最適輸送を用いたポリゴンマッチングと複数の基盤モデルによる参照表現セグメンテーション

○雨宮佳音，西村喬行，杉浦孔明（慶應義塾大学）

本研究では，屋内環境の画像，その環境に対応する 3 次元点群及び物体操作に関する指示文をもとに，対象物のセグメンテーションマスクを生成するタスクを扱う．例えば，“Go to the living room and pick up the orange pillow closest the plant.” という指示文が与えられた場合，植物に最も近いオレンジ色の枕にセグメンテーションマスクを生成することが望ましい．既存手法ではモデルが視覚情報に過度に依存することにより，ある画像に対して異なる対象物を示す複数の指示文が与えられた際，いずれの指示文に対してもほぼ同一の領域にマスクを生成する場合があった．そこで，本研究では画像に過度に依存せずにセグメント情報を扱うため，物体の位置情報を意図的に削減した画像特徴量，及び指示文と 3D 点群由来のマルチモーダル特徴量を融合する，Multimodal Segment Attention を提案する．屋内環境の画像，3 次元点群，物体操作に関する指示文及び対象物に対するマスク画像が含まれるデータセットにおいて，テストセットを拡充して実験を行った．結果として，提案手法は mIoU 及び P@0.5 において，ベースライン手法を 3.40 ポイント及び 6.78 ポイント上回った．

1. はじめに

自然言語による指示文から環境画像中の対象物を特定する技術は，ロボットや自動運転車において，人間との自然なインタラクションを実現するために有用である．特に，生活支援ロボットに対して移動や物体操作などを自然言語で指示できれば，介助を必要とする高齢者が難しい操作を覚える必要なくロボットを扱える．しかし，自然言語による指示には複雑な参照表現が含まれている場合があり，そのような指示文から対象物を適切に特定する性能は，現状不十分である．

本研究では，物体操作に関する自然言語の指示文が与えられた際，対象物に対するセグメンテーションマスクを生成する Object Segmentation from Manipulation Instructions 3D（OSMI-3D）タスク [1] を扱う．例えば，寝室を写した画像に対して，“Go to the bedroom and fluff up the smallest pillow” という指示文が与えられた際，画像中の最も小さい枕にセグメンテーションマスクを生成することが望ましい．

本研究で扱う OSMI-3D タスクでは，物体操作の命令に加えて移動の命令を含んだ 2 文以上の指示文や，対象物を修飾する参照表現を複数含んだ指示文が多く存在する．そのため，単純な RES タスクに比べて対象物の特定が困難である．例えば，枕が複数存在する画像に対して，“Go to the hallway area where there are three pictures side by side and get me the one on the right.” という指示文が与えられた場合，“the one on the right” のみでは対象物を特定できない可能性がある．この例では，“there are three pictures side by side” が対象物を間接的に示しているため，指示文全体の文脈を理解する必要がある．

OSMI-3D タスクと最も関連の深いタスクとして OSMI タスク [2] があり，MDSM [2] は OSMI タスクの複雑な参照表現を含む指示文から，対象物のマスクを適切に生成できたことを報告している．また，Nishimura ら [1] は画像の画角外に存在する物体に関連する参照表現を理解するため，画角外の物体に関するマルチモーダル特徴量を扱い，OSMI-3D タスクにおいて良好な結果を得ている．しかし，モデルが視覚情報に強く依存し，ある画像に対して異なる対象物を指す複数の指示文が与えられた場合，それぞれの指示文に対してほとんど同一の領域にマスクを生成する傾向があった．

本研究では，屋内環境の画像，3 次元点群及び複雑な参照表現を含む指示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．既存手法との主要な違いは，SAM [3] により生成したセグメンテーション画像に対する画像エンコーダとして CLIP [4] を用いる，Multimodal Segment Attention（MSA）を導入する点である．MSA を用いることにより，物体の位置に関する情報を意図的に欠落させ，物体の形状のみから言語と接地された特徴量抽出を行うことで，既存手法の課題であった画像への過度な依存を抑制することが期待される．提案手法の新規性は以下である．

図 1 OSMI-3D タスクの具体例．指示文：“Go to the bathroom on level 1 and refill the tissue box.”

