Evaluating VLMs' Spatial Reasoning Over Robot Motion: A Step Towards Robot Planning with Motion Preferences

Wenxi Wu[1] **Jingjing Zhang**[2] **Martim Brandão**[1]

[1]King's College London [2]University College London

ABSTRACT

Understanding user instructions and object spatial relations in surrounding environments is crucial for intelligent robot systems to assist humans in various tasks. The natural language and spatial reasoning capabilities of Vision-Language Models (VLMs) have the potential to enhance the generalization of robot planners on new tasks, objects, and motion specifications. While foundation models have been applied to task planning, it is still unclear the degree to which they have the capability of spatial reasoning required to enforce user preferences or constraints on motion, such as desired distances from objects, topological properties, or motion style preferences. In this paper, we evaluate the capability of 3 state-of-the-art VLMs, with 4 model variants in total, at spatial reasoning over robot motion, using four different querying methods. Our results show that, with the highest-performing querying method, Qwen2.5-VL achieves 71.4% accuracy zero-shot and 75% on a smaller model after fine-tuning, and GPT-4o leads to lower performance. We evaluate two types of motion preferences (object-proximity and path-style), and we also analyze the trade-off between accuracy and computation cost in number of tokens. This work shows some promise in the potential of VLM integration with robot motion planning pipelines.

1 INTRODUCTION

Spatial reasoning can enable robots to understand the relationships among objects in the environment, which is essential for robots to interact with the physical world. With the abundant semantic knowledge from internet-scale data, foundation models' ability to mimic natural language understanding and visual reasoning has to potential to provide an intuitive interface for users to give instructions to robots.

More specifically, our assumption in this paper is that Visual Language Models (VLMs) can enable users to express preferences over robot motion. For instance, preferences over motion could be about whether users would like a long or wavy path, or a smooth or rough path, or a path that passes close or far away from an object (e.g. "move the can while staying away from the lamp"). VLMs could be used to score images of diverse robot paths obtained by robot motion planning algorithms, so as to select the path that best describes the user's preference. However, we still need to better understand how good current VLMs are at this task.

In this paper we generate a dataset of images of diverse robot motions in home settings, and evaluate the capability of VLMs to select the images that correctly match a text description of the displayed robot motion. To integrate VLMs into motion planning pipelines, we propose to proceed as follows: given a motion planning task (start and goal position of the robot), we generate diverse robot trajectories that achieve the task using a sampling-based motion planner, then we use a VLM to score the trajectories based on how well they match the user's description of spatial object relations or path styles, and finally we select the highest-score trajectory as the output.

Our work allows to evaluate the accuracy of different VLMs and image querying approaches at selecting these highest-score trajectories, thus informing future human-centered robot motion planning pipelines.

2 RELATED WORK

The visual understanding capability of VLMs has been applied to various tasks such as visual question answering (Liu et al., 2024; Chen et al., 2024a), object detection (Chen et al., 2024b; Kuo et al., 2023) and optical character recognition (Nacson et al., 2024). In robotics, VLMs and Large-Language Models (LLMs) have shown potential in enhancing the generalization capability of robots (Brohan et al., 2023; Shah et al., 2022; Ahn et al., 2022; Driess et al., 2023; Bucker et al., 2024; Black et al., 2024) and assisting human-robot interactions (Bucker et al., 2023; Arkin et al., 2020), even though they have also shown to be prone to various issues of bias (Azeem et al., 2024; Hundt et al., 2022).

