OPEN ACCESS

Embodied Intelligence REVIEW

World models: Toward computationally usable embodied AI

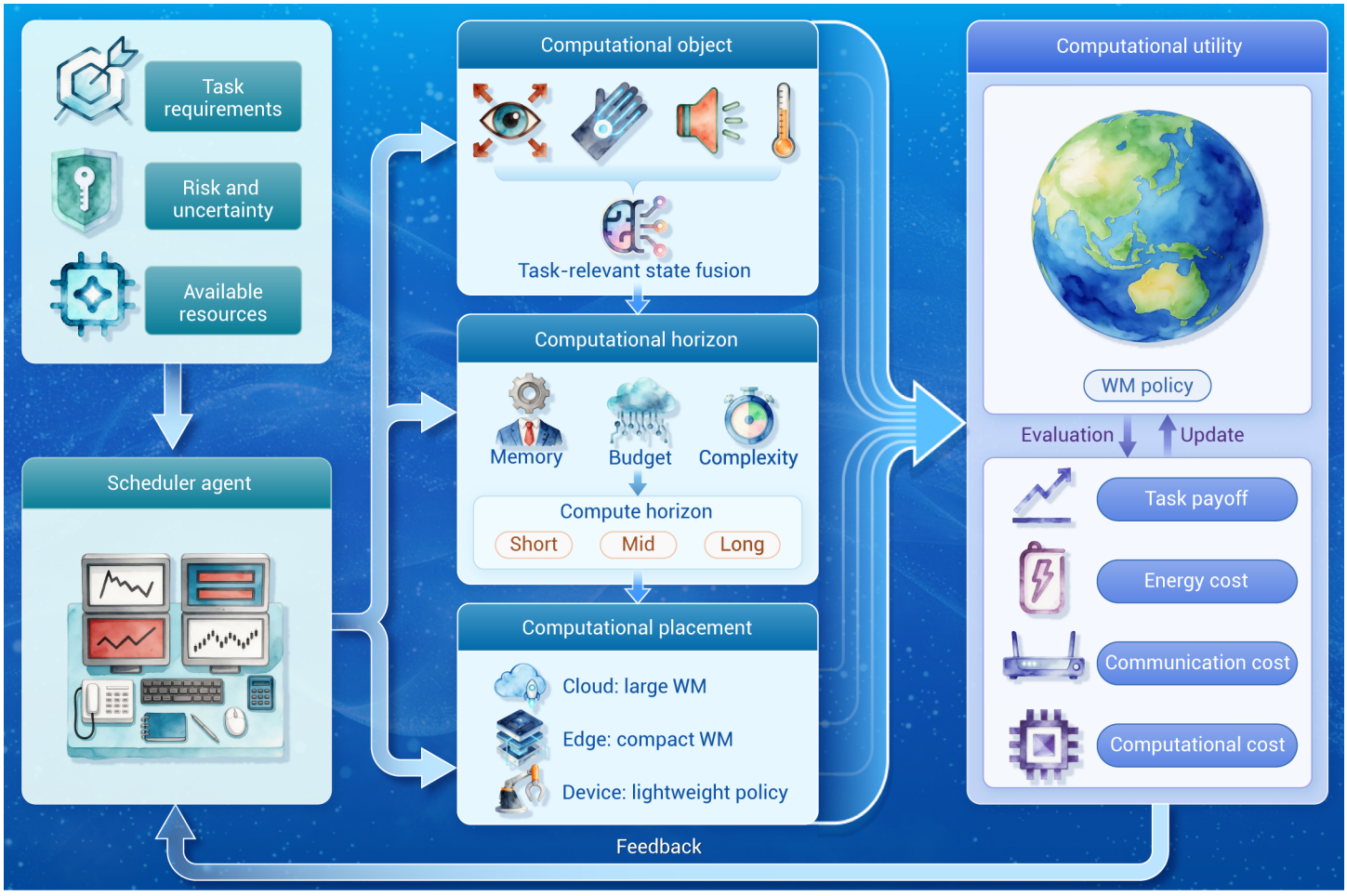
Yiyang Song,[1] Jianxiao Wang,[2,*] Hongshuo Liu,[2] Liang Lin,[3,4] and Jie Song[5]

*Correspondence: wang-jx@pku.edu.cn

Received: August 2, 2026; Accepted: October 19, 2026; Published Online: October 20, 2026; https://doi.org/10.59717/j.tiis.2026.100018

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Computational usability asks what to compute, how far to simulate, where to run it, and what value it adds.
- Action-oriented data help models prioritize the states that matter for decisions and safety.
- Risk-adaptive rollouts and device-edge-cloud computing may make world models more deployable.
- Action gain per computational cost measures improvements in decision-making relative to computing and resource costs.

REVIEW Embodied Intelligence

OPEN ACCESS

www.the-innovation.org/the-innovation-insights

World models: Toward computationally usable embodied AI

Yiyang Song,[1] Jianxiao Wang,[2,*] Hongshuo Liu,[2] Liang Lin,[3,4] and Jie Song[5]

[1]Department of Electrical and Computer Engineering, The University of Hong Kong, Hong Kong 999077, China
[2]National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871, China
[3]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China
[4]Joy Future Academy, JingDong Inc., Beijing 100871, China
[5]School of Advanced Manufacturing and Robotics, Peking University, Beijing 100871, China
*Correspondence: wang-jx@pku.edu.cn

Received: August 2, 2026; Accepted: October 19, 2026; Published Online: October 20, 2026; https://doi.org/10.59717/j.tiis.2026.100018

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Song Y., Wang J., Liu H., et al. (2026). World models: Toward computationally usable embodied AI. *The Innovation Insights* **1**:100018.

World models are evolving from latent dynamics modules in model-based reinforcement learning into critical infrastructure for embodied intelligence, autonomous driving, robot learning, and physical AI. However, current discussions often equate world models with larger-scale video generation or more realistic interactive simulation, overlooking a fundamental constraint: embodied systems must act under constraints on limited data, computation, and communication under real-time safety requirements. This paper frames this challenge as the "computational usability gap." An actionable world model must not only predict the future, but also determine what to compute, how far to roll out, where to compute, and whether the computation delivers sufficient action value. We review recent progress in representation, training and rollout, deployment, and evaluation, and argue that the focus is shifting from visual fidelity to action utility per unit of computation.

