Open6DOR: Benchmarking Open-instruction 6-DoF Object Rearrangement and A VLM-based Approach

Yufei Ding[1,2*], Haoran Geng[1,3*], Chaoyi Xu[2], Xiaomeng Fang[4], Jiazhao Zhang[1,4], Songlin Wei[1,2], Qiyu Dai[1], Zhizheng Zhang[2], He Wang[1,2,4†]

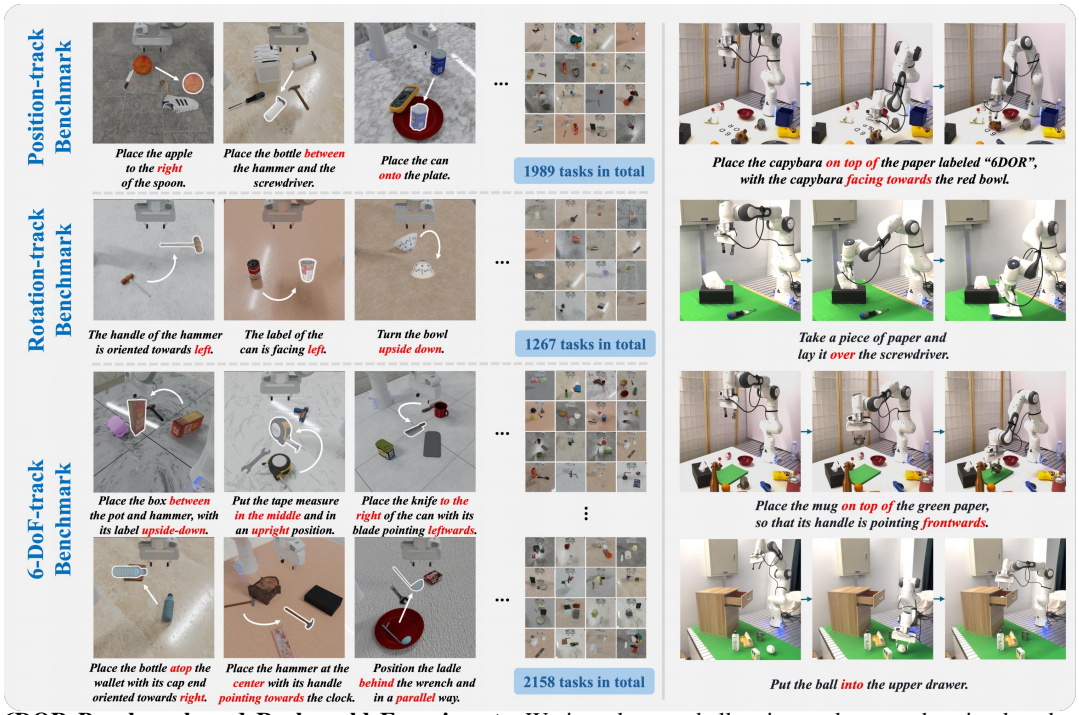

Fig. 1: **Open6DOR Benchmark and Real-world Experiments.** We introduce a challenging and comprehensive benchmark for Open-instruction 6-DoF object rearrangement tasks, termed Open6DOR. Following this, we propose a zero-shot and robust method, Open6DOR-GPT, which proves effective in demanding simulation environments and real-world scenarios.

***Abstract*— The integration of large-scale Vision-Language Models (VLMs) with embodied AI can greatly enhance the generalizability and the capacity to follow open instructions for robots. However, existing studies on object manipulation are not up to full consideration of the 6-DoF requirements, let alone establishing a comprehensive benchmark. In this paper, we propel the pioneer construction of the benchmark and approach for Open-instruction 6-DoF Object Rearrangement (Open6DOR). Specifically, we collect a synthetic dataset of 200+ objects and carefully design 5400+ Open6DOR tasks. These tasks are divided into the Position-track, Rotation-track, and 6-DoF-track for evaluating different embodied agents in predicting the positions and rotations of target objects.**

Besides, we also propose a VLM-based approach for Open6DOR, named Open6DOR-GPT, which empowers GPT-4V with 3D-awareness and simulation-assistance while exploiting its strengths in generalizability and instruction-following. We compare the existing embodied agents with our Open6DOR-GPT on the proposed Open6DOR benchmark and find that Open6DOR-GPT achieves the state-of-the-art performance. We further show the impressive performance of Open6DOR-GPT in diverse real-world experiments.

