Article

Not peer-reviewed version

Physical AI: The Next Frontier in AI and Robotics to Build Truly Autonomous Machines

Ankit Parag Shah *, Minghao Liu, Jinrui Huang, Bhavya Pranav Tandra, Yu Wang, Sakshi Agarwal, Paul Wu, Aishik Konwer, Rhett Rozga, Selli P. Kondamuri, Alper Halbutogullari, Qin Zhang, Wei Wei

Posted Date: 8 April 2026

doi: 10.20944/preprints202604.0549.v1

Keywords: physical AI; embodied intelligence; foundation models; vision-language-action models; sim-to-real transfer; digital twins; world models; robot learning

 doi:10.20944/preprints202604.0549.v1

Article

Physical AI: The Next Frontier in AI and Robotics to Build Truly Autonomous Machines

Ankit Parag Shah *, Minghao Liu, Jinrui Huang, Bhavya Pranav Tandra, Yu Wang, Sakshi Agarwal, Paul Wu, Aishik Konwer, Rhett Rozga, Selli P. Kondamuri, Alper Halbutogullari, Qin Zhang and Wei Wei

Center for Advanced AI, Accenture
* Correspondence: ankit.parag.shah@accenture.com

Abstract

For decades, artificial intelligence transformed digital domains while the physical world, with its unforgiving dynamics and infinite variability, remained largely beyond reach. That boundary is now dissolving: humanoid robots work full shifts on assembly lines, autonomous vehicles make millions of safety-critical decisions daily, and robotic surgical platforms have performed over twenty million procedures worldwide. The convergence of foundation models, high-fidelity simulation, and embodied control is producing *Physical AI*, machines that perceive, reason, and act reliably in open environments. However, existing surveys treat individual components of this stack in isolation, covering vision-language-action architectures, world models, or sim-to-real transfer separately, and none traces the full pipeline from sensor to deployment or grounds analysis in commercial outcomes. Here we synthesise the end-to-end Physical AI technology stack: multimodal sensing and fusion; edge hardware and accelerators; world modeling and simulation; vision-language-action models; learning paradigms from reinforcement to imitation; and deployment infrastructure spanning safety assurance, fleet learning, and governance. Our analysis reveals three cross-cutting findings. First, foundation-model generalisation, not task-specific engineering, is the economic lever that separates scalable deployments from those that stall at pilot stage. Second, digital twins have become prerequisites rather than accelerants: no deployment at fleet scale proceeds without high-fidelity virtual rehearsal. Third, regulatory and assurance barriers, not algorithmic limitations, now gate the transition from pilot to production across every domain. We document these patterns through commercial case studies in manufacturing, logistics, autonomous mobility, agriculture, and healthcare, and propose a four-phase maturity taxonomy characterising adoption from research prototype to fleet-scale operation. We conclude with six coupled research challenges and a roadmap anchored to regulatory milestones through 2030, offering researchers, industry leaders, and policymakers a practical map for scaling embodied intelligence from controlled environments to the complexity of the real world.

Keywords: physical AI; embodied intelligence; foundation models; vision-language-action models; sim-to-real transfer; digital twins; world models; robot learning

1. Introduction

Intelligence that merely thinks is no longer enough. For decades, artificial intelligence achieved remarkable feats in digital domains: mastering board games, generating fluent prose, and recognising objects in photographs. Yet the physical world, with its unforgiving dynamics, infinite variability, and millisecond deadlines, remained largely beyond reach. Moravec's paradox [1] captured the irony: the sensorimotor skills that a toddler performs effortlessly demand far greater computational sophistication than chess or theorem-proving. Today, that paradox is beginning to yield. Humanoid robots walk onto factory floors and work full shifts alongside people, performing dexterous assembly tasks that only human hands could accomplish a few years ago [2]. Autonomous vehicles navigate

 doi:10.20944/preprints202604.0549.v1

dense city traffic on multiple continents, making millions of safety-critical decisions daily [3]. Robotic surgical platforms have carried out over twenty million procedures worldwide [4], while autonomous tractors till entire fields without a human aboard [5]. These advances share not a single algorithm but a convergence: foundation models trained on internet-scale data are merging with high-fidelity simulation, multimodal perception, and embodied control to produce machines that can perceive, reason, and act in the open world. We call this convergence *Physical AI* [6], and it marks the moment when artificial intelligence steps out of the screen and into reality.

1.1. What Is Physical AI?

Physical AI enables autonomous systems to perceive, understand, reason, and act in the physical world, adapting to situations never encountered during training [6]. The concept builds upon decades of research in embodied intelligence [7] and situated cognition, which emphasised that intelligence emerges from the interaction between agents, their bodies, and their environments. Four core functions define Physical AI systems:

Perceive: gathering and interpreting real-time data from cameras, LiDAR, radar, force-torque sensors, and other modalities to build a rich picture of the surrounding environment.

Understand: comprehending spatial relationships, object affordances, and the physical behaviour of the 3D world—going beyond pattern recognition to situational awareness.

Reason: making informed, goal-driven decisions under uncertainty, including planning multi-step actions that respect dynamics, contact, and safety constraints.

Act: executing or orchestrating complex physical movements through actuators—motors, grippers, wheels, and whole-body controllers—to interact with and reshape the environment.

Physical AI extends generative AI—which produces text and images in digital domains—to real-world tasks through embodied interaction. Unlike traditional robotics, which relies on fixed programs in controlled settings, Physical AI replaces hand-crafted routines with adaptive policies that generalise, learn from experience, and handle novelty in real time. This transition from passive computation to physical agency raises the bar for reliability, safety, and real-time performance beyond what digital-only systems require.

Fundamental challenges.

Several deep technical problems distinguish Physical AI from other AI domains. *Moravec's paradox* [1] observes that sensorimotor skills trivial for humans—grasping objects, navigating uneven terrain—require enormous computational resources for machines, reflecting millions of years of evolutionary optimisation that current AI lacks. The *reality gap* causes policies trained in simulation to fail on physical hardware due to unmodelled dynamics, sensor noise, and environmental variation [8]. *Data scarcity* compounds these difficulties: even the largest robot demonstration datasets contain orders of magnitude fewer examples than the trillion-token corpora available for language models, limiting the learning approaches that have transformed NLP and vision. *Real-time constraints* demand that perception-to-action loops complete within milliseconds, precluding the iterative refinement possible in offline settings. Finally, *contact-rich manipulation*—handling deformable objects, assembling parts with tight tolerances, manipulating cables—remains challenging because accurate contact modelling is computationally expensive and small errors compound into task failure.

1.2. Evolution Toward Physical AI

The notion of intelligent physical systems has evolved through distinct paradigm shifts. Early robots such as Shakey (1966–1972) embodied the *sense-plan-act* paradigm, with symbolic planners like STRIPS searching for action sequences achieving specified goals [9,10]. Brooks' *subsumption architecture* challenged this approach, demonstrating that robust behaviour could emerge from layered reactive controllers without explicit symbolic reasoning [11]. Both paradigms struggled with generalisation—classical planners required hand-crafted domain models; behaviour-based systems required hand-designed behaviours—and neither could leverage the statistical regularities in large datasets. The

 doi:10.20944/preprints202604.0549.v1

2010s brought deep reinforcement learning into robotics: algorithms such as PPO [12] enabled robots to learn locomotion and manipulation through trial-and-error in continuous action spaces, while simulation tools like MuJoCo [13] and domain randomisation [8] began to bridge the gap between virtual and physical environments.

The current transformation is driven by the convergence of large language models (LLMs), vision-language-action (VLA) policies, and simulation-driven training pipelines. Vision-language models such as CLIP [14] bridged perception and language, and systems like SayCan [15] and Code-as-Policies [16] demonstrated that pre-trained LLMs could decompose high-level instructions into feasible robot actions. VLA models represent the latest evolution, directly integrating perception, language understanding, and motor control: RT-2 [17] co-trains on web-scale data and robotic demonstrations to generalise across tasks, PaLM-E [18] fuses multiple sensor modalities with language reasoning, and π_0 [19] uses flow matching to produce continuous motor commands at 50 Hz across seven platforms. Beyond task-specific models, generalist agents such as Gato [20] treat text, images, and motor commands as tokens in a single sequence, providing early evidence that cross-modal transfer can benefit physical control. The 2023–2025 period has seen unprecedented industrial adoption, with low-code deployment platforms [21], foundation models for industrial grasping [22], and commercial-scale fleets in logistics, manufacturing, and urban mobility [23,24]. Physical AI companies attracted over $7 billion in venture funding in 2024 alone [25], with mega-rounds signalling that investors view embodied foundation models as a generational opportunity. These deployments share a common enabler: high-fidelity simulation for safe exploration and rapid iteration at a fraction of the cost of physical prototyping, making digital twins and GPU-accelerated simulators essential infrastructure for the entire development cycle.

Figure 1 illustrates this end-to-end workflow. While the pipeline superficially resembles the classical sense-plan-act loop critiqued by Brooks [7], modern Physical AI systems differ in three fundamental ways: (1) learned policies replace symbolic planning, enabling reactive behaviour without hand-crafted world models; (2) continuous feedback loops link simulation, deployment, and model retraining so that fleet experience refines the next generation of policies; and (3) foundation models provide implicit representations that blur the boundary between perception and action, allowing a single architecture to serve multiple embodiments and tasks.

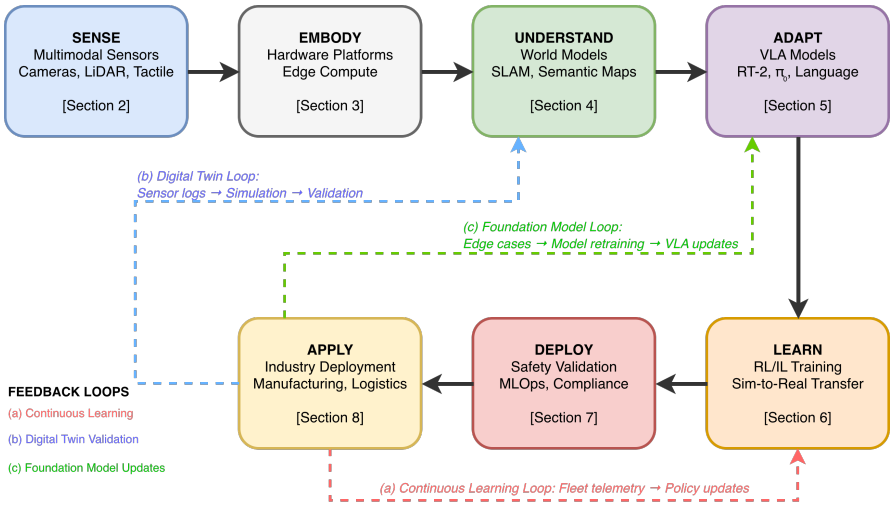

Figure 1. Physical AI pipeline overview. **Solid dark arrows** (left to right) denote sequential data flow through seven stages: *SENSE* → *EMBODY* → *UNDERSTAND* → *ADAPT* → *LEARN* → *DEPLOY* → *APPLY*. **Dashed coloured arrows** (curving backward) denote three continuous feedback loops: (a) Continuous Learning—fleet telemetry feeds policy updates; (b) Digital Twin—deployment logs validate policies in simulation; (c) Foundation Model—edge cases trigger VLA retraining. Each stage maps to a numbered survey section.

 doi:10.20944/preprints202604.0549.v1

1.3. Scope, Contributions, and Organisation

Existing surveys address individual components of the Physical AI stack in isolation—VLA architectures [26], world models for embodied AI, robot learning paradigms, or sim-to-real transfer—but none traces the full pipeline from sensor to deployment or grounds its analysis in commercial outcomes. Table 1 positions this survey relative to recent related work.

Table 1. Scope comparison with recent related surveys. ✓ indicates substantial coverage; ○ indicates partial or incidental treatment; — indicates not covered.

Survey	Sensors	HW	World Models	VLAs	Learning	Safety	Deploy.
Kawaharazuka et al. 2025 [26]	—	—	○	✓	○	—	—
Liu et al. 2025 [27]	○	—	○	○	✓	—	—
Sapkota et al. 2025 [28]	—	—	—	✓	○	—	—
This survey	✓	✓	✓	✓	✓	✓	✓

This survey provides an end-to-end treatment of the field at a critical juncture. Its principal contributions are:

1. **Full-stack integration perspective.** We trace the complete Physical AI pipeline from multimodal sensing through edge hardware, world models, foundation policies, and fleet deployment (Figure 1), exposing cross-cutting dependencies—between simulation fidelity and policy robustness, between hardware cost curves and deployment breadth, between governance frameworks and scalability—that single-topic surveys cannot capture.
2. **Deployment-grounded analysis.** We systematically document commercial deployments with measured outcomes across logistics, manufacturing, autonomous vehicles, and emerging domains, and propose a four-phase maturity taxonomy (Section 7) that characterises adoption trajectories from research prototypes to fleet-scale operations.
3. **Critical gap identification and roadmap.** We identify six coupled challenges—data ecosystems, sim-to-real resilience, lifelong adaptation, safety assurance, workforce integration, and sustainable hardware—and anchor a concrete research roadmap to regulatory milestones through 2030 (Section 8).

Survey methodology.

We surveyed literature from Google Scholar, Semantic Scholar, arXiv, and IEEE Xplore, focusing on 2020–2025 with selective inclusion of foundational work from earlier decades. Search queries combined terms from each pipeline stage (e.g., "vision-language-action models," "sim-to-real transfer robotics," "robot fleet learning") with deployment-oriented terms ("warehouse," "manufacturing," "autonomous driving"). We supplemented peer-reviewed publications with industry white papers, press releases, and verified deployment reports where academic sources were unavailable for recent commercial systems. All citations were verified against primary sources; arXiv preprints that have since appeared at peer-reviewed venues are cited with their published metadata.

The paper follows the pipeline of Figure 1. Appendix A surveys multimodal sensing and fusion. Section 2 examines hardware platforms, edge accelerators, and the co-design of morphology with control. Section 3 covers world modelling—SLAM, semantic mapping, intuitive physics, simulation infrastructure, and the emerging role of world foundation models. Section 4 analyses VLA models and generalist policies from RT-2 through π_0 to Gemini Robotics. Section 5 reviews learning paradigms: reinforcement learning, imitation, self-supervised methods, and sim-to-real transfer. Section 6 addresses safety assurance, governance, fleet deployment, and societal considerations. Section 7 presents industry case studies with measured outcomes. Section 8 synthesises open challenges into a research roadmap through 2030.

 doi:10.20944/preprints202604.0549.v1

2. Hardware Platforms and System Architectures

Physical AI systems depend on tightly coupled hardware stacks that span electromechanical design, sensing, compute and software. Progress in any layer matters only when it improves the combined system: lighter end-effectors enable faster motion planning, more accurate sensors reduce the burden on perception models, and efficient accelerators make it feasible to run large policies on-device. This section reviews how platform costs, sensing, compute and simulation infrastructure are evolving together to enable practical Physical AI deployments at scale.

2.1. Platform Landscape and Cost Curves

Today's robot platforms can reliably execute repetitive, well-defined tasks such as welding, depalletising, and single-SKU picking, but cannot yet generalise across environments, adapt to novel objects, or recover from unexpected failures without human intervention. This capability boundary, not component cost, explains why the market remains dominated by specialised form factors rather than general-purpose platforms [29]. Global shipments are nevertheless accelerating: IFR estimates 575 000 industrial units in 2025 and 700 000 by 2028 [30], with growth led by electronics and automotive. Traditional industrial arms from ABB, KUKA, and Fanuc dominate high-payload applications, while collaborative robots such as the UR series (Universal Robots) and GoFa (ABB) enable safer human-robot interaction in assembly and inspection tasks. Humanoid platforms are nevertheless accelerating: Figure announced a long-term agreement with BMW in 2024 to pilot its Figure 02 humanoids on assembly tasks, while Chinese manufacturer Unitree set an aggressive starting price from $4 900 for its R1 general-purpose humanoid aimed at researchers and service work [31,32]. Tesla's Optimus program targets retail pricing in the $20 000–$30 000 range for eventual manufacturing deployment [33]. These examples illustrate a bifurcated hardware roadmap where niche automation continues reducing cost today while generalist platforms mature through targeted pilots.

2.2. Integrated Sensing and Actuation Stacks

Integrated sensor suites can now fuse vision, depth, and force-torque data well enough for a single robot to switch from gross handling to fine assembly without retooling, but only when the sensing rig, calibration pipeline, and task domain are co-designed; migrating a proven configuration to a new robot or environment still requires months of re-integration [29]. Warehouse automation similarly benefits from integrated sensor suites: Zebra Technologies' vision systems and RFID readers combine with mobile robot bases from companies like Fetch Robotics to enable real-time inventory tracking and dynamic path planning in crowded fulfillment centers. Digital twin infrastructure now ingests heterogeneous sensor data—RGB cameras, LiDAR, radar and force-torque signals—allowing engineers to validate sensing layouts before hardware is fabricated. NVIDIA's Omniverse Cloud Sensor RTX stack renders physically accurate camera, radar and LiDAR returns in real time, enabling teams to design sensing rigs and validate fusion algorithms against synthetic edge cases that would be difficult to stage on the shop floor [34]. The same simulation pipelines drive continuous calibration: when the real robot drifts, the twin highlights which sensor or actuator needs maintenance, tightening the perception–control loop and minimizing downtime.

Three architectural patterns recur across these integrated stacks. First, the sensing modality mix is converging: most production deployments now fuse at least a camera stream, a depth or range sensor, and a force-torque channel, yet the calibration and fusion pipelines remain bespoke to each integrator, creating a fragile $N \times M$ coupling between N sensor types and M task domains. Second, simulation-in-the-loop calibration is replacing periodic offline routines, but its effectiveness depends on digital-twin fidelity that few operators can verify independently. The critical open challenge is *sensor-stack portability*: because perception performance is validated against a specific mechanical mounting, cable routing, and timing budget, migrating a proven sensing configuration to a new robot form factor still requires months of re-integration, an obstacle that no amount of individual component improvement can overcome without standardised electromechanical and temporal interfaces.

 doi:10.20944/preprints202604.0549.v1

2.3. Edge Compute and Neuromorphic Accelerators

A stark capability gap defines edge compute for Physical AI: current hardware runs classical detection and SLAM at real-time rates but cannot execute the vision-language models needed for open-ended reasoning at control frequencies physical tasks demand. On the widely deployed Jetson Orin (275 TOPS, 60 W), YOLO-class detectors reach ∼100 FPS, yet VLMs achieve only 0.1–0.4 FPS, far below the ≥10 FPS threshold for closed-loop perception [35]. Architects therefore split workloads: lightweight detectors handle the fast control loop while VLMs run asynchronously for scene understanding, or heavy inference is offloaded to the cloud at the cost of latency and reliability. Matching model architectures to heterogeneous accelerators is thus a first-order systems design problem, not a standalone algorithmic choice.

The next generation of edge silicon aims to close this gap. NVIDIA's Jetson Thor (∼2 000 TOPS, 130 W), shipping since August 2025, integrates a transformer engine sized to run multiple VLA policies concurrently on a mobile base [35,36]; OEMs including Boston Dynamics, Amazon Robotics, Figure, and Agility Robotics are early adopters [35]. Qualcomm's RB5 (15 TOPS, 15 W) targets drones and service robots where 5G offload compensates for limited on-device capacity [37]. At the opposite end of the power spectrum, neuromorphic accelerators such as Intel's Loihi 2 achieve orders-of-magnitude energy savings on event-driven workloads (<1 W), but cannot run dense transformers, restricting their role to spiking tactile controllers and reflexive loops [38,39]. Table 2 summarises what each platform tier can and cannot run today. Figure 2 illustrates how these platforms integrate with the complete Physical AI hardware stack.