- 画像に過度に依存せずセグメント情報を扱うため，SAM により生成したセグメンテーション画像に対して CLIP を適用して得た画像特徴量，及び指示文と 3D 点群由来のマルチモーダル特徴量を融合する，MSA を導入する．

2. 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おり [5, 6] マルチモーダル大規模言語モデルは著しい成果を挙げている [7, 8]．マルチモーダル言語処理の分野は、扱うモダリティの組み合わせにより様々な分野に細分化され，言語と画像を扱う例として参照表現理解タスクがある．参照表現理解タスクは，画像中の特定の領域を指すテキストをもとにその領域を予測し，矩形領域やセグメンテーションマスクを生成するタスクである [9, 10]．参照表現理解タスクのうちセグメンテーションマスクを生成する RES タスクについて，対象物のマスクを画素単位で予測する研究が広く行われている [7, 11]．一方，SeqTR [12] のように対象物のポリゴンの頂点群を予測するアプローチも存在す

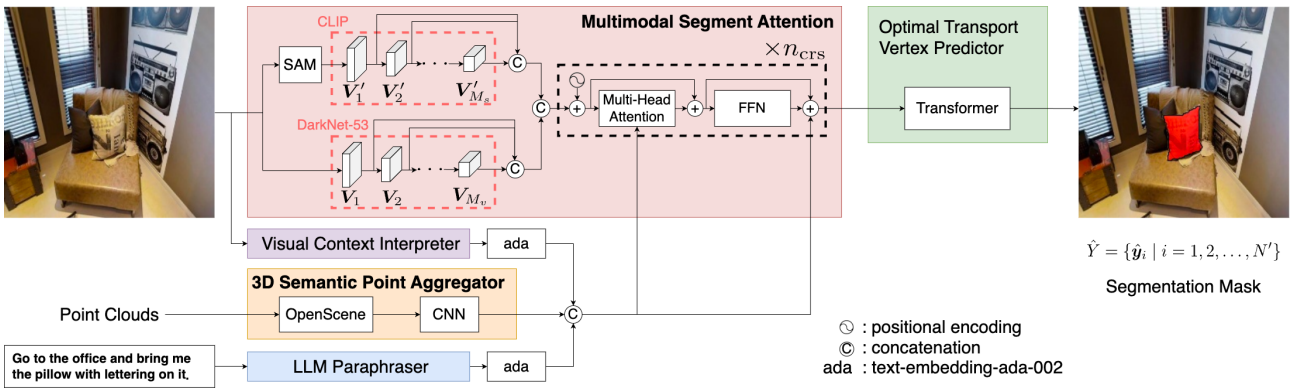


図 2　提案手法のネットワーク構造

る．また，生活支援ロボットの参照表現理解を目的として，物体操作に関する指示文から対象物を特定する問題に取り組んだ研究も存在する [2, 13]．Nishimura ら [1] は OSMI-3D タスクに初めて取り組み，3 次元点群の open-vocabulary マルチモーダル特徴量を扱うセグメンテーションモデルを提案した．

3.　問題設定

本論文では，屋内環境の画像，その環境に対応する 3 次元点群及び物体操作に関する指示文をもとに，対象物のセグメンテーションマスクを生成する OSMI-3D タスク [1] を扱う．本タスクでは，与えられた指示文で指定された対象物に対するセグメンテーション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 to the bathroom on level 1 and refill the tissue box.” という指示文が与えられた際，赤色の領域で示すマスクを生成することを目標とする．本タスクにおける入力は画像，3 次元点群及び指示文であり，出力は指示文で指定された対象物に対するセグメンテーションマスクである．本研究では，画像中に対象物が複数ある場合や全くない場合を扱わない．評価尺度には，RES タスクにおいて標準的である mean IoU(mIoU) 及び precision@K(K=0.5, 0.7) を用いる．

