実世界対話における フレーズグラウンディングモデルの評価と分析

植田 暢大 [1,2] 波部 英子 [2] 松井 陽子 [2] 湯口 彰重 [3,2] 河野 誠也 [2,4]
川西 康友 [2,4] 黒橋 禎夫 [1,2,5] 吉野 幸一郎 [2,4]
[1] 京都大学 大学院情報学研究科 [2] 理化学研究所 ガーディアンロボットプロジェクト
[3] 東京理科大学 先進工学部 [4] 奈良先端科学技術大学院大学 情報科学領域
[5] 国立情報学研究所
{ueda,kuro}@nlp.ist.i.kyoto-u.ac.jp akishige.yuguchi@rs.tus.ac.jp
{hideko.habe,yoko.matsui,seiya.kawano,yasutomo.kawanishi,koichiro.yoshino}@riken.jp

概要

実世界の共同作業における対話では，物体に対する参照表現が多く出現する．本研究ではフレーズグラウンディングの枠組みで，実世界対話における既存モデルの物体参照表現に対するフレーズグラウンディング性能を評価・分析し，課題を明らかにする．また，分析に基づいた改善手法の検討を行う．

1 はじめに

実世界で人間と協働する対話システムの実現には，対話中の参照表現の実世界へのグラウンディングが不可欠である．例えば「そのコップに注いで」という発話では，「コップ」のテキスト上の意味を理解するだけでは不十分であり，実世界において参照しているコップの実体を知る必要がある．

フレーズが参照している実体を視覚情報，特に画像中の物体矩形の形で特定するタスクはフレーズグラウンディングとよばれる [1, 2]．このタスクではベンチマークとして Flickr30K Entities データセット [3] が広く用いられており，代表的な解析モデルである MDETR [1] や GLIP [4, 5] は 80%以上の精度[1)]を達成している．しかし，Flickr30K Entities は人間が恣意的に撮影した静止画とそのキャプションから構成されており，これらのモデルが実世界対話における参照表現を正しくグラウンディングできるかは明らかではない．対話におけるフレーズグラウンディングを扱ったデータセットとして SIMMC 2.0 [6] がある．SIMMC 2.0 は代名詞の使用といった画像キャプションに現れにくい対話特有の現象も扱っているが，含まれる画像は CG の静止画であり実世界特有の視点の移動や物体の操作を含まない．

この問題に対処するため，我々は実世界における対話を扱うデータセット（Japanese Conversation dataset for Real-world Reference Resolution; J-CRe3）を構築した [7]．本研究ではこのデータセットを使用し，既存のフレーズグラウンディングモデルの実世界対話における性能を評価し，課題を明らかにする．また，明らかとなった課題に対する改善手法を提案し，評価する．

2 J-CRe3

J-CRe3 は実世界における 2 者対話で 1 人称視点動画と音声を収録し，フレーズグラウンディングタグを含む種々の視覚的および言語的アノテーションを付与したデータセットである（図 1）．対話内容は，家庭内の人間とお手伝いロボットの対話が想定されている．視覚的アノテーションに関しては，動画から抽出された 1 秒ごとのフレームそれぞれに，物体矩形およびその物体のクラス名，インスタンス ID が付与されている（図 1 左）．言語的アノテーションに関しては，日本語の発話書き起こしテキスト（同図中央）中のフレーズに対して述語項構造，共参照関係，橋渡し照応関係（同図右）が付与されている．また，フレーズの視覚的表現へのグラウンディングのためにテキスト中のそれぞれのフレーズに対して対応する物体矩形が付与されている．フレーズと物体矩形の関係には，「ここ」と「glass_3」のような直接の参照関係だけでなく，「注いで」と「glass_3」のような述語項構造におけるニ格に対応する間接的な関係も含まれる．本データセットの統

