RAIL: Robot Affordance Imagination with Large Language Models

Ceng Zhang[1*], Xin Meng[1*], Dongchen Qi[2], Gregory S. Chirikjian[1,2]

***Abstract*— This paper introduces an automatic affordance reasoning paradigm tailored to minimal semantic inputs, addressing the critical challenges of classifying and manipulating unseen classes of objects in household settings. Inspired by human cognitive processes, our method integrates generative language models and physics-based simulators to foster analytical thinking and creative imagination of novel affordances. Structured with a tripartite framework consisting of analysis, imagination, and evaluation, our system "analyzes" the requested affordance names into interaction-based definitions, "imagines" the virtual scenarios, and "evaluates" the object affordance. If an object is recognized as possessing the requested affordance, our method also predicts the optimal pose for such functionality, and how a potential user can interact with it. Tuned on only a few synthetic examples across 3 affordance classes, our pipeline achieves a very high success rate on affordance classification and functional pose prediction of 8 classes of novel objects, outperforming learning-based baselines. Validation through real robot manipulating experiments demonstrates the practical applicability of the imagined user interaction, showcasing the system's ability to independently conceptualize unseen affordances and interact with new objects and scenarios in everyday settings.**

I. INTRODUCTION

In domestic settings and healthcare facilities, unstructured environments and dynamic daily tasks necessitate human-level intelligence for robots to automatically and adaptively engage in novel objects and scenarios. Recently, Large Language Models (LLMs) have showcased their impressive conversational and logical abilities. Trained on vast amounts of data, LLMs can distill essential information from ambiguous and unspecific requests to create coherent narratives. Recent studies have demonstrated that LLMs can assist robots in making high-level decisions applicable to real-world scenarios. A key challenge lies in the fact that LLMs lack a practical grasp of physics, which hinders their comprehension of the physical world and their ability to make grounded assessments and feasible plans. For example, when provided with a description and asked "Is it a table?", LLMs may provide a logical response that lacks physical plausibility.

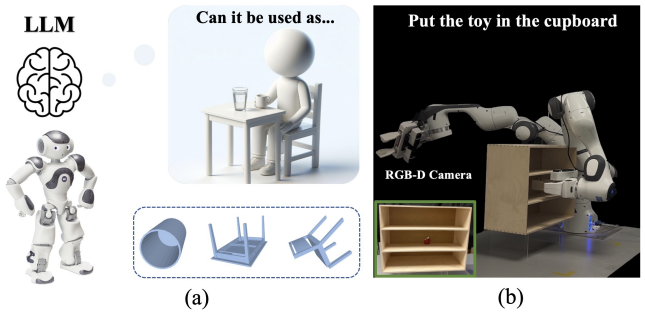

Fig. 1. Robot imagination with LLMs. (a) The robot imagines the affordances of randomly placed novel objects assisted with LLMs. (b) The robot performs novel tasks based on affordance reasoning.

Integrating physical properties, the concept of robot imagination assesses the object affordances from an interactive perspective, enriching the information for robot manipulation. To approach the affordance reasoning from a user-centric perspective, we define the affordance of an object by *Interaction-Based Definition (IBD)*. Illustrated by the examples of chairs and containers in [1]–[3], IBD defines the potential user and feasible user-object interactions, providing extensive instructions for the robot to imagine the scenario and assess the resultant interaction. We define the pose of the object that allows the expected interaction as the *functional pose* for the target affordance. The object that has at least one functional pose is recognized as functional, *i.e.*, fulfilling such affordance.

Current robot imagination methods require the development of customized imagination systems for different affordances, in which developers analyze the definition of a class of objects from the dictionary and decompose it into an IBD. To translate the affordance request into the programming language that the robot can understand and execute, in our previous works [2], [3], we propose the simplified agent model to describe the potential user and encode the interaction into feasible motions and expected outcomes. We define the applicable agent models and motions as the *imagination profile*. An evaluation matrix is tuned to determine the successful interactions, allowing the robot to recognize the functional poses of the object and, therefore, classify the objects. However, it is still an open challenge to imagine novel affordances without complicated implementations by human developers.

In this paper, we tackle this challenge by developing an automatic imagination pipeline that is only conditioned on the name of the requested object affordance, by employing LLMs to replace human analysis and heuristics. When approaching novel affordances, humans read the definition from

* denotes equal contribution.

This work was supported by NUS Startup grants A-0009059-02-00, A-0009059-03-00, CDE Board account E-465-00-0009-01, and National Research Foundation, Singapore, under its Medium Sized Centre Programme - Centre for Advanced Robotics Technology Innovation (CARTIN), sub award A-0009428-08-00.

