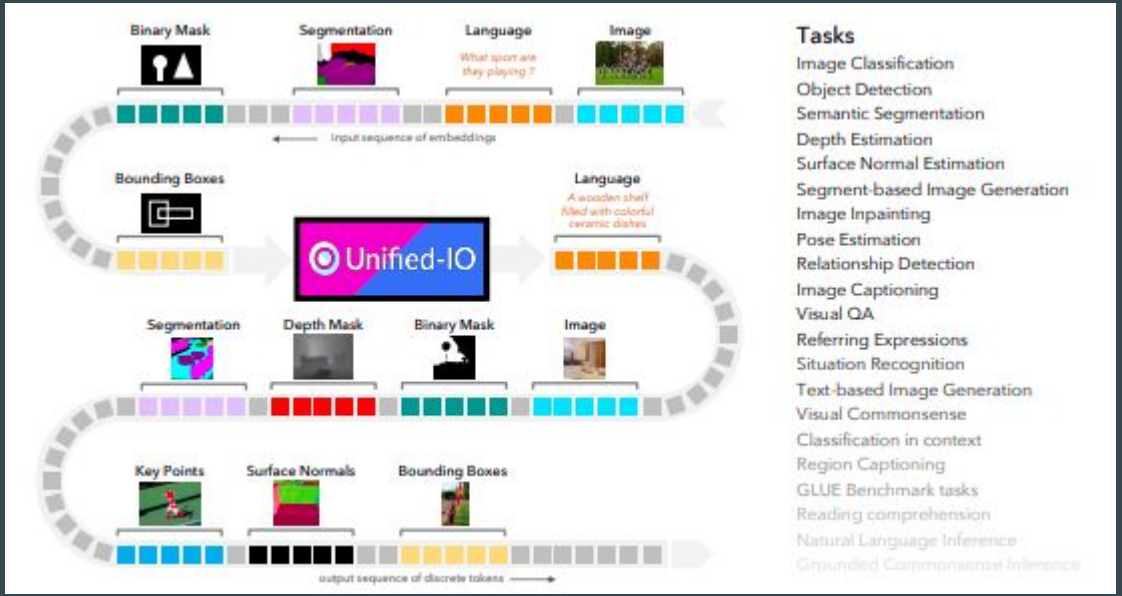
Multimodal NLP
Applying UnifiedIO to a Robotics Domain

•••

Gregory LeMasurier

UnifiedIO



Lu, J., Clark, C., Zellers, R., Mottaghi, R., & Kembhavi, A. (2022).
Unified-io: A unified model for vision, language, and multi-modal tasks.
arXiv preprint arXiv:2206.08916.

VIMA

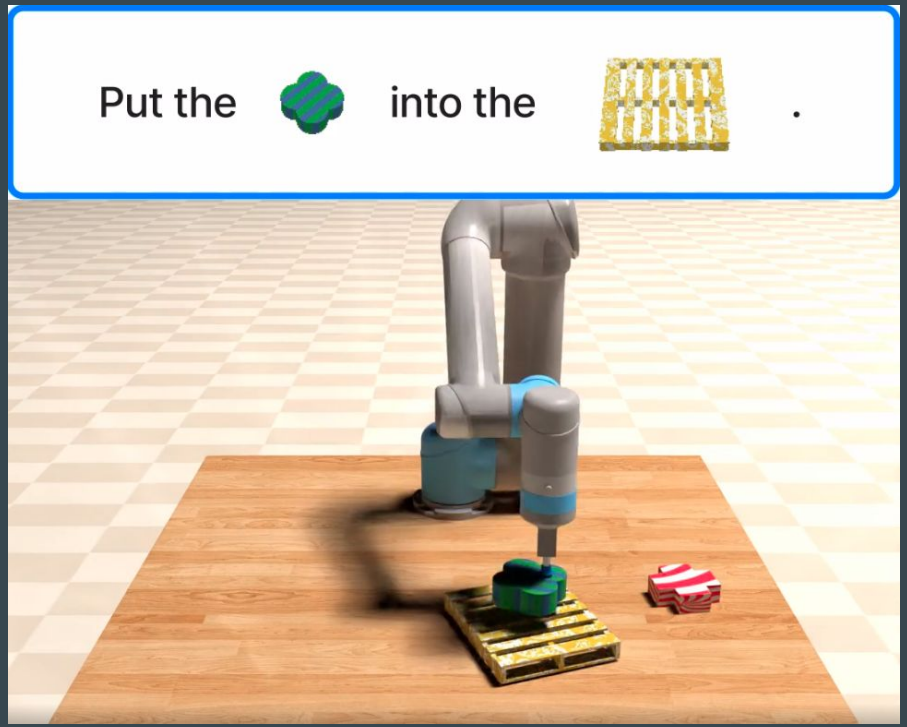
Generalizability in Robotics Tasks

- **Simple Object Manipulation**
- Visual Goal Reaching
- Novel Concept Grounding
- One-shot Video Imitation
- Visual Constraint Satisfaction
- Visual Reasoning

