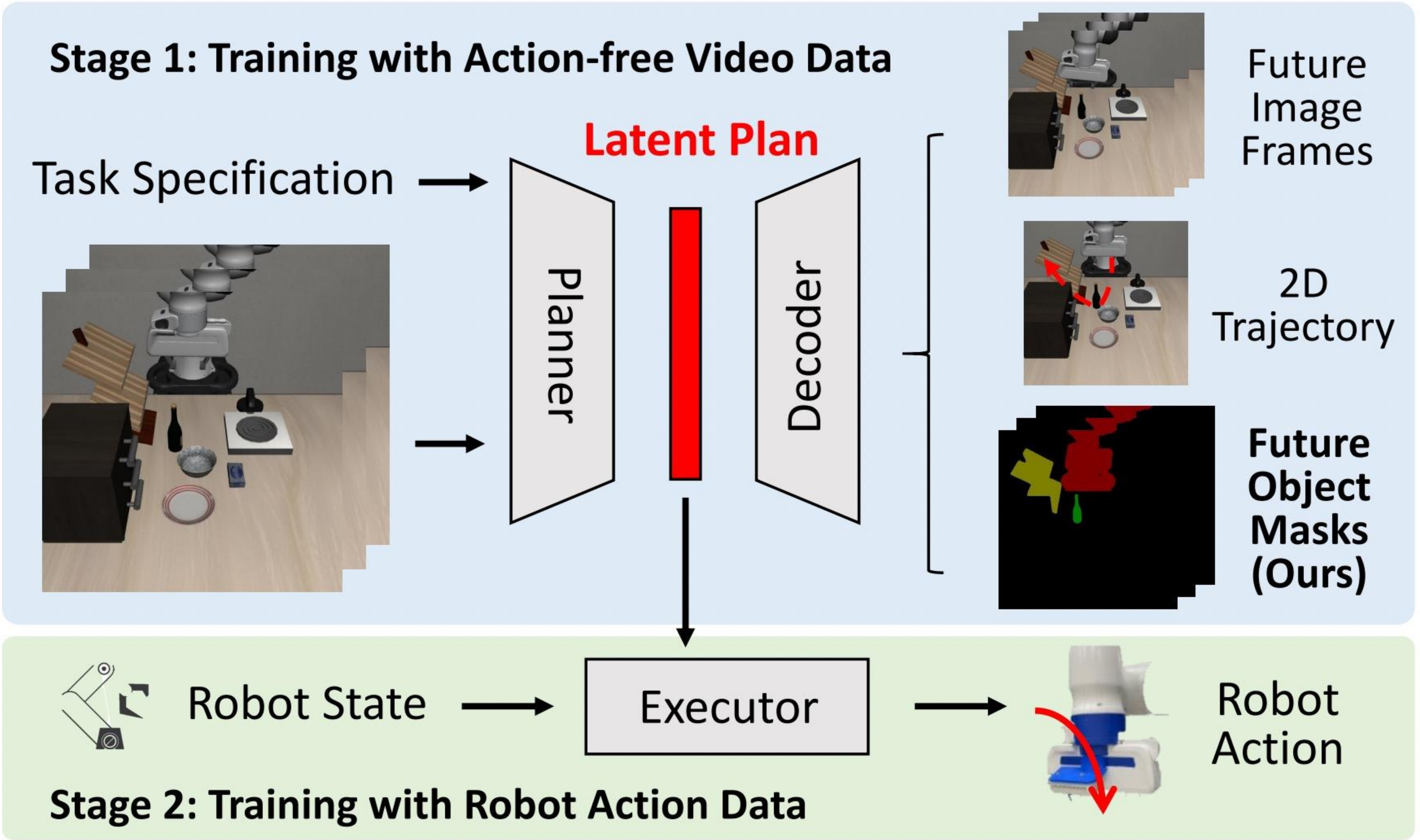
Mask2Act: Predictive Multi-Object Tracking as Video Pre-Training for Robot Manipulation