1） Recall@1

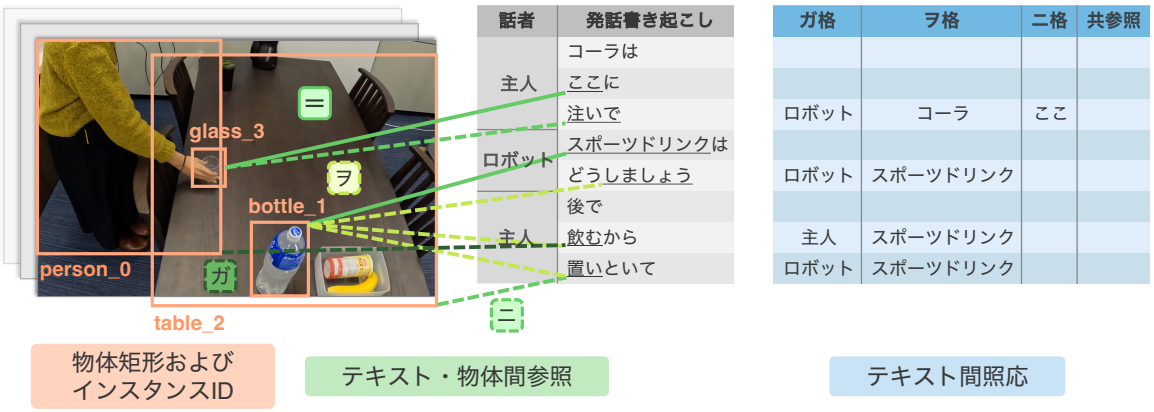

図 1　J-CRe3 の例

表 1　J-CRe3 の統計値

	学習	開発	テスト	合計
対話数	72	6	9	87
発話数	1594	103	230	1927
文数	1975	135	279	2389
形態素数	12683	889	1721	15293
タグ付きフレーム数	7947	671	1360	9978
物体矩形数	50072	3888	9184	63144
物体インスタンス数	1204	89	187	1480
直接の参照関係数	1196	76	160	1432

計値を表 1 に示す．

3　フレーズグラウンディングモデルの評価と改善

本章では既存のフレーズグラウンディングモデルにおける参照解析の結果と分析を示す．さらに，分析に基づいて共参照および複数物体追跡を考慮したフレーズグラウンディング手法を提案する．

3.1　タスク設定

J-CRe3 では，フレーズと物体矩形間に様々な種類の関係が付与されているが，本研究では最も重要である直接的な参照関係に注目する．また，システムへの入力は発話の書き起こしテキストとあらかじめ動画から 1 秒ごとに切り出した画像系列とする．画像とテキストが与えられたときテキスト中のフレーズに対応する画像中の物体矩形を推定するタスクはフレーズグラウンディングとよばれる [1, 2]．このタスクでは一般に，フレーズのグラウンディング対象は 1 枚の画像である．本研究では動画フレームが複数存在するため，フレーズが含まれる発話の開始時刻から次の発話の開始時刻までの間のフレームをそれぞれグラウンディング対象とする．

以降で述べるいずれの手法においても，リアルタイム対話での活用のためオンラインでの解析を想定する．すなわち，ある発話とフレームの解析結果を得る際に，それ以降の発話や画像フレームに関する結果は使用しない．

3.2　ベースライン手法

ベースラインとして既存のフレーズグラウンディングモデルの一つである GLIP [4] を用いる．GLIP は物体検出をフレーズグラウンディングと同一の枠組みで扱うことで事前学習において大量の画像テキストペアの使用を可能にしたモデルである．GLIP を用いることで，広範囲にわたる物体を検出することができると期待される．

GLIP は事前学習済みモデル[2)]が公開されているもののいずれも英語テキストで訓練されており，直接利用できない．そこで，J-CRe3 を含む日本語のフレーズグラウンディングデータセット等でこのモデルを fine-tuning する．

3.3　共参照および複数物体追跡の考慮

4.2 節にて詳しく述べるが，ベースライン手法では，曖昧な表現や小さな物体に対する精度が低いことがわかった．そこで共参照解析と複数物体追跡を組み合わせることでフレーズグラウンディングの精度向上を試みる．

共参照解析は実世界において同一の実体を指すテキスト中のフレーズを特定するタスクである．例えば，「テーブルの上にコップがあるからそこに注いで。」という文では「コップ」と「そこ」が同一の実体を指す．共参照解析を利用することで「コップ」に対するフレーズグラウンディング結果を「そこ」

