Evaluation of Multilingual Ability to Use Spatial Deictic Expressions in Vision-Language Models

Kaito Watanabe[1,2], Taisei Yamamoto[1,2], Tomoki Doi[1,2], Hitomi Yanaka[1,2,3]

[1] The University of Tokyo, [2] Riken, [3] Tohoku University

{nglhdf,yamamo96,doi-tomoki701,hyanaka}@is.s.u-tokyo.ac.jp

Abstract

One of the expected abilities of vision-language models (VLMs) is spatial reasoning ability based on a given text and image. To evaluate the spatial reasoning abilities of VLMs, we focus on the use of spatial deictic expressions, which are defined as spatial expressions whose referent is determined by their situational context, such as "this" and "that". To handle spatial deictic expressions, VLMs must jointly reason over language and visual space, grounding context-dependent references in the image's spatial structure. In addition, selecting appropriate spatial deictic expressions across languages requires VLMs to understand the language-specific spatial distinctions encoded by these expressions. In this paper, we develop a benchmark[1] to evaluate the multilingual ability of VLMs to use spatial deictic expressions in four languages. Our experiments using this benchmark reveal that the tested models use demonstratives in a manner different from that of humans, particularly in selecting the appropriate demonstratives based on the distance to the object.

1 Introduction

In recent years, large language models (LLMs) and vision-language models (VLMs) have achieved remarkable progress due to their scalability. Furthermore, a key feature of LLMs and VLMs is their ability to handle a wide range of tasks not only in English but also in multiple languages. The development prompted vigorous attempts to evaluate their reasoning ability. Previous studies have investigated the abilities of VLMs to capture spatial relations and spatial expressions, such as frames of reference[2] (Zhang et al., 2025b; Khemlani et al., 2025).

Despite efforts, the ability of VLMs to utilize spatial deictic expressions, an important type of spatial expression, has not been explicitly studied. Deixis is the usage of expressions whose referent is dependent on the situation of utterance. Spatial deictic expressions are expressions that depend on the space in which the utterance occurred, such as "here" or "that". For example, suppose there is a pen in front of the speaker. If the pen is far away, especially at an unreachable point, the speaker tends to say "that pen", while if it is near the speaker, the speaker may indicate it by "this pen". As this example shows, deictic functions are fundamental to human language because they connect linguistic expressions with the physical environment. Therefore, the evaluation of a VLM's proficiency in using spatial deictic expressions has linguistic significance for benchmarking its capacity for spatial reasoning.

We argue that the use of spatial deixis poses two major challenges for VLMs: cross-linguistic variation and inherent ambiguity. First, to use spatial deictic expressions appropriately, multilingual VLMs need to understand their semantic differences across languages. For example, while in English we use two demonstratives, proximal [3] "this" and distal "that", in Japanese we use three demonstratives, proximal *kono*, distal *ano*, and medial *sono*. As such, across languages, the number of kinds of spatial deictic expressions varies. In addition, as in another example, although Spanish and Japanese both have three demonstratives: proximal, distal, and medial, the meaning of each word is not the same. According to Diessel (1999), Spanish medial *ese* signifies a relatively intermediate place between the speaker and the addressee, while Japanese medial *sono* indicates a place near the addressee. Therefore, multilingual VLMs need

[1] Our benchmark is available in https://github.com/ynklab/multilingual-demonstratives-eval

[2] Frames of reference(FoR) are frameworks which are used to express the relative position of an object from the perspective of the other object.

[3] Proximal demonstratives refer to the objects located closer to the speaker, while distal demonstratives refer to the objects located further.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1203–1211

to capture such subtle differences in the semantic nuances of spatial distances across languages to employ demonstratives in a manner similar to humans. Second, Coventry et al. (2023), a study investigating demonstrative usage among approximately 30 participants per language, reported that individuals do not necessarily use the same demonstrative in the same situation; there are individual differences in capturing the semantic nuances of spatial distance. Thus, if multilingual VLMs succeed in learning spatial deixis expressions, they are expected to recognize differences across languages and reproduce the distribution of human performance in the use of demonstratives.

Based on these points, we developed a benchmark to measure the ability to use spatial deictic expressions in various languages, such as English and Japanese, and evaluated VLMs. The task included in the benchmark is designed to investigate how the absolute distance of an object influences the choice of demonstratives of VLMs. We also analyzed the performance of VLMs by comparing the results of experiments on human spatial deictic expressions reported in Coventry et al. (2023).

