Review

A Survey of Physical AI: A History from ChatGPT to World Models and Embodied Agents

Haichao Zhang *, Mingfei Chen, Shwai He, Zhengtong Xu, Yifan Shen, Yiyang Huang, Jianglin Lu, Yijiang Li, Yu She, Yun Fu

Posted Date: 2 June 2026

doi: 10.20944/preprints202606.0173.v1

Keywords: physical AI; ChatGPT; large language models; LLM-based world knowledge; world models; multimodal large language models; vision-language-action models; embodied agents; embodied AI; robotics; physical reasoning; multimodal grounding; policy learning; sim-to-real transfer; closed-loop evaluation

 doi:10.20944/preprints202606.0173.v1

Review

A Survey of Physical AI: A History from ChatGPT to World Models and Embodied Agents

Haichao Zhang [1,*], Mingfei Chen [2], Shwai He [3], Zhengtong Xu [4], Yifan Shen [5], Yiyang Huang [1], Jianglin Lu [1], Yijiang Li [6], Yu She [4] and Yun Fu [1]

[1] Northeastern University
[2] University of Washington
[3] University of Maryland, College Park
[4] Purdue University
[5] University of Illinois Urbana-Champaign
[6] University of California, San Diego
* Correspondence: zhang.haich@northeastern.edu

Abstract

Physical AI aims to extend artificial intelligence from digital reasoning to perception, prediction, simulation, planning, and action in the physical world. While recent progress has advanced through vision-language models, vision-language-action models, world models, policy learning, and embodied agents, existing discussions are often organized from robotics-centric, vision-centric, or cyber-physical perspectives. This survey instead studies Physical AI through the lens of *LLM-based world knowledge*. We argue that LLMs encode implicit semantic, commonsense, procedural, and causal priors through large-scale pretraining, making them useful high-level sources for physical reasoning, multimodal grounding, action grounding, and embodied decision making. However, language-mediated knowledge is sparse and lossy for dense physical states, such as geometry, motion, contact, force, high-frequency dynamics, and long-horizon temporal evolution. VLMs and MLLMs ground LLM-derived priors into perception, but often expose physical understanding through language outputs. VLAs connect perception and language to executable actions, yet typically lack predictive models of how the world evolves under actions. This motivates a roadmap from LLM-based world knowledge to multimodal grounding, action grounding, world modeling, policy learning, and embodied deployment. We review how these components provide perceptual, actionable, predictive, and simulative substrates for Physical AI, and discuss open challenges in grounding, world modeling, closed-loop evaluation, safety, and generalization. A curated repository of related papers and resources is available at https://github.com/Hai-chao-Zhang/Awesome-Physical-AI.

Keywords: physical AI; ChatGPT; large language models; LLM-based world knowledge; world models; multimodal large language models; vision-language-action models; embodied agents; embodied AI; robotics; physical reasoning; multimodal grounding; policy learning; sim-to-real transfer; closed-loop evaluation

1. Introduction

Large language models (LLMs), including frontier systems such as GPT-4 [1], Gemini [2], and Claude [3], have evolved from text generators into general-purpose reasoning engines capable of instruction following, task decomposition, tool use, planning, and agentic interaction [4,5]. A central reason for this transition is that large-scale pretraining allows LLMs to encode broad forms of *world knowledge*: semantic knowledge about objects and events, commonsense knowledge about everyday situations, procedural knowledge about how tasks are performed, and causal knowledge about likely consequences of actions [6–9]. Although this knowledge is implicit and language-mediated, it provides

 doi:10.20944/preprints202606.0173.v1

useful high-level priors for systems that must reason about, interact with, and eventually act in the physical world.

However, LLM-based world knowledge alone is insufficient for *Physical AI*. The physical world is dense, continuous, temporal, and governed by geometry, dynamics, contact, force, uncertainty, and embodiment-specific constraints. Language is an effective abstraction interface, but it is too sparse and lossy to represent physical states such as trajectories, velocities, contacts, deformations, occlusions, high-frequency dynamics, and long-horizon temporal evolution. For the NLP community, the central question is not whether robots can act, but how language-derived semantic, commonsense, procedural, and causal priors should be represented, grounded, verified, and coupled with predictive models when language becomes an interface to physical agency.

Vision-language models (VLMs) and multimodal large language models (MLLMs) provide a first step in this grounding process. By connecting language with images, videos, regions, objects, spatial relations, and affordances, they allow LLM-derived world knowledge to become situated in perceptual observations [2,10–13]. For Physical AI, this grounding is crucial: an agent must not only know that a glass is fragile or that a handle affords pulling, but also determine whether the relevant object is present, where it is, and whether the current configuration makes an action feasible. Yet many VLMs still express physical understanding primarily through language outputs [14], making them effective for high-level description but limited for dense physical prediction [15], continuous control, and action-conditioned future modeling.

Vision-Language-Action (VLA) models move one step further, from perception and description toward executable behavior. By mapping visual observations and language instructions to actions, VLAs provide an action-facing interface between multimodal reasoning and embodied control [16–19]. Nevertheless, VLAs alone are not sufficient for general Physical AI. Their action representations are often embodiment-specific, their robot data are limited compared with web-scale language and vision corpora, and their generalization across tasks, environments, and robot bodies remains fragile. More importantly, although VLAs connect perception and language to action, they typically do not provide an internal predictive model of how the physical world evolves under actions.

This motivates the role of *world models*. In contrast to LLM-based world knowledge, which captures what is semantically, procedurally, or causally plausible, world models aim to predict or simulate what will happen next under physical dynamics [20,21]. They mark a transition beyond purely LLM-centric backbones: instead of only relying on language-encoded priors, world models directly learn predictive and simulative knowledge from videos, trajectories, and embodied interactions. Video world models provide visual imagination and future prediction; latent world models learn compact predictive representations for planning and control; and interactive or action-conditioned world models allow agents to reason about counterfactual futures before acting. Thus, LLMs and world models are complementary: LLMs provide high-level knowledge about what typically happens and what actions may be meaningful, whereas world models estimate what is likely to happen next.

1.1. Definition, Roadmap, Scope, and Relation to Existing Surveys.

In this survey, we study Physical AI through the lens of *LLM-based world knowledge*. From a language-centered perspective, we define Physical AI as AI systems that ground language-based semantic, commonsense, procedural, and causal knowledge into multimodal perception, physical prediction, simulation, planning, policy learning, and embodied action. We include work that helps transform language-derived world knowledge into perception, action, prediction, planning, or embodied deployment, rather than attempting to cover all robotics, control, simulation, or video generation. This scope distinguishes our survey from existing perspectives on physical AI. Compared with robotics-centric surveys, our organizing variable is the interface between language-mediated priors and physical agency; compared with vision-centric surveys, our focus is not perception or generation alone but action-conditioned prediction and deployment; compared with VLA surveys, our roadmap separates action grounding from predictive world modeling. As shown in Figure 1, we organize recent progress from LLM world knowledge to multimodal grounding, action grounding, world modeling, policy

 doi:10.20944/preprints202606.0173.v1

learning, and embodied deployment, followed by open challenges in sim-to-real transfer, closed-loop evaluation, safety, reproducibility, and frontier systems.

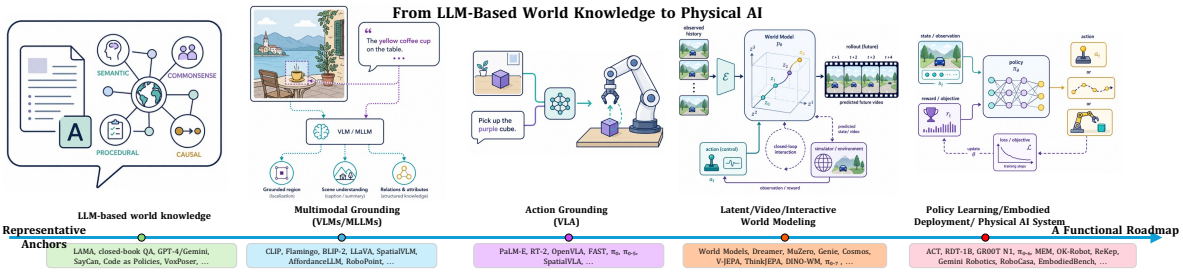

Figure 1. A roadmap from LLM-based world knowledge to Physical AI. LLMs provide semantic, commonsense, procedural, and causal priors, but these language-mediated priors must be grounded into multimodal perception, executable action, predictive world models, policy learning, and closed-loop embodied deployment. The organization is functional rather than chronological; examples are representative anchors.

1.2. Contributions.

To the best of our knowledge, this is the first world-knowledge-centered roadmap survey of Physical AI. Unlike robotics-centric, vision-centric, VLA-centric, or world-model-centric surveys, we organize recent progress around how LLM-derived world priors are grounded into perception, action, predictive world modeling, policy learning, and embodied deployment. This survey makes the following four contributions:

1. **A world-knowledge-centered formulation of Physical AI.** We provide, to the best of our knowledge, the first survey formulation of Physical AI in the current foundation-model era from the perspective of LLM-based world knowledge and world-model-based predictive knowledge. This formulation views Physical AI as systems that transform world priors into perception, prediction, simulation, planning, policy learning, and real-world action.
2. **A language-centered organization of recent advances.** We organize recent progress around the interfaces through which language-mediated priors become perceptual, actionable, predictive, and deployable physical intelligence, rather than treating Physical AI only as a robotics, vision, or cyber-physical systems problem.
3. **A roadmap from world knowledge to embodied agency.** We provide a layered roadmap from LLM-based world knowledge to multimodal grounding, action grounding, world modeling, policy learning, and embodied deployment, clarifying how these components jointly support Physical AI.
4. **A deployment-oriented discussion of open challenges.** We identify key gaps toward deployable Physical AI, including dense physical representation, language-to-action grounding, long-horizon world modeling, sim-to-real transfer, closed-loop evaluation, safety, reproducibility, and frontier systems.

2. LLM-Based World Knowledge for Physical AI

LLM-based world knowledge refers to language-mediated world priors stored as parametric regularities in large language models [6,7]. Prompts and context expose these priors as task hypotheses about objects, goals, actions, procedures, and constraints for Physical AI [4,5].

2.1. World Knowledge as Coupled Priors.

World knowledge is better treated as a set of coupled priors than as a monolithic store: semantic knowledge grounds object categories and references; commonsense knowledge supplies default assumptions and safety constraints; procedural knowledge decomposes tasks; causal knowledge suggests outcomes and risks; spatial knowledge supports layout and manipulation preconditions; and affordance knowledge links objects or parts to possible interactions [22–24]. In grounded Physical AI systems, spatial and affordance priors are often operationalized by models that connect linguistic

 doi:10.20944/preprints202606.0173.v1

relations to perceptual structure: recent spatial VLMs map relations such as "inside" or "aligned with" to metric or 3D structure [25–27]. Affordance-oriented work similarly identifies which objects, parts, or regions support a requested action [28–30].

2.2. Parametric World Knowledge in LLMs.

LLMs encode factual associations, procedural patterns, commonsense defaults, and task constraints through large-scale pretraining on corpora that repeatedly describe object functions, task procedures, safety constraints, and likely outcomes [4,6,7,31]. Because these descriptions recur across documents, they become parametric regularities whose strength tracks pretraining exposure [8,9]. Recent recall and materialization studies further probe which factual priors can be surfaced from LLMs [32,33]. Embodied planning work provides a task-level diagnostic: LLMs can often retrieve plausible steps, object uses, and constraints for situated tasks [34–36].

2.3. LLMs as Priors and Controllers in Physical AI.

LLMs contribute to Physical AI in four main ways. For *task planning*, SayCan [37] grounds LLM-generated step sequences with learned affordance values to filter physically infeasible actions, and Inner Monologue [38] enables closed-loop replanning by feeding back environmental observations as language [39]. Moving from prose to code, *skill generation* approaches such as Code as Policies [40] and ProgPrompt [41] prompt LLMs to produce executable robot programs, while Voyager [42] uses LLMs to iteratively build open-ended skill libraries. At the level of *goal and reward specification*, VoxPoser [43] composes spatial value maps from language instructions, reducing manual reward engineering. Finally, as *agentic orchestrators*, LLMs coordinate tools and sub-modules across multi-step tasks [35,36], acting as closed-loop decision engines rather than one-shot planners. Together, these uses establish LLMs as the semantic and procedural scaffolding of Physical AI pipelines—a foundation whose limits are examined next.

