PEER-REVIEWED JOURNAL OF COMPUTER SCIENCE (PRJCS)

ISSN Online: 3139-5082 Open Access Scholarly Journal

Physical AI: World Models and Embodied Intelligence

Rasmi M

Assistant Professor, Department of Computer Applications, Bethania Institute of Management Studies, Kunnamkulam, Kerala, India.

Article information

Received:15th June 2026
Received in revised form: 26th June 2026
Accepted: 6th July 2026
Available online: 13th July 2026

Volume: 1
Issue: 7
DOI: https://doi.org/10.5281/zenodo.21332611

Abstract

Physical AI names a shift in artificial intelligence away from purely digital tasks and toward systems that perceive, reason about, and act in the physical world. Between 2024 and 2026 this shift accelerated as robot foundation models, learned simulators, and embodied agents moved from isolated demonstrations to shared datasets, open policies, and reusable world models. This review organizes the area around two technical pillars. The first is world models: neural networks that learn predictive models of environment dynamics and support planning or behavior learning through imagined rollouts, from the early recurrent world model of Ha and Schmidhuber through the Dreamer family and joint-embedding predictive architectures. The second is generalist action policies, especially vision-language-action models such as RT-1, RT-2, Octo, and OpenVLA trained on cross-embodiment corpora like Open X-Embodiment. We connect these pillars to the simulation and data infrastructure that feeds them, including GPU-accelerated physics, domain randomization, and video generation framed as world simulation. We summarize benchmarks and evaluation practice, survey humanoid and manipulation applications, and discuss open challenges: data scarcity, the sim-to-real gap, safety, generalization, long-horizon control, and reproducible evaluation. All quantitative figures in the tables and charts are representative values collated from disparate reports and are intended to illustrate trends rather than to support controlled comparison. We close with directions toward neuro-symbolic integration and foundation world models.

Keywords: - Physical AI, World Models, Embodied Intelligence, Vision-Language-Action Models, Robot Foundation Models, Sim-to-Real, Reinforcement Learning, Generative Simulation

.

I. INTRODUCTION

For much of its modern history, artificial intelligence advanced fastest in disembodied settings such as text, images, and board games, where data is abundant and mistakes are cheap. The current frontier is shifting toward systems whose competence is measured by the consequences of their actions on real matter. This emerging direction is often grouped under the banner of Physical AI: artificial intelligence that perceives its surroundings through sensors, reasons about how those surroundings will change, and acts through a body such as a robot arm, a mobile base, or a humanoid platform. The distinguishing feature is not a single algorithm but a closed loop with the physical world, in which perception, prediction, and control are bound together by real dynamics, contact, and time [20].

The period from 2024 to 2026 gave this direction unusual momentum. Cross-embodiment robot datasets were assembled and released, generalist policies trained on them began to transfer across robot bodies, and large video and physics models were reframed as general-purpose simulators of the physical world [8], [16], [18]. Industry attention turned to humanoids and dexterous manipulation, while research consolidated around reusable foundation models for robotics rather than task-specific pipelines [21]. The result is a field that increasingly resembles the foundation-model era

of language and vision, but with the added constraints of embodiment: partial observability, safety, and the high cost of real-world data.

It is worth being precise about why a new label is useful. Digital agents that summarize documents or write code also perceive inputs and produce outputs, yet their outputs do not move mass, dissipate energy, or provoke contact forces, and their mistakes can be undone by resampling. Physical AI is defined by irreversibility and embodiment: an action that knocks over a cup cannot be retracted, sensing is partial and noisy, and the agent shares a continuous timeline with the world rather than proceeding turn by turn. These properties change what learning must achieve. A system must respect physical constraints, tolerate calibration drift, and remain safe under states never seen in training. The 2024 to 2026 wave of robot foundation models and world foundation models is best understood as an attempt to bring the scaling recipe of digital AI into this harder regime without discarding the constraints that embodiment imposes [18], [20].

