Generative Computer Vision for Physical Intelligence

Ruoshi Liu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

© 2025

Ruoshi Liu

All Rights Reserved

Abstract

Generative Computer Vision for Physical Intelligence

Ruoshi Liu

The fundamental purpose of human perception is to guide action. For this, my research is driven by the goal of building computer vision systems that can interact with the physical world. My approach toward this goal surrounds the use of generative models. Training generative models on unlabeled data is a powerful, scalable, and comprehensive way to learn about the world, and the knowledge learned through this process can be leveraged for revolutionary applications such as general-purpose industrial robots, autonomous search and rescue agents, and home robots that do your housework. However, deploying generative models in embodied AI systems for physical interaction poses several fundamental challenges. First, current generative models lack spatial understanding of the dynamic world, and second, they are not physically grounded, as evidenced by the hallucination problem. Third, interaction skills learned through human data are upper bound by human-level performance, which is often suboptimal.

This dissertation presents three approaches to address these challenges, each with examples and real-world deployable systems. Collectively, my dissertation advances the capabilities of generative models in physical intelligence by integrating techniques in computer vision, computer graphics, and robotics. It has had significant practical impacts, as evidenced by the widespread adoption of the open-source tools derived from this work in both research and industry settings. Looking forward, my research aims to further explore multimodal and multisensory perception, perception-action representation learning, and self-supervised robot learning.

Table of Contents

Acknowledgments xv

Dedication xvii

Chapter 1: Introduction and Background 1

1.1 Learning the Visual Prior for Embodied Intelligence 2

1.2 Grounding Vision Generative Models with Physical Structures 3

1.3 Embodied Self-Learning 5

1.4 Dissertation Overview 6

I Learning the Visual Prior 7

Chapter 2: Learning Geometric Prior for 3D Perception 8

2.1 Introduction 8

2.2 Related Work 10

2.3 Method 12

2.3.1 Learning to Control Camera Viewpoint 14

2.3.2 View-Conditioned Diffusion 15

2.3.3 3D Reconstruction 16

2.3.4 Dataset 17

2.4 Experiments 18

2.4.1 Tasks 19
2.4.2 Baselines 20
2.4.3 Benchmarks and Metrics 21
2.4.4 Novel View Synthesis Results 21
2.4.5 3D Reconstruction Results 23
2.4.6 Text to Image to 3D 24
2.5 Discussion 24
2.5.1 Future Work 24

Chapter 3: Learning Motion Prior for Robot Planning 25
3.1 Introduction 25
3.2 Related Work 27
3.3 Approach 29
3.3.1 Overview 29
3.3.2 Video Generation 29
3.3.3 Track then Act 30
3.4 Experiments 31
3.4.1 Experimental Setup 32
3.4.2 Object Rotation Task 32
3.4.3 Granular Material Scooping Task 34
3.4.4 Table Top Sweeping Task 34
3.4.5 Push-Shape Task (Long Horizon) 35
3.5 Limitations 36

3.6 Summary 37

II Physical Grounding of Generative Models 38

Chapter 4: Grounding Generative Priors with Differentiable Rendering 39

4.1 Introduction 40

4.2 Related Work 41

4.3 Methods 44

4.3.1 Problem Formulation 44

4.3.2 Overview of Approach 45

4.3.3 Generative Models 46

4.3.4 Differentiable Rendering of Reflection 46

4.3.5 Optimization for Inference 50

4.4 Experiments 51

4.4.1 3D Object Reconstruction 51

4.4.2 3D Human Reconstruction 52

4.4.3 Cars as Infrared Mirrors 53

4.4.4 Ablation Studies 53

4.5 Summary 55

Chapter 5: Grounding Generative Motion Prior with Kinematic Constraints 56

5.1 Introduction 56

5.2 Approach 58

5.2.1 Overview and Formulation 58

5.3 Experiments 61

5.3.1 Self Model Quality 61
5.3.2 Robot Pose Reconstruction 63
5.3.3 Applications 64
5.4 Related Works 67
5.5 Summary 68

III Embodied Self-Learning 69

Chapter 6: Self-Supervised Policy Learning with Generative Prior 70
6.1 Introduction 70
6.2 Related Works 73
6.2.1 Autonomous Underwater Manipulation 73
6.2.2 Sensorimotor Policy 74
6.3 Methods 75
6.3.1 Hardware / Software System 75
6.3.2 Manipulation Policy with Behavior Cloning 76
6.3.3 Self Learning for Policy Acceleration 77
6.4 Experiments 79
6.4.1 Behavior Cloning 79
6.4.2 Self Learning for Policy Acceleration 82
6.5 Summary 84

