NaviTrace: Evaluating Embodied Navigation of Vision-Language Models

Tim Windecker[1,2], Manthan Patel[1], Moritz Reuss[2], Richard Schwarzkopf[3], Cesar Cadena[1], Rudolf Lioutikov[2,4], Marco Hutter[1] and Jonas Frey[1]

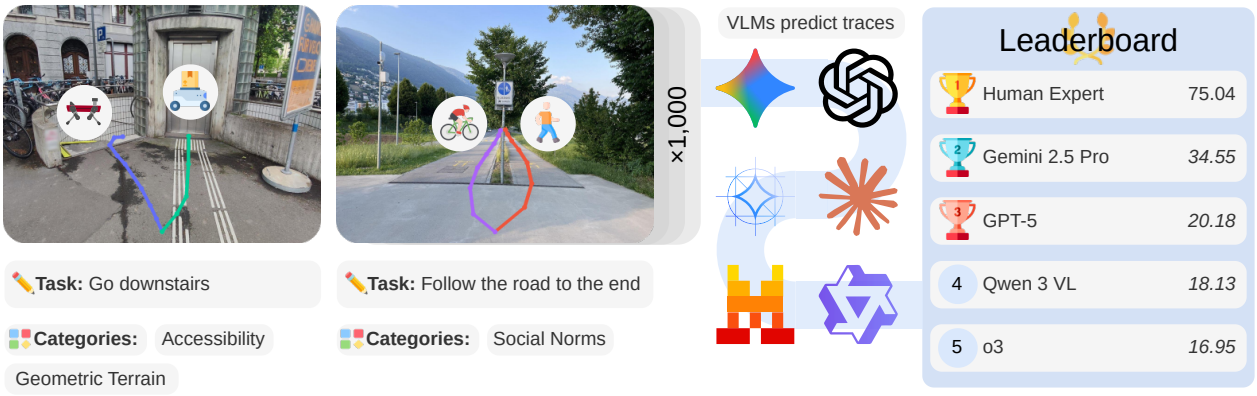

Fig. 1: We introduce **NaviTrace**, a novel VQA benchmark for VLMs that evaluates models on their embodiment-specific understanding of navigation across challenging real-world scenarios.

***Abstract*— Vision–language models demonstrate unprecedented performance and generalization across a wide range of tasks and scenarios. Integrating these foundation models into robotic navigation systems opens pathways toward building general-purpose robots. Yet, evaluating these models' navigation capabilities remains constrained by costly real-world trials, overly simplified simulations, and limited benchmarks. We introduce NaviTrace, a high-quality Visual Question Answering benchmark where a model receives an instruction and embodiment type (human, legged robot, wheeled robot, bicycle) and must output a 2D navigation trace in image space. Across 1000 scenarios and more than 3000 expert traces, we systematically evaluate eight state-of-the-art VLMs using a newly introduced semantic-aware trace score. This metric combines Dynamic Time Warping distance, goal endpoint error, and embodiment-conditioned penalties derived from per-pixel semantics and correlates with human preferences. Our evaluation reveals consistent gap to human performance caused by poor spatial grounding and goal localization. NaviTrace establishes a scalable and reproducible benchmark for real-world robotic navigation. The benchmark and leaderboard can be found at `https://leggedrobotics.github.io/navitrace_webpage/`.**

I. INTRODUCTION

We introduce NaviTrace, a VQA benchmark specially designed to evaluate embodiment-specific navigation performance across 1,000 diverse scenarios and four different embodiments. Each task within NaviTrace consists of a single real-world image paired with a high-quality language instruction, enabling efficient data collection while capturing challenging navigation tasks. Following the most intuitive approach to answering navigation questions, we provide solutions per embodiment as 2D paths in image space, which we refer to as traces. This carefully chosen formulation is more expressive than low-level commands such as "Forward" [1] and can also support longer-horizon planning. It can be seen as an extension of pointing—a common task that is evaluated and optimized in current foundation models [2], [3] and widely used to assess the visual grounding of VLMs [4]. Furthermore, traces have proven beneficial for addressing manipulation tasks [5]–[8]. NaviTrace tests VLMs for instruction following, spatial understanding, and physical understanding of varying embodiments (human, legged robot, wheeled robot, and bicycle), and categorizes scenarios based on the type of navigation challenges.

