生活支援タスクにおける大規模視覚言語モデルと拡散確率モデルを用いた参照表現セグメンテーション

Referring Expression Segmentation With Large-Scale Visual Language Model and Diffusion Probabilistic Model in Household Tasks

飯岡 雄偉 *1 Yui Iioka　吉田 悠 *1 Yu Yoshida　和田 唯我 *1 Yuiga Wada　畑中 駿平 *1 Shumpei Hatanaka　杉浦 孔明 *1 Komei Sugiura

*1慶應義塾大学
Keio University

In this paper, we propose the Multimodal Diffusion Segmentation Model (MDSM), which generates a mask in the first stage and refines it in the second stage. We introduce a crossmodal parallel feature extraction mechanism and extend diffusion probabilistic models to handle crossmodal features. Our proposed MDSM surpasses that of the baseline method by a large margin of +10.13 mean IoU.

1. はじめに

高齢化が進行する現代社会では，日常生活において介助支援の必要性が高まっている．それに伴って，在宅介助者の不足が社会問題となっており，これを解決するため，被介助者を物理的に支援可能な生活支援ロボットが注目されている [Yamamoto 19]．生活支援ロボットにとって，自然言語による命令を理解し，物体操作を行うことは必須である．しかし，ロボットの自然言語の理解能力は現状不十分である．

本研究では，自然言語を理解し，対象物のセグメンテーションマスクを生成するモデルの構築を目標とする．そのために，対象物を特定する自然言語による命令文が与えられた際，正しく内容を解釈し，対象物を特定する Object Segmentation from Manipulation Instructions (OSMI) タスクを扱う．図 1 に OSMI タスクの例を示す．具体的には，“Go to the living room and fetch the pillow closest to the radio art on the wall.” という命令文が与えられたときに，ロボットは，ラジオの絵に最も近い枕に対してマスクを生成することが求められる．対象物を特定する際，画素単位のマスクを用いた対象物の予測はバウンディングボックスを用いた予測と比較して，物体の位置や形状を正確に把握するのにはるかに有用である．

OSMI タスクでは，複数の参照表現を含む命令文理解が求められる．しかし，ロボットは 1 つの参照表現を理解することさえ難しい．例えば，[Qi 20] では先進技術での参照表現理解の精度が約 50%であり，人間による精度の 90.76%とは乖離があることが報告されている．さらに，OSMI タスクは，(1) 命令文に含まれる複数の物体に対する参照表現の理解，(2) 複数の物体候補から対象物の予測，(3) バウンディングボックスではなく画素単位でのマスク生成が必要であり，単純な参照表現セグメンテーション (RES) タスクよりも難しい．そのため，RES タスクを扱うモデルの多くは，OSMI タスクには不十分である．

本研究では，2 段階でマルチモダリティにおけるセグメンテーションマスクの生成を行う Multimodal Diffusion Segmentation Model (MDSM) を提案する．図 1 に MDSM の概要を示す．提案手法は，多数の参照表現を含む自然言語表現によって対象物のセグメンテーションマスクの生成を行う．MDSM は，1 段階目でセグメンテーションマスクを生成し，2 段階目ではそれを修正する．我々は拡散確率モデル [Ho 20] を拡張し，言語情報を処理することで OSMI タスクに応用する．多くの既存研究 [Yang 22, Liu 19] では，対象物と関係のない領域を予測することが多い．それに対して，MSDM は従来の 1 段階モデルよりも対象物と関係する領域を予測する．

本論文における主な貢献は以下である．

連絡先: 飯岡雄偉，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，kmngrd1805@keio.jp

