Institute for Interdisciplinary Information Sciences, Tsinghua University
清華大學 交叉信息研究院

ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION
ATLANTA 2025

SARO: Space-Aware Robot System for Terrain Crossing via Vision-Language Model

Shaoting Zhu* Derun Li* Linzhan Mou Yong Liu Ningyi Xu Hang Zhao†

清華大學 交叉信息研究院 Institute for Interdisciplinary Information Sciences, Tsinghua University
上海交通大學 SHANGHAI JIAO TONG UNIVERSITY
浙江大学 ZHEJIANG UNIVERSITY
Penn Engineering | GRASP Laboratory General Robotics, Automation, Sensing & Perception Lab
上海期智研究院 SHANGHAI QI ZHI INSTITUTE

Motivation and Introduction

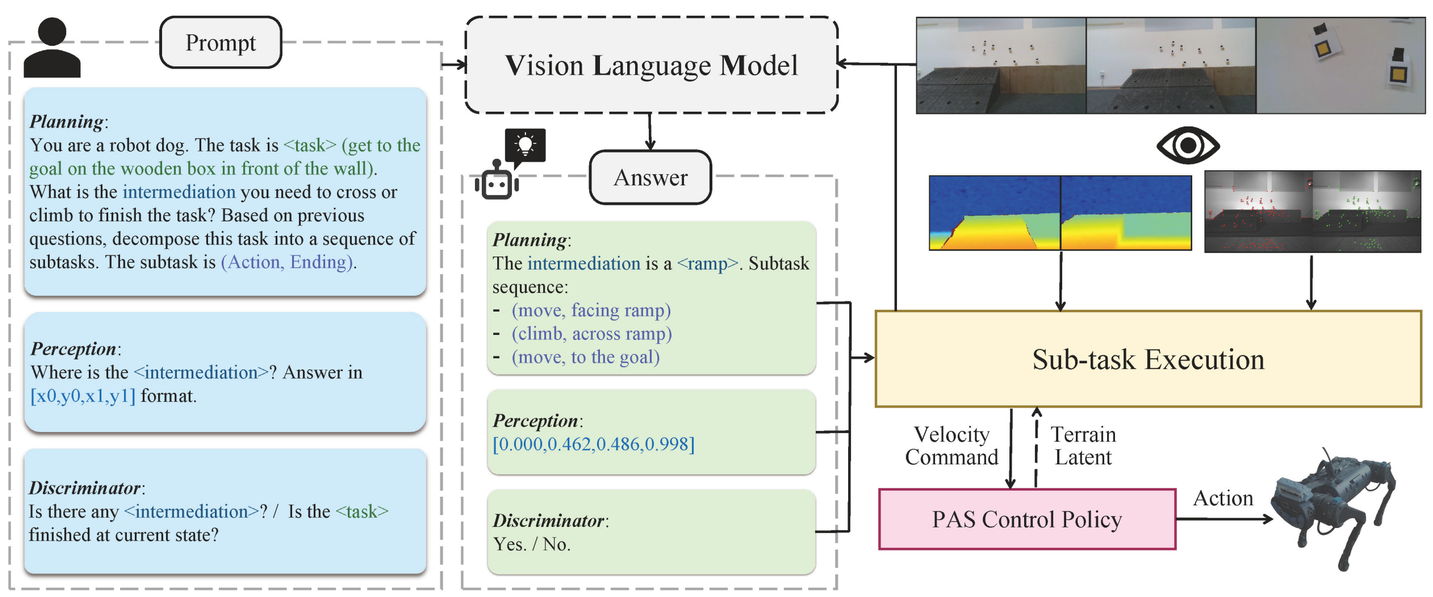
Vision-language models (VLMs) have shown advancements in common sense reasoning and remarkable generalization in vision tasks, significantly boosting the progress of robotic learning. However, VLMs suffer from the limitations of training data perspectives and the lack of a memory information bank, which is believed to curtail their usage in robot navigation tasks. Our motivation originates from this important question: ***How can we design a system to fully activate the potential of VLMs' visual common sense on robots to enable them to observe and understand and travel in the 3D world?***

In this work, we design a system called **SARO**, composed of a ***high-level reasoning module***, a ***close loop sub-task execute module***, and ***a low-level control policy***. The system enhances the 3D reasoning, motion planning, and locomotion ability of the robots.

