SpatialActor: Exploring Disentangled Spatial Representations for Robust Robotic Manipulation

AAAI 26 **Oral**

Hao Shi (石昊) @THU

https://shihao1895.github.io

Hao Shi[1], Bin Xie[2], Yingfei Liu[2], Yang Yue[1], Tiancai Wang[2], Haoqiang Fan[2], Xiangyu Zhang[3,4], Gao Huang[1]✉

[1]LeapLab, Tsinghua University, [2]Dexmal, [3]MEGVII, [4]StepFun

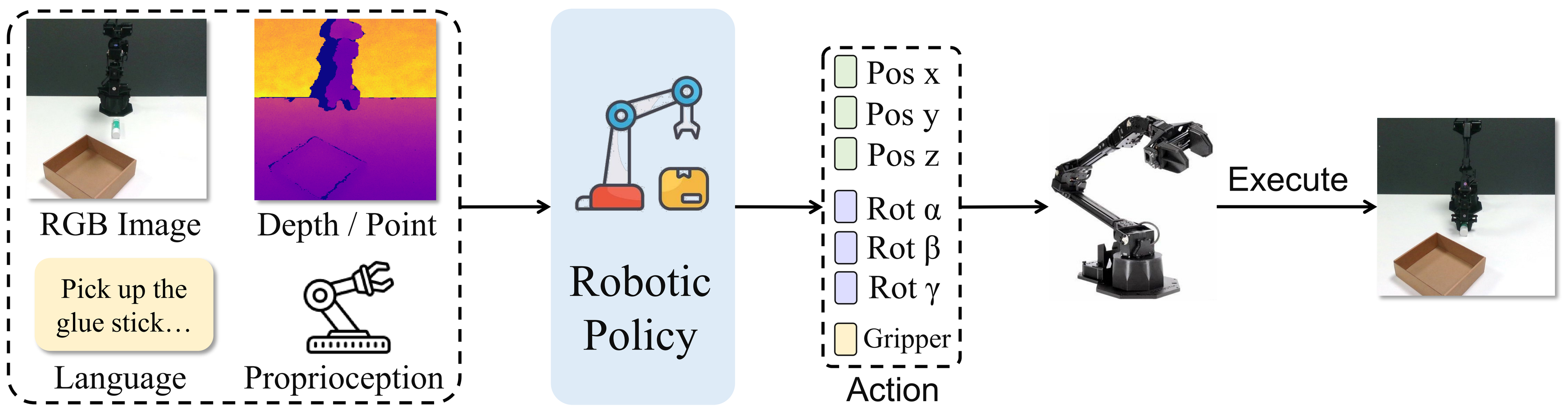
shi-h23@mails.tsinghua.edu.cn

Tsinghua University · DEXMAL

Page · Code

1 Introduction

① What is Robotic Manipulation Task?