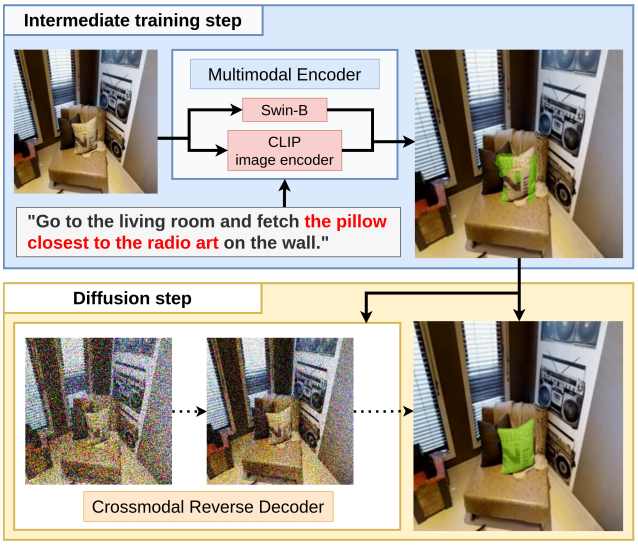


図 1: 提案手法の具体例.

- 2 段階のマルチモーダルセグメンテーションモデルである MDSM を提案する．
- 並列クロスモーダル特徴抽出機構を新規モジュールである Multimodal Encoder に導入する．
- 拡散確率モデルを拡張し，クロスモーダル特徴量を新規モジュールである Crossmodal Reverse Decoder に導入する．

2. 関連研究

[Uppal 22] では，視覚言語モダリティを扱う研究の動向がまとめられている．RES タスクでは，参照表現を含む文および画像を入力とし，画素単位での参照領域予測を目的とする．LAVT [Yang 22] は，RES を扱う代表的な手法であり，画像をエンコードする段階で，注意機構を介して言語特徴量とマージされたクロスモーダル特徴量を抽出する．また，石川らは、画像全体ではなく，画像内の関係領域に注目することで，対象物と他の物体との関係を直接的に学習する Target-dependent UNITER を提案した [Ishikawa 21]．Text-to-Image Generation では，言語情報に基づいたもっともらしい画像の生成を行うことを目的とする．Rombach らは，拡散確率モデルを，文やバウンディングボックスなどの条件に対応した柔軟な生成器を提案し，高解像度での合成を実現している [Rombach 22]．また，REVERIE データセット [Qi 20] は，屋内環境における家事作業に関するデータセットであり，複数の対象物に対する参照表現を含む自然言語の指示が付与された画像が含まれている．

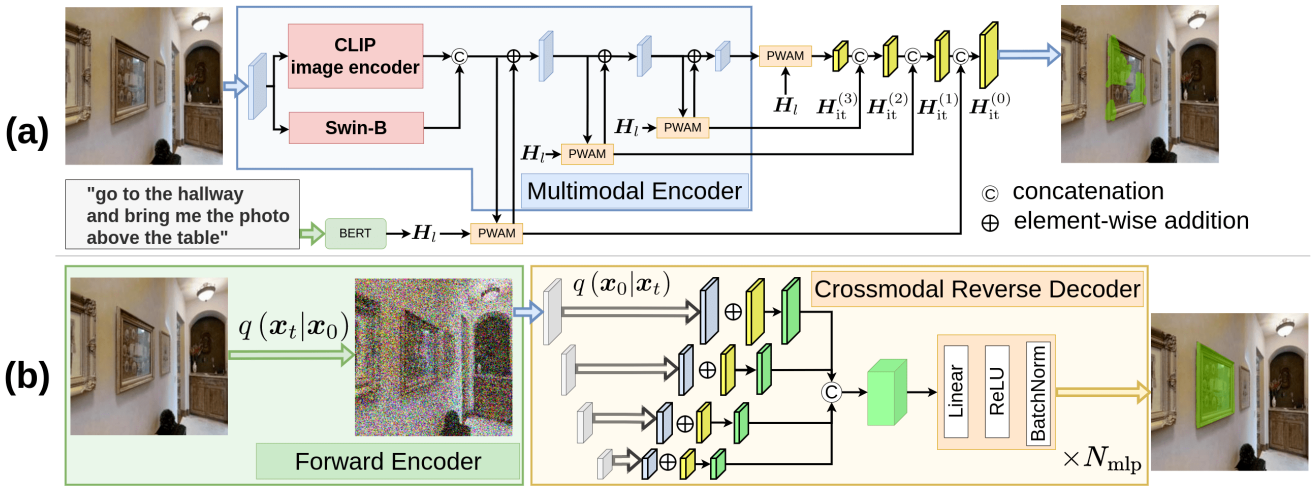


図 2: 提案手法 MDSN のネットワーク構造. (a) intermediate training step, (b) diffusion step.

