Learning Object-Centric Local Navigation from RGB Demonstrations

Tzu-Hsien Lee* , Fidan Mahmudova*
Karthik Desingh
*equal contribution

Introduction:

Object-centric local navigation, which guides a robot to a precise object-relative pose with **centimeter-level** translation and degree-level rotation, is critical for downstream manipulation tasks such as dexterous handover, object placement, and door operation. We present a compact, end-to-end, vision-based imitation learning framework that relies **exclusively on RGB** sensor observations - no maps, 3D reconstruction, object models, depth sensing, or LiDAR are required. The approach is validated on the Boston Dynamics **Spot robot** in real-world scenarios using a lightweight architecture that combines a frozen DINOv2[1] encoder with a simple MLP-based action decoder.

Network or Framework details:

Our policy uses a shared visual encoder (either ResNet-18[2] or a frozen DINOv2 ViT[1]) to process all current images and goal images through the same backbone. The attention-refined[3] embeddings are concatenated and passed into a compact MLP decoder that outputs the low-level displacement commands (Δx, Δy,Δθ).

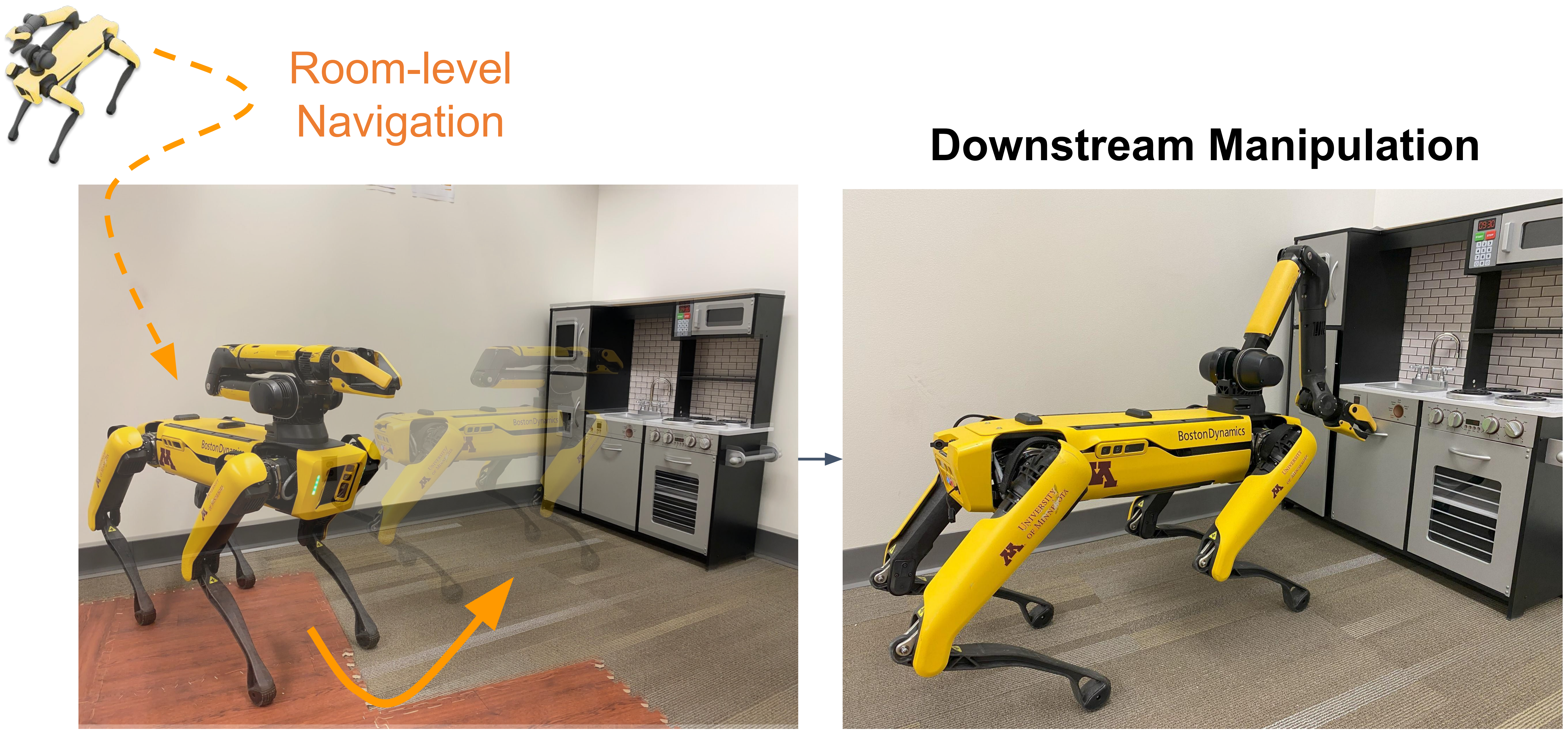
The dataset collection was automated utilising SPOT's onboard navigation system. 297 autonomous trajectories(~3,300 image-action pairs) were collected.

