SHURAN SONG Research Statement

From tossing objects into a trash can, to flinging a blanket to unfold it – humans frequently make use of various dynamic phenomena to manipulate things in the world. In these examples, dynamic manipulation (often with high-speed actions) allows us to leverage an object's momentum to greatly improve task efficiency and/or expand our physical reach range. In contrast, the majority of robot manipulators today are programmed to manipulate objects in a quasi-static manner – assuming the objects are *rigidly* grasped by the end-effector, moving *slowly* along a kinematic trajectory to avoid any unexpected dynamics. It's not that these robots can't move fast. In fact, many of our robots today already exhibit the hardware capabilities to move reliably with precision at high speeds. So what prevents robots from using dynamic actions for manipulation?

The key challenge remains in modeling the dynamics of the unstructured external world (i.e., the object that the robots interact with), where the diversity and complexity are significantly higher than those of the robot itself. As a result, classical methods that rely on an accurate and detailed physical models will fall short. In this context, our research goal can be characterized by three key aspects:

- **Leveraging dynamics:** Instead of trying to avoid dynamics (e.g., using slow actions), we want to enable robots to make use of the rich dynamic phenomena to achieve efficient manipulation.
- **In unstructured environments:** Instead of being limited to a few well-modeled instances, the robot should work with a wide variety of objects with unknown physical properties, including non-rigid objects.
- **From raw sensory input:** Instead of relying on accurate state information (e.g., from simulation or QR code), the whole manipulation process should operate directly from raw sensory input (e.g., images).

Each of the above goals poses unique challenges that amplify each other. For example, the outcome of any dynamic action is directly influenced by the objects' physical properties (e.g., friction, stiffness, or aerodynamics), which can no longer be ignored as in quasi-static cases. Meanwhile, many of these physical properties are notoriously hard to measure and model, especially under passive observation. Moreover, the visual input can be high-dimensional and noisy, providing only a partial observation of the objects and their properties, requiring the robot to jointly reason about its actions and the objects' motion to uncover the underlying dynamics.

Fig. 1: **Comparison.** Quasi-static manipulation (left) uses slow actions to avoid system dynamics, while dynamic manipulation (e.g., tossing; right) purposefully leverages the dynamics to extend the system's workspace and improve its efficiency.

My research aims to leverage self-supervised robot learning to tackle the aforementioned challenges and enable real-world applications. The key idea is to allow robots to learn about the dynamics of objects and their actions directly from data. When successful, such a framework will allow robots to automatically adapt and improve their manipulation skills with more data collected in new environments. However, a naive end-to-end learning approach would require a large amount of training data that is impractical to obtain, especially when considering all the combinations of objects, tasks, and robot embodiments. Therefore, when developing such a framework, it is critical to consider two fundamental questions:

- First, what to learn? In other words, what kind of inductive bias can we incorporate so that the algorithm does not need to learn "everything" while maintaining the flexibility to generalize to new scenarios?
- Then, how to learn it? For example, how to enable the system to self-collect data and create meaningful self-supervisory signals with minimal human intervention?

By tightly integrating perception and action through self-supervised learning, our approach relaxes the need for accurate dynamical models, and allows robots to i) learn dynamic skills from visual input, ii) improve the skills' precision using visual feedback, and iii) use their dynamic interactions to improve their understanding of the world. By changing the way we think about dynamics – from avoiding it to embracing it – we can simplify many classically challenging problems, leading to new robot capabilities.