3. 問題設定

本論文では，実世界の室内環境において，物体操作に関する命令文の対象物をセグメントする OSMI タスクを扱う．本タスクでは，与えられた命令文が指す画像内の対象物に対して，セグメンテーションマスクを生成することが望ましい．本論文では，以下のように用語を定義する．

- **命令文:** ロボットに家事作業を命令する文.
- **対象物:** 命令文を生活支援ロボットが実行するときに，動作の対象となる物体．ただし，各画像における対象物は 1 つしか存在しないことを前提とする.

また，入力は画像及び命令文であり，出力は画素単位での対象物に対するセグメンテーションマスクである．

本タスクと関連の深い RES タスクにおいて，標準的である mIoU, oIoU, Precision@k (P@k) を評価尺度とする．

4. 提案手法

図 2 に，MSDM のネットワーク構造を示す．MSDM は，RES タスクにおいて代表的な手法である LAVT [Yang 22] を拡張した，OSMI タスクを扱うモデルである．

本研究では，家庭内環境での命令文理解において，提案手法の有用性を検証する．ただし，本手法における CLIP[] を並列に用いたクロスモーダル特徴抽出機構や，拡散確率モデル [Ho 20] に基づいた視覚言語セグメンテーションは，より幅広いドメインにも適用可能だと考えられる．

4.1 Intermediate Training Step

intermediate training step での入力は，画像 $\boldsymbol{x}_0 \in \mathbb{R}^{H\times W\times 3}$ および $v \times l$ 次元の one-hot ベクトル集合である命令文 $\boldsymbol{x}_L$ である．ただし，H, W, v, l はそれぞれ画像の高さ，幅，命令文の語彙サイズ，トークン数を表す．

また，$\boldsymbol{x}_L$ を入力とし，BERT [Devlin 19] を用いて言語特徴量 $\boldsymbol{H}_l \in \mathbb{R}^{C\times l}$ を抽出した．ここで，C は各トークンにおける特徴量の次元数を示す．

4.1.1 Multimodal Encoder

第 1 層目の Multimodal Encoder は，$\boldsymbol{x}_0$ を入力とし，画像特徴量 $\boldsymbol{V}_1 \in \mathbb{R}^{H_1\times W_1\times C_1}$ を出力する．ここで，H_i, W_i, C_i は，それぞれ第 i 層目の画像サイズおよびチャネル数を示す．

まずは，$\boldsymbol{x}_0$ をそれぞれ Swin-B [Liu 21] および CLIP image encoder [Radford 21] に入力し，画像特徴量 $\boldsymbol{V}_{\mathrm{sw}}^{(1)} \in \mathbb{R}^{H_1\times W_1\times C_1}$ および，$\boldsymbol{V}_{\mathrm{clp}}^{\prime(1)} \in \mathbb{R}^{C_{\mathrm{clp}}}$ を得る．ここで，C_{clp} は CLIP image encoder によって出力される画像特徴量の次元を表す．次に，得られた $\boldsymbol{V}_{\mathrm{clp}}^{\prime(1)}$ を整形し，$\boldsymbol{V}_{\mathrm{clp}}^{(1)} \in \mathbb{R}^{H_{\mathrm{clp}}^{(1)}\times W_{\mathrm{clp}}^{(1)}\times C_1}$ を得る．ただし，$H_{\mathrm{clp}}^{(1)} = W_{\mathrm{clp}}^{(1)} = \sqrt{C_{\mathrm{clp}}/C_1}$ とする．最後に，$\boldsymbol{V}_{\mathrm{sw}}^{(1)}$ と $\boldsymbol{V}_{\mathrm{clp}}^{(1)}$ をチャネル方向に結合して，3×3 の畳み込みを行うことで，$\boldsymbol{V}_1$ を出力する．

