

From Static Perception to Physical Grounding: A Survey of Visual Reasoning in Physical AI

Anonymous ECCV 2026 Submission

Paper ID #13

Abstract. Vision-language models (VLMs), world foundation models (WFMs), and vision-language-action (VLA) systems are rapidly converging toward Physical AI: systems that must perceive, predict, reason, and act under physical constraints. Existing surveys typically treat these threads separately, emphasizing multimodal perception, predictive world modeling, or embodied control in isolation. This survey instead examines their interface through a capability-centric and verification-aware lens. We organize the Physical AI landscape along two axes: *reasoning depth*, from pattern recognition to multi-step planning and verification, and *physical grounding*, from static web media to embodied interaction, predictive rollouts, and closed-loop control. This framing yields four model families: perception-centric VLMs, reasoning-centric VLMs, WFMs, and physically grounded reasoning VLMs. Across these families, we analyze how data sources, learning paradigms, system interfaces, and evaluation protocols shape capabilities for dynamics prediction, feasibility assessment, long-horizon planning, and process-level validity. We further use Predict-Reason-Act as an organizing lens for emerging design patterns. The survey identifies key gaps in current systems and benchmarks, arguing that reliable Physical AI requires not only stronger perception, reasoning, or action policies, but tighter coupling among them and explicit verification of physical validity.

Keywords: Physical AI · Vision-Language Models · World Foundation Models · Embodied Reasoning · Verification

1 Introduction

VLMs have become central to modern AI, with strong performance on image recognition, visual question answering, document understanding, and multimodal instruction following. Yet deploying multimodal models in physical settings such as robotics, autonomous driving, embodied navigation, and simulation requires more than static perception: agents must reason over time, anticipate action consequences, evaluate physical feasibility, and revise decisions under uncertainty. These demands expose a gap between standard VLM benchmarks and reliable physical decision making.

In parallel, WFMs learn environment dynamics from large-scale data, while embodied agents and VLA systems demonstrate closed-loop control. These

threads are usually surveyed separately: VLM surveys emphasize perception and static reasoning, world-model surveys emphasize dynamics without language-grounded reasoning, and embodied-agent surveys emphasize execution. As a result, physical feasibility, process-level validity, and verification are rarely treated as shared evaluation targets. Recent unified systems such as Cosmos 3 [35] reinforce this convergence by combining multimodal generation, reasoning, world simulation, and action-oriented modeling.

We adopt a *capability-centered* perspective on *Physical AI*: systems that perceive, reason, predict, and act while respecting physical dynamics, constraints, and causal structure. We organize the literature along two axes, *reasoning depth* and *physical grounding*, to clarify why progress along one dimension does not guarantee robustness in physical settings and to synthesize perception-centric VLMs, reasoning-centric VLMs, WFMs, and more tightly integrated systems.

Scope and methodology. We cover recent work (primarily 2020–2026) connecting VLMs, reasoning-centric multimodal models, WFMs, VLAs, embodied agents, simulators, and benchmarks for physical grounding, dynamics prediction, feasibility, and process-level validity. Earlier foundational work is included when it remains central. We exclude language-only reasoning, low-level control without multimodal reasoning, and classical planning/control unless directly connected to modern Physical AI. Representative works are chosen to span capability regimes, training signals, and evaluation protocols rather than to exhaustively catalog every model.

2 Why a New Survey?

As shown in Table 1, recent surveys have made substantial progress in organizing related parts of this landscape, including VLMs, world models, VLAs, embodied agents, and robot learning. However, these threads are often treated as separate subfields. VLM surveys typically emphasize multimodal perception, instruction following, and static or short-context reasoning. World-model surveys focus on predictive dynamics and latent environment simulation, often without treating language-grounded reasoning or action feasibility as central concerns. VLA and embodied-agent surveys emphasize policy learning, control, and task completion, but often leave process-level reasoning validity and physical verification underdeveloped.

This separation obscures a central issue for Physical AI: robust physical-world intelligence depends not only on strong perception, strong reasoning, or strong dynamics modeling in isolation, but on the interface between them. A system may reason fluently without being physically grounded, predict plausible futures without understanding goals or constraints, or execute actions without exposing whether its intermediate reasoning is causally and physically valid. Existing surveys therefore leave open a cross-cutting question: how should we understand systems that jointly combine reasoning depth, physical grounding, and verification?

Table 1: Comparison with Existing Surveys.

S1 [10], S2 [30], S3 [52], S4 [18], S5 [53], S6 [49], S7 [26]. The symbols indicate coverage strength: ○ = not covered or only briefly mentioned, ◐ = partially covered, and ● = substantially covered. S1–S7 are selected as recent surveys from the closest adjacent areas: world models, embodied AI, robotic manipulation, VLMs, multimodal reasoning, VLAs, and autonomous driving.

