# PORTER: 最適輸送を用いた Polygon Matching に基づく参照表現セグメンテーション

九曜克之　飯岡雄偉　杉浦孔明
慶應義塾大学
{katsukuyo,kmngrd1805,komei.sugiura}@keio.jp

## 概要

自然言語による指示で生活支援ロボットが操作できれば便利であるが，現状，生活支援ロボットの命令文理解性能は十分ではない．そこで，本研究では，複数の参照表現を含む命令文および画像から動作対象の物体のマスクを予測するタスクを扱う．既存研究では，頂点の順番は異なるが同じ多角形を表す場合を区別して扱っているため，正解と近い多角形を予測しているにも関わらず，異なる多角形と判断してしまい，不適切な学習を促してしまうという問題がある．そこで本論文では，複雑な参照表現を含む命令文から対象物のマスクを生成するセグメンテーションモデルを提案する．本手法の新規性は，最適輸送を用いた Polygon Matching Loss の導入である．命令文，室内環境の画像，対象物のマスクで構成されるデータセットによる評価の結果，提案手法は標準的な評価尺度である mean IoU において，ベースライン手法を 10.41 ポイント上回った．

## 1　はじめに

現代社会で高齢化が進む中，日常生活における介助支援の重要性が高まっているが，その介助を担う在宅介助者は不足している．この解決策として，被介助者に物理的な支援が可能な生活支援ロボットが注目されている [1]．生活支援ロボットは，被介助者からの自然言語による指示で物体の把持や移動に関する操作を行うことが期待されている．しかし，現状ではその命令文理解性能は不十分である [2]．

本研究では，物体操作に関する命令文が与えられた際, 対象物のマスクを生成する Object Segmentation from Manipulation Instructions (OSMI) タスクを扱う．例えば，“Go into the living room and give me the pillow nearest the plant.” という命令文が与えられたとき，植物に最も近い枕のマスクを生成することが望ましい．本タスクはロボットによる物体把持において重要である．なぜなら，ロボットが物体を把持する場面では，その形状や位置を特定することが重要であり，マスクによる把持物体の領域予測のほうが，矩形領域による予測よりも望ましいからである．

Instruction : “Go into the living room and give me the pillow nearest the plant.”

図 1: OSMI タスクの具体例

OSMI タスクと関係が深いタスクとして，Referring Expression Segmentation （RES）タスク [3] がある．RES タスクに比べ，本研究で扱う OSMI タスクでは，ナビゲーション命令から始まる 2 文以上の命令文が多く，対象物を修飾する文が複数含まれている場合があり，単純な RES タスクよりも困難である．例えば，複数の物体の上にキャンドルがおいてある画像に対して，命令文 “Go to the table. And pick up the candle on the right.” が与えられているとする．その際，“the candle on the right” のみでは対象物体を特定できない可能性がある．この例では，“the table” が対象物を間接的に修飾しているため，その表現の理解が重要になる．

RES タスクを扱う既存研究は多く存在する [4–8]．しかし，これらの既存手法では，部分的なマスクしか生成されない場合や，複雑な参照表現を理解できず，対象と同カテゴリの異なる物体のマスクを生成する場合があるため，OSMI タスクのためのモデル

としては不十分である．

多角形に基づくマスクを用いて予測する手法である SeqTR [7] は，対象物を表す多角形の頂点を予測するため，画素単位でマスクを予測する手法よりも断片的なマスクを生成する可能性は低い．しかし，頂点の順番は異なるが同じ多角形を表す場合を区別して扱っているため，正解と近い多角形を予測しているにも関わらず，異なる多角形と判断し，非効率的な学習を促してしまう問題がある．

そこで，本研究では複雑な参照表現を含む命令文から対象物のマスクを生成するセグメンテーションモデル **P**olygon **O**ptimal t**R**ansport **T**ransform**ER** (PORTER) を提案する．既存手法との主要な違いは，最適輸送を用いた Polygon Matching Loss を導入する点である．これにより，予測マスクの頂点順序に関係なく正解マスクに近づくような学習が期待される．

