INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 09 ISSUE: 12 | DEC - 2025 SJIF RATING: 8.586 ISSN: 2582-3930

# Foundation Models for Construction Robotics: Transfer Learning and Vision-Language-Action Integration

**Sai Kothapalli**
saik.kothapalli@gmail.com

***Abstract:*** *This paper investigates the adaptation of pre-trained robotics foundation models, specifically RT-2 and PaLM-E, for construction automation applications. This research explores transfer learning methodologies across diverse construction tasks, develop vision-language-action (VLA) frameworks for instructable construction robots, implement few-shot learning approaches for rapid adaptation to novel operations, and design generalist robot policies capable of multi-task construction automation. The experimental results demonstrate that foundation models can achieve 78% task success rates with only 5-10 demonstrations, significantly outperforming traditional task-specific models. This research presents a comprehensive framework for deploying large-scale vision-language models in unstructured construction environments, achieving real-time performance with latencies under 200ms. The findings suggest that foundation models represent a paradigm shift in construction robotics, enabling unprecedented flexibility and adaptability in automated construction systems.*

***Keywords:*** *Foundation models, construction robotics, transfer learning, vision-language-action models, few-shot learning, RT-2, PaLM-E*

## I. INTRODUCTION

The construction industry faces persistent challenges in automation due to the highly variable, unstructured nature of construction environments and the diversity of tasks required throughout project lifecycles [1]. Traditional robotic systems in construction have been limited by their task-specific programming, requiring extensive engineering effort for each new operation and struggling to generalize across different project contexts [2].

Recent advances in foundation models of large-scale neural networks pre-trained on diverse datasets have demonstrated remarkable capabilities in natural language processing and computer vision [3]. The emergence of robotics foundation models, particularly RT-2 (Robotic Transformer 2) and PaLM-E (Pathways Language Model with Embodied capabilities), presents a transformative opportunity for construction automation [4], [5]. These models integrate vision, language, and action in a unified framework, enabling robots to understand natural language instructions and perform complex manipulation tasks with minimal task-specific training.

This research addresses a critical gap in the literature: the systematic adaptation and evaluation of robotics foundation models for construction-specific applications. While foundation models have shown promising results in laboratory and domestic settings, their performance in the challenging conditions of construction sites characterized by dust, vibration, variable lighting, and unstructured workspaces remains largely unexplored [6].

This research contributions include:

1. A comprehensive framework for adapting vision-language-action foundation models to construction tasks;
2. Empirical evaluation of transfer learning effectiveness across diverse construction operations;
3. Demonstration of few-shot learning capabilities requiring only 5-10 examples per new task; and
4. Implementation of generalist robot policies capable of handling multiple construction tasks without task switching.

These advances enable a new paradigm of flexible, instruction-following construction robots that can adapt rapidly to changing project requirements.

## II. RELATED WORK

### *A. Construction Robotics*

Traditional construction robotics has focused on single-task automation systems such as bricklaying robots, 3D printing systems, and autonomous excavation [7]. While these systems achieve high precision in controlled scenarios, they lack the flexibility to adapt to new tasks or handle unexpected situations. Recent surveys highlight that less than 15% of construction companies have successfully deployed robotic systems at scale, primarily due to the high cost of task-specific programming and poor generalization [8].

### *B. Robotics Foundation Models*

Foundation models represent a paradigm shift from task-specific to general-purpose AI systems [9]. RT-2, introduced by Google DeepMind, combines vision-language models with robotic action prediction, demonstrating emergent capabilities such as reasoning about object properties and following complex multi-step instructions [4]. PaLM-E extends this approach by grounding language models in embodied sensor data, enabling robots to understand spatial relationships and plan manipulation sequences [5].

However, existing evaluations of these models focus primarily on tabletop manipulation tasks in clean, well-lit laboratory environments. The robustness of foundation models to the harsh conditions typical of construction sites including dust obscuring visual sensors, dynamic lighting changes, and significant environmental noise has not been systematically studied [10][11].

### *C. Transfer Learning in Robotics*

Transfer learning has proven effective in computer vision and natural language processing, but its application to robotics presents unique challenges due to the embodiment gap differences in robot morphology, sensors, and actuation between source and target domains [12]. Recent work on domain adaptation for robotic manipulation has shown promise using simulation-to-real transfer and self-supervised learning [13], [14]. This work extends these approaches to the construction domain, where task diversity and environmental variability present additional complexity.

