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

Generating Descriptions for Possible Collisions in Object Placement Tasks

小松 拓実 *1 Takumi Komatsu
神原 元就 *1 Motonari Kambara
畑中 駿平 *1 Shumpei Hatanaka
松尾 榛夏 *1 Haruka Matsuo
平川 翼 *2 Tsubasa Hirakawa
山下 隆義 *2 Takayoshi Yamashita
藤吉 弘亘 *2 Hironobu Fujiyoshi
杉浦 孔明 *1 Komei Sugiura

*1慶應義塾大学 Keio University
*2中部大学 Chubu University

In this paper, we propose Nearest Neighbor Future Captioning Model that introduces Nearest Neighbor Language Model for future captioning of possible collisions, which enhances the model output by employing a nearest neighbors retrieval mechanism. Moreover, we introduce Collision Attention Module, which extracts attention regions of possible collisions, which enables our model to generate descriptions that adequately reflect the objects associated with possible collisions. The experimental results demonstrated that our method outperformed baseline methods, based on the standard metrics.

1. はじめに

高齢化が進行する現代社会において，在宅介護従事者の不足が深刻な社会問題となっている．生活支援ロボットはこうした社会問題に対する有望な解決策の一つとなり得る [Yamamoto 19]．生活支援ロボットにとって，散らかった環境下で日常品を操作するタスクは重要なタスクである．一方，生活支援ロボットが物体を操作する際に，他の物体と衝突し，ロボットアームや物体が破損する危険性がある．そのため，衝突の危険性を事前に予測し，自然言語でユーザに説明し，警告する機能は有用である．しかし，そのような機能は未だ不十分である．

本研究では，生活支援ロボットが物体を配置する際に発生する衝突を予測し，それに関する説明を生成することに焦点を当てる．例えば，生活支援ロボットが、机の上にペットボトルを置く際に，ケチャップの容器と衝突する可能性がある場合,「把持中のペットボトルがケチャップの容器と衝突する」のような説明を動作実行前にユーザに提示することが望ましい．将来に発生する衝突に関する説明生成を扱った先行研究の多くは，発生する衝突の領域に関して明示的にモデル化していないため，低品質な説明が生成される傾向がある [Yi 19, Kambara 22]．

本論文では，V&L の研究において十分に適用されていない Nearest Neighbor Language Model [Khandelwal 19] を導入し，最近傍検索機構を用いてモデルの出力を強化する Nearest Neighbor Future Captioing Model (NNFCM) を提案する．さらに,，衝突に関する配置領域の注目箇所を抽出する Collision Attention Module を導入する．このことにより，衝突の可能性のある領域を適切に反映した説明文を生成することができる．

本研究の主な新規性を以下に示す．

- 衝突に関する future captioning に NNLM を導入した NNFCM を提案する．
- 衝突に関する配置領域の注意箇所を抽出する Collision Attention Module を導入する．
- 配置領域および対象物体の関係性をモデル化する Cross Attentional Image Encoder を導入する．

2. 関連研究

キャプション生成分野には多くの研究がある [Xu 15,Lei 20]．過去の情報を用いたキャプション生成は，future captioning タスクや video captioning タスクなどのサブタスクに分けられる．

連絡先: 小松拓実，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，tak3k_1999@keio.jp

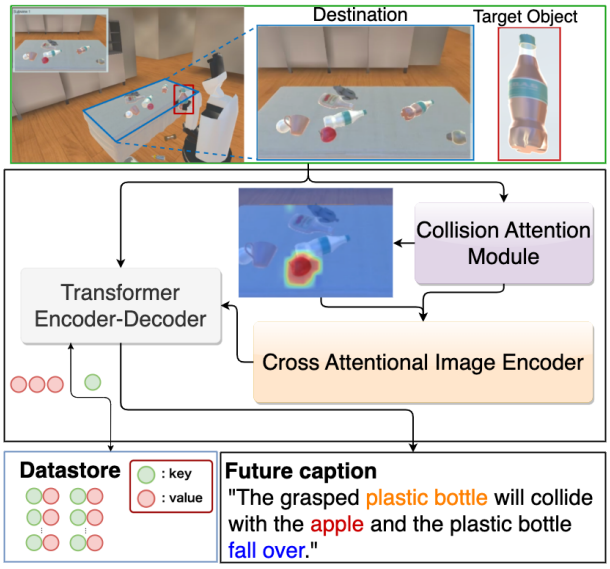


図 1: 提案手法の概要.

