A ROBOTIC TESTBED FOR AUTONOMOUS DUMP POCKET CLEANING USING IMITATION LEARNING

*Brik Henrry Meza Pinedo[1,3], Brian Pajares Correa[1,2]

[1]Faculty of Science and Engineering, Pontifical Catholic University of Peru (PUCP), Lima, Peru
[2]Antamina, Ancash, Peru
[3]NONHUMAN, Lima, Peru
(*Presenting author: brik.meza@pucp.edu.pe)

ABSTRACT

Dump pocket blockages at primary crushers cause production downtime and require manual clearing in confined spaces, leading to multiple fatalities annually from engulfment. Current autonomous approaches in mining remain limited, and the feasibility of state-of-the-art imitation learning (IL) for excavation tasks has not been systematically evaluated. This paper introduces an experimental testbed and benchmark for evaluating end-to-end IL architectures on autonomous dump pocket cleaning under controlled laboratory conditions. We benchmark four IL architectures: Action Chunking with Transformers (ACT), Diffusion Policy, and Vision-Language-Action models ($\pi_{0.5}$ and SmolVLA), using a low-cost SO-ARM100 platform ($250) with granular bentonite material. In a preliminary single-session evaluation (10-minute continuous autonomous operation per model), $\pi_{0.5}$ removes 404 g (40.4 g/min, approximately 65% of the expert teleoperation estimate of 620 g), while ACT removes 169 g, SmolVLA 113 g, and Diffusion Policy 57 g. Notably, ACT removes more material than SmolVLA despite lacking pretrained representations, suggesting that pretraining benefit is architecture- and scale-dependent rather than universal in this domain. Diffusion Policy is notably the slowest, consistent with its iterative denoising inference process. This work establishes a reproducible benchmark and open dataset (162 demonstrations) to support future research on autonomous confined-space operations. **Project page:** `https://brikhmp18.github.io/dump-pocket-il`

KEYWORDS

Autonomous Excavator, Dump Pocket Cleaning, Imitation Learning, Mining 4.0, VLA Models, Operational Safety, Robot Learning.

1. CONTEXT AND PROBLEM STATEMENT

Mining is a major economic driver in Peru, contributing around 10% of national output and approximately two-thirds of export value [1]. In large-scale operations, the primary crushing station is a throughput-critical node: haul trucks discharge run-of-mine material into the primary crusher dump pocket (hopper), and any loss of dump pocket availability directly propagates to upstream hauling and downstream processing.

A persistent operational challenge is ore accumulation, bridging, and blockage in and around dump pockets, which can force partial or full stoppages to restore material flow. In large-scale operations, blocked-crusher events can cause 50–200 hours of unplanned downtime annually, with clearing procedures requiring 2–6 hours per incident depending on severity. From an operations perspective, improving process availability has an outsized profitability impact in crushing circuits; industry handbooks explicitly highlight that even small increases in process availability materially affect overall profitability [6].

Current dump pocket cleaning and blocked-crusher clearing procedures are frequently performed under constrained access and harsh sensing conditions (dust, poor visibility, irregular ore geometry), and are often only partially mechanized (e.g., breaker booms or excavator-assisted probing). However, authoritative safety guidance emphasizes that blockage clearance should be performed from a position of safety and should *not* involve anyone entering or being lowered into the crushing area due to the potential for sudden, uncontrolled release of stored energy and falling material [2]. Consistent guidance for crushing operations also stresses eliminating routine presence on crusher access platforms and emphasizes prevention and safer clearing practices [3]. Despite these directives, fatal accidents continue to occur when personnel enter hoppers or confined spaces to clear obstructions; MSHA reports include cases where a worker entered a hopper to clear blocked material and was engulfed by dumped material [4], and MSHA safety alerts document multiple fatalities from engulfment while clearing obstructions in hoppers/bins/crushers [5].

While partial solutions exist (e.g., rock-breaker booms, probing, or localized "softening"/clearing strategies), they do not provide an integrated pipeline for *autonomous* detection and removal of heterogeneous ore build-up under mine-realistic variability.