2.4. Limits of Language-Only Physical Reasoning.

The systems reviewed above reveal a common pattern: LLMs are effective at proposing goals, plans, and constraints, but their outputs become more reliable when paired with grounding, verification, or action interfaces [35,37,38]. The core limitation is the abstraction gap between language and physics. Language compresses continuous physical states into sparse descriptions and often omits pose, velocity, contact, deformation, uncertainty, and embodiment constraints. In this sense, LLM-based priors remain linguistic: they may suggest that glass is fragile or that a handle can be grasped, but they do not estimate geometry, contact, force, friction, or future trajectories [44,45]. LLMs can therefore hallucinate objects, propose infeasible plans, or assume that actions succeed. Recent planning analyses, including LLM-modulo studies and PlanBench updates, sharpen this point by showing that language-only models remain unreliable without external models or verifiers [46,47]. Physical reasoning benchmarks further expose weaknesses in dynamics, vision-grounded physics, and tool use [48,49]. These limitations motivate the remainder of this survey, which progressively addresses perceptual grounding, action grounding, and predictive world modeling. LLMs are therefore useful interfaces and coordinators, but they form a crucial yet insufficient layer of Physical AI.

3. Grounding World Knowledge into Perception

To act in the physical world, an AI system must connect language-based world knowledge to perceptual observations. VLMs and MLLMs are the first major bridge in this transition: they align language with images, videos, regions, objects, and scenes [10,12,50,51]. Language-only knowledge is underspecified with respect to the current world state: an LLM may know that glass is fragile or that a handle affords pulling, but an embodied system must determine whether the relevant object is present, where it is, and whether the current configuration makes that prior actionable. Thus, perception makes world knowledge situated, task-relevant, and actionable.

 doi:10.20944/preprints202606.0173.v1

3.1. Vision-Language and Multimodal Models.

The development of VLMs can be viewed as a progression from image-text representation alignment to multimodal instruction following. Contrastive and cross-modal pretraining made language an open-vocabulary interface for visual concepts, allowing models to retrieve, classify, and reason over images beyond fixed label spaces [50,52,53]. More recent MLLMs connect visual encoders to large language backbones through cross-attention, query transformers, projection layers, or instruction tuning, enabling few-shot visual reasoning, visual dialogue, and open-ended image understanding [11,54,55]. For Physical AI, their importance is that instructions, observations, and task context can be interpreted through a shared language interface. A VLM can identify task-relevant objects and connect them to commonsense priors, making it useful as a high-level perceptual reasoner in open-world environments. Yet this grounding remains mostly semantic rather than physical: VLMs do not automatically recover metric state, physical parameters, or control-relevant uncertainty [56–59].

3.2. Spatial, Temporal, and Affordance Grounding.

Physical AI requires grounding beyond object categories: agents must localize objects, resolve spatial relations, and infer action affordances. Recent MLLMs provide stronger grounded outputs through region, mask, and point supervision [60,61], unified detection-grounding-segmentation [62], and explicit pointing supervision [63]. For robotics, recent work extends grounding toward robotics-relevant spatial reasoning and language-conditioned affordance prediction [25,27,64]. These works improve perceptual grounding, but Physical AI still requires metric state, reachability, uncertainty, and embodiment-specific constraints. Temporal grounding moves perception from static recognition to event localization and state change. Recent video MLLMs explicitly model timestamps and grounded moments, including TimeChat [65], VTG-LLM [66], Grounded-VideoLLM [67], VQToken [68], and TimeSuite [69]; VideoGLaMM further extends grounding to pixel-level video regions [70]. However, temporal grounding is still not physical dynamics: Video-MME and PhysBench show that long-video understanding and physical-world reasoning remain difficult for current MLLMs [71,72]. Affordance grounding connects perception to possible actions. Recent work uses VLM/LLM priors to predict interaction regions, 3D affordances, or language-conditioned affordance points, as in AffordanceLLM, PAVLM, Palm, and RoboPoint [26,28,73,74]. These models make perception more action-relevant, but feasibility still depends on embodiment, geometry, reachability, & low-level control.

3.3. The Language Bottleneck for Dense Physical States.

A major limitation of VLMs is that physical understanding is often mediated through language rather than dense state estimation. Text can describe a scene, but it does not encode pose, depth, motion, contact, uncertainty, or action-conditioned dynamics. Recent evaluations show this gap: BLINK probes low-level visual perception [75], Video-MME tests long-horizon temporal understanding [76], and PhysBench, QuantiPhy, and MASS-Bench evaluate physical reasoning, quantitative physics, and motion-aware spatiotemporal grounding [72,77,78]. Thus, VLMs are better treated as perceptual grounding layers, not complete physical world models.

4. Grounding World Knowledge into Action

Grounding world knowledge into perception allows an agent to understand what is happening in a scene, but Physical AI additionally requires deciding what to do next. Vision-Language-Action (VLA) models address this step by mapping visual observations, embodiment states, and language goals into action outputs. By connecting semantic understanding with actionable control, VLAs provide a key bridge between foundation-model reasoning and physical-world interaction. This section first reviews action space design and policy learning, followed by language-grounded VLA policies, hybrid reasoning-to-action frameworks, and emerging directions toward reliable Physical AI systems.

 doi:10.20944/preprints202606.0173.v1

4.1. Action Spaces and Policy Learning.

The action space defines how a Physical AI system acts. Depending on the embodiment and task, actions can be represented as poses, joint states, gripper commands, contact targets, waypoints, trajectories, action chunks, or high-level skills. This action representation determines how pretrained world knowledge becomes executable behavior: high-level skills align better with language and planning, while low-level commands provide precise but embodiment-specific control.

Recent VLA systems instantiate this spectrum through several action interfaces. One line discretizes robot actions into tokens, enabling autoregressive prediction with transformer-based VLM backbones, as in RT-2 [17] and OpenVLA [18]; FAST further improves this direction by tokenizing high-frequency action sequences in frequency space, and can be combined with models such as π_0 to form efficient autoregressive VLA policies [79]. A second line predicts continuous action chunks or trajectories, following action-chunking and diffusion or flow-based visuomotor policies such as ACT [80], π_0 [19], $\pi_{0.5}$ [81], DexVLA [82], and RDT-1B [83]. A third line spatially structures the action interface beyond raw motor commands. SpatialVLA [84] represents actions with adaptive 3D action grids, grounding candidate motions in egocentric coordinates. 3D-VLA [85] uses 3D interaction tokens to predict task-relevant targets or motion primitives in scene space. Such spatial action representations expose geometry to the policy and can improve transfer across objects, scenes, and embodiments.

4.2. Grounding Language in Action.

Early VLAs asked whether a VLM could act by treating actions like language. PaLM-E showed that embodied multimodal representations can support robot reasoning and planning [16]. RT-2 made the link to control explicit by co-fine-tuning a pretrained VLM on web-scale vision-language tasks and robot trajectories, representing robot actions as text-like tokens, and de-tokenizing outputs back into commands [17]. This matters because semantic and commonsense priors learned from Internet-scale data can be aligned with physical action data. Scaling this recipe requires broader data and transferable policy interfaces. Open X-Embodiment and RT-X standardize data across many robot embodiments [86]. Octo learns an open generalist policy from heterogeneous data [87], and OpenVLA trains on large-scale real robot demonstrations [18]. These models take image observations, proprioception, and language instructions as input, and output action tokens. They align well with language-model training, but remain data-dependent and limited in physical resolution, especially for high-frequency motion, contact timing, force control, and recovery.

4.3. Reasoning to Action.

Recent systems increasingly use hybrid VLA architectures that separate world knowledge and reasoning from motion generation. A VLM grounds instructions, object semantics, spatial relations, task history, and commonsense priors, while a specialized policy converts this reasoning into high-frequency actions under embodiment constraints. This shifts VLA from action-as-language toward a reasoning-to-control interface for Physical AI. π_0 combines a pretrained VLM with a flow-matching action expert [19], and $\pi_{0.5}$ adds heterogeneous co-training over robot data, web data, and semantic prediction tasks [81]. DexVLA uses a VLM with a diffusion expert [82], RDT-1B scales diffusion for bimanual manipulation [83], and GR00T N1 combines vision-language reasoning with a diffusion transformer action generator for humanoids [88]. TinyVLA [89], SmolVLA [90], Xiaomi-Robotics-0 [91], and StarVLA-α [92] show that efficiency and real-time execution are also part of the roadmap. The trend is also moving beyond offline imitation. $\pi^*_{0.6}$ uses reinforcement learning from real deployments with experience and corrective interventions [93], while MEM adds video and text memory for longer-horizon behavior and adaptation from history [94].

4.4. Physical AI Roadmap.

VLA models provide an action-facing layer for Physical AI, but action prediction alone is not sufficient. Even when VLAs use visual-language reasoning to interpret goals, objects, spatial relations, and task history, this reasoning is still coarse and semantic rather than grounded in physical dynamics.

 doi:10.20944/preprints202606.0173.v1

One potential direction is to connect multimodal LLM world knowledge with physical prediction and control through world models. These models can capture factors that are difficult to describe in language, such as friction, compliance, contact geometry, force, uncertainty, and timing, enabling physically grounded planning, control, and recovery beyond direct data-driven imitation.

5. World Models for Physical AI

The previous sections describe how LLMs provide high-level world knowledge, how VLMs ground such knowledge into perception, and how VLAs connect perception and language to executable actions. However, VLA models alone are insufficient for general Physical AI. Even when they align language, observations, and actions, they typically lack an internal predictive model of how the physical world evolves under actions. This limitation motivates a transition from using LLMs as semantic priors toward learning *world models* that directly acquire predictive and simulative knowledge from videos, trajectories, and embodied interactions.

We use *world model* to refer to a model that predicts or simulates future observations, latent states, rewards, values, or action consequences from current states and possible actions. This definition distinguishes world models from LLM-based world knowledge. LLMs encode semantic, commonsense, procedural, and causal priors about what usually happens and what actions may be meaningful; world models estimate what is likely to happen next under physical dynamics. In short, LLMs provide knowledge *about* the world, whereas world models provide predictive mechanisms for acting *in* the world. This idea has a long history in model-based learning and control: early neural world models learned compressed spatiotemporal representations for agents [20], latent dynamics models such as PlaNet and Dreamer learned to plan and act through latent imagination [95–97], and MuZero showed that decision-centric models can support planning by predicting value, reward, and policy-relevant quantities without explicitly reconstructing observations [98]. For Physical AI, recent world models can be organized by their prediction target and interaction interface. A more detailed taxonomy with additional representative works is provided in Appendix A.4.

5.1. Video-Space World Models.

Video-space world models generate or predict future visual observations, providing an intuitive interface for visual imagination, synthetic data generation, and future scene simulation. Representative examples include generative driving and real-world simulators such as GAIA-1 and UniSim, as well as interactive environment models such as Genie [99–101]. Cosmos further frames world foundation models as general-purpose world models that can be adapted to Physical AI applications such as robotics, autonomous driving, and synthetic data generation [56]. However, photorealistic video generation alone is not sufficient for Physical AI. A useful video-space world model must also support temporal consistency, controllability, action conditioning, and physically plausible dynamics.

5.2. Latent World Models.

Latent world models predict future states in representation space rather than reconstructing pixels. This is important because dense video generation can be computationally expensive and may allocate capacity to visual details that are irrelevant to control. Latent models instead focus on task-relevant dynamics, making them attractive for efficient planning, policy learning, and long-horizon imagination. Beyond latent dynamics in model-based reinforcement learning, JEPA-style predictive architectures argue for learning world models directly in representation space [21,102,103]. Recent V-JEPA 2 further connects self-supervised video representation learning with robot planning by post-training an action-conditioned latent world model from robot trajectories [104].

