June 11 – 15, 2025 | Nashville TN, USA

Embodied AI Workshop @ CVPR 2025

ARNOLD CHALLENGE

1st Place

DynScene: Scalable Generation of Dynamic Robotic Manipulation Scenes for Embodied AI



Sangmin Lee, Sungyong Park, Heewon Kim

Soongsil University

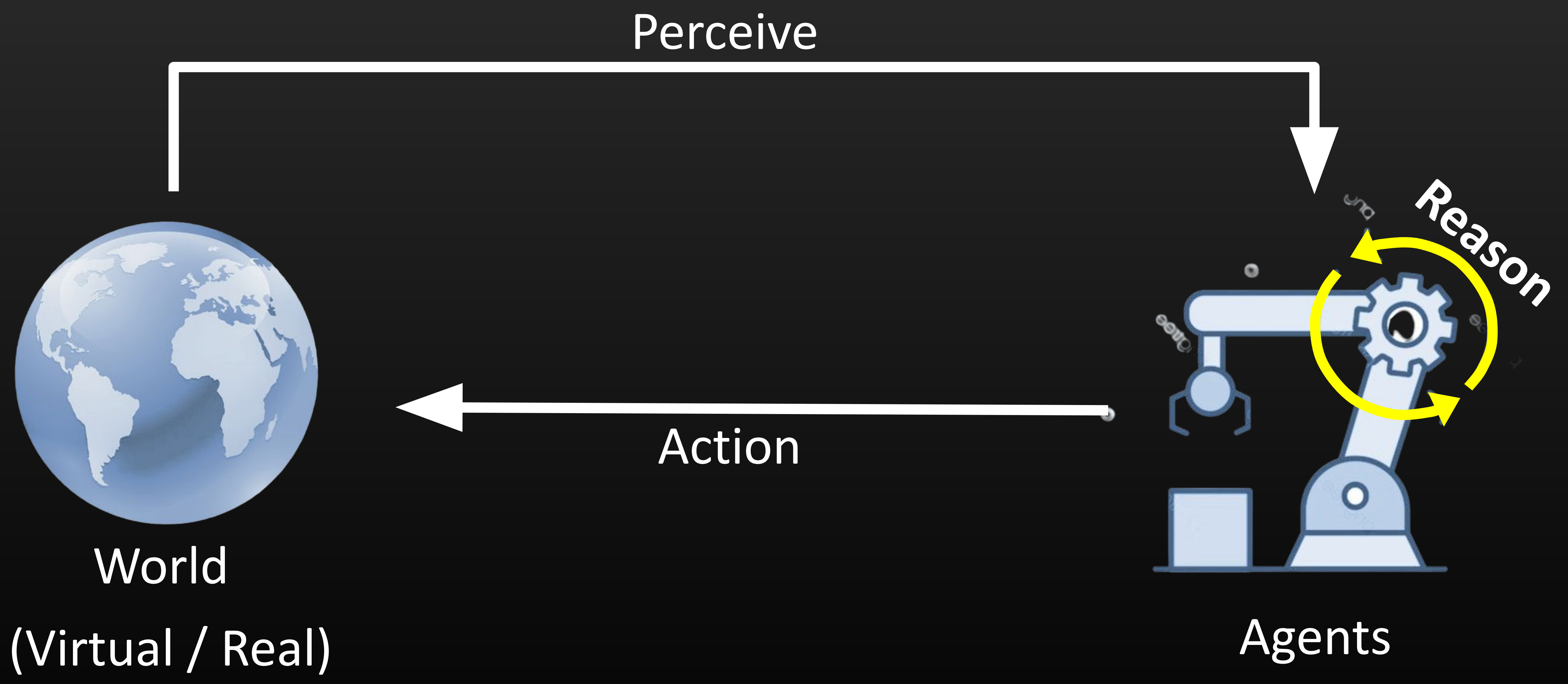
{sm32289, ejqdl010}@gmail.com, hwkim@ssu.ac.kr

Reality Lab
Soongsil University

https://reality.ssu.ac.kr/

Introduction

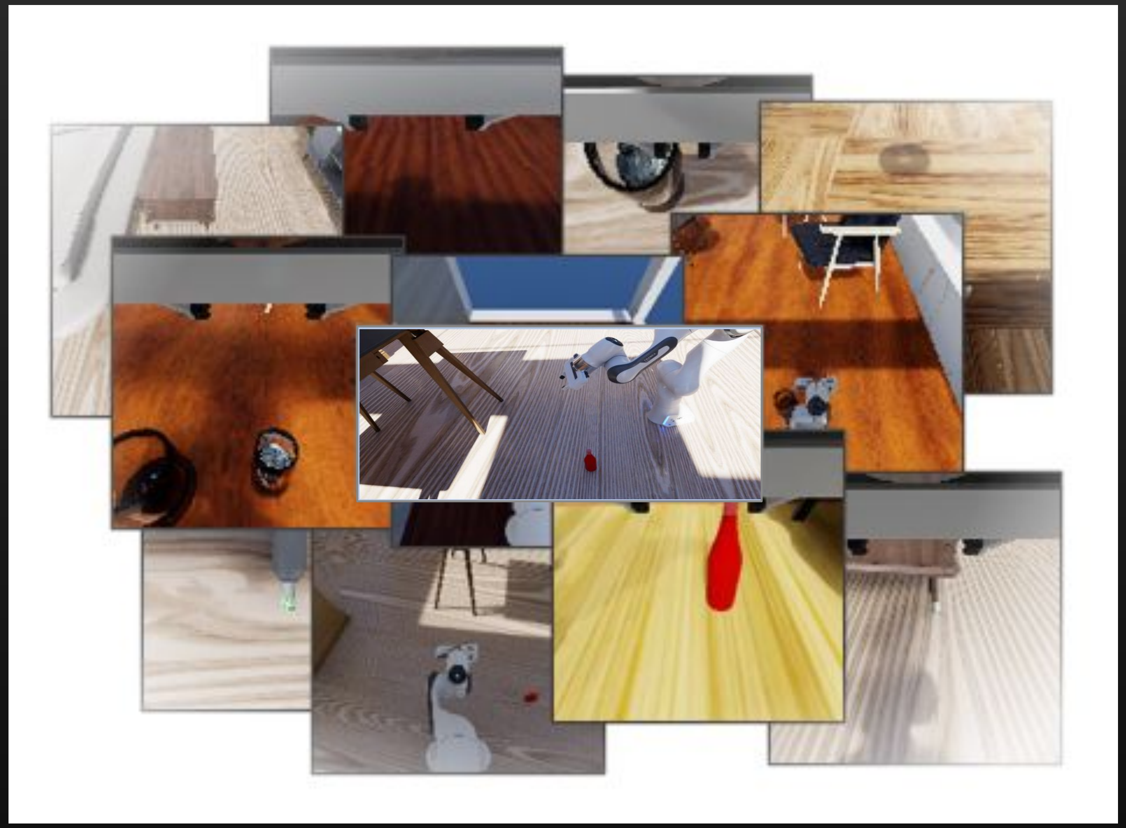
- Embodied AI aims to train agents that can **perceive**, **reason**, and **act** within **physically grounded environments**, ultimately enabling robots to perform complex tasks in the real world.



