When Agents Get Lost: Dissecting Failure Modes in Graph-Based Navigation Instruction Evaluation

Farzad Shami, Kimia Abedini, Seyed Hossein Hoessini, and Henrikki Tenkanen

Aalto University (Built Environment; Computer Science)
{farzad.shami, kimia.abedini, seyed.h.hosseini, henrikki.tenkanen}@aalto.fi

Abstract

Vision-and-Language Navigation (VLN) requires agents to interpret natural language instructions for spatial reasoning, yet evaluating instruction quality remains challenging when agents fail. This gap highlights a critical need for a principled understanding of why navigation instructions fail. Addressing this question requires a systematic analysis of failure patterns in spatial reasoning tasks. To address this, we first present a taxonomy of navigation instruction failures that clusters failure cases into four categories: (i) linguistic properties, (ii) topological constraints, (iii) agent limitations, and (iv) execution barriers. We then introduce a dataset of over 450 annotated failure navigation traces collected from GROKE, a vision-free evaluation framework that utilizes OpenStreetMap (OSM) data. Our dataset outlines the failure dynamics in spatial grounding to guide the development of better instruction generation, evaluation systems, and navigation agents. Our analysis of failure traces across GROKE demonstrates that agent limitations (74.2%) constitute the dominant error category, with stop-location errors and planning failures as the most frequent subcategories. The dataset and taxonomy together provide actionable insights that enable instruction generation systems to identify and avoid under-specification patterns while allowing evaluation frameworks to systematically distinguish between instruction quality issues and agent-specific artifacts.

https://fuzsh.github.io/lost/

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse language understanding tasks, yet their spatial reasoning abilities remain fundamentally limited. Recent benchmarks reveal that even state-of-the-art (SOTA) models experience performance degradation ranging from 42% to over 80% as spatial task complexity increases [1]. This limitation becomes particularly critical in embodied Artificial Intelligence (AI) applications, where models must translate natural language navigation instructions into executable spatial behaviors. The REM [20] benchmark shows that advanced reasoning models still underperform humans on spatial tasks, particularly when requiring viewpoint-independent reasoning and object tracking across trajectories. These findings indicate that the grounding problem between linguistic descriptions and spatial actions remains largely unresolved.

In VLN, this gap manifests through systematic failures in instruction following. Contemporary VLN systems achieve moderate success rates on standard benchmarks, yet comprehensive evaluations reveal persistent weaknesses. EmbodiedBench [22] demonstrates that multimodal language models excel at high-level planning but struggle with low-level spatial control, achieving only 28.9% average success across embodied tasks. More concerning, recent studies show that SOTA VLN methods

1st International Conference on Geospatial Artificial Intelligence (GeoAI 2026). June 3-6, 2026, Ghent, Belgium.

experience up to 25% success rate drops when evaluated with erroneous instructions [19], revealing fundamental fragility in instruction-following systems. This fragility raises a critical question: when navigation fails, is the failure attributable to agent limitations, instruction quality, or their interaction?

Existing failure analysis frameworks predominantly focus on agent-level diagnostics. Recent work has introduced taxonomies for perception errors, reasoning failures, and planning breakdowns in embodied systems [9, 22]. VLN evaluation has begun incorporating step-level error analysis to distinguish between navigation decision failures and generation errors [25]. However, these approaches assume that navigation instructions themselves are correct and complete. The instruction-level failure modes, such as linguistic ambiguity, topological inaccuracy, or insufficient spatial detail, remain largely unexamined. This represents a significant gap in understanding navigation system failures, as instruction quality fundamentally determines navigability independent of agent capability.

Vision-free evaluation frameworks offer an opportunity to isolate instruction properties from perceptual challenges. GROKE [16], a graph-to-text evaluation framework that converts OSM topology into structured textual representations for LLM-based reasoning, shows that graph-based reasoning over OSM representations can evaluate navigation instructions without visual perception. In GROKE, the navigation instruction is divided into multiple navigation steps, and the landmarks are detected using a specialized agent. For each step, the visible area is constructed from the OSM graph and fed as context to the LLM, which then attempts to execute the instruction step by navigating through the graph representation. By removing computer vision from the evaluation loop, such approaches enable direct assessment of whether instructions contain sufficient, correct, and actionable spatial information. This capability underpins systematic instruction-level failure analysis that cannot be achieved through vision-dependent evaluation methods.

