ENCHANT: 大規模言語モデルを用いた仮説生成に基づくクロスモーダル説明文生成

○平野慎之助 †，小松拓実 †，和田唯我 †，神原元就 †，畑中駿平 †, 平川翼 ‡，山下隆義 ‡，
藤吉弘亘 ‡，杉浦孔明 †
† 慶應義塾大学　　‡ 中部大学

本研究では，生活支援ロボット物体が物体を配置する際に生じる衝突を事前に予測し，衝突に関する説明文生成を行う．既存手法では衝突物体の特徴抽出が不適切であり，衝突に関係する物体の特定が不十分である．本研究では，大規模言語モデルを用いてデータ拡張を行う Enhanced Nearest-neighbor Captioning with Hypothesis AugmeNTation (ENCHANT) を提案する．提案手法の評価を行うため，一つの動画につき平均 3.2 人によってアノテーションされた 4,042 サンプルからなる新たなデータセットを構築した．実験の結果，提案手法がベースライン手法をすべての評価指標で上回った．

1. はじめに

高齢化が進行する現代社会において，在宅介助者の不足が問題化しており，生活支援ロボットはその解決策となり得る [1]．物体が無秩序に配置された環境において日常品を操作するタスクは，生活支援ロボットにとって重要なタスクである．一方，物体を配置する際に他の物体と衝突してロボット自身や物体が破損する危険性がある．そこで，タスク実行に伴う危険性を事前に予測し，自然言語を用いてユーザに説明する機能は有用であるが，そのような機能は不十分である．

そこで本研究では，生活支援ロボットが物体を配置する際に生じる衝突を事前に予測し，衝突に関する説明文を生成することを目的とする．生活支援ロボットの配置タスクにおける衝突予測には Nearest Neighbor Future Captioning (NNFC) [2] をはじめとする先行研究があるが，既存手法は衝突物体の特徴抽出が不適切であり，衝突する物体に関する生成品質が不十分である．

本研究では，衝突に関する説明文を生成する future captioning モデルである Enhanced Nearest-neighbor Captioning with Hypothesis AugmeNTation (ENCHANT) を提案する。ENCHANT は大規模言語モデルによる生成文を用いてデータ拡張を行う Nearest Neighbor Augmentation Module (NNAM) を導入することにより，豊富な表現を持つ出力を得ることができる．また，画像および言語の特徴抽出を対称的に行う Parallel Cross Attentional Decoder (PCAD) を導入することにより，衝突に関連する物体を適切に反映したキャプションを生成することができる．さらに，Segment Feature Extractor (SFE) は配置領域の特徴量をセグメンテーションを用いて抽出することにより，配置領域の物体特徴を適切に抽出することができる．提案手法における新規性は以下の通りである．

- 大規模言語モデルによる生成文を用いてデータ拡張を行う NNAM を導入し，生成された仮説に基づき衝突に関する説明文を生成する ENCHANT を提案する．
- 画像および言語の特徴抽出を対称的に行う PCAD を導入する．
- 衝突リスクの注意マップ [3] と，セグメンテーションモデル [4] を用いて配置領域の特徴量を抽出する SFE を導入する．

図 1　対象タスクの例．左：実験環境．右：配置領域．

2. 関連研究

キャプション生成分野では多くの研究が行われている [5] [6]．[5] では，画像キャプション生成に関する研究を，Encoder-Decoder フレームワーク，注意機構および学習戦略の 3 つに分類し網羅的にまとめている．加えて，標準的なデータセットおよび評価尺度についての包括的な総括を行い，MS COCO [7] で学習された最新手法の比較を行っている．

RFCM [8] や NNFC [2] は，日常タスクにおける危険性を予測し，説明文を生成する future captioning に取り組んだ手法である．また，CRT [9] は Transformer [10] を用いた代表的な指示文付与手法であり，対象物体および目標領域を含んだ物体移動指示文を生成する．

3. 問題設定

本論文では，衝突に関する future captioning を扱う．ここで，future captioning とは，動作前の画像から将来の状況の説明文を生成するタスクである．本タスクでは，物体配置時における物体間の衝突を予測し，予測された衝突に関する説明文を出力することが望ましい．