INTRODUCTION

The early paradigm of world models stems from model-based reinforcement learning, where agents learn compact internal representations of the environment, predict future states in latent space, and optimize policies accordingly. World Models showed that agents can train policies in self-generated "dreams".[1] PlaNet and the Dreamer series further extended latent dynamics to pixel-based planning and control, [2-3] with DreamerV3 solving more than 150 tasks under a unified configuration and demonstrating the potential of world models as general-purpose control backbones.[4]

Recently, the scope of world models has expanded rapidly. Video generation models such as Sora popularized the narrative of "video models as world simulators", [5] while JEPA, World Labs, Genie, and Cosmos have pushed this vision toward spatial interaction, real-time rollout, and physical AI platforms. V-JEPA 2 learns representations from over one million hours of internet video and trains an action-conditioned world model with limited robot video, enabling zero-shot robot planning. [6] World Labs further extends world models toward spatial intelligence and interactive 3D world generation. [7] Its RTFM framework reformulates world models as real-time generative renderers, emphasizing efficiency, scalability, and persistence, while noting that interactive high-resolution world generation imposes computational demands far beyond those of current large language models.[8] DeepMind's Genie 3 can generate 720p interactive dynamic worlds at 24 fps, support real-time navigation, and maintain consistency for several minutes. [9-10] NVIDIA Cosmos embeds world foundation models into a physical AI platform covering data processing, post-training, evaluation, synthetic data generation, and closed-loop simulation.[11] Functionally and architecturally, existing world models can be grouped into four categories, as shown in Figure 1: representation-oriented models, which focus on latent states and task-relevant features; generative models, which produce future videos, scenes, and synthetic data; action-coupled models, which integrate world prediction with policies or VLA models to compare action consequences and support decision-making; and system-level models, which emphasize hierarchical deployment and coordinated rollout across the device, edge, and cloud. Together, these categories reflect the expansion of world models from state learning and future generation to action decision-making and system deployment.

These advances suggest that world models are moving beyond systems that merely "watch and predict videos" toward data and simulation infrastructure for robotics, autonomous driving, and industrial agents. Real embodied systems, however, cannot generate futures offline with unlimited computation. They must act within milliseconds to seconds under partial observations, limited energy, constrained communication bandwidth, and safety requirements. Thus, the central question is shifting from how to generate more realistic worlds to how to predict action-relevant futures at an affordable computational cost. We refer to this shared challenge as the "computational usability gap" of world models. Whereas computational efficiency, adaptive inference, real-time computing, and edge AI primarily optimize the execution of a predefined computation, computational usability addresses a higher-level question: whether and how world-model computation should be invoked and configured for downstream decision-making. It therefore treats world-model computation itself as a decision variable, including whether to invoke it, what to predict, how far and at what fidelity to roll out, where to execute it, and whether the resulting action value justifies the associated cost.

THE COMPUTATIONAL USABILITY GAP OF WORLD MODELS

The bottleneck facing the deployment of world models in real-world systems is not a single architectural issue, but an end-to-end challenge spanning data processing, training and rollout, communication and deployment, and evaluation standards. We characterize this challenge as the "computational usability gap": even with strong generative capabilities, a model may still fail to deliver actionable, verifiable, and deployable rollouts under limited data, computation, communication, and real-time safety constraints. In other words, the key question is not merely whether a world model can predict the future, but whether it can predict action-relevant futures at an affordable computational cost.

$$\begin{aligned} &\min_{\phi,\theta,\rho} \mathbb{E}_{\xi\sim\mathcal{D}_{\text{phys}}}\left[\mathcal{L}_{\text{wm}}\left(\hat{\mathbf{y}}_\xi,\mathbf{y}_\xi\right)+\lambda_c C(u_\xi)\right] \\ &\text{s.t.}\quad z_\xi=\phi(\mathbf{X}_\xi), \\ &\qquad u_\xi=\rho\left(z_\xi,g_\xi,\Omega_\xi\right)=\left(M_\xi,H_\xi,q_\xi,p_\xi\right), \\ &\qquad \hat{\mathbf{y}}_\xi=F_\theta\left(z_\xi,\mathbf{a}_\xi,g_\xi,K_\xi;u_\xi\right), \\ &\qquad u_\xi\in U(\Omega_\xi),\Omega_\xi=\left(\eta_\xi,B_\xi,K_\xi\right). \end{aligned} \tag{1}$$

The objective consists of the world-model rollout loss and the computational cost. Specifically, L_{wm} measures the discrepancy between predicted and actual futures using the token-level cross-entropy of physical-state predictions, while $C(u_\xi)$ captures the computation, latency, energy, and communication costs induced by the scheduling decision u_ξ. The coefficient λ_c controls the trade-off between rollout performance and resource consumption. The constraints characterize state representation, computation scheduling, future rollout, and resource and safety feasibility. The encoder ϕ compresses the multimodal observations X_ξ along trajectory ξ into a task-relevant state z_ξ. Given z_ξ, the task goal g_ξ, and the scheduling context Ω_ξ, the scheduler ρ jointly determines the modality set M_ξ, rollout horizon H_ξ, computational fidelity q_ξ, and deployment location p_ξ. The scheduling context is defined as $\Omega_\xi=(\eta_\xi,B_\xi,K_\xi)$, where η_ξ captures task complexity, risk, or importance, B_ξ is the available resource budget, and K_ξ denotes historical memory or a knowledge base. The world model F_θ generates the predicted future state or observation sequence $\hat{y}_\xi$, conditioned on z_ξ, the candidate action sequence a_ξ, task goal g_ξ, memory K_ξ, and scheduling decision u_ξ. The constraint $u_\xi\in U(\Omega_\xi)$ restricts the scheduling decision to the feasible set defined by the current context, thereby satisfying

