Pointing Gesture Understanding via Visual Prompting and Visual Question Answering for Interactive Robot Navigation

Kosei Tanada[1], Shigemichi Matsuzaki[1], Kazuhito Tanaka[1],
Shintaro Nakaoka[1], Yuki Kondo[1], and Yuto Mori[1]

***Abstract*— In this paper, we explore a method of visual robot navigation that interprets human's gesture pointing toward desired directions and moves following the instructions with Vision Language Models (VLMs). In this method, we provide rating scales for Visual Question Answering (VQA) in visual or text prompts to VLMs to measure ambiguous pointing gestures. A VLM takes prefix texts and an observation image of a human's pointing with visual prompts and outputs the pointing scale that can be utilized for robot navigation. We validate two gesture rating scales and three visual clues with a pointing gesture dataset. The results demonstrate the difficulty of reliably accomplishing the targeted tasks and show the future direction of our research.**

I. Introduction

Large Language Models (LLMs) and Vision-Language Models (VLMs) have made significant progress and demonstrate broad capabilities using their adaptability via in-context learning in a variety of robotic tasks. The existing work has shown that LLMs and VLMs can ground robot-specific concepts from text instructions and visual representations, such as map or scene graph constructions and scene navigation [1]. However, an interpretation of interactive visual instruction has not been broadly considered, remaining a challenge to understand ambiguous spatial instructions, such as pointing gestures. In this study, we seek effective methods to exploit common knowledge of VLMs to infer the pointing direction with text and visual prompts.

Our ultimate goal is to achieve a general visual navigation policy that translates spatial and semantic cues, such as human gestures and arrow signs, into robot actions. We believe a combination of low-level control policy and high-level decision-making is effective. As an initial step towards this goal, we explore how to help the VLMs understand ambiguous gesture instructions of humans and generate plausible high-level reasoning, enabling intuitive and interactive robot control by users.

A challenge for this goal is quantifying ambiguous pointing directions and making it understandable for VLMs. To this challenge, we search for prompting methods to give discrete degrees that enable the evaluation of spatial pointing. We introduce the discrete scales to measure the distinctiveness of pointing gestures by text or text and visual prompting. We examine two types of rating scales with the state-of-the-art (SOTA) VLMs and investigate their performances.

Recent studies show that visual prompting enhances not only 2D scene understanding [2] but also the spatial reasoning capabilities of VLMs [3]. Inspired by those findings, we introduce some visual clues that represent abstracted information of human's pointing in a given image.

In this paper, we validate two gesture scales and further explore three visual prompts on one of them. We use text prompts with visual question answering (VQA) and some visual prompts to express gesture scales and visual clues. We demonstrate success rates and failure cases in our experiment with SOTA VLMs and provide future directions to our ultimate goal.

[1]Authors are with Frontier Research Center, Toyota Motor Corporation, Japan. kosei_tanada@mail.toyota.co.jp

II. Related Work

A. Large Language Models and Vision-Language Models

Large Language Models (LLMs) and Vision-Language Models (VLMs) have achieved significant progress [4], [5], and adapt to various domains and real-world applications, such as an embodied agent [6] and scene graph understanding [7]. However, more advanced capabilities, such as spatial gesture understanding in a zero-shot manner are still challenging. Existing work has shown that text [8] and visual [2], [9] prompting techniques encourage LLMs or VLMs to improve their reasoning performance. Inspired by these studies, we investigate what types of visual and text prompting are effective in understanding spatial pointing gestures.

B. Robot Navigation with Foundation Models

Recently, machine learning models trained with a variety and huge amount of data, called *Foundation Models*, are introduced to various robotic navigation from high-level reasoning [1], [10], [11] to low-level control[12]. Some pieces of existing work apply the general knowledge of LLMs and VLMs to robot navigation [1]. Recently, Nasiriany et al. [3] proposed a prompting method that applies iterative optimization of visual representation pre-processed on the image, and showed that their approach enables general robotic control including navigation in a zero-shot manner. While these methods reveal that SOTA VLMs contain rich knowledge for robotic tasks, effective visual prompting is still less explored, especially for the interpretation of visual semantics in robot navigation.