future captioning タスク [Hosseinzadeh 21] は，過去と現在の情報を基に，将来に発生する事象に対する説明を生成することを目的とする．既存手法として [Hosseinzadeh 21, Mahmud 21] などが挙げられる．video captioning タスクにおいて，多くの手法が提案されている [Wang 18, Sun 19]．近年の研究ではビデオキャプショニングにおける大規模事前学習モデルの有効性が示されている [Sun 19, Li 20]．例えば HERO [Li 20] は検索タスクを含む様々なタスクで既存手法を上回る性能を示している．

他にも，既存研究はロボティクス分野におけるキャプショニングタスクに取り組んでいる [Ogura 20, Magassouba 20]。例えば、[Ogura 20] は画像キャプショニングの一つである指示文付与タスクに取り組んでいる．このタスクは，与えられたシーン画像に対して，指示された物体を指定された目的地に移動させるための指示文を生成することを目的とする．ABEN [Ogura 20] および Multi-ABN [Magassouba 20] は ABN [Fukui 19] を用いて生成した指示文に詳細な説明を生成する手法である．

本提案手法は，RFCM [Kambara 22] などの衝突に関連し

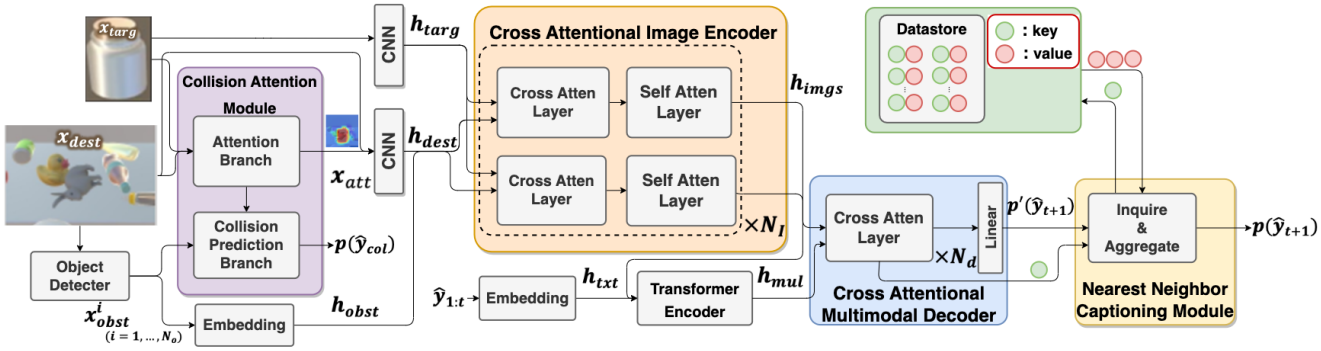


図 2: 提案手法のネットワーク構造

た future captioning モデルをベースとしている．本手法は，RFCM と異なり，タスク実行前の配置領域の画像を入力とする．また，future captioning に NNLM [Khandelwal 19] を導入した Nearest Neighbor Future Captioning Model を提案する．さらに，衝突に関する配置領域の注目箇所を抽出する Collision Attention Module を導入する．

3. 問題設定

本論文では，衝突に関する future captioning を扱う．ここで，futue captioning は，動作実行前の画像から将来の状況に説明を生成するタスクである．本論文で使用する用語を以下のように定義する．

- **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配置する領域
- **障害物:** 配置領域に既に置かれている物体

評価指標として，自然言語生成タスクにおける標準的な尺度である BLEU-4，METEOR，ROUGE-L および CIDEr-D を用いる．また，本研究では，実機ロボットではなく，シミュレーション環境を利用する．