本研究の独自性は以下である．

- 多角形において頂点の順番が異なっても同じ多角形を表す場合を扱うために，最適輸送を用いた Polygon Matching Loss を導入する．
- 大規模言語モデルを用いて，命令文に対して対象物に関する修飾関係を明瞭にする言い換えを行う Paraphraser を導入する．

## 2 問題設定

本研究で扱う OSMI タスク [6] では，命令文が指す対象物に対して，マスクを生成することが望ましい．図 1 に本タスクの具体例を示す．例えば，“Go into the living room and give me the pillow nearest the plant.” という命令文が与えられた際，紫色の領域で示すマスクの生成を目標とする．入力は画像および命令文である．出力は対象物に対するマスクである．本研究では，対象物は 1 つであることを前提とする．評価尺度として，RES において標準的である mean IoU(mIoU)，Precision@$k$(P@$k$) を用いる．

## 3 提案手法

提案手法は多角形に基づくマスクを扱う手法 [7] を拡張した，OSMI タスクを扱うモデルである．既存手法では，頂点の順番が異なるが同じ多角形を表す場合を考慮していない．このような場合に対応するため，提案手法は最適輸送に基づく Polygon Matching Loss を導入する．これは，頂点集合以外の追加情報を必要としないため，提案手法は既存の多角形に基づくマスクを扱う手法一般に適用可能であると考えられる．

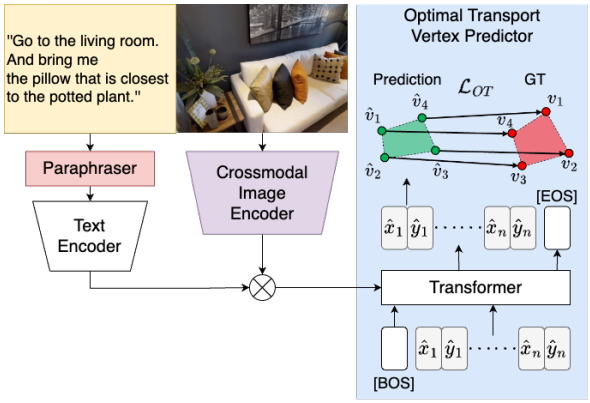


図 2: 提案手法のネットワーク構造．

既存手法との主な違いは以下である．

- 多角形において頂点の順番が異なっても同じ多角形を表す場合を扱うために，頂点集合全体として正解マスクに近づける最適輸送に基づく Polygon Matching Loss を導入する．
- 大規模言語モデルを用いて，命令文に対して対象物に関する修飾関係を明瞭にする言い換えを行う Paraphraser を導入する．

図 2 に提案手法のモデル構造を示す．提案手法は大きく分けて Paraphraser, Image Encoder および Optimal Transport Vertex Predictor の 3 つのモジュールから構成される．

ネットワークの入力 $\boldsymbol{x}$ は $\boldsymbol{x} = \{\boldsymbol{x}_{\text{img}}, \boldsymbol{x}_{\text{inst}}\}$ である．ここに，$\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{H\times W\times 3}$, $\boldsymbol{x}_{\text{inst}} \in \{0,1\}^{v\times l}$ はそれぞれ画像，命令文を表す．また，$H$, $W$, $v$ および $l$ はそれぞれ画像の高さ，幅，命令文の語彙サイズおよび最大トークン数を表す．

### 3.1 Paraphraser

OSMI タスクでは，対象物を修飾する文が複数含まれている場合があるが，既存手法はこのような場合を考慮していない．そのため，Paraphraser は命令文に対して対象物に関する修飾関係を明瞭にする言い換えを行う．入力は $\boldsymbol{x}_{\text{inst}}$ であり，出力は言い換えを行った命令文 $\boldsymbol{x}_{\text{p-inst}}$ である．本モジュールでは，大規模言語モデルである gpt-3.5-turbo [9] を用いる．例として，$\boldsymbol{x}_{\text{inst}}$ が “Go to the dining table. Then pick up the candle on the right.” であるとき，gpt-3.5-turbo から “#Pick up the right candle on the dining table.#” とい