4.　提案手法

提案手法は Nishimura らの手法 [1] を拡張し，物体操作に関する指示文をもとに，対象物のセグメンテーションマスクを生成する OSMI-3D タスクを扱うモデルである．本研究で行う拡張は，SAM [3] を用いて生成したセグメンテーション画像への CLIP [4] の適用である．CLIP を用いることで，物体の位置に関する情報を意図的に削減し，物体の形状のみから言語と接地された特徴量抽出を行う．これにより，画像への過度な依存を抑制する効果が期待されるため，セグメンテーション画像を用いる手法全般に適用可能であると考えられる．図 2 に提案手法のネットワーク構造を示す．主要モジュールは，3D Semantic Point Aggregator（3D-SPA）及び Multimodal Segment Attention（MSA）である．

4.1　入力

モデルの入力を $\boldsymbol{x} = \{\boldsymbol{x}_{\mathrm{img}}, X_{\mathrm{pcl}}, \boldsymbol{x}_{\mathrm{inst}}\}$ と定義する．ここで，$\boldsymbol{x}_{\mathrm{img}} \in \mathbb{R}^{3\times H\times W}$，$X_{\mathrm{pcl}} = \{\boldsymbol{\xi}_i \mid i = 0, 1, 2, \ldots, N_{\mathrm{pcl}}\}$ 及び $\boldsymbol{x}_{\mathrm{inst}} \in \{0, 1\}^{v\times l}$ はそれぞれ，画像，3 次元点群及び one-hot ベクトルとしてトークン化された指示文を表す．また，H，W，$\boldsymbol{\xi}_i$，N_{pcl}，v 及び l はそれぞれ，画像の高さ，幅，3 次元点群における i 番目の点、点の総数、指示文の語彙サイズ、最大トークン長である．

4.2　3D Semantic Point Aggregator

既存手法では，画像中に存在しない物体の情報が得られないため，画角外の物体に関する参照表現が与えられた際に対象物を特定することが困難であった．そこで，画角外の物体に関する参照表現の理解を強化するため，3D-SPA を導入する．3D-SPA は $\boldsymbol{x}_{\mathrm{img}}$ の画角外に存在する物体に関して，3 次元点群の open-vocabulary マルチモーダル特徴量を取得するモジュールである．これにより，別角度からの画像を必要とせず，画角外に存在する物体に関する情報を得ることが期待される．

本モジュールにおいて，入力は X_{pcl}，出力は中間特徴量 $\boldsymbol{h}_{\mathrm{spa}} \in \mathbb{R}^{d_{\mathrm{spa}}}$ である．ここで，d_{spa} は次元数を表す．まず、X_{pcl} のうち、$\boldsymbol{x}_{\mathrm{img}}$ が撮影された位置から水平方向に最も近い N_{near} 点である，X_{near} を抽出する．N_{near} 点のみを用いる理由は，参照表現は対象物の周囲の物体に関連する場合が多く，遠隔地の点を利用することは効率的でないためである．次に，OpenScene [15] の事前学習済みモデルを用いて，X_{near} からマルチモーダル特徴量 $\boldsymbol{h}'_{\mathrm{spa}}$ を抽出する．OpenScene は CLIP を使用しており，3D 点群の各点に open-vocabulary マルチモーダル特徴量が埋め込まれる．最終的に，$\boldsymbol{h}'_{\mathrm{spa}}$ についてアップサンプリング及び最大プーリングを行い，特徴量 $\boldsymbol{h}_{\mathrm{spa}}$ を得る．

