一人称視点に基づく テキスト駆動型アフォーダンス及び軌跡の学習

吉田智哉 [1,2]　栗田修平 [2]　西村太一 [3]　森信介 [1]
[1] 京都大学　[2] 理化学研究所　[3] LINE ヤフー株式会社
yoshida.tomoya.25h@st.kyoto-u.ac.jp　shuhei.kurita@riken.jp
tainishi@lycorp.co.jp　forest@i.kyoto-u.ac.jp

概要

視覚的アフォーダンス学習は，入力画像中のインタラクションするべき領域を局限する課題である．先行研究によって，この課題がロボットの把持課題に有効であることが示されている一方で，多様な行動に対するアフォーダンスの学習が次の問題となっている．本研究では，この問題の解決のためにテキスト駆動型視覚的アフォーダンス及び軌跡の学習を提案する．この課題は，テキストで表現された様々な行動に対して，それを達成するために，画像中のどの領域に，どのようにインタラクションするべきかを学習することを目的とした課題である．この課題を解決するために，一人称視点動画データセットから自動で擬似教師データセットを構築し，それを利用したモデルを提案する．実験の結果，提案手法が先行研究のモデルと比べて優れた性能を示したことを報告する．

1　はじめに

人間は物体のどの領域に，どのようにインタラクションするべきかを理解しており，その知識を利用することで安全に，かつ効率的に作業を行うことができる．例えば，ドアを開けようとする際には，ノブに触れそれを回転させることにより達成する．こうした人間と物体間のインタラクションは短いテキストとして表すことが可能であり，それらで条件づけられた行動のために，視覚中のどの領域に，どのようにインタラクションするべきかを推論可能なモデルは，ロボット工学やヒューマンコンピュータインタラクションの領域において有用なツールになると思われる．

近年，Gibson [1] が提案した，環境が動物に提供する行動の可能性を表すアフォーダンスの概念をロボットに適用する試みが盛んに行われており，これらの試みは視覚的アフォーダンス学習課題として定式化されている [2, 3, 4, 5, 6, 7]．視覚的アフォーダンス学習とは，図 1（左）に示されるように，入力画像（もしくは動画）中の物体に対して，インタラクションすべき領域を局限する課題である．先行研究 [2, 3] では，主に把持に焦点が当てられており，いくつかの研究によって，それらがロボットに応用可能であることが示されている [8, 9]．先行研究により，把持などの手と物体間のインタラクションにおいてアフォーダンスの学習が有効であることが示された一方で，手で把持された道具と物体間のインタラクション [5, 6] や目的に応じたインタラクションなどのより高次なアフォーダンス学習が次の問題となっている．

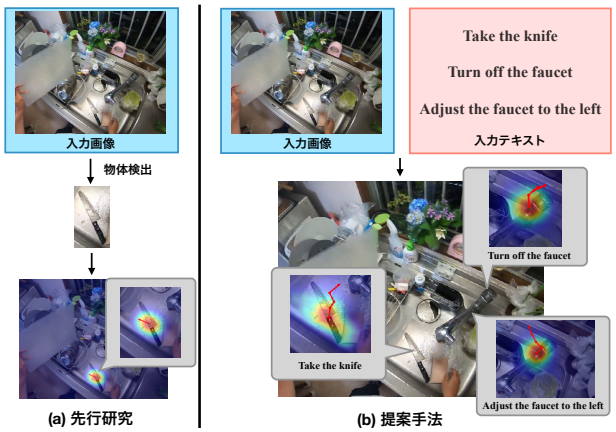

図 1　テキスト駆動型視覚的アフォーダンス学習．

本研究では，この問題を解決するためにテキスト駆動型視覚的アフォーダンス及び軌跡の学習を提案する．この課題は，図 1（右）に示されるように，テキストで表現された様々な行動に対して，それを達成するために，画像中のどの領域に，どのような方法でインタラクションするべきかを学習することを目的とした課題である．この課題の解決のために，

