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 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. 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]. Commonly, robotic behaviors have mainly relied on manually designed features through scalar values [2]–[4]. 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. The recent advances in Multimodal Large Language Models (MLLMs) have been in the integration of direct sensory perception into the reasoning processes, improving the ability to interpret and generate human-like responses [5]–[7]. Furthermore, MLLMs have achieved human-like reasoning performances in a variety of robotic tasks [8]–[10].

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. Specifically, 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 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. In particular, **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.

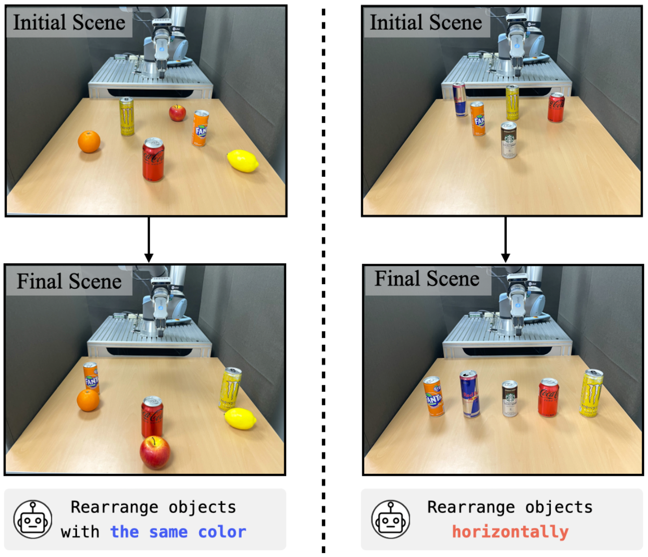

Initial Scene
Final Scene
Rearrange objects with the same color
Initial Scene
Final Scene
Rearrange objects horizontally

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 that provide which object to move and where to place it.

II. Proposed Method

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

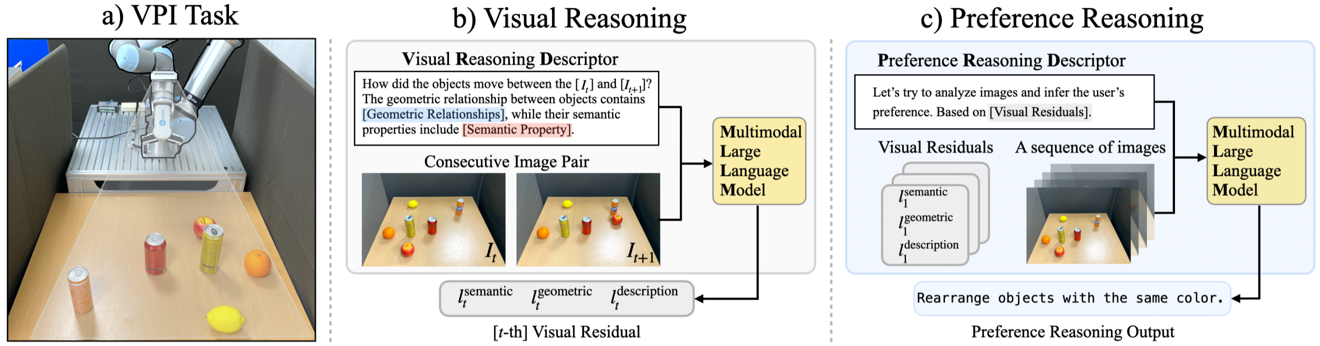
This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00277060, Development of open edge AI SoC hardware and software platform)

[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


a) VPI Task
b) Visual Reasoning
Visual Reasoning Descriptor
Consecutive Image Pair
I_t
I_{t+1}
Multimodal Large Language Model
l_t^{semantic} $l_t^{\text{geometric}}$ $l_t^{\text{description}}$
[t-th] Visual Residual
c) Preference Reasoning
Preference Reasoning Descriptor
Visual Residuals
A sequence of images
l_1^{semantic} $l_1^{\text{geometric}}$ $l_1^{\text{description}}$
Multimodal Large Language Model
Preference Reasoning Output

How did the objects move between the [I_t] and [I_{t+1}]? The geometric relationship between objects contains [Geometric Relationships], while their semantic properties include [Semantic Property].

