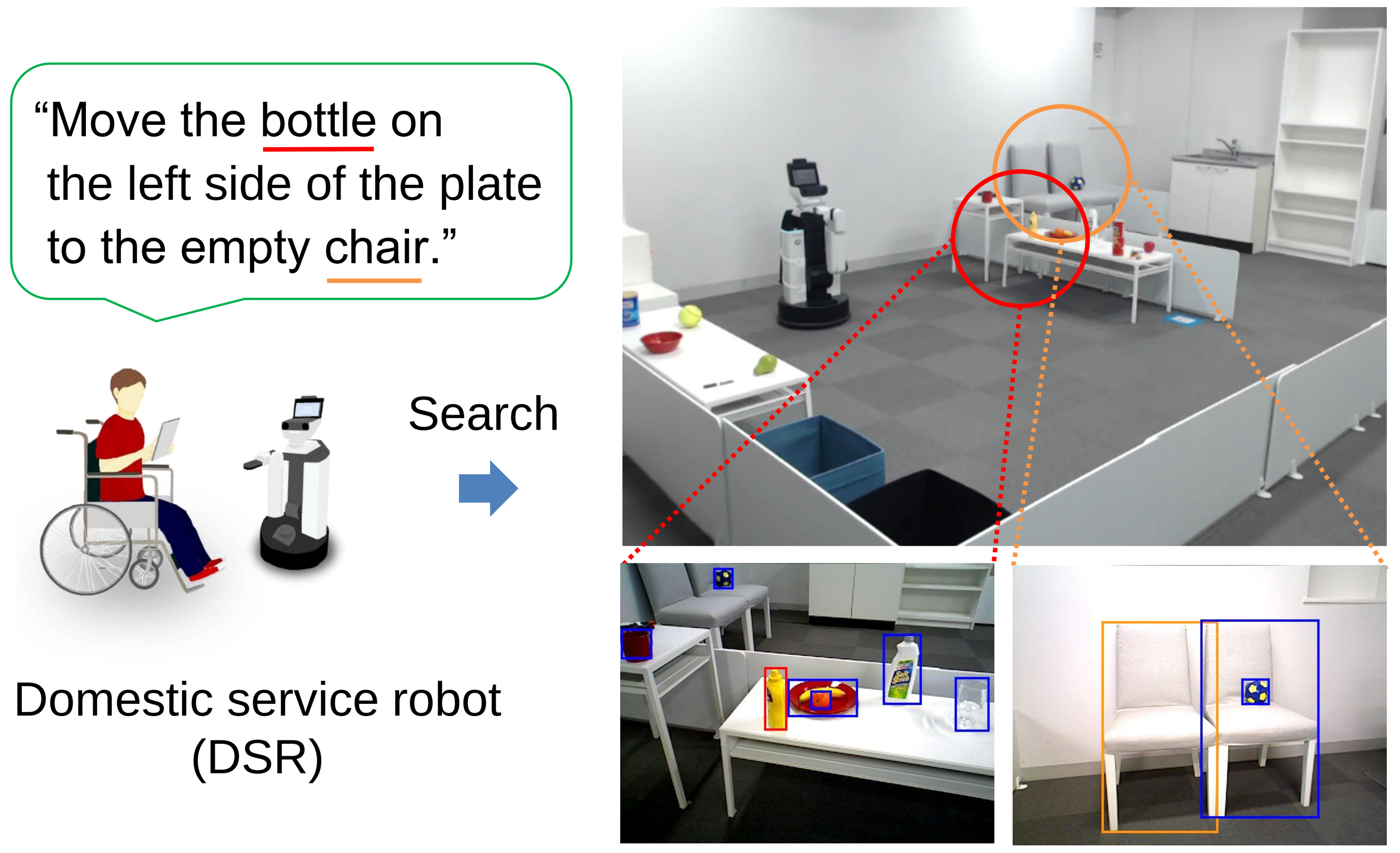
Switching Head–Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks

Ryosuke Korekata, Motonari Kambara, Yu Yoshida, Shintaro Ishikawa, Yosuke Kawasaki, Masaki Takahashi, and Komei Sugiura
(Keio University)

OCTOBER 1 - 5, 2023
IEEE/RSJ International Conference on Intelligent Robots and Systems

