Physically Ground Commonsense Knowledge for Articulated Object Manipulation with Analytic Concepts

Supplementary Material

Table of Contents

This is the supplementary material of the submission. In the supplementary material, we provide comprehensive details for better understanding of our main paper and offer more evidence to prove the effectiveness of our approach. The supplementary material is organized as follows:

Sec. A: Details and discussions about the **neural network structure** in our approach.

Sec. B: Details of our main experiments, including **experiment settings**, **experiment results**, and the **analysis of failure cases**.

Sec. C-G: **Ablation studies and analysis** about the **effectiveness of modules and design** in our approach.

Sec. H: Discussions about the **scalability of analytic concepts** on expanding to novel object categories and manual construction.

Sec. I: Analysis of **real-world experiments**, the discussions about the **sim2real gap**, and **robustness** of our approach in real world.

Sec. J: Discussions about the **generalization capability** of our approach.

Sec. K: Discussions about **current limitations and the future work**.

Sec. L-N: **Computer resources** , discussions about **societal impacts**, and statement on usage of LLMs.

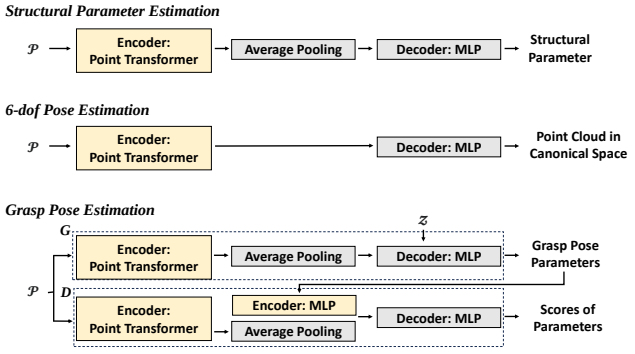

Figure 5. Network structure regarding Structural Parameters Estimation, 6-DoF Pose Estimation and Grasp Pose Estimation. $\mathcal{P}$ denotes the input point cloud, and G, D denote the generator and discriminator respectively.

A. Neural Network Structure

In this section, we show the structure of the neural networks used in **Structural Parameter Estimation**, **6-DoF Pose Estimation** and **Grasp Pose Estimation**.

As shown in Fig. 5, the structural parameter estimation network takes the point cloud of the detected part with 2,048 points as input. Then it utilizes a Point-Transformer [46] as encoder, which extracts 128 groups of points with size 32 from the input point cloud and has 12 6-headed attention layers. Subsequently, an average pooling layer is introduced to extract the global feature of the entire point cloud. Then, an MLP with three linear layers and accompanied ReLU activation outputs the structural parameters.

For the 6-DoF pose estimation network, the Point-Transformer encoder is directly linked to an MLP that has three linear layers with ReLU activation, which outputs a point cloud of the detected part in canonical space.

The grasp pose estimation network builds on the concept of conditional GAN [29]. The generator G employs Point-Transformer as the point cloud encoder, uses an average pooling layer to aggregate features and generates plausible grasp pose parameters from Gaussian noise z using an MLP decoder. The discriminator D obtains the point cloud feature using a Point-Transformer encoder along with an average pooling layer and encodes the generated grasp pose parameters with MLPs. The two features are then concatenated and passed through another set of MLPs to compute scores for the generated grasp pose parameters.

As our main purpose is to provide a baseline to demonstrate the feasibility and effectiveness of using analytic concepts to physically ground semantic-level commonsense knowledge, we do not delve into complex network designs.

B. Details of Experiments

B.1. Training Data Preparation

To prepare data for training the models in our approach, we first provide analytic concept annotations for real objects. Particularly, we annotate the concept parameters for training objects from PartNet-Mobility [44]. Then, we place each object in the SAPIEN [44] simulator and use a camera to capture the RGB image and depth map of this object. According to the URDF and analytic concept annotation of an object, we can easily obtain the ground-truth bounding box, point cloud, 6-DoF pose and structural parameters of each actionable part. To collect positive and negative samples for grasp pose parameters, we perform extensive random sampling for the parameters within its domain, then calculate the grasp pose based on the sampled parameters and test it in the simulator. If the grasp succeeds, the parameter is considered a positive sample; otherwise, it is a negative sample.

Table 6. Detailed statistics of the data split for manipulation tasks. "–" indicates that the category is unseen during training.

