Imitation Learning–Based Physical AI for Small-Scale Construction Robotics Tasks

Abdullah Rasul[1] and Daeho Kim[1]

[1]Department of Civil & Mineral Engineering, University of Toronto, Canada
abdullah.rasul@mail.utoronto.ca, civdaeho.kim@utoronto.ca

Abstract –

Robotic automation in construction remains limited by the complexity of physically interactive tasks and the variability of real job sites. This study presents a hardware-validated proof-of-concept for imitation learning–based Physical AI on a small-scale excavation task. A modified SO-101 robotic arm was equipped with a 3D-printed excavation bucket, and 50 expert demonstrations were collected using a leader–follower teleoperation setup with multi-camera observations. Two policy models, ACT and SmolVLA, were trained to map joint and visual states to follower arm actions. Experimental results demonstrate successful task execution, with ACT producing smoother trajectories and faster completion than SmolVLA, validating the feasibility of learning physically interactive behaviors from limited data. The framework supports extensions to richer sensory modalities, advanced learning architectures, and scalable humanoid data collection, providing a practical foundation for trade-specific construction automation.

Keywords –

Construction Robotics; Imitation Learning; Physical AI; Robotic Excavation; Humanoid Skill Acquisition

1 Introduction

The construction industry can benefit substantially from robotic automation in productivity, quality, and safety. However, adoption of advanced autonomous systems remains limited compared to other engineering sectors. Existing robots for niche tasks, such as bricklaying or site inspection, are typically pre-programmed or operate under controlled conditions, limiting their ability to handle the variability of real job sites [1]. As a result, current systems lack the adaptive, nuanced behaviors of human experts, particularly in tasks requiring force control, trajectory adaptation, and sequential decision-making. While the long-term goal is to enable humanoid robots to assist in high-risk or dangerous construction tasks, progress is expected trade by trade, with foundational skills first demonstrated on representative tasks.

Reviews of construction robotics research show that learning-based methods, including reinforcement and imitation learning, remain underexplored, with few systems experimentally validated on physically interactive tasks representative of real trades [2]. Unlike simulation-heavy domains, construction requires physical demonstrations on real hardware to handle diverse materials, terrain, and workflows. To address these challenges, frameworks integrating perception, reasoning, and control are needed to enable adaptive, learning-driven systems capable of operating in unstructured environments. Physical AI provides such a framework, supporting embodied intelligence, dexterous manipulation, and human–robot collaboration [3]. In construction, this approach improves robustness, generalization, and efficiency when interacting with materials and equipment.

Leveraging Physical AI's capabilities for perception, reasoning, and adaptive control, imitation learning enables robots to acquire complex task behaviors by observing and replicating expert demonstrations [4]. Unlike reinforcement learning, which relies on trial-and-error, imitation learning encodes task-specific strategies directly from skilled operators, which are essential for handling the variability and physical demands of real construction sites [5]. Despite its potential, imitation learning remains largely untested in construction robotics, with few validated systems deployed on real hardware performing representative tasks [2].

To address this gap, this study demonstrates a

practical, hardware-validated proof-of-concept for imitation learning–based Physical AI, using a small-scale excavation task as a representative construction trade. Unlike prior demonstrations that focus primarily on simulated environments or assembly-style manipulation tasks, this study evaluates imitation learning on a physically interactive excavation activity involving contact with granular material and sequential manipulation, which introduces challenges in perception, control, and task reproducibility not addressed in prior small-scale manipulation work.

The primary contribution is a systems-level framework comprising a structured demonstration dataset, a multi-camera observation pipeline, and a hardware validation protocol for imitation learning in construction-related manipulation tasks. A standard robotic arm was modified with a custom end effector, and approximately 50 expert demonstrations of digging and dumping were collected to train models. Excavation is used as a representative task toward broader construction automation. The experiments evaluate feasibility of learning from expert demonstrations, measure task success rates, identify implementation challenges, and provide insights for scaling to more complex construction tasks and platforms.

By combining Physical AI with imitation learning and expert demonstration, this study presents a practical pathway for construction robots to acquire dexterous and adaptive behaviors, bridging theoretical learning methods and real-world applications while laying the groundwork for future trade-specific automation and humanoid deployment.