This review is organized around two technical pillars that recur across the literature. The first is the world model, a learned predictive model of environment dynamics that lets an agent plan or improve its behavior by imagining possible futures rather than acting only from immediate observations [1], [2]. The second is the generalist action policy, most prominently the vision-language-action (VLA) model, which maps camera images and natural-language instructions directly to robot actions and inherits broad semantic knowledge from internet-scale pretraining [6], [7]. We trace the background that produced these pillars, examine each in turn, connect them to the simulation and data infrastructure that feeds them, and then treat benchmarks, applications, challenges, and future directions. Throughout, numerical values in tables and figures are labelled as representative, because reported robot results are rarely produced under matched conditions.

II. BACKGROUND: EMBODIED AI, REINFORCEMENT LEARNING, AND SIM-TO-REAL

A. From internet AI to embodied AI

Embodied AI studies agents that learn through interaction with an environment from a first-person, or egocentric, point of view, in contrast to internet AI that learns from static curated datasets [20]. A useful map of the area divides it into simulators, tasks, and methods. Simulators provide the worlds in which agents act; canonical tasks include visual exploration, visual navigation, and embodied question answering; and methods range from classical planning to end-to-end learning. Surveys of the field emphasize that the choice of simulator strongly shapes what can be studied, because fidelity, speed, and the diversity of scenes and objects bound the behaviors an agent can acquire [20].

B. Reinforcement learning and its data cost

Reinforcement learning (RL) provides the formal language for sequential decision making, in which an agent selects actions to maximize expected cumulative reward under unknown dynamics. Model-free RL learns a policy or value function directly from experience and has produced striking results in simulation, but it is notoriously sample-hungry, often requiring millions of interactions. For physical robots each interaction consumes real time, causes wear, and risks damage, so the raw sample cost of model-free RL is frequently prohibitive. This tension motivates two complementary strategies that dominate Physical AI: learning models of the environment so that experience can be reused in imagination, and pretraining on large offline datasets so that online interaction is minimized.

C. Imitation learning as the practical default

In practice, most deployed robot policies are trained not by reinforcement but by imitation, where a model learns to reproduce human teleoperated demonstrations through supervised behavior cloning. Imitation sidesteps the exploration and reward-design problems of reinforcement learning and turns robot learning into a data problem that resembles supervised learning in vision and language. This framing explains why the field invests so heavily in collecting demonstrations and why the generalist policies discussed later are overwhelmingly trained by imitation on large demonstration corpora [6], [9]. The weakness of pure imitation is compounding error: because the policy only sees expert states during training, small deviations at test time push it into unfamiliar states where mistakes accumulate. Remedies include collecting corrective data, adding interaction, or coupling the policy to a predictive model that can anticipate and recover from drift, which returns the discussion to world models.

D. The sim-to-real gap and domain randomization

Because simulation is fast, safe, and cheap, a common recipe trains policies in simulation and transfers them to hardware. The obstacle is the sim-to-real gap: differences in dynamics, sensing, lighting, and contact between the simulator and the world degrade transferred policies. Domain randomization addresses this by randomizing simulator parameters such as textures, lighting, masses, and friction during training, so that the real world appears to the model as one more variation of a broad distribution [14]. The original demonstration trained an object detector purely on non-realistic randomized renders yet localized objects in the real world to within about one and a half centimeters, illustrating that variability, rather than photorealism alone, can drive transfer [14]. Domain randomization remains a foundational tool, and it connects directly to the simulation infrastructure discussed later.

III. WORLD MODELS: LEARNING ENVIRONMENT DYNAMICS

A world model is a learned function that predicts how an environment will evolve, typically in a compact latent space rather than in raw pixels. Given past observations and a candidate action, it forecasts the next latent state and often the expected reward. With such a model an agent can search, plan, or train a policy inside its own predictions, a process sometimes called imagination or dreaming. The appeal is data efficiency and safety: rollouts in a learned model cost no real interaction and carry no physical risk.

A. Recurrent world models and latent imagination

The recurrent world model of Ha and Schmidhuber established a template that still organizes the field [1]. A variational autoencoder compresses high-dimensional frames into a small latent vector, a recurrent network with a mixture-density output predicts the next latent given the current latent and action, and a small controller selects actions from these features. Remarkably, the authors trained a controller entirely inside the dream generated by the model and then transferred it back to the real task, demonstrating that behavior can be learned without further environment interaction once the model is good enough [1]. The Dreamer line generalized this idea: Dreamer learns a latent dynamics model and improves a policy by propagating analytic value gradients back through imagined trajectories, solving long-horizon visual control tasks with strong data efficiency [2].