Robotic Manipulation in Embodied AI

- Goal: Perceive and manipulate objects to complete designated tasks.
- Challenge: Data scaling is one of the key issues in robotic manipulation.


Human-teleoperated data collection



Simulation datasets

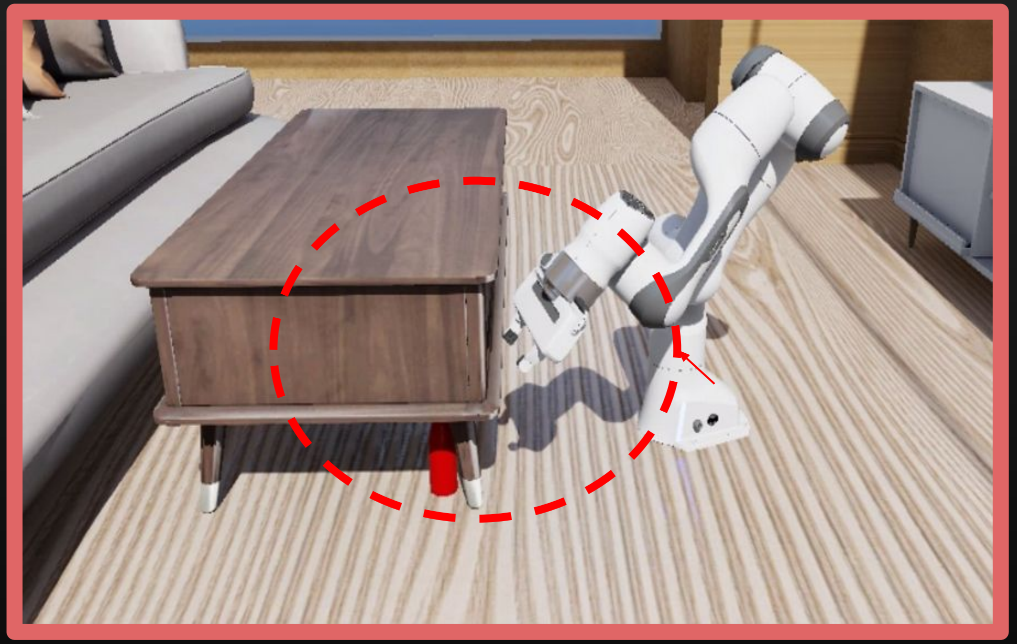
➔ *This approach remains costly, slow, and labor-intensive to scale.*

Comparison of Data Generation Methods

- Previous research has developed methods for generating **either static scenes or robot actions**, advancing embodied AI data creation.

Static Scene Generation

“The room has a sofa, table, red bottle, and Franka robot”



Inadequate object position for task execution

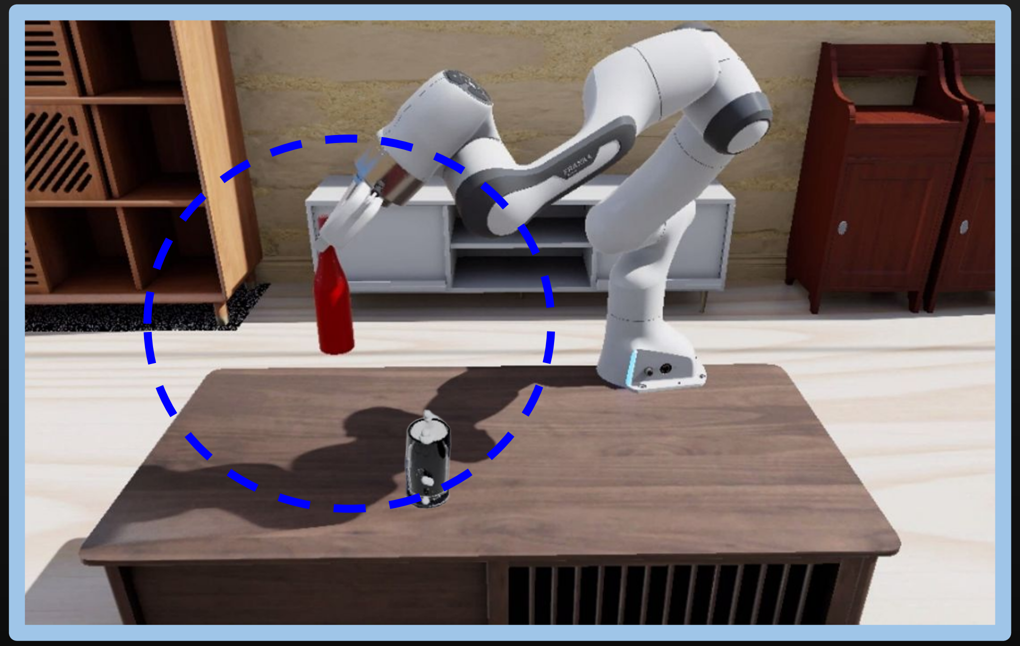
Action Generation

“Pick up the bottle”

Limited data diversity for interaction

Dynamic Scene Generation

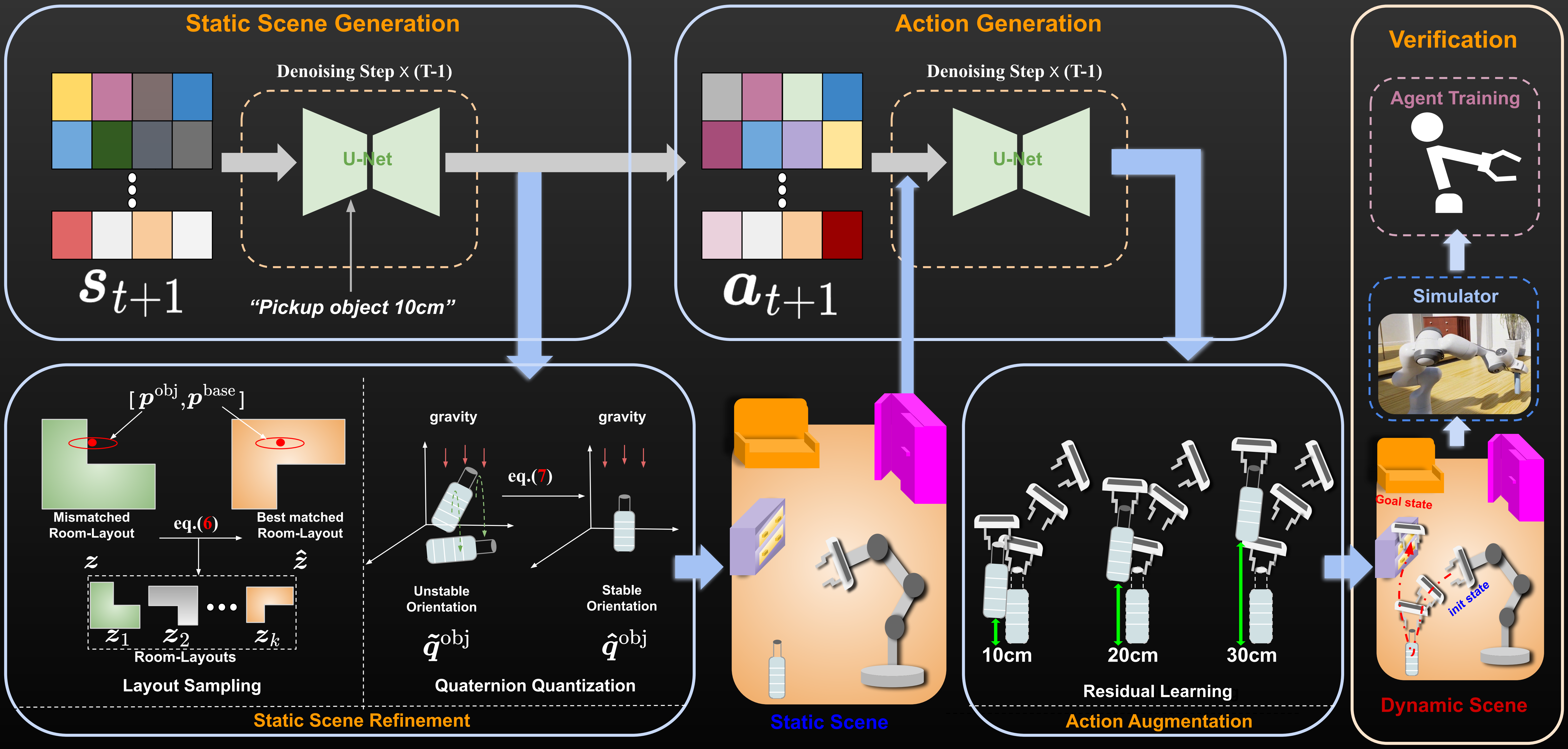
“Pick up the bottle ten centimeter”