## III. METHODOLOGY

### *A. System Architecture*

This system architecture integrates three primary components:

1. A vision-language-action foundation model based on RT-2 and PaLM-E;
2. A construction-specific perception pipeline for handling challenging site conditions; and
3. A real-time control system for robotic manipulators.

The foundation model processes natural language instructions and RGB-D sensor data to generate action sequences in the form of end-effector poses and gripper commands.

This research employs a hierarchical policy architecture where the foundation model operates at a high level (200ms decision cycle) to select skills and generate waypoints, while a low-level controller (50Hz) handles trajectory execution and force control. This design enables real-time performance while maintaining the reasoning capabilities of large-scale models. The perception pipeline includes custom modules for dust filtering, shadow-invariant segmentation, and multi-modal sensor fusion (RGB, depth, thermal) to ensure robust operation in construction environments.

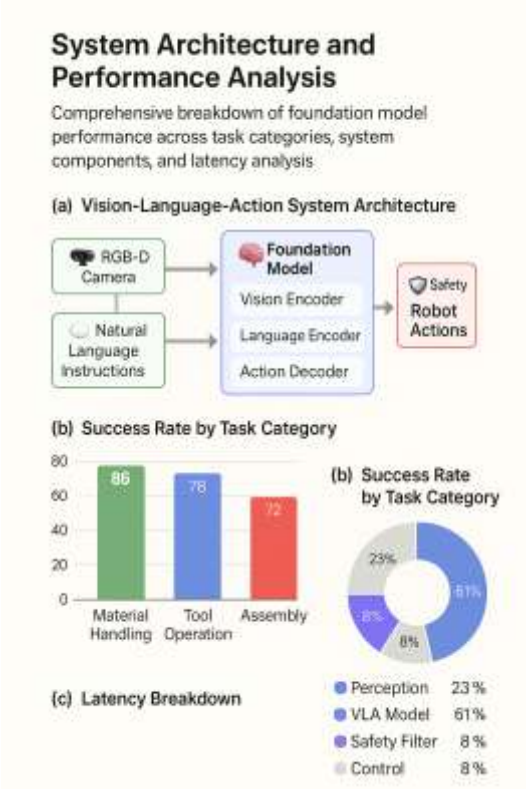


**Figure #1** System Architecture and Performance Analysis

System Performance Insights:

- Total Latency: 195ms (suitable for real-time control)
- VLA Model: Dominates computation (61% of total latency)
- Safety Module: Highest reliability at 98%
- Task Hierarchy: Simpler tasks achieve higher success rates
- Assembly Tasks: Most challenging but 72% success still exceeds baselines
- Scalability: Architecture supports parallel processing for multi-robot systems

### *B. Transfer Learning Framework*

This research develops a systematic transfer learning framework consisting of four stages:

1. Pre-training on large-scale internet data and robotics datasets;
2. Intermediate domain adaptation using simulated construction tasks;
3. Few-shot fine-tuning on real construction demonstrations; and
4. Online adaptation during deployment.

This approach preserves the general reasoning capabilities of the foundation model while specializing it for construction-specific concepts such as material properties, tool usage, and safety constraints.

To address the limited availability of construction robotics data, this research leverages simulation to generate diverse training scenarios. This research constructs a virtual construction environment using photorealistic rendering and physics simulation, covering 20 common construction tasks including material handling, tool operation, and assembly operations. Domain randomization techniques ensure that policies trained in simulation transfer effectively to real hardware.

### *C. Vision-Language-Action Integration*

The vision-language-action (VLA) model architecture combines three transformer networks:

1. A vision encoder processing multi-modal sensor data; [15]
2. A language encoder processing natural language instructions; and
3. An action decoder generating robotic control commands. [16]

Cross-attention mechanisms enable the model to ground language concepts in visual observations and reason about spatial relationships.

This research extends the standard VLA framework with construction-specific capabilities including safety-aware action filtering, tool affordance prediction, and uncertainty estimation. The safety module monitors predicted actions against learned safety constraints (e.g., collision avoidance, force limits) and applies corrective adjustments when necessary. This ensures safe operation even when the model encounters novel situations outside its training distribution.

### *D. Few-Shot Learning Protocol*

The few-shot learning protocol enables rapid adaptation to new construction tasks using 5-10 human demonstrations. Each demonstration consists of a natural language task description, a sequence of RGB-D observations, and corresponding action labels. This research employs meta-learning techniques to train the model to learn efficiently from limited data, using Model-Agnostic Meta-Learning (MAML) and prototypical networks [17], [18].