4. 提案手法

提案手法のネットワーク構造を図2 に示す．ネットワークは主要なモジュールは 4 つであり，それぞれ Collision Attention Module (CAM)， Cross Attentional Image Encoder (CAIE)， Cross Attentional Multimodal Decoder (CAMD) および Nearest Neighbor Captioning Module (NNCM) である．各モジュールの詳細については本章で後述する．

4.1 入力

ネットワークの入力を以下のように定義する．

$$\{\boldsymbol{x}_{\text{dest}}, \boldsymbol{x}_{\text{targ}}, \boldsymbol{x}^i_{\text{obst}} | i = 1, ...N_o\},$$

ここで， $\boldsymbol{x}_{\text{dest}} \in \mathbb{R}^{c\times h\times w}$, $\boldsymbol{x}_{\text{targ}} \in \mathbb{R}^{c\times h\times w}$, $\boldsymbol{x}^i_{\text{obst}} \in \mathbb{R}^{1024}$ はそれぞれ配置領域の RGBD 画像，対象物体の RGBD 画像化 i 番目の障害物の物体領域の特徴量を表す．また，N_o は検出された障害物の数を表す．Faster R-CNN [Ren 15] を使用して，$\boldsymbol{x}_{\text{dest}}$ から $\boldsymbol{x}^i_{\text{obst}}$ を抽出した．ここで，ResNet50 [He 16] で構成される RoI Pooling layer における FC 第 6 層の出力を矩形領域の特徴量とした．さらに，特徴量の位置符号化として 7 次元ベクトル $x_1, y_1, x_2, y_2, w, h, w \times h$ を用いる．ここで，(x_1, y_1) および (x_2, y_2) はそれぞれ各矩形領域の左上と右上の頂点の座標を表す．また，w および h はそれぞれ幅と高さを表す．その後，1 層の全結合層と正規化層を用いて特徴量 $\boldsymbol{h}_{\text{obst}}$ を取得する．$\boldsymbol{x}_{\text{dest}}$ および $\boldsymbol{x}_{\text{targ}}$ は 224×224 にリサイズ後，標準化を行った．提案手法は，シミュレーション環境と同様の入力を用いることで，実機ロボットにも適用可能である．

4.2 CAM

本モジュールは衝突予測における配置領域の注目箇所を抽出するモジュールである．本モジュールは [Magassouba 21] を拡張した Collision Prediction Branch および Attention Branch から構成される．Collision Prediction Branch は対象物体配置時の衝突予測を行うモジュールであり，Attention Branch は衝突予測に対する attention map を生成するモジュールである．ここで，attention map は，各画素の重要度を可視化した画像を表す．詳細は [Magassouba 21] に示されている．

4.3 CAIE

本エンコーダーは対象物体および配置領域の関係をモデル化するモジュールである．本モジュールは N_I 層で構成される．本論文では，$N_I = 2$ とした．各層は Cross-Attention 層および Transformer [Vaswani 17] と同様の Multi-Head Attention 層，Feed-Forward Network 層で構成される．cross-attention 機構は任意の行列 X_A および X_B を用いて以下のように定義される．

$$\text{CrossAttn}(X_{\text{A}}, X_{\text{B}}) =$$

$$\text{softmax}(d^{-\frac{1}{2}}(W_{\text{Q}}X_{\text{A}})(W_{\text{K}}X_{\text{B}})^{\top})(W_{\text{V}}X_{\text{B}}),$$

ここで，W_Q, W_K, W_V は学習可能な重みを表す．各層の入力は以下で与えられる．

$$\boldsymbol{h}^{(i)} = \begin{cases} (\boldsymbol{h}_{\text{dest}}, \boldsymbol{h}_{\text{obst}}, \boldsymbol{h}_{\text{targ}}) & (i = 0) \\ (\boldsymbol{h}^{(i-1)}_{\text{img}}, \boldsymbol{h}^{(i-1)}_{\text{obst}}) & (i = 1, ..., N_I) \end{cases}$$