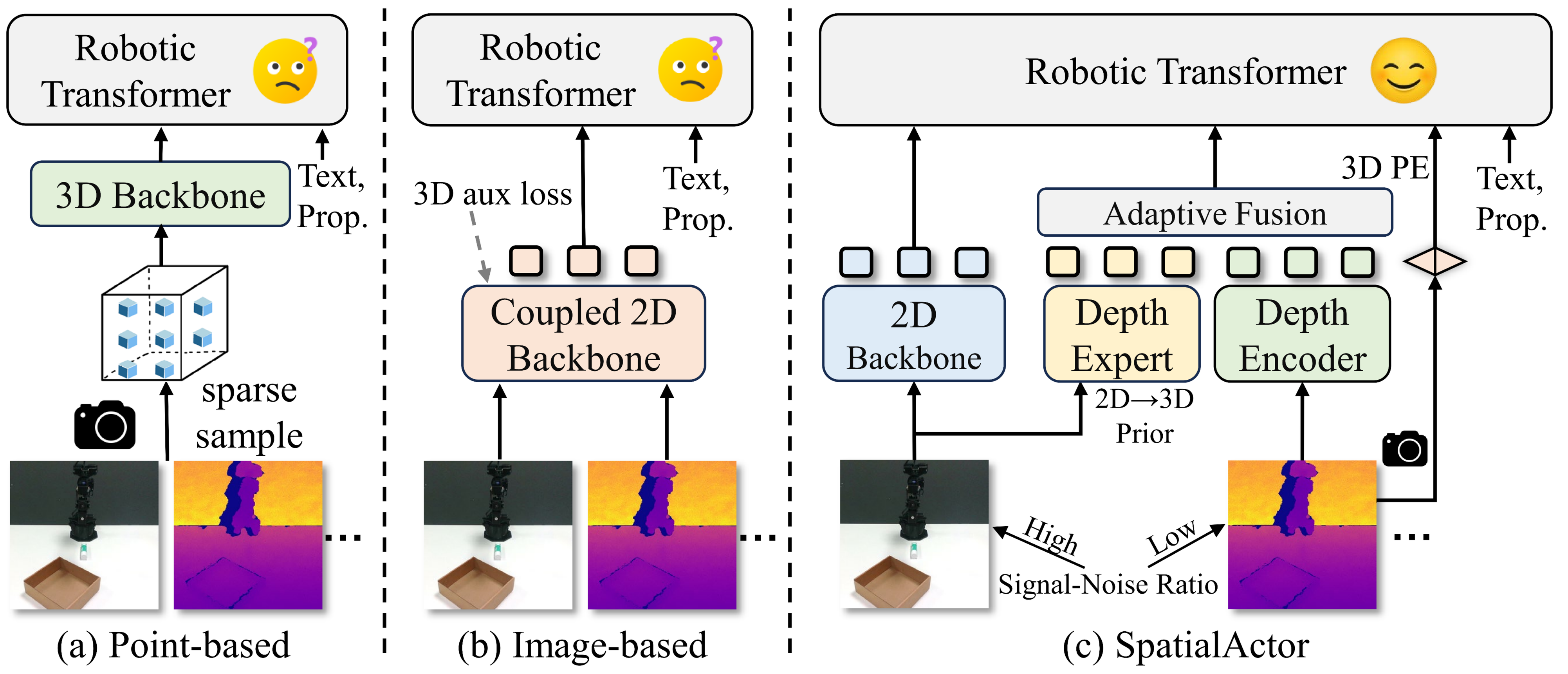


② Some Challenges in Robotic Manipulation

- ✓ Fine-grained spatial semantic understanding.
- ✓ Robustness to sensor noise.
- ✓ Low-level spatial structure inductive bias.

How to build robust spatial representations for robotic manipulation?

