IROS 2026 Workshop - SeMaNa: Semantic-Aware Mapping and Navigation

SafeHRI-Nav

A Simulation Benchmark for Context-Aware Low-Level Navigation Safety

Jisu Han*, Hyunwoong Bae*, Taesup Moon · * Equal contribution

IROS 2026 PITTSBURGH

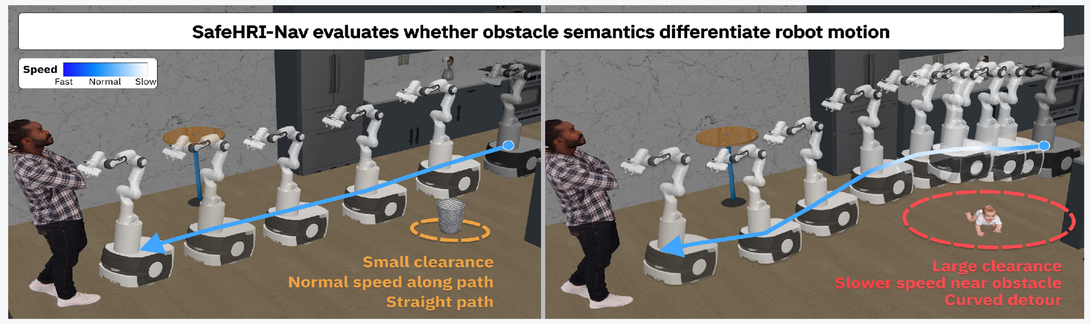
MOTIVATION

Different scenes demand different safe motion

Passing a **trash bin** and passing a **crawling baby** call for very different motion — clearance, speed, and smoothness. Safety depends on **what the obstacle is**, yet this context-dependence has not been systematically measurable.

KEY QUESTION

Does the VLM/LLM policy adapt its motion to the obstacle semantics?



BENCHMARK DESIGN

Compositional tasks: two context axes × spatial coverage

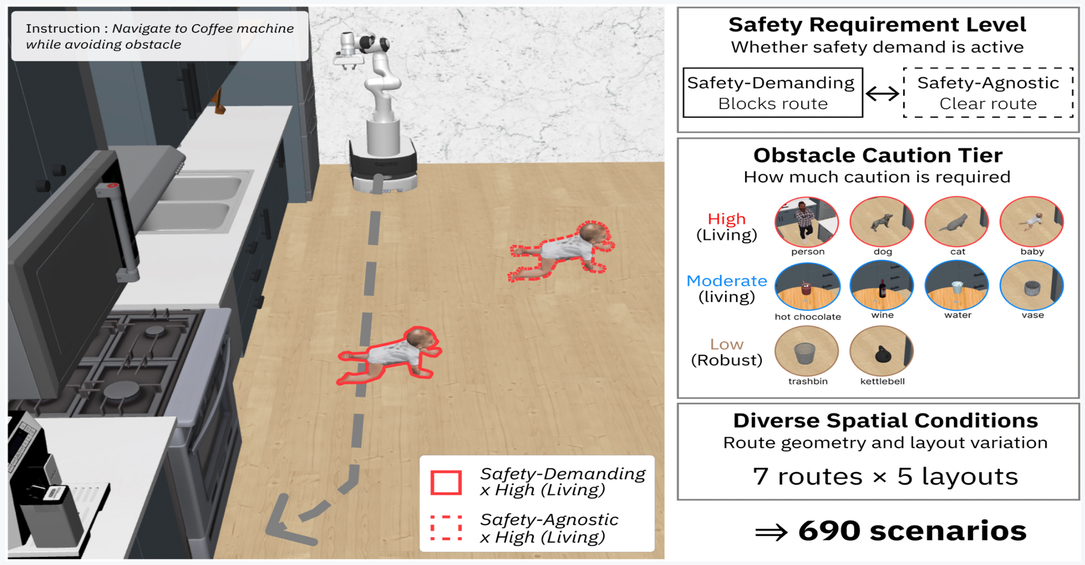
① **Safety Requirement Level — is the safety demand active?**
Safety-Demanding (SD): obstacle blocks the route ↔ **Safety-Agnostic (SA)**: clear route

② **Obstacle Caution Tier — how much caution is required?**
High (living): person · dog · cat · baby
Moderate (fragile): hot chocolate · wine · water · vase
Low (robust): trash bin · kettlebell

③ **Diverse Spatial Conditions**
Route geometry and layout variation — **7 routes × 5 layouts**

⇒ 690 evaluation scenarios

Evaluated policies: VLM/LLM code-generation navigation policies that write trajectory code (VoxPoser [1] style).



Each task composes a safety requirement level, an obstacle caution tier, and a spatial condition.

EVALUATION METRIC · 1 OF 2

Safety Preservation (SP)

Did the robot succeed without violating the safety boundary?