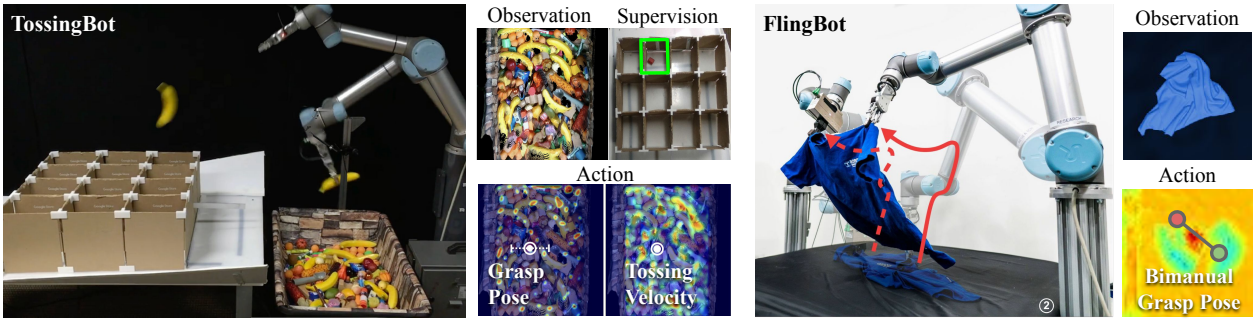

Fig. 2: **Learning Dynamic Skills from Pixels.** Our research allow robots to learn dynamic manipulation skills from visual inputs for a diverse set of objects. Left: TossingBot [1] learns to accurately throw different rigid objects into target bins **Best Paper in TR-O'20, Best System Paper in RSS'19**. Right: FlingBot [2] efficiently unfolds a piece of random fabric using highspeed fling actions. **Best System Paper in CoRL'21**.

1 Learning Dynamic Skills from Pixels

First introduced by Mason and Lynch in 1993 [3], dynamic manipulation has been extensively studied in the field of mechanics and control, where the focus is often to derive an analytical model for a specific object and then optimize the control parameters based on this model and state information (i.e., known object pose and geometry). However, accurately modeling the dynamical process can be challenging, and detailed state information is often impractical to obtain for unstructured objects. As a result, these analytical models have difficulty generalizing to slight perturbations or new objects.

Our research advances vision-based dynamic manipulation by allowing robots to learn new dynamic skills for diverse objects from their visual inputs [1, 2, 4]. By learning from data, the policy could compensate for the dynamics that are not explicitly accounted for in the analytical models. Fig. 2 shows two representative works: TossingBot [1] which learns to accurately throw different rigid objects into target bins, and FlingBot [2] which efficiently unfolds a piece of fabric using highspeed flinging actions. Leveraging dynamics, both systems drastically improve the efficiency and performance compared to their quasi-static counterparts, and extend the robots' effective workspace (e.g., unfold a piece of fabric longer than the robot arm).

The key challenge in applying machine learning in robotics is how to acquire the large amounts of training data necessary for a specific robotics task and hardware. To this end, our work uses a self-supervised framework with a spatial action map formulation that addresses this data issue from both ends:

- **Leveraging equivariance for sample-efficient learning.** In the spatial action map formulation, the policy infers the action parameters (e.g., grasp pose or swing velocity) *densely* for each pixel that is spatially anchored on the visual inputs. The policy network architecture is then designed to fully exploit the spatial equivariance between the input observation and output actions – applying translation/rotation(s) in the input should result in the same operation being applied to the output actions. This equivariant relation exists for many manipulation tasks [5], and by using an equivalent network, the policy can easily leverage this structural bias and improve learning efficiency.
- **Self-supervised learning for scaling-up data collection.** To automate the learning process, both system use self-supervised rewards automatically computed from the visual input (ie., objects' landing location for TossingBot, and cloth coverage area for FlingBot). When combined with an automatic reset mechanism, the system can continue its training for days with minimal human intervention. Through this automated learning process, the policy can self-adapt to new objects and situations on the fly and compensate for the dynamics not explicitly accounted for in the analytical models.

This self-supervised spatial action map framework (first proposed in our work [6]) has been used by many other research groups for a variety of applications that could benefit from sample-efficient real-world learning, such as pick and place [7], mobile manipulation [8], folding [9], sweeping [10], and has been extended to 3D visual representations (e.g., point cloud or voxels) with SE(3) equivariant networks [11].