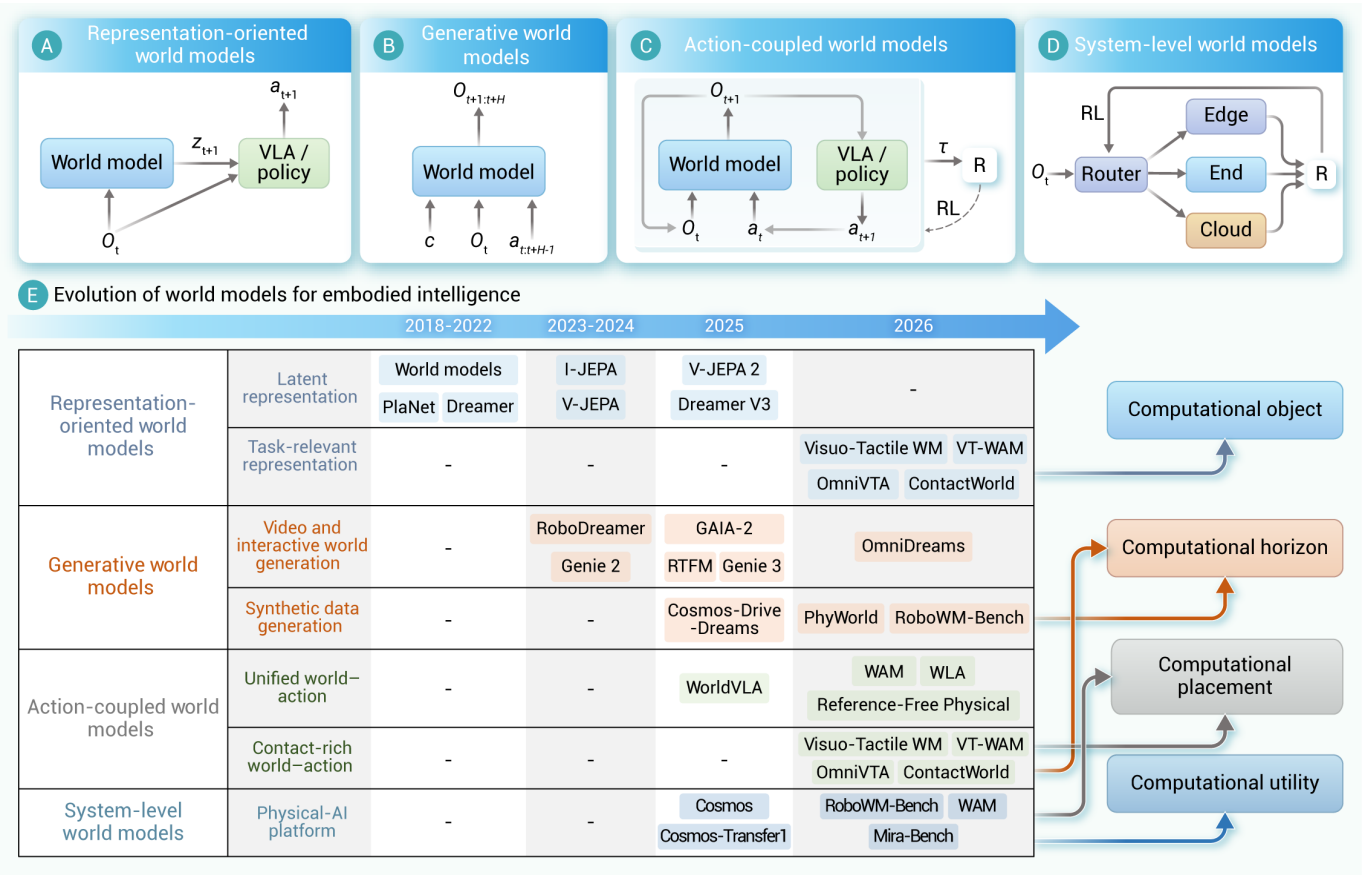


Figure 1. Architectural taxonomy, evolution, and computational-usability dimensions of world models for embodied intelligence. (A) Representation-oriented world models encode observations into compact latent or task-relevant representations for state prediction and policy planning. (B) Generative world models produce future observations, interactive environments, or synthetic data conditioned on the current observation, context, and optional action sequences. (C) Action-coupled world models integrate world prediction with a VLA or policy to evaluate action consequences and optimize decision-making through reinforcement learning. (D) System-level world models employ a routing mechanism to dynamically allocate world-model inference across end devices, edge servers, and the cloud under different latency and resource requirements. (E) Representative studies are organized according to architectural category and publication year, together with their primary connections to the four computational-usability questions: computational target, computational horizon, computational placement, and computational utility.

constraints on computation, latency, energy, communication bandwidth, and safety. By jointly optimizing the encoder parameters ϕ, world-model parameters θ, and scheduler parameters ρ, the formulation unifies four dimensions of computational usability: what to compute, how far to roll out, where to compute, and how to allocate computation according to its value.

The formulation indicates that computational usability is not merely a matter of improving prediction accuracy or reducing inference cost. Rather, it is an action-oriented computation-scheduling problem. Specifically, the scheduler adapts the rollout configuration to task context, predictive uncertainty, and resource constraints, allocating additional computation when uncertainty is high and triggering human intervention or safe fallback when confidence becomes insufficient. The objective then balances predictive performance and computational cost according to the expected value of the resulting rollout for downstream decision-making. Within this formulation, the four dimensions form a coupled closed-loop framework rather than independent categories, as shown in graphical abstract. Together, they cover the computational content, rollout scope, execution location, and value assessment of world-model computation. Given task requirements, risk or uncertainty, and available resources, the scheduler jointly configures the computational object, rollout horizon and fidelity, and computational placement. The resulting rollout is then evaluated through computational utility, which assesses its downstream action value relative to the associated resource cost and feeds this information back to subsequent scheduling decisions. The first gap lies in the coupling between physical consistency and long-horizon stability, that is, what to compute and how far to roll out. Video world models may generate visually realistic and temporally coherent short-term futures, yet still misrepresent physical details such as contact, collision, deformation, object permanence, and material changes. Over longer horizons, small errors accumulate into identity drift, structural collapse, and nonphysical motion. IntPhys 2 shows that most leading models still lag behind humans in intuitive-physics tests involving object permanence, immutability, spatiotemporal continuity, and solidity.[12] RoboWM-Bench further demonstrates that visual realism does not guarantee robotic executability: translating generated behaviors into actions often reveals errors in spatial reasoning, unstable contact prediction, and nonphysical deformation.[13] Thus, expanding computation to generate a future does not necessarily yield a stable and actionable physical state. For actionable embodied AI, computational usability therefore requires detecting degraded rollout reliability and adapting computation or triggering safe fallback before prediction errors propagate into downstream actions.

