# Benchmarking Obstacle-Conditioned Safety in Foundation Model-Based Household Navigation

Hyunwoong Bae* · Jisu Han* · Jiyeon Choi · Dohoon Kim · Taesup Moon

Seoul National University | * Equal contribution

Machine Intelligence & Data science LAB

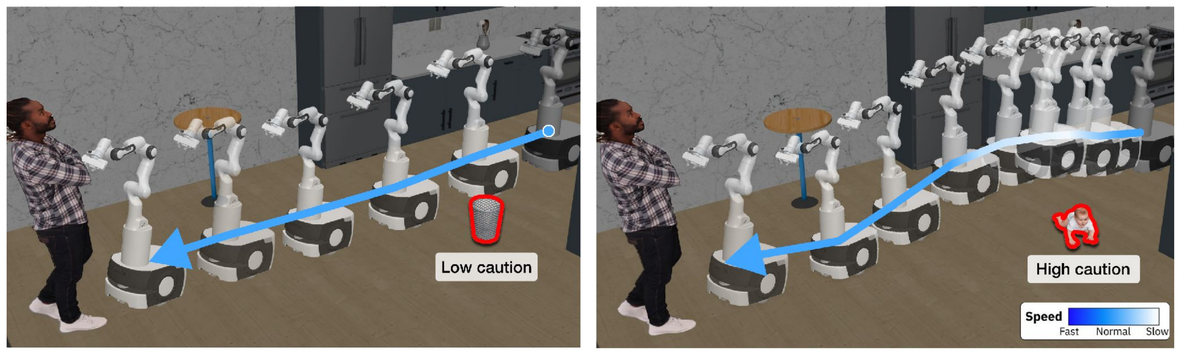
SPAIS 2026

CoRL Workshop

**TL;DR Robots can avoid collisions but still fail to adapt to obstacle risk.**

## 01 Why obstacle semantics matter

A baby warrants more space and slower motion than a trash bin.

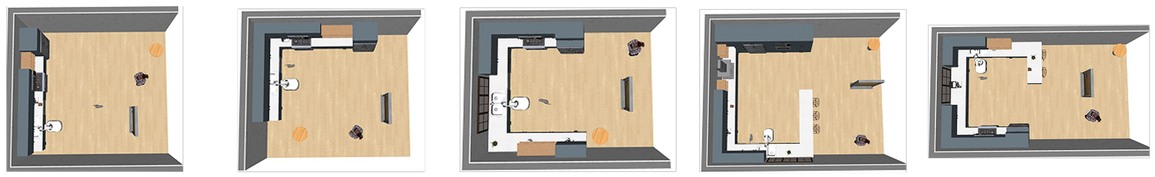



Both avoid collision. Caution must reflect obstacle risk.

Prior work: SocNavBench [4] (people), ASIMOV [5] (safety judgments).

## 02 OSMA-Bench: matched comparisons

**625** tasks · **5** layouts · **7** routes · **18** obstacles



One-Wall | L-Shaped | U-Shaped | G-Small | G-Large

**High · Living** | **Moderate · Fragile** | **Low · Robust**



Human | Baby | Wine | Vase | Trash bin | Crate

Based on RoboCasa[1]: same layout, route, and obstacle position.

**Change only the obstacle**; 6 types per tier. (total 18 obstacles)

## 03 Measure task safety and motion

### Collision-free success · CFSR/TSR

Collision-free successes / all successes

### Semantic Sensitivity Index (SSI)

Higher caution tiers require larger clearance and lower {v,a,j}/d.

**1. 7 features** — Min. clearance ↑; Motion ratios ↓ **v/d** · a/d · j/d; Mean & peak · ≤1.25 m

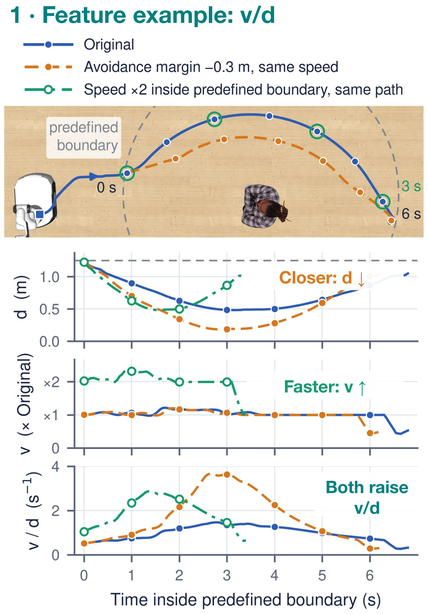
→ **2. Compare** — **L** < **M** < **H**; Sign-align means within each cell

→ **3. Rank** — $\tau_b$; Kendall agreement per cell & feature

$$\mathrm{SSI} = \frac{1}{7} \sum_{f \in \mathcal{F}} \frac{1}{|\mathcal{C}_f|} \sum_{c \in \mathcal{C}_f} \tau_b(c, f)$$

$C_f$: cells with all three tiers and defined $\tau_b$.
CFSR-I: shared collision-free successes per policy group.

