ManiGaussian++: General Robotic Bimanual Manipulation with Hierarchical Gaussian World Model

Tengbo Yu*,1, Guanxing Lu*,1, Zaijia Yang*,2, Haoyuan Deng3, Season Si Chen1,
Jiwen Lu4, Wenbo Ding1, Guoqiang Hu3, Yansong Tang†,1, Ziwei Wang3

***Abstract*— Multi-task robotic bimanual manipulation is becoming increasingly popular as it enables sophisticated tasks that require diverse dual-arm collaboration patterns. Compared to unimanual manipulation, bimanual tasks pose challenges to understanding the multi-body spatiotemporal dynamics. An existing method ManiGaussian [30] pioneers encoding the spatiotemporal dynamics into the visual representation via Gaussian world model for single-arm settings, which ignores the interaction of multiple embodiments for dual-arm systems with significant performance drop. In this paper, we propose ManiGaussian++, an extension of ManiGaussian framework that improves multi-task bimanual manipulation by digesting multi-body scene dynamics through a hierarchical Gaussian world model. To be specific, we first generate task-oriented Gaussian Splatting from intermediate visual features, which aims to differentiate acting and stabilizing arms for multi-body spatiotemporal dynamics modeling. We then build a hierarchical Gaussian world model with the leader-follower architecture, where the multi-body spatiotemporal dynamics is mined for intermediate visual representation via future scene prediction. The leader predicts Gaussian Splatting deformation caused by motions of the stabilizing arm, through which the follower generates the physical consequences resulted from the movement of the acting arm. As a result, our method significantly outperforms the current state-of-the-art bimanual manipulation techniques by an improvement of 20.2% in 10 simulated tasks, and achieves 60% success rate on average in 9 challenging real-world tasks. Our code is available at `https://github.com/April-Yz/ManiGaussian_Bimanual`.**

I. Introduction

General robotic bimanual manipulation agent is trending for their immeasurable potential in completing diverse complex tasks across houses [50], [49], hospitals [23], and factories [2]. A bimanual system is more than a naive combination of separate single-arm agents, as it enables challenging task that entails specific collaboration patterns including simultaneously manipulating and stabilizing target objects [28], [14]. Additionally, bimanual systems are often outperform unimanual agents when multiple action steps can be performed in parallel by different arms [15]. As a result, general bimanual manipulation system that can interact with diverse objects and environments across tasks is highly-desired in recent years. Typically, a robot manipulation agent [42], [15] comprises a perception module that encodes the visual clues into latent representation, and a policy head that maps these representation to the robotic action space. However, conventional visual representations [15], [28] usually suffer from insufficient generalization ability to multi-task bimanual manipulation scenarios, which is required to mine the unstructured scene geometry across diverse tasks and objects.

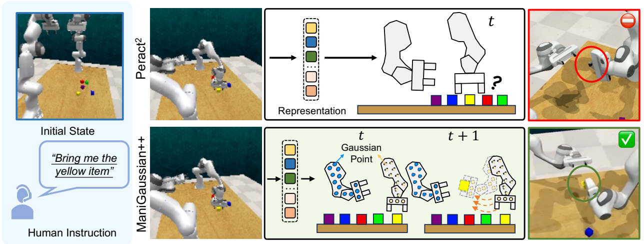


Fig. 1. Consider the human instruction *"Bring me the yellow item"*, where the task is considered successful if the right arm handover the yellow block to the left arm. The previous method (Peract^2 [15]) attempts to pick up the yellow block but fails to do so, while our ManiGaussian++ completes the task successfully by explicitly encoding the scene dynamics via future scene reconstruction in Gaussian embedding space.

To address this, existing works [30], [52], [47] propose to leverage self-supervised learning to enhance the generalization ability of the visual representation for robot learning, Earlier works [34], [33], [36] first propose to leverage pretrained 2D visual representation, which have shown initial success of visual representation learning but are constrained to relatively simple tasks due to the lack of geometric understanding like occlusion. To apply to more complex manipulation tasks that require 3D scene understanding, preceding methods [46], [52] attempt to model the workspace via 3D reconstruction methods like Neural Radiance Fields (NeRFs) [32] or Gaussian Splatting [26]. For example, ManiGaussian [30] pioneers to explicitly encode the scene dynamics via future scene reconstruction in Gaussian embedding space, which shows impressive performance in single-arm setting [25]. However, learning multi-body spatiotemporal dynamics in the dual-arm system poses challenges for existing methods, thereby leads to severe performance drops in bimanual manipulation scenarios.

In this paper, we propose a general bimanual manipulation agent named ManiGaussian++, which leverages hierarchical Gaussian world model to encode the multi-body spatiotemporal dynamics in dual-arm systems. Different from the

*: Equal Contribution
†: Corresponding Author
[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, Emails: {ytb23@mails.,lgx23@mails.,season.chen@, tang.yansong@, ding.wenbo@}sz.tsinghua.edu.cn
[2] School of Computer Science and Technology, Hainan University, Emails: 20213002625@hainanu.edu.cn
[3] School of Electrical and Electronic Engineering, Nanyang Technological University, Emails: {E230112@e., gqhu@, ziwei.wang@}ntu.edu.sg
[4] Department of Automation, Tsinghua University, Emails: lujiwen@tsinghua.edu.cn

prior ManiGaussian which only encodes coarse scene-level spatiotemporal dynamics, ManiGaussian++ concentrates on the multi-body spatiotemporal dynamics of the dual-arm system for bimanual manipulation tasks. Therefore, the complex interactions between the two manipulators and targets are considered to accomplish diverse collaboration patterns. More specifically, ManiGaussian++ first generates a task-oriented Gaussian radiance field from intermediate visual representations supervised by pre-trained vision-language models (VLMs), allowing us to assign distinct roles to the robot arms including stabilizing and acting arms for multi-body spatiotemporal dynamics modeling. Subsequently, we mine the multi-body spatiotemporal dynamics for intermediate visual representations via future scene prediction, where a hierarchical Gaussian world model with the leader-follower architecture is utilized. The leader predicts Gaussian Splatting deformation caused by motions of the stabilizing arm, through which the follower generates the physical consequences resulted from the movement of the acting arm. ManiGaussian++ demonstrates significant improvements over existing bimanual manipulation techniques across 10 simulated and 9 real-world tasks by sizable margins in terms of success rate. The contributions are as follows:

- We propose a general robotic bimanual manipulation agent named ManiGaussian++, which extends prior ManiGaussian by introducing the hierarchical Gaussian world model to learn the multi-body spatiotemporal dynamics for bimanual tasks.
- We generate task-oriented Gaussian Splatting to differentiate acting and stabilizing arms for multi-body dynamics modeling, and we propose a hierarchical Gaussian world model with future scene prediction to mine the multi-body dynamics.
- We perform comprehensive experiments on 10 tasks from RLBench2. The results indicate that our method surpasses the state-of-the-art approaches by large relative margins of 131.17%. We also evaluate ManiGaussian++ in real bimanual settings and obtain 60% success rates across 9 real-world tasks.