う文を得る．次に文頭と文末の “#” を除去することで，最終的に $\boldsymbol{x}_{\text{p-inst}}$ として “Pick up the right candle on the dining table.” を得る．その後，OpenAI が提供する text-embedding-ada-002 [10] を用いて $\boldsymbol{x}_{\text{p-inst}}$ から言語特徴量 $L \in \mathbb{R}^C$ を抽出した．ここに $C$ は $L$ の次元数を示す．

### 3.2 Image Encoder

実世界のオブジェクトについて，単一の解像度で特徴を抽出するだけでは，小さなオブジェクトや遠くのオブジェクトに対する性能が劣化する場合がある．Image Encoder では，このような場合を扱うために，異なる解像度を組み合わせて画像特徴の抽出を行う．本モジュールは，DarkNet-53 [11] に基づく畳み込みネットワークで構成される．入力は $\boldsymbol{x}_{\text{img}}$ であり， 出力は解像度の違う $M_V$ 種類の中間層における画像特徴量 $\{V_i\}_{i=1}^{M_V}$ である．ここに $V_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ であり，$H_i$，$W_i$，$C_i$ は $V_i$ の画像サイズおよびチャネル数を示す．得られた $V_i$ に畳み込みを行い，チャネル方向に結合することで，最終的な画像特徴量 $V \in \mathbb{R}^{H_V \times W_V \times C_V}$ を得る．$L$ および $V$ からマルチモーダル特徴量 $h_{MM} = \tanh(V) \odot \tanh(L)$ を得る．ただし，$\odot$ は要素ごとの乗算を表す．

### 3.3 Optimal Transport Vertex Predictor

多角形に基づくマスクを扱う既存手法 [7, 8] では，頂点の順番が異なっても同じ多角形を表す場合を考慮できていない．Optimal Transport Vertex Predictor (OTVP) では，このような場合を扱うために最適輸送を用いてマッチングを行い，多角形の頂点を予測する．このような場合を扱うために最適輸送を用いてマッチングを行い，多角形の頂点を予測する．

OTVP は，$h_{MM}$ および頂点の埋め込み表現 $E$ を入力とし，予測マスクの頂点集合を $\hat{\mathcal{Y}} = \{\hat{\boldsymbol{v}}_i \in \mathbb{R}^2\}_{i=1}^{M}$ を出力する．ここに，$\hat{\boldsymbol{v}}_i$ および $M$ はそれぞれ多角形を構成する頂点の座標および頂点数を表す．

OTVP は transformer encoder および transformer decoder で構成される．transformer encoder は $n_{\text{enc}}$ 層の transformer layer から構成される．$h_{MM}$ を transformer encoder に入力し，$h_{\text{enc}} \in \mathbb{R}^{H_{MM} W_{MM} \times d_{\text{enc}}}$ が得られる．ここで，$d_{\text{enc}}$ は次元数を示す．transformer decoder は $n_{\text{dec}}$ 層の transformer layer から構成される．$h_{\text{dec}}$ 及び $E$ を transformer decoder に入力し，$h_{\text{dec}} \in \mathbb{R}^{2M \times d_{\text{dec}}}$ が得られる．ここで，$d_{\text{dec}}$ は次元数を示す．最終的に，$h_{\text{dec}}$ に線形変換を行うことで $\hat{\mathcal{Y}}$ が得られる．

表 1: 各手法における定量的結果．

| method | mIoU [%] | P@0.5 [%] | P@0.7 [%] |
|---|---|---|---|
| MDSM [6] | 24.36 ± 3.87 | 22.49 ± 5.46 | 13.71 ± 3.34 |
| LAVT [4] | 28.16 ± 2.85 | 26.46 ± 4.01 | 18.75 ± 3.29 |
| Ours | **38.57** ± **2.77** | **48.96** ± **3.04** | **26.15** ± **1.90** |

