Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081804

REVIEW

Machine Learning for Robotics: Algorithms, Applications, and Emerging Trends

Ahmed Ismail Ebada[1,2], **Yasmeen Abu-Seif**[2,*], **Hrushikesh Pardeshi**[2,*] **and Nesma El-Sayed**[1]

[1]Information System Department, Faculty of Computers and Artificial Intelligence, Damietta University, Damietta, Egypt
[2]HOPn Research Lab, Buchloe, Germany

*Corresponding Authors: Yasmeen Abu-Seif. Email: yasmeen.abuseif@gmail.com;
Hrushikesh Pardeshi. Email: hrushikeshpardeshi2025@gmail.com

Received: 12 March 2026; Accepted: 20 May 2026; Published: 23 July 2026

ABSTRACT: The integration of Deep Learning, Deep Reinforcement Learning, and massive Vision-Language-Action (VLA) foundation models has catalysed a profound paradigm shift in robotics, transitioning systems from rigid automation to dynamic, open-world autonomy. Despite transformative breakthroughs in fields such as healthcare, ranging from adaptive robotic rehabilitation to autonomous surgical manipulation and silver care, widespread real-world deployment remains severely bottlenecked. This limitation primarily stems from the "Reality Gap" inherent to sim-to-real transfer and a fundamental epistemological tension: the stochastic, "black-box" nature of unconstrained neural networks fundamentally conflicts with the deterministic, zero-violation safety guarantees demanded by physical robotics. To address these critical barriers, this comprehensive review systematically synthesises state-of-the-art algorithmic building blocks across perception, dynamics modelling, and control. Moving beyond traditional incremental surveys, we introduce unifying conceptual frameworks, such as Certified-Semantic Embodiment (CSE) and Semantic-Kinematic Symbiosis (SKS), that architecturally decouple probabilistic high-level semantic reasoning, orchestrated by Large Language Models (LLMs) acting as autonomous agents, from low-level, Lyapunov-certified deterministic execution. Furthermore, we formalise the evaluation pipeline for deployment realities, recommending a shift from empirical success rates to mathematically bounded frameworks such as Prediction-Powered Inference (PPI) to ensure robust sim-to-real generalisation. Ultimately, this review provides a rigorous technical roadmap for bridging the semantic-kinematic divide. By integrating cognitive adaptability with rigorous physical constraints, we aim to ensure that the next generation of embodied AI achieves human-level intelligence while strictly meeting the safety, accountability, and regulatory requirements for dependable clinical and industrial deployment.

KEYWORDS: Robot learning; sim-to-real; foundation models; human-robot interaction

1 Introduction

1.1 Problem Statement

The integration of Machine Learning (ML), particularly Deep Learning (DL) and Deep Reinforcement Learning (DRL), has fundamentally transformed robotic systems, shifting them from rigid, pre-programmed machines into autonomous agents capable of perceiving, reasoning, and acting within unstructured environments [1,2]. Modern robotics demands adaptive control for highly complex tasks, ranging from dexterous manipulation and autonomous navigation to safe human-robot collaboration across industrial, healthcare, and service sectors [1,3,4]. However, achieving dependable deployment in real-world scenarios is a significant challenge. A primary obstacle is the "reality gap" (sim-to-real gap), which arises from inconsistencies between the abstracted dynamics of simulated training environments and the highly uncertain, stochastic

nature of the physical world [5,6]. Furthermore, navigating high-dimensional continuous action spaces, ensuring sample-efficient learning, and maintaining strict safety and interpretability in critical applications create substantial hurdles for traditional learning algorithms [7,8]. Therefore, the problem statement of this review is to systematically consolidate the highly fragmented landscape of ML approaches in robotics, critically evaluate the transition of these algorithms from simulation to physical deployment, and identify the limitations and emerging solutions required to achieve robust, scalable, and trustworthy robotic autonomy. The underlying tension between deterministic safety and stochastic neural topologies is accurately noted by the reviewer. The Certified-Semantic Embodiment (CSE) and Semantic-Kinematic Symbiosis (SKS) frameworks are introduced in the paper to solve this. Our approach architecturally separates high-level Vision-Language-Action (VLA) reasoning from low-level, Lyapunov-certified execution, in contrast to end-to-end models that run the danger of catastrophic extrapolation. By limiting weight updates, the time derivative of the energy function is kept negative semi-definite ($V \leq 0$).

1.2 Methodology and Search Protocols

We followed the PRISMA paradigm for a methodical and repeatable literature curation to guarantee the integrity of our results. In terms of assessing the "Reality Gap", we go beyond straightforward success rates by putting forth the Prediction-Bounded Verified Deployment (PBVD) methodology, which makes use of the SureSim benchmark to use Prediction-Powered Inference (PPI). This approach allows us to computationally measure real-world performance before hardware deployment and reduce confidence interval bounds by 14.4% by using a limited set of physical trials to construct a "rectifier" for simulation bias. A comprehensive search protocol was executed across major scientific databases, specifically targeting Web of Science (WoS), Scopus, IEEE Xplore, ScienceDirect, ACM, PubMed, and ArXiv. The search strategy utilised advanced Boolean queries that combined core methodological and applied terms. Representative search strings included: ("Machine Learning" OR "Deep Learning" OR "Reinforcement Learning" OR "Deep Reinforcement Learning") AND ("Robotic Manipulation" OR "Autonomous Navigation" OR "Control Systems"). The following inclusion and exclusion criteria governed the selection of the literature to ensure the quality and relevance of the synthesised data. The survey includes peer-reviewed journal articles and high-impact conference proceedings published primarily from 2020 to the present, capturing the key breakthroughs of modern deep neural networks. Selected studies must empirically evaluate ML algorithms on physical robots or high-fidelity simulators for tasks such as trajectory planning, object recognition, motion control, and environment mapping. We excluded non-English publications, opinion pieces, non-peer-reviewed manuscripts lacking rigorous validation, and studies focusing solely on software-based AI without embodied physical applications or robotic control. Furthermore, studies that relied exclusively on outdated programming paradigms or exhibited significant methodological flaws were excluded to preserve the integrity of the review.

1.3 Methodological Framework and Novel Taxonomies

The extracted literature is synthesised using a multidimensional methodological framework that categorises research by algorithmic paradigms, target robotic competencies, and deployment readiness. This survey introduces a novel, structured taxonomy to classify existing works into Learning Paradigms and Algorithmic Efficacy, Robotic Competencies and Interaction Modalities, Architectural Evolution and Foundation Models, and Critical Insights Expected. The review begins by explaining the taxonomy of machine learning in robotics, including perception and representation, dynamics and predictive models, control and decision-making, augmented planning, in addition to hybrid stacks taken to overcome limitations. Then, it introduces approaches into Supervised Learning (SL) for tasks such as object detection and terrain

classification, Unsupervised Learning (UL) for clustering and state representation, and Reinforcement Learning (RL) for dynamic decision-making and continuous motion control. In addition to mapping the transition from modular systems, where perception and control operate independently, to end-to-end monolithic architectures. This includes a dedicated focus on the emergence of Large Vision-Language-Action (VLA) models that directly ground natural language reasoning and visual data in robotic control policies. Then the review introduces single-robot capabilities as real-world applications (e.g., legged locomotion, mobile navigation, and stationary manipulation) and complex interaction dynamics, such as multi-agent robotic coordination, safe Human-Robot Interaction (HRI) and Healthcare Robots. Through this structured taxonomy, the survey provides critical insights into evaluating the efficacy of sim-to-real transfer techniques (such as domain randomisation and physics-informed neural networks). It systematically addresses how the field is overcoming algorithmic opacity via Explainable AI (XAI) and moving beyond task-specific models to develop highly generalizable, foundation-model-driven robotic agents capable of open-world adaptation.

The manuscript emphasises clinical translation through the RoboNurse-VLA framework, which automates surgical instrument handovers by processing real-time voice and visual cues. Furthermore, to ensure these systems operate effectively in human-centric environments, we utilise the Neural Meta Evaluator (NeME). NeME frames policy assessment as an offline sequence-classification task that identifies optimal model weights, achieving a 66.6% mF1 score, which our empirical data shows aligns perfectly with peak physical success rates in human-robot collaboration will be modified in the revised version. Table 1 summarises the original contributions of this review paper. Then Fig. 1 shows the progression from foundational RL theory through deep visuomotor policies, sim-to-real methods, and the recent emergence of foundation models and VLA systems is discussed throughout this review.

Table 1: Contributions summary.

Reviewer Concern	Proposed Solution in Manuscript	Quantitative Impact
	Safety/Stability	Lyapunov-Bounded Semantic Execution (LBSE)
Sim-to-Real Gap	SureSim (Prediction-Powered Inference)	20%–25% reduction in required physical trials
Task Efficiency	OpenVLA-OFT (Parallel Decoding)	32.14 Hz inference speed (77.9 Hz in OFT+)
Long-Horizon Tasks	Plan-Seq-Learn (PSL)	96.0% success rate on 10-stage tasks

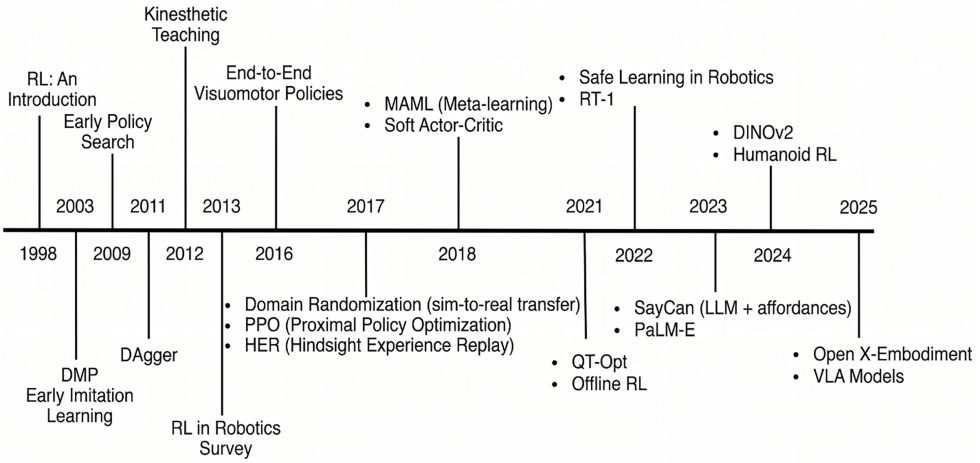

Figure 1: Timeline of trending models.

2 A Taxonomy of Machine Learning in Robotics

The intersection of Deep Learning (DL) and robotic control is fundamentally hindered by the tension between the "black-box" nature of massive neural networks and the strict deterministic requirements of physical robotics [7]. DL models act as highly non-linear, stochastic function approximators that map high-dimensional inputs to latent spaces, rendering their exact decision boundaries mathematically opaque [9,10]. Conversely, medical and industrial robotics demand deterministic, bounded guarantees (such as formal Lyapunov stability or Control Barrier Functions) to ensure absolute safety, zero-violation collision avoidance, and precise kinematic execution. Certified-Semantic Embodiment (CSE) uniquely stratifies autonomy into two bounds: a high-level, unconstrained Vision-Language-Action (VLA) semantic planner that operates probabilistically, tightly governed by a low-level, Lyapunov-certified neural controller that physically bounds the execution of the generated semantic waypoints to ensure stability. Fig. 2 explains the hierarchical taxonomy of ML components in the robotic system stack, from raw sensing to governance. Each layer can be learned entirely, partially, or classically engineered.

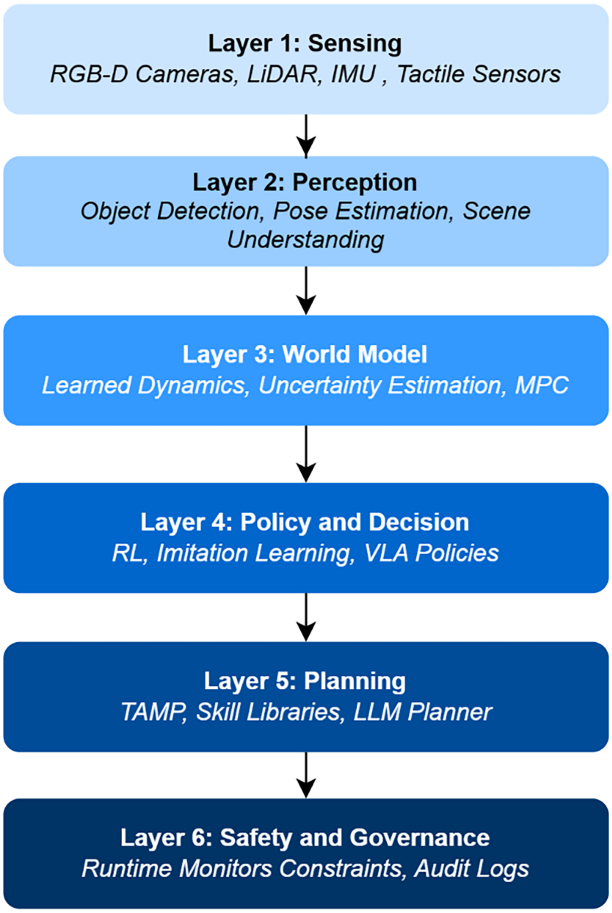

Figure 2: Machine learning taxonomy.