A second gap concerns a substantial mismatch between real-time requirements and deployment efficiency, that is, where to compute. Generative world models typically rely on large-scale computing resources and low-frequency offline inference, whereas drones, autonomous vehicles, and industrial systems require high-frequency closed-loop control, low-latency feedback, and safety redundancy. A model capable of generating high-quality futures in the cloud may still fail to allocate computation effectively across millisecond-level on-device control, second-level edge risk assessment, and long-horizon cloud rollout. As world models move toward real-world deployment, they can no longer assume unlimited computation; instead, they must perform selective rollout across different risk levels, timescales, and deployment locations.

Finally, a unified evaluation framework remains lacking, that is, how to assess the value of computation. If the field continues to prioritize metrics such as FVD,[14] reconstruction error, or subjective video quality, models will be optimized as better visual generators rather than effective action simulators. Evaluating world models for embodied intelligence should therefore incorporate physical consistency, long-horizon stability, task

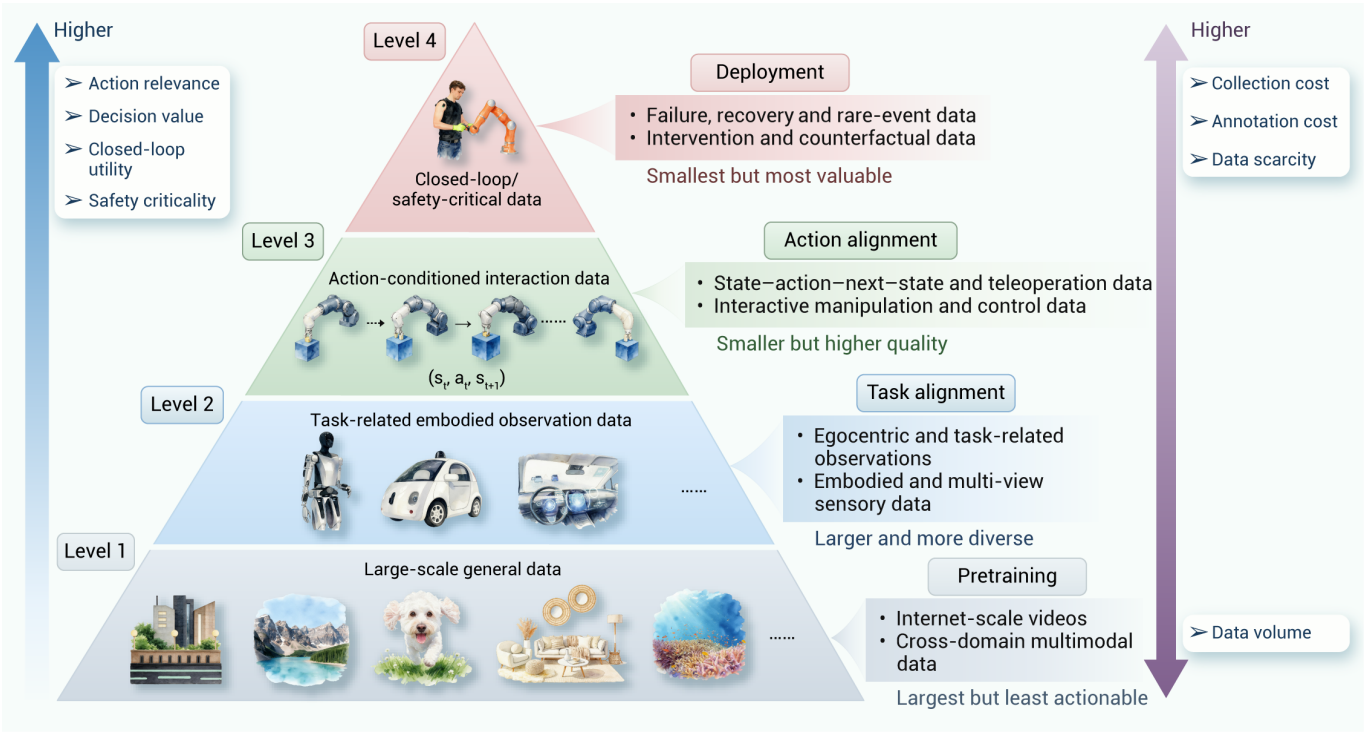


Figure 2. Action-oriented data pyramid for computationally usable world models. From general pretraining data to task-related observations, action-conditioned interactions, and closed-loop safety-critical experiences, the pyramid illustrates the transition from data scale toward action value, with increasing decision relevance and deployment utility but greater scarcity and acquisition cost.

success rate, risk-prediction accuracy, failure and risk prediction, closed-loop control gains, inference latency, energy consumption, and communication overhead. The key metric is not generation quality alone, but the action utility achieved per unit of computational cost.

COMPUTATIONAL OBJECT: DATA AND TASK-RELEVANT STATES

Data acquisition, annotation, and state representation jointly determine what a world model should compute. A computationally usable world model should not preserve every detail of the environment. Instead, it should extract action-relevant, task-specific states from high-quality data and achieve an effective abstraction between pixel-level reconstruction and opaque latent representations. Such abstractions should suppress irrelevant textures and appearance redundancy while retaining variables critical to action, including contact, load, friction, occlusion, safety boundaries, and task progress. The fundamental question is therefore not how finely the future should be predicted, but which data should be collected and which states should be represented, transmitted, and rolled out. In Eq. (1), this dimension is reflected by z_ξ and M_ξ, which determine which task-relevant information is represented and involved in subsequent rollout.

