How Should World Models Be Evaluated for Embodied Decision-Making? A Decision-Making-Centric Position

Yang Yu[1,2,*], **Shiyuan Zhang**[1,2], **Yifei Sheng**[1,2], **Haoxiang Ren**[1,2], **Haoxin Lin**[1,2,3]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] School of Artificial Intelligence, Nanjing University, Nanjing, China
[3] Cirquar Technologies, Nanjing, China

Abstract

World models have become a central abstraction in modern AI. The term now refers to several different objects: action-conditioned environment models, latent imagination models, future-video predictors, interactive neural simulators, latent predictive representations, and synthetic-data engines. Evaluation has broadened along with the term. Recent papers measure video realism, perceptual similarity, instruction following, physical plausibility, policy ranking, executability, planning success, and downstream policy improvement. This produces both metric diversity and a recurring problem of *claim/evidence mismatch*: papers sometimes make a stronger claim about what their model is useful for than their evaluation can establish.

This paper surveys the recent literature and argues that, for models presented as world models for embodied decision-making, the more decisive issue is not whether the model generates visually convincing videos, but whether it supports reliable interventional reasoning, policy evaluation, planning, and policy optimization under intervention, policy-induced distribution shift, and long-horizon rollout. This principle is actually not new: it is the *objective-mismatch* and *decision-aware model learning* lesson from model-based reinforcement learning, which already established that predictive accuracy is often a poor proxy for control utility. This paper presents (i) a survey of *evaluation practice* in the recent generative-world-model literature, (ii) a diagnosis, read off that survey, that this literature has re-encountered objective mismatch without adopting the tools developed to address it, and (iii) an operational evaluation framework and benchmark protocol. We organize the survey using an L0–L7 ladder spanning visual plausibility to policy optimization utility, noting that the levels cut across several orthogonal axes and so form an evidential hierarchy rather than a single scalar. The framework foregrounds interventional action fidelity, closed-loop rollout validity, reward/value prediction, policy-ranking agreement, optimization lift, model exploitability, and uncertainty calibration, with a minimal feasible reporting set for real-robot settings.

1 Introduction

World models have become an active theme in contemporary AI. In one line of work, inherited from model-based reinforcement learning, a world model is a learned dynamics model used for planning, imagination, policy evaluation, or policy optimization [25, 29, 26, 59]. In another line, recent embodied video-generation models are described as world models because they generate plausible future observations from text, images, videos, or actions [64, 10, 48, 36, 46]. A third line studies latent predictive representations, where the model predicts future embeddings rather than pixels [1, 45, 43]. A fourth line uses generative models as synthetic-data engines or executable video planners for robot learning [28, 2, 16, 31].

This proliferation has been productive, but it has also made evaluation ambiguous. Some papers evaluate pixel reconstruction or distributional video quality using MSE, PSNR, SSIM, LPIPS, FID, and FVD. Others evaluate instruction

*Corresponding Author: yuy@nju.edu.cn

following or physical plausibility using VLM judges, physical QA, or human preference. Others use final policy success after training inside the model. Still others use the correlation between world-model-estimated policy success and real or simulator success [49, 52, 38, 50]. These evaluations are not interchangeable. A model can be a strong video generator while being a poor environment model for control; conversely, a latent predictive model may be useful for planning without ever producing photorealistic pixels.

This observation is not new in spirit. In model-based reinforcement learning, the *objective-mismatch* problem already documented that one-step predictive likelihood is frequently uncorrelated with downstream control performance [35], and the *decision-aware model learning* line, including value-aware model learning and the value-equivalence principle, argued that models should be evaluated and trained against their eventual decision use rather than against reconstruction [30, 44, 21, 22, 58]. We therefore aim at carrying this underlying principle into the modern generative-world-model setting, where the dominant evaluation culture has drifted back toward reconstruction- and perception-style metrics, and specifically to (i) survey what the recent literature actually measures, (ii) show, from that survey, that the field has re-encountered objective mismatch without using the existing decision-aware vocabulary or diagnostics, and (iii) provide an operational framework and protocol that make the missing evidence concrete and reportable.

Our thesis is therefore conditional rather than universal. We do *not* claim that every system called a world model should be judged by policy optimization. If the intended use is future-video generation, then video quality and semantic plausibility are legitimate primary targets. The difficulty arises when evidence appropriate for one claim is used to support a stronger one. If a model is presented as a world model for embodied decision-making, then the question we find most informative is:

What would happen, in task-relevant terms, if the agent took these actions from this history?