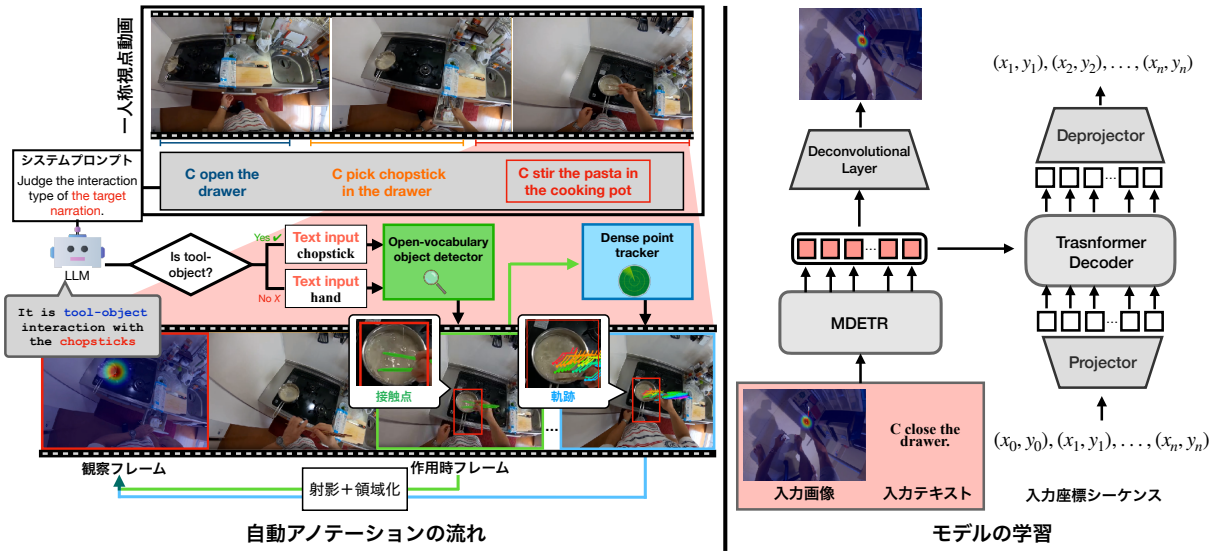

図 2　自動で擬似教師データを作成するための手法（左）及び，モデルの概要（右）．

大規模な一人称視点動画データセットから自動で擬似教師データを作成する手法及び，構築したデータセットを利用したモデルを提案する．実験の結果，提案手法が先行研究におけるモデルと比べて優れた性能であることが確認できた．

2　テキスト駆動型視覚的アフォーダンス及び軌跡の学習

テキスト駆動型視覚的アフォーダンス学習は，テキストで表現された様々な行動に対して，それを達成するために，画像中のどの領域に，どのような方法でインタラクションするかを学習することを目的とした課題であり，領域は確率分布，方法は軌跡として表現される．この課題の解決のためのデータを人手で作成するには，多くの時間と費用を必要とするため，本研究では自動で擬似教師データを作成する手法を提案する．手法の流れを図 2（左）に示す．

2.1　擬似教師データの作成

大規模一人称視点動画データセット [10, 11] には，短いクリップ内の撮影者の行動を表す短いテキスト及び，それらの行動の開始時刻 t_{obs} と終了時刻 t_{inter} が付与されている[1)]．各 t_{obs}, t_{inter} に対応する画像フレームを F_{obs}, F_{inter} と表記する．

インタラクションの分類．　人間と物体間のインタラクションの種類は大きく，“手と物体間のインタラクション（HOI）” あるいは “道具と物体間のインタラクション（TOI）” に分類可能である．HOI は，手が対象の物体と接触するインタラクションを表し，TOI は，把持された道具が対象の物体と接触するインタラクションを表す．各クリップがどちらに属すかの分類，さらに TOI の場合は利用されている道具を特定することを狙いとする．これを達成するために，あるクリップとその過去のクリップに付与されているテキスト列を入力として，大規模言語モデル，LLaMA2 [12] を利用する．適切な出力を得るために，システムプロンプトを与え，数例のサンプルを提示する．プロンプトの詳細は付録 A.1 に添付する．

