Decoupling Vision and Motion: Object-Centric Representations for Enhanced Manipulation

Tongmiao Xu [1]

[1]Department of Computer Science, National University of Singapore

INTRODUCTION

A core challenge in robot **imitation learning** lies in the entanglement between a robot's policy and its specific camera perspectives and object placements. This makes **policy generalization** difficult. We tackle this by proposing an **object-centric learning** framework that decouples observed **object motions (vision) and robot actions (motion)**, allowing generalization over varying **viewpoints and object positions.**

METHODS

We aim to learn a robot policy $\pi^*(a \mid o)$ using a demonstration dataset $D = \{(o_t, a_t)\}_{t=1}^{N}$. At each time step t, the policy takes the current observation o_t and predicts a sequence of future actions $(a_t, \ldots, a_{t+k-1})$.

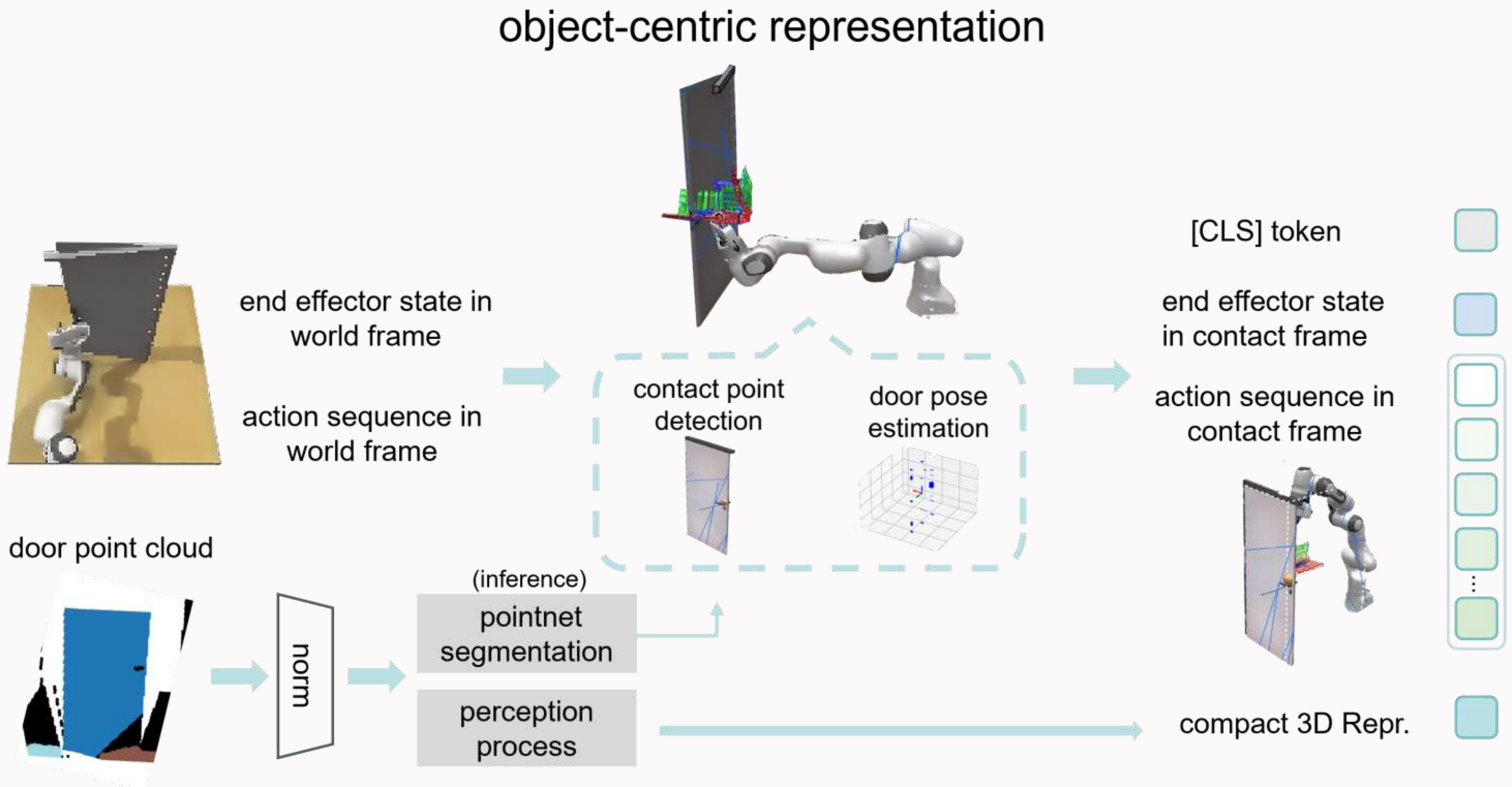
object-centric representation
end effector state in world frame
action sequence in world frame
door point cloud
norm
(inference)
pointnet segmentation
perception process
contact point detection
door pose estimation
[CLS] token
end effector state in contact frame
action sequence in contact frame
compact 3D Repr.

