Position: Scaling Simulation is Neither Necessary Nor Sufficient for In-The-Wild Robot Manipulation

Homanga Bharadhwaj [1]

Abstract

In this paper, we develop a structured critique of robotic simulations for real-world manipulation, by arguing that scaling simulators is neither necessary nor sufficient for making progress in general-purpose real-world robotic manipulation agents that are compliant with human preferences. With the ubiquity of robotic simulators, and recent efforts to scale them for diverse tasks, and at the same time the interest in generally capable real-world manipulation systems, we believe it is important to address the limitations of using simulation for real-world manipulation, so that as a community, we can focus our collective resources, energy, and time on approaches that have more principled odds of success. We further demonstrate the unique challenges that real-world manipulation presents, and show through examples and arguments why scaling simulation doesn't get us closer to solving these challenges required for diverse real-world deployment.

"The world is its own best model."

- Rodney Brooks

1. Introduction

Imagine a world where robots are everywhere, interacting with humans in the most basic everyday tasks like cooking and cleaning, while remaining compliant with human preferences. In such a world, we would expect robot manipulators to exhibit out-of-the-box *zero-shot* generalization capabilities. When prompted to do a task, the robot should *just do it!*. It shouldn't require any form of self-practice or fine-tuning for the specific task it's asked to perform. In addition, it should perform the task while respecting human preferences and without creating any unsafe intermediate scenarios.

Recent advances in machine learning and robot hardware design has helped move the world described above from a Utopian existence a couple of decades ago, to a practical possibility in the near future. As very prominently witnessed in computer vision and natural language processing (NLP), a number of advances including but not limited to availability of large datasets, ubiquity of compute resources, and improvements in deep learning based algorithms has positively impacted all research areas where machine learning is widely used. The story of robot learning has been no different. However, unlike computer vision and NLP, the success stories of robot learning approaches, in particular for manipulation have been significantly underwhelming. Approaches that *just work zero-shot* in diverse unseen scenarios are far from being realized in manipulation, compared to systems like Segment-Anything (Kirillov et al., 2023) and GPT-4 (Achiam et al., 2023) in vision and NLP.

Anecdotally, as observed by Jitendra Malik in a talk, until recently, experiments in robot manipulation papers were in such restrictive scenarios that just by looking at the background color of video results, one could form a reasonable guess of where the paper is from - for example, green background for Berkeley and red for CMU (Levine et al., 2016; Shaw et al., 2023; Maitin-Shepard et al., 2010). The lack of generalization in robotics results should not come as a surprise because most robotics tasks involve reasoning over time horizons much longer than that of typical vision tasks, and the state-space is much higher-dimensional than that in NLP tasks. If this weren't enough of a challenge, the real world is also very dynamic, and in several practical manipulation tasks, the environment might change *during* execution of the task itself.

We consider two desiderata of widely usable real-world manipulation systems: 1) zero-shot deployability and 2) compliance with human preferences. Machine learning has the potential to transform robot manipulation by helping learn policies that can generalize across diverse real-world scenarios, and can adapt to human preferences. Hence the question of *how* should machine learning be used for manipulation

[1]The Robotics Institute, School of Computer Science, Carnegie Mellon University. Correspondence to: Homanga Bharadhwaj <hbharadh@cs.cmu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

- including the type of datasets and learning mechanisms, is very pertinent. Understanding the limitations of existing approaches will help us make progress towards solving the hard problem of manipulation, and in democratizing robotics more broadly for human assistance.

In order to simplify the complexity of real-world robot manipulation, several papers have sought to learn manipulation policies in *simulated* environments and then transfer them to the real-world typically through approaches that bridge the simulation-to-reality (sim2real) gap (Akkaya et al., 2019; Sadeghi & Levine, 2016; Yan et al., 2017). This approach of sim2real transfer has become a leading paradigm in training manipulation policies. While simulation in and of itself may be helpful for understanding sequential-decision making, in the context of real-robot manipulation that generalizes to diverse scenarios and is compliant with human preferences, this paper takes the position that simulation as a tool is neither necessary nor sufficient.

Our position is that scaling simulation, in the form of designing more realistic simulated robots and environments, improving the speed and parallelism of simulation environments, and attempting to develop more realistic simulators are not crucial in the path to developing widely usable real-robot manipulation systems. We argue that over-indexing on simulations for real-robot manipulation can be detrimental to the eventual goal of generalizable manipulation in the real-world and lead to misleading notions of progress.

Given that robotics, particularly robot manipulation holds immense potential to significantly impact humanity and stands to benefit greatly from advances in machine learning, it is essential for the community to reconsider where and how we prioritize our efforts. In this paper we demonstrate why scaling simulation is probably not the right approach, and provide an alternate vision for how we can develop generalizable and compliant robot manipulation systems that can be beneficial in diverse scenarios.

2. Desiderata for Widely Usable Real-World Manipulation

In this section, we describe two key criteria for robot manipulation systems that are widely usable in diverse scenarios. Our focus is on manipulation systems that are aimed to be deployed *in-the-wild* in scenarios like homes, offices, and kitchens for helping us in everyday tasks. We do not consider structured manipulators that may be useful only for specific industrial automation tasks.

