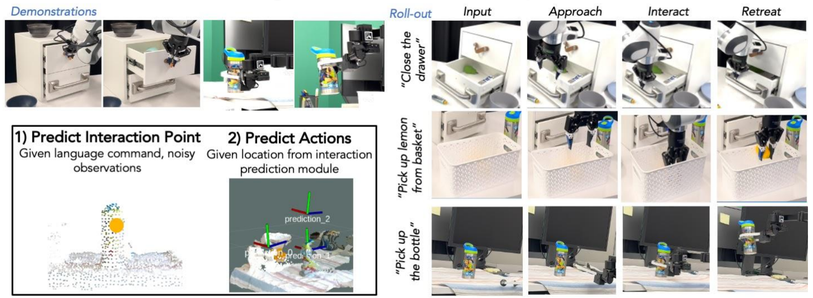
Spatial-Language Attention Policies for Efficient Robot Learning

Priyam P.*, Jay V.*, Sam P.‡ and Chris P.*
*Meta AI, ‡CMU

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

- Robotics decision-making involves reasoning over high-dimensional visual and spatial properties of a scene. We present **a method to tokenize the visual world**, while preserving the spatial information, such that we can **leverage Transformers to learn robust and generalizable action representations**

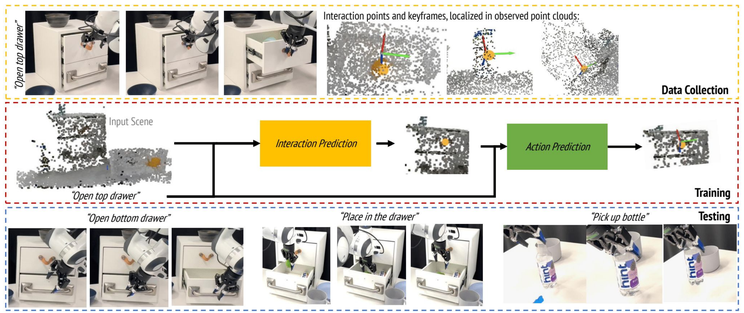


- No assumptions are made on camera placement, methodology works for both static and mobile manipulation cases
- We trained a multi-task model for eight tasks involving four manipulation skills
  - 80% success rate in the real world, and a 47.5% success rate when unseen clutter and unseen object configurations are introduced
  - **Improvement of 30% over prior work**[1] (+20% with clutter)

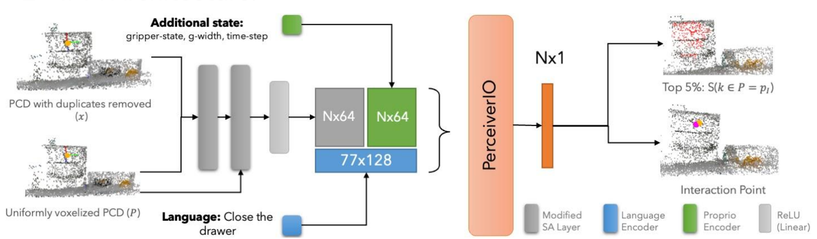
Prior Work

- ViT-style features [2] and their 3D extensions [1] when combined with attention-based learning mechanisms have shown great promise. However, these representations are still on a static grid.
- [3] introduces a way to interweave linguistic and multi-media information, while [4] presents a way to learn free-form spatial relations in 3D space

Approach



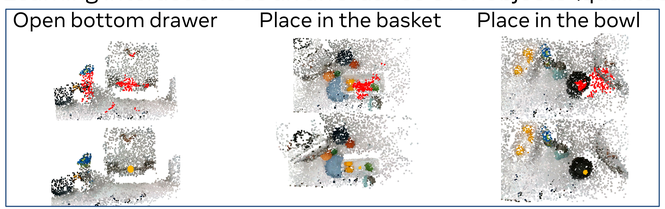
- Given user-provided skill demonstrations, with language describing the skill, we take a hybrid approach to predict robot actions:
  - The **Interaction Prediction Module (IPM)** uses the PerceiverIO Transformer and our tokenized spatial representation to predicts an interaction point on the object for the given skill (can be thought of as affordance)
  - Interaction point is used by the **Action Prediction Module (APM)** to predict the robot actions relative to this point to fulfill the skill
- IPM Architecture:



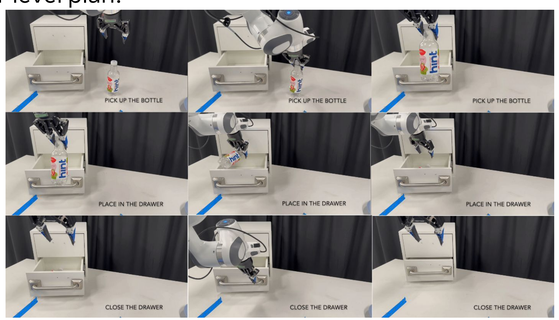
- Point-cloud is tokenized using PointNet++ layers which are sequenced with language tokens and an affordance representation is learnt over this point-cloud based on expert demonstration

Results

- Learning interactions as an affordance over objects (qualitative)



- Since skills are language conditioned, we can also chain them given a high-level plan!



References

[1] Shridhar, M., Manuelli, L. and Fox, D., 2023, March. Perceiver-actor: A multi-task transformer for robotic manipulation. In Conference on Robot Learning (pp. 785-799). PMLR.

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale

[3] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model

[4] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems, 30, 2017