Successful interaction and diverse data

Proposed Method : DynScene

- Framework Overview

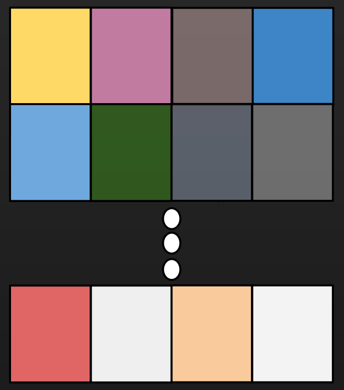


Data Representation for Dynamic Scene

Reality Lab
Soongsil University

- A dynamic scene pairs static scene $\boldsymbol{s}$ with residual action $\boldsymbol{a}$ for scalable augmentation.
- Enables diverse and coherent environment-behavior combinations.

Static Scene



$\boldsymbol{s}_{t+1}$

$\rightarrow [\boldsymbol{o}, \boldsymbol{r}, \boldsymbol{z}]$

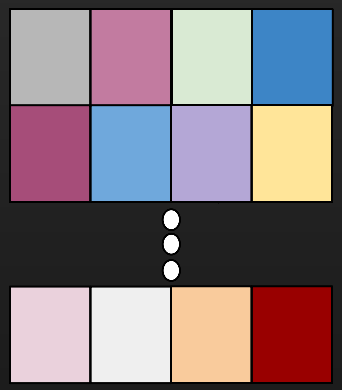
$\boldsymbol{o}$: Target object

$\boldsymbol{r}$: Robot base

$\boldsymbol{z}$: Room layouts

Absolute Coordinates

Robot Action



$\boldsymbol{a}_{t+1}$

$\rightarrow [\Delta \boldsymbol{p}_k^{\mathrm{ee}}, \Delta \boldsymbol{q}_k^{\mathrm{ee}}, \boldsymbol{g}_k]$

