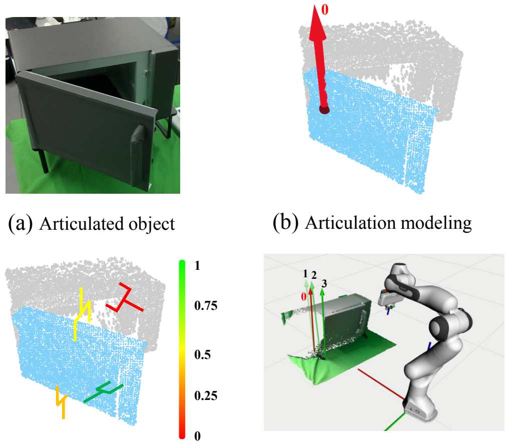
2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10610652

GAMMA: Generalizable Articulation Modeling and Manipulation for Articulated Objects

Qiaojun Yu[1], Junbo Wang[1], Wenhai Liu[1], Ce Hao[2], Liu Liu[3], Lin Shao[2], Weiming Wang[4] and Cewu Lu[1]*

***Abstract*— Articulated objects like cabinets and doors are widespread in daily life. However, directly manipulating 3D articulated objects is challenging because they have diverse geometrical shapes, semantic categories, and kinetic constraints. Prior works mostly focused on recognizing and manipulating articulated objects with specific joint types. They can either estimate the joint parameters or distinguish suitable grasp poses to facilitate trajectory planning. Although these approaches have succeeded in certain types of articulated objects, they lack generalizability to unseen objects, which significantly impedes their application in broader scenarios. In this paper, we propose a novel framework of Generalizable Articulation Modeling and Manipulating for Articulated Objects (GAMMA), which learns both articulation modeling and grasp pose affordance from diverse articulated objects with different categories. In addition, GAMMA adopts adaptive manipulation to iteratively reduce the modeling errors and enhance manipulation performance. We train GAMMA with the PartNet-Mobility dataset and evaluate with comprehensive experiments in SAPIEN simulation and real-world Franka robot. Results show that GAMMA significantly outperforms SOTA articulation modeling and manipulation algorithms in unseen and cross-category articulated objects. Images, videos and codes are published on the project website at: `sites.google.com/view/gamma-articulation`.**



(a) Articulated object

(b) Articulation modeling

(c) Grasp pose affordance

(d) Adaptive manipulation

Fig. 1. GAMMA framework in real microwave task. **(a)** Real microwave image to generate point clouds. **(b)** Articulated structure modeling, where blue points are segmented revolute joint and red arrow is the estimated but inaccurate joint axis and origin. **(c)** Grasp pose affordance evaluates the actionability and chooses ideal grasp poses. **(d)** Adaptive manipulation to iteratively update joint parameters. The initial joint axis(red) constantly gets closer to the ground truth after several iterations.

I. Introduction

Articulated structures like doors and drawers are commonly used in daily environments to store objects and divide spaces. They comprise interconnected parts governed by specialized joints, such as revolute and prismatic joints, which constrain the degrees of freedom according to their kinematic structure[1, 2]. When humans manipulate articulated objects, we can recognize the physical structure like a pivot with our prior knowledge and easily apply such skills to unseen circumstances. Likewise, artificial intelligence (AI)-empowered robots also show the potential to mimic humans' intuition and actively interact with articulated objects in awareness of kinetic constraints[3, 4].

Prior works focused on robot manipulation tasks by directly imitating end-to-end demonstrations [5, 6] or identifying joint parameters of specific instances or categories [2], however, they are less effective in the interaction and manipulation with novel articulated objects due to ignored geometrical and physical constraints [7]. Despite their success, building general-purpose robots to manipulate a diverse range of articulated objects in a large variety of environments in the physical world at the human level is extremely challenging. Therefore, recognizing and modeling the articulations are crucial for robots to generalize across-category objects with complex 3D geometries and kinematics.

In this paper, we propose a novel, adaptive learning framework **G**eneralizable **A**rticulation **M**odeling and **M**anipulation for **A**rticulated objects (GAMMA) that leverages articulation structures model and grasp pose affordance for generalizable cross-category manipulation as depicted in Fig. 1. Initially, GAMMA utilizes the point cloud to segment the whole object as several articulated parts and identify the physical structure and parameters of each part. Then, GAMMA proposes a series of potential grasp poses using GraspNet [8] and estimates the part-aware grasp pose affordance according to the identified articulation structure. Finally, GAMMA guides the robot manipulation by planning optimal trajectories based on the articulation model and grasp pose affordance. GAMMA constantly improves the articulation model by re-estimating the articulation parameters based on actual trajectories and updates optimal manipulation trajectories adaptively. In addition, GAMMA proficients at task

[1]Qiaojun Yu, Junbo Wang, Wenhai Liu, Cewu Lu are with Department of Computer Science, Shanghai Jiao Tong University, China. *Cewu Lu is the corresponding author. {`yqjllxs`, `sjtuwjb3589635689`, `sjtu-wenhai`, `lucewu`}`@sjtu.edu.cn`

[2]Ce Hao, Lin Shao are with Department of Computer Science, National University of Singapore, Singapore, `cehao@u.nus.edu`, `linshao@nus.edu.sg`

[3]Liu Liu is with Department of Computer Science and Information Engineering, Hefei University of Technology, China, `liuliu@hfut.edu.cn`

[4]Weiming Wang is with Department of Mechanical Engineering, Shanghai Jiao Tong University, China, `wangweiming@sjtu.edu.cn`

generalization by leveraging the learned physics-informed prior of articulated objects. When transferred to unseen tasks, GAMMA can quickly recognize the articulation structure, identify joint parameters and predict grasp pose affordance from the prior knowledge, which substantially improves the success rate in cross-category articulation-associated tasks.