2 Background and Related Work

Robotic automation in construction can benefit from learning-driven methods that enable adaptive task execution. Reinforcement learning (RL) and imitation learning (IL), supported by Physical AI, provide frameworks for acquiring complex behaviors from interaction and demonstration [3–6].

Reinforcement learning enables robots to optimize actions through environmental feedback. Early applications in construction include hierarchical RL and vision-language action models for multi-stage installation tasks, improving learning efficiency and policy generalization [7]. Multi-agent RL frameworks have been explored to coordinate multiple robots across sequential workflows, enhancing collaborative task execution [8,9]. Curriculum-based RL approaches support stepwise multi-task learning in simulated construction environments [10]. Despite these advances, RL in real-world construction remains limited by safety constraints, high data requirements, and challenges in transferring policies from simulation to hardware [7]. These limitations suggest that RL alone may not be sufficient for adaptive construction automation.

Imitation learning offers a practical alternative by training robots directly from expert demonstrations, bypassing trial-and-error learning. Although IL in construction is still emerging, studies have demonstrated its potential in simulation and virtual environments. Generative adversarial IL and VR-based demonstrations have been applied to long-horizon collaborative tasks, outperforming RL baselines, reducing reliance on hand-crafted reward functions, and enabling hybrid IL–RL training [11]. Other work integrates IL with environmental rewards to account for demonstration quality and improve performance [12].

Recent studies have also explored IL in small-scale tasks. For example, mobile robotic rebar cage assembly has used visual servoing combined with expert demonstrations to perform autonomous insertion and tying without explicit geometric modeling [13]. Force-integrated IL frameworks capture expert force feedback to guide motion in assembly tasks, improving precision and task success [14]. Hierarchical IL approaches decompose expert skills into sub-skills, enabling scalable skill transfer and reusable demonstration libraries [15]. These studies illustrate the potential of IL on physical platforms, but implementations remain largely at proof-of-concept scale.

Despite these advances, IL in construction remains largely limited to simulation or small-scale physical platforms, leaving a gap in understanding its feasibility, performance, and adaptability for representative construction activities. This motivates experimental demonstrations on accessible robotic platforms to evaluate skill acquisition, task execution, and adaptability in controlled scenarios.

3 Experimental Setup and Methods

3.1 Hardware Setup

The experimental platform is built around an SO-101

robotic arm [16], an open-source articulated manipulator for robotics research. The SO-101 provides motor encoder readings for joint positions, and motion states are recorded using these encoders combined with visual observations. As shown in Figure 1, the original gripper was replaced with a custom 3D-printed excavation bucket designed to approximate a typical digging tool while remaining within the arm's payload and kinematic limits.

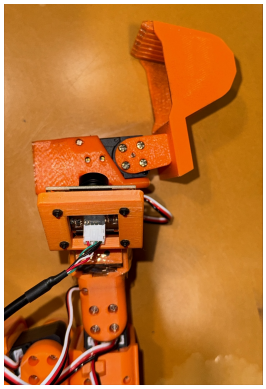

Figure 1. Modified SO-101 robotic arm with 3D-printed excavation bucket.

To capture state information for imitation learning, the setup includes three cameras: a wrist-mounted camera integrated with the SO-101 kit to record bucket–material interaction, an Intel RealSense camera providing a side view of the workspace, and a Logitech webcam mounted above for a top-down perspective. These visual streams, combined with joint encoder readings, define the robot's state representation. Expert demonstrations were collected using a leader–follower teleoperation setup, where the operator moves the leader arm in the air while the follower executes the dig-and-dump task, recording joint positions and camera observations. The teleoperation and data collection pipeline were implemented using the LeRobot library. The workspace consists of a contained area with two bowls for digging and dumping material, providing a controlled environment for the excavation task, as illustrated in Figure 2.

3.2 Task Design and Data Collection

The dig-and-dump task was selected as a representative construction activity for evaluating imitation learning on a physical robotic platform. Each episode consisted of the robot starting from its initial pose, digging material from the source bowl, transferring it, dumping into the target bowl, and returning to its original position. A total of 50 expert demonstrations were collected, with episodes repeated if motion was not smooth or dumping was inaccurate.