	S1	S2	S3	S4	S5	S6	S7	Ours
Perception	●	●	●	●	●	●	●	●
Reasoning	◐	◐	●	◐	●	●	●	●
Dynamics Modeling	●	●	●	○	○	○	◐	●
Physical Feasibility	○	●	◐	○	○	○	◐	●
Long-Horizon Planning	◐	●	●	○	◐	○	●	●
Process-Level Validity	○	○	○	○	●	○	◐	●
System Integration (Predict-Reason-Act)	◐	◐	●	○	◐	●	●	●
Verification	○	○	○	○	○	○	◐	●

This survey addresses that gap by providing a *capability-centric* and *verification-aware* synthesis of Physical AI. Rather than organizing the field only by architecture or application domain, we analyze how perception, reasoning, prediction, feasibility assessment, and action are coupled. Our two-axis landscape, reasoning depth $\times$ physical grounding, is designed to make this interface explicit, while the Predict–Reason–Act lens, defined in Section 3.2, connects VLMs, WFMs, VLAs, and physically grounded reasoning systems into a unified design space.

Key differentiators. Our synthesis is *capability-first* and *verification-aware*: it provides a unified reasoning-depth $\times$ physical-grounding landscape, elevates feasibility and process-level validity as first-class requirements, analyzes how reasoning and world models are coupled in practice, and identifies measurable challenges for moving from semantic grounding to verifiable Physical AI.

3 Physical AI: Definition and Landscape

Physical AI refers to systems that perceive, reason, predict, and act in physical environments while respecting dynamics, constraints, and causal structure. Unlike static VLMs, such systems must anticipate action consequences and operate under temporal and safety constraints. We analyze Physical AI through a two-axis landscape defined by reasoning depth and physical grounding.

3.1 Reasoning Depth and Physical Grounding

We organize Physical AI along two axes: *reasoning depth* and *physical grounding*, as illustrated in Figure 1.

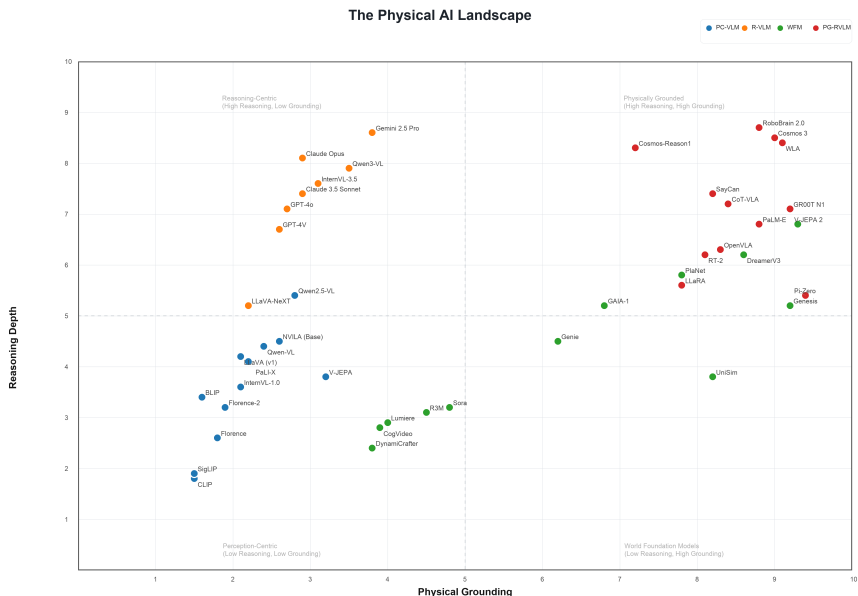


Fig. 1: Landscape of Physical AI systems by reasoning depth and physical grounding, highlighting complementary strengths across model families rather than a single dominant paradigm. Coordinates are qualitative placements using the anchored rubric in Table 2.

- **Reasoning depth** measures multi-step inference and task decomposition, from shallow pattern recognition to explicit reasoning trained via chain-of-thought supervision [46] or reinforcement learning [40]. Depth alone does not guarantee physical validity.
- **Physical grounding** measures coupling to physical state, actions, and dynamics, from static web-scale media (no interventions or state transitions) to embodied trajectories, controllable simulators, or learned world models that enable action-conditioned prediction, counterfactual rollouts, and feasibility checks.

These axes separate common model regimes: PC-VLMs (low depth/grounding), R-VLMs (higher depth, weak grounding), WFMs (strong grounding, implicit reasoning), and PG-VLMs (high depth and grounding). This view motivates the capability and design-space analysis that follows.

Scoring protocol. To position models within the diagnostic landscape of Figure 1, we assign each system a score on two orthogonal dimensions: reasoning depth and physical grounding. These scores are not intended as absolute measures of model quality. Instead, they characterize the relative capabilities most relevant to Physical AI.

