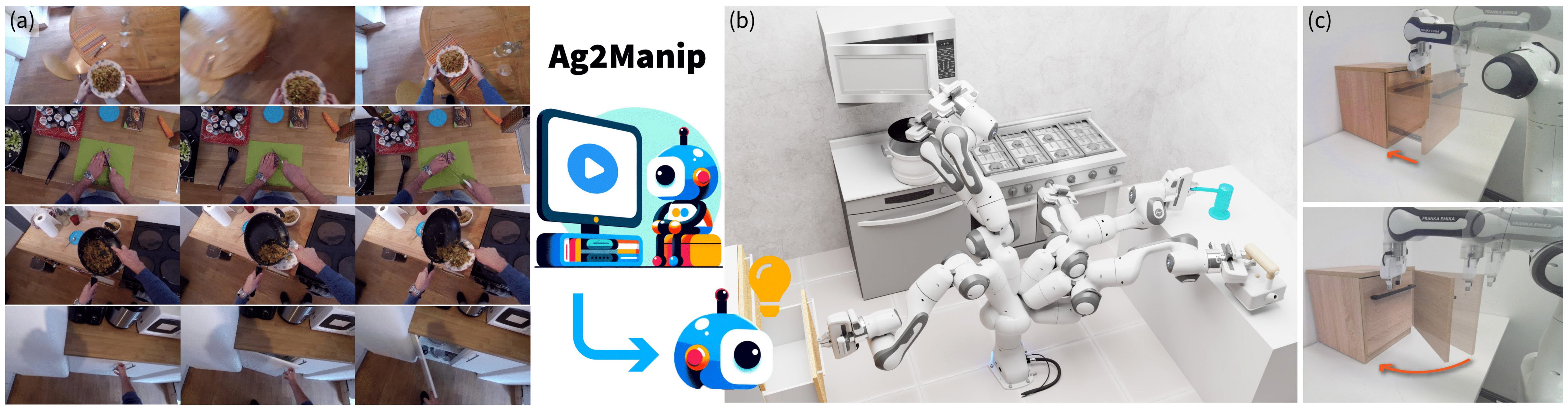
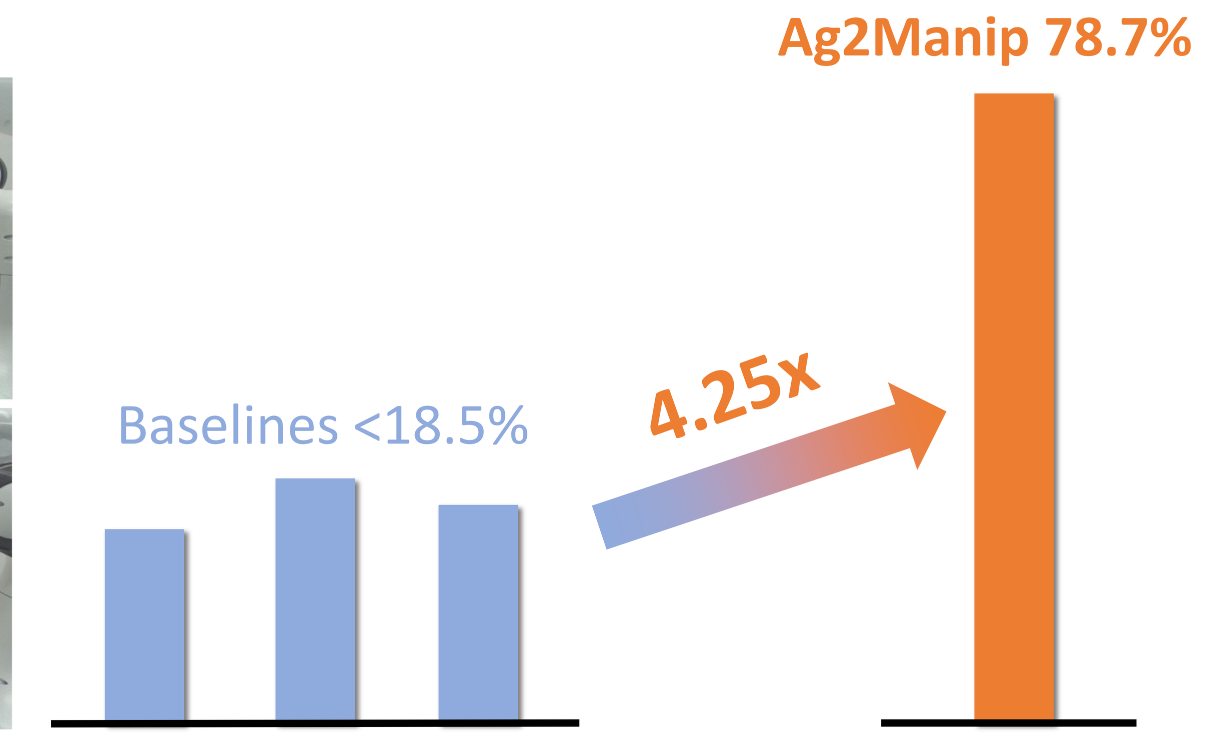
Ag2Manip: Learning Novel Manipulation Skills with Agent-Agnostic Visual and Action Representations