By removing computer vision from the evaluation loop, these approaches enable direct assessment of whether instructions contain sufficient and actionable spatial information, supporting systematic failure analysis not achievable through vision-dependent methods.

Contributions. We propose instruction-level failure understanding by systematically analyzing navigation instruction failures. Our contributions are as follows: (i) we collect and analyze 492 navigation failure traces from GROKE's evaluation of Map2Seq [14] test sets, representing 35.14% of evaluated instructions; (ii) we develop a hierarchical four-axis categorization framework spanning linguistic properties, topological constraints, agent-specific limitations, and execution-level breakdowns; (iii) we reveal that agent limitations represent the dominant error dimension, with stop-location errors and planning failures as the most frequent subcategories; and (iv) we derive six actionable design implications for improving vision-free navigation systems based on the identified failure patterns.

The paper proceeds as follows: Section 2 reviews related work, Section 3 presents our taxonomy and annotation methodology, Section 4 discusses results, and Section 5 concludes.

2 Related Work

Language-Guided Navigation and Spatial Reasoning. The effectiveness of LLMs in navigation is fundamentally dependent on how spatial information is encoded and represented. Contemporary approaches have attempted to address this through explicit reasoning and specialized architectures. NavGPT [26] emphasizes explicit reasoning processes, while VELMA [15] focuses on the verbalization embodiment of LLM agents in street-view environments. Building on these foundations, other methods have optimized spatial representations: MapGPT [5] demonstrates the value of map-guided prompting combined with adaptive path planning, and STMR [7] employs a semantic-topo-metric representation that combines semantic labels with topological connectivity to guide aerial navigation.

However, recent studies indicate that spatial reasoning capabilities deteriorate rapidly as problem scale and compositional complexity increase. Martorell [12] reveals that while models exhibit moderate competence in simple, direct spatial tasks, they struggle with complex compositional reasoning in grid-world contexts. Similarly, REM [20], a benchmark evaluating embodied spatial reasoning through multi-frame trajectories, highlights systematic limitations in spatial understanding. Addressing these representation gaps, GROKE [16] utilizes a vision-free evaluation framework on OSM-derived spatial graphs to demonstrate that structured JSON and textual formats substantially outperform the grid-based representations often used in earlier iterations.

Despite these advances, understanding of failure modes in navigation instruction evaluation remains less explored, limiting our ability to improve instruction generation and evaluation frameworks.

Related Datasets and Failure Studies. Systematic failure analysis has gained increasing attention across AI research domains [3, 13, 17]. However, in robotics and embodied AI, existing VLN research provides only anecdotal failure examples or limited error categorization. Researchers analyze failure modes in VLN or Vision-and-Language Action (VLA) models for robotic manipulation [10, 21, 22]. FailSafe [10] proposes a systematic failure generation and recovery framework that categorizes manipulation failures into three fundamental modes (translation, rotation, and no-ops failures) and enables VLA models to reason about and recover from errors during task execution. EmbodiedBench [22], a benchmark for evaluating multi-modal foundation models on embodied tasks, reveals systematic limitations in spatial understanding and action execution.

VLM4VLA [23] shows that initializing models with VLMs provides consistent advantages compared to training from scratch. However, their findings reveal that a VLM's performance on general tasks does not reliably indicate its effectiveness for downstream tasks. The research establishes that the vision encoder represents the primary performance bottleneck in this domain. Feng et al. [6] further reinforce this bottleneck by exposing the fragility of visually prompted benchmarks. They show that minor non-semantic factors can drastically alter accuracy, including visual marker design characteristics (such as color and shape) and low-level inference details like JPEG compression.