The 50 expert demonstrations collectively comprise 34,297 frames captured at 30 fps, with each frame containing joint positions for six degrees of freedom and synchronized images from three cameras (top, side, front). Each episode represents one complete dig-and-dump task, providing sufficient coverage of motion variability for training and evaluation.

Each demonstration targeted a full bucket excavation, with minor variations in digging motion due to operator behavior. Two bowls were used in the workspace, and their positions were slightly adjusted between episodes to introduce controlled variability while ensuring sufficient material depth for consistent dig-and-dump tasks.

During test executions, lighting and material distribution were kept consistent to ensure reproducible evaluation, as both ACT and SmolVLA are sensitive to visual changes. Small differences in initial arm pose and natural trajectory deviations provided additional variability, balancing realism in demonstrations with controlled repeatability in testing. Broader robustness experiments involving larger variations in workspace configuration, material quantity, or lighting were outside the scope of this study and are left for future work.

Figure 2. Experimental setup showing the robotic arm, bowls, and multi-camera configuration for data capture.

The workspace used black soil for visual contrast and was well-lit for observation by three cameras: the wrist-mounted camera captured bucket–material interaction, the Intel RealSense camera provided a side view, and the Logitech webcam recorded a top-down perspective. These visual streams, combined with joint encoder

readings from the SO-101 arm, formed the state representation for imitation learning, while actions corresponded to the follower arm's joint commands. The operator manipulated the leader arm as a guide while the follower executed the physical dig-and-dump task.

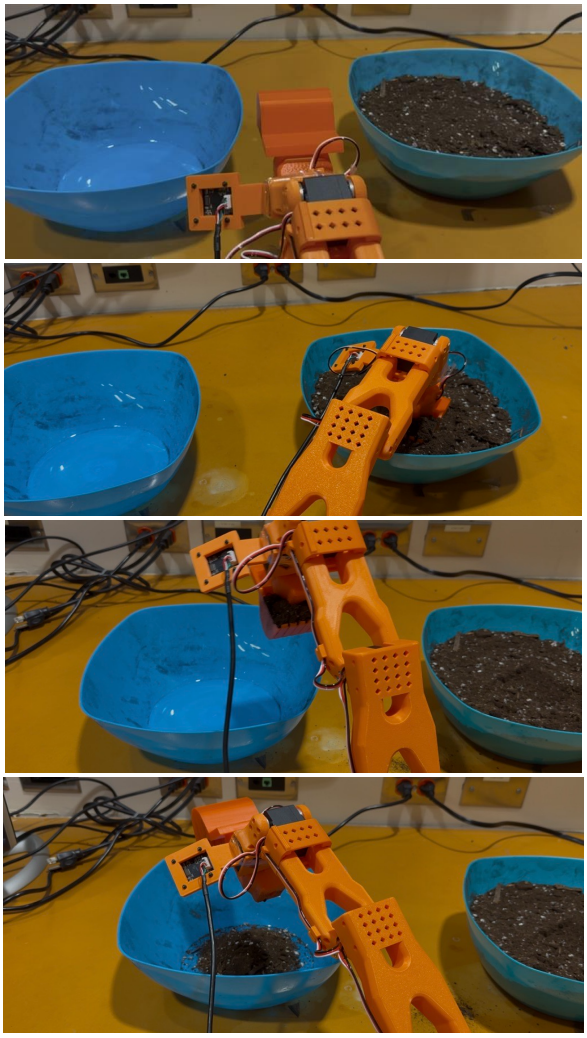

Figure 3. Four-stage sequence of a single dig-and-dump episode: (a) robot at starting pose, (b) digging from source bowl, (c) transferring material, and (d) dumping into target bowl.