Results:

Success Criteria: reaching target pose within **20 cm translation** & **5° orientation**.

Model	Success Rate
ResNet18 + MLP	36% (18/50 episodes)
DinoV2 + MLP	54% (27/50 episodes)

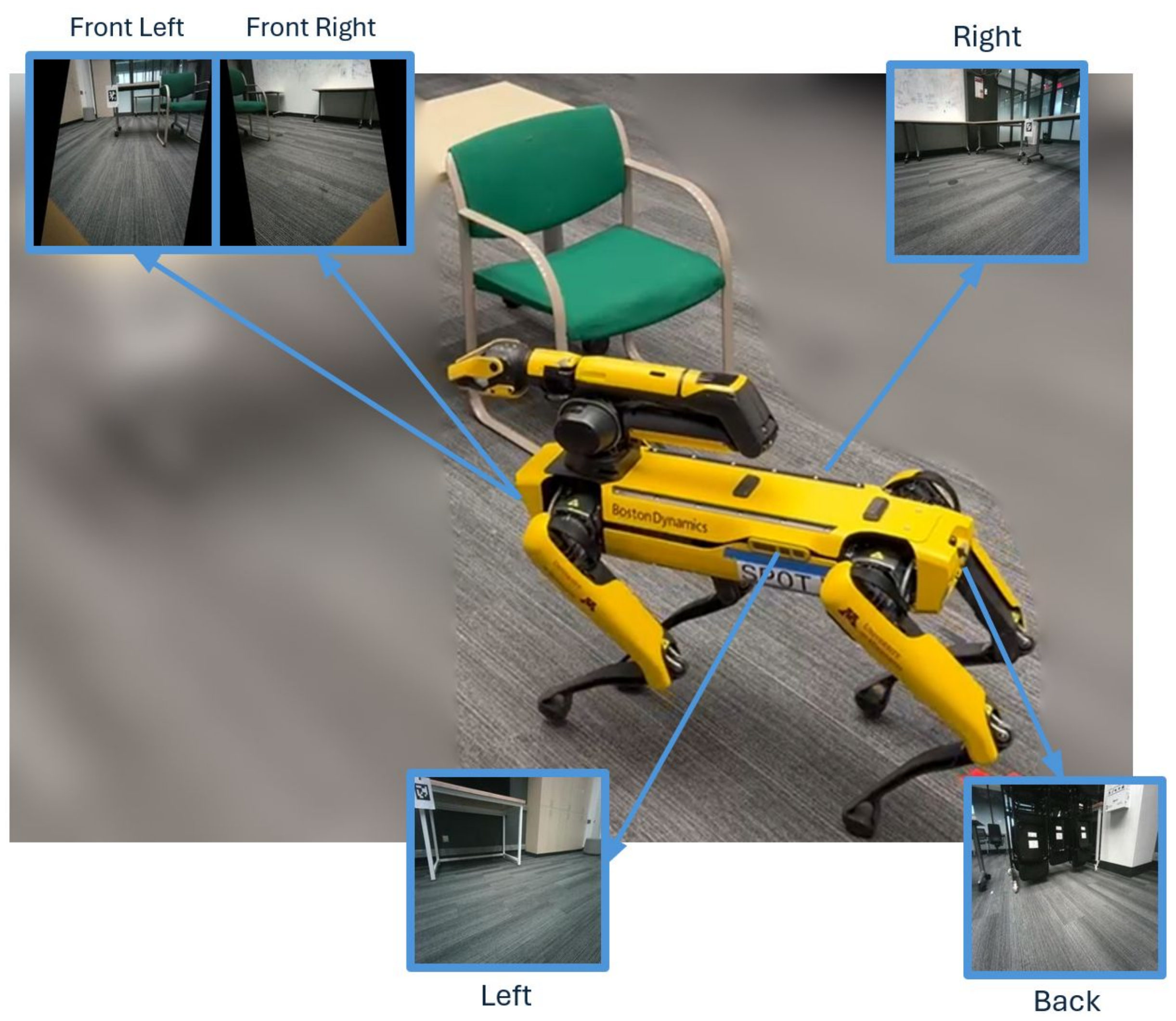
Object-centric local navigation is important for successful downstream manipulation