Chapter 7: From Policy Learning to Tool Design 86
7.1 Introduction 86
7.2 Related Work 90

7.2.1 Origami-Inspired Robot and Tool Design 90
7.2.2 Hardware-Software Co-Design 91
7.2.3 Dynamic Manipulation of Deformable Objects 92
7.3 Methods 93
7.3.1 Inverse Design Framework 93
7.3.2 Paper Airplane 96
7.3.3 Kirigami Gripper 98
7.3.4 Design Optimization 100
7.4 Experiments 103
7.4.1 Quantitative Experiments 103
7.4.2 Size of Design Space 104
7.4.3 Adaptation of Kirigami Gripper to Different Objects 105
7.4.4 Grasping with Designed Gripper 106
7.5 Summary 106

Chapter 8: Future Directions 107
8.1 Spatial Intelligence and Multimodal Perception for Embodied AI 107
8.2 Superhuman Perception and Interaction Skills 107

Conclusion or Epilogue 109

References 110

List of Figures

1.1 Overview of my research goals and approaches and their connections. 1

1.2 (a) Learning visual prior from the internet by equipping large-scale pretrained generative models with a spatial and physical understanding of the world. 2

1.3 (b) Physically grounding generative models mitigates hallucination and enables generalization to novel data modality. 4

1.4 Built upon the foundation of physically grounded generative models, (c) Embodied self-learning is used to improve the perception and interaction skills learned from humans through self-supervised interaction and exploration 5

2.1 Given a single RGB image of an object, we present Zero-1-to-3, a method to synthesize an image from a speci ed camera viewpoint. Our approach synthesizes views that contain rich details consistent with the input view for large relative transformations. It also achieves strong zero-shot performance on objects with complex geometry and artistic styles. 8

2.2 Viewpoint bias in text-to-image models. We show samples from both Dall-E-2 and Stable Diffusion v2 from the prompt "a chair". Most samples show a chair in a forward-facing canonical pose. 10

2.3 Zero-1-to-3 is a viewpoint-conditioned image translation model using a conditional latent diffusion architecture. Both the input view and a relative viewpoint transformation are used as conditional information. 12

2.4 3D reconstruction with Zero-1-to-3. Zero-1-to-3 can be used to optimize a neural eld for the task of 3D reconstruction from a single image. During training, we randomly sample viewpoints and use Zero-1-to-3 to supervise the 3D reconstruction. 13

2.5 Novel view synthesis on Google Scanned Objects [60]. The input view shown on the left is used to synthesize two randomly sampled novel views. Corresponding ground truth views are shown on the right. Compared to the baselines, our synthesized novel view contain rich textual and geometric details that are highly consistent with the ground truth, while baseline methods display a signi cant loss of high-frequency details. 14

2.6 Novel view synthesis on RTMV [64]. The input view shown on the left is used to synthesize 2 randomly sampled novel views. Corresponding ground truth views are shown on the right. Our synthesized view maintains a high delity even under big camera viewpoint changes, while most other methods deteriorate in quality drastically. 16

2.7 Novel view synthesis on in-the-wild images. The 1st, 3rd, and 4th rows show results on images taken by an iPhone, and the 2nd row shows results on an image downloaded from the Internet. Our method works are robust to objects with different surface materials and geometry. We randomly sampled 5 different viewpoints and directly showcase the results without cherry-picking. We include more uncurated results in the supplementary materials. 18

2.8 Diversity of novel view synthesis. With an input view, we x another viewpoint and randomly generate multiple conditional samples. The different results re ect a range of diversity in terms of both geometry and appearance information that is missing in the input view. 19

2.9 Qualitative examples of 3D reconstruction. The input view images are shown on the left. For each method, we show a rendered view from a different angle and the reconstructed 3D mesh. The ground truth meshes are shown on the right. 21

2.10 Novel View Synthesis from Dall-E-2 Generated Images. The composition of multiple objects (1st row) and the lighting details (2nd row) are preserved in our synthesized novel views. 23

3.1 Real-World Visuomotor Policy Learning via Video Generation. Dreamitate is a visuomotor policy learning framework that ne-tunes a video generative model to synthesize videos of humans using tools to complete a task. The tool's trajectory in the generated video is tracked, and the robot executes this trajectory to accomplish the task in the real-world. 25

