# Final Year Project
# Embodied AI for Navigation with Quadruped Robots

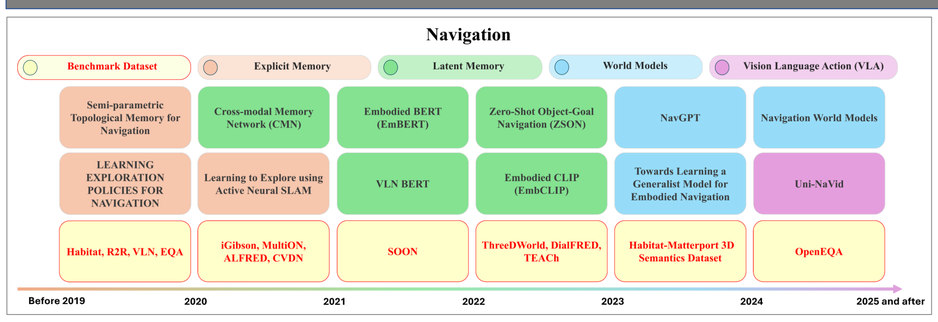
Kenny Lik Hang Wong, JunHui Hou
Department of Computer Science, City University of Hong Kong

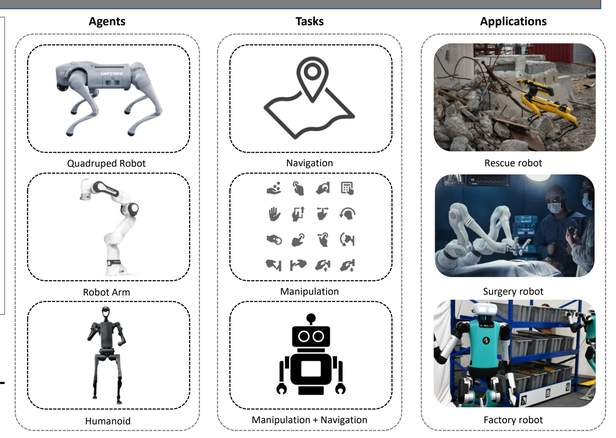
Department of Computer Science
香港城市大學
City University of Hong Kong

## ABSTRACT

This project advances Vision-Language Navigation (VLN) in Embodied AI (EAI) by introducing **Recurrent-VLN-Bert-Isaac**, a novel imitation learning-based model designed for quadruped robots, and the **VLN-Go2-Matterport dataset**, a new resource for high-fidelity indoor navigation. These contributions enhance accessibility for researchers in Computer Vision and Natural Language Processing while bridging the gap between research and real-world deployment. Promising experiments show the model effectively learns transferable navigation policies, with additional **LLM-based approaches** explored in the appendix.