Further related efforts aim to catalog challenges in human-agent interaction [2] and summarize failures for specific task items by attributing them to particular agents and error steps [24], but focus on interaction patterns or task-specific code debugging. While MAST [4] pioneered empirically derived datasets and taxonomies for multi-agent system failure patterns, our work represents, to our knowledge, the first such effort focused specifically on navigation instruction failures.

3 Failure Taxonomy & Annotation.

This section describes the hierarchical taxonomy we developed to categorize navigation failures and the annotation protocol we followed to label the dataset. First, we explain the data source and failure instance collection from GROKE's evaluation of Map2Seq. Second, we present the taxonomy structure and its four main dimensions. And third, we describe the annotation process, inter-annotator agreement analysis, and disagreement resolution procedures.

Data Collection. We analyzed navigation failure traces from GROKE, representing the environment as a spatial graph where nodes correspond to decision points and edges represent navigable paths. The system processes natural language instructions through multi-step reasoning, extracting landmarks and spatial relations to construct an executable route plan. We define failures as cases where the agent's stopping location exceeds 25 meters from the target destination, a threshold commonly used in embodied navigation benchmarks to distinguish successful task completion from spatial errors.

Hierarchical Taxonomy. We developed a hierarchical taxonomy by analyzing failure patterns and agentic system design methodologies reported in prior navigation research [4, 11, 15, 16, 18, 22, 25]. By examining documented error cases and failure modes across different navigation frameworks, we identified recurring patterns that informed our category design. The resulting taxonomy represents an original classification scheme that consolidates insights from the literature while introducing new subcategories specific to vision-free navigation with graph-based spatial reasoning.

The taxonomy covers four dimensions: Linguistic (`L`), Topological (`T`), Agent (`A`), and Execution (`E`). Each dimension includes multiple subcategories designed to capture different types of failure modes and ambiguities that occur during navigation tasks. Figures 1–2 present the complete breakdown of categories and subcategories. Note that we excluded the vision-related error category from the Agent (`A`) dimension because our traces operated in a vision-free setting.

Data Annotation Three expert annotators (i.e., authors) with backgrounds in computational linguistics and spatial reasoning participated in the annotation process. For the data annotation process, annotators received the reasoning for each step, the identified sub-goals, all extracted POIs, the map with the annotated path, the graph network showing the traversed path with marked POIs, and the hierarchical taxonomy.