The contributions of this paper are as follows:

1. We constructed the first benchmark to evaluate the ability to employ demonstratives in VLMs.

2. We revealed that open VLMs fail to use demonstratives in a human-like way, and the differences between humans vary across models. In particular, VLMs do not show the shift in the selection of demonstratives with distance, as observed in humans.

2 Background

2.1 Analysis of Spatial Reasoning Ability in VLMs

A large number of evaluations have been conducted on spatial reasoning ability in VLMs (Liu et al., 2025). Those previous benchmarking works revealed that VLMs still lack sufficient human-level ability (Khemlani et al., 2025; Zhang et al., 2025a), especially for 3D world spatial reasoning tasks such as distance estimation (Yang et al., 2025; Zhang et al., 2025a) and positional relationships (e.g. *front* or *left*) (Liu et al., 2023; Khemlani et al., 2025).

Moreover, many benchmarks were developed only in English (the works cited in the preceding paragraph are all for English evaluation). Only a few benchmarks contain tasks to evaluate the multilingual ability of spatial reasoning in VLMs (Liu et al., 2024; Haller et al., 2025; Zhang et al., 2025b).

Liu et al. (2024) aims to evaluate the ability of VLMs robustly and holistically. To realize this goal, the benchmark, MMBench, includes a task called "spatial relationship" that contains 2D spatial relationships, and "physical relation" that contains 3D spatial relationships. The authors created a Chinese version of the benchmark, called MMBench-CN. They demonstrated that most of the tested VLMs performed worse on MMBench-CN, but the gap in performance between English and Chinese for a model that achieves a high English score may be smaller.

Haller et al. (2025) constructed a benchmark called PISA-BENCH by collecting questions from the PISA test, an international assessment of the academic performance of students. It contains "spatial and geometric reasoning" tasks. As the name implies, it contains a task that asks for the shape of an object viewed from behind. As a result, they found a significant gap in performance between English and the others.

Zhang et al. (2025b) developed an evaluation protocol called COMFORT. The authors focused on frames of reference (FoR) to evaluate the robustness, consistency, and flexibility of VLMs' spatial reasoning. They constructed images of datasets using Blender (Blender Online Community, 2016). The benchmark asks VLMs to answer whether one object has an indicated relationship to the other, for example, "From the camera's viewpoint, is the ball behind the car?" using an image of a ball and a car. They concluded that VLMs lack the ability to use indicated FoR and have a bias to the English method of expression rather than a method specific to the language.

However, previous work does not focus on the extent to which VLMs handle spatial deictic expressions in spatial reasoning tasks. In this paper, we analyze multilingual abilities to use spatial deictic expressions in VLMs.

2.2 Spatial Deixis

2.2.1 Deixis and Demonstratives

Deixis is the usage of spatial expressions whose referent is dependent on the situation of utterance (Saito et al., 2015). Deictic expressions are actual expressions of deixis. For example, given

the sentence *John is here.*, the word *here* is a spatial deictic expression, so we cannot decide what the word *here* refers to without considering the situational context of the utterance. In this situation, the speaker doesn't need to explain the referent of *Here*. Instead, the listener has to determine the referent of *here*. In this way, the accurate recognition of the meaning of spatial deictic expressions requires precisely determining the referent of the word.

Deictic expressions are classified into several classes (Diessel, 1999; Fillmore, 1997), and spatial deictic expressions are one of them. Spatial deictic expressions are deictic expressions whose referent is related to space. In this paper, we focus on spatial deictic expressions, especially demonstratives.

Coventry et al. (2023) investigated how people with various native languages use different demonstratives. In their experiments, most languages have two or three demonstratives. The differences among demonstratives vary across languages, but, in general, they tend to depend on distance. Especially, a demonstrative used to refer to nearby and faraway objects from the speaker is called *proximal* and *distal*, respectively.

Some languages also have another criterion to differentiate demonstratives. For example, in Japanese language, *ano* and *sono* both refer to the object not near the speaker, but the difference depends on the distance from the addressee. Coventry et al. (2023) revealed that this kind of dependency on distance from the addressee is also observed in some languages, such as Finnish or Korean.