In summary, the contributions of our paper are threefold: (1) We propose the Generalizable Articulation Modeling and Manipulation for Articulated Objects (GAMMA) algorithm, which can generalize manipulation with cross-category articulated objects by estimating the articulation model and grasp pose affordance from the point cloud. (2) We employ physics-guided adaptive manipulation to generate articulation-feasible trajectories and iteratively estimate joint parameters to constantly improve the modeling accuracy and enhance manipulation performance. (3) We conduct comprehensive experiments in both simulation and real-world. Results reveal that GAMMA has strong generalizability and achieves great success in improving modeling accuracy and manipulation success rate on unseen, cross-category articulated objects.

II. Related Work

Articulated Object Modeling. The development of large-scale datasets of articulated objects, such as Shape2motion [9], PartNet-Mobility [10], and AKBNet [11], significantly promoted advanced research in the part segmentation and modeling of articulation structure. Part segmentation [12–15] is the preliminary of articulated object modeling and part-level manipulation, which separates different part entities of the same object category at the mask level. Early stage methods [1, 2, 4] heuristically segment the well-explored articulated objects with a fixed number of parts. The following methods, RPM-Net [16] and Shape2Motion [9] leverage point-wise motion prediction to separate articulated parts from unknown objects. For articulation structure modeling, ANCSH [2] and ReArtNet [1] exploit densely normalized coordinate space to effectively model articulated objects with similar kinematic parameters. They achieved accurate per-part and per-joint pose estimation for unseen objects within the same category and similar kinematics but lacked generalizability for multi-tasks.

Affordance. In the manipulation tasks, affordance [17] indicates potential interaction modalities between the robot and objects. In particular, visual affordance utilizes visual information observed from objects and robots to predict the probability of successful execution of each contact pose. Recently, extensive research has focused on learning grasping affordance [18–22] and manipulation affordance [23–27] to facilitate robot-object interaction. For manipulation tasks with articulated objects, Adaafford [28] and Where2act [29] utilize dense affordance maps as actionable visual representations, indicating the success rate of manipulation at each point on the 3D articulated object. VAT-MART [30] proposes visual prior as representations to estimating actionable grasp pose considering geometrical constraints of articulated objects. However, the previous methods directly learn affordance estimation without knowledge of articulation structure and have difficulty generalizing across various categories of articulated objects.

3D Articulated Object Manipulation. The preliminary works of articulation manipulation focused on imitation learning [31–33] that leverages demonstrations from experts to learn a manipulation policy. However, imitation learning has distribution shift problems, and collecting diverse demonstrations is time-consuming and expensive. In recent years, visual recognition [34–36] has emerged to estimate instance-level or category-level articulation parameters to generate manipulation trajectories. For instance, VAT-Mart [30], a pure learning-based method, employs 3D visual affordance to predict the open-loop task-specific motion trajectory of each point. On the contrary, a series of optimal control methods [37, 38], which directly use dynamic articulation models as constraints, optimize the manipulation trajectory based on recognized joint type and parameters. However, inaccurate parameter estimation may also cause repeated failure.

III. Problem formulation

We formulate the robot manipulation task T as follows. An unknown articulated object M consists of K movable parts as $M = \{m_i\}_{i=1}^{K}$. We observe such object M via point cloud P with N points as $P = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{N}$. Further, we model the object structure J from the points cloud by estimating the articulation parameters ψ_i for each part as $J = \{\psi_i\}_{i=1}^{K}$. As most object only consist of one-dimensional prismatic and revolute joints [2, 4, 39], we can simplify the articulation parameters as $\psi_i = \{\mathbf{u}_i, \mathbf{q}_i, c_i\}$, where $\mathbf{u}_i \in \mathbb{R}^3$ is a unit vector of joint axis, $\mathbf{q}_i \in \mathbb{R}^3$ represent the origin and c_i is the joint type. In addition, we formulate two variables $\mathbf{o_i} \in \mathbb{R}^3$ and $\mathbf{v_i} \in \mathbb{R}^3$ representing the vectors from point $\mathbf{p_i}$ to the centroid of articulated and to the joint axis $\mathbf{d_i}$, which are used in the articulation modeling section (Sec. IV-A).