Provide VIMA-Data and VIMA-Bench

Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., ... & Fan, L. (2022). *Vima: General robot manipulation with multimodal prompts.* arXiv preprint arXiv:2210.03094.

VIMA Simple Manipulation Task-01-L1

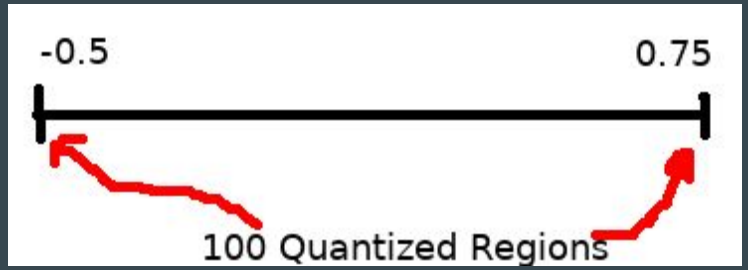


Research Question

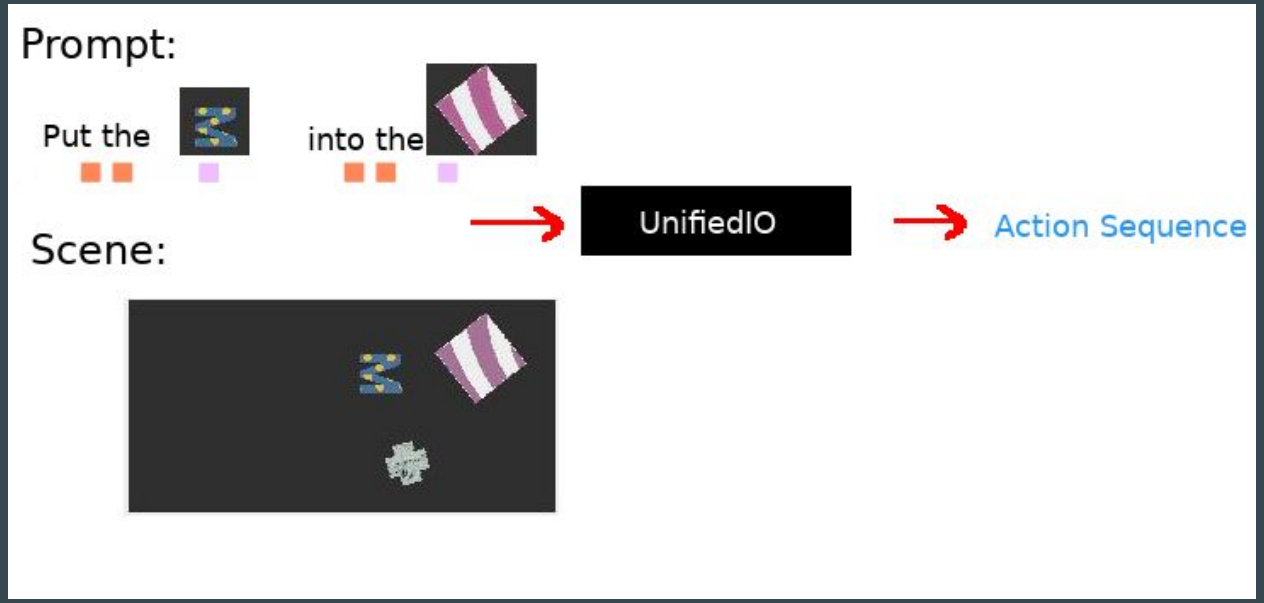
Can Unified-IO be applied to the robotics domain, using sensor and language input to generate motor commands to enable a robot to complete VIMA benchmark tasks?