Table 2. Edge AI capability gap: what today's accelerators can and cannot run at real-time control frequencies. Specs are compressed into the platform column; the two capability columns capture the deployment reality that raw TOPS numbers obscure.

Platform (specs)	Can run today	Cannot run / key gap
Jetson Thor ∼2000 TOPS / 130 W	Multiple VLA policies concurrently; transformer-class models on a mobile base at control-loop rates	Unproven at scale; supply-constrained; requires significant thermal engineering
Jetson Orin 275 TOPS / 60 W	YOLO-class detection at ∼100 FPS; lightweight SLAM; backbone of most deployed AMRs	VLMs at only 0.1–0.4 FPS (25−100× below real-time threshold); forces split architecture
Qualcomm RB5 15 TOPS / 15 W	On-device object detection; 5G enables cloud offload; drones and service robots	Zero foundation-model inference on-device; entirely cloud-dependent for VLA workloads
Loihi 2 Event-driven / <1 W	Spiking tactile controllers; event-driven reflexes at μW power	Cannot run dense transformers; no mature toolchain for VLA-class models; research-only

 doi:10.20944/preprints202604.0549.v1

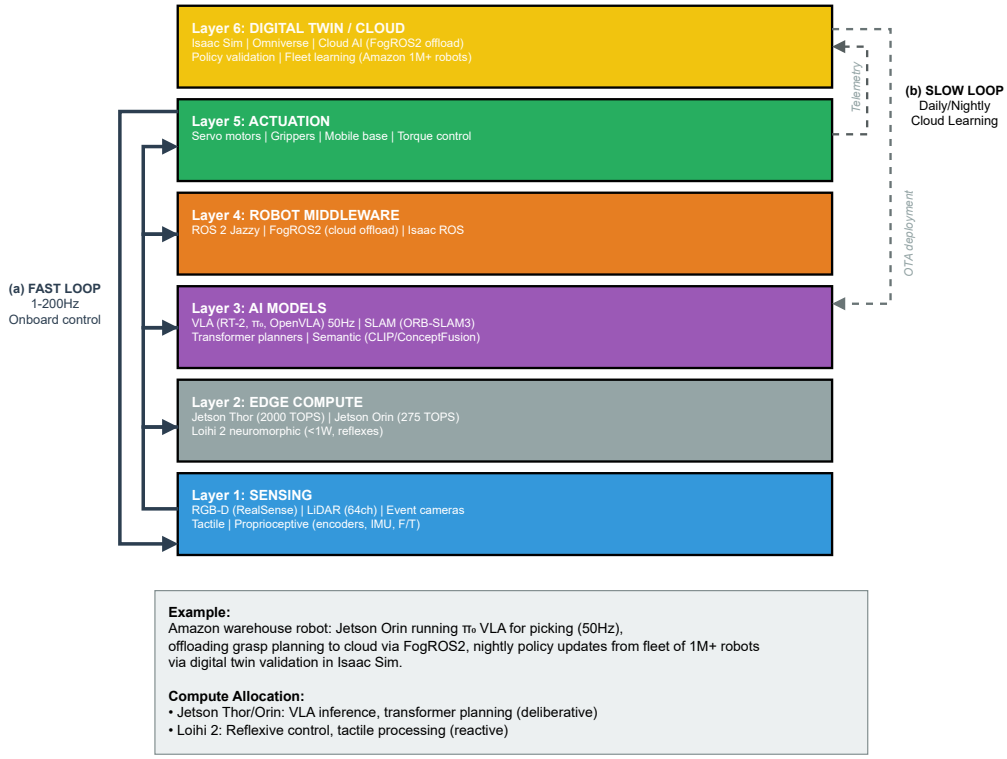

Figure 2. Physical AI hardware and compute stack. The six layers are read bottom-to-top: Sensing → Edge Compute → AI Models → Middleware → Actuation → Digital Twin/Cloud. **Left arrows** denote the fast onboard control loop (1–200 Hz); **right arrows** denote the slow cloud-based policy update loop (daily/nightly). Colour coding distinguishes layer function; labels within each layer list representative platforms.

2.4. Simulation Infrastructure and Digital Twins

Digital twins are now central to both hardware design and fleet operations; Section 3.5 provides the full treatment. The critical hardware constraint is a fidelity–scalability trade-off: high-physics-fidelity twins that model contact dynamics and sensor noise demand substantial GPU compute, limiting most organisations to twinning individual work cells rather than entire facilities [40]. A second under-addressed problem is twin *drift*: as hardware wears and layouts shift, the virtual replica diverges from reality, yet no standardised metric exists to quantify when a twin's safety guarantees have degraded below an acceptable threshold [34].

2.5. Middleware: ROS 2 and Alternatives

ROS 2 is now viable as fleet-scale middleware: its `ros2_control` framework supports over 50 platforms [41], and containerised deployment enables cloud-native fleet management [42]. However, it cannot yet replace proprietary hard-real-time stacks for sub-millisecond control loops such as force-controlled insertion. Most production robots therefore run a dual-stack architecture: ROS 2 for perception, planning, and fleet orchestration atop a vendor-locked real-time bus for actuation, with an impedance mismatch at the boundary that complicates debugging, certification, and latency accounting [43]. Closing this gap, either by pushing deterministic guarantees into the open stack or by standardising narrow bridge interfaces between the real-time and best-effort domains, remains a first-order middleware challenge.

2.6. Hardware-Accelerated Perception Pipelines

GPU-accelerated perception libraries such as NVIDIA's Isaac ROS [44,45] provide CUDA-accelerated SLAM and pose estimation integrated with ROS 2, achieving up to 10× throughput via zero-copy transfer [46]. Vendor-agnostic alternatives—Intel OpenVINO, Google Intrinsic, and community packages such as MoveIt—remain critical because NVIDIA dominates multiple layers of

 doi:10.20944/preprints202604.0549.v1

the Physical AI stack (simulation, edge compute, perception, world models), creating supply-chain fragility when a single vendor's roadmap changes; Section 8 discusses these concentration risks in detail.

3. World Modeling and Reasoning

The sensors and compute platforms described in previous sections generate vast streams of multimodal data—point clouds from LiDAR, RGB-D images, force-torque readings, and proprioceptive measurements—at rates exceeding gigabytes per second. Yet raw data alone does not enable intelligent action. A robot must synthesize these observations into structured internal representations: maps that encode navigable space, object models that capture affordances and semantics, and physics models that predict the consequences of actions. This transformation from perception to understanding is the domain of world modeling.

World modeling encompasses mapping, localization, intuitive physics, and semantic understanding, combining classical probabilistic methods with foundation models to build actionable representations for planning and decision-making in dynamic environments.

3.1. Mapping and Localization

Simultaneous Localization and Mapping (SLAM) algorithms construct a map of the environment while estimating the robot's pose within it. Classic SLAM approaches combine LiDAR, cameras and inertial measurements using probabilistic filters such as extended Kalman filters and particle filters [47]. These methods excel at incrementally building consistent maps for navigation tasks. Modern visual SLAM systems leverage deep neural networks to jointly estimate depth and camera pose, achieving real-time performance even in challenging lighting conditions [48,49]. ORB-SLAM3 [48] supports monocular, stereo, RGB-D, and visual-inertial configurations, making it versatile for warehouse mobile robots and delivery drones. DROID-SLAM [49] uses learned optical flow and bundle adjustment to surpass classical methods on difficult sequences. Radar-based SLAM is gaining traction in industrial settings because radio waves penetrate dust, fog, and smoke [50], enabling reliable navigation in factories and warehouses where visibility varies.

3.2. 3D Scene Understanding and Semantic Mapping

Beyond geometry, robots must recognize objects, affordances and semantics to interact intelligently with their environment. Vision-language models (VLMs) such as CLIP [14] allow robots to ground natural-language instructions in perception by learning visual concepts from text supervision. GPT-4V [51] extends this capability to complex scene understanding, enabling queries like "find the red bin on the top shelf" without task-specific training. Recent work such as ConceptFusion [52] fuses CLIP features into 3D semantic maps, supporting open-vocabulary queries in real time. Foundation VLA models such as π_0 [53] take this further by integrating VLM perception directly with motor control, enabling cross-embodiment manipulation from text prompts (Section 4). In warehouse environments, Fetch Robotics uses VLM-based semantic mapping to identify and pick novel SKUs without retraining, adapting to inventory changes on the fly. Similarly, BMW employs semantic bin picking systems that combine RGB-D depth with VLM scene understanding to achieve human-level versatility in assembly tasks.

3.3. Intuitive Physics and Causal Reasoning

Physical intelligence requires an understanding of how forces, friction and dynamics affect objects. Traditional robotics relies on explicit physics engines such as MuJoCo [13] and PyBullet [54] for simulation and control. MuJoCo provides efficient contact simulation using a convex approximation that enables fast, stable computation, though its soft contact model simplifies certain physical phenomena such as material deformation history and complex friction dynamics. PyBullet's open-source accessibility has made it a staple for research prototyping. Deep learning methods can learn implicit representations of physics from data without hand-crafted models [55]. Graph neural networks, for

 doi:10.20944/preprints202604.0549.v1

instance, predict object interactions and trajectory outcomes purely from visual observations. Liquid Time-constant Networks (LTCs) [56], developed by Hasani, Lechner, Amini, Rus, and Grosu across MIT, IST Austria, and TU Wien, are time-continuous neural models that adapt their dynamics based on input, enabling generalization to new environments and conditions. In drone navigation experiments, liquid network-powered systems demonstrated robust generalization to previously unseen seasonal and geographic conditions [57], outperforming standard recurrent neural networks. Manufacturing applications leverage predictive physics models to plan manipulation sequences for deformable objects—such as cable routing or fabric handling—where classical rigid-body assumptions fail.

3.4. Planning and Decision-Making

World models feed into planning algorithms that determine how to achieve goals. Classical planning uses graph search over discrete states; for example, A* [58] finds optimal paths in known environments by combining cost-to-come and heuristic estimates. Sampling-based methods such as Rapidly-exploring Random Trees (RRT) [59] excel in high-dimensional continuous spaces, making them ideal for robot arm motion planning. Trajectory optimization approaches like CHOMP [60] refine paths by minimizing a cost functional that balances smoothness, obstacle avoidance and dynamics constraints. Model-based reinforcement learning [61] leverages learned dynamics to plan in continuous spaces, simulating candidate actions internally before execution. Hierarchical approaches [62] decompose long-horizon tasks into subgoals; for example, a high-level policy might decide that a box should be moved from a shelf to a table, while a low-level controller handles the grasping and transport maneuvers.

Recent surveys of embodied AI categorize motion planning methods into four groups [27]:

- **Hierarchical planners** break down tasks into abstract skills and low-level execution, yielding interpretable plans and allowing reuse of primitives across tasks. However, hierarchical planners may struggle with continuous refinement when subtask boundaries are ambiguous.
- **Optimization-based planning** formulates motion generation as a constrained optimization problem. Methods such as RRT [59] and CHOMP [60] can find feasible trajectories while satisfying dynamics and collision constraints. These approaches provide smooth trajectories but require accurate models and can be computationally expensive. ABB's collaborative robots use CHOMP-based planning to safely navigate shared workspaces with human operators.
- **Morphology-based planning** accounts for changes in the robot's physical configuration. Some tasks benefit from altering morphology—for example, a robot that can switch between walking and rolling or adjust limb length. Morphology planning has been explored with adaptive shape-shifting robots and multi-modal locomotion.
- **Vision- or transformer-based planning** leverages large neural networks for end-to-end policy generation. CLIPort [63] combines CLIP's semantic understanding with Transporter Networks to map visual and language inputs directly to spatial action primitives for tabletop manipulation. Vision-language transformers operate on latent representations rather than explicit state spaces and are trained with large datasets, enabling generalization across tasks and environments.

Each category trades off interpretability, sample efficiency, and generalization. Combining the strengths of different planners—for instance, using a high-level vision-language planner to propose candidate behaviors and an optimization-based planner to refine trajectories—remains an active research direction. In warehouse settings, Amazon uses hierarchical planning to coordinate fleet-wide robot traffic while individual robots run local optimization for obstacle avoidance. Figure 3 illustrates how these planning components integrate with perception, world modeling, and action execution in a complete Physical AI pipeline.

 doi:10.20944/preprints202604.0549.v1

Table 3. Comparison of planning approaches for Physical AI systems. Each method offers distinct trade-offs for interpretability, computational cost, and generalization.

Planning Method	Key Strengths	Main Limitations	Primary Applications
Hierarchical	Interpretable plans, reusable skills	Ambiguous subtask bounds	Task decomposition, long-horizon
Optimization-based (RRT, CHOMP)	Smooth trajectories, constraints	Accurate models required, compute-heavy	Manipulation, motion planning
Morphology-based	Adaptive form, multi-modal locomotion	Complex control	Shape-shifting, terrain adaptation
Vision-Transformer (CLIPort)	End-to-end learning, generalizes	Black-box, data-hungry	Vision-language, manipulation

Figure 3. World modeling pipeline for Physical AI systems. **Arrows** denote data flow from sensory input (left) through four complementary representation types—geometric (maps, point clouds), semantic (scene graphs, VLM features), physics-based (dynamics models, causal graphs), and digital twin (virtual replicas, fleet-level simulation)—to planning and action execution (right). The **gear icon** denotes physics-based reasoning (dynamics engines); the **cloud/loop icon** denotes the digital twin layer (simulation and fleet optimisation). Feedback arrows show how execution outcomes refine each representation.

3.5. Simulation and Digital Twins

Simulation remains essential for validating robot behaviors at scale and producing synthetic training data without physical risk or cost. Modern digital twin pipelines combine photorealistic rendering, accurate physics, and tight integration with robotics middleware. NVIDIA Isaac Sim 5.0 [64], released as open source in 2025 and built atop the Omniverse platform, exemplifies this approach: USD-based scene composition [65] allows CAD designers, controls engineers, and ML practitioners to collaborate on a single source of truth, while PhysX [66] supplies contact-rich dynamics and native ROS 2 Jazzy interfaces ease deployment. Amazon Robotics maintains operational twins of fulfilment centres for rehearsing workflows before fleet-wide rollout [67]; deployment outcomes are discussed in Section 7.

Siemens' Xcelerator platform provides comprehensive digital twins that integrate mechanical design, electrical systems, and multi-physics simulation with robotics [68]. The platform incorporates physics-based human models for safety simulation, enabling virtual commissioning of factory systems before physical deployment. In 2024, Siemens expanded its partnership with NVIDIA, certifying industrial PCs for GPU-accelerated automation tasks from robotics to quality inspection [69].

 doi:10.20944/preprints202604.0549.v1

Academic simulators complement these commercial stacks: Habitat 3.0 [70] targets indoor embodied tasks with dynamic human avatars, MuJoCo [13] and CARLA [71] remain standard benchmarks for contact-rich manipulation and autonomous driving respectively, and open tools such as Gazebo [72] and PyBullet [54] lower the entry barrier. Across all platforms, key trends include differentiable physics for gradient-based learning, libraries of "SimReady" assets encoding physical materials, and tight coupling with dataset curation pipelines so that real-world logs continuously refine the twin.

Three architectural fault lines emerge across this design space. The first is a *fidelity–throughput trade-off*: platforms that invest in photorealistic rendering and multi-physics coupling achieve high visual and dynamic realism but struggle to generate the millions of rollouts per hour that reinforcement learning demands, whereas lightweight contact engines sacrifice material accuracy for speed. The second is a *scope–openness tension*: vertically integrated commercial stacks provide end-to-end workflows from CAD import through synthetic data export, yet their proprietary scene graphs and asset formats raise lock-in concerns; conversely, open-source engines offer extensibility and reproducibility but fragment the toolchain, requiring researchers to stitch together rendering, physics, and middleware bridges themselves. The third is a *single-agent–fleet gap*: most simulators model one robot interacting with a static or scripted environment, but real-world deployments require coordinating dozens of heterogeneous agents sharing dynamic spaces with humans. Notably absent is a principled *sim-to-real feedback loop*: no existing simulator automatically quantifies its own reality gap, ingests field-deployment logs, and re-calibrates its physics, rendering, or traffic models without manual effort.

Differentiable rendering enables a complementary class of inverse problems: optimising scene parameters (geometry, materials, lighting) through gradient descent to match target observations. Systems such as Mitsuba 3 [73] support physically accurate light transport with automatic differentiation, enabling three key Physical AI use cases: synthetic data generation with accurate reflectance, scene reconstruction that recovers material properties classical SLAM cannot, and sim-to-real transfer by gradient-based minimisation of the visual gap between rendered and real sensor observations [74,75]. As differentiable physics and rendering mature, their integration promises to accelerate sim-to-real gap closure across the pipeline.

3.6. World Foundation Models

World foundation models (WFMs) extend the capabilities of Physical AI systems by providing predictive models of how the world evolves under different actions. Rather than relying solely on hand-engineered physics or low-level policies, WFMs offer a learned "imagination" layer that allows robots to forecast the consequences of candidate plans, evaluate safety, and adapt to new environments with limited fine-tuning. These models complement the perception and control stacks by supplying long-horizon reasoning signals grounded in multimodal data.

NVIDIA's Cosmos platform [76] exemplifies deployable WFMs, combining action-conditioned video generation, sim-to-real domain adaptation (boosting navigation success from 54% to 91% in hybrid pipelines [77]), and physics-grounded commonsense reasoning [78]. Google's Genie 2 pursues similar goals with interactive environment generation from single images.

However, a critical "Gen2Real gap" persists across all current WFMs: generative models produce statistically plausible but physically inconsistent outputs—objects vanish under occlusion, gravity is violated, and scaling model size improves visual fidelity without reliably improving physical correctness [76,77]. Hybrid pipelines that combine generative augmentation with physics-based simulation are emerging as the practical response.

The Genesis simulator [79] represents a complementary approach: an open-source generative physics engine combining learned generative models with physics-based simulation. This hybrid neuro-symbolic approach aims to reduce hallucinations in world models while maintaining generative diversity, enabling safer deployment of predictive models in safety-critical applications such as warehouse automation and collaborative robotics. As WFMs mature, the integration of predictive world modeling with system-level digital twins—such as discrete-event simulators used in manufac-

 doi:10.20944/preprints202604.0549.v1

turing—will enable end-to-end optimization of robot fleet deployments, from layout design to failure recovery strategies.

3.7. Systems-Level Simulation and Operations Optimization

At the plant level, discrete-event simulation platforms such as Simio [80] complement robotics-specific simulators by modelling whole-plant workflows. Procter & Gamble's "Control Tower" virtual twin reduced deadhead truck movements by approximately 15% [81], illustrating how coupling fleet-level digital twins with robot-level simulation enables end-to-end optimisation from layout design to failure recovery. As WFMs mature, the next frontier is real-time predictive control: running inference at 10+ Hz during deployment so that model-predictive loops can forecast the consequences of candidate actions milliseconds before execution. The critical barrier remains verification—generative world models may hallucinate physically impossible outcomes, requiring hybrid architectures that pair fast physics-based simulation for short-horizon dynamics with generative models for longer-horizon semantic predictions.

4. Perception, VLMs/VLAs, and Generalist Policies

World models provide robots with structured representations of spatial layout, object properties, and physical dynamics—the geometric and mechanical understanding necessary for navigation and manipulation. Yet this traditional approach relies on explicit state representations and hand-crafted perception pipelines that struggle with open-ended reasoning: How should a robot respond to the instruction "tidy the kitchen"? What does "fragile" imply about grasping strategy? Which object is "the red one on the left" when multiple candidates exist? Answering such questions requires not just geometric understanding but semantic reasoning grounded in language and visual context. This capability emerges from vision-language-action models that unify perception, language understanding, and control within a single neural architecture.