Fig 1. **Object-centric policy learning.** The object-centric frame is defined using the detected contact point of the trained segmentation model and the estimated object pose. During training, we transform both the robot actions and oracle states into the object-centric frame. Similarly, object motion is normalized with respect to the object's orientation.

OBJECTIVES

We consider the problem of enabling a robot to imitate a door-opening task from a limited number of expert demonstration videos. The goal is to develop a policy that generalizes to unseen camera viewpoints and object placements that are out of distribution.

Object-centric Representation. Specifically, in the door-opening task, the robot must first establish and maintain *contact* with the door before *manipulating* its plane along the pivot axis. Therefore, we define an object-centric frame where the rotation aligns with the door plane and the translation corresponds to the contact point on the handle.

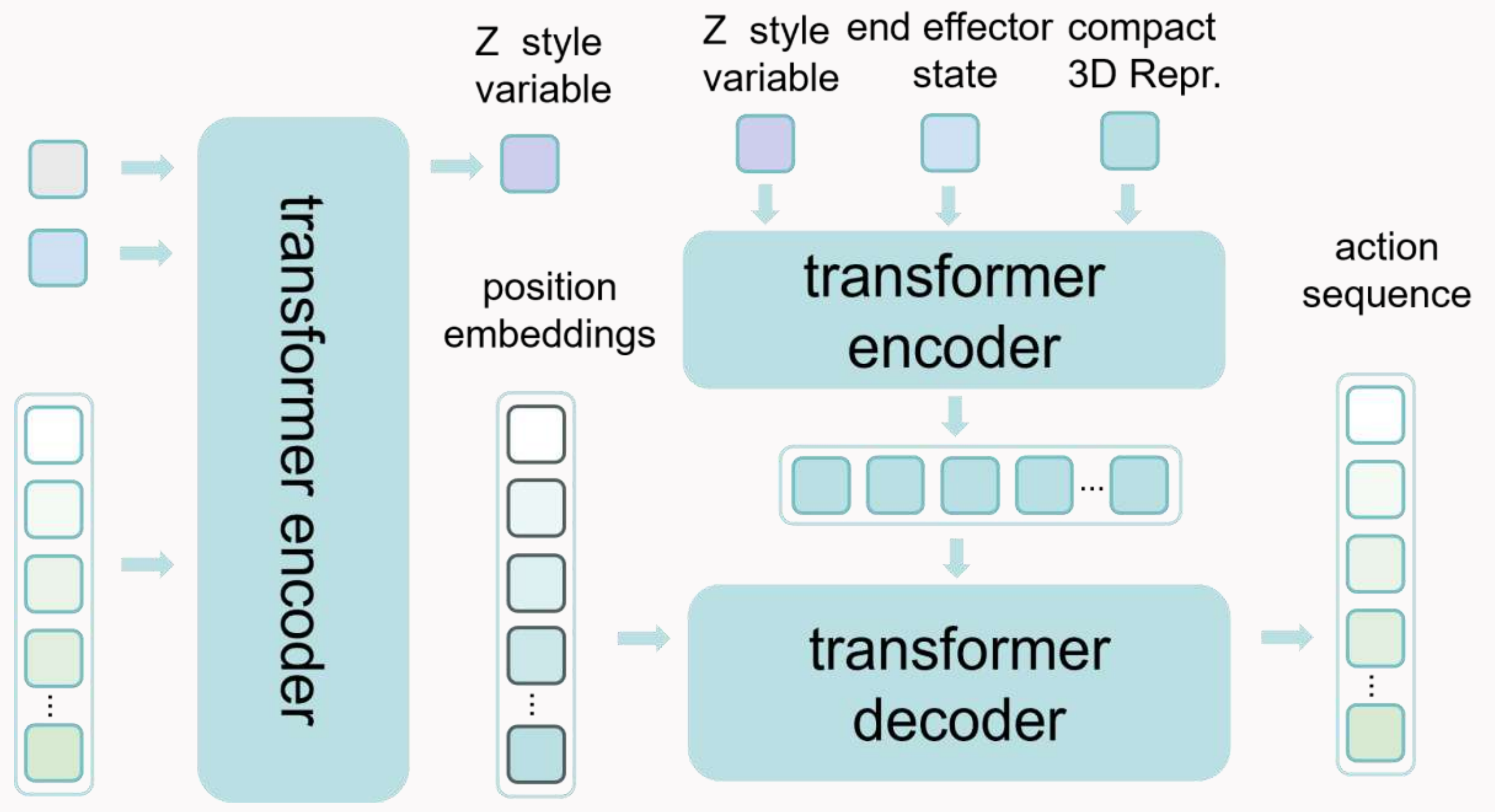

Fig 2. **Training pipeline**. We train an end-to-end policy as a Conditional VAE (CVAE) with an encoder-decoder architecture. The encoder processes object-centric observations and actions to train the decoder and the CVAE decoder is used to predict the action sequence.

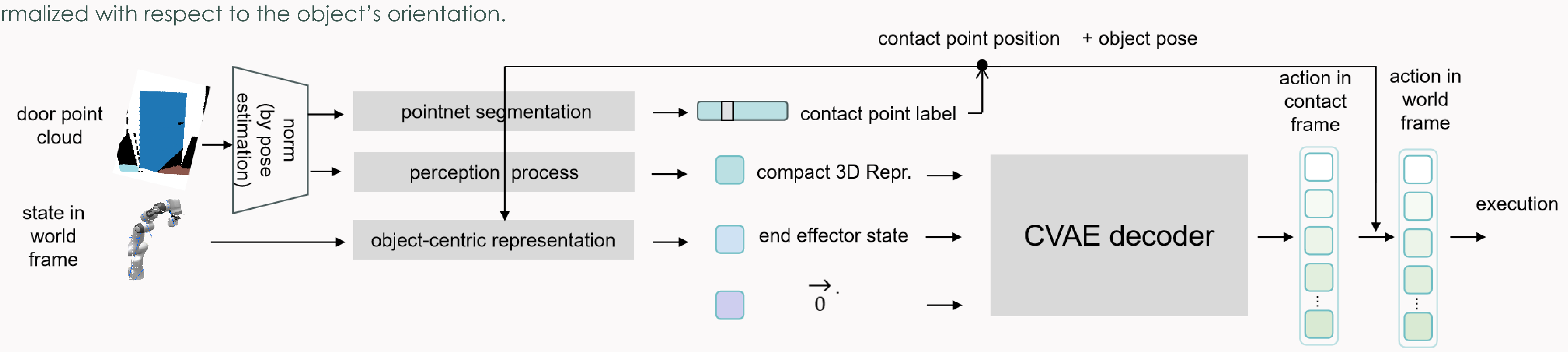

Fig 3. **Inference pipeline**. During inference, we first establish object-centric representation from the predicted contact point and estimated object pose. The decoder conditions on the normalized point cloud, end effector states, and latent variable z to predict object-centric end effector trajectory during inference. The predicted actions are transformed back to the world frame for execution based on frame.

RESULTS

We test our policy's ability to generalize across different door positions and camera views in a door-opening task in simulation, compared with another two baselines.

