Switching Head–Tail Funnel UNITER による 対象物体および配置目標に関する指示文理解と物体操作

Switching Head–Tail Funnel UNITER:
Multimodal Instruction Comprehension for Object Manipulation Tasks

是方 諒介 *1 Ryosuke Korekata
神原 元就 *1 Motonari Kambara
吉田 悠 *1 Yu Yoshida
石川 慎太朗 *1 Shintaro Ishikawa
川崎 陽祐 *1 Yosuke Kawasaki
高橋 正樹 *1 Masaki Takahashi
杉浦 孔明 *1 Komei Sugiura

*1 慶應義塾大学
Keio University

This paper describes a domestic service robot (DSR) that fetches everyday objects and carries them to specified destinations according to free-form natural language instructions. We propose Switching Head–Tail Funnel UNITER, which solves the task by predicting the target object and the destination individually using a single model. We conduct physical experiments in which a DSR delivers standardized everyday objects in a standardized domestic environment as requested by instructions with referring expressions. The experimental results show that our method outperforms the baseline method in terms of language comprehension accuracy and the object grasping and placing actions are achieved with success rates of more than 90%.

1. はじめに

高齢化が進行する現代社会において，日常生活における介助支援の需要は高まっている．これに伴い，在宅介助者不足が社会問題となっており，一つの解決策として被介助者を物理的に支援することが可能な生活支援ロボットに注目が集まっている [Yamamoto 19]．しかし，人間からの自然言語による指示をロボットが理解する能力についてはいまだ不十分である．

本研究では，物体の把持および配置に関する物体操作指示文を生活支援ロボットが理解し実行するための手法の構築を目的とする．図 1 に，提案手法の概要を示す．具体的には，“Move the bottle on the left side of the plate to the empty chair.” という指示文が与えられるとする．このとき，ロボットが周囲の物体および家具の中からボトルを対象物体として，椅子を配置目標として認識したうえで，ボトルを把持して椅子へ配置することが望ましい．

人間の発する指示はしばしば曖昧であり，対象となる物体やその配置目標をロボットが特定することは困難である．実際に，物体操作を含む Vision-and-Language Navigation (VLN) における標準ベンチマークである ALFRED [Shridhar 20] では，人間の精度は 91.0%と報告されている一方，最先端の手法 (e.g., [Inoue 22]) では 46%以下しか達成できていない．

物体操作指示文のためのマルチモーダル言語理解モデルは広く研究されている [Hatori 18, Magassouba 19, Ishikawa 21]．しかし，これらの既存手法に配置目標候補の入力を追加することで本研究で扱うタスクに拡張した場合，計算量の点で非実用的であることが多い．これは，環境中に多数存在する対象物体候補および配置目標候補に関するすべての組合せについて推論を行う必要があるためである．例えば，対象物体候補および配置目標候補がそれぞれ 100 個存在する場合を想定すると，最尤の組を探索するために合計 10000 回もの推論が必要となる．1 回の推論時間を 0.004 秒と仮定すると，ロボットの判断に要する時間が 40 秒と見込まれるため，非実用的である．

連絡先: 是方諒介，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，rkorekata@keio.jp