We develop a semantic-aware score to measure how well the predicted navigation trace aligns with human preferences. To achieve this, we combine the Dynamic Time Warping distance to a ground-truth trace, goal endpoint error, and pixelwise embodiment-conditioned penalties derived from a semantic segmentation model. We show that our metric, while inexpensive to compute and annotate, is competitive with more expensive human-derived metrics in aligning with human preferences.

Specifically, our main contributions are:

1) **NaviTrace:** A novel high-quality benchmark for evaluating the ability of VLMs to predict how different embodiments navigate in 1000 diverse and challenging real-world scenarios.
2) **Semantic-aware Score**: A new metric to measure the

[1]Robotic Systems Lab, ETH Zurich, Zurich, Switzerland
[2]Intuitive Robots Lab, KIT, Karlsruhe, Germany
[3]FZI Research Center for Information Technology, Karlsruhe, Germany
[4]Robotics Institute Germany

accuracy of 2D traces for real-world images. We test the score for alignment with human preferences by showcasing its correlation to human expert judgments.

3) **Evaluation of VLMs:** Comprehensive assessment of current state-of-the-art VLMs on our benchmark.

II. NAVITRACE BENCHMARK

We introduce NaviTrace, a benchmark for evaluating the ability of VLMs to predict navigation strategies for different embodiments in real-world scenarios (see Figure 1). To ensure relevance, diversity, and high-quality annotation, we manually collect real-world images and perform all labeling by hand. The dataset contains 1,000 scenarios with more than 3,000 traces, divided evenly into validation and test splits. The test set annotations remain secret and are used to evaluate the public leaderboard. During evaluation, a VLM receives a structured prompt with an image, a task description, and an embodiment type. The model must predict a path that solves the task, and its output is measured using our novel task-specific score function.

A. Data Collection

Each scenario in NaviTrace combines images, instructions, traces, and embodiment types to capture realistic navigation challenges (see Figure 1 for examples).

Image. Each scenario includes a distinct first-person image of a real-world environment. Most images are crowd-sourced and captured with consumer devices such as phones or GoPros, complemented by 164 curated samples from the publicly available GrandTour dataset [9]. To preserve privacy, we anonymize all personal data using EgoBlur [10] to blur faces and license plates.

Task Instruction. Each image is paired with a manually written instruction solvable purely from the visual information. These instructions emphasize cases where different embodiments behave differently, while still reflecting everyday scenarios. They are formulated either as goals (e.g., "Go to the red car") or as directional instructions (e.g., "Go forward, then turn left at the traffic light.").

Task Categories. To classify capabilities of models according to navigation-relevant attributes, we tagged each scenario with one or more categories, describing the main challenges of the navigation task:

- Geometric Terrain Property Assessment
- Semantic Terrain Property Assessment
- Accessibility
- Visibility
- Social Norms
- Dynamic Obstacle Avoidance
- Stationary Obstacle Avoidance

Ground-Truth Trace. We define a trace as a sequence of 2D points given as image coordinates that describes a navigation path. This representation is detached from robot-specific controls, ensuring compatibility with diverse model architectures. We draw one trace per suitable embodiment and multiple traces if there are equally valid and fast alternatives (e.g., avoiding an obstacle from the left or right).

Embodiments. We model four embodiment types to capture various real-world navigation behaviors:

- Human
- Legged Robot
- Wheeled Robot
- Bicycle

We deliberately exclude cars, since their viewpoint differs fundamentally from the embodiments above.