Puhao Li[1,2,★], Tengyu Liu[1,★], Yuyang Li[1,2,3], Muzhi Han[4], Haoran Geng[1,5], Shu Wang[4], Yixin Zhu, Song-Chun Zhu, Siyuan Huang[1,✉]

★Equal Contribution, ✉Corresponding Author.
[1]State Key Laboratory of General AI, BIGAI [2]Dept. of Automation, THU [3]Institute for AI, PKU [4]UCLA. [5]School of EECS, PKU





1. Autonomous Skill Acquisition

Autonomous skill acquisition is pivotal as they adapt to changing tasks and environments, moving robot from factory into our daily lives.

Task: Learning **novel** manipulation skills **without any expert input**.

Challenges: Visual and kinematics gaps prevent learning from human demos. Dexterous manipulation requires high precision in planning and execution.

Ag2Manip as the solution:

- Generalizable agent-agnostic visual and action representations for robotic manipulation.
- **78.7% success**, far surpassing baselines (18.5%) across 24 tasks.
- **Real-world experiments** validate our representation in few-shot IL.

2. Framework of Ag2Manip

(a) Learning agent-agnostic visual representation from Epic-Kitchen.

We first mask and inpaint humans in Epic-Kitchen videos to **eliminate the visual bias towards human**. Then, we learn an encoder $\mathcal{F}_\phi$ that maps RGB images to latent embeddings that **capture task progress** through time-contrastive loss:

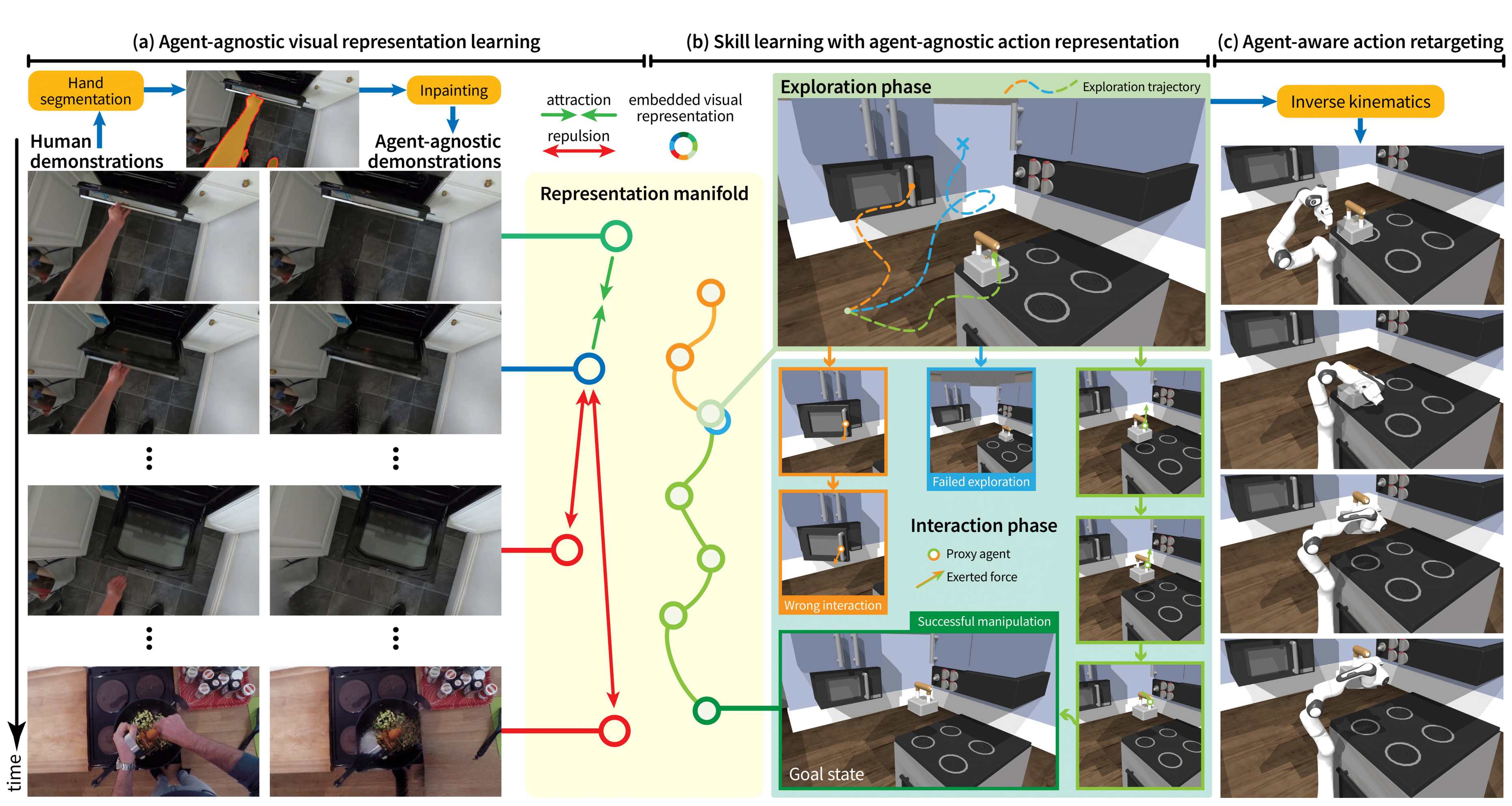
$$\mathcal{L} = \lambda_1 \mathbb{E}_{o_i^c, o_j^c, o_k^c, o_l^{\neq c} \sim \mathcal{D}^a} \mathcal{L}_{\text{tcn}} + \lambda_2 \mathbb{E}_{o \sim \mathcal{D}^a} \mathcal{L}_{\text{reg}}, \quad \mathcal{L}_{\text{tcn}} = -\log \frac{e^{\mathcal{S}(z_i^c, z_j^c)}}{e^{\mathcal{S}(z_i^c, z_j^c)} + e^{\mathcal{S}(z_i^c, z_k^c)} + e^{\mathcal{S}(z_i^c, z_l^{\neq c})}}, \quad \mathcal{L}_{\text{reg}} = \|\mathcal{F}_\phi(o)\|_1 + \|\mathcal{F}_\phi(o)\|_2.$$

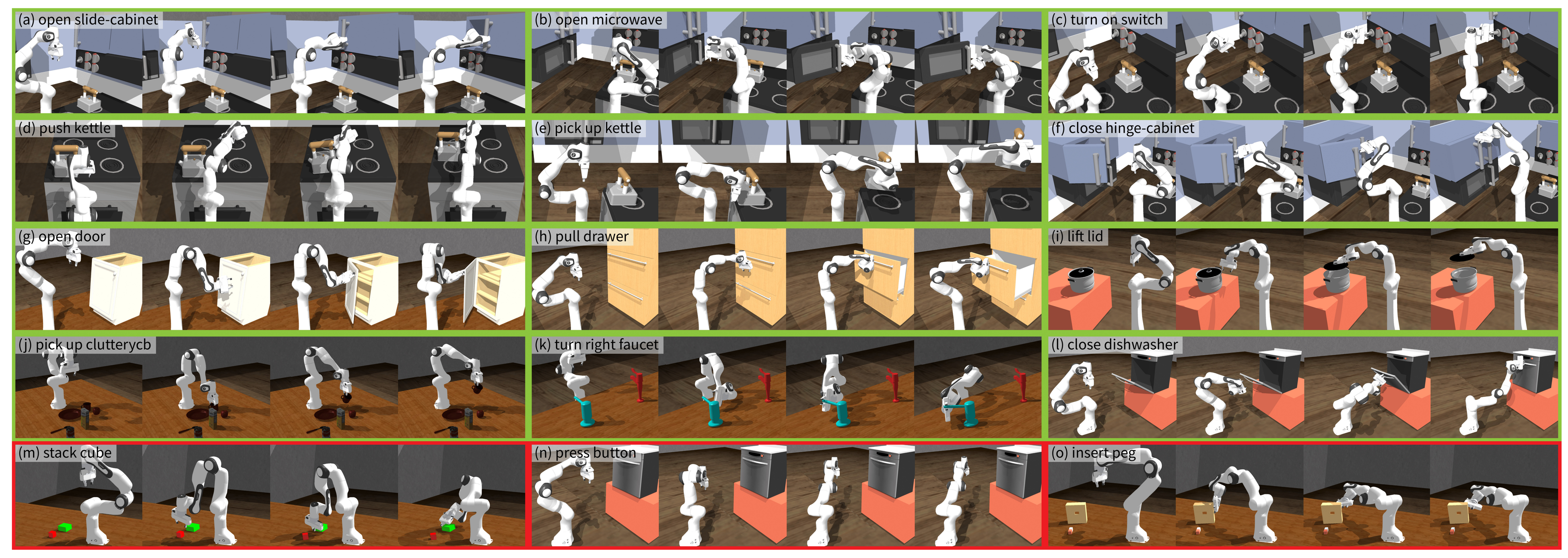
(b) Learning abstract skill with agent-agnostic action representation.