Let's try to analyze images and infer the user's preference. Based on [Visual Residuals].

```
Rearrange objects with the same color.
```

Fig. 2: **Overview of Chain-of-Visual-Residuals:** CoVR prompting involves generating visual reasoning descriptions of consecutive images and chaining these descriptions for interpreting human preferences from the scene sequences.

color, shape, and category, and the geometric property corresponds to inferring the relative positions between objects, and displacements.

We propose **C**hain-**o**f-Visual-**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. II-A and Preference Reasoning Descriptor (PRD) in Sec. II-B.

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, $l^{\text{geometric}}$ corresponds 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

[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.

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:

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

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.
> ...

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.

III. 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?

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 [6]).

- 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.

Real-world Experiment: Household			
Model	$\mathrm{SR}_{\mathrm{VRD}}$ ↑	$\mathrm{SR}_{\mathrm{PRD}}$ ↑	
		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 I: The number of $\mathrm{SR}_{\mathrm{VRD}}$ (mean±standard deviation) indicates the success rate of predicting visual residuals in between images. The $\mathrm{SR}_{\mathrm{PRD}}$ metric measures the effectiveness of preference reasoning, evaluating the ability of the method to infer user preferences. A higher number indicates better performance.

- MLLM-L2R: Inspired by Language-to-Reward (L2R) [11], 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 [12], [13]. Within the framework of 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 ($\mathrm{SR}_{\mathrm{VRD}}$) and the success rate of Preference Reasoning Descriptor ($\mathrm{SR}_{\mathrm{PRD}}$) for the given image sequences. In particular, $\mathrm{SR}_{\mathrm{VRD}}$ is calculated based on the visual residual between the image sequences and is defined as follows:

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

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, $\mathrm{SR}_{\mathrm{PRD}}$ is measured according to the predicted preference that matches the ground truth. The preference criteria for each scene are manually designed. We formulate $\mathrm{SR}_{\mathrm{PRD}}$ as follows:

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

where the indicator function $\mathbb{I}$ checks for a match between predicted description and ground truth preferences. $\mathrm{SR}_{\mathrm{PRD}}$ evaluates whether the predicted preferences $l_i^{\text{preference}}$ are in