Related work. Kim & Choi [16] demonstrated laboratory-scale hopper unblocking using RGB-D vision and a modular detect-then-clear pipeline. ExACT [8] applied Action Chunking with Transformers to control a full-scale excavator with LiDAR+camera fusion for general digging tasks. Our work differs by: (i) benchmarking four IL architectures under identical conditions rather than a single method, (ii) targeting dump pocket cleaning, a safety-critical confined-space task with documented fatalities, and (iii) providing a low-cost ($250), reproducible testbed enabling systematic IL research in mining contexts where full-scale excavators are cost-prohibitive for academic experimentation.

2. OBJECTIVES AND SCOPE

The primary objective is to develop and benchmark an autonomous excavator system in a controlled dump-pocket testbed. This research addresses the "Technology" and "Transformation" themes of the conference by:

- Establishing a reproducible benchmark to quantify removal rate and autonomy reliability.
- Eliminating human exposure in hazardous confined spaces.

Target KPIs and evaluation criteria are defined in Table 1.

Table 1 - Target KPIs and evaluation criteria for autonomous dump pocket cleaning.

Metric	Definition	Target / Rationale
Removal rate	Material removed (g/min)	Maximize autonomous throughput
Success rate	Autonomous completion without intervention	Reliability for deployment
Safety proxy	Human exposure time in hazard zone	Achieve zero exposure

3. METHODOLOGY OR APPROACH

The system (Figure 1) is built around two SO-ARM100 robot arms [10] in a leader-follower teleoperation configuration. The SO-ARM100 is a low-cost (sub-$250), open-source, 6-DoF robot arm with 3D-printed PLA+ components and off-the-shelf STS3215 servo motors (7.4V), developed as an accessible platform for robotics research. The leader arm enables intuitive human demonstration via backdriving – the operator physically moves the leader, and joint angles map to the follower in real-time at 30 Hz. The follower arm executes the task equipped with a custom excavator bucket gripper. The teleoperation interface displays dual-camera streams (wrist and overhead) alongside real-time motor data to enable precise demonstration collection. This setup produced the 162-episode visuomotor dataset. Data collection and policy training leverage an open-source imitation learning library [9], which provides unified infrastructure for data collection and policy training.

3.1 Task definition

Workspace: Dump pocket testbed with internal box dimensions $L \times W \times H = 0.454 \times 0.112 \times 0.310\,m$. Walls are vertical with inclination $\theta = 90^\circ$ and maximum wall height $H_{\max} = 0.112\,m$.

Target region: The target region is not pre-defined as a fixed ROI; instead, the VLA policy learns the cleaning objective implicitly from teleoperation demonstrations.

Material: Natural bentonite clay (commercial cat litter, 0.5–2.5 mm particle size). The material exhibits granular behavior with slight interparticle cohesion, providing stable scooping dynamics. Bulk density $\rho_{\text{bulk}} \approx 1500$–$1700\,kg/m^3$; testing conducted under dry indoor conditions.

Initial conditions: Each run begins with 3000 g of material in the dump pocket (± 20 g variation, $<1\%$). The excavator bucket has 18 g max capacity per scoop. The evaluation measures continuous removal rate (g/min); complete evacuation in 1-minute is not required.

Task cycle: The policy detects regions requiring cleaning, approaches with the bucket, scoops material (18 g max capacity), transports to the primary crusher at the center of the dump pocket, and deposits the material. This cycle repeats continuously to maximize removal rate.

Success criteria: We define *success* operationally as (i) fully autonomous operation

without human intervention during the 10-minute session, (ii) no safety-limit violations, and (iii) measurable material removal. All methods start from the same initial state (3 kg material mass), and the primary performance metric is total *material removed* (g/10 min).

Failure modes: partial scoop; wall adhesion; bucket slip; collision; controller saturation; perception dropouts; and insufficient material removal within run duration.

The control is governed by four end-to-end imitation learning paradigms: **ACT** (Action Chunking with Transformers) uses a CVAE to predict action chunks trained from scratch with an $L_1 + \lambda KL$ objective [7, 8]; $\pi_{0.5}$ ($\sim$3.7B-parameter VLA) employs hierarchical inference – high-level subtask prediction followed by flow-matching action generation – pretrained on heterogeneous multi-source data (household manipulation, cross-embodiment, web data) [12, 11]; **SmolVLA** (450M-parameter compact VLA) uses a flow-matching Action Expert pretrained on 481 open-source community manipulation datasets [13]; and **Diffusion Policy** formulates action generation as conditional denoising diffusion via an MSE noise-prediction objective with iterative Stochastic Langevin Dynamics at inference [15].