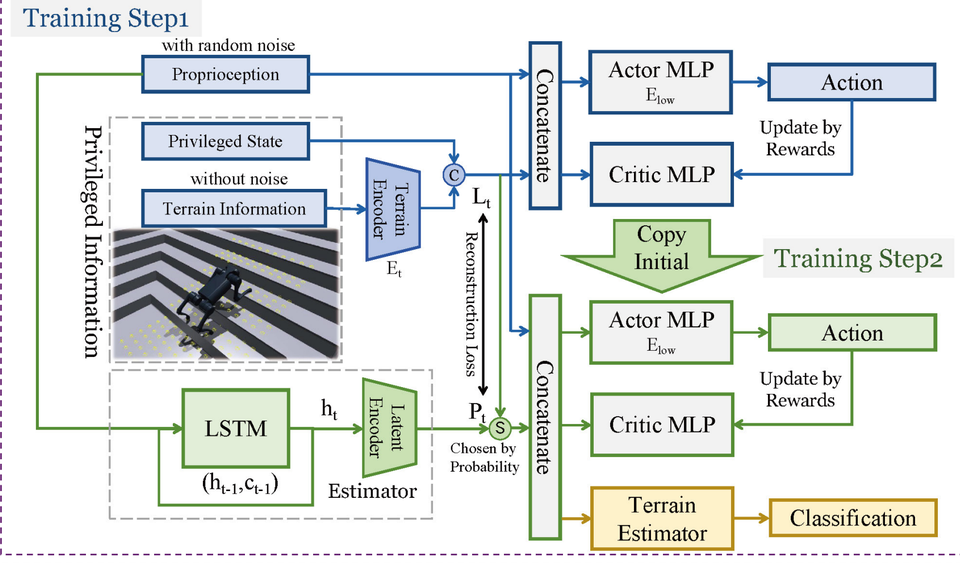
Task Definition:

Find the way and navigate through: $\{\mathcal{P}_1 \rightarrow \mathcal{I} \rightarrow \mathcal{P}_2\}$ *under the condition of* $\mathcal{G}$ *and* $\mathcal{L}$

High-level Reasoning and Task Execution



Low-level Locomotion Control Policy



We conduct two-stage training paradigm to obtain robust low-level locomotion policy. In the first training step, we train an oracle policy using proprioception, privileged state, and terrain information. In the second training step, we use the probability annealing selection (PAS) method to train the final actor network, which only uses proprioception as input. After the policy training process is finished, we exclusively train a terrain estimator to classify whether the robot is on the plane or is climbing the intermediation.

Probability Annealing Selection (PAS):

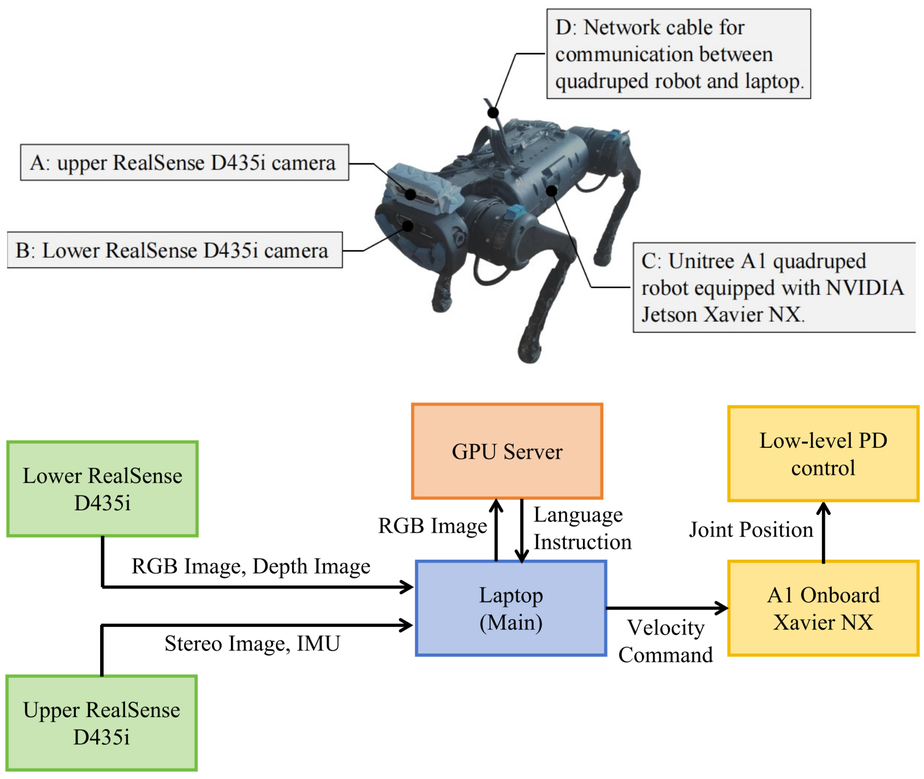
$$\boldsymbol{p}_t = E_e(\boldsymbol{o}_t^p), \tag{1}$$