* Equal contribution.
[1] Peking University.
[2] Galbot
[3] University of California, Berkeley
[4] Beijing Academy of Artificial Intelligence.
Corresponding author: hewang@pku.edu.cn

I. Introduction

The advent of large-scale embodied models, exemplified by the RT series [1, 2, 4] and VoxPoser [5], has demonstrated considerable progress in mobile or fixed-station pick-and-place operations. While these models are capable of rearranging the object positions following human instructions, they fall short of satisfying full 6-DoF object placement instructions that involve specified 3D rotations. This limitation renders them incompetent at many practical robotic applications, where both object position and orientation are essential. For instance, in our daily life we often need a water bottle to be placed upright, while on the shelves in retail stores, goods should face the same direction. Moreover, previous works [2, 4, 5] are often evaluated on their own robots in their own scenes with self-reported performance and nonstandard evaluation metrics. The absence of a standard evaluation protocal condone cherry-picking, obstruct comparative assessment, and thus, hinder the iterative enhancement of effective approaches.

In this paper, we target the task of Open-instruction 6-DoF Object Rearrangement, referred to as Open6DOR,

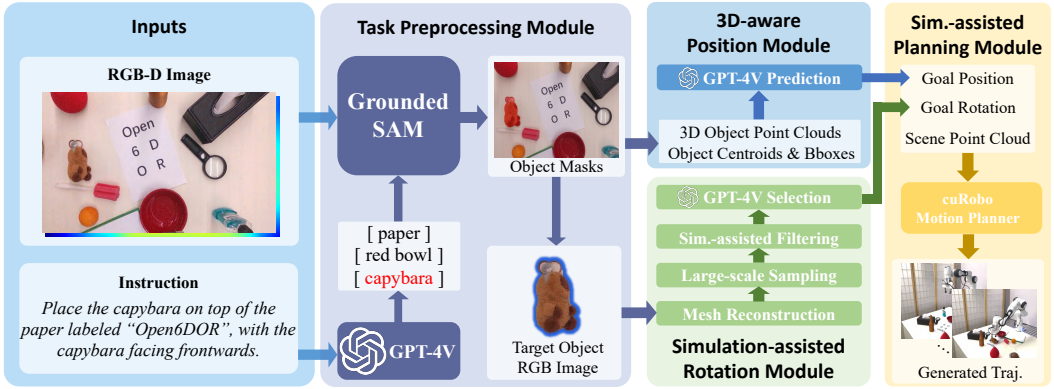

Fig. 2: **Method Overview.** Open6DOR-GPT takes the RGB-D image and instruction as input and outputs the corresponding robot motion trajectory. Firstly, the preprocessing module extracts the object names and masks. Then, two modules simultaneously predict the position and rotation of the target object in a decoupled way. Finally, the planning module generates a trajectory for execution.

which requires embodied agents to move the target objects according to open instructions that specify its 6-DoF pose. Open6DOR represents a fundamental skill for robotic manipulation tasks, presenting significant challenges in integrating instruction comprehension, 3D visual perception, and motion planning capabilities. Specifically, we promote the envelope of Open6DOR from two perspectives:

1) **Benchmark construction:** We construct a standardized benchmark, namely Open6DOR Benchmark, which comprises 5414 tasks designed with more than 200 objects across diverse categories in simulation environments. For comprehensive evaluation, we divide the Open6DOR benchmark into the position-track, rotation-track, and 6-DoF-track, each providing manually configured tasks along with comprehensive and quantitative 3D annotations. These tracks enable independent or combined assessments of translational, rotational, and overall performance.