5.3. Interactive and Action-Conditioned World Models.

Interactive and action-conditioned world models are the most directly relevant to Physical AI because embodied agents must evaluate counterfactual futures before acting. Rather than passively predicting what may happen next, these models estimate what would happen if a particular action

 doi:10.20944/preprints202606.0173.v1

were taken. This capability supports planning, policy learning, simulation-based training, safety evaluation, and recovery. It also clarifies why world models are complementary to VLAs: VLAs provide an action-facing interface, while world models provide the predictive substrate needed to evaluate actions before execution. The long-term challenge is therefore to connect LLM-derived world knowledge, multimodal grounding, VLA-style action interfaces, and world-model-based prediction into embodied agents that can act reliably in real or interactive environments.

6. From Models to Physical AI Systems

The roadmap from LLM based world knowledge to Physical AI ultimately requires systems that can act in real or interactive environments. While earlier sections focus on representation, reasoning, world modeling, and VLA models, this section shifts the focus from model capability to system deployment.

6.1. Embodied Agents.

Embodied agents connect high level task understanding with physical execution. A typical system may parse an instruction, infer task structure, ground objects and states, choose actions and trajectories, execute them through policies or controllers, and verify whether the expected outcome has been achieved [37,105–113]. This differs from a pure VLA formulation because the action model is only one component of the full system. System performance also depends on state estimation, execution interfaces, controller behavior, and online verification. Many embodied agents therefore use modular interfaces between reasoning and control. High level modules may operate over language, symbolic states, object relations, keypoint constraints, value maps, or robot programs, while low level modules execute motion primitives, grasp planners, visuomotor policies, or controllers [40,43,114,115]. This modularity can make long horizon behavior easier to specify and diagnose, but it also raises interface design questions. The system must decide what information should be passed between modules, how uncertainty should be handled, and when the agent should replan instead of continuing execution.

6.2. Deployment and System Integration.

A central challenge in Physical AI is the gap between offline evaluation and interactive deployment. Offline evaluation measures whether a model matches reference labels on fixed data, whereas deployment tests whether these predictions can drive closed-loop progress in the physical world. Open-loop accuracy can hide compounding errors, poor recovery, and sensitivity to state deviations, so robust deployment requires systems to observe the environment, execute actions, verify outcomes, and update plans when reality differs from expectation. Recent work explores this direction through real-world agent platforms [105], open-vocabulary planning and grounded decoding [116], LLM/VLM-based robot decision frameworks [117], and deployment-oriented action reasoning models [115,118]. Another key design choice is the interface between learned models and robot controllers. Frontier models may produce language plans, symbolic predicates, object constraints, value maps, or robot code, while robots execute controller references. An effective interface should be expressive enough to support diverse tasks while remaining constrained enough for stable execution. This motivates systems that connect learned perception and reasoning to symbolic world models [119], task-and-motion planning [120], optimization and constraint-based motion generation [43,106,114], and robot-program or policy-level execution interfaces [121,122].

6.3. Benchmarks and Evaluation.

Evaluation remains a central challenge for Physical AI. Static language and vision benchmarks measure recognition, reasoning, or prediction under fixed inputs, whereas Physical AI requires evaluation under interaction, where actions change future observations and outcomes. Benchmarks should therefore test whether systems can ground knowledge into perception, planning, execution, and recovery, rather than only answer questions or predict actions offline. Existing efforts study embodied task execution in simulation and household environments [123–126], embodied reasoning and agent

 doi:10.20944/preprints202606.0173.v1

evaluation [127,128], and embodied spatial intelligence with closed perception-action loops [129]. Closed-loop execution is especially important because open-loop prediction cannot fully capture compounding error, contact dynamics, safety failures, physical constraints, or recovery behavior. Benchmarks should evaluate task completion, failure modes, intervention counts, robustness to perturbations, and generalization across objects, scenes, instructions, initial states, and embodiments. Recent efforts support cross-task policy learning [130–132], deployment-oriented action evaluation [118], perception-action loop evaluation [129], physical reasoning [133], large-scale simulation and household evaluation [134–137], and automated task and data generation for sim-to-real policy learning [138]. However, real-robot evaluation remains necessary because implementation details can substantially affect performance. Overall, progress in Physical AI requires system-centric evaluation, where systems are compared by what they can reliably do in the world, not only by what they can predict from static inputs.

7. Challenges and Future Directions

Despite rapid progress, current systems remain far from general Physical AI. We summarize four challenges along the roadmap , with additional discussion in Appendix A.8.

7.1. From Implicit World Knowledge to Dense Physical Grounding.

LLMs encode broad but implicit and language-mediated world knowledge, which is difficult to convert into metric physical states [6,7]. Future systems must ground semantic, procedural, and causal priors into geometry, motion, contact, force, uncertainty, and temporal dynamics.

7.2. From Multimodal Grounding to Physical Perception.

VLMs and MLLMs ground language into images and videos, but semantic descriptions alone are insufficient for Physical AI [12,50]. They must move toward spatial, temporal, affordance-aware, and quantitative physical perception that supports action and prediction.

7.3. From VLAs to Generalist Embodied Policies.

VLAs connect perception and language to actions, yet remain limited by embodiment-specific action spaces, scarce robot data, and brittle cross-task generalization [17,18]. Future work should develop scalable action representations, cross-embodiment transfer, and policies augmented with memory, world models, or agentic components.

7.4. From World Models to Deployable Physical AI.

Video generation is not necessarily world modeling: Physical AI requires action-conditioned, temporally consistent, controllable, and physically plausible prediction [20,56,139]. World models must also balance efficiency and fidelity, since pixel-space simulators are costly while latent models often need task heads, decoders, or policy interfaces for action. Deployment further requires sim-to-real transfer, robustness to noise and latency, safety, closed-loop recovery, and reproducible evaluation [128].

8. Conclusion

This survey reframes Physical AI through the lens of LLM-based world knowledge, moving beyond purely robotics-centric, vision-centric, or cyber-physical views. We organize recent progress as a roadmap from language-derived priors to multimodal grounding, action grounding, world modeling, policy learning, and embodied agents. This perspective clarifies the complementary roles of LLMs and world models: LLMs provide semantic, commonsense, procedural, and causal priors, while world models provide predictive and simulative mechanisms for physical dynamics. By connecting language, multimodality, VLAs, world models, and embodied systems, this roadmap highlights the interfaces needed for more general Physical AI.

 doi:10.20944/preprints202606.0173.v1

9. Limitations

This survey focuses on a language-centered roadmap from LLM-based world knowledge to Physical AI. It is therefore not an exhaustive survey of all robotics, control, simulation, tactile sensing, audio perception, or cyber-physical systems. We emphasize LLMs, VLMs/MLLMs, VLAs, world models, and embodied agents because they form the main pathway from language-derived world knowledge to physical AI. Other important physical modalities, such as tactile sensing, force feedback, audio, material properties, mass estimation, and fluid or deformable-object dynamics, are discussed only when they directly relate to the roadmap. A broader Physical AI survey could extend this work by covering these modalities and domain-specific robotic systems in greater depth.

10. Broader Impact

The definition of Physical AI has evolved with the progress of artificial intelligence. Earlier discussions often framed Physical AI through specific tasks, modalities, or systems, such as computer vision, robotics, embodied agents, autonomous driving, or cyber-physical systems. While these perspectives are important, they can make Physical AI appear as a collection of downstream applications rather than a unified research problem. This survey provides a complementary perspective by organizing Physical AI around *LLM-based world knowledge*: the implicit semantic, commonsense, procedural, and causal priors encoded in large language models, and their grounding into perception, action, prediction, simulation, and embodied deployment.

The broader impact of this work is therefore mainly conceptual and organizational. By treating world knowledge as a central lens, the survey helps identify the underlying mechanisms that connect recent advances in LLMs, VLMs/MLLMs, VLAs, world models, policy learning, and embodied agents. This perspective clarifies why language models are useful for Physical AI, why language-mediated knowledge is insufficient for dense physical states, and why world models and closed-loop evaluation are necessary for moving from reasoning to reliable action in the physical world. We hope this roadmap can help researchers compare existing approaches, locate missing components, and develop future Physical AI systems that are more grounded, predictive, generalizable, and deployable.

At the same time, Physical AI may affect safety-critical domains, including robotics, autonomous driving, embodied assistants, industrial automation, and human–robot interaction. Misgrounded language priors, hallucinated plans, physically inconsistent world models, brittle action policies, or insufficient closed-loop evaluation may lead to unsafe behavior in real or simulated environments. For this reason, our survey emphasizes limitations and open challenges such as dense physical grounding, language-to-action reliability, physical consistency, sim-to-real transfer, safety, reproducibility, and transparent evaluation of closed or partially disclosed frontier systems. This work does not release new models, datasets, robot policies, controllers, or deployment systems; its impact is limited to taxonomy, analysis, and research guidance.

Appendix A. Supplementary Material

This appendix expands the roadmap and taxonomy used in the main paper. The main text focuses on the conceptual transition from LLM-based world knowledge to Physical AI under the page limit, while this appendix provides the additional context needed to make the roadmap verifiable and less table-only. In particular, we clarify the boundaries of the survey, contrast our organizing lens with existing survey perspectives, expand the taxonomy of roadmap stages, provide a more detailed world-model taxonomy, summarize evaluation protocols, list representative frontier systems, and discuss failure modes that motivate the challenges in the main paper.

The supplementary material is designed to serve two purposes. First, it makes explicit what is included and excluded in our survey. Since Physical AI overlaps with robotics, control, simulation, embodied AI, cyber-physical systems, multimodal learning, and model-based reinforcement learning, an unrestricted survey would be too broad and would obscure our central contribution. We therefore define the scope around the pathway from LLM-based world knowledge to grounded perception,

 doi:10.20944/preprints202606.0173.v1

grounded action, predictive world modeling, policy learning, and embodied deployment. Second, it provides additional references and categorizations that are not central enough to fit into the eight-page main paper but are useful for readers who want to trace the roadmap in more detail.

Across all tables, we use the same organizing principle: each component is described by its representational interface, its role in Physical AI, and its limitations. This makes the appendix complementary to the main paper rather than a separate literature catalogue. The tables are not intended to rank methods or claim that the listed systems are exhaustive. Instead, they provide representative anchors for the conceptual categories used throughout the survey.

Appendix A.1. Survey Scope and Boundary

Our survey is not intended to be an exhaustive review of all robotics, control, simulation, or cyber-physical systems. Instead, it studies Physical AI through the roadmap from LLM-based world knowledge to multimodal grounding, action grounding, world modeling, policy learning, and embodied deployment. This distinction is important because the term Physical AI is increasingly used across different communities with different assumptions: robotics work often emphasizes embodiment and control, vision work often emphasizes physical perception and generation, cyber-physical work often emphasizes deployment and sensing infrastructure, while language-centered work emphasizes reasoning, grounding, and agentic coordination.

The scope of this survey is therefore defined by whether a line of work contributes to the grounding of world knowledge into physical perception, prediction, planning, and action. For example, we include VLMs and MLLMs when they support spatial, temporal, or affordance grounding, but we do not attempt to cover all image captioning or visual question answering systems. Similarly, we include world models when they support prediction, simulation, planning, or policy learning for Physical AI, but we do not attempt to cover all video generation or all model-based reinforcement learning. This scope boundary is intended to reduce ambiguity for readers and reviewers: the paper is a roadmap survey centered on LLM-derived world knowledge, not a comprehensive encyclopedia of all physical intelligence.

Table A1 summarizes the intended scope. The middle column lists the categories we treat as part of the roadmap, while the right column identifies neighboring areas that are related but not exhaustively reviewed. This boundary also explains why some classical robotics, control, hardware, tactile sensing, and simulation topics are discussed only when they directly interact with foundation-model-based grounding or predictive modeling.

 doi:10.20944/preprints202606.0173.v1

Table A1. Scope boundary of this survey. We organize Physical AI as a roadmap from LLM-based world knowledge to grounded perception, action, world modeling, and embodied deployment, rather than as an exhaustive survey of all robotics or physical intelligence.

