視覚言語モデルにおける空間ダイクシス表現の運用能力の評価

渡部魁人 [1] 山本泰成 [1,2] 土井智暉 [1,2] 谷中瞳 [1,2,3]
[1] 東京大学 [2] 理化学研究所 [3] 東北大学
{nglhdf,yamamo96,doi-tomoki701,hyanaka}@is.s.u-tokyo.ac.jp

概要

本研究では視覚言語モデルの空間推論能力の評価の観点として，コンテクストを参照しなければ指し示すものが決定できない空間表現である空間ダイクシス表現（例:「この」）に着目する．視覚言語モデルがさまざまな言語において人間の言語使用に近い形で空間ダイクシス表現を運用できているかどうかを測る多言語のベンチマークを作成し，空間ダイクシス表現の使い分け方について視覚言語モデルと人間のパフォーマンスの比較を行う．分析の結果，視覚言語モデルは人間が使用する一部の指示語をほとんど使用しない場合があるなど，人間の言語使用とは異なった形で空間ダイクシス表現を使用していることが確認された．

1 はじめに

視覚言語モデルに期待される能力の一つに，与えられたテキストと画像に基づく空間推論の能力がある．言語における空間表現の例として，空間参照枠[1)]が関わる表現や動きの表現などがある [1]．その中で本研究では**空間ダイクシス表現**に着目する．

ダイクシスとは，時間や場所に関する表現が場面などのコンテクストに依存して使われる用法のことを指す [2]．空間ダイクシス表現はダイクシス表現の中でも，「ここ」「あの」といった，発話される空間に依存する表現のことを指す．そのため，空間ダイクシス表現を適切に扱うには，空間に関するコンテクストを適切に認識しつつ処理する必要がある．

また，多言語視覚言語モデルがさまざまな言語において適切に空間ダイクシス表現の選択を行うためには，それぞれの言語によって異なる空間ダイクシス表現の使い分けを理解する必要がある．例えば，英語には近称の “this” と遠称の “that” という 2 種類の指示語の使い分けがあるのに対して，日本語では近称の「この」，中称の「その」，遠称の「あの」という 3 種類の指示語の使い分けがあるというように，言語によって表現の種類の数が異なる．さらに，Coventry ら [3] は，同じ場面に対して使用する指示語が必ずしも一意には定まらない場合があることを報告している．そのため，人間の言語使用に近い形で視覚言語モデルが指示語を使用するためには，言語間での表現の違いを理解し，かつその使い分けの分布を適切に再現できることが望ましい．

以上の点を踏まえ，本研究では英語や日本語といった様々な言語での空間ダイクシス表現の運用能力を測るベンチマークを作成し，視覚言語モデルの評価を行った．そして，視覚言語モデルの結果と [3] において行われた人間の空間ダイクシス表現の使い分け方に関する実験の結果を比較することでモデルの性能を分析した．本研究で構築したベンチマークは研究利用可能な形で公開予定である．実験の結果，一部のモデルがある場面に対して人間が使用する指示語をほとんど使用しない傾向や，物体との距離によって指示語を使い分けることをしない傾向といった，人間の言語使用と異なった言語使用を行うことが確認された．これにより，視覚言語モデルは人間とは異なった形で空間ダイクシス表現を運用していることが示唆された．

2 関連研究

視覚言語モデルの空間理解能力の評価は英語での評価を中心に数多く行われてきた [4, 5]．特に，[5] では Blender を利用して作成した画像を利用して，視覚言語モデルが論理的に同等な内容のプロンプトに対して一貫した空間推論を行えていないことを示した．

多言語で視覚言語モデルの空間理解能力を分析する先行研究として，[6] では VLM の空間参照枠を扱う性能を測る COMFORT という評価プロトコルを提案し，実際に多言語で評価を行なっている．その研究では，VLM は指定された空間参照枠を使用す