B. Score

To fairly evaluate VLM-generated navigation traces, we design a score function that balances three factors: (i) how closely the path follows the ground truth, (ii) whether it reaches the intended goal, and (iii) whether it avoids unsafe or irrelevant regions. Later, we describe how we make the score range easier to interpret and show that our score formulation aligns well with human preferences. Formally, a trace is a sequence of points $T = [(x_1, y_1), \ldots, (x_n, y_n)]$ in image pixel space. We compare it against ground-truth traces across modalities $T' = [(x'_1, y'_1), \ldots, (x'_m, y'_m)] \in \mathcal{G}$ and select the trace with the lowest error:

$$\text{Score}(T, \mathcal{G}) = \min_{T' \in \mathcal{G}} \text{DTW}(T, T') + \text{FDE}(T, T') + \text{Penalty}(T) \quad (1)$$

Trace Similarity: We utilize Dynamic Time Warping (DTW) [11] with the Euclidean distance as the error metric, to measure trace similarity. DTW aligns sequences by stretching or compressing the time axis and can be computed using dynamic programming:

$$DTW(T, T') = D(n, m) \quad (2)$$

$$D(0, 0) = 0 \quad (3)$$

$$D(i, 0) = D(0, j) = \infty \ \ (i, j > 0) \quad (4)$$

$$D(i, j) = d((x_i, y_i), (x'_j, y'_j)) + \min\{D(i-1, j),\ D(i, j-1),\ D(i-1, j-1)\} \quad (5)$$

Goal Reaching: To reward reaching the correct target, we add the Final Displacement Error (FDE), which measures the Euclidean endpoint distance:

$$\text{FDE}(T, T') = \text{d}((x_n, y_n), (x'_m, y'_m)) \quad (6)$$

Semantic Penalty: Finally, we introduce embodiment-specific semantic costs that penalize traces crossing undesired regions. Using a Mask2Former model [12] trained on Mapillary Vistas [13], we infer semantic masks and map each class to manually tuned penalty values $\text{m}_e(S_i)$ depending on embodiment e. Classes representing more dangerous areas or obstacles are assigned higher penalty values. To allow for small deviations, we exclude a tolerance band around the ground-truth. The penalty is averaged pixel-wise along the predicted trace:

$$\text{Penalty}(T) = \frac{1}{|\text{Pixels}(T)|} \sum_{i \in \text{Pixels}(T)} \text{m}_e(S_i) \quad (7)$$

