Visual Preference Inference: An Image Sequence-Based Preference Reasoning in Tabletop Object Manipulation

Joonhyung Lee[1], Sangbeom Park[1], Yongin Kwon[2], Jemin Lee[2], Minwook Ahn[3] and Sungjoon Choi[1*]

***Abstract*— In robotic object manipulation, human preferences can often be influenced by the visual attributes of objects, such as color and shape. These properties play a crucial role in operating a robot to interact with objects and align with human intention. In this paper, we focus on the problem of inferring underlying human preferences from a sequence of raw visual observations in tabletop manipulation environments with a variety of object types, named Visual Preference Inference (VPI). To facilitate visual reasoning in the context of manipulation, we introduce the Chain-of-Visual-Residuals (CoVR) method. CoVR employs a prompting mechanism that describes the difference between the consecutive images (i.e., visual residuals) and incorporates such texts with a sequence of images to infer the user's preference. This approach significantly enhances the ability to understand and adapt to dynamic changes in its visual environment during manipulation tasks. Furthermore, we incorporate such texts along with a sequence of images to infer the user's preferences. Our method outperforms baseline methods in terms of extracting human preferences from visual sequences in both simulation and real-world environments. Code and videos are available at: https://joonhyung-lee.github.io/vpi/**

I. Introduction

Recent research has actively focused on aligning the behaviors of a robot or AI systems to match user preferences, thereby enhancing interaction and task performance efficiency [1]. For example, when a robotic arm performs the task of pouring the liquid inside a cup into a bowl, the trajectory planning can be optimized by aligning with a user's intuitive control preferences [2]. Commonly, robotic behaviors have mainly relied on manually designed features through scalar values, such as walking motions for quadruped robots [3], velocity for autonomous driving [4], and gait patterns for exoskeletons [5]. However, these approaches are limited in that preferences have yet to be extended to visual features from images that enable capturing the context of the current scene intuitively.

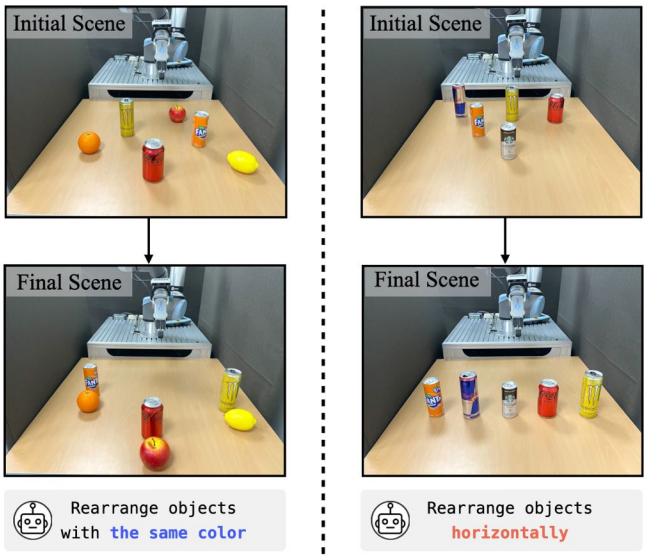


Fig. 1: **Visual Preference Inference (VPI) Tasks.** We define VPI tasks as reasoning user preferences based on an image sequence. Specifically, the task involves a robot that moves objects to target locations, following user instructions via mouse clicks which provide which object to move and where to place it.

Multimodal large language models (MLLMs) have advanced by integrating direct sensory perception into the reasoning processes, enhancing the ability to interpret and generate human-like responses [6]–[8]. These models have the ability to adapt contextually based on visual observations and respond to language descriptions appropriately. Furthermore, MLLMs have achieved human-like reasoning performances in a variety of robotic tasks (e.g., understanding complex multi-object relations [9] and long-horizon planning in manipulation tasks [10]). Human-like reasoning based on visual information has great potential as it can predict human preferences without manually defining features. In this work, our goal is to extract human preferences from raw visual information such as semantic (e.g., color and shape) or spatial (e.g., arrangement pattern) features.

In particular, we focus on inferring the human preferences that require visual understanding aligned with the user's intentions within robotic manipulation tasks. Hence, we introduce the task of extracting user's preferences solely

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079 , Artificial Intelligence Graduate School Program (Korea University)), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00277060, Development of open edge AI SoC hardware and software platform.)

[1]Joonhyung Lee, Sangbeom Park, and Sungjoon Choi are with the Department of Artificial Intelligence, Korea University, Seoul, Korea (email: {dlwnsgud8823, sangbeom-park, sungjoon-choi}@korea.ac.kr)

[2]Yongin Kwon, and Jemin Lee are with the Electronics and Telecommunications Research Institute, Daejeon, Korea (email: {yongin.kwon, leejaymin}@etri.re.kr)

[3]Minwook Ahn is with the Neubla, Seoul, Korea (email: minwook.ahn@neubla.com)