II. Related Work

Robotic Bimanual Manipulation. Generalizable bimanual manipulation agent enables complex task completion by introducing complex collaboration patterns, which is of great importance in various applications, such as housekeeping [49], healthcaring [23], [27], and manufacturing [2]. Existing methods [15], [10], [8], [14] attempt to train a foundation model for bimanual manipulation by manual-crafting a large set of demonstrations for imitation learning. However, due to the precise coordination between two high-degree-of-freedom arms required by bimanual tasks, teleoperating demonstrations for training generalizable policies is costly [50], [5], [4], [45], which presents challenges for the bimanual manipulation policy model to generalize to unseen tasks. For instance, by predicting the keyframe action and leveraging the 3D-aware voxel observation, a recent work PerAct^2 [15] shows initial potential in single-task bimanual manipulation while notably mitigating the demand of large-scale expert demonstrations. Though PerAct^2 improves the expressiveness of the policy network part by leveraging scalable multi-modal transformer [24], the visual representation that bottlenecks the generalizability of bimanual manipulation is neglected, which is the main focus of this paper.

Visual Representations for Robot Learning. To enhance the generalizability of the robotic manipulation agents, prior arts propose more powerful network architectures [12], [11], [42], [15] or employ self-supervision [33], [36], [22], [30] in visual representation learning. These methods aim to leverage diverse visual observations like mutli-view images [13], [12], point cloud [3], [11], and voxel [42], [15], [28], but often struggle with limited labeled data. To address this problem, self-supervised methods [34], [33], [36] enhances generalization ability through auxiliary tasks with prior knowledge. Earlier studies [34], [33], [36] have focused on enhancing 2D visual representations via self-supervised techniques such as time-contrastive learning [33] and masked modeling [36], but are limited to simpler tasks. Recent approaches tackle 3D scene comprehension with techniques like Neural Radiance Fields (NeRFs) [32] and Gaussian Splatting [26], with method like ManiGaussian [30] encoding scene dynamics for single-arm tasks. However, multi-body dynamics in complex tasks require more advanced approaches to handle significant challenges.

World Models. A world model simulates the future scene according to the current state and agent action, which often serves as an self-supervised objective to encode the underlying scene dynamcis for autonomous agent of various applications like autonomous driving [43], [9], gaming [17], [18], [20] and robotic manipulation [21], [44]. Early research [16], [17], [18], [19] focus on learning a latent space for future predictions with recurrent state-space models, demonstrating significant effectiveness in both simulated and real-world environments. With the evolution of cutting-edge network architectures, world models that predict high-dimensional representations are often used to predict the future image [7], [39], [31], [38], voxel [51] and Gaussian [1], [30], [48] domains. However, the multi-agent nature of bimanual manipulation poses challenges to model the mutual interactions between two manipulators and targets, and thus we introduce a hierarchical Gaussian world model to overcome this.

Gaussian Splatting. Gaussian Splatting [26] is termed by representing scenes using a collection of 3D Gaussian functions that can be 'splatted' onto 2D planes with rasterization, compared to implicit models like Neural Radiance Fields (NeRF) [32], [6], [40]. Recent works [35], [41], [29], [30] begin to notice the great potential of the Gaussian radiance field in robotic manipulation, which is able to consistently tracking the manipulator and target with its explicit and editable nature. Although these methods depicts impressive performance in unimanual settings by incorporating Gaussian-based representation, bimanual manipulation that involves complex multi-body dynamics is still unexplored.