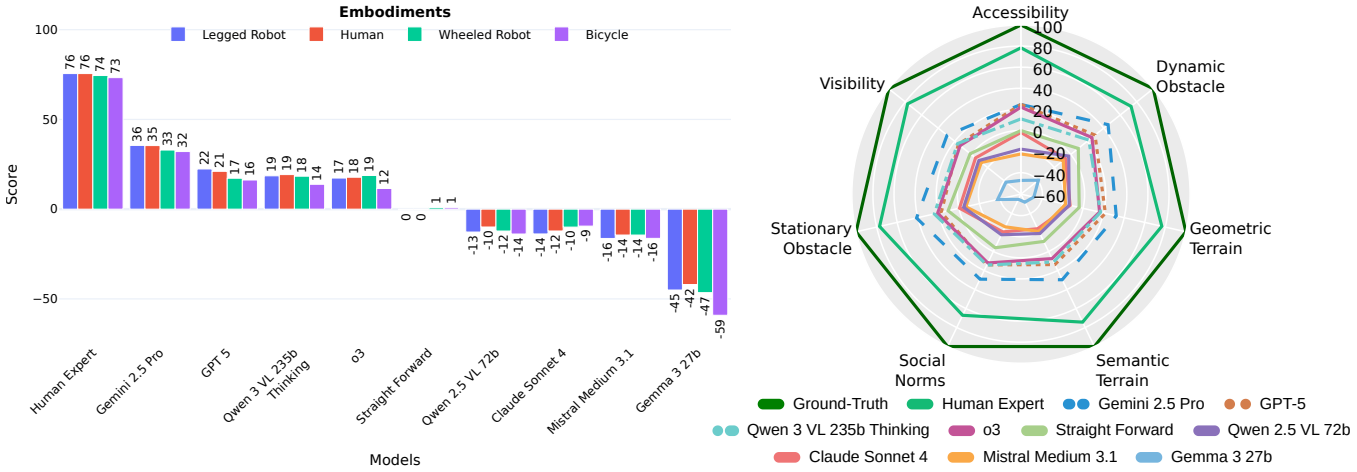

Fig. 2: **Left:** Ranking of VLMs, the uninformed baseline Straight Forward, and human expert performance split into each embodiment. Note that a higher score is better. **Right:** Performance per task category for the same models.

Scaling: In order to make the score values easier to interpret, we scale them to a range where a naive baseline is set to 0 and the best score at 100. We achieve this by setting the ground-truth performance 0 as the lower bound of the raw score. For the upper bound, we select the performance of just drawing a vertical line through the image center, which corresponds to the Straight Forward baseline performance of 3234.75. This results in the scaled score function:

$$\widehat{\text{Score}}(T, \mathcal{G}) = \frac{3234.75 - \text{Score}(T, \mathcal{G})}{3234.75} \cdot 100 \qquad (8)$$

Note that negative values are possible and do occur in our later experiments as some models perform worse than the Straight Forward baseline.

III. EXPERIMENTS

Our experiments aim to address three key questions:

1) How well do current VLMs predict navigation traces?
2) Does performance vary with embodiment or task category?
3) Which aspects of the tasks pose the greatest challenges?

To answer these questions, we first establish two baselines that give insight into the core difficulties of predicting navigation traces. Next, we outline our deployment of state-of-the-art VLMs, before presenting and analyzing the benchmark results for the test split.

A. Baselines

We compare VLM performance against two baselines:

- **Human:** Multiple participants collectively solve all test split scenarios, providing an upper bound for model performance.
- **Straight Forward:** Places a vertical line through the image center.

B. Models

We evaluate all VLMs by querying each model through API calls. After preliminary testing, we select five representative proprietary models: Gemini 2.5 Pro [14], GPT-5 [15], o3 [16], Claude Sonnet 4 [17], and Mistral Medium 3.1 [18]. We also include three open-weight models: Qwen 2.5 VL 72B [19], Qwen 3 VL 235B A22B Thinking [19], and Gemma 3 27B [20]. Among these models, Gemini 2.5 Pro, GPT-5, o3, and Qwen 3 VL automatically generate reasoning steps. Each model receives a carefully crafted prompt specifying the task, output format, expected embodiment behavior, and embodiment type. Models are instructed to return navigation traces as lists of normalized 2D points in JSON format, which we parse to compute performance scores.

C. Performance

We first analyze performance across embodiment types for both VLMs and human experts (see Figure 2). As a naive uninformed reference, we include the Straight Forward baseline. Human experts clearly outperform all VLMs, highlighting the gap between model capabilities and task difficulty. However, human experts also fall short of a perfect score as this requires to accurately match the ground-truth traces, as well as other human errors, such as occasional misunderstandings of the task. Among the models, Gemini 2.5 Pro ranks best, followed by GPT-5, Qwen 3 VL, and o3 with the Straight Forward baseline ranking unexpectedly close behind o3. Example predictions of the top four models are shown in Figure 3. Generally, we do not observe significant differences between embodiment types for all the models.

Turning to task categories in Figure 2 on the right, we again observe only minor variation. This uniformity should not be mistaken for balanced competence. Rather, the overall weakness of the models masks whether category and embodiment-specific differences exist. The competitiveness of the naive Straight Forward baseline highlights this deficit.