さらに，対象物に関する記述の理解を向上させるため，LLM Paraphraser [16] を用いる．これにより， 複数の文から構成されることの多い $\boldsymbol{x}_{\mathrm{inst}}$ を 1 つの文に結合し，対象物に関する参照表現を要約した文 $\boldsymbol{L}_{\text{p-inst}}$ を得る．また，Visual Context Interpreter [1] を用いて， $\boldsymbol{x}_{\mathrm{img}}$ について画像中の物体の属性や空間関係などを述べた説明文 $\boldsymbol{L}_{\mathrm{vci}}$ を得る．その後，text-embedding-ada-002 [17] を用いて，$\boldsymbol{L}_{\text{p-inst}}$ 及び $\boldsymbol{L}_{\mathrm{vci}}$ の文埋め込みである $\boldsymbol{h}_{\text{p-inst}} \in \mathbb{R}^{d_{\text{p-inst}}}$ 及び $\boldsymbol{h}_{\mathrm{vci}} \in \mathbb{R}^{d_{\mathrm{vci}}}$ を得る．ここで，$d_{\text{p-inst}}$ 及び d_{vci} はそれぞれ $\boldsymbol{h}_{\text{p-inst}}$ 及び $\boldsymbol{h}_{\mathrm{vci}}$ の次元数を表す．$\boldsymbol{h}_{\mathrm{vci}}$，$\boldsymbol{h}_{\mathrm{spa}}$ 及び $\boldsymbol{h}_{\text{p-inst}}$ をチャネル方向に結合し，ダウンサンプリングして $\boldsymbol{h}_{\mathrm{mix}}$ を得る．

4.3　Multimodal Segment Attention

既存手法 [1] では， モデルが視覚情報に過度に依存することにより，ある画像に対して異なる対象物を示す複数の指示文が与えられた際，いずれの指示文に対してもほぼ同一の領域にマスクを生成する場合があった．そこで，視覚情報に過度に依存することなくセグメン

表 1　ベースライン手法との比較及び Ablation Study の定量的結果

手法	x_{sam} に対する画像エンコーダ	mIoU	P@0.5	P@0.7
LAVT [11]	-	23.51 ± 3.36	23.50 ± 5.76	**16.17 ± 4.68**
SeqTR [12]	-	20.72 ± 0.67	17.34 ± 2.95	3.71 ± 1.02
Nishimura ら [1]	-	22.08 ± 0.71	22.75 ± 1.32	10.13 ± 1.40
提案手法 (a)	DarkNet-53 [14]	26.74 ± 1.22	29.30 ± 2.44	14.66 ± 2.26
提案手法 (b)	CLIP [4]	**26.91 ± 1.15**	**30.28 ± 2.14**	11.06 ± 1.81

ト情報を扱うため，MSA を導入する．MSA は，SAM を用いて生成したセグメンテーション画像を CLIP でエンコードして得た画像特徴量及び指示文と 3D 点群由来のマルチモーダル特徴量を融合するモジュールである．セグメンテーション画像に CLIP を適用することで，物体の位置に関する情報が欠落するため，画像への過度な依存を抑制する効果が期待される．

本モジュールにおいて，入力は $\boldsymbol{x}_{\text{img}}$ 及び $\boldsymbol{h}_{\text{mix}}$ である．まず，$\boldsymbol{x}_{\text{img}}$ に対して DarkNet-53 [14] を適用し，解像度の異なる M_v 種類の中間層における画像特徴量 $\{\boldsymbol{V}_k \in \mathbb{R}^{H_k \times W_k \times C_k}\}_{k=1}^{M_v}$ を得る．ここで，H_k，W_k 及び C_k は $\boldsymbol{V}_k$ の画像特徴量の高さ，幅及びチャネル数を表す．また，事前学習済みの SAM を用いて，$\boldsymbol{x}_{\text{img}}$ からセグメンテーション画像 $\boldsymbol{x}_{\text{sam}} \in \mathbb{R}^{3 \times H \times W}$ を生成する．この $\boldsymbol{x}_{\text{sam}}$ に対して CLIP ResNet-50 を適用し，M_s 種類の中間層における画像特徴量 $\{\boldsymbol{V}_l' \in \mathbb{R}^{H_l \times W_l \times C_l}\}_{l=1}^{M_s}$ を得る．ここで，H_l，W_l 及び C_l は V'_l の画像特徴量の高さ，幅及びチャネル数を表す．得られた中間特徴量 $\boldsymbol{V}_k$ 及び $\boldsymbol{V}_l'$ をダウンサンプリングし，それぞれをチャネル方向に結合することで $\boldsymbol{V}_{\text{mix}}$ を得る．画像特徴量 $\boldsymbol{V}_{\text{mix}}$ とマルチモーダル特徴量 $\boldsymbol{h}_{\text{mix}}$ に対して cross-attention を適用し，$\boldsymbol{S}_a = f_a(\boldsymbol{V}_{\text{mix}}, \boldsymbol{h}_{\text{mix}})$ を計算する．$f_a(\cdot,\cdot)$ は cross-attention を表し，任意の行列 $\boldsymbol{X}_A$ 及び $\boldsymbol{X}_B$ に対して以下のように定義する．