A skeptical reader might object that visual quality, semantic plausibility, or human preference can correlate with downstream utility in some settings, or that some systems called world models are simply not intended for control. We agree on both points. Our claim is therefore comparative rather than eliminative. We do not deny the usefulness of lower-level metrics or of purely generative world-model research. We argue that, for models whose stated aim is embodied decision-making, action-, outcome-, and policy-level evaluations usually provide stronger evidence than artifact quality alone.

This view is reflected in environment-model work on counterfactual learning, policy-conditioned models, full-horizon rollout, and generalizable embodied decision-making [8, 7, 41, 69], and in recent benchmarks that explicitly separate perceptual quality from functional utility [49, 38, 50, 31]. (Throughout, where cited titles use "counterfactual", for example ACEM's "counterfactual environment-model learning", we read it as the interventional notion we adopt in Section 2.6, not as the abduction-based, fixed-noise sense.)

This paper focuses on *world models claimed for embodied decision-making*: policy evaluation, planning, policy optimization, safety testing, and related uses. It is therefore neither anti-video-metric nor anti-VLM-judge. Our claim is that, for this use case, video and semantic metrics are often better interpreted as lower-level or auxiliary diagnostics unless the claimed use is itself purely generative. We make four contributions in this paper:

1. We provide a paper-by-paper survey of the recent world-model literature, organized by *what each paper actually evaluates*, and connect it explicitly to the older objective-mismatch and decision-aware model learning literature.
2. Reading off that survey, we quantify and document a recurring failure mode: *claim/evidence mismatch*, in which lower-level evidence is informally taken to support stronger decision-making claims, while interventional and exploitability diagnostics are largely absent.
3. We organize the literature using an L0–L7 world-model evaluation ladder, ranging from visual plausibility to policy optimization utility, while clarifying that the levels span several orthogonal axes and therefore constitute an evidential hierarchy rather than a single scalar scale.
4. We propose a decision-making-centric evaluation framework and benchmark protocol built around interventional branches, policy-induced distribution shift, full-horizon outcome fidelity, policy-ranking agreement, optimization lift, exploitability, and uncertainty calibration, including a minimal feasible variant for real-robot settings.

2 Background and Symbols

2.1 A brief genealogy of the term

In the setting most relevant to this paper, the closest ancestor of the phrase "world model" is the *environment model* tradition in model-based control and reinforcement learning. In that tradition, the core object is an action-conditioned

predictive model of dynamics, rewards, and sometimes uncertainty, used to answer questions of the form: if the agent takes action a from state or history h, what is likely to happen next, and what consequences will this have for return or task completion? In this narrower sense, the conceptual target is already interventional and decision-theoretic: the model is valuable because it supports planning, policy evaluation, or policy improvement [25, 29, 26, 59].

The label "world model" became visible through work that paired compact representation learning with latent dynamics and a controller [25]. In that formulation, the world model did not mean a photorealistic simulator of everything in the environment. It referred to an internal predictive model of environment evolution, usually in a latent state space, sufficient to support control. Dreamer-style methods and real-robot extensions such as DayDreamer preserved this interpretation: the world model is primarily a tool for imagined rollouts, value estimation, and policy learning [26, 59].

The term broadened as embodied AI and large-scale generative modeling matured. A first shift came from *future-observation prediction*. In many embodied and web-scale settings, video is the most accessible supervisory signal, while explicit action or reward labels may be scarce or heterogeneous. Models that predict future frames, videos, or multimodal continuations began to be described as world models, especially when they were used to forecast how scenes evolve under language, image, video, or action conditions [10, 48, 36, 46]. In this broader usage, the word "world" often refers to the model's ability to generate plausible futures, even if the model is not directly evaluated as a policy-evaluation or policy-optimization tool.

A second shift came from *interactive neural simulators*. Once action-conditioned video models became capable of autoregressive rollout, it became natural to reuse them as surrogate environments. Systems such as UniSim, Vid2World, IRASim, WorldGym, and WorldArena sit in this intermediate region: they are still generative models of future observations, but they are also queried as if they were interactive environments [64, 27, 72, 49, 50]. This blurs the boundary between "video predictor" and "world simulator," and it is one reason evaluation becomes ambiguous: the same model can be evaluated visually in one paragraph and functionally in the next.