3.2 Experimental Protocol

Baseline (Expert Teleoperation): Expert human teleoperation serves as the gold standard. During demonstration collection (162 episodes, 73,944 frames at 30 Hz; mean 15.18 s/ep, range 10.4–26.6 s, total 40.99 min), the expert removed 2542 g from 3000 g initial mass (15.69 g/episode, 62.04 g/min removal rate), establishing the upper bound of demonstration-phase performance under current hardware constraints.

Evaluation protocol: Each learned policy was evaluated in a single 10-minute continuous autonomous session under identical initial conditions (3 kg initial material mass, fixed lighting, fixed camera pose). The policy outputs joint-position targets q_{target} at 30 Hz; we log RGB images, joint states, actions, and timestamps. The primary metric is total material removed (grams per 10-minute session), measured by weighing the outlet collection tray before and after each session. Given the preliminary nature of this evaluation (single session per model), we report observed total removed mass and equivalent removal rate (g/min) without statistical aggregation. The expert teleoperation baseline is estimated from demonstration data (62.04 g/min average across 162 episodes), yielding an estimated 620 g over a 10-minute session.

Offline training: All models were trained on the 129-episode dataset using an open-source IL library [9]. Best checkpoints selected by validation loss: ACT at step 9205 (val loss 0.2746), $\pi_{0.5}$ at step 25774 (0.0745), SmolVLA at step 14728 (0.0254), Diffusion at step 7252 (0.0141). VLA models ($\pi_{0.5}$, SmolVLA) leverage pretrained backbones [12, 13]; ACT and Diffusion trained from scratch. Since ACT optimizes a composite objective ($L_1 + \lambda KL$) while other policies use MSE/flow-matching losses, Figure 2(a) plots ACT as $(\text{val/loss})^2$ for visual scale comparability only (monotone transformation, checkpoint ranking preserved); this panel serves as a convergence diagnostic, not a cross-architecture ranking.