③ Comparison of Robotic Spatial Representations Solution



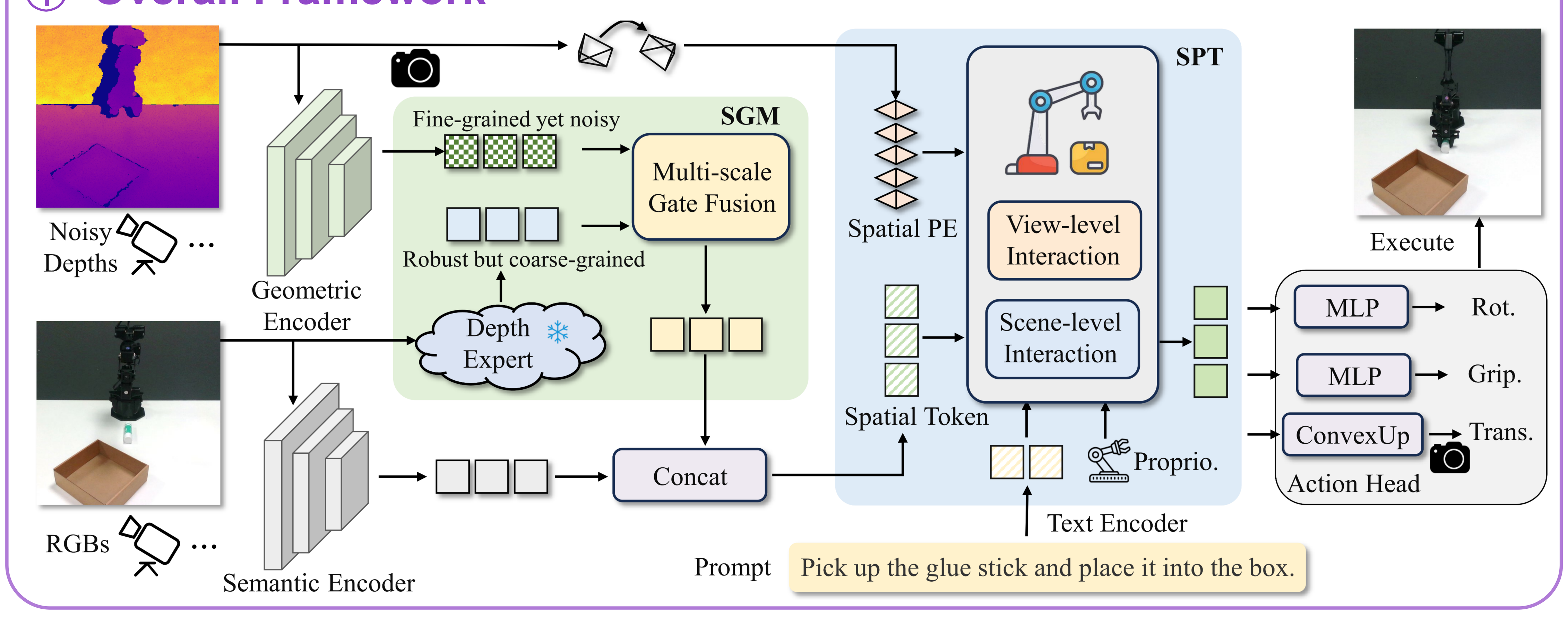
- ✓ **Point-based**
  Sparse fusion → Loss fine-grained semantics.
- ✓ **Image-based** (e.g., RVT)
  Coupled encode → Noisy depth interfere semantics.
- ✓ **Ours:** Disentangled framework
  ① Visual semantics
  ② Complementary high-level geometry
  - Fine-grained yet noisy: from real sensor depth
  - Robust but coarse-grained: from depth expert

  (High-SNR semantics + 2D-to-3D priors help smooth noisy depth)

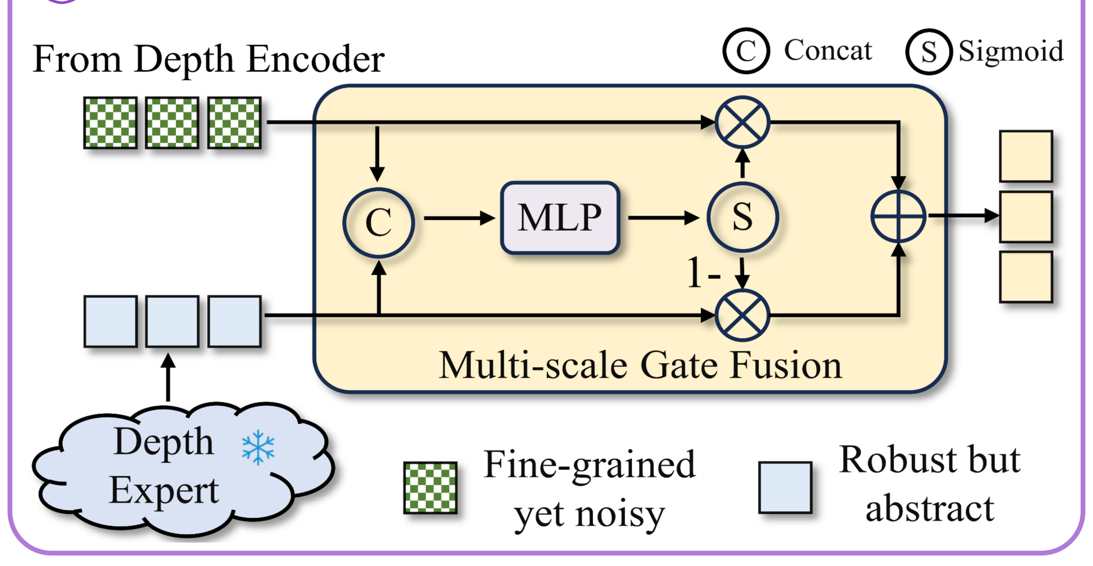
  ③ Low-level geometry: explicit 2D-3D correspondence