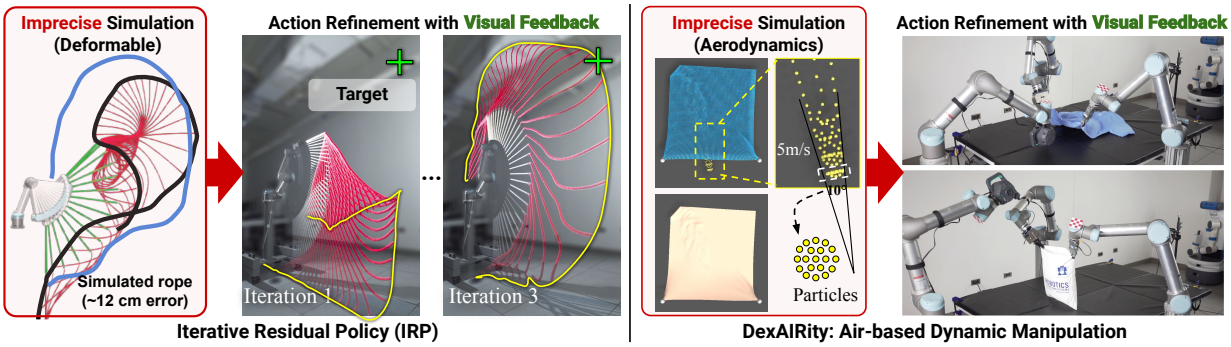

Fig. 3: **Precise Skills from Imprecise Models.** In both example, the system learns an imprecise model from an approximate simulator (for deformable objects or aerodynamics) and adapt it for precise real-world manipulation using closed-loop visual feedback. Left: Iterative Residule Policy [4] **Best Paper in RSS'22**. Right: DexAIRity [12] **Best System Paper Finalist in RSS'22**.

2 Precise Manipulation Skills from Imprecise Models

While showing promising results, learning actions directly from visual input is still challenging for tasks that have *a high precision requirement*. For example, while FlingBot can effectively unfold a cloth to increase its overall coverage, it cannot precisely swing the cloth to a target configuration. Naively extending the same learning framework will not work since the random policy will have a near-zero success rate for these high-precision tasks and therefore fail to provide useful training signal.

In fact, *precise dynamic manipulation* of deformable objects is extremely challenging even for humans. Imagine trying to hit a target with an unfamiliar rope – it is unlikely to succeed on the first try. However, as humans, we can build and use an intuition of physics (i.e., an approximate dynamical model) to correctly infer how to *adjust* our actions based on observations (e.g., swinging harder will make the rope reach higher). While our model is not perfect, we can adjust our actions in the right direction and quickly drive down errors. To impart this ability onto robots, we introduced Iterative Residual Policy (IRP) [4], a general formulation for goal-conditioned dynamic manipulation that highlights the following key features:

- **Learning delta dynamics from imprecise simulation.** Delta dynamics predict updated trajectories from an observed trajectory with small action perturbations. The hypothesis is that although the simulator is inaccurate in modeling the entire dynamical system, the "delta dynamics" (or at least its general direction) is a good approximation for many objects with different physical properties and, therefore, real-world scenarios as well.
- **Action adjustment with real-world visual feedback**: Instead of directly inferring the optimal action, IRP starts with an average action and iteratively refines it with visual feedback. This iterative approach makes it *robust* against noise in actions, observations, and model predictions, and achieves high precision.
- **Generalization.** We validate the IRP framework on both a rope-whipping task and a cloth placement task. Despite being trained only in simulation, IRP is able to efficiently generalize to noisy real-world dynamics, new objects with unseen physical properties, and even different robot hardware embodiments.

With Iterative Residual Policy, we hope to provide a **new perspective** on how to distill relevant knowledge from *inaccurate* simulators and achieve precise manipulation using *visual feedback*. This perspective is important for a wide range of systems that involve complex dynamics that are hard to model or simulate.