3.2 Method Overview. For each task, (Step 1) we capture stereo camera recordings of human demonstrations of tool use from a top and a side view. (Step 2) The video model is ne-tuned to recreate these demonstrations from an initial scene image. (Step 3) In a new, unseen scenario, a stereo image pair is provided to the video model to generate the manipulation. The tool's trajectory in the generated video is tracked and executed by the robot to complete the task. 28

3.3 Aligning Human and Robot Behavior via Trackable Tools. We use custom tools with known CAD models for human demonstrations, which facilitate precise 3D tracking, across four real-world manipulation tasks. We design unique tools (e.g., a spoon for scooping) to show the generalizability of our approach for varying tasks. 30

3.4 Rotation Qualitative Results. For this task, we provide as input the image on the left for each row, with an example of a successful rotation overlaid. Our video generation-based approach succeeds, while Diffusion Policy fails to select stable grasping points on the object, causing it to slip during manipulation. 31

3.5 Scooping Qualitative Results. We provide the input image on the left (without the overlays provided for illustration) to our model and Diffusion Policy, and show the output trajectory on the right. Our approach succeeds in most trials, while Diffusion Policy is often distracted by other objects in the scene, placing material in the wrong container and failing to scoop accurately. 33

3.6 Sweeping Qualitative Results. We provide the input image on the left (minus the white and green circles overlaid for illustration) to our model and Diffusion Policy, and show the output trajectories on the right. Our approach again achieves high success rate even on this challenging, multi-modal task, whereas Diffusion Policy generates trajectories that often collide with obstacles and fail to sweep to the target. 35

3.7 Push-Shape Qualitative Results. We provide the image on the left, including the gray overlay indicating the target object position and orientation, to our model and Diffusion Policy, and show output trajectories on the right. Our method produces appropriate pushing actions to adjust the object's position. In contrast, Diffusion Policy fails to effectively rotate the object to match the mask. 36

4.1 We introduce a method to reconstruct the 3D position and pose of a person from their thermal re ections in everyday objects with non-planar surfaces. Given an RGBD image and a thermal image in the middle of the gure, our method is able to recover the 3D mesh of the person (blue) as well as the objects (pink), even though they are not within the eld of view of the camera system. Our system never sees the scene on the left, which is only shown for visualization purposes. . 39

4.2 High-level overview of our analysis-by-synthesis framework. We sample random initializations from the latent space of pretrained generative models of humans and objects in 3D. Through a differentiable rendering process, we synthesize a re-ection image of a human body on object surfaces. This synthesized re ection is compared with the observed re ection with an L_1 loss. Gradients are backpropagated through differentiable rendering and generative models to the latent variables. 42

4.3 Differentiable Rendering of Re ection. Ray direction shown reverses the physical propagation direction of light by Helmholtz reciprocity. 45

4.4 3D Object Reconstruction from RGBD 47

4.5 3D Human Reconstruction (visualized from another camera view). From an RGBD image, we recover the 3D location $T_{>19}$, pose $q_{>19}$, and shape $z_{>19}$ of each object. A marching cube visualization of the reconstructed 3D objects is shown in pink. With reconstructed objects, we recover 3D location T, pose q, and shape z of the human from a denoised thermal input showing re ections of the human on object surfaces, which we visualize in blue. We also include the original scene and our reconstruction from a calibrated third-camera view for comparison. This image is not seen by our system during reconstruction. The black mesh where the objects are located is the depth pointclouds captured by the RGBD camera. 48

4.6 Real-world 3D human reconstruction from thermal re ections of cars. A diverse set of human poses can be reconstructed by using the surfaces of different types of cars as infrared mirrors. RGB input (1st row) and thermal input (2nd row) captured by a depth camera and a thermal camera are used as input to our method. Our reconstruction (3rd row) is compared with the original scene (4th row), both rendered/captured from another camera viewpoint. 49

4.7 Ablation Studies. Visualization of reconstruction obtained from ablated variations of our full model. While all variations can still nd the 3D location of humans relatively accurately, the ne-grained details of human poses are signi cantly improved in our full model. 50

4.8 Visualization of DeepSDF Surface Normals. Within a 1.6 cm 1.6 cm area. From the visualization, we can clearly see an improvement in surface smoothness at a small scale, which is bene cial to the differentiable rendering process. The X-Y plane (horizontal) indicates the location on a surface with a step size of 0.2mm. Given the unit surface normal vector at a point on the grid $\hat{n}_{grid}$, we compute its dot product with the unit surface normal vector at $\theta = 0^{\circ}$, and plot this value on the Z-axis. This shows the curvature of the surface as well as its level of smoothness. . 55