Existing studies support this view through two complementary directions: state representation and data construction. PlaNet and Dreamer showed that latent-space dynamics can substantially reduce the cost of planning and control.[2-3] I-JEPA showed that semantic representations can be learned without reconstructing pixels, whereas V-JEPA extended this principle to spatiotemporal prediction.[15-16] V-JEPA 2 suggests that large-scale videos without action labels can provide broad world priors, while a small amount of robot-interaction data can align these priors with action-conditioned prediction.[6] Complementarily, TWM demonstrates that world models can also serve as multimodal data engines, generating controllable roadside images, videos, safety-critical events, and LiDAR-style point clouds to enrich long-tail traffic data for autonomous driving.[17] Recent visuotactile world models also show that internet video alone is insufficient for learning real physical interactions. By incorporating tactile signals, multimodal interactions, and contact events, VT-WAM,[18] Visuo-Tactile World Models,[19] OmniVTA,[20] and ContactWorld[21] improve prediction in contact-rich settings and demonstrate that high-value physical data can enhance action capability more effectively than simply scaling video datasets.

World-model data systems should therefore shift from scale-first to value-first design and adopt an action-oriented data pyramid, as illustrated in Figure 2. Large-scale internet videos and simulated data form the base by providing broad world priors, while higher-value layers progressively incorporate robot interactions, failure cases, rare hazards, multimodal physical observations, and, at the top, intervention and counterfactual data. These high-level data help world models move beyond temporal correlation by learning how alternative actions change future states and outcomes. Annotation should likewise extend beyond class labels and textual descriptions to event-, state-, and risk-level supervision, covering critical events such as contact establishment, slip onset, collision boundaries, thermal anomalies, and policy interruption. Ultimately, world models should not pursue more data indiscriminately. They should organize high-quality, action-relevant data into task-specific states that are predictable, compressible, and transmissible, ensuring that limited computation directly supports subsequent rollout, control, and safety.

COMPUTATIONAL HORIZON: ACTION-CONSEQUENCE PREDICTION

The training objective of a world model should extend beyond fitting future observations to learning the action-conditioned relationships among actions, consequences, and risks. Passive video prediction asks what is most likely to happen next, whereas embodied intelligence asks what will happen under a given action and whether the associated risk is acceptable. The key issue in determining the computational horizon is therefore not simply to lengthen the rollout, but to decide—under task requirements and resource constraints—how far, at what fidelity, and over which possible futures to simulate. In this sense, world models are evolving from visual generators into action simulators. In Eq. (1), this dimension corresponds to H_ξ and q_ξ, which determine the rollout horizon and computational fidelity.

Recent studies are advancing this transition. World Action Models unify state prediction and action generation by learning their joint distribution.[22] WorldVLA integrates the world model and vision-language-action model within a unified autoregressive framework, allowing action generation and future-image prediction to reinforce each other.[23] WLA further organizes textual subtasks, subgoal images, and robot actions into a unified state-prediction framework, and enables world prediction on demand at inference time for test-time scaling.[24] In autonomous driving, GAIA-2

Table 1. Quantitative deployment trade-offs of representative methods relevant to embodied world models.

Deployment method	Computation / resource configuration	Runtime efficiency	Task performance	Evaluation setup
Low-bit quantization (OpenVLA)[31]	VRAM [GB]: 16.8 → 7.0	Throughput [actions/s]: 6.2 → 6.0	Success [%]: 71.3 → 71.9	BridgeData V2, 8 tasks; RTX 4090
Dynamic early exit (DeeR-VLA)[32]	Compute [GFLOPs/action]: 31.2 → 6.0	Latency [ms]: 55.0 → 17.5	Avg. successful sequence length [/5]: 4.07 → 4.08	CALVIN ABCD→D; NVIDIA V100
Parallel decoding & action chunking (OpenVLA-OFT)[33]	Actions/query: 1 → 8	Throughput [actions/s]: 4.2 → 109.7; latency [ms]: 239.6 → 72.9	Avg. success [%]: 76.5 → 95.3	LIBERO, 4 suites; NVIDIA A100
Latent rollout (V-JEPA 2-AC vs. Cosmos-7B)[6,11]	CEM samples/iteration: 80 → 800	Planning latency [s/action]: 240 → 16	Reach [%]: 80 → 100; P&P cup/box [%]: 0/0 → 80/50	Lab-2 manipulation; RTX 4090; CEM; H=1
On-demand world prediction (WLA-0)[24]	Active parameters: 2B	Inference latency [ms/step]: 40	RoboTwin2.0 success [%]: 92.94; RMBench success [%]: 56.5	RTX 5090; simulation + real-world evaluation
Communication-aware split inference (ComVLA)[34]	TX data [KB/step]: 1088 → 68; compute [TFLOPs]: 8.8 → 2.3	Latency [ms]: 54.8 → 42.7	Avg. success [%]: 96.9 → 95.4	LIBERO, 4 suites; H100; OpenVLA-OFT

generates controllable driving videos conditioned on ego-vehicle dynamics, traffic participants, road semantics, and environmental factors.[25] Cosmos-Drive-Dreams and OmniDreams similarly use world foundation models for long-tail scenario generation, closed-loop simulation, and interaction with driving policies.[26-27] Together, these studies show that world models are not merely visualization modules external to the policy, but decision components for comparing the future consequences of alternative actions.