また，各層は以下の処理を行う．

$$\boldsymbol{\alpha}^{(i)}_{\text{img}} = \text{CrossAttn}(\boldsymbol{h}^{(i-1)}_{\text{img}}, \boldsymbol{h}_{\text{targ}}),$$

$$\boldsymbol{h}^{(i)}_{\text{img}} = \text{FFN}(\text{MHA}(\text{FFN}(\boldsymbol{\alpha}^{(i)}_{\text{img}}))),$$

ここで，FFN(·) および MMHA(·) はそれぞれ Feed-Forward Netowork，Masked Multi-Head Attention [Vaswani 17] を表す．$\boldsymbol{h}^{(i)}_{\text{obst}}$ は同様の処理で求める．また，$i(i = 1, ..., N_I)$ 番目の層の各 FFN 層後には，残差結合および層正規化を行った．最終的に，出力 $\boldsymbol{h}_{\text{imgs}}$ を以下のように計算する．

$$\boldsymbol{h}_{\text{imgs}} = (\boldsymbol{h}_{\text{img}}, \boldsymbol{h}'_{\text{obst}}) = (\boldsymbol{h}^{(N_I)}_{\text{img}}, \boldsymbol{h}^{(N_I)}_{\text{obst}}).$$

4.4 CAMD

本デコーダーは次のトークンを自己回帰的に予測する．本モジュールは N_d 層で構成される．本論文では，$N_d = 2$ とした．各層は，Cross Attention 層および Feed-Forward Network 層 [Vaswani 17] で構成される．各層の入力は以下で与えられる．

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

初めに，Masked Multi-Head Attention 層および Feed-Forward Network 層で構成される transfomer encoder に $(\boldsymbol{h}_{\text{img}}, \boldsymbol{h}_{\text{txt}})$ を入力とし，出力 $\boldsymbol{h}_{\text{mul}}$ を得る．デコーダーの各層は以下の処理を行う．

$$\boldsymbol{h}^{(d)}_{\text{dec}} = \text{FFN}(\text{CrossAttn}(\boldsymbol{h}^{(d-1)}_{\text{dec}}, \boldsymbol{h}_{\text{imgs}})),$$

表 1: **定量的結果および ablation study.**

Methods	BLUE-4	METEOR	ROUGE-L	CIDEr-D
SAT [Xu 15]	11.13 ± 1.12	16.98 ± 0.66	27.83 ± 0.68	41.67 ± 4.83
RFCM [Kambara 22]	12.27 ± 0.72	18.92 ± 0.89	28.43 ± 1.11	46.64 ± 2.98
Ours (w/o NNCM)	14.26 ± 0.72	21.92 ± 0.42	31.14 ± 0.58	59.57 ± 4.48
Ours (w/o CAM)	13.61 ± 0.88	20.77 ± 0.88	30.69 ± 0.71	56.94 ± 5.47
Ours (w/o CCIM)	13.43 ± 0.43	20.58 ± 1.05	30.58 ± 0.55	54.68 ± 3.05
Ours (full)	$\mathbf{14.26 \pm 0.23}$	$\mathbf{21.31 \pm 0.37}$	$\mathbf{31.40 \pm 0.24}$	$\mathbf{61.06 \pm 3.15}$

また，各 $i(i = 1, ..., N_d)$ 層の後に，$\boldsymbol{h}_{\mathrm{dec}}^{(d-1)}$ の残差結合および層正規化を行う．最終的に，N_d 層の出力 $\boldsymbol{h}_{\mathrm{dec}}^{(N_d)}$ に対して，全結合層およびソフトマックス関数を用いて次のトークンの予測確率 $p(\hat{\boldsymbol{y}}_{t+1})$ を計算する．

