Keep the Flow: Continuous High-Precision Bimanual Manipulation in Constrained Spaces

Hojae Jeong[1*] Juhyoung Lee[1*] Eunwoo Song[1] Jaerim Kim[1] Heewon Kim[1,2†]
[1]Soongsil University [2]Kairoba
{98eddiechung, i0179, song200348, kkjjrr0504}@soongsil.ac.kr, hwkim@ssu.ac.kr

Abstract

Recent robotic datasets have scaled generalist policies, but most lack strict metric constraints for precise object placement. To address this, we introduce Keep the Flow, a real-world bimanual humanoid dataset explicitly designed for spatially constrained manipulation. We collect these demonstrations via VR teleoperation on a Unitree G1 humanoid equipped with three-fingered Dex3-1 hands, covering both Single-row and Multi-row task formations. To ensure high-precision annotations, an automated overhead RGB-D verification loop filters out demonstrations whose final placement error exceeds 2 cm. Furthermore, we develop an action-state-preserving visual augmentation pipeline that combines SAM2-based object recoloring with workspace grid removal.

1. Introduction

Practical arrangement tasks require more than semantic object placement. Goal-directed manipulation can define final scenes through object poses and inter-object relations [1], while shelf replenishment requires insertion space, low disturbance, and stability [2]. Thus, final-scene spatial configuration should be treated as an explicit evaluation target.

Meanwhile, robot learning has shifted toward scalable data collection and general-purpose manipulation. Teleoperation systems support bimanual, mobile, dexterous, and humanoid demonstrations [3–9], while large corpora and high-DoF datasets cover diverse tasks, environments, embodiments, and multimodal observations [10–16]. Generalist robot policies exploit this data for language-conditioned control across platforms [17–20]. However, many datasets record trajectories without explicitly specifying where objects should end up or how they should relate.

Continuous-state and spatial-grounding work addresses part of this issue: ARNOLD [21] moves beyond discrete object-state labels toward continuous object and goal states, and spatial grounding methods study metric spatial reasoning, 3D-aware VLA representations, and robotics-oriented supervision [22–24]. Yet real-world humanoid bimanual demonstrations for precise dexterous arrangement remain underexplored (Table 1).

Table 1. **Dataset comparison.** Representative manipulation datasets and benchmarks are compared by robot data setting, object final-state goals, and number of trajectories.

		Robot Data Setting		Object Final-State Goals	
Dataset / Benchmark	**Trajectories**	**Real-world Humanoid**	**Dexterous Hand**	**Continuous Goal State**	**Spatial Constraints**
ARNOLD [21]	10k	×	×	✓	×
AgiBot World [16]	1M	✓	✓*	×	×
RoboMIND [15]	19k†	✓	✓*	×	×
ActionNet [13]	13k	✓	✓	×	×
Humanoid Everyday [14]	10.3k	✓	✓	×	×
Keep the Flow (Ours)	10.8k	✓	✓	✓	✓

*Mostly grippers and only a few dexterous hands.
†19k represents the humanoid-related data out of the total dataset.

We introduce **Keep the Flow**, a real-world humanoid bimanual manipulation dataset collected via VR teleoperation on a Unitree G1 humanoid equipped with three-fingered Dex3-1 hands for Single-row and Multi-row formations. An overhead RGB-D verification loop removes demonstrations exceeding a 2 cm placement error, and SAM2-based [25] object recoloring with workspace grid removal expands the data 8-fold to 10.81k trajectories and roughly 7.79M frames. **Keep the Flow** provides curated data for spatially precise bimanual manipulation under goal-directed obstacle rearrangement and metric inter-object spacing.

2. Dataset

2.1. System Setup

Our hardware platform consists of a Unitree G1 humanoid robot equipped with Dex3-1 hands. The vision system integrates three RGB-D cameras: a head-mounted RealSense D435i for an egocentric view, a wrist-mounted D405 for close-up manipulation tracking, and an overhead D435i for global workspace observation. For teleoperation, wrist poses are processed using a Pinocchio-based IK solver [26],