2 Method

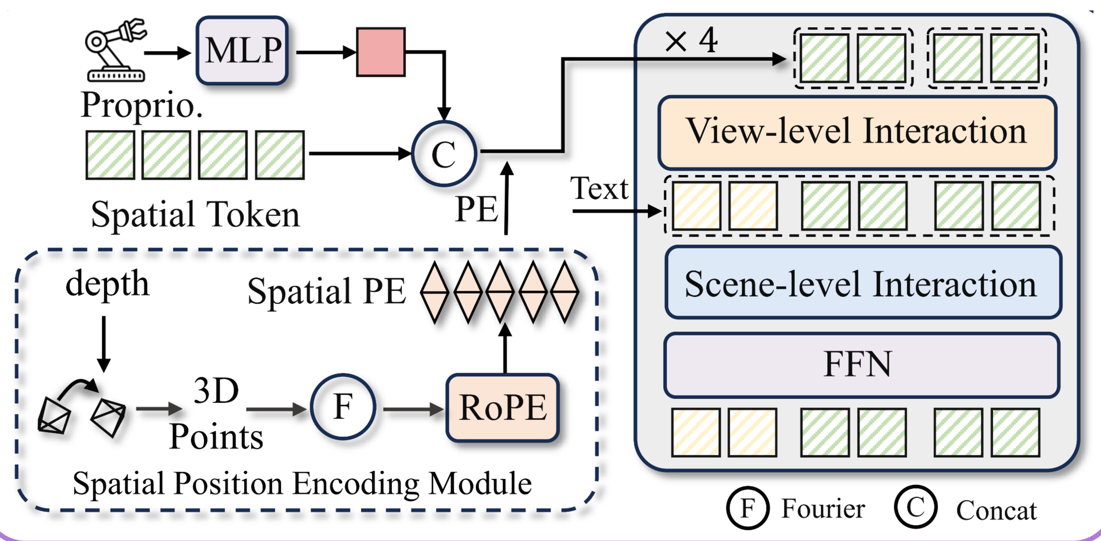
① Overall Framework



② Semantic-Guided Geometric Module



③ Spatial Transformer



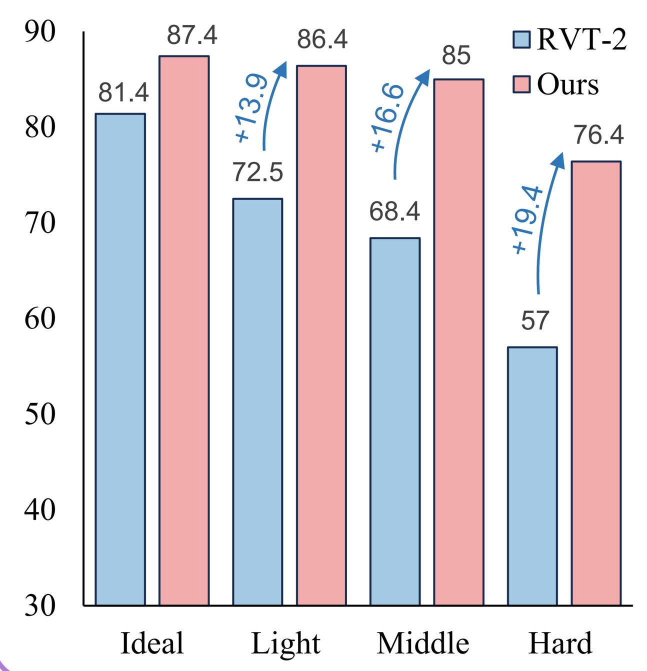
3 Experiments

① RLBench SOTA: 87.4% Success Rate

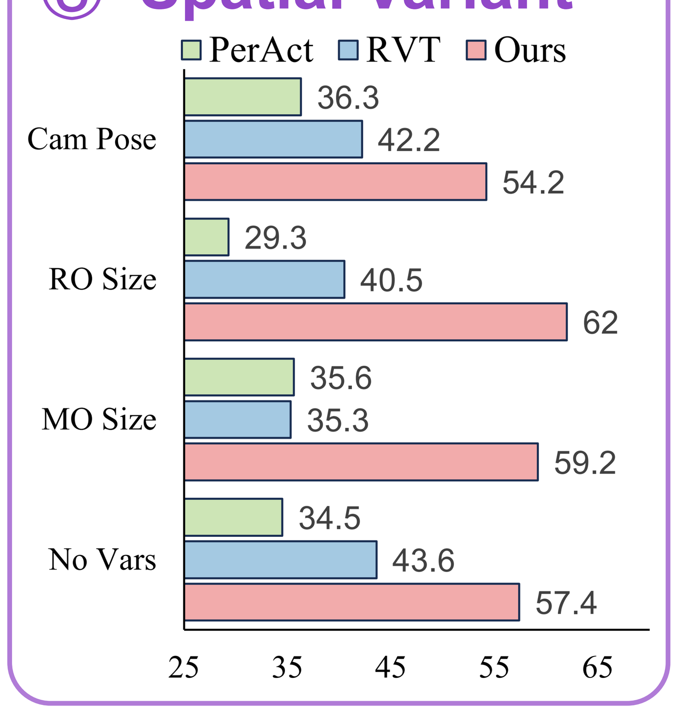
Models	Avg. Success ↑	Avg. Rank ↓	Close Jar	Drag Stick	Insert Peg	Meat off Grill	Open Drawer	Place Cups	Place Wine	Push Buttons
C2F-ARM-BC	20.1	9.5	24.0	24.0	4.0	20.0	20.0	0.0	8.0	72.0
HiveFormer	45.3	7.8	52.0	76.0	0.0	**100.0**	52.0	0.0	80.0	84.0
PolarNet	46.4	7.3	36.0	92.0	4.0	**100.0**	84.0	0.0	40.0	96.0
PerAct	49.4	7.1	$55.2_{\pm 4.7}$	$89.6_{\pm 4.1}$	$5.6_{\pm 4.1}$	$70.4_{\pm 2.0}$	$88.0_{\pm 5.7}$	$2.4_{\pm 3.2}$	$44.8_{\pm 7.8}$	$92.8_{\pm 3.0}$
RVT	62.9	5.3	$52.0_{\pm 2.5}$	$99.2_{\pm 1.6}$	$11.2_{\pm 3.0}$	$88.0_{\pm 2.5}$	$71.2_{\pm 6.9}$	$4.0_{\pm 2.5}$	$91.0_{\pm 5.2}$	$\mathbf{100.0}_{\pm 0.0}$
