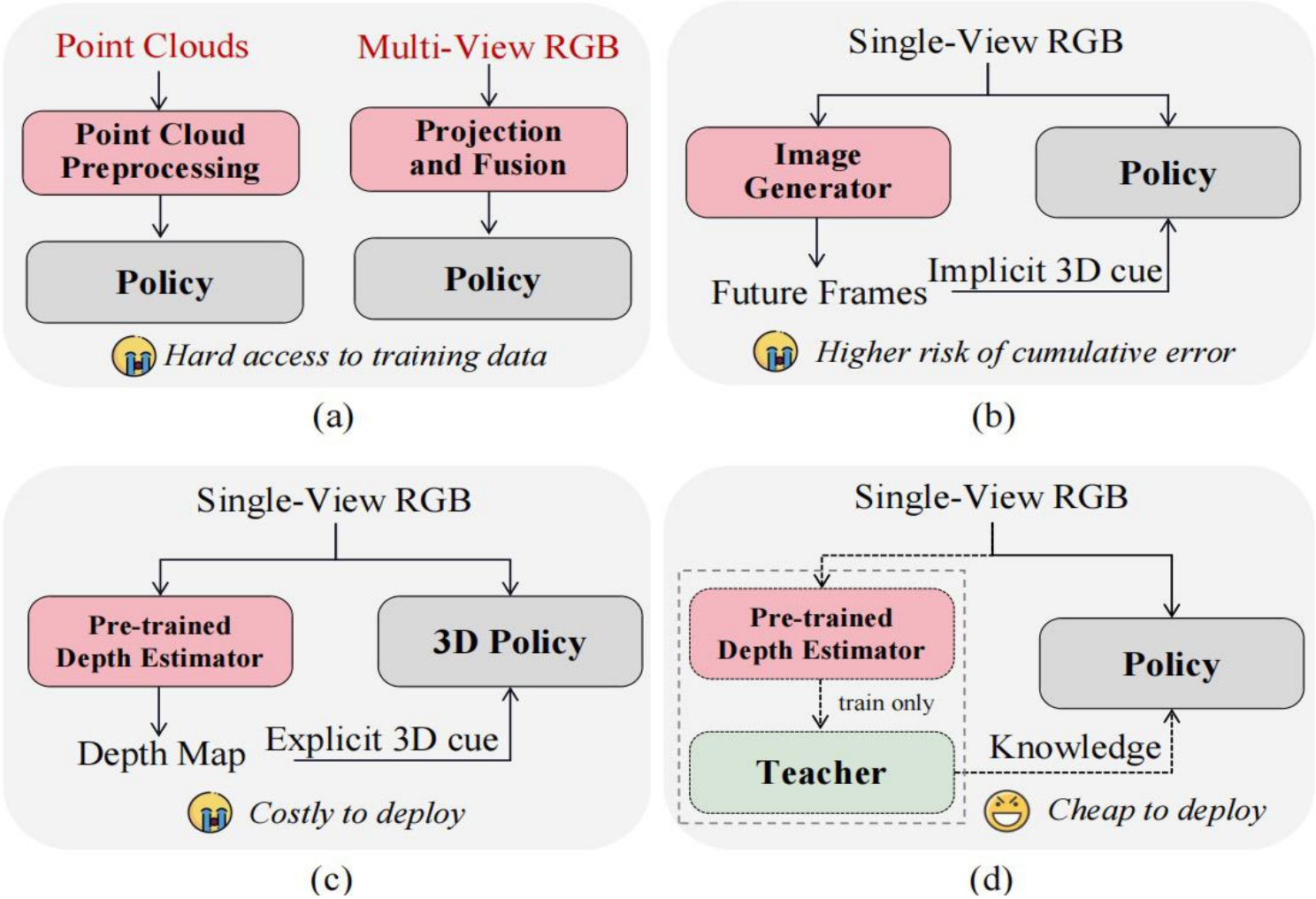
MonoLift: Learning 3D Manipulation Policies from Monocular RGB via Distillation

NeurIPS 2025

Ziru Wang, Mengmeng Wang ✉, Guang Dai, Yongliu Long, Jingdong Wang

Recent work learns 3D manipulation from monocular RGB, but 2D similarity often masks distinct 3D actions. Existing methods add structural cues or depth estimators, yet remain inefficient.