[1] Ceng Zhang, Xin Meng and Gregory S. Chirikjian are with the Department of Mechanical Engineering, National University of Singapore, Singapore. {tmy_zc, mengxin, mpegre}@nus.edu.sg

[2] Dongchen Qi and Gregory S. Chirikjian are with the Department of Mechanical Engineering, University of Delaware, Newark, DE 19716, USA. {dcqi, gchirik}@udel.edu

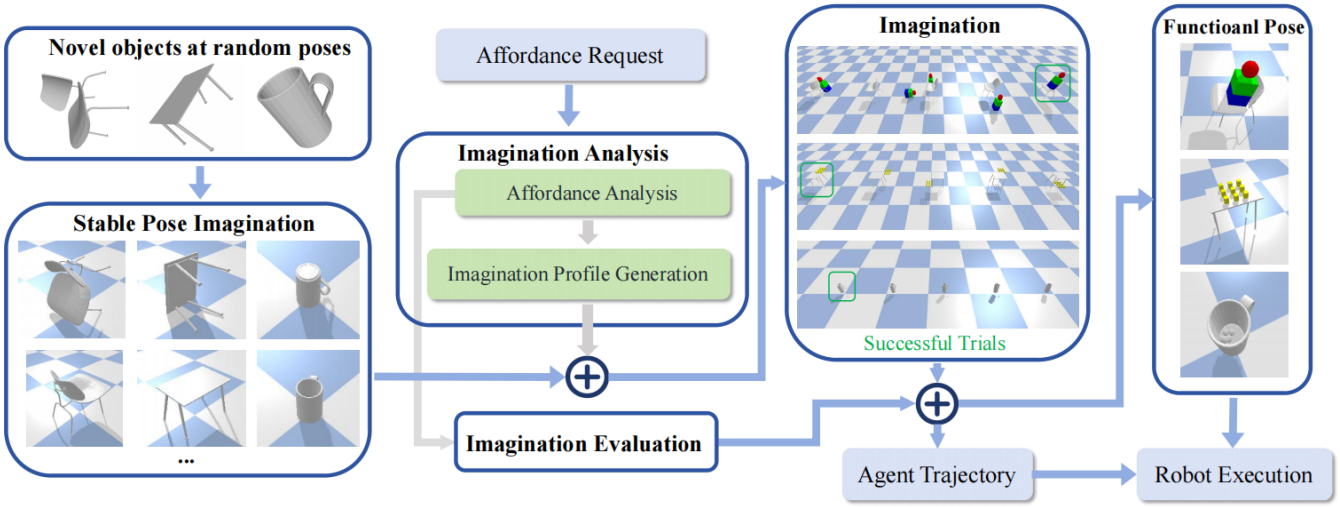

Fig. 2. Pipeline. Given an object model in a random pose, the algorithm first imagines its stable poses. The Imagination Analyzer analyzes the requested affordance and generates an executable imagination profile. The algorithm simulates the imagination profile with the object and loop for all stable poses. The Imagination Evaluation determines whether the object has the requested affordance. If the object is functional, the functional pose and agent trajectories are recorded for potential real robot execution.

the dictionary, analyze IBD, construct an imagination profile, and put it into the brain to imagine it, thereby analyzing the conclusion and generating executable actions. With robots having cameras as eyes, imagination as brains, and end-effectors as hands, we use LLMs as a powerful dictionary that provides detailed profiles.

Instead of asking LLMs to directly reason about the environment and plan actions to manipulate the object, we only require them to answer affordance-related semantic questions that are not conditioned on the specific objects and environments. With the imagination profile generated, the robot puts the real object in its brain and imagines the interactions proposed by LLMs (Fig. 1). To reach a physically applicable conclusion, LLMs also assist in comparing imagination outcomes with expected ones, providing practical proposals for affordance classification, functional pose prediction, and manipulation.

Evaluated with 301 novel synthetic data, our method showcases a robust 88.2% accuracy in identifying new affordances and an impressive 92.7% success rate in determining functional poses. In real robot experiments, the system recognized both the affordances and the functional poses of 18 previously unseen objects. It achieves a 100% success rate in executing novel tasks by accurately parsing semantic requests and reasoning novel affordances. Comparing it with leading learning-based approaches and an ablation study baseline, we empathize the effectiveness, generality, and practical applicability of our method. The main contribution of this paper lies in:

- An affordance reasoning pipeline that only requests target affordance names.
- An imagination framework simulates customized profiles for multi-class affordances.
- A real robot manipulation system for performing novel tasks on unseen objects based on affordance reasoning.

II. Related Work

A. Learning-based Object Classification

Object classification has been a crucial area of research in computer vision, a fundamental approach in this field is the application of Convolutional Neural Networks (CNNs) [4], based on which later works have made improvement and achieved better performance in tasks with large datasets [5]–[8]. Despite the speed and efficacy of these methods, they impose strict requirements on images. One challenge is that visual occlusions can greatly impact result accuracy, and potential flaws may go unnoticed by the method, leading to objects lacking their intended functionality (e.g., a cup without a bottom cannot hold liquid). To address this, studies have employed sophisticated techniques to integrate multiple viewpoints for 3D object classification [9]–[11]. However, they still encounter difficulties with object poses, resulting in a significant drop in classification accuracy when objects are not in their functional poses.