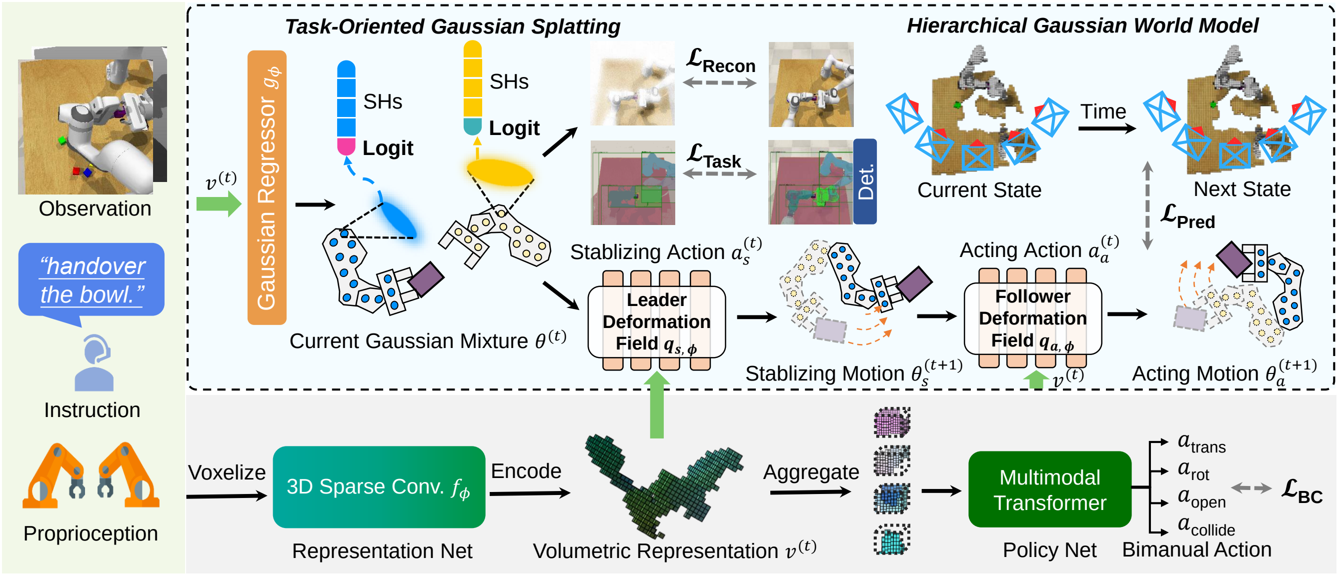

Task-Oriented Gaussian Splatting
Hierarchical Gaussian World Model
Observation
"handover the bowl."
Instruction
Proprioception
$v^{(t)}$
Gaussian Regressor g_ϕ
SHs
Logit
SHs
Logit
Current Gaussian Mixture $\theta^{(t)}$
$\mathcal{L}_{\text{Recon}}$
$\mathcal{L}_{\text{Task}}$
Det.
Current State
Time
Next State
$\mathcal{L}_{\text{Pred}}$
Stablizing Action $a_s^{(t)}$
Acting Action $a_a^{(t)}$
Leader Deformation Field $q_{s,\phi}$
Follower Deformation Field $q_{a,\phi}$
Stablizing Motion $\theta_s^{(t+1)}$
$v^{(t)}$
Acting Motion $\theta_a^{(t+1)}$
Voxelize
3D Sparse Conv. f_ϕ
Encode
Representation Net
Volumetric Representation $v^{(t)}$
Aggregate
Multimodal Transformer
Policy Net
a_{trans}
a_{rot}
a_{open}
a_{collide}
$\mathcal{L}_{\text{BC}}$
Bimanual Action

Fig. 2. **The overall pipeline of ManiGaussian++.** The task-oriented Gaussian radiance assigns unique labels to task-relevant agents and objects, and the hierarchical Gaussian world model upon it predicts future scenes in a leader-follower manner to encode the multi-body dynamics for bimanual manipulation.

III. Approach

In this section, we first briefly introduce preliminaries on bimanual manipulation, and then we depict an overview of our pipeline. Subsequently, we present the task-oriented Gaussian Splatting for multi-body modeling, and introduce the hierarchical Gaussian world model to encode the multi-body spatiotemporal dynamics. Finally, We describe the learning objectives for supervisions.

A. Problem Formulation

Language-conditioned bimanual manipulation is essential for next generation's general intelligent robot. To carry out various bimanual manipulation tasks, the agent must interactively predict the next best poses of both end-effectors to accomplish the human instruction, based on observations including visual input and the robot's proprioception. The observation at the t_{th} time step is $\mathbf{o}^{(t)} = (C^{(t)}, D^{(t)}, P^{(t)})$, where $C^{(t)}$ and $D^{(t)}$ respectively represent the RGB image and the depth image. $P^{(t)}$ is the proprioception that contains current time and gripper states. The action $\mathbf{a}^{(t)}$ for each end-effector at the t_{th} step contains the position $a_{\text{trans}}^{(t)} \in \mathbb{R}^{100^3}$, orientation $a_{\text{rot}}^{(t)} \in \mathbb{R}^{(360/5)\times 3}$, openness $a_{\text{open}}^{(t)} \in [0,1]$ and an indicator of whether to invoke collision avoidance of the motion planner $a_{\text{collide}}^{(t)} \in [0,1]$. For robot learning, we assume access to K offline trajectories composed of observation, left-arm action and right-arm action triplets paired with human instructions. To address the issue of limited demonstrations, traditional arts [30] enhance the visual representation by mining the spatiotemporal dynamics via self-supervised future scene reconstruction. However, the multi-body nature of the bimanual manipulation poses challenges to modeling the spatiotemporal dynamics precisely, and the decoded actions based on the ineffective visual representation fail to complete human instructions with incorrect collaboration patterns.

B. Overall Pipeline

The overall pipeline of our ManiGaussian++ method is shown in Figure 2. We first develop a task-oriented Gaussian radiance field to distinguish stabilizing and acting arms, where a stabilizing arm secures the object in hand while an acting arm performs the task to mitigate the multi-modality of bimanual manipulation. Subsequently, we build a hierarchical Gaussian world model in leader-follow architecture upon it to forecast the future scene, through which the multi-body spatiotemporal dynamics can be encoded to discover complex collaboration patterns. More specifically, we transform the visual input from RGB-D cameras to voxel space based on the calibrated camera parameters, which is then encoded by a sparse convolutional network as the visual representation in a volumetric format. For task-oriented Gaussian Splatting, we design a feed-forward Gaussian regressor to infer the task-oriented Gaussian radiance field from the visual representation, where each Gaussian particle is assigned with a task-oriented instance logit distilled from pretrained VLMs to differentiate acting and stabilizing arms for multi-body interactions. We create a hierarchical Gaussian world model with a leader-follower architecture to learn visual representation through future scene prediction. The leader anticipates deformation from the motion of the stabilizing arm, allowing the follower to generate the effects of the movement of the acting arm. Finally, we employ multi-modal transformer PerceiverIO [24] to predict the optimal robot actions based on the enhanced volumetric representation, which comprehends the multi-body spatiotemporal dynamics and thus can complete human instructions with precise collaboration patterns.