Abstract

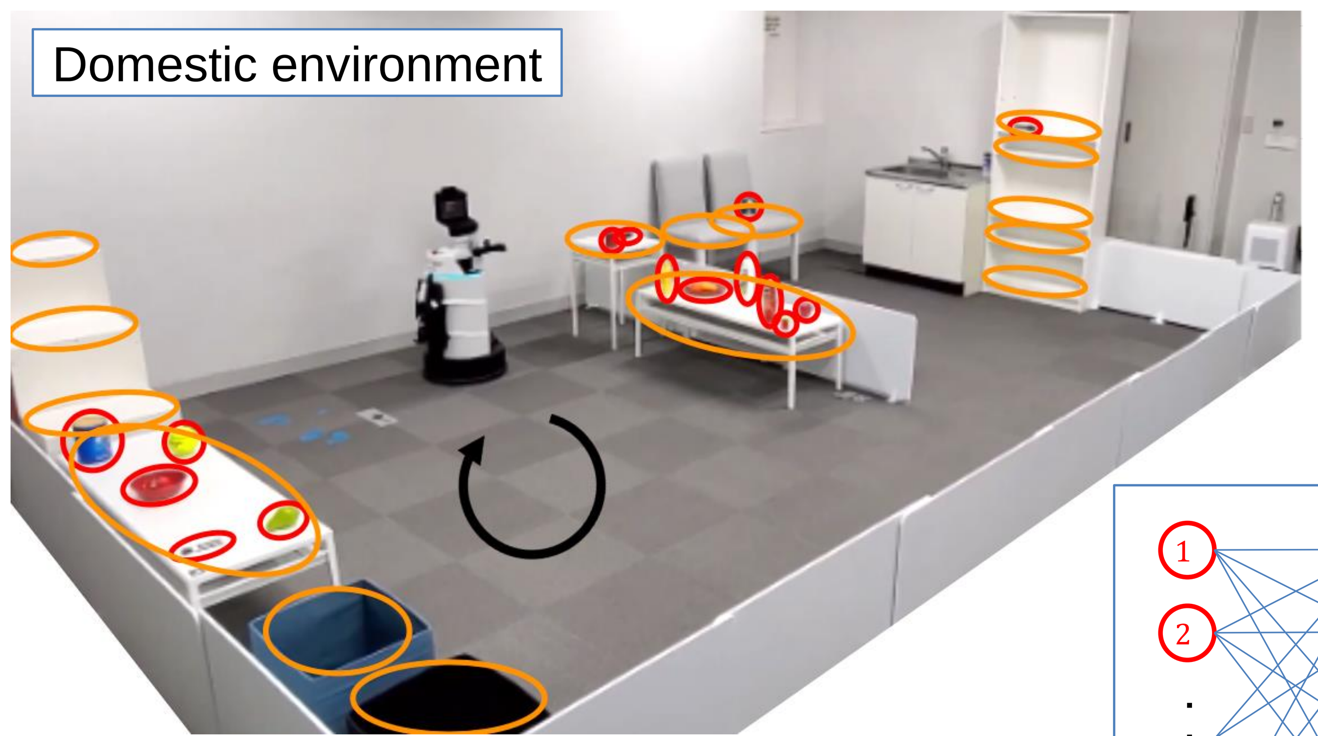
Target task	Multimodal language understanding method that comprehends object fetching and carrying instructions
Novelty	Introduce a **Switching Head–Tail** mechanism so that both target objects and destinations can be predicted individually using a single model
Results	Outperformed the baseline method in terms of language comprehension accuracy on the newly-built dataset and physical experiments

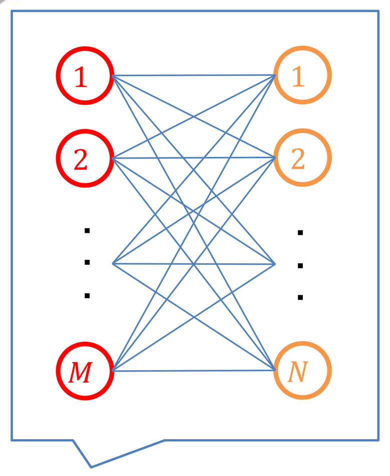


Related Work: Large Computational Cost

MTCM [Magassouba+, RA-L19]	Identifies target object from instruction and whole image
Target-dependent UNITER (TDU) [Ishikawa+, RA-L21]	Introduced the transformer attention mech-anism based on UNITER [Chen+, ECCV20]