Recently, the incorporation of methods such as attention mechanisms and transformer models, as demonstrated in Vision Transformer (ViT) [12], indicates a promising avenue to further enhance classification precision by utilizing global context and long-distance dependencies in images. Nevertheless, based on visual information, these methods face similar challenges. In addition, requiring substantial datasets and significant computational resources and time for training such a substantial model remain issues. In contrast, our approach does not depend on visual cues, but infers the object's affordance through physical interaction in simulation. And by employing a pre-trained frozen LLM, our system releases the need for training process and data requirements.

B. Affordance Recognition by Physical Reasoning

The detection of affordances has recently been a highly popular subject [13]–[16]. This area of study helps robots

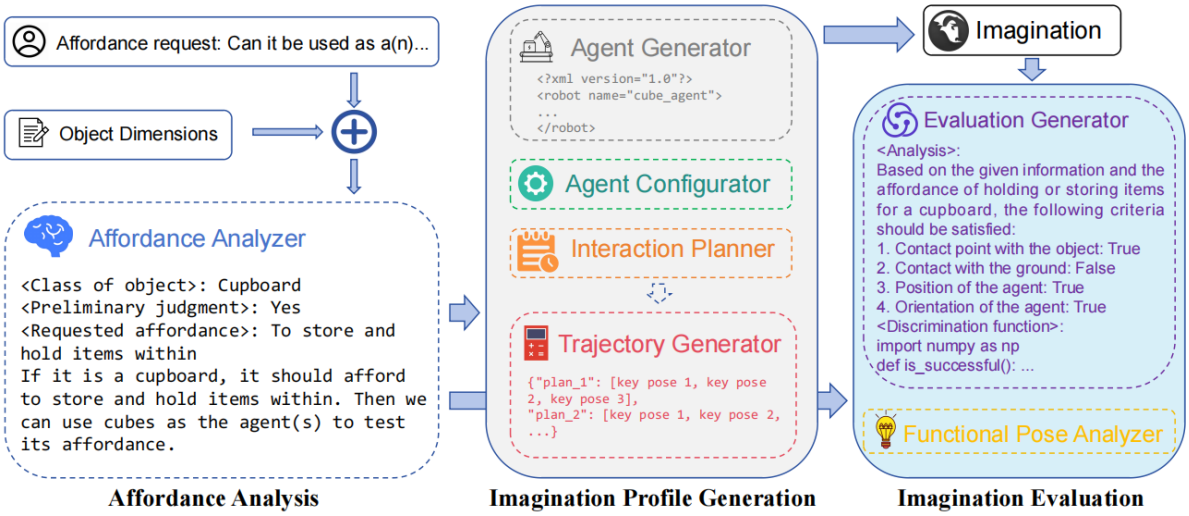

Fig. 3. Imagination analysis and evaluation framework. The Affordance Analyzer creates the IBD and an abstract imagination outline. The Imagination Profile Generator then develops detailed agent model and action trajectories. Subsequently, the Affordance Evaluator uses a scoring function generated to assess each imagined plan, determining the functional pose.

better identify objects in their environment, understand their functionalities, and learn to interact with them appropriately. One of the approaches to affordance detection involves physical interaction, which is seen as an intuitive and natural behavior of human [17]. Research works such as [18] and [19] suggest that robots can explore the functionalities of new objects by interacting with them, potentially using them as tools to perform tasks that were previously inaccessible. Our previous research also focuses on envisioning the affordances of objects, such as determining the sitability of a chair [2] and the containability of a cup [1], and utilizing the functional poses of these objects for robot manipulation tasks in real-world settings. While these methods are intuitive, setting up the simulation environment involves extensive hard coding and only identifies a specific affordance. In this study, we streamline the analysis and evaluation process by leveraging the reasoning capabilitiy of LLMs and applying it to *robot imagination* with various novel affordances.

C. Robot Planning with Large Language Models

As interest in large language models grows, the reasoning capability of these models has provided new opportunities for robot planning, a problem that previously required complex algorithm development [20]–[22]. Furthermore, their language analysis skills can help robots better understand user needs based on abstract instructions [23]–[25]. A unified model suggested in [21] can be implemented across various robots for diverse task executions. Fan et al. highlight the potential of LLMs in industrial robot applications in [26], with the aim of autonomous completion of production tasks on different assembly lines. One work similar to ours is [27], where a tool recognition system for robot task completion is introduced that comprises multiple LLM modules including an analyzer, planner, encoder, etc. In contrast to these instances, our focus is on utilizing a framework with several LLM modules to create an autonomous system that classifies objects by imagining the affordances of novel objects with minimal human intervention.