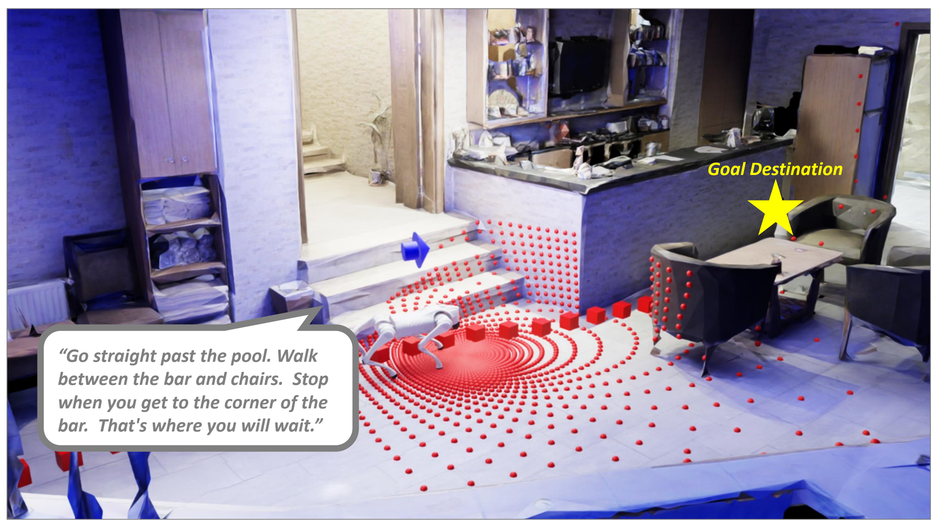
## INTRODUCTION



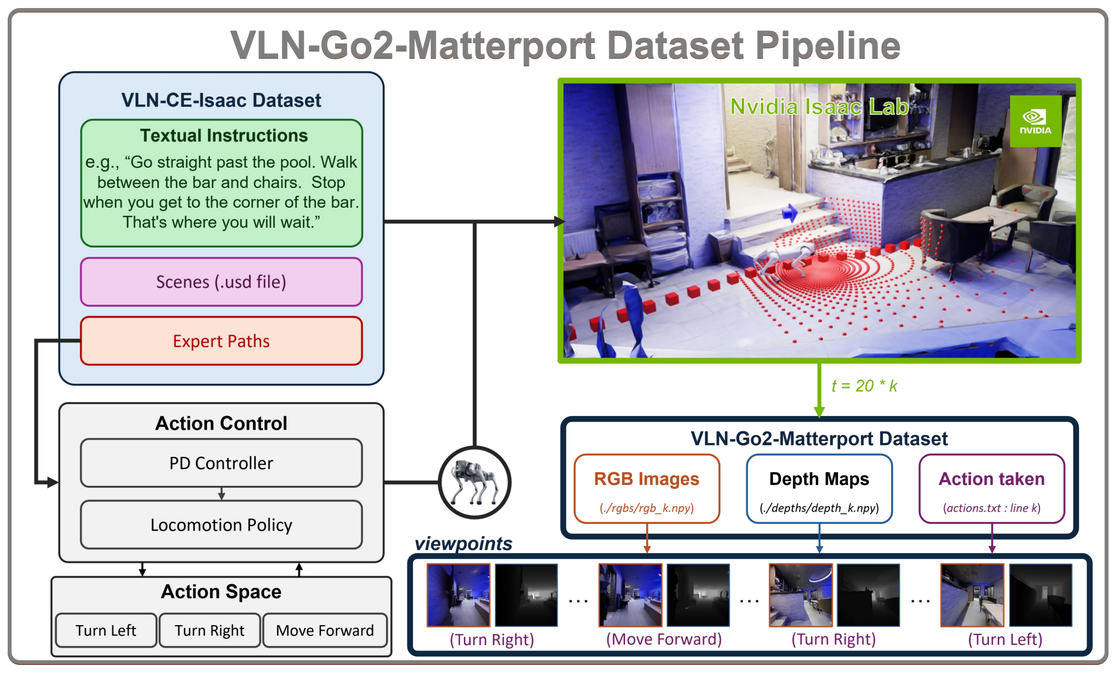

In **EAI**, agents with different embodiment can be used in various tasks. This project advances **VLN task** by targeting agile robots in high-fidelity physics simulators and providing resources for EAI research.