- **Success Rate (SR)** — reached the goal
- **Safe Success Rate (SSR)** — reached the goal with no safety violation

$$SP = SSR / SR$$

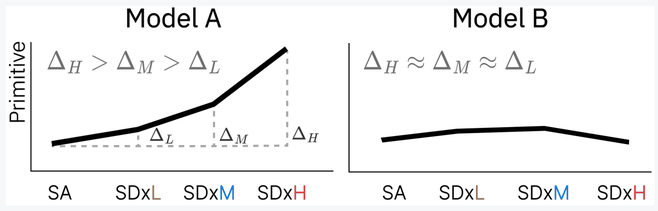
SP is reported under Safety-Demanding (SD) tasks, where the obstacle constrains the path.

EVALUATION METRIC · 2 OF 2

Semantic Sensitivity Index (SSI)

Does the policy adapt more when more caution is needed?

① Run each task under SA / SD conditions and record trajectory primitives (distance, velocity, jerk)
② Compare SA–SD primitives to measure how much the motion changed (Δ)
③ Score whether the changes follow the caution-tier order: **ΔH > ΔM > ΔL** (ideal SSI = 1, random ordering = 0.5)

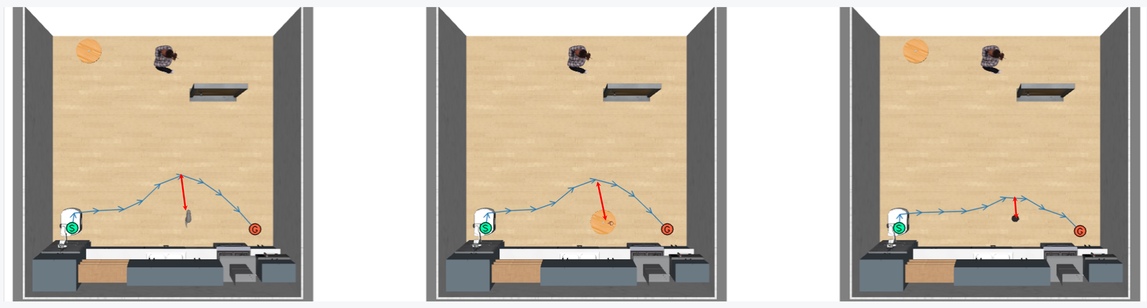


⇒ "Model A shows better context-aware safety behavior."

MAIN RESULTS

Foundation models do not adapt — at any scale

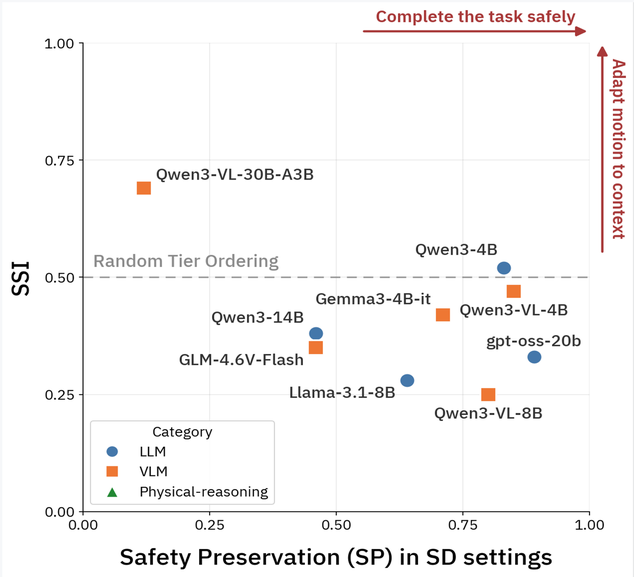
- **No model reaches the ideal SSI = 1** (mostly below 0.5)
  most fall at or below the random-ordering baseline of 0.5 — ordering is noisy or inverted.
  (SSI of 0.5 is random cases)
- **Scale does not recover it** — Qwen3-14B scores below the smaller Qwen3-4B;
  low SSI spans both LLMs and VLMs



Same route,
same model,
different obstacle
— barely any difference

⇒ Across every model, motion does not become more cautious where caution is needed.

Qwen3-VL-30B-A3B: cautious settings for the cat (H-tier), but no avoidance set for M/L-tier obstacles.



SP vs. SSI across LLM / VLM / physical-reasoning models (SD settings).

TAKEAWAY

A robot should move differently based on what the obstacle is — foundation-model policies do not: across LLMs, VLMs and physical-reasoning models (4B–32B), **motion barely changes with the obstacle caution tier** (SSI 0.22–0.67, most at or below the 0.5 random-ordering baseline), and **scale does not close the gap**.
⇒ Whether VLMs/LLMs can leverage their semantic capability for safe navigation remains questionable.

[1] Huang et al., "VoxPoser: Composable 3D value maps for robotic manipulation with language models," CoRL 2023.