

Do Multimodal Large Language Models Truly See What We Point At? Investigating Indexical, Iconic, and Symbolic Gesture Comprehension

Noriki Nishida¹, Koji Inoue², Hideki Nakayama³, Mayumi Bono⁴, Katsuya Takanashi⁵

¹RIKEN, ²Kyoto University, ³The University of Tokyo,

⁴National Institute of Informatics, ⁵The University of Shiga Prefecture

noriki.nishida@riken.jp, bono@nii.ac.jp

Abstract

Understanding hand gestures is essential for human communication, yet it remains unclear how well multimodal large language models (MLLMs) comprehend them. In this paper, we examine MLLMs' ability to interpret indexical gestures, which require external referential grounding, in comparison to iconic gestures, which depict imagery, and symbolic gestures, which are conventionally defined. We hypothesize that MLLMs, lacking real-world referential understanding, will struggle significantly with indexical gestures. To test this, we manually annotated five gesture type labels to 925 gesture instances from the Miraikan SC Corpus and analyzed gesture descriptions generated by state-of-the-art MLLMs, including GPT-4o. Our findings reveal a consistent weakness across models in interpreting indexical gestures, suggesting that MLLMs rely heavily on linguistic priors or commonsense knowledge rather than grounding their interpretations in visual or contextual cues.

1 Introduction

Human communication is inherently multimodal and extends beyond language; nonverbal expressions, particularly hand gestures (hereafter, gestures), are fundamental in conveying meaning and enhancing interaction (McNeill, 1992; Goldin-Meadow, 2003; Kendon, 2004; Kita, 2000). In recent years, multimodal large language models (MLLMs) have gained significant attention across various domains (Yin et al., 2023; Lu et al., 2024; Liu et al., 2023; Sun et al., 2024; Li et al., 2023; Alayrac et al., 2022; Maaz et al., 2024; Su et al., 2023; Zhang et al., 2023). These models excel at integrating textual, auditory, and visual information. However, their ability to accurately interpret gestures, particularly in dynamic real-world communication, remains underexplored.

In this paper, we investigate the extent to which MLLMs can comprehend the meaning and intent

behind gestures in real-world communication. We hypothesize that MLLMs, which do not acquire knowledge through direct interaction with their environment, will struggle significantly with *indexical* gestures—gestures that rely on external referents. Compared to *iconic* gestures (which depict imagery) and *symbolic* gestures (which are conventionally defined by cultural norms), indexical gestures require an understanding of external grounding, posing a challenge for MLLMs.

To test this hypothesis, we first constructed a benchmark dataset based on the Miraikan Science Communication (SC) Corpus (Bono et al., 2014; Sakaida et al., 2018), which contains Japanese dialogue transcripts, videos, and gesture descriptions. We defined and manually assigned five gesture type labels (Indexical, Iconic, Symbolic, Mixed, and Others) to 925 gesture instances in this corpus. Figure 1 illustrates examples of indexical, iconic, and symbolic gestures and their corresponding dialogue contexts and human-written descriptions.

Then, using state-of-the-art MLLMs, including GPT-4o (OpenAI, 2024), Gemini 1.5 Pro (Gemini Team, 2024), Qwen2.5-VL (Bai et al., 2025), and LLaVA-NeXT-Video (Li et al., 2024), we generated gesture descriptions based on both video frames and dialogue contexts. These generated descriptions were then evaluated against human-written reference descriptions to assess their validity. Finally, we analyzed performance differences across gesture types to determine whether MLLMs exhibit systematic weaknesses in interpreting certain gesture types.