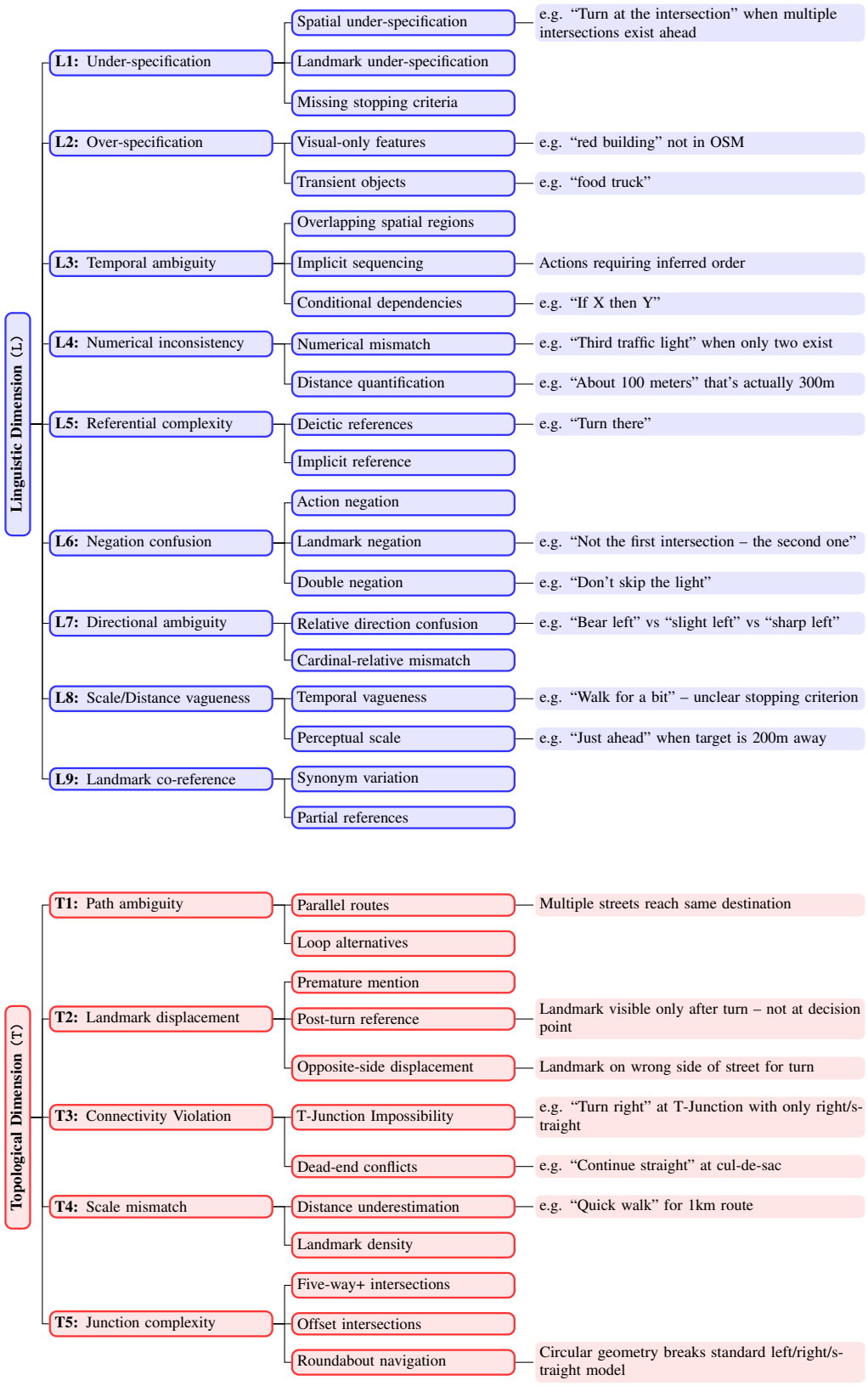

Figure 1: Linguistic and topological failure categories with subcategory definitions.