B. Scaling and generality

DreamerV3 pushed toward generality, applying a single configuration of hyperparameters across more than one hundred fifty tasks spanning continuous and discrete actions, visual and low-dimensional inputs, and dense and sparse rewards [3]. A widely cited result is that DreamerV3 collected diamonds in the game Minecraft from raw pixels and sparse rewards without human demonstrations or curricula, a task that demands long-horizon exploration [3]. DayDreamer then carried the Dreamer approach onto physical robots, learning locomotion, manipulation, and navigation directly in the real world across several platforms, with a walking policy acquired in roughly an hour of real interaction, a representative figure that signals the practical promise of model-based learning on hardware [19].

C. Joint-embedding predictive architectures

A distinct philosophy argues that predicting raw future pixels wastes capacity on unpredictable detail and that agents should instead predict abstract representations. The joint-embedding predictive architecture (JEPA) formalizes this: rather than reconstruct an output, it predicts the representation of a target from the representation of a context, capturing dependencies without generating every pixel [4]. The image-based instantiation, I-JEPA, predicts representations of masked target blocks from a single context block and learns semantic features without hand-crafted data augmentation, scaling efficiently with vision transformers [5]. This representation-first stance is influential because it frames world modeling as learning predictable structure rather than photorealistic rendering, a theme that reappears in debates about video generators as simulators [4], [5].

D. How world models are used

World models serve two distinct purposes, and confusing them causes much of the disagreement in the literature. The first purpose is behavior learning, where the model is a source of cheap synthetic experience: a policy is trained inside imagined rollouts and then deployed, as in the Dreamer family and its physical-robot descendant [2], [3], [19]. The second purpose is planning, where the model is queried at decision time to evaluate candidate action sequences and select the best, closer to model-predictive control. Each choice carries trade-offs. Planning at decision time is flexible and can incorporate new goals without retraining, but it is computationally expensive and sensitive to model error over long horizons. Training a policy in imagination amortizes that cost but bakes in whatever biases the model contains. A further design axis is what the model predicts: raw observations, compact latent states, or abstract representations. Predicting latent states or representations, rather than pixels, tends to improve robustness because the model is not penalized for failing to render unpredictable texture, which is a central argument for the joint-embedding view [4], [5]. Table II summarizes representative world-model approaches along these axes.

Table 2. Representative World-Model Approaches.

Approach	Year	Prediction target	Key idea
World Models [1]	2018	Latent frame + reward	VAE plus recurrent mixture-density model; train controller inside a dream
Dreamer [2]	2020	Latent state + value	Learn behaviors by backpropagating value gradients through imagined rollouts
DreamerV3 [3]	2023	Latent state + reward	Single configuration generalizes across 150+ tasks; Minecraft diamonds from scratch
DayDreamer [19]	2022	Latent state + reward	Dreamer on physical robots; online model-based learning on hardware
I-JEPA [5]	2023	Latent representation	Predict masked-region representations, not pixels; semantic features
Genie [17]	2024	Next video frame	Action-controllable interactive environments learned from unlabeled video

IV. VISION-LANGUAGE-ACTION MODELS AND ROBOT FOUNDATION MODELS

The second pillar is the generalist policy that maps perception and instruction to action. The idea is to import the recipe that worked for language and vision, namely pretraining a large model on broad data and adapting it to downstream tasks, into robotics, where data is scarcer and more heterogeneous [21]. The clearest expression is the vision-language-action model, which conditions on images and a natural-language command and outputs robot actions, usually as discretized tokens.

A. From RT-1 to RT-2

RT-1 is a transformer that takes camera images and a language instruction and emits discretized arm and base actions, running at interactive rates despite a modest thirty-five million parameters [6]. Trained on a large corpus of teleoperated demonstrations, it showed that a single model could absorb diverse task-agnostic data and still perform specific skills reliably. RT-2 changed the backbone by starting from a vision-language model pretrained on web-scale image and text data, then co-fine-tuning it to output actions expressed as text tokens [7]. This co-training let the model inherit semantic and reasoning abilities from the web, producing emergent generalization to novel objects and instructions that were absent from the robot data [7]. PaLM-E pursued a related fusion by injecting continuous sensor and image embeddings directly into a large language model for embodied planning, visual question answering, and manipulation [12].