Data Preprocessing - VIMA Data

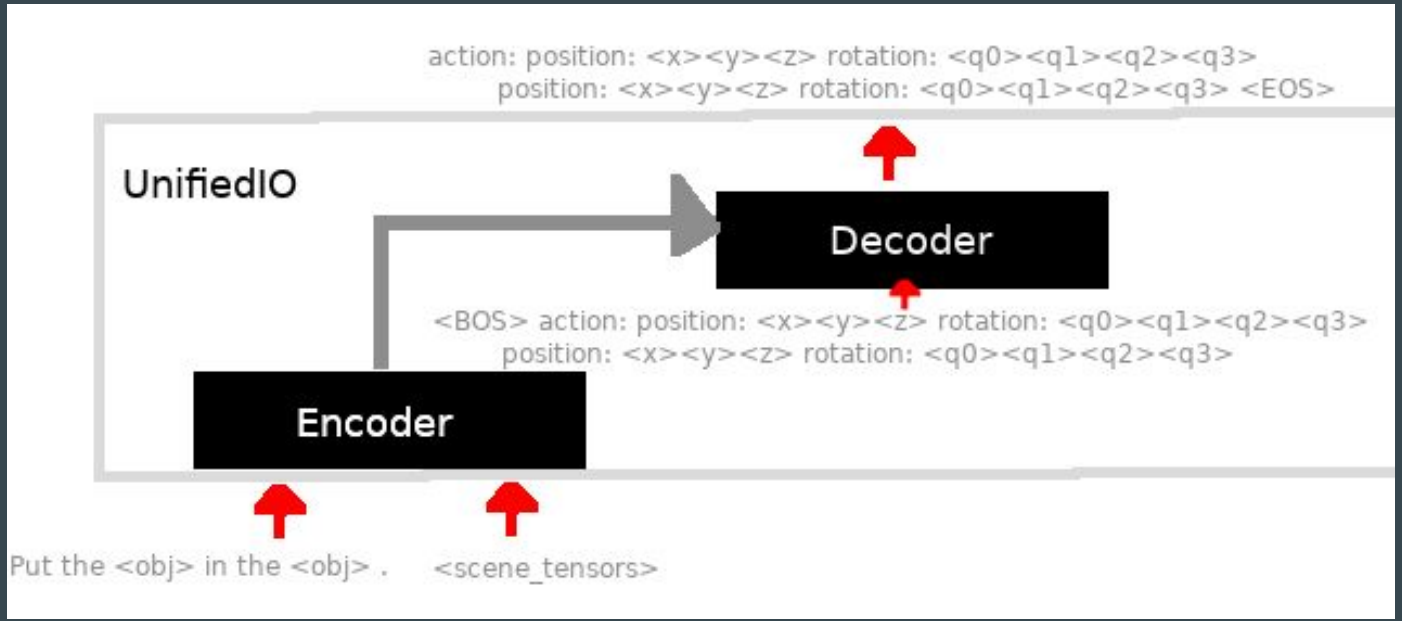
- rgb_top
  - First image is the initial state
    - Scene
    - Object tokens
- trajectory.pkl
  - Prompt
  - Object properties
  - Action bounds
  - Other misc. information
- action.pkl
  - Action poses to complete the task
    - Position(x,y,z), Rotation(q0,q1,q2,q3)
    - Quantized
  - 41/~50k samples had two positions for each action. These were excluded