2)　https://huggingface.co/GLIPModel/GLIP/tree/main

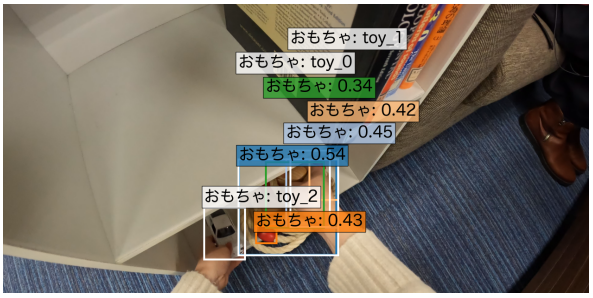

図 2　「あそうそう、おもちゃを置いた後で二段目の本を取ってきてほしいの。」という発話中の「おもちゃ」に対するベースライン手法の解析結果．白色の物体矩形は正解を表し，その他の物体矩形が上位 5 件のシステム出力を表す．システム出力には予測確率が記載されている．

表 2　物体矩形の画面占有率と Recall の関係．テストセットと開発セットを合わせた 15 対話で評価した結果を示す．括弧内の数値は正解した物体矩形数を表す．Recall@1 のみ分母に全正解数を示す．

画面占有率	Recall@1	Recall@5	Recall@10
0.0 – 0.005	0.276 (82/297)	0.582 (173)	0.737 (219)
0.005 – 0.05	0.488 (147/301)	0.744 (224)	0.801 (241)
0.05 – 1.0	0.683 (99/145)	0.855 (124)	0.869 (126)

に対する結果と同一視でき，曖昧な表現である「そこ」に対する精度の向上が期待できる．

物体追跡は動画中の物体矩形の系列から同一の物体を特定するタスクである [8]．動画フレーム中の複数の物体を対象とする場合，複数物体追跡とよばれる．1 人称視点動画では同一の物体でもフレームによってその大きさなど見え方は様々である．複数物体追跡によって，一部のフレームで小さく写っている物体が，他のフレームで大きく写っている場合の情報を利用できることを期待する．具体的には，あるフレーズについてベースライン手法によって参照先物体矩形が得られたとき，別フレームに存在する得られた物体と同一の物体を複数物体追跡によって特定する．対象フレーズと特定された物体矩形集合との参照確率の中から最も高い確率を現在のフレームにおける参照確率とする．

4　実験

4.1　実験設定

ベースライン手法として GLIP-T モデル[3] を用いた．英語テキストで訓練されたこのモデルの言語エンコーダを多言語で訓練されたモデルに置き換え，fine-tuning した．fine-tuning には Visual Genome [9], GQA [10], Flickr30k Entities JP [11], J-CRe3 [7] を使用した．

3)　https://huggingface.co/GLIPModel/GLIP/resolve/main/glip_tiny_model_o365_goldg.pth

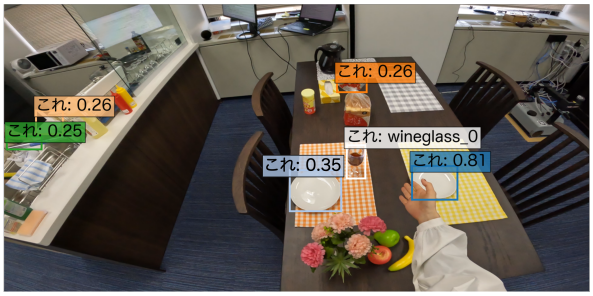

図 3　「これはまだ飲みますか？洗っちゃいますか？」という発話中の「これ」に対するベースライン手法の解析結果

表 3　参照表現の品詞と Recall の関係．テストセットと開発セットを合わせた 15 対話で評価し，頻度上位 5 種類の品詞を示す．

品詞	Recall@1	Recall@5	Recall@10
普通名詞	0.450 (259/575)	0.699 (402)	0.795 (457)
指示詞	0.300 (24/80)	0.675 (54)	0.738 (59)
接頭辞	0.649 (24/37)	0.865 (32)	0.919 (34)
サ変名詞	0.880 (22/25)	1.000 (25)	1.000 (25)
形式名詞	0.429 (9/21)	0.571 (12)	0.667 (14)