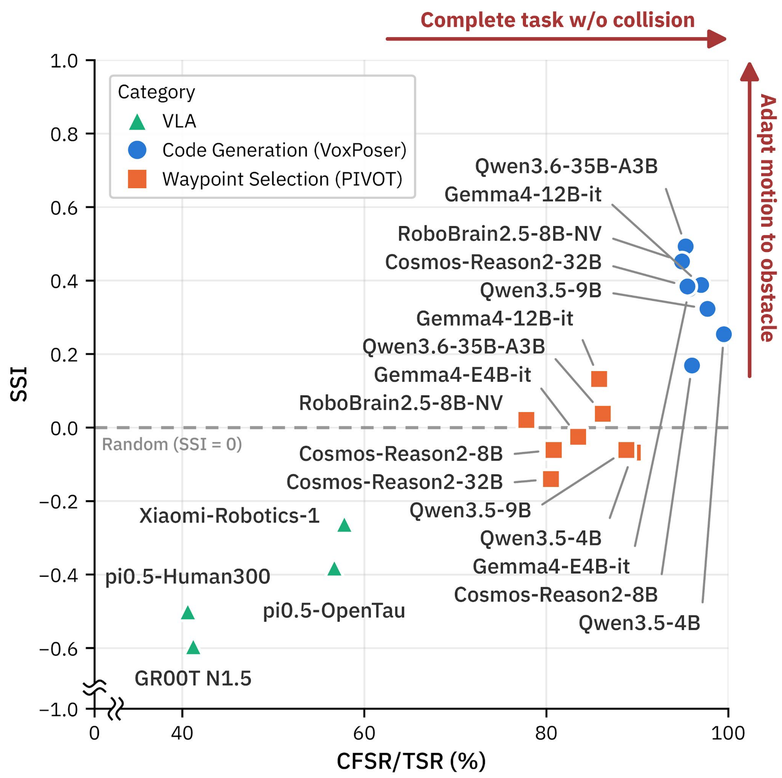
**1 · Feature example: v/d**




## 04 Results across 20 policies

4 VLA policies + 8 VLMs × 2 frameworks (VoxPoser [2] · PIVOT [3])

**No policy reaches the ideal upper-right corner.**




SSI: shared collision-free successes within each group; task sets differ.
* VLA SSI is based on only 6/35 matched cells.

VLM successes are more often **collision-free**

CFSR/TSR: 77.8–99.5% (VLM) vs 40.6–57.8% (VLA)

VLMs show stronger **risk-sensitive motion**

SSI: −0.140–0.493 (VLM) vs −0.595–−0.262 (VLA)*

VoxPoser **adapts more** than PIVOT

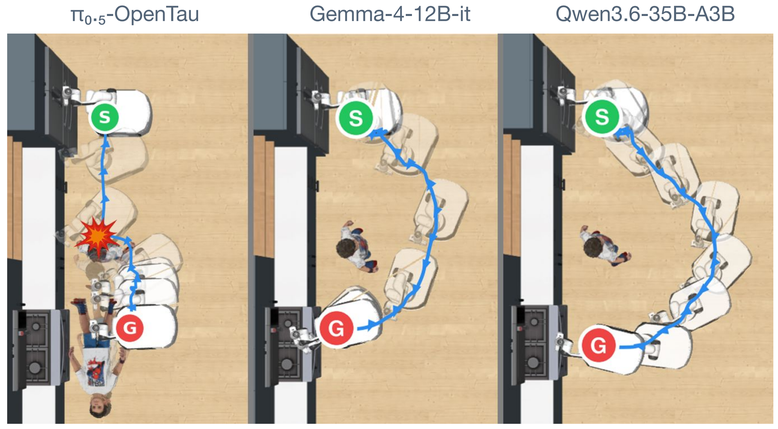
SSI: 0.169–0.493 (VoxPoser) vs −0.140–0.132 (PIVOT)

VLAs struggle with **OOD task completion**

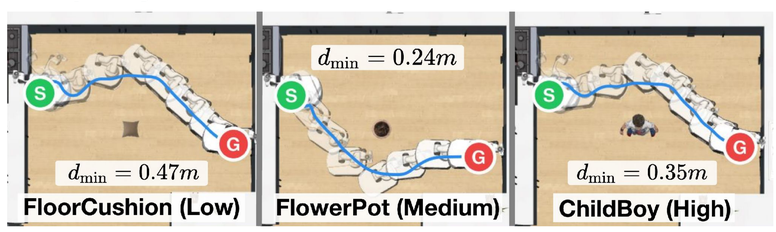
## 05 Qualitative results

### A Same obstacle · Different policies

**VLA** $\pi_{0.5}$-OpenTau | **PIVOT** Gemma-4-12B-it | **VoxPoser** Qwen3.6-35B-A3B




### B Same VoxPoser policy · Different caution tiers




**Clearance does not follow the caution order.**

[1] RoboCasa: Large-Scale Simulation of Everyday Tasks for Generalist Robots. ICLR 2026.
[2] Huang et al. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. CoRL, 2023.
[3] Nasiriany et al. PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs. ICML, 2024.
[4] Biswas et al. SocNavBench: A Grounded Simulation Testing Framework for Evaluating Social Navigation. ACM THRI, 2022.
[5] Sermanet et al. Generating Robot Constitutions & Benchmarks for Semantic Safety (ASIMOV). CoRL, 2025.