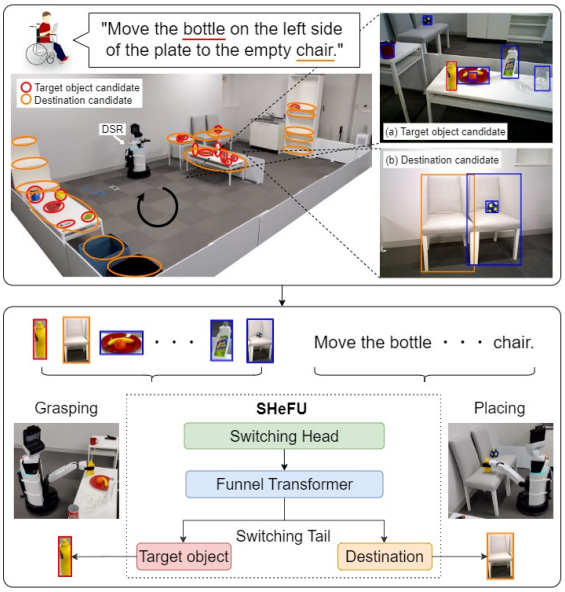


図 1: 提案手法の概要

本論文では，対象物体および配置目標に関する予測を単一モデルで個別に行う方法でタスクを解くことが可能な Switching Head–Tail Funnel UNITER (SHeFU) を提案する．これにより，対象物体候補および配置目標候補がそれぞれ M および N 個存在する状況において，推論回数を $O(M \times N)$ ではなく $O(M + N)$ とすることが可能になる．既存手法と異なる点は，Switching Head–Tail 機構を導入することで，単一モデルで対象物体候補および配置目標候補のどちらも入力として扱う．Switching Head 機構では，対象物体および配置目標を予測するためのパラメータを暗黙的に共有するモデルの条件付けを行う．一方で，Switching Tail 機構ではマルチタスク学習を可能にする．これらにより，対象物体を予測する際に配置目標に関する視覚および言語の情報を活用することが期待され，その逆もまた然りである．また，単一モデルでの学習が可能になるため，別々のモデルを用意する必要がなくなる．

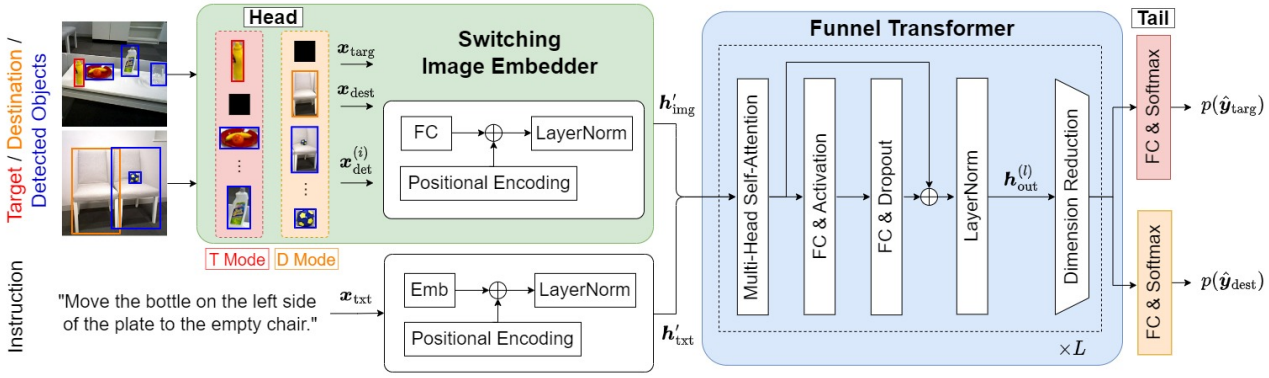


図 2: 提案手法のモデル構造

2. 関連研究

Embodied AI 分野は現在にいたるまで多くの研究が行われており，例えば標準化された家庭環境における生活支援ロボットのベンチマーク競技会 (e.g., World Robot Summit [Okada 19]) は本研究で扱うタスクと関連が深い．ただし，本研究ではこれらの競技会とは異なりテンプレートに基づく指示文は用いない．Embodied AI 分野の代表的なタスクとして，VLN [Anderson 18] や Object Goal Navigation [Fukushima 22] が存在する．ALFRED データセット [Shridhar 20] は，自然言語による指示文とロボットのカメラ画像から，家事タスクにおけるロボットの行動を訓練するためのデータセットである．

マルチモーダル言語処理分野のサーベイ論文として，[Uppal 22] などが挙げられる．画像および言語を扱う分野としては，Referring Expression Comprehension (e.g., [Wang 22])，Referring Expression Segmentation (RES)，Multimodal Language Understanding for Fetching Instructions (MLU-FI) (e.g., [Ishikawa 21])，および Dual Referring Expression Comprehension (DREC) (e.g., [是方 22]) などが挙げられる．本手法は，TDU [Ishikawa 21] のような MLU-FI タスクを扱う手法と関連が深い．

3. 問題設定

本研究で扱うタスクは，Dual Referring Expression Comprehension with fetch-and-carry (DREC-fc) タスクである．DREC-fc タスクは，日用品および家具が写る複数の画像から，参照表現を含む指示文の対象物体および配置目標の両方を特定し，ロボットが対象物体を配置目標まで運搬するタスクである．すなわち，本タスクは言語理解および動作実行という二つのサブタスクに分けられる．本論文で使用する用語は，[是方 22] に従う．入出力を以下のように定義する．