1) 空間参照枠とは，ある参照物から見た別の物体の相対的な位置を表現する際に用いられる空間表現の枠組み．

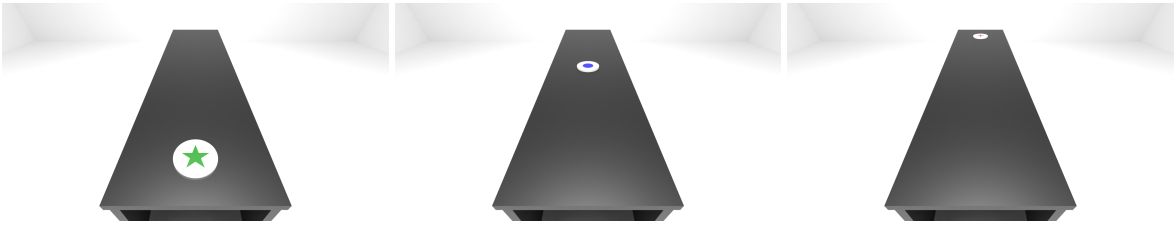

図 1　ベンチマークに含まれる画像データの例．指示対象との距離は左から順に 0.25m，1.50m，2.75m である．

る能力を欠いており，また，言語に固有な空間参照枠の使われ方よりもむしろ英語での使われ方に偏る傾向があると分析された．

言語モデルの参照表現理解能力を評価する研究としては，[7] が PUB という大規模言語モデルの語用論的理解能力を測るベンチマークのタスクの一つとして空間に限らないダイクシス表現に関するタスクを構成し，英語のダイクシス表現を大規模言語モデルが正確に理解できるかどうかが分析されている．これらの研究に対し，本研究では多言語での性能評価という観点で視覚言語モデルの空間ダイクシス表現の運用能力に着目した分析を行う．

3　Memory Game に基づくベンチマーク構築

本研究では Memory Game [3, 8] と呼ばれる人間を対象とする実験を元に視覚言語モデルの評価を行う．

3.1　Memory Game

Memory Game とは，言語学的な調査を行っているということを被験者に悟らせずに指示語の使われ方を調査する実験手法である．被験者は長机の一端に座り，実験者から実験の内容について，言語が記憶に与える影響に関する実験を行うという趣旨の説明を受ける．その後，長机の上に配置した物体に対して，被験者は指差しを行った状態で，その物体を命名するように指示される．この命名の際，被験者は “that black cross” のように指示語，色，物体の形の 3 語で命名を行わなければならない．そして，実験者は，この際に使用された指示語が何であるのかを記録して，実験の一試行が完了するという流れである．実験は物体の位置や物体の形を変更して複数の試行が行われる．

[3] は Memory Game をさまざまな言語に対して行い，言語によって指示語の使われ方がどのように異なるか分析を行っている．この研究では，日本語や英語をはじめとした 29 の言語と，それらのいずれかを母語とする被験者 874 人を対象とした実験が行われた．実験では，それぞれの言語，それぞれの距離において被験者がどの指示語を選択するのかを調査し，割合をまとめている．また，実験では実験者である聞き手が，被験者である話し手の近くにいる場合と長机の反対側に座る場合とで指示語の選択に変動があるか，という点についても調査された．

3.2　ベンチマーク構築

前述の Memory Game に基づき，本研究では視覚言語モデルの空間ダイクシス表現の運用能力を評価するベンチマークを構築した．まず，視覚言語モデルの空間理解能力を評価する先行研究 [5, 6] を参考に，Blender [9] を使用して各言語につき合計 60 件の画像データを構築した．

画像データの例を図 1 に示す．Memory Game で規定されている実験設定と合わせるために，柄のない単色の長机を使用し，机の周りには距離を推定させる原因となるような別の物体を置かない，といった設定を行った．長机には図 1 に示すように，手前から 0.25m，1.50m，2.75m の位置のいずれかに対象物を配置する．本研究では，長机に配置する対象物に描かれた図形を 5 種類（円，十字，正方形，星形，三角形），図形の色を 4 種類（黒，赤，青，緑）用意し，合計で各距離について 20 件ずつの画像を作成した．

評価対象の言語は，言語間で人間を対象とした先行研究 [3] で空間ダイクシス表現の運用に違いの見られた日本語，韓国語，英語，中国語の 4 言語とした．先行研究では，日本語と韓国語は 3 種類の指示語の使い分け，英語，中国語は 2 種類の指示語の使い分けがあると報告されている．

4　実験設定

本節ではベンチマークを用いた視覚言語モデルの評価実験の設定について述べる．

表 1 指示語使い分け評価のためのプロンプト（日本語）

画像を分析し、ディスク上の図形を特定してください。以下の 3 語のテンプレートに正確に従って説明してください。[指示詞][色][図形] 制約事項：指示詞には「この」または「あの」または「その」を使用してください。色は 1 単語を使用してください。図形は 1 単語を使用してください。3 つの単語のみを出力してください。句点や導入文は含めないでください。

4.1 プロンプト

Memory Game と同様に，与えられた入力画像中の長机の上に配置した物体に対して指示語，色，物体の形の 3 語で答えさせるようにプロンプトを設計した．評価に用いたプロンプトを表 1 に示す．

このプロンプトはまず英語のプロンプトを作成した後，Google Gemini[2] を用いて各言語に翻訳する．

4.2 評価対象のモデル

評価対象のモデルは，オープンな視覚言語モデルとして Qwen3-VL[3] [10] のモデルサイズが 8B と 32B の指示学習済みのモデルと，Gemma 3[4] [11] のモデルサイズが 4B と 12B の指示学習済みのモデルを使用する．いずれのモデルにおいても貪欲法による出力トークンのデコーディングを行う．