In order to output spatial deictic expressions appropriately, multilingual VLMs have to recognize the differences in demonstratives among languages.

2.2.2 Deixis and Language Models

To the best of our knowledge, whether VLMs can handle spatial deictic expressions has not been sufficiently explored. Existing attempts to evaluate the deixis understanding ability of language models are prone to focus on discourse deixis, a kind of deixis, because LLMs do not have access to external visual contextual information. A benchmark called PUB (Sravanthi et al., 2024) evaluates the pragmatic reasoning abilities of LLMs, including understanding of deixis. Their results show that almost all tested LLMs perform worse than humans, and the authors conclude that the poor scores are not due to world knowledge but to a lack of pragmatic reasoning. While several studies of deixis and language models exist in the field of human-computer interaction (Lai et al., 2025; Han and Isaacs, 2025), there are few studies of evaluation in vision-language models.

3 Method

To construct our benchmark, we refer to the linguistic experiments called "memory game" (Gudde et al., 2018). A detailed description of the "memory game" is provided as follows.

3.1 Memory Game

The memory game is a method to investigate the usage of demonstratives of humans without subjects noticing that linguistic investigation is conducted (Gudde et al., 2018)[4]. This method aims to elicit demonstratives from subjects of the experiment in experimentally controllable and naturalistic situations. To ensure the naturalistic setting, the experiment is presented as a memory experiment, and the purpose of investigating demonstrative usage is concealed from participants until the experiment is completed.

The procedure of the memory game in Gudde et al. (2018) is as follows. First, subjects sit on one side of a long desk and are told that the experiments examine the influence of language on memory. Then they are instructed to name the shape drawn on the disk, which is placed on the long desk, by pointing their finger at the disk. When naming the object, the subjects must use three words: a demonstrative, a color, and the shape of the object, like "that black cross". Experimenters record which demonstrative is used, and one trial of the experiment is done. Experiments are conducted multiple times with changes in some variables, such as the disk's location. The places of the disk are classified into three "regions" by distance, region 1 for 25-75 cm, region 2 for 100-150 cm, and region 3 for 175-225 cm. The result of the experiment showed that these parameters have an influence on demonstrative use.

Coventry et al. (2023) conducted a memory game across various languages, and they analyzed how the usages of demonstratives are different among languages. In their research, experiments were conducted with 29 languages (e.g., Japanese and English) and 874 subjects whose native language was one of them. The memory game experiment is conducted by using a white disk with a

[4]In the original paper (Gudde et al., 2018), another version of the memory game is also suggested, called "memory version".

shape drawn on it. The variables are places, shapes, colors, and the addressee's location. Places of the disk are classified into three "regions" by distance, like Gudde et al. (2018), but they set regions in a slightly different way: region 1 for 25-50 cm, region 2 for 150-175 cm, and region 3 for 275-300 cm. Shape is a shape drawn on the disk (e.g, a star, a cross, and a triangle). Color is the color of the shape. Through extensive experiments, they showed that native speakers use different demonstratives (distal, medial, and proximal) depending on the distance to the target. Furthermore, they found that the usage patterns of these demonstratives across distances vary across languages; for example, Japanese speakers tend to use medial demonstratives (*sono*) for objects at an intermediate distance, whereas Korean speakers use distal ones (*jeo*).

The memory game enables the evaluation of human demonstrative use across multiple languages, accounting for subtle ambiguities that arise depending on the distance between the speaker and the object. In this study, we construct a benchmark for evaluating demonstrative use in VLMs based on the memory game paradigm.

3.2 Task Setting

Based on the memory game, we composed a VQA task to analyze the extent to which VLMs capture spatial deixis across languages. We chose four languages for evaluation, Japanese, Korean, English, and Chinese, in which there were differences with respect to the usage of spatial deictic expressions. There exist three demonstratives in Japanese (proximal *kono*, distal *ano* and medial *sono*) and Korean (proximal *i*, distal *jeo* and medial *geu*), and two demonstratives in English (proximal *this*, and distal *that*) and Chinese (proximal *zhè ge*, and distal *nà ge*).

We follow the task setting used in the memory game. In each image (see Figure 1), one white disk with a colored shape is put on the black desk. VLMs are instructed to describe the shape drawn on the disk, using a fixed format consisting of a demonstrative, a color, and the shape, such as "this red circle."