共参照解析システムについては Ueda ら [12] に倣い，事前学習済み日本語 DeBERTa モデル[4] を京都大学テキストコーパス [13] 等[5] で fine-tuning したものを使用した．J-CRe3 のテストセットにおける F 値は 0.75 だった．

複数物体追跡システムについては，まず一般物体認識器 Detic [14][6] を動画各フレームに適用し，その結果を複数物体追跡器 BoTSORT [15] に入力した．J-CRe3 のテストセットにおける IDR [16] は 0.20 だった．その他設定の詳細は付録 A に示す．

評価指標は Recall@k を使用した．GLIP は，それぞれのフレーズについて複数の物体矩形とその予測確率を出力する．Recall@k は正解の物体矩形のうち，出力された予測確率上位 k 件の物体矩形に含まれるものの割合である[7]．

4)　https://huggingface.co/ku-nlp/deberta-v2-large-japanese
5)　https://github.com/ku-nlp/KyotoCorpus, https://github.com/ku-nlp/KWDLC, https://github.com/ku-nlp/AnnotatedFKCCorpus, https://github.com/ku-nlp/WikipediaAnnotatedCorpus
6)　https://github.com/facebookresearch/Detic/blob/main/docs/MODEL_ZOO.md
7)　先行研究 [1] に倣い，予測された物体矩形が正解の物体矩形と 0.5 以上の Intersection-over-Union (IoU) を持つ場合に両者が一致すると判断した．

表 4　共参照（Coref.）および複数物体追跡（MOT）を考慮した場合のフレーズグラウンディングの結果．oracle は解析器の出力ではなく正解タグを使用した結果を表す．

手法	テストセット（9 対話）			開発セット（6 対話）		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
ベースライン	0.477 (194/407)	0.700 (285)	0.764 (311)	0.416 (144/346)	0.711 (246)	0.824 (285)
+ Coref. (oracle)	0.474 (193/407)	0.700 (285)	0.764 (311)	0.416 (144/346)	0.697 (241)	0.783 (271)
+ MOT (oracle)	0.477 (194/407)	0.742 (302)	0.813 (331)	0.457 (158/346)	0.754 (261)	0.876 (303)
+ Coref. (oracle) + MOT (oracle)	0.479 (195/407)	0.759 (309)	0.826 (336)	0.465 (161/346)	0.832 (288)	0.954 (330)
+ Coref.	0.474 (193/407)	0.700 (285)	0.764 (311)	0.416 (144/346)	0.702 (243)	0.812 (281)
+ MOT	0.474 (193/407)	0.700 (285)	0.764 (311)	0.405 (140/346)	0.711 (246)	0.824 (285)
+ Coref. + MOT	0.472 (192/407)	0.700 (285)	0.764 (311)	0.405 (140/346)	0.702 (243)	0.812 (281)

4.2　ベースライン手法の結果および分析

ベースライン手法の J-CRe3 のテストセットにおける結果は，Recall@{1, 5, 10} がそれぞれ 0.477, 0.700, 0.764 だった．同モデルの Flickr30k Entities JP のテストセットにおける結果はそれぞれ 0.695, 0.881, 0.915 であり，J-CRe3 におけるフレーズグラウンディングの難しさがわかる．

解析事例の一つを図 2 に示す．このフレームには 3 つのおもちゃがアノテーションされているが，システムは最も左に写っているミニカーの検出に失敗している．図のように 1 人称視点動画では物体の一部のみが小さく写っている場合があり，解析失敗の一因になっている．物体のサイズごとに Recall を評価した表 2 からも，小さく写っている物体に対する精度が低いことがわかる．

図 4 に他の対話における解析事例を示す．図ではワイングラスが正解の物体となるが，物体名が言及されておらず皿に高い確率が与えられている．表 3 は参照表現の品詞ごとの Recall を示す．具体的な物体名を含む普通名詞に対する精度に比べ，指示詞や形式名詞など曖昧な表現に対する精度が低い．これは GLIP の訓練に画像キャプションを元にしたデータセットが多く使用されているためと考えられる．

4.3　共参照および複数物体追跡を考慮した手法の結果および分析

ベースライン手法に対し，共参照解析と複数物体追跡および両者を組み合わせた場合の結果を表 4 に示す．共参照の考慮について，ベースライン手法に対するスコアの向上は見られなかった．これは，解析単位である 1 発話内で，複数のフレーズが共参照関係にある場合が稀であるためと考えられる．