Residual position

Gripper state

Residual quaternion

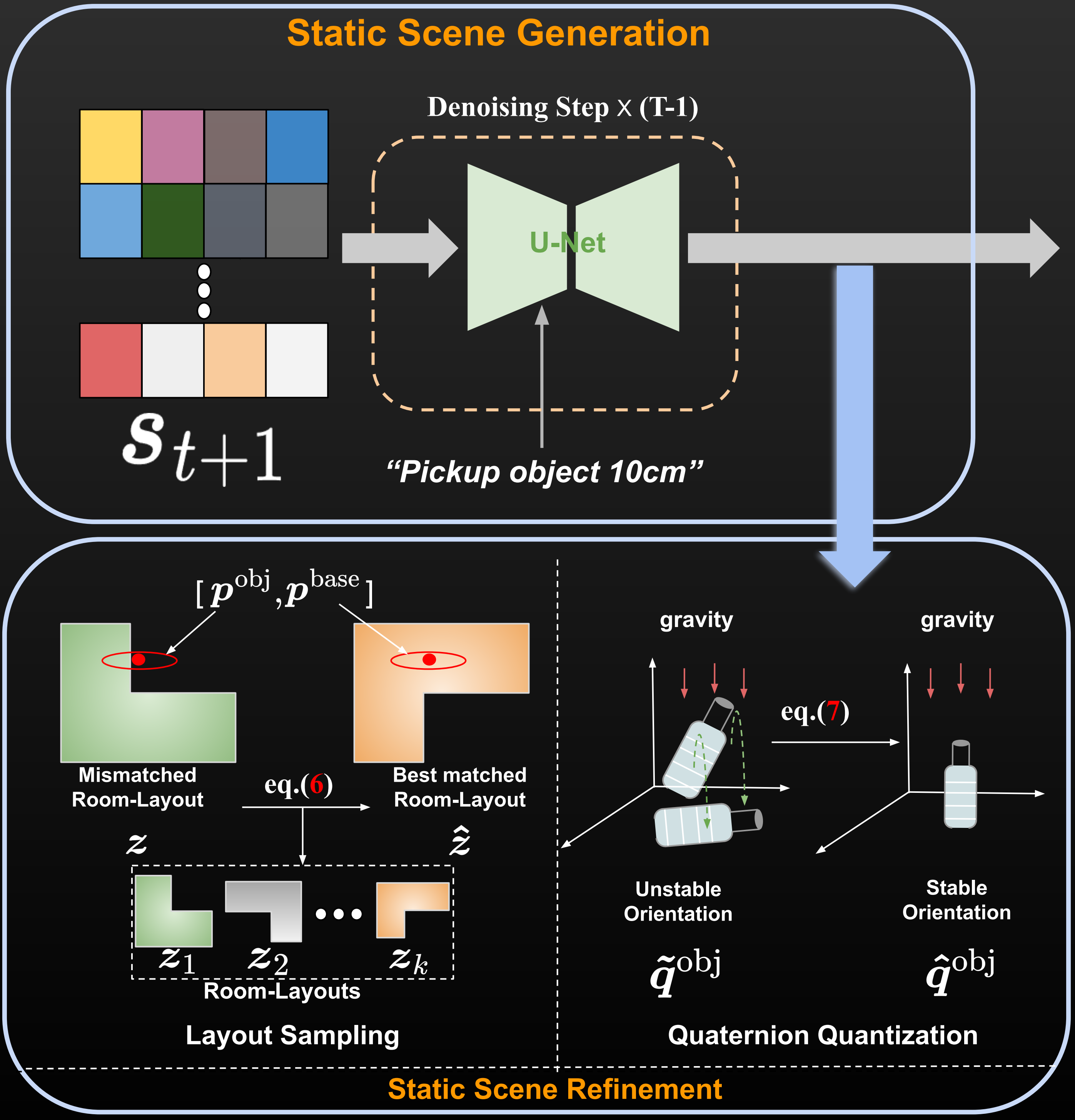
Residual Coordinates

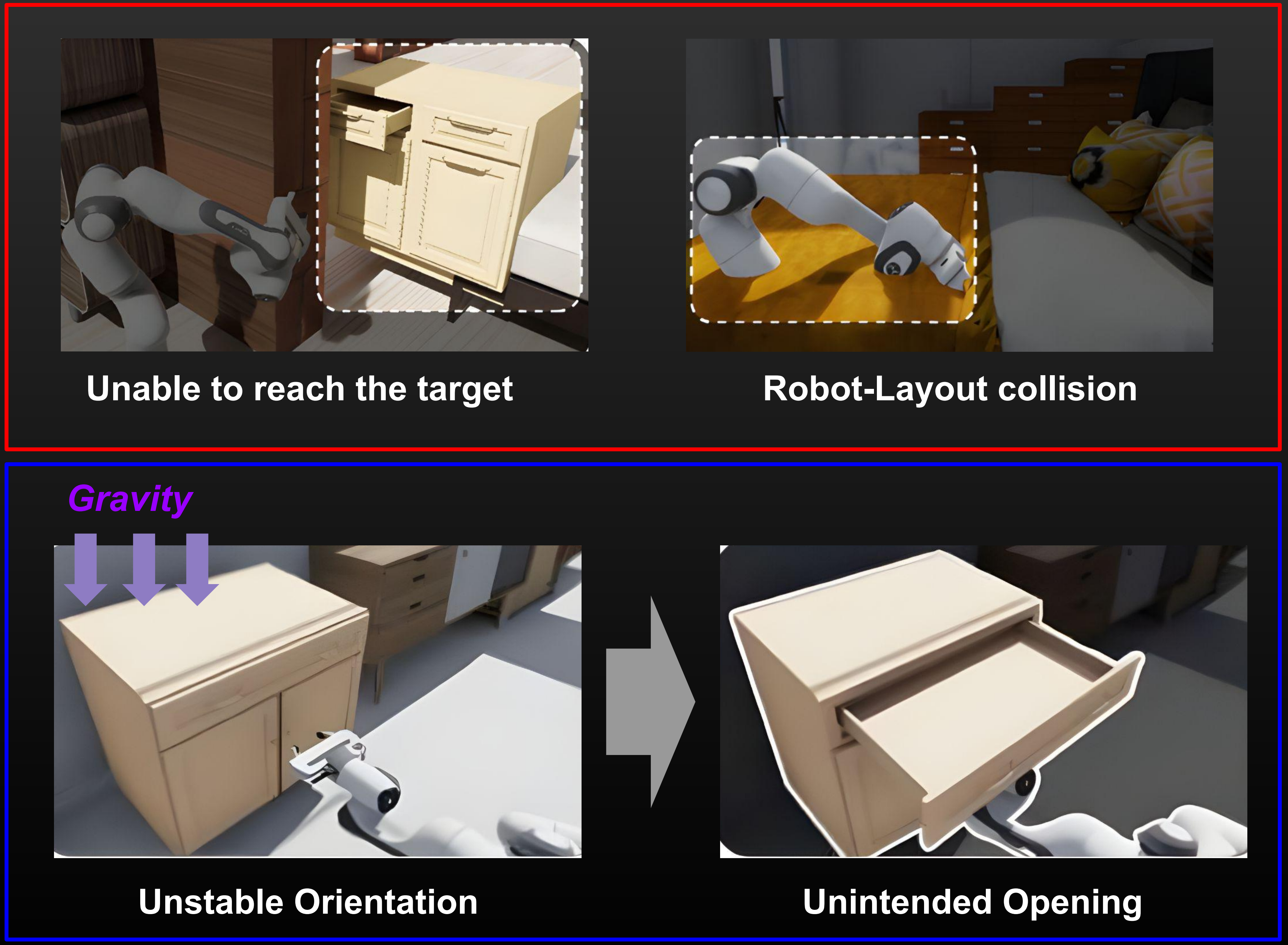
- ❑ Ensures consistent learning across identical tasks, independent of static scenes.
- ❑ Applicable to any static scene for action augmentation.

Static Scene Generation

Reality Lab
Soongsil University

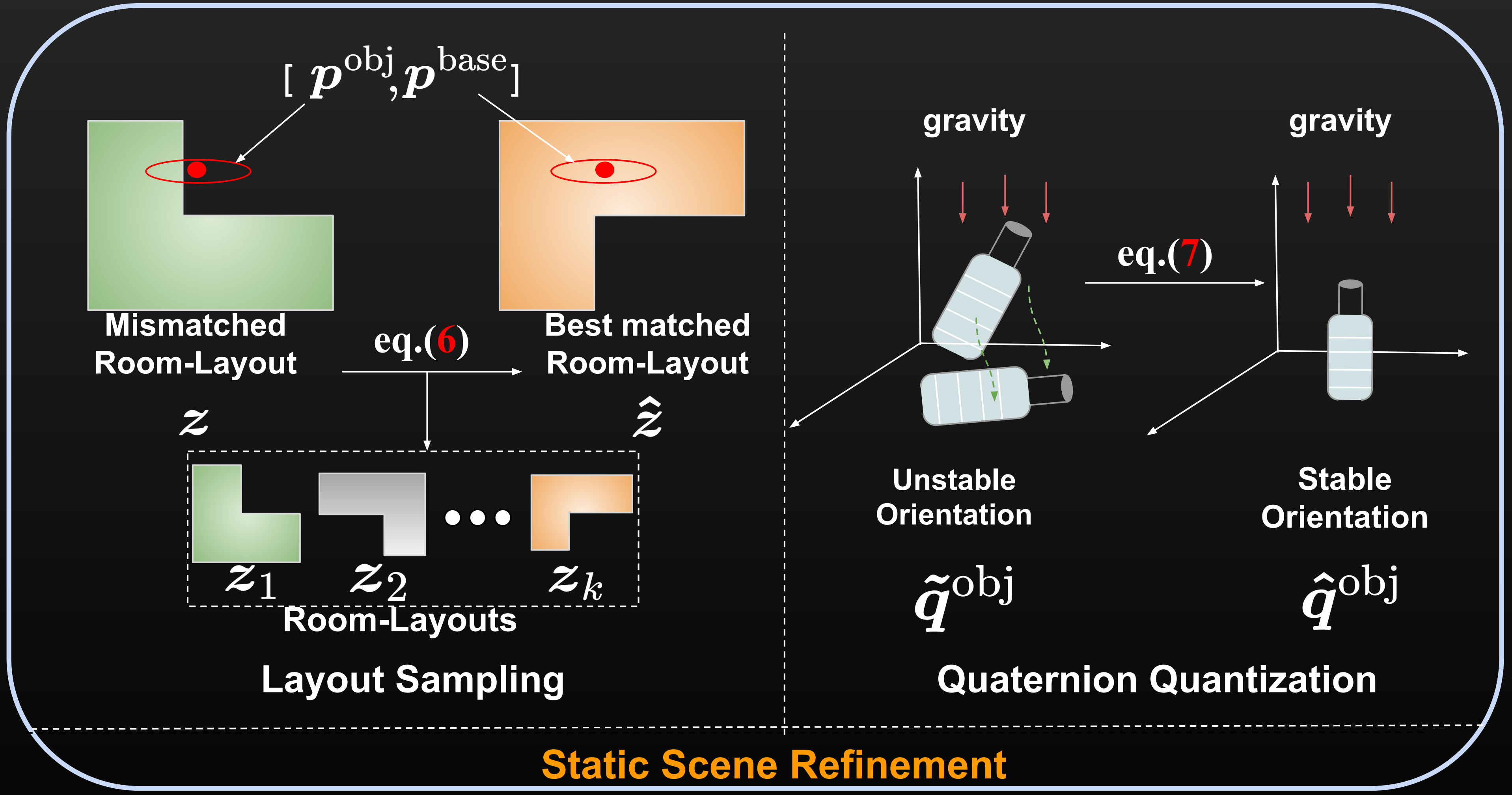
- Diffusion model effectively generates **realistic** static scenes aligned with instructions.
- However, requires refinement for ***physical plausibility*** and ***task accessibility***.