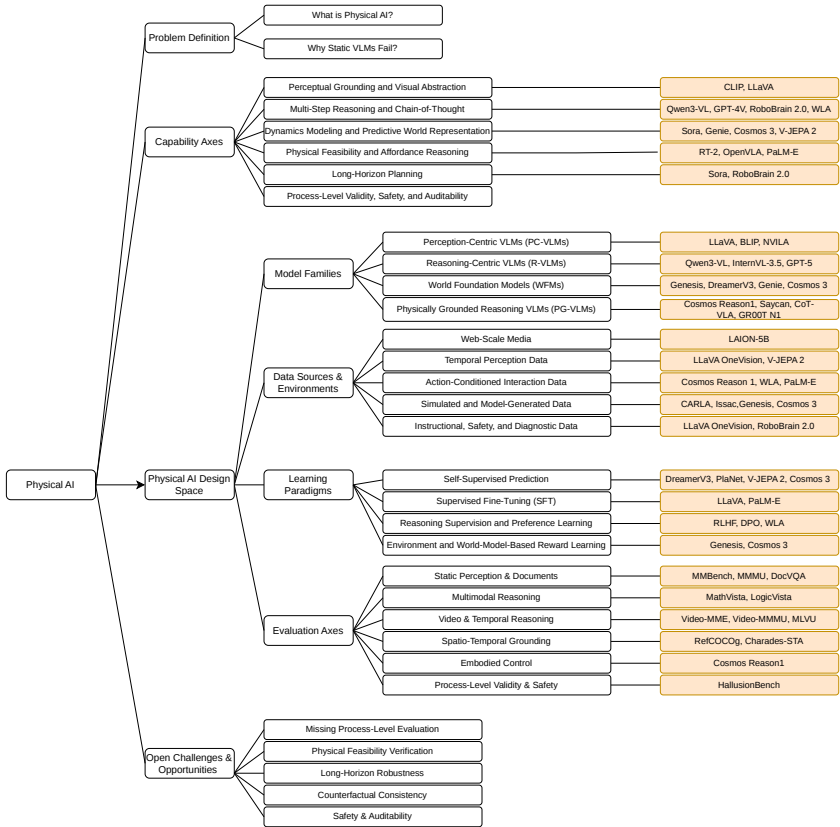


Fig. 2: Overview of the Physical AI design space. The taxonomy maps where physical capability enters a system: model families define computational roles, data sources provide physical signals, learning paradigms transform signals into behavior, and evaluation axes validate resulting capabilities.

Reasoning depth reflects explicit multi-step inference and structured intermediate reasoning: pattern recognizers score lower, while systems using CoT, task decomposition, search, planning, process supervision, reinforcement learning, or verification score higher. Physical grounding reflects coupling to physical state, actions, dynamics, simulators, embodied interaction, or predictive world models. Static image-text systems score low; systems with action-conditioned trajectories, counterfactual rollouts, or closed-loop embodied control score higher. Scores are based on reported training objectives, supervision, architecture, reasoning interfaces, and environment interaction rather than benchmark performance alone.

When a system spans multiple bands, we place it according to the strongest capability central to the method rather than an auxiliary demonstration.

Table 2: Anchored scoring rubric for Figure 1. Scores indicate relative capability bands rather than precise absolute measurements.

Score	Reasoning depth	Physical grounding
1–2	Pattern recognition, retrieval, or Static image-text or document shallow perception with little explicit inference.	data with no temporal, action, or environment signal.
3–4	Basic visual or temporal reasoning, usually outcome-only and weakly structured.	Passive video or weak temporal supervision without action-conditioned dynamics.
5–6	Multi-step reasoning, decomposition, or planning interfaces with limited external validation.	Predictive video, latent dynamics, or embodied data that models physical evolution but weakly supports intervention.
7–8	Structured reasoning with search, planning, process supervision, or verifier-style checks.	Action-conditioned rollouts, simulators, world models, or embodied feedback for feasibility assessment.
9–10	Closed-loop reasoning with explicit validation, correction, or policy-level decision making.	Interactive, counterfactual, or closed-loop embodied control with rollout-based evaluation and physical constraints.

Example placements. CLIP [41] anchors the lower-left region: it supports recognition and retrieval but lacks multi-step reasoning, action conditioning, learned dynamics, or environment feedback. DreamerV3 [23] anchors the world-model region: latent dynamics and imagination rollouts give it high grounding, while the absence of language-native task decomposition or explicit verification keeps reasoning depth moderate. The axes therefore capture complementary capabilities rather than a single notion of model strength.

3.2 The Predict–Reason–Act Loop

We use *Predict–Reason–Act* as an organizing lens, not a fixed architecture. *Predict* models future states, action-conditioned rollouts, or counterfactual scene evolutions using a simulator, world model, or predictive video model. *Reason* infers goals, constraints, causal relations, feasibility, and candidate plans from observations and predicted futures. *Act* selects, executes, or revises actions. Recent World-Language-Action models [45] make this coupling explicit by jointly predicting future states, language-level subtasks, and executable actions.

