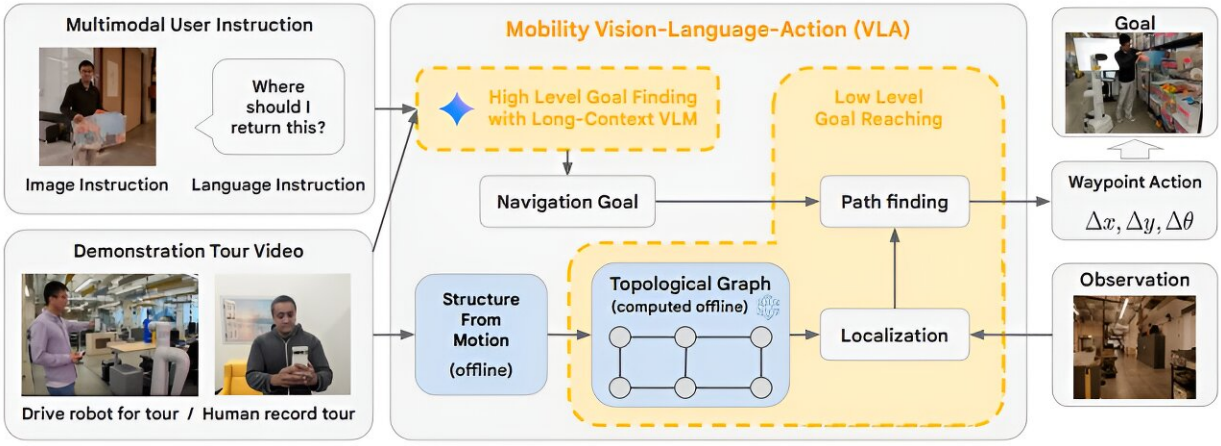
DeepMind demonstrates a robot capable of giving context-based guided tours of an office building

July 12 2024, by Bob Yirka



Mobility VLA architecture. The multimodal user instruction and a demonstration tour video of the environment are used by a long-context VLM (high-level policy) to identify the goal frame in the video. The low-level policy then uses the goal frame and an offline generated topological map (from the tour video using structure-from-motion) to compute a robot action at every timestep. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2407.07775

A team of roboticists and AI specialists at Google's DeepMind have demonstrated a robot capable of giving context-based guided tours around its offices. They have posted a paper describing their work, along with demonstration videos, on the *arXiv* preprint server.

AI applications have come a long way over just the past decade, and LLMs such as ChatGPT are now familiar to users around the globe. In this new effort, the research team gave RT-2 robots AI capabilities via Gemini 1.5 Pro and used it to allow the robot to perform sophisticated activities.

The robot can listen to a person it is guiding, parse a request and translate it into behavior. As an example, one researcher asked the robot to take it to a place in the office where writing or drawing could be done. The robot thought about the request for approximately 30 seconds and then guided the person to a place where a whiteboard had been attached to the wall in one of the offices.

The robot is able to carry out such tasks, the researchers explain, because its Gemini 1.5 Pro application was trained to understand the layout of the 850-square-meter office workspace using its long context window as it gathered data while watching videos of locations in the office.

The researchers describe such learning experiences as multimodal instruction navigation with demonstration tours—as the robot watched the videos, it was able to process different parts of the office scenery simultaneously, allowing it to generate associations.

By adding voice and text processing along with other AI features, the team at DeepMind was also able to give the robot the ability to perform inferential processing. As an example, a researcher asked the robot if there was any more of his favorite beverage in the refrigerator. The robot noted that there were several empty Coke cans near where the researcher was sitting, and used that information to guess that Coke was his favorite beverage. It then rolled itself to the refrigerator and looked inside of it to see if there were any cans of Coke. It then rolled itself back and reported what it had found.

More information: Hao-Tien Lewis Chiang et al, Mobility VLA: Multimodal Instruction Navigation with Long-Context VLMs and Topological Graphs, *arXiv* (2024). DOI: 10.48550/arxiv.2407.07775