Various studies have used VLMs in robot task planning. For example, SayCan (Ahn et al., 2022) leverages the semantic knowledge of LLMs to estimate the probability that actions make progress towards a given task. PaLM-E (Driess et al., 2023) has demonstrated that large multimodal embodied models trained on robotics tasks and internet-scale data can address various tasks in the real world. PaLM-E can generalize zero-shot on tasks with unseen objects, while SayCan is evaluated in kitchen tasks in a zero-shot manner. Our method also uses VLMs zero-shot on robot tasks in home settings, but focuses on motion planning tasks. Various other methods have been proposed since, also for task planning (Honerkamp et al., 2024; Wang et al., 2022; Chen et al., 2022; 2023; Lin et al., 2023). LLMs can provide the spatial relationships between objects for task and motion planners in object rearrangement tasks (Ding et al., 2023), which a planner can in turn use to compute task and motion plans. Wang el al. generates natural language narration using multimodal inputs including environment observations, robot states and task planner (Wang et al., 2024b). The natural language grounding is shown to improve recovery efficiency. NARRATE (Ismail et al., 2024) uses language to generate sequences of sub-tasks formulated with safe constraints and combine them with model predictive control. Similarly, we use VLMs to process requirements specified in textual descriptions, but focus on generating robot trajectories from instructions. Some methods use LLMs to break the long-horizon and complex tasks into sub-tasks and sub-goals (Yang et al., 2024; Lin et al., 2023). A recent study uses the reasoning capability of VLMs on raw images in trajectory planning (Tang et al., 2025). The previous works focus on using language models to compute goals and subgoals, but do not consider language to describe the style or desired properties of the path itself.

Language models have shown potential to generate motion from textual instructions as well. MotionGPT (Jiang et al., 2024) generates consecutive human motions by using multimodal signals as input tokens in LLMs. Researchers have also used Generative Pre-trained Transformer (GPT) with discrete representations to generate human motions from textual instructions (Zhang et al., 2023). Compared to these works, in this paper we use VLMs to generate robot motion instead of human motion, and our focus is on the use of VLMs (i.e. using images judges of proposed motion candidates).

Finally, another type of work in the literature using VLMs for robot planning is IMPACT (Ling et al., 2025), which uses these models to identify objects that tolerate contact with robots. While the approach is similar to our work in spirit, our work targets another problem—of modulating motion style or constraints using text input.

3 METHOD

3.1 GENERATING DIVERSE PATH CANDIDATES WITH MOTION PLANNING

To evaluate the VLMs on robot planning tasks, we first use a heuristic method to obtain a diverse set of candidate paths and visualize them on an image of the robot and environment. In particular, we use a combination of the Bidirectional Rapidly-exploring Random Trees (BiRRT) (Qureshi & Ayaz, 2015; Karaman & Frazzoli, 2011) and Probabilistic RoadMaps(Kavraki et al., 1996) to obtain diverse paths that satisfy start-state and goal-state constraints, though other diverse-planning methods could be used.

BiRRT operates by growing two trees: one from the start and one from the goal. It alternates between extending these trees and connecting them when they are sufficiently close. In order to obtain a variety of path candidates with different geometric and topological properties, we run BiRRT n times using different random seeds, therefore leading to n candidate trajectories. Similarly, we run PRM n times using different random seeds, each time building a graph from scratch and running Dijkstra

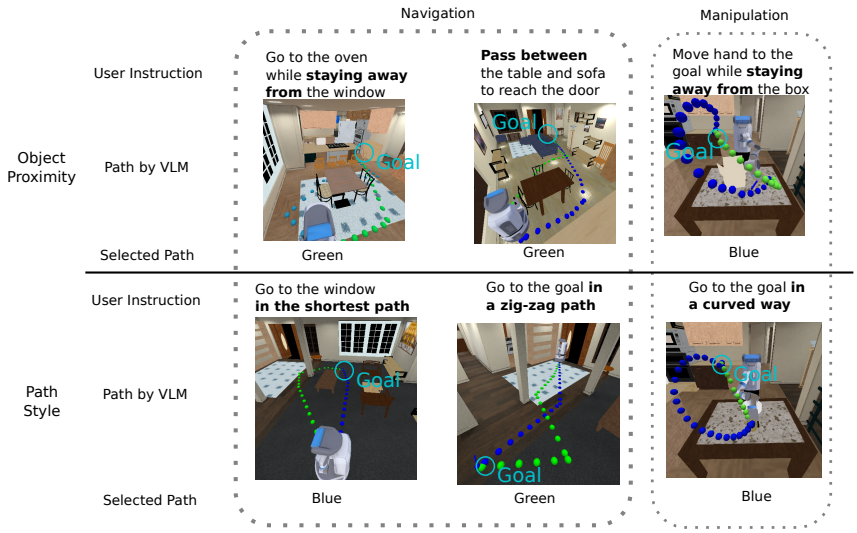


Figure 1: Examples of language-constrained robot motion planning problems, and the solutions scored highest by Qwen2.5-VL (dotted trajectories).