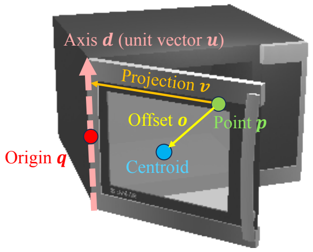

Axis $\boldsymbol{d}$ (unit vector $\boldsymbol{u}$)
Projection $\boldsymbol{v}$
Offset $\boldsymbol{o}$
Point $\boldsymbol{p}$
Origin $\boldsymbol{q}$
Centroid

Fig. 2. Illustration of points and vectors on articulated objects.

IV. Approach

We propose a novel robot manipulation algorithm, GAMMA that promotes generalizable manipulation with articulated objects. GAMMA consists of three main modules (Fig. 3): articulation modeling, part-aware grasp pose affordance and physics-guided adaptive manipulation. First, the articulation modeling (Sec. IV-A) leverages articulation parameters estimated from visual observation to accurately predict kinetic constraints and improve the manipulation policy. Then we employ part-ware grasp pose affordance (Sec. IV-B) to estimate suitable grasping poses to improve

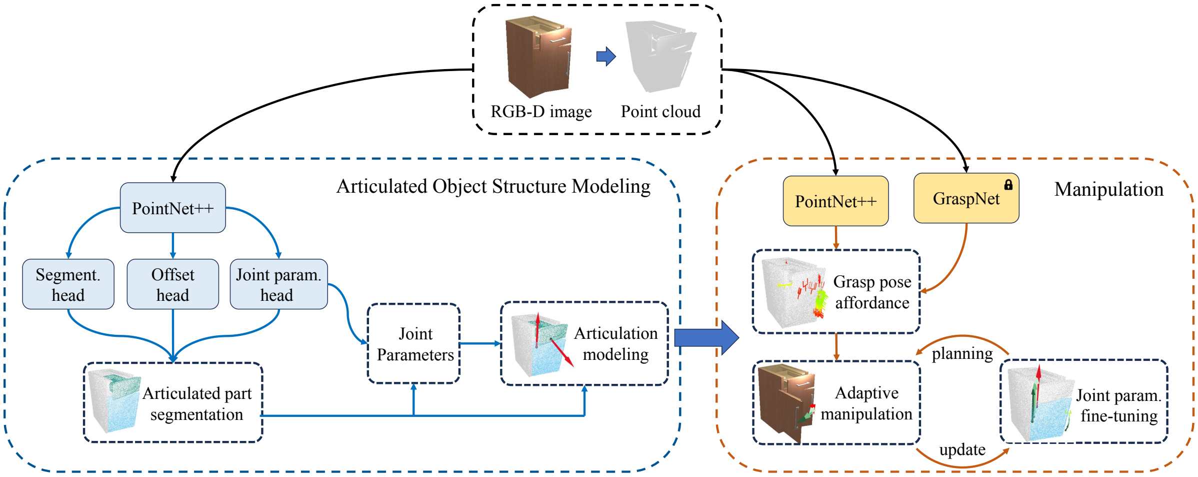


Fig. 3. **Pipeline of GAMMA.** We collect RGB-D images of articulated objects like a cabinet to generate point clouds. The articulation modeling block segments the articulated parts and estimates the joint parameters. The grasp pose affordance block estimates the actionability of each grasp pose and chooses the ideal ones. In the adaptive manipulation, the articulation model provides open-loop trajectory planning and we iteratively update the joint parameters with actual trajectory to improve modeling accuracy and grasping success rate.

grasping quality. Finally, the physics-guided adaptive manipulation (Sec. IV-C) incorporates prior knowledge of joint parameters to acquire optimal manipulation trajectories and dynamically updates articulation parameters from executed trajectories, which constantly enhances the modeling accuracy. In short, GAMMA aims to understand the physical structure of articulated objects to facilitate manipulation with generalized cross-category articulated objects.

A. Articulation Modeling

Understanding the structure of articulated objects helps infer the kinetic constraints of movable sections in the manipulation tasks. Therefore, we segment the whole object into several distinct rigid parts and estimate the articulation parameters for each part. We first extract point-wise features from a single partial point cloud observation $P = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^N$ with segmentation-style PointNet++ [40] backbone. Then we process the raw features with segmentation, offset and joint parameters heads.

The segmentation head predicts the per-point segmentation class $\hat{c}_i \in \{0, 1, 2\}$ corresponding to static, revolute and prismatic parts. The offset head generates offset vectors (Fig. 2) $\hat{\mathbf{o}}_i \in \mathbb{R}^3$ that shift each point towards the centroid of the corresponding part. The joint parameter head estimates the joint parameters by regressing every point $\mathbf{p}_i$ in the point cloud as the projection vector (Fig. 2) of each point to the joint axis $\hat{\mathbf{v}}_i \in \mathbb{R}^3$ and the estimated joint axis direction $\hat{\mathbf{d}}_i \in \mathbb{R}^3$ to the axis $\mathbf{u}$.