4.5 NNCM

本モジュールはデコーダーの出力を最近傍検索機構を用いて補強する．入力は $(\boldsymbol{z}_t, p(\hat{\boldsymbol{y}}_{t+1}))$ である．ここで，$\boldsymbol{z}_t$ は CAMD の最後の FFN 層に入力される潜在表現を表す．はじめに，$\boldsymbol{z}_t$ を基に距離関数 $d(\cdot, \cdot)$ に従い，以下で定義されるデータストアから k 近傍法を利用して N_{knn} のペア $\{(\boldsymbol{k}_n, \boldsymbol{v}_n)|n = 1, ..., N_{\mathrm{knn}}\}$ を取得するここで，距離関数は二乗誤差を用いた．また，$N_{\mathrm{knn}} = 64$ とした．データストアは以下のように定義される．

$$\{(\boldsymbol{z}_{i,t}, \hat{\boldsymbol{y}}_{i,t+1})|i = 1, ..., N, t = 1, ..., T - 1\},$$

ここで，N および T はそれぞれ学習集合のサンプル数および i 番目の説明文におけるトークン長を表す．また，$\boldsymbol{y}_{i,t+1}$ は i 番目の生成文 $\boldsymbol{y}_{i,1:t}$ における t 番目のトークンを表す．$\boldsymbol{z}_{i,t}$ は i 番目の正解文の 1 から t 番目のトークンである $\boldsymbol{y}_{i,1:t}$ から得る．

続いて，取得した N_{knn} のペア $\{\boldsymbol{k}_n, \boldsymbol{v}_n|n = 1, ..., N_{\mathrm{knn}}\}$ を以下の式に従い，集計する．

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

ここで，$\boldsymbol{k}_{\mathrm{dist}}$ は $\{d(\boldsymbol{k}_n, \boldsymbol{z}_t)|n = 1, ..., N_{\mathrm{knn}}\}$ を表す．また Z および V' はそれぞれ規格化定数，$\boldsymbol{v}_i(i = 1, ..., N_{\mathrm{knn}})$ を one-hot ベクトル化して並べたものを表す．最後に，再集計した予測確率 $p_{\mathrm{total}}(\hat{\boldsymbol{y}}_{t+1})$ を以下に従い計算する．

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

ここで，λ_{knn} は重みを表す．本論文では，$\lambda_{\mathrm{knn}} = 0.25$ とした．

4.6 損失関数

損失関数は以下で定義される．

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

ここで，λ_{CE} および λ_{NCE} はそれぞれ損失の重みを表し，それぞれ 0.9，5 とした．また，L_{CE} および L_{NCE} はそれぞれ交差エントロピー損失，InfoNCE 損失 [Radford 21] を表す．

5. 実験

5.1 データセット

本論文では，シミュレータ環境を用いてデータセットを構築した．具体的には，WRS2018 パートナーロボットチャレンジ／バーチャルスペースコンペティション [Okada 19] において使用されたシミュレータである SIGVerse [Inamura 14] を拡張したものを使用した．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，10 種類の環境および 6 種類の配置領域を用いた．続いて，配置領域に障害物をランダムに配置した．ここで，25 種類の障害物を使用した．次に，対象物体を 14 種類からランダムに選択した．ロボットに対象物体の RGBD 画像を取得させた後，対象物体を把持させた．ロボットは十分な空間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．続いて，衝突が衝突が発生したシーンのみを抽出した．最後に，アノテータに，ロボット視点および第三者視点から撮影した映像を提供したうえで，物体の配置に伴う衝突や転倒に関する説明を付与するように指示した．

BILA-caption 2.0 データセットのサイズは 1275 である．1 サンプルは，配置領域および対象物体の RGBD 画像および衝突に関する説明文から構成される．説明文の語彙サイズは 291 であり，平均文長は 20 である．BILA-caption 2.0 データセットをランダムに 8:1:1 となるように分割し，それぞれを訓練集合，検証集合，およびテスト集合とした．従って，訓練集合，検証集合，テスト集合はそれぞれ 1020 サンプル，128 サンプル，127 サンプルとした．訓練集合を用いてモデルのパラメータを更新し，検証集合を用いてハイパーパラメータの選択を行った．また，テスト集合を用いてモデルの性能の評価を行った．

5.2 実験設定

早期終了の条件として，以下の式で定義される k エポック目における generalization [Prechelt 98] を用いた．

$$GL(k) = \frac{100 \times E_{va}(k)}{E_{opt}(k) - 1},$$

ここで，$E_{va}(k)$ および $E_{opt}(k)$ はそれぞれ，k エポック目の検証集合における損失，k エポック目までの検証集合における損失の最低値を表す．また，本論文において，エポック数は 30，バッチサイズは 16 とした．