図 1 にシミュレーション環境および配置領域を示す．例えば，図 1 の例において,「把持しているペットボトルをカメラの上に配置しようとして衝突し，ペットボトルが倒れる」といった説明文を生成することが望ましい．本研究では，以下の入出力を想定する．

- **入力:** 対象物体および配置領域の RGBD 画像
- **出力:** 対象物体配置時の衝突に関する説明文

本論文で使用する用語を以下に定義する．

- **対象物体:** ロボットが把持する日常的な物体
- **配置領域:** ロボットが対象物体を置く場所
- **障害物:** 配置領域にすでに配置されている物体

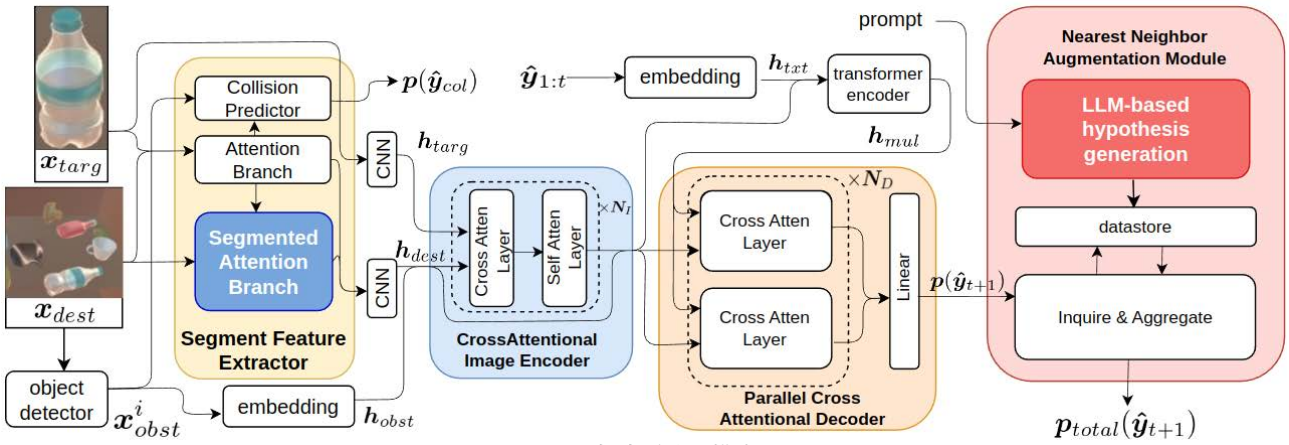


図 2　提案手法の構成

```
A: アームがペットボトルを置こうとして、砂糖の容器に衝突して倒れる
B: アームがペットボトルを置こうとして、砂糖の容器に衝突して弾き飛ばされる
A: アームが置いたコップが机の上の砂時計にぶつかって少し移動する
B: アームが置いたコップが机の上の砂時計にぶつかって転がる
A: {caption}
B:
```

図 3　データ拡張に使用したプロンプト

本タスクにおいて，ロボットは対象物体を把持した状態で配置領域の前にいることを前提とする．また，ロボットは十分なスペースがある領域に配置するという方策に基づいて対象物体を配置する．さらに，入力として画像のみを扱い，動画は扱わないものとする．本タスクでは、自然言語生成タスクにおける標準的な尺度を用いる．具体的には，BLEU4，METEOR，ROUGE-L，CIDEr-D，JaSPICE [11] の 5 つの指標を用いて，人間による付与文との比較を評価する．

本研究は衝突を伴うタスクを扱うため，実機ロボットを扱う際にロボット自身および物体の故障を招くリスクがある．そのため，実機ロボットの故障のリスクがないシミュレーション環境を用いる．