Our experiments reveal a consistent difficulty across all tested MLLMs in accurately interpreting indexical gestures compared to iconic and symbolic gestures. Further analysis suggests that MLLMs tend to rely on their internal knowledge, derived from text and pretraining, rather than visually recognizing referential grounding of gestures in dynamic environments. These findings indicate that

Indexical Gesture	Iconic Gesture	Symbolic Gesture
 <p>Dialogue Context: scA: Yes, yes, that's right. scA: This is the Subaru Telescope, a Japanese telescope. v01 (woman): Yeah. scA: Do you remember where it is? scA: Have you heard about it before?</p> <p>Human-Written Description: Indicates that the question is directed at v02.</p>	 <p>Dialogue Context: scA: And when it comes to uncovering these mysteries, in the past... scA: People like Da Vinci or Galileo Galilei... v02: Yeah. scA: They observed things by themselves using telescopes.</p> <p>Human-Written Description: Makes a gesture of looking through a telescope.</p>	 <p>Dialogue Context: scA: Earlier, we spread out the sun using a red sheet. scA: The Subaru Telescope, however, uses a single mirror</p> <p>Human-Written Description: Emphasizes that it is a single mirror.</p>

Figure 1: Examples of indexical, iconic, and symbolic gestures, along with their corresponding videos, dialogue contexts, and human-written descriptions. While the original dialogue and descriptions are in Japanese, we provide English translations for clarity.

MLMs have yet to fully internalize the role of external reference in human communication. Furthermore, unlike most prior work on multimodal language models, which focuses on English inputs and outputs, our study investigates gesture understanding and description generation entirely in Japanese, offering a perspective from a non-English interaction context. We publicly release the annotations along with the source code for data processing and experimentation¹.

2 Dataset Construction

2.1 Building on the Miraikan SC Corpus

We constructed a benchmark dataset by manually annotating *gesture types* to the Miraikan SC Corpus (Bono et al., 2014; Sakaida et al., 2018), a multimodal dataset of video-recorded Japanese conversations between science communicators (SCs) and visitors at the Miraikan science museum in Japan. The corpus contains 35 dialogue sessions, of which 18 sessions include manually-annotated gesture descriptions. Each dialogue session consists of the following data streams synchronized based on timestamps: (1) utterance transcripts, (2) videos captured from 5 fixed cameras, and (3) gesture descriptions. The Miraikan SC Corpus adopts a descriptive approach to gesture description annotation, with a focus on the relevance of gestures

to participants’ understanding (Bono and Sunaga, 2016). The gesture descriptions are structured into two levels: *interpretation-level descriptions*, which focus on how the gesture intent is interpreted in the communicative context, and *physical-level descriptions*, which detail the physical movements of body parts (face, body, hand, foot)². In our experiments, we used the interpretation-level descriptions of hand movements as references.

2.2 Gesture Type Definitions

We manually defined and assigned one of five gesture types to each of the 925 hand gestures annotated in the Miraikan SC Corpus.

- **Indexical:** Gestures that point to specific referents (e.g., people, objects). Example: Pointing at an exhibit; using hand movements to guide a visitor’s gaze.
- **Iconic:** Gestures that visually depict shapes, motions, or spatial configurations of objects or concepts. Example: Drawing the shape of a planet with hands; indicating a mountain’s height; mimicking running motions with alternating hand movements.
- **Symbolic:** Gestures defined culturally or socially with conventional meanings. Example: Giving a thumbs-up to indicate “good”; waving to greet

²For example, “The hand is extended forward and then raised upward” is a physical-level description, while “The gesture indicates the direction of the planet” is an intentional-level description.

¹<https://github.com/norikinishida/mlm-gesture-eval>

Gesture Type	# Examples	Avg. Len. [sec]
Indexical	309 (33.4%)	7.40
Iconic	169 (18.3%)	7.39
Symbolic	8 (0.9%)	6.90
Mixed	20 (2.2%)	9.00
Others	185 (20.0%)	7.30
Uncertain	234 (25.3%)	7.42
Overall	925 (100%)	7.41

Table 1: Dataset Statistics. We show the number of gesture examples for each gesture type. The average duration of each gesture type is also shown.

someone; making a “no” gesture by waving a hand; counting with fingers.