Roadmap Component	Included in This Survey	Not Exhaustively Covered
LLM-based world knowledge	Semantic, commonsense, procedural, causal, spatial, and affordance priors encoded in LLMs [6–9,140]	General factual recall, knowledge editing, or memory analysis unrelated to physical reasoning
Multimodal grounding	VLMs/MLLMs that ground language-derived knowledge into images, videos, regions, objects, spatial relations, and affordances [2,10–12,50]	Generic captioning, VQA, or multimodal dialogue not tied to physical grounding or interaction
Action grounding	VLA models, action representations, policy learning, and language-conditioned embodied control [16–19,81]	Classical robot control, motion planning, or manipulation methods without foundation-model grounding
World models	Video, latent, interactive, and action-conditioned models that support prediction, simulation, planning, or policy learning [20,56,95,96,98]	All video generation, all simulators, or all model-based RL methods outside the Physical AI roadmap
Embodied systems	Systems that close the loop between perception, planning, execution, recovery, and evaluation [88,107,108,141]	Hardware-specific robot engineering, robot design, and domain-specific control stacks

The table should be read as a boundary rather than a separation. Many excluded areas remain important to Physical AI, but they are not the organizing focus of this paper. For instance, low-level manipulation control and hardware design are indispensable for deployment, yet our discussion emphasizes how foundation models and world models interface with such systems. Likewise, video generation is relevant when it becomes temporally consistent, controllable, and action-conditioned, but generic video synthesis is not equivalent to physical world modeling.

Appendix A.2. Comparison with Existing Survey Perspectives

Existing survey perspectives cover important parts of the Physical AI landscape, but they usually begin from different assumptions. Broad Physical AI or PAI surveys often define the field from cyber-physical systems, robotics, sensing, and industrial deployment. Vision-centric generative Physical AI surveys emphasize physically grounded visual generation, physics-aware simulation, and visual understanding. VLA and robot foundation model studies focus on action spaces, robot policies, demonstrations, and embodiment-specific control. World-model-centered studies emphasize dynamics prediction, latent imagination, model-based planning, and simulation.

Our survey is complementary to these lines, but it starts from a different question: how can world knowledge encoded in LLMs be progressively grounded into perception, action, prediction, and deployment? This LLM-centered lens matters because many recent Physical AI systems use language models not merely as text interfaces, but as sources of semantic priors, task decompositions, commonsense constraints, tool orchestration, and agentic reasoning. At the same time, LLMs cannot model dense physical dynamics by themselves, which motivates the later transition toward VLA models and world models.

Table A2 clarifies this distinction. The goal of the comparison is not to claim that prior surveys are incomplete in their own scope. Rather, it shows that their organizing axes differ from ours. By making LLM-based world knowledge explicit, our survey connects the NLP and multimodal reasoning literature to Physical AI in a way that is not captured by purely robotics-centric, vision-centric, or world-model-only discussions.

 doi:10.20944/preprints202606.0173.v1

Table A2. Comparison with existing perspectives. Our survey is distinguished by using LLM-based world knowledge as the organizing lens and connecting it to multimodal grounding, VLA-style action interfaces, world models, policy learning, and deployable Physical AI systems.

Perspective	Main Focus	LLM World Knowledge	VLA / Action Grounding	World Models	Closed Systems
Broad Physical AI / PAI surveys [142, 143]	Concepts, applications, industrial systems, and cyber-physical perspectives	Limited	Partial	Limited	Partial
Vision-centric Generative Physical AI [57]	Physics-aware generation, visual simulation, and physically grounded computer vision	Limited	Limited	Partial	Partial
VLA / robot foundation model studies [18,144, 145]	Robot policies, action representations, and embodied control	Partial	Strong	Limited	Partial
World-model-centered studies [20,56, 96,98]	Prediction, latent dynamics, model-based planning, simulation, and policy learning	Limited	Partial	Strong	Partial
Ours	Roadmap from LLM-based world knowledge to Physical AI	Strong	Strong	Strong	Strong

The comparison also motivates why a roadmap structure is more appropriate than a flat taxonomy. A flat taxonomy would list LLMs, VLMs, VLAs, world models, and agents as independent families. Our view instead treats them as progressively more physically grounded interfaces: language priors are grounded into perception, perception and language are grounded into action, and action must ultimately be supported by predictive models and closed-loop deployment.

Appendix A.3. Extended Roadmap Taxonomy

The roadmap in the main paper compresses a large amount of literature into a small number of stages. Table A3 expands this roadmap by identifying the dominant representation at each stage, its role in Physical AI, and representative works. The table is intended to make explicit the hidden continuity between fields that are often discussed separately: NLP world knowledge, multimodal representation learning, robot action modeling, model-based prediction, policy learning, and embodied deployment.

A key design choice in this taxonomy is to organize methods by the interface through which knowledge becomes physically useful. LLM-based world knowledge is primarily textual and parametric. Multimodal grounding introduces visual and spatial representations. Action grounding introduces action tokens, trajectories, chunks, skills, and continuous controls. World modeling introduces future states, latent dynamics, rewards, values, or action-conditioned transitions. Policy learning then con-

 doi:10.20944/preprints202606.0173.v1

verts these representations into executable behavior, and embodied deployment tests whether the full stack can operate under feedback, noise, and real-world constraints.

This taxonomy also explains why no single model family currently solves Physical AI. LLMs provide broad priors but not dense physical state. VLMs and MLLMs ground perception but often remain language-mediated. VLAs provide an action-facing interface but struggle with cross-embodiment generalization and long-horizon prediction. World models provide predictive and simulative mechanisms but must still be connected to semantic goals, action interfaces, and reliable policies. Physical AI therefore emerges from the composition of these layers rather than from any one layer alone.

Table A3. Extended taxonomy of the roadmap from LLM-based world knowledge to Physical AI.

Stage	Main Representation	Role in Physical AI	Representative Works
LLM-based world knowledge	Textual and parametric knowledge	Provides semantic, commonsense, procedural, causal, spatial, and affordance priors	LAMA, closed-book QA, factual recall, procedural knowledge [6–8,140]
Multimodal grounding	Image/video-language representations	Grounds world knowledge into objects, scenes, spatial relations, temporal events, and affordances	CLIP, Flamingo, BLIP-2, LLaVA, Gemini [2,10–12,50]
Action grounding	Action tokens, trajectories, chunks, skills, or continuous controls	Maps perception and language instructions to executable actions	PaLM-E, RT-2, OpenVLA, π_0, $\pi_{0.5}$ [16–19,81]
World modeling	Future pixels, latent states, rewards, values, or action-conditioned transitions	Predicts and simulates possible futures for planning, policy learning, and counterfactual reasoning	World Models, Dreamer, MuZero, Genie, Cosmos, V-JEPA [20,56,96,98,101,103]
Policy learning	Learned policies, action experts, diffusion/flow policies, or controllers	Converts perception, reasoning, and prediction into behavior	ACT, FAST, RDT-1B, GR00T N1 [79,80,83,88]
Embodied deployment	Closed-loop systems with sensing, planning, execution, verification, and recovery	Tests whether models can reliably act in real or interactive environments	Gemini Robotics, RoboCasa, LIBERO, EmbodiedBench [107, 108,128,130,135]

The representative works in the last column are selected as anchors rather than exhaustive lists. Many systems occupy multiple stages: for instance, a VLA system may combine multimodal grounding, action tokenization, policy learning, and real-world evaluation. We place each representative work according to its most salient role in the roadmap, while acknowledging that frontier systems increasingly blur these boundaries.

Appendix A.4. Extended Taxonomy of World Models

World models are used differently across reinforcement learning, video generation, robotics, autonomous driving, and Physical AI. In model-based reinforcement learning, a world model often refers to a learned transition or reward model used for planning. In video generation, the term is increasingly used for models that generate plausible future frames or interactive visual environments. In robotics and embodied AI, a world model should support action-conditioned prediction, counterfactual reasoning, recovery, and policy learning. These meanings overlap but are not identical.

To avoid treating all generative video models or all model-based policies as the same type of world model, Table A4 organizes world models by prediction target and function in the Physical AI roadmap. This organization is important for the main paper's argument: world models are the stage

 doi:10.20944/preprints202606.0173.v1

where Physical AI begins to move beyond language-mediated priors and toward directly learned predictive or simulative knowledge about physical dynamics.

Table A4. Extended taxonomy of world models for Physical AI. The categories are organized by prediction target and their function in the roadmap from LLM-based world knowledge to embodied action.

Category	Prediction Target	Role in Physical AI	Representative Works
Classical / model-based RL world models	Future states, rewards, values, or policy-relevant quantities	Planning, latent imagination, decision making, and policy improvement	World Models, PlaNet, Dreamer, DreamerV3, MuZero [20,95–98]
Video-space world models	Future pixels, frames, or video tokens	Visual imagination, future scene prediction, synthetic data, and simulated experience	GAIA-1, UniSim, Genie, Cosmos [56,99–101]
Latent / representation-space world models	Future latent states, embeddings, or masked spatiotemporal representations	Efficient long-horizon prediction, compact planning, and control-relevant representation learning	PlaNet, Dreamer, I-JEPA, V-JEPA, V-JEPA 2 [95,96,102–104]
Interactive / action-conditioned world models	Future observations or latent states conditioned on candidate actions	Counterfactual reasoning, simulation-based policy learning, safety evaluation, and recovery	MuZero, UniSim, Genie, GAIA-1, V-JEPA 2, Cosmos [56,98–101,104]
World foundation models for Physical AI	General-purpose predictive or generative world representations	Adaptable substrate for robotics, autonomous driving, embodied agents, and synthetic data	Cosmos, Genie-style models, V-JEPA-style models [56,101,103,104]

Appendix A.4.1. Classical and Decision-Centric World Models.

Classical world models are rooted in model-based reinforcement learning and planning. They learn transition, reward, value, or policy-relevant predictions that allow agents to plan before acting. World Models, PlaNet, Dreamer, DreamerV3, and MuZero establish this foundation by showing that agents can learn compact internal models and use them for imagination, planning, and policy improvement [20,95–98]. For Physical AI, this line provides the decision-making substrate: agents should not only react to observations, but also evaluate possible futures.

Appendix A.4.2. Video-Space World Models.

Video-space world models predict future visual observations. They are attractive for Physical AI because videos expose motion, temporal evolution, scene changes, and possible future outcomes. GAIA-1 models autonomous-driving futures from video, text, and action inputs [99]; UniSim learns an interactive real-world simulator from heterogeneous data and uses it for policy training [100]; Genie learns generative interactive environments from unlabelled videos [101]; and Cosmos positions world foundation models as adaptable world models for Physical AI [56]. The main limitation is that visual realism does not guarantee physical correctness. A generated rollout may look plausible while violating object permanence, contact constraints, controllability, or causal consistency.

Appendix A.4.3. Latent and Representation-Space World Models.

Latent world models predict in compact representation spaces rather than pixel space. This makes them more efficient for planning and control because they can focus on task-relevant dynamics instead of reconstructing every visual detail. JEPA-style models further argue that predictive modeling should happen in representation space rather than through full generative reconstruction [21,102,103]. V-JEPA

 doi:10.20944/preprints202606.0173.v1

2 extends this idea to video-scale learning and post-trains an action-conditioned latent world model for robot planning [104]. For Physical AI, latent prediction is especially useful when the agent needs fast rollouts, uncertainty-aware planning, or long-horizon reasoning under limited computational budget.

Appendix A.4.4. Interactive and Action-Conditioned World Models.

Physical AI requires models that respond to actions, not only models that passively predict future frames. Interactive and action-conditioned world models estimate counterfactual futures under candidate actions, enabling planning, safety checking, policy learning, and recovery. This requirement separates physical world models from generic video generators: a Physical AI world model should be controllable, temporally consistent, action-conditioned, and useful for closed-loop decision making. Such models also make it possible to evaluate actions before executing them in the real world, reducing reliance on costly or unsafe trial-and-error deployment.

Appendix A.4.5. Relation to LLM-Based World Knowledge.

LLM-based world knowledge and world models are complementary. LLMs provide semantic, commonsense, procedural, and causal priors; world models provide predictive and simulative mechanisms for physical dynamics. The former tells an agent what actions may be meaningful; the latter estimates what is likely to happen if the agent acts. This complementarity explains why world models occupy a central position in the roadmap from LLM-based world knowledge to deployable Physical AI. In practice, future systems may combine LLMs for high-level goals, instructions, and commonsense constraints with world models for action-conditioned rollout, physical feasibility checking, and policy optimization.

