2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

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

The rise of foundation models with general-purpose capabilities has sparked a push to develop robots that are equally general-purpose—capable of flexible, wide-ranging behavior in real-world environments. Given their significant potential, it is crucial to rigorously assess how these models perform in real-world robotic applications. Such evaluations inform model development, reveal limitations, and establish benchmarks for comparing approaches. However, assessing the navigation capabilities of these models remains challenging. Navigation is critical for numerous robotic applications, including last-mile delivery, industrial inspection, search and rescue, and assistive tools for visually impaired people. Yet only limited evaluation methods currently exist for these capabilities.

Applying existing Vision Language Models (VLMs) to the navigation task itself can be straightforward: A text instruction and an image observation can be provided as input, prompting the language model to generate a text description that the robot's control system can translate into motor commands [11]–[15]. While various flavors of such systems exist, it remains unclear which VLM system performs best.

To evaluate this, there exist three main directions in the literature: The first is to perform real-world closed-loop rollouts on a set of navigation tasks and measure the success rate. However, such experiments are expensive, time-consuming, and inherently do not scale well for evaluating performance across diverse operating environments, while also lacking reproducibility. The second is to run the same closed-loop experiments in simulation. This approach improves reproducibility but still faces significant limitations. The diversity of tasks is constrained by the scenarios created in simulation, which are inherently simplified in terms of dynamics, mostly feature static scenes, and have limited semantics. Important factors, such as varying terrain properties or social norms, that should influence an agent's navigation

[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

TABLE I: An overview of vision language navigation and VLM datasets.

	Task			Data Source		Annotations			
Name	Type	VLM	Nav.	Description	Sim / Real	Description	Automatic / Manual	Scoring	Embodiments
R2R [1]	VLN	✗	✓	MP3D [2]	Sim	21,567 language navigation instructions	Auto/Manual	Success Rate, Navigation Error	Human
REVERIE [3]	VLN	✗	✓	MP3D [2]	Sim	21,702 language instructions that require navigating and identifying an object	Auto/Manual	Success Rate, SPL	Human
RxR [4]	VLN	✗	✓	MP3D [2]	Sim	126k multilingual time-aligned language instructions, 126k demonstration paths	Auto/Manual	Success Rate, SPL, Navigation Error, Normalized Dynamic Time Warping	Human
OctoNav-Bench [5]	Embodied Navigation	✗	✓	Habitat (MP3D, HM3D, Gibson, ProcTHOR)	Sim	45k+ annotated instructions with trajectories that combine the task types: object goal, point goal, image goal, instance-image goal, and VLN, 10k+ instruction-think-action pairs	Auto	Success Rate, SPL	Human
EgoWalk [6]	VLN	✗	✓	50 h of egocentric navigation recordings	Real	Automatic traversability region and language goals with extracted odometry trajectories	Auto	MSE, Absolute Displacement Error, and Final Displacement Error	Human
CityWalker [7]	Point Goal Nav.	✗	✓	2000 h of city walking videos from the internet	Real	With visual odometry extracted trajectory poses	Auto	Average Orientation Error	Human
SocialNav-SUB [8]	Social Navigation VQA	✓	✗	SCAND [9]	Real	4968 unique questions, 24840 human responses	Auto/Manual	Probability of Agreement, Consensus-Weighted Probability of Agreement	Quadruped, Wheeled
Social-LLaVA [10]	Social Navigation VQA	✓	✗	SCAND [9]	Real	40k questions fully annotated by humans	Manual	Human judges	Quadruped, Wheeled
NaviTrace (ours)	Nav. Traces for VLMs	✓	✓	1000 diverse real images	Real	1k language instructions, 3k+ traces that describe 2D paths	Manual	Human-preference aligned nav. metric	Human, Quadruped, Wheeled, Bicycle

behavior, are difficult to encode. Lastly, Visual Question Answering (VQA) benchmarks can overcome some of these limitations by making use of high-quality human annotations, which can incorporate semantics, social preferences, and geometric cues. While several navigation-focused VQA datasets exist, they typically (i) constrain outputs to text answers rather than trace-level plans, and (ii) evaluate only legged or wheeled robot embodiments [8], [10].

To fill this gap, 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" [11] 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 [16], [17] and widely used to assess the visual grounding of VLMs [18]. Furthermore, traces have proven beneficial for addressing manipulation tasks [19]–[22]. 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 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. RELATED WORK

Table I provides an overview of benchmarks evaluating the navigation performance of vision-based agents and VLMs.