* Corresponding Author

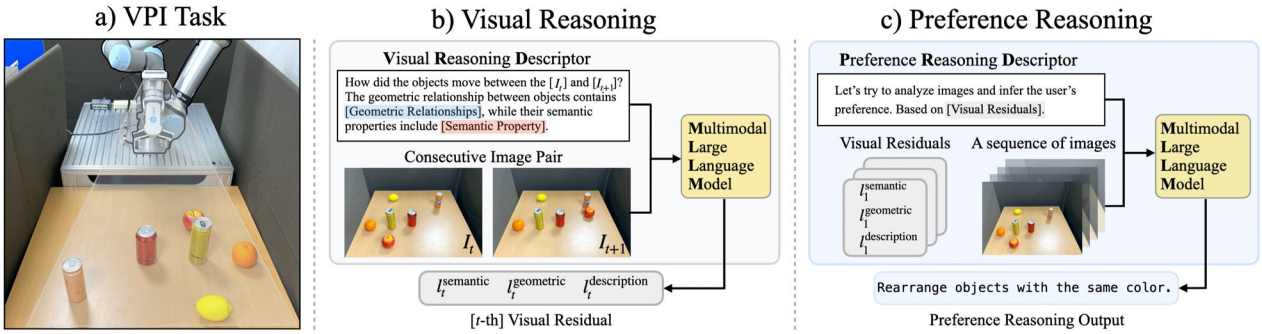


Fig. 2: **Overview of Chain-of-Visual-Residuals:** (a) We introduce a Visual Preference Inference (VPI) task, which extracts users' preferences solely from visual representations in tabletop manipulation environments. Our approach, CoVR prompting, involves generating (b) visual reasoning descriptions of consecutive images and (c) chaining these descriptions for interpreting human preferences from the scene sequences.

from visual representations, referred to as **VPI** which stands for **V**isual **P**reference **I**nference. To this end, we propose **C**hain-**o**f-**V**isual-**R**esiduals (**CoVR**) prompting, a method that involves a series of intermediate visual reasoning steps leading to the end response. Specifically, **CoVR** consists of two phases: 1) Visual Reasoning Descriptor (VRD), which maps user interaction with image inputs into scene descriptions that focus on capturing both semantic attributes of objects and changes in the geometric relationships between objects, and 2) Preference Reasoning Descriptor (PRD), which predicts a suitable preference considering the interactions of object manipulation.

We demonstrate the versatility of our method across a number of object manipulation tasks, as illustrated in Fig. 1. The results indicate that our method adeptly infers appropriate interaction patterns in response to dynamic changes of moved objects, including semantic reasoning tasks (e.g., the rearrangement of identically colored objects). Furthermore, the CoVR prompting is suitable for spatial reasoning tasks, which are notably difficult to articulate using textual descriptions alone (e.g., the horizontal rearrangement of objects).

Our contributions are summarized as follows: First, we present a visual reasoning task for inferring human preferences from a sequence of images. Second, we propose the CoVR prompting that enables inferring user preferences while understanding the dynamic changes from visual information. Third, we show that our proposed method can predict diverse user preferences in multiple manipulation tasks with various object types.

II. Related Work

A. Human Preferences in Robotics

Research on predicting user preferences has been actively conducted in robotics [11], [12]. In particular, these approaches have been utilized for a furniture assembly task where the robot assists the human by using a screwdriver on a small screw [13], [14]. In addition, the quadruped robot's behavior can naturally be tailored to one's preference, such as waving the left front leg [3]. Humanoid motion can also be optimized with respect to the user's preferences [15], [16]. Similar to the aforementioned approach, preferences can be designed by hand-crafted features. Moreover, a self-driving car has diverse features, including distance to other vehicles, speed, and heading angle [17]. Other studies have explored the exoskeleton domain to extract the gait pattern [5], [18]. Additionally, research on preference prediction has also been extended to using visual features for determining mobile robot path-planning preferences [19]. However, [19] also mentions that this paper is limited to handling the preference features that require interpreting visual information contextually. In contrast, we focus on reasoning visual information based on common knowledge for human-like prediction.

B. Multimodal Large models for Robotics

Large language models have been used for task planning [20], code generation [21], reward design [22], contact scheduling [23], and reasonableness reasoning [24] in robotics. Moreover, these foundation models can reason the common knowledge for understanding the geometric relationship between objects for placement tasks in the real-world [25]. However, relying solely on textual descriptions can make it difficult to connect with common knowledge. In contrast, visual representations make it easier to infer visual attributes, particularly when linking common knowledge to the physical world through sensory inputs. Accordingly, PG-VLM [26] leverages a multimodal LLM as a scalable way of providing high-level physical reasoning that humans use to interact with the world. Extensive work on enhancing the visual understanding capacity of multimodal LLMs by simply overlaying spatial and speakable marks on the images are mentioned in [27]. Similarly, 3DAP [28] has utilized the visual prompt method to maximize 3D understanding capabilities. Inspired by the visual reasoning capabilities of the foundation model, our method uses visual chain-of-thought prompts to infer user preferences.

III. Problem Formulation

In this section, we address the problem of extracting human preferences from visual representations (i.e., a set

of RGB images), referred to as **VPI**: **V**isual **P**reference **I**nference. In particular, we focus on tabletop manipulation tasks and formulate VPI to interpret human preferences using a sequence of n images $\mathcal{I} = \{I_1, I_2, \cdots, I_n\}$ obtained from a camera mounted on the end effector.