4.　提案手法

提案手法では，future captioning において，大規模言語モデルを用いてデータ拡張を行う．図 2 に提案手法の概要図を示す．提案手法の主要なモジュールは 4 つであり，それぞれ Segment Feature Extractor (SFE), Cross Attentional Image Encoder (CAIE), Parallel Cross Attentional Decoder (PCAD), Nearest Neighbor Augmentation Module (NNAM) である．本論文で提案する大規模言語モデルによる生成文を用いたデータ拡張は，既存のキャプション生成タスクを中心に幅広い手法に適用できると考えられる．

モデルの入力は以下である．

$$\{\boldsymbol{x}_{dest}, \boldsymbol{x}_{targ}, \boldsymbol{x}^{(i)}_{obst} \mid i = 1, ...N_o\}$$

ここで，$\boldsymbol{x}_{dest} \in \mathbb{R}^{c\times h\times w}, \boldsymbol{x}_{targ} \in \mathbb{R}^{c\times h\times w}, \boldsymbol{x}^{(i)}_{obst} \in \mathbb{R}^{1024}$ はそれぞれ，配置領域の RGBD 画像，対象物体の RGBD 画像，i 番目の障害物における物体領域の特徴量を表す．また，N_o は Faster R-CNN によって検出した障害物の個数を表す．$\boldsymbol{x}^{(i)}_{obst}$ は，$\boldsymbol{x}_{dest}$ から Faster R-CNN を用いて抽出した各矩形領域の特徴量とする．ここで，矩形領域の特徴量には RoI Pooling layer における FC 第 6 層の出力を採用する．また，Region Proposal Network には ResNet50 [12] を用いる．$\boldsymbol{x}^{(i)}_{obst}$ および $\boldsymbol{x}_{targ}$ は 224x224 にリサイズ後，標準化を行い，それぞれ $\boldsymbol{h}^{(i)}_{obst}$ および $\boldsymbol{h}_{targ}$ とする．

4.1　SFE

本モジュールは [3] を拡張した Collision Predictor, Attention Branch および Segmented Attention Branch で構成される．Collision Prediction Branch は対象物体配置時の衝突予測を行い，Attention Branch は衝突予測に対する attention map を生成する．また，Segmented Attention Branch は入力画像から SAM [4] を用いて得られたセグメンテーション画像と attention map の重畳を行う．

本モジュールにおける入力は，$\{\boldsymbol{x}_{dest}, \boldsymbol{x}_{targ}, \boldsymbol{x}^{(i)}_{obst} | i = 1, ...N_o\}$ であり，出力は衝突確率の予測値 $p(\hat{y}_{col})$ および attention map とセグメンテーション画像を重畳した画像 $\boldsymbol{i} \in \mathbb{R}^{c_2\times w_2\times h_2}$ である．ここで，$\hat{y}_{col}$ は衝突の有無を表す 2 値ラベルを表す．また，c_2, h_2 および w_2 はそれぞれ $\boldsymbol{i}$ のチャンネル数，縦幅および横幅を表す．$\boldsymbol{i}$ について，$\boldsymbol{x}_{dest}$ と同様のサイズにリサイズすることで出力 $\boldsymbol{h}_{dest}$ とする．

4.2　CAIE

本エンコーダは同一の構造を持つ N_I 層から構成される．各層の入力は $\boldsymbol{h}^{(e)}_{enc} = \{\boldsymbol{h}^{(e-1)}_{img}, \boldsymbol{h}_{targ}\}$ である．ただし，$\boldsymbol{h}^{(0)}_{enc} = \{\boldsymbol{h}_{dest}, \boldsymbol{h}_{targ}, \boldsymbol{h}^{(i)}_{obst} | i = 1, ...N_o\}$ とする．各層は，Transformer [10] における encoder layer の自己注意を相互注意に変更した層に対して $\boldsymbol{h}^{(e-1)}_{img}$ を query, $\boldsymbol{h}_{targ}$ を key および value として入力し，$\boldsymbol{s}^{(e)}_{img}$ を得る．その後，$\boldsymbol{s}^{(e)}_{img}$ に Multi-Head Attention，残差結合および layer normalization を適用し，各層の出力 $\boldsymbol{h}^{(e)}_{img}$ を得る．最終的に，$\boldsymbol{h}_{imgs} = \{\boldsymbol{h}^{(N_I)}_{img}, \boldsymbol{h}_{targ}\}$ を出力とする．