Finally, we jointly optimize the losses of segmentation, offset, and joint parameters as,

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N}\Big[\mathcal{L}_c(\hat{c}_i, c_i) + \mathcal{L}_o(\hat{\mathbf{o}}_i, \mathbf{o}_i) + \mathcal{L}_v(\hat{\mathbf{v}}_i, \mathbf{v}_i) + \mathcal{L}_d(\hat{\mathbf{d}}_i, \mathbf{d}_i)\Big], \tag{1}$$

where, c_i, $\mathbf{o}_i$, $\mathbf{v}_i$, and $\mathbf{d}_i$ are the ground-truth values of $\hat{c}_i$, $\hat{\mathbf{o}}_i$, $\hat{\mathbf{v}}_i$, and $\hat{\mathbf{d}}_i$. $\mathcal{L}_c(\hat{c}_i, c_i)$ denotes the focal loss to balance semantic distribution [41]; $\mathcal{L}_o(\hat{\mathbf{o}}_i, \mathbf{o}_i)$ and $\mathcal{L}_v(\hat{\mathbf{v}}_i, \mathbf{v}_i)$ impose constraints on both the L1 distance and direction of the point offsets; $\mathcal{L}_d(\hat{\mathbf{d}}_i, \mathbf{d}_i)$ optimize the axis direction estimation. In detail, offset loss is as

$$\mathcal{L}_o(\hat{\mathbf{o}}_i, \mathbf{o}_i) = \|\hat{\mathbf{o}}_i - \mathbf{o}_i\| - \Big(\frac{\mathbf{o}_i}{\|\mathbf{o}_i\|_2} \cdot \frac{\hat{\mathbf{o}}_i}{\|\hat{\mathbf{o}}_i\|_2}\Big) \tag{2}$$

By minimizing the loss function, we primarily model the articulated object with predicted segmentation $\hat{c}_i$, offset $\hat{\mathbf{o}}_i$, point projection $\hat{\mathbf{v}}_i$ and joint axis estimation $\hat{\mathbf{d}}_i$.

After completing the training phase, we freeze all parameters during feature inference. Then, we perform semantics-aware and axis-aware clustering to group the points into separate clusters. Specifically, the points of revolute or prismatic parts are first selected by the predicted semantics label. Then, they are shifted by offset vectors to form a more compact 3D distribution $\{(\mathbf{p}_i + \hat{\mathbf{o}}_i)\}$ where the intra-part points are spatially closer, and they are shifted by projection vector to form a more compact 3D distribution $\{(\mathbf{p}_i + \hat{\mathbf{v}}_i)\}$ where the intra-part points are spatially linear, involving the projection of points onto a common axis. Considering the density of feature sets $\{(\mathbf{p}_i + \hat{\mathbf{o}}_i), (\mathbf{p}_i + \hat{\mathbf{v}}_i)\}$, we adopt DBSCAN [42] to group these points into a part cluster as $M = \{m_i\}_{i=1}^K$. For each segmented part, we utilize the estimated per-point projection vector and axis direction to vote the joint parameters.

B. Part-Aware Grasp Pose Affordance

We extend the articulation model with grasp pose affordance, which interprets the high-dimensional visual information as practical interactive positions on the objects to facilitate robot manipulation. In a specific manipulation task T, we tackle each articulated part with the following steps.

First, we generate a series of grasp poses using GraspNet [8]. Secondly, we concatenate features of global geometrical feature f_{gg} from PointNet++ backbone [40], articulation parameter feature f_{ap}, and the contact pose feature f_{gp} from an MLP. Finally, we utilize the concatenated features to predict the actionability score $s_{g|P,\psi} \in [0, 1]$ to evaluate the feasibility of each grasp pose.

C. *Physics-Guided Adaptive Manipulation*

In the previous sections, we built the articulation model and generated grasp pose affordance from point cloud to provide kinetic constraint in open-loop planning. However, the uncertainty and errors in observation make the raw model inaccurate in the manipulation. To address this issue, we employ physics-guided adaptive manipulation to iteratively update the articulation model by minimizing the errors between planned and actual trajectories. This process helps improve model accuracy in diverse manipulation tasks with unseen articulation objects. In the following section, we detail the trajectory generation method for revolute and prismatic joints and parameter adjustment.

In the physics-guided trajectory planning, a *revolute joint* is defined by the joint axis $\mathbf{u}$ and origin $\mathbf{q}$. The trajectory of contact point $\mathbf{p}$ with respect to a rotation angle θ is

$$\mathcal{P} = \cos(\theta)\mathbf{I}\cdot\mathbf{p} + (1-\cos(\theta))\mathbf{u}\mathbf{u}^T\cdot\mathbf{p} + \sin(\theta)\mathbf{R}\cdot\mathbf{p} + \mathbf{q} \quad (3)$$

where $\mathbf{I}$ denotes an identity matrix and $\mathbf{R}$ denotes the skew-symmetric matrix of $\mathbf{u}$. Therefore, we can plan a feasible trajectory with rotation angle θ from the initial value to the target. Different from the revolute joint, the *prismatic joint* only translates along the joint axis $\mathbf{u}$. We can obtain the trajectory of contact points $\mathbf{p}$ with varying translation distance δ as,