2) **VLM-based approach:** We propose a VLM-based approach for Open6DOR tasks. Due to the aforementioned challenges of Open6DOR, all prior works, such as VoxPoser [5] and Dream2Real [7], fail to fulfill Open6DOR's 6-DoF requirements adequately. Among these efforts, Dream2Real [7] attempts to consider position and rotation dimensions simultaneously by imagining randomly rearranged scenes and leveraging VLM as an evaluator. This leads to almost intolerable time costs resulted from numerous renderings and VLM inferences, as well as unsatisfactory results due to the VLM's limited 3D perception, which renders it an incompetent critic. In contrast, we propose Open6DOR-GPT, which explicitly integrates 3D information from the initial scene into GPT-4V with equipped auxiliary modules and decomposes the translational and rotational determinations. In this way, we augment GPT-4V with 3D understanding capabilities and improve efficiency by reducing the determination space with decoupled modeling and simulation-assistance. Open6DOR-GPT achieves state-of-the-art performance in both benchmark evaluation and real-world experiments.

II. Open6DOR Benchmark

A. Open6DOR Task Formulation

We aim to identify a shared, fundamental component within complex embodied problems and, based on that, concisely formulate a elementary task, which we name Open6DOR. Open-instruction object rearrangement refers to the process wherein an embodied agent repositions objects within a scene from an initial state, following specific instructions. Specifically, a 6-DoF (Degrees of Freedom) object rearrangement task focuses on repositioning objects in a 6-DoF space, which includes both orientational and translational movement. We define each of these pick-and-place processes as an Open6DOR task, where a single target object is moved from its initial pose to a goal pose, guided by an open-vocabulary instruction. The input includes a single-view RGB-D image of the initial scene, denoted as I_{rgbd}, alongside a task instruction $\tilde{I}$ that specifies the desired pose of a target object in the scene. Based on these, the model is required to output the goal position P_{goal} and goal rotation R_{goal} of the target object. The Open6DOR task lies at the core of various long-horizon or complex-scene problems, simultaneously evaluating a model's capabilities in instruction following, 3D perception, and semantic understanding.

B. Open6DOR Benchmark Overview

The Open6DOR Benchmark is specifically designed for Open6DOR tasks grounded in simulation environment. To ensure comprehensive evaluation, we provide three specialized tracks of benchmark: Rotation-track $\mathcal{B}_r$, Position-track $\mathcal{B}_p$, and 6-DoF-track $\mathcal{B}_{6DOR}$.

Overall, the Open6DOR Benchmark consists of 5k+ tasks, featuring intricate configurations, realistic scenes, comprehensive annotations, and interactive environment.

Asset collection. The synthetic object dataset $\mathcal{O}_s$ comprises 200+ items spanning 70+ distinct categories.

Task configuration. Each task in the Open6DOR Benchmark is set within a table-top environment, where multiple objects are placed randomly. Initial object poses are carefully configured to avoid issues such as model clipping, range exceeding or unstable placement. Moreover, we manually design diverse instructions based on the target object, including positional and rotational requirements. The tasks are further reviewed to prevent occlusion or infeasible settings, ending with a total of 5k+ tasks.

Annotation and evaluation. For rotational assessment, we manually annotate each goal pose of a specific object