4.3 評価指標

モデルごとに指示語のトークンの確率分布を出力させ，実験設定ごとに各指示語の選択確率を計算する．また，人間に対して行われた実験である [3] の結果を，それぞれの言語に内在する確率分布から得られた経験的な分布とみなし，モデルから得られた確率分布が人間の分布からどの程度離れているのかを調べるため，両分布間の距離を計測する．ここでは，確率が 0 になるクラスを持つ分布に対しても分布間距離の計測が可能である Jensen–Shannon 距離を採用する．具体的には，対象物との距離ごとに得られたモデルの確率分布と，人間に対する実験の結果の Jensen-Shanon 距離をとり，モデルが対象物の色と形状を正しく認識できた数で重みつき平均をとったものを計算する．

2） https://gemini.google.com/
3） https://huggingface.co/collections/Qwen/qwen3-vl
4） https://huggingface.co/collections/google/gemma-3-release

5 実験結果と分析

5.1 モデルの実験結果と分析

モデルの出力の結果をまとめた表を表 2 に示す．各行は実験設定に対応し，それぞれの実験設定のもとでの指示語の出力確率の確率分布の平均を表している．最下段の人間に対する実験結果は [3] の結果である．斜線は，モデルが対象物の色と形状を正しく出力できなかったなどの理由により該当する設定の結果が得られなかったことを示す．また，モデルが対象物の色と形状を正しく認識できていた画像の数を表 3 に示した．

全体の傾向として，モデルによらず，指示語が 3 種類ある言語 (日本語と韓国語) の遠称をほとんど使用しない傾向が見られた．人間は物体との距離が離れると遠称を選択する割合が増えるのに対し，視覚言語モデルは Qwen3-VL-32b の日本語設定の場合を除いて遠称の選択確率がどの場合も 5%を切っている．また，Qwen3-VL-8b の韓国語の結果といった一部の実験設定では指示語が 3 種類ある言語の遠称以外の指示語についても，人間は一定程度選択しているにもかかわらず，モデルはほとんど選択しないという場合が見られた．

個別のモデルの傾向として，Gemma-3 は距離によらず似た確率分布を示すことが観察される．人間は物体の距離が遠くになるにつれて近称を選択する割合が減り，遠称を選択する割合が増えるのに対し，Gemma-3-4b も Gemma-3-12b も距離による確率分布の変動は人間の結果よりも小さいことが分かる．

また，表 3 の結果から，一部の実験設定においては対象物に描かれた図形やプロンプトの指示内容を，視覚言語モデルは認識できていないことを示している．特に，Gemma-3-4b においてその結果が顕著であり，指示内容をモデルが正しく認識できていない場合が多いということは，色と形状が正しく認識された場合であっても，人間と同様の指示語の運用を行っていないという結果に影響を与えている可能性がある．

以上から，視覚言語モデルの指示語の使用の特徴として，一部の指示語が全く使われない場合があり，特に指示語が 3 種類ある言語の遠称についてはモデルを問わずその傾向が見られることが挙げられる．指示語が 3 種類ある言語において遠称を視覚言語モデルがほとんど使用していないということは，

表 2　各距離における指示語（近称，中称，遠称）の使い分けの分布．

		日本語			韓国語			英語		中国語	
		この	あの	その	i	jeo	geu	this	that	zhè ge	nà ge
モデル	距離 (m)	近称	遠称	中称	近称	遠称	中称	近称	遠称	近称	遠称
	0.25	16.44	1.95	81.61	1.83	1.06	97.12	94.46	5.54	84.65	15.35
Gemma-3-4b	1.50	20.80	1.91	77.29	3.61	1.99	94.40	92.28	7.72	88.33	11.67
	2.75	18.46	1.08	80.46	2.66	1.38	95.96	88.29	11.71	＼	＼
	0.25	34.20	1.94	63.86	95.74	0.42	3.84	97.39	2.61	85.00	15.00
Gemma-3-12b	1.50	34.00	2.10	63.90	94.69	0.21	5.10	98.09	1.91	81.01	18.99
	2.75	32.80	3.60	63.60	97.65	0.12	2.24	97.23	2.77	78.46	21.54
	0.25	23.20	0.00	76.80	0.57	0.14	99.29	29.80	70.20	100.00	0.00
Qwen3-VL-8b	1.50	12.90	0.00	87.10	0.26	0.38	99.36	24.83	75.17	100.00	0.00
	2.75	3.46	0.03	96.52	0.22	0.54	99.24	12.40	87.60	99.99	0.01
	0.25	99.13	0.02	0.85	99.47	0.32	0.21	95.52	4.48	7.96	92.04
Qwen3-VL-32b	1.50	90.41	1.07	8.53	97.57	1.96	0.47	73.31	26.69	1.42	98.58
	2.75	67.07	10.20	22.73	98.28	1.29	0.43	69.38	30.62	1.58	98.42
	0.25	94.12	1.47	4.41	93.43	6.57	0.00	72.40	27.60	95.59	4.41
人間	1.50	1.96	36.76	61.27	2.53	89.39	8.08	17.19	82.81	15.20	84.80
	2.75	1.47	87.75	10.78	2.53	92.93	4.55	9.38	90.62	3.43	96.57