This lens connects Figure 1’s axes: prediction and environment feedback increase grounding, while structured inference, constraint reasoning, and plan revision increase reasoning depth. It differs from sense–plan–act because prediction explicitly models possible futures and action consequences, and it differs from

pure VLM reasoning because plans are checked against predicted physical consequences before action.

4 Capability-Centric Analysis

We analyze Physical AI through *capabilities*: perception, multi-step reasoning, dynamics prediction, feasibility assessment, planning, and process-level validity. Robust behavior requires these capabilities to be jointly present and coherently integrated; failures often come from missing interfaces rather than weak components in isolation.

4.1 Perceptual Grounding and Visual Abstraction

Perceptual grounding transforms sensory inputs into visual abstractions for recognition, localization, grounding, OCR, and long-context image or video understanding. It is typically realized through large vision encoders trained with self-supervised or contrastive objectives and aligned with language models, yielding PC-VLM backbones such as CLIP [41] and LLaVA [31]. However, static or weakly grounded objectives remain brittle under occlusion, noise, viewpoint shift, and temporal aliasing; perception benchmarks therefore do not guarantee intuitive physics, counterfactual reasoning, or action-conditioned prediction.

4.2 Multi-Step Reasoning and Chain-of-Thought

Multi-step reasoning enables task decomposition, evidence integration, and coherent inference. In VLMs it is often instantiated as CoT, where models generate intermediate steps before a final output. Systems such as GPT-4V [38] and Qwen3-VL [2] extend this pattern to images and videos. CoT improves interpretability and provides an optimization interface for supervision, self-training, or preference learning [40, 47], but it does not ensure physical grounding: traces are rarely validated against simulators, rollouts, or environment feedback. Physical AI therefore requires reasoning steps to be constrained or verified by predictive dynamics.

4.3 Dynamics Modeling and Predictive World Representation

Dynamics modeling predicts how environments evolve under natural dynamics and actions, capturing motion, persistence, interactions, and transitions. WFMs learn this capability from long temporal sequences, extending model-based RL ideas to video and sensor logs. V-JEPA 2 [20] connects self-supervised video learning with prediction and robot planning, while open-loop generators such as Sora [39] show visual dynamics but often lack explicit state or action representations. Robustness under distribution shift and long horizons still requires tighter coupling to reasoning and feasibility verification.

4.4 Physical Feasibility and Affordance Reasoning

Physical feasibility asks whether a proposed action can be executed safely given object properties, embodiment, and environmental constraints. It filters plans that violate physical laws, safety constraints, or task requirements. SayCan [1] explicitly constrained language proposals using skill affordances, while modern VLA systems such as OpenVLA [27] often absorb affordances through demonstrations. These signals generalize poorly under distribution shift, and benchmarks often conflate feasibility with perception, planning, and control, motivating feasibility as a first-class verifiable capability.

4.5 Long-Horizon Planning

Long-horizon planning requires decomposing goals, maintaining state, and adapting under partial observability. VLMs can propose abstract plans, while world models provide rollouts for lookahead. VLA policies such as RT-1, OpenVLA, Octo, π_0 , and GR00T N1 [5, 7, 27, 37, 44] map observations and instructions to executable robot actions, but plans remain brittle when language-generated subgoals are weakly grounded, generative models lack action semantics, or action-conditioned models accumulate error. Progress requires evaluations that isolate planning structure, recovery, and replanning rather than only final task success.

4.6 Process-Level Validity, Safety, and Auditability

Process-level validity asks whether intermediate traces and plans are causally grounded, physically consistent, and safe, not merely whether the final output is plausible. This is critical when intermediate errors compound into unsafe behavior. Current VLM agents rely on opaque states and free-form plans; structured reasoning states, world-model verifiers, and neuro-symbolic checkers offer promising hooks, but are not yet standardized in training or benchmarks.

5 Physical AI Design Space

We analyze Physical AI through four design axes: *model families*, capturing architectural roles; *data sources and environments*, determining physical signals; *learning paradigms*, governing how signals become behavior; and *evaluation axes*, specifying how capabilities are validated (Figure 2). The taxonomy is not a hierarchy of quality but a map of where physical capability enters a system and where it can fail. For example, a reasoning-centric VLM trained only on image-text data may excel at static reasoning but fail under action-conditioned dynamics, while a world model may predict plausible futures but lack goal-conditioned planning unless coupled to a reasoning module.

5.1 Model Families as Architectural Strategies

We first organize systems by their dominant role in the Predict–Reason–Act loop. The grouping is analytical rather than exclusive: many systems combine roles, but their dominant design reveals which capability they provide directly and which they borrow from other components.

5.2 Perception-centric VLMs.

PC-VLMs couple strong visual encoders with language models to represent images, documents, and short videos. CLIP-style models [41], LLaVA [31], and Gemini-Vision [16] excel at recognition, grounding, OCR, and spatial layout understanding, typically through visual pretraining, multimodal alignment, and instruction tuning. In Physical AI they provide perceptual abstractions, but weak dynamics and action grounding limit intuitive physics, feasibility reasoning, and long-horizon prediction.