We abstract a robot's actions into a proxy's **motion and exerted force**. We then use PPO for policy optimization within this **agent-agnostic action space**.

$$\mathcal{R}(o_t, g; \phi) = \exp\left(\left(1 + \alpha \cdot \mathbf{1}_{\mathcal{S}(z_t, z_g) - \beta > 0}\right) \frac{\mathcal{S}(z_t, z_g) - \beta}{\beta}\right) - 1, \qquad \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \mathcal{R}(o_t, g; \phi)\right]$$

(c) Retargeting the Abstracted Skills to a Specific Robot with IK.





3. Extensive Evaluations

Performance: Ag2Manip has a 78.7% overall success rate (**3x baseline increase**), benefiting from its agent-agnostic representations → **See Tab. I**.

Task Progress Consistency: The proposed agent-agnostic visual representation demonstrates **greater consistency in capturing task progress** → **See Tab. II**.

Real-world Imitation: In **few-shot real-world imitation learning**, our visual representation shows superior efficiency over the baselines → **See Tab. III**.

Table I: Simulation studies on Ag2Manip

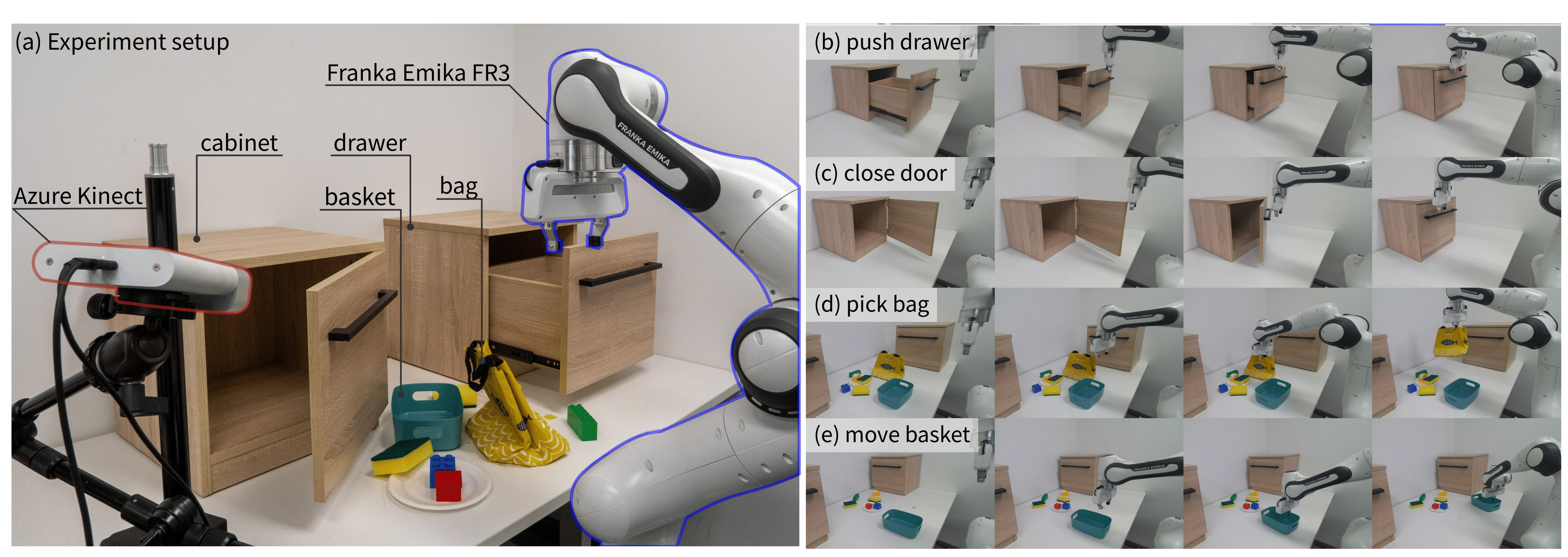
Method	FrankaKitchen											ManiSkill									PartManip							Overall
	a	b	c	d	e	f	g	h	i	j	Avg.	k	l	m	n	o	p	q	r	Avg.	s	t	u	v	w	x	Avg.	
R3M [8]	0	0	0	3	2	0	1	0	0	0	6.7%	0	6	0	0	0	0	0	0	8.3%	0	0	3	9	0	0	22.2%	11.1%
VIP [9]	0	0	0	2	6	0	3	0	0	0	12.2%	0	6	0	0	0	0	0	0	8.3%	0	0	0	9	0	0	16.7%	12.0%
Eureka [10]	0	0	0	7	3	2	3	0	0	0	16.7%	0	9	0	0	0	0	0	1	13.9%	0	0	3	6	0	0	20.0%	18.5%
Ours w/o Act.Repr.	4	1	8	9	9	9	9	1	7	2	65.6%	0	9	0	0	0	0	1	8	25.0%	0	0	8	9	0	0	31.5%	43.5%
Ours w/o Rew.Shp.	8	7	7	9	9	9	7	9	1	0	73.3%	9	9	8	0	3	1	4	5	54.2%	9	6	8	9	0	9	75.9%	67.6%
Ours	7	8	8	8	8	9	8	6	9	9	**88.9%**	7	9	6	0	7	2	8	8	**65.3%**	9	7	9	9	0	9	**79.6%**	**78.7%**
Ours (Proxy)	8	9	9	8	9	9	9	9	9	9	97.8%	7	9	5	5	7	3	8	9	73.6%	9	9	9	9	0	8	81.5%	85.7%

Table II: Task progress consistency of visual representations

Method	FrankaKitchen	ManiSkill	PartManip	Overall
ResNet50 [51]	$0.535^{\pm.169}$	$0.407^{\pm.182}$	$0.202^{\pm.197}$	$0.418^{\pm.199}$
CLIP [53]	$0.627^{\pm.086}$	$0.381^{\pm.139}$	$0.347^{\pm.151}$	$0.490^{\pm.134}$
R3M [8]	$0.498^{\pm.190}$	$0.393^{\pm.191}$	$0.525^{\pm.123}$	$0.474^{\pm.177}$
VIP [9]	$0.496^{\pm.246}$	$0.251^{\pm.178}$	$0.386^{\pm.121}$	$0.401^{\pm.208}$
Ag2Manip	$\mathbf{0.828^{\pm.082}}$	$\mathbf{0.696^{\pm.182}}$	$\mathbf{0.618^{\pm.227}}$	$\mathbf{0.740^{\pm.153}}$

Table III: Experimental Results

Method	PushDrawer	CloseCabinet	PickBag	MoveBasket
ResNet50 [51]	1/10	5/10	1/10	1/10
CLIP [53]	2/10	3/10	0/10	0/10
R3M [8]	4/10	5/10	4/10	3/10
VIP [9]	6/10	6/10	2/10	6/10
Ag2Manip	**7/10**	**8/10**	**8/10**	**8/10**



Conclusion

- Ag2Manip can acquire various robot manipulation skills **without expert demonstrations**.
- Ag2Manip leverages **innovative agent-agnostic visual and action representations** to bridge domain gaps and address precision challenges in robotic manipulation learning.
- Extensive simulated and real-world experiments **show its effectiveness in autonomous skill acquisition**.

Paper

Page

Code