Data were recorded using the LeRobot record pipeline, which synchronizes multi-modal observations during each episode and stores them in a standardized format. High-frequency robot states are tabulated, visual streams stored as videos, and metadata describes schema, episode segmentation, and configuration parameters. Temporal alignment between sensor modalities is maintained, and episodes with interruptions or motion artifacts were excluded. LeRobot tools allow refinement, splitting, merging, or conversion of dataset features and efficient video encoding, supporting scalable learning pipelines and analysis workflows.

To illustrate the task sequence, a four-panel image captures a complete episode, showing the robot starting from its initial pose, digging from the source bowl, transferring material, and dumping into the target bowl, as illustrated in Figure 3. This visualization highlights both the motion trajectory and interaction with the environment during a representative demonstration.

3.3 Learning Framework

Imitation learning was employed to enable the robot to acquire the dig-and-dump skill directly from expert demonstrations, providing a proof-of-concept baseline for learning physically interactive behaviors on a small-scale robotic platform. This approach bypasses trial-and-error exploration and avoids the need for hand-crafted reward functions, which is critical for safely learning manipulation in unstructured construction tasks.

Two policy models were selected to demonstrate feasibility rather than benchmark performance: Action Chunking with Transformers (ACT) [17] and SmolVLA [18]. ACT was chosen for its ability to model sequential dependencies in joint actions and to capture temporal correlations in expert trajectories using a transformer-based architecture. SmolVLA was selected as a compact vision-language-action model that integrates visual perception and action prediction efficiently, enabling multi-modal learning from limited demonstration data.

Inputs to both models included joint encoder readings and synchronized RGB frames from the wrist-mounted, side-view, and top-down cameras. Outputs were the follower arm joint commands, providing a direct mapping from observed expert states to executed actions.

Language instructions were included as a fixed task description for all episodes (e.g., "dig from the source bowl and dump into the target bowl") to maintain consistent dataset annotations. Multi-modal policies, including SmolVLA, rely primarily on visual observations and joint states to generate actions, while the constant language input serves only as a uniform contextual reference and does not influence policy behavior or artificially bias performance metrics. This design ensures that the model can handle multi-modal inputs without confounding the evaluation.

Training parameters were configured as follows:

ACT was trained for 100,000 epochs with a batch size of 32 and a fixed learning rate of 0.00001. SmolVLA was trained for 20,000 epochs with a batch size of 32 and an adaptive learning rate varying from 0.00001 to 0.0000025. Early stopping was applied based on validation loss to prevent overfitting.

Table 1 Summary of the training configurations

Model	Epochs	Batch Size	Learning Rate	Architecture
ACT	100,000	32	0.00001	Transformer-based action chunking
SmolVLA	20,000	32	Adaptive: 0.00001 →0.0000025	Vision-language-action integration

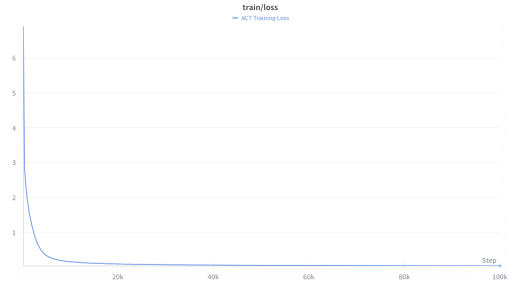

Figure 4. ACT training loss

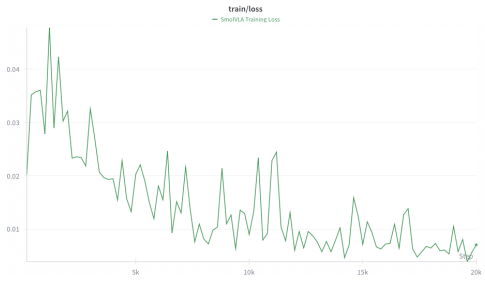

Figure 5. SmolVLA training loss

Training loss was monitored to assess convergence. ACT exhibited a smooth and consistently decreasing loss, indicating stable learning of sequential actions. In contrast, SmolVLA displayed fluctuating loss values, reflecting the model's adaptive learning rate and sensitivity to multi-modal inputs. Figure 4 shows ACT training loss, and Figure 5 shows SmolVLA training loss over the respective epochs.