5.3 Reasoning-centric VLMs.

R-VLMs emphasize multi-step inference, symbolic abstraction, and task decomposition over multimodal inputs. Systems such as GPT-4V [38], Gemini-Pro [16], PaLM-E [13], Qwen3-VL [2], and LLaVA-family models [31] primarily implement the *Reason* stage by interpreting goals and generating plans or tool calls. Their core trade-off is strong task decomposition with weak physical validation: without action–effect supervision, environment dynamics, or predictive rollouts, reasoning remains largely unconstrained by feasibility.

5.4 World Foundation Models.

WFMs learn environment dynamics from videos, embodied trajectories, or simulated rollouts. Dreamer [21], PlaNet [22], V-JEPA-style models [3, 20], and domain-specific dynamics learners implement the *Predict* stage through rollouts, forecasting, and consistency checks. Their core trade-off is strong prediction with weak goal-conditioned reasoning and instruction following; they are most useful when coupled with reasoning modules that propose actions for evaluation.

5.5 Physically Grounded Reasoning VLMs

PG-VLMs couple multimodal reasoning with predictive dynamics, structured state, or learned world-model interfaces so intermediate steps can be checked against plausible futures. Cosmos-Reason1 [36], Cosmos 3 [35], World-Language-Action models [45], RoboBrain 2.0 [8], and Qwen3-VL [2] illustrate this pattern. Cosmos 3 is especially aligned with our convergence thesis because it treats generation, reasoning, world simulation, and action-conditioned modeling as parts of a unified Physical AI backbone. The trade-off is improved integration but immature interfaces, sparse verification protocols, and limited closed-loop evidence.

5.6 Data Sources and Environments: Where the Signal Comes From

Data regimes determine which physical signals a system can learn and therefore which capabilities among perception, reasoning, prediction, and planning can emerge.

Static Semantic Multimodal Data. Web-scale image-text corpora [43], documents, and instructional images underpin PC-VLMs and R-VLMs, but the absence of temporal continuity and action-effect structure limits dynamics, feasibility, and long-horizon behavior.

Temporal Perception Data. Passive video [29] exposes motion and temporal coherence, supporting WFMs and temporal VLMs, but lack of action conditioning restricts causal reasoning and planning.

Action-Conditioned Interaction Data. Robotics and driving logs pair observations with actions and state transitions [7, 9, 36], directly supervising dynamics, affordances, and feasibility, but remain costly and limited in coverage.

Simulated and Model-Generated Data. Simulators such as CARLA and Genesis [12, 17] provide controllable dynamics for intervention, feasibility checks, and safety testing. Learned world models generate rollouts for scalable imagination and rare-event synthesis. Simulators offer stronger guarantees but limited realism; model-generated data trades fidelity for coverage.

Instructional, Safety, and Diagnostic Data. Instruction data [29], rule-annotated scenarios, counterfactual queries, and safety-critical edge cases target constraint satisfaction, failure detection, and robustness.

5.7 Learning Paradigms: How the Model Learns

Physical AI capabilities depend on how physical structure is induced during learning. Current systems combine representation learning, supervised or distilled alignment, reasoning supervision, preference optimization, and environment-grounded feedback.

Representation and Predictive Learning. Self-supervised objectives underpin VLMs and WFMs. Contrastive and masked objectives such as CLIP, SigLIP, and MAE [24, 41, 51] yield semantic and spatial representations, while predictive objectives over sequences support dynamics, forecasting, and rollout-based reasoning.

Supervised and Distilled Alignment. SFT and distillation align pre-trained representations with task-specific perception, reasoning, or control [25, 31]. They transfer semantics, action mappings, reasoning behaviors, or policies, but do not by themselves guarantee long-horizon consistency or physical feasibility.

Reasoning Supervision and Preference Optimization. CoT supervision trains on intermediate traces rather than final answers, while preference optimization refines outputs using comparative judgments [42]. These methods improve coherence and alignment but require physical verification signals to optimize feasibility.

Environment- and World-Model-Based Reward Learning. Environment feedback uses simulators or learned world models to score plans or intermediate steps for feasibility, safety, or outcomes. It offers a path beyond preference alignment, but remains difficult to scale and standardize.

5.8 Evaluation Axes: How We Judge Physical AI Systems

Evaluation remains fragmented: benchmarks often isolate perception, reasoning, or prediction while rarely testing their interaction. We argue for loop-level diagnosis of Predict–Reason–Act systems that separately measures *prediction validity* (physical and temporal constraints), *reasoning validity* (goals, constraints, causal claims), *feasibility validity* (executable actions for the embodiment and scene), and *recovery validity* (revision after failed predictions or actions). This decomposition attributes failures to perception, dynamics, reasoning, action grounding, or verification rather than only final task failure.