- 入力：指示文，対象物体候補が写る画像，配置目標候補が写る画像
- 出力：対象物体候補および配置目標候補が，対象物体および配置目標にともに一致する確率の予測値 $p(\hat{\boldsymbol{y}})$

transformer [Vaswani 17] などの大規模モデルの訓練には，大量のデータが必要であることが多い．しかし，実機ロボットを用いたデータ収集は人間が物体の配置を行う必要があるため多くの時間を要する．そこで，大量の訓練データを短時間で収集可能なシミュレーション環境で訓練し，実環境へゼロショット転移することでコストの削減を図る．

ロボットの移動，把持，および配置に関する軌道生成はヒューリスティックに行われるものとする．詳細は 5.1 節で後述する．

4. 提案手法

図 2 にモデルの構造を示す．モデル全体は 2 つの主要モジュールから構成され，それぞれ Switching Image Embedder および Funnel Transformer である．本手法は，自然言語によって指示を与えられる fetch-and-carry タスク [Okada 19] と関連が深い．なお，本研究ではテンプレートに基づく指示文は用いない．図において，Target，Destination，Detected Objects，および Instruction はそれぞれ対象物体候補，配置目標候補，画像中の各物体または家具，および指示文を表す．また，“FC”，“Emb”，“$\oplus$”，および角の丸い矢印の合流はそれぞれ全結合層，埋め込み，加算，および連結を示す．

本研究では，2 つの参照表現を含むマルチモーダル言語理解タスクにおいて提案手法を検証する．一方で，本手法における Switching Head–Tail 機構は，3 つ以上の参照表現を含む指示文理解タスクや RES タスクに対しても広く適用可能であると考えられる．また，提案手法は同様の入力であればシミュレーション環境および実機環境のいずれに対しても適用可能である．詳細については，[是方 22] を参照されたい．

4.1 入力

モデルへの入力を $\boldsymbol{x} = \{\boldsymbol{x}_{\text{targ}}, \boldsymbol{x}_{\text{dest}}, \boldsymbol{x}_{\text{txt}}\}$ と定義する．ここに，$\boldsymbol{x}_{\text{targ}} \in \mathbb{R}^{1024}$，$\boldsymbol{x}_{\text{dest}} \in \mathbb{R}^{1024}$，および $\boldsymbol{x}_{\text{txt}} \in \{0,1\}^{D_v \times D_l}$ はそれぞれ対象物体候補の領域，配置目標候補の領域，および指示文を表す．また，D_v および D_l はそれぞれ語彙サイズおよび指示文中のトークン数の最大値を表す．

4.2 Switching Image Embedder

Switching Image Embedder では，Switching Head 機構を用いてモードに応じて入力を切り替えながら対象物体候補，配置目標候補，および画像中の各物体または家具の領域に対する埋め込み処理を行う．ここで，対象物体について予測を行うことを target mode，配置目標について予測を行うことを destination mode と定義する．本モジュールへの入力は，$\boldsymbol{x}_{\text{targ}}$ および $\boldsymbol{x}_{\text{dest}}$ から構成される．

$\boldsymbol{x}_{\text{targ}}$ および $\boldsymbol{x}_{\text{dest}}$ について，以下に示す式により切り替え処理を行う．

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

すなわち，各モードにおいて不要な入力を 0 埋めする．これにより，0 埋めが予測対象を切り替える条件付けとして機能すると期待される．また，target mode においては対象物体候補が写る画像，destination mode においては配置目標候補が写る画像に対して物体検出を行い，画像中の周辺物体または家具

(a) (b)

図 3: (a) HSR および (b) 実機実験で使用した物体

の領域 $\{\boldsymbol{x}_{\text{det}}^{(i)} \in \mathbb{R}^{1024} \mid i = 1, \ldots, K\}$ を得る．ここで，K は Faster R-CNN [Ren 16] により検出された画像中の領域の数を示す．なお，画像特徴量抽出や positional encoding については $\boldsymbol{x}_{\text{targ}}$ および $\boldsymbol{x}_{\text{dest}}$ と同様に行う．

