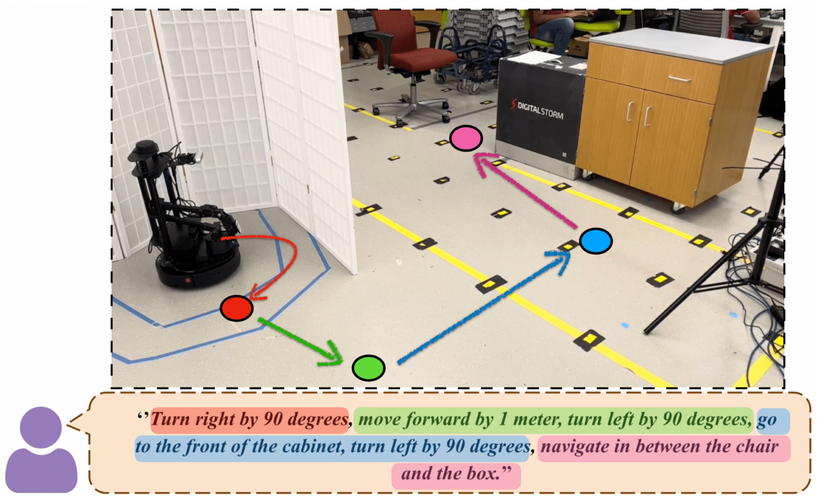
Vision and Language Navigation in the Real World via Online Visual Language Mapping

Chengguang Xu Hieu T. Nguyen Christopher Amato Lawson L.S. Wong

Khoury College of Computer Sciences
Northeastern University

Vision and Language Navigation in Continuous Environments

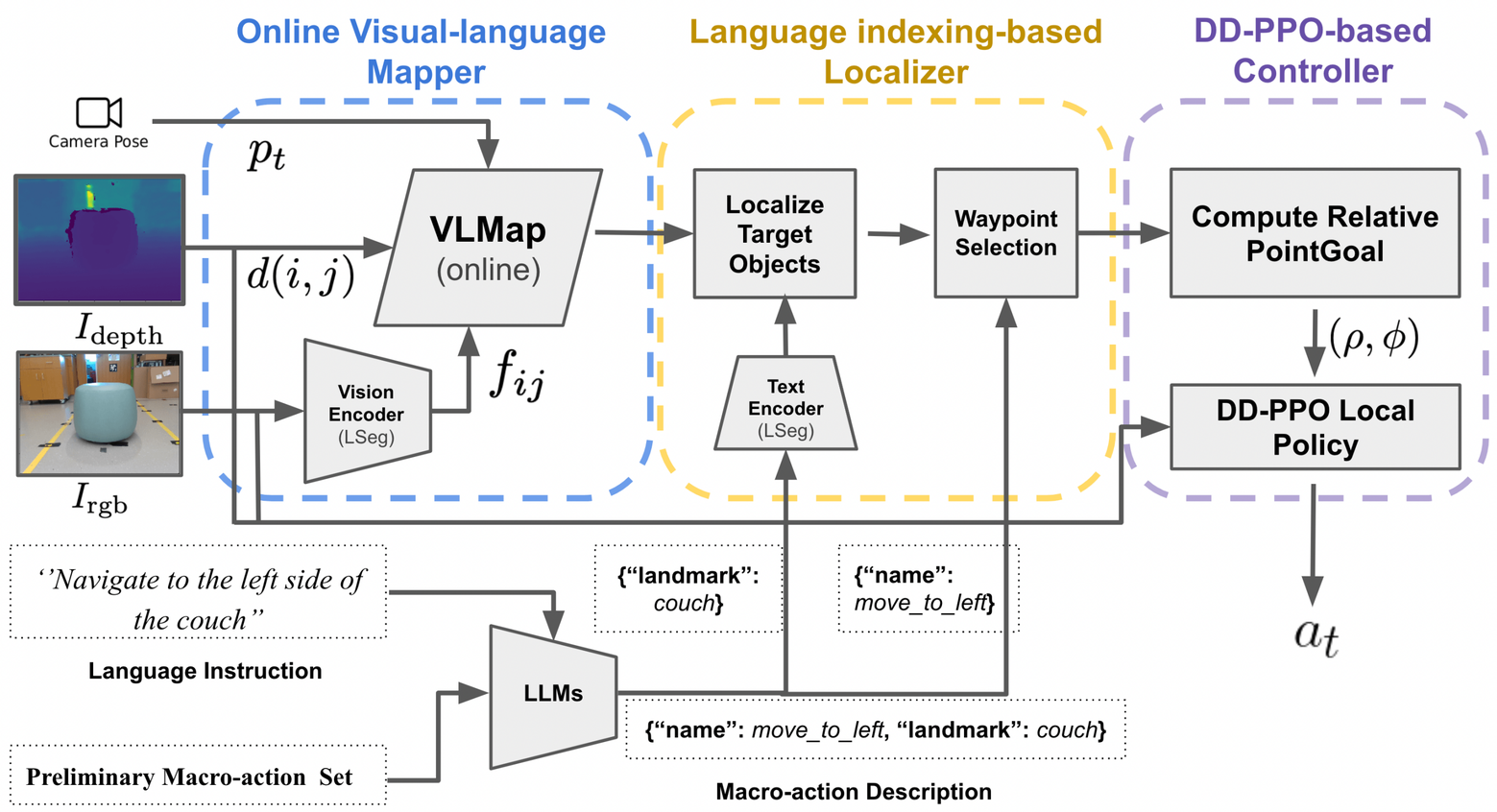


The robot is asked to navigate in unseen environments by following instructions in natural languages. The robot takes *RGB-D* images and *camera poses* as observations and outputs one of four discrete actions (i.e. *move_forward*, *turn_right*, *turn_right*, and *stop*).