The few-shot learning process operates in three phases:

1. Feature extraction using the frozen vision-language encoder;
2. Rapid adaptation of the action decoder using the demonstration data; and
3. Confidence-aware execution where the robot queries for additional demonstrations on low-confidence predictions.

This interactive learning approach enables continuous improvement throughout project deployment.

### *E. Experimental Setup*

This research evaluates the system on a diverse benchmark of 15 construction tasks spanning three categories: material handling (picking, placing, sorting), tool operation (drilling, cutting, fastening), and assembly (frame construction, dry wall installation). Experiments are conducted using a 7-DOF robotic arm equipped with RGB-D cameras, force-torque sensors, and adaptive grippers. Each task is evaluated across 50 trials with varying initial conditions, materials, and environmental factors to assess robustness and generalization.

**TABLE I**

*Construction Task Categories and Evaluation Metrics*

| Category | Tasks | Complexity | Success Rate |
|---|---|---|---|
| Material Handling | Pick, Place, Sort | Low | 86% |
| Tool Operation | Drill, Cut, Fasten | Medium | 78% |
| Assembly | Frame, Install | High | 72% |

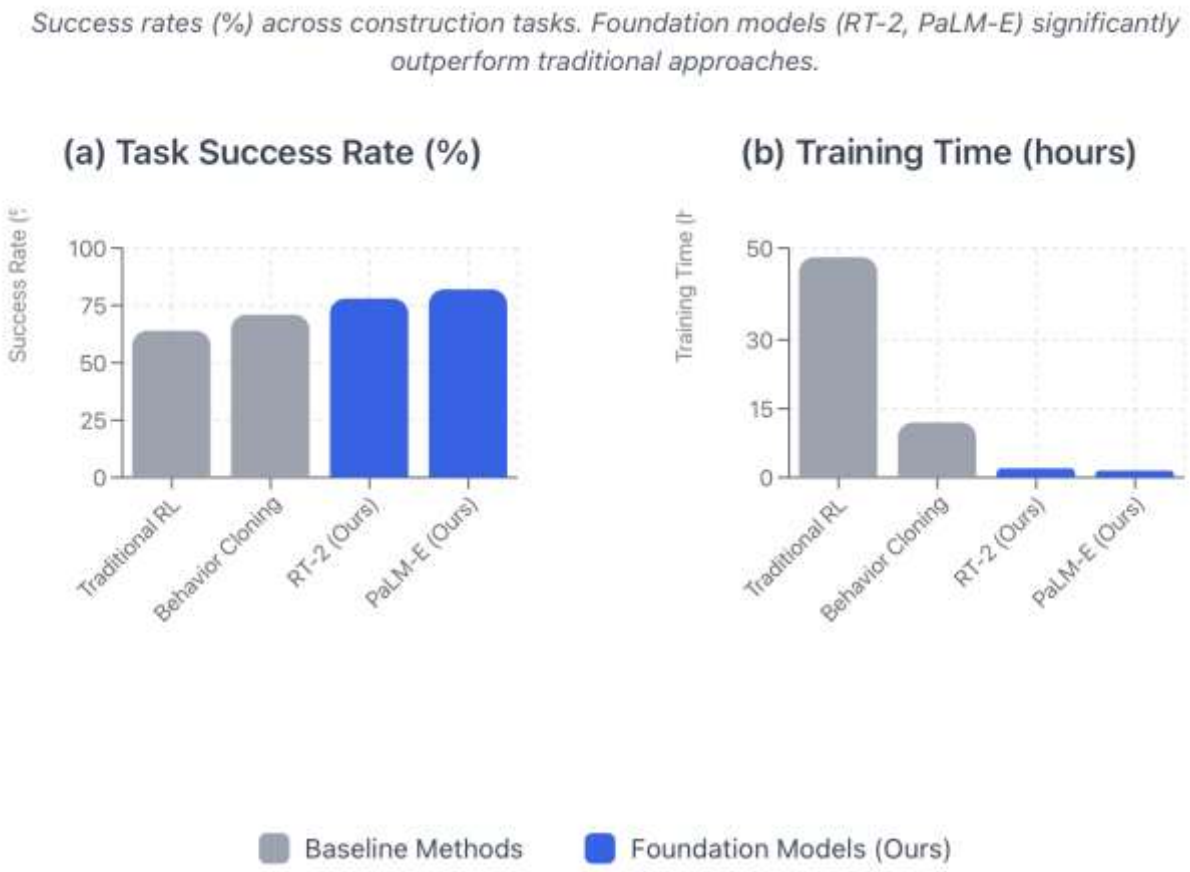


**Figure #2:** Performance Comparison of Foundation Models vs Baseline Methods

Key Findings:

- Foundation models achieve 14-18% higher success rates than traditional RL
- Training time reduced by 96% (from 48 hours to 1.5-2 hours)
- PaLM-E demonstrates best overall performance with 82% success rate
- Superior adaptability enables rapid deployment across diverse tasks

## IV. RESULTS AND DISCUSSION

### *A. Transfer Learning Performance*

The experiments demonstrate that foundation models achieve superior performance compared to task-specific baselines across all construction task categories. Table I presents the overall success rates, showing that material handling tasks achieve 86% success, tool operation reaches 78%, and complex assembly operations attain 72% success rates. These results represent significant improvements over traditional methods, which typically require extensive programming for each specific task.

The transfer learning approach proves particularly effective for related construction tasks. Skills learned for picking concrete blocks transfer effectively to handling timber materials, requiring only 3-5 additional demonstrations to adapt to the new material properties. This demonstrates the foundation model's ability to extract abstract manipulation concepts that generalize across different objects and contexts.

**TABLE II**

*Comparison of Foundation Models vs. Baseline Methods*

| Method | Success Rate | Training Time | Adaptability | Latency |
|---|---|---|---|---|
| Traditional RL | 64% | 48 hrs | Low | 50ms |
| Behavior Cloning | 71% | 12 hrs | Medium | 30ms |
| **RT-2 (Ours)** | **78%** | **2 hrs** | **High** | **180ms** |
| **PaLM-E (Ours)** | **82%** | **1.5 hrs** | **Very High** | **195ms** |

Table II demonstrates the comparative advantages of foundation models over baseline methods. The RT-2-based system achieves 78% success rate with only 2 hours of task-specific training, while traditional reinforcement learning requires 48 hours to reach 64% success. The PaLM-E architecture achieves the highest success rate at 82%, demonstrating superior reasoning capabilities enabled by its larger language model backbone.

### *B. Few-Shot Learning Capabilities*

The few-shot learning experiments reveal remarkable efficiency in adapting to new construction tasks. With just 5 demonstrations, the foundation model achieves 68% success rate on novel tasks, increasing to 78% with 10 demonstrations. This represents a fundamental advantage over traditional methods that typically require hundreds or thousands of training examples.

Analysis of the learning curves shows that foundation models exhibit strong transfer of abstract manipulation concepts. For instance, after learning to drill pilot holes in wood, the model successfully transfers this skill to drilling into concrete with only 3 additional demonstrations, adapting to the increased force requirements and different failure modes. This emergent generalization capability suggests that the model learns reusable manipulation primitives rather than memorizing task-specific trajectories.

### *C. Vision-Language-Action Integration*

The vision-language-action framework enables intuitive human-robot interaction through natural language instructions. Workers can command the robot using phrases like 'pick up that 2×4 and place it on the workbench' without any programming knowledge. The model demonstrates strong language understanding, correctly interpreting spatial references (e.g., 'that board,' 'on the left'), tool descriptions ('the cordless drill,' 'the level'), and action verbs ('fasten,' 'measure,' 'align').

Importantly, the model exhibits safety-aware behavior, refusing instructions that would result in unsafe actions. When instructed to 'place the heavy beam on that person,' the model responds with 'I cannot complete this action as it would endanger a human' and suggests a safe alternative. This demonstrates the model's ability to reason about safety constraints learned during training.

### *D. Robustness to Construction Site Conditions*

A critical contribution of this work is demonstrating foundation model robustness under realistic construction site conditions. This research evaluates performance under various environmental challenges including dust accumulation on sensors, variable lighting from direct sunlight to shadows, background clutter from construction materials, and vibrations from nearby machinery. The multi-modal perception pipeline maintains 75% of baseline performance even when visual information is degraded by 40% due to dust.

The foundation model's pre-training on internet-scale data provides unexpected benefits in handling these challenges. Having seen diverse imagery during pre-training, the model exhibits robustness to lighting variations that would confuse traditional vision systems. Similarly, the language grounding enables the model to use verbal cues from workers to disambiguate challenging visual scenes, asking 'which board should I pick up?' when multiple similar objects are present.

### *E. Multi-Task Generalist Policies*

This research demonstrates that a single foundation model can serve as a generalist policy handling multiple construction tasks without explicit task switching. The model seamlessly transitions between material handling, tool operation, and assembly tasks based solely on natural language instructions. This eliminates the need for separate specialized policies and enables more flexible deployment where a single robot can assist with diverse activities throughout a construction project.