Num of Obj	Training Categories										Testing Categories				
	Box	Door	Faucet	Kettle	Microwave	Fridge	Storage	Switch	TrashCan	Window	Bucket	KitchenPot	Safe	Table	Washing
Train Obj	20	23	65	22	9	32	270	53	52	40	-	-	-	-	-
Test Obj	8	12	19	7	3	11	75	17	17	18	36	23	29	95	16

Table 7. Evaluation results of manipulation experiments on system limitations. All values represent the average success rate as a percentage.

Module	Categories										
	Box	Dor	Fct	Fdr	Ket	Mcw	Stf	Swt	Tcn	Win	**AVG**
None	15.9	58.4	28.0	46.7	24.3	52.2	58.6	32.7	22.9	22.3	42.5
Actionable Part Segmentation	19.8	62.5	33.8	53.0	26.1	54.9	67.2	40.9	32.9	28.9	49.8
Concept Identification	26.1	64.4	35.0	55.4	30.8	62.9	67.8	42.2	35.1	29.8	51.2
Structural Param Estimation	61.6	81.9	69.6	74.2	64.4	75.3	78.2	65.3	66.1	64.5	72.0
6-DoF Pose Estimation	92.2	87.4	81.2	85.8	83.4	91.1	89.3	85.4	81.6	84.2	86.3
Grasp Pose Estimation	95.6	90.2	92.5	90.6	92.2	98.5	95.3	93.5	91.8	93.0	93.6
Force Direction	99.0	99.5	98.2	99.0	98.8	99.0	98.4	96.2	99.2	99.4	98.6

B.2. Detailed Manipulation Experiment Settings

Data Statistics. In Sec. 5.1.1 of the main paper, we introduced our data settings for manipulation experiments in simulation environment. Here we provide more details as shown in Tab. 6. We have collected 15 categories of objects from PartNet-Mobility [44], after removing objects that are either too small (*e.g.*, Pen, USB), requiring multi-gripper collaboration (*e.g.*, Plier, Scissors), or not making sense for robot gripper to manipulate (*e.g.*, Fan, Clock). The 15 categories are further divided into two groups, 10 categories for training and the rest 5 categories for testing. The 10 categories involved in training consist of 773 objects, which are further divided into 586 training objects and 187 testing objects. The testing categories consist of 199 objects.

Evaluation Settings. As discussed in Sec. 5.1.1 of the main paper, we adopt the commonly used metric, success rate, for evaluating manipulation tasks. Following Where2Act [30], a manipulation is considered successful if the target joint on an articulated object moves more than 0.01 unit length or 0.5 relative to its maximum motion range. We use a 5-interaction budget. The success rate is computed as the ratio of successful manipulations to the total number of manipulation trials. 22 analytic concepts are used in our experiments.

We compare our method with five baselines. **1) Where2Act [30]** predicts per-pixel action likelihoods and selects the point with highest score. We follow Where2Act to generate 100 action proposals at that point and select the one with highest score for manipualation. The action proposals are generally categorized into pushing and pulling. A task is considered successful if either type of action succeeds. The training data are collected by placing articulated objects in SAPIEN [44] and interacting with them using a two-finger gripper. **2) Where2Explore [32]** also predicts actions and proposals on point level, which employs a few-shot framework to enhance its generalization to novel object categories. We use the same training data, test data, and evaluation metric as Where2Act. **3) GAPartNet [8]** predicts the 6-DoF pose of the target part and defines interaction policies based on the predicted pose. We collect training data according to the GAParts definition in GAPartNet. **4) ManipLLM [23]** employs an MLLM to predict the contact point on a given RGBD featuring the target object, guided by a language prompt describing the manipulation task. For fair comparison with other approaches, we use a two-finger gripper as the end-effector instead of the originally used suction. **5) A3VLM [14]** is the current state-of-the-art method in MLLM-driven affordance learning. It employs an MLLM to predict the 6-DoF pose and motion axis of the target part and interacts with it based on the predicted information.

For fair comparison, we use a two-finger gripper as the end-effector instead of the original suction, and further adapt ManipLLM [23] and A3VLM [14] for gripper-based manipulation. For ManipLLM [23], we collect gripper-compatible affordance maps and conduct hyper-parameter search during LLaMa-Adapter fine-tuning and Test-time Adaptation to get the optimal performance. For A3VLM [14], it is designed to be end-effector agnostic, therefore no additional tuning of the model is needed. Instead, we propose an independent generic grasp-pose proposer according to A3VLM [14] to produce gripper-compatible grasp poses.

