CVPR Nashville JUNE 11-15, 2025
June 11 – 15, 2025 | Nashville TN, USA

Embodied AI Workshop @ CVPR 2025
ARNOLD CHALLENGE

Fine-tuning Robotic Foundation Model Using Dynamic Scene Generation

Challenge Team: **Reality Lab**





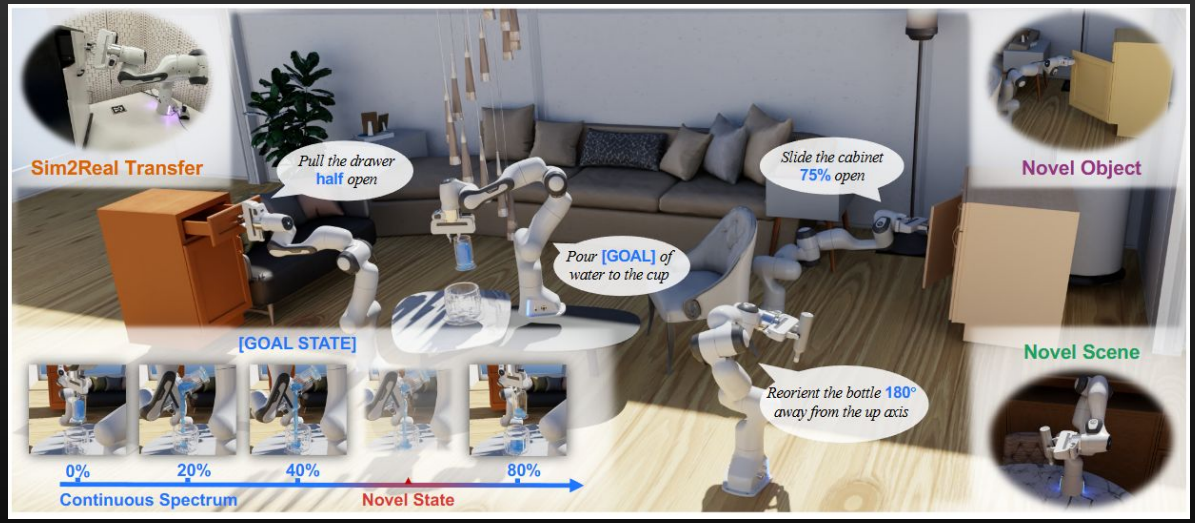
Dowon Kim, Chaewoo Lim, Sungyong Park, Sangmin Lee, Heewon Kim
Soongsil University

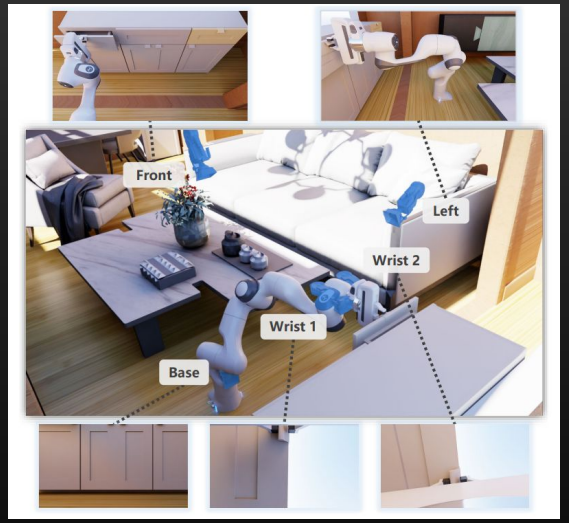
Reality Lab
Soongsil University
https://reality.ssu.ac.kr/

Robotic Manipulation in Embodied AI

Reality Lab
Soongsil University

- Goal: Perceive and manipulate objects to complete designated tasks.
- Challenge: Real-world manipulation requires *generalization* to unseen scenarios.