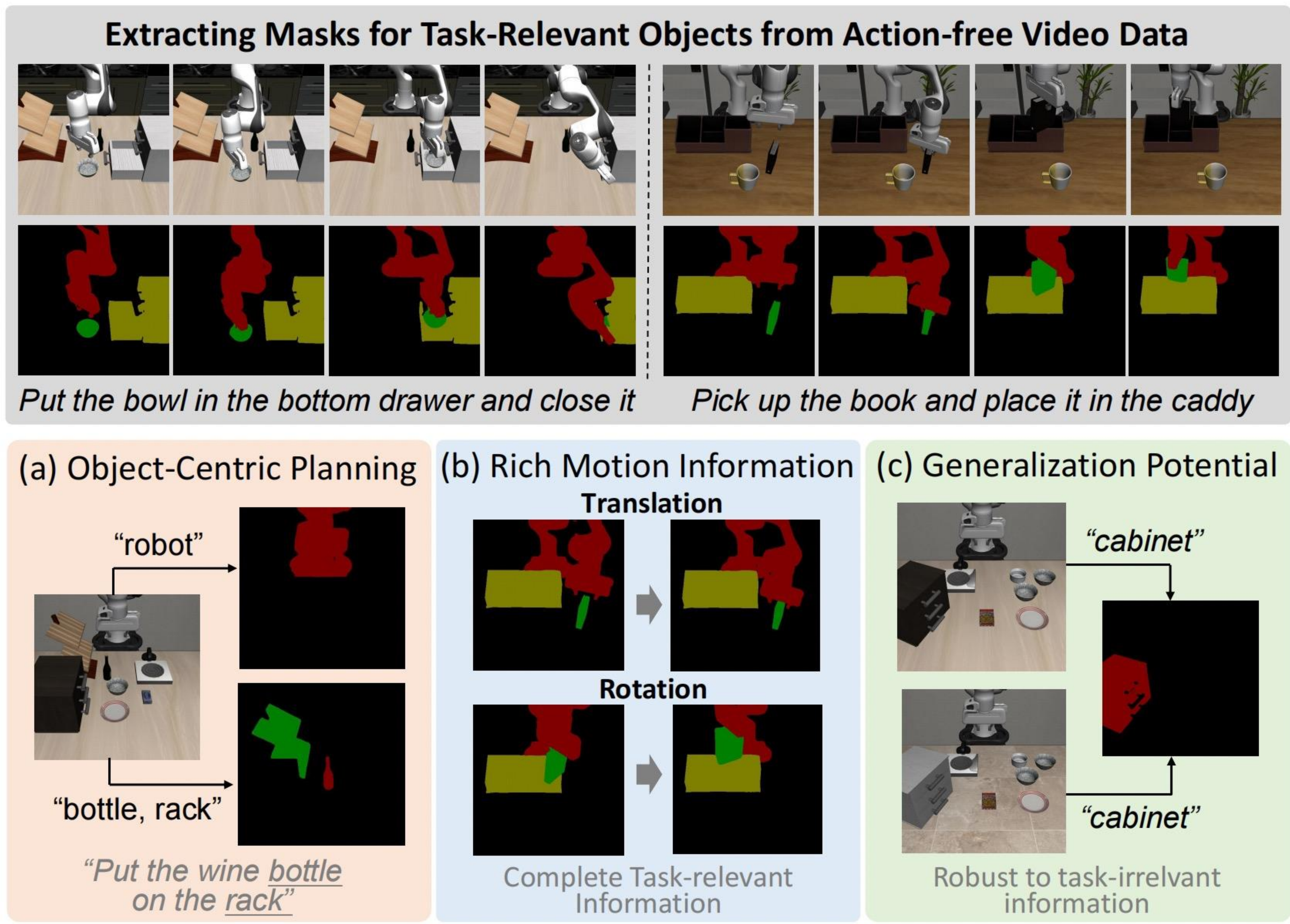
1 2

Junbo Zhang[1,2], Kaisheng Ma[1,2*]

BMVC 2025

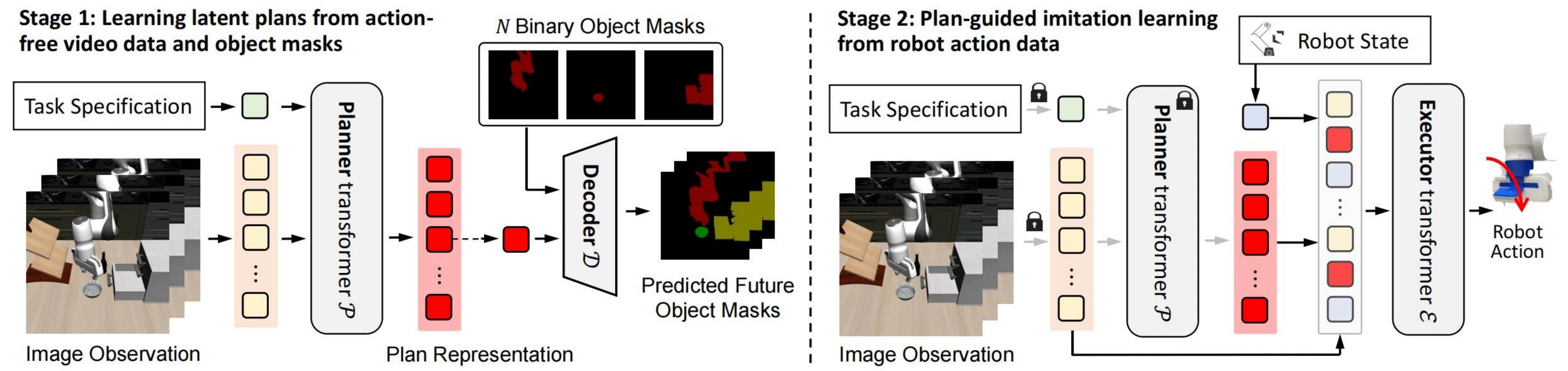
Motivation:





- ➢Leveraging action-free videos for robotic pre-training.
- ➢Planner-executor structure for two-stage training.
- ➢Novel intermediate representation between image frames and robot's motion: **object mask**.
- ➢Extracting task-relevant object masks from video with Grounded-DINO, SAM, XMem.
- ➢Three benefits of using object masks.