B.3. Analysis of Failure Cases

In most failure cases, we observe that the collision volume of the end-effector is a critical factor. For example, a cylindrical handle attached to a door may allow valid grasps from

Table 8. The first row shows the average accuracy of concept identification by the MLLM in percentage. The second row shows the average distance from the point cloud of the detected part to the mesh rendered from the estimated analytic concept parameters in millimeters.

Evaluation	Training Categories											Testing Categories					
	Box	Dor	Fct	Fdr	Ket	Mcw	Stf	Swt	Tcn	Win	**AVG**	Bkt	Pot	Saf	Tab	Wsm	**AVG**
Accuracy	97.5	95.0	91.7	82.6	96.4	92.9	96.7	94.8	94.8	90.0	94.2	97.6	88.0	95.9	97.8	88.9	95.6
Distance	9.24	3.66	1.92	1.85	2.65	5.67	5.42	1.84	9.80	8.62	5.18	9.73	3.41	9.75	8.65	12.4	6.55

various directions around its axis. However, in some orientations, the gripper may be obstructed by nearby parts when approaching the handle, leading to failed manipulation. In future work, we will take different types and brands of end-effectors into consideration to design more effective action policies for handling such tasks.

C. Detailed Results of System Error Breakdown

In Sec. 5.3, we analyzed the limitations of our proposed system and provided preliminary experimental results on the effect of introducing ground truth to each module. Here, we present detailed results of the system error breakdown to further illustrate the impact of incorporating ground truth across different categories of object manipulation tasks. Because ground truth is required for incorporation, the analysis is only conducted on training categories. The results are shown in Tab. 7.

D. Analysis of Concept Identification and Parameter Estimation

As discussed in Sec. 5.3 of the main paper, structural parameter estimation remains the major bottleneck in overall performance. In addition, concept identification also affects parameter estimation, since parameters align with the identified concept. To further analyze these modules, we provide a detailed quantitative evaluation of both concept identification and parameter estimation, with results shown in Tab. 8.

Concept Identification. In the concept identification stage, we provide the MLLM with multiple analytic concept identities along with their corresponding synopses in natural language, prompting it to select the concept that best matches the target part for manipulation. The accuracy of the concept identification is measured by comparing its selections to those made by humans. Accurate identification occurs when the concept chosen by the MLLM matches the human-chosen one. The results demonstrate that our concept identity as well as the corresponding synopsis are well-defined to be accurately understood by MLLMs.

The high concept identification accuracy (94.2%/95.6%) also indicates that the MLLM practically demonstrates few errors or hallucinations when reasoning about the synopses of analytic concepts. In addition, analytic concepts demonstrate strong adaptability by adjusting their parameters to represent diverse object structures, implying that our method is able to tolerate certain concept identification inaccuracies. Specifically, when concept identification exhibits certain inaccuracies due to MLLM errors or hallucinations, subsequent parameter estimators may still yield parameters that reasonably represent the target structure and lead to effective physical affordances. This indicates the robustness of our method against MLLM errors or hallucinations.

Parameter Estimation. The evaluation metric for parameter estimation is the Point2Face distance, measured in millimeters. This metric computes the average distance from the points in a point cloud to their closest counterparts on a corresponding mesh. In this context, the point cloud represents the detected part, and the mesh is rendered by the analytic concepts. As shown in Tab. 8, all distances are notably small, with an average of 5.18 mm across train categories and 6.55 mm across test categories. These results underscore the accuracy of our parameter estimator in capturing both the structural and pose parameters of the objects.

E. Further Analysis of Grasp Pose Knowledge

In Sec. 4.3 of the main paper, we have introduced grasp pose as an important type of analytic manipulation knowledge. Here, we qualitatively analyze the effectiveness of such knowledge.

To intuitively illustrate the difference between *estimated* and *sampled* grasp pose parameters, we provide examples of the manipulation knowledge "grasp around" applied to the analytic concept *"curved_handle"*. As shown in Fig. 6-Left, the knowledge function is parameterized by two factors: 1) ω, which defines the angle of the gripper turning along the handle, sampled within $(-\theta/2, \theta/2)$, where θ is the maximum allowable angle; 2) φ, which defines the angle of the gripper rotating around the handle, sampled within $(-\pi, \pi)$. In Fig. 6-Right, we provide ten examples of specific implementation of "grasp around" parameterized differently, and categorize them into four types. Grasp poses with blue background are physically plausible and are suitable to complete the task due to proper position and ro-