Our method focuses on analyzing visual signals from images to understand both the semantic and geometric properties of the objects in the scene. Specifically, the semantic property contains object color, shape, and category, and the geometric property corresponds to inferring the relative positions between objects, displacements, and the initial and final locations of the objects. As our method does not rely on the manual design of features, such properties are obtained solely from visual information utilizing an MLLM with proper prompting method which will be described in the next section.

IV. Proposed Method

We propose **C**hain-**o**f-**V**isual-**R**esiduals (**CoVR**) prompting, a method that connects visual understandings to reason about preferences from a long-horizon image sequence. Our proposed approach is comprised of two key components: Visual Reasoning Descriptor (VRD) in Sec. IV-A and Preference Reasoning Descriptor (PRD) in Sec. IV-B. Specifically, VRD is a template-based prompting method designed to understand semantic properties and identify changes in how the objects are geometrically related in consecutive image pairs. After the aforementioned process has been completed, PRD connects the previously obtained textual scene descriptions with input images to enhance the effectiveness of predicting a user's intention. The overall pipeline is depicted in Fig. 2.

A. Visual Reasoning Descriptor

Our goal is to identify *which* object has moved between two consecutive images and *how* the geometric relationship of objects has changed, while simultaneously inferring the semantic properties of each object. To this end, we present Visual Reasoning Descriptor (VRD) which translates input images into natural language scene descriptions (referred to as visual residuals). Visual residual V contains both the semantic properties of the objects and the difference in the objects' configurations between consecutive image pairs and consists of three components: $\{l^{\text{semantic}}, l^{\text{geometric}}, l^{\text{description}}\}$. l^{semantic} describes the semantic property of two objects (i.e., source object and target object) that have moved in between the image pairs and have been involved in this movement, $l^{\text{geometric}}$ refers to the spatial arrangement of the resulting relationship between two objects, and $l^{\text{description}}$ refers to the scene description of consecutive image pairs. VRD process can be formulated as:

$$\text{VRD}(I_{n-1}, I_n) \rightarrow V_{n-1} := \{l^{\text{semantic}}_{n-1}, l^{\text{geometric}}_{n-1}, l^{\text{description}}_{n-1}\},$$

where n represent image sequence index and V_n describes visual residual for the image pair (I_{n-1}, I_n).

Building upon this formulation, we provide the MLLM with the instructions along with the whole image sequence $\mathcal{I}$. However, when handling a large number of images, MLLMs tend to suffer from a lack of accuracy in interpreting scene information. To handle this issue, VRD first extracts the given image sequence into consecutive pairs and computes visual residual V by utilizing few-shot prompting where the prompts are given as follows:[1]

> I will give you a set of images [image1, image2, . . . , imageN].
> The goal is to reason about the geometric and semantic properties of objects in an image sequence.
> Format:
> - geometric property
> - semantic property
> - description
> [Examples]
> How did the objects move between the [image1] and [image2]?
> The geometric relationship between objects contains [Geometric Relationships], while their semantic properties include [Semantic Property].

As above, the VRD recognizes both the geometric and semantic properties of objects and provides a textual scene description to identify which objects have been moved and the corresponding relationships in between images.

For example in the case of Fig. 2-(I_1, I_2), scene description can be prompted by VRD to link object names ("apple", "orange drink"), semantic attribute ("sphere-shaped"), and geometric relationship ("in front of"). The generated response (highlighted) is as follows:

> geometric property: in_front_of
> semantic property: source object: apple, red, sphere_shaped,
> target object: orange drink, orange, cylinder_shaped
> description: Move the apple in front of the orange drink.

Here, the source object describes the object that has moved in between these image pairs, and the target object refers to the object that has been involved in this movement. We repeat this process iteratively to obtain visual residuals from all the successive image pairs.

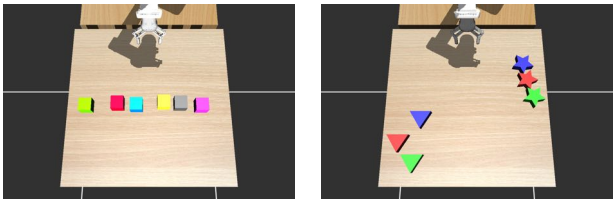
B. Preference Reasoning Descriptor

To interpret the overall preference from the obtained sequence of visual residuals $\mathcal{V} = \{V_1, \cdots, V_{n-1}\}$ between an image sequence $\mathcal{I}$ of length n, we propose Preference Reasoning Descriptor (PRD) to interpret user preferences described in natural language descriptions. To this end, we propose Preference Reasoning Descriptor (PRD) to interpret user preferences described in natural language descriptions. The visual residual information (obtained from VRD) along with the original image sequence is fed into PRD to reason about the underlying human preferences.

Given a set of images $\mathcal{I}$ and a sequence of visual residuals $\mathcal{V}$ for each image pair in $\mathcal{I}$, we formulate PRD that infers preferred objectives within the set of predefined preferences:

$$\text{PRD}(\mathcal{I}, \mathcal{V}) \rightarrow l^{\text{preference}},$$

[1]Geometric relations and semantic property include {to the left of, to the right of, in front of, behind of} and {color, shape, category}, respectively.