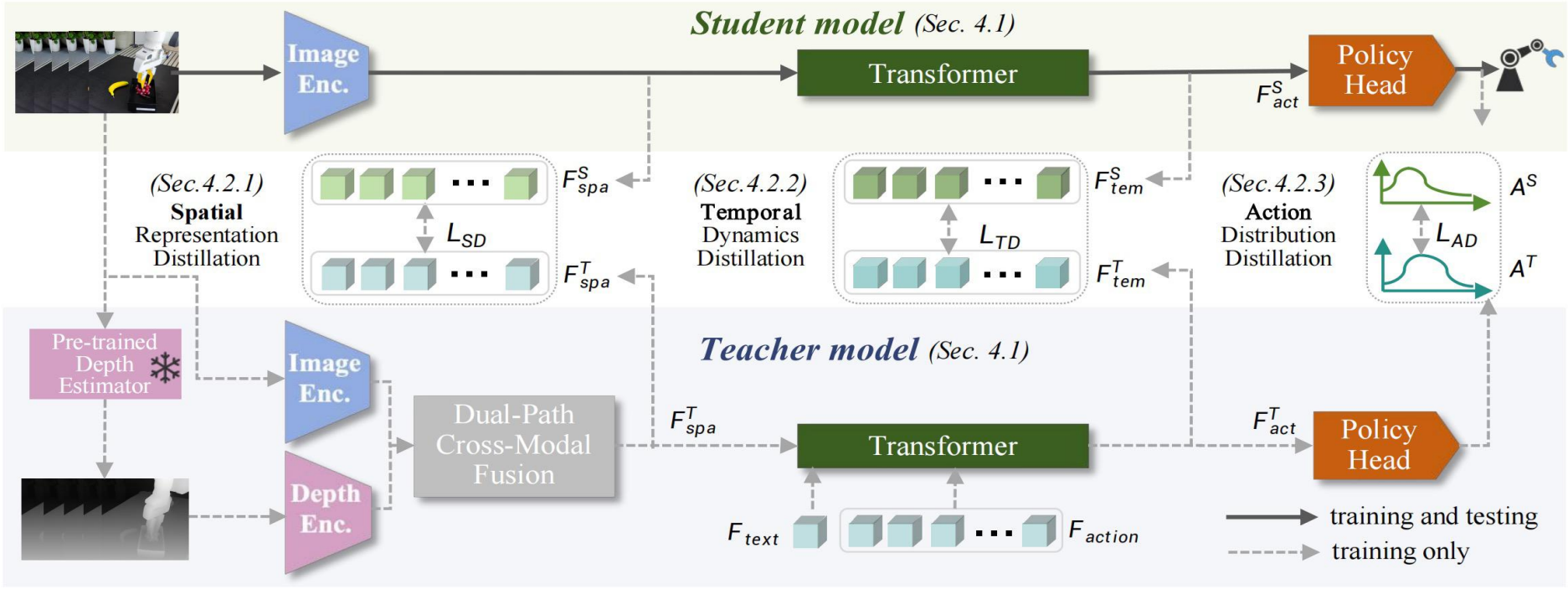
Comparison of frameworks for 3D policy learning.

Goal of the work:

Can we retain the benefits of depth-guided 3D reasoning without incurring inference-time cost?

We answer this with MonoLift, which lifts monocular RGB inputs into 3D-aware perception and control by distilling knowledge from a depth-guided teacher (see(d)).

Overview framework of MonoLift



Overview of the proposed MonoLift framework.

MonoLift consists of two main components:

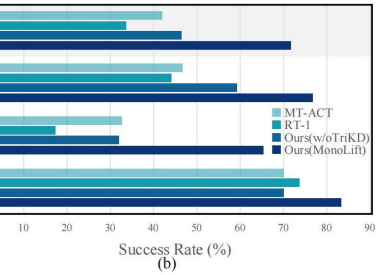
- (i) **Spatial Representation Distillation**: Transfers fused RGB–depth features from the teacher to help the student disambiguate visually similar yet structurally different observations.
- (ii) **Temporal Dynamics Distillation**: Aligns temporal feature trajectories to enable the student to capture motion patterns that reflect underlying 3D structural changes.
- (iii) **Action Distribution Distillation**: Transfers action distributions shaped by the teacher's 3D understanding, guiding the student to generate geometry-aware behaviors.

Performance comparison

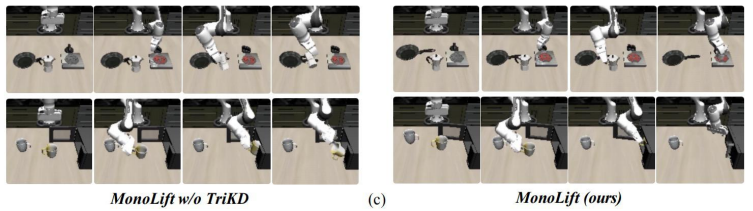
Table 1: **Quantitative results on 8 tasks from two Libero scenes.** Input types: **S-RGB** = Single-view RGB, **M-RGB** = Multi-view RGB, **RGB-D** = RGB+Depth. Task names are abbreviated: R/W = Red/White mug, Y/W = Yellow/White mug, C = Chocolate, L = Left, R = Right, P = Plate.