- **Mixed:** Gestures that combine multiple types simultaneously or sequentially. Example: Pointing at an exhibit while drawing a circle around it; pointing at one’s eyes while mimicking light entering them.
- **Others:** Gestures outside the above types.

2.3 Annotation Procedure and Statistics

We assigned 3 external annotators³ to label the same set of 925 gestures with gesture types. All 925 gesture instances were drawn from the 18 sessions in the Miraikan SC Corpus that include manually-annotated gesture descriptions.

We measured inter-annotator agreement across the 3 annotators. Out of the 925 annotated samples, 691 samples (74.7%) had full agreement among all 3 annotators; 220 samples (23.8%) had partial agreement, with 2 annotators assigning the same label and the third assigning a different label; 14 samples (1.5%) had no agreement, with all 3 annotators assigning different labels.

To ensure label reliability, we retained only gesture type labels that were consistently assigned by all annotators. For instances with annotation discrepancies, we assigned a new label, “Uncertain”.

Table 1 presents the statistical distribution of the annotated gesture types. Notably, indexical and iconic gestures appear more frequently than symbolic and mixed gestures. This trend aligns with the nature of the Miraikan SC Corpus, which primarily captures exhibit-centered conversations, where pointing and illustrative gestures are commonly used. Furthermore, the average duration of each gesture instance shows no significant variation

³All annotators were native Japanese speakers familiar with the context of the Miraikan SC Corpus, reducing potential cultural mismatches in interpreting symbolic gestures.

across indexical, iconic, and symbolic types. This suggests that any observed differences in MLLM-generated gesture descriptions across these types are not due to differences in temporal length, ensuring fair evaluation conditions. Multi-person gestures were extremely rare in the dataset, as the recorded interactions followed structured turn-taking protocols in a science museum setting.

3 Experimental Setup

3.1 Multimodal Large Language Models

A variety of multimodal large language model (MLLM) architectures have been proposed in recent years (Yin et al., 2023). While commercial API models such as GPT-4o and Gemini 1.5 Pro have demonstrated superior performance across multiple datasets (Lu et al., 2024; Fu et al., 2024), open-source MLLMs are increasingly being adopted for academic and reproducible research. To ensure broader model coverage, we included the following proprietary models: GPT-4o (OpenAI, 2024), GPT-4o-mini, Gemini 1.5 Pro (Gemini Team, 2024), Gemini 1.5 Flash. In addition, we included the following open-source models: Qwen2.5-VL-7B-Instruct (Bai et al., 2025; Wang et al., 2024; Bai et al., 2023) and LLaVA-NeXT-Video (Liu et al., 2023; Li et al., 2024).

3.2 Gesture Description Generation

To generate gesture descriptions, we provided each MLLM with both dialogue context and video frames leading up to the target gesture. Specifically, the input consisted of video frames and dialogue utterances covering the annotated gesture span, along with the 5-second window of preceding context. A prompt was used to instruct the models to describe the gesture’s meaning and intent. Figure 3 in Appendix A illustrates the prompt template used for generation. The same prompt was used across all MLLMs to ensure consistency in evaluation. We focused only on 486 examples labeled as indexical, iconic, or symbolic.

3.3 Evaluation

To evaluate the validity of MLLM-generated gesture descriptions, we used the gesture descriptions from the Miraikan SC Corpus (hand-movement annotations with intentional-level descriptions) as reference descriptions. We employed GPT-4o-mini as the evaluator, prompting it to assess the accuracy and validity of the generated descriptions based on