Appendix A.5. Benchmarks and Evaluation Protocols

Evaluation is a central difficulty for Physical AI because the roadmap spans several different kinds of competence. Static language benchmarks can test whether a model encodes commonsense or procedural knowledge, but they cannot determine whether that knowledge is grounded in a physical state. Vision-language benchmarks can test perception and grounding, but they often stop at recognition or description. VLA benchmarks can test whether actions are predicted from observations and instructions, but open-loop action accuracy does not fully capture closed-loop execution. World-model benchmarks can test prediction, but prediction quality must ultimately be judged by whether it supports planning and control.

Table A5 summarizes evaluation protocols along the roadmap. The key shift is from static recognition or offline prediction to closed-loop task completion, robustness, recovery, and cross-embodiment generalization. This is aligned with the main paper's argument that Physical AI should be evaluated by what a system can reliably do in the world, not only by what it can answer or predict from fixed inputs.

 doi:10.20944/preprints202606.0173.v1

Table A5. Evaluation protocols along the roadmap. Physical AI evaluation should shift from static recognition or offline prediction to closed-loop task completion, robustness, safety, recovery, and cross-embodiment generalization.

Roadmap Stage	Benchmark / Evaluation Type	What to Evaluate	Representative Works
LLM world knowledge	Physical commonsense, tool understanding, factual/procedural knowledge	Whether LLMs encode usable semantic, commonsense, procedural, and causal priors	PHYBench, PhySense, PhysToolBench [44,45,49]
VLM/MLLM grounding	Spatial, temporal, affordance, and physical reasoning benchmarks	Whether language-derived knowledge is grounded into perception, spatial relations, and physical states	BLINK, Video-MME, PhysBench, QuantiPhy, MASS-Bench [72,75–78]
VLA / action grounding	Robot manipulation, navigation, and action-prediction benchmarks	Whether models can map instructions and observations to executable actions	RT-2, OpenVLA, LIBERO, LIBERO-Pro [17,18,130,132]
World models	Video prediction, latent prediction, action-conditioned simulation, planning evaluation	Whether models predict physically plausible, controllable, temporally consistent futures	World Models, Dreamer, Genie, Cosmos, V-JEPA 2 [20,56,96,101,104]
Embodied agents	Closed-loop simulated or real-world tasks	Whether systems complete tasks, recover from errors, and generalize across environments and embodiments	BEHAVIOR, EAI, EmbodiedBench, RoboSuite, RoboCasa [123,127,128, 134,135]
Closed frontier systems	Black-box or product-level evaluation	Capability, reliability, reproducibility, safety, and transparency under limited disclosure	Gemini Robotics, Gemini Robotics 1.5, GR00T N1, π-series systems [19,81, 88,107,108,141]

A useful evaluation suite should therefore include both stage-specific and system-level metrics. Stage-specific metrics diagnose where a system fails: factual or physical commonsense, perceptual grounding, action prediction, world-model rollout, or closed-loop execution. System-level metrics evaluate whether these components work together under deployment constraints. For example, a strong VLA may still fail if its actions accumulate error, if its world model produces visually plausible but physically inconsistent futures, or if its controller cannot recover from perturbations. This is why task success, intervention count, robustness, safety, and recovery should be reported alongside conventional accuracy or prediction metrics.

Appendix A.6. Frontier Systems and Closed Models

Many influential Physical AI systems are released as frontier products, platforms, or partially documented technical reports rather than fully open academic artifacts. This creates a gap between real-world usage and academic evaluation. Closed or partially disclosed systems may demonstrate important capabilities, shape the terminology of the field, and influence user expectations, but their training data, architecture details, evaluation protocols, and failure cases are often unavailable.

Table A6 summarizes representative examples and their roles in the roadmap. We include them not as endorsements or as exhaustive comparisons, but because they represent the kinds of systems that motivate black-box evaluation, product-level benchmarking, and reproducibility discussions. In a

 doi:10.20944/preprints202606.0173.v1

survey of Physical AI, ignoring such systems would leave out a major part of the current landscape; however, treating them like fully open academic models would also be misleading. We therefore categorize them by role and openness.

Table A6. Representative frontier systems and closed or partially disclosed models. These systems motivate product-level and black-box evaluation protocols in addition to conventional academic benchmarks.

System	Category	Openness / Citation Type	Role in the Roadmap
GPT-4 / ChatGPT-style agents [1]	LLM / agentic assistant	Closed / technical report or product documentation	High-level world knowledge, planning, tool use, and task decomposition
Claude-style assistants [3]	LLM / agentic assistant	Closed / product documentation	Reasoning, tool use, coding, and agentic orchestration
Gemini Robotics and Gemini Robotics 1.5 [107,108]	Robotics foundation model	Closed or partially disclosed / technical report	Multimodal reasoning, embodied control, and real-world robot interaction
Cosmos [56]	World foundation model platform	Partially open / technical report	World modeling, synthetic data, simulation, autonomous driving, and robotics
GR00T N1 [88]	Humanoid foundation model	Partially open / technical report	Generalist humanoid policies and cross-embodiment action learning
π-series systems [19,81,141]	Generalist VLA / robot foundation models	Partially disclosed / technical reports	Action grounding, open-world generalization, policy learning, and embodied deployment

The main challenge posed by closed systems is not only that they are difficult to reproduce. It is also that they may combine several roadmap stages into a single product-level stack, making ablation and attribution difficult. A system may appear to have strong physical reasoning because of its LLM prior, its perception module, its action policy, its retrieval system, its simulator, or its human-feedback pipeline. Without transparent interfaces and standardized black-box tests, it is difficult to identify which component contributes to success or failure. This motivates evaluation protocols that separate capability testing, robustness testing, safety testing, and reproducibility reporting.

Appendix A.7. Failure Modes Along the Roadmap

The roadmap is useful not only because it organizes progress, but also because it localizes failures. A Physical AI system may fail at the level of world knowledge, perception, action, prediction, policy learning, deployment, or evaluation. These failures are qualitatively different. An LLM hallucination produces a plausible but ungrounded plan; a VLM grounding failure misidentifies the state of the world; a VLA failure maps a correct goal to the wrong action; a world-model failure predicts an implausible future; and a deployment failure can arise from sensing, latency, calibration, or controller mismatch.

Table A7 summarizes representative failure modes. These failures motivate our deployment-oriented discussion of challenges in the main paper. The table also clarifies why Physical AI cannot be evaluated by a single benchmark: each roadmap stage requires different diagnostics, and end-to-end task success alone may hide the source of failure.

 doi:10.20944/preprints202606.0173.v1

Table A7. Representative failure modes along the roadmap from LLM-based world knowledge to Physical AI.

Component	Typical Failure Mode	Why It Matters for Physical AI
LLMs	Hallucinated or ungrounded physical knowledge; overconfident plans; missing metric state	The model may propose plausible language plans that violate geometry, contact, force, or object-state constraints
VLMs / MLLMs	Correct semantic description but weak dense grounding	The model may identify objects but fail to estimate pose, depth, uncertainty, reachability, or action-conditioned dynamics
VLAs	Poor cross-embodiment generalization; data-dependent policies; brittle recovery	The same instruction may require different grasps, trajectories, or control strategies across robots and environments
World models	Visually plausible but physically inconsistent futures	Photorealistic generation may still violate object permanence, contact, gravity, controllability, or causal dynamics
Policy learning	Offline success but closed-loop failure	A model may predict correct actions under dataset states but fail under compounding errors or real-time perturbations
Embodied systems	Sensor, calibration, latency, controller, or hardware failures	Physical performance depends on the full system stack, not only model accuracy
Closed frontier systems	Limited reproducibility and incomplete disclosure	Strong product-level systems can shape the field while being difficult to benchmark, ablate, or compare fairly

The failure-mode view also suggests a practical debugging strategy. If a system fails before action, the issue may lie in world knowledge or perceptual grounding. If it fails during action, the issue may lie in action representation, embodiment transfer, or controller design. If it fails after several steps, the issue may lie in world modeling, compounding error, memory, or recovery. If it succeeds in simulation but fails in the real world, the issue may lie in sim-to-real transfer, sensing, calibration, latency, or hidden deployment assumptions. This decomposition turns the roadmap into an evaluation tool rather than just a taxonomy.

Appendix A.8. Extended Discussion of Challenges and Future Directions

The main paper summarizes the challenges along the roadmap with a small number of representative citations. Here we provide a more detailed discussion of the evidence behind each challenge and connect it to related work. The central point is that each stage in the roadmap exposes a different interface mismatch: LLMs expose world knowledge through sparse language; VLMs ground language into perception but often remain semantic; VLAs output actions but are tied to embodiment-specific action spaces; world models provide prediction but must be controllable and physically faithful; deployed systems must integrate all components under closed-loop constraints.

Appendix A.8.1. Implicit World Knowledge and Dense Physical Grounding.

LLM-based world knowledge is broad, but it is not an explicit symbolic database. It is stored as parametric regularities and exposed through prompting, context, plans, programs, or tool calls. Knowledge-probing and closed-book QA studies show that language models can store factual and relational knowledge in parameters [6,7], while later studies analyze long-tail factual acquisition, factual recall, and pretraining dynamics [8,9,32]. Materialization and mechanistic analyses further show that parts of such knowledge can be extracted or traced through model computations [33,146,147]. For Physical AI, however, these priors must be converted into dense physical variables such as pose,

 doi:10.20944/preprints202606.0173.v1

reachability, contact, force, friction, uncertainty, and temporal dynamics. Procedural knowledge and task-level planning provide useful priors [34,35,37–41,140], but they remain insufficient without perceptual grounding and physical verification.

Appendix A.8.2. Multimodal Grounding and Physically Faithful Perception.

VLMs and MLLMs provide the first major bridge from language-mediated priors to perceptual observations. Contrastive and multimodal pretraining align images or videos with language [2,10–12,50], enabling models to connect objects, scenes, and instructions to visual inputs. However, Physical AI requires more than captioning or visual QA. It requires spatial grounding, temporal grounding, affordance estimation, quantitative reasoning, and dense frame-level understanding. Recent benchmarks and evaluations expose gaps in low-level visual perception, long-video understanding, physical reasoning, quantitative physics, and motion-aware spatiotemporal grounding [72,75–78]. Related work on dense physical perception and intermediate-feature grounding suggests that VLM representations may be reused for downstream Physical AI tasks, but their outputs must be transformed into action-relevant or prediction-relevant representations [14,15,144,148].

Appendix A.8.3. VLA Generalization and Action-Interface Bottlenecks.

VLA models connect visual observations and language instructions to action outputs, making them a key interface between multimodal reasoning and embodied control. Representative systems such as PaLM-E, RT-2, OpenVLA, and the π-series demonstrate the promise of transferring web-scale or foundation-model knowledge into robotic action [16–19,81]. Nevertheless, VLA policies face three persistent bottlenecks. First, action spaces vary across embodiments, including action tokens, end-effector poses, trajectories, action chunks, and continuous controls. Second, robot data remain much smaller and more heterogeneous than language or vision data. Third, imitation-trained policies can be brittle under distribution shift and may lack recovery behavior. Recent work on VLA learning, action abstraction, and generalist robot policies explores scalable action representations, data mixtures, and richer policy architectures [79,80,83,88,141,144,145]. A promising direction is to augment VLA policies with memory, LLM agents, or world models so that policies do not only map observations to actions, but also reason over goals, histories, and possible futures.

Appendix A.8.4. World Models Beyond Video Generation.

World models provide the predictive and simulative substrate that VLAs often lack. Classical and decision-centric world models learn transition, reward, value, or policy-relevant quantities for planning and policy improvement [20,95–98]. Video-space world models generate or predict future visual observations and can support visual imagination, synthetic data, and interactive simulation [56,99–101]. Latent and JEPA-style world models instead predict in representation space, trading pixel-level reconstruction for efficiency and planning-relevant abstraction [21,102–104]. This distinction is important: photorealistic generation does not guarantee physical correctness, while latent prediction may be efficient but difficult to interpret or use directly without task heads, decoders, or policy interfaces. For Physical AI, a useful world model should be temporally consistent, action-conditioned, controllable, physically plausible, and useful for planning or closed-loop decision making. Recent work on physically grounded world-model evaluation and language-guided latent prediction provides early steps toward this goal [139,148].