(a) **Block** Task (b) **Polygon** Task

Fig. 3: Examples of preferences for simulation scenario: (a) spatial pattern preferences arranged within a horizontal line, and (b) semantic preferences grouped with the same shaped objects.

where $l^{\text{preference}}$ denotes the inferred preferences based on the visual residual information. Specifically, we define a preference set containing nine elements for the few-shot prompting method[2]:

> Rearrange objects with the same color.
> Group objects by the same shape.
> Make objects into a horizontal line.
> Sort objects vertically.
> ...

Based on the above preference set, PRD infers the user preferences (highlighted) with previously obtained visual residual components (in gray):

> Let's try to analyze images and infer the user's preference.
> Based on previous visual residuals: [Visual Residuals]
> Preference: Rearrange objects with the same color.

We note that our method is capable of inferring human preferences in an open-ended manner without giving a predefined preference set. However, explicitly giving the preference set is more effective in terms of evaluating the performance of our method and baselines.

V. Experiments

In this section, we designed our experiments to address the following questions: (1) Can our proposed visual reasoning descriptor (VRD) capture both semantic and geometric properties from images during tabletop manipulation tasks? (2) Can our proposed preference reasoning descriptor (PRD) accurately predict human preferences by utilizing visual residuals extracted from raw observations in multiple manipulation scenarios?

To validate the effectiveness of VRD, we conducted a set of experiments focusing on visual reasoning performance. In addition, we designed further experiments to evaluate the ability to infer human preferences, which we divided into two categories: those based on semantic property and those based on spatial patterns. Specifically, the performance of our proposed method is evaluated across three different environments: Block Task, Polygon Task, and Household Task.

[2]The whole prompts can be found on https://joonhyung-lee.github.io/vpi/

Fig. 4: **Household objects:** We use various daily objects to test our approach, some of which can be categorized by terms of color, shape, or category.

First, we conduct experiments to demonstrate both the visual reasoning capability and the semantic preference reasoning capability in the Block Task where a robot moves blocks as shown in Fig. 3(a). For the Polygon Task in Fig. 3(b), we validate the performance of visual reasoning and spatial pattern preference reasoning with three triangular and three star-shaped polygons by rearranging the polygons. Finally, the Household Task is designed to demonstrate the applicability of our proposed method in real-world scenarios where various daily objects are used, as shown in Fig. 4.

A. Baselines & Metrics

We compare our method with other baselines, including large language models and a linear preference extractor. For fair comparisons on visual reasoning, we utilize the same visual reasoning module (i.e., GPT-4V [7]).

- MLLM-Naive: An ablation of our approach that does not use the Visual Reasoning Descriptor and Preference Reasoning Descriptor. MLLM-Naive infers scene descriptions for consecutive image pairs in a similar way to our method but without using the VRD template. Then, this baseline interprets the preference directly, using only an entire image sequence in a single interaction.
- MLLM-L2R: Inspired by Language-to-Reward (L2R) [29], this baseline extracts normalized object 2D position (ranging from 0.0 to 1.0) information for feature computation. Subsequently, we integrate a code snippet generation module that produces a piece of code to compute preference weights using the obtained object positions.
- Mutual-Distance-based Preference Extractor (MDPE): This baseline assumes that human preferences are deterministic, following a linear user model as discussed in prior works [11], [30]. Within the framework of

Siimulation Experiments: Block & Polygon		
Model	SR_{VRD} ↑	
	Block	Polygon
MLLM-CoVR (Ours)	**0.72±0.11**	**0.79±0.13**
MLLM-Naive	0.40±0.15	0.56±0.23

TABLE I: Table for visual reasoning experiments. The number (mean±standard deviation) indicates the success rate of predicting visual residuals in between images. A higher number indicates better performance.

Simulation Experiment: Block			
Model	SR_{PRD} ↑		
	quadrant	vertical	horizontal
MLLM-CoVR (Ours)	0.65	**0.90**	**0.80**
MLLM-Naive	**0.85**	0.70	0.60
MLLM-L2R	0.25	0.70	**0.80**
MDPE	0.45	0.40	0.40

TABLE II: **Block: Spatial Pattern Preference Reasoning.** Preferences are set for spatial patterns (i.e., quadrant, vertical, and horizontal alignments).

linear models, MDPE computes the preference weights for each specific feature based on pre-defined functions using the mutual distances between objects and then derives the preference from these weights.

Our evaluation metrics are the success rate of Visual Reasoning Descriptor (SR_{VRD}) and the success rate of Preference Reasoning Descriptor (SR_{PRD}) for the given image sequences. In particular, SR_{VRD} is calculated based on the visual residual between the image sequences and is defined as follows:

$$\text{SR}_{\text{VRD}} = \frac{1}{N-1}\sum_{k=1}^{N-1}\left(\frac{\sum_{l\in V_k}\mathbb{I}(l=\hat{l})}{|V_k|}\right)$$

where $|\cdot|$ means the number of elements in the set and $\mathbb{I}$ represents an indicator function that checks whether each element l within the predicted response V_k matches its corresponding element $\hat{l}$ in the ground truth visual residual $\hat{V}_k$ for each consecutive image pair. The elements of V_k and $\hat{V}_k$ include (l_k^{semantic}, $l_k^{\text{geometric}}$, $l_k^{\text{description}}$), and their respective ground truth counterparts ($\hat{l}_k^{\text{semantic}}$, $\hat{l}_k^{\text{geometric}}$, $\hat{l}_k^{\text{description}}$).