Large language and vision-language models are increasingly embedded throughout robotics stacks: they translate natural-language goals into symbolic procedures, ground perception into action, and even generate code that stitches together traditional controllers. By tapping internet-scale priors, these models move robots beyond scripted behaviours toward systems that communicate, interpret intent, and adapt in situ. The same flexibility, however, introduces new risks around hallucinated plans, unsafe tool use, or brittle performance when the model faces embodiments it has never seen.

Addressing these challenges requires tight integration with the components discussed in earlier sections—world models, digital twins, and safety monitors. Rich simulation environments supply the counterfactual experience necessary to fine-tune large models, while runtime guardrails constrain their suggestions to verifiable behaviours. Within this section we review three complementary trends: language models for planning, multimodal models that join perception and action, and generalist policies trained across diverse robot fleets. For a comprehensive architectural survey of VLA models covering vision encoders, language models, action decoders, and training methodologies, we refer readers to Kawaharazuka et al. [26], which provides systematic analysis of modality integration techniques and cross-embodiment learning paradigms.

4.1. Perception Foundations

Before language can ground action, robots must parse raw sensor streams into structured scene representations. Three developments have reshaped robotic perception. First, *open-vocabulary detection and segmentation*: foundation models such as Segment Anything (SAM) [82] and Grounding DINO [83] enable robots to localise and segment arbitrary objects from natural-language queries without task-specific training, replacing hand-crafted detection pipelines. Second, *3D scene representations*: neural radiance fields and 3D Gaussian splatting now serve as compact, differentiable world representations for manipulation planning; Distilled Feature Fields (F3RM) [84] embed CLIP features into NeRF volumes for few-shot 6-DOF grasping, while Gaussian splatting variants enable real-time navigation with safety guarantees [85]. Third, *point cloud backbones*: PointNet++ [86] remains a widely used

 doi:10.20944/preprints202604.0549.v1

architecture for 3D object recognition in manipulation, though transformer-based alternatives are rapidly closing the accuracy gap. These perception modules provide the visual grounding that downstream VLA models rely on for open-ended task execution.

4.2. LLMs for Robot Planning

One of the most profound shifts in robotics is the integration of large language models (LLMs) into the decision-making loop. Classical controllers were hard-coded or trained on narrow datasets; in contrast, LLM-based planners exploit broad commonsense priors to reason about new tasks and environments. Early experiments such as PIGLeT combined text-driven reasoning with simulated household environments, showing that language-conditioned agents can acquire intuitive physics and object semantics without exhaustive manual programming.

PaLM-SayCan [15] popularised the idea of pairing a pre-trained language model with a grounded affordance model. The LLM decomposes a natural-language goal into candidate high-level skills, while the affordance model evaluates whether the robot can execute each skill given its sensors, actuators, and current state. Filtering proposed plans through feasibility scores helps maintain safety—commands that would violate kinematic limits or required preconditions are discarded before reaching low-level controllers. Variants of this architecture now support long-horizon task planning, hierarchical control, and interactive replanning as the environment changes.

Interactive dialogue and code as action.

Another branch of work treats the LLM as a just-in-time programmer. Systems such as Code-as-Policies [16] prompt the model with API documentation and a textual task description, allowing it to emit Python or ROS code that stitches together existing primitives. Chain-of-thought prompting improves reliability by forcing the LLM to articulate intermediate reasoning and safety checks. Human overseers can inspect, edit, or veto generated code before execution, creating a collaborative loop in which the model proposes novel behaviours while operators retain ultimate authority. Although these pipelines remain semi-autonomous, they point toward robots that can extend their own repertoires on demand by composing libraries of perception, planning, and control routines.

4.3. Vision-Language-Action Models

Vision-language models (VLMs) address the grounding problem in robot planning by integrating visual perception with language understanding, enabling more accurate and context-aware control. OpenAI's GPT-4V [51] demonstrates vision-language capabilities in embodied settings: researchers have shown that GPT-4V can interpret visual scenes and generate natural language instructions translatable into robot actions through code-as-policies frameworks. The Robotic Vision-Language Planning (ViLa) approach [87] leverages GPT-4V to generate sequences of actionable steps based on visual observations and high-level language instructions, integrating perceptual information directly into reasoning and planning. PaLM-E [18], an embodied multimodal language model, integrates the PaLM LLM with multiple sensor encoders—including Vision Transformers (ViT), Object Scene Representation Transformers (OSRT), and state estimation vectors—enabling rich scene interpretation and natural language instruction generation for downstream controllers.

Vision-Language-Action (VLA) models represent a further evolution, unifying perception, instruction, and control within a single architecture. Google's Robotics Transformer series exemplifies this progression: RT-1 [88] demonstrated single-task learning from demonstrations, while RT-2 [17] extended the approach by co-training on web-scale visual and textual data alongside robotic trajectory data. This integration enables RT-2 to generalize to unseen tasks and objects with minimal fine-tuning, leveraging analogies from internet-scale priors to execute novel instructions without explicit training. Open-source efforts have accelerated VLA research: OpenVLA-7B [89] provides a 7-billion parameter model achieving strong cross-embodiment performance; Octo [90] explores generalist policies for manipulation; RDT-1B [91] introduces diffusion-based action prediction for smooth trajectory synthesis; ManipLLM [92] and ReasonManip [93] investigate object-centric manipulation and multimodal

 doi:10.20944/preprints202604.0549.v1

reasoning. These systems demonstrate that VLA architectures can follow natural instructions and adapt to new environments through efficient fine-tuning on limited robot-specific data.

4.4. Generalist Agents and Cross-Embodiment Policies

The transition from task-specific controllers to generalist agents marks a fundamental shift in robotics. DeepMind's Gato [20] provided early evidence by training a single 1.2B-parameter transformer on 604 tasks spanning text, images, and robot control. Despite modest per-task performance, Gato demonstrated that cross-modal transfer—visual understanding improving motor control—is achievable when training diversity is sufficient.

Building on this foundation, Physical Intelligence introduced π_0 (pi-zero) as a robot foundation model specifically designed for continuous control across multiple embodiments [19]. The architecture employs a flow matching approach built upon PaliGemma (3 billion parameters) augmented with 300 million additional parameters for continuous action generation at 50 Hz. Unlike discrete action models that output categorical decisions, π_0's flow matching enables smooth, real-time trajectory generation essential for dexterous manipulation. Training on data from seven robotic platforms across 68 unique tasks—including autonomous laundry folding and cardboard box assembly—the model demonstrates cross-embodiment transfer: skills learned on one robot morphology transfer to different hardware configurations without architecture modification. The February 2025 open-source release of π_0 and its successor π_0.5 democratized access to state-of-the-art robotic intelligence, with π_0.5 extending generalization capabilities to entirely novel environments not represented in the training distribution [94].

NVIDIA's GR00T N1 [95] targets humanoid embodiments specifically: a 3B-parameter model combining a frozen Eagle vision-language backbone with a flow-matching action transformer, trained on a mixture of real humanoid demonstrations, synthetic trajectories from the GR00T-Dreams simulation blueprint, and internet-scale video. The successor GR00T N1.5 triples task success rates over N1 on the DreamGen benchmark (38.3% vs. 13.1% across 12 manipulation tasks), demonstrating that synthetic data generated in Isaac Sim and augmented with Cosmos Transfer can substantially reduce the months-long data-collection campaigns that have historically bottlenecked humanoid learning.

Figure AI's Helix [96] addresses the latency-capability trade-off through a dual-system architecture: System 2 (a 7B VLM at 7–9 Hz) handles scene understanding and planning, while System 1 (a visuomotor policy at 200 Hz) generates precise whole-body motor commands. This decoupling enables zero-shot manipulation of novel objects from natural language prompts while running entirely on embedded GPUs—a prerequisite for economically viable humanoid deployment. Helix-powered robots are piloting material handling at BMW Spartanburg, demonstrating industrial readiness.

Google DeepMind's Gemini Robotics [97], built on Gemini 2.0, is the broadest VLA platform reported to date, doubling prior VLA performance on generalisation benchmarks. Its successor, Gemini Robotics 1.5 [98], introduces transparent reasoning that surfaces intermediate decision steps, cross-embodiment motion transfer that enables skills learned on one platform to transfer to different hardware, and an Embodied Reasoning (ER) variant providing state-of-the-art spatial understanding. Partnerships with Boston Dynamics, Apptronik, and Agility Robotics signal rapid transition from research to commercial deployment across leading hardware platforms.

These developments indicate convergence toward unified architectures that integrate perception, reasoning, and action. The success of dual-system architectures—combining fast reactive control with slower deliberative reasoning—suggests this paradigm may become dominant for general-purpose robotics, analogous to how transformer architectures unified natural language processing. Figure 4 illustrates the evolution of VLA model architectures from early single-task systems through current generalist agents, showing the progression in model scale, training data diversity, and control capabilities that enable cross-embodiment deployment. Table 4 provides a systematic architectural comparison across these models, revealing the design choices that differentiate each system.

 doi:10.20944/preprints202604.0549.v1

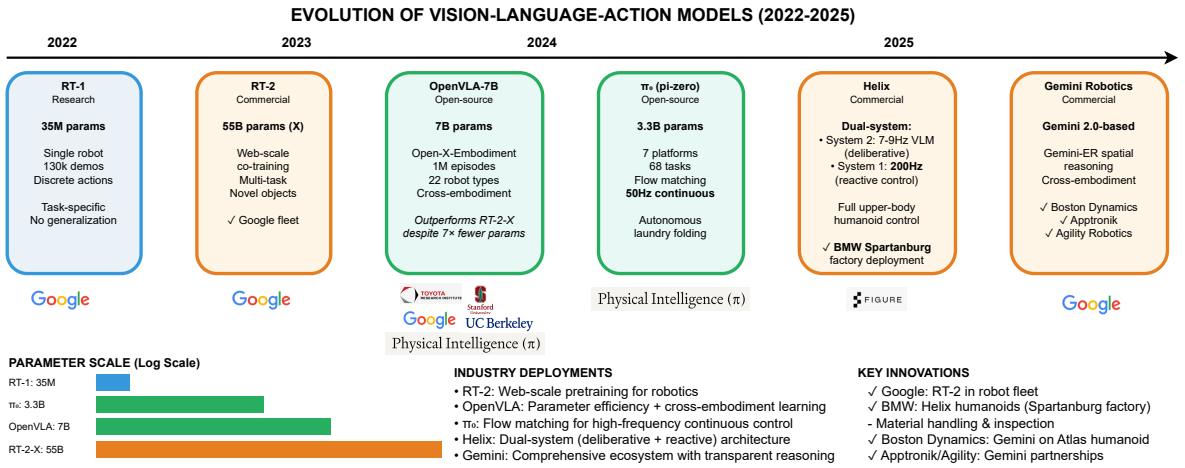

Figure 4. Evolution of Vision-Language-Action model architectures from 2022 to 2025, illustrating the progression from single-task systems to generalist agents capable of cross-embodiment control. The timeline shows increasing model scale, training data diversity, and control sophistication enabling industrial deployment.

Table 4. Architectural comparison of representative VLA models (2022–2025), ordered chronologically within three development phases. The table highlights the progression from discrete-action, single-embodiment systems to continuous-control, cross-embodiment generalist agents. Freq. denotes effective control frequency; Emb. is the number of distinct robot morphologies used during training.

Model	Params	Vision Enc.	LM Backbone	Action Decoder	Hz	Training Scale	Emb.	Open
Early VLA Systems (2022–2023)								
Gato [20]	1.2B	ResNet	Unified xfmr	Autoregressive tokens	—	604 tasks (multi-modal)	—[a]	No
RT-1 [88]	35M	EffNet-B3	FiLM cond.	Discrete (256 bins)	3	130K eps, 700+ tasks	1	Yes
RT-2 [17]	55B	ViT-22B	PaLI-X	Discrete text tokens	1–3	130K eps + web VLM	1	No
Open-Source Cross-Embodiment Models (2024)								
Octo [90]	93M	CNN patches	T5-Base	Diffusion (MLP)	—	800K eps (OXE)	9	Yes
OpenVLA [89]	7B	DINOv2+SigLIP	Llama 2	Discrete tokens	5–6	970K eps (OXE)	22	Yes
RDT-1B [91]	1.2B	SigLIP	T5-XXL	Diffusion xfmr	6	1M+ eps, 46 datasets	46	Yes
Industrial Generalist Agents (2024–2025)								
π_0 [19]	3.3B	SigLIP	Gemma 2B	Flow matching	50	10K hrs, 68 tasks	7	Yes
GR00T N1 [95]	3B	SigLIP-2 (Eagle)	SmolLM2	DiT + flow match	10/120[b]	7.4K hrs (real+sim)	5+	Yes
Helix [96]	∼7B	Undisclosed	7B VLM	Dual-system[c]	9/200[b]	500 hrs teleop	1	No
Gemini Rob. [97]	—	Gemini 2.0	Gemini 2.0	Cloud + local dec.	50	Thousands of hrs	3+	No

[a]Gato spans games, text, and limited robot control; included as the first generalist agent, not a dedicated VLA. [b]VLM / action-loop frequencies for dual-system architectures. [c]System 2 (7B VLM, 7–9 Hz) for planning; System 1 (80M visuomotor, 200 Hz) for motor control.

Design tradeoff analysis.

Three architectural trends emerge from this comparison. First, the action decoder has evolved from discrete tokenisation (RT-1/2, OpenVLA) through diffusion models (Octo, RDT-1B) to flow matching (π_0, GR00T N1): discrete tokens leverage pretrained language-model weights but cap control bandwidth at 3–6 Hz; diffusion enables multi-modal action distributions at the cost of inference latency; flow matching provides continuous, low-latency generation at 50+ Hz. Second, dual-system architectures have emerged independently in Helix and GR00T N1, decoupling slow deliberative reasoning (VLM at 7–10 Hz) from fast reactive control (visuomotor loop at 120–200 Hz)—an echo of the System 1/System 2 distinction from cognitive science that appears necessary for contact-rich manipulation where millisecond-level responses are critical. Third, parameter efficiency consistently outweighs raw scale: OpenVLA (7B) surpasses RT-2-X (55B) with 7× fewer parameters, while Octo (93M) provides competitive generalisation at a fraction of the compute, suggesting that training data diversity across embodiments matters more than model size alone. Notably, six of the ten models are fully open-source, enabling rapid community iteration and lowering the barrier to entry for new research groups.

 doi:10.20944/preprints202604.0549.v1

4.5. Benchmarking and Evaluation

Systematic benchmarking across 20 Open-X-Embodiment datasets reveals that parameter efficiency outweighs raw scale: OpenVLA (7B) outperforms RT-2-X (55B) by 16.5% with 7× fewer parameters and exceeds Diffusion Policy by 20.4% [89,99]. GPT-4o achieves the most consistent cross-task performance through prompt engineering, but all models struggle with tasks requiring multi-step planning or adaptation to significant environmental changes [99]. The field has grown from a handful of VLA systems to over 45 specialised implementations between 2022–2025 [28], yet standardised metrics for safety, interpretability, and failure recovery remain underdeveloped—aspects critical for real-world deployment but underrepresented in current evaluation frameworks.

5. Learning in Physical AI

The previous section examined state-of-the-art vision-language-action models such as RT-2, π_0, and Gemini Robotics—systems that can follow natural language instructions, manipulate diverse objects, and generalize across embodiments. These capabilities did not emerge from hand-coded rules but from learning: algorithms that enable robots to acquire skills through interaction, demonstration, and experience. Understanding how these models achieve such performance requires examining the learning paradigms that underpin them.

Learning enables physical AI systems to adapt, generalize and improve over time without explicit reprogramming. Major paradigms include reinforcement learning (RL), imitation learning, self-supervised learning, and hybrid methods. Modern deployments increasingly combine multiple learning approaches to balance sample efficiency, robustness, and generalization. This section surveys key algorithms, benchmarks, and real-world deployments that demonstrate how learning transforms robots from executing fixed routines into adaptive intelligent agents.

5.1. Reinforcement Learning

Reinforcement learning has produced some of the most striking demonstrations in Physical AI—from 96% grasp success across novel objects [100] to dynamic parkour on low-cost quadrupeds [101]—yet its impact on deployed systems remains narrower than these headline results suggest. The field has evolved from model-free methods that require millions of interactions toward model-based and offline approaches that drastically reduce sample requirements. Table 5 summarises the principal algorithms and their trade-offs; below we analyse the deployment patterns and bottlenecks driving this evolution.

Table 5. Deep RL algorithms for robotic control. The trajectory from model-free to model-based and offline methods reflects the field's response to sample-complexity bottlenecks in real-world deployment.

Algorithm	Type	Efficiency	Stability	Key Innovation & Limitations
PPO [12]	On-policy	Medium	High	Clipped surrogate objective; high sample cost limits real-robot use
SAC [102]	Off-policy	High	High	Maximum-entropy exploration; sensitive to reward shaping
TD3 [103]	Off-policy	High	Medium	Twin Q-networks, delayed updates; supersedes DDPG [104]
QT-Opt [100]	Off-policy	High	Medium	Fleet-scale Q-learning; requires >500k real interactions
TD-MPC2 [105]	Model-based	Very High	High	Latent trajectory optimisation; single 317M-param agent learns 80 tasks
Q-Transformer [106]	Offline	High	High	Autoregressive Q-learning over action tokens; offline data only
Diffusion-QL [107]	Offline	High	High	Diffusion model as policy class; SoTA on D4RL benchmarks

Critical analysis.

Early model-free methods (PPO, SAC, TD3) established that deep RL can solve individual robotic tasks, but their sample complexity—typically 10^5–10^7 environment steps per task—makes multi-task real-world training impractical without simulation or fleet-scale data collection. PPO achieves approximately 90% success on simulated navigation benchmarks and hybrid IL+RL methods reach ∼95% on AGV navigation [108], but these results rarely transfer to unstructured settings without significant

 doi:10.20944/preprints202604.0549.v1

adaptation. This bottleneck has driven three responses: (i) model-based methods like TD-MPC2 that learn latent dynamics, achieving strong performance across 104 tasks with a single hyperparameter set [105]; (ii) offline RL methods like Q-Transformer that leverage large demonstration datasets without requiring online interaction [106]; and (iii) diffusion-based policies that combine generative model expressiveness with value-guided optimisation, achieving state-of-the-art performance on the majority of D4RL tasks [107]. Standardised benchmarks—RoboSuite [109] for manipulation and Meta-World [110] for multi-task learning—have accelerated algorithmic comparison, but the community still lacks widely adopted real-world evaluation protocols that capture the full complexity of deployed Physical AI.

5.1.1. Scaling RL: Fleet Learning and Closed-Loop Control

Kalashnikov et al.'s QT-Opt [100] demonstrates that large-scale off-policy RL can achieve 96% grasp success on unseen objects using over 580,000 real-world attempts collected autonomously by a fleet of seven robots. Unlike pipeline grasping systems, QT-Opt performs closed-loop control—continually updating its strategy from live camera observations—and learns emergent behaviours such as regrasping and object repositioning without explicit programming. The key insight that off-policy replay enables continuous improvement from fleet data has since been adopted in industrial picking systems that learn from millions of daily interactions.

5.1.2. Multi-Robot Coordination and Multi-Agent Learning

Fleet learning scales single-agent policies across robot replicas that share a training objective; multi-robot coordination poses a fundamentally harder problem in which heterogeneous agents must learn complementary behaviours while communicating under bandwidth and latency constraints. The dominant paradigm is *centralised training with decentralised execution* (CTDE): during training a shared critic or mixing network has access to the joint state, but each agent conditions only on its own observations at deployment. QMIX [111] introduced monotonic value decomposition that factors the team reward into per-agent utilities, enabling tractable off-policy learning for cooperative tasks; MAPPO [112] later showed that a simpler approach—independent PPO with a shared centralised critic—matches or exceeds value-decomposition methods on standard benchmarks, lowering the implementation barrier for robotics applications. Communication adds another dimension: graph neural networks enable agents to share spatial features selectively, improving collaborative perception in cluttered or occluded environments [113]. More recently, foundation models are being applied to multi-robot task allocation. COHERENT [114] uses an LLM-based proposal-execution-feedback loop to decompose long-horizon goals into subtasks for heterogeneous fleets of quadrotors, legged robots, and manipulators, surpassing prior planners in success rate on a 100-task benchmark. A comprehensive survey of LLM integration into multi-robot systems [115] categorises emerging work across task allocation, motion planning, and human intervention, noting that scalability to dozens of agents and real-time replanning remain open challenges. These multi-agent techniques complement the fleet-scale data pipelines discussed in Section 6 and the warehouse deployments in Section 7, where coordinating hundreds of mobile units in shared spaces is an operational necessity.