複数物体追跡を考慮した場合については，oracle の設定でのみスコアが向上した．図 4 は MOT (oracle) の設定で解析に成功した例である．正解である包丁について，ベースラインモデルの確率は 0.45 だったが，包丁がより大きく写っている前のフレームの結果から 0.71 の確率を割り当てることができた．しかし，一般物体認識器を使用した場合は包丁の検出に失敗しており，確率は 0.45 のままだった．今回利用した一般物体認識器の Recall は 0.76，複数物体追跡器の IDR は 0.20 程度であり，これら解析器の性能向上が必要である．

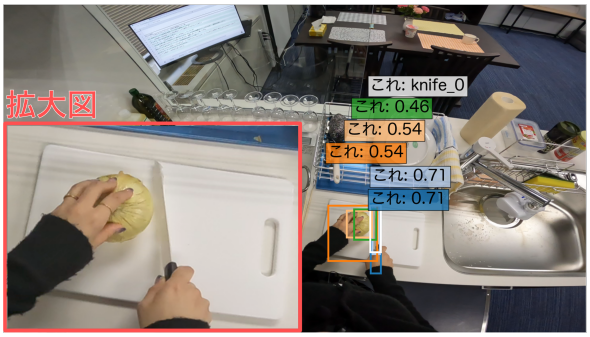

図 4　「これは良く切れますね。綺麗に研いでいて偉いですね。」という発話中の「これ」に対する MOT (oracle) の解析結果

5　おわりに

本研究では，実世界において人間と対話しつつ協働するシステムの実現を目指し，実世界対話における既存のフレーズグラウンディングモデルの性能を分析した．さらに分析に基づき，共参照および複数物体追跡を考慮したフレーズグラウンディング手法を提案した．いずれもスコアの向上は見られなかったが，正解タグを用いた設定では効果が見られたため複数物体追跡器などの解析モジュールの性能向上が今後の課題である．また，今回はフレーズと物体の直接の参照関係のみを扱ったが，述語と項のような間接的な参照関係の解析にも取り組みたい．

謝辞

本研究は，京都大学科学技術イノベーション創出フェローシップ事業の助成を受けたものである．本研究の一部は JSPS 科研費 22H03654 の支援を受けたものである．

参考文献

[1] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 1780–1790, 2021.

[2] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, **Computer Vision – ECCV 2020**, pp. 752–768, Cham, 2020. Springer International Publishing.

[3] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. **International Journal of Computer Vision (IJCV)**, Vol. 123, No. 1, pp. 74–93, 2017.

[4] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 10965–10975, June 2022.

[5] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 36067–36080. Curran Associates, Inc., 2022.

[6] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 4903–4912, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] 植田暢大, 波部英子, 湯口彰重, 河野誠也, 川西康友, 黒橋禎夫, 吉野幸一郎. 実世界における総合的参照解析を目的としたマルチモーダル対話データセットの構築. 言語処理学会 第 29 回年次大会, pp. 2990–2995, 沖縄, 2023.3.

[8] Mk Bashar, Samia Islam, Kashifa Kawaakib Hussain, Md. Bakhtiar Hasan, A. B. M. Ashikur Rahman, and Md. Hasanul Kabir. Multiple object tracking in recent times: A literature review, 2022.

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. **International Journal of Computer Vision (IJCV)**, Vol. 123, No. 1, pp. 32–73, 2017.

[10] Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. **Conference on Computer Vision and Pattern Recognition (CVPR)**, 2019.

[11] Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. A visually-grounded parallel corpus with phrase-to-region linking. In **Proceedings of The 12th Language Resources and Evaluation Conference**, pp. 4204–4210, Marseille, France, May 2020. European Language Resources Association.

[12] Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. BERT-based cohesion analysis of Japanese texts. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 1323–1333, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[13] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. Construction of a Japanese relevance-tagged corpus. In **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)**. European Language Resources Association (ELRA), 2002.

[14] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, **Computer Vision – ECCV 2022**, pp. 350–368, Cham, 2022. Springer Nature Switzerland.