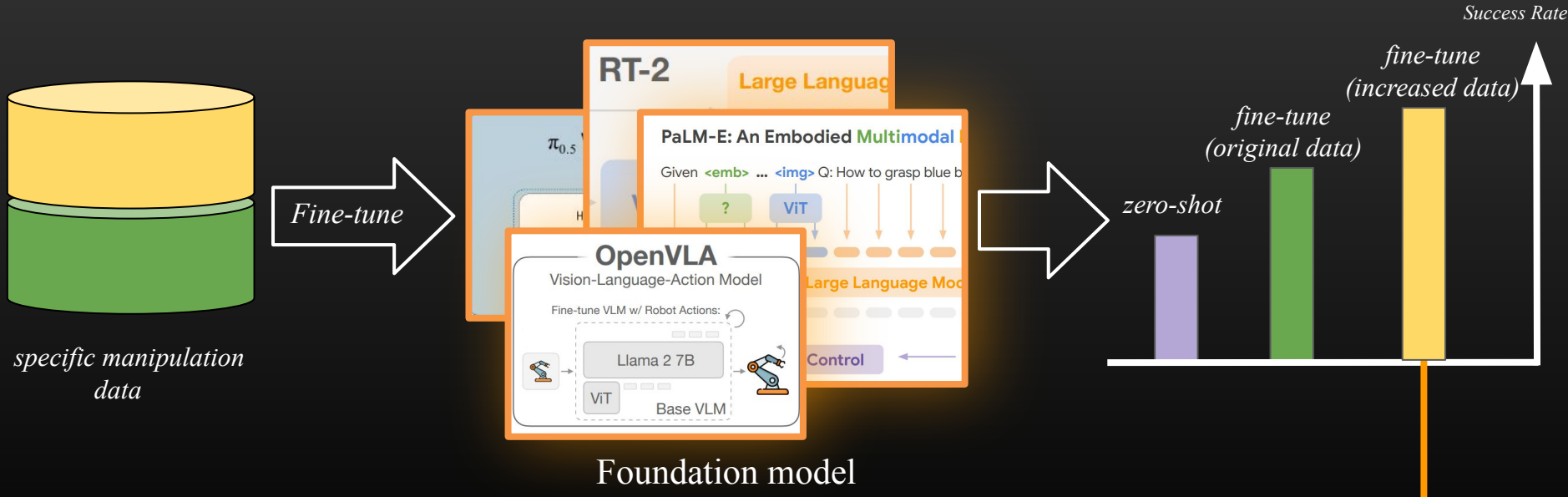


→ *To support this, ARNOLD provides novel objects, scenes, and goal states in evaluation.*

Foundation Model for Robotic Manipulation

Reality Lab
Soongsil University

- Foundation models are a promising approach to ***generalization*** in robotic manipulation.
- However, even these models ***require fine-tuning*** data to optimize for specific tasks.