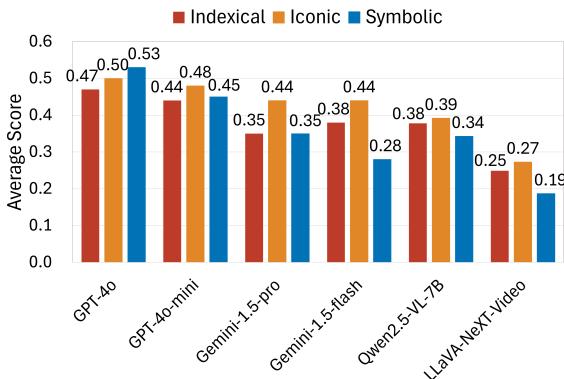


Figure 2: Performance of gesture descriptions generated by MLLMs. Evaluation scores were assigned using GPT-4o-mini, and the average score was computed for each gesture type.

the human-written reference descriptions (Zheng et al., 2023; Chen et al., 2024; Son et al., 2024). The scores range from 0.0 to 1.0, with higher scores indicating greater alignment with the reference descriptions. Figure 6 in Appendix A shows the evaluation prompt. It is worth noting that the human-written reference descriptions occasionally refer to speakers using transcript labels such as “v02” (visitor 02). These speaker labels are included in the dialogue context provided to the models, but are not explicitly linked to any visual entities in the video. This lack of alignment between textual speaker labels and visual grounding may affect the model’s ability to resolve referents in some cases.

4 Results and Discussion

4.1 Do MLLMs Struggle with Indexical Gestures?

To evaluate differences in MLLMs’ gesture comprehension across types, we averaged the evaluation scores of generated descriptions within each gesture type.

The results (Figure 2) show a clear trend: Indexical gestures consistently received lower scores than iconic gestures across all test models. This pattern held across both proprietary models and open-source models, indicating a consistent weakness in interpreting indexical gestures. Scores for symbolic gestures varied significantly.

These findings confirm our hypothesis that MLLMs struggle with indexical gestures, which require external referential grounding beyond linguistic priors and commonsense knowledge. This limitation underscores a key challenge in current

multimodal AI: while MLLMs excel at text-based reasoning, they struggle with context-aware, visually grounded interpretations of referential gestures. Such gestures are essential for human-like communication in dynamic environments, highlighting the need for models that better integrate vision and situational grounding in real-world interactions.

In contrast, iconic gestures, which are often self-contained and can be inferred through linguistic context and commonsense knowledge, are interpreted more reliably. The inconsistency in symbolic gesture scores may suggest that their comprehension is highly model-dependent, influenced by the distribution of cultural conventions in their training data. However, this variability should be interpreted with caution, as the number of symbolic gesture instances is very limited ($n=8$).

4.2 What Information is Missing for Indexical Gesture Comprehension?

To investigate why MLLMs struggle with indexical gestures and identify the contextual information they may be missing, we conducted additional experiments using GPT-4o. We tested whether augmenting prompts with additional cues would improve the quality of generated descriptions. We explored three modifications: (1) expanding the preceding dialogue window from 5 to 10 seconds for extended dialogue context, (2) incorporating physical-level descriptions of hand movements, and (3) explicitly specifying the gesture type labels. The prompts used for these settings are detailed in Appendix A.

Table 2 presents the results. Extending the dialogue context had minimal effect, suggesting that a longer textual context alone does not significantly improve indexical gesture interpretation. In contrast, providing physical-level descriptions and explicit gesture type labels substantially improved performance, indicating that these gesture-related cues contribute essential information that MLLMs otherwise fail to infer.

These findings suggest that MLLMs’ difficulty with indexical gestures is not merely due to insufficient conversational context but rather a lack of understanding of physical motion and referential grounding. While iconic and symbolic gestures can often be self-contained and interpreted using linguistic context and commonsense knowledge, indexical gestures require direct grounding, which MLLMs fail to achieve without external cues.

Additional Cues	Score
No augmentation	0.47
Extended dialogue context	0.48
Physical-level gesture description	0.60
Gesture type label	0.54

Table 2: Impact of additional cues on indexical gesture description generation.