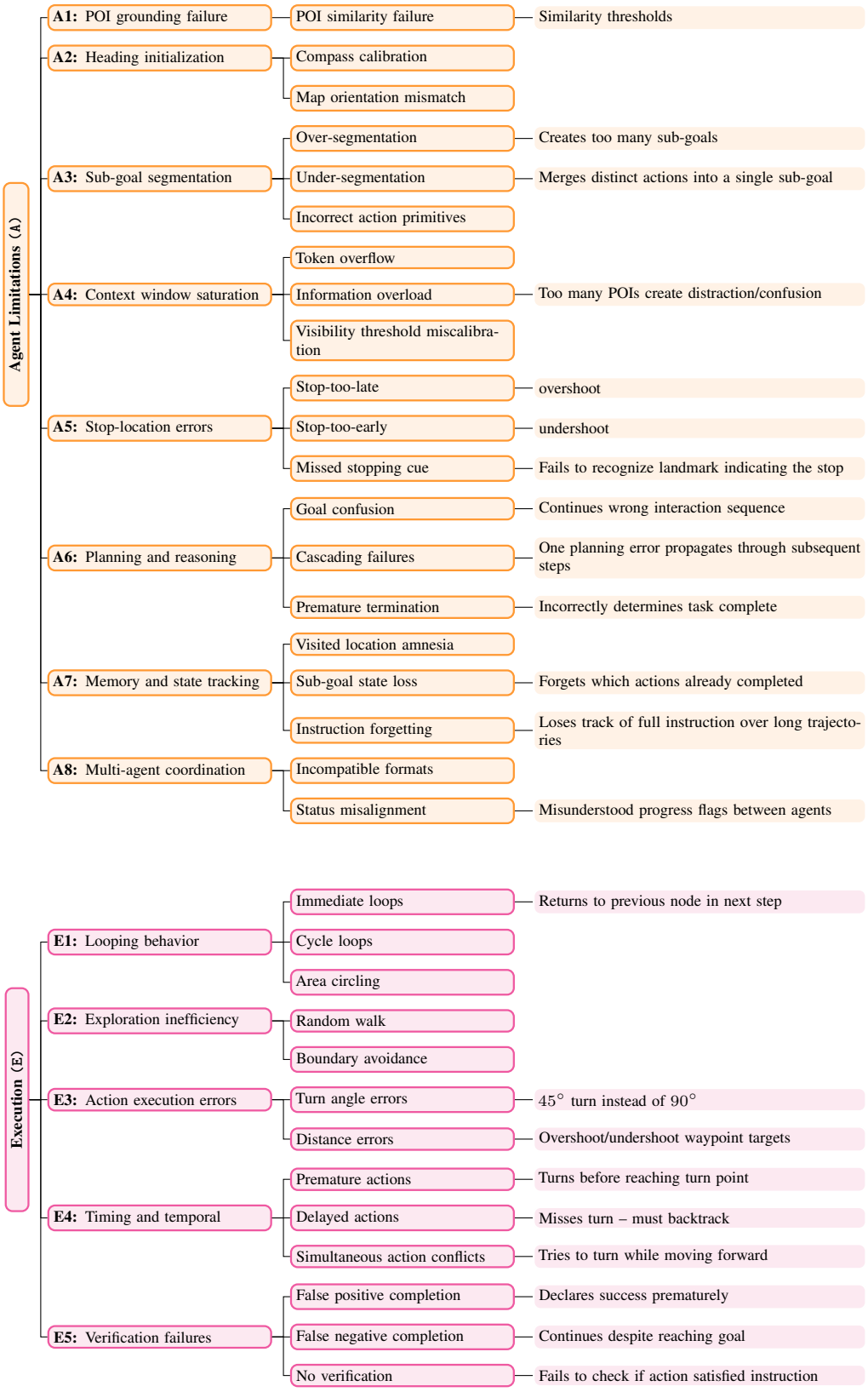

Figure 2: Agent limitations and execution failure categories with subcategory definitions.

We categorized navigation instructions at three levels: the top level distinguishes among `L`, `T`, `A`, and `E` error types; the mid-level groups errors into primary subcategories within each main type (e.g., `L1` and `L2` for linguistic errors); and the leaf level provides the most fine-grained classification of specific error instances (e.g., `L1.1` and `L1.2` as subcategories of `L1`). We measured inter-annotator agreement using Cohen's kappa (κ) as the evaluation metric. The mid-level category achieved $\kappa = 0.68$, which falls into the *"Substantial"* agreement range according to [8]. This demonstrates strong annotation reliability at this critical hierarchical level.

For the resolution of annotation disagreements, we applied a hierarchical approach based on the disagreement level. When annotators disagreed at the leaf level of the hierarchy, we assigned the deepest common ancestor as the final label. When disagreements occurred at the top-level or mid-level categories, we used a third expert adjudicator to resolve them. The final annotated dataset can be explored through our project page https://fuzsh.github.io/lost/explorer-tool.

4 Results & Discussions

As shown in Table 1, our analysis of 492 failed navigation instances reveals that errors are predominantly attributed to agent limitations, appearing in 74.2% of all cases. Execution and behavioral failures follow with 46.5%, while linguistic properties account for 23.2%, and topological constraints contribute to 12.4% of failures. Notably, half of all samples exhibit errors spanning multiple dimensions, indicating that navigation failures frequently result from compounded issues rather than isolated problems. This distribution suggests that while instruction quality and environmental complexity play important roles, the primary bottleneck lies in the agent's internal processing capabilities.