On the other hand, SR_{PRD} is measured according to the predicted preference that matches the ground truth. The preference criteria for each scene are manually designed. We formulate SR_{PRD} as follows:

$$\text{SR}_{\text{PRD}} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{I}(l_i^{\text{preference}} = \hat{l}_i^{\text{preference}})$$

where the indicator function $\mathbb{I}$ checks for a match between predicted description and ground truth preferences. SR_{PRD} evaluates whether the predicted preferences $l_i^{\text{preference}}$ are in alignment with the ground truth preferences $\hat{l}_i^{\text{preference}}$ across a defined set of scenes.

B. Block Task: Spatial Pattern Preference Reasoning

a) Setup: In this task, we evaluate the performance of visual reasoning and preference reasoning from image sequences, assuming that only one object is moved in between sequential images. This task contains six blocks of different colors that are rearranged in specific spatial patterns. These patterns can be defined within specific regions of a tabletop, such as the right side, left side, or upper side of a table, or by forming specific patterns, such as vertical or horizontal alignments. For example, as illustrated in Fig. 3(a), the blocks can be arranged to form a spatial pattern, such as a horizontal line along the center of the table. We compared our approach with a naive model, performing 10 times for the spatial pattern preference set.

b) Results: Firstly, Table I compares the visual reasoning performance of our method against the ablation of our method. The results indicate that our method outperforms the naive approach in understanding both the semantic and geometric properties in between an image sequence. Especially the results of 0.72±0.11 demonstrate the superior visual reasoning ability of our method. In contrast, the MLLM-Naive model shows limited ability in extracting visual signals between images with SR_{VRD} of 0.40±0.15 for the same task. This result highlights the effectiveness of our VRD template-based approach in recognizing the visual residuals within image sequences.

From the results in Table II, we compare the spatial pattern preference reasoning performance of our method to three other baseline approaches. Our method shows outstanding reasoning preference performance in the Block Task. This highlights the benefits of our prompting method as compared to the other baselines. Although the MDPE approach was expected to get a perfect score, MDPE did not achieve a score of 1.0 (indicating that this model is always correct). This poor performance of MDPE is mainly due to the parameter sensitivity used in its function function. Such sensitivity often leads to the recognition of multiple preferences, resulting in erroneous preference inferences. MLLM-L2R also exhibits limited effectiveness primarily because relying solely on the responses of the MLLM for position information is impractical; they do not account for specific geometric locations. From the experiment results, we would like to emphasize that our method has a high potential for visual reasoning tasks, taking into account spatial pattern preferences.

C. Polygon Task: Semantic Preference Reasoning

a) Setup: We aim to show the performance of visual reasoning and semantic preference reasoning from image sequences in the Polygon Task. This task consists of three triangular polygons and three star-shaped polygons. Our focus is on semantic preferences, specifically in grouping the objects based on either color or shape. The objective is to assess the ability to interpret semantic preferences derived from a sequence of images. For example, as depicted in Fig. 3(b), objects are sorted based on the shape attributes.

b) Results: On the Polygon task, as detailed in Table I, our methodology significantly outperforms the baseline

Simulation Experiment: Polygon		
Model	SR_{PRD} ↑	
	color	shape
MLLM-CoVR (Ours)	**1.0**	0.90
MLLM-Naive	0.20	0.70
MLLM-L2R	0.30	0.80
MDPE	**1.0**	**1.0**

TABLE III: **Polygon: Semantic Preference Reasoning.** Semantic (i.e., color and shape) preferences are evaluated. A score of 1.0 is considered a perfect performance, indicating that the model consistently makes accurate predictions.

Real-world Experiment: Household			
Model	SR_{VRD} ↑	SR_{PRD} ↑	
		Spatial Pattern	Semantic
MLLM-CoVR (Ours)	**0.63±0.08**	**0.67**	**0.67**
MLLM-Naive	0.28±0.19	0.17	0.33
MLLM-L2R	-	0.17	0.33
MDPE	-	0.50	**0.67**

TABLE IV: **Household: Real-world demonstration.** Each preference type is evaluated six times for both spatial pattern preference and semantic preference.

model, indicating superior visual reasoning accuracy. Specifically, our approach achieves a visual reasoning accuracy of 0.79±0.13 while the baseline records a lower accuracy of 0.56±0.23. This performance gap shows an enhanced ability of our method to accurately predict visual contexts within image sequences, especially in tasks involving complex geometric shapes such as polygons.

The results in Table III show that our method consistently outperforms other MLLM-based approaches for semantic preference reasoning in the Polygon task. In comparison, the MLLM-Naive method performs poorly with 0.20 for color and 0.70 for shape, indicating the limitations of naive models in capturing semantic properties. On the other hand, the MLLM-L2R model shows slight improvements, achieving a score of 0.30 for color and 0.80 for shape. However, these results still do not reach our reasoning preference performance. Notably, MDPE achieves the perfect scores (=1.0) both on color and shape criteria. However, it is important to know that the performance of MDPE depends on the manual tuning of the preference feature functions. These results imply that our method can successfully capture semantic preferences without any manual feature engineering, such as predefining a list of object attributes.