Single Instruction Following

Complex Instruction Following

Navigation Framework Overview



We first use large language models (**LLMs**) to parse the instruction into a sequence of preliminary macro-action descriptions containing both the macro-action name and the related landmark. Then, the online visual-language mapper maintains a visual-language map from the front-view RGB-D observations using visual-language models (**VLMs**). Based on the latest map and the macro-action description, the language-indexing-based localizer outputs the waypoint location, represented as a point goal, on the map. The controller takes in both the RGB-D observation and a relative point goal, computed from the waypoint location and the agent location on the map, and predicts the next action.

Quantitative Results

Table 1: Results of Pure Motion Task

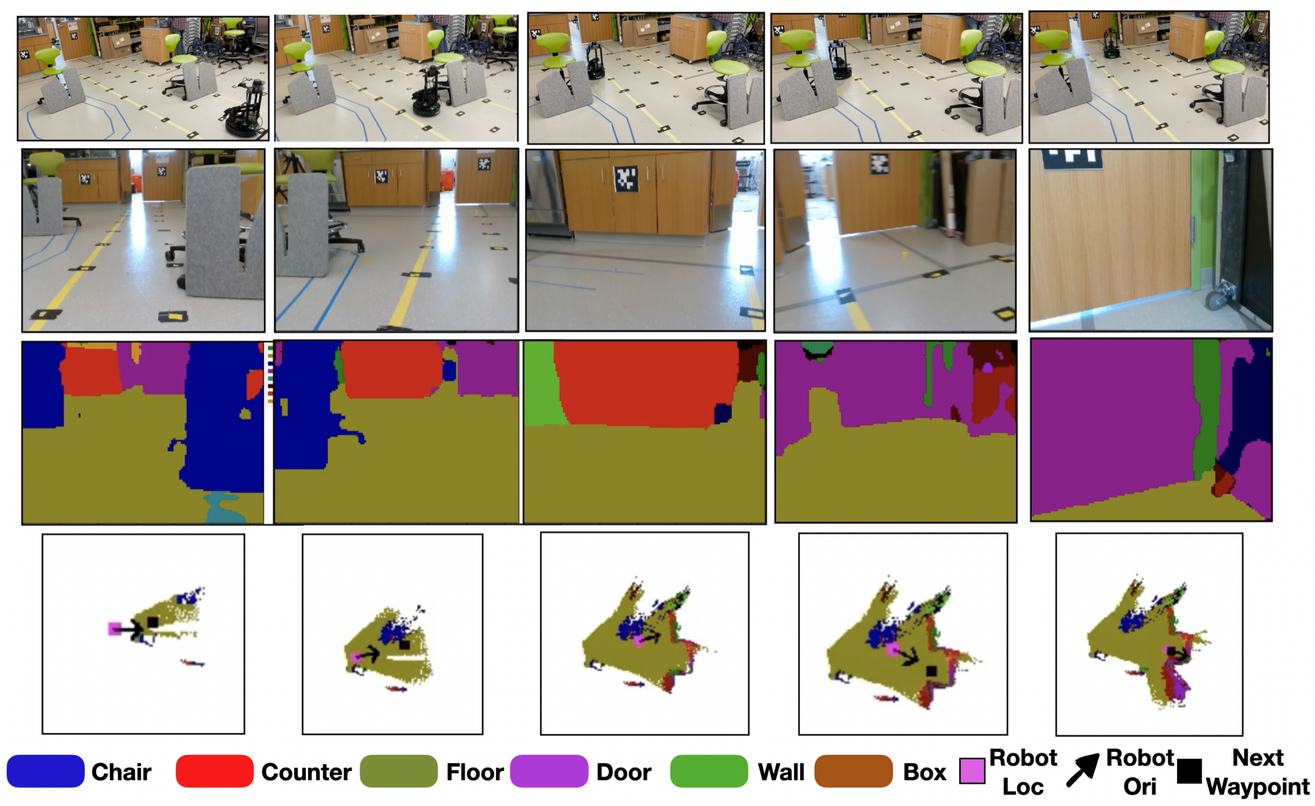
Target Dist (m)	Actual Dist (m)	Est Dist (m)	Err Dist (m)
0.5	0.426	-	-
1.0	0.748	0.238	0.014
2.0	1.678	0.308	0.014

Table 2: Results of Landmark-associated Motion Task

Method	CM2 [7]	
Instruction	SR (%)	Dist to Goal (m)
"Navigate to the *left* side of the chair"	60	0.88
"Navigate to the *right* side of the chair"	40	0.97
"Navigate to the *front* of the chair"	20	1.37
"Move *in between* the box and the chair"	0	2.06
Average	30	1.32
Method	Ours	
"Navigate to the *left* side of the chair"	100	0.79
"Navigate to the *right* side of the chair"	100	0.83
"Navigate to the *front* of the chair"	100	0.81
"Move *in between* the box and the chair"	80	0.20
Average	**95**	**0.66**

Method	SR (%)	Dist to Goal (m)	Time steps
CM2 [7]	0	4.9	203.4
Ours	**100**	**0.256**	**88.6**

Qualitative Results



With no *fine-tuning*, the proposed framework generalizes well to the unseen lab environment.

xu.cheng@northeastern.edu