4.3　PCAD

本デコーダは次のトークンを自己回帰的に予測する．本デコーダは同一の構造を持つ N_d 層で構成される．各層は Parallel Cross Attention 層および FeedForward-Network(FFN) 層 [10] から構成される．また，各層の入力は以下で与えられる．

$$\boldsymbol{h}^{(d)}_{dec} = \begin{cases} (\boldsymbol{h}_{mul}, \boldsymbol{h}_{imgs}) & (d = 0) \\ (\boldsymbol{h}^{(d-1)}_{dec}, \boldsymbol{h}_{imgs}) & (d = 1, ..., N_d) \end{cases}$$

はじめに，transformer encoder [10] に $\{\boldsymbol{h}_{img}, \boldsymbol{h}_{txt}\}$ を入力し，$\boldsymbol{h}_{mul}$ を得る．ここで，$\boldsymbol{h}_{img}$ および $\boldsymbol{h}_{txt}$ はそれぞれ transformer encoder および言語特徴量の出力である．Parallel Cross Attention 層では $\boldsymbol{h}_{imgs}$ および $\boldsymbol{h}_{mul}$ に対する Multi-Head Attention の計算を対称的

に行う．具体的には，$\boldsymbol{a}_{mul} = \mathrm{CrossAttn}(\boldsymbol{h}_{mul}^{(d)}, \boldsymbol{h}_{imgs})$ および $\boldsymbol{a}_{imgs} = \mathrm{CrossAttn}(\boldsymbol{h}_{imgs}, \boldsymbol{h}_{mul}^{(d)})$ を計算する．$\boldsymbol{h}_{dec}^{(d)} = \mathrm{FFN}(\{\boldsymbol{a}_{mul}^{(d)}, \boldsymbol{a}_{imgs}^{(d)}\})$ を本デコーダの d 層目における出力とする．続いて，N_d 層目の出力 $\boldsymbol{h}_{dec}^{(N_d)}$ に対して，全結合層およびソフトマックス関数を用いて予測確率 $p(\hat{\boldsymbol{y}}_{t+1})$ を計算する．

4.4 NNAM

本モジュールは大規模言語モデルによる生成文を用いて訓練データの拡張を行い，PCAD の出力 $p(\hat{\boldsymbol{y}}_{t+1})$ に対して，k 近傍法による rescoring を行う．本モジュールでは，大規模言語モデルにプロンプトを入力し，訓練データを拡張する．図 3 に本研究で使用したプロンプトを示す．ここで，{caption} は各サンプルに対応するキャプションを表す．入力は $\boldsymbol{z}_t$ および $p(\hat{\boldsymbol{y}}_{t+1})$ であり，出力は $p_{total}(\hat{\boldsymbol{y}}_{t+1})$ である．ここで，$\boldsymbol{z}_t$ は $\boldsymbol{y}_{i:t}$ に対応する PCAD の最終層に入力される埋め込み表現を表す．

はじめに，訓練データの各サンプルに対応するキャプションを，図 3 に示すプロンプトを用いて大規模言語モデルに入力し，キャプションの衝突後の事象に関する表層表現を変更した新たなサンプルを得る．続いて，$\boldsymbol{z}_t$ をもとに距離関数 $d(\cdot,\cdot)$ に従い，データストアから k 近傍法を用いて N_{knn} のペア $\{\boldsymbol{k}_n, \boldsymbol{v}_n | n = 1, ...N_{knn}\}$ を得る．ここで，距離関数には二乗誤差を用いた．データストアは，大規模言語モデルから得られる文を加えた新たなサンプル群に対して以下のように定義される．

$$\{\boldsymbol{z}_{i:t}, \hat{\boldsymbol{y}}_{i,t+1} \mid i = 1, ..., N, t = 1, ..., T-1\}$$