C. Task-Oriented Gaussian Splatting

In order to capture the multi-body spatiotemporal dynamics for general bimanual manipulation tasks, we start with a task-oriented Gaussian Splatting that disentangles the acting and stabilizing arms from cluttered scenes for representing the visual scene. Gaussian Splatting [26] is trending with its explicit nature that enables rapid rendering via rasterization. A Gaussian radiance field represents a scene with multiple Gaussian primitives, which can be parameterized by $\theta_i = (\mu_i, c_i, r_i, s_i, \sigma_i)$, which respectively represent the position, color, orientation, scaling, and opacity for the i-th Gaussian primitive. To render a novel image C, those 3D Gaussian primitives can be projected onto the 2D camera plane via differential tile-based rasterization. In this process, a typical pixel $\mathbf{p}$ can be colored by the alpha-blend rendering:

$$C(\mathbf{p}) = \sum_{i=1}^{N} \alpha_i c_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{1}$$

where N is the number of Gaussians in a tile, α_i represents the 2D density of the Gaussian points that can be computed by μ_i, r_i and s_i of the parameters. Though the vanilla Gaussian Splatting shows effectiveness in reconstructing 3D appearance and geometry, it struggles to generate high-quality task-oriented labels for relevant instances, which is of significance to focus on learning precise multi-body dynamics for bimanual manipulation tasks. To this end, we modify the Gaussian parameters and parameterize a Gaussian regressor to construct a task-oriented Gaussian radiance field, where the instance labels that distinguish the stabilizing and acting manipulators are learned simultaneously with the appearance and geometry. Besides, we enable the Gaussian particles to move with discrete time to account for the spatiotemporal dynamics of the scene, where the parameter of the i_{th} Gaussian at the t_{th} is:

$$\theta_i^{(t)} = (\mu_i^{(t)}, c_i^{(t)}, r_i^{(t)}, s_i^{(t)}, \sigma_i^{(t)}, l_i^{(t)}). \tag{2}$$

The positions, colors, orientations, scales, and opacities with the superscript t represent their counterparts at the t_{th} step in the movement. We append a $l_i^{(t)} \in \mathbb{R}^3$ variable as the instance-level logit to the Gaussian parameter set, which represents the probability of the Gaussian point belonging to a specific task-relevant instance including different manipulators or target objects. The instance map can also be rendered by projecting the instance-level logits of Gaussian primitives to the 2D camera plane. To be specific, the expected instance logits L of a typical pixel $\mathbf{p}$ can be written as:

$$L(\mathbf{p}) = \sum_{i=1}^{N} \alpha_i l_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{3}$$

where we omit the t superscript for brevity. To obtain the ground-truth instance map for both manipulators and target objects, we prompt the pretrained VLMs such as the open-vocabulary detector GroundedSAM [37] based on the keywords from human instructions.

D. Hierarchical Gaussian World Model for Bimanual Manipulation

Though the multi-body nature of bimanual manipulation enables complicated task completion, it also introduces novel challenges for the world model to learn multi-body spatiotemporal dynamics beyond unimanual manipulation. In order to capture the multi-body spatiotemporal dynamics for the visual representation in general bimanual manipulation tasks, we propose a hierarchical Gaussian world model with a leader-follower architecture for precise future scene prediction. The leader predicts Gaussian Splatting deformation caused by motions of the stabilizing arm, through which the follower generates the physical consequences resulted from the movement of the acting arm. The hierarchical Gaussian world model models the movement of explicit Gaussian points conditioned on the robot's action. For future prediction of our hierarchical Gaussian world model, the dominant movement can be regarded as rigid-body transform. Therefore, we only predict the SE(3) movement of Gaussian particle following the Newton-Euler equation, while keeping the inherent properties including color, scaling, opacity, and instance logits the same along the Markovian transition. The Gaussian particle changes caused by arm motions can be formulated as:

$$(\boldsymbol{\mu}^{(t+1)}, \mathbf{r}^{(t+1)}) = (\boldsymbol{\mu}^{(t)} + \Delta\boldsymbol{\mu}_{\mathrm{s}}^{(t)} + \Delta\boldsymbol{\mu}_{\mathrm{a}}^{(t)}, \mathbf{r}^{(t)} + \Delta\mathbf{r}_{\mathrm{s}}^{(t)} + \Delta\mathbf{r}_{\mathrm{a}}^{(t)}), \tag{4}$$

where we denote the changes of positions and orientation by $\Delta\boldsymbol{\mu}_{\mathrm{s}}^{(t)}$, $\Delta\boldsymbol{r}_{\mathrm{s}}^{(t)}$ and $\Delta\boldsymbol{\mu}_{\mathrm{a}}^{(t)}$, $\Delta\boldsymbol{r}_{\mathrm{a}}^{(t)}$ for stabilizing and acting manipulator, respectively. The subscript i for each Gaussian particle is omitted here for brevity. To predict the movements described in Equation (4) for multi-body spatiotemporal dynamics learning, we parameterize a hierarchical Gaussian world model that takes the visual representation as input, and outputs the future multi-view images for photometric supervision. More specifically, the hierarchical Gaussian world model contains a representation network f_ϕ that infers intermediate visual representation from the voxel observation, where ϕ refers to the learnable parameters. Then, a Gaussian regressor g_ϕ is utilized to reconstruct the current task-oriented Gaussian radiance field in a feed-foward manner. For future scene reconstruction, a leader deformation model $q_{\mathrm{s},\phi}$ interprets the preliminary movements imposed by the stabilizing arm as a Gaussian deviation $\theta_{\mathrm{s}}^{(t+1)}$, a follower deformation model $q_{\mathrm{a},\phi}$ predicts the physical consequences $\theta_{\mathrm{a}}^{(t+1)}$ by concluding both the acting and stabilizing arm. At last, a Gaussian renderer $\mathcal{R}$ in Equation (1) projects the predicted Gaussian radiance field onto the 2D camera plane:

$$\begin{cases} \text{Representation: } v^{(t)} = f_\phi(o^{(t)}) \\ \text{Gaussian regressor: } \theta^{(t)} = g_\phi(v^{(t)}) \\ \text{Leader model: } \theta_r^{(t+1)} = q_{\mathrm{s},\phi}(\theta^{(t)}, \mathbf{a}_{\mathrm{s}}^{(t)}, v^{(t)}) \\ \text{Follower model: } \theta_l^{(t+1)} = q_{\mathrm{a},\phi}(\theta_r^{(t+1)}, \mathbf{a}_{\mathrm{s}}^{(t)}, \mathbf{a}_{\mathrm{a}}^{(t)}, v^{(t)}) \\ \text{Gaussian renderer: } C^{(t+1)}, L^{(t+1)} = \mathcal{R}(\theta^{(t+1)}), \end{cases} \tag{5}$$

where $v^{(t)}$ denotes the enhanced visual representation, $\mathbf{a}_{\mathrm{s}}^{(t)}$ and $\mathbf{a}_{\mathrm{a}}^{(t)}$ are the stabilizing and acting actions at the t_{th} step. In order to effectively forecast future scenes with the hierarchical Gaussian world model, the visual representation is enhanced to capture multi-body spatiotemporal dynamics of the environment, which is crucial for generating appropriate bimanual actions with complex collaboration patterns.