A third shift came from *latent predictive representations*. In JEPA-style and related approaches, the modeling target is not pixel reconstruction but future latent structure. These methods argue, implicitly or explicitly, that a useful world model may be one that predicts abstract, planning-relevant representations rather than photorealistic images [1, 45, 43]. This line breaks the assumption that the natural output of a world model is a video, and it sharpens the distinction between *observational fidelity* and *decision-relevant sufficiency*.

A fourth shift came from *world models as synthetic-data engines or executable planners*. Here the model is often not used as a general-purpose simulator in the classical sense. Instead, it may generate robot videos that are converted into actions, augment datasets with imagined trajectories, or produce demonstrations that improve downstream learning [28, 2, 16, 31]. The phrase "world model" still refers to a model of environmental evolution, but the operational role is neither pure dynamics learning nor pure video generation. It is instrumental: generate useful training signal, executable plans, or interventional data.

Viewed this way, the current literature contains not one but several partially overlapping world-model traditions. The same term now covers at least six research objects: action-conditioned environment models, latent imagination models, future-video predictors, interactive neural simulators, latent predictive representations, and synthetic-data engines. These objects overlap, but they are not identical, and their evaluations need not be identical either.

2.2 Objective mismatch and decision-aware model learning

Our position is best understood as a continuation of an existing debate in model-based reinforcement learning, rather than as a new claim. We make this lineage explicit, because the recent generative-world-model literature has largely re-derived the question without connecting to the prior answers.

Objective mismatch. The objective-mismatch problem observes that the objective used to *train* a dynamics model, typically one-step predictive likelihood or reconstruction error, is different from, and often uncorrelated with, the objective that actually matters, namely downstream control performance [35]. A globally accurate model is sufficient for planning but is neither necessary nor, under finite capacity, achievable; conversely, a model that is accurate only where it matters for the task can support strong control. The modern reliance on FID, FVD, PSNR, or VLM physics scores as primary world-model metrics is, in our reading, a re-instance of exactly this mismatch.

Decision-aware and value-equivalent model learning. A complementary line argues that model learning should be made aware of its decision role. Value-aware model learning weights model errors by their effect on value estimates rather than treating all prediction errors equally [44]. The value-equivalence principle formalizes a related intuition: two models are equivalent for a set of policies and value functions if they induce the same Bellman updates, so a

Phase	Object commonly called a world model	Why this usage emerged	Representative works	Typical evaluation emphasis
I	Action-conditioned environment model	Planning, control, off-policy evaluation, imagination-based learning	[25, 29, 26, 59]	Policy return, sample efficiency, model-based planning, value estimation
II	Latent imagination model	Need for compact long-horizon predictive state under partial observability	[25, 26, 59]	Latent rollout quality, return prediction, downstream control
III	Future-video predictor	Availability of large-scale video data; embodied tasks naturally expressed as future visual prediction	[10, 48, 36, 46]	Video fidelity, semantics, physical plausibility
IV	Interactive neural simulator	Autoregressive action-conditioned video models reused as surrogate environments	[64, 27, 72, 49, 50]	Closed-loop rollout quality, policy ranking, planning success
V	Latent predictive representation	Reaction against pixel-centric evaluation; emphasis on abstraction and planning relevance	[1, 45, 43]	Planning, probing, transfer, dense correspondence, value-relevant features
VI	Synthetic-data engine / executable planner	Generative models used instrumentally to create trajectories, demonstrations, or plans	[28, 2, 16, 31, 55]	Downstream policy lift, executability, action recovery, imitation gains

Table 1: A brief genealogy of the term "world model." The same label now covers several partially overlapping objects, which helps explain why evaluation practices have diverged.

model need only be accurate in the value-relevant subspace, which generally shrinks as the policy/function set grows [21, 22]. We use this principle in its precise form where a reward/value set is defined, and only as an informal analogy in settings (such as decoder-free latent models) where no such set is specified. A recent survey unifies these threads under the heading of decision-aware MBRL [58]. The practical consequence for evaluation is direct: the right notion of model fidelity is relative to a declared set of policies, value functions, and outcomes, which is precisely the "decision contract" we propose in Section 5.