5.1 We introduce Differentiable Rendering of Robots (Dr. Robot), a robot self-model which is differentiable from its visual appearance to its control parameters. With it, we can perform control and planning of robot actions through image gradients provided by visual foundation models. 56

5.2 Rendering Pipeline. Our robot model is composed of 3 differentiable components. Forward kinematics projects a pose vector into a skeleton, Implicit LBS projects 3D Gaussians to the robot surface, and Appearance Deformation adjusts appearance of 3D Gaussians. 59

5.3 Visual Quality of Robot Model. Here, we showcase the learned robot model's visual quality by comparing it with results obtained from Deformable Gaussians [206]. Due to the complicated kinematic structure of a robot, [206] is unable to fit the deformation while ours can. 61

5.4 Robot Pose Estimation from a Single Image. From an input image, we perform optimization to reconstruct the joint angles of the robot and overlay the final rendering of the robot on top of the input image. These results show that accurate robot poses can be reconstructed from only a single image through our robot model. This also demonstrated that our robot model provides high-quality gradients for action optimization. 62

5.5 Text-to-Robot Hand Gestures We perform optimization of joint angles of a Shadow Hand to maximize the CLIP similarity between the rendered image and text prompt. We show the optimization process (left) as well as final outputs of different prompts (right). 64

5.6 Dr. Robot for text-conditioned robot control. We show that robot actions can be extracted from videos predicted by a fine-tuned text-to-video model and be executed for robot control. 65

5.7 Motion Retargeting with Video Point Tracker. We perform optimization on robot action trajectories to minimize the chamfer distance between the point tracks in the rendered video and a demonstration video, allowing us to transfer motion across the embodiment gap. 68

6.1 AquaBot combines behavior cloning with self-learning to optimize fully autonomous end-to-end visuomotor policies to achieve efficient manipulation skills across a wide range of tasks, including generalization to unseen objects (Rock Grasping), long horizon tasks (Trash Sorting) and robustness against large perturbations from unmodelled deformable and articulated objects (Rescue Retrieval). 71

6.2 Hardware Setup. Our accessible hardware platform ($2000 USD) consists of 2 cameras and a parallel jaw gripper for research and development of underwater visuomotor policy learning. 74

6.3 Learning Framework. In the rst stage (a), we train our base policy by learning from human demonstrations from of ine data. In the second stage (b), we roll out the behavior-cloned policy to collect more self-learning data to learn a surrogate model in an online fashion, which optimizes the motor speed in an online fashion. 75

6.4 Dynamic and Robust Manipulation. We plot policy outputs below their corresponding third-person views (a) shows how the robot decelerates by applying backward forces when it is close to the object, demonstrating pro ciency in underwater dynamics (b) shows how the policy will retry after unstable grasps, demonstrating robustness. 77

6.5 Self-supervised Optimization of Manipulation Ef ciency. By self-learning (SL), AquaBot learns to accelerate a manipulation policy learned from Behavior Cloning (BC) through trial-and-error. After only 100 iterations, it can perform the same manipulation task more ef ciently than vanilla BC policy and human experts. 80

6.6 Human Manipulation Ef ciency vs. Motor Speed. High motor speed leads to unstable manipulation and mistakes, and low motor speed leads to slow robot movements, both hurting the manipulation ef ciency. There exists a sweet spot for the motor speed, balancing stability and speed. 83

7.1 Self-supervised Design of Paper Tools. Through trial-and-error, PaperBot autonomously learns how to design and use paper tools directly in the real world. Given only 100 trials (3 hours), our fully autonomous system discovers a paper airplane folding and throwing strategy that ew further than the best human design after the same number of trials (top), and learns how to cut and actuate a kirigami gripper that exerts 0.93N of force (bottom), equivalent to the weight of over four strawberries. For our system's nal design, please see Fig. 7.5. 87

7.2 Paper Tools. Paper is an affordable and versatile medium for constructing a variety of different paper tools. 89

7.3 Approach Overview. Our framework samples paper designs for a tool, builds them, actuates a robot to perform on a task with the tool, and perceives its performance. By learning a surrogate model to predict the utility of a design, we obtain a differentiable model that allows us to solve inverse design tasks with gradient-based optimization. The above gure shows how our framework applies to two design tasks of paper airplanes (top) and kirigami grippers (bottom). 90