提案する最適輸送に基づくポリゴンマッチングでは，予測マスクの頂点集合と正解マスクの頂点集合とのマッチングのために最適輸送問題を解く．すなわち，予測マスクの頂点集合 $\hat{\mathcal{Y}}$ と正解マスクの頂点集合 $\mathcal{Y} = \{\boldsymbol{v}_i \in \mathbb{R}^2\}_{i=1}^{M}$ の 2 つが与えられたとき，最小の輸送コストで $\hat{\mathcal{Y}}$ を $\mathcal{Y}$ に移動させる輸送計画を求める．$\hat{\mathcal{Y}}$ および $\mathcal{Y}$ を 2 つの離散分布 $\alpha = \sum_{i=1}^{M} \boldsymbol{a}_i \delta_{\hat{\boldsymbol{v}}_i}$ および $\beta = \sum_{j=1}^{M} \boldsymbol{b}_j \delta_{\boldsymbol{v}_j}$ とみなす．ここに，$\delta_{\boldsymbol{v}_i}$ は $\boldsymbol{v}_i$ を中心とするディラックのデルタ関数を表す．重みベクトル $\boldsymbol{a}$ および $\boldsymbol{b}$ は，$\sum_{i=1}^{M} \boldsymbol{a}_i = \sum_{j=1}^{M} \boldsymbol{b}_j = 1$ を満たす．このとき，$\hat{\mathcal{Y}}$ と $\mathcal{Y}$ の間の最小輸送コスト，すなわち $\mathcal{L}_{\text{OT}}$ は次のように定義される．

$$\mathcal{L}_{\text{OT}} = \min_{\boldsymbol{P} \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})} \sum_{i=1}^{M} \sum_{j=1}^{M} C(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_j) \boldsymbol{P}_{ij} \tag{1}$$

ここに，

$$\mathcal{U}(\boldsymbol{a}, \boldsymbol{b}) = \left\{ \boldsymbol{P} \in \mathbb{R}^{M \times M} | \boldsymbol{P}_{ij} \geq 0, \boldsymbol{P}\mathbf{1}_M = \boldsymbol{a}, \boldsymbol{P}^\top \mathbf{1}_M = \boldsymbol{b} \right\} \tag{2}$$

であり，$C$ および $\boldsymbol{P}_{ij}$ は $\hat{\boldsymbol{v}}_i$ から $\boldsymbol{v}_i$ への輸送コストおよび輸送計画である．また，$\mathbf{1}_M$ は成分が全て 1 である $M$ 次元ベクトルを表す．$C$ は $C\left(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_j\right) = \|\hat{\boldsymbol{v}}_i - \boldsymbol{v}_j\|_2$ であり，$\|\cdot\|_2$ は $L^2$ ノルムを表す．本研究では，式 (1) を効率的に計算するためにエントロピー正則化を行い，Sinkhorn アルゴリズム [12] を用いる．

## 4 実験設定

本研究では，SHIMRIE データセット [6] に基づき，SHIMRIEv2 データセットを構築した．SHIMRIE データセットには，指示文，対象物に関連する画像，対象物の画素単位のマスクが含まれている．SHIMRIEv2 データセットを構築したのは，SHIMRIE データセットには次の 2 つの問題があるためである．(i) 多角形に基づくマスクが必要である．(ii)SHIMRIE データセットには，十分な精度のマスクを持たないサンプルが含まれている．(i) の問題に対して，SHIMRIEv2 データセットでは画素単位のマスクに加えて多角形に基づくマスクを導入した．(ii) の問題に対して，この原因は SHIMRIE データセットが Matterport3D データセット [13] に含まれ

“Take down the photo collage of flowers above the black shelf stand in the hallway with a basket of sports balls.”

図 3: 成功例の定性的結果. (a) LAVT [4], (b) MDSM [6], (c) 提案手法, (d) 正解マスク画像.