5.1.3. RL in Dynamic and Contact-Rich Settings

RL excels where analytical controllers struggle: highly dynamic or contact-rich tasks with under-specified models. A single neural policy trained entirely in simulation enables a low-cost quadruped to perform parkour—jumping onto obstacles twice its height and across gaps twice its length—through massive domain randomisation [101]. Liquid Time-Constant Networks [56] trained via RL allow drones to generalise across seasons and terrain without retraining [57]. For manipulation, RialTo [116] constructs digital twins from minimal real data, fine-tunes policies via RL with sparse rewards, and distils them back to the real robot, improving robustness by over 67% on contact-rich tasks such as stacking dishes while requiring minimal human supervision.

 doi:10.20944/preprints202604.0549.v1

5.1.4. Legged Locomotion and Whole-Body Control

Legged locomotion exemplifies how simulation-trained RL transfers to some of the most dynamic physical tasks. The dominant paradigm is *teacher-student distillation*: a privileged teacher policy with access to ground-truth terrain maps and contact forces is trained via RL in simulation, then distilled into a deployable student that relies solely on proprioceptive history. Miki et al. [117] demonstrated this on ANYmal, training an attention-based recurrent encoder that integrates proprioception and exteroception to traverse alpine trails, stairs, and rubble with zero-shot sim-to-real transfer. Rapid Motor Adaptation (RMA) [118] introduced an online adaptation module that infers latent environment parameters in fractions of a second, enabling a single policy to handle sand, mud, and deformable surfaces without explicit terrain classification. Building on these quadruped results, Hoeller et al. [119] extended the approach to agile parkour navigation at speeds up to 2 m/s on ANYmal, complementing the low-cost quadruped parkour results of Cheng et al. [101].

Humanoid locomotion has advanced rapidly: Radosavovic et al. [120] trained a causal transformer on proprioceptive history with large-scale model-free RL, achieving zero-shot sim-to-real bipedal walking across varied outdoor terrain with emergent human-like arm swing. Cheng et al. [121] extended this to whole-body control by leveraging human motion-capture data: the upper body imitates reference motions while the legs track velocity commands, enabling a full-sized humanoid to walk, dance, and shake hands. These locomotion advances underpin the humanoid deployment programmes discussed in Section 7, where platforms such as Figure 02 and Optimus must navigate factory floors while performing manipulation tasks.

5.1.5. Dexterous Manipulation and In-Hand Skills

Dexterous manipulation—grasping, reorienting, and using tools with multi-fingered hands—remains among the hardest contact-rich control problems because it demands simultaneous reasoning about friction, gravity, and high-dimensional finger coordination. Hardware advances have lowered the barrier to entry: the LEAP Hand [122] provides a fully actuated, anthropomorphic 16-DoF hand that can be assembled for $2,000 (roughly one-eighth the cost of the widely used Allegro Hand), while the Shadow Dexterous Hand offers 24 DoF for research requiring human-level kinematic fidelity. On the learning side, AnyRotate [123] trains a single sim-to-real policy for gravity-invariant multi-axis in-hand rotation by combining dense tactile sensing with domain randomisation, achieving zero-shot transfer on ten diverse objects. Data collection remains a bottleneck: DexCap [124] addresses this with a portable motion-capture glove that records human hand trajectories and retargets them to robot morphologies, enabling diffusion-policy training for bimanual dexterous tasks without teleoperation rigs. Together, these advances are closing the gap between coarse parallel-jaw grasping and the fine-grained manipulation required for assembly, tool use, and household tasks.

5.2. Imitation Learning and Learning from Demonstrations

Imitation learning (IL) trains policies from expert demonstrations, bypassing the need for reward engineering and enabling rapid bootstrapping of robot behaviors. Behavioral cloning—supervised learning from state-action pairs—is the simplest IL approach but suffers from distribution shift when the learned policy encounters states unseen during training. DAgger (Dataset Aggregation) [125] addresses this by iteratively collecting data under the learner's policy with expert corrections, stabilizing training for navigation and manipulation. Comprehensive surveys [126] document the spectrum of IL techniques, from kinesthetic teaching to passive observation of human activities.

Modern IL systems scale to real-world complexity by combining demonstrations with self-supervised refinement. Covariant's RFM-1 system learns to grasp over 10,000 SKU types in warehouse environments by starting from human demonstrations and continuously refining through autonomous interaction. Tesla's Optimus humanoid uses teleoperated demonstrations from human operators wearing motion-capture suits, learning assembly tasks deployed at BMW factories. Figure AI's Figure 02 combines IL with online RL to learn cabinet assembly, wire harness installation, and parts handling

 doi:10.20944/preprints202604.0549.v1

from a mix of demonstrations and autonomous practice. Recent work [127] emphasizes that data diversity matters more than quantity: policies trained on varied demonstrations generalize better than those trained on large but homogeneous datasets.

Hybrid IL+RL methods offer the best of both paradigms. RialTo [116] uses IL to initialize a policy from demonstrations, then refines it via RL to handle edge cases and recover from perturbations. This approach achieves 67% higher robustness than pure IL while requiring far fewer interactions than pure RL. In warehouse automation, Fetch Robotics employs similar hybrid strategies: human demonstrations provide coarse picking motions, while autonomous RL fine-tuning adapts to specific object properties and lighting conditions encountered during deployment.

Recent advances in visuomotor policy learning demonstrate remarkable sample efficiency. Diffusion Policy [128] represents policies as conditional denoising diffusion processes, achieving an average 46.9% improvement over prior methods across 12 manipulation tasks by gracefully handling multimodal action distributions in high-dimensional spaces. The Action Chunking Transformer (ACT) [129] predicts k-step action sequences rather than single actions, reducing the effective horizon and mitigating compounding errors; paired with low-cost ALOHA hardware (under $20,000), it achieves 80–90% success on dexterous bimanual tasks with only 10 minutes of demonstrations. Mobile ALOHA [130] extends this to whole-body mobile manipulation, reaching 90% success on household tasks with 50 demonstrations per task.

5.3. Self-Supervised and Unsupervised Learning

In self-supervised learning, robots create their own training labels by interacting with the environment, eliminating the need for manual annotation. Time-Contrastive Networks (TCN) [131] learn visual representations from multi-view robot videos by using temporal correspondence as supervision: frames close in time should have similar embeddings. TCN enables cross-embodiment transfer—a policy trained on one robot can transfer to another by aligning learned visual features. World Models [132] take this further by learning compact predictive models of environment dynamics, enabling agents to train in the latent space of the learned model rather than directly in the real world. This reduces the real-world interactions required, a critical advantage for expensive robotic hardware.

Self-supervised approaches [133] enable robots to autonomously generate training data through exploratory interaction. Examples include predicting the outcome of random actions, reconstructing missing sensory modalities (e.g., predicting tactile feedback from vision), or learning inverse models that infer which action led to an observed state transition. These techniques are particularly valuable for representation learning: unsupervised features often transfer better to downstream manipulation and navigation tasks than features trained on narrow supervised objectives. Tesla's Full Self-Driving system exemplifies large-scale self-supervised learning, extracting supervision from millions of miles of human driving to learn scene representations, object permanence, and trajectory prediction without manual labels. In warehouse settings, Amazon robots use self-supervised learning to discover stable grasp affordances by attempting thousands of grasps per day and learning which visual features predict success.

Learning from human video.

A promising route to data-scalable robot learning is to pre-train visual representations on large corpora of human activity video, then transfer them to robotic manipulation with minimal in-domain data. R3M [134] combines time-contrastive learning with video–language alignment on the Ego4D dataset, improving manipulation success by over 20% compared to training from scratch. VIP [135] casts representation learning as offline goal-conditioned RL, producing embeddings whose distances serve as dense visual reward signals—enabling few-shot real-robot learning from as few as 20 trajectories without task-specific fine-tuning. Voltron [136] unifies both paradigms through language-conditioned visual reconstruction and visually-grounded language generation, capturing low-level spatial features and high-level semantics simultaneously. Beyond frozen representations, UniSim [137] learns an interactive video world model from diverse datasets, allowing vision-language planners and RL policies to

 doi:10.20944/preprints202604.0549.v1

be trained entirely in a learned simulator and transferred zero-shot to real robots. These approaches are significant because they tap into orders-of-magnitude more data than robot-only corpora, offering a path toward closing the data gap identified in Section 5.

Table 6. Learning paradigms for Physical AI systems. Each approach offers distinct trade-offs between data requirements, sample efficiency, and the need for reward engineering.

Paradigm	Data Source	Efficiency	Reward	Best Suited For
Reinforcement Learning	Autonomous Interaction	Low	Required	Optimization, dynamics
Imitation Learning	Expert Demos	High	Not needed	Fast bootstrap, complex behavior
Self-Supervised	Autonomous Interaction	Medium	Not needed	Representation learning
Hybrid (IL + RL)	Demos + Interaction	Med-High	Minimal	Robust deployment

5.4. Sim-to-Real Transfer and Domain Adaptation

The reality gap—the discrepancy between simulated and real-world dynamics—remains a central challenge in deploying learned policies. Domain randomization [8] addresses this by varying visual rendering, object properties, and physics parameters during simulation training, producing policies robust to distribution shift. OpenAI demonstrated that policies trained with aggressive randomization in simulation transfer to real robots without any real-world fine-tuning. System identification [138] takes a complementary approach: using small amounts of real-world data to calibrate simulation parameters (friction coefficients, actuator dynamics, sensor noise) before policy training, enabling more accurate sim-to-real transfer with 10× less real data than purely real-world training.

Adversarial domain adaptation [139] learns to align simulated and real visual distributions through adversarial training, making policies invariant to visual appearance differences while preserving task-relevant information. Meta-learning methods such as MAML (Model-Agnostic Meta-Learning) [140] learn policy initializations that adapt quickly to new environments with only a few gradient steps, enabling rapid deployment across varied real-world conditions. MAML has been applied to robot pushing, grasping, and locomotion, demonstrating few-shot adaptation to novel objects, terrains, and lighting conditions.

Industrial deployments increasingly rely on the digital twin infrastructure described in Section 3.5 for sim-to-real transfer: policies are validated in high-fidelity replicas before reaching physical fleets, and live telemetry continuously recalibrates the twin as conditions drift. The trend is toward tight integration of simulation, deployment monitoring, and online adaptation. Figure 5 illustrates this cycle.

 doi:10.20944/preprints202604.0549.v1

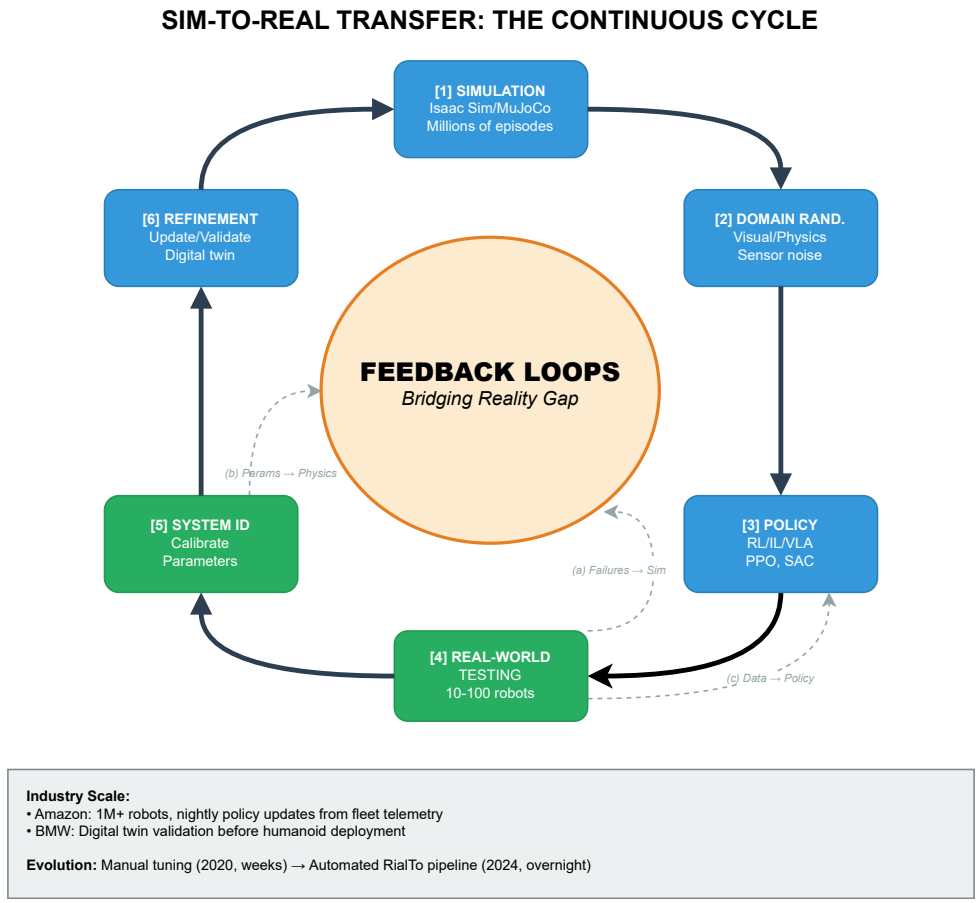

Figure 5. Sim-to-real transfer pipeline. **Solid dark arrows** (clockwise around the perimeter) trace the iterative cycle: Simulation Training → Domain Randomisation → Policy Learning → Real-World Testing → System Identification → Refinement, then back to simulation. **Dashed grey arrows** (pointing inward) denote three feedback signals that close the reality gap: (a) failure cases replayed in simulation, (b) calibrated parameters updating the physics model, and (c) aggregated fleet data retraining the deployed policy.

5.5. Generalization and Transfer Learning

Generalization—performing well on tasks and environments unseen during training—remains the ultimate test of Physical AI. Policies trained in narrow simulation environments often fail when deployed in the wild due to distributional shift: different lighting, novel object geometries, unexpected occlusions, or co-workers moving through the workspace. Techniques to improve generalization span the entire learning pipeline. Data augmentation and domain randomization broaden the training distribution. Multi-task learning trains a single policy on diverse tasks simultaneously, encouraging the emergence of reusable skills. Meta-learning optimizes for rapid adaptation rather than performance on any single task.

Digital twins and sim-to-real transfer, as used in RialTo [116], provide an efficient generalization strategy: by building a physics-accurate simulation of the specific deployment environment, policies fine-tuned in simulation can transfer back to the real world with high robustness. Curriculum learning—gradually increasing task difficulty during training—helps policies develop hierarchical skills that generalize better than end-to-end training on hard tasks. Foundation models (discussed in Section 4) offer another path: pre-training on massive diverse datasets followed by fine-tuning on specific deployment tasks. Waymo's autonomous driving system, for instance, combines simulation-based RL with millions of miles of real-world driving data and targeted fine-tuning for geographic regions, achieving robust generalization across cities, weather conditions, and traffic patterns.

 doi:10.20944/preprints202604.0549.v1

Continual learning and lifelong learning address generalization over time: how can a robot maintain performance on old tasks while learning new ones, avoiding catastrophic forgetting? Elastic Weight Consolidation and related techniques constrain updates to preserve critical parameters for previous tasks. In practice, industrial deployments often maintain task-specific policy snapshots and selectively fine-tune, accepting some memory overhead to ensure reliability. As robots move from controlled factory floors to open-world environments—homes, hospitals, outdoor construction sites—generalization will become the defining challenge, requiring tight integration of learning, world modeling, and real-time adaptation.

5.6. Cross-Embodiment Learning and Large-Scale Robot Datasets

The scaling laws observed in language models—larger datasets and models yield better generalization—are beginning to manifest in robot learning, though data scarcity remains a bottleneck. The Open X-Embodiment dataset represents a milestone in this direction: assembled through collaboration between 21 institutions, it pools over 1 million real robot trajectories from 22 different robot embodiments—from single arms to bi-manual robots and quadrupeds—spanning 60 existing datasets from 34 robotics labs worldwide [141]. The dataset demonstrates 527 skills across more than 150,000 tasks, enabling research on generalist cross-embodiment policies that can adapt efficiently to new robots, tasks, and environments.

RT-X models trained on this data outperform single-embodiment baselines substantially: RT-1-X improved 50% over the original RT-1 on academic tasks, while RT-2-X achieved 3x the success rate of RT-2 for emergent skills. This work validates a fundamental hypothesis: co-training with data from diverse platforms imbues models with additional skills and robustness, shifting robot learning from training separate models for every application toward generalist policies. The challenge ahead is scaling from 1 million to billions of robot interaction episodes—approaching the 2 trillion tokens used to train modern language models—while addressing the unique challenges of embodied learning: kinematic diversity, sensor heterogeneity, and the physical irreversibility of real-world exploration. Industry players are investing heavily in data flywheels: Tesla collects teleoperation logs from Optimus deployments, Physical Intelligence aggregates multi-robot demonstrations across customer sites, and Amazon uses its warehouse fleet as a continuous learning corpus. The company or consortium that curates the first 100 million+ episode robot dataset may achieve decisive advantages in generalist control.

The data bottleneck: quantity, quality, and modality

Data scarcity represents the consensus challenge in Physical AI. The gap is stark across multiple dimensions. **Quantity**: Open X-Embodiment's 1 million episodes represent a 2000× shortfall compared with the 2 trillion tokens used to train modern LLMs, reflecting the cost and difficulty of collecting real-world robot interaction data versus web-scraped text. **Quality**: Recent empirical work demonstrates that diverse demonstrations outperform large homogeneous datasets—policies trained on varied trajectories across different scenes, objects, and lighting conditions generalize better than those trained on repetitive tasks in identical settings [127]. Label noise in pre-training data can significantly impact downstream task performance [142], while learning frameworks must handle various imprecise label configurations including partial labels and noisy annotations [143]. This implies the data challenge is not purely about volume but about curating rich, representative experience. **Modality**: Effective robot learning requires synchronized multi-modal streams—vision, proprioception, force-torque sensing, and natural language annotations describing intent—creating significant curation and storage overhead compared with text-only or image-only datasets. A single hour of multi-camera, force-sensing teleoperation can generate hundreds of gigabytes of data, far exceeding the infrastructure requirements of language model training.

Addressing this bottleneck requires systemic solutions. The most promising near-term approach combines teleoperated demonstrations (expensive at $50–200/hr but high-quality), simulation-generated synthetic data (infinite but reality-gapped), and conservative online fine-tuning (sample-

 doi:10.20944/preprints202604.0549.v1

efficient but risks hardware). Automated dataset construction pipelines that handle collection, curation, quality assessment, and privacy-preserving aggregation are becoming critical infrastructure [144]. As the field matures, standardised data formats, quality metrics, and federated learning protocols will be essential for scaling robot learning to internet-scale data regimes.

6. Safety, Ethics, and Deployment

As foundation models gain physical agency, ensuring safe, predictable behaviour aligned with human values becomes a socio-technical challenge. Failures can emerge from flawed perception, incorrect reasoning, degraded hardware or unexpected human interaction, so safety engineering must span verification, governance, deployment infrastructure, and social adoption. Understanding the systematic differences between human and AI perception helps identify potential failure modes in vision-based systems [145], informing the design of complementary verification strategies.