4.3 Ablation on Modal Contributions

To better understand the respective contributions of visual and linguistic context, we conducted ablation experiments under “dialogue-only” and “vision-only” conditions using GPT-4o. For this analysis, we focused on indexical and iconic gestures.

Table 3 presents the average evaluation scores for each condition. For indexical gestures, removing the visual input led to a noticeable drop in performance ($0.47 \rightarrow 0.38$), indicating that visual information is essential for resolving referential grounding. In contrast, removing dialogue context had a milder effect ($0.47 \rightarrow 0.43$), suggesting that indexical interpretation is heavily reliant on visual cues. Iconic gestures, by contrast, were relatively robust to the absence of visual input ($0.50 \rightarrow 0.49$), but showed a significant drop when dialogue was removed ($0.50 \rightarrow 0.29$). This suggests that the meaning of iconic gestures is more closely tied to the accompanying linguistic context. For example, a gesture illustrating the height of a mountain is likely to co-occur with utterances about tall mountains, which provides complementary cues that support the interpretation of the gesture’s meaning. In contrast, indexical gestures remain difficult to interpret without visual grounding, as dialogue alone often lacks sufficient information to resolve what is being pointed to.

These findings reinforce the hypothesis that indexical gesture comprehension requires visually grounded, referential understanding, while iconic gesture interpretation relies more on context and commonsense associations in language.

4.4 How Are Indexical Gestures Interpreted by MLLMs?

To better understand how MLLMs interpret indexical gestures, we analyzed gesture descriptions generated by GPT-4o alongside their evaluation scores. We found that while GPT-4o often recognized pointing motions as indexical gestures, it

Input Condition	Indexical	Iconic
Full Input	0.47	0.50
Dialogue Only	0.38	0.49
Vision Only	0.43	0.29

Table 3: Evaluation scores for indexical and iconic gestures under different input conditions. “Full Input” uses both video and dialogue, while the other two are ablations.

frequently misinterpreted their referential intent. For instance, in one case (Figure 6), the human-written description indicated that the pointing gesture serves to disambiguate the (omitted) subject of the utterance, “(...) might have a chance to see through a telescope in the future,” without which the subject would remain unclear⁴. However, GPT-4o inferred that the pointing gesture referred to a celestial object on display, likely relying on text-based reasoning rather than external grounding.

These findings suggest that MLLMs prioritize linguistic context and commonsense knowledge over real-world referential resolution. While this strategy suffices for iconic and symbolic gestures, where meaning is largely self-contained, indexical gestures require explicit situational grounding, which MLLMs struggle to achieve.

5 Conclusion

This study investigated MLLMs’ ability to comprehend gestures, revealing a consistent weakness in interpreting indexical gestures, which require external referential grounding. Our findings suggest that current models may struggle with interpreting relational gestures such as pointing, possibly due to the absence of dynamic and embodied interactions in their training data. While not conclusive evidence of a fundamental architectural limitation, the results highlight a meaningful gap in multimodal grounding capabilities. A more comprehensive understanding of gesture interpretation may benefit from broader evaluations across diverse conversational contexts, non-English languages, and cultural settings, using datasets that reflect varied interaction styles and referential behaviors. We see our work as a step toward more inclusive and linguistically diverse benchmarks for multimodal understanding.

⁴In Japanese, it is common to omit the subject of a sentence when it can be inferred from context.