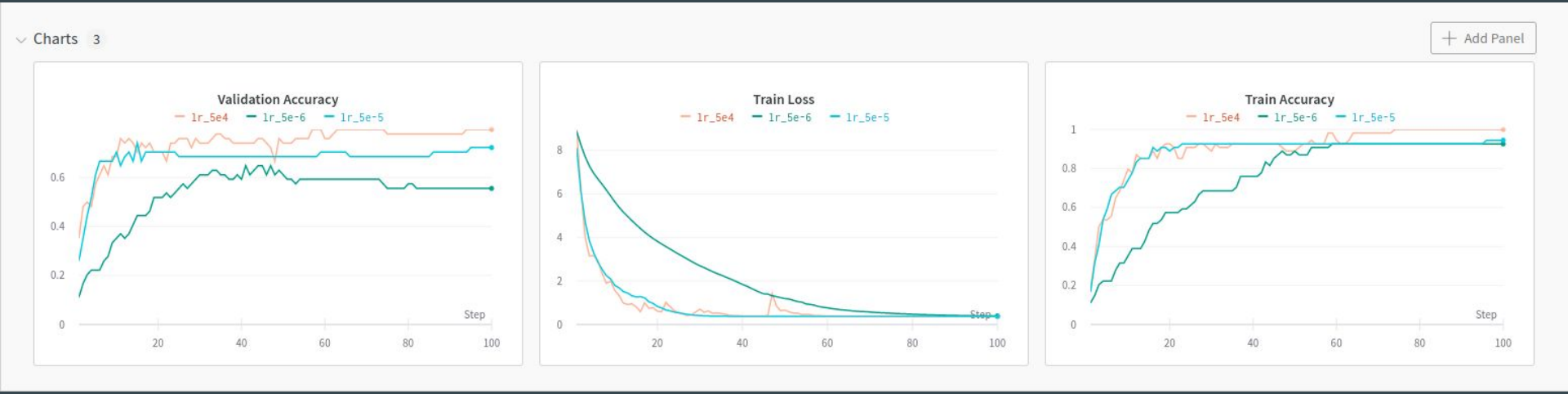
Model Inputs and Outputs



Model Inputs and Outputs



Identifying Parameters - Learning Rate



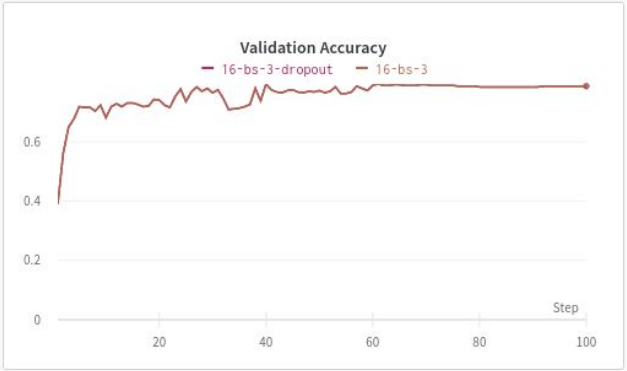
Model (small)

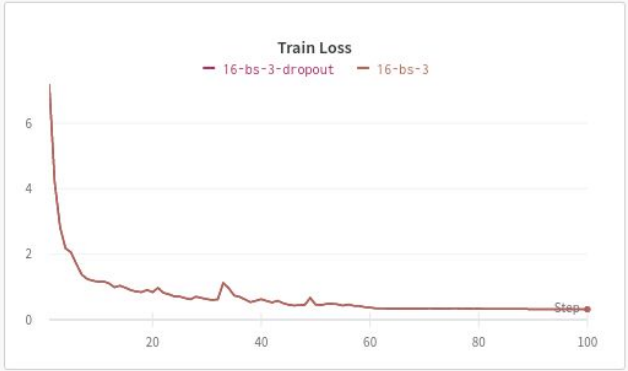
Batch size (3)*

16 train, 16 val

*The parameter evaluations were done on my personal computer

Identifying Parameters - Dropout





Model (small)

Batch size (3)*

16 train, 16 val

*The parameter evaluations were done on my personal computer

Identifying Parameters - Batch Size

Small (~14 M)

- Max batch size: 32

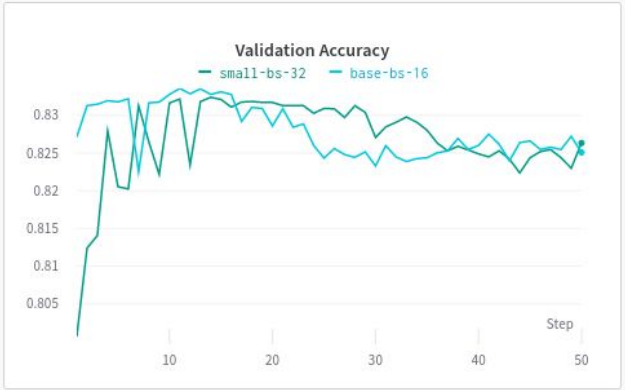
Base (~31M)