TABLE I

MULTI-TASK TEST RESULTS. MEAN SUCCESS RATES (%) OF MULTI-TASK AGENTS TRAINED WITH 100 DEMONSTRATIONS PER TASK AND EVALUATED OVER 25 EPISODES.

Method / Task	pick laptop	straighten rope	lift tray	push box	handover easy	put in fridge	press buttons	handover item	sweep to dustpan	take out tray	**Average Success ↑**	**Average Rank ↓**
PerAct^2 [15]	**12**	24	1	6	**41**	3	47	11	0	9	15.4	2.5
ManiGaussian [30]	8	28	4	24	36	4	36	12	24	12	18.8	2.2
ManiGaussian++ (Ours)	**12**	**40**	**8**	**48**	40	**28**	**48**	**20**	**92**	**16**	**35.6**	**1.1**

E. Learning Objectives

Current Scene Reconstruction. To encode the spatial consistency into the visual representation for further multi-body spatiotemporal dynamics digesting, we impose a multi-view photometric loss to regularize the current Gaussian Splatting generated from the Gaussian regressor:

$$\mathcal{L}_{\text{Recon}} = \sum_{\mathbf{p}} \|C^{(t)}(\mathbf{p}) - \hat{C}^{(t)}(\mathbf{p})\|_2^2, \tag{6}$$

where $C^{(t)}$ and $\hat{C}^{(t)}$ stand for the predicted and ground-truth 2D images from a randomly selected view at t_{th} time step, respectively.

Task-oriented Embodiment Mask Prediction. The task-oriented Gaussian radiance field differentiates acting and stabilizing arms for the follow-up multi-body spatiotemporal dynamics modeling. To optimize this, we first aggregate the logits embedded in the Gaussian particles by rendering them to the camera plane via rasterization, and then implement a cross-entropy objective:

$$\mathcal{L}_{\text{Task}} = -\sum_{\mathbf{p}} \sum_{l} \hat{B}^l(\mathbf{p}) \log B^l(\mathbf{p}) \tag{7}$$

where B^l and $\hat{B}^l$ are the predicted and ground-truth probability. The predicted probability is computed by normalizing the rendered instance logit map L via softmax, while the ground-truth probability is a discrete label obtained by prompting the VLM of a specific label l in the 2D camera plane.

Future Scene Prediction. To embed the multi-body spatiotemporal dynamics in the visual representations, we encourage the predicted scene based on the learned Gaussian parameters to get close to the ground-truth one. As we can not directly access the ground-truth future Gaussian radiance field, the training goal is to align predicted future images from multiple views with the ground-truth images obtained by actually taking the bimanual action, as follows:

$$\mathcal{L}_{\text{Pred}} = \|\hat{\mathbf{C}}^{(t+1)} - \mathbf{C}^{(t+1)}\|_2^2 \tag{8}$$

where $\mathbf{C}^{(t+1)}$ and $\hat{\mathbf{C}}^{(t+1)}$ represent the predicted and ground-truth future image, respectively.

Behavior Cloning. Following [42], [15] for fair comparisons, we leverage a multi-modal transformer PerceiverIO [24] to select the best candidates from discretized action bins based on enhanced volumetric representation and language instruction, where we leverage the cross-entropy loss CE to optimize action prediction as a classification problem:

$$\mathcal{L}_{\text{BC}} = CE(\mathbf{a}_{\text{left}}^{(t)}, \hat{\mathbf{a}}_{\text{left}}^{(t)}) + CE(\mathbf{a}_{\text{right}}^{(t)}, \hat{\mathbf{a}}_{\text{right}}^{(t)}), \tag{9}$$

where $\hat{\mathbf{a}}_{\text{left}}^{(t)}$ and $\hat{\mathbf{a}}_{\text{right}}^{(t)}$ are ground-truth actions of the left and right manipulators from provided expert demonstrations, respectively. The overall objective for our ManiGaussian++ agent at each time step is written as:

$$\mathcal{L} = \mathcal{L}_{\text{BC}} + \lambda_{\text{ Recon}}\mathcal{L}_{\text{Recon}} + \lambda_{\text{ Task}}\mathcal{L}_{\text{Task}} + \lambda_{\text{Pred}}\mathcal{L}_{\text{Pred}}, \tag{10}$$

where the hyperparameters $\lambda_{\text{Recon}}, \lambda_{\text{Task}}, \lambda_{\text{Pred}}$ can be tuned to balance various objectives.

IV. EXPERIMENTS

In this section, we first introduce the simulated and real-world experiment setups. Then, we report the simulated performance of our method compared with the state-of-the-art approaches. We conduct a comprehensive ablation study to validate the proposed task-oriented Gaussian radiance field and the hierarchical Gaussian world model. We interpret the proposed techniques by visualization. Finally, we present qualitative results to depict the effectiveness of our ManiGaussian++ in real-world settings.