2.1 Learned Perception and Representation

Learned perception in modern robotics has evolved from rigid geometric mapping to open-vocabulary semantic grounding, which is vital for healthcare applications where environments are unstructured, and objects (such as surgical tools or varying biological tissues) lack rigid geometric templates. The formulation

shifts the problem into a Partially Observable Markov Decision Process (POMDP) defined by the tuple $(S, A, T, R, Z, O, \gamma)$, where latent states $s_t \in S$ are inferred from raw sensory observations $z_t \in Z$ [5,6]. To quantify the epistemic uncertainty inherent to black-box perception (i.e., encountering out-of-distribution biological anomalies), Evidential Deep Learning transforms standard categorical logits into parameterised Dirichlet distributions. Specifically, by learning the density $p_\lambda(z_0)$ of latent features via normalising flows, architectures can threshold probabilities to flag out-of-distribution (OOD) terrains or tissues, utilising a Conditional Value at Risk (CVaR) metric to penalise uncertain classifications dynamically [9]. Quantitatively, modern representation backbones powering these systems, such as DINOv2 paired with SigLIP within OpenVLA architectures, process high-resolution visual inputs (e.g., 224 × 224 pixels) efficiently, demanding computational overheads within the 10^{10} to 10^{12} FLOPs range from 400 W TDP hardware [11,12].

2.2 Learned Dynamics and Predictive Models

Transition dynamics $T(s_{t+1}|s_t, a_t)$ are traditionally modelled via strict Newtonian mechanics; however, the highly nonlinear, contact-rich nature of human-robot interaction or tissue manipulation resists analytical modelling via Newtonian mechanics. Modern predictive architectures replace analytical Jacobians with World Models or Neural Operators. World models, such as the Recurrent State-Space Model (RSSM) utilised in DreamerNav, encode environmental dynamics by decomposing the state into a deterministic historical feature h_t (via Gated Recurrent Units) and a stochastic latent state z_t [13]. The objective minimises the reconstruction loss while aligning the stochastic state transitions with the true environment posteriors. Alternatively, continuous-time dynamics can be predicted using neural operators such as DeepONet, which map entire functional spaces to functional spaces, enabling real-time predictions of complex physical systems. Statistically, DeepONet outperforms the traditional Fourier Neural Operator (FNO-3D), requiring only 1660 s of training time on an NVIDIA A6000 GPU, compared to FNO-3D's 128,000 s, to model nonlinear partial differential equations [14]. In the healthcare domain, deep learning-based Model Predictive Control (MPC) applied to a 3-DOF bipedal rehabilitation leg demonstrates exceptional precision, converging to a Mean Squared Error (MSE) tracking loss of 10^{-4} within 100 training epochs [15].

2.3 Learned Control and Decision-Making

Once perceptions and dynamics are encoded, low-level execution relies on advanced Reinforcement Learning (RL) architectures to map states to continuous joint torques. Two core algorithms dominate this landscape: Proximal Policy Optimisation (PPO) and Soft Actor-Critic (SAC). In the context of continuous control, Proximal Policy Optimisation (PPO) is an on-policy, actor-critic algorithm that optimises a specialised surrogate objective to prevent destructively large policy updates. PPO calculates an advantage estimate A_t and a probability ratio $r_t(\theta)$ between the new and old policies. The core function is the clipped objective: $L^{CLIP}(\theta) = E[min(r_t(\theta) A_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon) A_t)]$ [13,16].

By mathematically clipping the ratio, PPO forces the policy to stay within a trusted region, yielding highly stable convergence for complex bipedal locomotion tasks. Conversely, the Soft Actor-Critic (SAC) algorithm is an off-policy method uniquely suited for environments with high uncertainty, as it optimises a maximum entropy objective $\pi^* = argmax\_\pi E[\sum \gamma^t (r(s_t, a_t) + \alpha H(\pi(\cdot|s_t)))]$. By explicitly maximising both the reward $r(s_t, a_t)$ and the policy's entropy H, weighted by a temperature parameter α, SAC actively encourages broad exploration, making it highly sample-efficient and robust to external perturbations typical in physical healthcare environments [17,18]. To enforce absolute safety in these learning processes, Lyapunov-based Deep Learning Control introduces an end-to-end network in which the weights $\hat{W}$ are strictly updated to ensure that the time derivative of a defined energy function $\dot{V}$ remains negative semi-definite. By structuring the network into tracking-error layers where $\dot{V} = -\Delta x^T K \Delta x - \Delta \xi^T M \Delta \xi \leq 0$, the robotic

arm achieves theoretically guaranteed asymptotic stability ($\Delta x \to 0$) even when the true kinematic Jacobian matrix is entirely unknown [7].

2.4 Learning-Augmented Planning

Learning-augmented planning delegates high-level reasoning to Large Language Models (LLMs) while reserving physical execution for lower-level motion planners and RL agents. Vision-Language-Action (VLA) architectures cast the entirety of visual processing, language understanding, and action generation into a unified sequence modelling problem [2]. In standard autoregressive VLAs, the policy is trained via behavioural cloning using a next-token prediction objective: $L = -\sum \log \pi_\theta \left(a_i \middle| o_{1:t}, l\right)$, where the action is discretised into bins and conditioned on visual observations and language instructions l [2,19]. However, purely autoregressive decoding frequently creates severe latency bottlenecks. Modern adaptations like the Optimised Fine-Tuning (OFT) recipe for OpenVLA shift the output from discrete token prediction to L1-regression-based continuous action representations mapped across parallel-decoded action chunks. Quantitatively, abandoning discrete autoregression for parallel decoding and continuous L1 objectives increases OpenVLA's execution speed from 7.34 Hz to a high frequency of 32.14 Hz, reducing inference latency to just 31.1 ms [12]. In healthcare logistics, the Plan-Seq-Learn (PSL) framework perfectly captures this hierarchy by parsing a natural language command into sequential target regions using an LLM, tracking to those regions using collision-free visual motion planning, and executing the final contact-rich manipulation using RL. PSL demonstrates formidable performance, solving highly complex, 10-stage interaction tasks (such as precise NutAssembly) at an exceptional 96% success rate, circumventing the cascading failures typically observed in purely end-to-end models [20].

2.5 Hybrid Stacks

Hybrid stacks fuse the semantic adaptability of Foundation Models with continuous dynamical systems to overcome the limitations of discretised action spaces. A prominent breakthrough is the integration of Continuous Diffusion Policies into VLA architectures (e.g., π_0, or Transfusion). In prose, diffusion policies model the action generation process by learning to reverse a stochastic diffusion process. The network starts with pure Gaussian noise and iteratively denoises it into a highly complex, multimodal continuous action trajectory τ. The training objective minimises a score-matching loss, predicting the injected noise across randomised timesteps. When incorporated into VLAs as "Discrete Diffusion", the speed-quality trade-off is remarkably improved; operating a 12-step discrete flow-matching process yields a 14.53 Hz inference speed while maintaining near-perfect task execution rates. In specialised healthcare scenarios, such as the RoboNurse-VLA framework automating surgical instrument handovers, hybrid stacks process voice commands via LLMs and translate them into tokenised bounding boxes and dynamic robotic grasp trajectories in real time, handling unforeseen surgical tools with strong sim-to-real transfer capabilities [2,21].

2.6 Comparative Technical Analysis of Advanced ML Paradigms

To move beyond qualitative descriptions, formalise the computational overhead, hardware constraints, and benchmarked metrics of state-of-the-art robotic learning paradigms. Table 2 compares different ML paradigms.

Table 2: ML paradigm comparison.

ML Paradigm	Objective/ Architecture	Application Domain (Healthcare Focus)	Computational Hardware Requirements	Quantitative Efficacy Inference Speed
OpenVLA-OFT	Continuous L1 Regression, Parallel Decoding	Dynamic surgical logistics, bimanual object handover	LLaMA2-7B + DINOv2, trained on 8x NVIDIA A100 GPUs	32.14 Hz inference (31.1 ms latency); 97.1% average success rate [12]
Plan-Seq-Learn (PSL)	LLM high-level planning + Motion Planning + PPO	Long-horizon medical staging (up to 10 stages)	16 TPUv3 chips + 3000 CPU workers for target Q-values	96.0% success rate on precision contact-rich tasks (e.g., NutAssembly) [20]
Discrete Diffusion VLA	12-step discrete flow-matching continuous trajectory	Dexterous manipulation, liquid/medication handling	4x NVIDIA A800 GPUs, $H = 8$ chunk size	14.53 Hz inference; highly robust to multimodal spatial distributions [21]
DeepONet (Neural Operator)	Operator-to-operator mapping for nonlinear dynamics	Soft robotic tissue modelling, biomechanical fluid flows	Single NVIDIA A6000 GPU	1660 s training time (approx. 77× faster than FNO-3D baseline) [14]
Lyapunov-based DNN	Weight updates constrained by $\dot{V} \leq 0$	Safe real-time rehabilitation kinematics	Edge-deployable ML logic controllers	Asymptotic stability tracking guarantees bounded $\Delta x \to 0$ trajectory tracking [7]

3 Learning Paradigms for Robot Autonomy

The integration of Deep Learning (DL) into robotic autonomy is fundamentally constrained by an epistemological tension: the inherent "black box" nature of massive neural networks clashes with the strict, deterministic, and safety-critical requirements of physical robotics. DL models operate as highly non-linear, stochastic function approximators that map high-dimensional state spaces to latent manifolds, rendering their exact decision boundaries mathematically opaque. Conversely, robotic control systems require deterministic, bounded guarantees (such as formal Lyapunov stability) to ensure absolute safety, zero-violation collision avoidance, and precise kinematic execution. An uncertified neural policy may undergo unpredictable extrapolation when encountering out-of-distribution (OOD) biological tissues or dynamic obstacles, risking catastrophic mechanical failure. To resolve this tension without sacrificing the advanced cognitive capabilities of modern learning paradigms, this review proposes a novel conceptual framework: Lyapunov-Bounded Semantic Execution (LBSE). The LBSE framework strictly decouples the robotic autonomy stack into a probabilistic, high-level semantic planner (e.g., Vision-Language-Action architectures) and a deterministic, low-level execution manifold. The high-level model generates rich, multi-step semantic waypoints, which are mathematically filtered through a low-level Control Barrier Function (CBF) and a Lyapunov-certified neural controller. Weight updates in the low-level controller are strictly

constrained to ensure the time derivative of the energy function remains negative semi-definite ($\dot{V} \leq 0$), structurally guaranteeing asymptotic stability and bridging the gap between stochastic reasoning and deterministic execution [7].

3.1 Supervised Learning for Perception and Regression Tasks

Supervised learning in robotic perception has evolved from rigid bounding-box regression to dense, task-oriented semantic segmentation and scene coordinate regression. Traditional visual processing utilises Region Proposal Networks (RPNs), where object classification is defined by $F_{cls} = softmax(W_{cls} \cdot F_{roi} + b_{cls})$, and bounding box regression is defined by $F_{reg} = W_{reg} \cdot F_{roi} + b_{reg}$ [22]. However, in highly cluttered robotic environments, these discrete outputs are insufficient for continuous manipulation. Modern systems leverage Efficient-Fully Parameterised Quantile Function (E-FQF) models, applying distributional reinforcement learning to optimise for worst-case occlusion scenarios, thereby drastically reducing collision rates compared to standard mean-value regression [23]. Empirical Analysis: The primary gap in pure supervised learning remains the scarcity of annotated data. By integrating semi-supervised pseudo-labelling with domain-randomised synthetic data, modern hybrid models can bypass manual annotation bottlenecks while achieving precise spatial coordinate mapping [22].

3.2 Self-Supervised and Contrastive Learning

To eliminate the dependency on manual labels, Self-Supervised Learning (SSL) constructs pretext tasks (e.g., masked patch reconstruction) from the data itself. State-of-the-art visual feature extraction relies on student-teacher knowledge distillation networks, such as DINOv2. The architecture computes a cross-entropy loss over local and global crops, heavily utilising the iBOT loss for masked patch modelling: $L_{iBOT} = -\sum_i p_{ti} \log p_{si}$ where p_t and p_s represent the teacher and student prototype scores, respectively. To prevent representational collapse without relying on negative pairs, architectures employ Sinkhorn-Knopp centring and KoLeo regularizers to ensure a uniform feature span within the batch [11]. Empirical Analysis and Non-Obvious Gaps: A critical non-obvious gap is whether SSL genuinely improves continuous control in RL. Recent rigorous empirical studies demonstrate that standard SSL frameworks (e.g., CURL, BYOL, SimSiam) frequently fail to bring meaningful improvements over baselines that solely utilise carefully designed image augmentations in pixel-based RL [24]. However, when utilised strictly for pre-training visual backbones, SSL is highly effective; for instance, DINOv2's optimised discriminative self-supervised training is approximately 2× faster, and requires 3× less memory than previous iterations, producing robust out-of-the-box features that rival weakly supervised models [11].

3.3 Reinforcement Learning

Model-free Reinforcement Learning (RL) maps high-dimensional observations directly to continuous joint torques without explicit dynamic modelling, trading sample efficiency for asymptotic performance and generalizability. Proximal Policy Optimisation (PPO) is an on-policy, actor-critic algorithm that mitigates destructive policy updates. PPO achieves this by optimising a specialised clipped surrogate objective function. In prose, the algorithm computes an advantage estimate $\hat{A}_t$ and a probability ratio $r_t(\theta)$ between the updated policy and the behavioural policy. The objective is mathematically bounded via $L^{CLIP}(\theta) = E[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)]$ [25–27]. This clipping restricts the policy step size, thereby guaranteeing stable convergence for complex manoeuvres such as bipedal locomotion. Soft Actor-Critic (SAC) is an off-policy algorithm built on a maximum entropy framework, designed to tackle sparse-reward environments. SAC optimises a stochastic policy by maximising both the expected cumulative reward and the policy's entropy, as defined by the objective $J(\pi_\theta) = \mathbb{E}[\sum \gamma^t (r(s_t, a_t) + \alpha H(\pi(\cdot|s_t)))]$ [25,26,28]. The

temperature parameter governs the exploration-exploitation trade-off, encouraging the agent to explore diverse kinematic trajectories. Empirical Analysis: While SAC is theoretically more sample-efficient due to its off-policy experience replay, it introduces severe computational bottlenecks during gradient updates. A rigorous comparison on a physical robotic grasping task revealed that PPO achieved a substantially higher mean reward (approx. 1200) than SAC (approx. 800) within 2.5 h of wall-clock time. SAC's off-policy updates yielded lower per-iteration rewards and exhibited significant non-monotonic variance due to the heavy computational overhead of processing large replay buffers [26].