4 Experiments and Results

The trained policies were deployed on the SO-101 robotic arm with the custom excavation bucket and evaluated in the same workspace used during data collection. Each model was tested over 20 independent executions of the dig-and-dump task to assess feasibility under real hardware conditions.

While the primary evaluation compares ACT and SmolVLA policies, we acknowledge that minimal baselines such as direct teleoperation playback, simple scripted trajectories, or single-view models could provide additional context. In this proof-of-concept study, we focused on comparing two representative imitation learning architectures to demonstrate feasibility. Inclusion of minimal baselines is suggested as future work to further characterize performance differences under controlled variations.

4.1 Evaluation Criteria

Performance was evaluated using the following criteria:

1. **Task success rate,** defined as successful transfer of material from the source bowl to the target bowl followed by a return to the initial pose. No strict numerical threshold was applied because precise quantification is challenging due to the small bucket size and contact-rich interaction. Minor incidental material displacement due to small trajectory deviations or control noise was tolerated, whereas significant loss of material was considered a failure.
2. **Execution time,** measured from task initiation to completion.
3. **Motion smoothness,** assessed qualitatively by observing trajectory continuity and presence of oscillatory or corrective motions.

These metrics were selected to reflect practical execution quality for manipulation rather than optimality. All 50 expert demonstrations were used for training due to the limited dataset size. In small-scale hardware-based imitation learning, policies are typically evaluated through independent real-world executions rather than a held-out dataset. Policy performance was therefore assessed over 20 independent dig-and-dump executions

on the physical robot, measuring task success rate, execution time, and motion characteristics.

ACT achieved a higher task success rate and shorter execution time, producing trajectories that are visually consistent with expert demonstrations. Motion execution is generally continuous, with limited corrective behavior during digging and dumping.

4.2 Results

Table 2 Evaluation results for both models.

Model	Task Success Rate (%)	Avg. Execution Time (s)	Motion Smoothness
ACT	85	20	Consistent
SmolVLA	70	35	Variable

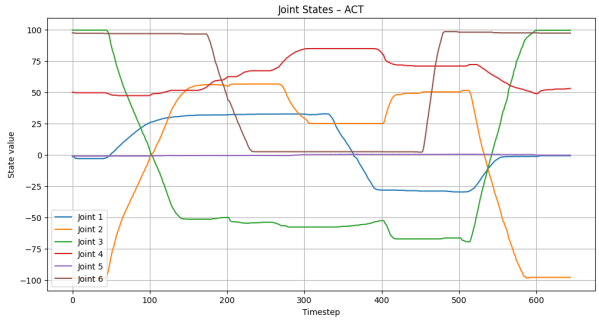

Figure 6. Joint motor trajectories - ACT.

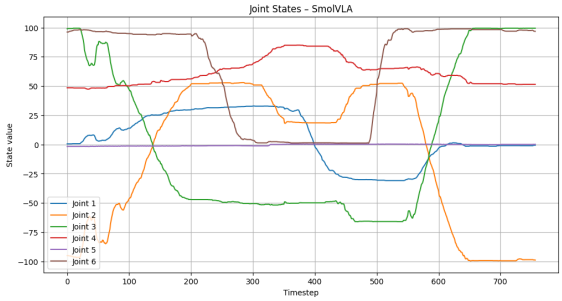

Figure 7. Joint motor trajectories – SmolVLA.

SmolVLA completed the task in most trials but exhibited greater variability in execution. The longer average execution time was primarily due to intermittent corrective motions, particularly during material transfer. This behavior is consistent with the fluctuating training loss observed during learning, indicating sensitivity to multi-modal inputs and adaptive learning dynamics.

Representative executions for both policies are shown in Figure 6 and 7, illustrating differences in motion continuity and task timing.

Overall, both models successfully demonstrated the feasibility of imitation learning for small-scale excavation on real hardware, with differences in performance reflecting architectural and training characteristics rather than task failure.

5 Discussion and Limitations

5.1 Feasibility of Imitation Learning for Construction Manipulation Tasks

The experimental results demonstrate that imitation learning can be successfully applied to a physically interactive excavation task on real hardware using a limited number of expert demonstrations and an accessible robotic platform. Both ACT and SmolVLA reproduce the dig-and-dump behavior without explicit modeling of contact dynamics or task constraints, confirming the feasibility of learning manipulation skills directly from demonstrations in construction-relevant settings.