Vision Language Navigation Benchmarks. Several relevant works focus on the evaluation of vision-language navigation (VLN) tasks, where agents follow natural language instructions using visual input. Well-established benchmarks include Room-to-Room (R2R) [1], REVERIE [3], and Room-Across-Room (RxR) [4]. R2R and RxR feature fine-grained instructions, while REVERIE uses coarser descriptions that also require object identification (e.g., "Bring me the bottom picture next to the top of the stairs on level one" [3]). All three benchmarks rely on Matterport3D (MP3D) scenes [2] in simulation, which provides realistic indoor environments but restricts navigation to discrete viewpoint transitions without realistic physics simulation. OctoNav-Bench [5] extends VLN benchmarks by combining multiple task types into free-form instructions. It leverages the Habitat simulator [23] that supports continuous action spaces. While simulators enable the training of reinforcement learning policies, they remain constrained to the underlying indoor training environments and often fail to accurately model the physical interactions corresponding to visual observations, which is one of the main causes of the visual sim-to-real gap. Other benchmarks address this limitation by directly collecting real-world data. EgoWalk [6] records egocentric navigation with annotated language goals and extracted trajectories. CityWalker [7] uses internet videos and visual odometry to extract trajectories, and therefore does not have language-conditioned tasks.

Existing VLN benchmarks present several limitations for VLM evaluation. Most require trajectory predictions in

specialized action spaces that VLMs cannot natively predict. They focus exclusively on human navigation, overlooking cross-embodiment challenges. To address these gaps, NaviTrace uses manually collected real-world images and VLM-accessible 2D trace prediction across multiple embodiment types.

Vision-Language Model Benchmarks. There exists a variety of VLM benchmarks, which adapt standard computer vision tasks into VLM-compatible formats [25]–[27], test embodied skills such as pointing [18], spatial understanding [28], or embodied reasoning [16]. Most relevant to NaviTrace are benchmarks in social navigation. SocialNav-Sub [8] evaluates models through VQA on videos of robot-human interactions, asking about spatial relations, robot and pedestrian motion, as well as interaction dynamics. Similarly, the VQA dataset SNEI [10] contains social scenarios in crowded spaces, asking models to describe perceptions, predict future movements, reason about robot actions, and give a general explanation of what is happening. However, to the best of our knowledge, no existing benchmark directly evaluates VLMs on navigation tasks or their understanding of differences between embodiments when navigating.

III. 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 [24]. To preserve privacy, we anonymize all personal data using EgoBlur [29] 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:** Decisions based on the shape, structure, or 3D geometry of permanent terrain features (e.g., stairs, a cliff, or closed doors).
- **Semantic Terrain Property Assessment:** Decisions requiring semantic understanding of properties (e.g., sidewalk, or road), or physical qualities (e.g., terrain stiffness, or friction).
- **Accessibility:** Barrier-free access for embodiments such as wheelchairs or delivery robots (e.g., wheelchair ramps, or automatically opening doors).
- **Visibility:** Scenarios with occlusions, poor lighting, or ambiguous information (e.g., blocked lines-of-sight, or unclear signage).
- **Social Norms:** Normative constraints from rules or signage (e.g., crosswalks, walking on a pedestrian walkway, or following a sign to not step on grass).
- **Dynamic Obstacle Avoidance:** Reacting to and planning around moving obstacles (e.g., humans, or vehicles).
- **Stationary Obstacle Avoidance:** Navigation around fixed obstacles not part of the general terrain structure (e.g., debris, or road closures).

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:** A regular pedestrian unable to climb tall obstacles.
- **Legged Robot:** A quadruped (e.g., ANYmal [30]) with behavior similar to humans but shorter in stature.
- **Wheeled Robot:** A small, wheelchair-like delivery robot that favors walkways and ramps.
- **Bicycle:** A cyclist following traffic rules, preferring bike lanes or streets, and avoiding stairs.

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

B. Data Quality

To ensure scenario diversity, we analyze the dataset along five factors: (i) geographical location, (ii) urban vs. rural setting, (iii) natural vs. structured environment, (iv) lighting conditions, and (v) weather. The geographic distribution is shown on the left in Figure 2. While the dataset is geographically concentrated in Switzerland, it also includes samples from several other countries to provide broader international representation.