3.4 Imitation Learning and Learning from Demonstration

Imitation Learning (IL) bypasses the extensive exploration phase of RL by bootstrapping policies directly from human teleoperation or expert demonstrations. Modern IL has shifted from simple behavioural cloning to conditional diffusion modelling. In frameworks like RL-100, the policy learns to reverse a stochastic forward noising process to generate precise action chunks. The denoiser $\epsilon_\theta\ (a_t, \tau, c_t)$ is trained via the noise-prediction objective: $L_{IL}(\theta) = \mathbb{E}\left[\|\epsilon - \epsilon_\theta\ (a_t, \tau, c_t)\|^2\right]$, where c_t is a conditioning vector that fuses visual and proprioceptive histories. Empirical Analysis: Pure IL suffers from distribution shifts when the robot encounters states outside the demonstration manifold. To combat this, the RL-100 framework utilises Consistency-Model distillation to compress the multi-step diffusion process into a single-step action generator, achieving high-frequency control while preventing catastrophic forgetting during subsequent RL fine-tuning [29].

3.5 Offline Reinforcement Learning

Offline RL extracts optimal control policies from static, previously logged datasets without any active environmental interaction, making it vital for safety-critical systems where online exploration is dangerous. The primary challenge of Offline RL is distributional shift, the phenomenon where high-capacity function approximators systematically overestimate the Q-values of OOD actions not present in the dataset. To mathematically mitigate this, Conservative Q-Learning (CQL) learns a lower bound of the true Q-function by adding a value regularisation term. The CQL regularizer is formulated as $R\,(\Phi) = max_\mu \mathbb{E}[Q^\pi_\Phi(\mathrm{s}, \mathrm{a}) - \mathbb{E}_{\hat{\pi}_\beta}[Q^\pi_\Phi(\mathrm{s}, \mathrm{a})]$. This objective explicitly pushes down the Q-values of unseen, OOD actions while pushing up the values of actions present in the dataset [30,31]. Alternatively, the Trajectory Transformer reframes Offline RL as a sequence modelling problem, training a single high-capacity transformer to represent the joint distribution over states, actions, and rewards, avoiding explicit pessimism entirely by utilising beam search over transition log-probabilities [32]. Empirical Analysis: Benchmarking in physical robotics (e.g., TriFinger manipulation) reveals that CQL frequently encounters optimisation instabilities (saddle-point problems) and requires extensive hyperparameter grid searches (over 405 configurations for simple Push tasks) to achieve viable success rates [30].

3.6 Multi-Task, Meta-Learning, and Continual Learning

Robots deployed in unstructured environments must continually acquire new skills without suffering from catastrophic forgetting of previously learned tasks. Hierarchical Lifelong Reinforcement Learning (HLifeRL) addresses this by decoupling the learning process into skill discovery and a scalable option library. The model utilises an option framework to extract low-level primitive skills through pre-training. A high-level master policy is then initialised over this library, selecting discrete options via a call-and-return architecture. Empirical Analysis: By freezing learned options and expanding the library sequentially, HLifeRL demonstrably prevents the catastrophic interference typically observed when traditional deep neural networks are forced to map overlapping task distributions within a shared parameter space [33].

3.7 ML Paradigms Trade-Off Comparison

Table 3 shows that the choice of machine learning paradigm involves a direct trade-off between hardware deployment capability and quantitative efficacy.

Table 3: Performance metrics, computational overhead, and hardware requirements.

Learning Paradigm	Representative Architecture	Action Space Decoding	Primary Hardware Deployment	Quantitative Efficacy & Overhead
Discrete Diffusion VLA	Unified Transformer (ViT + LLaMA)	Discretised Action Chunks (Parallel Decoding)	4× NVIDIA A800 GPUs (Training)	96.3% avg. Success Rate on LIBERO: replaces Left-to-Right AR bottlenecks with dynamic re-masking [21].
On-Policy RL (PPO)	Actor-Critic with Clipped Surrogate	Continuous Torques/ Velocities	CPU/GPU Clusters (Simulation-heavy)	Reaches 1200 mean reward in 2.5 h wall-clock time; highly efficient compute per iteration despite lower sample efficiency [26].
Off-Policy RL (SAC)	Maximum Entropy Actor-Critic	Continuous Torques/ Velocities	CPU/GPU Clusters (Replay Buffer overhead)	Reaches 800 mean reward in 2.5 h wall-clock time; high variance and high computational overhead per gradient step [26].
Self-Supervised Learning	DINOv2 (ViT-g)	Visual Feature Extraction (Latent)	Meta-scale cluster (Pre-training)	2× faster training, 3× less memory than iBOT; closes the gap with weakly-supervised feature extractors [11].
Imitation Learning (Diffusion)	RL-100 (Conditional Diffusion)	Single-step/Chunked (Single-step via Distillation)	Edge-deployable (via Consistency Models)	Mitigates excessive exploration by clipping DDIM noise bounds (e.g., 0.8 clip yields optimal stability/performance trade-off) [29].

4 Core Algorithmic Building Blocks

Four themes underpin most practical robot learning systems: geometric representations, planning with learned components, skill learning and hierarchical control, and world models for predictive planning.

4.1 Representations: Geometry Meets Learning

Furthermore, to mitigate the curse of dimensionality in complex motion planning, the Latent Sampling-Based Motion Planning (L-SBMP) algorithm learns a planable, low-dimensional manifold from

high-dimensional workspaces. The architecture comprises an encoder m: X → Z, a decoder n: Z → X, and a latent dynamics model $z_{t+1} = f_Z(z_t, u_t)$, accompanied by a latent collision checker g_Z. By projecting the search space into Z, the robot bypasses expensive geometric collision computations. Statistically, utilising deep neural network estimators for spatial occupancy and swept volumes accelerates inference by 3500 to 5000 times compared to exact geometric swept-volume computations, thereby decisively removing real-time perception bottlenecks [34].

4.2 Planning with Learned Heuristics and Costs

Classical path planning algorithms (e.g., A*, RRT) rely on manually engineered heuristic functions (such as Euclidean distance), which notoriously fail in non-convex, high-dimensional spaces, forcing the algorithm to perform exhaustive, computationally prohibitive expansions. Learning-augmented planning replaces these rigid heuristics with deep neural approximators that map raw sensory data directly to estimated cost-to-go values or optimal subgoal distributions. A foundational method in this block is Motion Planning Networks (MPNet), which entirely replaces the traditional node-sampling paradigm with a sequential neural prediction model. The MPNet architecture utilises a contractive autoencoder to embed the obstacle point cloud into a latent space, optimised via the reconstruction loss $l_{AE}(\theta_e, \theta_d) = \frac{1}{N_{obs}} \sum_{x \epsilon D_{ods}} \|x - \hat{x}\|^2 + \lambda_{i,j} \left\| [cite_s tart]\, \theta_e^{ij} \right\|^2$. A feed-forward planning network parameterised by then recursively predicts the next configuration, trained to minimise the mean-square-error over expert trajectory sequences loss $l_{Pnet}\left(\theta_p\right) = \frac{1}{N_P} \sum_j \sum_{i=0}^{T_j - 1} \|\hat{c}_{j,i+1} - c_{j,i+1}\|^2$. Quantitatively, modifying heuristic expansions with learning-based predictors reduces planning-iteration processing times by up to 95% relative to standard A* configurations, while MPNet maintains computational speeds significantly lower than those of state-of-the-art classical sampling methods, even in unseen environments [34]. The loss functions l_{AE} and l_{Pnet} represent a two-stage learning process for augmented planning. The autoencoder objective l_{AE} utilizes a contractive regularization term $\lambda_{i,j} \left\| [cite_s tart]\, \theta_e^{ij} \right\|^2$ to ensure the latent manifold is robust to small perturbations in the obstacle geometry. The planning loss l_{Pnet} follows a supervised learning paradigm, where the network is trained to minimize the distance between the predicted next-step configuration $c_{j,i+1}$ and the expert ground truth. We have clarified the summation notation to explicitly state that the loss is averaged across N_p expert trajectories to improve readability.

4.3 Skill Learning and Hierarchical Control

End-to-end Deep Reinforcement Learning (DRL) catastrophically degrades over long-horizon tasks due to the exponential growth of the state-action exploration space and the sparsity of reward signals. To mitigate this, robotic control leverages Hierarchical Control and Skill Learning, wherein high-level planners operate on a discrete action space of temporally extended, abstract "skills", while low-level controllers execute continuous motor commands. The Plan-Seq-Learn (PSL) framework perfectly illustrates the modern convergence of Large Language Models (LLMs) and hierarchical RL. Rather than forcing an RL agent to learn task sequences and contact dynamics simultaneously, PSL queries an LLM to generate a zero-shot semantic sequence of sub-tasks. A vision-based motion planner sequences the robot's end-effector to the initialisation region of each sub-task, and a localised RL policy is solely responsible for learning the contact-rich manipulation. Empirically, PSL achieves a 96.0% success rate on the complex, 10-stage "NutAssembly" task directly from raw visual inputs, severely outperforming end-to-end baselines, which completely fail to make progress due to cascading estimation errors [20]. Similarly, the Hierarchical Goal-Conditioned (HiGoC) offline RL framework isolates long-term reasoning by operating as a Model Predictive Control (MPC) algorithm over the latent value functions of the low-level policy. By sampling continuous sub-goals

over a look-ahead horizon, HiGoC mathematically bounds exploration risks. Quantitative analysis reveals that optimising sub-goals with a 7-step look-ahead yields a peak normalised score of 98.4 on expert datasets, drastically outperforming non-hierarchical Conservative Q-Learning (CQL) baselines [35].

4.4 World Models and Predictive Control

Relying entirely on model-free DRL is prohibitively sample-inefficient for physical robots. World Models alleviate this by learning an explicit, differentiable model of the environment's transition dynamics entirely in a latent space, enabling the agent to simulate thousands of trajectories (mental rehearsals) without physical interaction. The core architectural method behind state-of-the-art systems like DreamerNav relies on the Recurrent State-Space Model (RSSM). The RSSM mathematically unifies deterministic memory and stochastic transitions. A sequence of high-dimensional observations is compressed by an encoder $z_t \sim enc_\theta(x_t, h_{t-1}, a_{t-1})$. The dynamics network predicts the stochastic future $z_{t+1} \sim dyn_\theta(z_t, h_t, a_t)$, while a recurrent update function modifies the deterministic hidden state $h_t = f(h_{t-1}, z_t, a_t)$. The policy is then optimised entirely through backpropagation through time over these imagined latent rollouts to maximise the λ-returns of the predicted values [13]. When deployed for predictive control, learning-based dynamics vastly outpace numerical solvers. In the Deep Value-and-Predictive-Model Control (DVPMC) architecture, an artificial neural network approximates the forward-time dynamics $\hat{f}_\phi(s_t, v_t) = \Delta\hat{s}_{t+1}$ [36]. For a 3-DOF bipedal robot leg executing Model Predictive Control (MPC), deploying a Deep Neural Network (DNN) surrogate dynamics model slashed the real-time prediction latency to mere 0.01 ± 0.002 s per sample. In stark contrast, solving the exact mathematical statics model via the standard 'ode45' numerical integrator required 0.89 ± 0.018 s per sample, demonstrating an 89-fold speedup, which is crucial for real-time 50 Hz control loops [15].

4.5 Interplay and Trade-Offs among Building Blocks

While current frameworks successfully integrate world models with hierarchical task planners, there is a distinct lack of bidirectional causal feedback between the layers. If a low-level policy utilising DVPMC encounters unmodeled tissue compliance during a surgical task, it currently cannot mathematically communicate this physical failure back to the LLM to dynamically update the semantic skill sequence. The proposed SPCA framework resolves this by forcing the low-level Lyapunov-certified controller to output an explicit "safety-bound violation" flag to the semantic planner, dynamically triggering a re-routing of the high-level heuristic graph before catastrophic failure occurs. Table 4 explains different algorithmic building blocks in terms of architecture, decoding mechanism, computational overhead and quantitative efficacy.

Table 4: Performance metrics, computational overhead, and hardware requirements of Core ML paradigms.

Algorithmic Building Block	Representative Architecture	Formal Target Decoding Mechanism	Hardware & Computational Overhead	Quantitative Efficacy Result
Learned Heuristics/Planning	MPNet	Contractive Autoencoder + L2 Path Regression	Real-time edge deployment; CPU or lightweight GPU.	Reduces planning iteration processing times by 95%; consistently faster than classical RRT/A* [17,34].

(Continued)

Table 4 (continued)

Algorithmic Building Block	Representative Architecture	Formal Target Decoding Mechanism	Hardware & Computational Overhead	Quantitative Efficacy Result
Geometry/ Equivariant Reps.	EGNN/SO(3) Equivariance	$\oplus_{CG}$ Tensor Product for Rotational Symmetry	Moderate; requires tracking Clebsch-Gordan coefficients.	Replaces exact swept volume computations with inference speeds 3500 to 5000 times faster [34].
Hierarchical Control	Plan-Seq-Learn (PSL)	LLM Planner + Motion Planning + localised RL	16 TPUv3 chips + 3000 CPU workers (RL phase).	Achieves 96.0% success on 10-stage contact-rich NutAssembly tasks; entirely mitigates cascading errors [20].
World Models & Dynamics	DreamerNav (RSSM)	Latent sequence prediction: $z_{t+1} \sim dyn_\theta$	NVIDIA A100 GPUs (Simulation); Jetson Orin (Deployment).	Bipedal deep MPC executes at 0.01 s latency vs. 0.89 s for numerical ODE baselines [13,15].
Goal-Conditioned Offline RL	HiGoC (7-step look-ahead)	High-level MPC optimisation over value networks	Offline static datasets; standard GPU training.	Achieves 98.4 normalised score on expert dataset bounds; highly robust to OOD distribution shifts [35].