D. Household Task: Real-world Demonstration

a) Setup: The Household Task includes three object types: Fruits, Snacks, and Beverages. This task focuses on placing objects based on the semantic properties or within the spatial patterns, including preferences for both semantic and spatial arrangements. We use 12 real-world testing objects, as shown in Fig. 4, to evaluate the ability to identify both the semantic and the spatial preferences. We evaluate our approach in real-world tabletop environments with a 6-DoF UR5e manipulator with an OnRobot RG2 gripper.

b) Results: The metric of SR_{VRD}, presented in Table IV compared the visual reasoning performance of our approach against the ablation of our method, which was evaluated six times respectively. In particular, the results of 0.63±0.08 demonstrated the higher visual reasoning performance of our method. In contrast, the MLLM-Naive model showed a limited ability to extract visual signals between images with SR_{VRD} of 0.28±0.19 for the same task. These results support the effectiveness of our VRD template-based approach in recognition of visual residuals within image sequences.

Each type of preference was evaluated six times, and performance was measured in terms of SR_{PRD}. As illustrated in Fig. 6, in each step, the robot performs to move objects and captures images. The results of our method in Table IV are consistent with our simulation results, indicating the balanced performance of our method in spatial pattern and semantic preference reasoning. Compared to other MLLM-based approaches, they showed subpar performance in recognizing spatial patterns and semantic properties. We can notice that MLLM tends to misunderstand the spatial arrangements or semantic properties of objects without explicit annotation by VRD. While MDPE performs as effectively as our approach for both types of preference, it remains highly dependent on the need for handcrafted features. These results support the practical effectiveness of our method and support its successful application in real-world scenarios.

E. Limitation

Although our proposed method shows its effectiveness in multiple scenarios, our method is heavily reliant on the image sequence, which makes it highly dependent on perception and the order of images. For instance, a cylinder-shaped beverage might appear as a round object depending on the viewpoint of the camera. Therefore, our method may generate inaccurate scene descriptions, such as mismatching properties of objects or misunderstanding spatial relationships. We believe that one way to address this issue is to facilitate MLLM's understanding of human correction by incorporating human feedback.

VI. Conclusions

In this paper, we introduce a Visual Preference Inference (VPI) task designed to infer user preferences using visual reasoning from a series of images in the context of tabletop object manipulation. We have demonstrated the effectiveness of our method in interpreting spatial relations from image sequences and inferring preferences in both simulation and real-world tabletop environments. While our approach shows potential performance, it is critical to recognize its reliability on image sequences. Such dependence may result in sensitivity to perceptual biases and the order of images, which can lead to inaccuracies. Despite these limitations, our work presents a significant step toward enhancing the ability to understand preferences in manipulation tasks, opening avenues for further research in the field of reasoning preferences in the robotics domain.