Zero-Shot Deployment. For a manipulation system to be widely adopted in homes, offices and other generic environments, we posit that it needs to work out-of-the-box. When a user specifies a task, the system should just execute the task, without requiring any exploration or fine-tuning through interaction for this task. This is important for efficiency in automation, so that a manipulator can be repeatedly used for different tasks, without having to wait more than the time it would take a human to complete the task. For truly reliable zero-shot deployment, the manipulator must be able to generalize to diverse scenarios that are previously unseen. For example, a home robot built to assist with household tasks such as cooking and cleaning should generalize to arbitrary kitchens and homes, without requiring any fine-tuning through interactions in the test environment.

Zero-Shot deployment capabilities are a key requirement for making robot manipulators ubiquitous in everyday life for helping humans in different tasks. If we draw analogies with modern appliances like computers, vacuum cleaners, televisions etc., they all work out of the box, sometimes quite literally. It is only reasonable to necessitate that robots that are as ubiquitous as these current appliances should have similar properties. Note that zero-shot deployment does not preclude fine-tuning that doesn't involve unsafe/exploratory interactions with the environment. For example, we customize our televisions and computers to suit our specific needs. Similarly, through mechanisms like human preference elicitation, we could fine-tune the robot manipulator to suit our specific needs. An example of this in cooking could be, we would want the robot to always wipe the counter top after handling meat, before proceeding with other tasks. Zero-shot deployment simply means that the robot can perform all the tasks it might ever be asked to perform, in a reasonably safe way, but not necessarily in the way an end-user wants. For the latter, it might require some minor fine-tuning to become compliant with human preferences.

Examples of recent real-world systems that perform diverse manipulation tasks but do not really fit the zero-shot deployable criteria are Bahl et al. (2022) (requires online exploration for 1-2 hours given a new task) and Mahi Shafiullah et al. (2023) (requires collecting demonstrations for a new task to train a policy that works only for that specific task in that specific scene). Some recent systems that are closer in principle to this criteria are (Zitkovich et al., 2023; Brohan et al., 2022; Padalkar et al., 2023b; Bharadhwaj et al., 2023c; Shridhar et al., 2023; Bharadhwaj et al., 2024) (given a new task specified through a goal, the respective policies can directly execute the task in one go without any fine-tuning).

Compliance with Human Preferences. In addition to being zero-shot deployable, robot manipulators should be safe and compliant with human preferences. Broadly, compliance refers to comformity with a rule, specification, or policy. In machine learning, it is an implied understanding that algorithms should align with human preferences. This alignment is assumed to be a key goal. We anticipate that the algorithm, once trained, will act in accordance with the

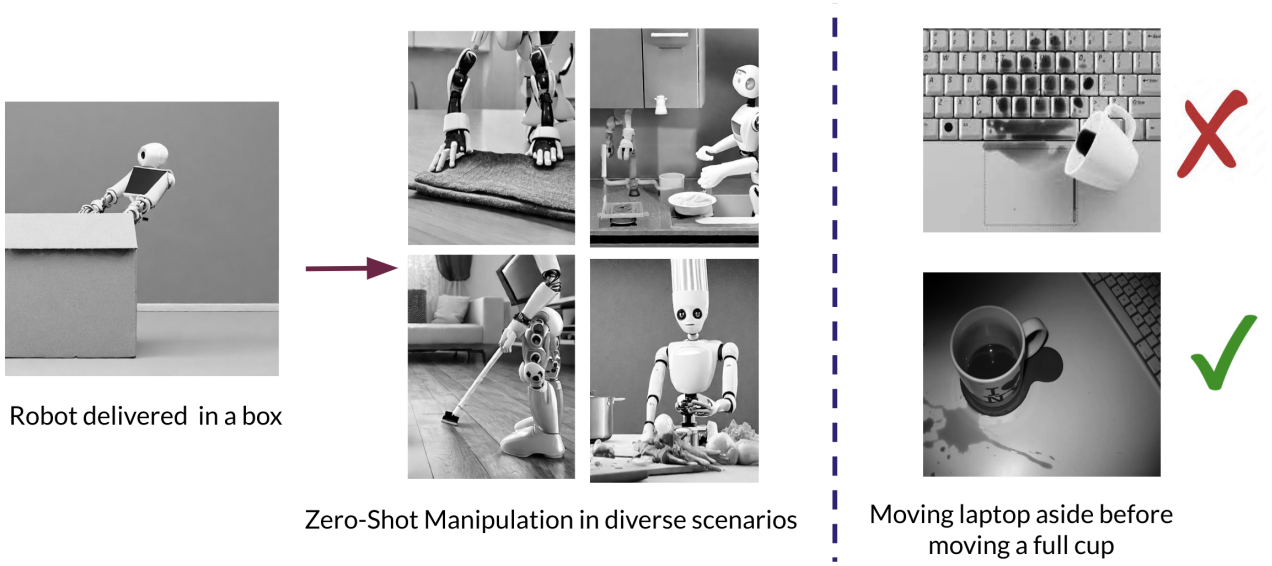

Figure 1. Illustration of the desiderata for widely usable real-world manipulation: zero-shot deployment [left] and compliance with human preferences [right] as discussed in section 2.

objective function set during its training. This objective function is assumed to represent the preferences of the person who designed the algorithm, serving as a stand-in for their desires. There are two key challenges: firstly, optimizing algorithms based on a given objective function, and secondly, the initial creation of that objective function.