るボクセルレベルのクラス情報と，REVERIE データセット [2] に含まれる対象物を囲む矩形領域を用いて, 半自動的に対象物の 2 次元セグメンテーションマスクを抽出していたためだと考えられる．そこで，SHIMRIEv2 データセットでは手動で精緻化したマスクを導入した.

SHIMRIEv2 データセットには，4,341 枚の画像に対応する，11,371 の命令文および対象物のマスクのペアが含まれている. 命令文の語彙サイズは 3,558, 全単語数は 196,541 語，平均文長は 18.8 である. 本研究では，計算量削減のため，640 × 480 の元画像を 256 × 256 にリサイズした.

# 5 実験結果

## 5.1 定量的結果

表 1 に MDSM [6]，LAVT [4] と提案手法との比較に関する定量的結果を示す． 各スコアは，5 回実験における平均値および標準偏差を表す. ベースライン手法として，LAVT [4] および MDSM [6] を使用した．LAVT は OSMI タスクと関連の深い RES タスクにおいて，MDSM は OSMI タスクにおいて良好な結果が得られているモデルであるためベースライン手法として選択した.

評価尺度には，mIoU，Precision@$k$(P@$k$) を用いた. mIoU および P@$k$ は，OSMI タスクと関連の深い RES タスクにおける標準的な尺度であるため使用した．また，本実験の主要尺度は mIoU とした.

表 1 より, 主要尺度である mIoU において, MDSM, LAVT および提案手法はそれぞれ 24.36, 28.16 および 38.57 であり，提案手法は MDSM より 14.21 ポイント，LAVT より 10.41 ポイント上回った. さらに, P@0.5 および P@0.7 においても提案手法は MDSM, LAVT を上回る性能であった. 以上より，提案手法が最も良好な性能であったといえる．主要尺度の mIoU において LAVT と提案手法の性能差は統計有意であった ($p < 0.05$).

## 5.2 定性的結果

図 3 に定性的結果を示す. 図において，(a), (b), (c) および (d) はそれぞれ，LAVT の予測マスク，MDSM の予測マスク，提案手法の予測マスクおよび正解マスクを示す. 図 3 における命令文は，“Take down the photo collage of flowers above the black shelf stand in the hallway with a basket of sports balls.” であり，対象物は画像の右側にある花のコラージュ写真である. LAVT および MDSM では，写真の一部のマスクのみを生成しており，さらに，誤って隣の展示物のマスクを生成している．それに対して，提案手法では正確に花のコラージュ写真のマスクを生成している.

# 6 おわりに

本研究では，物体操作に関する命令文の対象物のマスクを生成する OSMI タスクを扱った. 本研究の貢献は以下である.

- 多角形において頂点の順番が異なっても同じ多角形を表す場合を扱うために，最適輸送を用いた Polygon Matching Loss を提案した.
- 大規模言語モデルを用いて，命令文に対して対象物に関する修飾関係を明瞭にする言い換えを行う Paraphraser を提案した.
- 本タスクと関連の深い RES タスクにおいて標準的な評価尺度である mIoU，Precision@$k$ について，提案手法がベースライン手法を上回った.

将来研究として，実機による物体把持タスクへの適用が挙げられる.

## 謝辞

本研究の一部は，JSPS 科研費 23H03478，JST ムーンショット，NEDO の助成を受けて実施されたものである．

## 参考文献

[1] Takashi Yamamoto, et al. Development of Human Support Robot as The Research Platform of A Domestic Mobile Manipulator. **ROBOMECH**, Vol. 6, No. 1, pp. 1–15, 2019.

[2] Yuankai Qi, Qi Wu, Peter Anderson, et al. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In **CVPR**, pp. 9982–9991, 2020.

[3] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from Natural Language Expressions. In **ECCV**, pp. 108–124, 2016.

[4] Zhao Yang, Jiaqi Wang, et al. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In **CVPR**, pp. 18155–18165, 2022.

[5] Peng Wang, et al. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In **ICML**, pp. 23318–23340, 2022.