また，問い合わせた N_{knn} のペア $\{\boldsymbol{k}_n, \boldsymbol{v}_n \mid n = 1, ..., N_{knn}\}$ を以下の式に従い，集計する．

$$p_{knn}(\hat{\boldsymbol{y}}_{t+1}) = \frac{1}{Z} V' \mathrm{softmax}(\boldsymbol{k}_{dist})$$

ここで，$\boldsymbol{k}_{dist} = \{d(\boldsymbol{k}_n, \boldsymbol{z}_t) \mid n = 1, ..., N_{knn}\}$ である．また，Z, V' はそれぞれ，規格化定数，$\boldsymbol{v}_i (i = 1, ..., N_{knn}$ を one-hot ベクトル化して並べたものを表す．最後に，以下の式に従い，$p_{total}(\hat{\boldsymbol{y}}_{t+1})$ を得る．

$$p_{total}(\hat{\boldsymbol{y}}_{t+1}) = \lambda_{knn} p_{knn}(\hat{\boldsymbol{y}}_{t+1}) + (1 - \lambda_{knn}) p(\hat{\boldsymbol{y}}_{t+1})$$

ここで，λ_{knn} はハイパーパラメータである．

4.5 損失関数

損失関数は以下で定義される

$$L = \lambda_{CE} L_{CE}(\boldsymbol{y}_{t+1}, p(\hat{\boldsymbol{y}}_{t+1})) + \lambda_{NCE} L_{NCE}(\boldsymbol{h}_{img}, \boldsymbol{h}_{txt})$$

ここで，L_{CE}, L_{NCE} はそれぞれクロスエントロピー損失，InfoNCE 損失 [13] を表す．また $\lambda_{CE}, \lambda_{NCE}$ はハイパーパラメータであり，損失の重みを表す．

5. 実験設定

5.1 データセット

衝突に関する future captioning を扱ったデータセットとして BILA-caption 2.0 [8] が提案されている．BILA-caption 2.0 では，サンプル動画一つに対して一文のアノテーションのみが付与されており，これは複数の衝突が生じる本タスクにおいて不十分である．そのため，本研究では，一つの衝突動画につき複数のアノテーションが付与された BILA-caption 3.0 を構築した．

本データセットは，PonNet [3] と同様の生活支援ロボットシミュレータを利用してデータを収集した．すなわち，WRS2018VS [14] において使用されたシミュレータである SIGVerse [15] を拡張したシミュレータを使用した．シミュレーション環境を用いて，以下の手順でデータセットを作成した．

はじめに，環境および配置領域をランダムに選択した．ここで，環境および配置領域はそれぞれ 10 種類および 6 種類を用いた．続いて，シミュレータを用いて配置領域上に障害物をランダムに配置した．この時，障害物として 25 種類の物体を用いた．生活支援ロボットは対象物体を把持してから配置領域に配置するものとし、対象物体は予め定めた 14 種類の物体群からランダムに選択するものとする．また，対象物体は平面上に置かれているものとし，カメラとの距離は一定とする．対象物体の RGBD 画像を生活支援ロボットのヘッドカメラを用いて観測させるとともに、対象物体を把持させた．生活支援ロボットが配置する際，十分なスペースがある領域に配置するという方策に基づき対象物体を配置させた．このとき，配置領域の RGBD 画像を，生活支援ロボットのヘッドカメラを用いて収集した．続いて，対象物体配置時，危険な衝突が発生したシーンのみを抽出した．最後に,「把持中の空き缶がペットボトルと衝突する」のような，対象物体配置時における危険に関する説明文を次に示す方法で付与した．

アノテータには，ロボット視点および第三者視点から撮影した動画を与えたうえで，対象物体配置時に発生する衝突および落下に関する説明文を付与させた．また，衝突が複数存在する場合は，最も危険性の高い衝突に関する説明文を付与させた．

本データセットは 4,042 サンプルで構成される．1 サンプルは，配置領域の RGBD 画像，対象物体の RGBD 画像および衝突に関する日本語の説明文からなる．説明文の語彙サイズは 1,254 であり，平均文長は 22 である．また，アノテータの数は 200 人であり，各動画につき平均 3.2 人のアノテータによりアノテーションされている．本実験では，データセットを訓練集合，検証集合，テスト集合に分割し，それぞれ 3,185 個，363 個，494 個のサンプルを含む．ここで，それぞれの集合では環境の割合が均一になるように分割を行った．