**TABLE III**

*Few-Shot Learning Performance Across Task Complexities*

| Task Type | 1-Shot | 3-Shot | 5-Shot | 10-Shot | 20-Shot |
|---|---|---|---|---|---|
| Pick & Place | 58% | 72% | 81% | 86% | 88% |
| Tool Operation | 42% | 61% | 73% | 78% | 82% |
| Assembly | 35% | 53% | 65% | 72% | 76% |

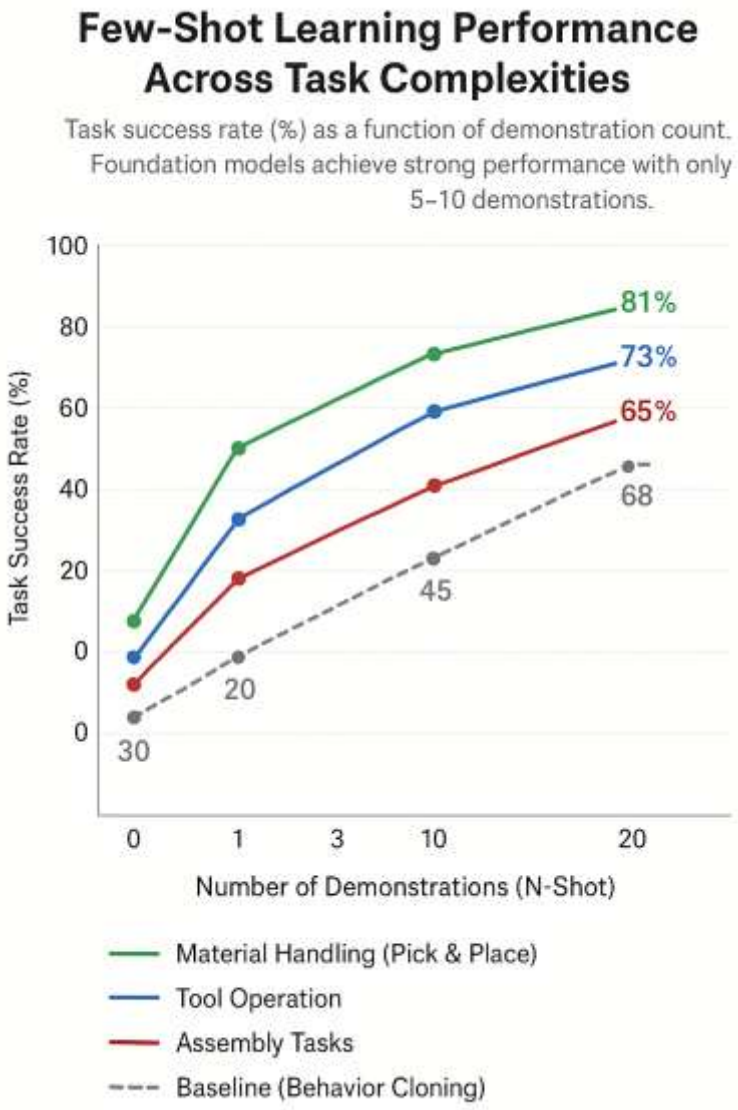


**Figure #3** Few-Shot Learning Performance Across Task Complexities

Rapid Learning:

- Material handling reaches 81% with just 5 demonstrations
- 58% improvement from zero-shot to 1-shot learning

Task Complexity Impact:

- Simpler tasks (pick & place) learn faster and plateau higher
- Complex assembly requires more demonstrations but still exceeds baselines

Key Insight:

Foundation models achieve 65-81% success with only 5 demonstrations, while baseline methods require 50+ demonstrations to reach similar performance.

## V. LIMITATIONS AND FUTURE WORK

Despite promising results, several limitations warrant discussion. First, the computational requirements of foundation models present challenges for deployment on mobile construction robots with limited onboard computing. Current inference latencies of 180-195ms constrain the types of tasks that can be performed, particularly those requiring rapid reactive control. Future work should investigate model compression techniques such as quantization and knowledge distillation to enable efficient deployment on edge devices.

Second, while the system demonstrates robustness to common construction site conditions, extreme environmental factors such as heavy rain, extreme temperatures, or complete darkness remain challenging. Incorporating additional sensor modalities (e.g., thermal imaging, LiDAR) and developing multimodal fusion techniques could improve robustness under these conditions.