A central challenge is to jointly address scale, physics, and horizon. Scaling video data may improve visual diversity, but does not necessarily ensure counterfactual consistency, physical plausibility, or long-horizon stability. PhyWorld improves physical realism through video-continuation fine-tuning and physics-preference post-training.[28] RoboWM-Bench and reference-free physical-consistency evaluation further indicate that training objectives must explicitly constrain physical executability rather than reward visual quality alone.[29] A more promising direction is risk-adaptive rollout: low-risk and familiar actions require only short-horizon, coarse simulations, whereas high-risk, uncertain, or irreversible actions should trigger longer-horizon, higher-fidelity, and multimodal prediction. A capable world model should not always compute further; it should know when additional computation is warranted, which futures to simulate, and whether the resulting predictions materially alter the decision.

COMPUTATIONAL PLACEMENT: CLOUD–EDGE–DEVICE COORDINATION

As world models move closer to real-world deployment, they can no longer assume unlimited computation. Robots, vehicles, and industrial agents already run perception, localization, planning, control, and safety monitoring concurrently. A world model implemented merely as an additional low-frequency, computationally intensive video generator is therefore unlikely to enter the control loop. A computationally usable world model should instead adopt a device–edge–cloud architecture that dynamically determines where, how often, and at what fidelity to perform rollout according to task frequency, risk level, bandwidth availability, and safety requirements. In Eq. (1), this dimension is represented by p_{ξ}, which specifies where the world-model computation is executed.

Recent studies reveal an emerging hierarchical deployment pattern. On-device computation is best suited to millisecond-level feedback loops, such as obstacle avoidance, contact regulation, slip detection, emergency stopping, and local state updates. These tasks require lightweight state prediction and confidence estimation rather than long-horizon video generation. OmniVTA integrates tactile prediction with a 60-Hz reflexive controller, demonstrating that world models can support high-frequency local control loops.[20] WLA reports an inference latency of approximately 40 ms on an RTX 5090 with 2 billion active parameters and allows world prediction to be enabled or disabled at inference time, providing a practical form of on-demand world modeling.[24] Edge systems are better suited to second-level local rollout, including task adaptation, short-term risk assessment, local map updates, and multi-robot coordination. Cloud systems can support long-horizon simulation, rare-scenario generation, data replay, model post-training, and collective experience aggregation. Cosmos integrates data curation, post-training, evaluation, and inference tools into a unified platform for robotics, autonomous driving, and industrial vision agents. Cosmos-Transfer1 further introduces an inference-scaling strategy for real-time world generation on a GB200 NVL72 rack.[30] These developments indicate that world models are evolving from individual neural networks into distributed embodied computing systems. This hierarchy favors a modular architecture rather than replicating the same large model across all levels. Lightweight encoders and short-horizon predictors can remain on device, heavier rollout modules run at the edge, and large generative models and long-horizon simulation stay in the cloud, with compact representations and adaptive routing enabling computation to shift with latency, bandwidth, and task demands.

To further quantify these deployment trade-offs, Table 1 summarizes representative methods that modify computation, rollout representation, or model placement, together with their reported resource use, runtime efficiency, and downstream task performance. The comparison shows that deployment usability cannot be characterized by resource reduction alone, but depends on how and where computation is performed and whether the resulting efficiency translates into task utility. More directly for world models, under the same RTX 4090 and CEM planner, V-JEPA 2-AC reduces planning latency from 240 to 16 s/action relative to pixel-generative Cosmos despite evaluating substantially more candidate actions, highlighting the importance of what is computed rather than computation volume alone. At the system level, ComVLA reduces transmitted data from 1088 to 68 KB/step and computation from 8.8 to 2.3 TFLOPs with only a 1.5-percentage-point decrease in average task success. Together, these results indicate that computational placement is inherently coupled with the computational object, workload, communication condition, and downstream utility rather than being an isolated device–cloud choice.

This coupling makes communication and computation scheduling central challenges for distributed world-model deployment. Real-world systems cannot continuously transmit all raw videos and sensor streams, while wireless links may further experience latency spikes, packet loss, or temporary disconnection. A more practical strategy is to communicate compressed states, anomalous events, failure segments, uncertainty tokens, and low-frequency global summaries, while allowing edge or cloud systems to trigger more intensive rollout, relabeling, and model updates when needed. When remote computation becomes temporarily unavailable, a lightweight on-device world model can serve as a short-horizon predictive buffer by rolling forward recent local states; if communication loss persists or prediction uncertainty increases, the system should fall back to a conservative local safety policy until synchronization is restored. World models therefore require the capacity for selective imagination: computing at the appropriate location and frequency for the relevant level of risk. Competition among deployable world models will ultimately depend not only on generation quality, but also on system-level optimization across rollout depth, communication overhead, on-device latency, energy budgets, and safety redundancy.

COMPUTATIONAL UTILITY: ACTION-VALUE EVALUATION

Evaluation criteria ultimately determine what world models are optimized to become. If evaluation remains centered on FVD, reconstruction error, or subjective video quality, world models will continue to evolve primarily as visual generators. They will become practical action-simulation systems only when evaluation shifts toward task success, risk reduction, closed-loop stability, and deployment cost. The central question is therefore not whether a generated future looks realistic, but whether each unit of computation

Table 2. Systematic evidence for the computational usability gap.

Computational-usability requirement	Quantitative evidence	Gap revealed
Actionable	Under MiraBench, 12/12 configurations have physical-rule adherence <26, while 5/12 achieve Task Completion Rate (TCR) >90[35]	Task success ≠ physically grounded and actionable prediction
Verifiable	The two highest-TCR configurations (100.0 and 96.0) achieve only 35.9 and 56.4 in Bias Resistance[35]	Successful-looking rollouts ≠ consequences of failure-inducing actions
Computationally affordable	Under the same RTX 4090 and CEM planner, Cosmos-7B → V-JEPA 2-AC reduces planning latency from 240 to 16 s/action, while Reach success increases from 80% to 100%[6,11]	More/heavier computation ≠ greater action utility

Note: Energy consumption remains rarely quantified together with task and closed-loop performance in the representative world-model studies considered here, indicating an important gap in end-to-end evaluation.

produces a measurable improvement in action. At the formulation level, Eq. (1) captures this dimension through the trade-off between rollout performance L_{wm} and computational cost $C(u_\xi)$, with λ_c controlling their balance.