[6] Yui Iioka, Yu Yoshida, et al. Multimodal Diffusion Segmentation Model for Object Segmentation from Manipulation Instructions . In **IEEE IROS**, pp. 7590–7597, 2023.

[7] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, et al. SeqTR: A Simple yet Universal Network for Visual Grounding. In **ECCV**, pp. 598–615, 2022.

[8] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, et al. PolyFormer: Referring Image Segmentation as Sequential Polygon Generation. In **CVPR**, pp. 18653–18663, 2023.

[9] https://platform.openai.com/docs/models/gpt-3-5.

[10] https://platform.openai.com/docs/models/embeddings.

[11] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. **arXiv preprint arXiv:1804.02767**, 2018.

[12] Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In **NIPS**, Vol. 26, pp. 2292–2300, 2013.

[13] Angel Chang, Angela Dai, Thomas Funkhouser, et al. Matterport3D: Learning from RGB-D Data in Indoor Environments. In **3DV**, pp. 667–676, 2018.

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

[15] Tao Mei, et al. Vision and Language: from Visual Perception to Content Creation. **APSIPA Transactions on Signal and Information Processing**, Vol. 9, p. E11, 2020.

[16] Rui Zhou, Cong Jiang, and Qingyang Xu. A Survey on Generative Adversarial Network-based Text-to-image Synthesis. **Neurocomputing**, Vol. 451, pp. 316–336, 2021.

[17] Fei Chen, Du Zhang, et al. VLP: A Survey on Vision-language Pre-training. **Machine Intelligence Research**, Vol. 20, No. 1, pp. 38–56, 2023.

[18] Jing Gu, Eliana Stefani, Qi Wu, et al. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. In **ACL**, pp. 7606–7623, 2022.

[19] Yanyuan Qiao, et al. Referring Expression Comprehension: A Survey of Methods and Datasets. **IEEE Transactions on Multimedia**, Vol. 23, pp. 4426–4440, 2020.

[20] Aishwarya Kamath, Mannat Singh, et al. MDETR-Modulated Detection for End-to-End Multi-Modal Understanding. In **ICCV**, pp. 1780–1790, 2021.

[21] Aly Magassouba, 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.

[22] Shintaro Ishikawa and Komei Sugiura. 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.

[23] Mohit Shridhar and David Hsu. Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction. In **RSS**, 2018.

[24] Oier Mees, Alp Emek, et al. Learning Object Placements for Relational Instructions by Hallucinating Scene Representations. In **IEEE ICRA**, pp. 94–100, 2020.

[25] Danny Driess, Fei Xia, Mehdi Sajjadi, Corey Lynch, et al. PaLM-E: An Embodied Multimodal Language Model. **arXiv preprint arXiv:1804.02767**, 2023.

[26] Jonathan Ho, et al. Denoising Diffusion Probabilistic Models. In **NeurIPS**, Vol. 33, pp. 6840–6851, 2020.

[27] Sahar Kazemzadeh, Vicente Ordonez, et al. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In **EMNLP**, pp. 787–798, 2014.

[28] Licheng Yu, Patrick Poirson, Shan Yang, Alexander Berg, and Tamara Berg. Modeling Context in Referring Expressions. In **ECCV**, pp. 69–85, 2016.

[29] Junhua Mao, Jonathan Huang, Alexander Toshev, et al. Generation and Comprehension of Unambiguous Object Descriptions. In **CVPR**, pp. 11–20, 2016.

[30] Tsung Lin, Michael Maire, et al. Microsoft COCO: Common Objects in Context. In **ECCV**, pp. 740–755, 2014.

[31] Jun Hatori, et al. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In **IEEE ICRA**, pp. 3774–3781, 2018.

[32] Mohit Shridhar, Jesse Thomason, et al. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In **CVPR**, pp. 10740–10749, 2020.