B. Cross-embodiment data and open generalist policies

A turning point was the recognition that robot data, though scarce per laboratory, is collectively large across the community. The Open X-Embodiment effort pooled demonstrations from twenty-two robot types contributed by twenty-one institutions, covering hundreds of skills, and trained RT-X models that showed positive transfer across robot bodies [8]. Building on this shared corpus, Octo released an open transformer policy trained on roughly eight hundred thousand trajectories that can be conditioned by language or goal images and fine-tuned to new sensors and action spaces [9]. OpenVLA followed with a fully open seven-billion-parameter model that reported strong generalist manipulation and efficient fine-tuning on consumer hardware through low-rank adaptation [10]. More recent systems such as pi-0 add a flow-matching action module to a vision-language backbone to produce smooth continuous control for dexterous tasks across multiple embodiments [11]. Earlier, the generalist agent Gato showed that one network with fixed weights could span hundreds of tasks including robot stacking, foreshadowing the multi-embodiment ambition [13]. Perception foundation models such as Segment Anything increasingly serve as reusable front ends that supply open-vocabulary object masks to these policies [22].

C. Action representation and inference speed

A subtle but consequential design choice is how a policy represents actions. Early VLA models discretized each continuous command into a small number of bins and emitted them as tokens, which let a language-model backbone treat control as another sequence-prediction problem and made web pretraining directly reusable [7], [10]. Discretization is simple and aligns action output with text output, but coarse bins limit precision and long token sequences slow inference, which matters because control loops must run at tens of hertz on hardware. Alternative designs predict continuous actions directly, for example through flow matching or diffusion-style action experts attached to a vision-language backbone, trading architectural simplicity for smoother and more precise control at high frequency [11]. These choices interact with model size: larger backbones improve semantic generalization but strain the latency budget, so practical systems balance capacity against real-time constraints, sometimes distilling or quantizing models for deployment [10]. Table I compares representative robot foundation and VLA models.

Table 1. Representative Robot Foundation and Vision-Language-Action Models.

Model	Year	Conditioning	Key idea
RT-1 [6]	2022	Image + language	Efficient transformer emitting discretized actions at interactive rate
PaLM-E [12]	2023	Image + language	Sensor embeddings injected into a large language model for embodied planning
RT-2 [7]	2023	Image + language	Web-pretrained VLM co-fine-tuned to output actions as text; emergent generalization
RT-X / OXE [8]	2023	Image + language	Cross-embodiment training on pooled 22-robot data; positive transfer
Octo [9]	2024	Language or goal image	Open transformer policy on ~800k trajectories; flexible fine-tuning
OpenVLA [10]	2024	Image + language	Open 7B VLA; efficient low-rank fine-tuning on consumer GPUs
pi-0 [11]	2024	Image + language	Flow-matching action expert for smooth dexterous control

V. SIMULATION, DATA, AND GENERATIVE WORLD SIMULATORS

Both pillars are only as strong as the data and simulators that feed them. Physical AI therefore depends heavily on infrastructure for generating experience at scale.

A. GPU-accelerated physics and domain randomization

Classical robotics simulators ran physics on the CPU and fed a separate GPU for learning, creating a bottleneck. Isaac Gym placed both physics simulation and policy optimization on the GPU and passed data directly into learning tensors, reporting throughput improvements of two to three orders of magnitude for some tasks and enabling thousands of environments to run in parallel on a single accelerator [15]. Such throughput makes domain randomization practical at scale, since a policy can experience an enormous distribution of randomized dynamics and appearances during training [14]. Together, fast parallel physics and randomized worlds form the workhorse pipeline for locomotion and, increasingly, manipulation.