4.3 Funnel Transformer

Switching Tail 機構では，モードに応じて最後のネットワークを切り替える処理を行う．Funnel Transformer モジュールの出力 $\boldsymbol{h}'_{\text{out}}$ について，以下に示す式により対象物体に関する予測確率 $p(\hat{\boldsymbol{y}}_{\text{targ}})$ を得る．

$$p(\hat{\boldsymbol{y}}_{\text{targ}}) = \text{softmax}(f_{\text{FC}}(\boldsymbol{h}'_{\text{out}}))$$

ここで，f_{FC} は全結合層を表す．また，配置目標に関する予測確率 $p(\hat{\boldsymbol{y}}_{\text{dest}})$ についても異なる全結合層を用いて同様に得る．target mode においては $p(\hat{\boldsymbol{y}}_{\text{targ}})$ を，destination mode においては $p(\hat{\boldsymbol{y}}_{\text{dest}})$ をモデル全体の最終的な出力とみなす．各モードにおける予測ラベル $\hat{y}_{\text{targ}}$ または $\hat{y}_{\text{dest}}$ は，予測確率を閾値 0.5 で二値化することで得られる．ただし，Switching tail 機構により，対象物体および配置目標について個別に推論を行う．したがって，以下に示す式により予測ラベル $\hat{y}$ を得る．

$$\hat{y} = \hat{y}_{\text{targ}} \cap \hat{y}_{\text{dest}} \tag{1}$$

5. 実機実験

5.1 設定

提案手法の訓練可能パラメータ数は約 3277 万である．訓練には，メモリ 24GB 搭載の GeForce RTX 3090 およびメモリ 64GB 搭載の Intel Core i9-10900KF を使用した．訓練には約 20 分，推論には約 4×10^{-3} 秒/sample を要した．合計 20000 ステップの訓練のうち，2000 ステップごとに検証集合における精度を測定した．検証集合においてもっとも高い精度を得たときの重みを用いて，実環境へのゼロショット転移を行った．その他のパラメータ設定は，[是方 22] と同様とした．

図 1 に，実機実験で使用した環境を示す．本環境は，家庭内の実環境における片付けタスクのベンチマークである国際的なロボット競技会 World Robot Summit 2020 Partner Robot Challenge/Real Space (WRS2020RS) [WRS 20] の標準環境に基づいている．環境の広さは 6.0×4.0 m^2 である．家具は WRS2020RS に準拠しており，6 種類存在する．家具の配置は図 1 に示された通りであり，収納箱は青色のものと黒色のもの，長机および椅子は同一のものがそれぞれ 2 つずつ存在する．このため，環境中に家具は合計 9 つ存在する．このうち，無作為に 1 つの家具を選択して配置目標として用いた．

実機実験においては，図 3 (a) に示すトヨタ自動車製の生活支援ロボット Human Support Robot (HSR) [Yamamoto 19] を用いた．図 3 (b) に，実機実験で使用した物体を示す．これらの物体は，WRS2020RS で標準物体として指定された YCB オブジェクト [Calli 15] から構成される．上部の 20 種類の物体群および下部の 19 種類の物体群は，それぞれ対象物体および背景オブジェクトとして用いた．対象物体については，[Calli 15] において “Food”，“Kitchen”，“Shape”，および “Task” カテゴリに属する物体から，HSR のエンドエフェクタで把持可能なものを選択した．また，背景オブジェクトについてはそれら以外から無作為に選択した．

表 1: 実機における言語理解精度

手法	精度 [%]
ベースライン手法 [Ishikawa 21]	52.0
提案手法	**55.9**

表 2: 実機におけるタスク成功率

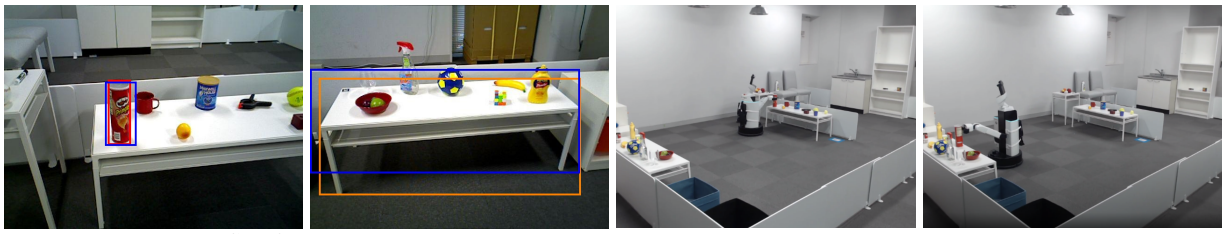
タスク	成功回数 / 試行回数	成功率 [%]
把持	60 / 63	95
配置	56 / 60	93