6.1. Assurance for Model-Mediated Control

LLM-driven controllers demand new verification tooling. Benchmarks such as SafeMindBench combine natural-language goal descriptions with simulated manipulation and navigation tasks to stress-test whether embodied agents can detect and avoid hazards, offering structured metrics for failure modes that previously went unnoticed [146]. Complementary analyses map emerging risks in embodied foundation models and argue for defence-in-depth, combining simulation stress tests, interpretable monitors, and conservative recovery policies [147]. These approaches shift assurance from ad hoc testing toward quantifiable coverage of safety requirements before deployment.

Table 7 categorises the principal failure modes observed across Physical AI deployments, mapping each to its root cause, observable symptoms, and established mitigation strategies.

Table 7. Taxonomy of Physical AI failure modes. Categories span the full stack from perception through planning, control, hardware, and human interaction. Mitigations draw on techniques discussed throughout this survey; no single defence is sufficient, motivating layered safety architectures.

Category	Failure Mode	Example / Symptom	Mitigation
Perception	Object hallucination	VLM detects nonexistent obstacle; grasps empty space	Multi-sensor fusion; confidence thresholds; redundant modalities
	Sensor degradation	LiDAR rain scatter; camera occlusion or glare	Graceful degradation; radar backup; digital twin replay for diagnosis
Planning	Infeasible plan	LLM proposes physically impossible action sequence	Affordance grounding (PaLM-SayCan); physics-aware verification
	Goal misinterpretation	Ambiguous NL instruction yields unintended behaviour	Clarification dialogue; conservative defaults; human-in-the-loop
Control	Distribution shift	Policy encounters unseen object geometry or dynamics	Domain randomisation; online adaptation (RMA); fleet retraining
	Latency-induced error	VLM response too slow for dynamic scene changes	Dual-system architecture (Helix); reactive safety fallbacks
Hardware	Actuator fault	Joint failure mid-task; gripper slip	Torque monitoring; redundant actuators; safe-stop protocols
	Compute overload	Edge GPU thermal throttle; inference timeout	Model distillation; workload partitioning; cloud offload (FogROS2)
Integration	Sim-to-real gap	Policy trained in simulation fails on real contacts	Iterative calibration (RialTo); hybrid sim-real training
	Fleet inconsistency	Model update degrades subset of heterogeneous fleet	Canary deployments; per-embodiment regression testing
Human–Robot	Intent misalignment	Robot optimises proxy metric, not user's true goal	Constitutional AI constraints; value alignment; audit logging
	Trust miscalibration	Operator over-relies on or under-trusts autonomy	Transparent confidence displays; graduated autonomy levels

 doi:10.20944/preprints202604.0549.v1

6.2. Governance, Oversight and Transparency

Regulators are codifying obligations for high-risk physical AI. The EU Artificial Intelligence Act, entering into force in August 2024, classifies autonomous robots that interact with people as high-risk systems and mandates risk management, post-market monitoring, and incident reporting as conditions for deployment [148,149]. Industry frameworks are evolving in parallel: Anthropic introduced an AI Safety Level taxonomy that requires explicit hazard analysis, capability limitations and staged deployment reviews before releasing higher-autonomy models, signalling a shift toward structured safety certification for AI systems [150]. Anthropic's Constitutional AI (CAI) framework [151] embeds value alignment directly into model training by encoding explicit rules that guide behaviour; for embodied agents, such constitutions can encode safety constraints, operational guidelines, and ethical boundaries. Programme-level oversight increasingly relies on constitutional prompting to enforce behavioural guardrails, though these soft constraints must be paired with auditable logging and human-in-the-loop approvals for irreversible actions.

As deployments scale, liability frameworks face a novel challenge: when an autonomous system causes harm, responsibility is distributed across the deploying organisation, VLA model provider, hardware manufacturer, simulation vendor, and sensor supplier. Traditional product liability struggles with learned policies that evolve through deployment rather than fixed engineered behaviours. The EU AI Act's phased compliance timeline (Figure 6), with full high-risk obligations by August 2027, provides initial structure, but regulatory approaches remain fragmented across jurisdictions. The path to ubiquitous Physical AI will require iterative refinement of governance structures—including third-party certification, operational design domains, and standardised incident reporting—informed by real-world operational experience.

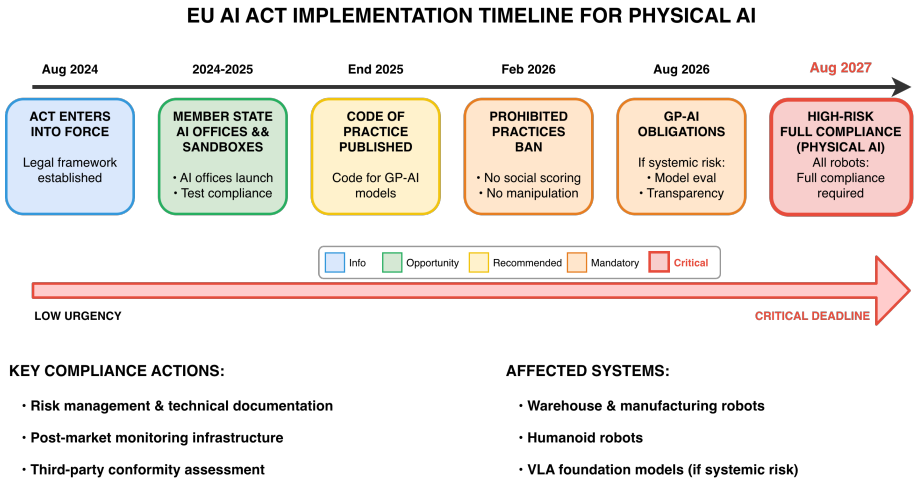

Figure 6. EU AI Act implementation timeline for Physical AI systems. Six regulatory milestones from August 2024 (Act enters force) to August 2027 (high-risk full compliance deadline) are color-coded by urgency level from *Info* (blue) to *Critical* (red). The urgency gradient reflects the escalating compliance requirements for Physical AI systems, including warehouse robots, humanoids, and VLA foundation models. Key compliance actions include risk management, post-market monitoring infrastructure, and third-party conformity assessment.

6.3. Operational Risk Management

Runtime assurance blends sensing, simulation and human supervision. Teams deploy shadow policies that run in parallel with production controllers and intervene when predicted risk exceeds a threshold, a pattern already used in autonomous vehicles and now migrating to warehouses and hospitals [147]. Digital twin infrastructure improves both detection and response: high-fidelity sensor models let engineers replay incidents, patch perception failures and re-validate policies before rolling updates back to the fleet [34]. Privacy and data-protection concerns persist because embodied systems continuously capture human environments. On-device processing, differential privacy and secure

 doi:10.20944/preprints202604.0549.v1

logging—all mandated under regional data laws—are becoming first-class requirements for platform certification under acts such as the EU AI Act [148].

6.4. Societal and Workforce Considerations

Physical AI will reshape labour markets as collaborative robots, autonomous mobile platforms and humanoids assume repetitive or hazardous tasks. The World Economic Forum's *Future of Jobs 2025* report anticipates that by 2030, automation and technological change will displace 92 million roles globally while creating 170 million new ones—a net increase of 78 million jobs—amplifying demand for technicians, data specialists and safety professionals who can manage robot fleets [152]. To build public trust, deployments must pair reskilling programs with transparent communication about failure handling and escalation paths. Human-centred design—clear affordances, explainable decision making and accessible override mechanisms—remains essential to achieve social acceptance and equitable benefits. Effective human-AI teaming frameworks become critical as Physical AI systems transition from fully autonomous operation to collaborative scenarios where humans and robots work in shared spaces [153]. These frameworks address task allocation, communication protocols, and trust calibration to ensure productive collaboration while maintaining safety.

6.5. MLOps and Deployment Infrastructure

Deploying learned policies to robot fleets borrows heavily from cloud-native software engineering—containerised builds for reproducibility, orchestrated rollouts for scale, and experiment-tracking platforms for model governance—but adds constraints unique to Physical AI: hard real-time scheduling on heterogeneous edge hardware, hardware-in-the-loop validation before each release, and federated learning protocols that improve fleet-wide performance without centralising sensitive operational data. These practices are table stakes for production robotics; the distinguishing challenges lie in integrating them with the safety assurance and fleet learning workflows described below.

Edge-cloud hybrid architectures such as FogROS2 [154,155] address the mismatch between on-robot compute budgets and foundation model requirements by offloading heavy inference to cloud providers with latency-aware failover.

6.6. Fleet Learning and Continuous Improvement

Fleet learning closes the loop between deployment and training. Wayve ingests petabytes of driving data across its L4 fleet, curating edge cases and fine-tuning models that are validated in simulation before over-the-air release [156,157]. Logistics operators follow the same pattern using the digital twin infrastructure discussed in Section 3.5: policies are staged virtually, regression tested, and rolled out to fleets via canary deployments. Research prototypes are exploring how mixed-reality data and digital twins can substantially reduce development costs for fleet-scale deployment.

6.7. Operations, Testing and Rollout Automation

MLOps pipelines for robotics extend beyond model registries to include hardware-in-the-loop testing, telemetry normalization and automated rollback. Hazard analyses from embodied foundation model studies recommend that each release include simulation-based regression, staged rollout gates and continuous compliance auditing [147]. Benchmarks such as SafeMindBench complement these processes by quantifying scenario coverage, while runtime monitors flag distribution shifts that may require retraining [146]. Organisations increasingly mirror cloud software practices—blue/green deployments, configuration as code, automated incident playbooks—but must account for physical safety, supply-chain constraints and regulatory reporting timelines. Successful programmes orchestrate these processes so that safety cases, compliance artefacts and performance dashboards update automatically as new firmware, sensors or models ship.

 doi:10.20944/preprints202604.0549.v1

7. Applications and Case Studies

Physical AI is already transforming production lines, warehouses, mobility services and scientific laboratories. Rather than pursuing a single "universal" robot, leading deployments pair specialised hardware with foundation models, simulation infrastructure, and human oversight. Three cross-cutting patterns recur across every domain examined below. First, *digital twins and simulation infrastructure* (Section 3.5) have become prerequisites, not accelerants: no deployment at fleet scale proceeds without a high-fidelity virtual rehearsal stage. Second, *foundation-model generalisation* (Sections 4 and 5) is the economic lever that separates scalable systems from bespoke integrations: deployments that rely on per-SKU or per-task engineering stall at pilot stage, whereas those that exploit VLA-style generalisation compress marginal deployment cost toward zero. Third, the *safety and governance stack* (Section 6) determines ceiling, not floor: every domain reports that regulatory and assurance barriers, not algorithmic ones, gate the transition from pilot to production. Readers should evaluate the case studies that follow through these three lenses, asking where each deployment sits on the continuum from constrained automation to open-world autonomy.

7.1. Humanoids and Flexible Assembly

Automakers are piloting humanoids to take over dexterous, ergonomically difficult tasks on mixed-model assembly lines. In 2024, Figure announced a multi-year agreement with BMW to deploy its Figure 02 humanoids at the Spartanburg plant, capable of autonomously performing precise, human-like production tasks to enhance efficiency and ergonomics on the factory floor [31]. The deployment focuses on repetitive material handling and inspection chores, freeing technicians for quality and customisation work while providing a real-world testbed for Figure's vision-language-action stack. Tesla's Optimus programme similarly targets manufacturing deployment, with plans to deploy humanoids in Tesla factories for tasks such as battery cell handling and parts sorting, eventually scaling to external customers once validation is complete [158]. At the other end of the market, Unitree's $5,900 R1 humanoid is designed for research and service applications, combining whole-body control, vision-based perception, and natural-language interaction in a compact platform capable of inspecting hazardous environments such as mines, pipelines, and substations. [32]. Sanctuary AI's Phoenix humanoid advances toward human-like intelligence with major improvements in dexterity, sensing, efficiency, and learning speed—automating new tasks in under 24 hours and setting a new benchmark for general-purpose robotics [159]. These pilots suggest that progress in generalist control models will be mediated by co-design with industrial partners who can supply high-quality task data and safety envelopes.

Cloud-based AI platforms like Microsoft's Azure Machine Learning increasingly support manufacturing companies by providing scalable infrastructure for training and deploying models across distributed facilities. For instance, Microsoft's Azure platform enables automotive suppliers like Denso to implement AI-powered quality inspection and predictive maintenance systems integrated with industrial IoT sensors and manufacturing execution systems [160]. This exemplifies the cloud-to-edge deployment pattern now common in Physical AI: models are trained centrally on large datasets aggregated across facilities, then deployed to local edge devices for low-latency inference while maintaining operational continuity during network outages. Such platforms accelerate the adoption of foundation models in manufacturing by handling the MLOps complexities of continuous model updates, version control, and compliance monitoring as production conditions evolve.

Three common enablers emerge across these humanoid and assembly deployments. First, every programme that has progressed beyond press releases anchors itself to an *industrial co-design partner* who supplies the constrained task envelope, high-quality demonstration data, and safety infrastructure that general-purpose platforms lack on their own. Second, the cloud-to-edge pattern (centralised training, local inference) described in Section 2 recurs universally, reflecting the reality that current edge accelerators can run inference but not training at the scale foundation models require. Third, the gap between a compelling laboratory demo and a multi-shift factory deployment remains dominated by

 doi:10.20944/preprints202604.0549.v1

reliability and certification, not perception or control: mean-time-between-failure targets in automotive manufacturing are measured in months, far exceeding what any current humanoid has publicly demonstrated. The pilots that advance fastest are those that start with narrowly scoped, ergonomically motivated tasks, where failure is recoverable and the economic case does not depend on full autonomy.

7.2. Warehousing, Logistics and Retail

Fulfilment centres increasingly rely on digital twins to orchestrate fleets of mobile robots, conveyors and human pickers. Amazon reports that coupling Isaac Sim with a high-fidelity twin of a fulfilment centre cut the time needed to validate new workflows and let engineers stage complex traffic and charging policies before pushing updates to its fleet of over one million robots [67]. The Amazon-Covariant partnership, announced in August 2024 with the acquisition of key personnel including co-founders Pieter Abbeel, Peter Chen, and Rocky Duan, represents a significant integration of foundation models into warehouse-scale operations [161]. Covariant's RFM-1 (Robotics Foundation Model 1), trained on tens of millions of trajectories from global warehouse deployments, enables picking robots to manipulate virtually any SKU without item-specific training—following text or image commands, answering questions about their environment, and requesting clarification when instructions are ambiguous [162,163]. This capability addresses the fundamental challenge of SKU diversity: warehouses handling millions of distinct products can no longer afford per-item vision training, making foundation model generalisation an economic necessity. DHL Supply Chain is expanding its robotics and AI capabilities globally, piloting advanced picking solutions that improve throughput for mixed-SKU orders and reduce errors through multimodal verification [164]. FedEx is integrating AI-driven sortation and robotics into its automated hubs, with facilities like Memphis processing tens of thousands of packages per hour while improving accuracy and reducing manual intervention [165]. Retailers are adopting similar stacks for back-room replenishment and returns processing, where long-tail object variation previously made automation uneconomical. The common pattern is to route sensor-rich telemetry (RGB-D, force torque, barcode scans) into shared training corpora so that each software update improves the entire fleet's performance.

A recurring pattern across these logistics deployments is the centrality of *fleet-level data flywheels*: every successful system routes operational telemetry back into a shared training corpus so that each robot's experience benefits the entire fleet, precisely the continual-learning loop analysed in Section 5. This architecture explains why warehouse automation has reached Phase 1 maturity ahead of other domains: the environment is sensor-rich, the task vocabulary is large but finite, and failure costs are low (a misplaced parcel is inconvenient, not dangerous), creating ideal conditions for the sim-to-real transfer and domain-randomisation techniques of Section 5 to close the reality gap incrementally. What distinguishes deployments that scale from those that stall is the transition from per-SKU engineered solutions to foundation-model generalisation. The remaining barriers are not algorithmic but infrastructural: integrating heterogeneous legacy warehouse-management systems, maintaining model freshness as SKU catalogues churn, and managing the workforce transition as throughput per human operator rises.

7.3. Autonomous Mobility and Field Robotics

Wayve exemplifies how self-driving stacks are converging with physical AI: its RouteDrive Foundry platform aggregates multi-modal data across electric delivery vans to train end-to-end driving policies, then verifies candidates in simulation before shipping over-the-air updates to partner fleets operated by companies such as Asda and Japan Post [156,157]. Research platforms in maritime logistics are leveraging digital twins and simulation to reduce data collection and integration overhead for multi-robot operations, while maintaining safety margins and improving operational efficiency. In agriculture, John Deere's autonomous tractors integrate vision-language models with precision farming workflows, enabling natural language task specification ("spray only infected crop sections") while maintaining centimeter-level GPS accuracy across hundreds of hectares [5]. Construction robotics firms such as Boston Dynamics and Built Robotics deploy autonomous quadrupeds and excavators

 doi:10.20944/preprints202604.0549.v1

for site inspection, earthmoving, and progress monitoring, improving safety in hazardous environments and reducing manual surveying requirements through automated data capture and integration [166]. Similar ideas are migrating to mining and energy, where ruggedised mobile manipulators inspect pipelines, wind turbines, and offshore platforms, leveraging digital twins to plan maintenance interventions before component failures occur.

The deployments in this subsection share a distinctive trait that separates them from warehouse automation: they operate in *open, unstructured environments* where the state space cannot be exhaustively enumerated, making the world-modelling and simulation techniques of Section 3.5 simultaneously more critical and less reliable. End-to-end learned policies validated in simulation underpin both urban driving and autonomous agriculture, yet the long tail of rare events exceeds what current digital twins can faithfully reproduce, exposing the "Gen2Real gap" identified in Section 3.6. A second cross-cutting pattern is the *over-the-air update cycle*: unlike factory robots bolted to a production line, mobile platforms receive continuous policy refinements, turning each deployed unit into a data-collection asset that narrows the sim-to-real gap over time. The deployments that progress fastest are those that combine a large deployed fleet (for data volume), a constrained operational design domain (for tractable safety cases), and a regulatory environment willing to grant incremental approvals rather than demanding full autonomy certification upfront.

7.4. Science, Healthcare and Hazardous Environments

Laboratories are turning to foundation model-controlled robots to accelerate experimentation while keeping humans out of hazardous conditions. Systems such as GAMORA use VR-guided gesture control and digital twin simulation to execute hazardous lab tasks remotely, ensuring precision and safety while enabling immersive training and real-time feedback [167].Human-robot collaboration in surgery is advancing rapidly, with autonomous surgical assistants increasingly supporting complex procedures, though challenges remain in aligning robotic actions with surgeon preferences and ensuring seamless interaction [168]. Across these domains, success depends on rigorous validation, sterile or clean-room compatibility, and the ability to integrate with legacy lab information systems or hospital IT infrastructure.

The gap between pilot and scale is widest in these domains, and the reasons are instructive. Unlike logistics or manufacturing, healthcare and laboratory environments impose *non-negotiable safety and sterility constraints* that cannot be relaxed during a learning phase; the assurance frameworks discussed in Section 6 are not future requirements but present-day gatekeepers. Consequently, the most successful deployments succeed precisely because they keep a human in the loop as a real-time supervisor rather than attempting full autonomy. A second barrier is *data scarcity*: surgical procedures and laboratory protocols generate far fewer training trajectories than warehouse picks or driving miles, starving the foundation-model pipelines of Section 4 of the volume they need to generalise. The path forward likely runs through high-fidelity simulation (Section 3.5) and synthetic-data augmentation rather than direct real-world data collection, but the fidelity requirements for biological tissue, chemical reactions, and patient variability far exceed those of rigid-body manipulation, a frontier challenge for the world foundation models surveyed in Section 3.6.

 doi:10.20944/preprints202604.0549.v1

Table 8. Representative Physical AI deployments across industries with measured outcomes. These examples illustrate how foundation models, digital twins, and specialized hardware combine to deliver economic value at scale.