$$\mathcal{P} = \mathbf{p} + \delta\mathbf{u}. \quad (4)$$

We plan the trajectory of grasp points in L steps in a manipulation task T with articulation parameters ψ at time step t as $\tau_t^{\text{plan}} = \{\mathcal{P}_i^{\text{plan}}\}_{i=1}^{L}$. In the actual manipulation, we adopt receding horizon control and only execute the first H $(H < L)$ steps of the planned τ_t^{plan}. As a result, the actual trajectory sampled from the real robot is $\tau_t^{\text{actual}} = \{\mathcal{P}_i^{\text{actual}}\}_{i=1}^{H}$. We finally optimize the articulation parameters ψ by minimizing the product of matching matrix $C_t \in \mathbb{R}^{H\times L}$ in Hungarian algorithm [43] and the errors between planned and actual trajectories as,

$$\mathcal{L}_t(\psi) = \frac{1}{H}\sum_{h=0}^{H}\sum_{l=0}^{L}\|\mathcal{P}_h^{\text{actual}} - \mathcal{P}_l^{\text{plan}}\|C_t[h,l]. \quad (5)$$

By iteratively planning trajectories and optimizing articulation parameters with actual trajectories sampled from the real world, we can constantly improve the modeling accuracy of the articulated object.

V. Experiments

In this section, we conduct comprehensive robot manipulation tasks with articulated objects in both simulated and real-world environments. We compare the performance of GAMMA with other baselines to answer the following question: 1) Can GAMMA effectively model articulated structure from point cloud by segmenting the articulation parts? 2) Can physics-guided adaptive manipulation based on grasp affordance improve the model accuracy and manipulation performance? 3) Can GAMMA effectively generalize skills learned from perception prior on unseen and cross-category articulation in both simulated and real-world environments?

A. *Experimental Setup*

Environments. We conduct simulation experiments in the SAPIEN simulator [10]. SAPIEN simulator provides physical simulation for robots, rigid bodies, and articulated objects, which also has photo-realistic rendering that facilitates sim-to-real generalization. We train GAMMA with 4 categories of objects: cabinet (revolute and prismatic joints), door (revolute joint), refrigerator (revolute joint), and microwave (revolute joint). We evaluate the performance and generalization ability of GAMMA in 3 unseen categories: safe (revolute joint), table (revolute and prismatic joints), and washing machine (revolute joint). In addition, we also set up real-world experiments with a 7-Dof Franka robot. We first mount an RGB-D camera RealSense L515 on the robot's wrist to sense point clouds of all articulated objects. Then we apply GAMMA to manipulate two unseen objects: cabinet (revolute and prismatic joints) and microwave (revolute joint).

Datasets. We train GAMMA with PartNet-Mobility dataset [29, 30], which has 562 different articulated objects in 7 categories. In detail, each articulated object contains one or more prismatic or revolute rigid parts. The initial joint states are uniformly distributed within the joint limit range. For observation, we used an RGB-D camera with 448×448 resolution to spherically sample camera viewpoints in front of the target object with yaw angle in $[-90°, 90°]$ and pitch angle in $[30°, 60°]$. In practice, we sample 20 different states for each object and 5 viewpoints for each state to generate point clouds. In total, we generate 32800, 9400, and 14000 samples for training, validation, and testing, respectively, using 328, 94, and 140 objects.

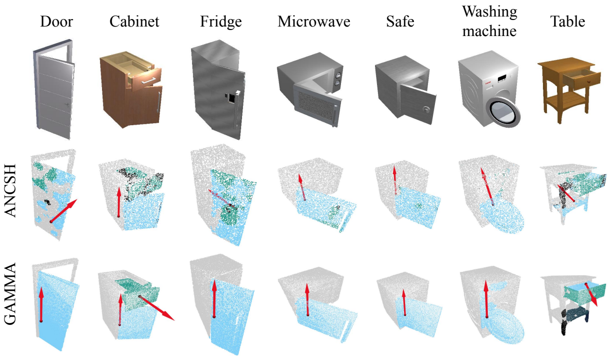


Fig. 4. We implement ANCSH and GAMMA to model articulated objects in 7 categories. The first rows are images in the simulation environment. The second and third rows are segmentation results of articulated parts, marked as blue, green and dark green points. Each color represents a separate modeled articulated part. The red arrow and dot denote the estimated joint axis direction and origin position.

Baselines of Articulation Modeling and Manipulation.

- ANCSH [2] is a state-of-the-art articulated object modeling method. It first segments the object as articulated parts with

TABLE I
ARTICULATION MODELING RESULT

	Category	AP75[2](%) ↑		Type Acc.[2](%) ↑		Axis error[2](°) ↓		Origin error[2](cm) ↓	
		ANCSH [2]	GAMMA	ANCSH	GAMMA	ANCSH	GAMMA	ANCSH	GAMMA
Unseen instances[1]	Cabinet	66.4	**80.9**	95.8	**98.9**	19.95	**9.48**	37.24	**15.63**
	Door	60.3	**97.1**	**100**	**100**	19.45	**2.27**	**4.90**	10.72
	Microwave	**100**	**100**	**100**	**100**	4.55	**0.87**	6.29	**2.62**
	Refrigerator	52.9	**98.3**	**100**	**100**	19.33	**6.99**	22.2	**1.69**
	Average	69.9	**94.1**	98.95	**99.7**	15.82	**4.90**	17.66	**7.66**
Unseen categories[1]	Safe	**100**	**100**	100	**100**	10.53	**3.75**	8.09	**1.36**
	Washing machine	81.8	**90.2**	**100**	**100**	24.71	**7.77**	6.24	**4.54**
	Table	16.1	**50.5**	56.9	**82.0**	45.25	**15.04**	34.57	**18.69**
	Average	66.0	**80.2**	85.6	**94.0**	26.83	**8.85**	16.3	**8.20**