A large part of the the machine learning community is engaged in the first challenge, striving to enhance optimization by developing more effective learning algorithms or creating superior neural network architectures. Generally, this is not problematic since, for numerous supervised learning tasks (like classification using cross-entropy loss) or unsupervised learning tasks (such as image generation with pixel-reconstruction error), the predefined objectives tend to effectively meet the intended outcomes. For instance, in sequential decision-making scenarios like game-based reinforcement learning, the game's rules often inherently dictate the reward function, eliminating the need for manual construction. However, the situation becomes more complex and less straightforward when applying these principles to intricate, real-world control problems involving robots, where the formulation of an appropriate reward function is less obvious.

This is difficult to achieve because oftentimes humans themselves cannot quantify their own choices precisely, and so constructing a *reward function* that can optimize for this is tricky. However, given two or more options, humans are good at choosing the the option they like. In other words, they are good at *relative* preference elicitation. For robotic systems that must interact with humans during deployment, it seems plausible that they should also be exposed to human interaction during training. Indeed recent works have attempted this (Hejna III & Sadigh, 2023; Sadigh et al., 2017) for developing manipulators that are more aligned with human preferences. Compared to LLMs, where RLHF (Reinforcement Learning from Human Feedback) is easy to achieve, for robotics this is non-trivial due to safety considerations in online improvement.

3. Preliminaries on Simulation for Manipulation

In this paper, by *simulation*, we mean physics simulations for robotic manipulation. Some examples of popular simulators for robot manipulation include MuJoCo (Todorov et al., 2012; Brockman et al., 2016), Isaac (Makoviychuk et al., 2021), Brax (Freeman et al., 2021), PyBullet (Coumans & Bai, 2016–2021) etc. These simulators serve as platforms for replicating real-world dynamics by making several approximations to the physical properties of objects and environments. The key idea across simulators is to *simulate* the forward dynamics of the world i.e. provide answer to the question *given a current state of the world and some force, what is the next state of the world?* The dynamics models are algorithms that compute the motion of components in the scene including robots, based on forces, torques, and interactions with the environment. They ensure that simulated movements closely replicate real-world physics, taking into account factors like friction, inertia, and material properties

of objects.

Most current robotic simulators are extremely inaccurate in properly simulating realistic physics (Afzal et al., 2020; Lidec et al., 2023; Yoon et al., 2023). This is especially true for robotic manipulation where contacts are high dimensional, the complexity of motions are much higher, and the range of motions and types of objects are much diverse, compared to typical locomotion or navigation scenarios. So, unlike locomotion and navigation (Kumar et al., 2021; Szot et al., 2021; Lee et al., 2020; Tan et al., 2018), simulation for manipulation has had much less success in enabling real-world generalization. Beyond considerations like modeling contacts and collisions, to be truly effective, simulation frameworks need to overcome other challenges like offering real-time feedback and high-fidelity visuals, which have also been difficult to achieve in practice.

Recognizing that any simulation of the real-world is likely to be imperfect, recent approaches have attempted to mitigate the sim to real gap through several techniques like domain adaption, domain randomization, reality check-pointing, and noise-modeling. Here, we briefly describe some of these techniques, which is in no way meant to be an exhaustive list.

Domain Randomization: This technique involves randomizing aspects of the simulation during the training phase. Parameters like lighting, object textures, and physical properties like friction, mass, damping, inertia etc. are varied. This creates a wide variety of training scenarios, encouraging the model to learn features and behaviors that are robust to the variations found in the real world.

Domain Adaptation: This approach focuses on adapting a model trained in a simulated environment to perform well in the real world. Techniques like fine-tuning the model on a small set of real-world data or using adversarial training to align the feature distributions of the simulated and real-world data are common.

Reality Checkpoints: Incorporating intermittent real-world interactions during the training phase can help in mitigating the domain gap. The model is primarily trained in simulation but is periodically updated or corrected based on real-world data or feedback through rollouts in the real world. This helps in gradually reducing the reality gap.

Sensor Calibration and Noise Modeling: Accurately modeling sensor noise and calibrating sensor readings in the simulation to match those in the real world can significantly improve the transferability of the learned behaviors. This is also called system identification and can be performed adaptively, similar to reality checkpoints.

Each of these techniques can be used independently or in combination, depending on the specific requirements of the task and the nature of the reality gap in a given application. The choice of technique is also influenced by factors such as the availability of real-world data, the computational resources, and the specific characteristics of the task and the morphology of the robots.

4. Scaling Simulation is not Sufficient

In order to achieve widely usable real-world manipulation systems, specifically those that satisfy the desiderata in section 2, we take the position that scaling robotic simulations is not sufficient. We believe it is important to address this because a large volume of recent works in scaling robot learning for manipulation are extensively trying to scale up simulation. In this section, we consider robot manipulation policies that are trained entirely in simulated environments i.e. without the use of any data or interactions with physical robots. Several recent works adopt this philosophy of training entirely in simulation and deploying either in purely simulated environments, or on a physical robot with limited real-world environment variations (Handa et al., 2023; Akkaya et al., 2019).

4.1. Even the best simulators cannot match reality

For a simulation system to be useful as a proxy for real-world manipulation, the physics of object-object interactions and object-robot interactions needs to be accurate across a ramge of real world scenarios. While it is now possible to create a near accurate physical simulation of the robot arm in isolation, it is difficult to manually create accurate simulations of contacts, and motions of most common objects, let alone all possible objects. A part of this difficulty is obtaining 3D assets of common objects in the real-world, which is necessary for creating simulation environments. Even large scale datasets of assets (Downs et al., 2022; Szot et al., 2021; Savva et al., 2019) only contain a few categories of objects, and are not comparable to real-world diversity. In addition to assets of objects, simulating all possible physical forces is tricky for several types of objects, and recreating the visual appearance of real scenes with diversity similar to that of real-world distributions is further non-trivial.