## Problem Definition



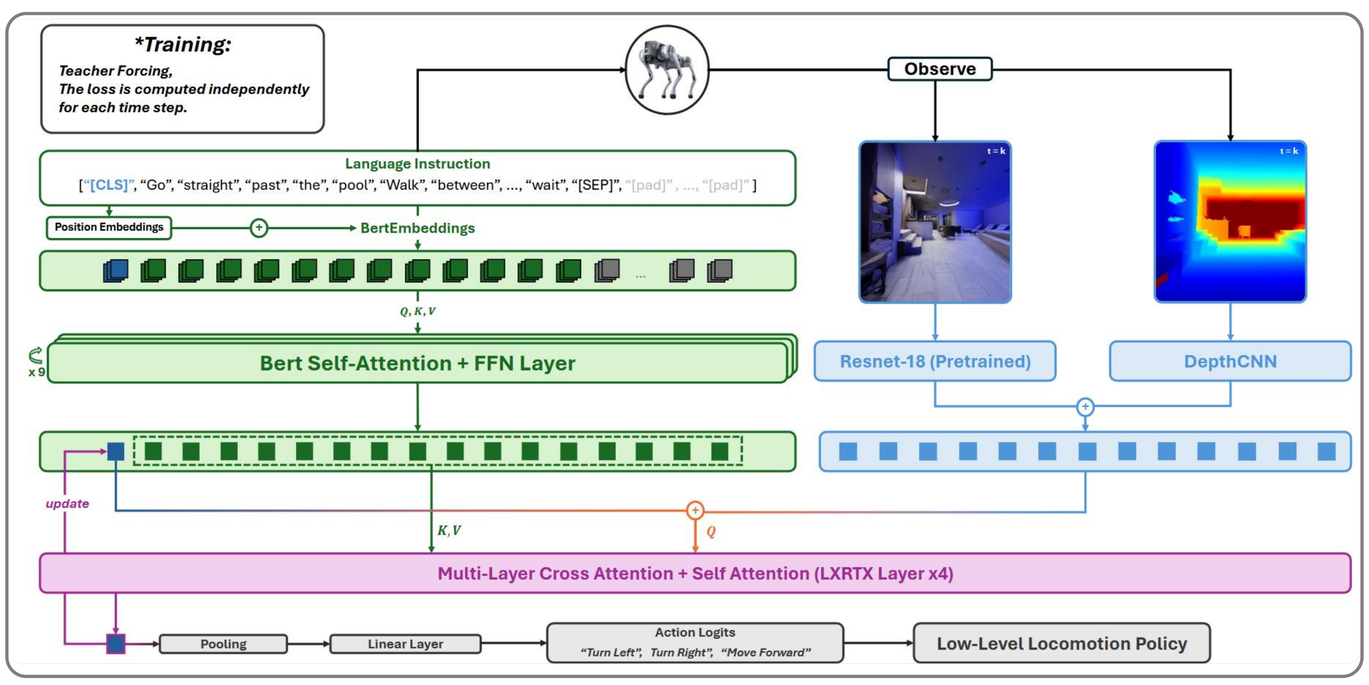

**Environment**: Nvidia Omniverse Isaac Sim Simulator
**Agent**: Unitree Go2 Quadruped Robot
**Perception**: RGB-D Camera Images
**Instruction**: e.g., "Go straight past the pool… you will wait."
**Success Condition**: the robot agent navigate to a position close (<= 1m) to the target location.

## VLN-Go2-Matterport Dataset




- **869 expert-guided episodes** across diverse Matterport indoor environments
- **Multimodal data streams**: RGB (1280×720), depth maps (640×480), discrete actions.
- **Quadruped-specific perspective**: Camera observations are Unitree Go2's body-mounted viewpoint.
- **Hybrid control**: Combines high-level discrete actions with low-level velocity execution.
- **Success-filtered**: Only contains episodes where the robot reached within 1m of the goal.
- **Temporal alignment**: Action-observation pairs recorded at 3Hz (Inference/20 sim steps).

## Recurrent-VLN-Bert-Isaac



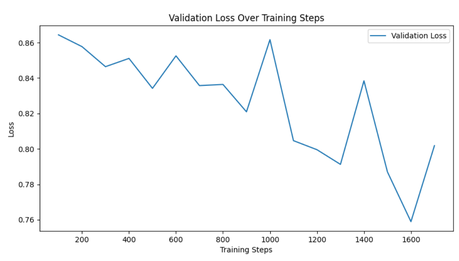

**Inputs:**
- Language instruction, RGB and depth images.

**Architecture:**
- Language embedded via BERT.
- Visual features from RGB (ResNet-18) and depth (CNN), fused and projected.
- State updated recurrently with cross-attention.

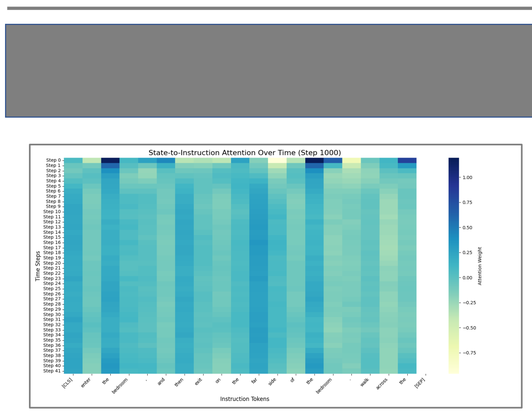
**Key Operations:**
- **Visual features**: $E_{vis,t} = Proj([ResNet18(RGB_t), DepthCNN(depth_t)])$
- **State update**: $S_t = [s_{\{t-1\}}, E_{\{vis,t\}}]$, $s_t = Transformer(S_t, E_{lang})[0]$
- **Action**: $a_t = argmax(softmax(Ws_t + b))$

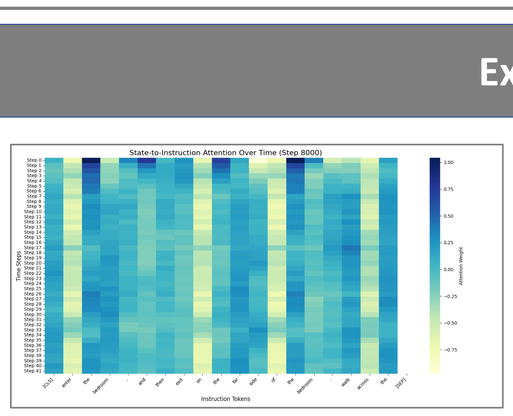
**Training**: Imitation learning with cross-entropy loss

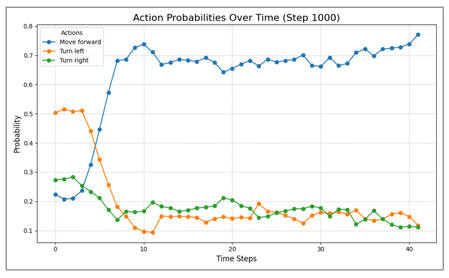
Best: 70.38% action accuracy across all 869 episodes.