- ***The scale and variety of training data for effective fine-tuning is still underexplored.***

Exploring Data Requirements for Fine-Tuning

Reality Lab
Soongsil University

- **In this challenge, we investigate:**

  We explore strategies to generate training data effective for agent learning.

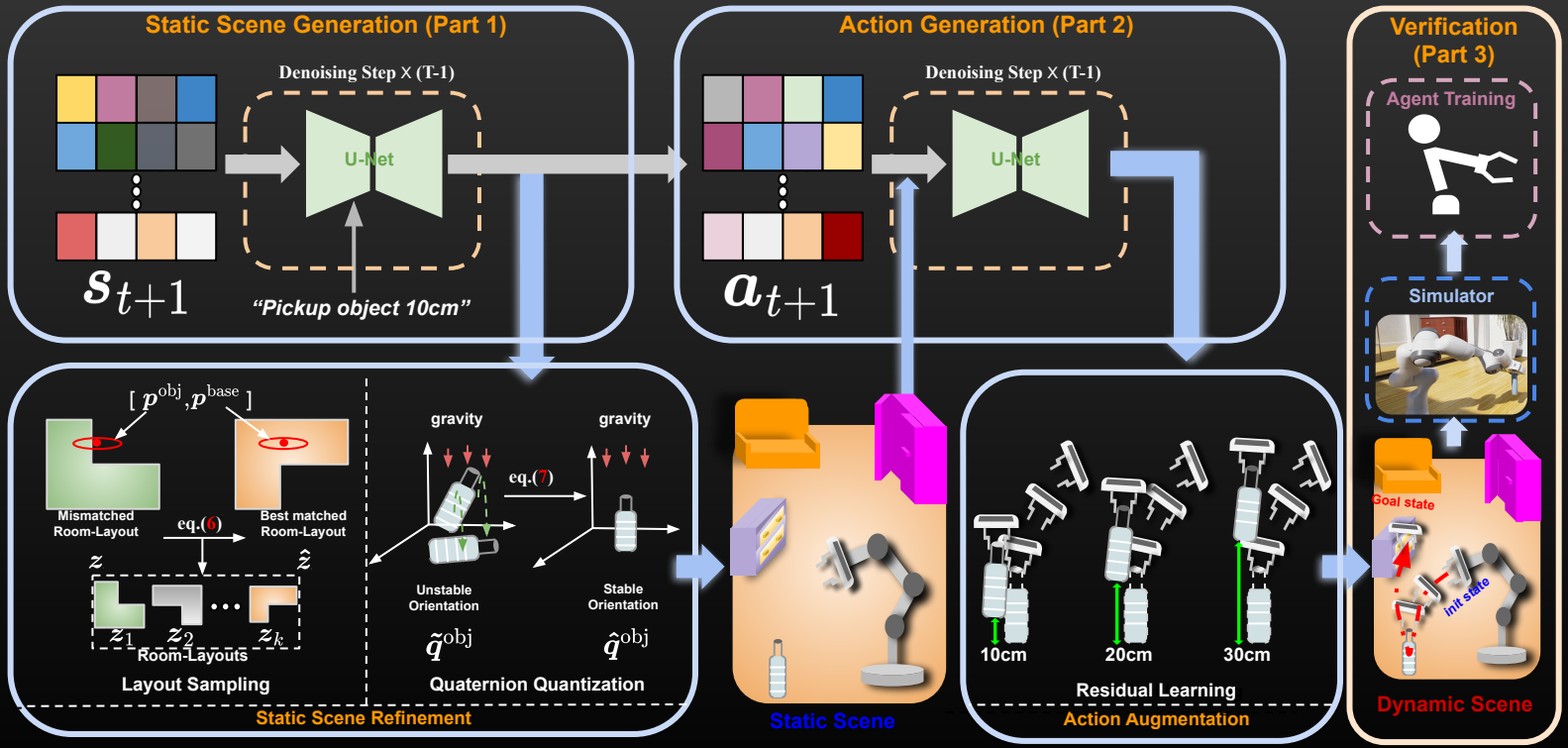
- **Our solution: DynScene [1] - robotic manipulation scene generation method enabling scalable dataset creation.**

