

# マルチモーダル大規模言語モデルはジェスチャーをどこまで理解しているのか：指標性・図像性・象徴性を問う

西田典起<sup>1</sup> 井上昂治<sup>2</sup> 中山英樹<sup>3</sup> 坊農真弓<sup>4</sup> 高梨克也<sup>5</sup>

<sup>1</sup> 理化学研究所 <sup>2</sup> 京都大学 <sup>3</sup> 東京大学 <sup>4</sup> 国立情報学研究所 <sup>5</sup> 滋賀県立大学  
noriki.nishida@riken.jp

## 概要

本研究では、マルチモーダル大規模言語モデル (MLLM) がジェスチャーの意味をどの程度理解できているのか調査する。特に、MLLM は外界への参照性・依存性が高い「指標的ジェスチャー」の理解を、イメージを描写する「図像的ジェスチャー」や常識によって定められる「象徴的ジェスチャー」の理解よりも苦手とするのではないかという仮説を検証する。未来館 SC コーパスの 925 件のジェスチャーに対して人手でタイプラベルを付与し、GPT-4o を含む MLLM によるジェスチャー説明文の生成と評価を行った。その結果、MLLM は一貫して指標的ジェスチャーの理解に困難があることを明らかにした。

## 1 はじめに

人間のコミュニケーションにおいて、言語だけでなく非言語的な表現が果たす役割は非常に大きい。非言語的な表現の中でもジェスチャーは、発話内容を補完し、視覚的かつ直感的な形で情報を伝える重要な手段である。例えば、特定の外界物体への指差しや、手の動きによるイメージの描写は、言葉では伝えきれないニュアンスや関係性を補完する役割を担っている。

近年、マルチモーダル大規模言語モデル (MLLM) が様々な領域で注目を集めている [1, 2, 3]。MLLM はテキストと視覚・音声情報などの統合に優れており、多岐にわたる応用が期待されている。しかし、こうしたモデルが、実世界の動的な環境において、ジェスチャーの意味や意図をどの程度正確に理解し、応用可能なのかについて十分に明らかになっているとは言い難い。

そこで本研究では、MLLM が実世界コミュニケーションにおいてジェスチャーの意味や意図をどの程度理解できているのか評価することを目的とする。

特に、環境との動的な相互作用を通して知識獲得を行っていない MLLM にとって、外界への参照性や依存性が鍵とする「指標的 (indexical/indicative) ジェスチャー」の理解は、イメージを描写する「図像的 (iconic/depictive) ジェスチャー」や文化・慣習によって定められた「象徴的 (symbolic/emblematic) ジェスチャー」の理解と比較して困難なのではないかという仮説を立て、それを検証する。

上記の検証のために、初めにデータセットの構築を行った。データセットの構築では、映像と対話文、ジェスチャー説明文 (レリバントアノテーション [4]) が同期されている未来館 SC コーパス [5] に対して、人手で 5 種類 (指標的、図像的、象徴的、混合、その他) のジェスチャータイプを付与した。その結果、925 件のジェスチャー事例を収集した。各事例はジェスチャータイプと、対応する時間スパンの映像、会話文、ジェスチャー説明文から構成される。

次に、GPT-4o [1] や Gemini-1.5-pro [2] を含む複数の MLLM によって、映像文脈と会話文脈に対するジェスチャー説明文を生成した。生成されたジェスチャー説明文の妥当性をリファレンス説明文と比較することで評価し、ジェスチャータイプごとに妥当性の傾向に差があるのか調べた。

実験の結果、すべてのモデルにおいて、指標的ジェスチャーに対する理解精度が他のジェスチャータイプに比べて一貫して低いことが示された。生成結果を分析し、MLLM はテキストと事前学習によって得た一般常識的な内部知識に基づいてジェスチャーを解釈する傾向があり、特に外界への参照を視覚的に認識することが必要なケースにおいて、ジェスチャーの適切な解釈を生成できないケースが見られた。これらの結果は、MLLM がコミュニケーションにおける外界参照の重要性を十分に学習していないことを示唆しており、今後のモデル改良にお