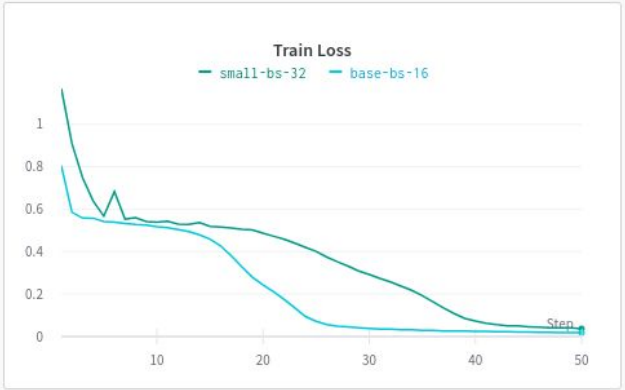
- Max batch size: 16

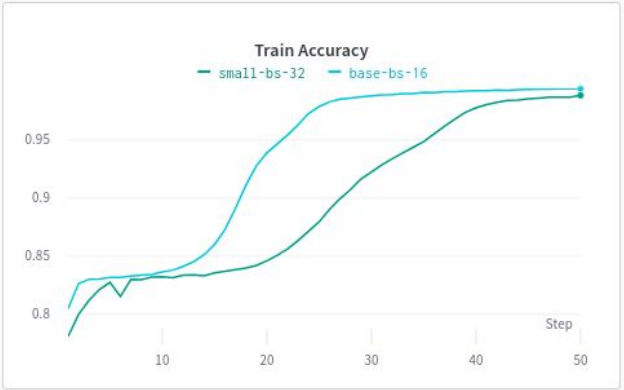
Training

Metrics:

- **Loss:** Cross-Entropy Loss
- **Accuracy:** % Token match

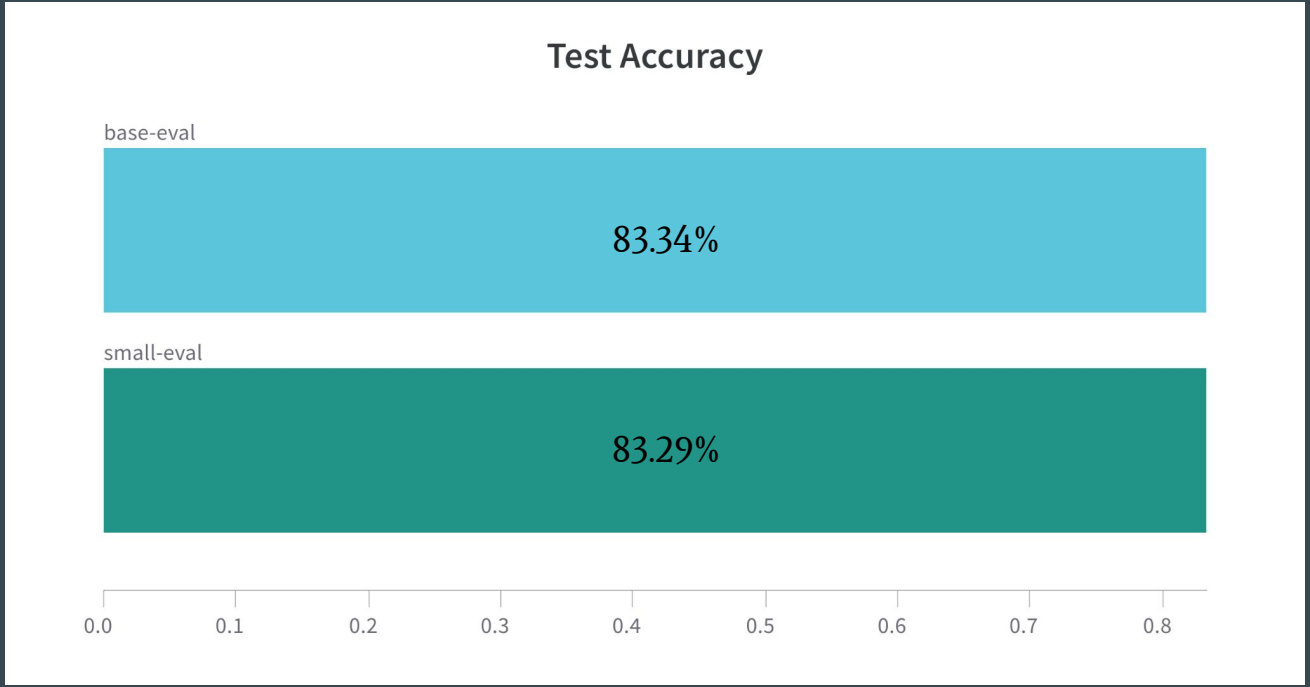




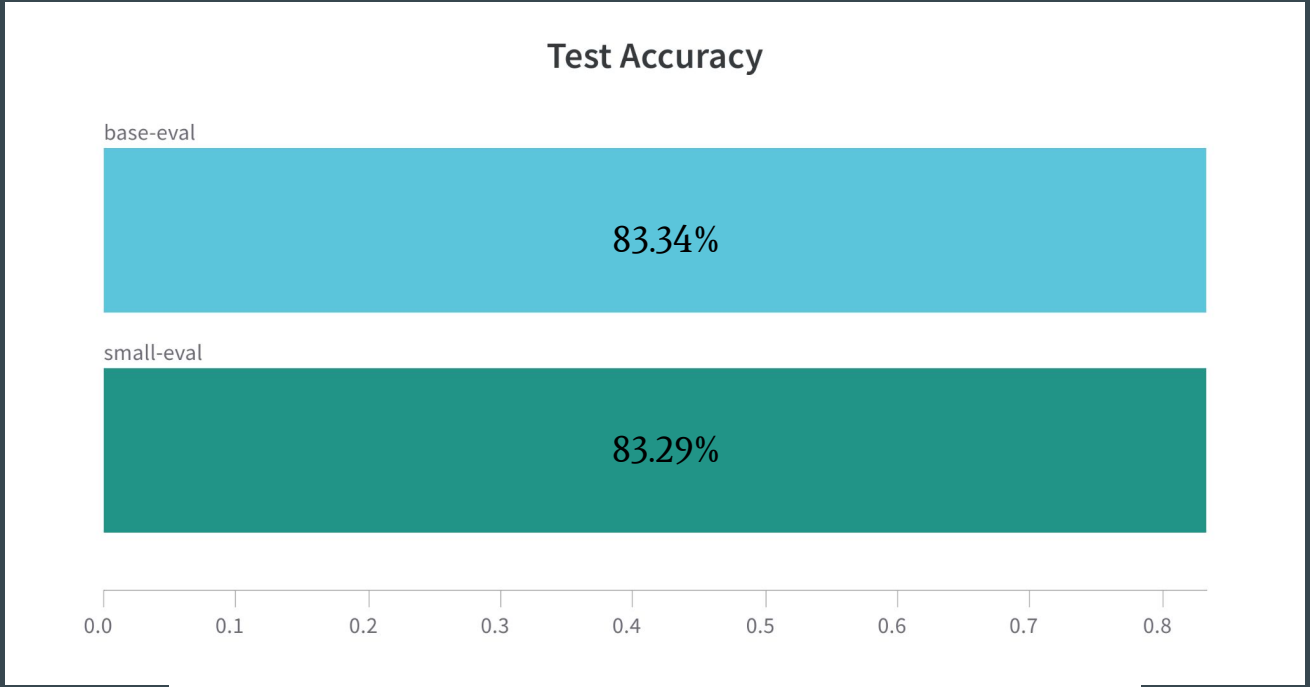


Checkpoints: Small (14) Base (11)