Recent benchmarks have begun to drive this transition. World Action Models extend evaluation beyond visual fidelity to physical commonsense and action feasibility, emphasizing that world models should be assessed by their ability to support both state prediction and action generation.[22] RoboWM-Bench further shows that visual realism does not imply robotic executability, evaluating task completion by translating generated behaviors into embodied action sequences.[13] MiraBench introduces action-conditioned reliability and decomposes evaluation into physical compliance, action following, and optimistic-bias detection under failed actions, providing a more diagnostic framework for action-conditioned world models.[35] Together, these studies suggest that future evaluation should encompass visual and geometric consistency, physical plausibility, long-horizon stability, counterfactual responsiveness, closed-loop task gains, and deployment cost, rather than relying on a single video-quality metric. These observations reveal a systematic pattern. Table 2 synthesizes evidence around three requirements of computational usability—actionability, verifiability, and computational affordability—while the deployment trade-offs are quantified separately in Table 1. Together, the two tables reveal coupled gaps across predictive reliability, action utility, and system-level cost. Specifically, high nominal task success can remain decoupled from physical grounding and failure-conditioned reliability, while greater computation does not necessarily translate into higher action utility.

Building on these observations and the formulation-level trade-off in Eq. (1), we further evaluate computational utility at the decision level by relating downstream action gain to the associated resource cost. We propose action gain per computational cost (AGC) as a core concept for comparing models and platforms: $\mathrm{AGC}=\frac{\Delta J_{\mathrm{task}}}{\lambda_1\cdot\mathrm{FLOPs}+\lambda_2\cdot\mathrm{Energy}+\lambda_3\cdot\mathrm{Comm}+\lambda_4\cdot\mathrm{Latency}}$. Here, ΔJ_{task} denotes the task improvement over a model-free or lightweight baseline, such as higher success, lower risk, or fewer human interventions, while the normalized computation, energy, latency, and communication costs are weighted according to deployment requirements. As a proof of concept, we apply AGC to matched real-robot results from Efficient-WAM,[36] with traditional Motus[37] as the comparison model. Efficient-WAM achieves an AGC about 40 times that of Motus. Across domains, AGC should compare normalized relative utility rather than raw task objectives. AGC is not intended to replace all existing metrics; rather, it highlights that the value of a world model lies not in how many resources it consumes to generate futures, but in whether those resources translate into actionable gains. A more appropriate evaluation framework should form a hierarchical closed loop: the offline level assesses visual quality, physical-violation rates, and long-horizon consistency; the simulation level evaluates counterfactual rollout, failure prediction, and policy improvement; the execution level measures task success, anomaly recovery, latency, and energy consumption; and the system level examines device–edge–cloud communication overhead, multi-agent sharing gains, and safety redundancy. Such evaluation is more demanding, but it better reflects the questions that must be answered before world models can be deployed in robotics, autonomous driving, and industrial systems.

CONCLUSIONS

The competitive focus of world models is shifting from generative fidelity to action utility. The next frontier is not to generate ever more complete visual worlds, but to selectively simulate futures that can materially alter decisions under constraints on data, computation, communication, and safety. This paper introduces computational usability—encompassing what to compute, how far to roll out, where to compute, and how to assess computational value—as a central perspective for understanding this transition. Large-scale generative models will remain an important foundation. Advances such as Genie, Cosmos, GAIA, and World Labs demonstrate that world priors, synthetic data, interactive simulation, and long-tail scenario generation can substantially expand the training space of embodied intelligence. Physical AI, however, requires not unconstrained imagination, but imagination that is schedulable, verifiable, and deployable. Only when a world model can determine what to compute, how far to simulate, where to execute the computation, and whether the result is worth its cost—and translate its generative capabilities into reduced risk, improved decisions, and greater deployment efficiency—can it become genuine infrastructure for real-world physical systems.

DECLARATION OF GENERATIVE AI USE

The authors declare that no generative AI or AI-assisted technologies were used in the preparation of this manuscript.

REFERENCES

1. Ha D. and Schmidhuber J. (2018). Recurrent world models facilitate policy evolution. *Adv. Neural Inf. Process. Syst.* **31**:2451–2463. DOI:10.48550/arXiv.1809.01999
2. Hafner D., Lillicrap T., Fischer I., et al. (2019). Learning latent dynamics for planning from pixels. *Proc. Mach. Learn. Res.* **97**:2555–2565. DOI:10.48550/arXiv.1811.04551
3. Hafner D., Lillicrap T., Ba J., et al. (2020). Dream to control: Learning behaviors by latent imagination. *International Conference on Learning Representations*. DOI: 10.48550/arXiv.1912.01603
4. Hafner D., Pasukonis J., Ba J., et al. (2025). Mastering diverse control tasks through world models. *Nature* **640**:647–653. DOI:10.1038/s41586-025-08744-2
5. Brooks T., Peebles B., Holmes C., et al. (2024). Video generation models as world simulators. *OpenAI Technical Report*
6. Assran M., Bardes A., Fan D., et al. (2025). V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. Preprint at arXiv:2506.09985. DOI: 10.48550/arXiv.2506.09985
7. World Labs. (2025). Marble: A multimodal world model. *World Labs Research Release*
8. World Labs. (2025). RTFM: A real-time frame model. *Research preview*
9. Parker-Holder J., Ball P., Bruce J., et al. (2024). Genie 2: A large-scale foundation world model. *Google DeepMind Research Report*
10. Parker-Holder J. and Fruchter S. (2025). Genie 3: A new frontier for world models. *Google DeepMind Research Report*
11. NVIDIA. (2025). Cosmos world foundation model platform for physical AI. *NVIDIA Technical Report* DOI: 10.48550/arXiv.2501.03575
12. Bordes F., Garrido Q., Kao J.T., et al. (2025). IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. Preprint at arXiv:2506.09849. DOI: 10.48550/arXiv.2506.09849
13. Jiang F., Chen Y., Xu K., et al. (2026). RoboWM-Bench: A benchmark for evaluating world models in robotic manipulation. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn* Workshops: 4455–4460. DOI: 10.48550/arXiv.2604.19092
14. Unterthiner T., van Steenkiste S., Kurach K., et al. (2019). Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717. DOI: 10.48550/arXiv.1812.01717
15. Assran M., Duval Q., Misra I., et al. (2023). Self-supervised learning from images with