Act3D	65.0	5.3	92.0	92.0	27.0	94.0	93.0	3.0	80.0	99.0
SAM-E	70.6	2.9	$82.4_{\pm 3.6}$	$\mathbf{100.0}_{\pm 0.0}$	$18.4_{\pm 4.6}$	$95.2_{\pm 3.3}$	$\mathbf{95.2}_{\pm 5.2}$	$0.0_{\pm 0.0}$	$94.4_{\pm 4.6}$	$\mathbf{100.0}_{\pm 0.0}$
3D-Diff-Actor	81.3	2.8	$96.0_{\pm 2.5}$	$\mathbf{100.0}_{\pm 0.0}$	$65.6_{\pm 4.1}$	$96.8_{\pm 1.6}$	$89.6_{\pm 4.1}$	$24.0_{\pm 7.6}$	$93.6_{\pm 4.8}$	$98.4_{\pm 2.0}$
RVT-2	81.4	2.8	$\mathbf{100.0}_{\pm 0.0}$	$99.0_{\pm 1.7}$	$40.0_{\pm 0.0}$	$99.0_{\pm 1.7}$	$74.0_{\pm 11.8}$	$38.0_{\pm 4.5}$	$\mathbf{95.0}_{\pm 3.3}$	$\mathbf{100.0}_{\pm 0.0}$
SpatialActor	$\mathbf{87.4}_{\pm 0.8}$	**2.3**	$94.0_{\pm 4.2}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{93.3}_{\pm 4.8}$	$98.7_{\pm 2.1}$	$82.0_{\pm 3.3}$	$\mathbf{56.7}_{\pm 8.5}$	$94.7_{\pm 4.8}$	$\mathbf{100.0}_{\pm 0.0}$