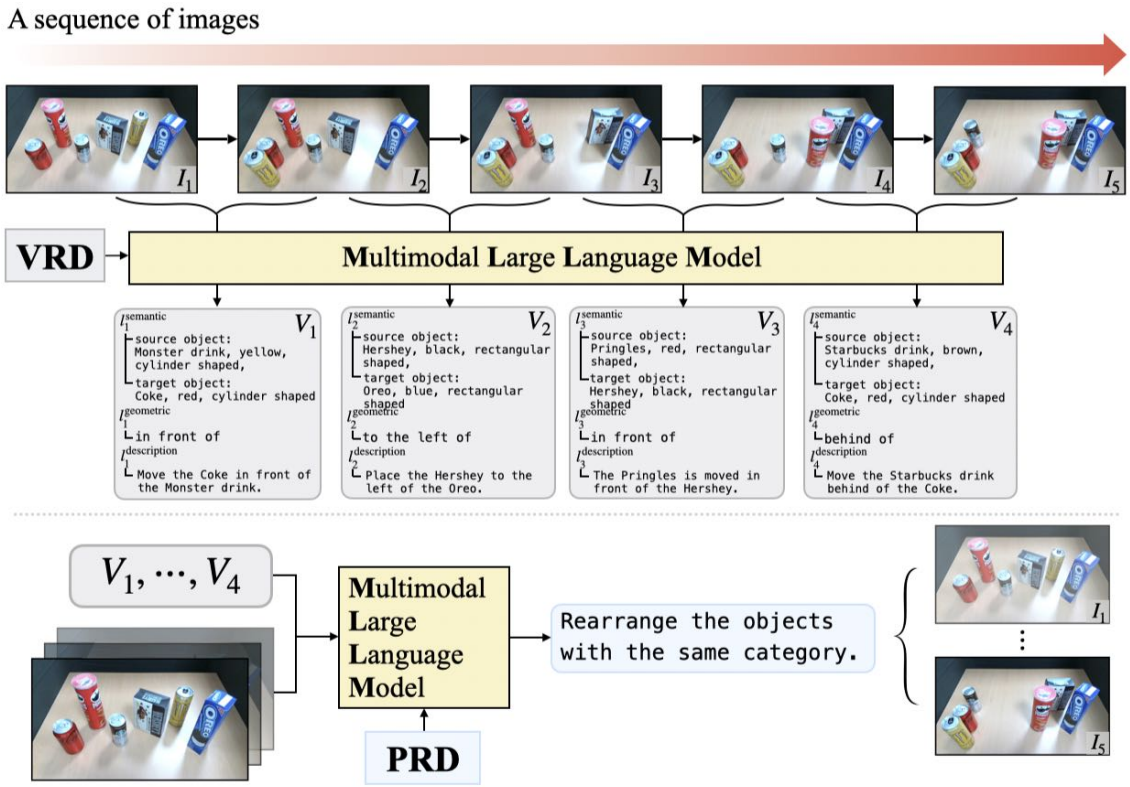


Fig. 5: This figure illustrates the application of CoVR in a scenario where objects are rearranged based on their category. The result of each visual residual shows the model's ability to identify semantic and geometric properties of objects, emphasizing the practical utility of CoVR in tasks that require a visual understanding of object properties and spatial relationships. See more videos and tasks at https://joonhyung-lee.github.io/vpi/

REFERENCES

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022.

[2] M. Li, D. P. Losey, J. Bohg, and D. Sadigh, "Learning user-preferred mappings for intuitive robot control," in *in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 960–10 967.

[3] K. Lee, L. Smith, and P. Abbeel, "Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training," *arXiv preprint arXiv:2106.05091*, 2021.

[4] N. Wilde, D. Kulić, and S. L. Smith, "Active preference learning using maximum regret," in *in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 952–10 959.

[5] U. H. Lee, V. S. Shetty, P. W. Franks, J. Tan, G. Evangelopoulos, S. Ha, and E. J. Rouse, "User preference optimization for control of ankle exoskeletons using sample efficient active learning," *Science Robotics*, vol. 8, no. 83, p. eadg3705, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.adg3705

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[7] Openai, gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[8] Y. Zhao, I. Misra, P. Krähenbühl, and R. Girdhar, "Learning video representations from large language models," in *in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6586–6597.

[9] I. Kapelyukh, Y. Ren, I. Alzugaray, and E. Johns, "Dream2real: Zero-shot 3d object rearrangement with vision-language models," *arXiv preprint arXiv:2312.04533*, 2023.

[10] Y. Hu, F. Lin, T. Zhang, L. Yi, and Y. Gao, "Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning," *arXiv preprint arXiv:2311.17842*, 2023.

[11] N. Wilde, D. Kulić, and S. L. Smith, "Learning user preferences in robot motion planning through interaction," in *in Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 619–626.

[12] A. Bobu, M. Wiggert, C. Tomlin, and A. D. Dragan, "Inducing structure in reward learning by learning features," *The International Journal of Robotics Research*, vol. 41, no. 5, pp. 497–518, 2022. [Online]. Available: https://doi.org/10.1177/02783649221078031

[13] T. Munzer, M. Toussaint, and M. Lopes, "Preference learning on the execution of collaborative human-robot tasks," in *in Proc. of IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 879–885.

[14] E. C. Grigore, A. Roncone, O. Mangin, and B. Scassellati, "Preference-based assistance prediction for human-robot collaboration tasks," in *in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4441–4448.

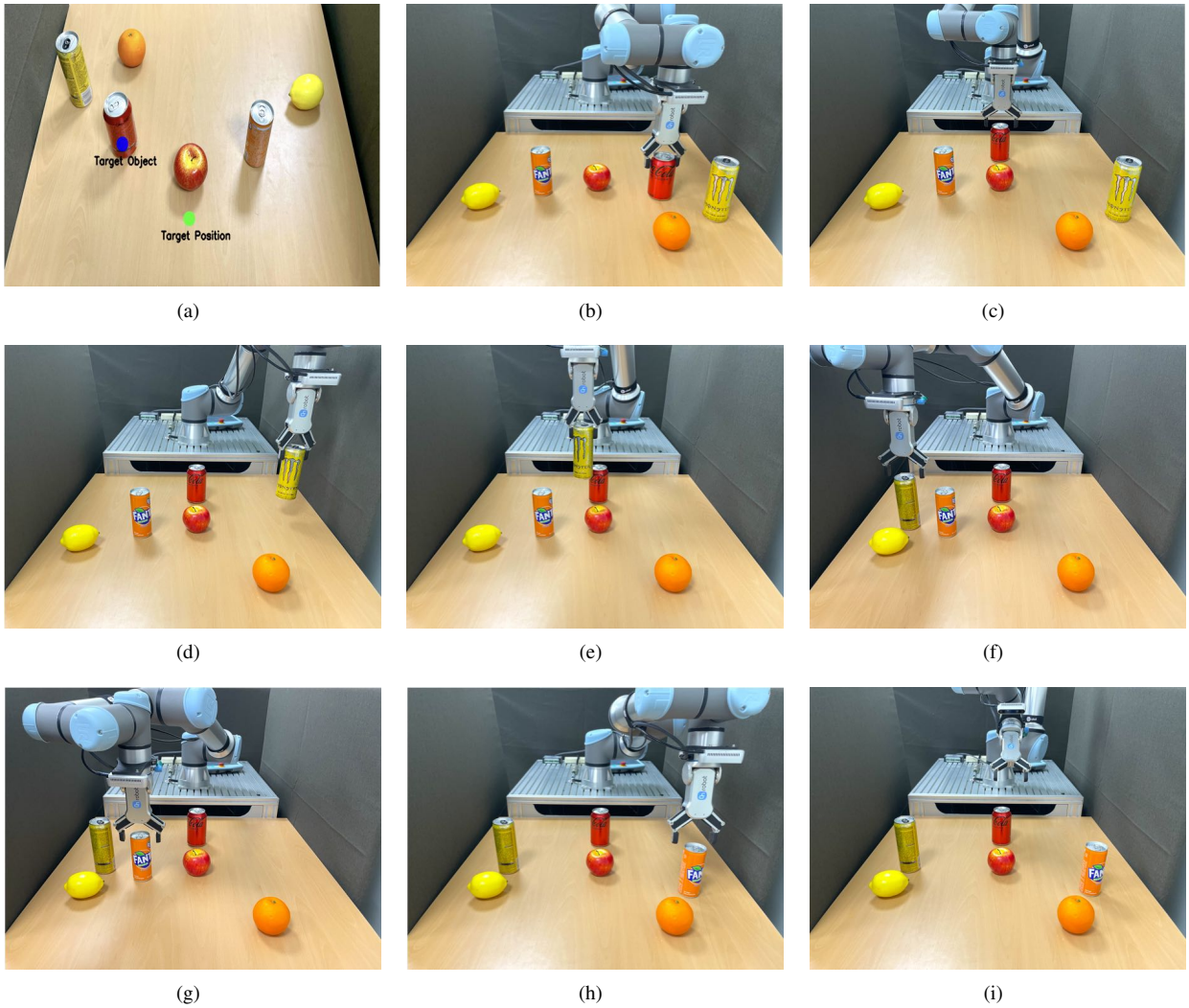
[15] L. Guan, K. Valmeekam, and S. Kambhampati, "Relative behavioral attributes: Filling the gap between symbolic goal specification and reward learning from human preferences," 2023.