$$f_a(\boldsymbol{X}_A, \boldsymbol{X}_B) = \text{softmax}\left(\frac{(\boldsymbol{W}_q \boldsymbol{X}_A)(\boldsymbol{W}_k \boldsymbol{X}_B)^T}{\sqrt{d}}\right)(\boldsymbol{W}_v \boldsymbol{X}_B)$$

ここで，$\boldsymbol{W}_q$，$\boldsymbol{W}_k$，$\boldsymbol{W}_v$ は学習可能な重み，d はスケーリングファクターを表す．その後，$\boldsymbol{S}_a$ に対して $\boldsymbol{h}_{\text{mix}}$ を加算し，MSA における出力 $\boldsymbol{S}_{\text{msa}}$ を得る．最終的に，OTVP [16] を用いて，$\boldsymbol{S}_{\text{msa}}$ から予測マスクの頂点集合 $\hat{\boldsymbol{y}} = \{\hat{\boldsymbol{v}}_i \in \mathbb{R}^2\}_{i=1}^{N'}$ を得る．ここで，N' は予測マスクの頂点数，$\hat{\boldsymbol{v}}_i$ は多角形を構成する頂点の座標を表す．

5. 実験設定

5.1 データセット

本研究では，SHIMRIE-3D [1] を用いた．データセットの詳細は，Nishimura らの論文 [1] を参照されたい．ただし，以下に説明する理由からテストセットを拡充した．SHIMRIE-3D には 1 つの画像に対して複数の指示文が存在するが，これらは全て同一の対象物を指す．そのため，同一画像に対して異なる対象物を指す指示文が与えられた際に，適切にマスクを生成できるか評価することが困難である．そこで，SHIMRIE-3D のテスト集合の各画像に対して，既存の指示文が指す物体とは異なる物体を指す指示文を 258 サンプル追加した．

SHIMRIE-3D には，4,289 枚の画像，11,508 の指示文及びそれに対応する対象物のマスクが含まれている．指示文の語彙サイズは 2,630，全単語数は 218,613，平均文長は 19.0 である．SHIMRIE-3D データセットでは全 11,508 サンプルのうち，10,068 サンプルを訓練集合に，829 サンプルを検証集合に，611 サンプルをテスト集合に使用した．これは，REVERIE データセットで行われた分割方法を踏襲した．SHIMRIE-3D に含まれる環境画像は 90 のフロアマップから収集されており，訓練集合において既知の環境である seen 集合，及び訓練集合において未知の環境である unseen 集合に分割される．検証集合は，558 サンプルの seen 集合，271 サンプルの unseen 集合で構成されている．テスト集合は全て unseen 集合で構成されている．ここで，訓練集合はモデルのパラメータ更新に，検証集合はハイパーパラメータの調整に使用した．また，テスト集合はモデルの評価に使用した．