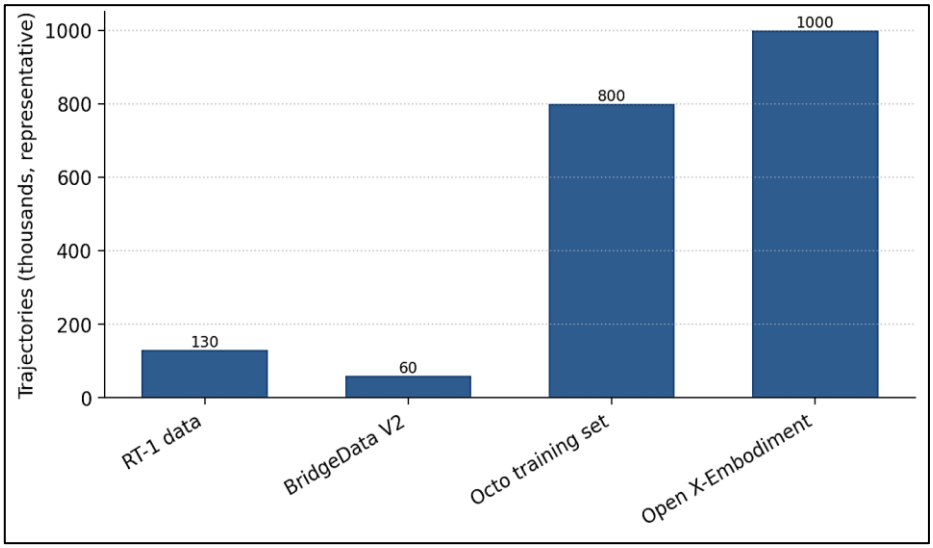
B. Generative simulation and video as world simulator

A newer thread treats large generative models as simulators. Video generation models trained on internet-scale footage produce coherent, temporally extended scenes, and one influential framing argues that scaling such models is a promising path toward general-purpose simulators of the physical world [16]. The systems can render plausible object motion and interaction from text or image prompts, though they do not enforce physical laws and can violate conservation or contact constraints. Genie extended this direction to action-controllable worlds, learning an interactive environment from unlabeled internet videos and a latent action model, so that a user can act frame by frame in a generated world without ground-truth action labels [17]. Platforms such as Cosmos package pretrained world foundation models, tokenizers, and data pipelines specifically for Physical AI, positioning a general world model that developers fine-tune into task-specific simulators [18]. These approaches promise abundant synthetic experience, but their fidelity to true dynamics remains an open question examined later.

C. Data scale across robot corpora

Even with generative simulation, real demonstrations anchor current generalist policies, and their scale has grown quickly. Figure 1 shows representative sizes of several robot-learning corpora, expressed in thousands of trajectories, to convey the trend from single-lab datasets toward pooled cross-embodiment collections [6], [8], [9]. The values are approximate and collated from separate reports, so they indicate order of magnitude rather than exact counts.

Fig 1. Representative scale of selected robot-learning corpora, in thousands of trajectories. Values are approximate and collated from separate reports; intended to show orders of magnitude, not exact counts.



VI. BENCHMARKS AND EVALUATION

Measuring progress in Physical AI is unusually hard because the environment is part of the experiment. Simulated benchmarks offer reproducibility, while real-robot evaluation offers validity, and the two rarely agree.

A. Simulated benchmarks

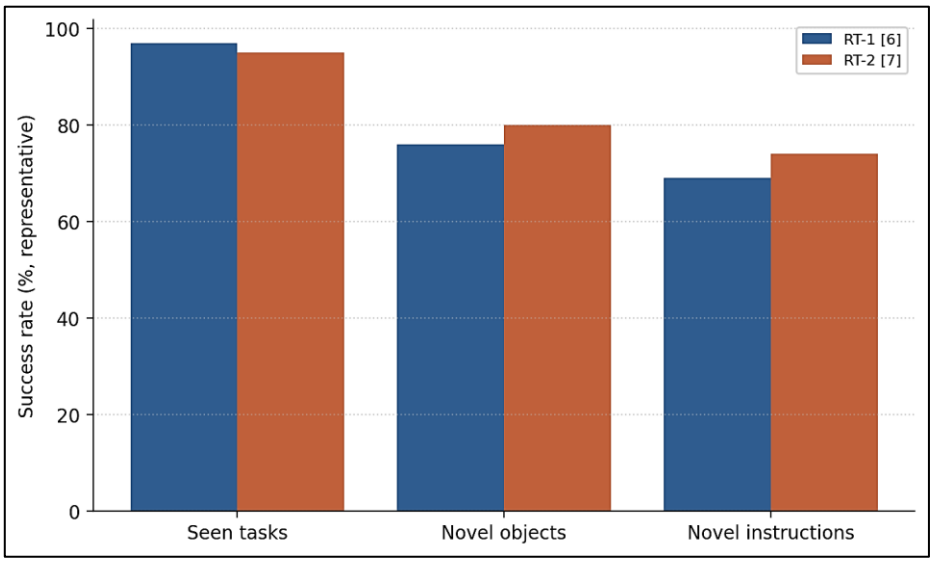
Meta-World provides a suite of distinct manipulation tasks with a shared robot and action space, designed for multi-task and meta reinforcement learning so that generalization across tasks can be measured rather than performance on a single skill [24]. CALVIN targets long-horizon, language-conditioned manipulation, requiring an agent to compose sequences of skills specified by free-form instructions from onboard sensors, and it reports success as the length of