インタラクション領域の射影．　短い動画クリップ内の隣接する 2 フレームの動きは基本的に小さいため，それらの関係はホモグラフィで表すことができる [13]．したがって，クリップ内の任意のフレーム同士の射影行列はその 2 フレーム間中のフレームらの射影行列を掛け合わせることによって得られる．SURF [14] アルゴリズムを用いて二つのフレーム間の特徴点マッチングを行い，4 つの特徴点ペアをサンプリングし，RANSAC [15] を適用することで F_{obs} と F_{inter} 間の射影行列を取得する．次に，任意のテキストを受け付ける物体検出モデル[2)]を利用し，F_{inter} フレーム中のインタラクションしている物体（道具もしくは手）のセグメンテーションマスクを取得する．得られたセグメンテーションマスクは，計算した F_{obs} と F_{inter} 間の射影行列により，F_{obs} へと射影される．

軌跡の射影．　点追跡モデルである CoTracker [16] を利用し，F_{inter} 中のインタラクション領域に対し

1）　Ego4D [10] は下流タスクである Short-term action acticipation 課題のためのデータセットを利用した．

2）　https://github.com/luca-medeiros/lang-segment-anything

	テキストの有無	領域			軌跡		
		Sim ↑	CC ↑	AUC-J ↑	D-Sim ↑	ADE ↓	DTW ↓
Hotspots [2]		6.4	6.5	56.5	-	-	-
VRB [8]		9.8	10.1	59.6	-	-	-
Lang-SAM	✓	9.4	11.2	57.1	-	-	-
提案手法		17.4	27.5	89.2	12.8	17.6	21.6
提案手法	✓	**19.3**	**29.4**	**90.4**	**34.9**	**11.1**	**16.3**

表 1 擬似テストセットにおける各ベースライン及び提案手法の結果．テキストの有無は，モデルが予測のためにテキスト情報を利用するかどうかを表す．

て 0.5 秒間の領域追跡を行う．各フレームにおける追跡結果の座標平均を計算し，それらの点を，領域の射影と同様の手法によって，F_{obs} に射影する．操作の詳細は付録 A.2 に添付する．

2.2 モデル

前節で構築したデータセットを用いて，チャンネル数 c, 幅 w, 高さ h からなる入力画像 $F_{obs} \in \mathbb{R}^{c \times w \times h}$ と入力テキストに対し，インタラクションするべき領域 $P \in \mathbb{R}^{w \times h}$ 及び，軌跡 $L = \{(x_0, y_0), ..., (x_n, y_n)\}$ を生成する．図 1（右）に示されるように，モデルは MDETR [17] における最終層の出力を利用し，P, L を生成する．MDETR は Transformer [18] を利用した物体検出器である DETR [19] をテキストを受け付けるように拡張したモデルであり，物体検出及び物体セグメンテーションが可能なモデルである．領域予測では，セグメンテーション課題 [20] のためにファインチューニングされた MDETR を利用し，目的関数 L_{reg} をバイナリ交差エントロピーとする．軌跡予測では，各入力座標は先行研究に倣い，線形層から成る座標エンコーダを介して高次の特徴量に変換される [21]．また，始点は領域 P の中心座標とし，その後の軌跡を学習する．各座標は MDETR から得られる特徴量を相互注意機構を介して，Transformer デコーダによって再帰的に生成される．目的関数 L_{traj} は二乗誤差とする．訓練時は，損失を $L = L_{reg} + \lambda L_{traj}$ として同時学習を実施する．また，λ はハイパーパラメータである．

3 実験

3.1 実験設定

データセット． 擬似教師データを作成するための一人称視点動画データセットとして Ego4D [10] を利用した．これにより 75,655 件の擬似データの作成を行い，それらを 69,655 件の教師データ，3,000 件の検証データ，3,000 のテストデータに分割した．