III. Language Model Assisted Imagination

The pose of a rigid body can be represented as $g = (R, \mathbf{p}) \in SE(3)$, where $R \in SO(3)$ is a rotation matrix, $\mathbf{p} \in \mathbb{R}^3$ describes the position. *Functional pose* g_f is therefore defined as a pose in which the object affords the functionality. Given an unseen object and a novel affordance, our goal is three-fold: verifying whether the object possesses such affordance, identifying the functional pose, and how to use the object if it is functional.

In this work, we approach the problem of affordance reasoning from an interaction-based perspective by employing a generalized automatic robot imagination pipeline shown in Fig. 2. The novel affordance can be assessed by a three-step stream: 1) imagination analysis: analyzing the IBD of the affordance and decomposing it into the potential interaction of agents and expected outcome; 2) imagination: simulating the generated interaction in a physics-based simulator; 3) imagination evaluation: evaluating the quantitative imagination results and proposing the object affordance. Instead of heuristic reasoning from human developers, we propose a framework that integrates LLMs for analysis that can be generalized to a wide scope of affordances.

Taking natural language a as the prompt from the corpus $\mathcal{V}$, LLMs output the optimal content a^* based on previous tokens t:

$$a^* = \arg\max_{a \in \mathcal{V}} P(a|t), \tag{1}$$

our framework employs multiple LLM modules to analyze affordances and imagination outcomes. Specifically, they take the user's prompt r and constraints c as input and generate output $\mathcal{O} = \mathcal{L}(r, c)$.

For a novel object, the input of the framework is a task description d_{task} which only includes the requested affordance and the dimensional information of the object, defined as $d_{\text{task}} = \{r_{\text{aff}}, g_{\text{obj}}\}$. The requested affordance r_{aff} is specified

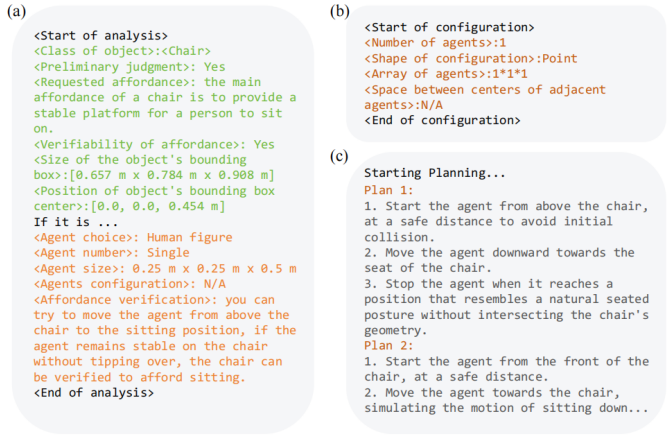

Fig. 4. (a) Affordance Analyzer, (b) Agent Configuration Generator, (c) Agent Motion Planner.

by natural language, and g_{obj} includes the dimension of the object's bounding box (OBB) and the object position.

In IBD, an object must be in a stable pose to be considered functional for any affordance. Therefore, the algorithm first takes the 3D model of the unseen object and finds a set of stable poses by stable pose imagination [2], referred to as $g_s \in G_s$. Setting the object into each stable pose as the candidate pose, by breaking the imagination analysis into a high-level affordance analysis and a low-level imagination profile generation, our approach proposes the four-step workflow: 1) **Affordance Analyzer** $\mathcal{A}(d_{\text{task}}) \rightarrow d_{\text{IBD}}, d_{\text{agent}}, d_{\text{interaction}}, d_{\text{evl}}$ parses the input d_{task} and generates the IBD d_{IBD} and an abstract imagination outline including the description of the agent d_{agent} and agent action $d_{\text{interaction}}$, and expected outcome d_{evl}. 2) **Imagination Profile Generator** $\mathcal{G}(d_{\text{agent}}, d_{\text{interaction}}) \rightarrow a, p$ takes the analyzed outline as input and outputs executable agent model a and trajectories $\mathcal{T}$ for imagination. 3) **Imagination** $\mathcal{I}(a, \mathcal{T}) \rightarrow \mathcal{R}$ simulates the generated agent motions and saves resultant configurations $\mathcal{R}$. 4) **Imagination Evaluator** $\mathcal{E}(d_{\text{agent}}, d_{\text{interaction}}, d_{\text{evl}}, \mathcal{R}) \rightarrow j$ outputs the summarized affordance judgments by generating a scoring function to evaluate the imagination outcomes. The workflow is shown in Fig. 3.