Appendix A.8.5. Deployment and System-Level Evaluation.

Even strong models can fail when deployed as Physical AI systems. World models and simulators can support training and planning, but real-world deployment introduces sensor noise, latency, calibration errors, embodiment mismatch, sim-to-real transfer, safety constraints, and recovery requirements. Robotics and embodied benchmarks such as LIBERO, RoboCasa, BEHAVIOR, RoboSuite, and EmbodiedBench evaluate different parts of this system-level challenge [123,128,130,134,135]. Frontier

 doi:10.20944/preprints202606.0173.v1

systems such as Gemini Robotics, GR00T N1, and π-series models further show that Physical AI is becoming a product-level systems category, but many such systems are closed or only partially disclosed [19,81,88,107,108,141]. Future evaluation should therefore report closed-loop task completion, recovery, intervention counts, robustness, safety, reproducibility, and cross-embodiment generalization rather than relying only on static model-level metrics.

Table A8. Extended analysis of challenges and future directions along the roadmap. The main paper summarizes these challenges with a small number of representative citations; this table provides additional evidence and connects each challenge to the corresponding interface mismatch.

Challenge	Interface Mismatch	Future Direction	Representative Evidence
Implicit world knowledge	LLM priors are language-mediated and difficult to convert into metric physical state	Extract, align, and ground semantic, procedural, and causal priors into dense physical representations	Knowledge probing, factual recall, procedural knowledge, language planning [6–8,37,140]
Physical perception	VLMs often output semantic descriptions rather than dense physical state	Move toward spatial, temporal, affordance, quantitative, and action-relevant grounding	VLM/MLLM grounding and physical reasoning benchmarks [12,50,72,75–77]
Generalist VLA policies	Actions are embodiment-specific and robot data are limited	Develop scalable action representations, cross-embodiment transfer, and policies augmented with memory or world models	PaLM-E, RT-2, OpenVLA, π-series, FAST, GR00T N1 [16–19,79,88]
Predictive world modeling	Video realism does not imply physical correctness; latent models may require task heads or policy interfaces	Build action-conditioned, controllable, efficient, and physically plausible world models	Dreamer, MuZero, Genie, UniSim, Cosmos, V-JEPA [56,96,98,100,101,103]
Deployment	Model-level accuracy does not capture sensing, control, latency, recovery, safety, or sim-to-real robustness	Evaluate integrated systems with closed-loop task completion, recovery, safety, and reproducibility	LIBERO, RoboCasa, EmbodiedBench, Gemini Robotics, GR00T N1 [88,107,128,130,135]

Appendix A.9. Terminology

We use several terms throughout the survey whose meanings vary across communities. Table A9 records the definitions used in this paper. These definitions are intentionally functional: they describe the role each concept plays in the roadmap rather than attempting to settle all terminology debates in Physical AI, robotics, or model-based learning.

In particular, we distinguish *world knowledge* from *world models*. World knowledge refers to implicit priors about objects, actions, environments, and likely consequences, often encoded in the parameters of LLMs and exposed through prompting or agentic reasoning. A world model, by contrast, is a predictive or simulative mechanism that estimates how observations, latent states, rewards, values, or action consequences evolve. This distinction is central to our argument: LLMs help an agent reason about what is meaningful or plausible, while world models help estimate what is likely to happen under physical dynamics.

 doi:10.20944/preprints202606.0173.v1

Table A9. Terminology used throughout the survey.

Term	Definition in This Survey
World knowledge	Semantic, commonsense, procedural, causal, spatial, and affordance priors about objects, agents, actions, environments, and likely consequences.
LLM-based world knowledge	Language-mediated world priors stored as parametric regularities in LLMs and exposed through prompting, context, or agentic reasoning.
World model	A predictive or simulative model that estimates future observations, latent states, rewards, values, or action consequences from current states and possible actions.
VLA model	A model that maps visual observations, language instructions, and sometimes embodiment states into executable actions or action-relevant representations.
Physical AI	AI systems that ground world knowledge into multimodal perception, physical prediction, simulation, planning, policy learning, and real-world or interactive action.

References

1. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
2. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* **2023**.
3. Anthropic. Claude 3.5 Sonnet Model Card Addendum, 2024.
4. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
5. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **2022**, *35*, 24824–24837.
6. Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A.H.; Riedel, S. Language Models as Knowledge Bases? In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 2463–2473.
7. Roberts, A.; Raffel, C.; Shazeer, N. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 5418–5426.
8. Kandpal, N.; Deng, H.; Roberts, A.; Wallace, E.; Raffel, C. Large Language Models Struggle to Learn Long-Tail Knowledge. In Proceedings of the Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 15696–15707.
9. Chang, H.; Park, J.; Ye, S.; Yang, S.; Seo, Y.; Chang, D.S.; Seo, M. How Do Large Language Models Acquire Factual Knowledge During Pretraining? In Proceedings of the Advances in Neural Information Processing Systems, 2024, Vol. 37.
10. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* **2022**, *35*, 23716–23736.
11. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International conference on machine learning. PMLR, 2023, pp. 19730–19742.
12. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Advances in neural information processing systems* **2023**, *36*, 34892–34916.
13. Lu, J.; Wang, H.; Xu, Y.; Wang, Y.; Yang, K.; Fu, Y. Representation Potentials of Foundation Models for Multimodal Alignment: A Survey. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2025, pp. 16669–16684.

 doi:10.20944/preprints202606.0173.v1

14. Zhang, H.; Lu, Y.; Wang, L.; Li, Y.; Chen, D.; Xu, Y.; Fu, Y. LinkedOut: Linking World Knowledge Representation Out of Video LLM for Next-Generation Video Recommendation. *arXiv preprint arXiv:2512.16891* **2025**.
15. Zhang, H.; Chai, W.; He, S.; Li, A.; Fu, Y. Dense video understanding with gated residual tokenization. *arXiv preprint arXiv:2509.14199* **2025**.
16. Driess, D.; Xia, F.; Sajjadi, M.S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. PaLM-E: an embodied multimodal language model. In Proceedings of the Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 8469–8488.
17. Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of the Conference on Robot Learning. PMLR, 2023, pp. 2165–2183.
18. Kim, M.J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.P.; Sanketi, P.R.; Vuong, Q.; et al. OpenVLA: An Open-Source Vision-Language-Action Model. In Proceedings of the Conference on Robot Learning. PMLR, 2025, pp. 2679–2713.
19. Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al. π_0: A Vision–Language–Action Flow Model for General Robot Control. *Robotics: Science and Systems XXI* **2025**.
20. Ha, D.; Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122* **2018**, *2*, 440.
21. LeCun, Y. A Path Towards Autonomous Machine Intelligence Version 0.9. 2, 2022-06-27 **2022**.
22. Arkin, R.C. Integrating behavioral, perceptual, and world knowledge in reactive navigation. *Robotics and autonomous systems* **1990**, *6*, 105–122.
23. Hagoort, P.; Hald, L.; Bastiaansen, M.; Petersson, K.M. Integration of word meaning and world knowledge in language comprehension. *science* **2004**, *304*, 438–441.
24. Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In Proceedings of the European conference on computer vision. Springer, 2022, pp. 146–162.
25. Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; Xia, F. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14455–14465.
26. Yuan, W.; Duan, J.; Blukis, V.; Pumacay, W.; Krishna, R.; Murali, A.; Mousavian, A.; Fox, D. RoboPoint: A Vision-Language Model for Spatial Affordance Prediction in Robotics. In Proceedings of the Proceedings of the 8th Conference on Robot Learning, 2025, Vol. 270, *Proceedings of Machine Learning Research*, pp. 4005–4020.
27. Song, C.H.; Blukis, V.; Tremblay, J.; Tyree, S.; Su, Y.; Birchfield, S. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 15768–15780.
28. Qian, S.; Chen, W.; Bai, M.; Zhou, X.; Tu, Z.; Li, L.E. AffordanceLLM: Grounding Affordance from Vision Language Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024, pp. 7587–7597.
29. Huang, S.; Ponomarenko, I.; Jiang, Z.; Li, X.; Hu, X.; Gao, P.; Li, H.; Dong, H. ManipVQA: Injecting Robotic Affordance and Physically Grounded Information into Multi-Modal Large Language Models. *arXiv preprint arXiv:2403.11289* **2024**.
30. Chu, H.; Deng, X.; Lv, Q.; Chen, X.; Li, Y.; Hao, J.; Nie, L. 3D-AffordanceLLM: Harnessing Large Language Models for Open-Vocabulary Affordance Detection in 3D Worlds. In Proceedings of the International Conference on Learning Representations, 2025.
31. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* **2021**.
32. Yuan, J.; Pan, L.; Hang, C.W.; Guo, J.; Jiang, J.; Min, B.; Ng, P.; Wang, Z. Towards a Holistic Evaluation of LLMs on Factual Knowledge Recall. *arXiv preprint arXiv:2404.16164* **2024**.
33. Hu, Y.; Nguyen, T.P.; Ghosh, S.; Razniewski, S. Enabling LLM Knowledge Analysis via Extensive Materialization. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 16189–16202.
34. Song, C.H.; Wu, J.; Washington, C.; Sadler, B.M.; Chao, W.L.; Su, Y. LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2998–3009.

 doi:10.20944/preprints202606.0173.v1

35. Rana, K.; Haviland, J.; Garg, S.; Abou-Chakra, J.; Reid, I.; Suenderhauf, N. SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning. In Proceedings of the Proceedings of the 7th Conference on Robot Learning, 2023, pp. 23–72.
36. Hua, P.; Liu, M.; Macaluso, A.; Lin, Y.; Zhang, W.; Xu, H.; Wang, L. GenSim2: Scaling Robot Data Generation with Multi-modal and Reasoning LLMs. In Proceedings of the Proceedings of the 8th Conference on Robot Learning, 2025, Vol. 270, *Proceedings of Machine Learning Research*, pp. 5030–5066.
37. Ichter, B.; Brohan, A.; Chebotar, Y.; Finn, C.; Hausman, K.; Herzog, A.; Ho, D.; Ibarz, J.; Irpan, A.; Jang, E.; et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In Proceedings of the Proceedings of the 6th Conference on Robot Learning, 2023, pp. 287–318.
38. Huang, W.; Xia, F.; Xiao, T.; Chan, H.; Liang, J.; Florence, P.; Zeng, A.; Tompson, J.; Mordatch, I.; Chebotar, Y.; et al. Inner Monologue: Embodied Reasoning through Planning with Language Models. In Proceedings of the Proceedings of the 6th Conference on Robot Learning, 2023, pp. 1769–1782.
39. Huang, W.; Abbeel, P.; Pathak, D.; Mordatch, I. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. In Proceedings of the Proceedings of the 39th International Conference on Machine Learning, 2022, pp. 9118–9147.
40. Liang, J.; Huang, W.; Xia, F.; Xu, P.; Hausman, K.; Ichter, B.; Florence, P.; Zeng, A. Code as policies: Language model programs for embodied control. In Proceedings of the 2023 IEEE International conference on robotics and automation (ICRA). IEEE, 2023, pp. 9493–9500.
41. Singh, I.; Blukis, V.; Mousavian, A.; Goyal, A.; Xu, D.; Tremblay, J.; Fox, D.; Thomason, J.; Garg, A. ProgPrompt: Program Generation for Situated Robot Task Planning using Large Language Models. *Autonomous Robots* **2023**, *47*, 999–1012.
42. Wang, G.; Xie, Y.; Jiang, Y.; et al. Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv preprint arXiv:2305.16291* **2023**.
43. Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; Fei-Fei, L. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In Proceedings of the Proceedings of the 7th Conference on Robot Learning, 2023, Vol. 229, *Proceedings of Machine Learning Research*, pp. 540–562.
44. Qiu, S.; Guo, S.; Song, Z.Y.; Sun, Y.; Cai, Z.; Wei, J.; Luo, T.; Yin, Y.; Zhang, H.; Hu, Y.; et al. PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems, 2025, Vol. 38.
45. Xu, Y.; Liu, Y.; Gao, Z.; Peng, C.; Luo, D. PhySense: Principle-Based Physics Reasoning Benchmarking for Large Language Models. *arXiv preprint arXiv:2505.24823* **2025**.
46. Kambhampati, S.; Valmeekam, K.; Guan, L.; Verma, M.; Stechly, K.; Bhambri, S.; Saldyt, L.; Murthy, A. LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning, 2024.
47. Valmeekam, K.; Stechly, K.; Kambhampati, S. LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's o1 on PlanBench. *arXiv preprint arXiv:2409.13373* **2024**.
48. Xiang, K.; Li, H.; Zhang, T.J.; Huang, Y.; Liu, Z.; Qu, P.; He, J.; Chen, J.; Yuan, Y.J.; Han, J.; et al. SeePhys: Does Seeing Help Thinking? Benchmarking Vision-Based Physics Reasoning. In Proceedings of the Advances in Neural Information Processing Systems, 2025, Vol. 38.
49. Zhang, Z.; Chen, K.; Lin, X.; Jiang, L.; Zheng, X.; Lyu, Y.; Guo, L.; Li, Y.; Chen, Y.C. PhysToolBench: Benchmarking Physical Tool Understanding for MLLMs. *arXiv preprint arXiv:2510.09507* **2025**.
50. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PmLR, 2021, pp. 8748–8763.
51. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 4904–4916.
52. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* **2019**, *32*.
53. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. In Proceedings of the Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 5100–5111.
54. Team, Q. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804* **2026**.