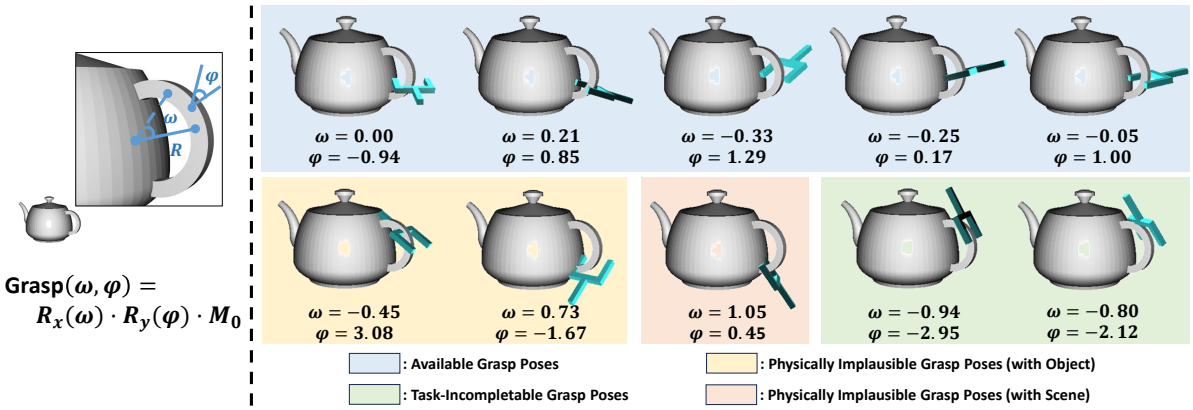

Figure 6. Examples of grasp poses for task *"pour the kettle"*. $\mathbf{M_0}$ denotes to the default pose of the gripper for grasping the handle, *i.e.* the grasp pose parameterized by $\omega = 0$ and $\varphi = 0$. $\mathbf{R_x}$ and $\mathbf{R_y}$ respectively represent rotations around the x-axis and the y-axis.

Table 9. Evaluation results of effectiveness of synopses in prompting.

	Concept Identification Accuracy	Overall Performance	Synopsis Example
Current Implementation	*94.2*	*42.5*	*A handle with a lever and an axis, perpendicular to each other, forming an L-shape. Can be turned around the axis, typically found on doors.*
More Descriptive	95.0	42.7	A handle consisting of a lever and a rotational axis arranged perpendicularly to form an L-shape. It allows turning around the axis and is commonly used for opening or closing doors.
Less descriptive (sufficient knowledge)	93.5	42.2	A handle with a lever and an axis perpendicular to each other, forming an L-shape.
Less descriptive (insufficient knowledge)	52.8	22.5	A type of handle forming an L-shape for turning.

tation. Grasp poses with green background are also physically plausible, but fail to hold the handle effectively. Grasp poses with yellow and red background are physically implausible, respectively colliding with the kettle itself and the tabletop (not shown in the figure). The grasp poses instantiated from *sampled* parameters could represent all possible candidates here, while only the parameters corresponding to the available grasp poses can be *estimated* by the grasp pose generator, which improves the effectiveness of grasping.

F. Analysis of Concept Synopses in Prompting

Here we provide an analysis on the effectiveness of synopses in prompting the MLLM to identify the concepts. We modify the synopses of analytic concepts to construct three controlled variants: 1) more descriptive synopses, 2) less descriptive synopses with sufficient semantic-level knowledge, and 3) less descriptive synopses lacking sufficient semantic-level knowledge. Specifically, we keep using GPT-4o [18] as the MLLM and remain other modules the same as our current implementation. Tab. 9 shows the performance using synopses with variations of descriptiveness, with "L_Handle" as an example. The results indicate that the concept identification accuracy and overall task performance remain at peak performance and are barely affected as long as the synopses provide sufficient knowledge, while synopses with insufficient knowledge result in significant performance degradation.

G. Distillation of Analytic Concepts

We have so far created 153 analytic concepts, while 22 of them are used for manipulation tasks. Specifically, we iteratively distill redundant analytic concepts by analyzing their impact on manipulation success rates, retaining only those with effective contributions.

We initially select 15 object categories and invite volunteers to create analytic concepts to represent objects in these categories, along with associated manipulation knowledge. After cleaning and filtering, 153 analytic concepts are retained, which sufficiently cover all 15 categories. We then conduct experiments using these 153 concepts, as shown in **Set 1** of Tab. 10. However, many of these concepts are never identified by the MLLM during manipulation (*e.g.*, *cuboidal_legs*, *cylindrical_body*). By removing unused concepts, we obtain **Set 2** with 59 concepts and observe

Table 10. Experiment results using different set of analytic concepts. Set 4 refers to the set of analytic concepts used in our main experiement.