A. Affordance Analysis

The Affordance Analyzer $\mathcal{A}$ processes the task description d_{task} and proposes analysis as the outline of the affordance reasoning. This module first assesses whether the object dimension matches the target affordance r_{aff}. If there is a mismatch, it proposes a potential alternative affordance and replaces it as r_{aff}. Given the target affordance, the analyzer summarizes the IBD d_{IBD}. Based on d_{IBD}, it proposes a physical figure that can act as an agent to verify the target affordance. Multiple agents are involved in simulating complicated physics or efficient parallel computing. The configuration of the agent d_{agent} is analyzed as a language description of the shape and the abstract layout, *i.e.*, the geometry shape of the agent, and how the agent(s) distribute. The analyzer plans the active motion $d_{\text{interaction}}$ of the agent(s) to interact with the object and forecast the outcome, d_{result}. If the resultant configuration of the planned agent(s) motion aligns with the expected outcome, the object is identified as functional.

B. Imagination Profile Generation

With the abstract analysis, the algorithm creates a set of configuration files to perform the planned agent motion, including the agent model and the numerical trajectories. The generator is composed of four LLM modules, acting in a step-wise manner.

1) Agent Model Generator: Taking d_{task} and d_{agent}, the generator outputs a simple representation of the agent model a in a unified robotics description format (URDF) file. The agent model maintains the essential characteristics related to the target affordance and employs an appropriate scale considering the object dimension.

2) Agent Configuration Generator: To handle complicated scenarios, the imagination analyzer proposes multiple agents to simulate joint or parallel physics. We organize the spatial distribution of the agent into a cube or planar grid pattern, with sides of the shape aligning with the world frames, and the orientations of all agents remaining the same, an example is shown in Fig. 4(b). The distribution is parameterized by the number of agents along each side and the distance between each pair of neighboring agents. Based on d_{task} and a, the LLM module proposes the distribution parameters, considering the object dimension and collision avoidance. Therefore, the pose of each agent relative to the geometric center of the distribution can be extracted. For cases where only one agent is employed, the agent geometric center is the center of the distribution.

3) Agent Motion Planner: Taking d_{task} and the abstract description of the interaction method $d_{\text{interaction}}$, the planner aims to produce high-level plans of agent motions, shown in Fig. 4(c). Specifically, each plan provides a sequence of actions d_{motion}, indicating the relative spatial relations between the agent and object, as well as the moving direction of the agent. With the object placed in each candidate pose, it generates multiple plans $\mathcal{D} = \{d_{\text{motion1}}, d_{\text{motion2}}, ..., d_{\text{motion}i}, ...\}$, enabling the exploration of a diverse range of interactions. $d_{\text{motion}i}$ is the language description of the i-th plan.

4) Trajectory Generator: With the initial language descriptions $\mathcal{D}$ of the plans, the generator converts them into a group of numerical trajectories, referred to as $\mathcal{T} = \{t_1, t_2, ..., t_i, ...\}$. t_i is the i-th trajectory of the center of the agent(s). Each trajectory is given as a sequence of via poses of the agent(s) relative to the object, with the object in a candidate pose. The algorithm calculates the trajectory of each agent as the executable plan saved in a JSON file imported for imagination.

C. Imagination

The algorithm imagines the proposed agent a and planned interactions $\mathcal{T}$ in a physics-based simulator. It first loads the objects in the candidate stable poses and the agents in the initial poses, then simulates the agent moving through the

planned via-poses and the agent is released after the trajectory is completely executed. In addition, collision checking is performed at each step, and when a collision occurs, all agents involved are released immediately.

For the simulation of each plan, the resultant configuration r is summarized from two aspects: 1) Agent poses: the position and orientation of each agent relative to the object; 2) Contact points: number of contact points of each agent with the object, other agents, and the ground, respectively.

The outcome of imagination is the combination of the resultant configuration of all generated plans $\mathcal{R} = \{r_1, r_2, ..., r_i, ...\}$.

D. Affordance Evaluation

In this part, the algorithm analyzes the results of the simulation $\mathcal{R}$ to evaluate the object affordance. It consists of two LLM modules to evaluate r of each plan and the object affordance, respectively.

1) Scoring Function Generator: Based on d_{IBD} and d_{evl}, the generator produces an executable function $\mathcal{F}$ that scores the outcome of each plan. The function takes in the resultant configuration of each imagined plan r and generates a weighted score S.

$$\mathcal{F}(r) \longrightarrow S \tag{2}$$

It considers criteria related to target affordance, including agent-object relative position, agent orientation, agent-object contact, agent-agent contact, and agent-ground contact. The success of the plan is determined by whether S exceeds 0. All successful interactions are selected as candidate functional interactions $(d_{\text{motion}f}, t_f, r_f)$. Compared to previous works where the evaluation matrix is manually defined and tuned, our method provides an automatic approach that enables generalization across various types of affordances.