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

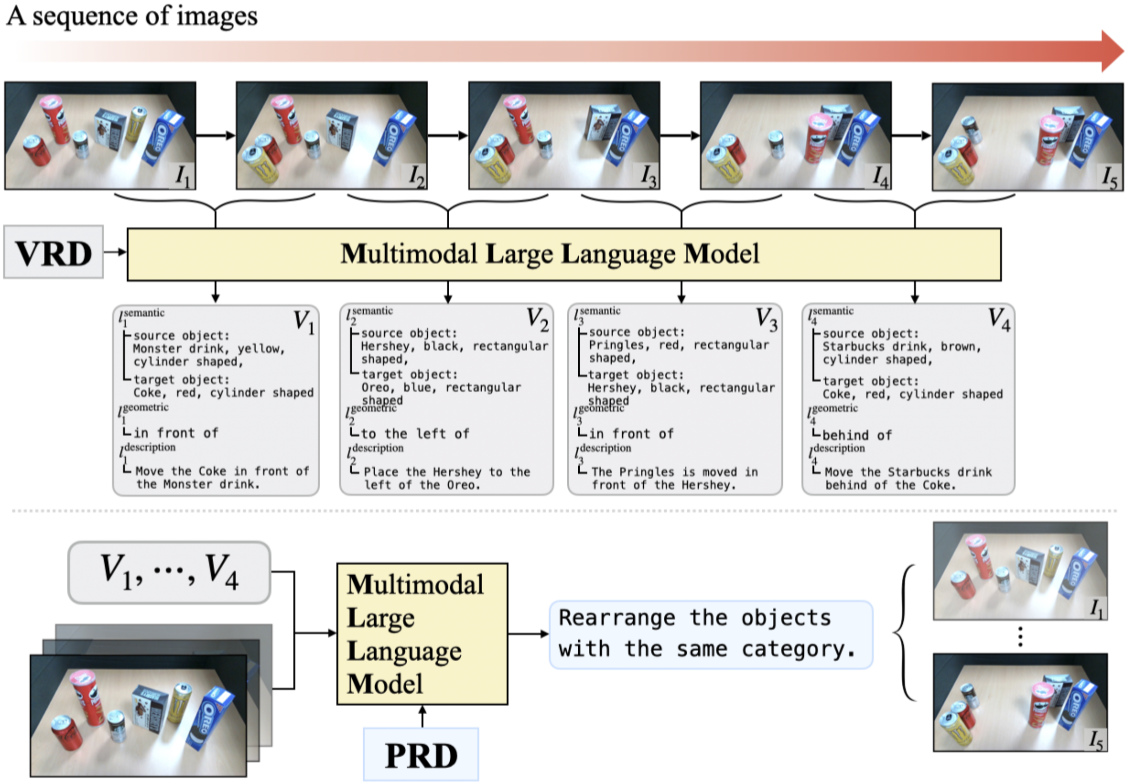

A sequence of images
I_1
I_2
I_3
I_4
I_5
VRD
Multimodal Large Language Model
V_1
l_1^{semantic}
source object: Monster drink, yellow, cylinder shaped,
target object: Coke, red, cylinder shaped
$l_1^{\text{geometric}}$
in front of
$l_1^{\text{description}}$
Move the Coke in front of the Monster drink.
V_2
l_2^{semantic}
source object: Hershey, black, rectangular shaped,
target object: Oreo, blue, rectangular shaped
$l_2^{\text{geometric}}$
to the left of
$l_2^{\text{description}}$
Place the Hershey to the left of the Oreo.
V_3
l_3^{semantic}
source object: Pringles, red, rectangular shaped,
target object: Hershey, black, rectangular shaped
$l_3^{\text{geometric}}$
in front of
$l_3^{\text{description}}$
The Pringles is moved in front of the Hershey.
V_4
l_4^{semantic}
source object: Starbucks drink, brown, cylinder shaped,
target object: Coke, red, cylinder shaped
$l_4^{\text{geometric}}$
behind of
$l_4^{\text{description}}$
Move the Starbucks drink behind of the Coke.
$V_1, \cdots, V_4$
Multimodal Large Language Model
PRD
Rearrange the objects with the same category.
I_1
I_5

Fig. 3: **Running Example of CoVR.** This example 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/

alignment with the ground truth preferences $\hat{l}_i^{\text{preference}}$ across a defined set of scenes.

B. 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 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 I compared the visual reasoning performance of our approach against the ablation of our method, which was evaluated six times respectively. Especially the results of 0.63±0.08 demonstrated the superior visual reasoning ability of our method. In contrast, the MLLM-Naive model showed a limited ability in extracting visual signals between images with SR_{VRD} of 0.28±0.19 for the same task. This result highlights the effectiveness of our VRD template-based approach in recognizing the visual residuals within image sequences.

In the preference reasoning experiment, each type of preference was evaluated six times and performance was measured in terms of SR_{PRD}. As illustrated in Fig. 3, in each step, the robot performed to move objects and captures images. The results of our method in Table I indicate 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.

IV. 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 real-world tabletop environments. 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.

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] 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.

[3] 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.

[4] 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

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

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

[7] 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.

[8] 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.

[9] J. Lee, S. Park, J. Park, K. Lee, and S. Choi, "Spots: Stable placement of objects with reasoning in semi-autonomous teleoperation systems," in *in Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 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] 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.

[12] 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.

[13] 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