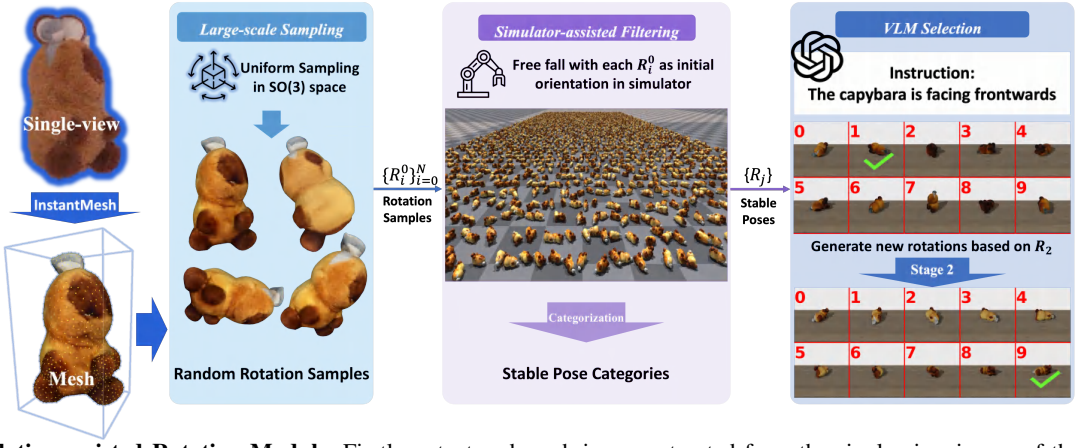

Fig. 3: **Simulation-assisted Rotation Module.** Firstly, a textured mesh is reconstructed from the single-view image of the target object. Then, we employ large-scale sampling to obtain multiple rotation samples. This sample set is then narrowed down through a simulation-assisted filtering process to derive several stable pose categories. Finally, we generate rendered images of the pose candidates, from which GPT-4V selects the optimal goal rotation.

as quaternions. The axis of symmetry, if present, is also specified to represent rotational equivalence. These annotations enable calculation of deviations between predicted and ground-truth rotations. For positional evaluation, we design heuristic functions to judge whether the spatial arrangement conforms to the instruction.

Simulation setting. The Open6DOR Benchmark is based on Isaac Gym [8], offering an interactive environment and executable platform. All tasks can be directly loaded into the simulator, in which users may control a robotic arm to complete the tasks. We also provide motion-planning APIs that generate actions based on a goal pose, implemented with cuRobo [10] and IsaacGym Motion Planing Library [8].

Rendering augmentation. To enhance observation realism, we propose a rendering API based on Blender [3]. Using such high-quality rendering, we generate a single-vew RGB-D image dataset, which serves as observation input for models. Additionally, the API enables customization of camera positions, lighting conditions, and background textures to accomodate personalized observation settings.

III. Open6DOR-GPT

A. Method Overview

As shown in Fig. 2, we enhance GPT-4V [9]'s capabilities to address the challenges of the Open6DOR task in a decomposed way. Initially, the Task Preprocessing Module deciphers $\tilde{I}$ based on the I_{rgbd} and feeds the resulting images to the Position Module and Rotation Module respectively. Within the two modules, we empower GPT-4V with 3D awareness and simulation assistance, thereby effectively outputting the predicted goal position P_{goal} and rotation R_{goal}. Finally, the Simulation-assisted Planning Module identifies a suitable grasping pose and plans out an optimal action trajectory to accomplish the task. We will first introduce each module of our proposed system in subsections *B-E* to explain how an Open6DOR task is accomplished. We then elaborate on how the system tackles long-horizon tasks with multiple rounds of operations.

B. Task Preprocessing Module

With the single-view RGB-D Image I_{rgbd} and the task instruction $\tilde{I}$ as input, this module leverages GPT-4V to interpret the instruction and identifies object names$\{O_i^{name}\}$, which in turn triggers GroundedSAM [6] to generate a set of labeled masks. Based on the masked Image I_{mask}, the RGB image of the target object I_{object} is extracted. These images are used in subsequent modules.

C. 3D-aware Position Module

Taking the masked RGB-D image I_{mask} and task instruction $\tilde{I}$ as input, the 3D-aware Position Module $\mathcal{M}_p$ determines and outputs the goal position.

To incorporate three-dimensional (3D) data into GPT-4V's understanding, our approach utilizes back-projection based on I_{mask} to generate a 3D masked point cloud, symbolized as PC_i^{3d}. This computation includes determining the centroid Center_i^{3d} and bounding box Bbox_i^{3d} of the point cloud associated with the queried object.

These spatial attributes are then integrated back into the prompt for GPT-4V, facilitating the model to accurately ascertain the goal position for the target object P_{goal}.

D. Simulation-assisted Rotation Module

As illustrated in Fig. 3, with the single-view RGB image of the target object I_{object} and the task instruction $\tilde{I}$ as input, the rotation module would output the goal rotation R_{goal} for the object. We first reconstruct the target object from I_{object} using InstantMesh [11] , resulting in a textured mesh denoted as M. The reconstruction process is followed by four phases: (1) large-scale sampling (2) simulation-assisted filtering (3) rotation categorization (4) GPT-4V selection.