3.3 Benchmark Construction

We constructed the datasets with Blender (Blender Online Community, 2016) following the construction process of previous benchmarks (Khemlani et al., 2025; Zhang et al., 2025b), because we can easily tweak the location of objects and the brightness with it. To be consistent with experimental settings regulated in Gudde et al. (2018), we produced images with the aligned settings. For example, we used a long desk without any patterns and did not place any objects around it, which can help to estimate the distance of the object. We also tweaked the camera height or the room brightness so that VLMs could obtain sufficient information to answer.

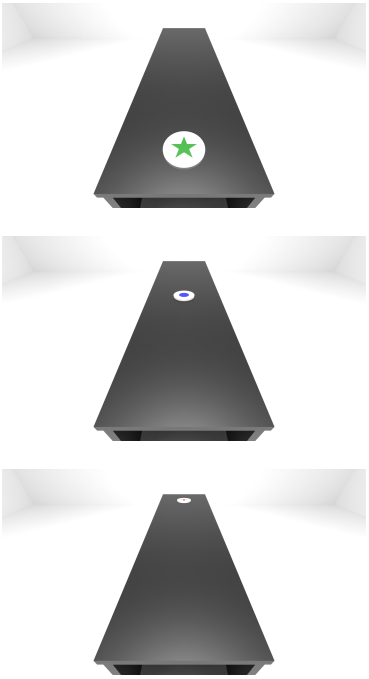

Figure 1: Examples of images included in the benchmark. Distance from the object is 0.25m, 1.50m, and 2.75m, from left to right.

Examples of the image included in the benchmark are displayed in Figure 1. On the desk, we put a disk with a solid-colored shape. The disk was placed at one of the distances 0.25 m, 1.50 m, and 2.75 m from the subject, as displayed in Figure 1, corresponding to region 1, 2, and 3 in Coventry et al. (2023) respectively. We used 5 figures (circle, cross, square, star, and triangle) and 4 colors (black, blue, green, and red) to create 20 images per region, resulting in 60 images in total (5 shapes $times$4 colors for the shape $\times$ 3 positions for the disk).

4 Experimental Setting

4.1 Model

We chose Gemma 3 4B [5] and Gemma 3 12B [6] from Gemma 3 series (Gemma Team, 2025), and Qwen3-

[5] https://huggingface.co/google/gemma-3-4b-it
[6] https://huggingface.co/google/gemma-3-12b-it

VL 8B [7] and Qwen3-VL 32B [8] from Qwen3-VL series (Bai et al., 2025) for evaluation. All of them are open, instruction-tuned, and multilingual VLMs. We avoided closed models such as GPT-5.2 because we cannot obtain raw logits for analysis from their APIs. We accessed trained models of them through Hugging Face Hub[9]. Output token decoding was performed using a greedy algorithm for all models.

4.2 Prompt

Since the original instructions used in Coventry et al. (2023) experiments are not publicly available, we design prompts for VLMs following the protocol described in Coventry et al. (2023). We configured a prompt to have VLMs output the same format as the memory game for humans: a demonstrative, the color of the object, and a shape drawn on the disk placed on the long desk. The prompt used in our experiments is shown below.

> Prompt: Analyze the image and identify the shape on the disk. Describe it by filling in the following three-word template exactly:
> [Demonstrative] [Color] [Shape]
> Constraints:
> Use "This" or "That" for the demonstrative.
> Use a single word for the color.
> Use a single word for the shape.
> Output only the three words. Do not include a period or any introductory text.

This prompt was written in English at first, and then translated into each language by Google Gemini[10]. After the translation, we added lacking demonstratives, especially medial demonstratives of Japanese and Korean, because the literal translation of the prompt contained only proximal and distal demonstratives.

4.3 Metrics

We primarily employ two metrics. One is a probability distribution of the use of demonstratives for each VLM. The other is the Jensen-Shannon distance between the probability distributions of demonstratives by models and humans. This allows us to analyze how precisely models reproduce human usage of the spatial deictic expressions.

The probability distribution of the use of demonstratives is calculated from the logits obtained from outputs of LLMs. We first calculate the logits of all target demonstratives, and then normalize them by applying the softmax function. For example, we calculate the probability for "this" in English as $P(this)/(P(this) + P(that))$.