to find a path to the goal, and each time using a different cost function for extra diversity: we used shortest distance, sinusoidal, and circular path costs.

Then, we use the K-means clustering algorithm to group trajectory candidates based on their waypoints, and partition the n candidates into k clusters. After clustering, the path that is closest to the centroid of each cluster is selected to be visualized in image inputs for VLMs, therefore leading to $k < n$ candidate paths (and k images representing the execution or overlay of those paths in the environment, as we will describe next).

3.2 Dataset Construction

Using the above method for obtaining diverse trajectories for a given motion planning problem, we then constructed a dataset of 558 language-constrained robot motion planning problems. Out of these, 126 were navigation problems, and 432 were arm-manipulation problems. Each problem consists of a virtual scene, a start and a goal location, and a text-description of the desired properties of the motion. We manually selected start and goal locations in the different scenes so as to allow for diverse ways of traveling between the two. More information about the scenes is provided in Section 4.1.

We consider the following two categories of user preferences:

- **Object proximity** preferences describe desired spatial relationships with objects in the environment, e.g. stay close to/ away from object A, pass between object B and C.
- **Path style** preferences describe desired stylistic and shape-related properties of the robot trajectories, e.g. straight line, curved, zigzag, shortest path.

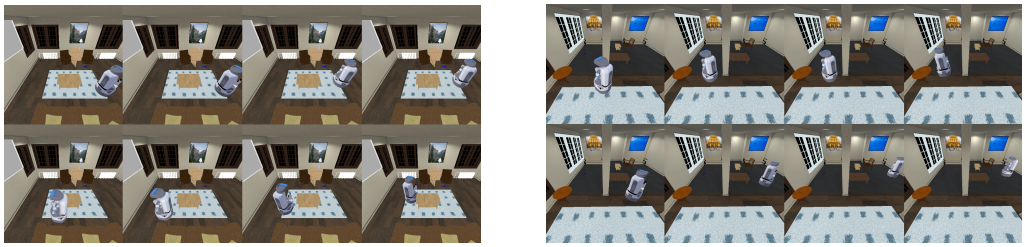
Examples are shown in Fig. 1. In order to obtain the ground-truth of the path that should be selected by a VLM for each problem, we applied our path-candidate generation and clustering procedure to generate a set of path candidates for each problem, and manually annotated each path with a description that was unique to it and could distinguish it from the others. In this way, each instruction is paired with a ground-truth path identifier. The curated dataset consists of sets of (image with multiple paths, text instruction, ground-truth path identifier for instruction). Using this dataset, we then evaluated the accuracy of the VLM at selecting the right path candidate for each problem.

3.3 VLM-based Path Selection

Given the candidate paths which are output by the clustering algorithm, we generate an image (or set of images) presenting those paths, and then use a Vision Language Model (VLM) to compute

(a) (b) (c) (d)

Figure 2: Examples of single-image trajectory trails in 4 scenes: a) move towards the door; b) move to the table in the kitchen; c) move to the table; d) move to the shelf behind the columns.



(a) (b)

Figure 3: Example of galleries of screenshots showing the robot moving along each trajectory, with each row presenting a sequence screenshots from one trajectory. (a) The robot goes around the table in 2 topologically different ways; (b) The robot goes around the table and column in 2 topologically different ways.

scores according to how well the image matches the instruction l, and obtain the path with the highest score. We use the following prompt in order to obtain the selected path: "This picture shows a robot in a simulated environment. Each trajectory of dots is a sequence of waypoints that guide the robot to move along. Please rate each of the trajectories out of 100 for how well they match the user instruction, by analyzing if they satisfy the constraints on object proximity and path style. User instruction: $\{l\}$. Give the color with the highest score at the end."