Refinement

- Two refinement techniques ensure physical feasibility and task execution.
  - Layout sampling to select collision-free room configurations using position difference.
  - Quaternion quantization stabilizes object orientations by discretizing rotations.



- Layout Sampling (eq.6)

$$\hat{r} = \arg\min_r \left(\|\boldsymbol{p}_r^{\text{obj}} - \tilde{\boldsymbol{p}}^{\text{obj}}\|^2 + \|\boldsymbol{p}_r^{\text{base}} - \tilde{\boldsymbol{p}}^{\text{base}}\|^2 \right)$$

$$\hat{\boldsymbol{z}} = \boldsymbol{z}_{\hat{r}}$$

- Quaternion Quantization (eq.7)

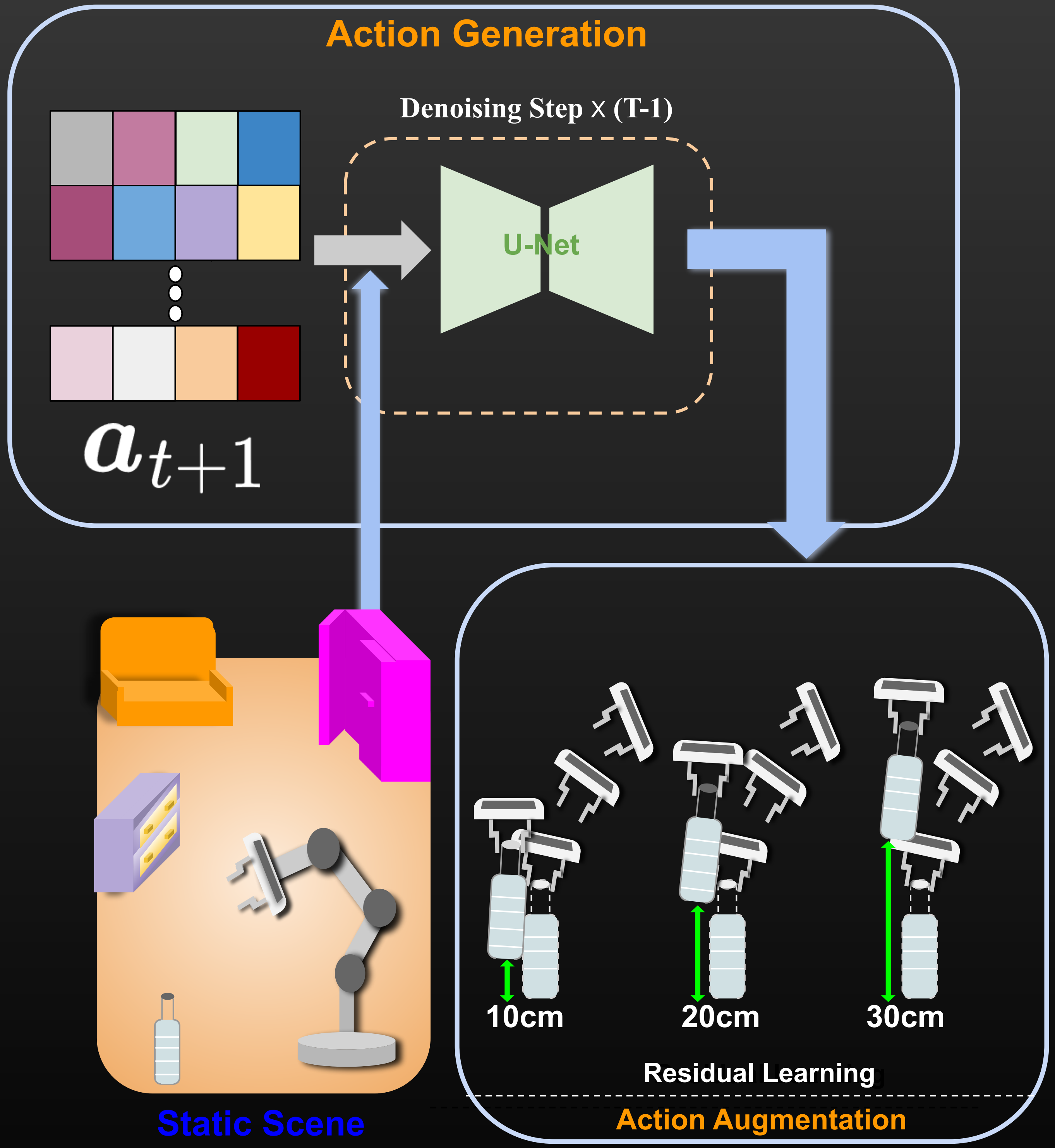
$$\hat{\boldsymbol{q}}^{\text{obj}} = \text{round}\left(\frac{\tilde{\boldsymbol{q}}^{\text{obj}}}{\delta} \right) \cdot \delta$$

Action Generation and Augmentation

- Action generation uses diffusion model conditioned on static scenes.