C. Gesture-based robot control

Gesture-based robot control has long been investigated [13], [14]. Lin et al. [15] proposed a framework to interpret gestures and language instructions with LLMs in the table-top manipulation setting. Cuan et al. [16] proposed

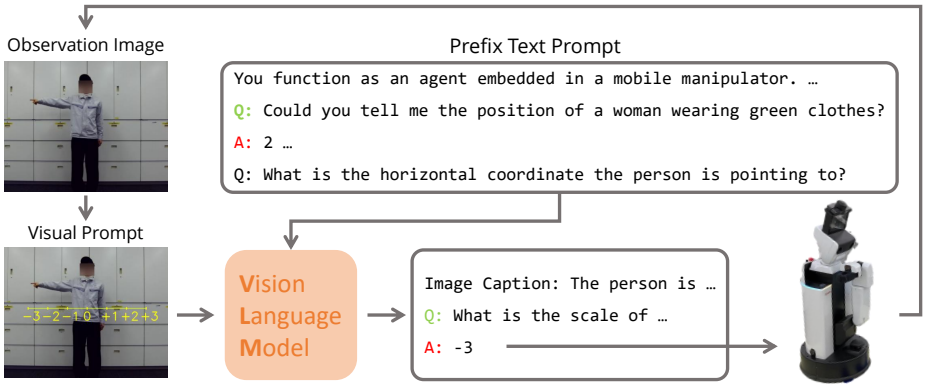

Fig. 1. Overview of the proposed method

an imitation learning method to train gesture-aware social navigation policies. In contrast, we aim to explore the zero-shot capabilities of VLMs to handle various spatially and semantically meaningful gestures.

III. Methodology

A. Overall architecture

In this paper, we consider the problem of understanding visual instruction given by human pointing gestures using VLMs. Fig. 1 shows the overview of our method. Formally, a VLM $\mathcal{F}$ takes a text sequence T^i and an image $I^i \in \mathbb{R}^{H\times W\times 3}$, and outputs text sequences T^o.

$$T^o = \mathcal{F}(I^i, T^i). \tag{1}$$

In our method, T^i is given by a prefix text prompt $\mathrm{p}_{\mathrm{pre}}$. $\mathrm{p}_{\mathrm{pre}}$ contains a role p_{r} of the embodied agent, an output example p_{e}, and a request to complete contexts composed of caption part and VQA part. I^i is given by an observation image o_t with a human representing visual instruction with a pointing gesture $g^i \in G$ at the current time t . Here, $o_t = f(g^i)$, where f is a function embedded in a camera sensor. The camera captures the scene in front of a robot, and f maps the scene with a human's pointing g^i to RGB image o_t. In our method, Eq. 1 is defined as follows:

$$T^o = \mathcal{F}(f(g_i), \mathrm{p}_{\mathrm{pre}}) \tag{2}$$

To provide rating scales of the pointing, we use additional visual and text prompts described in the next subsection.

B. Visual and Text Prompting for Gesture Understanding.

1) Rating Scales: We introduce two types of rating scales, *Ordinal Scale* and *Numerical Scale*.

(i) Ordinal Scale (OS) represents how gesture instruction is strongly indicated. The scale is given by a text prompt as follows:

```
The following three scales are available when you
    answer "Yes" for each question:
1. Strongly True
2. True
3. Weekly True
Select a scale after answering "Yes" for the given
    question.
```

This metric is similar to Likert scale [17], which generally has five or seven points, while we set a 3-point ordinal scale to quantify a degree of the answer "Yes". The ordinal scale is intended to recognize the pointing direction while keeping the ambiguity of the instruction to ensure the safety of subsequent behaviors of the robot.

