Fine-tuning Robotic Foundation Model Using Dynamic Scene Generation

Sungyong Park*[1,2], Sangmin Lee*[1†], Dowon Kim[1‡], Heewon Kim[1,2]
[1]Soongsil University, [2]Kairoba
{ejqdl1010, sm32289, a01066304767}@gmail.com, hwkim@ssu.ac.kr

Abstract

Continuous-state robotic manipulation requires an agent to interpret language goals, estimate object states, and execute physically valid actions toward fine-grained targets. However, training data for such tasks is sparse: collecting demonstrations for every object, scene, and target state is expensive, and synthetic data is useful only when it remains executable. In this paper, we present the methods developed for the ARNOLD Challenge [1] on continuous-state robotic manipulation. ***PSASI*** *improves continuous-state supervision through state interpolation and phase-specific control.* ***FRFM*** *further expands the training distribution with physics-verified dynamic scene generation for foundation-model fine-tuning. Experiments on ARNOLD show that PSASI improves continuous-state manipulation, while FRFM achieves stronger performance by fine-tuning a foundation policy with DynScene-generated data.*

1. Introduction

Robotic manipulation aims to enable agents to follow language instructions and interact with objects in the environment. Prior benchmarks have largely focused on discrete task completion [5–7, 9]. In contrast, ARNOLD [2] evaluates continuous-state manipulation, where agents reach fine-grained target states specified by language. The main challenge is the sparsity of continuous-state supervision. Training demonstrations cover only limited objects, scenes, and target states. During evaluation, agents are tested on novel combinations outside the training distribution. This gap is critical because small changes in the target state can require different action magnitudes or control strategies.

To address these limitations, we propose two complementary methods for improving continuous-state robotic manipulation through manipulation-aware data expansion. First, we introduce **PSASI** (*Phase-Specific Agent and State Interpolation*), which addresses the sparsity of continuous-state supervision. PSASI decomposes manipulation into phase-specific control regimes and synthesizes intermediate state-conditioned samples between sparse demonstrations. By separating heterogeneous action phases and densifying supervision along the continuous goal dimension, PSASI improves generalization to unseen target states. Second, we introduce **FRFM** (*Fine-tuning Robotic Foundation Model using Dynamic Scene Generation*), which extends data expansion beyond state interpolation. FRFM fine-tunes a vision-language-action foundation model with executable manipulation data generated by DynScene [4].

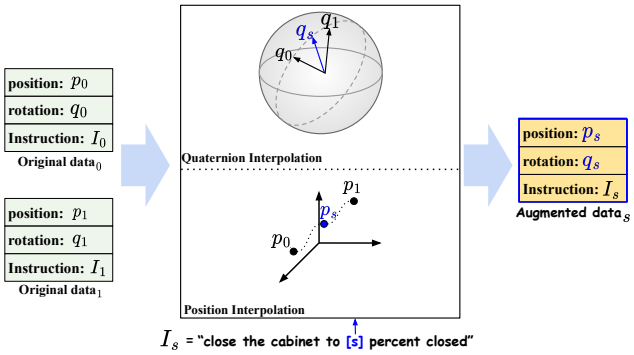

Figure 1. **State interpolation.** Given sparse demonstrations, we synthesize intermediate object states along the continuous goal dimension and pair them with updated goal-conditioned supervision.

Experiments on the ARNOLD dataset show that structured data expansion improves continuous-state manipulation and that generated executable data becomes more effective when combined with a stronger foundation policy.

2. Proposed Methods

2.1. Phase-specific agent and state interpolation

We introduced PSASI (*Phase-Specific Agent and State Interpolation*) for continuous-state robotic manipulation. The key observation was that a manipulation trajectory contains multiple phases with distinct control characteristics. For example, approaching an object, grasping it, and moving it toward a desired continuous state require different visual cues

*Equal contribution
†Now at Qualcomm
‡Now at LG AI Research

Table 1. **Evaluation on the ARNOLD test set.** We report success rates (%) for each task category and the overall average.

Policy	P.OBJ	R.OBJ	O.DRAWER	C.DRAWER	O.CABINET	C.CABINET	P.WATER	T.WATER	OVERALL
PerAct	88.81	3.90	26.05	33.78	11.59	20.39	34.33	14.29	29.14
PerAct + PSASI	**95.00**	25.00	43.33	43.33	**45.00**	38.33	46.67	28.33	45.62
FRFM	85.00	**85.00**	**50.00**	**90.00**	35.00	**55.00**	**75.00**	**65.00**	**57.00**

Table 2. **Ablation study of PSASI.** We report success rates (%) on the validation and test splits.