Collision Attention Module は 5.1 節で説明したデータセットを拡張したデータセットを使用して事前学習を行った．この時，[Magassouba 21] に従って，モジュールの学習を行った．

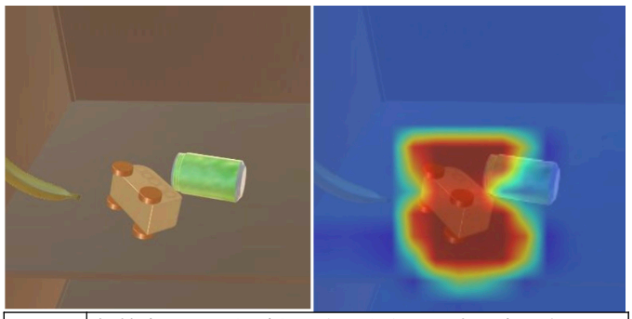
5.3 定量的結果

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．ベースライン手法として，RFCM [Kambara 22] および SAT [Xu 15] を以下の理由により選択した．RFCM は物体配置タスク時の衝突の関する future captioning において良好な結果が得られているためであり，SAT は代表的な画像キャプショニング手法であるため選択した．

表 1 より，提案手法は全ての評価尺度において，ベースライン手法を上回った．具体的には，主要評価尺度である CIDEr-D において，提案手法は 61.06 ポイントであった．一方，ベースライン手法である RFCM および SAT はそれぞれ 41.67 ポイント，46.64 ポイントであった．従って，提案手法がそれぞれ 19.39，14.42 ポイント上回った．BLUE-4，METEOR，および ROUGE-L においても同様に，提案手法は SAT を 3.13 ポイント，4.33 ポイント，2.39 ポイント上回った．また，RFCM を 1.99 ポイント，2.39 ポイント，および 2.97 ポイント上回った．全ての標準評価尺度において，提案手法及びベースラインの性能差は統計有意 ($p < 0.05$) であった．

5.4 定性的結果

定性的結果を図 3 に示す．図 3 の左図および右図はそれぞれ配置領域の RGB 画像および attention map を配置領域の RGB 画像に重畳した画像を表す．図 3 において，把持している物体は「ペットボトル」であり，衝突する物体は「おもちゃの木の車」である．SAT は衝突した物体を「ハンド」と誤って記述した．同様に，RFCM は衝突した物体を「哺乳瓶」と誤って記述した．一方で，提案手法は，対象物体および衝突した物体をそれぞれ「ペットボトル」「おもちゃの木の車」と適切に記述した．



Ref	把持中のペットボトルをおもちゃの木の車の上に配置しようとして、うまく置けずにペットボトルが倒れる
SAT	おもちゃの木の車とハンドが衝突する
RFCM	把持している空のペットボトルを哺乳瓶の上に配置しようとして倒れる
Ours	把持中のペットボトルをおもちゃの木の車の上に配置しようとして、うまく置けずに倒れる

図 3: 定性的結果

5.5 Ablation Study

Ablation 条件として，以下の 3 つの条件を用いた．

(a) w/o NNCM: NNCM の性能への寄与を調査するために，NNCM を取り除いた．
(b) w/o CAM: CAM の性能への寄与を調査するために，CAM を取り除いた．
(c) w/o CAIE: cross-attention 機構の有効性を調査するために，CAIE において Cross Attention 層の代わりに Self Attention 層を用いた．

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 CIDEr-D はそれぞれ 59.57, 56.94，54.68 および 61.06 であった．従って，CAIE における cross-atteniton 機構の影響が最も大きいことが示された．

6. おわりに

本研究では，動作実行前の情報をもとに将来の状況の説明を生成する future captioning タスクを扱った．特に，生活支援ロボットが障害物のある家具に物体を配置する際に発生する衝突に関する future captioning を扱った．本研究における貢献を以下に示す．本研究では，生活支援ロボットが障害物のある家具に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．本研究における貢献を以下に示す．