Table 1: Taxonomy of outdoor navigation agent failures across four dimensions.

	Linguistic		**Topological**		**Agent**		**Execution**	
	Code	**% of Ling.**	**Code**	**% of Topo.**	**Code**	**% of Agent**	**Code**	**% of Exec.**
Failure Taxonomy	L1	24.6%	T1	18.0%	A1	27.4%	E1	0.4%
	L2	42.1%	T2	4.9%	A2	3.3%	E2	32.8%
	L3	5.3%	T3	1.6%	A3	6.6%	E3	3.9%
	L4	6.1%	T4	4.9%	A4	8.8%	E4	49.8%
	L5	7.9%	T5	73.8%	A5	49.9%	E5	39.3%
	L6	1.8%		–	A6	32.1%		–
	L7	14.9%		–	A7	0.8%		–
	L8	14.0%		–	A8	3.0%		–
	L9	0.9%		–		–		–
Prevalence	**23.2%**		**12.4%**		**74.2%**		**46.5%**	

Note: **Critical** (dark red), **High** (orange), **Medium** (blue), **Low** (gray). Rate: % within dimension. Refer to Figures 1–2 for code details.

Agent limitations represent the most prevalent source of failures. Within this dimension, stop-location errors (`A5`) constitute the dominant failure mode, occurring in 49.9% of agent-related failures (182 instances). These errors manifest when agents either overshoot their destination, stop prematurely before reaching the described location, or fail to recognize landmarks indicating the stopping point. Planning and reasoning errors (`A6`) follow at 32.1% (117 instances), encompassing cases where agents exhibit goal confusion, cascading planning failures, or premature task termination. POI grounding failures (`A1`) account for 27.4% (100 instances), representing situations where fuzzy string matching or semantic similarity computations fail to correctly identify landmarks mentioned in the instructions.

Execution failures represent the second most common error dimension, appearing in 46.5% of all annotated samples. Timing and temporal errors (`E4`) dominate this category at 49.8% (114 instances within the execution dimension), occurring when agents perform correct actions in incorrect sequences or with inappropriate timing. A representative example involves an agent that correctly analyzed instructions and map data, noting that *"poi_x"* is to its left with a traffic light behind it, but failed because the relevant POIs had already been passed during navigation.

Linguistic errors contribute to 23.2% of navigation failures, highlighting the critical importance of instruction quality. Over-specification (`L2`) represents the most common linguistic failure mode at 42.1% (48 instances), occurring when instructions reference landmarks not visible or not present in the map data. This includes visual-only features such as details not captured in OSM.

Topological errors appear in 12.4% of failures, with junction complexity (T5) representing the majority with 73.8% (45 instances). Multi-way intersections and offset intersections with non-perpendicular streets create exponential action space complexity that challenges vision-free navigation systems. For example, one annotation noted that understanding *"This road will merge with another road. You will take a slight right to stay on the road"* requires a better understanding of the surrounding area, suggesting that visual input may be necessary for such topologically complex scenarios.

Implications for System Design. Our error analysis reveals several actionable insights for improving vision-free navigation systems. We organize these implications according to the primary failure dimensions identified in our taxonomy.

(1) Spatial Representation Limitations: A recurring issue involves the mismatch between how geographic features are represented in OSM and their real-world spatial extent. Large areas such as parks are often encoded as single point coordinates rather than polygonal regions, which leads agents to treat expansive landmarks as precise locations. Similar problems occur with corner landmarks located on the opposite side of the street, where the point representation fails to capture the spatial relationship between the navigator and the landmark.

(2) Ambiguous Terminology Interpretation: Certain navigation terms, such as *"block"* and *"T-intersection"* introduce systematic interpretation challenges. Our analysis indicates that even human annotators occasionally disagree on the intended meaning of these expressions. This suggests that navigation systems would benefit from explicit disambiguation mechanisms or contextual reasoning modules that can resolve such terminological ambiguity.