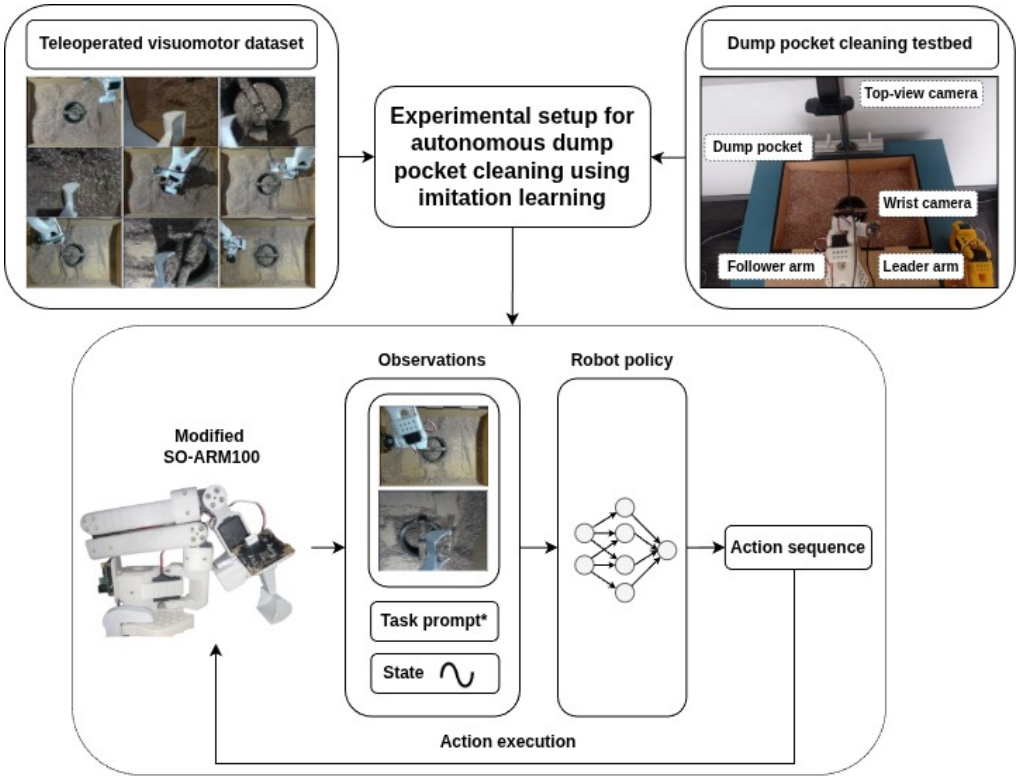

Figure 1 - Experimental setup for autonomous dump pocket cleaning using imitation learning. Top left: teleoperated dataset (162 demonstrations). Top right: physical testbed with dual RGB cameras (top-view, wrist-mounted), dump pocket, and SO-ARM100 platform (follower and leader arms). Bottom: imitation learning pipeline receiving observations (camera streams, task prompt*, proprioceptive state), processing through policy network, and executing actions in closed loop. *Task prompt: "use the scoop-like end effector to gather sand from the dump pocket and deposit it into the primary crusher at the center" – used only by VLA models ($\pi_{0.5}$, SmolVLA) for language conditioning; ACT and Diffusion operate on vision and proprioception only.

3.3 Observations and Action Space

We define the observation o_t as the concatenation of two RGB streams and proprioception:

- **Wrist camera:** 2MP camera (30 FPS) mounted on the follower arm end-effector (bucket), providing close-range perspective of material and manipulation area.
- **Overhead camera:** Webcam providing global context view of the dump pocket and arm workspace.
- **Proprioception:** Joint angles (6-DoF from STS3215 servos) and gripper state, synchronized with camera timestamps at 30 Hz.
- **Task prompt (VLA only):** Natural language instruction: "use the scoop-like end effector to gather sand from the dump pocket and deposit it into the primary crusher

at the center" – used by $\pi_{0.5}$ and SmolVLA for multimodal conditioning; ACT and Diffusion do not use language input.

The action a_t is a vector of joint-position targets q_{target} predicted by the policy and executed by the low-level servo controller. The teleoperation interface displays both camera streams simultaneously alongside real-time motor data (joint angles, velocities, torques) to enable precise demonstration collection and system monitoring. Actions are produced at 30 Hz control frequency. ACT uses action chunking with horizon $H = 100$ steps; $\pi_{0.5}$, SmolVLA, and Diffusion Policy operate in receding-horizon mode with model-specific temporal windows as configured in the training library [9]. Table 2 summarizes the platform specifications and dataset statistics.

Table 2 - Platform and dataset summary.

Platform	**Dataset**
SO-ARM100, 6-DoF, bucket 0–18 g	162 eps (129/33 split), 73,944 frames
Sub-$250, leader-follower	30 FPS, mean 15.18 s/ep (range 10.4–26.6 s)
Dual RGB (wrist + overhead)	40.99 min total, expert rate 62.04 g/min

4. RESULTS, OUTCOMES, OR PERFORMANCE

We conducted controlled evaluations on the SO-ARM100 testbed in a laboratory-scale dump-pocket environment (indoor setting, no mine deployment). Each model was evaluated in a single 10-minute continuous autonomous session under identical initial conditions (3 kg initial bentonite mass, fixed lighting). The primary metric is total material removed (g) over the session; results are summarized in Figure 2 and Table 3.

Note on statistical power: This evaluation reports single observed values per model (N=1 session of 10 minutes). Statistical inference (mean±std, confidence intervals) is not applicable; results represent exploratory observations pending systematic replication.

Table 3 - Benchmark results: total material removed in a single 10-minute autonomous session. Expert baseline estimated from demonstration data (62.04 g/min × 10 min). All learned policies evaluated under identical initial conditions (3 kg bentonite, fixed lighting). Given N=1 session per model, statistical aggregation (mean±std) is not applicable; values represent single observed measurements.