以降，第 i 層目の Multimodal Encoder では，後述する Pixel-Word Attention Module (PWAM) [Yang 22] よって抽出される特徴量 $\boldsymbol{F}_{i-1} \in \mathbb{R}^{H_{i-1}\times W_{i-1}\times C_{i-1}}$ および $\boldsymbol{V}_{i-1}$ から得られるマルチモーダル特徴量 $\boldsymbol{E}_{i-1} \in \mathbb{R}^{H_{i-1}\times W_{i-1}\times C_{i-1}}$ を入力とし，$\boldsymbol{V}_i$ を出力する．なお，入力 $\boldsymbol{E}_{i-1}$ は以下のようにして求められる．

$$\boldsymbol{S}_{i-1} = \tanh(\mathrm{Conv}(\boldsymbol{F}_{i-1})),$$

$$\boldsymbol{E}_{i-1} = \boldsymbol{F}_{i-1} \odot \boldsymbol{S}_{i-1} + \boldsymbol{v}_{i-1},$$

ただし，$\boldsymbol{S}_{i-1}$, Conv, $\odot$ はそれぞれ，$\boldsymbol{F}_{i-1}$ の重みマップ，1×1 の畳み込み，要素ごとの乗算を表す．

4.1.2 Output

intermediate training step では，バイリニア補間によるアップサンプリングを行った各 $\boldsymbol{F}_i$ をチャネル方向に結合し，最終的なセグメンテーションを行う．最終的な出力である $H\times W$ の 2 値予測マスク画像 $\hat{\boldsymbol{y}}_{\mathrm{it}}$ は，$p(\hat{\boldsymbol{y}}_{\mathrm{it}})$ において閾値 0.5 以上の値を 1，それ以外を 0 としたものとする．

4.2 Diffusion Step

4.2.1 Forward Encoder

diffusion step での入力は，$\boldsymbol{x}_0$，$\boldsymbol{H}_{\mathrm{it}}^{(i)}$，および $p(\hat{\boldsymbol{y}}_{\mathrm{it}})$ である．Forward Encoder は一般的な拡散確率モデル [Ho 20] と同様に，マルコフ過程に基づいて，ガウシアンノイズを徐々に加算する機構である．ここでは，$\boldsymbol{x}_0$ を入力とし，t 回ガウシアンノイズが加算された画像 $\boldsymbol{x}_t \in \mathbb{R}^{H\times W\times 3}$ を出力する．加算されるノイズは以下の正規分布で表される確率分布 $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ に従う．確率分布 $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ および $\boldsymbol{x}_t$ は以下のように表せる．

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) := \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \sqrt{1-\bar{\alpha}_t}\boldsymbol{I}\right),$$

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},$$

ここで，$\boldsymbol{I}$ は単位行列を表す．また，α_t および $\bar{\alpha}_t$ は以下のように定義する．

$$\alpha_t := 1-\beta_t, \quad \bar{\alpha}_t := \prod_{s=1}^{t}\alpha_s,$$

ここで，β_t は t 回目に加算されるノイズの重みである．

4.2.2 Crossmodal Reverse Decoder

Crossmodal Reverse Decoder では $(\boldsymbol{x}_t, \boldsymbol{H}_{\mathrm{it}}^{(i)}, p(\hat{\boldsymbol{y}}_{\mathrm{it}}))$ を入力とし，$H\times W$ の各ピクセルにおける対象物のクラスである確率 $p(\hat{\boldsymbol{y}}_{\mathrm{diff}})$ を出力する．