5.2 パラメータ設定

最適化手法や損失関数の詳細については，Nishimura らの論文 [1] を参照されたい．実験おいて，訓練可能なパラメータ数は約 350M，積和演算数は約 720G であった．訓練には 24GB の GPU メモリ搭載の GeForce RTX 4090，Intel Core i9-13900KF，及び 64GB の RAM を用いた．訓練は 90 エポック行い，訓練時間は 4 時間程度，1 サンプルあたりの推論時間は 7.4ms 程度であった．各エポック毎に検証集合において mIoU を計算し，その中で最も高い結果を得たモデルを用いてテスト集合において評価を行った．

6. 実験結果

6.1 定量的結果

ベースライン手法と提案手法の定量的比較結果を表 1 に示す．実験はそれぞれ 5 回ずつ行い，その平均及び標準偏差を示した．表中の太字の数値は各指標における最も高い数値を表す．LAVT [11], SeqTR [12] 及び Nishimura らの手法 [1] をベースライン手法とした．LAVT 及び SeqTR は，OSMI-3D タスクと関連の深い RES タスクにおいて，Nishimura らの手法は OSMI-3D タスクにおいて，良好な結果を得ているため選択した．評価尺度には，OSMI-3D 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 mIoU 及び P@K を使用した．表 1 より，主要尺度である mIoU において，LAVT，SeqTR，Nishimura らの手法及び提案手法はそれぞれ，23.51，20.72，22.08 及び 26.91 であり，提案手法はベースライン手法のうち最良の LAVT を 3.40 ポイント上回った．同様に，P@0.5 においても全てのベースライン手法を上回った．mIoU 及び P@0.5 において，ベースライン手法との性能差は統計有意であった（$p < 0.05$）．

6.2 Ablation studies

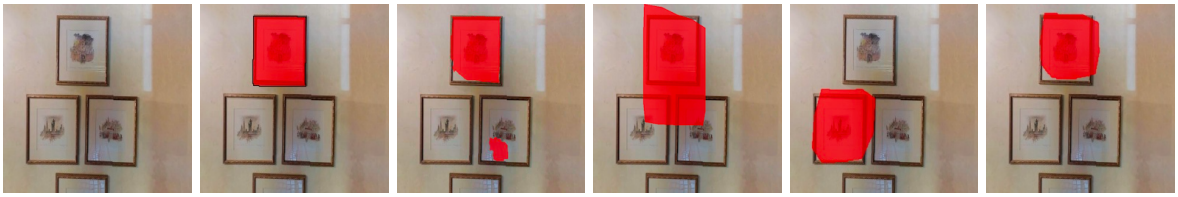
表 1 に，ablation study の定量的結果を示す．ablation 条件として以下を定めた．

Multimodal Segment Attention ablation

(i)

指示文：“Please go to the bathroom on level 3 and refill the tissue box.”

(ii)



指示文：“Go to the bathroom and take down the picture hung the highest”

図 3　各手法における定性的結果．左から順に $\boldsymbol{x}_{\text{img}}$，GT，LAVT [11]，SeqTR [12]，Nishimura ら [1]，提案手法．

Multimodal Segment Attention 内の $\boldsymbol{x}_{\text{sam}}$ の画像エンコーダについて，CLIP の代わりにベースライン手法である Nishimura らの手法と同様に DarkNet-53 を適用することで，有効性を調査した．DarkNet-53 を用いたモデル (a) における mIoU は 26.74 であり，CLIP を用いたモデル (b) よりも 0.17 ポイント減少した．同様に，P@0.5 においても減少した．

6.3　定性的結果

図 3 の (i) 及び (ii) に提案手法における成功例の定性的結果を示す．図は左から順に $\boldsymbol{x}_{\text{img}}$，正解マスク，LAVT，SeqTR，Nishimura らの手法及び提案手法における予測マスクである．(i) の指示文は “Please go to the bathroom on level 3 and refill the tissue box” であった．この例において，LAVT では洗面器に，SeqTR 及び Nishimura らの手法では特定の物体が存在しない領域に対して誤ってマスクを生成した．これに対し，提案手法ではティッシュボックスに対して適切にマスクを生成した．また，(ii) には指示文が “Go to the bathroom and take down the picture hung the highest” である例を示す．この例において，LAVT では正解以外の領域にも，SeqTR では正解より広範な領域に，Nishimura らの手法では誤った絵にマスクを生成した．一方，提案手法では上部の絵の領域のみに正しくマスクを生成した．