Static Perception and Grounding. Benchmarks such as MMBench [33] and DocVQA [34] form the backbone of PC-VLM evaluation, but do not probe temporal consistency, physical causality, or interaction-dependent behavior.

Static Multimodal Reasoning. Static reasoning benchmarks such as MMMU [50] measure multi-step inference and compositional reasoning, but remain outcome-oriented and do not assess causal validity, physical feasibility, or intermediate robustness.

Temporal Understanding versus Predictive Dynamics. Video-MME [14] evaluates temporal ordering over observed sequences, whereas Physion, IntPhys 2, MVP, and PAI-Bench [4, 6, 28, 55] probe physical prediction and causal interactions. The gap is that temporal reasoning over observations does not imply counterfactual or action-conditioned dynamics.

Spatial grounding and spatio-temporal localization. RefCOCO/g and Charades-STA [15, 48] test region or action localization, while RoboRefer and RefSpatial-Bench [54] probe embodied spatial reasoning. These evaluations remain mostly offline and rarely assess feasibility.

Embodied Planning and Execution. Embodied and simulator benchmarks use task completion, reward, or safety violations as end-to-end signals. Cosmos-Reason1 and FoMER [11, 36] further probe video-conditioned common-sense and step-by-step embodied reasoning, but still conflate perception, prediction, planning, and control.

Process-Level Faithfulness and Safety. HallusionBench [19] assesses hallucination and instruction faithfulness, while ERIQ and GenieReasoner [32] begin to isolate the reasoning–execution bridge. Process-level physical verification, however, remains far from standardized.

6 Challenges and Outlook for Physical AI

Despite rapid progress in multimodal perception and reasoning, current Physical AI systems remain brittle in settings that require long-horizon interaction,

counterfactual generalization, and safety-critical decision making. A central unresolved challenge is the transition from *semantic grounding* to *physical grounding*: while perception- and reasoning-centric VLMs perform well on static understanding, their inferences are often driven by language-native priors rather than dynamics-aware constraints, leading to physically implausible plans under intervention. Addressing this gap is tightly coupled to data regimes: web-scale multimodal corpora provide semantic breadth but weak physical signal, motivating greater reliance on embodied interaction data and simulated environments that expose models to rare and long-tail safety-critical events. Progress on this challenge should be visible not only as higher benchmark accuracy, but as improved counterfactual prediction, physical violation detection, and transfer under distribution shift.

A second major bottleneck is *process-level validity*. Current training and evaluation protocols remain largely outcome-based, rewarding plausible final answers while ignoring the physical validity, faithfulness, and causal necessity of intermediate reasoning steps. This misalignment motivates process-aware optimization and evaluation, where intermediate representations such as state abstractions, plans, or constraints are explicitly validated against physical evidence, feasibility constraints, or predictive rollouts. A meaningful advance would therefore show intermediate reasoning traces that can be checked against simulators, learned world models, or human-annotated causal constraints, not just more fluent explanations.

Long-horizon planning under uncertainty presents a further challenge. WFMs provide predictive substrates but accumulate error under distribution shift, while reasoning-centric VLMs generate multi-step plans without guarantees of executability. Robust autonomy requires closing the loop between *proposal* and *validation*, coupling plan generation with uncertainty-aware prediction, feasibility checks, and online revision using simulators or learned world models. Progress should be measured by lower compounding error, better recovery after failed actions, and robust transfer across scenes and embodiments.

Taken together, these challenges point toward a convergent trajectory for Physical AI: the integration of PC-VLMs, reasoning-centric models, and WFMs into coherent Predict–Reason–Act systems with stable interfaces and verifiable decision processes. Such integration should be evaluated through ablations that reveal which module or interface contributes to reliability, rather than through aggregate success rates alone. By reframing the field around capabilities, data, learning, and evaluation choices rather than architectures alone, this survey highlights the path toward physically grounded, process-aware, and deployable AI systems.

References

1. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., et al.: Do as i can, not as i say: Grounding language in robotic affordances. In: Conference on Robot Learning (2022), <https://arxiv.org/abs/2204.01691>