The fidelity of humans' selection of demonstratives is calculated referring to the probability distribution of humans reported on Coventry et al. (2023). We regarded the results of Coventry et al. (2023), which were conducted for humans, as the probability distribution of demonstratives inherent in each language. We measured the Jensen-Shannon distance between the distributions of humans and VLMs. We adopted the Jensen-Shannon distance because it can measure the distance for a distribution that has a class whose probability is 0.

We calculated the Jensen-Shannon distance using the following formula:

$$JSD(P||Q) = \sqrt{\frac{D_{KL}(P||M) + D_{KL}(Q||M)}{2}} \tag{1}$$

where P and Q is a probability distribution, M is $\frac{1}{2}(P + Q)$ (pointwise mean), and $D_{KL}(P||M)$ denotes the Kullback-Leibler divergence from M to P, defined as:

$$D_{KL}(P||M) = \sum_{d} P(d) \log \frac{P(d)}{M(d)} \tag{2}$$

for each d in demonstratives of the target language.

We compute the Jensen–Shannon distance between the distribution of demonstrative usage by the VLM and that by humans for each distance to the target, and compute a representative value for the model by averaging these distances across target distances.

Before the evaluation, we eliminated the inappropriate results that do not align with the format "[Demonstrative] [Color] [Shape]" and made some mistakes on colors and shapes, because VLMs fail to understand our instructions or to recognize the indicated object in those trials. As colors and shapes can be expressed in many ways (e.g. "cross" and "plus" for a cross shape), we permitted those varieties of expressions during the validation. The

[7] https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
[8] https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct
[9] https://huggingface.co/
[10] https://gemini.google.com/

judgment to determine whether the output corresponds to the synonym of the gold answer was performed manually by the author.

5 Results and Discussion

5.1 Results and Analysis

Models	Japanese	Korean	English	Chinese
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

Table 1: The number of images in which the color and shape of the object are precisely recognized by the model (out of 60 images).

The number of images whose color and shape the model could precisely recognize is presented in Table 1, and the probability distributions of demonstratives for each VLM per language and distance settings are shown in Figure 2. The probability distributions shown in Figure 2 are calculated based only on the cases where images were correctly recognized by models, as presented in Table 1. Each row of Figure 2 corresponds to an experimental setting and presents the average probability distribution of demonstrative for that setting. The bottom row is the result of the human investigations reported in Coventry et al. (2023). Green, orange, and light blue represent the probabilities of proximal, medial, and distal demonstratives, respectively.

Across models, a general trend is to avoid distal demonstratives in languages with three demonstratives, regardless of the model. Although the proportion of distal demonstratives increased as the distance from the object increased for humans, the proportions of distal demonstratives in Korean and Japanese, which have three demonstratives, are consistently under 5% except for the case of Qwen3-VL 32B for Japanese language. Due to unbalanced distributions, models often fail to select the same demonstratives as humans do, especially in Japanese and Korean.

Regarding the trends for each model, Gemma 3 exhibits similar probability distributions across distances. Although the proportion of proximal demonstratives decreases and that of distal demonstratives increases as the distance increases for humans, changes in the probability distribution of both Gemma 3 4B and Gemma 3 12B across different distances are smaller than the results of humans. The Qwen3-VL series does not show such irrelevance to distance, but it fails to capture the human-like probability distribution.

Table 1 presents that VLMs could not recognize the figures painted on the object and the content of the prompt in some experimental settings. This phenomenon is particularly prominent in Gemma 3 4B, which exhibits biased distributions in demonstrative usage. In contrast, Qwen3-VL achieves higher recognition accuracies than Gemma 3 models across all languages. This may be caused by differences in the pre-training dataset. While the pre-training dataset of Gemma 3 is not disclosed, that of Qwen3-VL includes tasks of visual question-answering with camera-captured images (Bai et al., 2025). The fact that Qwen3-VL was pre-trained with similar tasks to ours may cause the more accurate recognition of the object in Qwen3-VL. We can also observe that with Qwen3-VL 8B and 32B, English and Chinese scores of accurate recognition are higher than Japanese and Korean ones. This can also be explained by the fact that it was pre-trained with data in which English and Chinese account for a large proportion (Bai et al., 2025).