6. 実験結果

6.1 定量的結果

定量的結果を表 1 に示す．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．本研究ではベースライン手法として，NNFC [2] を用いた．ここで，NNFC は物体配置時の衝突に関する future captioing において良好な結果が得られているため選択した．生成文の評価には BLEU4, ROUGE-L, METEOR, CIDEr-D および JaSPICE を使用し，主要尺度を JaSPICE とした．

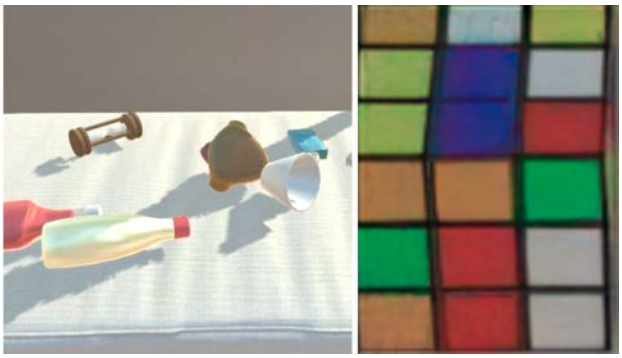
表 1 より，主要尺度である JaSPICE において，提案手法とベースライン手法はそれぞれ，19.37, 22.33 であり，提案手法が 2.96 ポイント上回った．BLEU4, METEOR, ROUGE-L および CIDEr-D においても同様に，提案手法がそれぞれ 25.92，28.98，45.60 および 39.85 であり，それぞれ 2.97 ポイント，1.64 ポイント，2.01 ポイントおよび 4.61 ポイント上回った．

6.2 定性的結果

図 4 に定性的結果を示す．図 4 において，左図および右図はそれぞれ配置領域及び対象物体の RGB 画像を表す．図 4 は，ルービックキューブおよびマヨネーズの容器が衝突し，衝突後，マヨネーズの容器が横に移動する例である．ベースライン手法は，衝突する物

表 1　定量的結果および Ablation Study

手法	BLEU4	METEOR	ROUGE-L	CIDEr-D	JaSPICE
NNFC [2]	22.95 ± 0.99	27.34 ± 0.36	43.59 ± 0.64	35.24 ± 2.05	19.37 ± 0.76
(a)	25.31 ± 0.96	28.90 ± 0.47	45.34 ± 0.65	37.48 ± 2.60	21.40 ± 0.67
(b)	24.71 ± 1.00	29.05 ± 0.14	45.27 ± 0.56	37.95 ± 2.20	21.60 ± 0.78
(c)	25.13 ± 1.13	**29.2 ± 0.35**	45.49 ± 0.77	38.03 ± 2.69	21.61 ± 0.39
Ours	**25.92 ± 0.55**	28.98 ± 0.40	**45.60 ± 0.51**	**39.85 ± 1.39**	**22.33 ± 0.60**



参照文	アームが掴んでいたルービックキューブをテーブルの上に置き、ルービックキューブとマヨネーズが衝突する
NNFC [2]	アームがルービックキューブを机の上に置こうとしたが、置こうとした場所にペットボトルと接触してしまい、ルービックキューブが棚の上で倒れる
Ours	アームがルービックキューブを机の上に置こうとしたが、マヨネーズの容器に衝突し、マヨネーズの容器が少し動く

図 4　定性的結果

体をペットボトルと誤って記述した．一方で，提案手法では，衝突する物体をマヨネーズの容器と適切に記述した．また，衝突後の危険性に関して「マヨネーズの容器が少し動く」と適切に記述した．