Limitations

While this study provides key insights into the limitations of MLLMs in gesture comprehension, several aspects remain to be addressed. (1) Our analysis is based on the Miraikan SC Corpus, which captures interactions in a science museum setting. While this dataset provides rich multimodal information, its domain specificity may limit the generalizability of our findings to other communicative contexts. Future work should explore more diverse datasets, including daily conversations, instructional videos, or multilingual interactions. (2) While our prompts were designed to elicit natural gesture interpretations from MLLMs, they did not fully replicate the annotation guidelines used for the original human-written descriptions in the Miraikan SC Corpus. This gap in prompting design may have contributed to certain discrepancies in interpretation. (3) We employed LLM-based evaluation (GPT-4o-mini) to assess the quality of gesture descriptions. While LLM-based evaluation has been shown to be reliable in many tasks, it remains a proxy measure and may not fully capture the nuances of human interpretation of gestures. We acknowledge this limitation and recommend that future work incorporate human evaluations using consistent rubrics to validate automated assessment more robustly. (4) Symbolic gestures, in particular, are culturally dependent, and the performance variability across models suggests that training data composition plays a major role. Expanding evaluations to different language models and cultural contexts would help clarify whether MLLMs truly internalize gesture meaning or simply reflect training biases.

Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful and insightful comments, which we found very helpful in improving the paper. This work was supported by JSPS KAKENHI Grant Numbers JP22B102, JP22H05013, JP22H05015, and JP21K17815.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne

Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA. Curran Associates Inc.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Mayumi Bono, Hiroaki Ogata, Katsuya Takanashi, and Ayami Joh. 2014. The practice of showing ‘who i am’: A multimodal analysis of encounters between science communicator and visitors at science museum. In *Universal Access in Human-Computer Interaction. Universal Access to Information and Knowledge*, pages 650–661, Cham. Springer International Publishing.

Mayumi Bono and Masashi Sunaga. 2016. A proposal of annotation scheme for body movement based on participants’ understandings. In *JSAI SIGs Conference Papers, SIGSLUD-B503*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal lilm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*.

Chaoyou Fu, Yuhua Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint, arXiv:2405.21075*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint, arXiv:2403.05530*.

Susan Goldin-Meadow. 2003. *Hearing Gesture: How Our Hands Help Us Think*. Harvard University Press, Cambridge, MA.

- Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge.
- Sotaro Kita. 2000. How representational gestures help speaking. In David McNeill, editor, *Language and Gesture*, pages 162–185. Cambridge University Press, Cambridge, UK.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. *Visual instruction tuning*. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Chaochao Lu, Chen Qian, Guodong Zheng, Hongxing Fan, Hongzhi Gao, Jie Zhang, Jing Shao, Jingyi Deng, Jinlan Fu, Kexin Huang, et al. 2024. 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*.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. *Video-ChatGPT: Towards detailed video understanding via large vision and language models*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.
- David McNeill. 1992. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press.
- OpenAI. 2024. *Gpt-4 technical report*. Preprint, arXiv:2303.08774.
- Rui Sakaida, Ryosaku Makino, and Mayumi Bono. 2018. Preliminary analysis of embodied interactions between science communicators and visitors based on a multimodal corpus of Japanese conversations in a science museum. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. 2024. Llm-as-a-judge & reward model: What they can and cannot do. *arXiv preprint arXiv:2409.11239*.
- Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. *PandaGPT: One model to instruction-follow them all*. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 11–23, Prague, Czech Republic. Association for Computational Linguistics.
- Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. 2024. video-salmonn: speech-enhanced audio-visual large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- Hang Zhang, Xin Li, and Lidong Bing. 2023. *Video-LLAMA: An instruction-tuned audio-visual language model for video understanding*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Prompts

This appendix presents the prompts used for gesture description generation and evaluation in our experiments.

A.1 Gesture Description Generation Prompts

To assess MLLMs’ ability to interpret gestures, we designed a baseline prompt and explored modifications incorporating external cues to improve Indexical gesture comprehension. Figure 3 presents the baseline prompt, where MLLMs generate gesture descriptions using only dialogue context and video frames without additional guidance. Figure 4 shows the prompt with physical-level descriptions, providing explicit details on hand movements to help MLLMs recognize gestures based on motion rather than relying solely on textual context. Figure 5 illustrates the prompt with explicit gesture

type labels (Indexical, Iconic, or Symbolic), guiding MLLMs to distinguish gestures based on their communicative function.