評価指標． 領域推定では，先行研究や顕著性マップ推定にて一般的に利用される評価指標である，Pearson's correlation coefficient (CC)，Similarity metrics (Sim) [22] 及び AUC-Judd (AUC-J) [23] を利用した [2, 13]．また，Kullback-Leibler divergence は分布の裾野に敏感であるであるため利用しない [24]．軌跡推定では，先行研究にて一般的に利用される評価指標である，Average displacement error (ADE) 及び Dynamic time warping (DTW) [25] を利用した．さらに，座標シーケンス間の偏角を評価する Direction similarity metrics (D-Sim) を利用した．各指標の詳細は付録 B に添付する．

ベースライン． 視覚的アフォーダンス学習課題に取り組んだ先行研究のモデルである，VRB [8], Hotspots [2] と比較を行う．どちらのモデルも言語を入力としてとらず，対象の物体のみが写っている画像に対して領域推定を行うモデルである．また，言語を介したモデルとの比較のために，Open-vocabrualy object detection において性能が最も優れている Grounding DINO [26] を利用したセグメンテーションモデルである Lang-SAM[3] と比較を行う．これらのモデルは領域予測におけるベースラインであり，軌跡予測におけるベースラインではないことを強調しておく．ベースラインの詳細は付録 C に添付する．

3.2 結果

表 1 に，擬似テストセットにおける提案手法とベースラインの結果を示す．ヒートマップ推定では，どの評価指標においても提案手法がベースラインよりも優れていることが確認できる．また，領域予測，軌跡予測のどちらにおいてもテキスト情報なしで学習した提案手法の性能が，ありのときよりも