[1] **Unseen instances** include data from four categories of objects used in the model training, which are used to validate and avoid overfitting. **Unseen categories** include data from three totally new categories, which are used to test the generalizability of models.
[2] Four evaluation metrics are **AP75**: average precision under IoU 0.75 of instance part segmentation; **Type Acc.**: joint type classification accuracy; **Axis error**: joint axis error; and **Origin error**: joint origin error.

single-view point clouds; then transforms the points into a normalized coordinate space to estimate joint parameters.

- Reinforcement learning (RL, TD3 [44]) is a widespread baseline of robot manipulation tasks. The observation includes point clouds and end-effector states, and the action is the incremental changes of the end-effector's state.
- Where2Act [29] selects grasping points with better actionability for manipulation and generates short-term manipulation actions (pushing or pulling) on each point. We test Where2Act in the long-term manipulation tasks by repeatedly executing selected actions multiple times to progressively manipulate the articulated parts.
- VAT-Mart [30] employs 3D object-centric actionable visual priors for manipulation tasks. This model predicts interaction-aware and task-aware visual action affordance and trajectory proposal for manipulation tasks.

B. Articulated Object Modeling Results

Tasks and Metrics. In the articulated object modeling task, we evaluate the performance of the articulation model from point clouds. We first train ANCSH [2] and GAMMA with 4 training categories in PartNet-Mobility dataset [29, 30] and evaluate in 3 unseen categories. Finally, we compare three modeling accuracy metrics: average precision under IoU 0.75 of instance part segmentation, joint type classification accuracy, joint axis error, and joint origin error.

Results. Fig. 4 and Table I presents the articulation modeling results of ANCSH and GAMMA. Results show that both ANCSH and GAMMA have high joint-type prediction accuracy in most categories. However, ANCSH cannot easily recognize tables because they are unseen and have extra stand part that interferes with the articulation structure. In addition, GAMMA significantly outperforms ANCSH in segmenting articulated parts. The reason is GAMMA not only abstracts visual representation from point cloud as ANCSH does, but also clusters points with the same semantics by shifting them along the axis direction and projecting them to a more compact and information-rich feature set. Therefore, GAMMA can easily segment the articulated parts since their features are more distinguishable in the transformed vector space.

In addition, although ANCSH has shown success in estimating joint parameters on seen articulated objects, it failed to estimate joint axis direction in the cross-category tasks and has larger origin errors than GAMMA in the unseen categories (Table I). In the experiments, we find a potential reason that ANCSH transforms all points to a normalized coordinate space to estimate joint parameters. This approach can work when all objects have similar sizes and structures, but in the cross-categories tasks, where the objects have diverse sizes and structures, the points normalized coordinate space might entwine heavily and lead to totally wrong axis estimation. On the contrary, GAMMA does not apply normalized coordinate space but utilizes all projected points and estimated axis direction to vote for the final joint parameters. This method leverages the geometrical structure of articulated objects and easily generalizes to unseen cross-category tasks.

C. Articulation Manipulation Results

Tasks and Metrics. We test 4 articulation manipulation tasks in Where2Act [29] and VAT-Mart [30]: pushing door, pulling door, pushing drawer and pulling drawer on both seen and unseen objects. Specifically, the pulling tasks are harder than pushing because they require accurate grasping of the handles or graspable edges. In addition, we especially emphasize the generalizability of manipulating cross-category articulated objects, so we extensively evaluate 4 tasks on 94 unseen objects with the same categories of training set and 140 objects in the unseen categories. Finally, we compare the success rate of 4 basic methods: RL(TD3) [44], Where2Act [29], VAT-Mart [30], GAMMA(ours) and two ablation methods: GAMMA without adaptive manipulation (w/o adpt.), GAMMA without grasp pose affordance (w/o afford.).

Results. We present the success rate of manipulation results in Table. II. RL(TD3) baseline learns from scratch and only achieves little success. Where2Act slightly improves the

TABLE II
ARTICULATED OBJECT MANIPULATION RESULTS

	Unseen instances[1]				Unseen categories[1]			
	door		drawer		door		drawer	
	pushing	pulling	pushing	pulling	pushing	pulling	pushing	pulling
RL(TD3) [44]	5.63	1.20	5.27	3.76	3.91	0.79	2.53	3.08
Where2Act [29]	31.59	7.53	30.58	9.30	34.52	5.04	22.69	8.61
VAT-Mart [30]	53.84	14.79	55.70	41.08	38.06	12.94	36.74	29.07
GAMMA(w/o afford.)[2]	51.79	35.05	54.46	54.55	59.26	53.66	46.20	35.62
GAMMA(w/o adpt.)[2]	69.62	41.62	61.29	42.86	**87.41**	58.15	63.49	38.10
GAMMA (ours)	**72.31**	**48.06**	**68.80**	**57.60**	86.86	**63.89**	**65.09**	**42.11**

[1] Unseen instances and unseen categories have the same definition in Table I. We perform experiments on each object with 5 trials and calculate the average **success rate** (%) in every task.