We propose four methods of querying VLMs for robot path selection from images:

1. Single-image trajectory: the candidate paths are shown as trails of dots in different colors, as shown in Fig. 2. For navigation problems, we visualize paths as dots along the route of the base of the robot. For manipulation problems, we visualize the trajectory of the end-effector with similar dots (e.g. Fig. 1). All the candidate paths are shown in one picture to query VLMs.

2. Multi-image trajectory trail: all the candidate paths are split into individual images, where each image contains one trajectory trail. The VLM is queried once for every image, to give scores to each image based on the instructions. Finally we select the image (trajectory) with the highest score.

3. Single-image trajectory trail with visual context: the VLM is first asked to generate visual context on the images (Maeda et al., 2024), which describes the detailed content of images with structured data, including objects, features, and relationships between them. Then the visual context is given to the VLM together with textual and visual prompts to output the highest-score path.

4. Single-image screenshot gallery: for each candidate path we generate a sequence of images of the robot moving along the path by simulating the execution of the path. Fig. 3 shows the screenshot gallery of 2 candidate trajectories, with each row representing one path. The VLM is then asked to select the row that satisfied the user description.

We will evaluate and compare these four methods in our experiments.

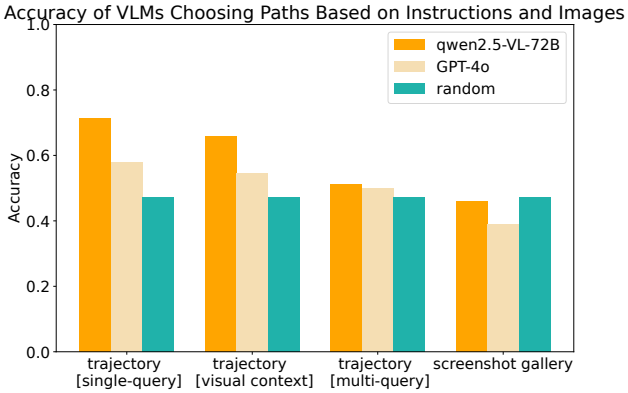


Figure 4: Accuracy of VLMs in selecting candidate paths in images with different query methods, averaged across two types of preferences (object proximity and path style).

4 RESULTS

4.1 EXPERIMENTAL SETUP

We used iGibson (Li et al., 2022), a simulation environment that contains 3D interactive scenes reconstructed from real homes, to generate robot trajectories in simulated home environments. As described in the previous section, we used the BiRRT motion planner (Qureshi & Ayaz, 2015; Karaman & Frazzoli, 2011) and Probabilistic RoadMap(Kavraki et al., 1996) to generate n =50 diverse paths from the start to goal in each planning problem in each simulated scene. We use a set of iGibson scenes called Ihlen_1_int, Pomaria_1_int, Beechwood_0_int, Benevolence_1_int and Merom_0_int which include objects from the BEHAVIOR dataset (Srivastava et al., 2022). We evaluate three different VLMs: 1) Qwen2.5-VL(Team, 2025)(Wang et al., 2024a): a VLM with visual localization ability. It uses bounding boxes and point-based representation for object grounding which provides a foundation for visual reasoning. We use the version Qwen2.5-VL-72B-Instruct in our experiments. 2) GPT-4o: a model that can take images as input and answer questions about the images. It achieves state-of-the-art performance on visual perception benchmarks including MMMU (Yue et al., 2024) and ActivityNet (Heilbron et al., 2015). 3) LLaVa1.5: a model that excels in conversational-style Visual Question Answering (Liu et al., 2023).