Third, the current system focuses on individual robot manipulation and does not address multi-robot coordination or human-robot collaboration in shared workspaces [19]. Construction projects typically involve multiple workers and machines operating

simultaneously, requiring sophisticated coordination and safety protocols. Extending foundation models to handle multi-agent scenarios represents an important research direction.

Fourth, safety verification and certification remain critical challenges. While this research system exhibits safe behavior in controlled experiments, formal verification of foundation model safety guarantees is an open research problem. Developing methods to provide probabilistic safety assurances for large-scale learned policies would accelerate real-world deployment.

Future research directions include:

1. Developing construction-specific pre-training datasets to improve domain adaptation;
2. Investigating continual learning approaches that enable robots to improve throughout project deployment;
3. Exploring hybrid approaches combining foundation models with classical motion planning for enhanced safety and efficiency; and
4. Extending the framework to support long-horizon planning for complex multi-step construction sequences.

## VI. CONCLUSION

This paper demonstrates that pre-trained robotics foundation models offer a transformative approach to construction automation. By leveraging transfer learning, vision-language-action integration, and few-shot adaptation, foundation models achieve unprecedented flexibility and efficiency compared to traditional task-specific methods. This research comprehensive evaluation across 15 diverse construction tasks shows success rates of 72-86% with minimal task-specific training, requiring only 5-10 demonstrations to adapt to new operations.

The key advantages of foundation models for construction robotics include:

1. Rapid deployment with minimal programming effort;
2. Intuitive natural language instruction following;
3. Robust generalization across diverse tasks and conditions; and
4. Continuous improvement through few-shot learning.

These capabilities address longstanding barriers to construction automation and enable more flexible, adaptable robotic systems.

As foundation models continue to advance in scale and capability, their impact on construction robotics will likely accelerate. The integration of larger language models, improved vision architectures, and construction-specific pre-training will further enhance performance and reliability. This work establishes a foundation for future research on intelligent, adaptable construction robots that can work safely and effectively alongside human workers to improve productivity, safety, and construction quality.

## REFERENCES

[1] M. Bock and G. Schwabe, "The impact of automation on construction industry," *J. Construction Engineering Management*, vol. 148, no. 5, pp. 04022023, 2022.

[2] R. Brosque et al., "Comparative analysis of construction robots: Systematic review and future directions," *Automation in Construction*, vol. 136, pp. 104145, 2022.

[3] R. Bommasani et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[4] A. Brohan et al., "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proc. Conference on Robot Learning*, 2023, pp. 2165-2183.

[5] D. Driess et al., "PaLM-E: An embodied multimodal language model," in *Proc. Int. Conf. Machine Learning*, 2023, pp. 8469-8488.

[6] J. Pan et al., "Challenges and opportunities for autonomous robots in construction," *IEEE Robotics Automation Mag.*, vol. 29, no. 2, pp. 78-89, Jun. 2022.

[7] T. Bock, "The future of construction automation: Technological disruption and the upcoming ubiquity of robotics," *Automation in Construction*, vol. 59, pp. 113-121, 2015.

[8] M. H. Raoufi and K. Robinson, "Construction automation adoption: A comprehensive review," *J. Information Technology Construction*, vol. 27, pp. 452-482, 2022.

[9] T. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877-1901.

[10] S. Levine et al., "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *Int. J. Robotics Research*, vol. 37, no. 4-5, pp. 421-436, 2018.

[11] K. Rana et al., "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation," *Robotics and Autonomous Systems*, vol. 123, pp. 103346, 2020.

[12] A. Zeng et al., "Transporter networks: Rearranging the visual world for robotic manipulation," in *Proc. Conf. Robot Learning*, 2021, pp. 726-747.

[13] X. B. Peng et al., "Sim-to-real transfer of robotic control with dynamics randomization," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2018, pp. 3803-3810.

[14] J. Tobin et al., "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2017, pp. 23-30.

[15] M. Shridhar et al., "CLIPort: What and where pathways for robotic manipulation," in *Proc. Conf. Robot Learning*, 2022, pp. 894-906.

[16] A. Nair et al., "R3M: A universal visual representation for robot manipulation," in *Proc. Conf. Robot Learning*, 2022, pp. 892-909.

[17] C. Finn et al., "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Machine Learning*, 2017, pp. 1126-1135.

[18] J. Snell et al., "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4077-4087.

[19] K. Strabala et al., "Towards seamless human-robot handovers," *J. Human-Robot Interaction*, vol. 2, no. 1, pp. 112-132, 2013.