Benchmarking Large Language Model Reasoning in Indoor Robot Navigation

Emirhan Balcı[1], advised by Mehmet Sarıgül[2] and Barış Ata[2]

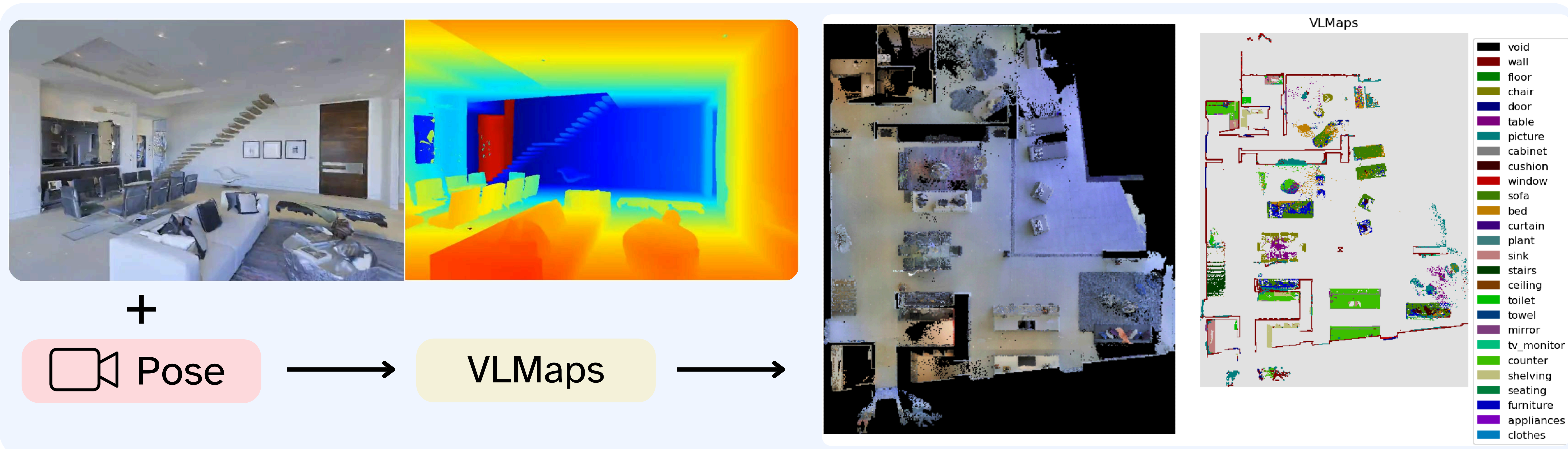
Adana Science and Technology University[1], Çukurova University[2]