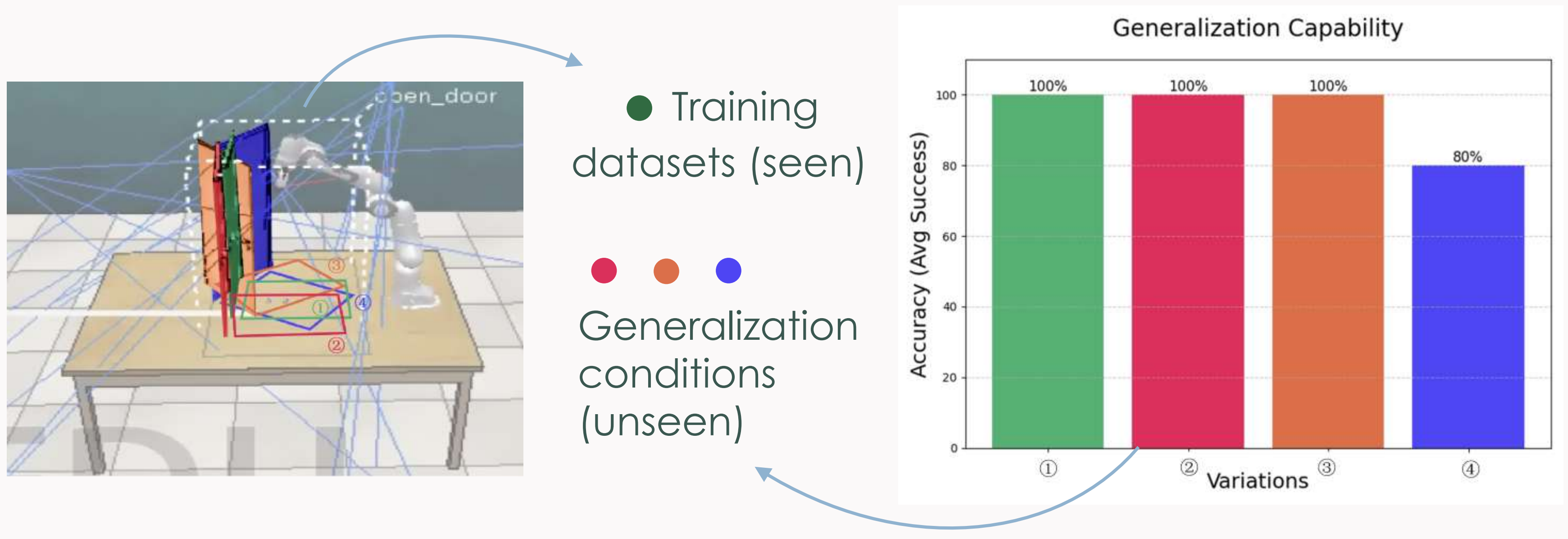

Fig 4. **Task variations and generalization results in simulation.** The average success rates indicate our approach's strong generalization ability across unseen camera viewpoints and object placements (variations ②, ③, ④)

Tbl 1. **Baseline comparison in simulation.** Our method is compared with ACT [1] and 3d diffusion policy[2].

Spatial Generalization	①	②	③	④	average (%)
ACT	✓	✗	✗	✗	30.0
DP3	✓	✗	✗	✗	15.0
Ours (Object-Centric)	✓	✓	✓	✓	95.0

🎯The results show that our method demonstrates strong generalization to unseen viewpoints and placements.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to Prof.David Hsu and Dr. Hanbo Zhang who gave me the golden opportunity to do this project and members of AdaComp@NUS for their invaluable comments and insightful discussions.

REFERENCES

[1]T. Z. Zhao et al., "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," *arXiv:2304.13705*, 2023.

[2]Y. Ze et al., "3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations," *arXiv:2403.03954*, 2024.

CONTACT

XU TONGMIAO | INCOMING UIUC MSCS STUDENT
EMAIL: XTONGMIAO@GMAIL.COM
HOMEPAGE: https://tamphie.github.io/
LINKEDIN: https://www.linkedin.com/in/xtongmiao/