 doi:10.20944/preprints202606.0173.v1

55. Shen, Y.; Liu, Y.; Zhu, J.; Cao, X.; Zhang, X.; He, Y.; Ye, W.; Rehg, J.; Lourentzou, I. Fine-grained preference optimization improves spatial reasoning in vlms. *Advances in Neural Information Processing Systems* **2026**, *38*, 17929–17960.
56. Agarwal, N.; Ali, A.; Bala, M.; Balaji, Y.; Barker, E.; Cai, T.; Chattopadhyay, P.; Chen, Y.; Cui, Y.; Ding, Y.; et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575* **2025**.
57. Liu, D.; Zhang, J.; Dinh, A.D.; Park, E.; Zhang, S.; Mian, A.; Shah, M.; Xu, C. Generative physical ai in vision: A survey. *arXiv preprint arXiv:2501.10928* **2025**.
58. Shridhar, M.; Manuelli, L.; Fox, D. Cliport: What and where pathways for robotic manipulation. In Proceedings of the Conference on robot learning. PMLR, 2022, pp. 894–906.
59. Bear, D.M.; Wang, E.; Mrowca, D.; Binder, F.J.; Tung, H.Y.F.; Pramod, R.; Holdaway, C.; Tao, S.; Smith, K.; Sun, F.Y.; et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261* **2021**.
60. Rasheed, H.; Maaz, M.; Mullappilly, S.S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R.M.; Xing, E.; Yang, M.H.; Khan, F.S. GLaMM: Pixel Grounding Large Multimodal Model, 2024, [arXiv:cs.CV/2311.03356].
61. Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; Jia, J. Lisa: Reasoning segmentation via large language model. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 9579–9589.
62. Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; Yuan, L. Florence-2: Advancing a unified representation for a variety of vision tasks. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4818–4829.
63. Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J.S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 91–104.
64. Pothiraj, A.; Stengel-Eskin, E.; Cho, J.; Bansal, M. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 8001–8010.
65. Ren, S.; Yao, L.; Li, S.; Sun, X.; Hou, L. Timechat: A time-sensitive multimodal large language model for long video understanding. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14313–14323.
66. Guo, Y.; Liu, J.; Li, M.; Cheng, D.; Tang, X.; Sui, D.; Liu, Q.; Chen, X.; Zhao, K. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 3302–3310.
67. Wang, H.; Xu, Z.; Cheng, Y.; Diao, S.; Zhou, Y.; Cao, Y.; Wang, Q.; Ge, W.; Huang, L. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290* **2024**.
68. Zhang, H.; Fu, Y. Vqtoken: Neural discrete token representation learning for extreme token reduction in video large language models. *Advances in Neural Information Processing Systems* **2026**, *38*, 32851–32869.
69. Zeng, X.; Li, K.; Wang, C.; Li, X.; Jiang, T.; Yan, Z.; Li, S.; Shi, Y.; Yue, Z.; Wang, Y.; et al. Timesuite: Improving mllms for long video understanding via grounded tuning. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 38057–38081.
70. Munasinghe, S.; Gani, H.; Zhu, W.; Cao, J.; Xing, E.; Khan, F.S.; Khan, S. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 19036–19046.
71. Fu, C.; Yuan, H.; Dong, Y.; Zhang, Y.F.; Shen, Y.; Hu, X.; Li, X.; Su, J.; Long, C.; Xie, X.; et al. Video-MME-v2: Towards the Next Stage in Benchmarks for Comprehensive Video Understanding. *arXiv preprint arXiv:2604.05015* **2026**.
72. Chow, W.; Mao, J.; Li, B.; Seita, D.; Campagnolo Guizilini, V.; Wang, Y. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In Proceedings of the International Conference on Learning Representations, 2025, Vol. 2025, pp. 97959–98108.
73. Liu, S.C.; Chen, W.; Cheng, W.L.; Huang, Y.L.; Liao, I.B.; Li, Y.H.; Zhang, J.; et al. PAVLM: Advancing point cloud based affordance understanding via vision-language model. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2025, pp. 4299–4306.

 doi:10.20944/preprints202606.0173.v1

74. Liu, Y.; Zhu, J.; Mo, Y.; Li, G.; Cao, X.; Jin, J.; Shen, Y.; Li, Z.; Yu, T.; Yuan, W.; et al. PALM: Progress-Aware Policy Learning via Affordance Reasoning for Long-Horizon Robotic Manipulation. *arXiv preprint arXiv:2601.07060* **2026**.
75. Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N.A.; Ma, W.C.; Krishna, R. Blink: Multimodal large language models can see but not perceive. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 148–166.
76. Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2025, pp. 24108–24118.
77. Puyin, L.; Xiang, T.; Mao, E.; Wei, S.; Chen, X.; Masood, A.; Fei-Fei, L.; Adeli, E. QuantiPhy: A Quantitative Benchmark Evaluating Physical Reasoning Abilities of Vision-Language Models. *arXiv preprint arXiv:2512.19526* **2025**.
78. Wu, X.; Li, Z.; Jin, J.; Shi, G.; KV, G.; Raj, V.; Sinha, N.; Chen, J.; Du, F.; Manocha, D. MASS: Motion-Aware Spatial-Temporal Grounding for Physics Reasoning and Comprehension in Vision-Language Models. *arXiv preprint arXiv:2511.18373* **2025**.
79. Pertsch, K.; Stachowicz, K.; Ichter, B.; Driess, D.; Nair, S.; Vuong, Q.; Mees, O.; Finn, C.; Levine, S. FAST: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747* **2025**.
80. Zhao, T.Z.; Kumar, V.; Levine, S.; Finn, C. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. *arXiv preprint arXiv:2304.13705* **2023**.
81. Physical Intelligence.; Black, K.; Brown, N.; Darpinian, J.; Dhabalia, K.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; et al. $\pi_{0.5}$: A Vision–Language–Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054* **2025**.
82. Wen, J.; Zhu, Y.; Li, J.; Tang, Z.; Shen, C.; Feng, F. DexVLA: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855* **2025**.
83. Liu, S.; Wu, L.; Li, B.; Tan, H.; Chen, H.; Wang, Z.; Xu, K.; Su, H.; Zhu, J. RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation. In Proceedings of the International Conference on Learning Representations, 2025.
84. Qu, D.; Song, H.; Chen, Q.; Yao, Y.; Ye, X.; Ding, Y.; Wang, Z.; Gu, J.; Zhao, B.; Wang, D.; et al. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model. *Robotics: Science and Systems* **2025**.
85. Zhen, H.; Qiu, X.; Chen, P.; Yang, J.; Yan, X.; Du, Y.; Hong, Y.; Gan, C. 3D-VLA: a 3D vision-language-action generative world model. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning. JMLR.org, 2024, ICML'24.
86. O'Neill, A.; Rehman, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlekar, A.; Jain, A.; et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 6892–6903.
87. Octo Model Team.; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Xu, C.; Luo, J.; et al. Octo: An Open-Source Generalist Robot Policy. In Proceedings of the Proceedings of Robotics: Science and Systems, Delft, Netherlands, 2024.
88. NVIDIA.; Bjorck, J.; Castañeda, F.; Cherniadev, N.; Da, X.; Ding, R.; Fan, L.; et al. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *arXiv preprint arXiv:2503.14734* **2025**.
89. Wen, J.; Zhu, Y.; Li, J.; Zhu, M.; Wu, K.; Xu, Z.; Liu, N.; Cheng, R.; Shen, C.; Peng, Y.; et al. TinyVLA: Towards fast, data-efficient vision-language-action models for robotic manipulation. In Proceedings of the IEEE Robotics and Automation Letters (RA-L), 2025.
90. Shukor, M.; Aubakirova, D.; Capuano, F.; Kooijmans, P.; Palma, S.; Zouitine, A.; Aractingi, M.; Pascal, C.; Russi, M.; Marafioti, A.; et al. SmolVLA: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844* **2025**.
91. Cai, R.; Guo, J.; He, X.; Jin, P.; Li, J.; Lin, B.; Liu, F.; Liu, W.; Ma, F.; Ma, K.; et al. Xiaomi-Robotics-0: An Open-Sourced Vision-Language-Action Model with Real-Time Execution. *arXiv preprint arXiv:2602.12684* **2026**.
92. Ye, J.; Gao, N.; Yang, S.; Zheng, J.; Wang, Z.; Chen, Y.; Chen, P.; Chen, Y.; Liu, S.; Jia, J. StarVLA-α: Reducing Complexity in Vision Language Action Systems. *arXiv preprint arXiv:2604.11757* **2026**.
93. Physical Intelligence.; Amin, A.; Aniceto, R.; Balakrishna, A.; Black, K.; Conley, K.; Connors, G.; Darpinian, J.; Dhabalia, K.; DiCarlo, J.; et al. $\pi^*_{0.6}$: a VLA That Learns From Experience, 2025, [arXiv:cs.LG/2511.14759].