■ Goal: Finding the maximum likelihood pair





M: Number of candidate target objects
N: Number of candidate destinations

☹ Time complexity for inference: $O(M \times N)$

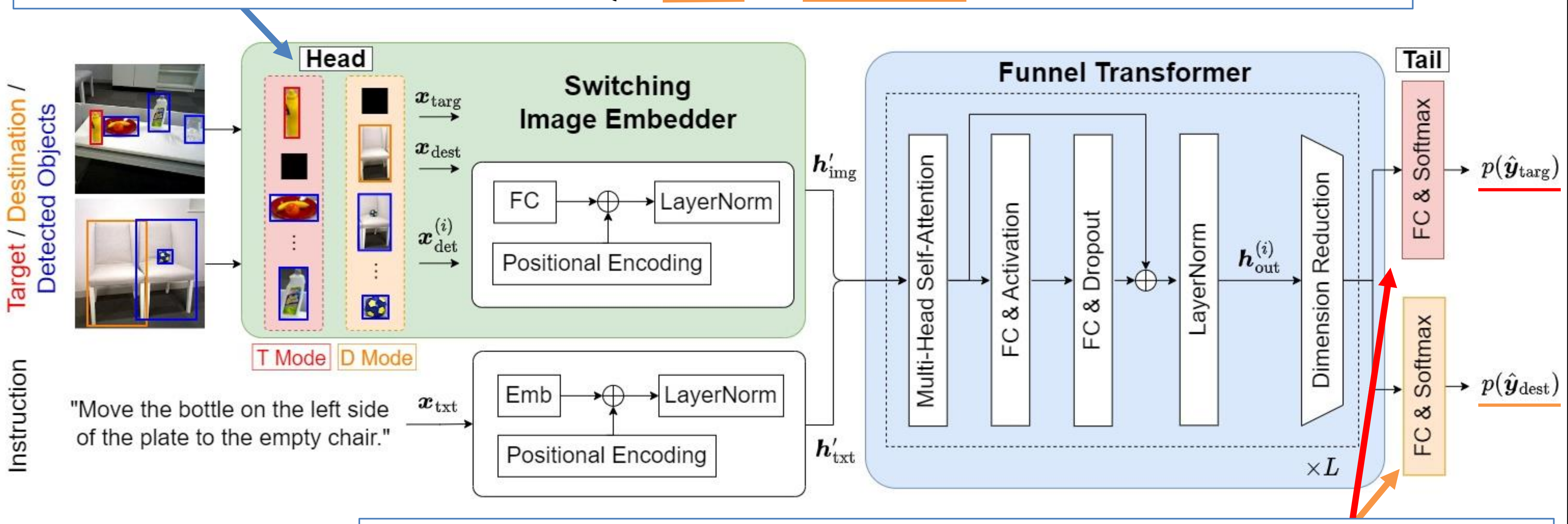
Method: Switching Head–Tail Funnel UNITER (SHeFU)

■ Both target objects and destinations can be predicted individually using a single model, which reduces the computational cost (= ☺ Time complexity for inference: $O(M + N)$)

Step 1: 1 2 · · · j · · · M Step 2: 1 2 · · · k · · · N

Switching Head mechanism: ✓ Conditions the model by partially zero-filling the input

$$(\boldsymbol{x}_{\mathrm{targ}}, \boldsymbol{x}_{\mathrm{dest}}) = \begin{cases} (\boldsymbol{x}_{\mathrm{targ}}, \boldsymbol{0}) & \text{if target mode} \\ (\boldsymbol{0}, \boldsymbol{x}_{\mathrm{dest}}) & \text{if destination mode} \end{cases}$$