はじめに，$\boldsymbol{x}_t$ を用いて，Forward Encoder における t 回

目に加算されたノイズ $\boldsymbol{\epsilon}_t \in \mathbb{R}^{H\times W\times 3}$ を求める．予測ノイズ $\hat{\boldsymbol{\epsilon}}_t^{(n_b)} \in R^{H_{n_b}\times W_{n_b}\times C_{n_b}}$ は，事前学習された拡散確率モデル [Ho 20] に対して，$\boldsymbol{x}_t$ を入力して抽出する．ここで，n_b は UNet 構造における各層のインデックス番号を表し，H_{n_b}, W_{n_b}, C_{n_b} はそれぞれ，第 n_b 層目の画像サイズおよびチャネル数を表す．次に，以下の式に従い，マルチモーダル特徴量 $\boldsymbol{H}'^{(n_b)}_{\text{seg}} \in \mathbb{R}^{H_{n_b}\times W_{n_b}\times C_{n_b}}$ を得る．

$$\hat{\boldsymbol{x}}_{t-1}^{(n_b)} = \boldsymbol{x}_t^{(n_b)} - \hat{\boldsymbol{\epsilon}}_t$$

$$\boldsymbol{H}'^{(n_b)}_{\text{seg}} = \hat{\boldsymbol{x}}_0^{(n_b)} + \boldsymbol{H}_{\text{it}}^{(n_b)}$$

ここで，$\hat{\boldsymbol{x}}_t^{(n_b)}$ は $\boldsymbol{x}_t^{(n_b)}$ の予測値を示す．次に，$\boldsymbol{H}'^{(n_b)}_{\text{seg}}$ に対してバイリニア補間を行うことで，リサイズされた特徴量 $\boldsymbol{H}_{\text{seg}}^{(n_b)} \in \mathbb{R}^{H_1\times W_1\times C_{n_b}}$ を抽出する．得られた各 $\boldsymbol{H}_{\text{seg}}^{(n_b)}$ をすべてチャネル方向に結合することで，マルチモーダル特徴量 $\boldsymbol{H}_{\text{seg}} \in \mathbb{R}^{H\times W\times C_{\text{seg}}}$ を求める．ここで，C_{seg} は特徴量の次元を表す．最後に，$\boldsymbol{H}_{\text{seg}}$ および $p\,(\hat{\boldsymbol{y}}_{\text{it}})$ を用いて，以下のように diffusion step におけるマスクである確率の推定値 $p\,(\hat{\boldsymbol{y}}_{\text{diff}})$ を求める．

$$\Delta p = f_{\text{BN}}(\text{ReLU}(f_{\text{FC}}(\boldsymbol{H}_{\text{seg}})))$$

$$p(\hat{\boldsymbol{y}}_{\text{diff}}) = p\,(\hat{\boldsymbol{y}}_{\text{it}}) + \Delta p,$$

ここで，$\boldsymbol{z}_{\text{diff}}$ は，$H \times W$ の 2 値正解マスク画像 $\boldsymbol{y}$ と $p\,(\hat{\boldsymbol{y}}_{\text{it}})$ との差分の予測値を表す．また，f_{BN}, f_{FC} はそれぞれ，バッチ正規化，線形結合を表す．最終的な出力である $H \times W$ の 2 値予測マスク画像 $p(\hat{\boldsymbol{y}}_{\text{diff}})$ は，$\hat{\boldsymbol{y}}_{\text{diff}}$ において 0.5 以上の値を 1，それ以外を 0 としたものとする．

intermediate training step における損失関数には交差エントロピー誤差を，diffusion step における損失関数には平均絶対誤差を用いた．

5. 実験

5.1 データセット

本研究で扱う OSMI タスクでは，対象物に対する家事作業の命令文と画像，および対象物の正解マスク画像が必要である．しかし，我々の知る限り，これらの要件を満たす既存のデータセットが存在しなかったため，全ての要件を満たす SHIMRIE データセットを構築した．本データセットに含まれるマスク画像は，Matterport3D データセットに含まれるボクセル単位でのクラス情報，および REVERIE データセットに含まれる対象物のバウンディングボックス情報を用いて収集した．命令文は REVERIE-dataset から収集することで，正解マスク画像と対応させた．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．