Industry	Organization	Technology Stack	Measured Outcomes
Manufacturing and Assembly			
Automotive	BMW + Figure	Figure 02, VLA	Multi-year pilot; material handling
Automotive	Tesla	Optimus	Battery handling; internal deploy
Retail/Service	Sanctuary AI	Phoenix, dexterous manip.	Back-of-store ops; SKU handling
Research	Unitree	R1 ($5.9k)	Research labs; whole-body control
Warehousing and Logistics			
Fulfillment	Amazon	Isaac Sim, FMs	1M+ robots; workflow validation
Picking	Covariant	RFM-1 VLA	Novel SKU recognition, cluttered bins
Distribution	DHL	VLA picking	40% throughput gain; low errors
Sortation	FedEx	AMRs + AI	100k+ pkg/hr; sub-1% misroutes
Autonomous Mobility and Field Ops			
Auto Driving	Wayve	RouteDrive, E2E	Fleet learning: Asda, Post, Nissan
Agriculture	John Deere	Auto tractors, VLM	cm-level GPS; NL task spec
Construction	Built Robotics	Quadrupeds, excavators	60% survey cost cut; safety gains
Laboratory and Healthcare			
Lab Auto	GAMORA	LLM liquid handlers	Closed-loop expts; NL goals

7.5. Deployment Maturity and Industry Trajectories

Physical AI adoption varies significantly across industries, reflecting differences in technical readiness, regulatory constraints, economic incentives, and workforce considerations. Rather than predicting specific deployment dates, we characterize maturity through phases that capture operational scale, autonomy level, and market penetration.

Phase 1: Operating at Scale. Logistics and warehousing lead Physical AI adoption, with systems already deployed across hundreds of facilities worldwide. Amazon operates over one million robots, Alibaba's Cainiao network coordinates 700+ autonomous units achieving 99% sortation accuracy, and European logistics providers report 40% throughput improvements from VLA-powered picking systems. These deployments demonstrate that Physical AI can deliver measurable value in semi-structured indoor environments with well-defined workflows, dense sensor coverage, and continuous infrastructure support. Success factors include controlled operating domains, tolerance for occasional failures (misplaced items can be corrected), and strong economic drivers (labor shortages, 24/7 operations).

Phase 2: Pilot Deployments and Validation. Manufacturing humanoids represent this phase: BMW pilots Figure 02 humanoids for material handling, Tesla develops Optimus for battery assembly, and automotive suppliers test collaborative robots on mixed-model production lines. These systems operate under human supervision with limited autonomy, focusing on repetitive tasks in structured environments. The value proposition—addressing ergonomic challenges, labor shortages, and production flexibility—is compelling, but technical and safety hurdles remain. Challenges include dexterous manipulation of varied parts, safe human-robot collaboration, integration with legacy manufacturing execution systems, and achieving reliability standards (mean time between failures measured in months, not hours). Industry observers note that widespread manufacturing deployment depends on demonstrating multi-year reliability, regulatory approval for collaborative operation, and cost-effectiveness compared with task-specific automation.

Phase 3: Regulatory Proving and Limited Commercial Operation. Autonomous vehicles exemplify technical capability constrained by regulatory and societal acceptance. Waymo operates commercial robotaxi service across multiple US cities; Baidu's Apollo Go completed 1.1 million rides in Q4 2024 across ten Chinese cities. These systems achieve Level 4 autonomy and demonstrate safer-than-human metrics, but broad deployment awaits liability frameworks, insurance mechanisms, and public acceptance. Geographic fragmentation—permissive regulation in parts of the US and China

 doi:10.20944/preprints202604.0549.v1

versus cautious approaches in the EU and Japan—creates regional disparities, with the EU AI Act mandating full high-risk compliance by August 2027 [148].

Phase 4: Early Research and Niche Applications. Healthcare, agriculture, and construction represent nascent domains where Physical AI shows promise but faces fundamental barriers: stringent safety requirements in healthcare, economic viability challenges outside large-scale farming, and environmental heterogeneity in construction. Deployment trajectories depend on robust perception in unstructured environments, safe physical human-robot interaction, and social acceptance—challenges that cut across the entire Physical AI stack surveyed in this paper.

Table 9 summarizes these maturity phases, key characteristics, representative industries, and primary deployment barriers. The transition between phases is not purely linear—breakthroughs in foundation model capabilities, sim-to-real transfer, or safety verification could accelerate progress, while regulatory setbacks or high-profile failures could slow adoption. The path forward requires coordinated advances across the Physical AI stack: better sensors and world models, larger training datasets, robust safety validation, and governance structures refined by operational experience.

Table 9. Physical AI deployment maturity phases across industries. Phases characterize operational scale and autonomy level rather than specific calendar dates, as deployment timing depends on regulatory developments, technological breakthroughs, and market dynamics.

Phase	Industries	Transition Criteria	Key Characteristics	Primary Barriers
Phase 1: At Scale	Logistics, warehousing	>100 units; >12 mo uptime; documented ROI	100s–1000s deployed; 24/7 operations	Infrastructure costs; skilled workforce
Phase 2: Pilot	Manufacturing (humanoids, cobots)	>10 units; >3 mo trial; partner-validated tasks	10s–100s units; supervised; focused tasks	Reliability, safety cert., integration
Phase 3: Regulatory	Autonomous vehicles	Safety case accepted; limited commercial licence	Technically ready; limited commercial ops	Liability, public acceptance
Phase 4: Research	Agriculture, construction, healthcare	Published demo; >1 real-world trial	Demos and niche deployments; high supervision	Unstructured envs, economics

8. Challenges, Open Problems and Outlook

Despite rapid advances, open questions remain about making physical AI systems reliable, scalable, and broadly beneficial. We summarise key technical, organisational and societal challenges that emerged repeatedly across the sections above, and outline a roadmap for the coming years.

8.1. Data, Benchmarks and Evaluation

Physical AI agents encounter long-tailed phenomena—rare lighting conditions, novel objects, unexpected human behaviours—that are underrepresented in existing corpora. New benchmarking efforts such as SafeMindBench curate multi-risk scenarios for embodied agents and provide quantitative metrics for how well language-to-action pipelines anticipate and mitigate hazards [146]. Closing the loop between fleets and evaluation suites will require common schemas for logging multimodal data, privacy-preserving sharing mechanisms, and third-party certification bodies that can attest to evaluation coverage.

8.2. Sim-to-Real Transfer and the Gen2Real Gap

Despite advances in domain randomisation and iterative calibration (Section 5), automating the creation of task-specific contact models and sensor noise profiles remains labour-intensive [116]. Gen-

 doi:10.20944/preprints202604.0549.v1

erative world foundation models introduce a related "Gen2Real gap" (Section 3.6): scaling improves visual fidelity without reliably improving physical correctness [76]. Principled methods for detecting physical inconsistencies in generated training data remain an open problem.

8.3. Robustness and Continual Learning

Physical AI platforms must learn continuously without catastrophic forgetting, all while running within tight power budgets. Neuromorphic accelerators such as Loihi 2 demonstrate how sparse, event-driven computation can support online learning for tactile feedback loops at milliwatt power levels, but integrating them with transformer-based planners remains an unsolved systems problem [38]. We lack principled methods for partitioning skills between dense and spiking networks, synchronising updates across heterogeneous accelerators, and guaranteeing stability during on-the-fly fine-tuning.

8.4. Safety Assurance and Governance

Safety cases for embodied foundation models remain ad hoc. Analyses of emerging hazards emphasise the need for layered mitigations that span simulation stress-testing, runtime monitoring and human override channels, yet there are few agreed-upon templates for documenting these controls [147]. Regulatory regimes such as the EU AI Act introduce obligations for high-risk physical AI—including risk management, post-market monitoring and incident reporting—but industry lacks mature tooling to produce the required evidence automatically [149]. Harmonising regulatory expectations across jurisdictions, while maintaining interoperability and respecting privacy, is an ongoing challenge.

8.5. Hardware, Energy and Supply Chains

Hardware costs and energy consumption still limit deployment breadth. Interact Analysis estimates that average industrial robot prices fell 12% since 2019, yet many SMEs still struggle to justify capital expenditure without shorter payback periods [169]. Emerging platforms such as Jetson Thor offer transformer-class throughput at 130 W, but supply remains constrained and integration requires significant thermal and power engineering [35]. A practical bottleneck is emerging at the edge: vision-language models currently achieve only 0.1–0.4 FPS on Jetson Orin hardware—far below the 10 FPS minimum required for real-time perception—forcing architects to run lightweight detection pipelines (e.g., YOLO at $\sim$100 FPS) for fast-loop perception and reserve VLMs for asynchronous reasoning and summarisation. Sustainable robotics will depend on recyclable materials, energy-aware planning and secure supply chains for critical components such as sensors, batteries and high-bandwidth actuators.

8.6. Limitations of Current Systems

Despite the progress surveyed above, several systemic weaknesses temper optimism. First, *evaluation remains anecdotal*: most VLA results are reported on proprietary benchmarks or single-lab setups, making cross-model comparison unreliable (Table 4 highlights how few models share a common evaluation protocol). Second, *long-horizon reasoning is fragile*: current VLA models excel at short, reactive behaviours but degrade sharply on tasks requiring multi-step planning, error recovery, or adaptation to significant environmental changes [99]. Third, *safety assurance lags capability*: no deployed Physical AI system yet meets the evidence standards that regulators will require under the EU AI Act's 2027 deadline, and industry tooling for continuous safety-case generation is nascent [147]. Fourth, *the open-source ecosystem, while growing, is uneven*: open models such as OpenVLA and π_0 democratise access to VLA research, but open datasets remain small ($\sim$1M trajectories) compared to the trillion-token corpora that power language models, and few open benchmarks test contact-rich or outdoor tasks. Finally, *vendor concentration* poses a systemic risk: much of the Physical AI stack—simulation (Isaac Sim, Omniverse), edge compute (Jetson), world models (Cosmos), and perception libraries (Isaac ROS)—depends heavily on NVIDIA's ecosystem, creating supply-chain fragility and limiting architectural diversity. Addressing these limitations requires coordinated community investment in

 doi:10.20944/preprints202604.0549.v1

shared benchmarks, open datasets, vendor-agnostic tooling, and transparent reporting of failure rates alongside success metrics.

8.7. Near-Term Outlook and Integration Priorities

The coming three years will be defined less by algorithmic breakthroughs than by the integration of data, hardware, and governance. Generalist models like π_0 highlight transferable skills across embodiments, yet their success hinges on curated datasets, efficient inference hardware, and rigorous safety certification [53]. Human-centred design will remain essential: workforce studies forecast simultaneous job displacement and demand for new technical roles, underscoring the importance of reskilling programmes and participatory deployment planning [152]. Finally, progress toward sustainable physical AI will require pairing energy-efficient computation with lifecycle-aware hardware design so that embodied intelligence can operate responsibly at scale.

8.8. Future Directions

We envision continued progress along several dimensions over the next few years. First, communities should institutionalise multi-risk benchmarks such as SafeMindBench and couple them with shared logging schemas so that fleet operators can quantify coverage of rare events before large-scale deployments [146]. Second, the workflow pioneered by RialTo—iteratively calibrating real-to-sim parameters and looping updates back to deployed policies—needs to become a turnkey capability embedded in commercial digital twin stacks, reducing the expert effort now required to bridge sim-to-real gaps [116]. Third, regulatory compliance must move from periodic paperwork to continuous evidence generation: the phased obligations under the EU AI Act culminate in August 2027, compelling robotics teams to maintain auditable safety cases, adopt the forthcoming code of practice for general-purpose models, and participate in AI Act sandboxes where available [149,170,171]. Finally, sustainability goals should guide hardware choices; with Interact Analysis forecasting industrial robot shipments to reach 716,000 units by 2028, researchers must pair neuromorphic-transformer hybrids and efficient edge platforms with lifecycle analyses that curb energy and material footprints [35,38,169].

Looking further ahead, six coupled objectives emerge from these themes:

- **Data ecosystems.** Establish privacy-preserving data trusts that fuse real and synthetic trajectories, covering the long-tailed edge cases that still elude today's datasets (Section 8); auditing bodies certify benchmark completeness before large-scale deployments [146].
- **Resilient sim-to-real.** Deliver self-healing simulation stacks that load telemetry from deployed fleets each night, tune contact and sensor parameters automatically, and push validated updates back to production robots with provable guarantees on performance drift [116].
- **Lifelong adaptation.** Architect heterogeneous compute planes where neuromorphic substrates handle reflexes and dense transformers handle deliberation, enabling continual learning without catastrophic forgetting or power spikes on untethered platforms [35,38].
- **Safety and assurance.** Transition from ad hoc safety cases to continuously updated "living dossiers" that fuse simulation stress tests, runtime monitoring and governance checkpoints; regulators accept these dossiers as evidence for high-risk certification under acts such as the EU AI Act [147,149,170].
- **Ethics and labour.** Embed participatory design and workforce reskilling into deployment roadmaps so that automation augments rather than displaces frontline teams, supported by transparent reporting on job transitions and access to new technical roles [152].
- **Sustainable hardware.** Achieve circular supply chains for actuators, batteries and sensors, with recycling and remanufacturing targets codified into procurement; pair energy-aware planning with recyclable materials to halve embodied carbon relative to 2024 installations [169].

Realising this roadmap will require cross-disciplinary consortia: simulation and hardware teams sharing models and failure data, ethicists and policymakers contributing to telemetry pipelines, and workforce specialists co-designing training interventions with automation engineers. By anchoring

 doi:10.20944/preprints202604.0549.v1

research investments to these milestones, the community can transform physical AI from promising prototypes into dependable infrastructure by the end of the decade.

9. Conclusion

Physical AI closes the knowing-to-doing gap, translating internet-scale knowledge into physical competence through multimodal perception, world models, foundation policies, and edge hardware. This survey has mapped a field at an inflection point: VLA models follow natural language commands across embodiments [17,53], digital twins enable continuous fleet learning [67], and edge accelerators make real-time transformer inference feasible on mobile platforms [35]. Our analysis identified three cross-cutting findings: foundation-model generalisation is the economic lever separating scalable deployments from stalled pilots; digital twins have become prerequisites for fleet-scale operation; and regulatory barriers, not algorithmic ones, now gate the transition from pilot to production.

Three bottlenecks will determine the pace of progress. *Data scarcity*: the largest robot datasets contain roughly one million trajectories [141], a 2,000× shortfall relative to language-model corpora, demanding scalable teleoperation, simulation, and privacy-preserving data sharing. *The reality gap*: despite advances in domain randomisation and real-to-sim calibration [8,116], policies trained in simulation still require significant adaptation for contact-rich environments; generalisation across tasks, embodiments, and conditions remains fragile. *Safety assurance*: embodied foundation models lack the auditable safety cases that regulators require under frameworks such as the EU AI Act [149], and tooling for continuous evidence generation is nascent [147]. Meanwhile, workforce projections of 92 million jobs displaced and 170 million created by 2030 [152] demand that deployment roadmaps embed reskilling from the outset.

The trajectory ahead will be shaped less by individual breakthroughs than by *systems integration*: coupling data ecosystems with hardware platforms, governance frameworks, and human-centred design. Open-source initiatives such as OpenVLA [89], Open X-Embodiment [141], and Genesis [79] are democratising access, but responsible scaling requires safety architected into the stack, from sensor fusion through policy inference to actuation, rather than bolted on post-deployment. The ultimate measure of Physical AI will be not benchmark scores but whether embodied intelligence augments human capability, distributes benefits equitably, and operates sustainably at planetary scale.

Acknowledgments: The authors would like to thank Eduardo Torres Jara for his valuable feedback and insightful discussions that contributed to this work.

Appendix A. Sensor Technologies and Perception

Physical AI systems convert raw sensory data into structured environmental representations by combining multiple complementary modalities [47,50]. Table A1 summarises the principal sensor types. Rather than catalogue every modality—standard references cover LiDAR, radar, ultrasonics, and RGB cameras in depth—we focus on the sensing frontiers most relevant to Physical AI: event-driven vision, tactile perception, and multi-modal fusion.

Table A1. Comparison of sensor modalities for Physical AI systems. Cost scale: $ (low) to $$$$ (very high).

Sensor Type	Range	Accuracy	Weather	Cost	Primary Applications
RGB Camera	0–50m	Medium	Poor	$	Object recognition, inspection
RGB-D Camera	0–6m	High	Poor	$$	Bin picking, manipulation
Event Camera	0–50m	High	Excellent	$$$	High-speed tracking, drones
2D LiDAR	0–30m	Very High	Good	$$	Navigation, SLAM, pallets
3D LiDAR	0–300m	Very High	Good	$$$$	Autonomous vehicles, mapping
Radar (77GHz)	0–300m	Medium	Excellent	$$	All-weather obstacle detection
Tactile Array	Contact	Very High	N/A	$$	Manipulation, force feedback

 doi:10.20944/preprints202604.0549.v1

Appendix A.1. Frontier Sensing Modalities

RGB-D and structured-light cameras

RGB-D cameras remain the workhorse for manipulation, providing colour and per-pixel depth from structured infrared patterns. Intel's RealSense D455, widely used in bin picking, achieves depth accuracy within 2% at 4 meters with an operating range up to 6 meters [172]. Microsoft's Azure Kinect adds skeletal tracking for collaborative robotics [173]. Time-of-flight variants complement these for short-range obstacle detection where GPS is unavailable.

Event cameras

Event-based sensors record per-pixel intensity changes asynchronously rather than capturing fixed-rate frames, achieving microsecond temporal resolution with dynamic range exceeding 120 dB [174]. These properties enable robust tracking of rapid motion under extreme lighting—conditions that defeat conventional cameras. Deep learning architectures are bridging the gap between event streams and standard vision pipelines, with applications in high-speed drone navigation and manufacturing inspection [175]. As solid-state designs from companies such as Prophesee reduce cost and power consumption, event cameras are poised to become standard components in mobile robot perception stacks.

Force-torque and tactile sensors

Six-axis force-torque sensors mounted at the wrist are standard in industrial assembly, polishing, and machine tending [176]. Tactile arrays—providing distributed contact pressure, vibration, and temperature across gripper surfaces—achieve spatial resolution exceeding human fingertips [177] and are critical for detecting grasp success and incipient slip in scenarios where vision alone is insufficient. Vision-based tactile sensors such as GelSight and DIGIT, which image deformation of a soft membrane under contact, have recently enabled learned tactile policies for dexterous manipulation, though durability in industrial settings remains an open challenge.

Appendix A.2. Sensor Fusion and Multi-Modal Architectures

Each modality has complementary strengths: vision captures rich semantics but is lighting-sensitive; LiDAR provides precise geometry without colour; radar penetrates fog and rain at lower resolution. Fusing these modalities yields perception that exceeds any single sensor [50]. Figure A1 illustrates a representative multi-modal fusion architecture.

Classical approaches use Kalman and particle filters to combine noisy measurements through recursive Bayesian estimation [47]. Modern architectures replace hand-crafted pipelines with end-to-end neural networks that learn optimal cross-modal feature combinations. BEVFusion exemplifies this trend: by projecting camera and LiDAR features into a unified bird's-eye-view representation, it achieves up to 13.6% higher mIoU on BEV map segmentation over single-modality baselines while maintaining real-time performance [178].