Simple everyday tasks like pouring coffee from a cup, washing dishes in the sink, chopping vegetables on a table, folding clothes etc. are extremely difficult to reliably simulate owing to high complexity of the respective state-spaces. Scaling simulation frameworks is unlikely to directly help with these tasks as each of these would require separate nuanced considerations for faithful simulation. Indeed, creating a very accurate simulation of reality is a painstaking process, and one that is unlikely to be possible with an accuracy high enough for direct real-world deployment. Even

if some aspects of reality like visual realism of scenes is achieved with high fidelity, simulating accurate physics and modalities beyond vision like tactile sensing will still remain non-trivial at scale. Since direct transfer from simulation to diverse real world scenarios is probably unlikely, hence simulation alone is not sufficient for training real-world manipulation policies.

4.2. Sim2Real requires solving a more difficult problem to solve a simpler problem

Consider the task of pouring water from a jug to a cup, shown in Fig. 2. Solving this task probably requires an approximate reasoning of how water flows, its viscosity, and based on that an estimate of how much to tilt the jug for water to flow slow enough without creating a big slash, and fast enough without waiting the entire day for it to gradually drip. This task probably doesn't require reasoning about how each *'particle'* of water is likely to interact under the effect of all possible forces with its neighboring particles, with gravity, the surface of the jug, and the base of the cup. Indeed, simulating the process of water being poured from a jug is arguably a more difficult problem than designing a policy to pour water from the jug.

This example of pouring in terms of difficulty in simulation is not a contrived edge case, but is abundant in everyday tasks (Fig. 2). In order to know *how* to do something, we do not need to know the 'what' of everything involved in the process. Even as humans, we do several everyday tasks without knowing how exactly the world will evolve is response to our actions. We chop onions without knowing how many pieces will be formed on each strike, and where on the chopping board they will fall, we throw trash into the bin without knowing where in the bin it will land, we pour ingredients into the soup without knowing the viscosity at each step of cooking - the list can be endless. In fact, for *most* everyday tasks, we do not have an accurate forward simulation of each intermediate step. Through experience we have developed models of what to focus on and what not to (or what cannot be focused on with the limited attention span we have). However, when designing physics based simulations, there is no way to factor in this likelihood of "salient" events worth focusing on. One can only hope to simulate *everything* accurately, thereby solving a much harder task than what is needed for manipulation.

4.3. Simulations provide a false notion of progress in manipulation

The flurry of easily usable computer simulations for manipulation has had an unintended consequence of researchers over-indexing on simulation experiments for drawing conclusions about manipulation in general. It is our position that simulation experiments, no matter how rigorous and diverse, are simply *not* sufficient for justifying anything substantial about real-world manipulation. Even if a task-specific policy for a task like "opening a drawer" works with 90% success rate across 100 different drawers in simulation, when this policy is deployed on an arbitrary real-world drawer with the same robot, the odds of success are not necessarily going to be close to 90%, simply because it is not possible to simulate all types of drawer that are likely to naturally exist in the world. When we extend this argument to a generalist agent capable of multi-task behaviors, which this paper focuses on, the conclusions from simulation become even weaker in the real world.

So, do simulation experiments not provide any useful indication of progress? Of course they do! Simulation studies are a great way for prototyping sequential decision making algorithms, for example those based on reinforcement learning (RL). If a particular RL algorithm succeeds in simulation, for example in Atari games, the takeaway is precisely that this algorithm is good for Atari games! If it also succeeds in simulated articulation manipulation tasks, then the takeaway is that it also succeeds in simulated articulated manipulation tasks! If it succeeds in a wide diversity of simulated control tasks ranging from video games to locomotion to manipulation, then the takeaway is that this algorithm is good for a wide diversity of simulated control tasks! Several recent *generalist* agents have been developed in simulation that fit this criteria (Hafner et al., 2019; 2023; Hansen et al., 2022) However, concluding that a policy trained with this algorithm is simulation will also be helpful directly for real-world control tasks is over-statement!

Some recent papers have attempted to demonstrate this point quantitatively through experiments. In (Dasari et al., 2023), the authors find that pre-trained visual representations that enable high success rates in simulated robot manipulation tasks, do not perform well in the real world. More generally, there is very less correlation between performance in simulation and in real tasks, when agents are deployed with the same pre-trained visual representation backbone. Similar observations about the connection between scaling robot policies in simulation and their applicability in the real world not being straightforward are made in anotehr recent work (Pumacay et al., 2024) In summary, while simulation environments may be useful for their own sake, for real-world manipulation they are not likely to be sufficient even at scale.

5. Scaling Simulation is not Necessary

In this section we argue that scaling simulation frameworks is not necessary for reliable real-world manipulation. We establish this by demonstrating that whatever benefits can be accrued from simulation for real-world manipulation can be achieved through alternate means in much easily scalable

Figure 2. Illustration of everyday tasks where simulating the task is a much difficult problem than actually executing the task. Tasks such as these are ubiquitous in everyday life, and only require a coarse reasoning of the forward dynamics of the scene for successful execution.

ways.

5.1. Priors for Interaction

A line of simulation based robotics research uses policy learning techniques in simulated environments to pre-train a manipulation policy that is fine-tuned in the real-world. This fine-tuning is done either through real-world reinforcement learning (Bahl et al., 2022) or through imitation learning with a few demonstrations (Ma et al., 2022). This pre-training followed by fine-tuning is akin to a type of *domain adaptation* described in section 3. For pre-training in simulation, several techniques like domain randomization, combined with reinforcement learning can be used.