E. Simulation-assisted Planning Module

Utilizing the predicted goal position P_{goal} and goal rotation R_{goal}, the planning module formulates an effective execution strategy with simulation assistance.

IV. Experiments

A. Results on Position-track Benchmark

We evaluate the performance of our position module and several baselines on the Position-track Benchmark. As shown in Table I, Comparatively, our approach markedly surpasses all these baselines by over 30 percent, demonstrating superior and consistent performance on the Position-track Benchmark.

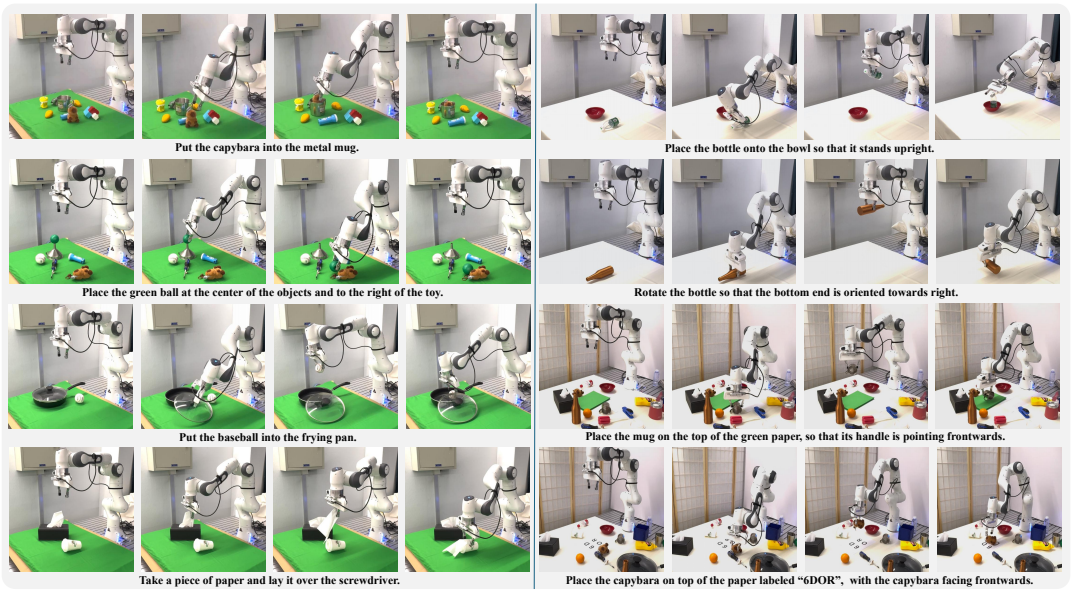

Fig. 4: **Real-world Experiments.** We ground Open6DOR-GPT in real-world settings and conduct various tasks including long-horizon ones, demonstrating its zero-shot generalization potential across challenging tasks.

Success Rate (%)	Level 0	Level 1	Level 2	Overall
GPT-4V [9]	46.8	39.1	50.0	45.2
Dream2Real* [7]	17.2	11.0	-	15.9
VoxPoser* [5]	35.6	21.7	0.0	32.6
VoxPoser(VLM)* [5]	37.2	19.9	0.0	33.5
Open6DOR-GPT	**78.6**	**60.3**	**80.0**	**74.9**

TABLE I: **Results on Position-track Benchmark.** We compared our approach against several benchmarks for positioning proposals. This includes: (1) GPT-4V [9], utilizing pixel input to predict object placement and employing depth for 3D location. (2) A tailored Dream2Real [7] baseline for our task. (3,4) VoxPoser [5] original and adapted versions, aligning with our goals. Our tests include GPT-4V's Large Language Model (LLM) and Vision-Language Model (VLM) setups, with an asterisk denoting ground-truth data usage as reference baselines.