7.4 System Setup. We visualize the workspace of PaperBot while designing paper airplanes (a-c) and kirigami grippers (d-e). (a) shows the full workspace, including two RealSense D435i cameras to measure travel distance over the runway. (b) shows two xArm7s folding a paper airplane. (c) shows the sticky holders. (d) shows the xArm 7s retrieving a gripper cut by the Cricut Maker 3 and actuating it. (e) shows load cells that we use to measure the force of the candidate gripper design. 91

7.5 Parameterization and Learning Process. Left shows the parameterization of the tools. Right shows the intermediate designs during the learning process as well as the best human-fold plane and adapted gripper designs for smaller and larger objects. These iterations correspond to the teaser gure. See Fig. 7.1 for the actual performance of these designs. 93

7.6 Quantitative Performance of paper airplanes (top) and kirigami grippers (bottom) per optimization iteration, designed by our method (blue) versus baselines. 94

7.7 Exploring Larger Design Spaces. When the paper airplane design space increases from 3 to 5, random and grid search performs signi cantly worse. In contrast, our approach can ef ciently navigate the larger design space to nd better design/throwing parameters. Further, regardless of the design space, our approach signi cantly outperforms all baselines. 96

7.8 Paper Airplanes Landing Positions. We visualize the top-down paper airplane mask over the course of training. Compared to SNES, PaperBot was able to better balance exploration and exploitation, escaping the local minima at around 2.5 meters, and eventually beating the human baseline. 97

7.9 Adapted Grippers Under Actuation. Changing the distance between load cells changes where gripping force is measured, and thus, the optimal bending point for the gripper designs. Given only 50 adaptation trials, PaperBot discovered subtle design changes which signi cantly impacted the mechanical structure of the paper, leading to grippers tailored to different object sizes. 99

7.10 Grasping with Adapted Gripper Design. We compare the grasping process of our gripper design before adaptation and after adaptation for larger and smaller objects. Note that the “before adaptation” gripper is optimized for 5cm objects. . . 100

7.11 Grasping of everyday objects using the optimized gripper design. We used the grippers with a design optimized by our method to perform grasping of objects with various sizes, surface materials, weights, and rigidity, as well as fresh fruits as shown in the bottom row. 101

List of Tables

2.1 Results for novel view synthesis on Google Scanned Objects. All metrics demonstrate that our method is able to outperform the baselines by a signi cant margin. 20

2.2 Results for novel view synthesis on RTMV. Scenes in RTMV are out-of-distribution from Objaverse training data, yet our model still outperforms the baselines by a signi cant margin. 22

2.3 Ablation studies on the size of the training dataset and training from scratch. We evaluate on Google Scanned Objects. These results indicate that pretraining with Stable Diffusion is critical to the success of our approach. 22

2.4 Results for single view 3D reconstruction on GSO. Note that our volumetric IoU is better than the compared methods by a large margin. 24

3.1 Tasks Summary. We detail the train and test setup for each task above. For each task, we use a distinct set of objects during testing than training. 32

3.2 Quantitative Results. We compare our method to Diffusion Policy on four tasks quantitatively. We report success rates (successful trials / total trials) for rotation, scooping, and sweeping tasks. For Push-Shape, we report mean intersection-over-union (mIoU) and average rotation error (Rot. Error). Our approach performs well across all tasks, whereas Diffusion Policy shows worse performance overall and degrades in the more challenging sweeping and Push-Shape scenarios. 33

4.1 Quantitative evaluation of our reconstructed 3D human. We used two evaluation methods by comparing the extracted 2D keypoints and 3D skeleton with a calibrated 3rd camera view. Object type indicates the type of objects serving as re ectors. Columns 3-5 are three variations of our full model with some parts ablated. Random shows the corresponding metric if a random sample were to be drawn from the HumanEva[200] dataset, which includes diverse poses in daily human activities. Numbers show the average normalized Euclidean distance between reconstruction and ground truth. 52

5.1 Visual and Geometric Quality of Rendering. The table shows PSNR (higher is better) and Chamfer distance (lower is better) metrics for different methods and systems. 62

5.2 Single-View Robot Pose Reconstruction. 63

6.1 Object Grasping Performance 79

6.2 Trash Sorting Performance 80

6.3 Success Rate Comparison for Rescue Retrieval 81

6.4 Performance vs. Action Horizon 81

6.5 Learned Speed Generalizes to Other Tasks. 84

7.1 Adapted Tool Design.Force measurement of original gripper optimized for grasping 5cm-sized objects compared with adapted grippers optimized for different object sizes. 102

Acknowledgements

I would like to thank my advisor, Carl Vondrick, for his guidance, inspiration, and encouragement throughout my years at Columbia. Carl has been my role model in research for his passion, his curiosity to explore, and his out-of-the-box thinking. But perhaps what truly sets him apart is his love for his students and the ability to give the right guidance at the right time. Looking back, there were perhaps a hundred different ways my PhD could have gone wrong, and somehow Carl managed to always steer me back to the right path. Above all else, Carl taught me to be humble, which I believe to be one of the most valuable yet underrated qualities in today's academia and society, and I hope to carry that to my future research endeavors.