At deployment scale, autonomous vehicles such as Waymo's sixth-generation platform fuses four LiDAR units, thirteen cameras, and six radar sensors into a 360° perception system processing over 1 GB/s of data [3]. Warehouse robots merge LiDAR-based obstacle detection with camera-based object recognition for reliable navigation through dynamic environments [50]. Manufacturing cells combine visual servoing for coarse positioning with force-torque feedback for contact-phase adjustments during assembly.

 doi:10.20944/preprints202604.0549.v1

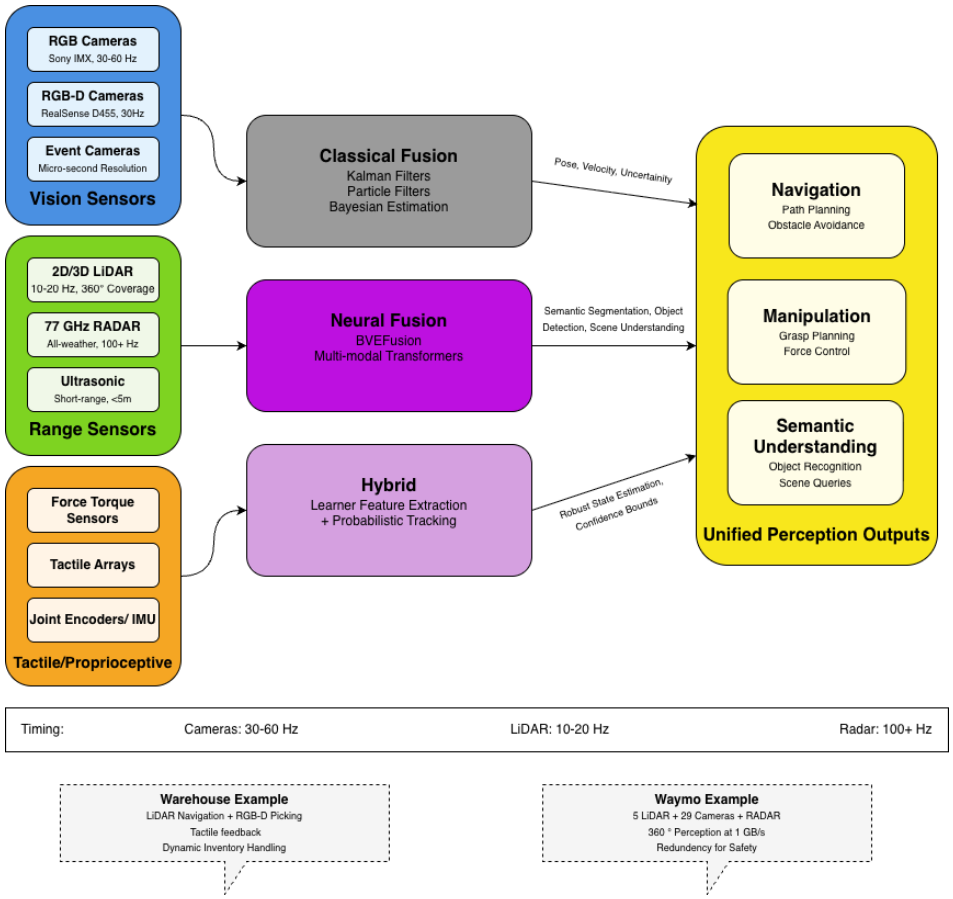

Figure A1. Multi-modal sensor fusion architecture for Physical AI, showing how diverse sensor inputs (vision, range, tactile) are combined through classical probabilistic methods and modern neural fusion approaches to produce robust, unified perception.

Key challenges persist: temporal alignment across modalities operating at different rates (cameras at 30–60 Hz, LiDAR at 10–20 Hz, radar faster still) demands precise calibration and synchronisation; computational budgets on mobile platforms constrain fusion complexity; and graceful degradation under sensor failure requires architectures that dynamically re-weight or exclude unreliable inputs. Learned fusion approaches increasingly outperform classical filters but introduce new difficulties for safety certification and adversarial robustness—issues addressed in Section 6.

References

1. Moravec, H. *Mind Children: The Future of Robot and Human Intelligence*; Harvard University Press: Cambridge, MA, 1988.
2. Figure AI. F.02 Contributed to the Production of 30,000 Cars at BMW, 2025. 11-month deployment trial at BMW Group Plant Spartanburg.
3. Waymo. Meet the 6th-generation Waymo Driver. https://waymo.com/blog/2024/08/meet-the-6th-generation-waymo-driver, 2024. Autonomous vehicle sensor suite with 13 cameras, 4 LiDAR, 6 radar sensors. Accessed: 2025-10-31.
4. Intuitive Surgical. 20 Million Patients Benefit from da Vinci Surgery Globally. GlobeNewsWire, 2026. Cumulative milestone of 20 million da Vinci procedures worldwide. Accessed: 2026-02-01.
5. John Deere. Autonomous Tractor. https://www.deere.com/en/autonomous/, 2024. Fully autonomous tractor with GPS and vision systems for precision agriculture. Accessed: 2025-10-31.
6. NVIDIA. NVIDIA Expands Omniverse With Generative Physical AI, 2025. Press release, January 6, 2025.
7. Brooks, R.A. Intelligence without representation. *Artificial Intelligence* **1991**, *47*, 139–159.
8. Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In Proceedings of the 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017, pp. 23–30.

 doi:10.20944/preprints202604.0549.v1

9. Nilsson, N.J. Shakey the Robot. Technical Note 323, SRI International, Artificial Intelligence Center, Menlo Park, CA, 1984.
10. Fikes, R.E.; Nilsson, N.J. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial intelligence* **1971**, *2*, 189–208.
11. Brooks, R.A. A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation* **1986**, *2*, 14–23.
12. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* **2017**.
13. Todorov, E.; Erez, T.; Tassa, Y. MuJoCo: A physics engine for model-based control. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 5026–5033. https://doi.org/10.1109/IROS.2012.6386109.
14. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 8748–8763.
15. Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; et al. Do as i can, not as i say: Grounding language in robotic affordances. In Proceedings of the Conference on Robot Learning (CoRL). PMLR, 2023, Vol. 205, pp. 287–318.
16. Liang, J.; Huang, W.; Xia, F.; Xu, P.; Hausman, K.; Ichter, B.; Florence, P.; Zeng, A. Code as policies: Language model programs for embodied control. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 9493–9500.
17. Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of the Conference on Robot Learning. PMLR, 2023, pp. 2165–2183.
18. Driess, D.; Xia, F.; Sajjadi, M.S.; Lynch, C.; Chowdhery, A.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; et al. PaLM-E: An Embodied Multimodal Language Model. In Proceedings of the Proceedings of the 40th International Conference on Machine Learning. PMLR, 2023, Vol. 202, pp. 8469–8488.
19. Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. π_0: A Vision-Language-Action Flow Model for General Robot Control. In Proceedings of the Robotics: Science and Systems (RSS), 2025.
20. Reed, S.; Zolna, K.; Parisotto, E.; Colmenarejo, S.G.; Novikov, A.; Barth-Maron, G.; Gimenez, M.; Sulsky, Y.; Kay, J.; Springenberg, J.T.; et al. A generalist agent. *Transactions on Machine Learning Research (TMLR)* **2022**.
21. Intrinsic. Unlocking New Value in Industrial Automation with AI. https://www.intrinsic.ai/, 2024. Accessed: 2025-10-30.
22. Intrinsic and NVIDIA. NVIDIA and Google's Intrinsic Developing Next-Generation Robots. https://roboticsandautomationnews.com/2024/05/16/nvidia-and-googles-intrinsic-developing-next-generation-robots/, 2024. Accessed: 2025-10-30.
23. Cainiao. Alibaba's Cainiao Launches Enterprise Smart Warehouse Solution. https://techwireasia.com/2022/03/alibabas-cainiao-launches-enterprise-smart-warehouse-solution/, 2022. Accessed: 2025-10-30.
24. Baidu. Baidu Apollo Launches 6th-Gen Robotaxi with 60% Lower Cost. https://cnevpost.com/2024/05/15/baidu-apollo-launches-6th-gen-robotaxi/, 2024. Accessed: 2025-10-30.
25. Glasner, J. The Year of Humanoid Robots. Crunchbase News, 2024. Reports $7.2 billion in robotics venture funding in 2024. Accessed: 2025-10-31.
26. Kawaharazuka, K.; Oh, J.; Yamada, J.; Posner, I.; Zhu, Y. Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications. *IEEE Access* **2025**, *13*, 162467–162504. https://doi.org/10.1109/ACCESS.2025.3609980.
27. Liu, Y.; Chen, W.; Bai, Y.; Liang, X.; Li, G.; Gao, W.; Lin, L. Aligning Cyber Space With Physical World: A Comprehensive Survey on Embodied AI. *IEEE/ASME Transactions on Mechatronics* **2025**, *30*, 7253–7274.
28. Sapkota, R.; Cao, Y.; Roumeliotis, K.I.; Karkee, M. Vision-Language-Action (VLA) Models: Concepts, Progress, Applications and Challenges. *arXiv preprint arXiv:2505.04769* **2025**.
29. Sriram, A. Function over flash: Specialized robots attract billions with efficient task handling. https://www.reuters.com/business/finance/function-over-flash-specialized-robots-attract-billions-with-efficient-task-2025-05-22/, 2025. Accessed: 2025-10-28.
30. International Federation of Robotics. Global Robot Demand in Factories Doubles Over 10 Years. https://ifr.org/ifr-press-releases/news/global-robot-demand-in-factories-doubles-over-10-years, 2025. Market forecast for industrial robot shipments. Accessed: 2025-10-31.

 doi:10.20944/preprints202604.0549.v1

31. BMW Group. BMW Group Invests in AI Robotics Start-Up Figure. https://www.bmwgroup.com/en/news/general/2024/humanoid-robots.html, 2024. Partnership for humanoid robot deployment in manufacturing. Accessed: 2025-10-31.
32. Unitree Robotics. Unitree R1 Humanoid Robot. https://www.unitree.com/R1/, 2025. Compact general-purpose humanoid robot (1.2 m, 25 kg, 26 joints) priced from $4,900 (R1 AIR). Unveiled July 2025. Accessed: 2025-10-31.
33. Shakir, U. Tesla's Optimus bot makes a scene at the robotaxi event. The Verge, 2024. Musk estimated Optimus retail price at $20,000–$30,000. Accessed: 2025-10-31.
34. NVIDIA Corporation. NVIDIA Announces Omniverse Microservices to Supercharge Physical AI. https://www.globenewswire.com/news-release/2024/06/17/2899696/0/en/NVIDIA-Announces-Omniverse-Microservices-to-Supercharge-Physical-AI.html, 2024. Sensor RTX microservices for physically accurate sensor simulation. Announced June 2024. Accessed: 2025-10-31.
35. NVIDIA Corporation. NVIDIA Blackwell-Powered Jetson Thor Now Available, Accelerating the Age of General Robotics. https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-thor/, 2025. Platform delivering 2,070 FP4 TFLOPS for humanoid and mobile robots. Announced August 2025. Accessed: 2025-10-31.
36. TrendForce. NVIDIA Jetson Thor Targets Advanced Humanoid Applications. https://www.trendforce.com/presscenter/news/20250826-12685.html, 2025. Market analysis of Jetson Thor performance improvements. Accessed: 2025-10-31.
37. Qualcomm. Qualcomm Launches World's First 5G and AI-Enabled Robotics Platform. https://www.qualcomm.com/news/releases/2020/06/qualcomm-launches-worlds-first-5g-and-ai-enabled-robotics-platform, 2020. Robotics RB5 platform with 5G connectivity and 15 TOPS AI performance. Accessed: 2025-10-31.
38. Intel Labs. Intel Loihi 2 Neuromorphic Research Chip. https://www.intel.com/content/www/us/en/research/neuromorphic-computing-loihi-2-technology-brief.html, 2024. Second-generation neuromorphic processor for energy-efficient AI. Accessed: 2025-10-31.
39. Intel Corporation. Intel Builds World's Largest Neuromorphic System to Enable More Sustainable AI. https://newsroom.intel.com/artificial-intelligence/intel-builds-worlds-largest-neuromorphic-system-to-enable-more-sustainable-ai, 2024. 1.15 billion neuron neuromorphic system deployed at Sandia National Laboratories. Accessed: 2025-10-31.
40. NVIDIA Corporation. NVIDIA Announces Omniverse Real-Time Physics Digital Twins With Industry Software Leaders. https://www.globenewswire.com/de/news-release/2024/11/18/2983079/0/en/NVIDIA-Announces-Omniverse-Real-Time-Physics-Digital-Twins-With-Industry-Software-Leaders.html, 2024. Digital twin platform transforming manufacturing with 1,200x faster simulations. Announced November 2024. Accessed: 2025-10-31.
41. ROS 2 Control Contributors. Welcome to the ros2_control documentation! ROS 2 Control Documentation (Rolling), 2025. Accessed: September 29, 2025.
42. Lumpp, F.; Panato, M.; Bombieri, N.; Fummi, F. A Design Flow Based on Docker and Kubernetes for ROS-based Robotic Software Applications. *ACM Transactions on Embedded Computing Systems* **2024**, *23*, 74:1–74:24. https://doi.org/10.1145/3594539.
43. Open Source Robotics Foundation. Jazzy Jalisco (jazzy) — ROS 2 Documentation. OSRF Technical Documentation, 2024.
44. NVIDIA. NVIDIA Brings Generative AI Tools, Simulation and Perception Workflows to ROS Developer Ecosystem. NVIDIA Newsroom, 2024. Announcement for ROSCon 2024.
45. NVIDIA. Isaac Perceptor for Autonomous Mobile Robot Development. Technical specification, NVIDIA Corporation, 2024.
46. NVIDIA. *NVIDIA Implementation of Type Adaptation and Negotiation (NITROS)*. NVIDIA Corporation, 2024. Isaac ROS Documentation.
47. Thrun, S. Probabilistic robotics. *Communications of the ACM* **2002**, *45*, 52–57.
48. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.; Tardós, J.D. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE transactions on robotics* **2021**, *37*, 1874–1890.
49. Teed, Z.; Deng, J. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021, Vol. 34, pp. 16558–16569.
50. Ušinskis, V.; Nowicki, M.; Dzedzickis, A.; Bučinskas, V. Sensor-Fusion Based Navigation for Autonomous Mobile Robot. *Sensors* **2025**, *25*, 1248. https://doi.org/10.3390/s25041248.

 doi:10.20944/preprints202604.0549.v1

51. OpenAI. GPT-4V(ision) System Card. https://openai.com/index/gpt-4v-system-card/, 2023. Accessed: 2025-10-30.
52. Jatavallabhula, K.M.; Kuwajerwala, A.; Gu, Q.; Omama, M.; Chen, T.; Maalouf, A.; Li, S.; Iyer, G.; Saryazdi, S.; Keetha, N.; et al. Conceptfusion: Open-set multimodal 3d mapping. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
53. Physical Intelligence. The π_0 Foundation Model. https://www.physicalintelligence.company, 2024. Accessed: 2025-10-27.
54. Coumans, E.; Bai, Y. PyBullet, a Python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021.
55. Battaglia, P.; Pascanu, R.; Lai, M.; Jimenez Rezende, D.; et al. Interaction networks for learning about objects, relations and physics. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2016, Vol. 29.
56. Hasani, R.; Lechner, M.; Amini, A.; Rus, D.; Grosu, R. Liquid time-constant networks. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2021, Vol. 35, pp. 7657–7666.
57. Chahine, M.; Hasani, R.; Kao, P.; Ray, A.; Shubert, R.; Lechner, M.; Amini, A.; Rus, D. Robust flight navigation out of distribution with liquid neural networks. *Science Robotics* **2023**, *8*, eadc8892.
58. Hart, P.E.; Nilsson, N.J.; Raphael, B. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics* **1968**, *4*, 100–107.
59. LaValle, S.M. Rapidly-exploring random trees: A new tool for path planning. Technical Report TR 98-11, Department of Computer Science, Iowa State University, 1998.
60. Ratliff, N.; Zucker, M.; Bagnell, J.A.; Srinivasa, S. CHOMP: Gradient optimization techniques for efficient motion planning. In Proceedings of the 2009 IEEE international conference on robotics and automation. IEEE, 2009, pp. 489–494.
61. Sutton, R.S. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin* **1991**, *2*, 160–163.
62. Kaelbling, L.P.; Lozano-Pérez, T. Hierarchical task and motion planning in the now. In Proceedings of the 2011 IEEE international conference on robotics and automation. IEEE, 2011, pp. 1470–1477.
63. Shridhar, M.; Manuelli, L.; Fox, D. Cliport: What and where pathways for robotic manipulation. In Proceedings of the Conference on robot learning. PMLR, 2021, pp. 894–906.
64. NVIDIA. Isaac Sim: GPU-Accelerated Robot Simulation. NVIDIA Developer Documentation, 2024.
65. Pixar Animation Studios. Universal Scene Description (OpenUSD). https://openusd.org, 2016. Open-source framework for 3D scene interchange and collaboration. Accessed: 2025-10-31.
66. NVIDIA. PhysX 5 SDK. https://github.com/NVIDIA-Omniverse/PhysX, 2022. Open-source GPU-accelerated physics engine for real-time simulation. Accessed: 2025-10-31.
67. Amazon Robotics and NVIDIA. Amazon Robotics Builds Digital Twins of Warehouses with NVIDIA Omniverse and Isaac Sim. https://resources.nvidia.com/en-us-omniverse-enterprise/amazon-robotics, 2024. Case study on warehouse digital twin deployment at scale. Accessed: 2025-10-31.
68. Siemens. Understanding Your Whole Factory with the Comprehensive Digital Twin. https://blogs.sw.siemens.com/thought-leadership/2024/12/26/understanding-your-whole-factory-with-the-comprehensive-digital-twin/, 2024. Accessed: 2025-10-30.
69. Siemens and NVIDIA. Siemens and NVIDIA Expand Partnership to Accelerate AI Capabilities in Manufacturing. https://press.siemens.com/global/en/pressrelease/siemens-and-nvidia-expand-partnership-accelerate-ai-capabilities-manufacturing, 2025. Accessed: 2025-10-30.
70. Puig, X.; Undersander, E.; Szot, A.; Cote, M.D.; Yang, T.Y.; Partsey, R.; Desai, R.; Clegg, A.W.; Hlavac, M.; Min, S.Y.; et al. Habitat 3.0: A co-habitat for humans, avatars and robots. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
71. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An open urban driving simulator. In Proceedings of the Conference on robot learning. PMLR, 2017, pp. 1–16.
72. Koenig, N.; Howard, A. Design and use paradigms for Gazebo, an open-source multi-robot simulator. In Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2004, Vol. 3, pp. 2149–2154.
73. Jakob, W.; Speierer, S.; Roussel, N.; Vicini, D. Dr.Jit: A Just-In-Time Compiler for Differentiable Rendering. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* **2022**, *41*. Core compiler of the Mitsuba 3 rendering system, https://doi.org/10.1145/3528223.3530099.

