

# Supplementary Material

## PACA: Perspective-Aware Cross-Attention Representation for Zero-Shot Scene Rearrangement

Shutong Jin<sup>1\*</sup>, Ruiyu Wang<sup>1\*</sup>, Kuangyi Chen<sup>2</sup>, Florian T. Pokorny<sup>1</sup>  
<sup>1</sup>KTH Royal Institute of Technology, <sup>2</sup>Graz University of Technology  
{shutong, ruiyuw, fpokorny}@kth.se, kuangyi.chen@tugraz.at

### A. User Study

#### A.1. Overview

A user study was conducted to evaluate satisfaction with rearrangement goals generated by two humans, *PACA*, *SG-Bot* [4], and a custom baseline called *Parallel*. The study involved 43 subjects, who rated 3 scenes from Section 5.2.2, providing a total of 559 ratings on a scale from 1 (Very bad) to 5 (Very good). During the study, subjects were asked, “If a robot made this rearrangement for you at home, how satisfied would you be?” The full results, including mean and standard deviations, are presented in *Table 1*. The following sections outline the implementation of the user study and the baseline methods.

#### A.2. Implementation of the User Study

Given the highly modular nature of the scene rearrangement task, this user study focuses solely on analyzing user satisfaction with the generated goals, independent of the matching and robot execution stages. To minimize the influence of users’ ratings based on the visual appearance of the goals, we recreated the object arrangements in the real world to match their positions in the simulated or synthetic images and presented only the unified real-world images to the users. For example, *SG-Bot*’s goals were originally generated in simulation. The corresponding recreated scenes are shown in *Fig. 8*.

For each subject, the study randomly sampled one goal from each baseline for each scene and anonymously numbered the rearrangements. The subject was then asked to rate the numbered images. An example of the user study sent to one subject is shown in *Fig. 1*. Due to the lack of trained models and simulated objects, Fruit and Office scenes are inapplicable to *SG-Bot*.

#### A.3. Implementation of the Baselines

##### A.3.1 Human 1 and Human 2

*Fig. 2* shows the initial setup of the three scenes, with objects randomly scattered on the table. Before each human rearrangement, the scene was reset to this state. The best-effort arrangements were carried out by the first two authors (Human 1 and Human 2), each working independently without seeing the other’s results. The outcomes are shown in *Fig. 3*.

##### A.3.2 PACA

Three goals for each scene were generated using the prompts listed in Section 5.4. An additional set was created with the prompt “fork, knife, plate, table” (excluding “spoon”) to align with the goals generated by *SG-Bot*, where the spoon is absent. The goals for each scene are shown in *Fig. 4*, *Fig. 5*, and *Fig. 6*, respectively.

##### A.3.3 Parallel

*Parallel* is a custom rearrangement baseline in which objects were manually aligned horizontally in random order without overlapping, as shown in *Fig. 7*.

##### A.3.4 SG-Bot

*SG-Bot* generates goals from scene graphs by training Graph-to-3D [2] and AtlasNet [3]. Trained in the PyBullet [1] environment, *SG-Bot* is not a zero-shot method and focuses specifically on Dining scenes, with no trained models or simulated objects for Fruit and Office scenes. Therefore, we only compared their results for the Dining scene. The examples used in the user study are shown in *Fig. 8*. We obtained the goals by running their open-source code for evaluation and demonstration without modifying model weights or inputs. For the user study, we selected

the demonstrations that are most closely aligned with our scenes.

#### A.4. Table

See page 2.

#### A.5. Figures

See page 2 and 3.