I also want to thank my two important mentors, Shuran Song and Shree Nayar for providing invaluable advice and support throughout my PhD. It is often easy to fall into a narrow mindset as a PhD student but working with them has opened me up to very different perspectives, research tastes, and advising styles, which has been consistently instrumental to my research journey. I would also like to thank my undergraduate research advisors Pengyu Hong, Seth Fraden, and Gabriella Sciolla, who introduced me to academic research.

PhD could be a very lonely journey because everyone ghts their own battles, but I was extremely lucky to be surrounded by amazing friends and collaborators. I want to thank my lab mates, Alper Canberk, Arjun Mani, Basile Van Hoorick, Chengzhi Mao, Dídac Surís, Ege Ozguroglu, Ishaan Chandratreya, Junbang Liang, Lennart Schulze, Mia Chiquier, Purva Tendulkar, Revant Teotia, Sachit Menon, Scott Geng, Sreehari Rammohan, Sruthi Sudhakar, and Utkarsh Mall, for the drinks we got after deadline pushes, and the occasional freak-outs from the mice and cockroaches we saw together at CEPSR.

I would also like to thank my close collaborators including Rundi Wu, Huy Ha, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, Cheng Chi, Matt Deitke, Andrew Owens, Mengxue Hou, Ludwig Schdmidt, Abdullah Hamdi, Achal Dave, Changxi Zheng, and Simon Stent for making my life a lot less miserable.

Outside of research, I want to thank my friends at Columbia including Colman Leung,

Honglin Chen, Jeremy Klotz, Jingxi Xu, Mandi Zhao, Matthew Beveridge, Mengda Xu, Qijia Shao, Ruilin Xu, Samir Gadre, Tao Long, Xiaofeng Yan, Yuhong Zhong, Zeyi Liu, Zhanpeng He, Zhenjia Xu, and Ziang Ren for their company and emotional support.

I want to thank Toyota Research Institute for generously supporting my PhD research both nancially and intellectually. I also would like to thank my thesis committee members, Richard Zemel, Jiajun Wu, and Derek Hoiem, as well as other mentors I encountered, including Alyosha Efros, Matei Ciorcarlie, Peter Belhumeur, Shangzhe Wu, and Yunzhu Li.

I want to thank my cat Pattie, who I decided to adopt the day I got my PhD offer, which turned out to be the one of the best decisions I've ever made in my life.

At the end, I want to thank my parents and my sister for their unconditional and continuous support for my career.

To world peace.

Chapter 1: Introduction and Background

The fundamental purpose of human perception is to guide action. For this, this dissertation is driven by the goal of building computer vision systems that can interact with the physical world. My approach toward this goal surrounds the use of generative models. Training generative models on unlabeled data is a powerful, scalable, and comprehensive way to learn about the world, and the knowledge learned through this process can be leveraged for revolutionary applications such as general-purpose industrial robots, autonomous search and rescue agents, and home robots that do your housework. However, deploying generative models in embodied AI systems for physical interaction poses several fundamental challenges. First, current generative models lack a spatial understanding of the dynamic world, and second, they are not physically grounded, evidenced by the hallucination problem. Third, interaction skills learned through human data are upper bound by human-level performance, which is often suboptimal.

To address these challenges, my work focuses on developing spatially and physically grounded perception systems that can improve themselves by self-guided interaction with the physical world. Breaking this down, my work tackles three fundamental aspects of embodied AI systems:

Figure 1.1: Overview of my research goals and approaches and their connections.

(a) Learning the visual prior by equipping large-scale vision generative models with a spatial and physical understanding of the world for interaction with the environment [1, 2, 3, 4, 5, 6].
(b) Physically grounding generative models by adding structures signi cantly mitigates hallucination and enables zero-shot knowledge transfer to novel data modality [7, 8, 9, 10, 11, 12].

Figure 1.2: (a) Learning visual prior from the internet by equipping large-scale pretrained generative models with a spatial and physical understanding of the world.

