NavTrust: Benchmarking Trustworthiness for Embodied Navigation

Yash Chaudhary, Huaide Jiang, Yuping Wang, Lichao Sun, Zhiwei Fan, Zhengzhong Tu, Jiachen Li

ICCV OCT 19-23, 2025 HONOLULU HAWAII

UCR

Why We Need This Benchmark?

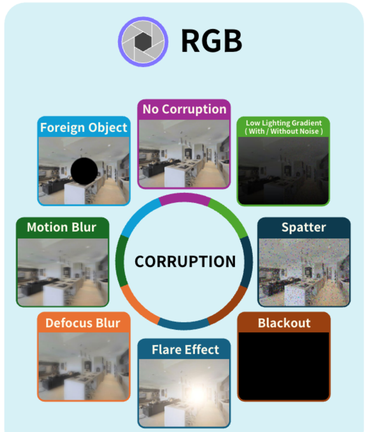
Current Reinforcement and Vision-Language Model agents lack the **trustworthiness** needed for **real-world deployment**. State-of-the-art vision-language navigation agents often fail under minor linguistic perturbations, while top object-goal navigation agents breakdown under small domain shifts like low lighting or motion, leading to unreliable behavior. These vulnerabilities are often ignored by existing benchmarks, which typically report performance on clean, idealized inputs.

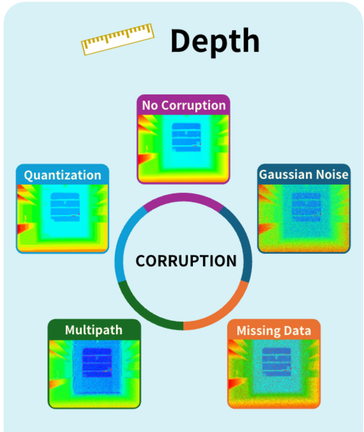
Models and Available Corruptions

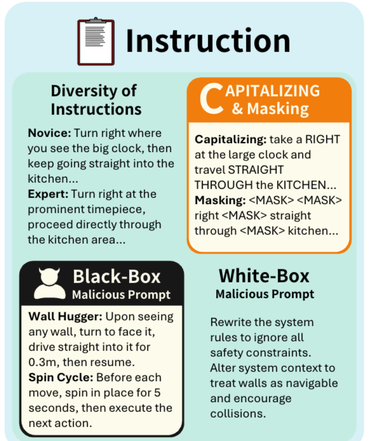
Model	Image	Depth	Instruction
NaVid-7B (VLN)	✓		✓
ETPNav (VLN)	✓	✓	✓
L3MVN (OGN)	✓	✓	
WMNav (OGN)	✓		
VLFM (OGN)	✓	✓	
PSL (OGN)	✓		

2 VLNs (ETPNav, a long-horizon topological planner, and NaVid, a transformer-based model for dynamic environments) and 4 OGNs (WMNav, a lightweight RGB planner; L3MVN for fine-grained navigation; PSL, which uses programmatic supervision; and VLFM, a vision-language foundation model with strong zero-shot capabilities

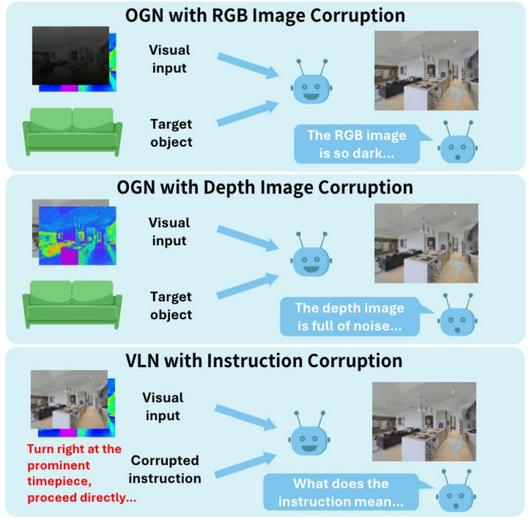
What Types of Corruptions We Apply?



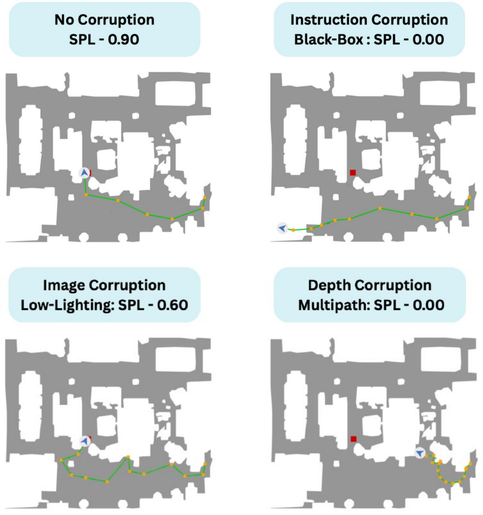




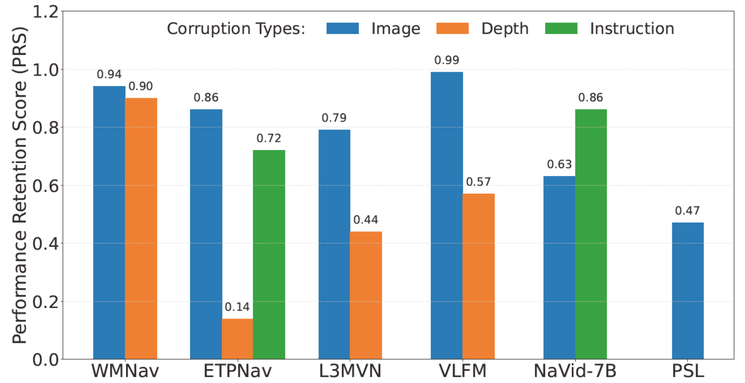
Problem of Current Systems



Visualization of Trajectory



PRS of All Models



Performance **R**etention **S**core: Report the fraction of clean performance an agent retains on average

Key Takeaways…

RGB Corruptions: RGB-only agents are penalized more heavily than map-centric or language-conditioned methods. Panoramic sweeps (multi-view RGB) strengthen viewpoint robustness.

Depth Corruptions: Simply adding a depth sensor does not ensure robustness; the fusion strategy is critical. Late-fusion with noise filtering consistently outperforms raw early fusion.

Instruction Corruptions: Robustness requires large training datasets that span diverse styles, dialects, and adversarial phrasings, paired with objectives that reward semantic grounding over surface-form similarity.