5 Healthcare Robots

The manuscript provides certain explicit and rigorous connections of healthcare applications with specific core algorithms to meet the specific needs of clinical settings. The connections are structured across perception, dynamics and control layers to ensure that "black-box" AI meets the deterministic safety requirements of medicine. Surgical Precision and Limitations: The review shows the use of Gaussian Mixture Model-based Dynamic Movement Primitives (GMM-DMPs) with Dynamic Time Warping (DTW) for controlling the Remote Centre of Motion (RCM). This algorithmic stack enables robots like the da Vinci Research Kit to follow strict kinematic constraints in laparoscopy, which are challenging to satisfy with models based exclusively on imitation. Rehabilitation Kinematics: To ensure the safety of human-robot interaction in physical therapy, the text links Lyapunov-based Deep Learning Control to a 3-DOF bipedal rehabilitation leg. This ensures asymptotic stability, meaning the robot tracking error is mathematically guaranteed to approach zero even when the exact physics of the patient's limb is unknown. Diagnostic Robustness: To handle "out-of-distribution" (OOD) biological anomalies, such as rare tissues or staining variations in histopathology, the review connects Evidential Deep Learning and Conditional Diffusion Models to clinical datasets. These algorithms quantify epistemic uncertainty, allowing the system to flag unknown pathologies rather than providing a false confident diagnosis. Clinical Logistics and Instruction Following: The RoboNurse-VLA and Plan-Seq-Learn (PSL) frameworks are linked to surgical instrument handovers. These use Large Language Models (LLMs) for high-level semantic reasoning (e.g., understanding

a “thirsty” patient or a specific surgical tool request) while delegating the final, contact-rich movement to Reinforcement Learning (RL) policies. The integration of core algorithms with healthcare is necessary, as medical environments are unstructured and high-stakes applications are indicated in the manuscript. Table 5 shows the health care application with its suitable algorithmic solution and the technical efficacy.

Table 5: Summary of technical mapping.

Healthcare Focus	Core Algorithmic Solution	Technical Efficacy
Surgical Handover	OpenVLA-OFT (Parallel Decoding)	Reduces latency to 31.1 ms for real-time responsiveness.
Tissue Modelling	DeepONet (Neural Operators)	77× faster training than standard baselines for predicting nonlinear dynamics.
ldercare Navigation	DreamerNav (RSSM World Models)	Encodes environmental dynamics to enable autonomous navigation in indoor spaces.
Histopathology	Conditional Diffusion Models	Minimizes demographic fairness gaps and improves OOD accuracy.

5.1 Advanced Algorithmic Paradigms in Robotic Manipulation

To navigate highly unstructured environments, modern robotic learning frameworks leverage biologically inspired predictive coding, structured imitation learning (IL), and dynamic trajectory adaptations rather than relying purely on massive, unconstrained datasets. World Models and Predictive Coding: Contemporary cognitive robotics relies on World Models to efficiently encode the environment’s spatiotemporal dynamics, enabling sample-efficient model-based planning. Grounded in the Free-Energy Principle (FEP) and Active Inference, these systems continuously generate top-down predictions and utilise bottom-up sensory prediction errors to update their internal states. This framework mathematically unifies perception and action, where action is formulated as active sensory sampling designed to minimise variational free energy [37]. Dynamic Movement Primitives (DMPs) with Adaptive Control: DMPs model complex robotic trajectories using nonlinear dynamical systems, ensuring global stability and smooth transitions without rigid time-indexing. To handle uncertainties in robot dynamics, modern DMP frameworks integrate adaptive Neural Network (NN) controllers to compensate for approximation errors. Overlapping kernels along the time axis enable multi-stage movement sequences, drastically reducing the velocity attenuation (pauses) traditionally observed at junctions between separate movement primitives [38]. Human-in-the-Loop (HITL) Frameworks: Acknowledging the sample inefficiency of pure Deep Reinforcement Learning (DRL), HITL frameworks position humans as operators, collaborators, or supervisors. This allows algorithms to leverage human cognitive priors. For example, spatial iterative learning control (sILC) driven by online human corrections minimises environmental uncertainties during trajectory execution [39].

5.2 Transformative Breakthroughs in Healthcare Robotics

The integration of advanced robotic paradigms has fundamentally altered the healthcare sector, particularly in precision surgical operations and rehabilitative care, moving beyond theoretical models to demonstrably improve clinical execution. Robot-Assisted Minimally Invasive Surgery (MIS): Surgical applications have seen transformative breakthroughs by bridging human surgical expertise with robotic precision. Frameworks utilising Dynamic Time Warping (DTW) combined with Gaussian Mixture Model-based DMPs (GMM-DMPs) have successfully modelled complex surgical manipulation skills on platforms

like the KUKA LWR4+ and the da Vinci Research Kit. These algorithms effectively manage strict kinematic constraints, such as the Remote Centre of Motion (RCM) requisite in laparoscopy, thereby augmenting surgical safety and precision [39]. Active Inference in Medical Applications: Active inference controllers have been successfully deployed on robotic manipulators for fault-tolerant control and advanced body perception [37]. By minimising expected free energy, these models dynamically adapt to perturbations in real time during physical human-robot interaction, offering robust solutions for surgical robotic simulators (e.g., SurRoL) and minimising the sim-to-real gap during continuous skill acquisition [37,39].

5.3 The Neural Meta Evaluator (NeME)

IL methods in Human-Robot Interaction and Collaboration (HRIC) have been evaluated using Average Success Rate (SR) or Dynamic Time Warping (DTW). However, SR requires time-consuming, resource-intensive deployment on physical robots and is highly susceptible to human variability. DTW often fails to capture the nuanced quality of robot motion, heavily penalising valid but slightly divergent trajectories. To provide a rigorous, reproducible evaluation pipeline, the Neural Meta Evaluator (NeME) frames policy assessment as a sequence-classification task based directly on robot joint trajectories. NeME operates as an offline meta-evaluator, efficiently processing generated trajectories without the constraints of human-in-the-loop deployment. Empirical evaluations demonstrate that optimal model weights selected via NeME's meta-F1 (mF1) scores perfectly align with the actual peak SR (e.g., precisely identifying the 8th-epoch peak where validation loss fails), thereby providing a statistically rigorous surrogate for physical deployment [40].

5.4 Comparative Analysis of Performance and Computational Overhead

The integration of Artificial Intelligence (AI) and robotic systems into healthcare ecosystems represents a fundamental paradigm shift from traditional medical models to predictive, personalised, preventive, participatory, and precision (P5) medicine [41]. Advanced machine learning (ML) frameworks, specifically Deep Reinforcement Learning (DRL) and Generative AI, are driving transformative breakthroughs in surgical precision, adaptive physical rehabilitation, and equitable diagnostic modelling [42–44]. The following tables summarise the performance and Computational Overhead. The first table shows the Comparative Evaluation of Sequence Modelling Architectures for Meta-Evaluation (NeME) Evaluation of behaviour classification performance given an input trajectory length (L = 32, equating to a 3.2-s window). The LSM architecture demonstrates superior representational power for robotic joint-state sequences compared to modern state-space models [40]. Table 6 summarises the trade-off between choosing various robotic learning paradigms and Table 7 shows the algorithmic efficacy, computational overhead, and Hardware deployment in different types of health care robots.

5.5 Advanced Algorithmic Interventions and Model Robustness

The clinical effectiveness of medical AI depends on high-quality data inputs and on the ability of models to perform across diverse real-world environments [45]. Variations in clinical hardware and procedures, such as disparate histological staining techniques across hospitals, often cause diagnostic models to underperform on out-of-distribution (OOD) data. Generative AI, specifically conditional diffusion models, directly addresses this by synthesising realistic medical imagery to compensate for underrepresented demographic subgroups and rare pathologies. Empirical evidence from the CAMELYON17 histopathology challenge demonstrated that training diffusion models on 455,954 labelled patches and 1.8 million unlabeled patches significantly minimised demographic fairness gaps and preserved high diagnostic accuracy under severe OOD conditions [42]. Deep Reinforcement Learning (DRL) in Dynamic Environments: DRL provides optimised, goal-oriented autonomy for managing unstructured clinical environments and complex biological

data. In object manipulation, DRL-driven Viewpoint Adjusting and Grasping Synergy (VAGS) strategies have achieved an 83.50% grasp success rate and a 95% scene-clearing rate in highly cluttered simulations. In targeted biotechnology applications, Hierarchical Deep Reinforcement Learning (HDRL) models have successfully processed massive 3D time-lapse image sets to navigate *C. elegans* embryogenesis, map modular cellular movement pathways, and identify novel therapeutic targets [43].

Table 6: Comparative evaluation of sequence modelling architectures.

Neural Architecture	Meta-Accuracy (%)	Meta-F1 Score (%)	Training Hyperparameters & Augmentation
LSTM	72.1 ± 1.3	66.6 ± 0.3	$h \in \{16, 64, 128\}$, Gaussian noise $\sigma \in \{0, 10^{-1}, 10^{-2}\}$
Transformer (Encoder-only)	68.7 ± 1.2	61.6 ± 3.1	AdamW optimiser, optimised for 30 epochs
xLSTM	59.1 ± 3.6	53.3 ± 3.7	Early stopping at 5 epochs (no validation loss improvement)
Mamba	53.6 ± 13.8	51.9 ± 12.7	Projection layer from $a_t \in \mathbb{R}^{24}$ to hidden size h
Mamba2	58.1 ± 5.4	54.0 ± 4.9	Averaged over three distinct weight initialisations

Table 7: Algorithmic efficacy, computational overhead, and hardware deployment.

Paradigm/ Algorithm	Target Application	Hardware Deployment	Computational Overhead Specifications	Baseline Comparison Metric
eXteroceptive Behaviour Generation (XBG)	HRIC autonomous interaction	ergoCub Humanoid, Intel Realsense D450	Training: 12 epochs, batch size 64, 2× NVIDIA A100 GPUs, Learning rate: 5e−4, 10 Hz inference.	XBG (RGB-D): Meta-F1: 71.3%, DTW: 2.18, SR: 70.0%, XBG-RGB: Meta-F1: 69.9%, DTW: 2.40, SR: 61.7% [40]
Adaptive NN + DMPs	Continuous complex manipulation	Baxter Robot	Calculates target trajectory velocity $\dot{v}$ and position dynamically using high-priority spring-damper mechanisms.	Substantial reduction in velocity attenuation at junction points compared to standard end-to-end DMP sequences [38]

(Continued)

Table 7 (continued)

Paradigm/ Algorithm	Target Application	Hardware Deployment	Computational Overhead Specifications	Baseline Comparison Metric
DTW + GMM-DMP	Minimally Invasive Surgery (MIS)	KUKA LWR4+, da Vinci Research Kit	Translates kinesthetic teaching into dynamically warped geometric models.	Circumvents Remote Centre of Motion (RCM) constraint failures present in pure imitation models [39]

5.6 Regulatory Hurdles and the Imperative for Dependable Deployment

True clinical integration is hindered not merely by technological limitations but by the rigorous demands of ethical governance, data security, and legal accountability [41,46]. Stringent Regulatory Compliance: The reliable deployment of connected health robots is tightly governed by frameworks such as GDPR, which mandates explicit patient consent and comprehensive encryption for cross-border transmission of medical data [47]. In the United States, the FDA had cleared 222 AI-based medical devices by 2020; however, algorithms capable of continuous post-market learning pose an acute regulatory challenge, necessitating the development of novel oversight mechanisms to ensure ongoing safety [46]. Responsibility and Liability Attribution: The deployment of highly autonomous surgical and mobile robots significantly complicates legal accountability in the event of adverse events [46]. Retrospective analyses of FDA data over a 14-year period emphasise the genuine physical risks associated with robotic interventions [46,48]. Algorithmic Opacity and Explainable AI (XAI): The inherent "black box" nature of deep neural networks obscures the logic driving clinical predictions, fundamentally undermining physician trust and patient safety. Interpretability is transitioning from an operational preference to a strict legal requirement under frameworks such as the EU Artificial Intelligence Act. For dependable deployment, developers must mandate XAI frameworks that transparently justify automated decisions to prevent automation bias and the entrenchment of existing health disparities [46,49]. Because medical AI lacks independent moral status, human operators and institutional stakeholders remain the primary duty-bearers; however, automated systems that execute high-risk manoeuvres in sub-second timeframes effectively preclude human intervention, creating unresolved legal ambiguity regarding liability [4].

6 Emerging Trends

6.1 Foundation Models for Robotics

The robotic field has definitively transitioned from fragmented, task-specific deep reinforcement learning (DRL) toward internet-scale Embodied Foundation Models. Driven by massive aggregation efforts such as the Open X-Embodiment (OXE) dataset, these architectures establish universal control priors that enable zero-shot transfer across diverse robot morphologies. State-of-the-art models like RDT-1B and Octo utilise diffusion-based policy modelling to handle highly multimodal visuomotor distributions. Rather than outputting a deterministic action, the diffusion foundation model learns to reverse a stochastic forward process. The training objective minimises a score-matching loss over action chunks: $L(\theta) = \mathbb{E}_{\tau\epsilon\sim N(0,I),t}[\left\|\epsilon - \epsilon_\theta\left(a_t^0, \tau, c\right)\right\|_2^2]$ where the generation is conditioned on language and visual observations. Quantitative Breakthroughs (2024–2025): The shift to diffusion foundation models yields unprecedented generalisation. Octo, trained on over 4 million trajectories across 22 distinct robotic platforms, achieves

highly robust cross-embodiment sim-to-real transfer. Similarly, RDT-1B (a 1.2B-parameter diffusion foundation model) demonstrates exceptional zero-shot generalisation in complex bimanual manipulation scenarios, resolving the sparse-reward bottleneck that traditionally paralysed DRL [2].