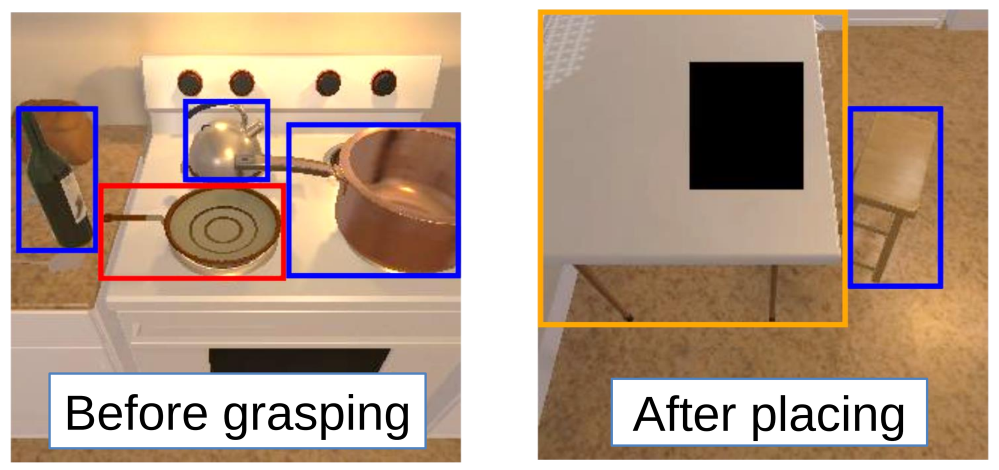


Switching Tail mechanism:
✓ Outputs the predicted probability according to the mode
✓ Multi-task learning: $\mathcal{L} = \lambda_{\mathrm{targ}} \mathcal{L}_{\mathrm{CE}}\left(\boldsymbol{y}_{\mathrm{targ}}, p(\hat{\boldsymbol{y}}_{\mathrm{targ}})\right) + \lambda_{\mathrm{dest}} \mathcal{L}_{\mathrm{CE}}\left(\boldsymbol{y}_{\mathrm{dest}}, p(\hat{\boldsymbol{y}}_{\mathrm{dest}})\right)$

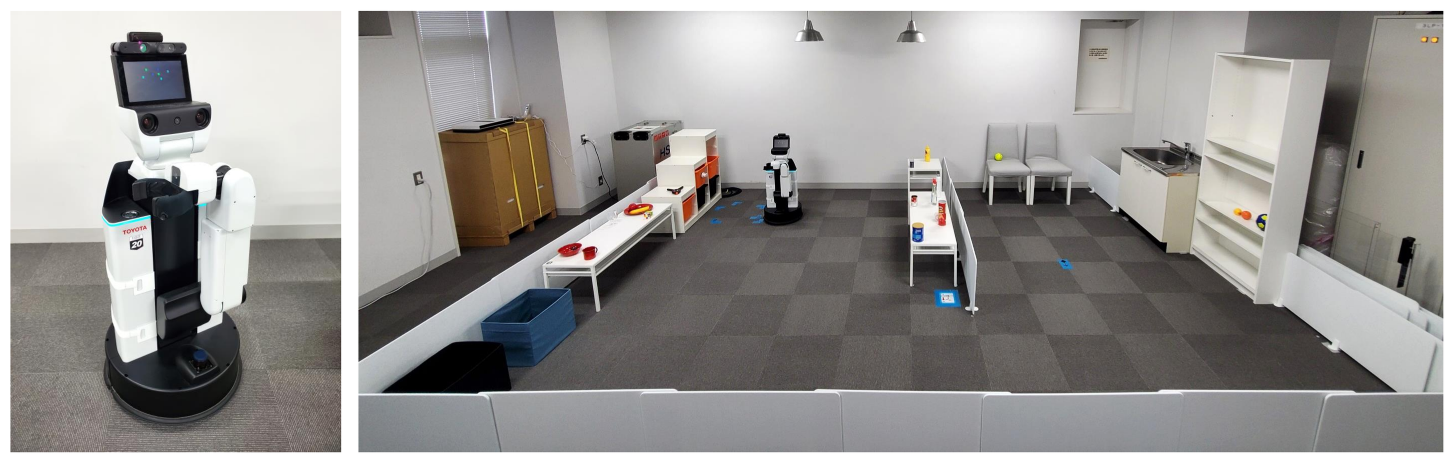
Settings: Simulation and Physical Experiments

1. ALFRED-fc: Based on ALFRED [Shridhar+, CVPR20] (= Standard Vision-and-Language Navigation benchmark)

Dataset size (train : valid : test)	5748 (4420 : 642 : 686)
# Images	1099
# Instructions	3452
# Unique words	646
# Average words	8.4



2. Language comprehension and Grasping/Placing actions (= Heuristic methods)



Robot and environment: WRS [Okada+, AR19]

Objects: YCB [Calli+, RAM15]

Quantitative Results

■ Metric: Language comprehension accuracy [%]

Method	ALFRED-fc	Real
extended TDU [Ishikawa+, RA-L21]	79.4 ± 2.76	52.0
Ours (W/o Switching Head)	78.4 ± 2.05	-
Ours (W/o Switching Tail)	76.9 ± 2.91	-
Ours (SHeFU)	**83.1 ± 2.00**	**55.9**

+3.7 +3.9

■ Metric: Task success rate (SR) [%]