The right side of Figure 2 summarizes the distribution of scenarios across the remaining factors. The scenarios are

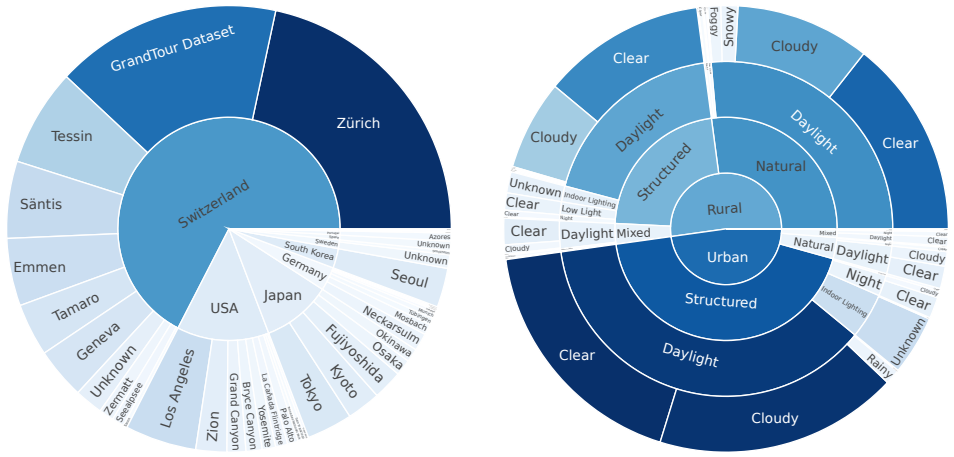

Fig. 2: **Left:** Geographic distribution of image sources, with the inner circle denoting countries and the outer circle specifying cities or regions. Images originating from the GrandTour Dataset [24] are explicitly marked in the outer circle. **Right:** Distribution of scenarios by setting (urban vs. rural), environment type (natural vs. structured), lighting, and weather.

balanced between an urban and rural setting. Structured environments appear more frequently than natural ones, because urban scenes rarely contain natural elements. Most images were captured in daylight under clear or cloudy weather, resulting in high visual quality. This shows a tendency toward favorable conditions for vision, however the benchmark primarily targets navigation challenges rather than visual perception under difficult conditions.

C. 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) [31] 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 [32] trained on Mapillary Vistas [33], 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:

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

Scaling: In order to make the score values easier to interpret, we scale them to a range where the worst score is at 0 and the best score at 100. We achieve this by setting the ground-truth performance 0 as the lower bound. 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:

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

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

Evaluation: We acknowledge that, for each term, within our proposed score function, different choices may be made. Complexities can arise from defining the exact start and end points, accounting for whether the path intersects hazardous terrain, and recognizing that distances measured in 2D image space do not directly translate to 3D navigation behavior. For

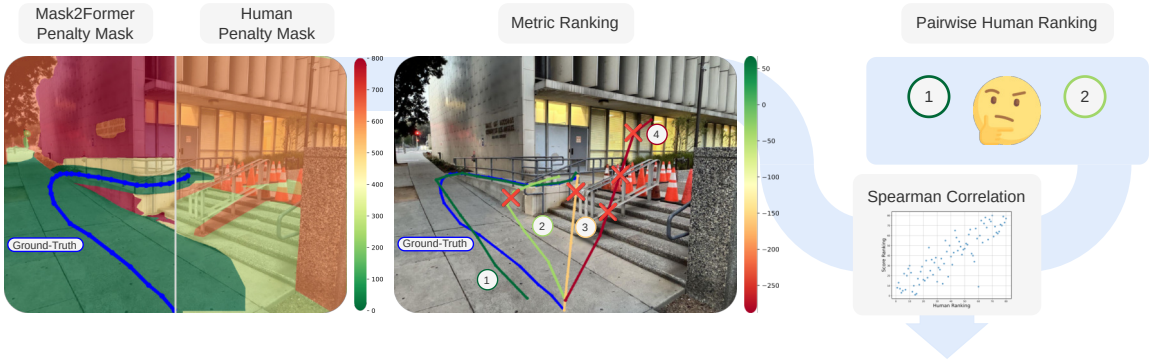

Fig. 3: **Left:** Comparison between penalty cost masks based on Mask2Former and manual segmentation. These masks are used to punish traces crossing unsafe or irrelevant areas. **Right:** We show that the score function aligns with human preference by calculating the correlation between the score ranking and a pairwise ranking created by a human.

example, at greater distances, higher accuracy is required to follow a path, although in practice, a receding-horizon control approach may be applied.