This recipe of incorporating priors for real-world manipulation, although plausible for narrow deployment scenarios, is unlikely to scale to diverse real-world tasks in diverse scenes. This is due to a number of issues ranging from the visual artifacts of scenes in simulation differing significantly from what real-world scenes look like (large visual domain gap), the challenges of designing reward functions for each task separately (for having reinforcement learning as the pre-training objective in simulation), the difficulty of collecting expert demonstrations (for imitation learning as the pre-training objective in simulation). Although some of these issues have been made more feasible by large models, for example reward design with LLMs (Chen et al., 2023a; Qi et al., 2023), and semantic augmentations for visual diversity (Yu et al., 2023; Bharadhwaj et al., 2023c; Mandi et al., 2022; Chen et al., 2023b), the sim2real gap for a diverse set of tasks is still likely to be significant to the extent that just a little fine-tuning in the real world is unlikely to be sufficient. Indeed, results from recent papers adopting this recipe have only been able to show successes in narrow domains (Qi et al., 2023; Chen et al., 2023a).

We argue that simulation is in fact not necessary for pre-training, and a much simpler, more scalable route exists with higher likelihood of reducing domain gap both in terms of visual representations and action trajectories. This approach makes use of pre-training with diverse passive videos on the web. Videos on the web, of humans doing everyday tasks are abundant, and are freely available. There are already several successes of this in recent robot manipulation works, both for visual pre-training and action pre-training (Nair et al., 2022; Ma et al., 2022; Bharadhwaj et al., 2023b; Xiao et al., 2022; Bahl et al., 2023). Scaling approaches to leverage these passive datasets is a much more tractable direction to pursue as it doesn't require the significant engineering efforts of designing realistic simulators.

Of course, leveraging passive video datasets for robotics is not trivial, due to the huge domain gap, but being abundant, diverse, and freely available is a major benefit compared to simulation. In addition, recent advances in computer vision techniques like 3D human pose estimation (Mehta et al., 2017; Kocabas et al., 2023; Ye et al., 2023a), hand-object pose estimation from RGB videos (Rong et al., 2020; Ye et al., 2022; Yang et al., 2022; Pavlakos et al., 2023), affordance prediction (Ye et al., 2023b; Goyal et al., 2022), and tracking points and objects in videos (Karaev et al., 2023; Doersch et al., 2022; Yang et al., 2023a) have made it possible to leverage useful priors for manipulation from

these passive datasets. Several of these approaches can be conveniently scaled with data and compute, are thus not tied to tedious engineering efforts involved in creating accurate physical simulations.

5.2. Safety Considerations

A generally important use case of simulations across research fields is to perform tests of algorithms that will be eventually deployed in real-world safety critical scenarios. Since failures in simulations are not fatal, trial and error based learning is especially amenable for simulation environments. Testing in simulation is a critical step for safety-critical systems where direct real-world experimentation can be hazardous, costly, or impractical. Such simulation-based testing has been particularly successful for closed systems like nuclear power plants, and electric grids, where safety is not conditioned on interaction with an ever changing and hard to predict environment. Beyond widely deployed systems, in robotics research, simulation based development has been helpful for navigation and locomotion (Kumar et al., 2021; Hwangbo et al., 2019; Yang et al., 2023b; Truong et al., 2023) where the degrees of freedom of the system are low and the factors of variation in the environment are enumerable. For locomotion with quadrupeds and bipeds, a recent wave of domain randomization based RL algorithms that change the structure, terrain, and friction of the ground in simulation has been successful for sim2real transfer (Kumar et al., 2022; Agarwal et al., 2023).

However, for manipulation, the notions of safety are much more subtle, task-dependent, and in many cases directly dependent on the human alongside whom the manipulator is deployed (Thananjeyan et al., 2021; Bharadhwaj et al., 2020; Thumm & Althoff, 2022). In many scenarios, a safe behavior is related to notions of compliance described in section 2. When a robot is moving a full cup of coffee from one end of the table to the other, we would not want it spill any on the laptop that is next to cup. A compliant robot might do something similar to what a human would in this scenario - simply move the laptop aside first before grasping the cup! Although simple to describe, such behaviors are difficult to simulate especially autonomously. Hence, simulations are not necessary for developing safe real-world manipulators compliant with human preferences.

A much simpler approach to safe real-world manipulation is to have certain basic constraints directly in the real-world and perform preference based optimization of compliance. Examples of simple task-agnostic constraints include having smooth trajectories without jerky motions, collision detection and avoidance with on-arm sensors, and a well-defined workspace around the robot beyond which any part of the arm doesn't venture. In addition to these physical constraints on behavior, the priors on the learned policy, especially action priors also help in constraining the manipulation behaviors to a constrained super-set of compliant behaviors. Further, fine-tuning with real-robot trajectories, and with human-in-the-loop feedback are easily usable techniques for further constraining the robot behaviors to lie within a subset of compliant trajectories for different use cases.