Finally, we provide qualitative insights into the reasoning process of models such as Gemini 2.5 Pro and o3. Figure 4 contains an example reasoning output of o3 where the task is to "go to the red car". While the model's textual reasoning correctly distinguishes between the available path options and identifies the correct solution, its predicted trace fails to align

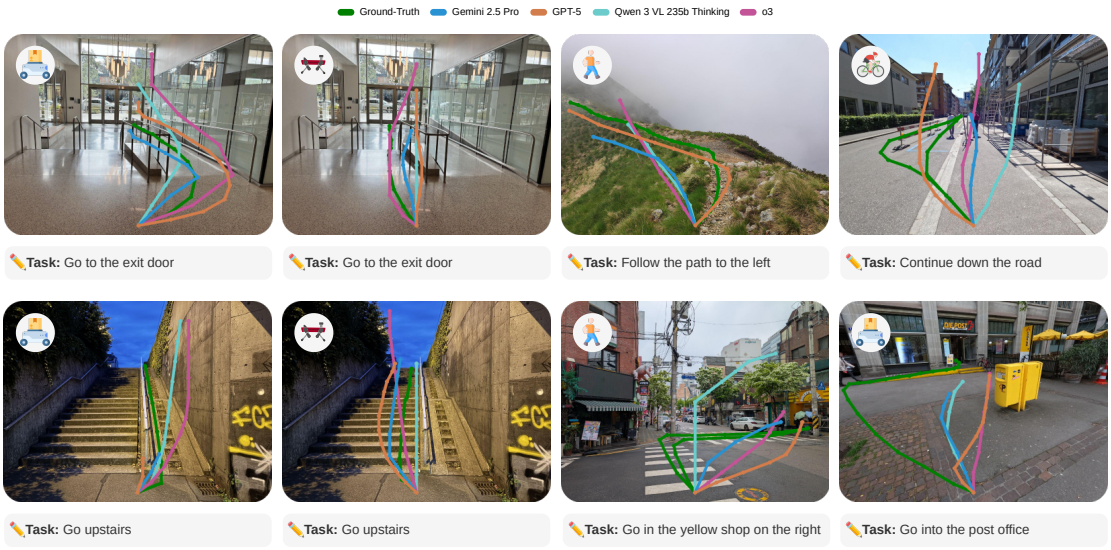

Fig. 3: Example predictions by the models Gemini 2.5 Pro, GPT-5, Qwen 3 VL, and o3.

with this reasoning. This is a common pattern we observe when qualitatively analyzing o3's reasoning and suggests a gap between linguistic reasoning and spatial grounding, particularly in localizing traversable structures within the image.

D. Summary of Key Findings

Our evaluation reveals two critical insights about current VLM navigation capabilities and areas of future work:

(1) Large human performance gap. Across all four embodiments and task categories, VLM scores are substantially worse than the human baseline, highlighting significant room for improvement (see Figure 2). **(2) Embodiment robustness.** Aggregate performance differences across Human, Legged Robot, Wheeled Robot, and Bicycle embodiments are small, suggesting general limitations in spatial grounding rather than embodiment-specific blind spots (see Figure 2).

Fig. 4: Example of o3's reasoning with the prediction in pink on the left and the steps on the right. The model reasons correctly but is unable to predict a corresponding trace.

IV. CONCLUSION

We presented NaviTrace, a novel benchmark for evaluating VLM navigation capabilities across different embodiments, along with a novel semantic-aware scoring function for fair evaluation of 2D navigation traces. NaviTrace provides the first systematic evaluation framework for embodied navigation in real-world scenarios, featuring $1,000$ diverse images from urban and rural environments and four embodiment types. Our benchmark extends pointing tasks to sequential navigation prediction, creating a natural bridge between high-level VLM reasoning and low-level robotic control.

To encourage future progress, we will make NaviTrace publicly available with test tasks, a leaderboard, and validation split for potential fine-tuning applications. NaviTrace establishes an essential testbed for developing and evaluating navigation-capable VLMs, enabling advances in embodied AI towards truly capable robotic navigation systems.