*Equal contribution
†Corresponding author

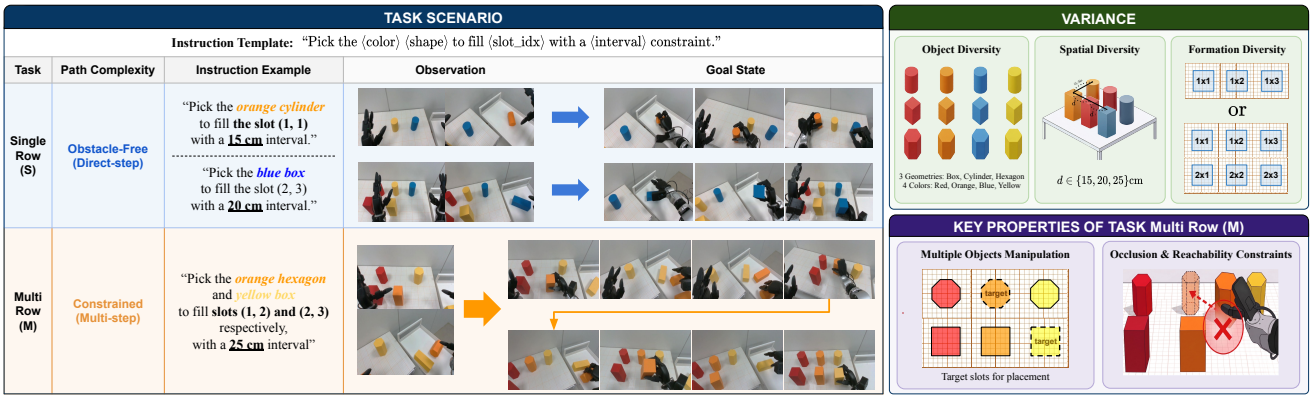


Figure 1. **Task Overview.** An overview of the task scenarios and dataset variations for the **Keep the Flow** dataset. The left panel illustrates the task configurations, while the right panel details the parameterized axes of dataset variance and key properties of the task.

and finger motions are mapped via Dex-Retargeting [9] using a Meta Quest 3. To ensure manipulation stability, the robot's lower body remains in a fixed standing posture, while the waist yaw tracks the operator's head orientation in real time.

2.2. Task Design

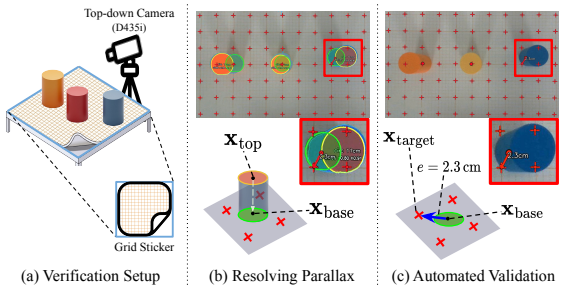
To capture complex, long-horizon trajectories, **Keep the Flow** emphasizes a workflow across Single-row (S) and Multi-row (M) setups, as illustrated in Fig. 1. First, *precise inter-object spacing* imposes uniform-spacing placement, requiring the system to manage relative distances between items. Second, *goal-directed obstacle rearrangement* in M-setups involves temporarily relocating front-row obstacles to place targets in the back.

To systematically scale the volume and diversity of the collected data under this task definition, we parameterize the configurations across multiple axes:

- **Object Diversity:** The target object set $\mathcal{O}$ encompasses 3 distinct geometric shapes (Box, Cylinder, Hexagon) across 4 colors (Red, Orange, Blue, Yellow).
- **Spatial Variations:** Spatial variations are rigorously regulated by modulating the target inter-object distance $d \in \{15, 20, 25\}$ cm within the grid layout.
- **Grid Formation:** Each data collection session requires transferring $N \in \{1, 2, 3\}$ objects from a cart to designated slots in a 1×3 or 2×3 grid formation.
- **Arrangement Logic:** Objects within the same row share the same geometry but have different colors.

2.3. Dataset Processing & Properties

Data Augmentation: To prevent overfitting to visual artifacts, we first generate grid-free images by inpainting grid lines segmented in HSV color space. We then segment target objects using SAM2 [25] and apply 3 distinct mask-based color perturbations. By combining the two environmental views (grid-present and grid-free) with the four object appearance states (one original and three color-swapped), we expand the physical dataset volume 8-fold. This encourages policies to rely on physically relevant object features rather than grid-specific visual artifacts or specific color biases.



(a) Verification Setup (b) Resolving Parallax (c) Automated Validation

Figure 2. **Overview of the Automated Verification Loop.** The system utilizes an overhead view and depth data to resolve parallax and filter out spatially inaccurate demonstrations.