5.3. Sub-optimal Data

An argument for scaling simulation is that it allows for generation of significantly high quantity of data at minimal cost (Mandlekar et al., 2023; Dalal et al., 2023). Typical approaches for automating such simulated robot data collection rely on scripted policies rolled out with scene variations, and pre-trained agents (for example agents trained with task-specific reward functions via RL). However, prior works relying on such automatic data generation strategies are typically bottle-necked by the diversity of generations, since it is intractable to automatically define reward functions for different manipulation tasks in diverse scenarios (Ma et al., 2023; Mandi et al., 2022). As such the behaviors are restricted to table-top manipulations in structured scenes (Shridhar et al., 2023; Mandi et al., 2022; Mandlekar et al., 2023). Hence, being able to collect an abundance of data through simulation is unhelpful if they are not representative of the real world diversity of tasks and scenes.

For generalizable real-world manipulation, we believe that recent approaches in democratizing tool use will play a key role in enabling diverse real world data collection in a scalable manner (Chi et al., 2024; Wang et al., 2024). Easily usable tools like these that humans can wear and perform everyday tasks will likely have a huge impact in collecting real world datasets for robot manipulation. In addition to these tools, widely adoptable teleoperation systems, like ALOHA (Fu et al., 2024) are going to further enable collecting expert datasets of high quality that can be used for imitation learning. Such real-world data collection systems can also be equipped with sensing modalities beyond vision like tactile sensing with easily available tactile sensors (Yuan et al., 2017; Bhirangi et al., 2021) for multi-modal real-world learning instead of relying on imperfect tactile simulators.

In order to learn from such diverse real-world datasets (of different robot embodiments in different scenarios), a popular approach is to define an action space that is embodiment-agnostic, and train policies across diverse datasets through a single model. A recent example of this is the Open-X system (Padalkar et al., 2023a). Further, going beyond robot datasets, some recent papers (Bharadhwaj et al., 2023a; Bahl et al., 2023) have also used human video datasets (web data) in conjunction with robot data to learn unified models, by

abstracting out relevant details from the human datasets (like hand and object poses). Overall, we believe such scalable approaches will be able to address the issue of data paucity for manipulation in a much more useful manner, compared to scaling simulation.

6. Discussion

In this paper we argued as to why scaling simulation is neither necessary nor sufficient for generalizable real-world manipulation that is compliant with human preferences. However something can be neither necessary nor sufficient, and still be "useful." So, is simulation useful in this context? Maybe! But significantly only in contrived and highly structured scenarios. The claims in this paper do not necessarily apply to structured settings like industrial automation tasks, where the task definitions, the environment, and all aspects of the scene are largely pre-determined and unchanging. Instead, we focused on manipulation systems that are aimed to be deployed *in-the-wild* in scenarios like homes, offices, and kitchens for helping us in diverse everyday tasks. A central theme of this position paper has been that the potential benefits of simulation are outweighed by the complexity of designing and scaling simulation frameworks in a manner that is aligned with the goals of diverse real-robot manipulation. Properly scaling simulation is not trivial, requires significant engineering efforts, and in the end is still likely to fall short of accurately modeling real-world physics.

In recent years, owing to the fact that all simulators are imperfect, approaches for sim2real transfer have become increasingly popular. The most popular among them is domain randomization, since it requires simulation-specific training and can in-principle be directly deployed in the real world. A common narrative around domain randomization is that if the simulation parameters are randomized "enough" during training, say for n different combinations where n is very large, then the trained policy should generalize to the real world, since reality is just the $n + 1^{\text{th}}$ simulation.

This couldn't further be from what happens because an imperfect simulation not just doesn't have the parameter values of the real world right, it doesn't have the parameters right to begin with! Especially for robotic manipulation, where the tasks are contact-rich, involve diverse objects including deformable objects, and a lot of visual and structural variations in the scene, estimating the right parameters to quantify factors of variation through simulation is hard, almost intractable. So, no amount of domain randomization is likely to be enough to capture the diversity of real-world variations.

Perhaps, more philosophically, over-reliance on simulation based benchmarks has distracted the robot learning community in recent years, to the point where there are significant differences in the types of tasks targeted by robotics researchers who work on non-learning methods, and robot learning researchers who demonstrate results on simple simulation benchmarks. Thankfully this trend is starting to change but still a lot of papers claim to improve robotics but demonstrate results only in simulation. This is harmful for the the machine learning community broadly as it doesn't bring us any closer to useful autonomous robots in the real world while providing us a misleading sense of progress in this direction. Only by deploying robots in the real world can we quantify the limitations of current machine learning algorithms for robotics, and develop solutions to improve them.

While this paper specifically targeted robotic manipulation, several of the arguments also apply to locomotion and navigation research. It is true that a lot of recent progress in developing autonomous quadrupeds that can walk in diverse terrains, and home robots that can perform goal directed navigation in generic homes has been fueled largely by advances in simulation-based training, the scope of real-world deployment is still far from what we should be comfortable with as end-users of these systems.

7. Conclusion

Machine Learning for robot manipulation poses unique challenges owing to high-dimensional state-action spaces and complex long-horizon decision making. In order to build robot manipulation systems that are useful in everyday life, we need to mitigate these challenges by developing scalable approaches that can enable diverse real-world generalization while being compliant with human preferences. In this paper we argue that scaling robotic simulators is unlikely to help towards this goal of generalizable real-world manipulation, and can in fact serve as a misleading metric of progress towards this goal. We specifically argued that scaling simulation is neither necessary not sufficient for developing generalizable and compliant real-world manipulation that can be used for diverse tasks in diverse scenarios. We showed that while this does not preclude simulation from being *helpful* or *convenient* in certain scenarios, for robotic manipulation, there exist simpler and more easily scalable techniques with higher odds of succcess. We believe that as community we should focus our resources on techniques and data sources that are more likely to yield big wins. Most of the critique in this paper, as is typical in position papers is based on conceptual arguments, and evidence of prior research works. We welcome arguments and research methodologies that challenge the takeaways of this position paper, and hope that this structured critique of scaling simulations for real-world manipulation will spark discussions in the community, grounded in specifics rather than philosophical disagreements.