Method	Input Type	R-L	R-R	W-L	Y/W-R	C-L	C-R	R-P	W-P	Avg. S.R. (%)
MT-ACT	S-RGB	43.3	20.0	70.0	36.7	20.0	20.0	33.3	40.0	35.4±2.6
RT-1	S-RGB	33.3	36.7	73.3	70.0	30.0	40.0	46.7	50.0	47.5 ±2.7
MT-R3M	S-RGB	36.7	46.7	73.3	60.0	50.0	50.0	36.7	40.0	49.2 ±1.6
GROUND	S-RGB	38.4	40.8	51.2	38.4	70.4	79.2	72.8	25.6	52.5±8.8
3D-VLA	RGB-D	56.7	46.7	73.3	66.7	86.7	83.3	60.0	76.7	68.7±1.0
SPA	M-RGB	40.0	36.7	70.0	76.7	70.0	76.7	60.0	60.0	61.2±2.7
MonoLift	S-RGB	63.3	83.3	83.3	73.3	80.0	100.0	83.3	80.0	**80.8±3.3**

Method	Avg. S.R. (%)
MT-ACT	42.0±2.2
RT-1	33.7±3.0
w/o TriKD	46.5±1.3
MonoLift	**71.7±1.1**

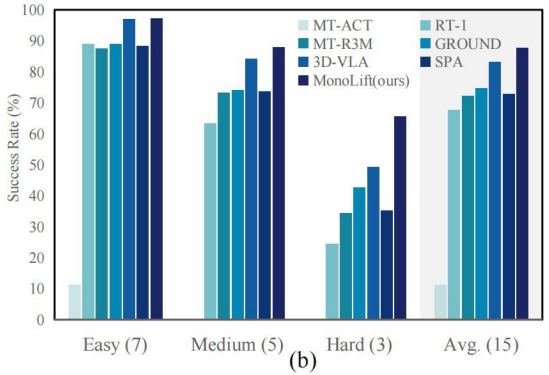


(a) (b)

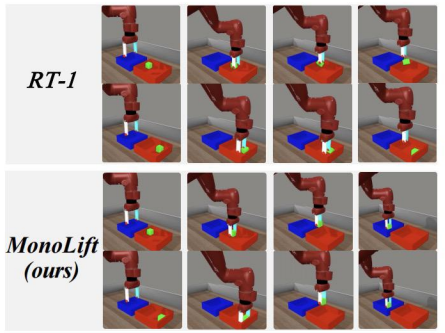


Method	Input Type	Avg. S.R. (%)
MT-ACT	S-RGB	11.3±0.5
RT-1	S-RGB	67.6±1.6
MT-R3M	S-RGB	72.2±2.7
GROUND	S-RGB	74.8±1.6
3D-VLA	RGB-D	83.2±2.4
SPA	M-RGB	72.8±2.1
MonoLift	S-RGB	**87.8±2.3**

(a)



(b)



(c)

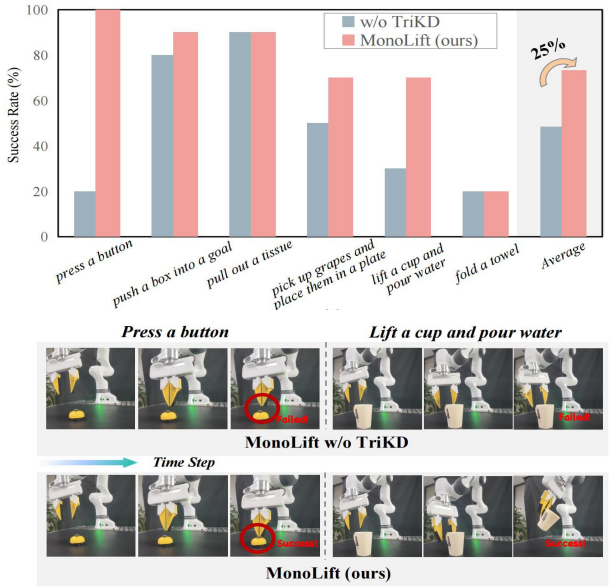


Figure 3: **Quantitative results and qualitative analysis on Meta-World.** Success rates including (a) overall average and (b) individual results for easy, medium, and hard tasks. (c) Visualized failure cases of the RT-1, and corresponding successful executions in `bin_picking` by MonoLift, both under monocular RGB input.

Conclusions

- (i) We propose MonoLift, a resource-efficient policy learning framework that learns from a 3D-aware teacher built on a pre-trained depth estimator, enabling monocular RGB agents to perform structured perception and control without extra 3D data or inference overhead.

- (ii) We design a tri-level knowledge distillation strategy that conveys spatial, temporal, and behavioral cues to improve the student's contextual understanding and decision-making under limited-modality constraints.

- (iii) We validate MonoLift across a wide range of simulated and real-world robotic manipulation tasks, demonstrating its ability to effectively learn 3D-aware policies while maintaining deployment efficiency.