It is also observable that while human preference for demonstratives varies according to distance, VLMs do not exhibit such sensitivity. For instance, in Japanese, the most frequently selected demonstratives for humans shifted from proximal *kono* at 0.25 m to medial *sono* at 1.50 m, and finally to distal *ano* at 2.75 m. But Qwen3-VL 32B chose the proximal demonstrative at the highest proportion regardless of distance, and the other three models chose the medial at the highest proportion regardless of distance. This lack of the shift of demonstratives by distance may be a feature of VLMs.

5.2 Distance from Human Distribution

Jensen-Shannon distances between the outputs of models and humans' results of Coventry et al. (2023) are presented in Table 2. For comparison, the Jensen-Shannon distribution between the human distribution and the uniform distribution, in which every demonstrative is chosen with the same probability, is provided. Diagonal lines in the table indicate that we cannot calculate the distance due to insufficient data resulting from a violation of the output format.

Table 2 shows Gemma 3 4B performs more differently from human distribution than uniform distribution in any language. In contrast, Qwen3-

Figure 2: Probability distributions across distances for each experimental setting. The red horizontal line separates results for VLMs (above) from those for humans (below). Columns represent different languages, while rows above the red horizontal line correspond to specific VLMs. Human results in the bottom row were calculated based on the results in Coventry et al. (2023).

VL 32B shows a similar distribution to humans in Japanese and Chinese, therefore, we can observe that Qwen3-VL 32B uses demonstratives in a relatively close way to humans.

We can also observe that the average distance in English is consistently smaller than that in Korean. This can be explained in two ways. First, Korean is a relatively low-resource language compared to English, and the shortage of training resources for Korean demonstratives may lead to low fidelity to humans' distribution. Second, the probability distribution of the use of demonstratives in English is closer to a uniform distribution than that in Korean. Thus, it is a relatively easier task to answer in a closer distribution in English than in Korean.

These results suggest that the lack of general object recognition ability in VLMs may cause biased demonstrative usage, even when color and shape are accurately recognized at evaluation time. To further analyze the relationship between model object recognition ability and demonstrative usage, we calculated the correlation between image recognition ability and the distance of the probability distribution across 15 experimental settings, encompassing various combinations of languages and model architectures. The analysis yielded a Pearson correlation coefficient of $r = -0.40$, indicating a modest negative trend between the two metrics. This result suggests that models with higher recognition proficiency tended to exhibit demonstrative usage distributions more similar to those of humans.

6 Conclusion

In this paper, we developed a multilingual benchmark to assess the extent to which VLMs can use spatial deictic expressions in a manner similar to humans based on the memory game paradigm. Using our benchmark, we investigated the probability distributions of demonstratives by object distance.

We discovered that all tested models fail to reproduce humans' probability distribution for demonstratives as a function of the absolute distance from the viewpoint to the target object. We also observed that, unlike humans, some VLMs tend not to adjust their probability distributions based on distance to the referent. These results suggest that current VLMs have significant room for improvement in their use of demonstratives in a human-like

Models	Distance(m)	Japanese	Korean	English	Chinese
Gemma 3 4B	0.25	0.7227	0.9554	0.2615	0.1597
	1.50	0.4591	0.8349	0.6830	0.6571
	2.75	0.8252	0.8880	0.7174	\
	Average	0.6080	0.8705	0.5041	\
Gemma 3 12B	0.25	0.5732	0.2067	0.3167	0.1557
	1.50	0.5104	0.9173	0.7641	0.5839
	2.75	0.7904	0.9367	0.8252	0.7018
	Average	0.5815	0.7313	0.6326	0.5063
Qwen3-VL 8B	0.25	0.6749	0.9849	0.3677	0.1495
	1.50	0.4821	0.8771	0.0799	0.8221
	2.75	0.8510	0.9143	0.0413	0.9437
	Average	0.6648	0.9200	0.1539	0.6499
Qwen3-VL 32B	0.25	0.1294	0.1638	0.2799	0.8175
	1.50	0.8424	0.9130	0.4940	0.2277
	2.75	0.7489	0.9247	0.5462	0.0510
	Average	0.5055	0.7446	0.4444	0.3725
Uniform distribution	0.25	0.5740	0.5869	0.1962	0.4626
	1.50	0.3907	0.5186	0.2998	0.3216
	2.75	0.5149	0.5551	0.3912	0.4791
	Average	0.4932	0.5535	0.2957	0.4211

Table 2: Distance from human distribution of demonstrative use per language.

way. We also clarified the difference in performance among languages. In particular, all tested models show weaknesses in Korean language. We also found that VLMs seldom use distal demonstratives in Japanese and Korean.