[16] Z. Dong, Y. Yuan, J. Hao, F. Ni, Y. Mu, Y. Zheng, Y. Hu, T. Lv, C. Fan, and Z. Hu, "Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model," 2023.

[17] E. Biyik and D. Sadigh, "Batch active preference-based learning of reward functions," in *Conference on robot learning*. PMLR, 2018, pp. 519–528.

[18] M. Tucker, E. Novoseller, C. Kann, Y. Sui, Y. Yue, J. Burdick, and A. D. Ames, "Preference-based learning for exoskeleton gait optimization," 2020.

[19] K. S. Sikand, S. Rabiee, A. Uccello, X. Xiao, G. Warnell, and J. Biswas, "Visual representation learning for preference-aware path



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 6: **Real-world demonstration snapshot.** This snapshot illustrates the application of semantic preference in object rearrangement based on color matching, which is demonstrated through a series of robot actions: (a) shows the user interface as perceived by the user, where the blue circle represents the object to be moved and the green circle represents the desired coordinates, (b)-(c) show the robot placing a red drink next to an apple to align by color, (d)-(f) depict the robot placing a yellow drink in front of a lemon, and (g)-(i) finally show the robot placing an orange drink close to an orange. We demonstrate our method on real robot hardware, with hard-coded procedures for picking and placing objects.

planning," in *in Proc. of IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 303–11 309.

[20] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*. PMLR, 2023, pp. 287–318.

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

[22] M. Kwon, S. M. Xie, K. Bullard, and D. Sadigh, "Reward design with language models," *arXiv preprint arXiv:2303.00001*, 2023.

[23] Y. Tang, W. Yu, J. Tan, H. Zen, A. Faust, and T. Harada, "Saytap: Language to quadrupedal locomotion," 2023.

[24] J. Lee, S. Park, J. Park, K. Lee, and S. Choi, "Spots: Stable placement of objects with reasoning in semi-autonomous teleoperation systems," 2023.

[25] Y. Ding, X. Zhang, C. Paxton, and S. Zhang, "Task and motion planning with large language models for object rearrangement," *arXiv preprint arXiv:2303.06247*, 2023.

[26] J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, and D. Sadigh, "Physically grounded vision-language models for robotic manipulation," *arXiv preprint arXiv:2309.02561*, 2023.

[27] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023.

[28] D. Liu, X. Dong, R. Zhang, X. Luo, P. Gao, X. Huang, Y. Gong, and Z. Wang, "3daxiesprompts: Unleashing the 3d spatial task capabilities of gpt-4v," *arXiv preprint arXiv:2312.09738*, 2023.

[29] W. Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T. L. Chiang, T. Erez, L. Hasenclever, J. Humplik, B. Ichter, T. Xiao, P. Xu, A. Zeng, T. Zhang, N. Heess, D. Sadigh, J. Tan, Y. Tassa, and F. Xia, "Language to rewards for robotic skill synthesis," 2023.

[30] N. Wilde, D. Kulic, and S. L. Smith, "Bayesian active learning for collaborative task specification using equivalence regions," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, p. 1691–1698, Apr. 2019. [Online]. Available: http://dx.doi.org/10.1109/LRA.2019.2897342