# A 付録

## A.1 関連研究

マルチモーダル言語処理に関する研究は広く行われている [14–18]．[19] では，参照表現理解タスクにおける各手法，標準データセット，標準評価尺度を包括的に紹介している．

Referring Expression Comprehension (REC) は参照表現を含む文および画像を入力とし，その対象物の矩形領域を予測するタスクである．MDETR [20] はこのタスクを扱う代表的な手法であり，モデルの初期段階において画像と言語のモダリティを融合する機構を導入し，良好な結果を得ている．Referring Expression Segmentation (RES) は，REC と同様に対象物の領域を予測するが，対象物の矩形領域ではなくマスクを予測をするタスクである．LAVT [4] はこのタスクを扱う代表的な手法であり，画像特徴量を抽出する段階で，注意機構を介して言語特徴量と融合した特徴量を抽出する．近年では，対象物を画素単位で予測するのではなく，対象物を表す多角形の頂点を予測することで RES を解くアプローチもが存在する [7,8]．

一方，生活支援ロボットの参照表現理解を目的として，物体操作の指示文から物体を特定する問題に取り組んだ研究も存在する [21–24]．[23,24] では，対話システムを導入することで指示文の曖昧さを取り除き，物体を特定する手法を提案している．また，PaLM-E [25] では画像と言語を扱う大規模言語モデルを用いて指示文をもとにタスクの分解および実行を行う．[6] は OSMI タスクに初めて取り組んだ論文であり，OSMI の評価のための SHIMRIE データセットを作成し，二段階のセグメンテーションモデルである MDSM を提案した．MDSM は，Encoder-Decoder モデルで生成したマスクに対し，DDPM [26] を用いて洗練を行っている．

RES における標準的なデータセットとして RefCOCO [27]，RefCOCO+ [28] および G-Ref [29] があげられる．これらはすべて MS COCO [30] から収集された画像に対して，自然言語表現でのアノテーションがされている．PFN-PIC [31] は画像および把持する対象物体に関する指示文から構成されるデータセットであり，約 20 種類の日用品を 4 つの箱に無作為に配置した物体に対する固定視点の画像を用いている．REVERIE データセット [2] は Remote Embodied Visual Referring Expression in Real Indoor Environments (REVERIE) タスクを行うためのデータセットであり，経路とその終着点にある物体に対して物体操作に関する命令文が与えられている．ALFRED [32] は，自然言語による命令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ベンチマークである．提案手法は SeqTR [7] と異なり，最適輸送を用いた多角形のマッチングによって対象物の予測を行う．

表 2: Ablation Study における定量的結果．

| Model | mIoU [%] | P@0.5 [%] | P@0.7 [%] |
|---|---|---|---|
| (i) | 37.28 ± 1.77 | 48.39 ± 3.13 | 21.36 ± 3.54 |
| (ii) | 36.97 ± 2.60 | 46.56 ± 5.31 | 25.52 ± 1.83 |
| (iii) | **38.57** ± **2.77** | **48.96** ± **3.04** | **26.15** ± **1.90** |

## A.2 Ablation Studies

表 2 に，提案手法の ablation studies の定量的結果を示す．ablation 条件として以下の 2 つを定めた．

- Polygon Matching Loss ablation

  $\mathcal{L}_{OT}$ を取り除き，$\mathcal{L}_{OT}$ の性能への寄与を調査した．表 2 より，モデル (i) における mIoU は 37.28 であり，モデル (iii) よりも 1.29 ポイント減少した．P@$k$ においても同様に減少した．このことから，Polygon Matching Loss が性能向上に寄与しているといえる．これは，Polygon Matching Loss により，モデルが頂点の順序に関係なく，頂点集合全体として正解マスクに近づくように学習したためであると示唆される．

- Paraphraser ablation

  Paraphraser を取り除くことで Paraphraser の有効性を調査した．表 2 より，モデル (ii) における mIoU は 36.97 であり，モデル (iii) よりも 1.60 ポイント低かった．このことから，Paraphraser が指示文の修飾関係を明瞭にし，言語理解の向上を促したということが示唆される．