3) https://github.com/luca-medeiros/lang-segment-anything

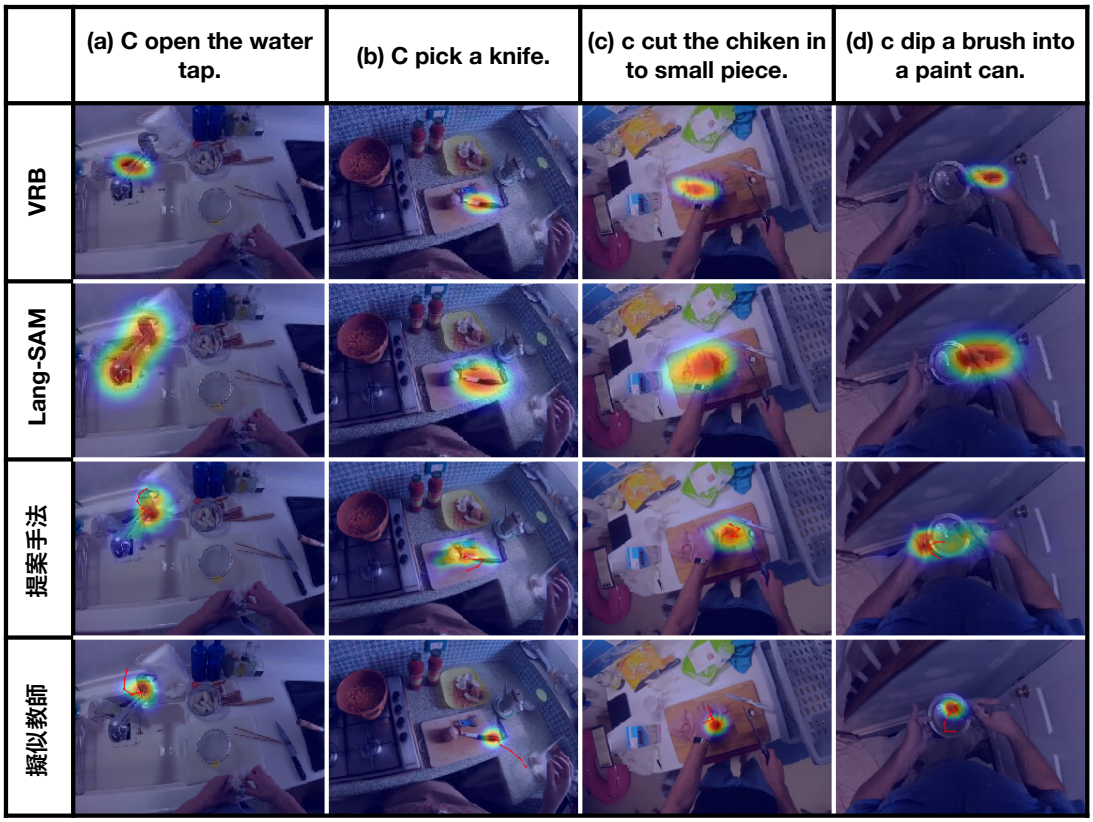

図 3　領域予測及び軌跡予測の例．(a)，(b) は手と物体間のインタラクションの例であり，(c)，(d) は道具と物体間のインタラクションの例を示す．赤い線は軌跡を表し，矢印は終点を表す．

低いことからテキスト中に表される行動を利用することが重要であることがわかる．セグメンテーションモデルである Lang-SAM は，評価指標 CC において他のベースラインよりも優れている一方で，AUC-J においては劣る結果となっている．これは，物体中の局所的な領域でなく物体全体に対して領域を予測しているため，偽陰性に敏感な AUC-J において低くなっていることが考えられる．

3.3　生成例

前節において性能が優れていたベースラインである VRB と Lang-SAM 及び提案手法の生成例を図 3 に示す．また，ベースラインは領域予測のためのモデルであるため軌跡はない．手と物体間のインタラクションである (a)，(b) では，ベースライン，提案手法ともに正しい領域を示していることが確認できる．一方で，道具と物体間のインタラクションである (c)，(d) ではベースラインが道具や物体全体など予測に失敗しているのに対して，提案手法では道具がインタラクションするべき領域を予測できている．しかし，予測された領域は広く，改善の余地がある．また，Lang-SAM の領域はどの例においても物体全体にかかっており，前節で考察した AUC-J が低くなっている原因が反映されている．

軌跡予測については，正解データの軌跡と比較して方向のずれは大きくないものの，始点のずれが目立つことが確認できる．始点のずれは領域予測の結果によるものであるため，改善のためには領域予測の精度を向上させる必要がある．

4　おわりに

本研究では，新規の課題であるテキスト駆動型視覚的アフォーダンス及び軌跡の学習を提案した．この課題の解決のために，自動で擬似教師データを作成する手法を提案し擬似データセットを構築した．また，構築したデータセットを利用し，領域と軌跡を予測するためのモデルを提案した．実験の結果，先行研究における手法と比べて提案手法が優れた性能であることが確認できた．今回の評価は擬似テストデータで行ったが，これはモデルが真に有用であることを示すためには不十分である．今後の課題として，人手で作成したデータで評価を行うとともに，モデルの改良を行っていきたい．

謝辞

本研究は JST さきがけ JPMJPR20C2 および JSPS 科研費 JP22K17983，Microsoft Accelerate Foundation Models Research の支援を受けたものです.

参考文献

[1] James J Gibson. **The ecological approach to visual perception: classic edition**. Psychology press, 2014.

[2] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In **ICCV**, 2019.

[3] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In **CVPR**, 2023.

[4] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In **CVPR**, 2018.

[5] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In **ICRA**, 2015.

[6] Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, and Yoichi Sato. Fine-grained affordance annotation for egocentric hand-object interaction videos. In **WACV**, 2023.

[7] Lorenzo Mur-Labadia, Jose J. Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In **ICCV**, 2023.

[8] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In **CVPR**, 2023.

[9] Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3d point clouds. In **IROS**, 2023.

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In **CVPR**, 2022.

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In **ECCV**, 2018.

[12] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[13] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In **CVPR**, 2022.