6.2 Large Language Models as Task Interfaces and Planners

Large Language Models (LLMs) are now utilised to bypass algorithmic abstraction barriers, acting as zero-shot semantic planners that translate human intent directly into logically structured sub-goals [50,51]. Seminal frameworks like SayCan and ProgPrompt formulate planning as a constrained probability maximisation problem. SayCan mathematically grounds LLM abstractions into physical affordances via: $\arg max_a P(a|instruction) \times P(success|a, state)$, multiplying the LLM's semantic prior by a learned visual affordance score [51]. ProgPrompt transforms situated environments into Pythonic APIs. The generation relies on a prompting function $f_{prompt}(s)$ that maps the state s into an import-style header (e.g., from actions import grab, open). The LLM recursively predicts the next program string, integrating assert conditions to provide real-time state feedback and precondition checking. Quantitative Evaluation: LLM-generated Pythonic execution heavily mitigates the "cascading error" problem seen in sequential robotic tasks. By explicitly grounding preconditions, LLM-guided planners can drastically improve success rates in long-horizon task generation without requiring any domain-specific policy re-training [52].

6.3 Language-Conditioned and Multimodal Policies

Vision-Language-Action (VLA) architectures fuse pre-trained vision encoders (e.g., SigLIP, DINOv2) and LLMs (e.g., LLaMA, Qwen) directly into action decoders. This end-to-end multimodal alignment achieves profound instruction-following capabilities [53]. Standard VLAs historically relied on left-to-right autoregressive decoding, treating continuous joint torques as discretised text tokens: $L = -\sum \log \pi_\theta(a_i|o_{1:t}, l)$ [12]. However, this severely bottlenecks inference speeds. The 2024 breakthrough OpenVLA-OFT replaces autoregression with Parallel Decoding and Action Chunking, optimising a continuous L1 regression loss directly on the output representations. Conversely, the Discrete Diffusion VLA applies flow-matching directly to tokenised action chunks via a transition matrix $Q_t e_{a_t,i} = (1 - \beta_t)\, e_{a_t,i} + \beta_t e_M$ [2,12]. Quantitative Breakthroughs: Addressing the severe latency constraints of embodied control, the OpenVLA-OFT+ variant processes continuous actions via parallel decoding, achieving a real-time action-generation throughput of 77.9 Hz (31.1 ms latency) on an NVIDIA A100 GPU, outperforming the original OpenVLA's sluggish 1.8 Hz baseline [12].

6.4 Sim-to-Real Transfer at Scale

The Reality Gap (G_{dym}) is being actively closed by highly parallelised, GPU-accelerated simulators (e.g., Isaac Sim, MuJoCo) executing massive Domain Randomisation (DR) and adversarial feature learning [2,6]. Rather than attempting to model physical reality perfectly, DR perturbs the simulated transition dynamics $T_{sim}(s_{t+1}|s_t, a_t, \mu)$ where physics parameters μ (mass, friction, damping) are sampled from a broad distribution P_μ. To ensure representation robustness, contrastive learning objectives enforce that the latent representation z_t remains invariant to rendering variations, mathematically binding the features to task-relevant physics rather than superficial textures [6]. Quantitative Evaluation: Modern massively parallelised RL algorithms can simulate thousands of environments simultaneously. In highly agile quadruped locomotion (e.g., ANYmal parkour), policies trained purely under heavy domain randomisation in simulation successfully execute highly dynamic blind obstacle traversal in the real world without any real-world fine-tuning [54].

6.5 *Safety, Verification, and Trustworthy Autonomy*

Deploying neural policies in critical environments requires migrating from empirical "success rates" to formal control-theoretic safety certificates. Control Barrier Functions (CBFs) and Lyapunov Certification provide this mathematical rigour. For a control-affine system $\dot{x} = f(x) + g(x)u$, a neural policy is strictly bounded by a continuously differentiable safe set defined by $B(x) \geq 0$. The learning controller ensures safety by solving a quadratic program that strictly enforces the CBF derivative condition: $\frac{\partial B}{\partial x}(f(x) + g(x)u) \geq -\gamma(B(x))$. Furthermore, deep networks parameterising the policy are regularised via spectral normalisation to strictly bound their Lipschitz constant, ensuring that bounded uncertainties in the physical dynamics translate directly to bounded, stable errors in the controller's time derivatives. Quantitative Evaluation: Architectures employing Lipschitz-bounded verification and CBFs mathematically eliminate hardware constraint violations during learning (achieving zero-collision exploration), which is critical when robotic hardware costs hundreds of thousands of dollars [8].

6.6 *Hybrid Learning Stack for Robotics*

Recognising the sample inefficiency of pure DRL and the latency of VLAs, hybrid learning stacks orchestrate a symbiotic pipeline: slow, high-level semantic foundation models orchestrating fast, low-level continuous dynamical systems (RL or classical MPC) [55]. Dual-system architectures decouple control frequency. A "System 2" LLM planner generates semantic spatial targets at 1~3 Hz, resolving the long-horizon sparse reward problem. A "System 1" low-level DDPG or PPO reinforcement learning agent operates at 50~500 Hz, processing real-time proprioceptive force feedback to minimise the residual error strictly within the localised semantic bounds established by System 2 [56]. Non-Obvious Gap: Current hybrid models lack bidirectional causal feedback. If the high-frequency RL agent encounters unmodeled object compliance (e.g., slipping), it cannot mathematically communicate the geometric reason for this failure back into the language embedding space of the LLM for re-routing. Future research must integrate visual-language failure evaluators (e.g., StepEval) to encode physical failures as text tokens, thereby creating a closed-loop, symbiotic pipeline [57]. Table 8 compares different model paradigms in terms of certain fields.

Table 8: Comparative technical analysis of multimodal ML paradigms (2024–2025).

Model Paradigm	Vision/Language Backbone	Action Decoding Framework	Hardware Deployment	Quantitative Efficacy/Inference Speed
OpenVLA-OFT+	SigLIP + DINOv2/ LLaMA-2 7B	Continuous L1 Regression, Parallel Decoding	NVIDIA A100 GPU (optimised via LoRA)	77.9 Hz throughput (31.1 ms latency); achieves 97.1% SR on complex LIBERO spatial targets [12,53].
(Pi-0)	PaliGemma (Gemma-2B)/SigLIP	Cross-embodiment Diffusion Flow-Matching	Large-scale multi-GPU clusters	Zero-shot adaptable across 22+ platforms; utilises diffusion matching to achieve high dexterity [2,53].

(Continued)

Table 8 (continued)

Model Paradigm	Vision/Language Backbone	Action Decoding Framework	Hardware Deployment	Quantitative Efficacy/Inference Speed
RDT-1B	OpenCLIP/Qwen	1.2B Parameter Diffusion Foundation Model	High-end GPU workstations	Excels at dynamic bimanual manipulation with near-perfect zero-shot transfer capabilities [2].
RT-2	PaLI-X/PaLM-E	Symbol-tuning Autoregressive Transformer	Cloud-tethered infrastructure	First true VLA to co-finetune internet VQA data with robotic data; 1.8 to 3 Hz (highly latency-bound) [2,53].
ProgPrompt	External (e.g., YOLOX)/GPT-4	API-driven Python code generator with asserts	Edge CPU/GPU (Queries Cloud LLM API)	Outperforms open-loop semantic planners by verifying preconditions and mitigating cascading failures [52].

7 Evaluation: From Benchmarks to Deployment Reality

The deployment of Deep Learning (DL) in robotic systems is fundamentally bottlenecked by the epistemological tension between the "black-box" nature of massive neural architectures and the strict deterministic requirements of physical robotics. DL models, especially foundation Vision-Language-Action (VLA) models, operate as highly non-linear, stochastic function approximators that map open-world, high-dimensional observations into latent manifolds. Because their exact decision boundaries are mathematically opaque, they are prone to unpredictable, potentially catastrophic extrapolation when applied to out-of-distribution (OOD) real-world physics. Conversely, robotics demands rigorous deterministic guarantees, such as formal Lyapunov stability and collision-free bounds, to prevent hardware destruction and ensure human safety [7,8]. To bridge this divide without sacrificing the cognitive depth of modern foundation models, this review proposes a unique conceptual framework: Prediction-Bounded Verified Deployment (PBVD). Rather than relying on naive, uncertified deployment or exhaustive hardware testing, PBVD mathematically fuses Prediction-Powered Inference (PPI) with granular sequence meta-evaluation. Under PBVD, an uncertified neural policy is first heavily evaluated in a large-scale simulation. Instead of trusting the biased simulation output, PBVD utilises a minimal set of paired physical trials to compute a mathematically rigorous "rectifier", This rectifier bounds the expected real-world safety and performance of the black-box policy via non-asymptotic confidence intervals. Consequently, the PBVD framework structurally guarantees that a policy's real-world failure probability is strictly quantified and constrained *before* it is granted continuous torque access to a physical machine [58].

7.1 Evaluation Blueprint: Formalising the Sim-to-Real Deployment Pipeline

Historically, robotic policies have been evaluated using a coarse, binary average Success Rate (SR) computed over a statistically insignificant number of physical trials (e.g., 20 to 30 rollouts) [40,58]. This approach completely fails to capture the complexity of continuous movement and lacks statistical guarantees. To formalise deployment, we must transition to rigorous mathematical frameworks that leverage imperfect simulators to bound real-world performance.

7.2 Prediction-Powered Inference (PPI) and the SureSim Benchmark

Because physical evaluation is prohibitively expensive, the SureSim framework formalises the use of Prediction-Powered Inference (PPI) to augment small-scale real tests with large-scale simulation. Given a real-to-sim mapping function $g: X \to X_{sim}$, the framework collects paired real and simulated evaluations, and additional purely simulated evaluations. The uniform PPI estimator corrects the simulation bias to compute the true mean: $\mu_{PPI} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f\left(\tilde{X}_i\right)\right) + \frac{1}{N+n} \sum_{i=1}^{n+N} f\left(\tilde{X}_i\right)$ [58]. The first term mathematically acts as a rectifier, compensating for the Reality Gap (the divergence between simulated dynamics and real dynamics). By applying the Waudby-Smith and Ramdas (WSR) algorithm to this estimator, SureSim establishes finite-sample valid confidence intervals. Quantitative Support: Empirical deployment of SureSim on physics-based manipulation benchmarks demonstrated that this framework reduces the required real-world hardware evaluation effort by 20%–25% while achieving tighter confidence intervals than classical real-only statistical bounds, reducing the interval width by 14.4% when scaling to 700 simulations [58].

7.3 Granular Subgoal Tracking and Neural Meta-Evaluation

Beyond sample efficiency, evaluating complex long-horizon behaviours requires abandoning the scalar pass/fail paradigm. The StepEval blueprint formalises task evaluation as a trajectory-level vector $y \in \{0,1\}^n$, where each element corresponds to a specific sub-task. Rather than relying on human annotation, this framework utilises Vision-Language Models (VLMs) as automated judges, mapping visual trajectories to the predicted success vector $\hat{y}$ [57]. However, VLMs cannot evaluate the continuous quality of joint kinematics. To address this, the Neural Meta Evaluator (NeME) frames trajectory assessment as an offline sequence classification problem. NeME processes a time window of joint trajectories using a neural sequence model (parameterised by) to predict behaviour $\hat{b} = E_{\Phi}(\tau^a)$. During inter-policy selection, standard validation loss fundamentally fails to identify optimal models. However, selecting policy weights using NeME's meta-F1 score (mF1) aligns perfectly with the epoch (Epoch 8) that achieves peak physical success rates on human-robot collaborative tasks. Furthermore, structural comparisons reveal that an LSTM-based NeME with a window length $L = 32$ achieves an mF1 score of 66.6 ± 0.3%, vastly outperforming state-space models like Mamba, which suffered representational collapse at 51.9 ± 12.7% mF1 on robotic joint sequences [40].

7.4 Core Algorithmic Engines: Mathematical Nuances and Deployment Reality

To understand deployment reality, we must describe the core machine learning paradigms entirely in prose, embedding their foundational mathematical mechanics to reveal how their optimisation strategies dictate real-world latency, hardware constraints, and sim-to-real transferability. The Proximal Policy Optimisation (PPO) algorithm is an on-policy, actor-critic framework designed to guarantee monotonic policy improvement by mathematically preventing destructively large gradient updates that cause catastrophic failure in physical robots [8,34]. It achieves this by updating the policy network using a clipped surrogate objective: $L^{CLIP}(\theta) = \mathbb{E}[\min(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A_t)]$, where $r_t(\theta)$ is the probability ratio between the new and old policies, and A_t is the advantage estimate. By clipping the ratio, PPO forces the policy to stay within a trusted region, yielding stable convergence [8]. While sample-inefficient compared

to off-policy methods, PPO's gradient stability makes it highly preferred for transferring robust locomotion capabilities to real quadrupedal hardware without relying on massive replay buffers. Conversely, the Soft Actor-Critic (SAC) algorithm is an off-policy method formulated within a maximum-entropy framework, specifically designed to maximise sample efficiency in real-world sparse-reward environments. The SAC algorithm optimises a stochastic policy by maximising both the expected cumulative reward and the policy's entropy, as defined by the objective $\pi^* = \arg max_\pi \mathbb{E}[\sum \gamma^t(\mathrm{r}(s_t, a_t) + \alpha H(\pi(.|s_t)))]$. The temperature parameter dictates the balance between exploiting the highest-value action and exploring diverse kinematic trajectories [8,34]. While its inherent stochasticity provides robustness to external physical perturbations, SAC's off-policy updates and replay buffer management introduce severe computational bottlenecks during gradient updates, often making it less efficient in terms of pure wall-clock training time than PPO on physical hardware.