Success Rate(%)	Level 0	Level 1	Level 2	Overall
GPT-4V [9]	9.1	6.9	11.7	9.2
Dream2Real* [7]	37.3	27.6	26.2	31.3
S-F + GPT-4V	41.1	30.7	30.4	38.4
Open6DOR-GPT (S-F + 2-Stage 4V)	**45.7**	**32.5**	**49.8**	**41.1**
Open6DOR-GPT*	**48.2**	**32.6**	**60.0**	**44.1**

TABLE II: **Results on Rotation-track Benchmark.** Quantitative comparison with a refined version of Dream2Real [7] method (replacing CLIP Model with GPT-4V), and ablation studies of different phases in the Rotation Module. 'S-F' stands for 'Sampling-Filtering'. '*' means using ground-truth mesh instead of reconstructed ones. The first three rows ablate Phase1-4, Phase3-4, and Stage2 in Phase 4, respectively.

B. Results on Rotation-track Benchmark

Our Rotation Module comprises four phases aimed at enhancing GPT-4V [9] through a simulation-assisted sample-and-filter mechanism. To evaluate the effectiveness of each phase, we conduct ablation studies using the Rotation-track of Open6DOR Benchmark, with results detailed in Table II.

Success Rate (%)	Rotation	Position	Overall	Time Cost(s)
Dream2Real [7]	-	-	-	>700
Dream2Real* [7]	18.7	26.2	13.5	358.3
Open6DOR-GPT	**40.0**	**84.8**	**35.6**	**126.3**

TABLE III: **Results on 6-DoF-track Benchmark.** We compare our method with an optimized version of Dream2Real [7] on the 6DoF Benchmark. The three columns depict the quality of the goal pose in terms of rotation, position, and overall performance.

C. Results on 6-DoF Benchmark

We evaluate our entire pipeline using the 6DoF-track of Open6DOR Benchmark. The evaluation of rotational, positional, and joint performance are presented in Table III. Our approach also demonstrates better efficiency compared to baseline approaches.

D. Real-world Experiments

As shown in Fig. 4, our zero-shot method is able to tackle challenging Open6DOR scenarios and demonstrates strong potential in long-horizon tasks.

V. Conclusion

In this paper, we pioneer the establishment of the Open6DOR benchmark and VLM-based approach, addressing the need for a comprehensive evaluation and a foregoing method exploration in open-instruction 6-DoF object rearrangement. Our synthetic benchmark, comprising over 200 objects and 5400 tasks, offers a standardized framework for evaluating the capabilities of embodied agents in simulation environments. Additionally, our Open6DOR-GPT approach achieves state-of-the-art performance, augmenting GPT-4V with 3D awareness and simulation assistance. As for the current limitations, while Open6DOR-GPT significantly improves position and rotation handling, it does not achieve real-time performance, and rotation understanding remains suboptimal. We look forward to future improvements to our benchmarks, especially for real-world extensions.

References

[1] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M.G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.W.E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: Rt-2: Vision-language-action models transfer web knowledge to robotic control (2023) 1

[2] 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) 1

[3] Foundation, B.: Blender. https://www.blender.org/ (2024), accessed: 2024-09-11 3

[4] Gu, J., Kirmani, S., Wohlhart, P., Lu, Y., Arenas, M.G., Rao, K., Yu, W., Fu, C., Gopalakrishnan, K., Xu, Z., et al.: Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. arXiv preprint arXiv:2311.01977 (2023) 1

[5] Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973 (2023) 1, 2, 4

[6] Huang, W., Xia, F., Shah, D., Driess, D., Zeng, A., Lu, Y., Florence, P., Mordatch, I., Levine, S., Hausman, K., Ichter, B.: Grounded decoding: Guiding text generation with grounded models for robot control (2023) 3

[7] Kapelyukh, I., Ren, Y., Alzugaray, I., Johns, E.: Dream2real: Zero-shot 3d object rearrangement with vision-language models. arXiv preprint arXiv:2312.04533 (2023) 2, 4

[8] Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., State, G.: Isaac gym: High performance gpu-based physics simulation for robot learning (2021), https://arxiv.org/abs/2108.10470 3

[9] OpenAI: Gpt-4 technical report (2023) 3, 4

[10] Sundaralingam, B., Hari, S.K.S., Fishman, A., Garrett, C., Wyk, K.V., Blukis, V., Millane, A., Oleynikova, H., Handa, A., Ramos, F., Ratliff, N., Fox, D.: curobo: Parallelized collision-free minimum-jerk robot motion generation (2023) 3

[11] Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024) 3