V. LIMITATIONS

NaviTrace has several key limitations. The dataset is geographically concentrated in Switzerland, which may limit generalizability to other regions with different infrastructure and navigation norms. The benchmark is restricted to single-image scenarios, preventing evaluation of temporal reasoning and multi-step planning required in dynamic environments. The current embodiment selection is limited to ground vehicles and excludes aerial drones. Additionally, our semantic scoring function relies on automated segmentation models that may introduce systematic evaluation biases.

While annotating traces has proven to be easy and efficient, the proposed scoring function—although shown to align effectively with human preferences and sufficient for evaluating VLM navigation capabilities—may still fail to capture more nuanced aspects of human preferences. For instance, while multiple ground-truth traces can capture ambiguities, they constrain the score function to recognize only specific points as goals rather than broader targets such as an entire doorway. Furthermore, it is not possible to take into account whether a trace works for precisely defined robot dimensions.

ACKNOWLEDGMENT

This work was supported by the German Research Foundation (DFG) – 448648559, Luxembourg National Research Fund (Ref. 18990533), and the Swiss National Science Foundation (SNSF) as part of the projects No.200021E_229503 and No.227617. We thank Kaiqi Qu, Omkar Jarande, and Qicai Tan for joining us in the labeling effort. We also thank all the people helping us collect images and participating in the evaluation of human performance.

REFERENCES

[1] A.-C. Cheng *et al.*, "Navila: Legged robot vision-language-action model for navigation," 2025. [Online]. Available: https://arxiv.org/abs/2412.04453

[2] G. R. Team *et al.*, "Gemini robotics: Bringing ai into the physical world," 2025. [Online]. Available: https://arxiv.org/abs/2503.20020

[3] M. Deitke *et al.*, "Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models," 2024. [Online]. Available: https://arxiv.org/abs/2409.17146

[4] L. Cheng *et al.*, "Pointarena: Probing multimodal grounding through language-guided pointing," 2025. [Online]. Available: https://arxiv.org/abs/2505.09990

[5] J. Yang *et al.*, "Magma: A foundation model for multimodal ai agents," 2025. [Online]. Available: https://arxiv.org/abs/2502.13130

[6] D. Niu *et al.*, "Llarva: Vision-action instruction tuning enhances robot learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.11815

[7] V. de Bakker *et al.*, "Scaffolding dexterous manipulation with vision-language models," 2025. [Online]. Available: https://arxiv.org/abs/2506.19212

[8] R. Zheng *et al.*, "Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," 2025. [Online]. Available: https://arxiv.org/abs/2412.10345

[9] J. Frey *et al.*, "Boxi: Design Decisions in the Context of Algorithmic Performance for Robotics," in *Proceedings of Robotics: Science and Systems*, Los Angeles, United States, June 2025.

[10] N. Raina *et al.*, "Egoblur: Responsible innovation in aria," 2023.

[11] P. Senin, "Dynamic time warping algorithm review," *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, vol. 855, no. 1-23, p. 40, 2008.

[12] B. Cheng *et al.*, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 1290–1299.

[13] G. Neuhold *et al.*, "The mapillary vistas dataset for semantic understanding of street scenes," in *International Conference on Computer Vision (ICCV)*, 2017. [Online]. Available: https://www.mapillary.com/dataset/vistas

[14] G. Comanici *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," 2025. [Online]. Available: https://arxiv.org/abs/2507.06261

[15] OpenAI, "Gpt-5 system card," OpenAI, Tech. Rep., August 2025, version updated August 13, 2025; PDF available at https://cdn.openai.com/gpt-5-system-card.pdf.

[16] ——, "Openai o3 and o4-mini system card," OpenAI, Tech. Rep., April 2025, pDF available at https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

[17] Anthropic, "Claude opus 4 & claude sonnet 4 system card," Anthropic, Tech. Rep., May 2025. [Online]. Available: https://www.anthropic.com/claude-4-system-card

[18] M. AI, "Mistral medium 3.1," August 2025. [Online]. Available: https://mistral.ai/models

[19] S. Bai *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[20] G. Team *et al.*, "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786