いて重要な示唆を与えるものとする。本研究の成果は、MLLMの現状の限界性の一つの側面を明らかにするとともに、ジェスチャー理解能力を評価するための新たなベンチマークと方法論を提供する。これにより、次世代のコミュニケーション理解モデルの開発に向けた道筋を示すことを目指す。

## 2 データセット構築

本研究では、未来館 SC コーパス [5] に収録される映像付き対話データに対して、人手でジェスチャータイプを付与することで、評価用データセットの構築を行った。本節では、その詳細を記述する。

### 2.1 未来館 SC コーパス

未来館 SC コーパスは、日本科学未来館の科学コミュニケーター (以下 SC) が、来館者に展示説明を行う場面を、収録したマルチモーダルコーパスである。本コーパスでは、全 35 セッションの対話データが収録されており、そのうち 18 セッションにジェスチャーに関するアノテーションが付与されている。各対話データは以下の要素で構成されており、時間軸に沿って同期されている: (1) 対話の書き起こしテキスト、(2) 5 台の固定カメラから取得された映像、(3) ジェスチャー (身体動作) に関する自然言語による行為記述 (レリバントアノテーション)。未来館 SC コーパスでは、事前に固定的なタグセットを定義するのではなく、対話参加者の志向性や理解に基づいて「参加者にとってレリバントな」行為記述が行われている [4]。具体的には、身体部位 (face, body, hand, foot) ごとに、参加者の理解に志向した記述 (上位層) と身体の物理的な動きに志向した記述 (下位層) を提供している。本研究では、hand 部位に関する上位層の記述をジェスチャー説明文として用いた。これにより、指標的、図像的、または象徴的なジェスチャーの効率的な収集と、対話参加者の理解に基づくジェスチャーの意味をモデルがどの程度認識できているのか評価することを可能にしている。

### 2.2 ジェスチャータイプの定義

本研究では、未来館 SC コーパスにおいて、身体部位が hand でありかつ上位層の説明文が付与されている 925 件のジェスチャーに対して、人手<sup>1)</sup>で以下

1) 論文の投稿段階では、第一著者による作業が完了しており、外部作業員によるアノテーション作業は継続中である。

表1 アノテーション結果の統計情報。

ジェスチャータイプ	事例数 (割合)	平均長 [秒]
指標的	276 (29.8%)	7.63
図像的	161 (17.4%)	7.51
象徴的	8 (0.9%)	7.27
混合	31 (3.4%)	7.69
その他	449 (48.5%)	7.23
全体	925 (100%)	7.41

の 3+2 タイプのラベルを付与した。

- **指標的:** 指標的 (indexical)、または直示的 (indicative) なジェスチャー。特定の対象物 (人、展示物など) を指し示したり、注意を引くためのジェスチャー。例: 展示物を指さす。観客に視線を向けさせるために手で方向を示す。
- **図像的:** 図像的 (iconic)、または描写的 (depictive) なジェスチャー。形状、動き、配置などの抽象的または具体的な概念を視覚的に表現するジェスチャー。例: 天体の形状を手で描く。山の高さを手で表現する。両手を交互に振って走っている様子を真似る。
- **象徴的:** 象徴的 (symbolic)、またはエンブレム的なジェスチャー。文化的・社会的に意味が定められたジェスチャー。例: 親指を立てて「グッド」の意味を示す。手を振って挨拶する。手を振って「NO」を表す。指を折り (or 伸ばし) ながらカウントや列挙をする。
- **混合:** 上記の 3 タイプへの分類が難しく、混合的なもの。複数タイプのジェスチャーが同時または逐次的に発生するもの。例: 指差しながら、指す対象物にむかって指先で円を描く。目を指差しながら、目に光線が入ってくる様子を手振りで表現する。
- **その他:** 上記に該当しないジェスチャー。

### 2.3 統計情報

アノテーション結果の統計情報を表 1 に示す。ジェスチャータイプごとの事例数には偏りが見られるが、これは対象コーパスの特性を考えると自然な現象である。特に指標的ジェスチャーと図像的ジェスチャーの事例数が「その他」を除く他のジェスチャータイプよりも多いことが特徴的であり、これは未来館 SC コーパスが展示物を中心とした対話を対象としていることに起因すると考えられる。この