Verification Loop: The verification setup integrates an external top-down camera and a 1 cm grid sticker, which defines the target coordinates $\mathbf{x}_{\text{target}}$ (Fig. 2a). Through this overhead view, the system detects the object's top center $\mathbf{x}_{\text{top}}$ via image moments [27] and resolves parallax using depth data to compute the true base center $\mathbf{x}_{\text{base}}$ (Fig. 2b). Samples with placing error $e = \|\mathbf{x}_{\text{base}} - \mathbf{x}_{\text{target}}\| > \tau = 2$ cm are automatically discarded (Fig. 2c).

Data Distribution. The dataset comprises 1.35k original physical trajectories. To accurately reflect the sequential skills the policy must learn, these trajectories are decomposed into 1,723 verified skill instances (980 from the Single-row task and 743 from the Multi-row task). Finally, applying the aforementioned visual augmentation expands these original physical trajectories into eight visual variants. This scales the final dataset to 10.81k trajectories and approximately 7.79M frames, ensuring extensive visual diversity for robust policy training while preserving the original action-state synchronization.

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] Z. Zeng, Z. Zhou, Z. Sui, and O. C. Jenkins, "Semantic robot programming for goal-directed manipulation in cluttered scenes," in *IEEE international conference on robotics and automation (ICRA)*, 2018. 1

[2] T. Motoda, D. Petit, T. Nishi, K. Nagata, W. Wan, and K. Harada, "Shelf replenishment based on object arrangement detection and collapse prediction for bimanual manipulation," *Robotics*, 2022. 1

[3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 1

[4] J. Aldaco, T. Armstrong, R. Baruch, J. Bingham, S. Chan, K. Draper, D. Dwibedi, C. Finn, P. Florence, S. Goodrich, *et al.*, "Aloha 2: An enhanced low-cost hardware for bimanual teleoperation," *arXiv preprint arXiv:2405.02292*, 2024.

[5] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *Conference on Robot Learning (CoRL)*, 2024.

[6] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[7] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," in *Conference on Robot Learning (CoRL)*, 2024.

[8] K. Shaw, Y. Li, J. Yang, M. K. Srirama, R. Liu, H. Xiong, R. Mendonca, and D. Pathak, "Bimanual dexterity for complex tasks," in *Conference on Robot Learning (CoRL)*, 2024.

[9] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," in *Robotics: Science and Systems (RSS)*, 2023. 1, 2

[10] O. X.-E. Collaboration, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *IEEE international conference on robotics and automation (ICRA)*, 2024. 1

[11] D. D. Team, "Droid: A large-scale in-the-wild robot manipulation dataset," in *Robotics: Science and Systems (RSS)*, 2024.

[12] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning (CoRL)*, 2023.

[13] Fourier ActionNet Team and Mu, Yao, "ActionNet: A dataset for dexterous bimanual manipulation." https://action-net.org/, 2025. 1

[14] Z. Zhao, H. Jing, X. Liu, J. Mao, A. Jha, H. Yang, R. Xue, S. Zakharor, V. Guizilini, and Y. Wang, "Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation," 2025. 1

[15] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, *et al.*, "Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation," in *Robotics: Science and Systems (RSS)*, 2025. 1

[16] A. W. Team, "Agibot world 2026." https://huggingface.co/datasets/agibot-world/AgiBotWorld2026, 2026. 1

[17] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning (CoRL)*, 2023. 1

[18] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," in *Conference on Robot Learning (CoRL)*, 2024.

[19] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *Robotics: Science and Systems (RSS)*, 2024.

[20] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π_0: A vision-language-action flow model for general robot control," in *Robotics: Science and Systems (RSS)*, 2025. 1

[21] R. Gong, J. Huang, Y. Zhao, H. Geng, X. Gao, Q. Wu, W. Ai, Z. Zhou, D. Terzopoulos, S.-C. 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 (ICCV)*, 2023. 1

[22] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "Spatialvlm: Endowing vision-language mod-

els with spatial reasoning capabilities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[23] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," in *Robotics: Science and Systems (RSS)*, 2025.

[24] C. H. Song, V. Blukis, J. Tremblay, S. Tyree, Y. Su, and S. Birchfield, "RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1

[25] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," in *International Conference on Learning Representations (ICLR)*, 2025. 1, 2

[26] J. Carpentier and N. Mansard, "Analytical derivatives of rigid body dynamics algorithms," in *Robotics: Science and Systems (RSS)*, 2018. 1

[27] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE transactions on information theory*, 1962. 2