(3) Landmark Detection and Matching: Groke's current architecture relies on LLM-based landmark extraction from navigation instructions and then fuzzy matching. A notable example involves traffic lights, which appear frequently in the Map2seq dataset. While instructions typically reference *"light"*, the corresponding OSM nodes use the tag *"traffic_signal."* Standard fuzzy string matching fails to bridge this lexical gap. Implementing improved semantic mappings between instruction vocabulary and OSM tag conventions could substantially reduce planning, reasoning, and stop-location errors.

(4) Action Timing at Intersections: We identified cases where agents misinterpret the spatial context for executing actions. When positioned at an intersection with a traffic light and receiving an instruction such as *"turn right at the light,"* agents sometimes execute the turn immediately at the current position. However, the intended interpretation often requires advancing to the next traffic light before performing the action. This finding highlights the need for better temporal grounding of action directives for the first action.

(5) Junction Complexity Handling: The current sub-goal detection mechanism discards information about turn sharpness and angle, which proves insufficient for complex junctions. Enhancing the sub-goal extraction to preserve angular information would enable more informed decision-making at multi-way intersections. Alternatively, developing a richer representation scheme for complex junctions could improve the model's ability to distinguish between multiple exit options.

(6) Stop-Location Refinement: Given that many stop-location errors result in positions that are approximately correct, a practical improvement would involve incorporating visual input specifically for the final navigation steps. Alternatively, providing richer contextual information about surrounding POIs during the last-step reasoning phase could improve stopping accuracy without requiring full visual perception throughout the navigation task.

5 Final Remarks

We introduced a hierarchical taxonomy for categorizing navigation instruction failures across four dimensions and presented an annotated dataset of failure traces from vision-free evaluation settings. Our analysis revealed that agent limitations, particularly stop-location and planning errors, constitute the dominant failure sources, while half of all failures exhibit multi-dimensional error patterns. These findings provide actionable insights for improving instruction generation systems and evaluation frameworks. Although our analysis is limited to a single evaluation framework and excludes vision-dependent failure modes, the taxonomy and dataset offer a foundation for developing more robust navigation systems and for future research on automated failure diagnosis.

References

[1] M. Bai, A. K. Cohen, E. Koss, and C. Lichtenbaum. Stuck in the matrix: Probing spatial reasoning in large language models. *arXiv preprint arXiv:2510.20198*, 2025.

[2] G. Bansal, J. W. Vaughan, S. Amershi, E. Horvitz, A. Fourney, H. Mozannar, V. Dibia, and D. S. Weld. Challenges in human-agent communication. *arXiv preprint arXiv:2412.10380*, 2024.

[3] P. Bryan, G. Severi, J. de Gruyter, D. Jones, B. Bullwinkel, A. Minnich, S. Chawla, G. Lopez, M. Pouliot, A. Fourney, et al. Taxonomy of failure mode in agentic ai systems. *Microsoft AI Red Team*, 2025.

[4] M. Cemri, M. Z. Pan, S. Yang, L. A. Agrawal, B. Chopra, R. Tiwari, K. Keutzer, A. Parameswaran, D. Klein, K. Ramchandran, M. Zaharia, J. E. Gonzalez, and I. Stoica. Why do multi-agent LLM systems fail? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=fAjbYBmonr.

[5] J. Chen, B. Lin, R. Xu, Z. Chai, X. Liang, and K.-Y. Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9796–9810, 2024.

[6] H. Feng, L. Lian, L. Dunlap, J. Shu, X. Wang, R. Wang, T. Darrell, A. Suhr, and A. Kanazawa. Visually prompted benchmarks are surprisingly fragile, 2026. URL https://arxiv.org/abs/2512.17875.

[7] Y. Gao, Z. Wang, L. Jing, D. Wang, X. Li, and B. Zhao. Aerial vision-and-language navigation via semantic-topo-metric representation guided llm reasoning. *arXiv preprint arXiv:2410.08500*, 2024.