4.2 QUALITATIVE ANALYSIS

Fig. 1 shows 6 example problems with user instructions and the motion scored highest by Qwen2.5-VL. The figure shows that the robot is instructed to go to the goal position with constraints on the movement with respect to certain objects in the environment or the overall path style. In the first example (top left), the robot is commanded to go to the oven. There are 2 ways to go around the table to reach the goal position. The VLM selected the trajectory that matches the instruction which is far away from the window. In the last example (bottom right), the robot receives the instruction to move its hand in a curved way. In this case, the blue curved path is chosen instead of the green straight line.

4.3 ACCURACY WITH DIFFERENT IMAGE QUERY METHODS

Fig. 4 shows overall accuracy on the navigation problems, using the 4 different image query methods. As the figure shows, the best-performing method was the single-query method, with an average accuracy over 70% when using Qwen2.5-VL. The multi-query method led to lower accuracy, even though it uses more queries. We hypothesize that the reason for this is that when a VLM is asked to score a trajectory individually, the metric is not consistent across multiple requests—because it is not able to compare the one in the image currently being processed with the others. On the other hand, when all the trajectories are shown together in the single-query method, the VLMs can compare which trajectory is further away from the painting and assign consistent relative scores to those trajectories.

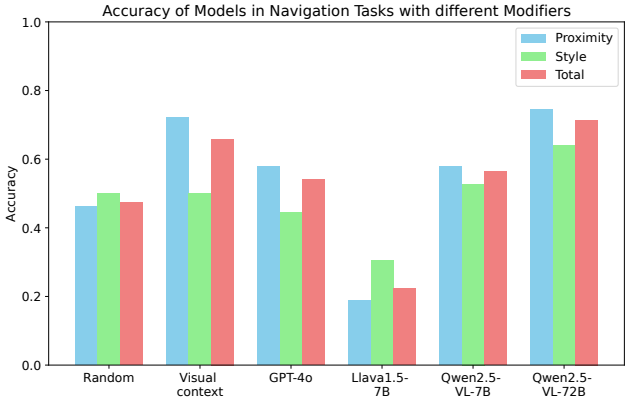


Figure 5: Accuracy of VLMs in selecting candidate paths in images, on navigation tasks.

The screenshot-gallery method led to an accuracy slightly higher than random selection. Compiling all screenshots in one image makes the size of each screenshot small, thus leading to a lack of detail. For the rest of the experiments, we therefore use the single-query method, which led to highest accuracy.

4.4 Accuracy on two types of preferences (object proximity and path style)

4.4.1 Navigation Tasks

The accuracy on different types of preferences on the 126 instruction-image pairs (90 object-proximity and 36 path-style) is shown in Fig. 5. All methods are evaluated zero-shot. Qwen2.5-VL performs better than GPT4-o, which is consistent with its reported precise object grounding capabilities (Team, 2025). With single-query trajectory trail image, Qwen2.5-VL achieves an accuracy of 74.4% on proximity problems and 63.9% on path-style problems. For all the tested models except Llava1.5, the accuracy of proximity problems is higher than path style problems.

From our experiment, adding a visual context step does not increase accuracy. This might be due to the fact that these large-sized models automatically generate and keep tokens of image context already, e.g. Qwen2.5 has a default setting to keep context with length up to 32,768 tokens. Although visual context can help smaller-sized models such as Llava (Liu et al., 2024) to keep track of detailed content of images, including objects, attributes, and relationships (Maeda et al., 2024), this extra step can add redundancy to the textual information and can potentially undermine the performance of models built with internal context tracking.

4.4.2 Manipulation Tasks

The accuracy of the visually matching the textual command to robot taks image on 432 manipulation problems (258 proximity and 174 path style) is shown in Fig. 6. In the manipulation tasks, the robot stands in front of a table with objects and is asked to move its arm in reference to the objects on the tabletop. Examples are shown in Fig. 1. Qwen2.5-VL-72B has the highest accuracy in proximity problems (66.3%) and GPT-4o has the highest in path-style problems (69.5%). The overall success rate in manipulation tasks is lower than in navigation (65.5% in manipulation vs 71.4% in navigation on Qwen2.5-VL-72B).