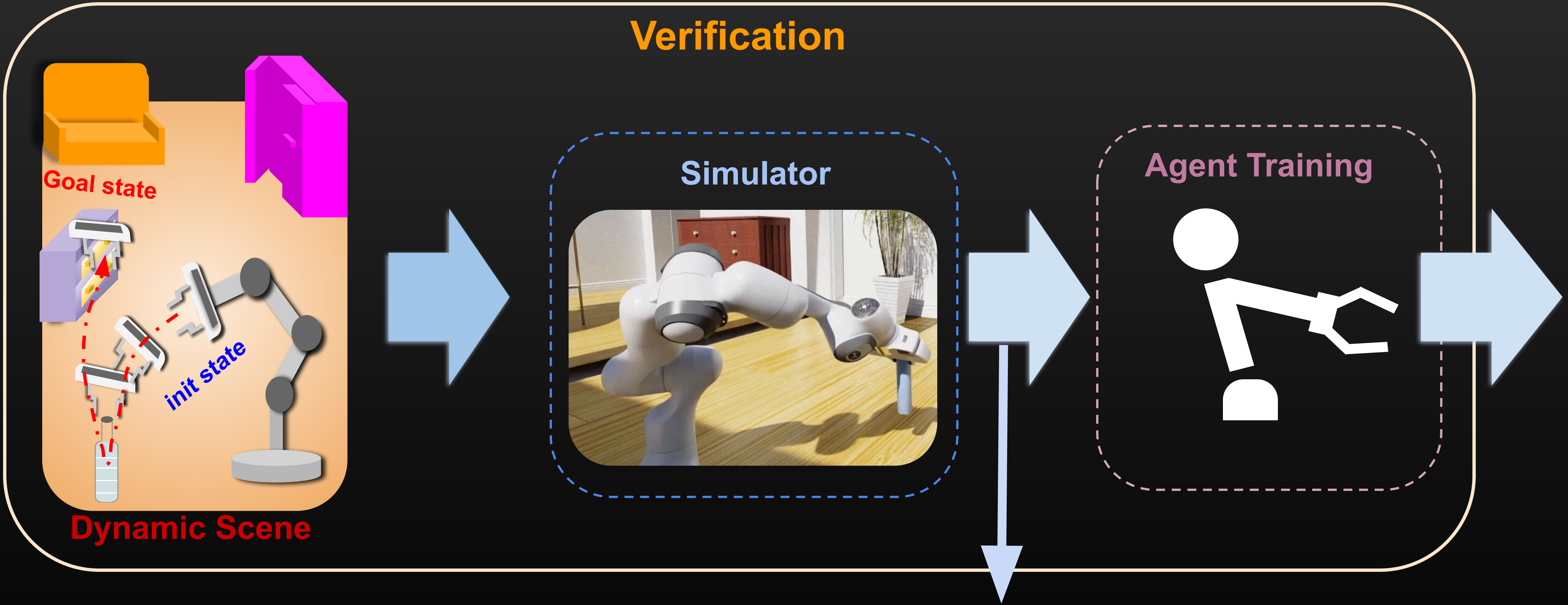
$$\mathcal{L}(\phi)_{\text{scene}} := \mathbb{E}_{\boldsymbol{a}_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\phi(\boldsymbol{a}_t, t; \mathbf{s})\right\|^2\right]$$

- Action augmentation generates multiple trajectories from single static scene.
  - Ex) 10 scenes × 10 actions = 100 diverse dynamic scenes.
  - Scalable augmentation
  - Spatial generalization



Filtering Invalid Dynamic Scenes

- **Inaccurate actions can degrade agent performance and cause task failures.**
  - **Generated dynamic scenes undergo physics simulation verification in *NVIDIA Isaac Sim*.**



High-quality data improves performance.

Only *successful* scenes pass

Experiments

- Comparison with Human Expert
  - ARNOLD benchmark contains human-demonstrated manipulation trajectories collected via Xbox controller.

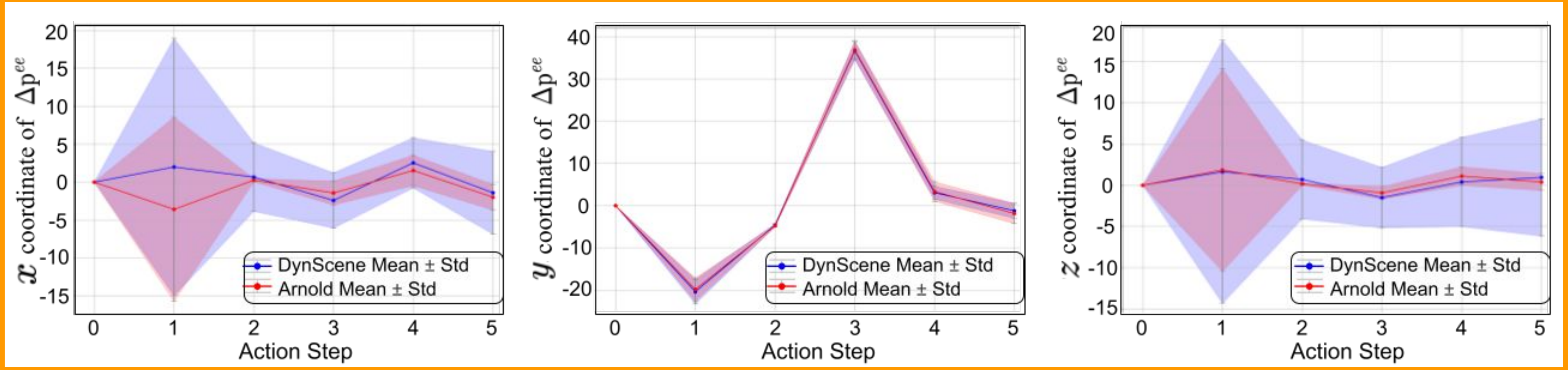
Method	Task	Time (sec)	Success Rate (%)
DynScene (Ours)	P.OBJECT	2.53 ± 0.02	92.00
	R.OBJECT	2.52 ± 0.02	88.00
	C.CABINET	2.50 ± 0.02	58.00
	O.CABINET	2.53 ± 0.02	37.00
	C.DRAWER	2.52 ± 0.02	95.00
	O.DRAWER	2.52 ± 0.02	41.00
	P.WATER	2.52 ± 0.02	83.00
	T.WATER	2.52 ± 0.02	62.00
	Average	**2.52** $\pm$ 0.02	**69.50**
ARNOLD† [7]	Average	67.50	37.50

➔ DynScene generates training data **26× faster** with **1.8× higher** success rate than human experts.

Experiments

Reality Lab
Soongsil University

- **Comparison with Human Expert**
  - **Analyzed end-effector position changes in 'pour water' task.**
  - **Y-axis remains similar due to task-specific vertical precision requirements**



→ Wider distributions along the **x-** and **z-** axes than ARNOLD, indicating **greater action diversity**.

Experiments

Reality Lab
Soongsil University

- **Comparison of Action Diversity**
  - **DynScene with ARNOLD training data, using the same number of valid dynamic scenes per task.**

Varied action paths — *Temporal variability* — *Broader gripper exploration*

Task	Fréchet Distance (FD) ↑		Dynamic Time Warping (DTW) ↑		Spatial Coverage (SC) ↑	
	ARNOLD [7]	**DynScene (Ours)**	ARNOLD [7]	**DynScene (Ours)**	ARNOLD [7]	**DynScene (Ours)**
Pickup Object	32.85	**36.38**	54.77	**61.23**	2.94	**3.29**
Reorient Object	23.49	**27.45**	27.98	**35.56**	2.30	**2.73**
Close Cabinet	37.77	**43.85**	76.72	**86.01**	**4.30**	3.21
Open Cabinet	**40.83**	40.59	81.71	**83.54**	2.92	**3.17**
Close Drawer	24.85	**25.55**	35.47	**37.86**	1.66	**3.78**
Open Drawer	**28.36**	24.36	**44.98**	39.51	2.67	**4.04**
Pour Water	20.46	**26.52**	31.09	**61.62**	2.96	**4.91**
Transfer Water	**18.52**	17.84	27.90	**36.65**	2.89	**3.85**
Average	28.39	**30.32**	47.58	**55.25**	2.83	**3.62**

➔ DynScene produces more diverse and spatially expansive actions than the training data.