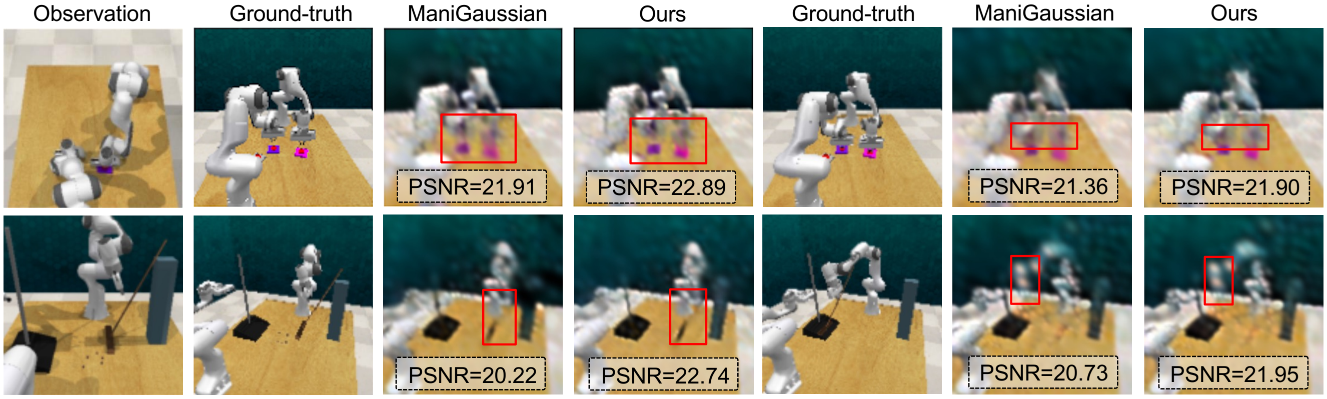
A. Experiment Setup

Simulation. For benchmarking, we conduct our simulation experiments on RLBench2 [15], a bimanual extension from the popular RLBench [25]. It contains 10 challenging languaged-conditioned manipulation tasks varying from different challenge levels. For agent observation, we employ RGB-D images from six cameras with a resolution of 256×256, in line with [15]. We use the same number of cameras as ManiGaussian [30] to provide multi-view supervision for fair comparisons. In the training phase, we provide 100 demonstrations for each task, which are generated by an Oracle scripted expert.

TABLE II

Comparison of Our Methods with Different Techniques. We manually categorize the 12 RLBench2 task to 3 groups for further interpretability, then we select one task from each category. For more details, please refer to the supplementary file.

Row ID	Gaussian Splatting	Task-oriented GS	Hierarchical GWM	`sweep to dustpan`	`handover item`	`push box`	Average ↑
1	✗	✗	✗	0	11	6	5.67
2	✓	✗	✗	24	12	24	20.00
3	✓	✓	✗	32	16	32	26.67
4	✓	✓	✓	**92**	**20**	**48**	**60.00**



(a) Front view (b) Novel view at current time step (c) Novel view at future time step

Fig. 3. **Novel View Synthesis Results.** Our ManiGaussian++ captures the multi-body spatiotemporal dynamics precisely, while ManiGaussian fails to model it. Note that we turn off the behavior cloning loss for better illustration.

Real Robot. The experimental setup consists of two Universal Robots UR5e arms equipped with Robotiq 2F-85 grippers, controlled via two Xbox controllers to collect demonstration data. Two RGB-D Realsense cameras capture 640 × 480 resolution images at 30 Hz. While multi-view cameras are utilized during the training phase, only a single camera is used during inference. We collect 30 real-world human demonstrations for training, while evaluating the trained policy for 10 episodes with a Nvidia RTX 4080 GPU.

Baselines. We compare our ManiGaussian++ with the state-for-the-art robotic bimanual manipulation methods, including PerAct^2 [15], which is a strengthened version of the widely-used single-arm agent PerAct [42]. Additionaly, we include the former version Manigaussian [30] by modifying the action dimension to be compatible with bimanual settings. The primary evaluation metric is the task success rate, which is calculated as the percentage of episodes in which the agent completes the task within a budget of 25 steps.

B. Comparisons with the State-of-the-Arts

We compare the proposed ManiGaussian++ with the state-of-the-art methods in the commonly-used bimanual manipulation benchmark RLBench^2 and report the performances. Our ManiGaussian++ achieves the best performance across 10 tasks ranging in different challenging levels, which demonstrates the superiority of the proposed techniques. Notably, by digesting the multi-body spatiotemporal dynamics in bimanual manipulation tasks, ManiGaussian++ outperforms its former version ManiGaussian by a sizable relative improvement of 89.36% (18.8% vs 35.6%). Even the enhanced version of PerAct^2 that leverages six cameras to ensure seamless observation is also defeated by the proposed ManiGaussian++, underling the importance of mining the consistency from the provided multi-view images and the multi-body spatiotemporal dynamics. The results prove the capacity of the proposed ManiGaussian++ to handle general robotic manipulation tasks.

C. Ablation Study

We propose task-oriented Gaussian Splatting to differentiate acting and stabilizing arms for multi-body spatiotemporal dynamics modeling, and hierarchical Gaussian world model for visual representation. Table II shows the effectiveness of each concept. We start with a vanilla PerAct^2 baseline (5.67% success rate), and then progressively include the proposed techniques. Using a Gaussian regressor to predict parameters improves performance by 14.33% (5.67% vs 20.00%). Task-oriented Gaussian Splatting advances the baseline by 21.00% (5.67% vs 26.67%), emphasizing the importance of distinguishing arm roles for better collaboration. Finally, the hierarchical Gaussian world model boosts the performance by 33.33% (26.67% vs 60.00%), which demonstrates the effectiveness of digesting the multi-body dynamics for bimanual manipulation.

D. Qualitative Analysis

Figure 3 shows the novel view image synthesis results. First, based on the front view observation where the gripper