SHIMRIE データセットには，4341 枚の画像に対応する，11371 の命令文とマスクとのペアが含まれている．命令文は英語で記述されており，語彙サイズは 3558，全単語数は 196541 語，平均文長は 18.8 である．SHIMRIE データセットでは，全 11371 サンプルのうち，10153 サンプルを訓練集合に，856 サンプルを検証集合に，362 サンプルをテスト集合にそれぞれ使用した．分割の方法は REVERIE-dataset を踏襲している．また，検証集合は 582 サンプルの seen 集合，および 274 サンプルの unseen 集合から構成されており，テスト集合は unseen 集合のみで構成されている．ここで，seen 集合，unseen 集合はそれぞれ，訓練集合において既知の環境，未知の環境を表す．

intermediate training step において，訓練集合はモデルの学習に，検証集合はハイパーパラメータを調整するために使用した．また，テスト集合はモデルの性能評価に使用した．また diffusion step においては，検証集合をモデルの学習，テスト集合をモデルの性能評価に使用した．

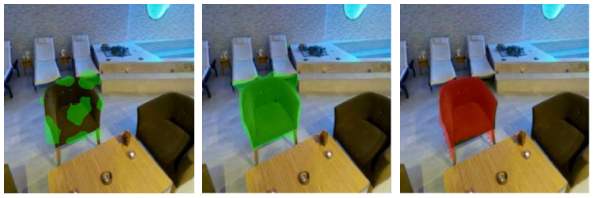
5.2 実験設定

表 1 に提案手法におけるハイパーパラメータ設定を示す．intermediate training step，diffusion step における訓練可能なパラメータ数はそれぞれ 123M，1.09M であり，積和演算数は 508G, 71.2G であった．提案手法の学習にはメモリ 24GB 搭載の GeForce RTX 3090 および Intel Core i9-10900KF，また 64GB の RAM を使用した．提案手法における訓練時間はそれぞれ，intermediate training step で約 1 時間 45 分，diffusion step で約 1 時間 25 分であった．また，1 サンプルあたりの推論時間はそれぞれ，intermediate training step で約 0.039 秒，diffusion step で約 0.36 秒であった．提案手法では，diffusion step での学習時において，1 ステップごとに検証集合による損失を算出し，3 エポック目以降で最も低い損失を記録した時のテスト集合における精度を，最終的な精度とした．

表 1: MDSM におけるパラメータ設定.

	Intermediate. step	Diffusion step
エポック数	11	5
バッチサイズ	16	1
学習率	5×10^{-5}	1×10^{-3}
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.99$)	Adam ($\beta_1 = 0.9, \beta_2 = 0.99$)

“Go to the laundry room and straighten the picture closest to the light switch.”



“Go to the lounge and remove the small brown chair facing the counter.”

(a) ベースライン手法　(b) 提案手法　(c) 正解マスク

図 3: 成功例の定性的結果. (a) LAVT [Yang 22], (b) 提案手法, (c) 正解マスク画像.

5.3 実験結果

5.3.1 定量的結果

表 2 にベースライン手法 (i) 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．また，提案手法は diffusion step の有無による 2 つの場合 (ii)(iii) で実験を行った．ベースライン手法として LAVT [Yang 22] を使用した．LAVT は OSMI タスクと深く関連する RES タスクにおいて，良好な性能が報告さ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．

表 2 より，主要尺度である mIoU において，(i)，(ii) はそれぞれ 24.27, 30.19%であり，(ii) が 5.92 ポイント上回った．さらに，oIoU および P@k のいずれにおいても (ii) は (i) と同等以上の性能であった．また，(iii) における mIoU は 34.40%であり，(i) を 10.13 ポイント上回った．さらに，oIoU および P@k のいずれにおいても (iii) は (i) を上回った．したがって，(iii) が最も良好な性能であった．この (i) と (iii) における性能差は統計有意であった（$p < 0.05$）

5.3.2 定性的結果

図 3 に定性的結果を示す．図において，パネル (a), (b), (c) はそれぞれ，ベースライン手法の予測マスク，提案手法の予測マスク，正解マスクを示す．上段における命令文は，“Go to the laundry room and straighten the picture closest to the