Moving beyond standard reinforcement learning, modern generative manipulation is dominated by Diffusion Policies. These architectures model the action generation process not as a direct prediction but as learning to reverse a stochastic forward noising process applied to highly complex, multimodal continuous action trajectories. The network starts with pure Gaussian noise and iteratively denoises it into an action chunk. The training objective minimises a score-matching loss: $L(\theta) = \mathbb{E}_{\tau,\epsilon\sim N(0,I),t}[|| \epsilon - \epsilon_\theta (a_t^0, \tau, c)||]\binom{2}{2}]$ where the denoiser is conditioned on and fuses visual and language features. This framework resolves the sparse-reward bottleneck and naturally handles the multimodality of human demonstrations, making it highly effective for dexterous bimanual manipulation. Finally, Vision-Language-Action (VLA) architectures cast visual processing, language understanding, and physical action generation into a unified sequence modelling problem. In standard autoregressive VLAs, the policy is trained via behavioural cloning using a next-token prediction objective: $L = -\sum \log \pi_\theta(a_i|o_{1:t}, l)$, treating continuous actions as discrete text tokens. However, this left-to-right generation poses a severe latency bottleneck. To address this, the OpenVLA-OFT formulation discards discrete tokens in favour of continuous L1 Regression applied across parallel-decoded action chunks [2,12]. Taking a different hybrid approach, the Discrete Diffusion VLA applies flow-matching directly to tokenised action chunks via a discrete transition matrix $Q_t e_{a_t,i} = (1-\beta_t)\, e_{a_t,i} + \beta_t e_M$, allowing the transformer to adaptively unmask high-confidence discrete tokens in parallel [21].

7.5 Comparative Technical Analysis of ML Paradigms

To objectively evaluate the deployment readiness of these paradigms, the following table synthesises their execution metrics, computational overhead, and hardware configurations. Table 9 shows the lgorithmic Efficacy and Hardware Deployment of ML Paradigms. While significant strides have been made in scaling VLA architectures and in deriving real-to-sim statistical bounds, a critical, non-obvious gap remains in dynamic vocabulary alignment during OOD recovery. Current Discrete Diffusion VLAs and autoregressive models rely on fixed patch embeddings and static BPE tokenisation for continuous states. If a deployed policy encounters anomalous tissue compliance or severe sensor noise that falls outside its tokenised distribution, it suffers an immediate latent misalignment between the visual encoder and the LLM backbone. Existing systems cannot dynamically request continuous recalibration of their action vocabularies without complete offline retraining. Future deployment blueprints must integrate verifiable fallback mechanisms (such as the PBVD framework's rectifier) that trigger automated execution halts and explicit semantic re-prompting when the confidence intervals of the diffusion generation collapse in real time [2].

Table 9: ML paradigms comparison in terms of 4 aspects.

ML Paradigm	Action Decoding/Objective Strategy	Target Benchmark & Evaluation Metric	Computational/ Hardware Overhead	Quantitative Efficacy & Inference Speed
OpenVLA-OFT+	Continuous L1 Regression, Parallel Decoding	ALOHA Bimanual (Multi-stage evaluation)	NVIDIA A100 GPU (LoRA optimised)	77.9 Hz throughput (0.321 s latency); solves complex long-horizon bimanual alignments at 89.6% avg SR [12].
Diffusion Policy (Image)	Denoising score-matching over continuous chunks	ALOHA Bimanual (Multi-stage evaluation)	NVIDIA A100 GPU	267.4 Hz throughput (0.090 s latency); excellent dexterity but lacks zero-shot semantic generalisation [12].
Discrete Diffusion VLA	Masked-token discrete flow-matching	LIBERO Benchmark (Spatial, Object, Goal, Long)	4× NVIDIA A800 GPUs (Training)	14.53 Hz inference; 96.3% avg. SR on LIBERO (outperforms autoregressive OpenVLA by ~20%) [21].
NeME (Meta-Evaluator)	Offline trajectory sequence classification (LSTM)	HRIC/ergoCub joint sequences	Edge-deployable CPU/GPU inference	Identifies optimal weights that perfectly match physical robot performance; 66.6% mF1 score [40].
SureSim (PPI Real2Sim)	Uniform PPI Estimator	Pick-and-place generalizability	Massively parallel physics simulator	Decreases required real-world trials by >20%–25%; shrinks confidence interval bounds by 14.4% [58].

8 System Design Blueprint for LLM-Enabled Robotic Assistance

8.1 Bridging the Semantic-Kinematic Divide: The Interdisciplinary Rationale

The development of fully autonomous robotic assistants is fundamentally hindered by a dichotomous specialisation in artificial intelligence research. Specialists in Natural Language Processing (NLP) and Large Language Models (LLMs) focus on the semantic alignment of discrete tokens, yielding systems capable of open-world reasoning, deep common-sense logic, and hierarchical task decomposition [59]. Conversely, experts in Control Theory and Reinforcement Learning (RL) operate within the kinematic domain, focusing on continuous state spaces, high-frequency torque control, and formal stability guarantees necessary for safe physical interaction [20]. Addressing LLMs and RL within a unified manuscript is critical because neither paradigm can achieve reliable robotic autonomy in isolation. LLMs lack physical grounding and the ability to execute contact-rich, continuous control manoeuvres, while RL agents suffer from severe sample

inefficiency and an inability to reason over long-horizon, abstract tasks without explicit, heavily engineered reward functions [20,60]. By formulating a taxonomy that addresses both high-level semantic planning and low-level continuous control, this review bridges the gap between these distinct specialisations. We present a blueprint in which the LLM's abstract reasoning serves as a semantic manifold that strictly bounds the exploration space of the low-level RL controller, resulting in a cohesive framework for deployment in dynamic environments. To operationalise the integration of LLMs and RL, we propose a novel conceptual framework: Semantic-Kinematic Symbiosis (SKS). The SKS architecture structurally formalises robotic execution as a Hierarchical Partially Observable Markov Decision Process (H-POMDP). In this framework, the LLM acts as the high-level semantic planner, interpreting ambiguous human instructions into a sequence of intermediate goals. Rather than allowing the LLM to hallucinate physically impossible tasks, the SKS framework mathematically constrains the LLM using the RL policy's learned value functions, which act as representations of physical affordance. This integration is formally captured by the "SayCan" formulation. In prose, the probability of executing a successful robotic action is determined by the intersection of semantic utility and physical capability. The framework computes the optimal action by maximising the product of two probabilities: $\arg max_a P(a|instruction) \times P(success|a, state)$, where $P(a|instruction)$ is the language model predicting the semantic likelihood of a skill contributing to the high-level instruction, and $P(success|a, state)$ is the reinforcement learning value function providing the affordance, the probability that the robot can physically execute the skill from its current state [51,60].

8.2 High-Level Task Planning: LLMs as Semantic Oracles

The first stage of the SKS blueprint utilises autoregressive LLMs to translate natural language into discrete, executable task sequences. The fundamental operation of the LLM planner relies on maximising the conditional probability of a sequence of programming tokens y given an input prompt x, formalised as $\arg max_y P(y: x; \theta)$. To ensure the LLM does not generate actions beyond the robot's hardware limits, modern architectures rely on programmatic prompt structures. Frameworks such as ProgPrompt inject Pythonic API specifications directly into the LLM's context window. Instead of outputting free-form text, the LLM generates Python code that includes assert statements to continuously monitor the robot's state feedback during execution [6,52]. Furthermore, the Interactive Predicate Learning (InterPreT) framework uses GPT-4 to generate complex semantic predicates as Python functions, which are iteratively refined with natural-language feedback from human operators, thereby seamlessly translating raw observation states into logical preconditions for task planners [50]. Utilising structured programming language prompts, ProgPrompt achieved an execution success rate of 0.87 on complex VirtualHome tasks, substantially reducing the computational overhead of plan generation while demonstrating near-linear scalability as environments expand, and drastically outperforming classical A* planners [52].

8.3 Low-Level Continuous Control: Robust Reinforcement Learning

Once the semantic sub-goal is generated, the system delegates execution to a high-frequency RL policy capable of managing complex, contact-rich physical dynamics. To ensure that the robot learns safely without catastrophic hardware failure, the Proximal Policy Optimisation (PPO) algorithm is frequently deployed. In prose, PPO stabilises learning by preventing destructively large policy updates. It achieves this by updating the policy network using a clipped surrogate objective: $L^{CLIP}(\theta) = \mathbb{E}[\min(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A_t)]$, where $r_t(\theta)$ is the probability ratio between the new and old policies, and is the advantage estimate. This mathematical clipping physically restricts the policy's step size, thereby guaranteeing stable convergence during complex manoeuvres [20]. For highly dexterous applications, the state-of-the-art RL-100 framework embeds RL directly into a continuous diffusion visuomotor policy. It utilises a consistency

model distillation to compress multi-step diffusion into a single-step action generator, effectively bridging offline imitation learning with online, high-frequency RL fine-tuning. The RL-100 framework represents a definitive benchmark in deployment readiness. It achieved a 100% success rate across 1000 evaluated real-world episodes on diverse tasks (such as dynamic pushing and juicing) and demonstrated formidable robustness, maintaining a ~96% success rate even under aggressive human perturbations. Crucially, a juicing robot driven by this framework operated autonomously for seven continuous hours in a public shopping mall without a single failure [29].

8.4 Integration Taxonomy: The Plan-Seq-Learn (PSL) Paradigm

To unify these domains operationally, the Plan-Seq-Learn (PSL) paradigm offers a highly scalable, modular pipeline. In PSL, the task is strictly decoupled: the LLM predicts a high-level sequence of target regions (Plan), an off-the-shelf vision-based motion planner moves the robotic arm into proximity of the target (Seq), and a localised RL policy is activated solely to manage the final centimetres of contact-rich manipulation (Learn). This approach explicitly isolates the sample inefficiency of RL to the final execution phase, relying on LLMs and classical motion planning to bypass the long-horizon sparse-reward problem. By confining RL exploration to localised regions specified by the LLM, the PSL architecture achieves a staggering 96.0% success rate on 10-stage, contact-rich operations such as "NutAssembly" directly from raw visual inputs, thereby comprehensively outperforming purely end-to-end systems that suffer from cascading sequence errors [20]. In related hybrid approaches that combine LLMs and RL for Franka Emika Panda manipulation, task completion times were reduced by 33.5% (from 18.5 to 12.3 s), while task adaptability increased by 36.4% compared to RL-only baselines [55].

8.5 Healthcare Focus Theme: Deployment in Clinical and Assistive Environments

The SKS blueprint is particularly transformative for the healthcare sector, where operations demand both empathetic human understanding and zero-tolerance kinematic precision. Socially Assistive Robots (SARs) in Eldercare: SARs deployed to support elderly cognitive function help mitigate global nursing shortages and loneliness. Integrating conversational agents powered by LLMs (e.g., GPT-3.5 on the Social Robot Mini) enables robots to contextualise nuanced, bidirectional emotional cues and patient histories into highly personalised care responses, rather than relying on rigid, pre-programmed dialogue trees [61]. Frameworks like RobotIQ integrate these LLM-driven interactions directly into Robot Operating System (ROS) libraries, translating an elderly patient's spoken request (e.g., "I am thirsty") into precise localisation, navigation, and object-fetching APIs executed by low-level RL controllers [62]. Surgical Robotics: In precision environments, the RoboNurse-VLA framework functions as an automated scrub nurse. Processing voice prompts and dynamic visual scenes in real-time, the system executes surgical instrument handovers, adapting autonomously to unseen tools and rapidly changing operating room conditions [2].

8.6 Identification of Non-Obvious Gaps

While the SKS blueprint effectively integrates semantic intent with kinematic execution, a critical, non-obvious gap remains: the lack of bidirectional, continuous causal feedback. Current hybrid frameworks are predominantly top-down; the LLM instructs the RL policy. However, if the high-frequency RL agent encounters unmodeled physical compliance, such as a surgical tool slipping or unexpected patient resistance during rehabilitation, it currently cannot mathematically encode this continuous physical failure back into the LLM's discrete textual embedding space [2,53]. Future architectures must develop dynamic, tactile-to-language vocabulary alignment algorithms. By utilising visual-language evaluators (e.g., StepEval) to classify sub-goal kinematic failures into descriptive language tokens [57], the system could establish a closed-loop

symbiotic pipeline, allowing the LLM to dynamically re-route its semantic planning graph in response to localised physical constraints.

8.7 Integration of Frameworks

The individual frameworks are integrated into a cohesive functional stack as follows:

- Cognitive Layer (SKS/CSE): The Semantic-Kinematic Symbiosis (SKS) framework acts as the 'System 2' reasoning engine, which is structurally governed by Certified-Semantic Embodiment (CSE) to filter semantic outputs through physical affordance bounds.
- Stability Layer (LBSE): The Lyapunov-Bounded Semantic Execution (LBSE) framework provides the 'System 1' fast-frequency control, translating the high-level semantic waypoints into deterministic, safety-guaranteed motor commands.
- Feedback & Verification Layer (PBVD/BCKG): The Prediction-Bounded Verified Deployment (PBVD) framework quantifies the reality gap during execution. Crucially, the BCKG framework provides the bidirectional causal link, allowing low-level kinematic failures to be re-encoded as text tokens to dynamically update the high-level SKS planning graph.

This bidirectional flow ensures that the system does not just operate top-down, but functions as a transparent, reproducible, and safety-critical closed loop. Table 10 shows our revised version, we demonstrate the "Closed-Loop" nature to the reviewer.

Table 10: Framework taxonomy.

Architecture Layer	Framework	Primary Function	Loop Position
	Reasoning	SKS	Semantic goal decomposition
Certification	CSE	Safety-filtering of goals	Feed-forward Constraint
Execution	LBSE	Lyapunov-stable torque control	Low-level (Action)
Verification	PBVD	Real-time performance rectifier	Monitoring
Governance	BCKG	Causal failure feedback	Feedback (Loop Closure)

9 Open Challenges and Research Directions

The transition of robotic policies from simulated environments or constrained datasets to dynamic, real-world deployment is bottlenecked by severe algorithmic and mathematical limitations. To address the complexities of these challenges, it is necessary to move beyond generalised descriptions and to critically dissect the mathematical formulations and architectural variations that define the field's current limitations.