Impact Statement

This is a position paper argument for re-thinking the role of simulation in real-world robotic manipulation. There are many potential societal consequences of our work, specifically the deployment of robotic manipulation systems at large scale in people's homes, offices, kitchens etc.

Acknowledgement

We thank Abitha Thankaraj, Swaminathan Gurumurthy, Shubham Tulsiani, Jason Ma, Akash Sharma, Abhinav Gupta, Mandi Zhao, Vikash Kumar, Soroush Nasiriany, Sergey Levine, Chuhan Chen, and Eugene Vinitsky for discussions that inspired the paper, disagreements, comments, and feedback.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Afzal, A., Katz, D. S., Goues, C. L., and Timperley, C. S. A study on the challenges of using robotics simulators for testing. *arXiv preprint arXiv:2004.07368*, 2020.

Agarwal, A., Kumar, A., Malik, J., and Pathak, D. Legged locomotion in challenging terrains using egocentric vision. In *Conference on Robot Learning*, pp. 403–415. PMLR, 2023.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Bahl, S., Gupta, A., and Pathak, D. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

Bahl, S., Mendonca, R., Chen, L., Jain, U., and Pathak, D. Affordances from human videos as a versatile representation for robotics. *arXiv preprint arXiv:2304.08488*, 2023.

Bharadhwaj, H., Kumar, A., Rhinehart, N., Levine, S., Shkurti, F., and Garg, A. Conservative safety critics for exploration. *arXiv preprint arXiv:2010.14497*, 2020.

Bharadhwaj, H., Gupta, A., Kumar, V., and Tulsiani, S. Towards generalizable zero-shot manipulation via translating human interaction plans. *arXiv preprint arXiv:2312.00775*, 2023a.

Bharadhwaj, H., Gupta, A., Tulsiani, S., and Kumar, V. Zero-shot robot manipulation from passive human videos. *arXiv preprint arXiv:2302.02011*, 2023b.

Bharadhwaj, H., Vakil, J., Sharma, M., Gupta, A., Tulsiani, S., and Kumar, V. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. *arXiv preprint arXiv:2309.01918*, 2023c.

Bharadhwaj, H., Mottaghi, R., Gupta, A., and Tulsiani, S. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. *arXiv preprint arXiv:2405.01527*, 2024.

Bhirangi, R., Hellebrekers, T., Majidi, C., and Gupta, A. Reskin: versatile, replaceable, lasting tactile skins. In *5th Annual Conference on Robot Learning*, 2021.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

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

Chen, T., Tippur, M., Wu, S., Kumar, V., Adelson, E., and Agrawal, P. Visual dexterity: In-hand reorientation of novel and complex object shapes. *Science Robotics*, 8 (84):eadc9244, 2023a.

Chen, Z., Kiami, S., Gupta, A., and Kumar, V. Genaug: Retargeting behaviors to unseen situations via generative augmentation. *arXiv preprint arXiv:2302.06671*, 2023b.

Chi, C., Xu, Z., Pan, C., Cousineau, E., Burchfiel, B., Feng, S., Tedrake, R., and Song, S. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

Coumans, E. and Bai, Y. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021.

Dalal, M., Mandlekar, A., Garrett, C., Handa, A., Salakhutdinov, R., and Fox, D. Imitating task and motion planning with visuomotor transformers. *arXiv preprint arXiv:2305.16309*, 2023.

Dasari, S., Srirama, M. K., Jain, U., and Gupta, A. An unbiased look at datasets for visuo-motor pre-training. In *Conference on Robot Learning*, pp. 1183–1198. PMLR, 2023.

Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., and Yang, Y.

Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35: 13610–13626, 2022.

Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T. B., and Vanhoucke, V. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Freeman, C. D., Frey, E., Raichuk, A., Girgin, S., Mordatch, I., and Bachem, O. Brax–a differentiable physics engine for large scale rigid body simulation. *arXiv preprint arXiv:2106.13281*, 2021.

Fu, Z., Zhao, T. Z., and Finn, C. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

Goyal, M., Modi, S., Goyal, R., and Gupta, S. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3293–3303, 2022.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Handa, A., Allshire, A., Makoviychuk, V., Petrenko, A., Singh, R., Liu, J., Makoviichuk, D., Van Wyk, K., Zhurkevich, A., Sundaralingam, B., et al. Dextreme: Transfer of agile in-hand manipulation from simulation to reality. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5977–5984. IEEE, 2023.

Hansen, N., Wang, X., and Su, H. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

Hejna III, D. J. and Sadigh, D. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*, pp. 2014–2025. PMLR, 2023.

Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., and Hutter, M. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., and Rupprecht, C. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Kocabas, M., Yuan, Y., Molchanov, P., Guo, Y., Black, M. J., Hilliges, O., Kautz, J., and Iqbal, U. Pace: Human and camera motion estimation from in-the-wild videos. *arXiv preprint arXiv:2310.13768*, 2023.

Kumar, A., Fu, Z., Pathak, D., and Malik, J. Rma: Rapid motor adaptation for legged robots. *arXiv preprint arXiv:2107.04034*, 2021.