[1] Sangmin Lee, Sungyong Park, and Heewon Kim, "DynScene: Scalable Generation of Dynamic Robotic Manipulation Scenes for Embodied AI", CVPR 2025.

Proposed Method : DynScene

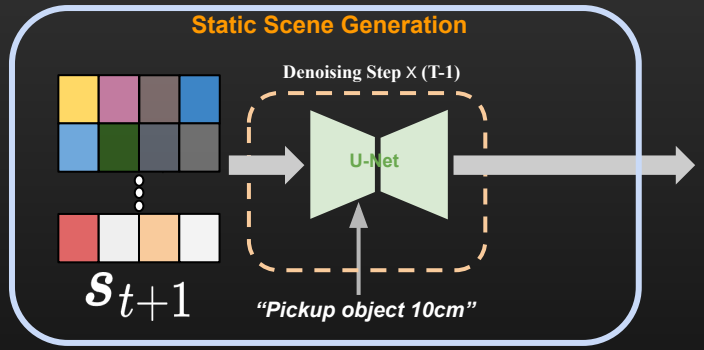
Reality Lab
Soongsil University

- **Framework Overview**



Part 1: Static Scene Generation

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



Unable to reach the target

Robot-Layout collision

Gravity

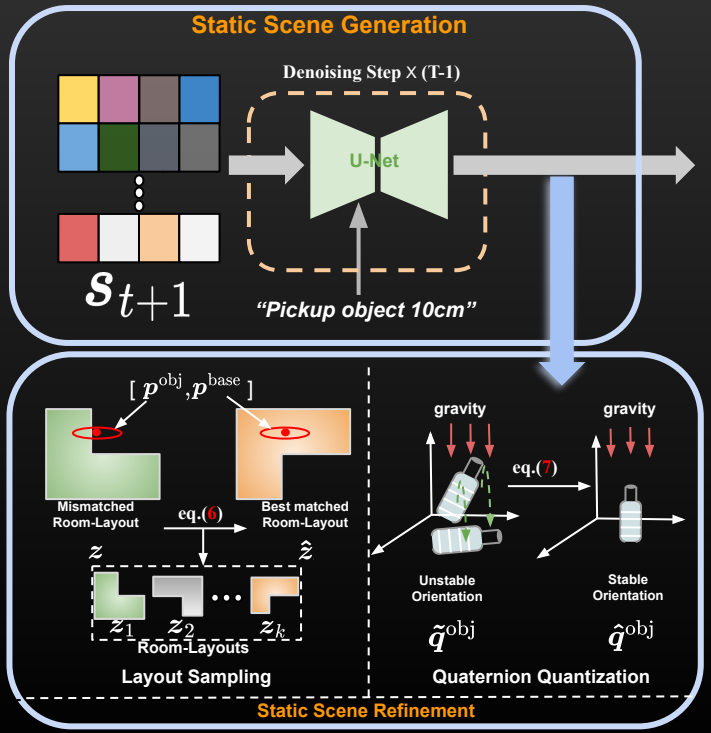
Unstable Orientation

Unintended Opening

Part 1: 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.***



- **Layout Sampling** (eq.6)