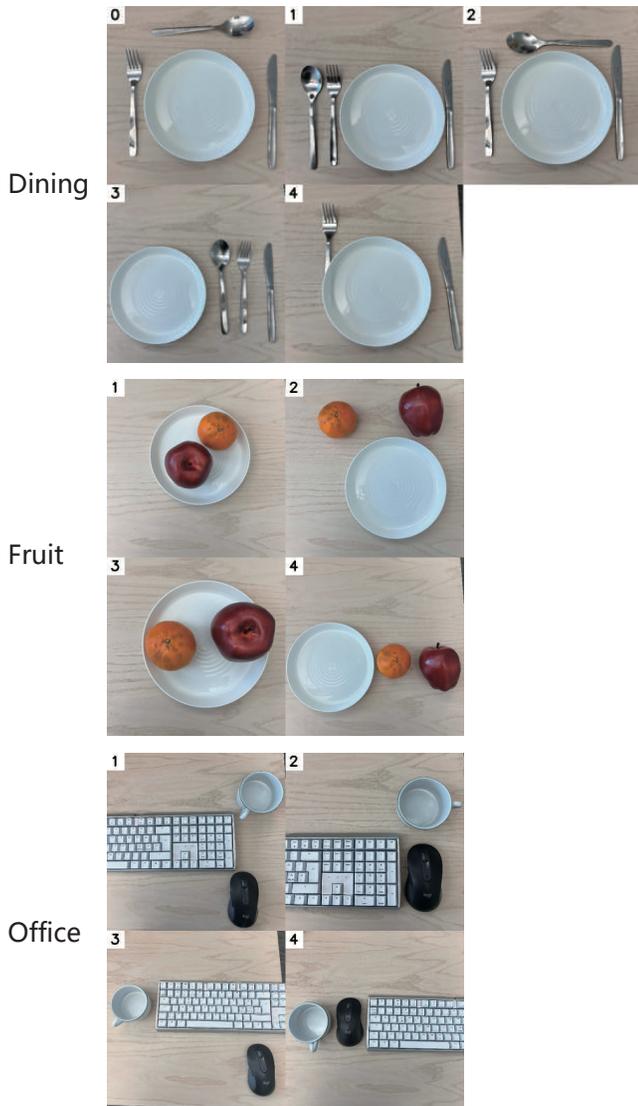


Figure 1. An example of a user study sent to a subject.

#### References

[1] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning, 2016. 1



Figure 2. Initial object states for three scenes.



Figure 3. Rearrangements of three scenes performed by two humans.



Figure 4. Scene Dining of PACA: The first column shows the generated goals and the second column shows the human rearrangements based on those goals.

[2] Helisa Dhama, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipu-

	Human 1	Human 2	<i>PACA</i>	<i>Parallel</i>	SG-Bot
Dining	$3.41 \pm 1.23$	<b><math>4.34 \pm 0.81</math></b>	$3.85 \pm 1.28$	$3.05 \pm 0.85$	$1.68 \pm 0.90$
Fruit	<b><math>4.63 \pm 0.67</math></b>	$1.89 \pm 1.12$	$4.42 \pm 0.82$	$2.26 \pm 1.25$	—
Office	$3.55 \pm 1.32$	<b><math>3.67 \pm 1.27</math></b>	$3.49 \pm 1.23$	$1.95 \pm 0.97$	—

Table 1. Mean and standard deviations of the ratings for *PACA* and four baselines, with the highest values in bold and the second highest underlined.

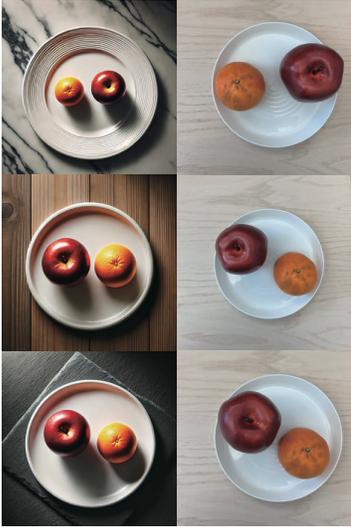


Figure 5. Scene Fruit of *PACA*: The first column shows the generated goals and the second column shows the human rearrangements based on those goals.



Figure 6. Scene Office of *PACA*: The first column shows the generated goals and the second column shows the human rearrangements based on those goals.



Figure 7. Rearrangements of three scenes in *Parallel*.



Figure 8. Scene Dining of SG-Bot: The first column shows the generated goals and the second column shows the human rearrangements based on those goals.

- lation of 3d scenes using scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1
- [3] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 1
- [4] Guangyao Zhai, Xiaoni Cai, Dianye Huang, Yan Di, Fabian Manhardt, Federico Tombari, Nassir Navab, and Benjamin Busam. Sg-bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs. *arXiv preprint arXiv:2309.12188*, 2023. 1