表 2: 比較実験の定性的結果．“ME” は intermediate training step における Multimodal Encoder モジュールを表す．

[%]	Diff. step	ME	mIoU	oIoU	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9
(i) LAVT [Yang 22]			$24.27_{\pm 3.15}$	$22.25_{\pm 2.85}$	$21.27_{\pm 5.66}$	$13.37_{\pm 3.74}$	$5.97_{\pm 2.50}$	$0.94_{\pm 0.38}$	$0.00_{\pm 0.00}$
(ii) Ours		✓	$30.19_{\pm 3.98}$	$27.08_{\pm 2.89}$	$31.66_{\pm 6.52}$	$23.04_{\pm 4.66}$	$10.33_{\pm 1.63}$	$1.55_{\pm 1.36}$	$0.00_{\pm 0.00}$
	✓	✓	$\mathbf{34.40}_{\pm 3.79}$	$\mathbf{31.59}_{\pm 3.03}$	$\mathbf{36.63}_{\pm 6.14}$	$\mathbf{27.79}_{\pm 5.28}$	$\mathbf{16.30}_{\pm 2.98}$	$\mathbf{6.41}_{\pm 1.19}$	$\mathbf{0.66}_{\pm 0.62}$

表 3: 条件 (b) における Ablation study．“Feat.” は intermediate training step で得られる特徴量を表す．

[%]	Cond.	Feat.	mIoU	oIoU	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9
	(b-1)	$\boldsymbol{H}_{\mathrm{it}}^{(0)}$, $\boldsymbol{H}_{\mathrm{it}}^{(1)}$, $\boldsymbol{H}_{\mathrm{it}}^{(2)}$, $\boldsymbol{H}_{\mathrm{it}}^{(3)}$	$34.09_{\pm 4.14}$	$31.57_{\pm 3.07}$	$35.52_{\pm 6.04}$	$27.29_{\pm 4.93}$	$\mathbf{16.35}_{\pm 3.16}$	$6.35_{\pm 1.22}$	$\mathbf{1.82}_{\pm 1.33}$
	(b-2)	$\boldsymbol{H}_{\mathrm{it}}^{(0)}$, $\boldsymbol{H}_{\mathrm{it}}^{(1)}$, $\boldsymbol{H}_{\mathrm{it}}^{(2)}$	$\mathbf{34.40}_{\pm 3.79}$	$\mathbf{31.59}_{\pm 3.03}$	$\mathbf{36.63}_{\pm 6.14}$	$\mathbf{27.79}_{\pm 5.28}$	$16.30_{\pm 2.98}$	$\mathbf{6.41}_{\pm 1.19}$	$0.66_{\pm 0.62}$
	(b-3)	$\boldsymbol{H}_{\mathrm{it}}^{(0)}$, $\boldsymbol{H}_{\mathrm{it}}^{(1)}$, $\boldsymbol{H}_{\mathrm{it}}^{(3)}$	$33.68_{\pm 4.37}$	$30.61_{\pm 4.17}$	$35.80_{\pm 6.73}$	$26.46_{\pm 4.66}$	$15.69_{\pm 3.88}$	$6.08_{\pm 1.58}$	$0.50_{\pm 0.41}$
	(b-4)	$\boldsymbol{H}_{\mathrm{it}}^{(0)}$, $\boldsymbol{H}_{\mathrm{it}}^{(2)}$, $\boldsymbol{H}_{\mathrm{it}}^{(3)}$	$33.44_{\pm 4.52}$	$30.51_{\pm 3.89}$	$35.03_{\pm 6.36}$	$26.85_{\pm 6.02}$	$15.91_{\pm 4.36}$	$5.25_{\pm 0.82}$	$1.66_{\pm 1.48}$
	(b-5)	$\boldsymbol{H}_{\mathrm{it}}^{(1)}$, $\boldsymbol{H}_{\mathrm{it}}^{(2)}$, $\boldsymbol{H}_{\mathrm{it}}^{(3)}$	$32.54_{\pm 4.97}$	$29.97_{\pm 4.14}$	$35.30_{\pm 6.72}$	$26.24_{\pm 6.26}$	$13.15_{\pm 3.77}$	$3.26_{\pm 1.95}$	$0.50_{\pm 0.41}$

