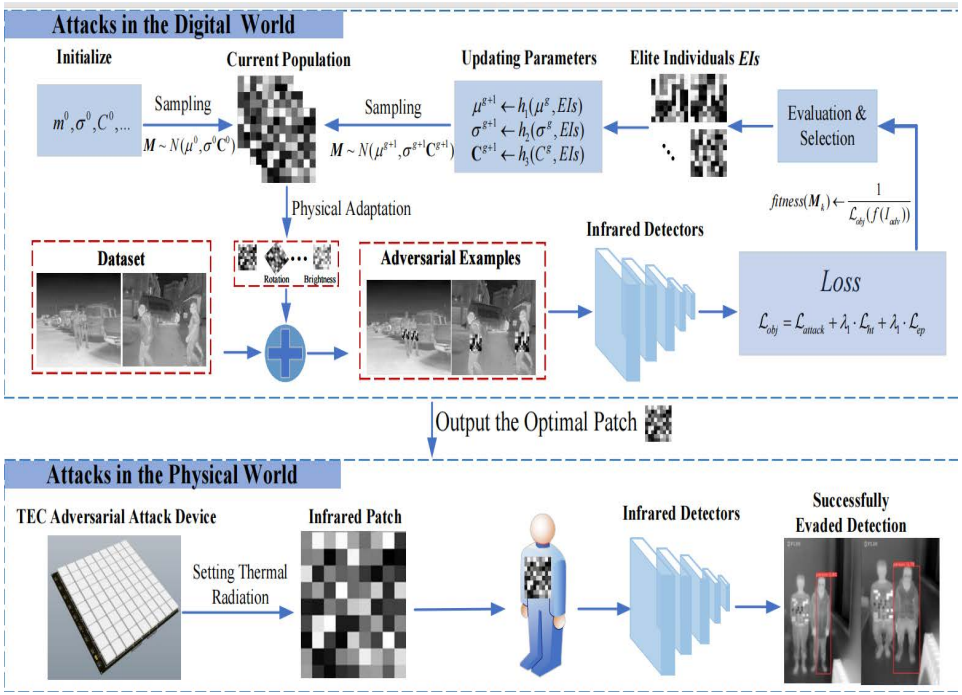
Method	Adroit		DexArt		MetaWorld							Avg
	Door	Pen	Laptop	Toilet	Bin Picking	Box Close	Hammer	Peg Insert Side	Disassemble	Shelfplace	Reach	
3D Dense Policy	**72** ± 3	**61** ± 0	**85** ± 4	**74** ± 3	**47** ± 10	**69** ± 8	**100** ± 0	**82** ± 4	**98** ± 1	**77** ± 4	**31** ± 3	**72** ± 4
DP3 [49]	62 ± 4	43 ± 6	81 ± 2	71 ± 3	34 ± 30	42 ± 3	76 ± 4	69 ± 7	69 ± 4	17 ± 10	24 ± 1	53 ± 7
2D Dense Policy	**59** ± 8	**65** ± 1	28 ± 7	**36** ± 8	**25** ± 2	**51** ± 3	**86** ± 4	**60** ± 7	**71** ± 6	**59** ± 6	**27** ± 4	**52** ± 5
DP [5]	37 ± 2	13 ± 2	**31** ± 4	26 ± 8	15 ± 4	30 ± 5	15 ± 6	34 ± 7	43 ± 7	11 ± 3	18 ± 2	25 ± 5

Method	Put Bread	Open Drawer	Pour Balls			Flower Arrangement	
	Succ. (%)	Succ. (%)	Poured (%)	Balls ↑	Complete(%)	Succ.(%)	Flowers ↑
3D Dense Policy	**85**	**45**	85	**7.30**/10	**60**	**70**	**1.0**/3.0
RISE [43]	75	40	**95**	6.85/10	25	50	0.6/3.0
2D Dense Policy	**55**	**20**	**35**	**3.30**/10	**25**	–	–
Diffusion Policy [5]	40	**20**	30	2.35/10	20	–	–
ACT [51]	35	10	30	2.75/10	20	–	–