[8] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[9] M. Li, S. Zhao, Q. Wang, K. Wang, Y. Zhou, S. Srivastava, C. Gokmen, T. Lee, L. E. Li, R. Zhang, W. Liu, P. Liang, L. Fei-Fei, J. Mao, and J. Wu. Embodied agent interface: Benchmarking LLMs for embodied decision making. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=iSwK1YqO7v.

[10] Z. Lin, J. Duan, H. Fang, D. Fox, R. Krishna, C. Tan, and B. Wen. Failsafe: Reasoning and recovery from failures in vision-language-action models. *arXiv preprint arXiv:2510.01642*, 2025.

[11] T. Ma, Y. Zhang, Z. Wang, and P. Kordjamshidi. Breaking down and building up: Mixture of skill-based vision-and-language navigation agents. In *NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI*, 2025. URL https://openreview.net/forum?id=vEL31CS2Wi.

[12] N. Martorell. From text to space: Mapping abstract spatial models in llms during a grid-world navigation task. In *World Conference on Explainable Artificial Intelligence*, pages 268–291. Springer, 2025.

[13] I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst. The fallacy of ai functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972, 2022.

[14] R. Schumann and S. Riezler. Generating landmark navigation instructions from maps as a graph-to-text problem. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 489–502, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.41. URL https://aclanthology.org/2021.acl-long.41/.

[15] R. Schumann, W. Zhu, W. Feng, T.-J. Fu, S. Riezler, and W. Y. Wang. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18924–18933, 2024.

[16] F. Shami, S. Dey, N. Van de Weghe, and H. Tenkanen. Groke: Vision-free navigation instruction evaluation via graph reasoning on openstreetmap. *arXiv preprint arXiv:2601.07375*, 2026.

[17] M. Srikumar, J. Pratt, K. Chmielinski, C. Ashurst, C. Bakalar, W. Bartholomew, R. Bommasani, P. Cihon, R. Crootof, M. Hoffmann, et al. Prioritizing real-time failure detection in ai agents. *Partnership on AI*, 2025.

[18] Y. Sun, Y. Qiu, Y. Aoki, and H. Kataoka. Outdoor vision-and-language navigation needs object-level alignment. *Sensors*, 23(13):6028, 2023.

[19] F. Taioli, S. Rosa, A. Castellini, L. Natale, A. Del Bue, A. Farinelli, M. Cristani, and Y. Wang. Mind the error! detection and localization of instruction errors in vision-and-language navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12993–13000. IEEE, 2024.

[20] J. Thompson, E. Garcia-Lopez, and Y. Bisk. Rem: Evaluating llm embodied spatial reasoning through multi-frame trajectories. In *Second Conference on Language Modeling*.

[21] K. Vo, T. Hanyu, Y. Ikebe, T. T. Pham, N. Chung, M. N. Vu, D. N. H. Minh, A. Nguyen, A. Gunderman, C. Rainwater, et al. Clutter-resistant vision-language-action models through object-centric and geometry grounding. *arXiv preprint arXiv:2512.22519*, 2025.

[22] R. Yang, H. Chen, J. Zhang, M. Zhao, C. Qian, K. Wang, Q. Wang, T. V. Koripella, M. Movahedi, M. Li, H. Ji, H. Zhang, and T. Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DgGF2LEBPS.

[23] J. Zhang, X. Chen, Q. Wang, M. Li, Y. Guo, Y. Hu, J. Zhang, S. Bai, J. Lin, and J. Chen. Vlm4vla: Revisiting vision-language-models in vision-language-action models, 2026. URL https://arxiv.org/abs/2601.03309.

[24] S. Zhang, M. Yin, J. Zhang, J. Liu, Z. Han, J. Zhang, B. Li, C. Wang, H. Wang, Y. Chen, et al. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *arXiv preprint arXiv:2505.00212*, 2025.

[25] X. Zhao, G. Zhou, and Q. Wu. Vln-mme: Diagnosing mllms as language-guided visual navigation agents. *arXiv preprint arXiv:2512.24851*, 2025.

[26] G. Zhou, Y. Hong, and Q. Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649, 2024.