6.3　Ablation Study

以下の 3 つの条件を Ablation study に定めた．

(a) w/o 大規模言語モデルによる拡張: 大規模言語モデルによる拡張を行わないことで，提案手法の性能への影響を調査した．
(b) w/o PCAD: PCAD において，画像および言語の特徴抽出を対称的に行わず，言語特徴量を query, 画像特徴量を key および value とした．
(c) w/o Segmented Attention Branch: Segmented Attention Branch を取り除き，提案手法の性能への影響を調査した．

表 1 に Ablation Study の結果を示す．Ablation 条件 (a), (b), (c) および提案手法における JaSPICE の値はそれぞれ 21.40，21.60，21.61 および 22.33 であった．このことから，大規模言語モデルによるデータ拡張が提案手法の性能へ最も寄与していると考えられる．

7.　結論

本論文では，生活支援ロボットが物体を配置する際に発生する衝突に関する future captioning タスクを扱った．

本研究の主要な貢献は以下である．

- 大規模言語モデルによる生成文を用いてデータ拡張を行う Nearest Neighbor Augmentation Module を導入し，生成された仮説に基づき衝突に関する説明文を生成する ENCHANT を提案した．
- 画像および言語の特徴抽出を対称的に行う Parallel Cross Attentional Decoder を導入した．
- 配置領域の特徴量をセグメンテーションを用いて抽出する Segment Feature Extractor を導入した．
- 提案手法は全ての評価指標において，ベースライン手法の性能を上回った．

謝辞

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

参　考　文　献

[1] Y. Takashi, et al., “Development Human Support Robot as the Research Platform of a Domestic Mobile Manipulator,” ROBOMECH journal, vol.6, no.1, pp.1–15, 2019.
[2] 小松拓実，神原元就，畑中駿平，松尾榛夏，平川翼，山下隆義他，“Nearest Neighbor Future Captioning: 物体配置タスクにおける衝突リスクに関する説明文生成,” JSAI，2023.
[3] A. Magassouba, K. Sugiura, A. Nakayama, et al., “Predicting and Attending to Damaging Collisions for Placing Everyday Objects in Photo-Realistic Simulations,” Advanced Robotics, vol.35, no.12, pp.787–799, 2021.
[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, et al., “Segment Anything,” arXiv:2304.02643, 2023.
[5] Y. Ming, N. Hu, et al., “Visuals to text: A comprehensive review on automatic image captioning,” IEEE/CAA Journal of Automatica Sinica, vol.9, no.8, pp.1339–1365, 2022.
[6] M. Stefanini, M. Cornia, , et al., “From Show to Tell: a Survey on Deep Learning-based Image Captioning,” IEEE TRANS. PAMI, vol.45, no.1, pp.539–559, 2022.
[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, “Microsoft COCO: Common Objects in Context,” ECCV, pp.740–755, 2014.
[8] M. Kambara and K. Sugiura, “Relational Future Captioning Model for Explaining Likely Collisions in Daily Tasks,” ICIP, pp.2601–2605, 2022.
[9] M. Kambara and K. Sugiura, “Case relation transformer: A crossmodal language generation model for fetching instructions,” RAL, vol.6, no.4, pp.8371–8378, 2021.
[10] A. Vaswani, N. Shazeer, Noam Parmar, J. Uszkoreit, et al., “Attention is all you need,” Advances in neural information processing systems, vol.30, pp.5998–6008, 2017.
[11] 和田唯我，兼田寛大，杉浦孔明，“JaSPICE: 日本語における述語項構造に基づく画像キャプション生成モデルの自動評価尺度,” 言語処理学会第 29 回年次大会，2023.
[12] K. He, X. Zhang, S. Ren, et al., “Deep Residual Learning for Image Recognition,” CVPR, pp.770–778, 2016.
[13] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, et al., “Learning Transferable Visual Models from Natural Language Supervision,” ICML, pp.8748–8763, 2021.
[14] H. Okada, et al., “What Competitions were Conducted in the Service Categories of the World Robot Summit?,” Advanced Robotics, vol.33, no.17, pp.900–910, 2019.
[15] T. Inamura, C. Tan, et al., “Development of Robocup@ Home Simulation Towards Long-Term Large Scale HRI,” Robot World Cup XVII 17, pp.672–680, 2014.