correctly completed instruction chains [23]. Such benchmarks make controlled comparison possible, but success in simulation does not guarantee success on hardware.

B. Real-robot and cross-embodiment evaluation

Generalist policies are increasingly evaluated on physical robots and, crucially, across different robot bodies, as in the Open X-Embodiment protocol where a model trained on many embodiments is tested for positive transfer to specific ones [8]. Real evaluation introduces confounds that simulated benchmarks avoid: hardware differences, human resets, scene variation, and small sample sizes inflate variance and hinder reproducibility. Figure 2 shows representative, illustrative success rates for a few generalist policies across coarse evaluation categories. Because these numbers come from different studies with different robots, scenes, and task definitions, the figure indicates qualitative trends only and must not be read as a controlled head-to-head comparison.

Fig 2. Illustrative, representative success rates for generalist policies across coarse evaluation categories. Numbers are collated from disparate reports with differing robots and protocols and are not a controlled comparison.



C. What the metrics measure

The headline metric in manipulation is task success rate, the fraction of trials in which a goal is achieved, but this single number hides important structure. Long-horizon benchmarks such as CALVIN report the average number of consecutive sub-tasks completed before failure, which rewards sustained competence rather than isolated success and exposes the compounding-error problem directly [23]. Multi-task suites such as Meta-World instead emphasize breadth, measuring how many distinct skills a single policy can master and how quickly it adapts to held-out tasks [24]. Data efficiency, the amount of interaction or the number of demonstrations needed to reach a target success rate, is a further axis that model-based methods highlight because their advantage often lies there rather than in peak performance [2], [19]. No single scalar captures embodied competence, and responsible reporting therefore combines success rate with horizon length, generalization category, sample budget, and the variance observed across trials.

The broader lesson is that evaluation practice has not kept pace with model capability. Community efforts increasingly call for standardized real-robot protocols, shared task definitions, and reporting of variance across seeds and trials, so that claimed improvements reflect genuine capability rather than favorable conditions [21].

VII. HUMANOID AND MANIPULATION APPLICATIONS

The two pillars meet in applications that demand both prediction and dexterous action. Manipulation is the central testbed, since grasping, placing, and tool use require contact-rich control that is difficult to model analytically. Generalist VLA policies now perform multi-step manipulation such as picking specified objects, opening containers, and following language instructions across varied scenes, and open models make these capabilities reproducible outside a single laboratory [9], [10]. Flow-based policies extend this to dexterous, high-frequency skills such as folding laundry, clearing a table, and bagging groceries, tasks that stress precision and long horizons [11].

Locomotion and whole-body control illustrate the model-based branch. DayDreamer learned a quadruped walking gait directly on hardware in about an hour by improving a policy inside a learned world model, avoiding the usual reliance on hand-tuned simulation transfer [19]. Massively parallel GPU simulation with domain randomization has likewise produced robust legged locomotion that transfers to real robots, and the same recipe is now being pushed toward humanoids, where high degrees of freedom and balance constraints make both data efficiency and safety acute [15]. Humanoid platforms attract particular attention in the 2024 to 2026 period because a human-shaped body can, in principle, reuse human-scale environments, tools, and demonstration data, aligning hardware with the abundant human video that generative simulators consume [16], [18].