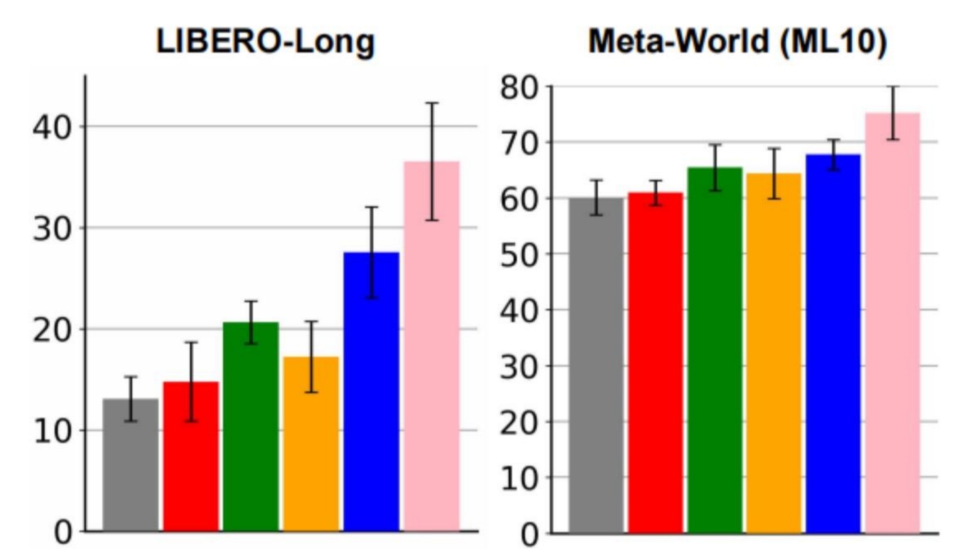
Method: Learning Object-Centric Latent Plans By Multi-Object Tracking



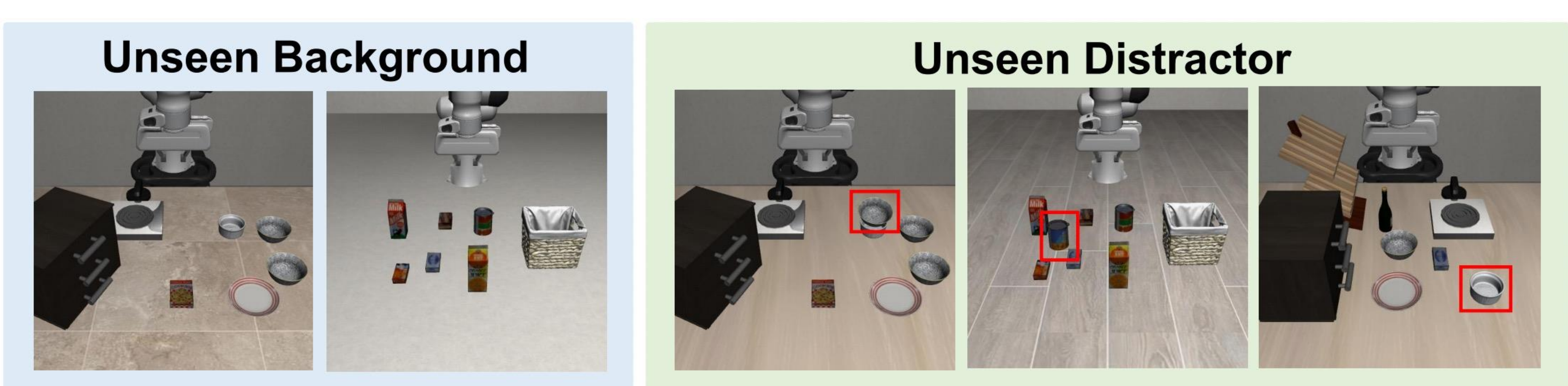
Experiments:

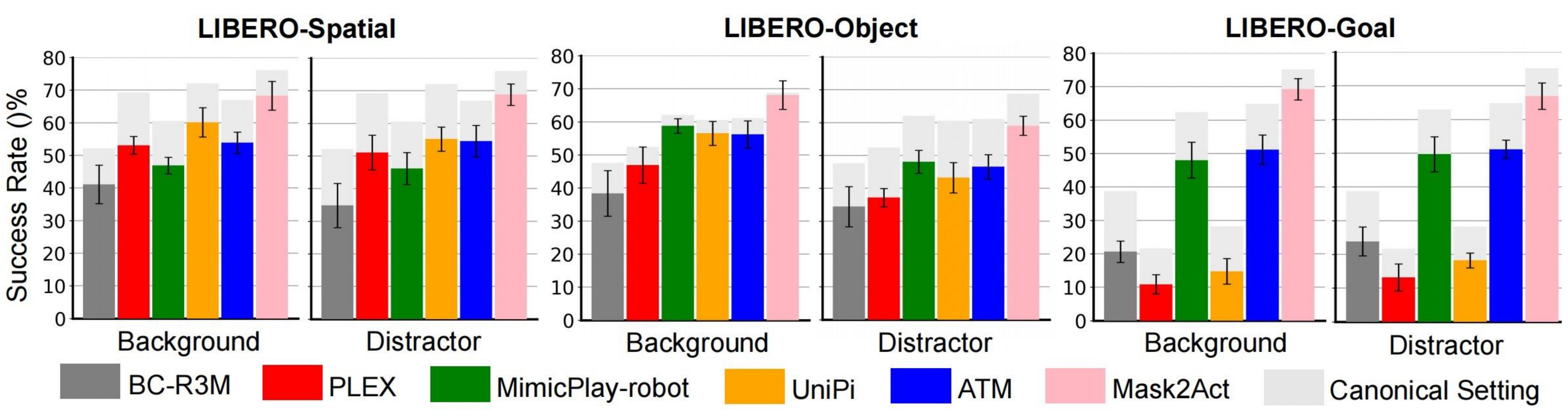
- Video Pre-training and Fine-tuning Results on LIBERO and Meta-World

Method	Latent Plan	LIBERO Spatial	LIBERO Object	LIBERO Goal	LIBERO Long	Meta-World MT10
BC-R3M	-	52.16 ± 7.9	47.56 ± 5.1	38.68 ± 3.3	13.08 ± 2.2	63.44 ± 3.3
PLEX	Image	69.24 ± 2.3	52.44 ± 6.8	21.68 ± 2.0	14.84 ± 3.1	67.92 ± 3.8
MimicPlay-robot	Trajectory	60.52 ± 2.6	62.04 ± 4.2	63.00 ± 5.1	17.88 ± 1.9	67.08 ± 3.1
UniPi	Image	72.04 ± 3.5	60.52 ± 2.7	28.20 ± 2.9	10.96 ± 1.8	70.12 ± 4.7
ATM	Trajectory	66.92 ± 1.6	61.12 ± 5.3	64.84 ± 4.4	24.36 ± 4.8	71.76 ± 3.9
Mask2Act (Ours)	Mask	**76.12** ± 2.3	**68.76** ± 5.7	**75.16** ± 3.8	**30.64** ± 6.1	**78.24** ± 5.1



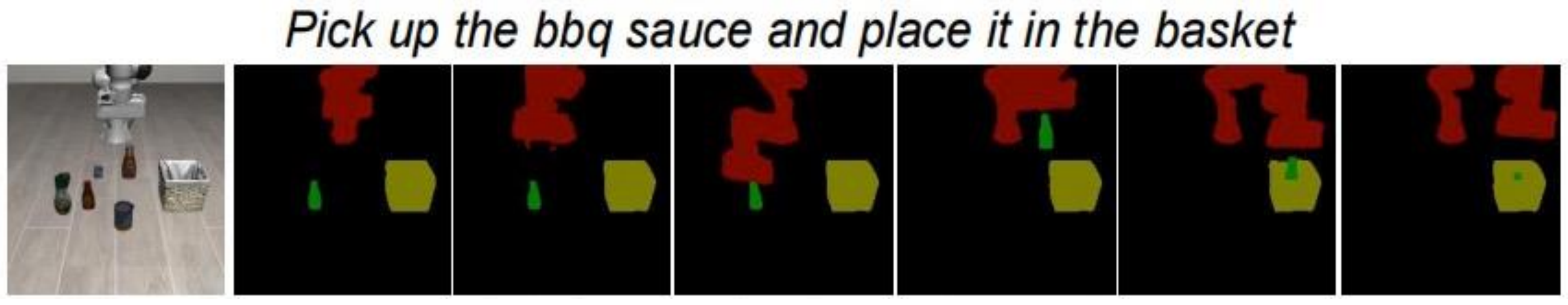
- Generalization Test: Unseen Background & Unseen Distractor





- Visualization Results:

Pick up the bbq sauce and place it in the basket



Put the wine bottle on the rack