(a) Zero123 [1] (b) Dreamitate [2]

(c) Embodied self-learning is used to improve the perception and interaction skills learned from human data through self-supervised interaction and exploration [13, 14].

1.1 Learning the Visual Prior for Embodied Intelligence

The Internet is a gigantic pool of unstructured data, with a vast amount of knowledge and information hidden within. In recent years, the computer vision community has developed powerful generative models trained on images and videos collected from the internet. Despite their immediate successes in content creation applications, a fundamental research question is how to deploy them on embodied AI systems in the physical world.

We tackle this problem step-by-step. In [1], I introduced a method to control the camera perspective in large-scale image diffusion models, giving a pretrained 2D diffusion model spatial understanding of the world. To achieve this, we first created the largest 3D dataset to date [6], a dataset composed of 10 million 3D assets crowdsourced from the entire internet. We then take the largest 2D diffusion model that's pretrained on 5 billion 2D images and fine-tune it to generate a novel view of a scene when queried for the target camera viewpoint. With this fine-tuned model, we further developed an optimization method to extract the full 3D structure of a scene from just a single input image. Due to the internet's prior learned from both 2D and 3D data, this model demonstrated unprecedented generalization performance, with the ability to reconstruct 3D structures from highly challenging images such as abstract paintings (see Fig. 1.2a) and line drawings. To further bridge the gap between spatially aware generative models and reality, we extended the

learning framework proposed in [1] to handle occlusion [5], lighting condition [15], and dynamic scenes [3], assisting future embodied AI systems to navigate through in-the-wild environments.

Building on a strong foundation in spatial understanding, a key challenge in embodied AI is advancing physical interaction with the environment. I hypothesized that a foundational video generative model, trained on a large-scale internet dataset, inherently learns prior knowledge for physical interaction tasks. In [2], we proposed using a video generative model as a robot manipulation planner by ne-tuning a pretrained video model on human demonstration videos. During deployment, the model, conditioned on current scene observations, generates an action-plan video (e.g., scooping beans from one bowl to another as shown in Fig. 1.2b). A 3D tracking model then translates this video into robot joint angles for execution. Empirically, this approach effectively leverages the rich priors from pretraining, enabling generalization to novel objects and diverse visual contexts. To enhance control during planning, we introduced in [4] a mechanism for predicting object interaction videos conditioned on hand movements. These approaches represent a novel integration of foundational video generative models with robotic planning, leveraging learned priors from large-scale human videos to enable adaptive, real-world, in-the-wild manipulation.

1.2 Grounding Vision Generative Models with Physical Structures

Despite the powerful prior knowledge about the world learned from internet-scale visual data, the current generation of vision generative models is still faced with many limitations.
(1) many problems involve understanding data where large-scale datasets are impossible to obtain.
(2) generative models tend to hallucinate outputs that are not physically plausible.

To address the rst limitation, my research focuses on combining vision generative models with differentiable rendering, a class of methods modeling the physical transportation of light as a differentiable function. Inspired by analysis by synthesis, a cognitive approach that combines hypotheses about input with computational re-creation of the input, we proposed a learning framework where a vision generative model provides different hypotheses by generating data from a latent probabilistic distribution and differentiable rendering selects and re nes these hypotheses

Figure 1.3: (b) Physically grounding generative models mitigates hallucination and enables generalization to novel data modality.

(a) Humans as Light Bulbs [7] (b) Dr. Robot [9]

to recreate a visual observation. A unique advantage of this framework is the ability to transfer the prior knowledge learned by pretrained generative models to a novel data modality where large-scale data is impossible or difficult to collect. In [7], we applied this learning framework to address the 3D perception problem in thermal imaging, a data modality that's not largely available on the internet yet crucial to many industries and applications such as autonomous driving. Specifically, we proposed a novel differentiable ray tracing technique to refine the generated 3D scene sampled from a 3D generative model, as shown in Fig. 1.3a. As we solve the optimization problem to recreate the thermal images, we are able to find the true 3D structure of the scene. This framework is general and can be applied to other visual cues. In [8], we learn to reconstruct the full 3D object geometry by looking at the shadow it casts on the ground. Due to the optimization nature of the inference process, this framework is often slow and thus impractical. Therefore, we developed a method based on amortized machine learning to accelerate the optimization process [10].