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

Unable to reach the target

Robot-Layout collision

Gravity

- **Quaternion Quantization** (eq.7)

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

Unstable Orientation

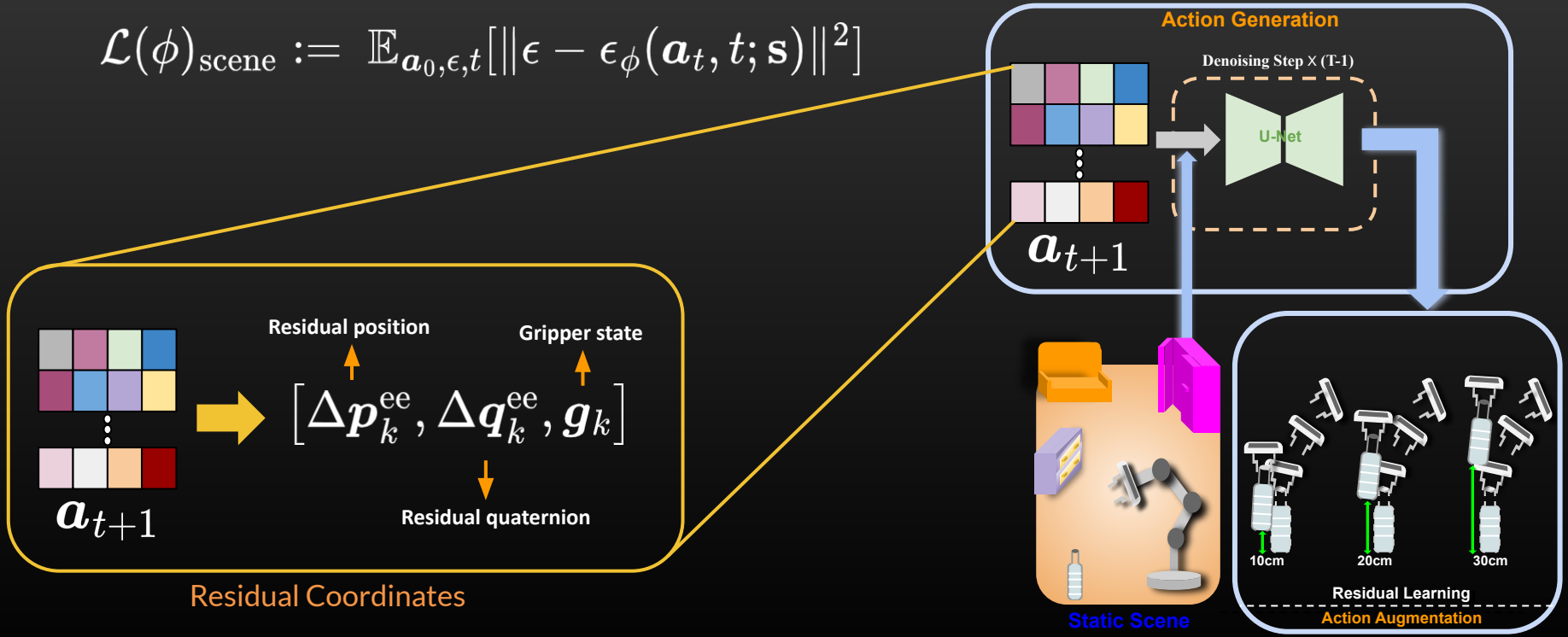
Unintended Opening

Part 2: Action Generation and Augmentation

Reality Lab
Soongsil University

- **Action generation uses diffusion model conditioned on static scenes.**

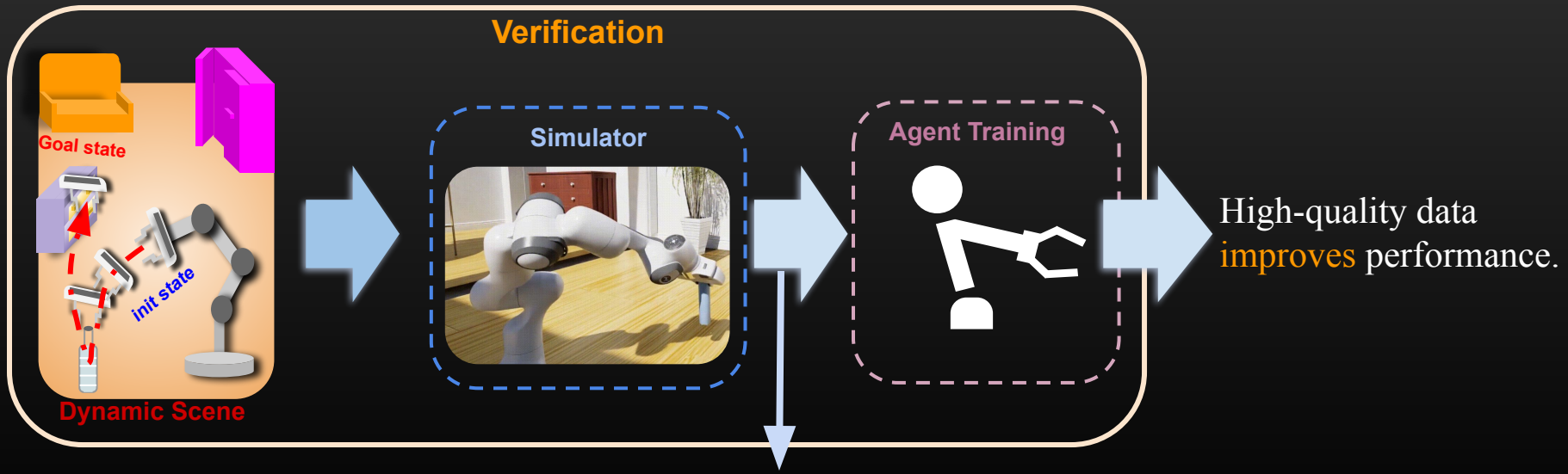
$$\mathcal{L}(\phi)_{\text{scene}} := \mathbb{E}_{\boldsymbol{a}_0,\epsilon,t}[\|\epsilon - \epsilon_\phi(\boldsymbol{a}_t, t; \mathbf{s})\|^2]$$