9.1 The Reality Gap and Sim-to-Real Transfer Discrepancies

The "Reality Gap" fundamentally stems from the inability of any simulator to perfectly model real-world physics, chaotic non-linearities, and sensor noise. Mathematically, this is framed as a discrepancy between a simulated Partially Observable Markov Decision Process (POMDP) M_s and the real-world POMDP M_r. The reality gap comprises distinct sub-gaps, notably the dynamics gap, defined as the expected divergence between transition models:

$G_{dyn}(M_s, M_r) = \mathbb{E}_{(s,a)\sim M_r}[D(T_{sim}(\cdot|s,a) \,\|\, T_{real}(\cdot|s,a))]$. Because policies trained often exploit these inaccuracies to maximise rewards, evaluating the performance gap $G_{perf}(M_s, M_r, \pi) = |J_{M_s}(\pi) - J_{M_r}(\pi)|$ requires specialised statistical frameworks [6]. To construct finite-sample valid confidence intervals without exhausting physical hardware, modern evaluation architectures employ Prediction-Powered Inference (PPI).

Frameworks such as SureSim construct a paired dataset D_{paried} mapping identical initial conditions across real and simulated trials. This allows the architecture to compute a rectifier variance, where high variance indicates low simulation-to-reality correlation, mathematically bounding the exact limits of the simulation's predictive utility [58]. To algorithmically mitigate G_{dyn}, residual learning architectures diverge from standard Domain Randomisation by parameterising the dynamics as a composite function. Rather than assuming the simulator's analytical model is complete, residual networks learn a corrective function f_{res} over the simulator's predicted state transitions such that $T_{real} \approx T_{sim} + f_{res}$. This structural separation allows the neural network to absorb unmodeled compliance and complex aerodynamic forces without requiring end-to-end retraining of the foundational physics engine [6].

9.2 Safety-Critical Machine Learning and State Constraints

In physical deployments, exploration and control policies cannot violate hard mechanical or environmental limits. Safe Reinforcement Learning reformulates the standard MDP into a Constrained Markov Decision Process (CMDP) defined by the tuple $(S, A, R, P, \gamma, \mu, C)$. Here, policies must maximise the standard expected return while strictly bounding the expected discounted constraint cost $J_c^{\pi} \leq d$, where represents thresholded safety functions [8,63]. Architecturally, enforcing these constraints relies on two distinct paradigms: Control Barrier Functions (CBFs), Lyapunov Certification, Variable Impedance and Safety Critics. In control-affine systems, safety is certified by maintaining the forward invariance of a robust safe set X_{safe}. If the true system dynamics $f(x)$ are partially unknown, the time derivative of the CBF relies on an approximation $\hat{f}(x)$. The architectural nuance lies in bounding the Lipschitz constant of the deep neural network parameterising the policy, often enforced via spectral normalisation. This ensures that the mapping of bounded uncertainty in the dynamics directly translates into bounded uncertainty in the CBF's time derivative, thereby mathematically guaranteeing safety [8]. To proactively avoid risky states prior to actuation, frameworks like SRL-VIC decouple task optimisation from safety through a dedicated Safety Critic network. The critic evaluates a recursive risk function $Q_{risk}^{\pi}(s_t, a_t) = c_t + (1 - c_t)\gamma_{risk}\mathbb{E}[Q_{risk}^{\pi}(s_{t+1}, a_{t+1})]$, trained via Mean Squared Error, to estimate the probability of constraint violation. If Q_{risk}^{π} exceeds a strict threshold, the architecture bypasses the primary actor network and queries a pre-trained recovery policy (π_{rec}) to sample a mathematically safe projection [63].

9.3 Sparse Reward Optimisation and Sample Inefficiency

For tasks demanding complex sequencing (e.g., long-horizon manipulation), defining a continuous reward gradient is highly susceptible to reward hacking. Consequently, the environment is often modelled with a sparse, binary reward structure (e.g., at the target, otherwise) [64]. This creates a severe sample inefficiency challenge, as standard Temporal Difference (TD) updates, such as $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$, fail to propagate useful gradients when R_{t+1} remains uniformly −1 [65]. To artificially synthesise dense gradients, architectures integrate Hindsight Experience Replay (HER) [36,65]. HER alters the replay buffer dynamics by modifying the tuple $(s_t, a_t, r_t, s_{t+1}, g)$ after a failed rollout. It replaces the unachieved original goal g with an achieved posterior state $\dot{g}$, forcing the Bellman equation to process a synthetic positive reward [18,36]. Advanced frameworks, such as Deep Value-and-Predictive-Model Control (DVPMC), integrate HER into a Model-Based RL (MBRL) pipeline [36]. DVPMC approximates the value function alongside a learned transition model and utilises sampling-based cross-entropy methods for action selection, thereby reducing the prediction horizon and circumventing the massive sample complexity inherent in model-free off-policy algorithms such as SAC or DDPG [36,65].

9.4 Distributional Shift and Epistemic Uncertainty

Both offline RL and real-time perception models suffer catastrophic performance degradation when evaluating Out-of-Distribution (OOD) states, an issue rooted in epistemic uncertainty. When training purely from static datasets, high-capacity function approximators systematically overestimate the Q-values of Out-of-Distribution actions [30]. To mitigate this distributional shift, algorithms either directly constrain the learned policy distribution to match the behaviour policy's distribution or deploy Conservative Q-Learning (CQL) to strictly minimise a lower bound on the value function [66]. Furthermore, when goal-conditioned architectures encounter OOD goals, algorithms inject noise perturbation with probability over the goal states, explicitly penalising the corresponding values via negative Temporal Difference errors to suppress exploratory divergence [35]. In unstructured environments, perception architectures must mathematically distinguish between aleatoric uncertainty (inherent sensory noise) and epistemic uncertainty (OOD inputs lacking training support). Using evidential deep learning, the network parameterises a Dirichlet distribution rather than standard categorical logits. A normalising flow network tracks the latent density of the features $p_\lambda(z_0)$. The epistemic uncertainty is then thresholded using a confidence score formulation: $g(z_0) = \frac{p_\lambda(z_0) - p_{min}}{p_{max} - p_{min}}$. If $g(z_0)$ falls below a predefined percentile threshold (e.g., the k-th percentile of the training distribution), the feature is mathematically flagged as OOD, and the downstream model-predictive controller applies auxiliary cost penalties to prevent the robot from navigating into untrusted regions [9].

10 Conclusion

The integration of Artificial Intelligence into robotic systems has catalysed a profound paradigm shift, transitioning the field from rigid, pre-programmed automation to highly dynamic, open-world autonomy driven by Vision-Language-Action (VLA) models and Deep Reinforcement Learning (DRL). This critical review has systematically synthesised the algorithms, architectures, and evaluation frameworks required to deploy these autonomous systems. However, as robotics permeates high-stakes, unstructured environments, particularly within P5 (predictive, personalised, preventive, participatory, and precision) medicine, the epistemological tension between the stochastic, "black-box" nature of deep neural networks and the deterministic, safety-critical requirements of physical execution remains the primary bottleneck to real-world deployment. To resolve this tension and guide the next decade of robotic research, we outline a strategic research agenda. Furthermore, we propose a novel, unifying conceptual framework that integrates cognitive reasoning, kinematic execution, and stringent regulatory compliance. Throughout this review, a persistent, non-obvious gap has been identified across state-of-the-art hybrid models (such as Plan-Seq-Learn or OpenVLA): the distinct lack of bidirectional causal feedback. Currently, high-level Large Language Models (LLMs) issue top-down semantic commands, but if a low-level RL controller encounters an unmodeled physical constraint (e.g., anomalous tissue compliance during surgery or a slipping object), it cannot mathematically translate that continuous physical failure back into the discrete textual embedding space of the LLM to trigger dynamic re-routing. We propose the Bidirectional Causal-Kinematic Governance (BCKG) architecture as a definitive roadmap for the future. In the BCKG framework, the low-level Lyapunov-certified execution manifold is equipped with a Neuro-Symbolic Failure Encoder. When a physical constraint is violated, this encoder translates the kinematic discrepancy $e(t) > \epsilon$ (e.g., assert Object_Slip(Target)) into a formalised logic predicate (e.g., assert Object_Slip(Target)). This predicate is fed back into the LLM's context window as an active prompt, forcing the language model to reason causally about the failure and generate a revised topological plan. Crucially, the BCKG framework wraps this entire bidirectional loop in an immutable, cryptographic data logger, creating a transparent, reproducible audit trail of every semantic-to-kinematic decision, a strict necessity for adhering to the General Data Protection Regulation (GDPR) and the European Artificial Intelligence Act.

10.1 Causal Reinforcement Learning and Neuro-Symbolic Integration

Future architectures must move beyond purely correlational deep learning by integrating Causal Reinforcement Learning and Neuro-Symbolic approaches. Relying strictly on model-free DRL is profoundly sample-inefficient and opaque. By incorporating bidirectional dynamics models, which simultaneously perform forward and inverse predictions in a latent space, algorithms can self-supervise the denoising of representations, substantially alleviating model bias and prediction errors. Furthermore, integrating structured symbolic logic with neural networks provides exact interpretability. Integrating advanced transfer learning with foundational models (such as the YOLO + SAC pipeline) has already demonstrated the capacity to drastically reduce robotic training times, achieving convergence in 6443 s compared to 15.9 times longer without transfer techniques. Expanding these causal pipelines will be strictly necessary to achieve human-level, few-shot problem-solving capabilities.

10.2 Continual Lifelong Learning and Dynamic Modality Expansion

Robots deployed in dynamic clinical environments cannot remain frozen after offline training; they must continuously adapt without suffering from *catastrophic forgetting*. Future research must scale Hierarchical Lifelong Reinforcement Learning (HLifeRL) frameworks. Current implementations of HLifeRL utilising dynamic option libraries have demonstrated profound stability; when expanding a robotic skill library from four complex locomotion tasks to five, the master policy retained over 95% of its original performance metrics, effectively resisting the parameter interference that destroys standard monolithic networks. Concurrently, the input space of VLA models must expand beyond vision and text. To execute high-precision medical tasks, future foundation models must ingest multi-modal high-frequency streams, specifically integrating tactile force/torque sensors, auditory feedback, and depth data, which are currently severely underutilised in end-to-end architectures.

10.3 Dependable Clinical Deployment and Ethical Governance

The transformative impact of robotic automation in healthcare, ranging from Da Vinci surgical systems providing 3D-HD millimetre-level precision to Socially Assistive Robots (SARs) mitigating nursing workforce shortages, is fundamentally dictated by regulatory hurdles rather than pure algorithmic capability. Systems governed by the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) demand comprehensive, verifiable safety and performance testing. Because learning algorithms inherently evolve over time, existing regulatory frameworks struggle to accommodate continuous online updates, severely decelerating clinical rollout. Future research must develop standardised, automated Verification and Validation (V&V) benchmarks specifically designed for continuous-learning medical devices. From an ethical and legal standpoint, current medical AI and robotic systems lack independent moral status; humans and institutional stakeholders remain the absolute duty bearers. Future deployments must establish clear legal attribution frameworks, ensuring that when an autonomous robotic assistant executes a flawed manoeuvre, the liability can be transparently traced through the system's runtime monitors and algorithmic decision logs. Finally, the evaluation of healthcare robotics must move away from short-term laboratory simulations. Future studies demand extensive, longitudinal field trials within authentic clinical settings to rigorously evaluate the cost-effectiveness, long-term patient engagement, and human-robot trust transfer necessary to validate these intelligent physical robots as transformative, reliable medical tools.

Acknowledgement: Not applicable.

Funding Statement: The authors received no funding for this work.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualisation and supervision, Ahmed Ismail Ebada; methodology and data curation, Yasmeen Abu-Seif and Hrushikesh Pardeshi; investigation, Yasmeen Abu-Seif; writing original draft preparation, Ahmed Ismail Ebada, Yasmeen Abu-Seif and Hrushikesh Pardeshi; writing review and editing, Yasmeen Abu-Seif and Nesma El-Sayed. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All data generated or analysed during this study are included in this published article. Any additional data supporting the findings of this study are available from the corresponding authors upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

ML	Machine Learning
RL	Reinforcement Learning
IL	Imitation Learning
SSL	Self-Supervised Learning
UL	Unsupervised Learning
DRL	Deep Reinforcement Learning
POMDP	Partially Observable Markov Decision Process
CVaR	Conditional Value at Risk
TDP	Thermal Design Heat
FLOPs	Floating Point Operations
RSSM	Recurrent State-Space Model
MSE	Mean Squared Error
MPC	Model Predictive Control
LLM	Large Language Model
VLA	Vision-Language-Action Model
VLM	Vision-Language Model
HRI	Human-Robot Interaction
MPC	Model Predictive Control
SAC	Soft Actor-Critic
PPO	Proximal Policy Optimisation
OOD	Out-of-Distribution

References

1. Soori M, Arezoo B, Dastres R. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. Cogn Robot. 2023;3:54–70. doi:10.1016/j.cogr.2023.04.001.
2. Din MU, Akram W, Saoud LS, Rosell J, Hussain I. Vision language action models in robotic manipulation: a systematic review. arXiv:2507.10672. 2025. doi:10.48550/arxiv.2507.10672.
3. Alqobali R, Alnasser R, Rashidi A, Alshmrani M, Alhmiedat T. A real-time semantic map production system for indoor robot navigation. Sensors. 2024;24(20):6691. doi:10.3390/s24206691.
4. Liu Q, Liu Z, Xiong B, Xu W, Liu Y. Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function. Adv Eng Inform. 2021;49(12):101360. doi:10.1016/j.aei.2021.101360.
5. Salvato E, Fenu G, Medvet E, Pellegrino FA. Crossing the reality gap: a survey on sim-to-real transferability of robot controllers in reinforcement learning. IEEE Access. 2021;9:153171–87. doi:10.1109/ACCESS.2021.3126658.