2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), <https://arxiv.org/abs/2511.21631>, arXiv preprint arXiv:2511.21631
3. 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 (2024), <https://arxiv.org/abs/2404.08471>, arXiv preprint arXiv:2404.08471
4. Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y.F., Pramod, R.T., Holdaway, C., Tao, S., Smith, K., Sun, F.Y., Fei-Fei, L., Kanwisher, N., Tenenbaum, J.B., Yamins, D.L.K., Fan, J.E.: Physion: Evaluating physical prediction from vision in humans and machines. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (2021), <https://arxiv.org/abs/2106.08261>
5. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Heess, N., et al.: π_0 : A vision-language-action flow model for general robot control (2024), <https://arxiv.org/abs/2410.24164>, arXiv preprint arXiv:2410.24164
6. Bordes, F., Garrido, Q., Kao, J.T., Williams, A., Rabbat, M., Dupoux, E.: Inthys 2: Benchmarking intuitive physics understanding in complex synthetic environments (2025), <https://arxiv.org/abs/2506.09849>, arXiv preprint arXiv:2506.09849
7. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., et al.: Rt-1: Robotics transformer for real-world control at scale. In: Robotics: Science and Systems (2023), <https://arxiv.org/abs/2212.06817>
8. Cao, M., He, C., Lin, H., et al.: Robobrain 2.0: See better. think harder. do smarter. (2025), <https://arxiv.org/abs/2507.02029>, arXiv preprint arXiv:2507.02029
9. Collaboration, O.X.E., O'Neill, A., Rehman, A., Gupta, A., Kekane, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models (2023), <https://arxiv.org/abs/2310.08864>, arXiv preprint arXiv:2310.08864
10. Ding, J., Zhang, Y., Shang, Y., Feng, J., Zhang, Y., Zong, Z., Yuan, Y., Su, H., Li, N., Piao, J., Deng, Y., Sukiennik, N., Gao, C., Xu, F., Li, Y.: Understanding world or predicting future? a comprehensive survey of world models (2025), <https://arxiv.org/abs/2411.14499>, arXiv preprint arXiv:2411.14499
11. Dissanayake, D., Heakl, A., Thawakar, O., Ahsan, N., Thawkar, R., More, K., et al.: How good are foundation models in step-by-step embodied reasoning? (2025), <https://arxiv.org/abs/2509.15293>, arXiv preprint arXiv:2509.15293
12. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Proceedings of the 1st Annual Conference on Robot Learning (2017), <https://arxiv.org/abs/1711.03938>
13. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., et al.: Palm-e: An embodied multimodal language model. In: Proceedings of the 40th International Conference on Machine Learning (2023), <https://arxiv.org/abs/2303.03378>
14. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X.,

- Chen, E., Shan, C., He, R., Sun, X.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025), <https://arxiv.org/abs/2405.21075>
15. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE International Conference on Computer Vision (2017), <https://arxiv.org/abs/1705.02101>
 16. Gemini Team, Google: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023), <https://arxiv.org/abs/2312.11805>
 17. Genesis Authors: Genesis: A generative and universal physics engine for robotics and beyond. <https://github.com/Genesis-Embodied-AI/Genesis> (December 2024), accessed: 2026-01-18
 18. Ghosh, A., Acharya, A., Saha, S., Jain, V., Chadha, A.: Exploring the frontier of vision-language models: A survey of current methodologies and future directions (2025), <https://arxiv.org/abs/2404.07214>, arXiv preprint arXiv:2404.07214
 19. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024), <https://arxiv.org/abs/2310.14566>
 20. Guizilini, V., Bardes, A., Bakir, B., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning (2025), <https://arxiv.org/abs/2506.09985>, arXiv preprint arXiv:2506.09985
 21. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. In: International Conference on Learning Representations (2020), <https://arxiv.org/abs/1912.01603>
 22. 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 36th International Conference on Machine Learning (2019), <https://arxiv.org/abs/1811.04551>
 23. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models (2024), <https://arxiv.org/abs/2301.04104>, arXiv preprint arXiv:2301.04104
 24. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022), <https://arxiv.org/abs/2111.06377>
 25. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015), <https://arxiv.org/abs/1503.02531>, nIPS Deep Learning and Representation Learning Workshop
 26. Jiang, S., Huang, Z., Qian, K., Luo, Z., Zhu, T., Zhong, Y., Tang, Y., Kong, M., Wang, Y., Jiao, S., Ye, H., Sheng, Z., Zhao, X., Wen, T., Fu, Z., Chen, S., Jiang, K., Yang, D., Choi, S., Sun, L.: A survey on vision-language-action models for autonomous driving (2025), <https://arxiv.org/abs/2506.24044>, arXiv preprint arXiv:2506.24044
 27. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: Openvla: An open-source vision-language-action model. In: Conference on Robot Learning (2024), <https://arxiv.org/abs/2406.09246>