以下の会話文と対話シーンの画像系列に基づいて、シーンの終盤で行われたジェスチャーの意味を説明してください。

会話文:

scA: きよ、星がキラキラ集まったのが、もっと実は、もっともっと外にあります。

v01: もっと(す)。

scA: そうそう、もっと

v04: ヘーエーエ。

v01: もっとあるだって、りゅうせい。

scA: もっと、そう、もっといっぱいあるんです。

v03: (あんどろめだー;アンドロメダ)(のね)

scA: あのね、望遠鏡が見る

scA: 機会あるかもしれません。

scA: あの

ジェスチャーの説明:

このシーンの終盤で行われたジェスチャーは、指を使って何かを指し示す動作です。このジェスチャーは、会話の内容に関連して、特定の場所や物体を示すために行われたものと考えられます。具体的には、星や天体に関する話題が出ているため、展示物やディスプレイ上の天体を指し示して、説明を補足している可能性があります。指差しは、視覚的な情報を強調し、聞き手に具体的なイメージを持たせるための効果的な手段です。



図1 ジェスチャー説明文の生成のために用いたプロンプトの例。赤文字は変数として入力事例ごとに変わる部分を、青文字はGPT-4oによって実際に生成された部分を示す。

傾向は、本コーパスが本研究の仮説を検証する上で適したデータセットであることを示唆している。また、各事例の前文脈(対象ジェスチャーの開始時点までの5秒間)を含む平均時間長にはジェスチャータイプ間で偏りが認められず、ジェスチャータイプごとのジェスチャー説明文生成の妥当性評価に差異がある場合、その原因は時間長によるものではないことが示されている。

### 3 実験設定

本研究では、以下の問いに答えるための実験を行った。(1) MLLM は動的な環境においてジェスチャーの意味や意図をどの程度正確に説明できるか? (2) ジェスチャーを指標的、図像的、象徴的というタイプに分けたとき、MLLM の指標的ジェスチャーの理解精度は他タイプに比べて低いのではないか?

**モデル:** これまで様々な MLLM アーキテクチャが提案されてきたが [6]、様々なデータセットにおいて、GPT-4o 等の商用 API モデルの精度の高さが報告されている [7]。そこで本実験でも、商用 API モデルである GPT-4o と GPT-4o-mini (GPT-4o の軽量版)、Gemini-1.5-pro、Gemini-1.5-flash (Gemini-1.5-pro の軽量版) を用いた。

**生成方法:** ジェスチャー説明文の生成では、各対象ジェスチャーに至るまでの会話文と映像を文脈として与え、プロンプトによってジェスチャーの意味や意図を説明するように指示した。ジェスチャー説明文の生成に用いたプロンプトを図1に示す。

以下は、あるジェスチャーの説明文のペアです。1つはモデルによって生成された説明文であり、もう1つは人手で作成された正解説明文です。生成された説明文が正解説明文の意味をどの程度含んでいるかを評価してください。評価は以下の基準に基づき、0.00から1.00の間のスコアで示してください。また、その理由を簡潔に説明してください。

評価基準:

- 1.00: 生成された説明文が正解説明文の意味を完全に含んでいる。
- 0.75: 生成された説明文が正解説明文の意味の大部分を含んでいるが、いくつかの細部が欠けている。
- 0.50: 部分的に一致しているが、重要な情報が欠けている。
- 0.25: わずかな関連性があるが、大部分が一致していない。
- 0.00: 全く意味が一致していない。

ジェスチャーのタイプ: 指標的

生成された説明文:

このシーンの終盤で行われたジェスチャーは、指を使って何かを指し示す動作です。このジェスチャーは、会話の内容に関連して、特定の場所や物体を示すために行われたものと考えられます。具体的には、星や天体に関する話題が出ているため、展示物やディスプレイ上の天体を指し示して、説明を補足している可能性があります。指差しは、視覚的な情報を強調し、聞き手に具体的なイメージを持たせるための効果的な手段です。

正解説明文:

「望遠鏡が見る機会があるかもしれません」の発話の主語がvであることを示す

評価スコア: 0.25

理由: 生成された説明文は、指を使って何かを指し示す動作についての説明を行っていますが、正解説明文の具体的な内容(「望遠鏡が見る機会があるかもしれません」の発話の主語がvであることを示す)とは関連性が薄いです。生成された説明文は、ジェスチャーの一般的な機能や目的について述べていますが、正解説明文の重要な情報(発話の主語や具体的な文脈)を含んでいないため、わずかな関連性があるものの、大部分が一致していないと評価しました。