Evaluation - Token Match Accuracy

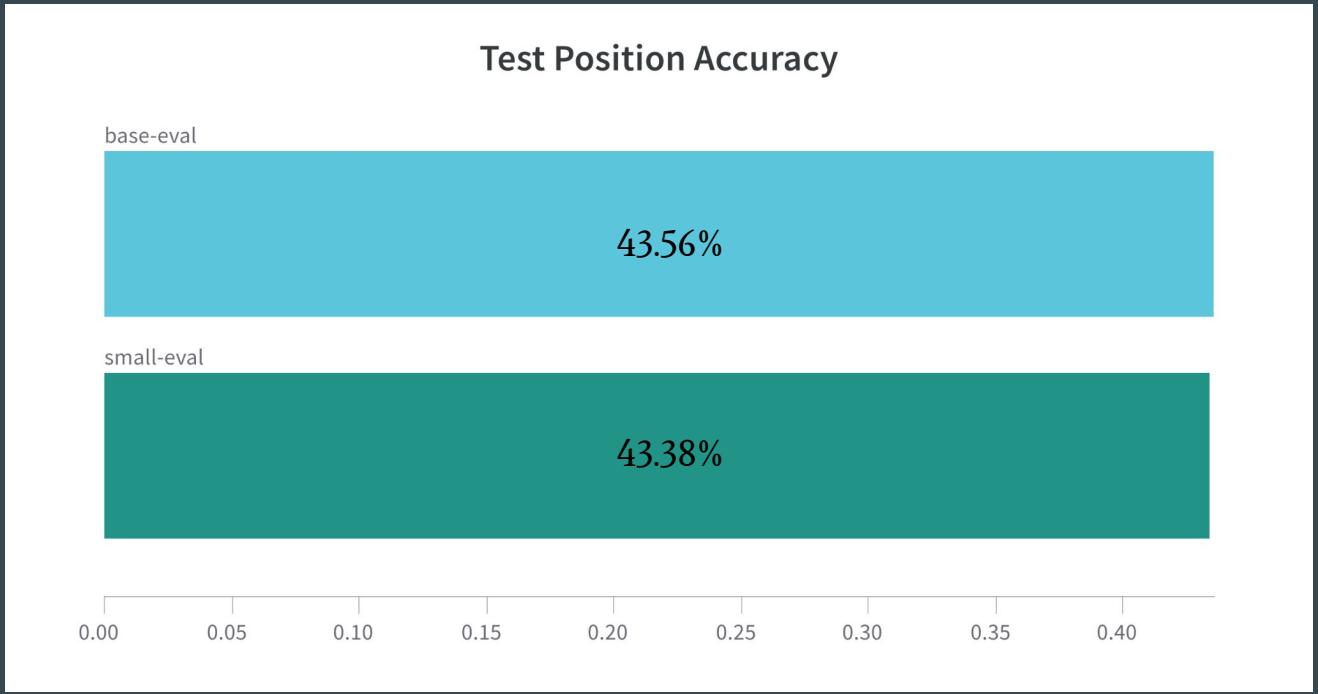


Evaluation - Token Match Accuracy



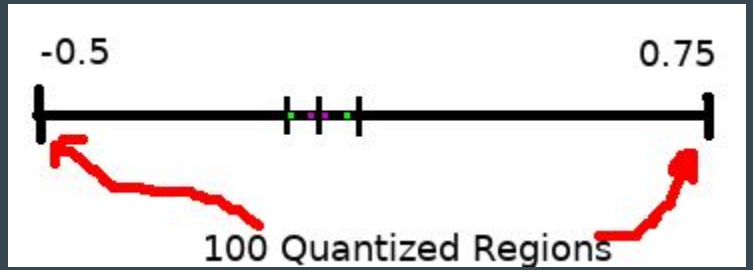
action: position: <x><y><z> rotation: <q0><q1><q2><q3>
position: <x><y><z> rotation: <q0><q1><q2><q3> <EOS>