(ii) Numerical Scale (NS) is given by text and visual prompts. Fig. 2 shows visual prompts providing numerical scales on an image. We use three types of numerical scales, axes (Fig. 2(a)) and horizontal scales (Fig. 2(b), 2(c)). The axes prompt overlays a 2.5D axis that represents a 2D coordinate with a depth axis with RGB colors. The axes prompt is inspired by [9], where they investigate the effectiveness of 3DAxiesPrompts on the objects for spatial reasoning tasks. Unlike [9], we use the 2D yz coordinate with the depth axis x at the origin of the coordinate, which needs less human annotation than a 3D axis along the object structure. The horizontal scale prompts show a scale with a range of either $[0, 6]$ or $[-3, 3]$ on the image. These approaches are less informative for 3D than the axes prompt, while they express minimum measures to measure horizontal pointing. The axes and horizontal scale prompts are also described by corresponding text prompts as well. For example, the axes prompt is explained as follows:

```
A human indicates a direction to move in the given
    image by a hand.
The red x-axis goes along the depth direction of
    the image.
The green y-axis extends across the left side of
    the image.
Do not consider the direction from the person's
    perspective.
The blue z-axis represents the height of the image.
```

To overlay a numerical scale on the image, we use an open-vocabulary detector [18] and get a "human face" position on the image. The origin or the center of the scale is determined by the center of x-pixel position of the human face.

2) Visual Clues for Pointing Gestures: Fig. 3 shows all visual clues that we explore in this paper. We investigate three types of visual prompts, visual controller, hand detection, and human pose detection that give clues to understand the ambiguous pointing gesture. *(i) Visual controller* shows triangles

(a) *Axes.* (b) *Scale* $[-3, 3]$ (c) *Scale* $[0, 6]$

Fig. 2. Visual prompting for numerical rating scales.

(a) *Controller.* (b) *Human hand.* (c) *Human pose.*

Fig. 3. Visual clues for pointing gestures.

and texts that represent the right and left direction in camera perspective, helping VLMs understand the concept of the directions. *(ii) Hand* and *(iii) human pose detection* overlay a detected hand or human pose on the image. Humans easily infer the pointed direction so that they identify where to pay attention (e.g. a directed hand or a rough representation of the gesture). Based on this consideration, we seek how VLMs can enhance their capabilities to understand spatial pointing with abstract representations of the body parts. These three visual clues are indicated by text prompts as well. For example, the human pose prompt is explained as follows:

```
A human indicates a direction to move in the given
    image by a hand.
The results of human pose detection are overwritten
     on the image.
```

We use open-vocabulary detector [18] to get the "human hand" position, and MediaPipe [19] to detect the human pose.

IV. Experiments

A. Setup

We test our approach to verify the inference performance with rating scales and visual clues described in Sec. III. We examine one ordinal scale and three numerical scales shown in Sec. III-B.1. To evaluate these methods, we collect images of a person making a pointing gesture in three office environments and apply each visual pre-processing to the images. We use ZED Mini (CM429) to capture images with a resolution of 1920×1080. We process raw RGB images with visual clues to validate the ordinal scale and visual prompts to validate the numerical scales. To check the effectiveness of the visual clues, we also use raw RGB images for the evaluation of the ordinal scale. To avoid confusion caused by visual information overload, we do not adapt the visual clues to the numerical scales. We use GPT-4V [4] and Gemini Pro [5] that are broadly utilized as SOTA VLMs. We define cases of "successful inference" for each rating scale as follows:

- For the ordinary scale, we make human-annotated image-and-text pairs that serve as the ground truth for inference results.

TABLE I

Success rates of OS and VC

Visual Clue	Success Rate	
	GPT	Gemini
None	21.7%	32.2%
Hand	32.8%	26.7%
Pose	**33.3%**	25.6%
Controller	27.8%	27.2%

TABLE II

Success rates of NS

Numerical Scale	Success Rate	
	GPT	Gemini
Axes	31.0%	43.0%
Scale $[-3, 3]$	**60.0%**	40.0%
Scale $[0, 6]$	6.6%	44.8%

- For the numerical scale, we set that the correct scale is the nearest scale of minimum and maximum x-pixel value derived from human-hand detection. For example, if the minimum and maximum x-pixel value of a pointing hand are the nearest to scales 4 and 5, the correct scales are set to 4 and 5. We use an open-vocabulary detector [18] to obtain the area of the x-pixel of a pointing hand.