		PerAct†	PSA	**PSASI**
Setting	Phase-Specific Agent (PSA)	✗	✓	✓
	State Interpolation (SI)	✗	✗	✓
Validation set	Grasp	46.00	**61.00**	**61.00**
	Manipulation	42.00	**58.00**	57.00
	Grasp + Manipulation	34.00	**45.00**	**45.00**
Test set	Grasp + Manipulation	22.00	28.00	**31.00**

Table 3. **Evaluation on the generalization splits.** SI denotes state-interpolation, and DS denotes DynScene-generated data.

	PerAct	OpenVLA-OFT	OpenVLA-OFT	OpenVLA-OFT
SI	✗	✗	✓	✓
DS	✗	✗	✗	✓
Novel object	22.84	27.92	31.88	**33.33**
Novel scene	29.73	37.71	41.67	**44.58**
Novel state	4.49	26.88	36.46	**40.21**
Any state	16.36	33.33	**45.63**	44.17

and action distributions. A 3D voxel-based policy [8] can mix heterogeneous action distributions across manipulation phases, leading to ambiguous supervision. To address this issue, we decompose the manipulation process into phase-specific agents, allowing each policy to specialize in a more focused control regime.

To further improve generalization over continuous target states, we introduce state interpolation, as illustrated in Fig. 1. Given two sparse demonstrations with different target states, we obtain an intermediate state s and its corresponding end-effector pose by interpolation: the position is interpolated linearly, and the orientation is interpolated using spherical linear interpolation (SLERP). We then use the interpolated pose to form a new goal-conditioned sample and update the instruction accordingly.

2.2. Fine-tuning foundation models with Dynscene

State interpolation expands the sparse demonstration set along the state dimension through interpolation. However, it does not explicitly increase scene or object diversity. To further broaden the training distribution, we propose **FRFM** (*Fine-tuning Robotic Foundation Model using Dynamic Scene Generation*), which combines state interpolation with DynScene-generated executable data and fine-tunes a vision-language-action foundation model, OpenVLA-OFT [3]. DynScene generates text-conditioned dynamic manipulation scenes, including diverse object-scene configurations and corresponding robot action sequences. Moreover, its residual action representation enables action augmentation, generating diverse actions from the same scene. By combining the original ARNOLD demonstrations with DynScene-generated data, the policy observes broader variations in scenes, objects, and executable state transitions. This provides richer supervision for foundation-model fine-tuning and improves generalization beyond the original demonstration distribution.

3. Experiments

Table 1 summarizes the ARNOLD test-set results. PerAct [8] obtains 29.14% overall success, showing the difficulty of continuous-state manipulation with a conventional voxel-based policy. By incorporating phase-specific control and state interpolation, PSASI improves the overall success rate to 45.62%. FRFM achieves the strongest overall performance, reaching 57.00% success. These results show that dense continuous-state supervision improves conventional manipulation policies, and diverse generated manipulation data further benefits foundation-model fine-tuning.

Table 2 validates the design of PSASI. We separately evaluate grasping, manipulation, and the full grasp-to-manipulation pipeline to analyze how phase-specific training affects each stage. The phase-specific agent improves the test-set success rate from 22.00% to 28.00% by separating grasping and manipulation into more focused control regimes. With state interpolation, PSASI further improves the success rate to 31.00%, indicating that intermediate state-conditioned samples help learning from sparse continuous-state demonstrations.

Table 3 shows a similar trend in the generalization split. OpenVLA-OFT [3] improves over PerAct across novel object, novel scene, novel state, and any-state settings, indicating the benefit of stronger model capacity. State interpolation mainly improves novel-state generalization, while DynScene further improves novel-object, novel-scene, and novel-state performance by expanding the data distribution beyond existing demonstrations. Overall, these results show that continuous-state manipulation benefits from both denser state supervision and broader scene-level variation.

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Advanced GPU Utilization Support Program, the National Program for Excellence in SW (2024-0-00071), and the graduate school support program (IITP-2026-RS-2024-00430997, IITP-2026-RS-2024-00426853) supervised by the IITP (Institute of Information & communications Technology Planning & evaluation) and the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (RS-2025-24803335) and the Ministry of Education (MOE) for the G-LAMP program (RS-2025-25441317), and Seoul Metropolitan Government for the RISE program (2025-RISE-01-020-04), and Basic Science Research Program (RS-2021-NR060140). We thank Corca AI for sponsoring the compute used in this work.

References

[1] https://embodied-ai.org/cvpr2026/. 1

[2] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, et al. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20483–20495, 2023. 1

[3] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 2

[4] Sangmin Lee, Sungyong Park, and Heewon Kim. Dynscene: Scalable generation of dynamic robotic manipulation scenes for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12166–12175, 2025. 1

[5] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023. 1

[6] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[7] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 1

[8] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023. 2

[9] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021. 1