Evaluation - Position Token Match Accuracy

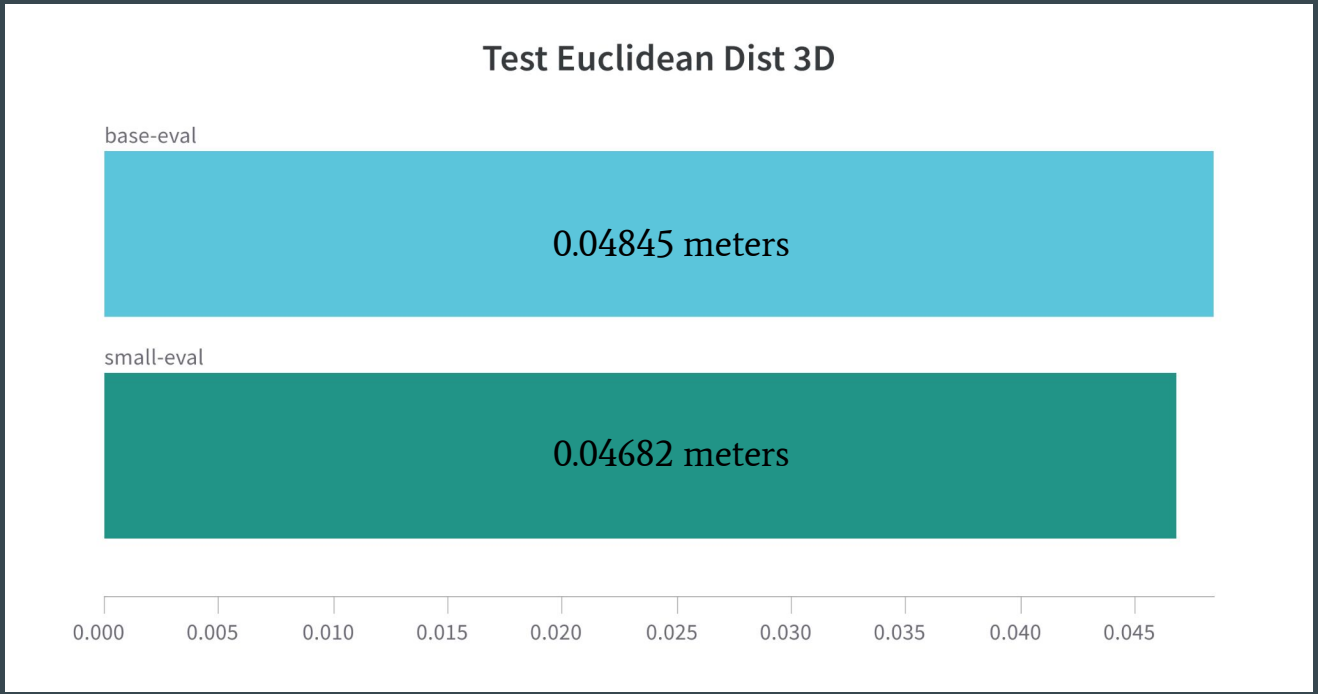


Evaluation - Position Token Match Accuracy

```
[[32016 32074 32056 32022 32042 32057]
 [32029 32032 32056 32020 32051 32057]
 [32012 32043 32056 32016 32068 32057]]
[[32029 32077 32056 32017 32038 32057]
 [32028 32031 32056 32018 32050 32057]
 [32013 32043 32056 32017 32069 32057]]
0.3888889
```

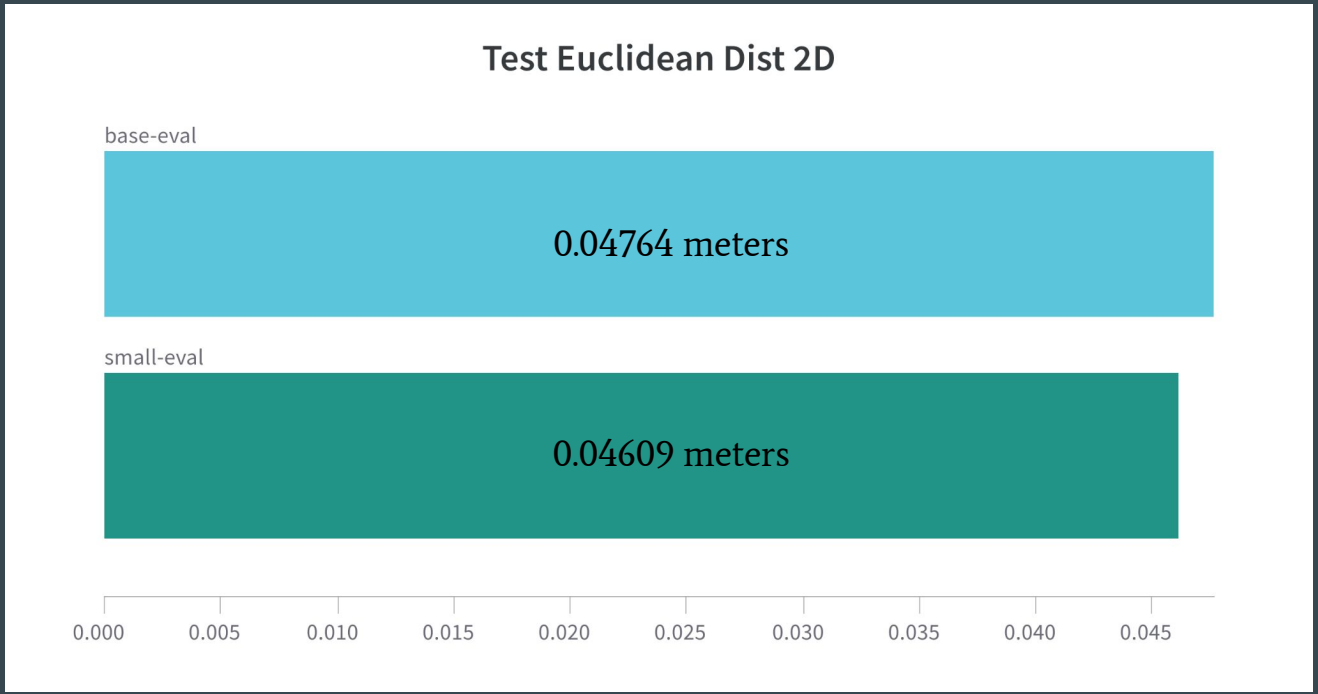


Evaluation - Euclidean Distance



Max: 2.165064

Evaluation - Euclidean Distance



Max: 1.767767

Evaluation - Quantization Error

Euclidean Distance from:

- Raw expected positions
- Quantized expected positions

Average error of: 0.01104 meters

Results

Method	Task 01 - L1 Generalization
VIMA (20M)	**100%**
Gato (20M)	62%
Flamingo (20M)	56%
Decision Transformer (20M)	59.5%
*UNIFIED-IO (small)(14M)	*Average Error: 0.04682 m*
*UNIFIED-IO (base)(31M)	*Average Error: 0.04845 m*