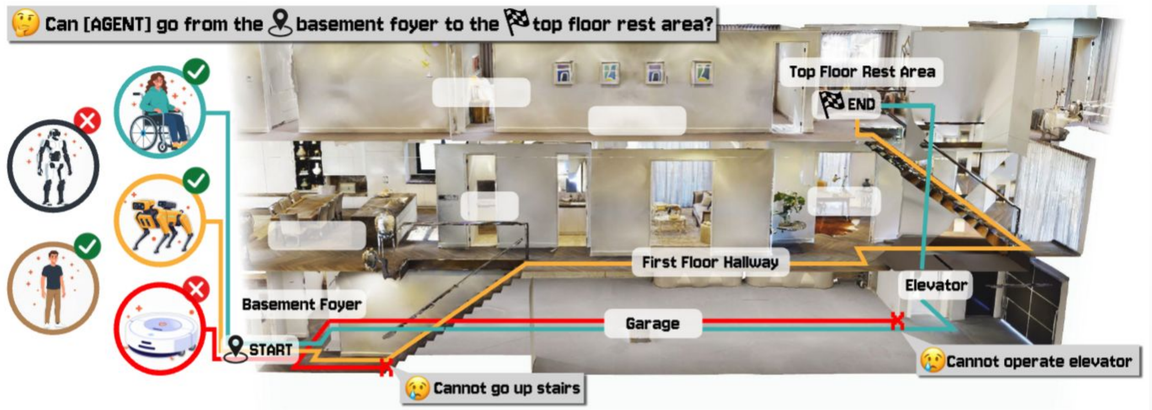
CapNav: Benchmarking Vision Language Models on Capability-conditioned Indoor Navigation

Xia Su[1]; Ruiqi Chen[1]; Benlin Liu[1]; Jingwei Ma[1]; Zonglin Di[2]; Ranjay Krishna[1]; Jon Froehlich[1]
[1]University of Washington, [2]University of California, Santa Cruz

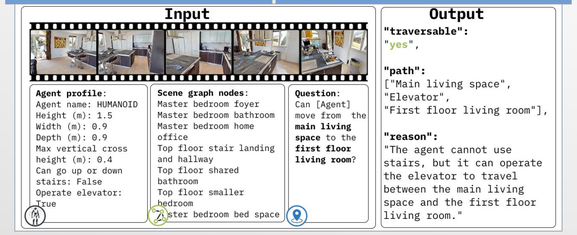
CVPR JUNE 3-7, 2026 DENVER COLORADO



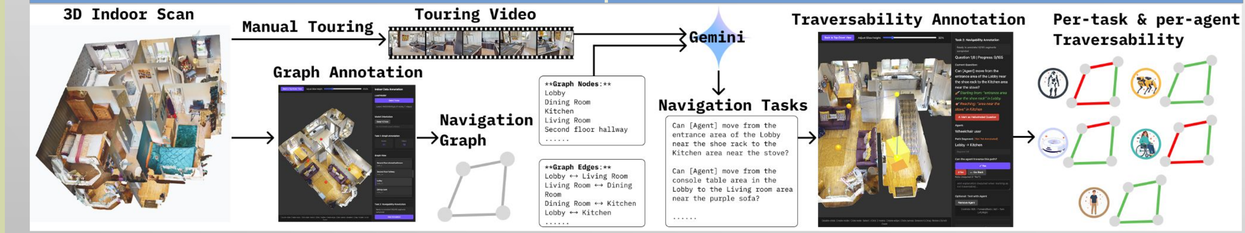
Abstract

Real-world navigation is conditioned by the agent's mobility constraints. For example, a sweeping robot cannot traverse stairs, while a quadruped can. We introduce **Capability-Conditioned Navigation** (***CapNav***), a benchmark designed to evaluate how well VLMs can navigate complex indoor spaces given an agent's specific physical and operational capabilities. CapNav defines five representative human and robot agents, and provides **45 real-world indoor scenes, 473 navigation tasks, and 2365 QA pairs**. We evaluate 13 SOTA VLMs and find that current VLM's navigation performance drops sharply as mobility constraints tighten, and all models struggle with obstacle types that require reasoning on spatial dimensions.