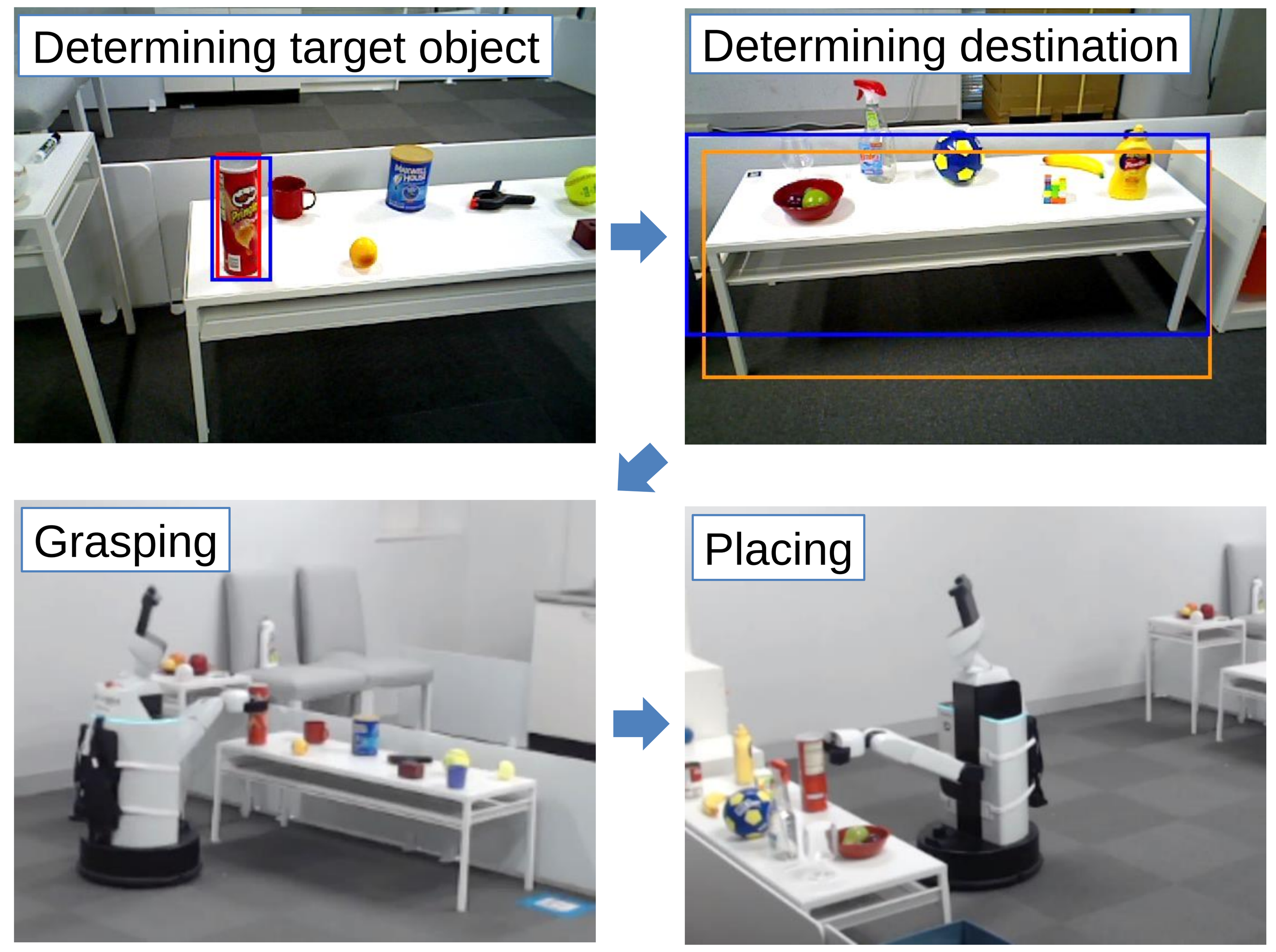
Task	SR↑
Grasping	95 (60/63)
Placing	93 (56/60)

Executed only when language comprehension succeeded

Qualitative Results

■ Successful case in the physical experiments

"Put the red chips can on the white table **with the soccer ball on it**."



References:
[Magassouba+, RA-L19] Magassouba, A., Sugiura, K., Quoc, T. A., & Kawai, H. (2019). Understanding Natural Language Instructions for Fetching Daily Objects Using GAN-Based Multimodal Target–Source Classification. IEEE RA-L, vol.4, no.4, pp.3884-3891.
[Ishikawa+, RA-L21] Ishikawa, S. & Sugiura, K. (2021). Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots. IEEE RA-L, vol.6, no.4, pp.8401-8498.
[Chen+, ECCV20] Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: UNiversal Image-TExt Representation Learning. ECCV, pp.104-120.
[Shridhar+, CVPR20] Shridhar, M., Thomason, J., Gordon, D., et al. (2020). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. CVPR, pp.10740-10749.
[Okada+, AR19] Okada, H., Inamura, T., & Wada, K. (2019). What competitions were conducted in the service categories of the World Robot Summit? Advanced Robotics, vol.33, no.17, pp.900-910.
[Calli+, RAM15] Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., & Dollar, A. (2015). Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set. IEEE RAM, vol.22, no.3, pp.36-52.