By removing the need for a precise model, this framework could open doors to **nontraditional robot hardware** that are often considered impractical to model. For example, in the project DextAIRity [12], we ask the robot to learn to manipulate deformable objects (e.g., bags and garments) by controlling the active airflow, Fig. 3. In this case, both the aerodynamic and deformable objects are extremely hard to simulate. To approximate these effects, we model the airflow with streams of invisible particles. This simulator is far from accurate. However, the algorithm can learn from this imperfect model and then adjust its action predictions

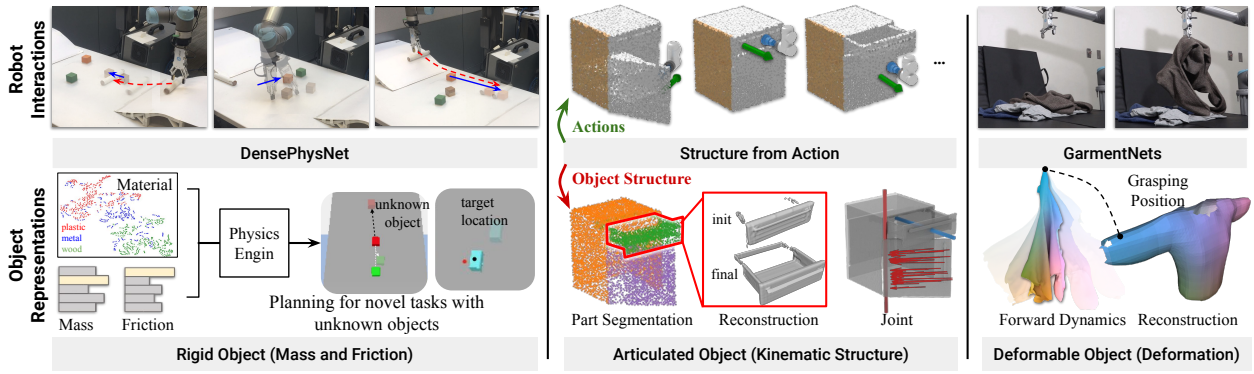

Fig. 4: **From Skills to World Model.** Our research studies how to leverage a sequence of strategic robot manipulation to acquire a rich object dynamic model that encodes their underlying physical properties [15], kinematic structure [16, 17], as well as occluded surfaces and possible deformations[18].

based on real-world visual feedback. During execution, the robot could observe the deformation of the cloth, which reveals useful properties of the *invisible* airflow and thereby informs action adjustments. By using airflow for manipulation, DextAIRity allows the system to apply dense forces on out-of-contact surfaces, expands the system's reach range, and provides safe high-speed interactions. These properties are particularly advantageous when manipulating under-actuated deformable objects. Moving forward, we aim to expand this framework to incorporate other sensing modalities beyond visual inputs (such as tactile, force [13] and sound [14]) in order to handle a wider range of objects and dynamics.

3 From Skills to World Models: Learning Object Models from Interactions

In the previous section, we show that having a dynamic model is beneficial, even though the model learned from the simulator is often limited by its imprecision. This begs the question: can a robot directly learn the dynamic model by observing the real world – i.e., learn a world model? It is an exciting direction with a few critical challenges: 1) To enable generalization, this world model needs to **uncover the underlying structure** of the object instead of memorizing all possible state and action pairs. 2) Many physical properties (e.g., friction, stiffness) can not be directly inferred from passive observations and require strategic physical interactions (e.g., to sense a rope's stiffness, the most effective way is to swing it).

We refer to this goal as *active scene understanding*. In contrast to the passive vision algorithms that are limited to "seeing what they are asked to see" (i.e., identifying pre-defined objects or parameters), this framework allows the agent to actively decide "what to see" and "how to see it" using its actions. Fig. 4 shows a few examples of our work in this direction, such as using robot interactions to decipher objects' physical properties [15], infer their kinematic structure [16, 17], and discover their hidden geometries [17–19]. Finally, the learned world model could be used to enable robot manipulation with these objects [15, 20].