2) Functional Pose Analyzer: The imagined pose is hypothesized as potentially functional if there exists at least one candidate functional interaction. To analyze the validity of the candidate functional pose, we introduce another LLM module. By analyzing the candidate functional interactions $(d_{\text{motion}f}, t_f, r_f)$, it determines if the evaluation provided by $\mathcal{F}$ is consistent with common sense.

The object is classified as not functional if none of the candidate interactions of all stable poses $g_s \in G_s$ is valid. If there exists at least one valid candidate interaction, the analyzer selects the best functional interaction. The corresponding object pose is therefore the optimal functional pose.

IV. Experiments

We implement our framework system with Python, using Pybullet [28] as the physics simulator. The algorithm is evaluated on a computer running an Intel Core i7-11370H @ 3.3GHz CPU and Nvidia GTX 3060 GPU. In a real experiment setting Fig 5 (b), a Franka Emika robot arm is used for manipulation. An RGBD camera is mounted on the end effector for reconstruction. The language model in this study is founded on GPT-4 [29]. To enhance response time and speed of generation, we opt for the GPT-4-turbo variant and set the parameter *Temperature = 0.1* to ensure optimality while providing a certain degree of generalization.

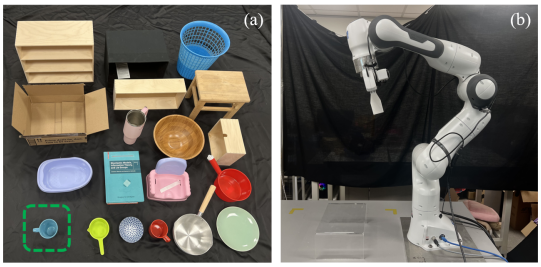

Fig. 5. Real world experiment details. (a) Snapshot for different classes of objects used for affordance imagination. The circled cup is used to tune the robot planning. (b) Real robot setting.

A. Data

Our dataset comprises 301 synthetic and 19 real-life objects. The synthetic objects, sourced from a subset of the Princeton ModelNet40 dataset [30], span 8 classes: cup, basket, bathtub, chair, plate, table, vase and bowl. To optimize the performance of LLM modules and the simulator, we utilize 6 synthetic objects across 3 classes, *i.e.*, cup, vase, and chair. The rest data, which are unseen by our system, are used as the test set. To assess affordance classification, non-functional objects from various classes are also included in the testing phase.

In real robot experiments, we use a cup to tune the parameters for the robot planning to shorten the sim-to-real gap. The remaining 18 novel objects, varied in size and appearance, are used to test the recognization of 15 classes of affordances and affordance-based manipulation. This includes 13 novel affordances previously unencountered by our system.

B. Real Robot Experiment

The object is placed on a transparent stand to enable comprehensive scanning of the bottom section, with it in a functional pose. The task is given to the system through a semantic request formatted as "put the (real agent) in/on the (requested affordance name)." The system has no prior knowledge of the object as well as the requested affordance.

The robot arm first moves to 12 predefined poses to scan the object. The object model is cropped and segmented from the scene point cloud reconstructed using TSDF-Fusion [31]. With the requested affordance name segmented and passed to the affordance reasoning module, the algorithm imagines the object model and classifies if the object can be used with the requested affordance in the placed pose. If the object is recognized as functional and in a functional pose, the algorithm proposes an optimal agent trajectory. The robot then requests a volunteer to hand over the real agent. Holding the real agent, the robot executes the imagined trajectory and positions the agent, transferring the affordance understanding to practical actions. Throughout this process, simple force control is used for robot manipulation, pausing its motion if external force surpasses a set limit, ensuring safety and precision in task execution.

TABLE I

ACCURACY (%) OF AFFORDANCE CLASSIFICATION

Method		Synthetic data								
		basket	bathtub	chair	cup	plate	table	vase	bowl	overall
Ours (random pose)		**94.6**	**82.6**	85.7	**91.2**	**91.7**	**85.7**	74.6	**93.2**	**87.7**
BLIP	functional pose	45.2	13.0	95.2	70.6	11.1	85.1	87.2	68.2	56.8
	random pose	12.9	8.7	42.9	35.3	5.6	31.9	38.5	36.4	30.2

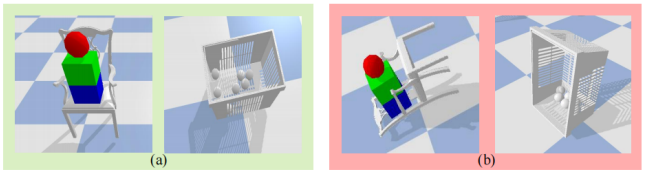

Fig. 6. Functional poses analysis. (a) True functional poses. (b) "Fake" functional poses in which the simulation result is scored as successful but is not validated by the Functional Pose Analyzer.