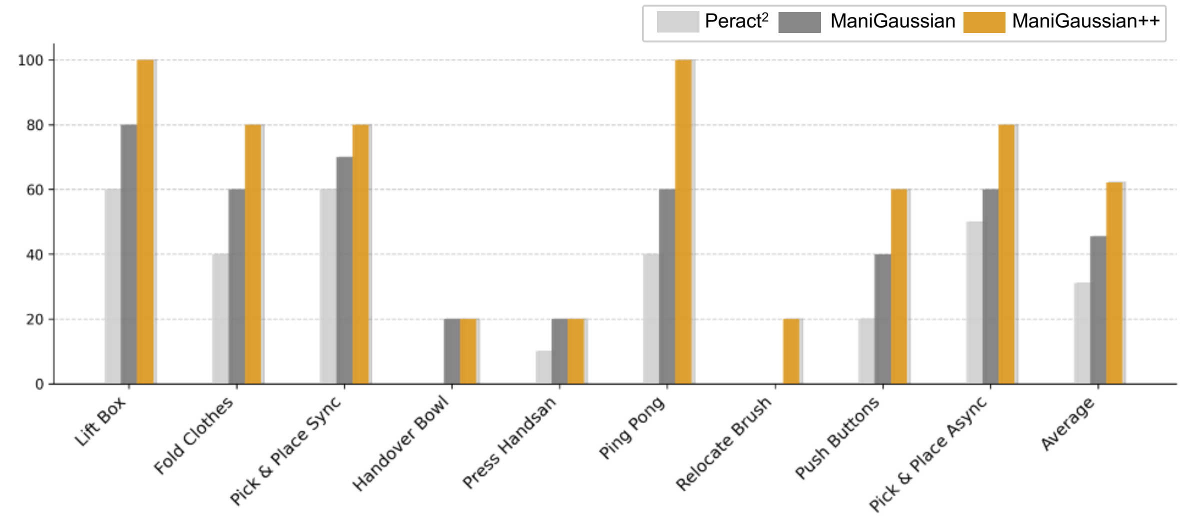
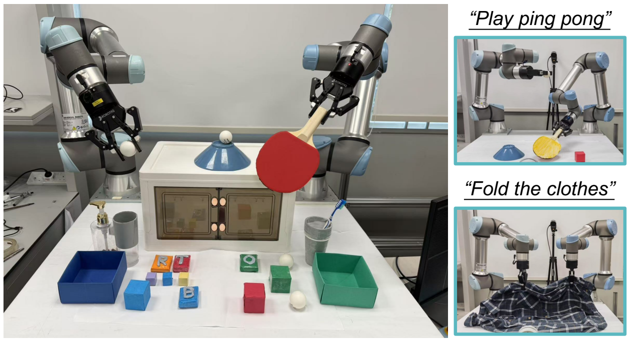


Fig. 4. **Real-world Results.** We train and evaluate Peract^2, ManiGaussian and ManiGaussian++ on 9 challenging real-world tasks.



(a) Experiment Setups (b) Complex Tasks

Fig. 5. **Real-World Experiments** with two UR5e manipulators.

shape cannot be seen, our ManiGaussian++ offers superior detail in modeling buttons and grippers in novel views. Second, our method accurately predicts future states based on the recovered details.

For example, in the top case of the `press buttons` task, our ManiGaussian++ is able to predict the current and the future gripper position. This qualitative result demonstrates that our ManiGaussian++ learns the intricate multi-body dynamics successfully. In the bottom case, for the task `sweep to dustpan`, ManiGaussian++ not only predicts the future position, but also predicts the future location of broom influenced by the gripper and the proper coordination of both robot arms. These cases illustrate the generation fidelity of the proposed hierarchical Gaussian world model.

E. Real-world Experiments

We validate Peract^2, ManiGaussian and ManiGaussian++ with 9 challenging real-robot tasks, which is depicted in Figure 4, our ManiGaussian++ outperform Peract^2 and ManiGaussian by a sizable relative improvement of 100% (31.11% vs 62.22%) and 36.57% (45.56% vs 62.22%), shows that our method is able to complete all 9 tasks simultaneously with only one model conditioned on natural human language from scratch, without any pertaining on the simulation or sim-to-real transferring. Notably, Figure 5 shows that ManiGaussian++ is able to complete complex tasks like `Play ping pong` and `Fold Clothes` that involve complex collaboration patterns, attributing to the proposed hierarchical Gaussian world model that encodes the multi-body spatiotemporal dynamics for the visual representation. Besides, ManiGaussian++ is robust to distractors like the lightning environment, which further validates the generalizability obtained by mining the multi-body spatiotemporal dynamics. Please refer to supplementary videos for more real-world qualitative results and details on the task setups.

V. Conclusion

In this paper, we have presented ManiGaussian++, a novel framework that addresses the challenges of multi-task bimanual manipulation through hierarchical Gaussian world modeling. Our approach extends the ManiGaussian framework by explicitly modeling multi-body spatiotemporal dynamics via a hierarchical Gaussian world model for dual-arm collaboration. Specifically, We use task-oriented Gaussian Splatting from visual features to differentiate acting and stabilizing arms for dynamics modeling. A hierarchical Gaussian world model employs a leader-follower architecture: the leader predicts deformation from the stabilizing arm, while the follower models the acting arm's effects. Through extensive experiments, ManiGaussian++ demonstrates significant improvements over state-of-the-art general bimanual manipulation methods, achieving an improvement of 20.2% in 10 simulated tasks and 60% success rate in 9 challenging real-world tasks. Limitations include the demand for calibrated multi-view cameras for supervisions, which increases the cost of real robot deployment.

References

[1] J. Abou-Chakra, K. Rana, F. Dayoub, and N. Sünderhauf, "Physically embodied gaussian splatting: Embedding physical priors into a visual 3d world model for robotics," in *CoRL*, 2023.

[2] J. F. Buhl, R. Grønhøj, J. K. Jørgensen, G. Mateus, D. Pinto, J. K. Sørensen, S. Bøgh, and D. Chrysostomou, "A dual-arm collaborative robot system for the smart factories of the future," *Procedia manufacturing*, vol. 38, pp. 333–340, 2019.

[3] S. Chen, R. Garcia, C. Schmid, and I. Laptev, "Polarnet: 3d point clouds for language-guided robotic manipulation," *arXiv preprint arXiv:2309.15596*, 2023.

[4] I. Chuang, A. Lee, D. Gao, and I. Soltani, "Active vision might be all you need: Exploring active vision in bimanual robotic manipulation," 2024.

[5] R. Ding, Y. Qin, J. Zhu, C. Jia, S. Yang, R. Yang, X. Qi, and X. Wang, "Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning," 2024.

[6] D. Driess, I. Schubert, P. Florence, Y. Li, and M. Toussaint, "Reinforcement learning with neural radiance fields," *NeurIPS*, 2022.

[7] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," *NeurIPS*, vol. 36, 2024.

[8] J. Gao, X. Jin, F. Krebs, N. Jaquier, and T. Asfour, "Bi-kvil: Keypoints-based visual imitation learning of bimanual manipulation tasks," 2024.

[9] Z. Gao, Y. Mu, R. Shen, C. Chen, Y. Ren, J. Chen, S. E. Li, P. Luo, and Y. Lu, "Enhance sample efficiency and robustness of end-to-end urban autonomous driving via semantic masked world model," *arXiv preprint arXiv:2210.04017*, 2022.