B. Results

1) Gesture scale validations: Table I shows success rates of inference with an ordinal scale and visual clues. Visual prompt overlaying pose detection gets the highest success rate (33.3%) with GPT-4V. The visual representation that abstracts a human's body state can help VLMs understand the pointing gesture. This perspective can be applied to hand detection as well, where it relatively shows a higher success rate (32.8%). However, we do not detect reliable success rates for a real-world robot application with each VLM and visual clue.

Table II shows inference success rates with numerical scales. The scale $[-3, 3]$ gets the highest success rate (60.0%). Compared with the axes prompt, since each scale from -3 to $+3$ is clearly expressed in the image, the scale $[-3, 3]$ provides understandable marks that help to clarify where the finger is pointing. In contrast, the scale $[0, 6]$ gets the lowest score with GPT-4V. We assume it is easier for VLMs to connect the concepts of left/right and $+/-$.

2) Failure case analysis: We further investigate the failure cases to understand how VLMs capture the pointing gesture in the image. We define two types of failure cases, *wrong direction* and *wrong labels*. (i) *Wrong direction* means that the direction indicated by the response is opposite to the actual direction. (ii) *Wrong label* shows the direction in the response is correct, while the scale in the response is different from the labeled scale. Tables III, IV show the result of the analysis. While both cases are similarly observed for the ordinal scale, *wrong direction* is detected more than *wrong label* ones about numerical scales.

C. Discussions

The validations with the VLMs commonly revealed that neither the ordinal nor the numerical scale is performing adequately for real-world robotic applications. GPT-4V got a higher score in the experiment of the ordinary scale than Gemini Pro, while Gemini Pro is more promising in some results of the numerical scales, such as the axes prompt.

TABLE III
FAILURE CASES OF AS AND VC

Visual Clues	Failure Cases			
	GPT		Gemini	
	WD	WL	WD	WL
Hand	43.6%	56.4%	51.1%	48.9%
Pose	52.5%	47.5%	40.3%	59.7%
Controller	46.2%	53.8%	32.1%	67.9%

TABLE IV
FAILURE CASES WITH NS

Numerical Scale	Failure Cases			
	GPT		Gemini	
	WD	WL	WD	WL
Axes	65.0%	25.0%	88.2%	11.8%
Scale $[-3, 3]$	66.7%	33.3%	75.9%	24.1%
Scale $[0, 6]$	88.2%	11.8%	72.9%	27.1%

However, we found that all approaches were far from a safe and reliable application. The results suggest that the SOTA VLMs do not adapt their latent knowledge to the pointing direction or the discrete scale interpretation in a zero-shot manner. The result of failure case analysis demonstrates that half and more misunderstanding cases are derived from *wrong direction* ones. This means that VLMs output opposite directions with less consideration of the gesture instruction, which can have less interactive navigation in real-world situations.

Qualitatively, we observed that GPT-4V understood ambiguous metrics better than Gemini Pro, as shown in Table III, while the scores of an ordinal scale were relatively unstable for numerical representation in the image. In contrast, Gemini Pro's outputs depend on where the person is pointing. If the person on the image points to the right side of the image, the output of Gemini is relatively reliable, while if the person is pointing to the left side of the image, the answer is relatively incorrect. This aspect affects the result of failure cases with "wrong direction" that most of them are derived from the images pointing to the left side.

V. CONCLUSIONS AND FUTURE WORK

We explored the effectiveness of rating scales and visual clues to understand human's pointing discretely, enabling intuitive control of the visual navigation of the robot. The preliminary results showed that it is still challenging for SOTA VLMs to understand human pointing gestures with visual discrete scales and VQA-based text prompts.