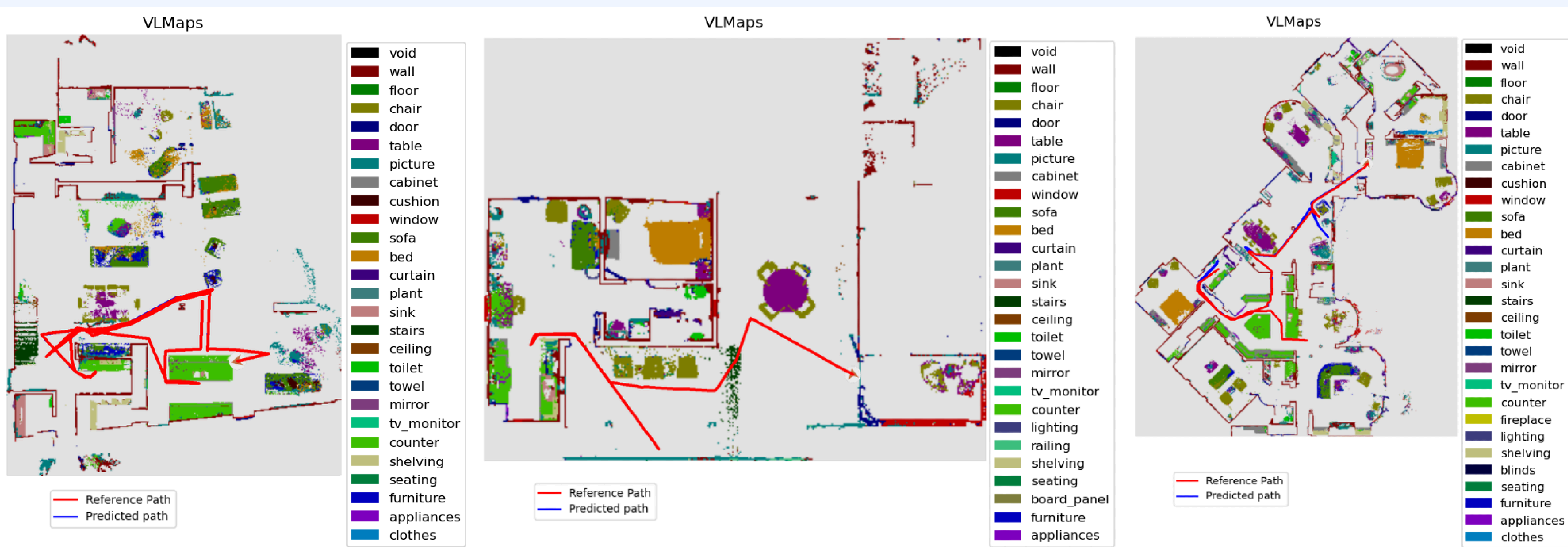
Introduction

Text-based generative large language models (LLMs) are capable of handling complex navigation tasks involving spatial understanding and sequential decision-making. Based on this context, we aimed to analyze the indoor navigation planning performances of recent LLMs and provide a realistic experiment baseline.

Methodology

We built a benchmark pipeline that uses VLMaps with Habitat-Sim to model realistic object- and spatial-based navigation scenarios. Habitat-Sim provides RGB-D images and camera poses for three Matterport3D scenes, while VLMaps generates their semantic maps. LLMs are prompted to generate robot code for object-goal, spatial-goal, and common-sense reasoning tasks within the mapped scenes. The resulting trajectories are extracted from Habitat-Sim. Performance of ten LLMs—from the ChatGPT, DeepSeek, Claude, and Gemini families—is measured using the trajectory evaluation metrics Success Rate, SDTW, CLS, and Navigation Error.





Sample Prompt: Spatial Goal Navigation

```
> I want you to perform multiple tasks in order. Move first to the left side of the chair in front of you, face the sofa, and then move to the west of the counter, later, with the counter on your right, go to the east of the window, face the chair in front of you and move to the south of the door. Finally, turn absolute 180 degrees.
```

Results

The findings indicate that while the models successfully executed object and spatial-based instructions, some models struggled with those requiring common-sense reasoning. Additionally, GPT-4o, DeepSeek-R1, and Claude 3.5 Sonnet models demonstrated superior navigation-planning performances relative to the other models, due to their commonsense reasoning capabilities.

Metrics / Models	Object Goal Navigation Test				Spatial Goal Navigation Test				Common-Sense Reasoning Navigation Test			
	SR ↑	NE ↓	SDTW ↑	CLS ↑	SR ↑	NE ↓	SDTW ↑	CLS ↑	SR ↑	NE ↓	SDTW ↑	CLS ↑
gpt-4-turbo	1.00	0.00	1.00	0.99	1.00	0.00	1.00	1.00	0.00	18.48	0.00	0.71
gpt-3.5-turbo	1.00	0.00	1.00	0.99	0.00	8.31	0.00	0.73	1.00	0.22	0.96	0.86
gpt-4o	1.00	0.00	1.00	0.99	1.00	0.00	1.00	1.00	1.00	0.05	0.99	0.88
o1-mini	1.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00	12.24	0.00	0.72
deepseek-r1	1.00	0.00	1.00	0.99	1.00	0.00	1.00	1.00	1.00	0.05	0.99	0.88
deepseek-v3	1.00	0.00	1.00	0.99	1.00	0.00	1.00	1.00	0.00	18.48	0.00	0.67
claude-3.5-haiku	1.00	0.00	1.00	0.99	1.00	0.00	1.00	1.00	1.00	0.08	0.97	0.87
claude-3.5-sonnet	1.00	0.00	1.00	0.99	1.00	0.00	1.00	1.00	1.00	0.05	0.99	0.88
gemini-flash-1.5	1.00	0.09	1.00	0.99	1.00	0.00	1.00	1.00	-	-	-	-
gemini-2.0-flash	1.00	0.00	1.00	0.99	1.00	0.00	1.00	1.00	0.00	14.88	0.00	0.40