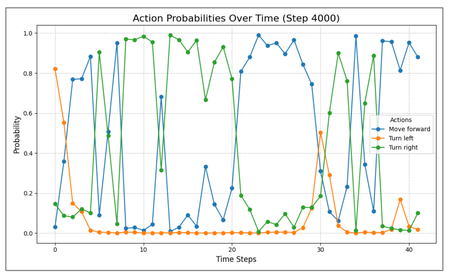



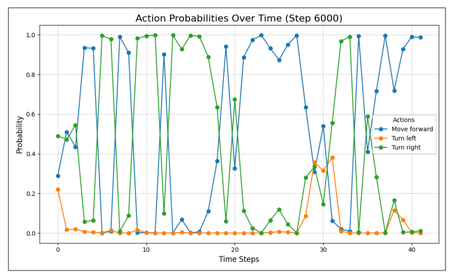

## Experiment And Analysis

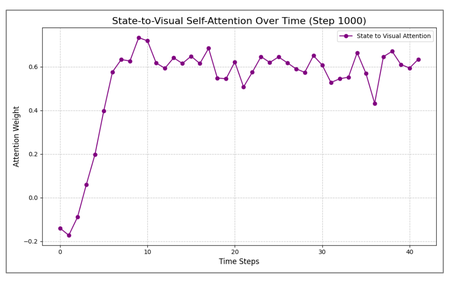


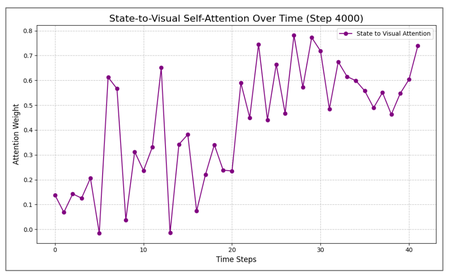



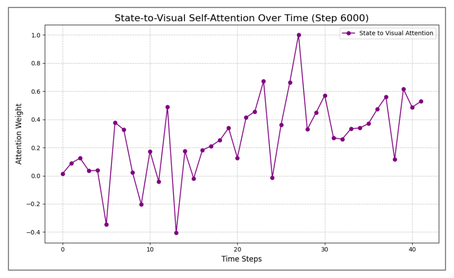




## Future Works

- Address **LLM hallucinations** in VLN by integrating robust spatial information (e.g., depth maps) for grounded reasoning. (Please see the LLM Case Study)
- Advance **spatial-temporal reasoning** techniques to improve model dynamic, real-world navigation tasks. (A view from Professor Li fei-fei)
- Explore **efficient training strategies** for **World Models and VLA models** to reduce reliance on massive datasets/GPU resources.
- Investigate **continual learning** paradigms for Embodied AI to adapt to ever-changing environments beyond static pre-training.
- Re-evaluate the scalability of **pre-trained models** in dynamic settings and propose alternatives to data-driven scaling.

- **Model Performance**: **70.38% dataset accuracy** but exhibited overfitting in later training stages, with a bias toward the "Move Forward" action. **40% success rate** in simulator tests, struggling with paths requiring multiple turns (*compounding errors*).
- **Key Insights**: Attention shifted to semantically rich words (e.g., "bedroom") and visual cues (e.g., depth patterns) as training progressed. Limited dataset size (**869 episodes**) caused overfitting, yet depth features enabled partial generalization.
- **Limitations**: **Action bias** from imbalanced data; **local memory** ([CLS] token) hindered global environment reasoning. **Hardware constraints** (3Hz action frequency) and sim-to-real gaps reduced real-world applicability.

Main Repository

Video Demo



Acknowledgement: I sincerely thank my supervisor, Prof. Junhui David Hou, for his guidance. — Kenny Lik Hang Wong (klhwong3-c@my.cityu.edu.hk)

Department of Computer Science
香港城市大學
City University of Hong Kong