Given these complexities and challenges, it is essential to evaluate whether the proposed score function aligns with human judgments. To do so, we compute Spearman correlations [34] between a human pairwise ranking of predictions and several score variants. To cover the full quality range of predictions, we compile an equal mix of human, model, baseline, and intentionally flawed predictions. Annotators perform pairwise comparisons to produce a human ranking, which we correlate with each score variant using Spearman's rank correlation. This correlation ranges from perfect agreement (± 1) to no relation (0), ignoring linear relationships in favor of rank order. The procedure is illustrated in Figure 3.

TABLE II: Spearman correlation between variants of the score function and human ranking.

Score Variant	Spearman Correlation [↑]
RMSE	0.8167
Fréchet	0.8310
DTW	0.8417
DTW + FDE	0.8656
DTW + FDE + Manual Penalty	**0.8723**
DTW + FDE + Mask2Former **(ours)**	0.8707

We begin by comparing plain DTW similarity with alternative measures such as root mean square error (RMSE) and discrete Fréchet distance (see Table II), and observe that DTW consistently achieves the highest performance across all three trace-similarity metrics. We then evaluate the additional contribution of the FDE term for goal-reaching and find a further, consistent improvement in performance. Extending the score with our Mask2Former-based semantic penalty leads to another clear performance gain. To assess how well these automatically derived semantic cost terms align with expert annotations, we asked human annotators to semantically segment images into task-relevant, irrelevant (but safe), and hazardous regions (see Figure 3). While dense manual semantic labeling is resource-intensive, it offers only limited performance gains over our Mask2Former-based strategy. Taken together, these findings validate our decision to combine all three terms.

IV. 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 five 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 five 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.
- **Oracle-Goal Straight Line:** Connects the given start and goal points with a direct line. In contrast to VLMs, this method has direct access to the goal and start point.
- **Only predict goal point:** To isolate the difficulty of identifying the goal, we use Gemini 2.5 Pro to predict only the goal location and connect it to the given start via a straight line.
- **Only predict path:** Conversely, given both start and goal, Gemini 2.5 Pro predicts only the navigation path.

Together, these baselines capture informed strategies, an upper bound with human performance, and provide context for assessing VLM performance.

B. Models

We evaluate all VLMs by querying each model through API calls. After preliminary testing, we select five representative

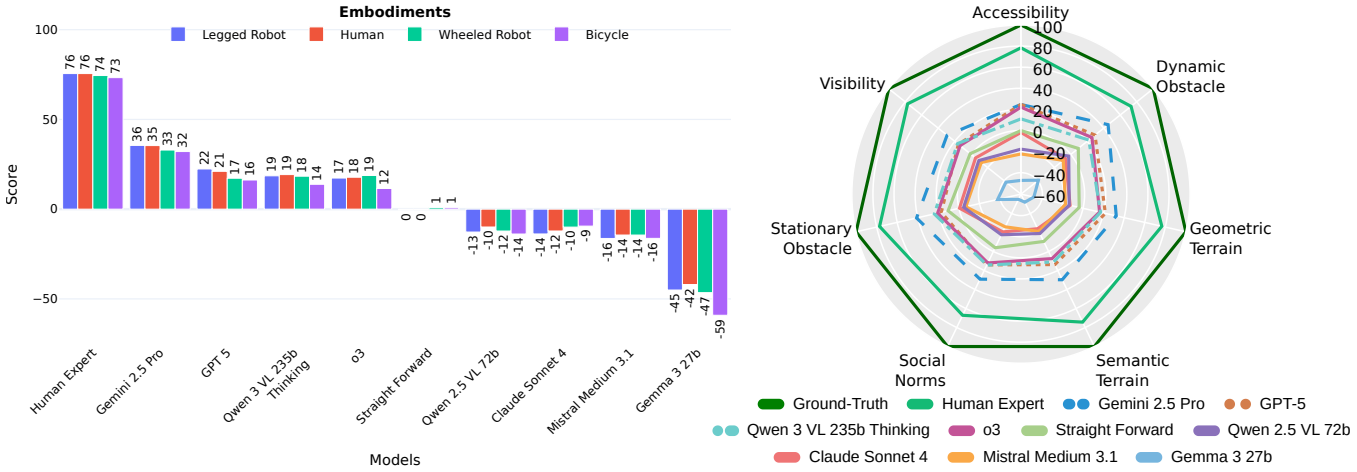

Fig. 4: **Left:** Ranking of VLMs, the uninformed baseline Straight Forward, and human expert performance split into each embodiment. Note that a higher score is better. **Right:** Performance per task category for the same models.

proprietary models: Gemini 2.5 Pro [35], GPT-5 [36], o3 [37], Claude Sonnet 4 [38], and Mistral Medium 3.1 [39]. We also include three open-weight models: Qwen 2.5 VL 72B [40], Qwen 3 VL 235B A22B Thinking [40], and Gemma 3 27B [41]. 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 4). 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. 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 5. Generally, we do not observe significant differences between embodiment types for all the models.