light switch” であり，対象物は壁の手前にかけられた絵画である．ベースライン手法では，誤って対象物ではない屋外の領域を予測している．それに対して，提案手法では正解マスクと同じ絵画の領域を予測している．同様に，下段における命令文は，“Go to the lounge and remove the small brown chair facing the counter” であり，対象物は左側の茶色い椅子である．提案手法では，ベースライン手法と比較して不足の少ない適切な領域を予測している．

5.3.3 Ablation studies

Ablation study として，以下の条件を定めた．

(a) 並列クロスモーダル特徴抽出機構の有無: intermediate training step において，並列クロスモーダル特徴抽出機構を取り除き，その有効性を調査する．

(b) Crossmodal Reverse Decoder における $\boldsymbol{H}_{\mathrm{it}}^{(i)}$ の有無: diffusion step において，$\boldsymbol{H}_{\mathrm{it}}^{(i)}$ の性能への寄与を調査する．

表 2，3 にそれぞれ条件 (a)，(b) における定量的結果を示す．表 2 より，条件 (a-1) における mIoU は，条件 (a-2) よりも 5.92 ポイント低かった．また，oIoU および P@k において，P@0.9 を除き大きく悪化した．これより，並列クロスモーダル特徴抽出機構は性能向上に寄与しているといえる．表 3 より，条件 (b-3)，(b-4)，および (b-5) における mIoU は条件 (b-1) と比較して，それぞれは 1.55，0.65，0.41 ポイント低かった．一方，条件 (b-2) では，条件 (b-1) よりも，mIoU が 0.31 ポイント高かった．これより，$\boldsymbol{H}_{\mathrm{it}}^{(0)}$ が最も性能向上のために重要だといえる．これは，解像度の高い特徴量である $\boldsymbol{H}_{\mathrm{it}}^{(0)}$ が画素単位のセグメンテーションに有効であったためだと考えられる．

6. おわりに

本論文では，実世界の室内環境において，物体操作に関する命令文の対象物をセグメントする OSMI タスクを扱った．本研究の貢献は以下である．

- 2 段階からなるマルチモーダルセグメンテーションモデル MDSM を提案した．
- CLIP [Radford 21] および Swin Transformer [Liu 21] による並列クロスモーダル特徴抽出機構を導入した．
- DDPM [Ho 20] を拡張し，マルチモーダル特徴量を扱う Crossmodal Reverse Decoder を導入した．
- 新しく構築した SHIMRIE データセット上で，標準的な尺度について MDSM はベースライン手法を上回った．

謝辞

本研究の一部は，JSPS 科研費 20H04269，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

[Devlin 19] Devlin, J., Chang, M., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *NAACL-HLT*, pp. 4171–4186 (2019)

[Ho 20] Ho, J., Jain, A., and Abbeel, P.: Denoising Diffusion Probabilistic Models, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 6840–6851 (2020)

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

[Liu 19] Liu, R., Liu, C., Bai, Y., and Yuille, A.: CLEVR-Ref+: Diagnosing Visual Reasoning with Referring Expressions, in *CVPR*, pp. 4185–4194 (2019)

[Liu 21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in *ICCV*, pp. 10012–10022 (2021)

[Qi 20] Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W. Y., Shen, C., and Hengel, A.: REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments, in *CVPR* (2020)

[Radford 21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., et al.: Learning Transferable Visual Models From Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021)

[Rombach 22] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models, in *CVPR*, pp. 10684–10695 (2022)

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

[Yamamoto 19] Yamamoto, T., Terada, K., Ochiai, A., Saito, F., 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)

[Yang 22] Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., and Torr, P.: LAVT: Language-Aware Vision Transformer for Referring Image Segmentation, in *CVPR*, pp. 18155–18165 (2022)