Models	Put in Cupboard	Put in Drawer	Put in Safe	Screw Bulb	Slide Block	Sort Shape	Stack Blocks	Stack Cups	Sweep Dustpan	Turn Tap
C2F-ARM-BC	0.0	4.0	12.0	8.0	16.0	8.0	0.0	0.0	0.0	68.0
HiveFormer	32.0	68.0	76.0	8.0	64.0	8.0	8.0	0.0	28.0	80.0
PolarNet	12.0	32.0	84.0	44.0	56.0	12.0	4.0	8.0	52.0	80.0
PerAct	$28.0_{\pm 4.4}$	$51.2_{\pm 4.7}$	$84.0_{\pm 3.6}$	$17.6_{\pm 2.0}$	$74.0_{\pm 13.0}$	$16.8_{\pm 4.7}$	$26.4_{\pm 3.2}$	$2.4_{\pm 2.0}$	$52.0_{\pm 0.0}$	$88.0_{\pm 4.4}$
RVT	$49.6_{\pm 3.2}$	$88.0_{\pm 5.7}$	$91.2_{\pm 3.0}$	$48.0_{\pm 5.7}$	$81.6_{\pm 5.4}$	$36.0_{\pm 2.5}$	$28.8_{\pm 3.9}$	$26.4_{\pm 8.2}$	$72.0_{\pm 0.0}$	$93.6_{\pm 4.1}$
Act3D	51.0	90.0	95.0	47.0	93.0	8.0	12.0	9.0	92.0	94.0
SAM-E	$64.0_{\pm 2.8}$	$92.0_{\pm 5.7}$	$95.2_{\pm 3.3}$	$78.4_{\pm 3.6}$	$95.2_{\pm 1.8}$	$34.4_{\pm 6.1}$	$26.4_{\pm 4.6}$	$0.0_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$
3D-Diff-Actor	$\mathbf{85.6}_{\pm 4.1}$	$96.0_{\pm 3.6}$	$\mathbf{97.6}_{\pm 2.0}$	$82.4_{\pm 2.0}$	$\mathbf{97.6}_{\pm 3.2}$	$44.0_{\pm 4.4}$	$68.3_{\pm 3.3}$	$47.2_{\pm 8.5}$	$84.0_{\pm 4.4}$	$99.2_{\pm 1.6}$
RVT-2	$66.0_{\pm 4.5}$	$96.0_{\pm 0.0}$	$96.0_{\pm 2.8}$	$88.0_{\pm 4.9}$	$92.0_{\pm 2.8}$	$35.0_{\pm 7.1}$	$\mathbf{80.0}_{\pm 2.8}$	$69.0_{\pm 5.9}$	$\mathbf{100.0}_{\pm 0.0}$	$99.0_{\pm 1.7}$
SpatialActor	$72.0_{\pm 3.6}$	$\mathbf{98.7}_{\pm 3.3}$	$96.7_{\pm 3.9}$	$\mathbf{88.7}_{\pm 3.9}$	$91.3_{\pm 6.9}$	$\mathbf{73.3}_{\pm 6.5}$	$56_{\pm 7.6}$	$\mathbf{81.3}_{\pm 4.1}$	$\mathbf{100.0}_{\pm 0.0}$	$95.3_{\pm 3.0}$