Turning to task categories in Figure 4 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.

Our experiments demonstrate that Gemini 2.5 Pro achieves the best overall performance in general navigation capabilities. To gain a clearer understanding of the challenges involved in navigation, we decompose the task into goal-point prediction and path-shape prediction. Therefore, we compare Gemini 2.5 Pro with baseline models that have access to privileged information as well as with human experts (see Table III). Using Gemini 2.5 Pro to only predict the goal point and then connecting it with a straight line yields only slightly worse results than having Gemini 2.5 Pro predict the full trace. While baselines with explicit access to the goal point perform significantly better, suggesting that locating the goal area is already a major challenge. In particular, predicting only a path shape with Gemini 2.5 Pro performs better than Oracle-Goal Straight Line, showing that the model possesses a basic understanding of the scenarios. However, even when providing the goal point, Gemini 2.5 Pro falls short of human expert performance without this advantage. Overall, Gemini 2.5 Pro struggles especially in recognizing goal areas but also underperforms in shaping meaningful paths, highlighting the dual difficulty of the task.

TABLE III: Comparison between informed baselines, a human and Gemini 2.5 Pro. Note that a higher score is better.

Model	Score [↑]
Only goal point with Gemini 2.5 Pro	29.65
Gemini 2.5 Pro	34.38
Oracle-Goal Straight Line	51.89
Only path with Gemini 2.5 Pro	56.55
Human Expert	**75.40**

Finally, we provide qualitative insights into the reasoning process of models such as Gemini 2.5 Pro and o3. Figure 6 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 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 four critical insights about current VLM navigation capabilities and areas of future work:

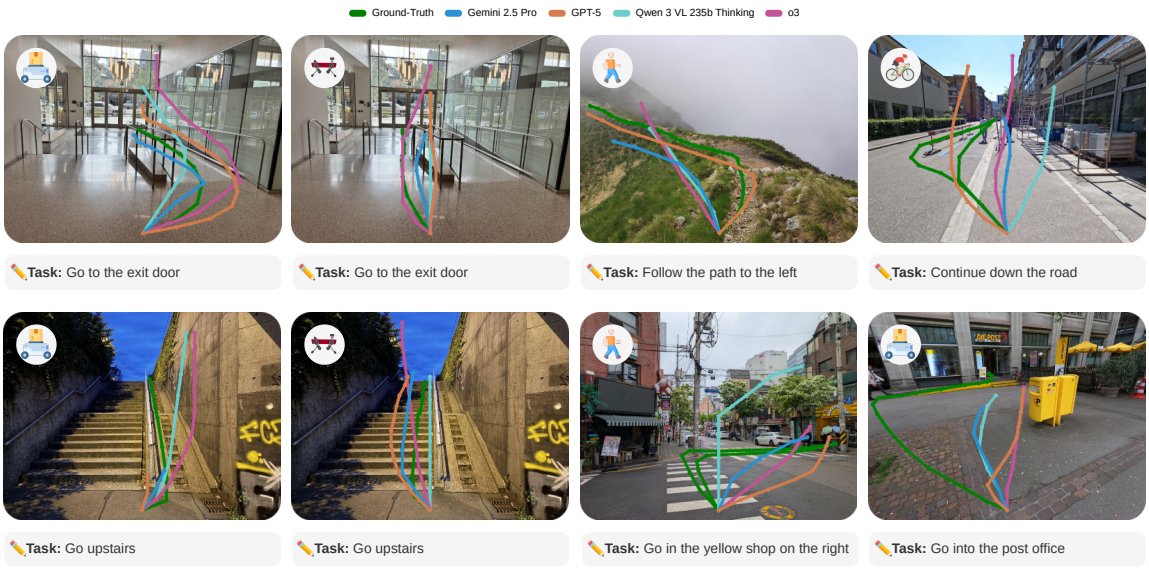

Fig. 5: Example predictions by the models Gemini 2.5 Pro, GPT-5, Qwen 3 VL, and o3.