a joint-embedding predictive architecture. *Proc. IEEE/CVF Conf. Comput. Vis.* Pattern Recogn.: 15619–15629. DOI: 10.1109/CVPR52729.2023.01499
16. Bardes A., Garrido Q., Ponce J., et al. (2024). Revisiting feature prediction for learning visual representations from video. Trans. Mach. Learn. Res. DOI: 10.48550/arXiv.2404.08471
17. Zhang Z., Wang J., Lin C., et al. (2026). TWM: Multimodal data generation engine via traffic world model. *Communications in Transportation Research* **6**(3):9640045. DOI:10.26599/COMMTR.2026.9640045
18. Tian S., Zheng Y., Zheng Y., et al. (2026). VT-WAM: Visual-tactile world action model for contact-rich manipulation. Preprint at arXiv:2607.02503. DOI: 10.48550/arXiv.2607.02503
19. Higuera C., Arnaud S., Boots B., et al. (2026). Visuo-tactile world models. Preprint at arXiv:2602.06001. DOI: 10.48550/arXiv.2602.06001
20. Zheng Y., Gu S., Li W., et al. (2026). OmniVTA: Visuo-tactile world modeling for contact-rich robotic manipulation. Preprint at arXiv:2603.19201. DOI: 10.48550/arXiv.2603.19201
21. Zhang Z., Zhou P., Zhang K., et al. (2026). ContactWorld: What matters in vision-tactile world models for contact-rich manipulation. Preprint at arXiv:2606.13877. DOI: 10.48550/arXiv.2606.13877
22. Wang S., Shi J., Fu Z., et al. (2026). World action models: The next frontier in embodied AI. Preprint at arXiv:2605.12090. DOI: 10.48550/arXiv.2605.12090
23. Cen J., Yu C., Yuan H., et al. (2025). WorldVLA: Towards autoregressive action world model. *Adv. Neural Inf. Process. Syst (NeurIPS)* 2026 DOI: 10.48550/arXiv.2506.21539
24. Yang Y., Liu Z., Kou S., et al. (2026). World-language-action model for unified world modeling, language reasoning, and action synthesis. *Conference on Robot Learning (CoRL)* 2026 DOI: 10.48550/arXiv.2606.05979
25. Russell L., Hu A., Bertoni L., et al. (2025). GAIA-2: A controllable multi-view generative world model for autonomous driving. Preprint at arXiv:2503.20523. DOI: 10.48550/arXiv.2503.20523
26. Ren X., Lu Y., Cao T., et al. (2025). Cosmos-Drive-Dreams: Scalable synthetic driving data generation with world foundation models. Preprint at arXiv:2506.09042. DOI: 10.48550/arXiv.2506.09042
27. NVIDIA. (2026). NVIDIA OmniDreams: Real-time generative world model for closed-loop autonomous vehicle simulation. Preprint at arXiv:2606.03159. DOI: 10.48550/arXiv.2606.03159
28. Zhao P., Lin J., Rupprecht T., et al. (2026). PhyWorld: Physics-faithful world model for video generation. Preprint at arXiv:2605.19242. DOI: 10.48550/arXiv.2605.19242
29. Oh Y. and Yun S. (2026). Reference-free assessment of physical consistency in world model-based video generation. Preprint at arXiv:2606.22363. DOI: 10.48550/arXiv.2606.22363
30. NVIDIA. (2025). Cosmos-Transfer1: Conditional world generation with adaptive multimodal control. Preprint at arXiv:2503.14492. DOI: 10.48550/arXiv.2503.14492
31. Kim M.J., Pertsch K., Karamcheti S., et al. (2025). OpenVLA: An open-source vision-language-action model. *Proc. Mach. Learn. Res.* **270**:2679–2713. DOI:10.48550/arXiv.2406.09246
32. Yue Y., Wang Y., Kang B., et al. (2024). DeeR-VLA: Dynamic inference of multimodal large language models for efficient robot execution. *Adv. Neural Inf. Process. Syst (NeurIPS)*. **37**:56619–56643. DOI:10.52202/079017-1803
33. Kim M.J., Finn C. and Liang P. (2025). Fine-tuning vision-language-action models: Optimizing speed and success. *Proc. Robot. Sci. Syst.* DOI: 10.15607/RSS.2025.XXI.017
34. Liu B., Poe W.Y., Di J., et al. (2026). ComVLA: Communication-aware split inference for VLA models in 6G-connected robotics. Preprint at arXiv:2609.07838. DOI: 10.48550/arXiv.2609.07838
35. Yang T., Shen Z., Mi Z., et al. (2026). MiraBench: Evaluating action-conditioned reliability in robotic world models. Preprint at arXiv:2605.29360. DOI: 10.48550/arXiv.2605.29360
36. Li J., Guo T., Ye Y., et al. (2026). Efficient-WAM: A 1B-parameter world-action model with low-cost future imagination. Preprint at arXiv:2606.10040. DOI: 10.48550/arXiv.2606.10040
37. Bi H., Tan H., Xie S., et al. (2026). Motus: A unified latent action world model. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*: 35101–35113. DOI: 10.48550/arXiv.2512.13030.

FUNDING AND ACKNOWLEDGMENTS

Science Foundation of China (No. 72422015, 72595832 and 72595830). The funders had norole instudy design, data collection and analysis, decision to publish, or preparation of themanuscript.

AUTHOR CONTRIBUTIONS

Y.S. conducted the literature review and drafted the manuscript. J.W. conceived and supervised the study. H.L. contributed to the literature review and manuscript revision; L.L. contributed to technical discussions and critical revision; J.S. provided academic guidance and reviewed the manuscript. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.