図2 ジェスチャー説明文の評価に用いたプロンプトの例。

すべての MLLM において同じプロンプトを使用した。本実験では指標的タイプ、図像的タイプ、象徴的タイプに属する 445 件のジェスチャー事例のみを対象とし、混合タイプと「その他」タイプの事例は除いた。

**評価方法:** MLLM によって生成されたジェスチャー説明文の妥当性を評価するために、未来館 SC コーパスのジェスチャー説明文(hand 部位かつ上位層の行為記述)をリファレンス説明文とし、生成された説明文とリファレンス説明文との一致度合いを求めた。そのための評価尺度としては、ROUGE や BLEU 等の n-gram ベースの評価尺度も考えられるが、近年では、LLM による評価の信頼性の高さが知られている [8, 9, 10]。そこで、本実験でも、生成された説明文とリファレンス説明文を LLM (GPT-4o-mini) に提示し、生成された説明文の妥当性を評価するように指示した。評価のために用いたプロンプトを図2に示す。モデル間での公平性を満たすために、MLLM の種類によらず、評価では一貫して GPT-4o-mini を用いた。

## 4 実験結果と考察

### 4.1 MLLM にとって指標的ジェスチャーの理解は難しいか?

MLLM のジェスチャー理解力にジェスチャータイプ間で差があるか調べるために、各ジェスチャー事例に対して生成した説明文の自動評価スコアをジェスチャータイプごとに平均化した。その結果を図3に示す。図からわかるとおり、MLLM の違いによらず、一貫して、指標的ジェスチャーにおける評

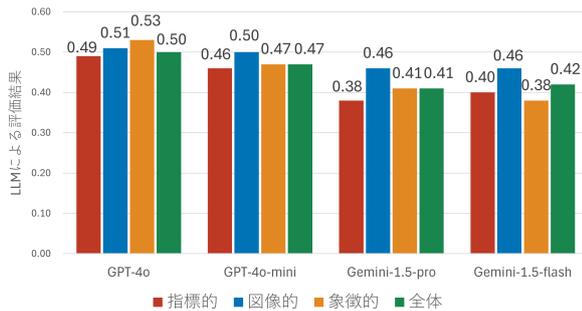


図3 LLM (GPT-4o-mini) によるジェスチャータイプ別評価結果。各ジェスチャータイプに属する事例に対する評価スコアの平均を示す。「全体」はすべての事例に対する評価スコアの平均を示す。

評価スコアが図像的ジェスチャーよりも低いことが観測された。また、Gemini-1.5-flashを除く他のすべてのモデルでは、指標的ジェスチャーの評価スコアは象徴的ジェスチャーおよび全体で求めたスコアよりも低かった。これらの結果は、MLLMにとって、環境への依存性が弱く事前学習で得た一般的な常識によってある程度推測できる図像的ジェスチャーおよび象徴的ジェスチャーの解釈に比べて、外界の動的な環境への参照性が鍵となる指標的ジェスチャーの解釈は難しいという、本研究の仮説を支持する。

#### 4.2 指標的ジェスチャーの理解のためには何が欠けているのか？

前節の結果から、MLLMにとって指標的ジェスチャーの理解が他タイプのジェスチャーに比べて難しいということがわかった。本節では、どのような情報が指標的ジェスチャーの理解のために欠けているのか調査するために、様々な追加情報をプロンプトに加えて、指標的ジェスチャーに関して生成された説明文の質が向上するか調べた。具体的には、(1) 前文脈を5秒から10秒に拡張した会話文脈、(2) 未来館SCコーパスに収録されているhand部位かつ下位層(物理的な動きに着目)のジェスチャー説明文、または(3) 本研究で付与したジェスチャータイプラベルを追加情報とした。結果を表2に示す。会話文脈を拡張した場合でも、モデルの評価スコアには顕著な向上は見られなかった。一方で、下位層説明文およびジェスチャータイプの情報を追加した場合には、評価スコアが大きく向上する結果が得られた。この結果は、GPT-4o(を含むMLLM)にとって、指標的ジェスチャーの理解を困難にしている要因が会話文脈の欠如ではなく、ジェスチャーそのものが有する視覚的特徴や環境・身体の物理的特性に関する深い理解に起因する可能性を示唆している。