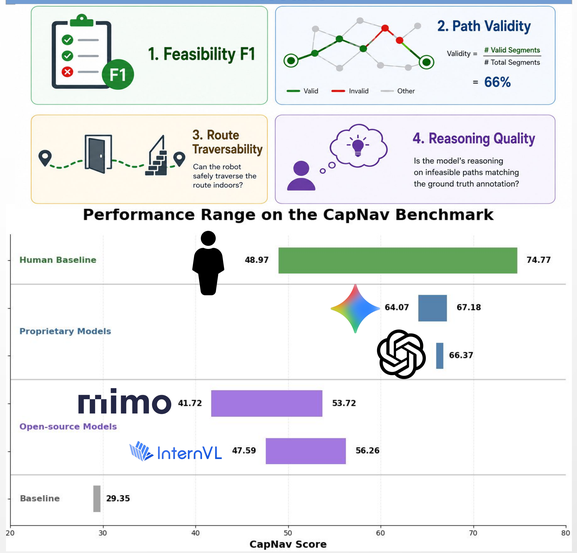
The CapNav Task



The CapNav Dataset

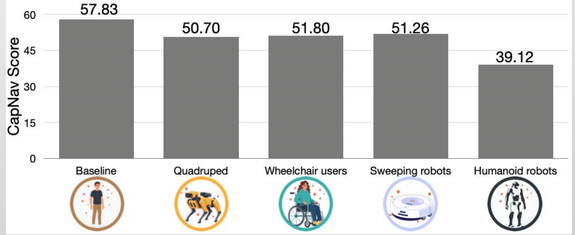


Results

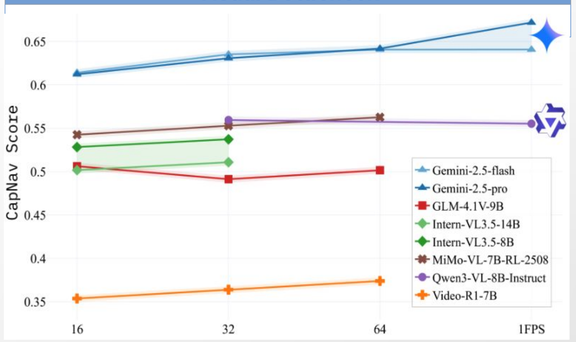


Capability Degradation

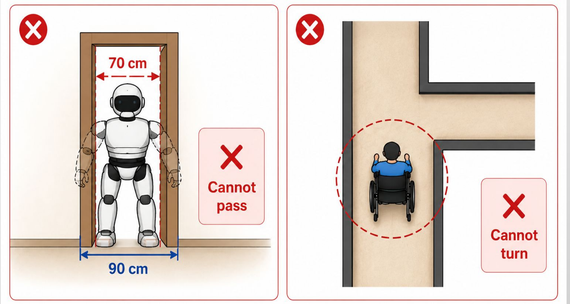
Across all tested models, performance is highest under the baseline embodiment without mobility limitations. As constraints tighten, performance degrades.



Visual Bottleneck



Dimension Neglect



Conclusions

Performance degrades systematically as mobility constraints tighten.
Increasing the inference budget yields inconsistent gains and expose a **visual integration bottleneck**.
Models exhibit **dimension neglect** and struggle with dimension reasoning (e.g., narrow clearances and turning radii).

Xia Su
University of Washington
Email: xiasu@cs.washington.edu
Website: https://xiasu.github.io/
Phone: +1 (206)-234-2908

Ruiqi Chen
Benlin Liu
Jingwei Ma
Zonglin Di
Ranjay Krishna
Jon E. Froehlich

Project Website & Dataset:

© POSTER TEMPLATE BY GENIGRAPH