6.4　被験者実験

OSMI-3D タスクを人間が解いた場合の性能を評価するため，被験者実験を実施した．実験には 5 人の被験者が参加した．テスト集合からランダムに 100 サンプルを抽出し，1 人に対して 20 サンプルの画像と指示文のペアを提示し，指示文に従って画像中の対象物にセグメンテーションマスクを生成するよう指示した．このとき，mIoU，P@0.5 及び P@0.7 はそれぞれ 84.83，91 及び 89 であった．

7.　おわりに

本研究では，屋内環境の画像，3 次元点群及び物体操作指示文から，対象物に対してセグメンテーションマスクを生成する OSMI-3D タスク [1] を扱った．

Multimodal Segment Attention の導入により画像への過度な依存は改善されたが，依然として画像への依存傾向がある．そこで，将来研究として視覚情報への依存をより強く抑制するため，予め言語情報を考慮したセグメント情報を与える手法の実現が挙げられる．

謝辞

本研究の一部は，JSPS 科研費 23K28168，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] T. Nishimura, K. Kuyo, M. Kambara, et al., “Object Segmentation from Open-Vocabulary Manipulation Instructions Based on Optimal Transport Polygon Matching with Multimodal Foundation Models,” IROS, 2024.

[2] Y. Iioka, Y. Yoshida, Y. Wada, et al., “Multimodal Diffusion Segmentation Model for Object Segmentation from Manipulation Instructions,” IROS, pp.7590–7597, 2023.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, et al., “Segment Anything,” ICCV, pp.4015–4026, 2023.

[4] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML, pp.8748–8763, 2021.

[5] S. Uppal, S. Bhagat, et al., “Multimodal Research in Vision and Language: A Review of Current and Emerging Trends,” Information Fusion, vol.77, pp.149–171, 2022.

[6] F. Chen, D. Zhang, et al., “VLP: A Survey on Vision-language Pre-training,” MIR, vol.20, no.1, pp.38–56, 2023.

[7] X. Lai, Z. Tian, et al., “LISA: Reasoning Segmentation via Large Language Model,” CVPR, pp.9579–9589, 2024.

[8] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, et al., “InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning,” NeurIPS, pp.49250–49267, 2023.

[9] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, et al., “MAttNet: Modular Attention Network for Referring Expression Comprehension,” CVPR, pp.1307–1315, 2018.

[10] G. Luo, Y. Zhou, X. Sun, et al., “Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation,” CVPR, pp.10031–10040, 2020.

[11] Z. Yang, J. Wang, Y. Tang, K. Chen, et al., “LAVT: Language-Aware Vision Transformer for Referring Image Segmentation,” CVPR, pp.18155–18165, 2022.

[12] C. Zhu, Y. Zhou, et al., “SeqTR: A Simple yet Universal Network for Visual Grounding,” ECCV, pp.598–615, 2022.

[13] R. Korekata, et al., “Switching Head-Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks,” IROS, pp.3865–3872, 2023.

[14] C. Wang, et al., “Scaled-YOLOv4: Scaling Cross Stage Partial Network,” CVPR, pp.13029–13038, 2021.

[15] S. Peng, et al., “OpenScene: 3D Scene Understanding with Open Vocabularies,” CVPR, pp.815–824, 2023.

[16] 九曜克之，飯岡雄偉，杉浦孔明，“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,” 第 30 回言語処理学会資料，2024.

[17] OpenAI, “text-embedding-ada-002,” 2024. Accessed: June 2024. [Online]. https://platform.openai.com/docs/models/embeddings