 doi:10.20944/preprints202606.0173.v1

94. Torne, M.; Pertsch, K.; Walke, H.; Vedder, K.; Nair, S.; Ichter, B.; Ren, A.Z.; Wang, H.; Tang, J.; Stachowicz, K.; et al. MEM: Multi-Scale Embodied Memory for Vision Language Action Models. *arXiv preprint arXiv:2603.03596* **2026**.
95. Hafner, D.; Lillicrap, T.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; Davidson, J. Learning latent dynamics for planning from pixels. In Proceedings of the International conference on machine learning. PMLR, 2019, pp. 2555–2565.
96. Hafner, D.; Lillicrap, T.; Ba, J.; Norouzi, M. Dream to Control: Learning Behaviors by Latent Imagination. In Proceedings of the International Conference on Learning Representations, 2020.
97. Hafner, D.; Pasukonis, J.; Ba, J.; Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104* **2023**.
98. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature* **2020**, *588*, 604–609.
99. Hu, A.; Russell, L.; Yeo, H.; Murez, Z.; Fedoseev, G.; Kendall, A.; Shotton, J.; Corrado, G. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080* **2023**.
100. Yang, S.; Du, Y.; Ghasemipour, S.K.S.; Tompson, J.; Kaelbling, L.P.; Schuurmans, D.; Abbeel, P. Learning Interactive Real-World Simulators. In Proceedings of the The Twelfth International Conference on Learning Representations, 2024.
101. Bruce, J.; Dennis, M.D.; Edwards, A.; Parker-Holder, J.; Shi, Y.; Hughes, E.; Lai, M.; Mavalankar, A.; Steigerwald, R.; Apps, C.; et al. Genie: Generative interactive environments. In Proceedings of the Forty-first International Conference on Machine Learning, 2024.
102. Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 15619–15629.
103. Bardes, A.; Garrido, Q.; Ponce, J.; Chen, X.; Rabbat, M.; LeCun, Y.; Assran, M.; Ballas, N. Revisiting Feature Prediction for Learning Visual Representations from Video. *Transactions on Machine Learning Research* **2024**.
104. Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; Zholus, A.; et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985* **2025**.
105. Guo, P.; Mai, Z.; Xu, Z.; Zhang, K.; Zhang, H.; Miao, Z.; Ajoudani, A.; Kingston, Z.; Qiu, Q.; She, Y. AgenticLab: A Real-World Robot Agent Platform that Can See, Think, and Act. *arXiv preprint arXiv:2602.01662* **2026**.
106. Liu, P.; Orru, Y.; Vakil, J.; Paxton, C.; Shafiullah, N.M.M.; Pinto, L. OK-Robot: What Really Matters in Integrating Open-Knowledge Models for Robotics. In Proceedings of the 2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024.
107. Team, G.R.; Abeyruwan, S.; Ainslie, J.; Alayrac, J.B.; Arenas, M.G.; Armstrong, T.; Balakrishna, A.; Baruch, R.; Bauza, M.; Blokzijl, M.; et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020* **2025**.
108. Team, G.R.; Abdolmaleki, A.; Abeyruwan, S.; Ainslie, J.; Alayrac, J.B.; Arenas, M.G.; Balakrishna, A.; Batchelor, N.; Bewley, A.; Bingham, J.; et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342* **2025**.
109. Liu, B.; Jiang, Y.; Zhang, X.; Liu, Q.; Zhang, S.; Biswas, J.; Stone, P. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477* **2023**.
110. Fu, M.; Yu, J.; El-Refai, K.; Kou, E.; Xue, H.; Huang, H.; Xiao, W.; Wang, G.; Li, F.F.; Shi, G.; et al. CaP-X: A Framework for Benchmarking and Improving Coding Agents for Robot Manipulation, 2026, [arXiv:cs.RO/2603.22435].
111. Zhang, H.; Xu, Y.; Fu, Y. Out-of-Sight Embodied Agents: Multimodal Tracking, Sensor Fusion, and Trajectory Forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2026**.
112. Zhang, H.; Xu, Y.; Lu, H.; Shimizu, T.; Fu, Y. Oostraj: Out-of-sight trajectory prediction with vision-positioning denoising. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14802–14811.
113. Zhang, H.; Xu, Y.; Lu, H.; Shimizu, T.; Fu, Y. Layout sequence prediction from noisy mobile modality. In Proceedings of the Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 3965–3974.

 doi:10.20944/preprints202606.0173.v1

114. Huang, W.; Wang, C.; Li, Y.; Zhang, R.; Fei-Fei, L. ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation. In Proceedings of the Conference on Robot Learning. PMLR, 2025, pp. 4573–4602.
115. Shen, W.; Kumar, N.; Chintalapudi, S.; Wang, J.; Watson, C.; Hu, E.; Cao, J.; Jayaraman, D.; Kaelbling, L.P.; Lozano-Pérez, T. TiPToP: A Modular Open-Vocabulary Planning System for Robotic Manipulation. *arXiv preprint arXiv:2603.09971* **2026**.
116. Huang, W.; Xia, F.; Shah, D.; Driess, D.; Zeng, A.; Lu, Y.; Florence, P.; Mordatch, I.; Levine, S.; Hausman, K.; et al. Grounded decoding: Guiding text generation with grounded models for embodied agents. *Advances in Neural Information Processing Systems* **2023**, *36*, 59636–59661.
117. Jin, Y.; Li, D.; Shi, J.; Hao, P.; Sun, F.; Zhang, J.; Fang, B.; et al. Robotgpt: Robot manipulation learning from chatgpt. *IEEE Robotics and Automation Letters* **2024**, *9*, 2543–2550.
118. Fang, H.; Duan, J.; Clay, D.; Wang, S.; Liu, S.; Huang, W.; Fan, X.; Tsai, W.C.; Chen, S.; Wang, Y.R.; et al. MolmoAct2: Action Reasoning Models for Real-world Deployment. *arXiv preprint arXiv:2605.02881* **2026**.
119. Athalye, A.; Kumar, N.; Silver, T.; Liang, Y.; Wang, J.; Lozano-Pérez, T.; Kaelbling, L.P. From Pixels to Predicates: Learning Symbolic World Models via Pretrained Vision-Language Models, 2026, [arXiv:cs.RO/2501.00296].
120. Kumar, N.; Shen, W.; Ramos, F.; Fox, D.; Lozano-Pérez, T.; Kaelbling, L.P.; Garrett, C.R. Open-World Task and Motion Planning via Vision-Language Model Generated Constraints, 2026, [arXiv:cs.RO/2411.08253].
121. Wu, J.; Antonova, R.; Kan, A.; Lepert, M.; Zeng, A.; Song, S.; Bohg, J.; Rusinkiewicz, S.; Funkhouser, T. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots* **2023**, *47*, 1087–1102.
122. Zhang, X.; Altaweel, Z.; Hayamizu, Y.; Ding, Y.; Amiri, S.; Yang, H.; Kaminski, A.; Esselink, C.; Zhang, S. Dkprompt: Domain knowledge prompting vision-language models for open-world planning. *arXiv preprint arXiv:2406.17659* **2024**.
123. Srivastava, S.; Li, C.; Lingelbach, M.; Martín-Martín, R.; Xia, F.; Vainio, K.E.; Lian, Z.; Gokmen, C.; Buch, S.; Liu, K.; et al. BEHAVIOR: Benchmark for Everyday Household Activities in Virtual, Interactive, and Ecological Environments. In Proceedings of the 5th Annual Conference on Robot Learning.
124. Li, C.; Zhang, R.; Wong, J.; Gokmen, C.; Srivastava, S.; Martín-Martín, R.; Wang, C.; Levine, G.; Lingelbach, M.; Sun, J.; et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In Proceedings of the Conference on Robot Learning. PMLR, 2023, pp. 80–93.
125. Li, C.; Xia, F.; Martín-Martín, R.; Lingelbach, M.; Srivastava, S.; Shen, B.; Vainio, K.E.; Gokmen, C.; Dharan, G.; Jain, T.; et al. iGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks. In Proceedings of the 5th Annual Conference on Robot Learning.
126. Shen, B.; Xia, F.; Li, C.; Martín-Martín, R.; Fan, L.; Wang, G.; Pérez-D'Arpino, C.; Buch, S.; Srivastava, S.; Tchapmi, L.; et al. iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021, pp. 7520–7527.
127. Li, M.; Zhao, S.; Wang, Q.; Wang, K.; Zhou, Y.; Srivastava, S.; Gokmen, C.; Lee, T.; Li, L.E.; Zhang, R.; et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems* **2024**, *37*, 100428–100534.
128. Yang, R.; Chen, H.; Zhang, J.; Zhao, M.; Qian, C.; Wang, K.; Wang, Q.; Koripella, T.V.; Movahedi, M.; Li, M.; et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560* **2025**.
129. Hong, Y.; Liu, J.; Yin, H.; Li, M.; Guibas, L.; Li, F.F.; Wu, J.; Choi, Y. ESI-Bench: Towards Embodied Spatial Intelligence that Closes the Perception-Action Loop. *arXiv preprint* **2026**.
130. Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; Stone, P. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems* **2023**, *36*, 44776–44791.
131. Fei, S.; Wang, S.; Shi, J.; Dai, Z.; Cai, J.; Qian, P.; Ji, L.; He, X.; Zhang, S.; Fei, Z.; et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626* **2025**.
132. Zhou, X.; Xu, Y.; Tie, G.; Chen, Y.; Zhang, G.; Chu, D.; Zhou, P.; Sun, L. LIBERO-PRO: Towards Robust and Fair Evaluation of Vision-Language-Action Models Beyond Memorization. *arXiv preprint arXiv:2510.03827* **2025**.
133. Huang, Y.; Li, B.; Saxena, V.; Liang, Y.; Mishra, U.A.; Ji, L.; Zha, L.; Wu, J.; Kumar, N.; Scherer, S.; et al. KinDER: A Physical Reasoning Benchmark for Robot Learning and Planning, 2026, [arXiv:cs.RO/2604.25788].

 doi:10.20944/preprints202606.0173.v1

134. Zhu, Y.; Wong, J.; Mandlekar, A.; Martín-Martín, R.; Joshi, A.; Lin, K.; Maddukuri, A.; Nasiriany, S.; Zhu, Y. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293* **2020**.
135. Nasiriany, S.; Maddukuri, A.; Zhang, L.; Parikh, A.; Lo, A.; Joshi, A.; Mandlekar, A.; Zhu, Y. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523* **2024**.
136. Nasiriany, S.; Nasiriany, S.; Maddukuri, A.; Zhu, Y. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. *arXiv preprint arXiv:2603.04356* **2026**.
137. Yang, X.; Dagli, R.; Zook, A.; Hadfield, H.; Goyal, A.; Birchfield, S.; Ramos, F.; Tremblay, J. RoboLab: A High-Fidelity Simulation Benchmark for Analysis of Task Generalist Policies. In Proceedings of the Robotics: Science and Systems (RSS), 2026.
138. Gong, R.; Zhang, X.; Shang, J.; Minniti, M.V.; Patel, J.; Pepe, V.; Yan, R.; Gundogdu, A.; Kapelyukh, I.; Abbas, A.; et al. AnyTask: an Automated Task and Data Generation Framework for Advancing Sim-to-Real Policy Learning. *arXiv preprint arXiv:2512.17853* **2025**.
139. Lin, J.; Akbari, A.; He, Y.; Zhao, L.; Zhang, H.; Akbari, A.; Xu, X.; Lu, Z.Y.; Nan, E.; Deng, H.; et al. PhyGround: Benchmarking Physical Reasoning in Generative World Models. *arXiv preprint arXiv:2605.10806* **2026**.
140. Ruis, L.; Mozes, M.; Bae, J.; Kamalakara, S.R.; Talupuru, D.; Locatelli, A.; Kirk, R.; Rocktäschel, T.; Grefenstette, E.; Bartolo, M. Procedural Knowledge in Pretraining Drives Reasoning in Large Language Models. In Proceedings of the International Conference on Learning Representations, 2025.
141. Intelligence, P.; Ai, B.; Amin, A.; Aniceto, R.; Balakrishna, A.; Balke, G.; Black, K.; Bokinsky, G.; Cao, S.; Charbonnier, T.; et al. $\pi_{0.7}$: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities. *arXiv preprint arXiv:2604.15483* **2026**.
142. Li, Y.; Li, Z.; Duan, Y.; Spulber, A.B. Physical artificial intelligence (PAI): the next-generation artificial intelligence. *Frontiers of Information Technology & Electronic Engineering* **2023**, *24*, 1231–1238.
143. Dewi, R.S.; Kawakib, A.N.; Laili, M.N.; Fauziah, A.L.; Sabrina, S.R.; Hana, R.L. A systematic review of physical artificial intelligence (Physical AI): concepts, applications, challenges, and future directions. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)* **2025**, *4*, 2246–2253.
144. Zhang, J.; Chen, X.; Wang, Q.; Li, M.; Guo, Y.; Hu, Y.; Zhang, J.; Bai, S.; Lin, J.; Chen, J. VLM4VLA: Revisiting Vision-Language-Models in Vision-Language-Action Models. *arXiv preprint arXiv:2601.03309* **2026**.
145. Gao, C.; Liu, Z.; Chi, Z.; Huang, J.; Fei, X.; Hou, Y.; Zhang, Y.; Lin, Y.; Fang, Z.; Shao, L. Vla-os: Structuring and dissecting planning representations and paradigms in vision-language-action models. *Advances in Neural Information Processing Systems* **2026**, *38*, 136705–136736.
146. Geva, M.; Bastings, J.; Filippova, K.; Globerson, A. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12216–12235.
147. Tao, Y.; Hiatt, A.; Haake, E.; Jetter, A.J.; Agrawal, A. When Context Leads but Parametric Memory Follows in Large Language Models. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 4034–4058.
148. Zhang, H.; Li, Y.; He, S.; Nagarajan, T.; Chen, M.; Lu, J.; Li, A.; Fu, Y. ThinkJEPA: Empowering Latent World Models with Large Vision-Language Reasoning Model. *arXiv preprint arXiv:2603.22281* **2026**.