The observed differences between the two models highlight trade-offs in policy design. ACT produces more consistent trajectories and smoother execution, reflecting its temporal modeling of action sequences. SmolVLA exhibits greater variability, likely due to its compact architecture and adaptive learning dynamics when integrating multi-modal inputs. Importantly, both policies remain task-complete in most trials, reinforcing that the primary contribution is demonstrating feasibility rather than optimizing performance.

Observed failures are generally minor and model dependent. ACT occasionally requires small corrective adjustments when transferring material near bowl boundaries, whereas SmolVLA shows higher trajectory variability with intermittent pauses or misalignment during digging and dumping. These patterns provide practical insight into policy limitations and inform potential improvements in training and control.

5.2 Limitations of the Current Setup

Several limitations of the current system should be noted. First, the SO-101 arm provides only joint encoder feedback and lacks direct force or torque sensing, so

contact interactions with soil are inferred indirectly from motion and visual observations, limiting precise control of digging forces.

Second, evaluation metrics are limited to task completion, execution time, and qualitative motion characteristics. While sufficient for a proof-of-concept, more comprehensive metrics such as force efficiency, transferred material volume, spillage rate, or repeatability would be required for scaled deployment; these were not measured in the present setup but represent important directions for future work.

Finally, the task environment is intentionally constrained, with limited bowl position variability and consistent material properties, limiting assessment of generalization across workspace layouts and soil conditions.

5.3 Extending Visual and State Representation

Future improvements can leverage richer visual representations and expanded sensing modalities. In the visual domain, object detection or segmentation could enable explicit reasoning about tool–material interaction regions, bowl boundaries, and material flow, improving spatial awareness and reducing reliance on implicit features.

Beyond vision, additional sensing modalities could further enhance performance, including force–torque sensing at the wrist, tactile sensing on the bucket, or depth sensing for estimating material volume. These provide complementary state information particularly valuable for contact-rich construction tasks.

5.4 Advanced Learning Methods and Policy Generalization

While this study focuses on behavioral cloning–based imitation learning, more advanced methods could improve robustness and generalization. Approaches such as Generative Adversarial Imitation Learning (GAIL) or Inverse Reinforcement Learning (IRL) could enable policies to infer underlying task objectives rather than strictly replicate demonstrated trajectories. Hybrid imitation learning and reinforcement learning approaches may further refine behaviors under physical interaction constraints. Additionally, hierarchical policy structures could decompose complex construction tasks into reusable sub-skills, supporting operations scability.

5.5 Implications for Scaling and Humanoid Data Collection

Although the experiments were conducted on a small-scale robotic arm, the methodology informs scalable data collection for larger platforms including construction machinery and humanoid robots. The leader–follower paradigm and multi-camera setup provide a template for structured, time-aligned demonstrations across embodiments, which can be complemented by motion capture systems to record skilled workers performing trade-specific tasks. By validating imitation learning on accessible hardware, this work demonstrates how teleoperation and multi-view observation can be applied to contact-rich construction tasks, providing a foundation for trade-specific skill acquisition and scalable learning pipelines across richer sensory modalities and policy architectures.

While the experiments are limited to a small-scale robotic arm, the evaluated models, ACT and SmolVLA, are general-purpose robotic policies that successfully learned contact-rich excavation behavior. This suggests that such models can be adapted to specialized construction tasks. By extension, general-purpose humanoid control systems may also be applicable to construction-relevant tasks, provided appropriate sensing and actuation, and with necessary adaptation for differences in kinematics, dynamics, and workspace scale.

6 Conclusion

This study presents a hardware-validated proof-of-concept for imitation learning–based Physical AI in construction, demonstrated on a small-scale excavation task using a modified SO-101 robotic arm. By collecting structured expert demonstrations via a leader–follower teleoperation setup and multi-camera observation, we show that both ACT and SmolVLA policies can successfully reproduce task behaviors, demonstrating the feasibility of learning physically interactive skills from a limited dataset.