Model	**Removed (g / 10 min)**	**Rate (g/min)**	**% of Expert (est.)**
Expert Teleoperation (est.)	≈620	62.0	100
$\pi_{0.5}$	404	40.4	65
ACT	169	16.9	27
SmolVLA	113	11.3	18
Diffusion Policy	57	5.7	9

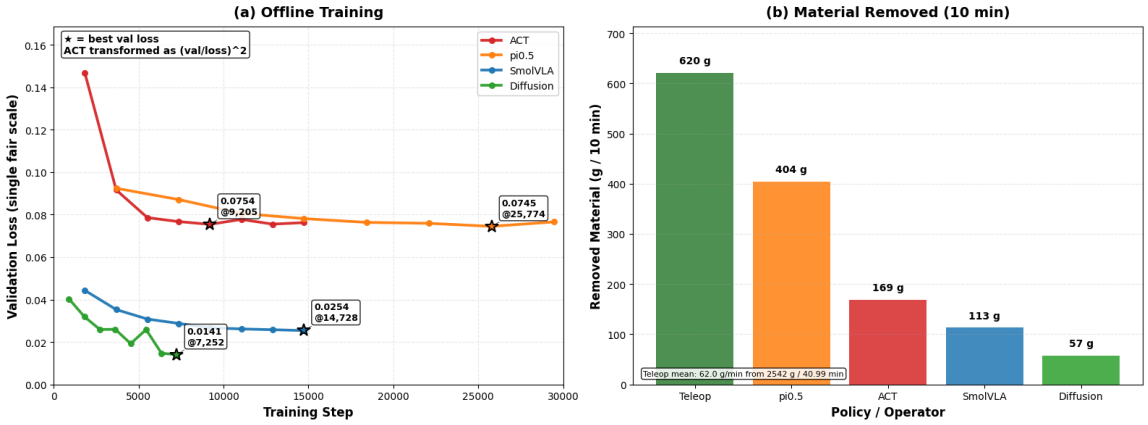

Figure 2 - Benchmark results and offline training metrics. (a) Validation loss trajectories; ACT is plotted as $(\text{val/loss})^2$ for visual scale comparability (see Section 3 for justification). Best checkpoints (stars): ACT 0.0754 @ 9,205 steps; $\pi_{0.5}$ 0.0745 @ 25,774 steps; SmolVLA 0.0254 @ 14,728 steps; Diffusion 0.0141 @ 7,252 steps. This panel is a convergence diagnostic; cross-architecture comparison uses real-robot results only. (b) Total material removed (g) in a single 10-minute autonomous session. Expert baseline (≈620 g) estimated from demonstration data (62.0 g/min, 2542 g over 40.99 min). $\pi_{0.5}$ removes 404 g, followed by ACT (169 g), SmolVLA (113 g), and Diffusion Policy (57 g).

Key findings: Under controlled laboratory conditions, $\pi_{0.5}$ achieves the highest removal performance (404 g in 10 minutes, 40.4 g/min), reaching approximately 65% of the expert teleoperation estimate (620 g). ACT (169 g, 27% of expert) ranks second despite having no pretrained representations, trained solely on 162 demonstration episodes. SmolVLA (113 g, 18% of expert) ranks third, removing less material than ACT despite being pretrained on 481 open-source robot manipulation datasets, suggesting this general-purpose pretraining does not transfer effectively to the excavation domain. Diffusion Policy (57 g, 9% of expert) is notably the slowest by a large margin, consistent with its iterative denoising process being less suited to fast reactive control in granular material manipulation.

5. DISCUSSION

The benchmark demonstrates feasibility of end-to-end learned policies for autonomous dump pocket cleaning under laboratory conditions. A key interpretive finding is that pretraining benefit is architecture- and scale-dependent rather than universal: ACT, trained from scratch, removes more material than pretrained SmolVLA in this single-session evaluation, while $\pi_{0.5}$'s larger scale and hierarchical design transfer more effectively to the excavation domain. Results reflect simplified laboratory conditions (bentonite, fixed lighting, N=1 session per model) that may not stress-test reliability differences emerging under operational challenges.