Experiments

- **Text-Conditioned Scene Generation Results**
  - **Generate diverse and dynamic scenes from identical text prompts.**
  - **Variations in object shapes, initial states, and room layouts.**
  - **Preserves physical plausibility and semantic alignment.**

Experiments

- Evaluation with manipulation agents
  - **ARNOLD-only *vs.* ARNOLD + DynScene combination performance comparison.**
  - **Particularly effective for complex manipulation tasks.**

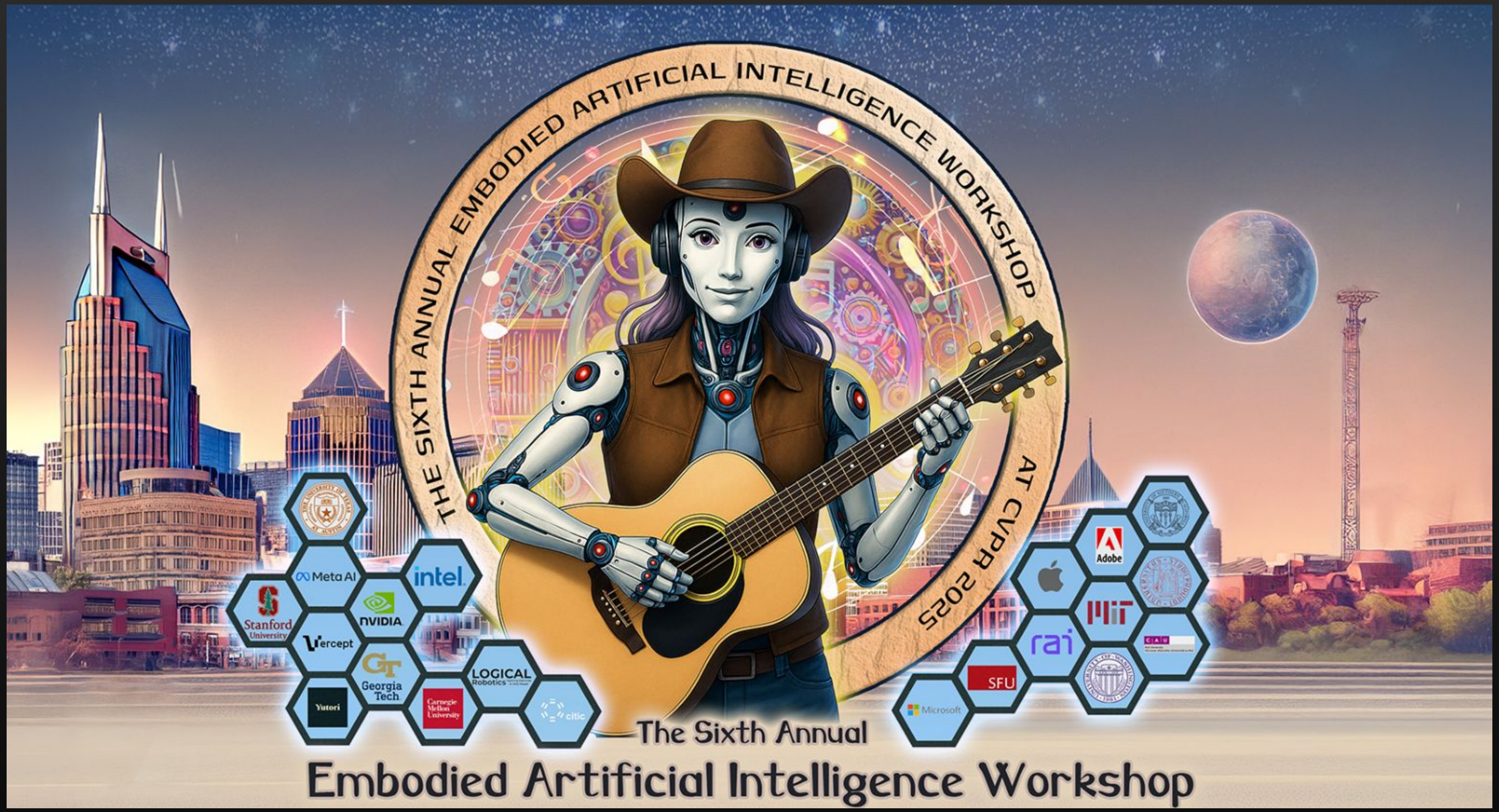
Method	P.Object	R.Object	O.Drawer	C.Drawer	O.Cabinet	C.Cabinet	P.Water	T.Water	Average
BC-Lang-CNN (ARNOLD)	**5.00**	0.00	0.00	20.00	0.00	10.00	0.00	0.00	4.35
BC-Lang-CNN (DynScene + ARNOLD)	**5.00**	0.00	0.00	**25.00**	0.00	**20.00**	0.00	0.00	**6.20**
BC-Lang-ViT (ARNOLD)	1.67	0.00	0.00	35.00	0.00	10.00	0.00	0.00	5.84
BC-Lang-ViT (DynScene + ARNOLD)	**5.00**	0.00	0.00	**45.00**	0.00	**33.33**	0.00	0.00	**10.42**
PerAct (ARNOLD)	88.81	3.90	26.05	**33.78**	11.59	20.39	**34.33**	14.29	29.14
PerAct (DynScene + ARNOLD)	**90.00**	**20.00**	**38.33**	30.00	**16.67**	**31.67**	30.00	**21.67**	**34.79**
PerAct-PSA (ARNOLD)	90.00	**30.00**	41.67	**51.67**	20.00	15.00	**63.33**	20.00	41.46
PerAct-PSA (DynScene + ARNOLD)	**95.00**	25.00	**43.33**	43.33	**45.00**	**38.33**	46.67	**28.33**	**45.62**