[2] GAMMA(w/o afford.) and GAMMA(w/o adapt.) are two ablation studies that remove grasp pose affordance and adaptive manipulation from the GAMMA framework.

success rate but it lacks long-horizon trajectory planning. VAT-Mart, on the other hand, substantially increases the success rate on all tasks by estimating visual action affordance and predicting trajectory proposals. As a comparison, we find GAMMA significantly outperforms all baselines and improves the success rate significantly which indicate that the cross-category performance of each module contributes to a significant improvement in the overall robotic manipulation.

In addition, we implement two ablation studies to analyze the influence of grasp pose affordance and adaptive manipulation. In the GAMMA without affordance (w/o afford) experiment, we randomly chose a grasp pose on the articulated object. The performance degrades to much lower success rates because the edges of articulated objects are not actionable. In the GAMMA without adaptive manipulation (w/o adapt.) experiment, we only apply open-loop trajectory planning to manipulate the articulated objects. The ablation causes obvious decreases in success rate in unseen instances and moderate or no decreases in the unseen categories. These two ablations reveal that grasp pose affordance and adaptive manipulation crucially contribute to the outstanding performance of the GAMMA algorithm.

D. Real-World Experiments

We apply our method to real-world objects to verify its generalization ability. In the real world, we set up 20 different camera viewpoints, five states for each object, and generated 100 point clouds for each object, totaling 200 point clouds. To reduce the gap between simulated and real point cloud data, we randomly select 10% to 30% of the points in the simulated point cloud and add Gaussian noise with a mean of 0 and a standard deviation of 0.03.

First, we use the model trained on simulated data to predict real-world point clouds for both the microwave and cabinet. The microwave has an average axis error of $6.14°$ and an origin error of $7.36cm$. For the cabinet, the door has an average axis error of $5.33°$and an origin error of $7.7cm$, while the drawer has an axis error of $6.21°$. More results can be found on supplementary materials. The results demonstrate our proposed method can generalize to real-world point clouds well. Then we estimate the grasp pose affordance and apply adaptive manipulation on three robotic manipulation tasks of pulling the door and drawer on a cabinet and the door on a microwave. The effectiveness of overall framework has been verified through real-world robotic manipulation as shown in Fig. 5 and supplementary video.

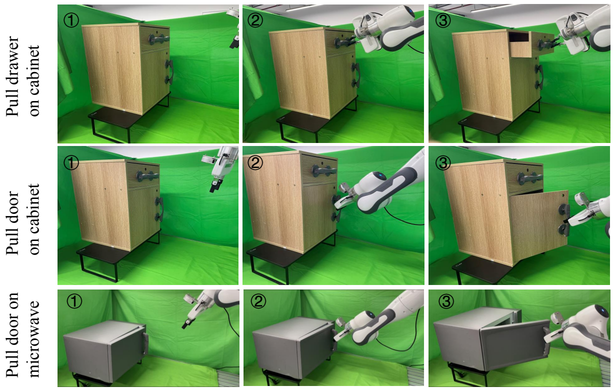


Fig. 5. We implement GAMMA in the real-world experiments. Manipulation tasks include pulling the drawer and door on a cabinet and pulling the door on a microwave.

VI. CONCLUSION

In this paper, we propose the generalized articulation modeling and manipulation (GAMMA) framework that estimates articulation parameters and grasp pose affordance from point clouds. In addition, GAMMA utilizes actual trajectory to iteratively update articulation parameters to improve manipulation performance. Experiments show that GAMMA significantly outperforms baselines in both articulated object modeling and manipulation, and has outstanding generalizability in cross-category tasks.

VII. ACKONWLEDGMENTS

This work was supported by the National Key Research and Development Project of China (No.2022ZD0160102, No.2021ZD0110704), Shanghai Artificial Intelligence Laboratory, XPLORER PRIZE grants, and National Natural Science Foundation of China (62302143, 52305030).

REFERENCES

[1] L. Liu, H. Xue, W. Xu, H. Fu, and C. Lu, "Toward real-world category-level articulation pose estimation," *IEEE Transactions on Image Processing*, vol. 31, pp. 1072–1083, 2022.

[2] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3706–3715.

[3] S. Y. Gadre, K. Ehsani, and S. Song, "Act the part: Learning interaction strategies for articulated object part discovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 752–15 761.

[4] Z. Jiang, C.-C. Hsu, and Y. Zhu, "Ditto: Building digital twins of articulated objects from interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5616–5626.

[5] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," in *The Eleventh International Conference on Learning Representations*, 2022.