4.5 Computation cost

We investigate the computation cost of each querying method by comparing the number of tokens used by each method and each VLM. Fig. 7 shows the average number of tokens used in each request to the VLMs. When all trajectories are shown in one image, an average token number of 687.3 is used by Qwen2.5-VL and 721.2 by GPT-4o. This querying method requires the fewest number of tokens among all methods we have tested, since it uses a single image, while also leading to the highest performance.

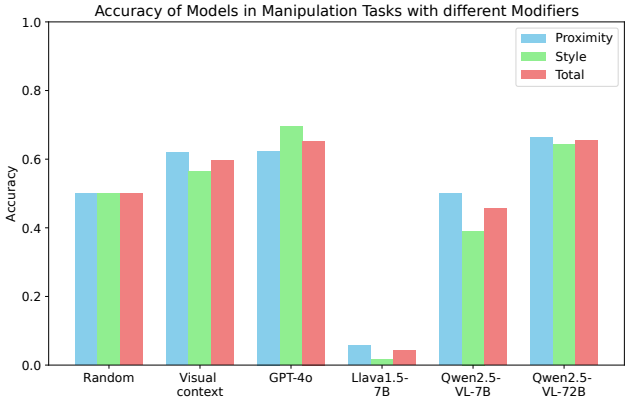


Figure 6: Accuracy of VLMs in selecting candidate paths in images, on manipulation tasks.

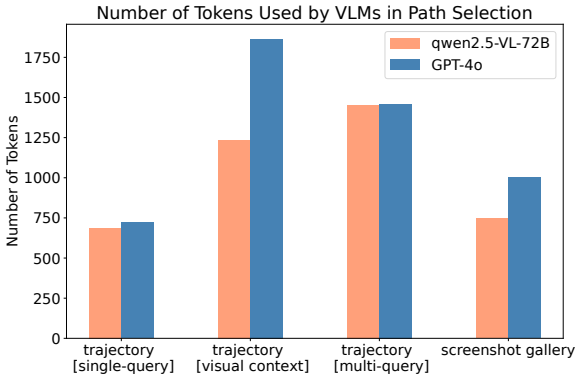


Figure 7: Number of tokens used by each querying method.

Next, we investigate the relation between the *financially-relevant* computation cost, in number of tokens, that VLMs consume and the accuracy of the path selection by limiting the maximum number of tokens used (i.e. resizing the images sent to the VLMs). We test the accuracy of the models by restricting the maximum allowed number of token from 200 to 800 per query on Qwen2.5-VL-7B and 72B. Fig.8 shows that, given a fixed querying method (single query trajectory image), the accuracy of path selection grows approximately linearly with the number of tokens used within the range between the worst and best performance.

4.6 Effect of Finetuning on Language Models with smaller size

Finally, we studied the performance improvement after finetuning. We used Supervied Fine-Tuning (SFT) with a training set that contained 98 examples on 2 models, LLaVa1.5-7B and Qwen2.5-VL-7B. We then tested them on a test set that contained 14 proximity problems and 14 path-style problems. The performance of the models before and after finetuning is shown in Fig. 9. For fair comparison purposes, values of zero-shot performance of the original models shown on the table were also computed on the test. The figure shows there is a prominent increase in accuracy in proximity problems for all models after finetuning. The accuracy improves more than 20% for Qwen2.5-VL-7B and more than 60% for LLaVa1.5-7B after exposure to a small dataset. Therefore, the architecture can adapt to new types of user instructions. Although the smaller model (Qwen2.5-VL-3B) takes less computation for finetuning and inference, its performance did not improve as much as other 2 models with 7B parameters.