Another major limitation of generative modeling is the problem of hallucination, making it hard to deploy in the physical world, such as when it is used by a robot. In [9], I proposed a robot self-model that is differentiable from its visual appearance to its control parameters, as shown in Fig. 1.3b. It achieves this by integrating the kinematic structure as well as the geometry of the robot body into the rendering process in a differentiable way. With these hard constraints, the robot model always moves in a physically grounded way, and with it, we can perform control and planning of robot actions through image gradients provided by visual foundation models. In addition to modeling robots, we also developed generative models with domain-specific inductive biases[12, 11] to constrain the generative process in order to generate data that are physically consistent with reality.

Figure 1.4: Built upon the foundation of physically grounded generative models, (c) Embodied self-learning is used to improve the perception and interaction skills learned from humans through self-supervised interaction and exploration

(a) AquaBot [14] (b) PaperBot [13]

1.3 Embodied Self-Learning

The current paradigm of AI research relies heavily on learning from humans, with successful examples including next-token prediction for LLM in NLP, contrastive learning for VLM in computer vision, and behavior cloning for manipulation in robotics. The fundamental limitation of this paradigm is that the skills and knowledge learned from human data are upper bound by human performance, which is often quite sub-optimal due to our biological constraints, such as the sensitivity of our sensors, muscle strength, reaction time, etc. On the other hand, intelligent robots don't have to be subject to human biological constraints. Instead of learning from humans, can we enable robots to learn by themselves through active interactions with the environment?

While self-learning in real world prevents the sim-to-real gap for robots, it presents major challenges: robots risk breaking objects, damaging themselves, or overheating during continuous operation. Faced with these obstacles, I explored an unconventional solution: self-learning underwater. Water dampens dangerous actions, adds buoyancy to reduce power demands, and provides cooling, enabling longer operational periods. This led to a paper [14] in which an underwater robot learns to improve a visuomotor policy for manipulation tasks autonomously (see Fig. 1.4a). In this work, we designed a self-learning underwater robot system using affordable commercial hardware. First, we trained a visuomotor policy based on human demonstrations to complete various underwater tasks, including object grasping, garbage sorting, and people rescuing. However, due to human limitations in motor control and reaction time—especially in dynamic underwater environments—this initial policy was sub-optimal and slow. To enhance it, we developed a self-learning

method that uses surrogate optimization to adjust motor speeds during deployment, accelerating manipulation tasks. This self-learning occurs entirely in the real world, enabled by underwater safety and cooling, and resulted in performance that exceeded a human operator's speed by 41%.

Robot self-learning can also be extended to the tool use and design problem, which is one of the hallmarks of human intelligence that differentiates us from most other animals. In our work [13], we apply robots to solve challenging paper-based tool design tasks, including (1) paper airplane design and (2) paper gripper design, both of which are traditionally extremely challenging to solve through simulation or physical analysis. Instead of relying on simulation, we created a robotic system to fold papers into airplanes with different shapes and cut papers into actuatable grippers using robot arms and other paper-related tools, as visualized in Fig. 1.4b. With an automated tool creation pipeline implemented with robots, we use the aforementioned surrogate optimization method to learn the design parameters of these tools, which eventually leads to tool design that outperforms the best human design. This superhuman performance is only possible thanks to the incredible precision and repeatability of robots and an ef cient learning algorithm.

1.4 Dissertation Overview

In this dissertation, we study an approach for achieving physical intelligence with generative modeling. We explore three different aspects of this approach:

Part I. Learning the Visual Prior [1, 2]

Part II. Physical Grounding of Generative Models [7, 9]

Part III. Embodied Self-Learning [13, 14].

Part I

Learning the Visual Prior

Chapter 2: Learning Geometric Prior for 3D Perception

The ability to perceive the 3D world is a fundamental aspect of physical intelligence. In this chapter, we hypothesize that a useful 3D visual prior can be learned by training a generative model on a massive amount of 2D visual data readily available on the internet. We present a system to demonstrate the powerful nature of this visual prior learned from internet-scale 2D datasets.

Figure 2.1: Given a single RGB image of an object, we present Zero-1-to-3, a method to synthesize an image from a speci ed camera viewpoint. Our approach synthesizes views that contain rich details consistent with the input view for large relative transformations. It also achieves strong zero-shot performance on objects with complex geometry and artistic styles.

2.1 Introduction

From just a single camera view, humans are often able to imagine an object's 3D shape and appearance. This ability is important for everyday tasks, such as object manipulation [16] and navigation in complex environments [17], but is also key for visual creativity, such as painting [18]. While this ability can be partially explained by reliance on geometric priors like symmetry, we seem to be able to generalize to much more challenging objects that break physical and geometric