Under this formulation, the role of robot manipulation is expanded – it becomes an important way to *"gather information and learn"* about the world, in addition to the traditional role of *"rearranging"* the world. Thus, a central focus of our research is to learn intelligent manipulation strategies for information gain. For example, in the process of discovering the kinematic structure of an unknown articulated object, observing the outcomes of simple actions (such as pushing) seldom provide new information about an object's articulation (e.g., pushing will just move the entire object rigidly). Hence, it is important for robots to learn strategic interactions to create informative object motions. For example, an informative action could expose initially occluded surfaces (e.g., opening a closed drawer), or trigger motions that indicate the segmentation of a new part (e.g., holding one part down, and pulling on another). In these frameworks [16–18, 20], we make minimal assumptions about the object, aiming at a single interaction and perception policy that generalizes to any object categories with unknown kinematics. Ultimately, this framework could change working processes for a number of domains where robots need to rapidly analyze their environments (e.g., new homes, unknown desert locations, or collapsed buildings) and swiftly react to evolving situations.

4 Ongoing and future directions: Skill Discovery with a Commonsense World Model

So far, we discussed how to learn the parameters for a pre-defined library of skills (e.g., grasp, toss, fling), and how to use those skills to improve the robots' internal model of dynamics (i.e., the world model). Moving forward, a research question I'm particularly excited about is how to automatically "expand" that library of skills using the common sense knowledge accumulated in the world model. To achieve this goal, we would need an efficient way to 1) scale up the knowledge stored in the model and 2) discover meaningful robot actions, in particular:

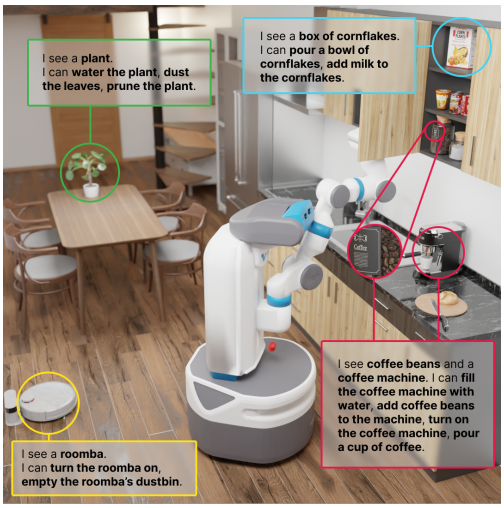

Fig. 5: **Commonsense-informed Exploration**

Language-informed Dynamic Models. With the active scene understanding framework in §3, robots can learn a detailed world model that describes the low-level dynamics of objects. However, this self-supervised process is limited to the objects that the robot has immediate physical access to. On the other hand, there is a large repository of commonsense knowledge about object dynamics provided by Internet data in the form of text and can be captured by the Large Language Models (LLMs) [21, 22]. This commonsense knowledge provides *complimentary* information about objects' properties that are not captured by the dynamic models learned through self-supervised interactions and can help inform the outcome of possible actions. For example, a language model could easily inform the robot that "If I toss an egg, the egg will break.", similarly "If I put an egg in hot water for 3 min, the egg will be soft-boiled."

This knowledge is particularly useful for informing dangerous or irreversible dynamics, where learning through trial and error is not desirable. However, to really use these language models on robots, we need to ground them in the *physical world*, which means connecting them to sensory inputs and low-level robot actions. Our recent work "Semantic Abstraction" [23] takes a step towards this goal by grounding open vocabulary visual language models onto 3D environments. As a result, the agents could now reason about their 3D environment with an open-vocabulary description and thereby unlocking the commonsense knowledge captured by the large language model. Moving forward, we hope this line of work could help bridge the currently "close-world" robotics tasks into "open-world" formulations that could generalize to new semantic labels, vocabulary, visual properties, and domains.

Commonsense-informed skill discovery. A common feature shared by many of our prior works is the assumption of pre-defined action primitives that map complex action trajectories to a few critical parameters. While these primitives helped in improving learning and planning efficiency, they can also be perceived as the biggest bottlenecks for such systems, since designing a comprehensive library of action primitives for every environment, task, and robot embodiment can be tedious and difficult to scale.