以下では，実機実験の環境について説明する．本実験では，12 種類の物体配置を作成した．各物体配置において，物体は無作為な位置に配置された．ただし，本実験ではすべての物体が家具の上に配置されていることを前提とする．各試行において，対象物体を図 3 (b) 上部の物体群から，配置目標を図 1 中の家具からそれぞれ無作為に決定した．そのうえで，ロボットに対して “Pick up the apple and put it down on the right-hand chair.” などの指示文を与えた．指示文は英語で，合計 418 文与えられた．

次に，ロボットの動作について説明する．はじめに，事前に定められた 16 個の waypoint をロボットが初期位置から順に巡回することで，環境中の画像収集を行った．ロボットの移動については，事前に与えられた地図を用いて標準的な手法で経路計画を行った．各 waypoint は，ロボットが各家具に正対して複数の視点角度から物体および家具を撮影できるように定めた．画像の取得には，HSR の頭部に搭載された Asus Xtion Pro カメラを用いた．これらの画像を提案手法への入力とし，推論には ALFRED-fc データセット [是方 22] で訓練されたモデルを利用することでゼロショット転移を行った．ロボットの把持動作については，深度画像および矩形領域を基に把持点を決定した．具体的には，深度画像の矩形領域内に対してカメラの内部パラメータを掛け合わせることでカメラ座標系に変換した点群を取得し，各座標軸における中央値を把持点とした．なお，ロボットが把持に成功した場合のみ配置動作を行うものとした．ロボットの配置動作については，家具を撮影した waypoint を利用しルールベースで行った．

5.2 定量的結果

表 1 および表 2 に，実機実験の定量的結果を示す．評価指標として，言語理解精度およびタスク成功率 $\text{SR} = \frac{N_{\text{success}}}{N_{\text{attempts}}}$ を用いた．ここで，N_{attempts} および N_{success} はそれぞれ試行回数および成功回数を表す．なお，言語理解性能を評価するには負例も必要であるため，ALFRED-fc データセットと同様の前処理により負例を作成した．正例と負例に偏りがないため，このような場合に標準的な精度を評価指標として採用した．ただし，提案手法は対象物体候補および配置目標候補について個別に推論を行う点でベースライン手法と異なるため，正解ラベルを $y = y_{\text{targ}} \cap y_{\text{dest}}$ と定義することで統一的に評価した．ここで，$y_{\cdot}$ は対象物体候補または配置目標候補がそれぞれの



指示文：“Put the red chips can on the white table with the soccer ball on it.”

図 4: 実機における定性的結果

ground truth に一致するかの真偽値を表す．

言語理解におけるベースライン手法として，TDU [Ishikawa 21] を DREC-fc タスクへ拡張した手法を用いた．これは，TDU は DREC-fc タスクと関連の深い MLU-FI タスクにおいて良好な結果が報告されているためである．ベースライン手法では，対象物体候補および配置目標候補の両方が入力に含まれる．

表 1 に，実機における言語理解精度を示す．表 1 より，ベースライン手法は精度が 52.0%であるのに対し，提案手法は 55.9%であり，提案手法が 3.9 ポイント上回った．

表 2 に，実機における把持および配置タスクの成功率を示す．なお，言語理解タスクにおいて予測が True Positive であった場合のみ，ロボットに把持および配置動作を行わせた．本研究は把持および配置動作について学習に基づく新規手法を提案するものではないが，表 2 から，実機において言語理解と動作実行を統合可能であることが示唆される．

5.3 定性的結果および考察

図 4 に，定性的結果として成功例を示す．左から順に，対象物体候補，配置目標候補，物体把持動作，および配置動作を示す．赤色，橙色，および青色の矩形領域はそれぞれ対象物体の ground truth，配置目標の ground truth，および対象物体候補または配置目標候補を表す．