Part3: Filtering Invalid Dynamic Scenes

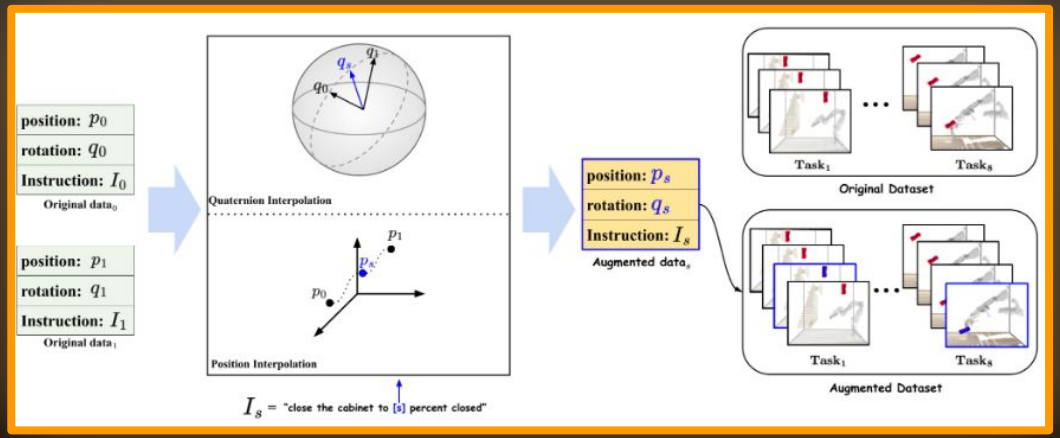
Reality Lab
Soongsil University

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

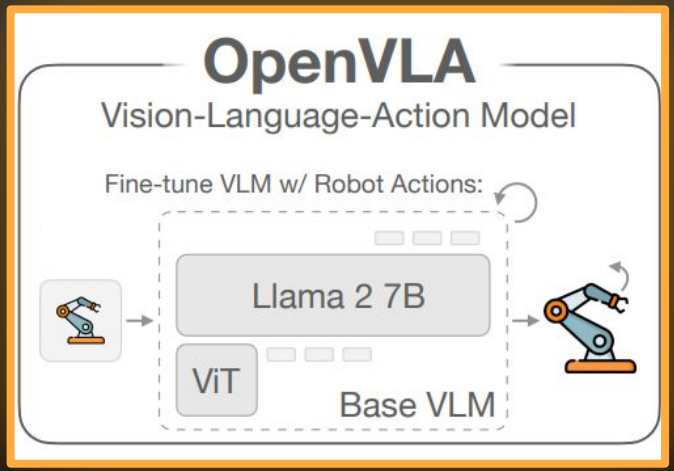


Implementation Details

Training OpenVLA with state interpolation augmentation



State Interpolation
(3rd place in 2024)



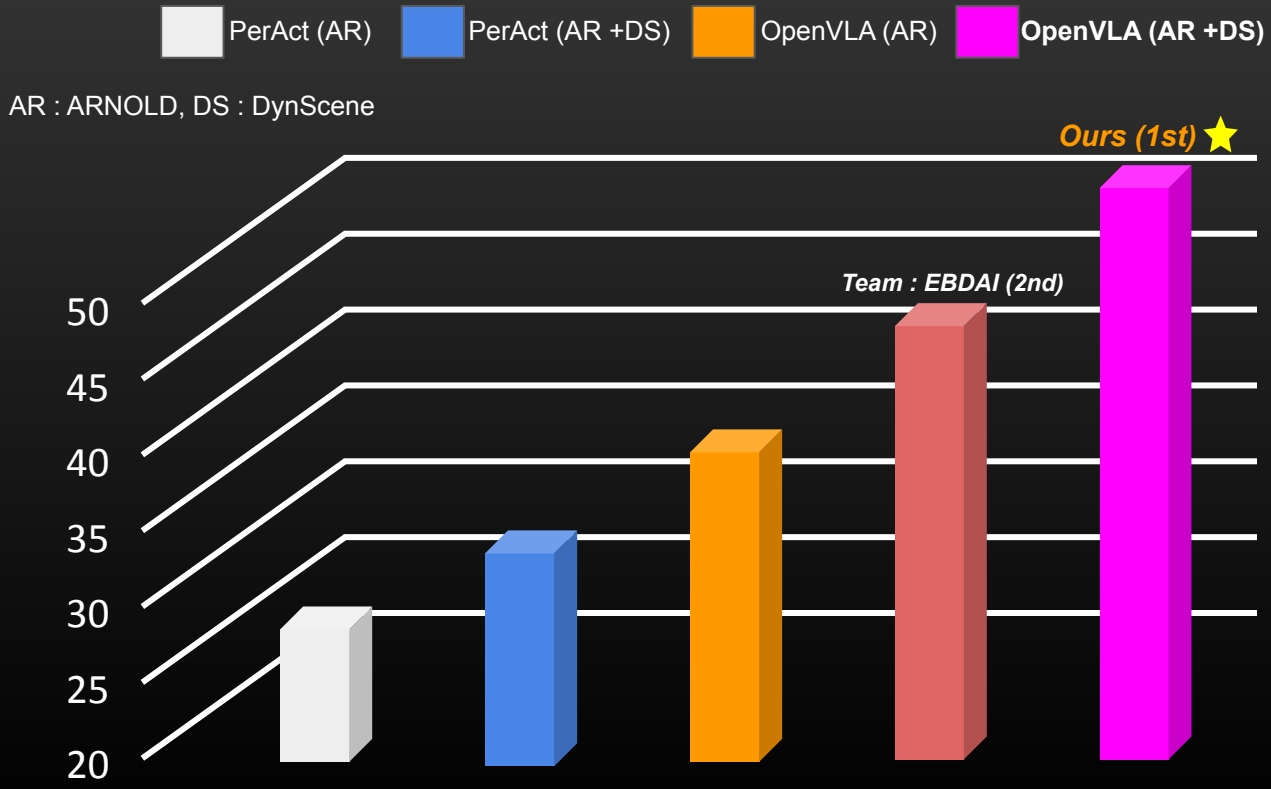
OpenVLA [2]

[2] Kim et al., "OpenVLA: An Open-Source Vision-Language-Action Model", arXiv:2406.09246 (2024).

Results

Reality Lab
Soongsil University

- **We trained foundation models using DynScene-generated data.**



Reality Lab
Soongsil University

Thank you

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