The results highlight differences in trajectory stability and execution time between models, illustrating how architecture and learning dynamics influence policy performance. The framework provides a practical foundation for extending learning pipelines to richer visual and sensory modalities, advanced policy architectures, and more complex construction tasks.

Finally, the methodology establishes a pathway toward humanoid-oriented data collection, including motion capture of skilled workers, enabling scalable acquisition of trade-specific skills and laying the groundwork for future humanoid-assisted construction automation.

References

[1] J.M. Davila Delgado, L. Oyedele, A. Ajayi, L. Akanbi, O. Akinade, M. Bilal, and H. Owolabi. Robotics and automated systems in construction: Understanding industry-specific challenges for adoption. Journal of Building Engineering, 26:100868, 2019.

[2] J.M. Davila Delgado and L. Oyedele. Robotics in construction: A critical review of the reinforcement learning and imitation learning paradigms. Advanced Engineering Informatics, 54:101787, 2022.

[3] H. Song, L. Wang, X. Qiao, Y. Chen, D. Sun, and Z. Sun. Embodied intelligence for robot manipulation: development and challenges. Vicinagearth, 2:8, 2025.

[4] P. Abbeel and A.Y. Ng. Apprenticeship learning via inverse reinforcement learning. In Proceedings of the International Conference on Machine Learning (ICML), pages 1–8, 2004.

[5] B.D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. Robotics and Autonomous Systems, 57(5):469–483, 2009.

[6] Z. Hu, H. Yu, V. Chandramouli, and C. Liang. Sample-efficient robot skill learning for construction tasks: Benchmarking hierarchical reinforcement learning and vision-language-action VLA model. arXiv preprint arXiv:2512.14031, 2025.

[7] X. Xu and B. García de Soto. Reinforcement learning with construction robots: A review of research areas, challenges and opportunities. In Proceedings of the 39th International Symposium on Automation and Robotics in Construction (ISARC), Bogotá, Colombia, pages 375–382, 2022.

[8] K. Duan and Z. Zou. MARC: Multi-agent robot control framework for enhancing reinforcement learning in construction tasks. arXiv preprint arXiv:2305.14586, 2023.

[9] W. Cai and Z. Zou. A reinforcement learning-based approach for conducting multiple tasks using robots in virtual construction environments. In Proceedings of ICRA 2022 Workshop on Future of Construction Robotics, 2022.

[10] L. Huang and Z. Zou. Deep reinforcement learning-based construction robot collaboration for sequential tasks. In Proceedings of ICRA 2022 Workshop on Future of Construction Robotics, 2022.

[11] R. Li and Z. Zou. Enhancing construction robot learning for collaborative and long-horizon tasks using generative adversarial imitation learning. Advanced Engineering Informatics, 58:102140, 2023.

[12] K. Duan, Z. Zou, and T.-Y. Yang. Training of construction robots using imitation learning and environmental rewards. Computer-Aided Civil and Infrastructure Engineering, 40(9):1150–1165, 2024.

[13] T. Sun, B. Han, J. Wu, S. Rusinkiewicz, and Y. Shao. Mobile robotic rebar cage assembly via imitation learning. Automation in Construction, 181:106671, 2026.

[14] H. You, Y. Ye, T. Zhou, and J. Du. Force-based robotic imitation learning: A two-phase approach for construction assembly tasks. arXiv preprint arXiv:2501.14942, 2025.

[15] H. Yu, V.R. Kamat, and C.C. Menassa. Cloud-based hierarchical imitation learning for scalable transfer of construction skills from human workers to assisting robots. arXiv preprint arXiv:2309.11619, 2023.

[16] SO-101 Robotic Arm. On-line: https://www.sobotrobot.com/so101, Accessed: 22/01/2026.

[17] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. On-line: https://arxiv.org/abs/2304.13705, Accessed: 22/01/2026.

[18] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, S. Alibert, M. Cord, T. Wolf, and R. Cadene. SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics. On-line: https://arxiv.org/abs/2506.01844, Accessed: 22/01/2026.