	Num of Concepts	Train Cat Results	Test Cat Results
Set 1	153	42.0	40.1
Set 2	59	42.2	39.8
Set 3	27	42.8	40.3
Set 4	*22*	*42.5*	*40.8*
Set 5	17	35.4	29.8
Set 6	7	10.3	6.4

close success rates, since the excluded concepts are not identified and thus causing no huge impact on performance. We further distill the 59 concepts to obtain **Set 3** through i) retaining only one concept for concepts representing highly similar parts (*e.g.*, keeping *curved_handle* and removing *curved_tophandle*), and ii) removing concepts that can be composed by others (*e.g.*, *symmetric_windows* can be constructed with multiple *regular_window*s). This progress has no significant impact on manipulation performance, as the knowledge of the removed concepts can still be captured by the remaining ones. In addition, we obtain **Set 4** with 22 concepts, *i.e.* the set we use for main experiments, by removing each of the 27 concepts individually, evaluating performance with the remaining concepts, and retaining only those whose removal causes significant performance drop. As a comparison, **Set 5** and **Set 6** respectively exclude the concepts representing lids/doors and handles/switches, showing significant drops in success rate, indicating that these concepts are crucial for completing manipulation tasks.

H. Scalability & Generalizability of Analytic Concepts

Analytic Concepts' General Coverage on Real-world Objects. As discussed in Sec. G, we use 22 analytic concepts to cover the actionable parts of 15 object types that are suitable for single-gripper manipulation in our experiments. We further analyze the number of concepts to cover the actionable parts of all 46 object categories in PartNet-Mobility [44], an articulated object dataset containing real-world manufacturers' products. By randomly sampling 20, 30, and 40 categories ten times, the average numbers of required concepts are 27, 34, and 37 respectively. Covering the actionable parts of all 46 categories requires 39 concepts. This shows a diminishing marginal increase in required concepts as the object type grows. This stems from the fact that analytic concepts describe commonsense structural and manipulation knowledge of objects, which is reusable across different object categories. For example, the "L_Handle" concept can be used in object categories including door, storage furniture, faucet, window, *etc.*

Scalability of Analytic Concepts to Novel Objects and Categories. As mentioned above, the strong cross-category generalizability enables analytic concepts to easily scale to novel object types. Given a novel object type, we can first find reusable ones from the existing collection of analytic concepts. If the current collection is sufficient, scaling to new object types requires no additional effort. If the current collection is insufficient to fully represent the new object type, a small number of new concepts can be conveniently defined at program level through code inheritance and invocation from existing concepts and basic geometry templates.

Scalability of Manually Constructing Analytic Concepts. Manual construction of analytic concepts will not significantly hinder their scalability, because 1) only a few (sometimes even zero) new concepts are needed to cover a new object category, since analytic concepts are highly generalizable and reusable across different object categories, and 2) scaling up analytic concepts is practically convenient at program level through code inheritance and invocation from existing concepts and basic geometry templates, which enables developing a new concept in just about 2 hours on average as mentioned in Sec. 3.4. We will also explore the possibility of using algorithms to automatically develop new concepts as our future work, which is a promising way to drastically reduce human effort.

I. Real World Experiment Analysis

I.1. Sim-to-real Gap

Although part of the training is conducted in a simulated environment, the gap between the simulated and real environments on our approach is relatively small for the following reasons: **1)** The Grounded-SAM [36] is trained on real data in real-world experiments to crop the target part's point cloud from the RGB image. **2)** The structural parameter estimation and 6-DoF pose estimation modules are trained in a simulated environment and performed on the target part's point cloud. For a target part suitable for manipulation, typically a few centimeters in size, the point clouds captured in the real environment and the simulator do not differ significantly to produce drastically different output for the modules. Thus the sim2real gap is rather small on the part-level. **3)** We adopt data augmentations to further narrow the sim2real gap, including adding noises and corruptions, and randomizing objects' poses and camera perspective.

I.2. Robustness in Real-World Experiments

In real-world experiments, we have identified several factors that may impact the effectiveness of our approach. We discuss these factors in this section.