[6] H. Shen, W. Wan, and H. Wang, "Learning category-level generalizable object manipulation policy via generative adversarial self-imitation learning from demonstrations," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 166–11 173, 2022.

[7] L. P. Kaelbling, "The foundation of efficient robot learning," *Science*, vol. 369, no. 6506, pp. 915–916, 2020.

[8] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[9] X. Wang, B. Zhou, Y. Shi, X. Chen, Q. Zhao, and K. Xu, "Shape2motion: Joint analysis of motion parts and attributes from 3d shapes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8876–8884.

[10] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 097–11 107.

[11] L. Liu, W. Xu, H. Fu, S. Qian, Q. Yu, Y. Han, and C. Lu, "Akb-48: a real-world articulated object knowledge base," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 809–14 818.

[12] E. Kalogerakis, M. Averkiou, S. Maji, and S. Chaudhuri, "3d shape segmentation with projective convolutional networks," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3779–3788.

[13] L. Yi, H. Su, X. Guo, and L. J. Guibas, "Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2282–2290.

[14] X. Lai, J. Liu, L. Jiang, L. Wang, H. Zhao, S. Liu, X. Qi, and J. Jia, "Stratified transformer for 3d point cloud segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8500–8509.

[15] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. Wang, "Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7081–7091.

[16] Z. Yan, R. Hu, X. Yan, L. Chen, O. Van Kaick, H. Zhang, and H. Huang, "Rpm-net: recurrent prediction of motion and parts from point cloud," *arXiv preprint arXiv:2006.14865*, 2020.

[17] J. J. Gibson, "The theory of affordances," *Hilldale, USA*, vol. 1, no. 2, pp. 67–82, 1977.

[18] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 1316–1322.

[19] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su, "S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes," in *Conference on robot learning*. PMLR, 2020, pp. 53–65.

[20] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies between affordance and geometry: 6-dof grasp detection via implicit representations," *arXiv preprint arXiv:2104.01542*, 2021.

[21] M. Kokic, D. Kragic, and J. Bohg, "Learning task-oriented grasping from human activity datasets," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3352–3359, 2020.

[22] P. Mandikal and K. Grauman, "Learning dexterous grasping with object-centric visual affordances," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 6169–6176.

[23] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim, "Demo2vec: Reasoning object affordances from online videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2139–2147.

[24] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.

[25] T. Nagarajan and K. Grauman, "Learning affordance landscapes for interaction exploration in 3d environments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2005–2015, 2020.

[26] T. Nagarajan, Y. Li, C. Feichtenhofer, and K. Grauman, "Ego-topo: Environment affordances from egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 163–172.

[27] D. Xu, A. Mandlekar, R. Martín-Martín, Y. Zhu, S. Savarese, and L. Fei-Fei, "Deep affordance foresight: Planning through what can be done in the future," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6206–6213.

[28] Y. Wang, R. Wu, K. Mo, J. Ke, Q. Fan, L. J. Guibas, and H. Dong, "Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*. Springer, 2022, pp. 90–107.

[29] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6813–6823.

[30] R. Wu, Y. Zhao, K. Mo, Z. Guo, Y. Wang, T. Wu, Q. Fan, X. Chen, L. Guibas, and H. Dong, "Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects," *arXiv preprint arXiv:2106.14440*, 2021.

[31] J. Wong, A. Tung, A. Kurenkov, A. Mandlekar, L. Fei-Fei, S. Savarese, and R. Martín-Martín, "Error-aware imitation learning from teleoperation data for mobile manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 1367–1378.

[32] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," *arXiv preprint arXiv:2302.04659*, 2023.

[33] P.-L. Guhur, S. Chen, R. G. Pinel, M. Tapaswi, I. Laptev, and C. Schmid, "Instruction-driven history-aware policies for robotic manipulations," in *Conference on Robot Learning*.

PMLR, 2023, pp. 175–187.

[34] B. Abbatematteo, S. Tellex, and G. Konidaris, “Learning to generalize kinematic models to novel objects,” in *Proceedings of the 3rd Conference on Robot Learning*, 2019.

[35] V. Zeng, T. E. Lee, J. Liang, and O. Kroemer, “Visual identification of articulated object parts,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2443–2450.

[36] A. Jain, R. Lioutikov, C. Chuck, and S. Niekum, “Screwnet: Category-independent articulation model estimation from depth images using screw theory,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 670–13 677.

[37] M. Mittal, D. Hoeller, F. Farshidian, M. Hutter, and A. Garg, “Articulated object interaction in unknown scenes with whole-body mobile manipulation,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1647–1654.

[38] Y. Karayiannidis, C. Smith, F. E. V. Barrientos, P. Ögren, and D. Kragic, “An adaptive control approach for opening doors and drawers under uncertainties,” *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 161–175, 2016.

[39] H. Zhang, B. Eisner, and D. Held, “Flowbot++: Learning generalized articulated objects manipulation via articulation projection,” *arXiv preprint arXiv:2306.12893*, 2023.

[40] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in neural information processing systems*, vol. 30, 2017.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[42] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[43] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[44] S. Fujimoto, H. Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.