 doi:10.20944/preprints202604.0549.v1

74. Liu, M.; Grabli, S.; Speierer, S.; Sarafianos, N.; Bode, L.; Chiang, M.; Hery, C.; Davis, J.; Aliaga, C. Controllable Biophysical Human Faces. *Computer Graphics Forum* **2025**, *44*, e70170.
75. Sang, S.; Zhi, T.; Song, G.; Liu, M.; Lai, C.; Liu, J.; Wen, X.; Davis, J.; Luo, L. Agileavatar: Stylized 3d avatar creation via cascaded domain bridging. In Proceedings of the SIGGRAPH Asia 2022 Conference Papers, 2022, pp. 1–8.
76. Agarwal, N.; Ali, A.; Bala, M.; Balaji, Y.; Barker, E.; Cai, T.; Chattopadhyay, P.; Chen, Y.; Cui, Y.; Ding, Y.; et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575* **2025**.
77. Ali, A.; Bai, J.; Bala, M.; Balaji, Y.; Blakeman, A.; et al. World Simulation with Video Foundation Models for Physical AI. *arXiv preprint arXiv:2511.00062* **2025**.
78. Azzolini, A.; Bai, J.; Brandon, H.; Cao, J.; Chattopadhyay, P.; et al. Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning. *arXiv preprint arXiv:2503.15558* **2025**.
79. Authors, G. Genesis: A Generative and Universal Physics Engine for Robotics and Beyond, 2024.
80. Simio LLC. Simio Digital Twin Simulation Software. https://www.simio.com, 2024. Process digital twin platform for discrete event simulation. Accessed: 2025-10-31.
81. Consumer Goods Technology. P&G Taps into AI and Automation for Faster, Smarter Operations, 2024. Describes P&G's Control Tower virtual twin reducing deadhead movements by 15%.
82. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
83. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2024.
84. Shen, W.; Yang, G.; Yu, A.; Wong, J.; Kaelbling, L.P.; Isola, P. Distilled Feature Fields Enable Few-Shot Language-Guided Manipulation. In Proceedings of the Conference on Robot Learning (CoRL), 2023. Best Paper Award.
85. Chen, T.; Shorinwa, O.; Bruno, J.; Swann, A.; Yu, J.; Zeng, W.; Nagami, K.; Dames, P.; Schwager, M. Splat-Nav: Safe Real-Time Robot Navigation in Gaussian Splatting Maps. *IEEE Transactions on Robotics* **2025**.
86. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017.
87. Hu, Y.; Lin, F.; Zhang, T.; Yi, L.; Gao, Y. Look Before You Leap: Unveiling the Power of GPT-4V in Robotic Vision-Language Planning. *arXiv preprint arXiv:2311.17842* **2023**.
88. Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; et al. Rt-1: Robotics transformer for real-world control at scale. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
89. Kim, M.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Sanketi, P.; Vuong, Q.; et al. OpenVLA: An Open-Source Vision-Language-Action Model. In Proceedings of the Conference on Robot Learning (CoRL). PMLR, 2024, Vol. 270, pp. 2679–2713.
90. Octo Model Team.; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Xu, C.; Luo, J.; et al. Octo: An Open-Source Generalist Robot Policy. In Proceedings of the Proceedings of Robotics: Science and Systems, Delft, Netherlands, 2024.
91. Liu, S.; Wu, L.; Li, B.; Tan, H.; Chen, H.; Wang, Z.; Xu, K.; Su, H.; Zhu, J. RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation. In Proceedings of the International Conference on Learning Representations (ICLR), 2025.
92. Li, X.; Zhang, M.; Geng, Y.; Geng, H.; Long, Y.; Shen, Y.; Zhang, R.; Liu, J.; Dong, H. ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18061–18070.
93. Tang, W.; Jing, D.; Pan, J.H.; Lu, Z.; Liu, Y.H.; Li, L.E.; Ding, M.; Fu, C.W. Incentivizing Multimodal Reasoning in Large Models for Direct Robot Manipulation, 2025, [arXiv:cs.AI/2505.12744].
94. Physical Intelligence Team. Open Sourcing π_0. Physical Intelligence Blog, 2025.
95. Bjorck, J.; Castañeda, F.; Cherniadev, N.; Da, X.; Ding, R.; et al. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *arXiv preprint arXiv:2503.14734* **2025**.
96. Figure AI. Helix: A vision-language-action model for generalist humanoid control. Figure AI Blog, 2025.
97. Team, G.R. Gemini Robotics: Bringing AI into the Physical World, 2025.

 doi:10.20944/preprints202604.0549.v1

98. Team, G.R. Gemini Robotics 1.5: Pushing the Frontier of Generalist Robots with Advanced Embodied Reasoning, Thinking, and Motion Transfer, 2025.
99. Guruprasad, P.; Sikka, H.; Song, J.; Wang, Y.; Liang, P.P. Benchmarking Vision, Language, & Action Models on Robotic Learning Tasks, 2024, [arXiv:cs.RO/2411.05821].
100. Kalashnikov, D.; Irpan, A.; Pastor, P.; Ibarz, J.; Herzog, A.; Jang, E.; Quillen, D.; Holly, E.; Kalakrishnan, M.; Vanhoucke, V.; et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In Proceedings of the Conference on robot learning. PMLR, 2018, pp. 651–673.
101. Cheng, X.; Shi, K.; Agarwal, A.; Pathak, D. Extreme parkour with legged robots. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 11443–11450.
102. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the International Conference on Machine Learning (ICML). PMLR, 2018, pp. 1861–1870.
103. Fujimoto, S.; van Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In Proceedings of the International conference on machine learning. PMLR, 2018, pp. 1587–1596.
104. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. In Proceedings of the International Conference on Learning Representations (ICLR), 2016.
105. Hansen, N.; Su, H.; Wang, X. TD-MPC2: Scalable, Robust World Models for Continuous Control. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
106. Chebotar, Y.; Vuong, Q.; Hausman, K.; Xia, F.; Lu, Y.; Irpan, A.; Kumar, A.; Yu, T.; Herzog, A.; Pertsch, K.; et al. Q-Transformer: Scalable Offline Reinforcement Learning via Autoregressive Q-Functions. In Proceedings of the Conference on Robot Learning (CoRL), 2023.
107. Wang, Z.; Hunt, J.J.; Zhou, M. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
108. Zhu, Y.; Wan Hasan, W.Z.; Ramli, H.R.H.; Norsahperi, N.M.H.; Mohd Kassim, M.S.; Yao, Y. Deep Reinforcement Learning of Mobile Robot Navigation in Dynamic Environment: A Review. *Sensors* **2025**, *25*, 3394. https://doi.org/10.3390/s25113394.
109. Zhu, Y.; Wong, J.; Mandlekar, A.; Martín-Martín, R.; Joshi, A.; Nasiriany, S.; Zhu, Y. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293* **2020**.
110. Yu, T.; Quillen, D.; He, Z.; Julian, R.; Hausman, K.; Finn, C.; Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In Proceedings of the Conference on robot learning. PMLR, 2019, pp. 1094–1100.
111. Rashid, T.; Samvelyan, M.; Schroeder de Witt, C.; Farquhar, G.; Foerster, J.; Whiteson, S. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In Proceedings of the Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 4295–4304.
112. Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; Wu, Y. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2022.
113. Zhou, Y.; Xiao, J.; Zhou, Y.; Loianno, G. Multi-Robot Collaborative Perception With Graph Neural Networks. *IEEE Robotics and Automation Letters* **2022**, *7*, 2289–2296.
114. Liu, K.; Tang, Z.; Wang, D.; Wang, Z.; Zhao, B.; Li, X. COHERENT: Collaboration of Heterogeneous Multi-Robot System with Large Language Models. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA), 2025. https://doi.org/10.1109/ICRA55743.2025.11127808.
115. Li, P.; An, Z.; Abrar, S.; Zhou, L. Large Language Models for Multi-Robot Systems: A Survey. *arXiv preprint arXiv:2502.03814* **2025**.
116. Torne, M.; Simeonov, A.; Li, Z.; Chan, A.; Chen, T.; Gupta, A.; Agrawal, P. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
117. Miki, T.; Lee, J.; Hwangbo, J.; Wellhausen, L.; Koltun, V.; Hutter, M. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics* **2022**, *7*.
118. Kumar, A.; Fu, Z.; Pathak, D.; Malik, J. RMA: Rapid Motor Adaptation for Legged Robots. In Proceedings of the Robotics: Science and Systems (RSS), 2021.
119. Hoeller, D.; Rudin, N.; Sako, D.; Hutter, M. ANYmal parkour: Learning agile navigation for quadrupedal robots. *Science Robotics* **2024**, *9*.

 doi:10.20944/preprints202604.0549.v1

120. Radosavovic, I.; Xiao, T.; Zhang, B.; Darrell, T.; Malik, J.; Sreenath, K. Real-world humanoid locomotion with reinforcement learning. *Science Robotics* **2024**, *9*.
121. Cheng, X.; Ji, Y.; Chen, J.; Yang, R.; Yang, G.; Wang, X. Expressive Whole-Body Control for Humanoid Robots. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
122. Shaw, K.; Agarwal, A.; Pathak, D. LEAP Hand: Low-Cost, Efficient, and Anthropomorphic Hand for Robot Learning. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
123. Yang, M.; Lu, C.; Church, A.; Lin, Y.; Ford, C.J.; Li, H.; Psomopoulou, E.; Barton, D.A.; Lepora, N.F. AnyRotate: Gravity-Invariant In-Hand Object Rotation with Sim-to-Real Touch. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
124. Wang, C.; Shi, H.; Wang, W.; Zhang, R.; Fei-Fei, L.; Liu, C.K. DexCap: Scalable and Portable Mocap Data Collection System for Dexterous Manipulation. In Proceedings of the Robotics: Science and Systems (RSS), 2024.
125. Ross, S.; Gordon, G.; Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.
126. Argall, B.D.; Chernova, S.; Veloso, M.; Browning, B. A survey of robot learning from demonstration. *Robotics and autonomous systems* **2009**, *57*, 469–483.
127. Mandlekar, A.; Xu, D.; Wong, J.; Nasiriany, S.; Wang, C.; Kulkarni, R.; Fei-Fei, L.; Savarese, S.; Zhu, Y.; Martín-Martín, R. What matters in learning from offline human demonstrations for robot manipulation. In Proceedings of the Conference on Robot Learning (CoRL). PMLR, 2021, pp. 1678–1690.
128. Chi, C.; Feng, S.; Du, Y.; Xu, Z.; Cousineau, E.; Burchfiel, B.; Song, S. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
129. Zhao, T.Z.; Kumar, V.; Levine, S.; Finn, C. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
130. Fu, Z.; Zhao, T.Z.; Finn, C. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. In Proceedings of the Conference on Robot Learning (CoRL), 2024.
131. Sermanet, P.; Lynch, C.; Chebotar, Y.; Hsu, J.; Jang, E.; Schaal, S.; Levine, S. Time-contrastive networks: Self-supervised learning from video. In Proceedings of the 2018 IEEE international conference on robotics and automation (ICRA). IEEE, 2018, pp. 1134–1141.
132. Ha, D.; Schmidhuber, J. Recurrent world models facilitate policy evolution. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2018, Vol. 31.
133. Levine, S.; Pastor, P.; Krizhevsky, A.; Ibarz, J.; Quillen, D. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research* **2018**, *37*, 421–436.
134. Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; Gupta, A. R3M: A Universal Visual Representation for Robot Manipulation. In Proceedings of the Conference on Robot Learning (CoRL), 2022.
135. Ma, Y.J.; Liang, W.; Du, G.; Jayaraman, D.; et al. VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
136. Karamcheti, S.; Nair, S.; Chen, A.S.; Kollar, T.; Finn, C.; Sadigh, D.; Liang, P. Language-Driven Representation Learning for Robotics. In Proceedings of the Robotics: Science and Systems (RSS), 2023.
137. Yang, S.; Du, Y.; Ghasemipour, K.; Tompson, J.; Kaelbling, L.; Schuurmans, D.; Abbeel, P. Learning Interactive Real-World Simulators. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
138. Tan, J.; Zhang, T.; Coumans, E.; Iscen, A.; Bai, Y.; Hafner, D.; Bohez, S.; Vanhoucke, V. Sim-to-real: Learning agile locomotion for quadruped robots. In Proceedings of the Robotics: Science and Systems (RSS), 2018.
139. James, S.; Wohlhart, P.; Kalakrishnan, M.; Kalashnikov, D.; Irpan, A.; Ibarz, J.; Levine, S.; Hadsell, R.; Bousmalis, K. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12627–12637.
140. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International conference on machine learning. PMLR, 2017, pp. 1126–1135.
141. Open X-Embodiment Collaboration. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024. Co-winner, Best Conference Paper Award.

 doi:10.20944/preprints202604.0549.v1

142. Chen, H.; Wang, J.; Shah, A.; Tao, R.; Wei, H.; Xie, X.; Sugiyama, M.; Raj, B. Understanding and mitigating the label noise in pre-training on downstream tasks. In Proceedings of the International Conference on Learning Representations (ICLR), 2024. Spotlight.
143. Chen, H.; Shah, A.; Wang, J.; Tao, R.; Wang, Y.; Li, X.; Xie, X.; Sugiyama, M.; Singh, R.; Raj, B. Imprecise label learning: A unified framework for learning with various imprecise label configurations. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2024, Vol. 37, pp. 59621–59654.
144. Liu, M.; Di, Z.; Wei, J.; Wang, Z.; Zhang, H.; Xiao, R.; Wang, H.; Pang, J.; Chen, H.; Shah, A.; et al. Automatic dataset construction (adc): Sample collection, data curation, and beyond. *arXiv preprint arXiv:2408.11338* **2024**.
145. Liu, M.; Wei, J.; Liu, Y.; Davis, J. Human and ai perceptual differences in image classification errors. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 14318–14326.
146. Chen, R.; Sun, Y.; Wang, J.; Lv, M.; Zhang, Q.; Zeng, Y. SafeMind: Benchmarking and Mitigating Safety Risks in Embodied LLM Agents. *arXiv preprint arXiv:2509.25885* **2025**.
147. Dalrymple, D.; Skalse, J.; Bengio, Y.; Russell, S.; Tegmark, M.; Seshia, S.; Omohundro, S.; Szegedy, C.; Goldhaber, B.; Ammann, N.; et al. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. *arXiv preprint arXiv:2405.06624* **2024**.
148. EU Artificial Intelligence Act. High-level Summary of the AI Act. https://artificialintelligenceact.eu/high-level-summary/, 2024. Summary of compliance obligations for high-risk AI systems. Accessed: 2025-10-31.
149. European Commission. AI Act: Regulatory Framework for Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, 2024. Official EU AI Act policy page with entry into force date August 1, 2024. Accessed: 2025-10-31.
150. Anthropic. Announcing our updated Responsible Scaling Policy. https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-policy, 2024. AI Safety Level framework and ASL-3 protections announced October 2024. Accessed: 2025-10-31.
151. Bai, Y.; et al. Constitutional AI: Harmlessness from AI Feedback. arXiv preprint arXiv:2212.08073, 2022.
152. World Economic Forum. The Future of Jobs Report 2025. https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf, 2025. Workforce impact analysis projecting 92M jobs displaced and 170M created by 2030. Accessed: 2025-10-31.
153. Vats, V.; Binta Nizam, M.; Liu, M.; Wang, Z.; Ho, R.; Sai Prasad, M.; Titterton, V.; Venkat Malreddy, S.; Aggarwal, R.; Xu, Y.; et al. A Survey on Human-AI Collaboration with Large Foundation Models. *arXiv preprint arXiv:2403.04931* **2024**.
154. Ichnowski, J.; Chen, K.; Dharmarajan, K.; Adebola, S.; Danielczuk, M.; Mayoral-Vilches, V.; Jha, N.; Zhan, H.; Llontop, E.; Xu, D.; et al. FogROS 2: An Adaptive and Extensible Platform for Cloud and Fog Robotics Using ROS 2. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 5493–5500. https://doi.org/10.1109/ICRA48891.2023.10161307.
155. Chen, K.; Wang, M.; Gualtieri, M.; Tian, N.; Juette, C.; Ren, L.; Ichnowski, J.; Kubiatowicz, J.; Goldberg, K. FogROS2-LS: A Location-Independent Fog Robotics Framework for Latency Sensitive ROS2 Applications. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 10581–10587. https://doi.org/10.1109/ICRA57147.2024.10610759.
156. Hu, A.; Russell, L.; Yeo, H.; Murez, Z.; Fedoseev, G.; Kendall, A.; Shotton, J.; Corrado, G. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080* **2023**.
157. Wayve. Nissan to Launch Next-Generation Autonomous Driving Technology in FY2027. https://wayve.ai/press/nissan-announcement/, 2025. Partnership to integrate Wayve's AI Driver software into Nissan ProPILOT. Accessed: 2025-10-31.
158. Tesla. Optimus: Tesla's General-Purpose Humanoid Robot. https://www.tesla.com/we-robot, 2024. General-purpose humanoid robot demonstrated at We, Robot event October 2024. Accessed: 2025-10-31.
159. Sanctuary AI. Phoenix Seventh Generation Humanoid Robot. https://www.sanctuary.ai/blog/sanctuary-ai-unveils-the-next-generation-of-ai-robotics, 2024. General-purpose humanoid with 24-hour task learning capability. Released April 2024. Accessed: 2025-10-31.
160. DENSO Robotics. DENSO and Microsoft Azure Partnership for Cloud-Connected Industrial Robots. DENSO Press Release, 2024.
161. staff, A. An update on how we're accelerating the use of AI in robotics at scale. Amazon, 2024.
162. Covariant. Introducing RFM-1: Giving robots human-like reasoning capabilities. Covariant Technical Blog, 2024.

 doi:10.20944/preprints202604.0549.v1

163. Covariant. The Covariant Brain: Powering the future of automation. Technical documentation, Covariant, 2024.
164. DHL Supply Chain. DHL Supply Chain Continues to Innovate With Orchestration, Robotics, and AI in 2024. https://www.dhl.com/us-en/home/press/press-archive/2024/dhl-supply-chain-continues-to-innovate-with-orchestration-robotics-and-ai-in-2024.html, 2024. Deployment of 7,000+ robots including AMRs and collaborative systems. Accessed: 2025-10-31.
165. FedEx and Nimble. FedEx Announces Expansion of FedEx Fulfillment With Nimble Alliance. https://newsroom.fedex.com/newsroom/global-english/fedex-announces-expansion-of-fedex-fulfillment-with-nimble-alliance, 2024. Partnership for autonomous fulfillment robots in warehouse operations. Announced September 2024. Accessed: 2025-10-31.
166. Built Robotics. Autonomous Construction Equipment. https://www.builtrobotics.com, 2024. AI-powered autonomous systems for bulldozers, excavators, and construction machinery. Accessed: 2025-10-31.
167. Wasay, F.A.; Rahman, M.A.; Ghouse, H. GAMORA: A Gesture Articulated Meta Operative Robotic Arm for Hazardous Material Handling in Containment-Level Environments. *arXiv preprint arXiv:2506.14513* **2025**.
168. Colan, J.; Davila, A.; Yamada, Y.; Hasegawa, Y. Human-Robot collaboration in surgery: Advances and challenges towards autonomous surgical assistants, 2025, [arXiv:cs.RO/2507.11460].
169. Interact Analysis. Industrial Robot Forecast Update: Wide Variations Across Robot Types, Regions and Industries. https://interactanalysis.com/insight/industrial-robot-forecast-update-wide-variations-across-robot-types-regions-and-industries/, 2024. Market forecast for industrial robot shipments. Accessed: 2025-10-31.
170. European Commission. EU AI Act Implementation Timeline. https://ai-act-service-desk.ec.europa.eu/en/ai-act/eu-ai-act-implementation-timeline, 2025. Accessed: 2025-10-28.
171. Chee, F.Y. Code of Practice to Help Companies with AI Rules May Come End 2025, EU Says. https://www.reuters.com/business/media-telecom/code-practice-help-companies-with-ai-rules-may-come-end-2025-eu-says-2025-07-03/, 2025. Accessed: 2025-10-28.
172. Intel. Intel RealSense Depth Camera D455, 2020.
173. Microsoft. Azure Kinect DK. https://azure.microsoft.com/en-us/products/kinect-dk, 2019. RGB-D camera with time-of-flight depth sensor for collaborative robotics. Accessed: 2025-10-31.
174. Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A.J.; Conradt, J.; Daniilidis, K.; et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* **2022**, *44*, 154–180.
175. Rebecq, H.; Ranftl, R.; Koltun, V.; Scaramuzza, D. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence* **2021**, *43*, 1964–1980.
176. Robotiq. Most Popular Uses for Force Torque Sensors in Industry, 2015. Survey of industrial force-torque sensor applications. Last updated 2025.
177. Analog Devices. A Technical Note on a Tactile Sensor Prototype. https://www.analog.com, 2025. Accessed: 2025-10-27.
178. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.; Han, S. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 2774–2781.