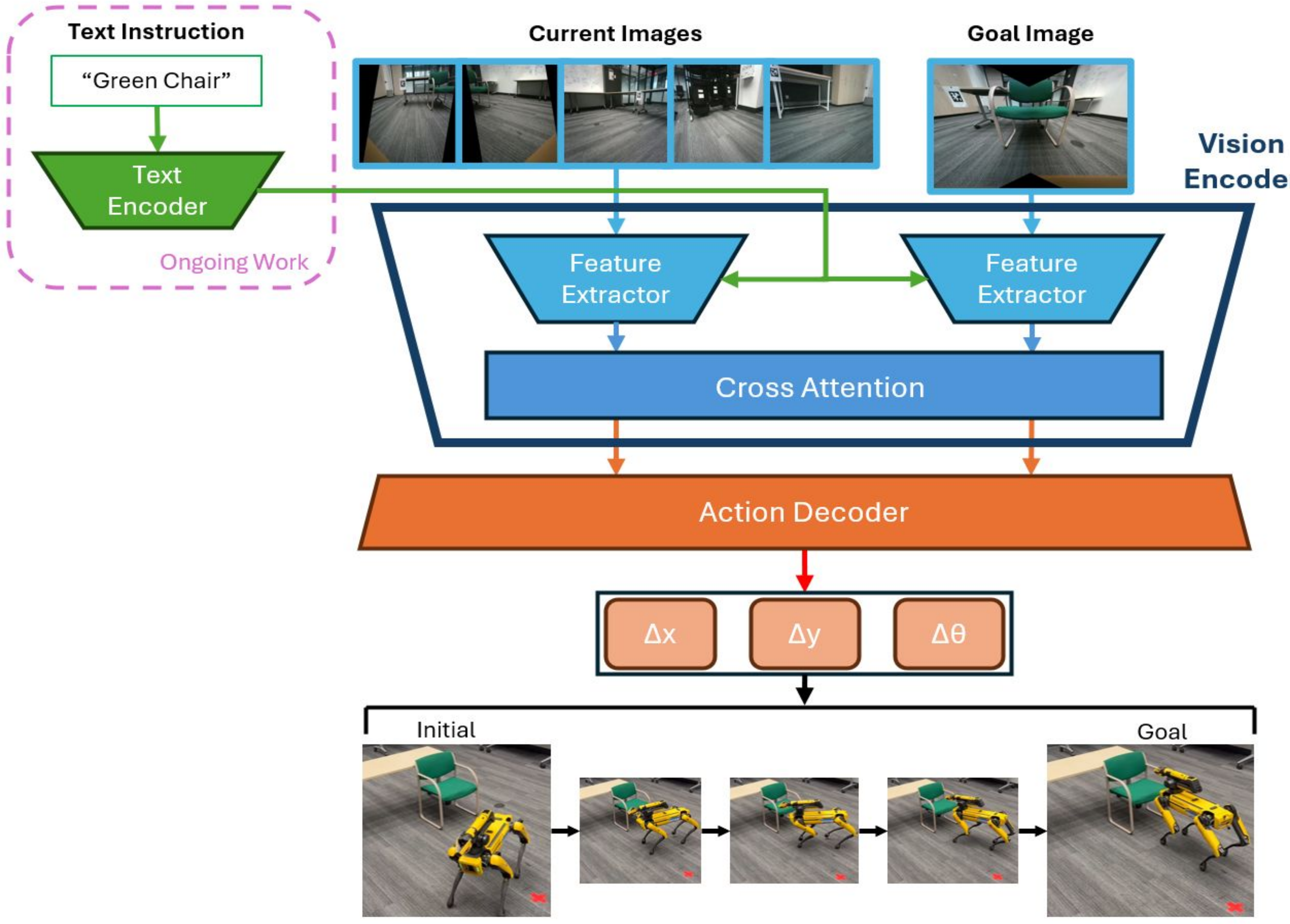
Object-Centric Local Navigation

"Door opening after local-navigation"

Motivation



SPOT onboard vision



Architecture

[1]: Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).
[2]: He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[3]: Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

MINNESOTA ROBOTICS INSTITUTE
UNIVERSITY OF MINNESOTA

Robotics: Perception & Manipulation Lab