[10] K. F. Gbagbe, M. A. Cabrera, A. Alabbas, O. Alyunes, A. Lykov, and D. Tsetserukou, "Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations," 2024.

[11] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki, "Act3d: 3d feature field transformers for multi-task robotic manipulation," in *CoRL*, 2023, pp. 3949–3965.

[12] A. Goyal, V. Blukis, J. Xu, Y. Guo, Y.-W. Chao, and D. Fox, "Rvt2: Learning precise manipulation from few demonstrations," *RSS*, 2024.

[13] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation," *arXiv preprint arXiv:2306.14896*, 2023.

[14] J. Grannen, Y. Wu, B. Vu, and D. Sadigh, "Stabilize to act: Learning to coordinate for bimanual manipulation," in *CoRL*. PMLR, 2023, pp. 563–576.

[15] M. Grotz, M. Shridhar, T. Asfour, and D. Fox, "Peract2: Benchmarking and learning for robotic bimanual manipulation tasks," 2024.

[16] S. Ha, David and Jurgen, "Recurrent world models facilitate policy evolution," *NeurIPS*, vol. 31, 2018.

[17] D. Hafner, K.-H. Lee, I. Fischer, and P. Abbeel, "Deep hierarchical planning from pixels," *NeurIPS*, vol. 35, pp. 26 091–26 104, 2022.

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

[19] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering atari with discrete world models," *arXiv preprint arXiv:2010.02193*, 2020.

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

[21] N. Hansen, H. Su, and X. Wang, "Td-mpc2: Scalable, robust world models for continuous control," *arXiv preprint arXiv:2310.16828*, 2023.

[22] N. Hansen and X. Wang, "Generalization in reinforcement learning by soft data augmentation," in *ICRA*, 2021.

[23] Z. J. Hu, Z. Wang, Y. Huang, A. Sena, F. Rodriguez y Baena, and E. Burdet, "Towards human-robot collaborative surgery: Trajectory and strategy learning in bimanual peg transfer," *RA-L*, vol. 8, no. 8, pp. 4553–4560, 2023.

[24] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *ICML*, 2021.

[25] S. James, Z. Ma, D. Arrojo, David Rovick, and A. J, "Rlbench: The robot learning benchmark & learning environment," *RA-L*, 2020.

[26] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *TOG*, vol. 42, no. 4, 2023.

[27] J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, and A. Krieger, "Surgical robot transformer (srt): Imitation learning for surgical tasks," *arXiv preprint arXiv:2407.12998*, 2024.

[28] I.-C. A. Liu, S. He, D. Seita, and G. Sukhatme, "Voxact-b: Voxel-based acting and stabilizing policy for bimanual manipulation," 2024.

[29] H. Lou, Y. Liu, Y. Pan, Y. Geng, J. Chen, W. Ma, C. Li, L. Wang, H. Feng, L. Shi, *et al.*, "Robo-gs: A physics consistent spatial-temporal model for robotic arm with hybrid representation," *arXiv preprint arXiv:2408.14873*, 2024.

[30] G. Lu, S. Zhang, Z. Wang, C. Liu, J. Lu, and Y. Tang, "Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation," in *ECCV*. Springer, 2025, pp. 349–366.

[31] R. Mendonca, S. Bahl, and D. Pathak, "Structured world models from human videos," *arXiv preprint arXiv:2308.10901*, 2023.

[32] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *CACM*, vol. 65, no. 1, pp. 99–106, 2021.

[33] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv*, 2022.

[34] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. Gupta, "The unsurprising effectiveness of pre-trained vision models for control," in *ICML*, 2022.

[35] A. Quach, M. Chahine, A. Amini, R. Hasani, and D. Rus, "Gaussian splatting to real world flight navigation transfer with liquid networks," *arXiv preprint arXiv:2406.15149*, 2024.

[36] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *CoRL*, 2023.

[37] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[38] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, "Masked world models for visual control," in *CoRL*, 2023, pp. 1332–1344.

[39] Y. Seo, K. Lee, S. L. James, and P. Abbeel, "Reinforcement learning with action-free pre-training from videos," in *ICML*, 2022, pp. 19 561–19 579.

[40] D. Shim, S. Lee, and H. J. Kim, "Snerl: Semantic-aware neural radiance fields for reinforcement learning," *ICML*, 2023.

[41] O. Shorinwa, J. Tucker, A. Smith, A. Swann, T. Chen, R. Firoozi, M. D. Kennedy, and M. Schwager, "Splat-mover: multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting," in *CoRL*, 2024.

[42] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *CoRL*, 2023.

[43] X. Wang, Z. Zhu, G. Huang, X. Chen, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *arXiv preprint arXiv:2309.09777*, 2023.

[44] P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg, "Daydreamer: World models for physical robot learning," in *CoRL*, 2023, pp. 2226–2240.

[45] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang, "Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation," 2024.

[46] Y. Ze, N. Hansen, Y. Chen, M. Jain, and X. Wang, "Visual reinforcement learning with self-supervised 3d representations," *RA-L*, 2023.

[47] Y. Ze, G. Yan, Y.-H. Wu, A. Macaluso, Y. Ge, J. Ye, N. Hansen, L. E. Li, and X. Wang, "Gnfactor: Multi-task real robot learning with generalizable neural feature fields," in *CoRL*. PMLR, 2023, pp. 284–301.

[48] M. Zhang, K. Zhang, and Y. Li, "Dynamic 3d gaussian tracking for graph-based neural dynamics modeling," in *CoRL*, 2024.

[49] T. Zhang, D. Li, Y. Li, Z. Zeng, L. Zhao, L. Sun, Y. Chen, X. Wei, Y. Zhan, L. Li, and X. He, "Empowering embodied manipulation: A bimanual-mobile robot manipulation dataset for household tasks," 2024.

[50] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," 2023.

[51] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," in *ECCV*. Springer, 2025, pp. 55–72.

[52] H. Zhu, H. Yang, Y. Wang, J. Yang, L. Wang, and T. He, "Spa: 3d spatial-awareness enables effective embodied representation," *arXiv preprint arXiv:2410.08208*, 2024.