$\pi_{0.5}$ **advantages:** $\pi_{0.5}$'s strong performance is attributable to: (i) *scale* – at ∼3.7B parameters with the longest training run (25,774 steps), it is the most compute-intensive model evaluated; (ii) *hierarchical reasoning* – it decouples high-level subtask prediction from low-level flow-matching action generation, enabling more robust adaptation to varying

material states; and (iii) *breadth of pretraining* – household and cross-embodiment data provide richer priors than compact community datasets alone.

Pretraining and domain mismatch: A notable finding is that ACT removed more material than SmolVLA in this session, despite lacking SmolVLA's pretrained vision-language backbone; neither VLA's pretraining includes mining excavation data ($\pi_{0.5}$ draws from heterogeneous household and cross-embodiment datasets; SmolVLA from 481 open-source manipulation datasets), yet $\pi_{0.5}$ transfers effectively while SmolVLA does not. This suggests pretraining benefit is architecture- and scale-dependent: $\pi_{0.5}$'s larger capacity, hierarchical design, and significantly longer training (25,774 vs. 14,728 steps) all likely contribute, making it difficult to isolate pretraining from compute advantages. **Deployment barriers:** (i) *Material variability* – bentonite testbed versus heterogeneous ore (dust to boulders); (ii) *Scale* – testbed 1000$\times$ smaller than industrial hoppers (0.016 vs 10–50 m^3); (iii) *Environment* – dust, poor lighting, vibration will degrade vision [18]; (iv) *Control latency* – Diffusion Policy exhibited visibly slower reactive behavior compared to other methods, consistent with the iterative denoising inference process requiring multiple forward passes per action, which may limit responsiveness in fast-changing manipulation scenarios. This lab-to-mine gap is the primary limitation.

Task-state ambiguity and visual activation: ACT, SmolVLA, and Diffusion Policy exhibited hesitation at rollout start: the initial camera observation closely resembles the terminal episode state (same arm rest pose, similar material distribution), creating visual ambiguity between "task just started" and "task already finished." $\pi_{0.5}$ did not exhibit this behavior, likely due to language-conditioned task-initiation signals. A practical recommendation is to introduce a visual activation cue – for example, a display that shows green when a rollout begins and switches to red once material has been deposited – enabling vision-dependent policies to unambiguously distinguish task states without architectural changes.

Dataset scale: The 129-episode training split is modest by IL standards. Scaling to several hundred or thousand demonstrations with greater material-state variability would likely improve performance across all architectures and is a natural step before drawing conclusions about their performance upper bounds on this task.

6. CONCLUSIONS AND IMPLICATIONS FOR INDUSTRY

This work provides the first systematic benchmark of modern imitation learning architectures for mining excavation tasks under controlled laboratory conditions. In a single preliminary 10-minute evaluation session, $\pi_{0.5}$ removes 404 g (65% of the expert teleoperation estimate) while ACT, SmolVLA, and Diffusion Policy remove 169, 113, and 57 g respectively. A notable observation is that ACT, trained from scratch on 162 demonstrations, removes more material than SmolVLA despite SmolVLA having a pretrained VLA backbone, suggesting that pretraining benefit depends strongly on architecture and domain proximity. These preliminary results establish an experimental foundation and open research benchmark for evaluating end-to-end learning methods on confined-space excavation tasks.

The motivation is a critical safety need: manual dump pocket clearing exposes workers to fatal engulfment risk [4, 5] while causing 50–200 hours of annual downtime; autonomous systems could eliminate human entry [2]. Substantial barriers remain: the testbed uses bentonite rather than heterogeneous ore, operates at 1000× smaller scale (0.016 vs. 10–50 m^3), and lacks dust and vibration – this lab-to-mine gap is the primary limitation.

From a research perspective, this work establishes a precedent for introducing IL methods to mining contexts. This benchmark provides: (i) an accessible platform ($250 hardware) for mining robotics research, (ii) an open dataset (162 episodes) for reproducible comparisons, and (iii) preliminary evidence that model scale and architecture matter more than pretraining alone when adapting foundation models to specialized excavation tasks. This research foundation supports long-term development of autonomous technologies that may eventually contribute to eliminating confined-space exposure in mining once sufficient reliability is demonstrated under realistic operational conditions.