Mobile manipulation and navigation stress the same pillars over longer spatial and temporal scales. Here an agent must move through an environment, locate task-relevant objects, and then act on them, so perception, mapping, and language grounding combine with control. Embodied AI benchmarks originally framed navigation and embodied question answering as core tasks, and these remain natural settings for evaluating whether a policy can follow instructions across rooms rather than within a single tabletop scene [20]. Language-conditioned formulations let a user specify goals in free-form text, and long-horizon suites test whether an agent can chain navigation and manipulation into coherent multi-step behavior [23]. The move toward humanoids sharpens these demands, because a legged, two-armed body must maintain balance while manipulating, coupling whole-body control with task execution and raising the stakes for both data efficiency and safety [15], [19].

Across these applications a common architecture is emerging: a perception front end that may draw on segmentation or vision-language features [22], a policy or planner that is either a VLA model or a world-model controller, and a simulation-and-data pipeline that supplies training experience. The applications differ less in structure than in the balance between imagined experience and real demonstration.

VIII. CHALLENGES AND OPEN PROBLEMS

A. Data scarcity and heterogeneity

Robot data is expensive to collect and fragmented across incompatible embodiments, sensors, and action spaces. Pooling efforts such as Open X-Embodiment mitigate scarcity but introduce heterogeneity that models must reconcile, and the total volume remains small compared with the text and image corpora behind language and vision foundation models [8], [21]. Whether generative simulation can supply the missing volume without introducing systematic errors is unresolved.

B. The sim-to-real gap and simulator fidelity

Domain randomization and high-fidelity physics narrow the sim-to-real gap but do not close it, particularly for contact-rich manipulation where friction, deformation, and slip are hard to model [14], [15]. Video and learned world simulators add expressive appearance but do not guarantee physical consistency, so policies trained inside them may exploit artifacts that do not exist in reality [16], [17]. Quantifying and bounding this fidelity gap is an active problem.

C. Safety, generalization, and long horizons

Acting in the world makes errors costly, so safety and uncertainty quantification are first-class concerns rather than afterthoughts [21]. Generalization remains uneven: models transfer well to nearby variations yet degrade under distribution shift in objects, lighting, or dynamics. Long-horizon tasks compound these difficulties, since small per-step errors accumulate and sparse rewards make credit assignment hard, which is precisely why world models that plan over imagined futures and benchmarks that score instruction chains are valued [2], [3], [23].

D. Evaluation and reproducibility

Finally, evaluation itself is a bottleneck. Real-robot results depend on hardware, resets, and scene setup that are difficult to standardize, and small sample sizes inflate variance, so reported gains can be fragile [21]. Progress in Physical AI will depend on evaluation protocols that are as reproducible as the simulated benchmarks yet as valid as real deployment.

IX. FUTURE DIRECTIONS AND CONCLUSION

Several directions are likely to shape the next phase. The first is convergence between the two pillars: unifying learned world models with generalist action policies so that a single system both imagines futures and acts, combining the data efficiency of model-based planning with the semantic breadth of VLA pretraining [2], [10]. The second is the foundation world model, a large reusable predictor of physical dynamics that is pretrained once and fine-tuned into many task-specific simulators or planners, a vision that platforms for Physical AI are beginning to pursue [18]. Representation-first approaches such as the joint-embedding predictive architecture may provide the abstract, predictable state spaces such foundation models need, avoiding the waste of pixel-level generation [4], [5].

A third direction is neuro-symbolic integration, pairing learned perception and control with explicit structure for goals, constraints, and safety, so that long-horizon plans can be verified and guarantees can be stated. This complements the data-driven pillars by supplying the compositional and verifiable reasoning that pure end-to-end policies lack. It also connects Physical AI to the broader agentic AI paradigm, in which teams of specialized agents coordinate multi-step workflows through explicit orchestration, a coordination layer that embodied systems will need as they combine perception, planning, and control modules [25]. In parallel, hard-won lessons from autonomous-vehicle perception, where deep learning must stay robust under occlusion, weather, and distribution shift, transfer directly to embodied perception under similar real-world stress [26]. Alongside these, the community will need standardized, reproducible evaluation and honest reporting of variance, without which apparent progress cannot be trusted [21].