図 4 の例において，対象物体および配置目標はそれぞれ赤いポテトチップス缶およびサッカーボールの乗った白いテーブルである．対象物体候補および配置目標候補はともに ground truth に一致しているため，$(y_{\text{targ}}, y_{\text{dest}}) = (1, 1)$ が成り立つ．この例に対して，提案手法は $(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}}) = (1, 1)$ であると正しく予測した．その上で，ロボットが正確にポテトチップス缶を把持し，机への配置に成功した．

典型的なシーンにおいて，対象物体候補および配置目標候補はそれぞれ平均して 73 個および 89 個検出された．1 回の推論時間は約 4×10^{-3} 秒であるため，ベースライン手法および提案手法の計算時間はそれぞれ 26 秒（6497 回の推論）および 0.6 秒（162 回の推論）であると考えられる．

6. おわりに

本研究では，複数の画像から指示文の対象物体および配置目標の両方を特定し，ロボットが対象物体を配置目標まで運搬する DREC-fc タスクを扱った．本研究の貢献を以下に示す．

- 対象物体候補および配置目標候補がそれぞれ M および N 個存在する場合に，推論回数を $O(M \times N)$ ではなく $O(M + N)$ とすることが可能な SHeFU を提案した．
- Switching Head–Tail 機構を導入することで，対象物体および配置目標について，単一モデルで個別に予測することを可能にした．
- 実機ロボットを用いて実験を行い，SHeFU がベースライン手法を言語理解精度で上回った．さらに，実機において言語理解と動作実行を統合可能であることを示した．

謝辞

本研究の一部は，JSPS 科研費 20H04269，JST CREST，NEDO の助成を受けて実施されたものである．

参考文献

[Anderson 18] Anderson, P., et al.: Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments, in *CVPR*, pp. 3674–3683 (2018)

[Calli 15] Calli, B., Walsman, A., et al.: Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set, *IEEE RAM*, Vol. 22, No. 3, pp. 36–52 (2015)

[Fukushima 22] Fukushima, R., Ota, K., Kanezaki, A., Sasaki, Y., and Yoshiyasu, Y.: Object Memory Transformer for Object Goal Navigation, in *ICRA*, pp. 11288–11294 (2022)

[Hatori 18] Hatori, J., Kikuchi, Y., Kobayashi, S., et al.: Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions, in *ICRA*, pp. 3774–3781 (2018)

[Inoue 22] Inoue, Y. and Ohashi, H.: Prompter: Utilizing Large Language Model Prompting for a Data Efficient Embodied Instruction Following, *arXiv preprint arXiv:2211.03267* (2022)

[Ishikawa 21] Ishikawa, S. and Sugiura, K.: Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots, *IEEE RA-L*, Vol. 6, No. 4, pp. 8401–8408 (2021)

[Magassouba 19] Magassouba, A., Sugiura, K., et al.: 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 (2019)

[Okada 19] Okada, H., Inamura, T., and Wada, K.: What competitions were conducted in the service categories of the World Robot Summit?, *AR*, Vol. 33, No. 17, pp. 900–910 (2019)

[Ren 16] Ren, S., He, K., et al.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *IEEE Trans. PAMI*, Vol. 39, No. 6, pp. 1137–1149 (2016)

[Shridhar 20] Shridhar, M., Thomason, J., Gordon, D., et al.: ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks, in *CVPR*, pp. 10740–10749 (2020)

[Uppal 22] Uppal, S., Bhagat, S., et al.: Multimodal Research in Vision and Language: A Review of Current and Emerging Trends, *Information Fusion*, Vol. 77, pp. 149–171 (2022)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention Is All You Need, *NeurIPS*, Vol. 30, (2017)

[Wang 22] Wang, P., et al.: OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, in *ICML*, pp. 23318–23340 (2022)

[WRS 20] World Robot Summit 2020 Partner Robot Challenge Real Space Rules & Regulations (2020)

[Yamamoto 19] Yamamoto, T., Terada, K., Ochiai, A., Saito, F., Asahara, Y., and Murase, K.: Development of Human Support Robot as the research platform of a domestic mobile manipulator, *ROBOMECH Journal*, Vol. 6, No. 1, pp. 1–15 (2019)

[是方 22] 是方諒介, 吉田悠 他：物体操作タスクにおける Switching Funnel UNITER による対象物体および配置目標に関する指示文理解, 第 40 回日本ロボット学会学術講演会, 4F3-05 (2022)