ACKNOWLEDGEMENTS

The authors thank NONHUMAN for providing laboratory facilities and infrastructure support that enabled this research. We are grateful to the open-source robotics community, particularly the developers of the SO-ARM100 platform and the open-source robot learning library [9], whose accessible tools and collaborative spirit made these experiments possible. We also acknowledge PUCP for institutional support.

REFERENCES

[1] BBVA Research. (2023). *Peru: Mining sector outlook 2022*. February 3, 2023. (PDF report).

[2] Health and Safety Authority (HSA). (n.d.). *Clearing Blocked Crushers*. Quarrying: Crushing, Sizing & Screening guidance.

[3] Health and Safety Executive (HSE). (n.d.). *Safe operation and use of mobile jaw crushers*. Quarries guidance (UK).

[4] Mine Safety and Health Administration (MSHA). (2023). *Fatality Report: September 8, 2023 Fatality - Final Report*. U.S. Department of Labor.

[5] Mine Safety and Health Administration (MSHA). (n.d.). *Confined Spaces Safety Alert*. U.S. Department of Labor.

[6] Metso. (2011). *Crushing and Screening Handbook* (5th ed.). (AusIMM-hosted handbook PDF).

[7] Zhao, T. Z., Kumar, V., Levine, S., & Finn, C. (2023). *Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware*. Proceedings of Robotics: Science and Systems (RSS) XIX, Daegu, Republic of Korea. DOI: 10.15607/RSS.2023.XIX.016

[8] Chen, L., Jin, S., Wang, H., & Zhang, L. (2024). *ExACT: An End-to-End Autonomous Excavator System Using Action Chunking With Transformers*. ICRA 2024 Future of Construction Workshop. arXiv:2405.05861

[9] LeRobot Team. (2026). *LeRobot: An Open-Source Library for End-to-End Robot Learning*. Under review at ICLR 2026. Available at: `https://github.com/huggingface/lerobot`

[10] Knight, R., Kooijmans, P., Cadene, R., Alibert, S., Aractingi, M., Aubakirova, D., Zouitine, A., Martino, R., Palma, S., Pascal, C., & Wolf, T. (2024). *Standard Open SO-100 & SO-101 Arms*. GitHub. `https://github.com/TheRobotStudio/SO-ARM100`

[11] Physical Intelligence (Black, K., Brown, N., Driess, D., et al.). (2024). π_0*: A Vision-Language-Action Flow Model for General Robot Control*. arXiv:2410.24164

[12] Physical Intelligence (Black, K., Brown, N., Darpinian, J., et al.). (2025). $\pi_{0.5}$*: A Vision-Language-Action Model with Open-World Generalization*. arXiv:2504.16054

[13] Shukor, M., Aubakirova, D., Capuano, F., Kooijmans, P., Palma, S., Zouitine, A., Aractingi, M., Pascal, C., Russi, M., Marafioti, A., Alibert, S., Cord, M., Wolf, T., & Cadène, R. (2025). *SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics*. arXiv:2506.01844

[14] Zitkovich, B., et al. (2023). *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control*. Proceedings of the 7th Conference on Robot Learning (CoRL), in Proceedings of Machine Learning Research, 229:2165-2183. arXiv:2307.15818

[15] Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., & Song, S. (2024). *Diffusion Policy: Visuomotor Policy Learning via Action Diffusion*. The International Journal of Robotics Research. DOI: 10.1177/02783649241273668. arXiv:2303.04137

[16] Kim, H., & Choi, Y. (2022). *Lab Scale Model Experiment of Smart Hopper System to Remove Blockages Using Machine Vision and Collaborative Robot*. *Applied Sciences*, 12(2), 579.

[17] Zhang, L., Zhao, J., Long, P., Wang, L., Qian, L., Lu, F., Song, X., & Manocha, D. (2021). *An autonomous excavator system for material loading tasks*. *Science Robotics*, 6(55), eabc3164. DOI: 10.1126/scirobotics.abc3164

[18] Cavieres, B., Cruz, N., & Ruiz-del-Solar, J. (2025). *Dust Filtering in LiDAR Point Clouds Using Deep Learning for Mining Applications*. *Sensors*, 25(20), 6441. DOI: 10.3390/s25206441