表 3　対象物の色と形状がモデルに正しく認識されていた画像の数（全 60 件中）．

モデル	日本語	韓国語	英語	中国語
Gemma-3-4b	12	18	34	12
Gemma-3-12b	18	19	45	28
Qwen3-VL-8b	44	38	50	48
Qwen3-VL-32b	22	26	49	44

表 4　各言語における人間による指示語の運用との乖離の程度．

	日本語	韓国語	英語	中国語
Gemma-3-4b	0.6080	0.8705	0.5041	＼
Gemma-3-12b	0.5815	0.7313	0.6326	0.5063
Qwen3-VL-8b	0.6648	0.9200	0.1539	0.6499
Qwen3-VL-32b	0.5055	0.7446	0.4444	0.3725
一様分布	0.4932	0.5535	0.2957	0.4211

モデルは指示語が 3 種類ある言語においても，中称にあたる指示語を，指示語が 2 種類ある言語における遠称のように用いることで，指示語が 2 種類ある言語と類似の仕組みで指示語を選択している可能性がある．

5.2　人間の分布との距離

モデルの出力と [3] で行われた人間に対する実験の結果の Jensen-Shanon 距離をまとめた表を表 4 に示す．また，比較のために，各言語のすべての指示語が等確率で出力されるような確率分布 (一様分布) と人間の分布との Jensen-Shanon 距離を併記する．

表 4 の結果から，Gemma-3-4b はどの言語でも，一様分布よりも人間の分布から離れた指示語の運用の確率分布を示すことが分かる．一方，Qwen3-VL-32b は今回実験したモデルのうち日本語と中国語において比較的人間に近い分布を示していることが分かり，Qwen3-VL-32b は相対的に人間と近い空間ダイクシス表現の運用をしていたといえる．

6　おわりに

本研究では，視覚言語モデルが空間ダイクシス表現をどの程度人間に近い形で運用できているかを測定するために，人間の運用と比較可能な多言語ベンチマークを作成した．実験の結果，人間と比較して，視覚言語モデルは日本語・韓国語において遠称の使用頻度が低い傾向が見られ，一部のモデルにおいては距離に応じた表現の使い分けが行われていないことが観察された．これらの結果は，空間ダイクシス表現という基本的な空間表現について，視覚言語モデルは人間とは異なった理解をしていることを示唆する．本研究の分析に用いたプロンプト，言語，モデルの数は限られているため，今後，これらの数を増やしてさらなる分析を進める．

謝辞

本研究は JST CREST，JPMJCR2565 の支援を受けたものである．

参考文献

[1] **Grammars of Space: Explorations in Cognitive Diversity**. Language Culture and Cognition. Cambridge University Press, 2006.

[2] 斎藤純男, 田口善久, 西村義樹. 明解言語学辞典. 三省堂, 2015.

[3] Kenny R Coventry, Harmen B Gudde, Holger Diessel, Jacqueline Collier, Pedro Guijarro-Fuentes, Mila Vulchanova, Valentin Vulchanov, Emanuela Todisco, Maria Reile, Merlijn Breunesse, et al. Spatial communication systems across languages reflect universal action constraints. **Nature human behaviour**, Vol. 7, No. 12, pp. 2099–2110, 2023.

[4] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In **2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 6693–6702, 2019.

[5] Sangeet Khemlani, Tyler Tran, Nathaniel Gyory, Anthony M. Harrison, Wallace E. Lawson, Ravenna Thielstrom, Hunter Thompson, Taaren Singh, and J. Gregory Trafton. Vision language models are unreliable at trivial spatial cognition, 2025.

[6] Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities, 2025.

[7] Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 12075–12097, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[8] Harmen B Gudde, Debra Griffiths, and Kenny R Coventry. The (spatial) memory game: testing the relationship between spatial language, object knowledge, and spatial cognition. **Journal of Visualized Experiments: JoVE**, No. 132, p. 56495, 2018.

[9] Blender Online Community. Blender - a 3d modelling and rendering package. **Blender Foundation, Blender Institute**, 2016.

[10] Qwen Team. Qwen3 technical report, 2025.

[11] Gemma Team. Gemma 3. 2025.