28. Krojer, B., Murdock, C., Li, M., Jain, U., Rohrbach, A., Darrell, T.: A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs (2025), <https://arxiv.org/abs/2506.09987>, arXiv preprint arXiv:2506.09987
29. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer (2024), <https://arxiv.org/abs/2408.03326>, arXiv preprint arXiv:2408.03326
30. Li, X., He, X., Zhang, L., Wu, M., Li, X., Liu, Y.: A comprehensive survey on world models for embodied ai (2025), <https://arxiv.org/abs/2510.16732>, arXiv preprint arXiv:2510.16732
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (2023), <https://arxiv.org/abs/2304.08485>
32. Liu, Y., Wang, S., Wei, D., Cai, X., Zhong, L., Yang, J., et al.: Unified embodied vlm reasoning with robotic action via autoregressive discretized pre-training (2025), <https://arxiv.org/abs/2512.24125>, arXiv preprint arXiv:2512.24125
33. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? In: European Conference on Computer Vision (2024), <https://arxiv.org/abs/2307.06281>
34. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2021), <https://arxiv.org/abs/2007.00398>
35. NVIDIA, Agarwal, N., Ali, A., Bai, J., et al.: Cosmos 3: Omnimodal world models for physical ai (2026), <https://arxiv.org/abs/2606.02800>, arXiv preprint arXiv:2606.02800
36. NVIDIA, Azzolini, A., Bai, J., Brandon, H., Cao, J., Chattopadhyay, P., Chen, H., Chu, J., Cui, Y., Diamond, J., Ding, Y., Feng, L., Ferroni, F., Govindaraju, R., Gu, J., Gururani, S., Hanafi, I.E., Hao, Z., Huffman, J., Jin, J., Johnson, B., Khan, R., Kurian, G., Lantz, E., Lee, N., Li, Z., Li, X., Liao, M., Lin, T.Y., Lin, Y.C., Liu, M.Y., Lu, X., Luo, A., Mathau, A., Ni, Y., Pavao, L., Ping, W., Romero, D.W., Smelyanskiy, M., Song, S., Tchapmi, L., Wang, A.Z., Wang, B., Wang, H., Wei, F., Xu, J., Xu, Y., Yang, D., Yang, X., Yang, Z., Zhang, J., Zeng, X., Zhang, Z.: Cosmos-reason1: From physical common sense to embodied reasoning (2025), <https://arxiv.org/abs/2503.15558>, arXiv preprint arXiv:2503.15558
37. NVIDIA, Bjorck, J., Leal-Taixe, F.C.M., et al.: Gr00t n1: An open foundation model for generalist humanoid robots (2025), <https://arxiv.org/abs/2503.14734>, arXiv preprint arXiv:2503.14734
38. OpenAI: Gpt-4v(ision) system card. <https://openai.com/research/gpt-4v-system-card> (2023), accessed: 2026-01-18
39. OpenAI, Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators. <https://openai.com/research/video-generation-models-as-world-simulators> (2024), accessed: 2026-01-18
40. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems (2022), <https://arxiv.org/abs/2203.02155>
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceed-

- ings of the 38th International Conference on Machine Learning (2021), <https://arxiv.org/abs/2103.00020>
42. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Advances in Neural Information Processing Systems (2023), <https://arxiv.org/abs/2305.18290>
 43. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (2022), <https://arxiv.org/abs/2210.08402>
 44. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., et al.: Octo: An open-source generalist robot policy (2024), <https://arxiv.org/abs/2405.12213>, arXiv preprint arXiv:2405.12213
 45. Wang, X., Zhang, X., Yuan, Z., Li, P., et al.: World-language-action model (2026), <https://arxiv.org/abs/2606.05979>, arXiv preprint arXiv:2606.05979
 46. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems (2022), <https://arxiv.org/abs/2201.11903>
 47. Yu, B., Yuan, H., Li, H., Xu, X., Wei, Y., Wang, B., Qi, W., Chen, K.: Long-short chain-of-thought mixture supervised fine-tuning eliciting efficient reasoning in large language models. arXiv preprint arXiv:2505.03469 (2025)
 48. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 69–85 (2016)
 49. Yu, Z., Wang, B., Zeng, P., Zhang, H., Zhang, J., Gao, L., Song, J., Sebe, N., Shen, H.T.: A survey on efficient vision-language-action models (2025), <https://arxiv.org/abs/2510.24795>, arXiv preprint arXiv:2510.24795
 50. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024), <https://arxiv.org/abs/2311.16502>
 51. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023), <https://arxiv.org/abs/2303.15343>
 52. Zhang, P.F., Cheng, Y., Sun, X., Wang, S., Li, F., Zhu, L., Shen, H.T.: A step toward world models: A survey on robotic manipulation (2025), <https://arxiv.org/abs/2511.02097>, arXiv preprint arXiv:2511.02097
 53. Zhou, C., Wang, M., Ma, Y., Wu, C., Chen, W., Qian, Z., Liu, X., Zhang, Y., Wang, J., Xu, H., Luo, F., Chen, X., Hao, X., Li, H., Zhang, A., Wang, W., Zhang, K., Jia, G., Li, L., Lu, Z., Lu, Y., Guo, Y.: From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models (2025), <https://arxiv.org/abs/2509.25373>, arXiv preprint arXiv:2509.25373
 54. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., Zhang, S.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics (2026), <https://arxiv.org/abs/2506.04308>, arXiv preprint arXiv:2506.04308

55. Zhou, F., Huang, J., Li, J., Ramanan, D., Shi, H.: Pai-bench: A comprehensive benchmark for physical ai (2025), <https://arxiv.org/abs/2512.01989>, arXiv preprint arXiv:2512.01989