$$\boldsymbol{i}_t = \text{Probability Selection}\ (\boldsymbol{P}_t, \boldsymbol{p}_t, \boldsymbol{l}_t), \tag{2}$$

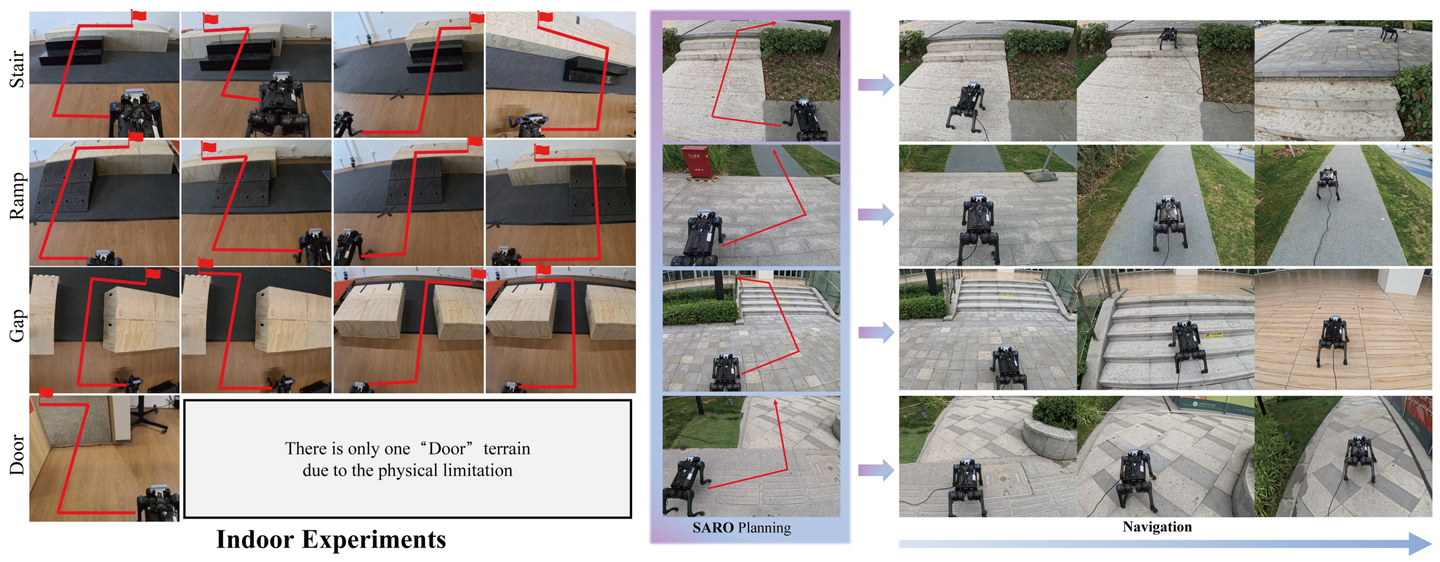
$$\boldsymbol{a}_t = E_{low}(\boldsymbol{i}_t, \boldsymbol{o}), \tag{3}$$

$$\text{Probability}\ \boldsymbol{P}_t = \boldsymbol{\alpha}^{iteration}. \tag{4}$$

Experiments



Intermediation	Overall	Stable Loc	w/o Closed-Loop	NoMaD	LSTM	Across Terrains	ViNT
Stair	60%	**88%**	10%	0%	0%	70%	0%
Ramp	25%	**67%**	10%	0%	0%	50%	0%
Gap	45%	**94%**	20%	20%	0%	80%	0%
Door	30%	**63%**	15%	70%	0%	50%	0%



More Demos on Websete Page: *https://saro-vlm.github.io/*