A.2 Evaluation Prompt

To evaluate the quality of generated gesture descriptions, we employed GPT-4o-mini as the evaluator. Figure 6 presents the evaluation prompt, where GPT-4o-mini was provided with both the generated description and the human-written reference description and instructed to assess their accuracy and validity on a 0.0 to 1.0 scale.

B Annotation Policy and Labeling Considerations

We used the “Mixed” category to account for cases where gestures could not be reliably separated into distinct types due to overlapping characteristics. While a multi-label annotation scheme (e.g., tagging both “indexical” and “iconic”) may offer a finer-grained representation, we prioritized high inter-annotator agreement and interpretability in this initial study. We consider multi-label annotation a promising direction for future work.

C Effect of Translation into English

To examine the extent to which language affects MLLM performance, we conducted an additional experiment in which all inputs were translated into English. Specifically, for each gesture instance, we translated the Japanese dialogue context, the reference description, and the instruction prompts (for description generation and evaluation) into English using GPT-4o-mini. We then prompted GPT-4o to generate gesture descriptions in English, using the same prompt structure as in the Japanese setting. Evaluation was performed using GPT-4o-mini, which assessed the generated English descriptions against the translated reference descriptions, using the same rubric as described in Section 3.3.

The results are summarized in Table 4. Interestingly, GPT-4o’s performance on indexical gestures improved in the English translation setting ($0.47 \rightarrow 0.56$). One possible explanation is that Japanese often omits subjects and relies heavily on context, making utterances and referents more ambiguous. When translated into English, these omissions are often filled in with explicit pronouns like “you” or “he,” reducing referential ambiguity. In contrast, the score for symbolic gestures declined in the English setting ($0.53 \rightarrow 0.44$), likely

Language	Indexical	Iconic	Symbolic
Japanese	0.47	0.50	0.53
English	0.56	0.52	0.44

Table 4: Comparison of gesture description accuracy in Japanese vs. translated English setting (GPT-4o).

due to cultural dependency and meanings being lost in translation. Symbolic gestures often presuppose culturally shared meanings that are implicitly understood within the original culture (e.g., placing hands together to express gratitude or prayer). Such associations are not always explicitly represented in English training corpora, making it difficult for models to infer the intended meaning when gestures are translated out of their cultural context. Note that symbolic gestures are underrepresented in our dataset ($n = 8$), making this category more susceptible to variance; the observed decline should therefore be interpreted with caution.

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。[Based on the following conversation and a series of images depicting the dialogue scene, explain the meaning of the gesture performed towards the end of the scene.]

会話文 : [Conversation :]

scA: きょ、星がキラキラ集まつのが、もっと実は、もっともっと外にあります。[The stars are actually gathered together in a sparkling way, and there are even more, much farther away.]

v01: もっと(す)。[More.]

scA: そうそう、もっと[Yes, yes, more]

v04: ハーーえ。[Oh.]

v01: もっとあるだって、りゅうせい。[There are even more meteors.]

scA: もっと、そう、もっといっぱいあるんです。[More, yes, there are many more.]

v03: (あんどろめだー;アンドロメダ)(のね) [Andromeda.]

scA: あのね、望遠鏡将来見る機会あるかもしません。[Well, in the future, (you) might have a chance to see them through a telescope.]

scA: あの[You know...]

ジェスチャーの説明: [Gesture Explanation :]

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示する動作です。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。[The gesture performed toward the end of this scene involves using a finger to point at something. This gesture is likely made to indicate a specific location or object in relation to the conversation. Specifically, given that the topic revolves around stars and celestial bodies, the gesture may be used to point at an exhibit or a celestial object displayed on a screen to supplement the explanation. Pointing gestures serve as an effective means to emphasize visual information and help the listener form a clearer mental image.]