C. Baseline

We compare our method with the baseline, BLIP [32], which is a Visiual Question Answering (VQA) model that makes responses to user's prompt, we make comparisons on two tasks: affordance classification and functional pose prediction.

For affordance classification, we present an object's image alongside the query "Is it a (requested affordance name)?", and it ascertains whether the model possesses the requested affordance. To study how the object's pose impacts its performance and demonstrate the versatility of our approach across different poses, for each data sample we capture two images in arbitrary and functional poses, respectively. In functional pose prediction, we change the query to "Can the (requested affordance name) function in this pose?" With the image input, it makes judgement whether the object is in its functional pose. By doing so, we evaluate its ability to make responses on object pose information by visual cues.

Additionally, we perform ablation experiments in functional pose prediction to assess the significance of the functional pose analyzer in the framework. By comparing the outcomes of the base method with and without the functional pose analyzer, we aim to elucidate its impact on the overall performance of the affordance reasoning process.

D. Evaluation

We recruit volunteers to annotate the experiment result. For each object, we present it to the volunteer and ask "Do you think it can be a (requested affordance name)?" For each predicted functional pose, with the result of imagination, we ask the volunteer "Do you think this is the functional pose for (requested affordance name)?" In addition, for each trial of the real robot experiment, we show the experiment video and ask "Do you think the task was successfully performed?" to ensure that the pipeline runs in a reasonable manner.

V. RESULTS

We test our method on synthetic and real object data, respectively. The synthetic data are placed in a random pose, while the real object is placed on the table with an arbitrary upright pose.

A. Affordance Classification

The affordance classification achieves high success rates on synthetic data, as shown in Tab. I. Notably, the classification of novel affordance classes demonstrates exceptional performance, achieving a 88.2% success rate. Failure examples are mainly due to improperly generated profiles, such as imagination outline analysis, agent's model, trajectory parameters of agent motion, etc. We notice that the performance is degraded on vases, which we hypothesize is because the geometry varies greatly and the openings are usually small. In the absence of specific information about geometry, the generated agent may only match the overall dimensions of the object and it might be oversized for the opening, leading to a failure of the affordance imagination. BLIP does not perform as well on synthetic data even with objects in a functional pose. It shows good performances only for classes with unique appearance features such as chairs and vases. When objects are placed in random poses, it results in an overall correctness of only 30.2%.

In real robot experiments, the success rate is achieved 100%, with the objects placed in functional poses. BLIP showed a significant increase in performance on real data, with a success rate of 84.2%, which we attribute to the fact that most of the data used to train the model come from pictures of real objects.

TABLE II

ACCURACY (%) OF FUNCTIONAL POSE PREDICTION

Method	Synthetic data	Real data
Ours	**92.7**	**100.0**
Ours w/o Functional Pose Analyzer	75.3	79.2
BLIP	55.1	60.5

B. Functional Pose Prediction

We evaluate the functional pose prediction results on objects that are successfully recognized as functional. In Tab II, our method achieves very high accuracy in functional pose prediction for synthetic data. When considering only

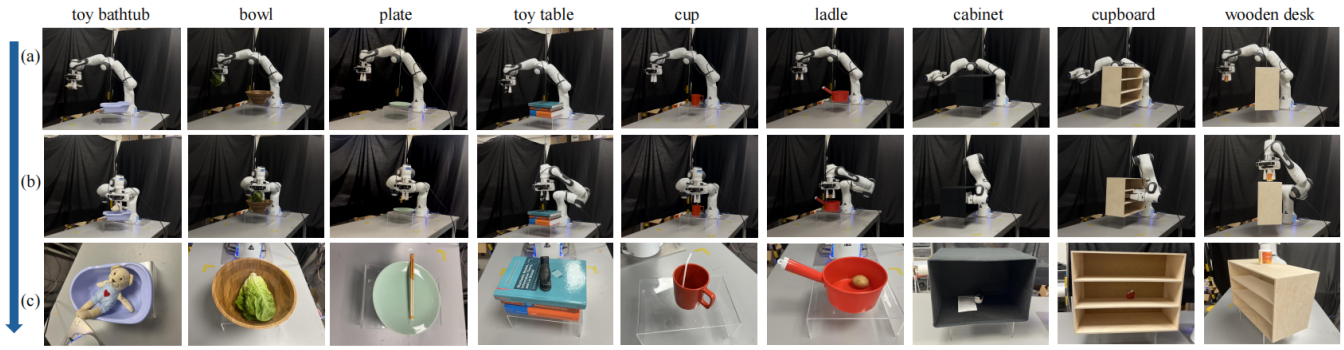

Fig. 7. Real Robot Experiment Results. (a)-(b)The robot positions the real agent according to the requested task. (c) Results.

the current pose, the robot successfully recognizes the object as being in a functional pose across all real-world trials.