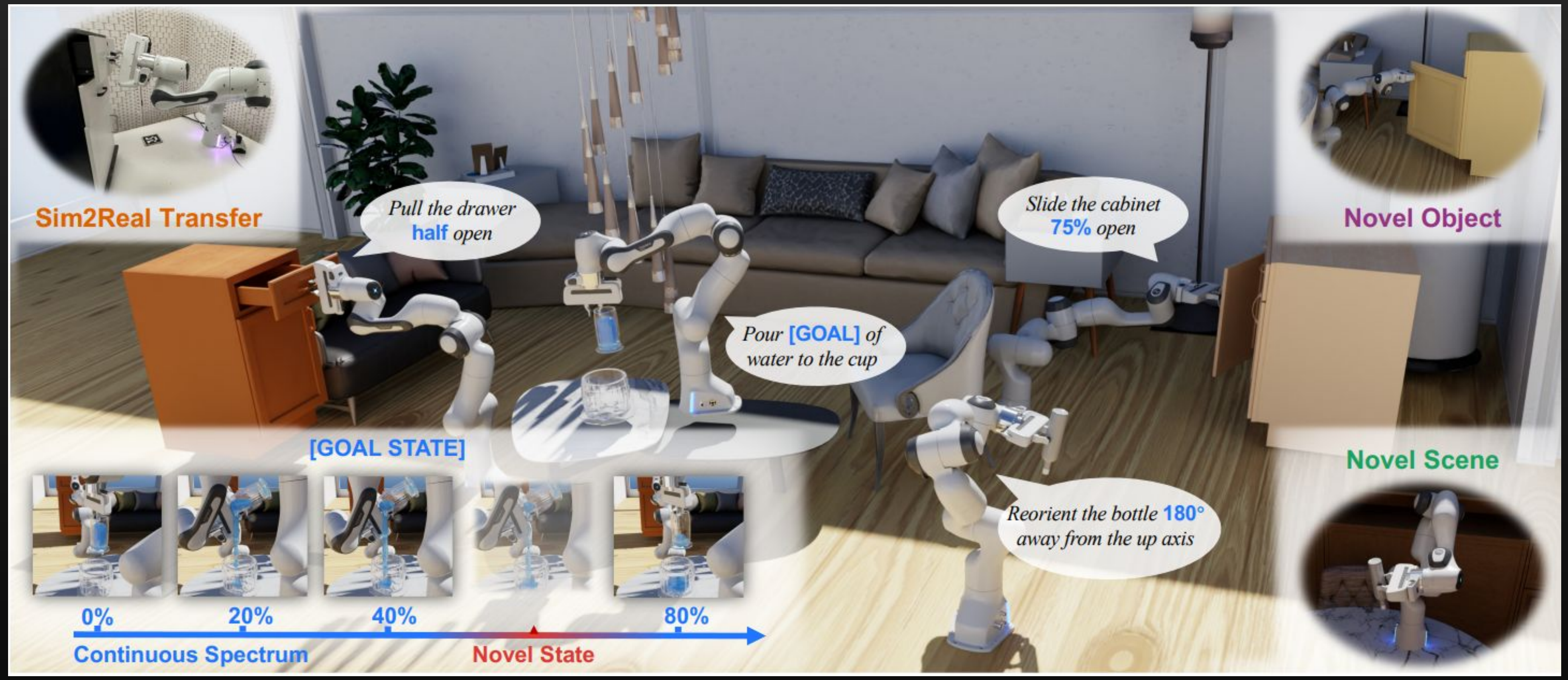
→ Integrating DynScene data **significantly boosts** robotic manipulation performance.

ARNOLD Challenge (CVPR2025 Embodied AI Workshop)

Reality Lab
Soongsil University

- **ARNOLD Challenge is a robotic manipulation challenge.**
  - ARNOLD Benchmark includes 8 manipulation tasks with continuous states and novel object/scene generalization.





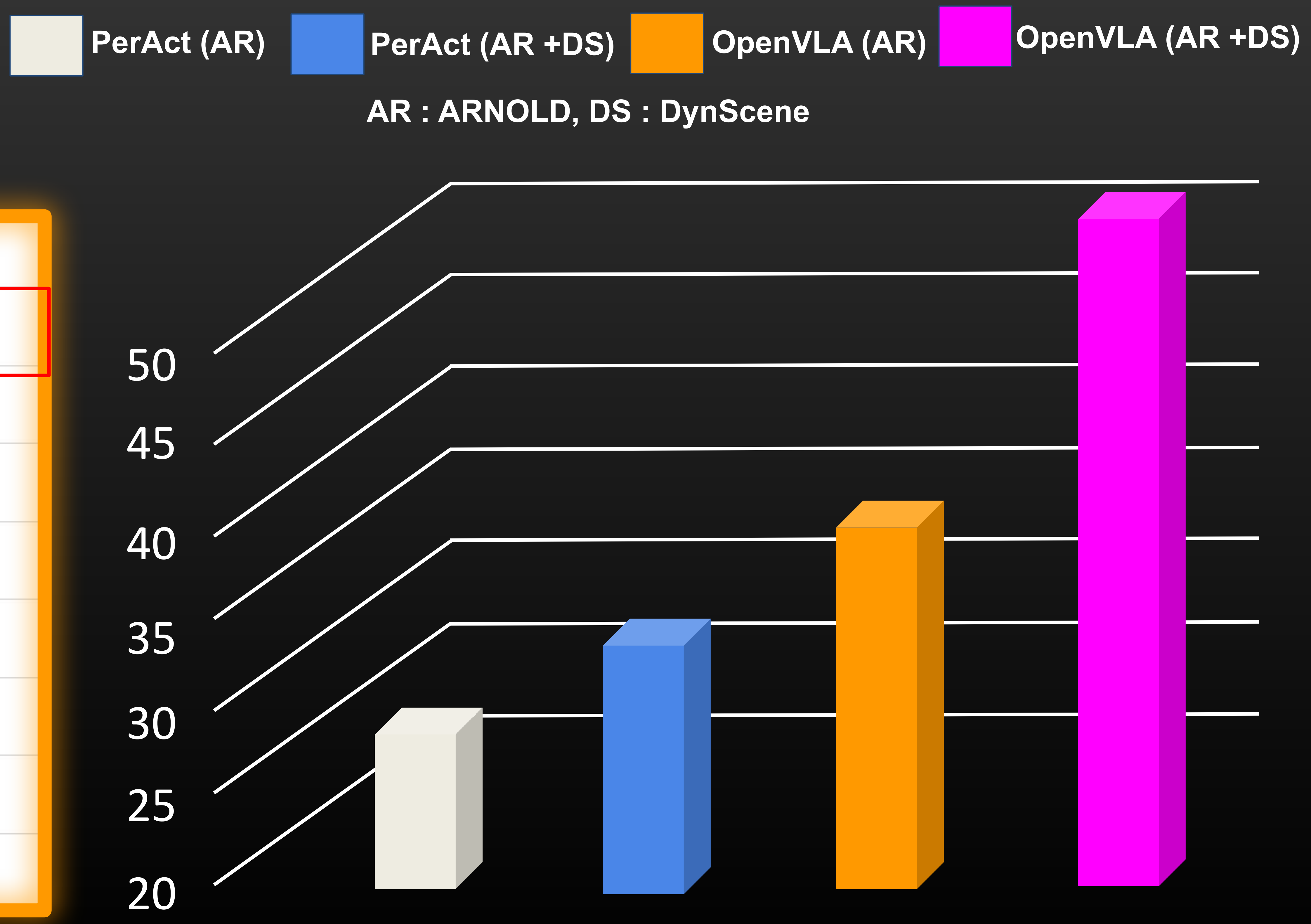
Results on ARNOLD Challenge

Reality Lab
Soongsil University

- We trained foundation models using DynScene-generated data.

1st Place

Rank	Participant team	SR (↑)
1	RealityLab	0.49
2	EBDAI	0.45
3	Fun Guy (Fusion(SD&PC))	0.39
4	larr (final)	0.32
5	Windboy (DT1)	0.25
6	MilkyWay	0.25
7	MCC-EAI	0.25
8	Host_31221_Team B	0.22



Summary

- Unified framework generates dynamic scenes from text instructions.
- Residual actions enable spatial generalization across configurations.
- Physics-based refinement techniques ensuring collision-free and stable scene generation.
- Efficient data generation achieves 26.8× faster speed with superior agent performance.

Thank You