Kumar, A., Li, Z., Zeng, J., Pathak, D., Sreenath, K., and Malik, J. Adapting rapid motor adaptation for bipedal robots. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1161–1168. IEEE, 2022.

Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., and Hutter, M. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):eabc5986, 2020.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Lidec, Q. L., Jallet, W., Montaut, L., Laptev, I., Schmid, C., and Carpentier, J. Contact models in robotics: a comparative analysis. *arXiv preprint arXiv:2304.06372*, 2023.

Ma, Y. J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., and Zhang, A. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Ma, Y. J., Liang, W., Wang, G., Huang, D.-A., Bastani, O., Jayaraman, D., Zhu, Y., Fan, L., and Anandkumar, A. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*, 2023.

Mahi Shafiullah, N. M., Rai, A., Etukuru, H., Liu, Y., Misra, I., Chintala, S., and Pinto, L. On bringing robots home. *arXiv e-prints*, pp. arXiv–2311, 2023.

Maitin-Shepard, J., Cusumano-Towner, M., Lei, J., and Abbeel, P. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *2010 IEEE International Conference on Robotics and Automation*, pp. 2308–2315. IEEE, 2010.

Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Mandi, Z., Bharadhwaj, H., Moens, V., Song, S., Rajeswaran, A., and Kumar, V. Cacti: A framework for scalable multi-task multi-scene visual imitation learning. *arXiv preprint arXiv:2212.05711*, 2022.

Mandlekar, A., Nasiriany, S., Wen, B., Akinola, I., Narang, Y., Fan, L., Zhu, Y., and Fox, D. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., and Theobalt, C. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4): 1–14, 2017.

Nair, S., Rajeswaran, A., Kumar, V., Finn, C., and Gupta, A. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

Padalkar, A., Pooley, A., Jain, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Singh, A., Brohan, A., et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023a.

Padalkar, A., Pooley, A., Jain, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Singh, A., Brohan, A., et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023b.

Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., and Malik, J. Reconstructing hands in 3d with transformers. *arXiv preprint arXiv:2312.05251*, 2023.

Pumacay, W., Singh, I., Duan, J., Krishna, R., Thomason, J., and Fox, D. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

Qi, H., Yi, B., Suresh, S., Lambeta, M., Ma, Y., Calandra, R., and Malik, J. General in-hand object rotation with vision and touch. In *Conference on Robot Learning*, pp. 2549–2564. PMLR, 2023.

Rong, Y., Shiratori, T., and Joo, H. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020.

Sadeghi, F. and Levine, S. (cad)^{2}rl: Real single-image flight without a single real image. *CoRR*, abs/1611.04201, 2016. URL http://arxiv.org/abs/1611.04201.

Sadigh, D., Dragan, A. D., Sastry, S., and Seshia, S. A. *Active preference-based learning of reward functions*. 2017.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9339–9347, 2019.

Shaw, K., Bahl, S., and Pathak, D. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pp. 654–665. PMLR, 2023.

Shridhar, M., Manuelli, L., and Fox, D. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D. S., Maksymets, O., et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021.

Tan, J., Zhang, T., Coumans, E., Iscen, A., Bai, Y., Hafner, D., Bohez, S., and Vanhoucke, V. Sim-to-real: Learning agile locomotion for quadruped robots. *arXiv preprint arXiv:1804.10332*, 2018.

Thananjeyan, B., Balakrishna, A., Nair, S., Luo, M., Srinivasan, K., Hwang, M., Gonzalez, J. E., Ibarz, J., Finn, C., and Goldberg, K. Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3):4915–4922, 2021.

Thumm, J. and Althoff, M. Provably safe deep reinforcement learning for robotic manipulation in human environments. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 6344–6350. IEEE, 2022.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Truong, J., Zitkovich, A., Chernova, S., Batra, D., Zhang, T., Tan, J., and Yu, W. Indoorsim-to-outdoorreal: Learning to navigate outdoors without any outdoor experience. *arXiv preprint arXiv:2305.01098*, 2023.

Wang, C., Shi, H., Wang, W., Zhang, R., Fei-Fei, L., and Liu, C. K. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

Xiao, T., Radosavovic, I., Darrell, T., and Malik, J. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

Yan, M., Frosio, I., Tyree, S., and Kautz, J. Sim-to-real transfer of accurate grasping with eye-in-hand observations

and continuous control. *arXiv preprint arXiv:1712.03303*, 2017.

Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., and Zheng, F. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023a.

Yang, L., Li, K., Zhan, X., Lv, J., Xu, W., Li, J., and Lu, C. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2750–2760, 2022.

Yang, Y., Shi, G., Meng, X., Yu, W., Zhang, T., Tan, J., and Boots, B. Cajun: Continuous adaptive jumping using a learned centroidal controller. *arXiv preprint arXiv:2306.09557*, 2023b.

Ye, V., Pavlakos, G., Malik, J., and Kanazawa, A. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21222–21232, 2023a.

Ye, Y., Gupta, A., and Tulsiani, S. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3895–3905, 2022.

Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., and Liu, S. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, pp. 22479–22489, 2023b.

Yoon, J., Son, B., and Lee, D. Comparative study of physics engines for robot simulation with mechanical interaction. *Applied Sciences*, 13(2):680, 2023.

Yu, T., Xiao, T., Stone, A., Tompson, J., Brohan, A., Wang, S., Singh, J., Tan, C., Peralta, J., Ichter, B., et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023.

Yuan, W., Dong, S., and Adelson, E. H. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.