[15] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Botsort: Robust associations multi-pedestrian tracking. **arXiv preprint arXiv:2206.14651**, 2022.

[16] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In **ECCV Workshops**, 2016.

[17] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In **2019 IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 8429–8438, 2019.

表 5 GLIP の fine-tuning 1 段階目におけるハイパーパラメータ

パラメータ名	値
オプティマイザ	AdamW
エポック数	2.5
学習率	0.0001
重み減衰	0.05
ウォームアップステップ数	2000
バッチサイズ	24

表 6 GLIP の fine-tuning 2 段階目におけるハイパーパラメータ

パラメータ名	値
オプティマイザ	AdamW
エポック数	1.0
学習率（言語エンコーダ以外）	0.0001
学習率（言語エンコーダ）	0.00001
重み減衰	0.05
ウォームアップステップ数	1000
バッチサイズ	24

A 実験設定の詳細

A.1 GLIP の fine-tuning

ベースライン手法において，我々は GLIP [4] を fine-tuning して使用した．GLIP のアーキテクチャは大きく言語エンコーダ，画像エンコーダ，そしてこれらエンコーダから得られた表現を統合するモジュールから構成される．我々は事前学習済みモデルの知識をできる限り活用できるように，これらのモジュールを 2 段階に分けて fine-tuning した．

1 段階目は，モデルの日本語への適応を目的とする．英語テキストのみで訓練された事前学習済み GLIP モデルの言語エンコーダ[8]を，多言語で訓練されたモデルである mDeBERTaV3 base[9]に置き換え，Visual Genome [9]，GQA [10]，Flickr30k Entities JP [11] を混合して fine-tuning する．このとき，英語テキストの過学習を防ぐため，言語エンコーダのパラメータは固定した．また，画像エンコーダの性能は言語非依存のため，画像エンコーダのパラメータも固定した．ハイパーパラメータを表 5 に示す．

2 段階目は，モデルの 1 人称視点画像および対話形式テキストへの適応を目的とする．予備実験にて J-CRe3 のみを使用して fine-tuning したところ GLIP の物体検出の性能が下がってしまったため，J-CRe3 と Flickr30k Entities JP を混合して使用した．この際いずれのモデルパラメータも固定しなかった．ハイパーパラメータを表 6 に示す．

表 7 共参照解析モデルのハイパーパラメータ

パラメータ名	値
オプティマイザ	AdamW
エポック数	16
学習率	0.00005
重み減衰	0.01
ウォームアップステップ数	1000
バッチサイズ	16

A.2 共参照解析システム

共参照解析システムについては，Ueda ら [12] の手法を使用した．本研究では，事前学習済み日本語 DeBERTaV2 large モデル[10]を以下のデータセットで fine-tuning した．

- 京都大学テキストコーパス[11]
- 京都大学ウェブ文書リードコーパス[12]
- Annotated FKC Corpus[13]
- Wikipedia Annotated Corpus[14]

表 7 にハイパーパラメータを示す．

A.3 複数物体追跡システム

複数物体追跡システムについては，一般物体認識器 Detic [14] と，複数物体追跡器 BoTSORT [15] を組み合わせて使用した．Detic は公開されているモデルの中で Objects365 データセット [17] における mAP が最も高い Detic_C2_SwinB_896_4x_IN-21K+COCO[15]を使用した．BoTSORT については，Re-identification モデルとして Torchreid[16]で公開されている osnet_ain_x1_0[17]を使用した．

8） https://huggingface.co/bert-base-uncased
9） https://huggingface.co/microsoft/mdeberta-v3-base
10） https://huggingface.co/ku-nlp/deberta-v2-large-japanese
11） https://github.com/ku-nlp/KyotoCorpus
12） https://github.com/ku-nlp/KWDLC
13） https://github.com/ku-nlp/AnnotatedFKCCorpus
14） https://github.com/ku-nlp/WikipediaAnnotatedCorpus
15） https://github.com/facebookresearch/Detic/blob/main/docs/MODEL_ZOO.md#cross-dataset-evaluation
16） https://github.com/KaiyangZhou/deep-person-reid
17） https://kaiyangzhou.github.io/deep-person-reid/MODEL_ZOO#msmt17-combineall-true-market1501-dukemtmc-reid