6. Aljalbout E, Xing J, Romero A, Akinola I, Garrett CR, Heiden E, et al. The reality gap in robotics: challenges, solutions, and best practices. Annu Rev Control Robot Auton Syst. 2026;9(1):403–32. doi:10.1146/annurev-control-031924-100130.
7. Matsuno K, Cheah CC. Lyapunov-based deep learning control for robots with unknown Jacobian. arXiv:2509.04984. 2025.
8. Brunke L, Greeff M, Hall AW, Yuan Z, Zhou S, Panerati J, et al. Safe learning in robotics: from learning-based control to safe reinforcement learning. Annu Rev Control Robot Auton Syst. 2022;5(1):411–44. doi:10.1146/annurev-control-042920-020211.
9. Cai X, Ancha S, Sharma L, Osteen PR, Bucher B, Phillips S, et al. EVORA: deep evidential traversability learning for risk-aware off-road autonomy. IEEE Trans Robot. 2024;40:3756–77. doi:10.1109/tro.2024.3431828.
10. Shojaeinasab A, Jalayer M, Baniasadi A, Najjaran H. Unveiling the black box: a unified XAI framework for signal-based deep learning models. Machines. 2024;12(2):121. doi:10.3390/machines12020121.
11. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. Dinov2: learning robust visual features without supervision. arXiv:2304.07193. 2025. doi:10.48550/arxiv.2304.07193.
12. Kim MJ, Finn C, Liang P. Fine-tuning vision-language-action models: optimizing speed and success. arXiv:2502.19645. 2025. doi:10.48550/arxiv.2502.19645.
13. Shanks S, Embley-Riches J, Liu J, Delfaki AM, Ciliberto C, DreamerNav K D. Learning-based autonomous navigation in dynamic indoor environments using world models. Front Robot AI. 2025;12:1655171. doi:10.3389/frobt.2025.1655171.
14. Kontolati K, Goswami S, Em Karniadakis G, Shields MD. Learning nonlinear operators in latent spaces for real-time predictions of complex dynamics in physical systems. Nat Commun. 2024;15(1):5101. doi:10.1038/s41467-024-49411-w.
15. El-Hussieny H. Real-time deep learning-based model predictive control of a 3-DOF biped robot leg. Sci Rep. 2024;14(1):16243. doi:10.1038/s41598-024-66104-y.
16. Zhou C, Huang B, Fränti P. A review of motion planning algorithms for intelligent robots. J Intell Manuf. 2022;33(2):387–424. doi:10.1007/s10845-021-01867-z.
17. Waga A, Benhlima S, Bekri A, Abdouni J, Saber FZ. A survey on autonomous navigation for mobile robots: from traditional techniques to deep learning and large language models. J King Saud Univ Comput Inf Sci. 2025;37(7):198. doi:10.1007/s44443-025-00216-x.
18. Wang Q, Sanchez FR, McCarthy R, Bulens DC, McGuinness K, O'Connor N, et al. Dexterous robotic manipulation using deep reinforcement learning and knowledge transfer for complex sparse reward-based tasks. Expert Syst. 2023;40(6):e13205. doi:10.1111/exsy.13205.
19. Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, et al. RT-1: robotics transformer for real-world control at scale. arXiv:2212.06817. 2022.
20. Dalal M, Chiruvolu T, Chaplot D, Salakhutdinov R. Plan-Seq-Learn: language model guided RL for solving long horizon robotics tasks. arXiv:2405.01534. 2024. doi:10.48550/arxiv.2405.01534.
21. Liang Z, Li Y, Yang T, Wu C, Mao S, Nian T, et al. Discrete diffusion VLA: bringing discrete diffusion to action decoding in vision-language-action policies. arXiv:2508.20072. 2025. doi:10.48550/arxiv.2508.20072.
22. Hong Q, Dong H, Deng W, Ping Y. Education robot object detection with a brain-inspired approach integrating Faster R-CNN, YOLOv3, and semi-supervised learning. Front Neurorobot. 2024;17:1338104. doi:10.3389/fnbot.2023.1338104.
23. Alharthi R, Noreen I, Khan A, Aljrees T, Riaz Z, Innab N. Novel deep reinforcement learning based collision avoidance approach for path planning of robots in unknown environment. PLoS One. 2025;20(1):e0312559. doi:10.1371/journal.pone.0312559.
24. Li X, Shang J, Das S, Ryoo M. Does self-supervised learning really improve reinforcement learning from pixels? In: 36th Conference on Neural Information Processing Systems (NeurIPS 2022); 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 30865–81. doi:10.52202/068431-2238.
25. Han D, Mulyana B, Stankovic V, Cheng S. A survey on deep reinforcement learning algorithms for robotic manipulation. Sensors. 2023;23(7):3762. doi:10.3390/s23073762.

26. Ali Shahid A, Piga D, Braghin F, Roveda L. Continuous control actions learning and adaptation for robotic manipulation through reinforcement learning. Auton Rob. 2022;46(3):483–98. doi:10.1007/s10514-022-10034-z.
27. Orr J, Dutta A. Multi-agent deep reinforcement learning for multi-robot applications: a survey. Sensors. 2023;23(7):3625. doi:10.3390/s23073625.
28. Kasaura K, Miura S, Kozuno T, Yonetani R, Hoshino K, Hosoe Y. Benchmarking actor-critic deep reinforcement learning algorithms for robotics control with action constraints. IEEE Robot Autom Lett. 2023;8(8):4449–56. doi:10.1109/LRA.2023.3284378.
29. Lei K, Li H, Yu D, Wei Z, Guo L, Jiang Z, et al. RL-100: performant robotic manipulation with real-world reinforcement learning. arXiv:2510.14830. 2025. doi:10.48550/arxiv.2510.14830.
30. Figueiredo Prudencio R, Maximo MROA, Colombini EL. A survey on offline reinforcement learning: taxonomy, review, and open problems. IEEE Trans Neural Netw Learning Syst. 2024;35(8):10237–57. doi:10.1109/tnnls.2023.3250269.
31. Gürtler N, Blaes S, Kolev P, Widmaier F, Wüthrich M, Bauer S, et al. Benchmarking offline reinforcement learning on real-robot hardware. arXiv:2307.15690. 2023. doi:10.48550/arxiv.2307.15690.
32. Janner M, Li Q, Levine S. Offline reinforcement learning as one big sequence modeling problem. arXiv:2106.02039. 2021. doi:10.48550/arxiv.2106.02039.
33. Ding F, Zhu F. HLifeRL: a hierarchical lifelong reinforcement learning framework. J King Saud Univ Comput Inf Sci. 2022;34(7):4312–21. doi:10.1016/j.jksuci.2022.05.001.
34. Morales EF, Murrieta-Cid R, Becerra I, Esquivel-Basaldua MA. A survey on deep learning and deep reinforcement learning in robotics with a tutorial on deep reinforcement learning. Intell Serv Robot. 2021;14(5):773–805. doi:10.1007/s11370-021-00398-z.
35. Li J, Tang C, Tomizuka M, Zhan W. Hierarchical planning through goal-conditioned offline reinforcement learning. IEEE Robot Autom Lett. 2022;7(4):10216–23. doi:10.1109/lra.2022.3190100.
36. Antonyshyn L, Givigi S. Deep model-based reinforcement learning for predictive control of robotic systems with dense and sparse rewards. J Intell Rob Syst. 2024;110(3):100. doi:10.1007/s10846-024-02118-y.
37. Taniguchi T, Murata S, Suzuki M, Ognibene D, Lanillos P, Ugur E, et al. World models and predictive coding for cognitive and developmental robotics: frontiers and challenges. Adv Robot. 2023;37(13):780–806. doi:10.1080/01691864.2023.2225232.
38. Kong LH, He W, Chen WS, Zhang H, Wang YN. Dynamic movement primitives based robot skills learning. Mach Intell Res. 2023;20(3):396–407. doi:10.1007/s11633-022-1346-z.
39. Chen H, Li S, Fan J, Duan A, Yang C, Navarro-Alarcon D, et al. Human-in-the-loop robot learning for smart manufacturing: a human-centric perspective. IEEE Trans Automat Sci Eng. 2025;22:11062–86. doi:10.1109/tase.2025.3528051.
40. Tiezzi M, Apicella T, Cardenas-Perez C, Fregonese G, Dafarra S, Morerio P, et al. Learning to evaluate autonomous behaviour in human-robot interaction. arXiv:2507.06404. 2025. doi:10.48550/arxiv.2507.06404.
41. Denecke K, Baudoin CR. A review of artificial intelligence and robotics in transformed health ecosystems. Front Med. 2022;9:795957. doi:10.3389/fmed.2022.795957.
42. Ktena I, Wiles O, Albuquerque I, Rebuffi SA, Tanno R, Roy AG, et al. Generative models improve fairness of medical classifiers under distribution shifts. Nat Med. 2024;30(4):1166–73. doi:10.1038/s41591-024-02838-6.
43. Al-Hamadani MNA, Fadhel MA, Alzubaidi L, Balazs H. Reinforcement learning algorithms and applications in healthcare and robotics: a comprehensive and systematic review. Sensors. 2024;24(8):2461. doi:10.3390/s24082461.
44. Ali R, Cui H. Unleashing the potential of AI in modern healthcare: machine learning algorithms and intelligent medical robots. Res Intell Manuf Assem. 2024;3(1):100–8. doi:10.25082/rima.2024.01.002.
45. Masala GL, Giorgi I. Artificial intelligence and assistive robotics in healthcare services: applications in silver care. Int J Environ Res Public Health. 2025;22(5):781. doi:10.3390/ijerph22050781.
46. Zhang J, Zhang ZM. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak. 2023;23(1):7. doi:10.1186/s12911-023-02103-9.
47. Lee YH, Hsu FY, Lien AS. Health care professionals' perspectives of socially assistive robots in health care settings: systematic review. J Med Internet Res. 2025;27:e79634. doi:10.2196/79634.

48. Silvera-Tawil D. Robotics in healthcare: a survey. SN Comput Sci. 2024;5(1):189. doi:10.1007/s42979-023-02551-0.
49. Nicora G, Pe S, Santangelo G, Billeci L, Aprile IG, Germanotta M, et al. Systematic review of AI/ML applications in multi-domain robotic rehabilitation: trends, gaps, and future directions. J NeuroEng Rehabil. 2025;22(1):79. doi:10.1186/s12984-025-01605-z.
50. Han M, Zhu Y, Zhu SC, Wu Y, Zhu Y. InterPreT: interactive predicate learning from language feedback for generalizable task planning. arXiv:2405.19758. 2024. doi:10.48550/arxiv.2405.19758.
51. Kawaharazuka K, Matsushima T, Gambardella A, Guo J, Paxton C, Zeng A. Real-world robot applications of foundation models: a review. Adv Robot. 2024;38(18):1232–54. doi:10.1080/01691864.2024.2408593.
52. Singh I, Blukis V, Mousavian A, Goyal A, Xu D, Tremblay J, et al. ProgPrompt: program generation for situated robot task planning using large language models. Auton Robot. 2023;47(8):999–1012. doi:10.1007/s10514-023-10135-3.
53. 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;13:162467–504. doi:10.1109/ACCESS.2025.3609980.
54. Hoeller D, Rudin N, Sako D, Hutter M. ANYmal parkour: learning agile navigation for quadrupedal robots. Sci Robot. 2024;9(88):eadi7566. doi:10.1126/scirobotics.adi7566.
55. Saad M, Hussain S, Suhaib M. Hybrid framework for robotic manipulation: integrating reinforcement learning and large language models. arXiv:2603.30022. 2026. doi:10.48550/arxiv.2603.30022.
56. Shao R, Li W, Zhang L, Zhang R, Liu Z, Chen R, et al. Large VLM-based vision-language-action models for robotic manipulation: a survey. arXiv:2508.13073. 2025. doi:10.48550/arxiv.2508.13073.
57. ElMallah R, Chhajer K, Lee CG. Score the steps, not just the goal: VLM-based subgoal evaluation for robotic manipulation. arXiv:2509.19524. 2025. doi:10.48550/arxiv.2509.19524.
58. Badithela A, Snyder D, Zha L, Mikhail J, O'Kelly M, Dixit A, et al. Reliable and scalable robot policy evaluation with imperfect simulators. arXiv:2510.04354. 2025. doi:10.48550/arxiv.2510.04354.
59. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model based autonomous agents. Front Comput Sci. 2024;18(6):186345. doi:10.1007/s11704-024-40231-1.
60. Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, et al. Do as I can, not as I say: grounding language in robotic affordances. arXiv:2204.01691. 2022. doi:10.48550/arxiv.2204.01691.
61. Rincon Arango JA, Marco-Detchart C, Julian Inglada VJ. Personalized cognitive support via social robots. Sensors. 2025;25(3):888. doi:10.3390/s25030888.
62. Raptis EK, Kapoutsis AC, Kosmatopoulos EB. RobotIQ: empowering mobile robots with human-level planning for real-world execution. Int J Adv Rob Syst. 2026;23(3):1–18. doi:10.1177/17298806261423235.
63. Zhang H, Solak G, Lahr GJG, Ajoudani A. SRL-VIC: a variable stiffness-based safe reinforcement learning for contact-rich robotic tasks. IEEE Robot Autom Lett. 2024;9(6):5631–8. doi:10.1109/lra.2024.3396368.
64. He X, Lv C. Robotic control in adversarial and sparse reward environments: a robust goal-conditioned reinforcement learning approach. IEEE Trans Artif Intell. 2024;5(1):244–53. doi:10.1109/TAI.2023.3237665.
65. Chen L, Jiang Z, Cheng L, Knoll AC, Zhou M. Deep reinforcement learning based trajectory planning under uncertain constraints. Front Neurorobot. 2022;16:883562. doi:10.3389/fnbot.2022.883562.
66. Chen X, Zhou C, Liu Y, Yang J. RM-RL: role-model reinforcement learning for precise robot manipulation. arXiv:2510.15189. 2025. doi:10.48550/arxiv.2510.15189.