Fig. 6: 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.

(1) Large human performance gap. Across all four embodiments and task categories, VLM scores are substantially worse than both human and oracle-like baselines, highlighting significant room for improvement (see Figure 4 and Table III). **(2) Goal localization is the dominant failure mode.** When models predict only the goal location and we connect it with a straight line, scores are similar to full-trace predictions. Yet even with the correct goal, path shaping lags behind human performance (see Table III). **(3) 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 4). **(4) Score function alignment with human preference.** Our semantic-aware trace score, that builds on the DTW distance [31] with endpoint error and embodiment-conditioned penalties using automated semantics [32], [33], correlates more strongly with human preference than DTW alone. Using manual segmentation yields an additional but modest gain.

V. 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.

VI. 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] P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] A. Chang *et al.*, "Matterport3d: Learning from rgb-d data in indoor environments," 2017. [Online]. Available: https://arxiv.org/abs/1709.06158

[3] Y. Qi *et al.*, "Reverie: Remote embodied visual referring expression in real indoor environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] A. Ku *et al.*, "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding," 2020. [Online]. Available: https://arxiv.org/abs/2010.07954

[5] C. Gao *et al.*, "Octonav: Towards generalist embodied navigation," 2025. [Online]. Available: https://arxiv.org/abs/2506.09839

[6] T. Akhtyamov *et al.*, "Egowalk: A multimodal dataset for robot navigation in the wild," 2025. [Online]. Available: https://arxiv.org/abs/2505.21282

[7] X. Liu *et al.*, "Citywalker: Learning embodied urban navigation from web-scale videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 6875–6885.

[8] M. J. Munje *et al.*, "Socialnav-SUB: Benchmarking VLMs for scene understanding in social robot navigation," in *ICRA 2025 Workshop: Human-Centered Robot Learning in the Era of Big Data and Large Models*, 2025. [Online]. Available: https://openreview.net/forum?id=cCuylmKVXq

[9] H. Karnan *et al.*, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 807–11 814, 2022.

[10] A. Payandeh *et al.*, "Social-llava: Enhancing robot navigation through human-language reasoning in social spaces," 2024. [Online]. Available: https://arxiv.org/abs/2501.09024

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

[12] J. Zhang *et al.*, "Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks," 2025. [Online]. Available: https://arxiv.org/abs/2412.06224

[13] R. Dang *et al.*, "Rynnbrain: Open embodied foundation models," *arXiv preprint arXiv:2602.14979v1*, 2026. [Online]. Available: https://arxiv.org/abs/2602.14979v1

[14] Z. Chu *et al.*, "Abot-n0: Technical report on the vla foundation model for versatile embodied navigation," 2026. [Online]. Available: https://arxiv.org/abs/2602.11598

[15] Z. Chen *et al.*, "Socialnav: Training human-inspired foundation model for socially-aware embodied navigation," *arXiv preprint arXiv:2511.21135*, 2025.

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

[17] 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

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

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

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

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

[22] 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

[23] M. Savva *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[24] 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.

[25] X. Fu *et al.*, *BLINK: Multimodal Large Language Models Can See but Not Perceive*. Springer Nature Switzerland, October 2024, p. 148–166. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-73337-6_9

[26] S. Tong *et al.*, "Cambrian-1: A fully open, vision-centric exploration of multimodal llms," in *Advances in Neural Information Processing Systems*, A. Globerson *et al.*, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 87 310–87 356. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/9ee3a664ccfeabc0da16ac6f1f1cfe59-Paper-Conference.pdf

[27] R. Ramachandran *et al.*, "How well does gpt-4o understand vision? evaluating multimodal foundation models on standard computer vision tasks," 2025. [Online]. Available: https://arxiv.org/abs/2507.01955

[28] M. Du *et al.*, "Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.05756

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

[30] M. Hutter *et al.*, "Anymal - a highly mobile and dynamic quadrupedal robot," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, October 2016, p. 38–44. [Online]. Available: http://dx.doi.org/10.1109/IROS.2016.7758092

[31] 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.

[32] 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.

[33] 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

[34] J. Hauke and T. Kossowski, "Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data," *Quaestiones geographicae*, vol. 30, no. 2, pp. 87–93, 2011.

[35] 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

[36] 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.

[37] ——, "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.

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

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

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

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