In contrast, the baseline method BLIP shows low accuracy on both synthetic and real data, which is in line with the out-of-distribution challenge faced by vision-based methods. The performance is hugely affected by the view points, object pose, and object appearances. The ablation trials exhibit a notable decline in accuracy, suggesting that the Functional Pose Analyzer plays an important role in accurately determining the correct functional pose. Failure is often caused by the incomprehension of the Scoring Function, leading to the identification of "fake" functional poses that we do not expect from human common sense, illustrated in Fig. 6 By incorporating the Functional Pose Analyzer, our method provides more adequate and reliable judgments, ensuring that the predictions align more closely with practical expectations.

C. Real Robot Manipulation

Qualitative results are shown in Fig. 7, our system achieves 100% across 20 novel trials encompassing 15 distinct tasks, utilizing 18 objects previously unknown to the system. Notably, the system exhibits impressive generalization capabilities by successfully recognizing and manipulating 13 new affordances. Furthermore, it can identify objects with multifunctional uses and adapts its interaction accordingly. For example, the rightmost two trials in Fig. 7 showcase an object that the system recognizes both as a shelf, by positioning the agent on top, and as a cupboard, by inserting the agent inside.

VI. Conclusion

In this paper, we introduced an intelligent real2sim2real affordance reasoning framework that enables robots to understand and interact with novel classes of objects based on semantic requests. This process involves analyzing the affordance, imagining the generated scenarios, and evaluating the outcome to classify object affordance, predict the functional pose, and propose the potential user interaction. Our system demonstrated a success rate of 88.2% in identifying the affordance of novel classes, and successfully performed 20 novel tasks in real-world settings, showing significant potential in a wide range of daily indoor applications. Future work aims to expand the framework to articulated and deformable objects and understand and execute more complicated task commands.

References

[1] H. Wu and G. S. Chirikjian, "Can i pour into it? robot imagining open containability affordance of previously unseen objects via physical simulations," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 271–278, 2020.

[2] H. Wu, D. Misra, and G. S. Chirikjian, "Is that a chair? imagining affordances using simulations of an articulated human body," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7240–7246, IEEE, 2020.

[3] X. Meng, H. Wu, S. Ruan, and G. S. Chirikjian, "Prepare the chair for the bear! robot imagination of sitting affordance to reorient previously unseen chairs," *arXiv preprint arXiv:2306.11448*, 2023.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[9] S. Ullman, "Three-dimensional object recognition based on the combination of views," *Cognition*, vol. 67, no. 1-2, pp. 21–44, 1998.

[10] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view cnns for object classification on 3d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2016.

[11] A. Kanezaki, Y. Matsushita, and Y. Nishida, "Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5010–5019, 2018.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] L.-F. Yu, N. Duncan, and S.-K. Yeung, "Fill and transfer: A simple physics-based approach for containability reasoning," in *Proceedings of the IEEE international conference on computer vision*, pp. 711–719, 2015.

[14] L. Hinkle and E. Olson, "Predicting object functionality using physical simulations," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2784–2790, IEEE, 2013.

[15] Z. Jia, A. Gallagher, A. Saxena, and T. Chen, "3d-based reasoning with blocks, support, and stability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2013.

[16] E. Ruiz and W. Mayol-Cuevas, "Where can i do this? geometric affordances from a single example with the interaction tensor," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2192–2199, IEEE, 2018.

[17] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proceedings of the National Academy of Sciences*, vol. 110, no. 45, pp. 18327–18332, 2013.

[18] Y. Zhu, Y. Zhao, and S. Chun Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2855–2864, 2015.

[19] K. P. Tee, J. Li, L. T. P. Chen, K. W. Wan, and G. Ganesh, "Towards emergence of tool use in robots: Automatic tool recognition and use without prior tool learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6439–6446, IEEE, 2018.

[20] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[21] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[22] Y. Chen, J. Arkin, Y. Zhang, N. Roy, and C. Fan, "Autotamp: Autoregressive task and motion planning with llms as translators and checkers," *arXiv preprint arXiv:2306.06531*, 2023.

[23] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," *arXiv preprint arXiv:2307.01928*, 2023.

[24] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530, IEEE, 2023.

[25] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500, IEEE, 2023.

[26] H. Fan, X. Liu, J. Y. H. Fuh, W. F. Lu, and B. Li, "Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics," *Journal of Intelligent Manufacturing*, pp. 1–17, 2024.

[27] M. Xu, P. Huang, W. Yu, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, and D. Zhao, "Creative robot tool use with large language models," *arXiv preprint arXiv:2310.13065*, 2023.

[28] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.

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

[30] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

[31] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 303–312, 1996.

[32] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, pp. 12888–12900, PMLR, 2022.