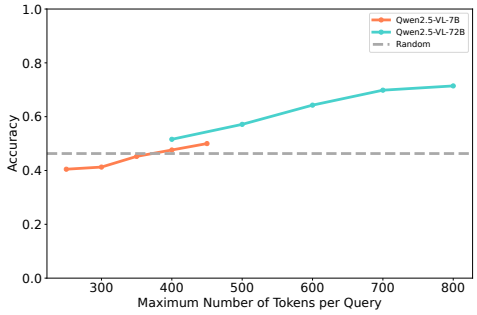
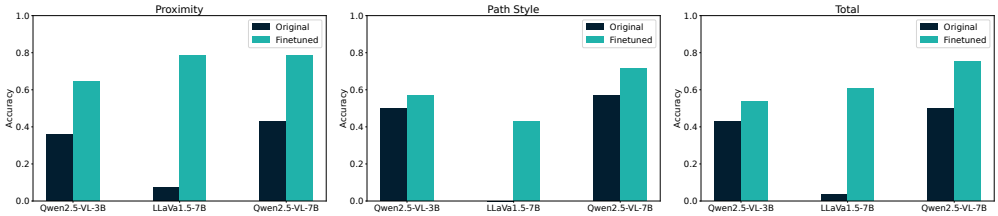


Figure 8: Accuracy of path selection with varying max. number of allowed tokens (image size).



(a) Proximity (b) Path style (c) Total

Figure 9: Accuracy of zero-shot and finetuned models on a test set of 28 problems.

4.7 Discussion and Limitations

As shown in Fig. 4, using different methods of visualization to query VLMs can lead to varied performance. We found that a single image showing all clustered candidate trajectories as colored dotted paths was an effective way to query state-of-the-art VLMs to select robot motion, given user descriptions of motion preferences. Different visualization methods could lead to further improvements.

We found that a common failure was that VLMs failed to recognize which path was the shortest or longest when requested, which is exactly the type of problem classical optimal planners (e.g. RRT*, PRM*) can efficiently solve. Another common failure was of "hallucination", where VLMs selected a candidate path that did not exist (e.g. selected 'red' as the best path, when none were of that color).

5 Conclusions

In this paper, we evaluated the spatial reasoning capabilities of VLMs to select robot trajectories that satisfy user preferences related to path topology and style. VLMs scored motion candidates with respect to how well they matched the user request, given images with candidate paths generated by robot motion planners. We showed that Qwen2.5-VL achieves an overall accuracy of 71.4%, with 74.4% on proximity and 63,9% on path style problems, which is considerably higher than GPT4-o. The evaluated VLMs performed better on object-proximity preferences, compared to path style (e.g. to obtain a "long" or "zigzag" path), but finetuning was able to obtain high accuracy gains even on small vision-language models with few examples, by between 10 and 57 percentage points. Finally, we investigated the tradeoff between accuracy and computation cost, showing that accuracy decreases approximately linearly with the decrease in number of tokens or image size. In the future, our goal is to integrate these capabilities robustly into robot motion planning pipelines, though this may require significant work on increasing accuracy or integrating users into the loop through appropriate interfaces.

ACKNOWLEDGMENTS

This work was supported by the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [EP/S023356/1].

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

Jacob Arkin, Daehyung Park, Subhro Roy, Matthew R Walter, Nicholas Roy, Thomas M Howard, and Rohan Paul. Multimodal estimation and communication of latent semantic knowledge for robust execution of robot instructions. *The International Journal of Robotics Research*, 39(10-11): 1279–1304, 2020.

Rumaisa Azeem, Andrew Hundt, Masoumeh Mansouri, and Martim Brandão. Llm-driven robots risk enacting discrimination, violence, and unlawful actions. *arXiv preprint arXiv:2406.08824*, 2024.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL https://arxiv.org/abs/2307.15818.

Arthur Bucker, Luis Figueredo, Sami Haddadin, Ashish Kapoor, Shuang Ma, Sai Vemprala, and Rogerio Bonatti. Latte: Language trajectory transformer. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 7287–7294. IEEE, 2023.

Arthur Bucker, Pablo Ortega-Kral, Jonathan Francis, and Jean Oh. Grounding robot policies with visuomotor language guidance, 2024. URL https://arxiv.org/abs/2410.06473.

Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning, 2022. URL https://arxiv.org/abs/2209.09874.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024a.