Our future work aims to alleviate this assumption by developing algorithms that enable robots to automatically discover a suitable library of action primitives that are 1) conditioned on both the embodiment and environment, meanwhile 2) interpretable and reusable so that these primitives can be used with a language-based planner to solve downstream tasks in a zero-shot manner. By studying how robots can autonomously acquire new dynamic skills in unstructured settings, we can better build decision-making systems that operate in a truly unstructured human-centered world – learning new dynamic skills that understand world dynamics and the relatively stochastic nature of it as a means to adapt to it. This form of exploration should also be safe so that the robot can continue to provide utility as it adapts over time. This idea goes beyond manipulation and applies to understanding human dynamics in collaborative settings. Doing so is key to getting robots to become an integral part of our everyday lives.

References

[1] Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Tossingbot: Learning to throw arbitrary objects with residual physics. In *Robotics: Science and Systems (RSS)*, 2019.

[2] Huy Ha and Shuran Song. Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding. *Conference on Robot Learning (CoRL)*, 2021.

[3] Matthew T Mason and Kevin M Lynch. Dynamic manipulation. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1993.

[4] Cheng Chi, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Iterative residual policy for goal-conditioned dynamic manipulation of deformable objects. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[5] Xiaolong Li, Yijia Weng, Li Yi, Leonidas Guibas, A Lynn Abbott, Shuran Song, and He Wang. Leveraging se(3) equivariance for self-supervised category-level object pose estimation. *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[6] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, 41(7):690–705, 2022.

[7] Xupeng Zhu, Dian Wang, Ondrej Biza, Guanang Su, Robin Walters, and Robert Platt. Sample efficient grasp learning using equivariant models. *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[8] Jimmy Wu, Xingyuan Sun, Andy Zeng, Shuran Song, Johnny Lee, Szymon Rusinkiewicz, and Thomas Funkhouser. Spatial action maps for mobile manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020. doi: 10.15607/RSS.2020.XVI.035.

[9] Robert Lee, Daniel Ward, Akansel Cosgun, Vibhavari Dasagi, Peter Corke, and Jurgen Leitner. Learning arbitrary-goal fabric folding with one hour of real robot experience. *Conference on Robot Learning (CoRL)*, 2020.

[10] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. *Conference on Robot Learning*, 2020.

[11] Shubham Agrawal, Yulong Li, Jen-Shuo Liu, Steven K Feiner, and Shuran Song. Scene editing as teleoperation: A case study in 6dof kit assembly. *Intelligent Robots and Systems(IROS)*, 2022.

[12] Zhenjia Xu, Cheng Chi, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. DextAIRity: Deformable Manipulation Can be a Breeze. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[13] Jingxi Xu, Shuran Song, and Matei Ciocarlie. Tandem: Learning joint exploration and decision making with tactile sensors. *IEEE Robotics and Automation Letters*, 2022.

[14] Alexis Burns, Siyuan Xiang, Daewon Lee, Larry Jackel, Shuran Song, and Volkan Isler. Look and listen: A multi-sensory pouring network and dataset for granular media from human demonstrations. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2519–2524. IEEE, 2022.

[15] Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B Tenenbaum, and Shuran Song. DensePhysNet: Learning dense physical object representations via multi-step dynamic interactions. In *Robotics: Science and Systems (RSS)*, 2019.

[16] Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Act the part: Learning interaction strategies for articulated object part discovery. *ICCV*, 2021.

[17] Neil Nie, Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Structure from Action: Learning Interactions for Articulated Object 3D Structure Discovery. *arXiv preprint*, 2022.

[18] Cheng Chi and Shuran Song. GarmentNets: Category-Level Pose Estimation for Garments via Canonical Space Shape Completion. *ICCV*, 2021.

[19] Zhenjia Xu, Zhanpeng He, Jiajun Wu, and Shuran Song. Learning 3d dynamic scene representations for robot manipulation. In *Conference on Robotic Learning (CoRL)*, 2020.

[20] Zhenjia Xu, He Zhanpeng, and Shuran Song. Umpnet: Universal manipulation policy network for articulated objects. *IEEE Robotics and Automation Letters*, 2022.

[21] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[23] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3D scene understanding from 2D vision-language models. In *Conference on Robot Learning*, 2022.