Low-quality Point Cloud. Although we strive for precise estimation of the target part's parameters and 6-DoF pose, the primary focus of our work is to reasonably ground the knowledge inferred by MLLMs in the physical world, and finally enabling more accurate, stable, and controllable robot manipulation. Therefore, although the low-quality point cloud of the target part collected in real world in some cases may introduce slight inaccuracies in parameter and 6-DoF pose estimation, our approach still ensures that the knowledge inferred by MLLMs is effectively grounded in physical world. As long as the estimated parameters and pose reasonably describe the target part, the robot can successfully complete manipulation tasks under the guidance by our approach.

Similar Shapes. As our study focuses on the spatial structure and function of target parts, we do not aim to fully represent all the details of the target part. While various real-world objects can share similar parts, they also share similar spatial structure and functions, which can be effectively captured by analytic concepts and manipulation knowledge. In this manner, the robot is guided to complete manipulation tasks on objects with similar target parts using similar knowledge. Thus similar shapes in target parts will not cause confusion.

J. Generalization Capability in Manipulation Tasks

The generalization capability in manipulation tasks of our approach stems from two aspects: 1) highly expressive capability of analytic concepts to describe objects' structural knowledge, and 2) the atomic nature of manipulation knowledge. In addition to the discussions of these aspects, we also provide a discussion of how our approach handles manipulation tasks beyond common practice.

Highly Expressive Capability of Analytic Concepts. Analytic concepts can comprehensively describe the structural knowledge of various articulated objects encountered in daily life, as the design and manufacturing processes of man-made objects are based on combining a series of basic geometric primitives (*e.g.*, CAD). Specifically, we introduce 10 basic geometries (*Cuboid, Cylinder, Sphere, Cone, Triangular Prism, Ring, Torus, Rectangular Ring, Cuboid with Circular Hole, Cylinder with Box Hole*) as the fundamental building blocks of analytic concepts to describe objects' structural knowledge. Each geometry encapsulates a category of shapes that share common spatial properties, while its associated parameters instantiate specific instances. Building on these basic geometries, analytic concepts describe an object's structural knowledge by combining these geometries, resulting in an exponential diversity in representing structural knowledge. As a reference, PartNet-Mobility [44] is an articulated object dataset consisting of objects collected from 3D Warehouse, encompassing models of real-world manufacturers' products. In our experiments, our analytic concepts effectively describe the structural knowledge of various parts of objects from Xiang et al. [44]. The diversity and complexity of the structural knowledge that analytic concepts can cover are comparable to real-world level, which demonstrates the highly expressive capability of our analytic concepts.

Atomic Nature of Manipulation Knowledge. The manipulation knowledge of robot articulated object manipulation tasks can be comprehensively defined due to the atomic nature of fundamental actions (*e.g.*, lift, turn, push, *etc.*) during manipulation, and these atomic actions are enumerable. An object manipulation process can be decomposed into a finite sequence of atomic actions (*e.g.*, *opening a door* can consist of *grasping* the knob, *turning* it around its axis, and *pulling* the knob). By systematically combining these atomic actions, we can define a broad spectrum of manipulation knowledge to accommodate various object manipulation tasks. Leveraging the MLLM, our approach can determine the most reasonable manipulation knowledge for a specific manipulation task among the numerous manipulation knowledge that analytic concepts are able to cover. This enables robots to successfully complete a wide range of manipulation tasks.

K. Limitations and Future work

Our current work focuses on **articulated object manipulation**, through aligning semantic-level knowledge with physical-level knowledge. We focus on gripper-based manipulation of articulated objects and hence develop manipulation knowledge, aiming to validate the feasibility and effectiveness of using analytic concepts to improve the performance of articulated object manipulation tasks. In the future, we will extend our approach to support a broader range of end-effectors (*e.g.*, dexterous hand), objects beyond articulated objects (*e.g.*, deformable ones), complex scenarios (*e.g.*, long-tasks involving multiple objects) to enable more comprehensive manipulation capabilities.

L. Experiment Computer Resources

Our experiments are conducted on single NVIDIA A100 GPU and Intel(R) Xeon(R) Platinum 8276M CPU @ 2.20GHz.

M. Societal Impacts

We propose analytic concepts as a bridge between semantic-level knowledge of articulated objects and the physical

world, extending the applicability of MLLMs to real-world understanding and interaction, which may improve the quality of human life. However, this approach depends on MLLMs, which may incur additional economic and energy costs due to computation.

N. Statement on Usage of Large Language Models

This paper employs Large Language Models solely for polishing the writing, including grammar correction, phrasing refinement, and improvements in fluency and readability. No scientific ideas, experimental designs, results, or conclusions were generated by LLMs. All conceptual and technical contributions are entirely the work of the authors.