- 衝突に関する future captioning に NNLM [Khandelwal 19] を導入した NNFCM を提案した．
- 衝突に関する配置箇所の注目領域を強調するための Collision Attention Module を導入した．
- 配置領域および対象物体の関係性をモデル化する Cross Attentional Image Encoder を導入した．
- 提案手法は全ての評価指標において，ベースライン手法を上回った．

将来研究として，衝突予測と危険性に関する説明文生成の同時予測が考えられる．

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

[Fukui 19] Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H.: Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, in *CVPR*, pp. 10705–10714 (2019)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in *CVPR*, pp. 770–778 (2016)

[Hosseinzadeh 21] Hosseinzadeh, M. and Wang, Y.: Video Captioning of Future Frames, in *WACV*, pp. 980–989 (2021)

[Inamura 14] Inamura, T., Tan, C., et al.: Development of Robocup@ Home Simulation Towards Long-Term Large Scale HRI, in *Robot World Cup XVII 17*, pp. 672–680 (2014)

[Kambara 22] Kambara, M. and Sugiura, K.: Relational Future Captioning Model for Explaining Likely Collisions in Daily Tasks, in *ICIP*, pp. 2601–2605 (2022)

[Khandelwal 19] Khandelwal, U., Levy, O., Jurafsky, D., et al.: Generalization through Memorization: Nearest Neighbor Language Models, in *ICLR* (2019)

[Lei 20] Lei, J., Wang, L., Shen, Y., Yu, D., et al.: Mart: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning, in *ACL*, pp. 2603–2614 (2020)

[Li 20] Li, L., Chen, Y.-C., Cheng, Y., et al.: HERO: Hierarchical Encoder for Video+ Language Omni-Representation Pre-Training, in *EMNLP*, pp. 2046–2065 (2020)

[Magassouba 20] Magassouba, A., Sugiura, K., and Kawai, H.: Multimodal attention branch network for perspective-free sentence generation, in *CORL*, pp. 76–85 (2020)

[Magassouba 21] Magassouba, A., Sugiura, K., Nakayama, A., et al.: Predicting and Attending to Damaging Collisions for Placing Everyday Objects in Photo-Realistic Simulations, *AR*, Vol. 35, No. 12, pp. 787–799 (2021)

[Mahmud 21] Mahmud, T., Billah, M., et al.: Prediction and Description of Near-Future Activities in Video, *CVIU*, Vol. 210, p. 103230 (2021)

[Ogura 20] Ogura, T., Magassouba, A., Sugiura, K., et al.: Alleviating the burden of labeling: Sentence generation by attention branch encoder–decoder network, *RA-L*, Vol. 5, No. 4, pp. 5945–5952 (2020)

[Okada 19] Okada, H., Inamura, T., et al.: What Competitions were Conducted in the Service Categories of the World Robot Summit?, *AR*, Vol. 33, No. 17, pp. 900–910 (2019)

[Prechelt 98] Prechelt, L.: Automatic Early Stopping Using Cross Validation: Quantifying the Criteria, *Neural networks*, Vol. 11, No. 4, pp. 761–767 (1998)

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

[Ren 15] Ren, S., He, K., Girshick, R., and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *NeurIPS*, Vol. 28, (2015)

[Sun 19] Sun, C., Myers, A., Vondrick, C., et al.: Videobert: A Joint Model for Video and Language Representation Learning, in *ICCV*, pp. 7464–7473 (2019)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is All You Need, *NeurIPS*, Vol. 30, (2017)

[Wang 18] Wang, X., Chen, W., Wu, J., Wang, Y.-F., and Wang, W. Y.: Video Captioning via Hierarchical Reinforcement Learning, in *CVPR*, pp. 4213–4222 (2018)

[Xu 15] Xu, K., Ba, J., Kiros, R., et al.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, in *ICML*, pp. 2048–2057 (2015)

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

[Yi 19] Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., et al.: CLEVRER: Collision Events for Video Representation and Reasoning, in *ICLR* (2019)