Figure 3: Example prompt used for gesture description generation. Red text indicates variables that change per input instance, while blue text represents the actual output generated by GPT-4o. Each line has been translated into English, with the translation appended in brackets ('[', ']') at the end of each line.

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。[Based on the following conversation and a series of images depicting the dialogue scene, explain the meaning of the gesture performed towards the end of the scene.]

会話文 : [Conversation :]

{utterances}

ジェスチャーの物理的な観点からの説明: [Explanation from the physical perspective of the gesture:]

{physical_level_description}

ジェスチャーの説明: [Gesture Explanation :]

Figure 4: Example prompt used for gesture description generation with physical-level descriptions as additional context. The physical descriptions provide details on the hand movements associated with the gesture, aiming to enhance MLLMs' ability to interpret indexical gestures by incorporating motion-related cues.

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。[Based on the following conversation and a series of images depicting the dialogue scene, explain the meaning of the gesture performed towards the end of the scene.]

会話文: [Conversation:
{utterances}]

ジェスチャータイプ: [Gesture Type:
{gesture_type_label}]

ジェスチャーの説明: [Gesture Explanation:]

Figure 5: Example prompt used for gesture description generation with explicit gesture type labels as additional context. The gesture type label (Indexical, Iconic, or Symbolic) is provided to guide MLLMs in interpreting the gesture's communicative function more accurately.

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。*[The following is a pair of descriptions for a gesture. One is a description generated by a model, and the other is a ground truth description created manually. Evaluate the extent to which the generated description includes the meaning of the ground truth description. The evaluation should be given as a score between 0.00 and 1.00 based on the following criteria, along with a brief explanation of the reasoning.]*

評価基準: [Evaluation Criteria:]

- 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。[The generated description fully includes the meaning of the ground truth description.]
- 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。[The generated description includes most of the meaning of the ground truth description but lacks some details.]
- 0.50: 部分的に一致しているが、重要な情報が欠けている。[The descriptions partially match, but important information is missing.]
- 0.25: わずかな関連性があるが、大部分が一致していない。[There is slight relevance, but most of the meaning does not match.]
- 0.00: 全く意味が一致していない。[The meanings do not match at all.]

ジェスチャーのタイプ: 指標的 [Gesture Type: Indexical]

生成された説明文: [Generated Description:]

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。*[The gesture performed toward the end of this scene involves using a finger to point at something. This gesture is likely made to indicate a specific location or object in relation to the conversation. Specifically, given that the topic revolves around stars and celestial bodies, the gesture may be used to point at an exhibit or a celestial object displayed on a screen to supplement the explanation. Pointing gestures serve as an effective means to emphasize visual information and help the listener form a clearer mental image.]*

正解説明文: [Ground Truth Description:]

「望遠鏡将来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[The gesture indicates that the (omitted) subject of the utterance "(...) might have a chance to see through a telescope in the future" is v.]

評価スコア: 0.25 [Evaluation Score: 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（「望遠鏡将来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。*[Explanation: The generated description explains a pointing gesture but is only loosely related to the meaning of the ground truth description. While the generated description discusses the general function and purpose of pointing, it does not include the key information from the ground truth description—specifically, that the gesture is indicating the subject of the utterance. As a result, while there is slight relevance, most of the meaning does not match, leading to a score of 0.25.]*

Figure 6: Example prompt used for gesture description evaluation. Red text indicates variables that change per input instance, while blue text represents the actual output generated by GPT-4o. Each line has been translated into English, with the translation appended in brackets ('[,']') at the end of each line. Note: In the original Japanese utterance “(...) might have a chance to see through a telescope in the future,” the subject is omitted, which is a common phenomenon in Japanese when the subject can be inferred from context. In this case, the pointing gesture serves to disambiguate the subject of the utterance, without which the referent would remain unclear. In our dataset, “v” denotes a visitor in the dialogue. Each participant is labeled in the transcript, with identifiers such as “sc” for the science communicator and “v01”, “v02”, etc., for individual visitors.