Our ultimate goal is a visual navigation policy that can interpret more diverse visual semantic cues, such as signs with arrows. We expect that VLMs can handle such arbitrary visual information and that this capability will be a crucial direction of our approach. Our future work is to enhance the capability of high-level decision-making to a variety of semantic information using common sense knowledge of VLMs and connect it to the low-level control policy, such as NoMaD[20], enabling robots to plan where to go and understand the environments more interactively.

ACKNOWLEDGMENT

We thank Masayuki Masuda, Kazuhiro Shintani, and Hiroshi Bito for fruitful discussions and support.

REFERENCES

[1] D. Shah, M. Equi, B. Osinski, F. Xia, B. Ichter, and S. Levine, "Navigation with Large Language Models : Semantic Guesswork as a Heuristic for Planning," in *Conference on Robot Learning*, 2023, pp. 1–11.

[2] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," 2023.

[3] S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, Q. Vuong, T. Zhang, T.-W. E. Lee, K.-H. Lee, P. Xu, S. Kirmani, Y. Zhu, A. Zeng, K. Hausman, N. Heess, C. Finn, S. Levine, and B. Ichter, "Pivot: Iterative visual prompting elicits actionable knowledge for vlms," 2024.

[4] OpenAI, "Gpt-4v(ision) system card," 2023. [Online]. Available: https://cdn.openai.com/papers/GPTVSystemCard.pdf

[5] G. Team *et al.*, "Gemini: A family of highly capable multimodal models," 2023.

[6] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," 2023.

[7] K. Rana, J. Haviland, S. Garg, J. Abou-Chakra, I. Reid, and N. Suenderhauf, "Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning," 2023.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[9] D. Liu, X. Dong, R. Zhang, X. Luo, P. Gao, X. Huang, Y. Gong, and Z. Wang, "3daxiesprompts: Unleashing the 3d spatial task capabilities of gpt-4v," 2023.

[10] D. Shah, B. Osinski, B. Ichter, and S. Levine, "LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action," in *Conference on Robot Learning*. PMLR, 2022. [Online]. Available: http://arxiv.org/abs/2207.04429

[11] H. Biggie, A. N. Mopidevi, D. Woods, and C. Heckman, "Tell Me Where to Go: A Composable Framework for Context-Aware Embodied Robot Navigation," in *Conference on Robot Learning*, 2023, pp. 1–27. [Online]. Available: http://arxiv.org/abs/2306.09523

[12] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine, "ViNT: A Foundation Model for Visual Navigation," in *Conference on Robot Learning*, 2023. [Online]. Available: http://arxiv.org/abs/2306.14846

[13] P. Rybski and R. Voyles, "Interactive task training of a mobile robot through human gesture recognition," in *IEEE International Conference on Robotics and Automation*, vol. 1, no. May. IEEE, 1999, pp. 664–669. [Online]. Available: http://ieeexplore.ieee.org/document/770051/

[14] R. Stiefelhagen, C. Fogen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2004, pp. 2422–2427. [Online]. Available: http://ieeexplore.ieee.org/document/1389771/

[15] L.-H. Lin, Y. Cui, Y. Hao, F. Xia, and D. Sadigh, "Gesture-informed robot assistance via foundation models," 2023.

[16] C. Cuan, E. Lee, E. Fisher, A. Francis, L. Takayama, T. Zhang, A. Toshev, and S. Pirk, "Gesture2Path: Imitation Learning for Gesture-aware Navigation," Tech. Rep., 2022. [Online]. Available: http://arxiv.org/abs/2209.09375

[17] R. Likert, *A Technique for the Measurement of Attitudes*, ser. A Technique for the Measurement of Attitudes. Archives of Psychology, 1932, no. nos. 136-165. [Online]. Available: https://books.google.co.jp/books?id=9rotAAAAYAAJ

[18] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, "Simple open-vocabulary object detection with vision transformers," 2022.

[19] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," 2019.

[20] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "Nomad: Goal masked diffusion policies for navigation and exploration," 2023.