表2 指標的ジェスチャーに対する説明文生成において、追加情報が与える影響。

モデル	追加情報	指標的
GPT-4o	なし	0.49
GPT-4o	会話文脈の拡張	0.50
GPT-4o	下位層説明文	0.61
GPT-4o	ジェスチャータイプ	0.58

#### 4.3 指標的ジェスチャーはMLLMによってどのように解釈されているか？

GPT-4oによって実際に生成されたジェスチャー説明文と、対応する評価結果を、それぞれ図1と図2の青文字部分に載せる。図1からわかるように、この事例では、会話文のみからは対象ジェスチャーが指標的であることを推測することは難しい。にもかかわらず、GPT-4oはジェスチャーが「指を使って何かを指し示す動作」であることを同定することに成功している。しかし、図2の正解説明文が示すように、このジェスチャーは「望遠鏡将来見る機会があるかもしれません」の省略された主語がvであることを示すためにvを指し示す表現であるが、GPT-4oは、このジェスチャーを「星や天体に関する話題が出ているため、展示物やディスプレイ上の天体を指し示している」と解釈してしまっている。つまり、GPT-4oは、ジェスチャーの外界参照性とその意図を十分に把握する前に、テキスト情報と一般常識を使ってジェスチャーを間違った形で推測してしまっている。図像的、または象徴的なジェスチャーの場合は、このようなテキスト情報に基づく推測や一般常識が有効に働くと予想されるが、指標的ジェスチャーの場合は、外界に強くグラウンディングされた解釈(説明文の生成)を要求されるため、指標的ジェスチャーに対する評価スコアが低くなったと考えることができる。

### 5 おわりに

本研究では、マルチモーダル大規模言語モデル(MLLM)のジェスチャー理解力を調べるために、ジェスチャーを指標的、図像的、象徴的のタイプに分類し、各ジェスチャータイプごとのジェスチャー説明文の生成結果を評価した。実験結果は、MLLMはモデルによらず一貫して指標的ジェスチャーの理解を苦手としており、それが外界への参照性や環境との相互作用、身体の物理的特性に関する深い理解の欠如に起因する可能性を示唆している。

## 謝辞

本研究は JSPS 科研費 22B102, JP22H05015, JP21K17815 の助成を受けたものです。

## 参考文献

- [1] OpenAI. Gpt-4 technical report, 2024.
- [2] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.
- [3] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 543–553, Singapore, December 2023. Association for Computational Linguistics.
- [4] 坊農真弓, 須永将史. 身体動作のアノテーション手法の提案. 人工知能学会 言語・音声理解と対話処理研究会 (第 76 回). SIG-SLUD-B503, 2016.
- [5] Mayumi Bono, Hiroaki Ogata, Katsuya Takanashi, and Ayami Joh. The practice of showing ‘who i am’: A multimodal analysis of encounters between science communicator and visitors at science museum. In Constantine Stephanidis and Margherita Antona, editors, **Universal Access in Human-Computer Interaction. Universal Access to Information and Knowledge**, pp. 650–661, Cham, 2014. Springer International Publishing.
- [6] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. **arXiv preprint arXiv:2306.13549**, 2023.
- [7] Chaochao Lu, Chen Qian, Guodong Zheng, Hongxing Fan, Hongzhi Gao, Jie Zhang, Jing Shao, Jingyi Deng, Jinlan Fu, Kexin Huang, et al. From gpt-4 to gemini and beyond: Assessing the landscape of mllms on generalizability, trustworthiness and causality through four modalities. **arXiv preprint arXiv:2401.15071**, 2024.
- [8] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. **Advances in Neural Information Processing Systems**, Vol. 36, pp. 46595–46623, 2023.
- [9] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. **arXiv preprint arXiv:2402.04788**, 2024.
- [10] Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. Llm-as-a-judge & reward model: What they can and cannot do. **arXiv preprint arXiv:2409.11239**, 2024.