[14] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). **Computer vision and image understanding**, Vol. 110, No. 3, 2008.

[15] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. **Commun. ACM**, Vol. 24, No. 6, 1981.

[16] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. **arXiv:2307.07635**, 2023.

[17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In **ICCV**, 2021.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **NeurIPS**, Vol. 30, 2017.

[19] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In **ECCV**, 2020.

[20] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-based Image Segmentation in the Wild. In **CVPR**, 2020.

[21] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In **ICPR**, 2021.

[22] Michael J Swain and Dana H Ballard. Color indexing. **International journal of computer vision**, Vol. 7, No. 1, 1991.

[23] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In **ICCV**, 2009.

[24] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? **PAMI**, Vol. 41, No. 3, 2019.

[25] Dynamic time warping. **Information Retrieval for Music and Motion**, 2007.

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. **arXiv preprint arXiv:2303.05499**, 2023.

[27] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In **CVPR**, 2020.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In **ICLR**, 2013.

A　擬似教師データ作成の詳細

A.1　インタラクションの分類

大規模言語モデルはオープンソースであり，高い性能を持つ LLaMA2-60B [12] を利用した．また，システムプロンプトは，“<<SYS>> Please analyze the provided sequence of actions, focusing primarily on the last action to determine the tool for the interaction. Context from previous actions can be considered if necessary to fully understand the last action. Sequence of actions: [’action1’, ’action2’, ’action3’, ...] <</SYS>>” とし，3 つの例を生成前に提示した．

A.2　領域・軌跡の射影

F_{obs} から F_{inter} までの射影行列を取得するために，hand-object detector [27] をクリップ中の各フレームに対して適用した．これにより，各フレーム中の手の矩形領域と，その手が触れている物体の矩形領域が得られる．これらは射影行列の計算時にマスクする領域として利用され，これを利用することで，射影結果のずれを軽減させる．射影後のセグメンテーションマスクを領域として表現するために，混合ガウスモデル（GMM）を得られたセグメンテーションマスクに適用し，GMM からサンプリングした点にガウシアンぼかしを適用することで，領域として取得した．これらの作業により，射影結果のずれを緩和させることが期待できる．

B　評価指標の詳細

領域推定では：

- Pearson’s Correlation Coefficient (CC): 2 つの分布間の相関を測定したもの．
- Similarity Metrics (Sim): 2 つの確率分布間の類似度を測る計算したものであり，2 つの確率分布間の各座標における最小値を取り，それらを合計で計算される．
- AUC-Judd (AUC-J): Area of Under Curve (AUC) 指標の派生であり， 異なる閾値における 2 つの分布の一致率を計算したもの．

を利用した．

軌跡推定では：

- Average Displacement Error (ADE): 2 つの座標シーケンス中の各座標における二乗誤差の平均．
- Dynamic Time Warping (DTW): 2 つの座標シーケンス間の距離を，総当たりで計算し，最小となる組み合わせの絶対誤差を計算したもの．
- Direction Similarity Metrics (D-Sim): 座標シーケンスを各座標間の偏角シーケンスに変換し，変換後の 2 つのシーケンスのコサイン類似度を計算したもの．

を利用した．

C　ベースラインの詳細

実験ではベースラインとして，VRB [8], Hotspots [2], Lang-SAM[4] と比較を行った．VRB と Hotspots は，対象の物体のみが写っている画像に対して領域予測をするモデルであるため，入力テキスト中の物体名を spaCy[5] を用いて抽出し，その結果を Lang-SAM に入力することにより，画像中の対象の物体の矩形領域を抽出し，対象の物体のみが写る画像を取得した．また，Hotspots は，事前に定義された数種類の行動ラベルを指定することで，その行動に対応する領域を出力するモデルであるため，入力テキスト中の動詞を抽出後，Word2Vec [28] を用いて，Hotspots の持つ最も類似度が高いラベルに置き換えることで推論を行った．

4） https://github.com/luca-medeiros/lang-segment-anything
5） https://spacy.io