Adversarial Attacks



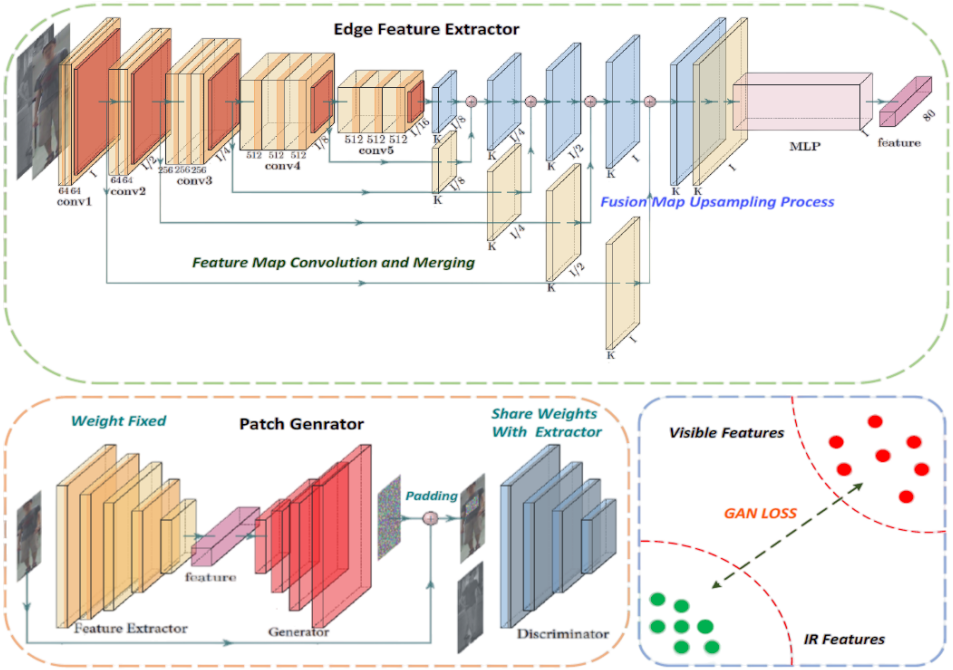
From optimization to generation



My early work explored physical adversarial attacks. Many optimization researchers now focus on this area, **as creating discrete physical attacks on black-box models is framed as an iterative optimization problem**. This work uses CMA-ES to optimize an infrared TEC plate, enabling pedestrians to evade detection.

While optimization is a convenient and effective strategy given black-box model access, I was often concerned by its limitations. **Many real-world deployed models are entirely closed, lacking APIs**, which makes iterative optimization infeasible. Furthermore, **scene and model updates can quickly render optimized attacks ineffective.**

From optimization to generation



Thus, I developed a GAN-based generative model for self-supervised, in-sample adversarial example generation. By using the semantic feature distance between adversarial and original samples as the evaluation metric, we completely bypassed model training. This approach achieved significant attack success on unseen pedestrian matching models.

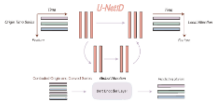
Research Interests

My research experiences have cultivated two key interests. The first is leveraging knowledge learning to enhance model reasoning. This mirrors MBA's approach of inferring actions from object motion and Dense Policy's inference of full trajectories from keyframes. Such research facilitates the evolution of intelligence in models, potentially even enabling active learning.

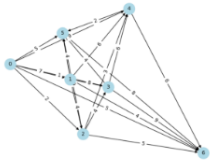
The second interest lies in modeling the learned knowledge itself. MBA, for instance, uses a generative model to learn object motion knowledge. Similarly, in my adversarial attack research, I employed a generative model to capture the semantic features of samples. These generative models serve not just for reconstruction, but also for understanding, paving the way for the development of world models.

Projects Overview

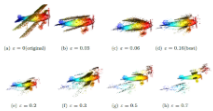
MetaPalace: Let you in a meta world of The Palace Museum
We've done what the Old Palace official website couldn't: offering 3D artifact views with single-view reconstruction and an interactive LLM-powered tour guider using RAG technology.
[website] [front-end code] [back-end code]



U-pre: U-Net is an excellent learner for time series forecasting
Time series forecasting is suited for U-Net's architecture due to its consistent input-output distributions and strong mathematical alignment. Combining U-Net with Bert-Encoder improved performance by incorporating both local and global attention.
[code] [report-cn]



AcoFlow: Heuristic Search for Maximum Flow Problem
The problem of finding the maximum flow lies in how to design better heuristic information to find the augmenting path. We boldly challenge this problem through the ant colony algorithm.
[code] [report-cn]



FGSM3D: Is the point cloud gradient perturbation attack feasible?
We tried to extend FGSM to the 3D field and achieved significant success within a certain gradient range, but the sampling method of 3D models tells us that things seem to be not that simple...
[code] [report-cn]

AgentCrossTalk: Performe a Crsosstalk between two LLM agents
This project uses the Google Gemini to create a simple chatbot application simulating two crosstalk performers performing based on user-provided topics.
[code] [website]

Beyond my core research, I've dedicated time to various projects, often outside my primary focus research, to broaden my expertise. These include MetaPalace (3D asset generation and RAG), AgentCrossTalk (LLM based multi-agent dialogue), U-Pre (time series forecasting), and ACOFlow (optimization algorithms). These diverse projects are all funny and meaningful~

Thanks for your Listening