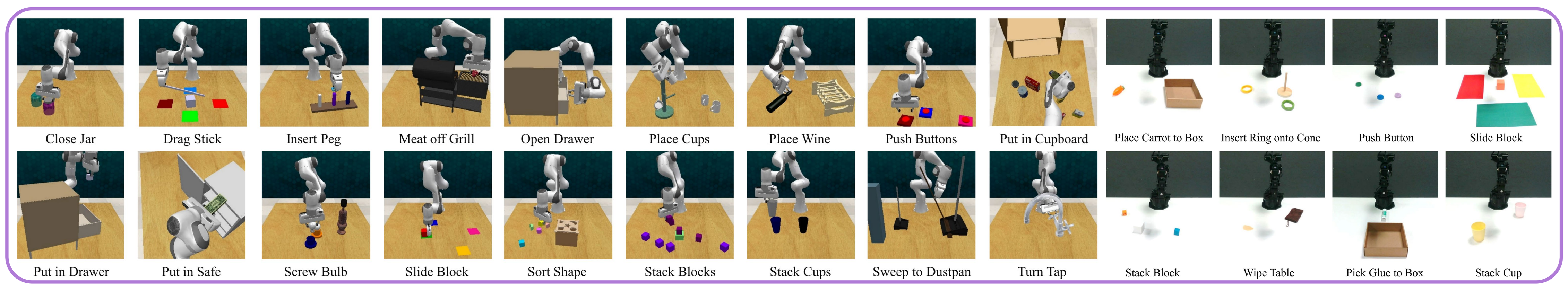
② Robust to Noise



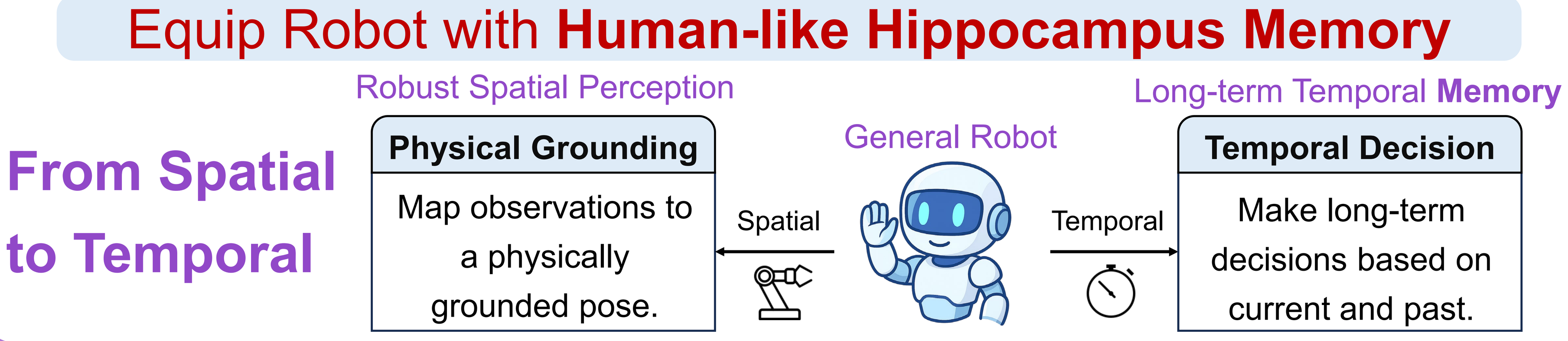
③ Spatial Variant



4 Visualization & Further Work



Equip Robot with **Human-like Hippocampus Memory**



MemoryVLA

④ Real-World Performance