Hanning Chen, Wenjun Huang, Yang Ni, Sanggeon Yun, Yezi Liu, Fei Wen, Alvaro Velasquez, Hugo Latapie, and Mohsen Imani. Taskclip: Extend large vision-language model for task oriented object detection, 2024b. URL https://arxiv.org/abs/2403.08108.

William Chen, Siyi Hu, Rajat Talak, and Luca Carlone. Leveraging large (visual) language models for robot 3d scene understanding, 2023. URL https://arxiv.org/abs/2209.05629.

Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. Task and motion planning with large language models for object rearrangement, 2023. URL https://arxiv.org/abs/2303.06247.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015. doi: 10.1109/CVPR.2015.7298698.

Daniel Honerkamp, Martin Büchner, Fabien Despinoy, Tim Welschehold, and Abhinav Valada. Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. *IEEE Robotics and Automation Letters*, 9(10):8298–8305, October 2024. ISSN 2377-3774. doi: 10.1109/lra.2024.3441495. URL http://dx.doi.org/10.1109/LRA.2024.3441495.

Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. Robots enact malignant stereotypes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 743–756, 2022.

Seif Ismail, Antonio Arbues, Ryan Cotterell, René Zurbrügg, and Carmen Amo Alonso. Narrate: Versatile language architecture for optimal control in robotics, 2024.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024.

Sertac Karaman and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning, 2011. URL https://arxiv.org/abs/1105.1186.

L.E. Kavraki, P. Svestka, J.-C. Latombe, and M.H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Transactions on Robotics and Automation*, 12(4):566–580, 1996. doi: 10.1109/70.508439.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models, 2023. URL https://arxiv.org/abs/2209.15639.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 455–465. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/li22b.html.

Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: from natural language instructions to feasible plans. *Autonomous Robots*, Nov 2023. ISSN 1573-7527. doi: 10.1007/s10514-023-10131-7. URL https://doi.org/10.1007/s10514-023-10131-7.

Yiyang Ling, Karan Owalekar, Oluwatobiloba Adesanya, Erdem Bıyık, and Daniel Seita. Impact: Intelligent motion planning with acceptable contact trajectories via vision-language models, 2025. URL https://arxiv.org/abs/2503.10110.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

Koki Maeda, Shuhei Kurita, Taiki Miyanishi, and Naoaki Okazaki. Vision language model-based caption evaluation method leveraging visual context extraction, 2024.

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. Docvlm: Make your vlm an efficient reader, 2024. URL https://arxiv.org/abs/2412.08746.

Ahmed Hussain Qureshi and Yasar Ayaz. Intelligent bidirectional rapidly-exploring random trees for optimal motion planning in complex cluttered environments. *Robotics and Autonomous Systems*, 68:1–11, June 2015. ISSN 0921-8890. doi: 10.1016/j.robot.2015.02.007. URL http://dx.doi.org/10.1016/j.robot.2015.02.007.

Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action, 2022. URL https://arxiv.org/abs/2207.04429.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on robot learning*, pp. 477–490. PMLR, 2022.

Zhipeng Tang, Sha Zhang, Jiajun Deng, Chenjie Wang, Guoliang You, Yuting Huang, Xinrui Lin, and Yanyong Zhang. Vlmplanner: Integrating visual language models with motion planning. *arXiv preprint arXiv:2507.20342*, 2025.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

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. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, Shih-Fu Chang, Mohit Bansal, and Heng Ji. Language models with image descriptors are strong few-shot video-language learners, 2022. URL https://arxiv.org/abs/2205.10747.

Zihan Wang, Brian Liang, Varad Dhat, Zander Brumbaugh, Nick Walker, Ranjay Krishna, and Maya Cakmak. I can tell what i am doing: Toward real-world natural language grounding of robot experiences, 2024b. URL https://arxiv.org/abs/2411.12960.

Zhutian Yang, Caelan Garrett, Dieter Fox, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Guiding long-horizon task and motion planning with vision language models, 2024. URL https://arxiv.org/abs/2410.02193.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations, 2023. URL https://arxiv.org/abs/2301.06052.