The results described above suggest that there are differences between VLMs and humans in handling spatial deictic expression, which is one of the fundamental expressions for spatial understanding. We believe that our benchmark serves as a new testbed for evaluating the spatial reasoning abilities of VLMs across languages.

7 Limitations

Though this benchmark introduces a novel methodology to evaluate the ability to use spatial deictic expressions, it is subject to several limitations. First, the models cannot recognize a number of pictures, resulting in a constrained number of samples for evaluation. Furthermore, the scope of our experimental setup is currently restricted. Since our evaluation relies on a single prompt, it is difficult to rigorously determine whether the observed trends reflect the models' intrinsic capabilities or are artifacts of the prompt's phrasing. The generalizability of these results is also constrained by the limited size of the dataset used and the limited number of languages and model architectures evaluated in this study. Increasing the dataset size in future studies is required to mitigate potential biases and ensure a more robust assessment of the results. In addition, since we needed to use a controlled prompt setting to ensure the VLM outputs satisfy the expected format (demonstrative, color, and shape), the variety of usage of demonstratives in natural situations is not sufficiently reflected. Establishing more naturalistic settings remains as future work. Finally, the evaluation process involved human judgment to verify model outputs. It introduces an element of subjectivity and limits the benchmark's scalability.

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR2565, Japan. We would also like to thank Anirudh Reddy Kondapally for his valuable comments and suggestions. The authors would also like to thank the anonymous reviewers for their constructive feedback, which helped to improve the quality of this paper.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, and 46 others. 2025. Qwen3-vl technical report. *ArXiv*, abs/2511.21631.

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

Kenny R Coventry, Harmen B Gudde, Holger Diessel, Jacqueline Collier, Pedro Guijarro-Fuentes, Mila Vulchanova, Valentin Vulchanov, Emanuela Todisco, Maria Reile, Merlijn Breunesse, and 1 others. 2023.

Spatial communication systems across languages reflect universal action constraints. *Nature human behaviour*, 7(12):2099–2110.

Holger Diessel. 1999. *Demonstratives : form, function, and grammaticalization*. Number v. 42 in Typological studies in language. J. Benjamins.

C.J. Fillmore. 1997. *Lectures on Deixis*. Center for the Study of Language and Information Publication Lecture Notes. Cambridge University Press.

Gemma Team. 2025. Gemma 3.

Harmen B Gudde, Debra Griffiths, and Kenny R Coventry. 2018. The (spatial) memory game: testing the relationship between spatial language, object knowledge, and spatial cognition. *Journal of Visualized Experiments: JoVE*, (132):56495.

Patrick Haller, Fabio Barth, Jonas Golde, Georg Rehm, and Alan Akbik. 2025. Pisa-bench: The pisa index as a multilingual and multimodal metric for the evaluation of vision-language models. *Preprint*, arXiv:2510.24792.

Chang Han and Katherine E. Isaacs. 2025. A deixis-centered approach for documenting remote synchronous communication around data visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):930–940.

Sangeet Khemlani, Tyler Tran, Nathaniel Gyory, Anthony M. Harrison, Wallace E. Lawson, Ravenna Thielstrom, Hunter Thompson, Taaren Singh, and J. Gregory Trafton. 2025. Vision language models are unreliable at trivial spatial cognition. *Preprint*, arXiv:2504.16061.

Yuzhi Lai, Shenghai Yuan, Youssef Nassar, Mingyu Fan, Atmaraaj Gopal, Arihiro Yorita, Naoyuki Kubota, and Matthias Rätsch. 2025. Natural multimodal fusion-based human–robot interaction: Application with voice and deictic posture via large language model. *IEEE Robotics & Automation Magazine*, pages 2–11.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. 2025. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods. *Preprint*, arXiv:2511.15722.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yoshio Saito, Yoshihisa Taguchi, and Yoshiki Nishimura. 2015. *[The Sanseido Dictionary of Linguistics] Meikai gengogaku jiten*. Sanseido.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. 2025a. SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11591–11609, Vienna, Austria. Association for Computational Linguistics.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025b. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*.