In summary, Physical AI marks a shift from disembodied prediction toward systems that perceive, predict, and act in the physical world. Its two organizing pillars, learned world models and generalist vision-language-action policies, are increasingly fed by shared cross-embodiment data, GPU-accelerated physics, and generative simulators, and they are

converging toward reusable foundation models for embodiment. The obstacles are real: data scarcity, the sim-to-real and fidelity gaps, safety, generalization, long horizons, and evaluation. The representative numbers gathered here should be read as signposts of direction rather than settled measurements. Yet the trajectory from isolated demonstrations toward open policies, shared datasets, and foundation world models suggests that embodied intelligence is becoming a systematic engineering discipline rather than a collection of bespoke results.

REFERENCES

[1]D. Ha and J. Schmidhuber, "Recurrent World Models Facilitate Policy Evolution," in Advances in Neural Information Processing Systems (NeurIPS), 2018, arXiv:1809.01999.

[2]D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to Control: Learning Behaviors by Latent Imagination," in International Conference on Learning Representations (ICLR), 2020, arXiv:1912.01603.

[3]D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering Diverse Domains through World Models," arXiv preprint arXiv:2301.04104, 2023.

[4] Y. LeCun, "A Path Towards Autonomous Machine Intelligence," OpenReview, ver. 0.9.2, 2022.

[5] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 15619-15629.

[6] A. Brohan et al., "RT-1: Robotics Transformer for Real-World Control at Scale," arXiv preprint arXiv:2212.06817, 2022.

[7] B. Zitkovich et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in Proc. 7th Conf. Robot Learn. (CoRL), PMLR, vol. 229, 2023.

[8] Open X-Embodiment Collaboration, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2024, pp. 6892-6903, arXiv:2310.08864.

[9] Octo Model Team, "Octo: An Open-Source Generalist Robot Policy," in Robotics: Science and Systems (RSS), 2024, arXiv:2405.12213.

[10] M. J. Kim et al., "OpenVLA: An Open-Source Vision-Language-Action Model," arXiv preprint arXiv:2406.09246, 2024.

[11] K. Black et al., "pi-0: A Vision-Language-Action Flow Model for General Robot Control," arXiv preprint arXiv:2410.24164, 2024.

[12] D. Driess et al., "PaLM-E: An Embodied Multimodal Language Model," in Proc. Int. Conf. Mach. Learn. (ICML), 2023, arXiv:2303.03378.

[13] S. Reed et al., "A Generalist Agent," Trans. Mach. Learn. Res. (TMLR), 2022, arXiv:2205.06175.

[14] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2017, pp. 23-30.

[15] V. Makoviychuk et al., "Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning," arXiv preprint arXiv:2108.10470, 2021.

[16] T. Brooks et al., "Video Generation Models as World Simulators," OpenAI Technical Report, 2024.

[17] J. Bruce et al., "Genie: Generative Interactive Environments," in Proc. Int. Conf. Mach. Learn. (ICML), 2024, arXiv:2402.15391.

[18] NVIDIA, "Cosmos World Foundation Model Platform for Physical AI," arXiv preprint arXiv:2501.03575, 2025.

[19] P. Wu, A. Escontrela, D. Hafner, K. Goldberg, and P. Abbeel, "DayDreamer: World Models for Physical Robot Learning," in Conf. Robot Learn. (CoRL), 2022, arXiv:2206.14176.

[20] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A Survey of Embodied AI: From Simulators to Research Tasks," IEEE Trans. Emerg. Topics Comput. Intell., vol. 6, no. 2, pp. 230-244, 2022.

[21] R. Firoozi et al., "Foundation Models in Robotics: Applications, Challenges, and the Future," arXiv preprint arXiv:2312.07843, 2023.

[22]A. Kirillov et al., "Segment Anything," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 4015-4026, arXiv:2304.02643.

[23] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks," IEEE Robot. Autom. Lett., vol. 7, no. 3, pp. 7327-7334, 2022.

[24] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning," in Conf. Robot Learn. (CoRL), PMLR, vol. 100, 2020, pp. 1094-1100.

[25] "Agentic Artificial Intelligence: Autonomous Multi-Agent Workflows and Orchestration," Peer-Reviewed Journal of Computer Science (PRJCS), E-ISSN 3139-5082, 2025.

[26] "Autonomous Vehicle Perception: Deep Learning Challenges and Solutions," Peer-Reviewed Journal of Computer Science (PRJCS), E-ISSN 3139-5082, 2025.