Model exploitation. A third established phenomenon is that planners and optimizers actively seek out regions where the model is overoptimistic, so that improving average accuracy does not prevent catastrophic failures at the optimizer's chosen points [29, 35]. Our later exploitability gap (XGap, Section 5) is a measurement of exactly this effect, which is not a new concept, only as an evaluation quantity that the recent generative-world-model literature rarely reports (Section 3.6 quantifies this).

Against this background, the present paper does not propose a new training principle. It argues that the *evaluation* practices of the recent generative-world-model literature have drifted away from these lessons, and it offers a concrete framework and protocol for realigning them.

2.3 A narrow and a broad reading of "world model"

The survey suggests it is useful to distinguish two readings of the term.

Narrow reading (decision-theoretic). A world model is an action-conditioned predictive model that supports interventional reasoning about trajectories, outcomes, and values. Here the primary questions are interventional: what would happen if the agent chose this action sequence, and how good would that be?

Broad reading (predictive or generative). A world model is any model that predicts or generates future states of the world, whether in pixels, video, latent space, symbolic form, or another representation. Here the model may or may not be directly used for planning or control.

Neither reading is illegitimate, but they place pressure on different evaluations. The narrow reading foregrounds policy relevance. The broad reading licenses a wider range of predictive or generative metrics. Much of the confusion in the literature arises when evidence collected under the broad reading is discussed as if it also settled questions posed by the narrow one.

The broadening of the term can be understood as the result of several converging pressures. First, *data availability* shifted: large, varied video corpora are easier to obtain than large corpora of reset-matched action branches with reliable action and reward annotation, which encourages observation-centric training and evaluation. Second, *modeling scale* shifted: video diffusion and transformer-based generative models became strong enough that predicting future observations began to look like a plausible route to modeling the world, especially in embodied settings where the world is primarily encountered through images and video [10, 48]. Third, *use cases diversified*: some researchers want a surrogate evaluator, some a data engine, some a planner, some a representation learner, and some an open-ended video generator, and the same family of models can serve more than one role but not always equally well. Fourth, *interfaces blurred*: a model that predicts future observations under actions can be treated as a simulator; one that predicts latent futures can be treated as a planning substrate; one that generates useful rollouts can be treated as a data engine. Once these uses became common, the older environment-model language and the newer generative-world-model language merged.

2.4 Controlled-process view

Let $\mathcal{E}$ denote the real environment, which may be an MDP, a POMDP, a real robot task family, or a trusted simulator used as ground truth. A trajectory is

$$\tau = (o_0, a_0, r_0, o_1, a_1, r_1, \ldots, o_H).$$

A history at time t is

$$h_t = (o_0, a_0, \ldots, o_t).$$

A policy π maps histories to actions. Its discounted return is

$$J_{\mathcal{E}}(\pi) = \mathbb{E}_{\tau \sim \mathcal{E}, \pi}\left[\sum_{t=0}^{H-1} \gamma^t r_t\right].$$

We use $\widehat{\mathcal{E}}$ as a deliberately broad abstraction for the learned object under evaluation:

$$\widehat{\tau}_{t:t+H} \sim \widehat{\mathcal{E}}(\cdot \mid h_t, a_{t:t+H-1}, c),$$

where c may include language instructions, goals, embodiment information, camera viewpoint, latent state, or policy context. Depending on the paper, the output may be pixels, latent states, rewards, values, success probabilities, uncertainty estimates, action proposals, or some combination of these.

This notation is more general than the classical transition model $P(s_{t+1}, r_t \mid s_t, a_t)$. Some surveyed models output only videos; some output only latent states; some output only scores or judgments. The notation abstracts over these cases so that differences in evaluation can be discussed in a common language. The key point is therefore not output modality. The key question is what decisions the model is meant to support, and what evidence has been provided for that use.

2.5 Decision uses and evidential burdens

We use U to denote the intended *decision use* of the model:

$$U \in \{\text{prediction}, \text{policy evaluation}, \text{planning}, \text{policy optimization}, \text{synthetic data}, \text{representation}\}.$$

Different uses place pressure on different properties.

- If $U =$ prediction, visual fidelity, semantics, and physical plausibility may be primary.
- If $U =$ policy evaluation, policy ranking and value calibration become central.
- If $U =$ planning, action sensitivity and long-horizon outcome fidelity become important.

- If $U =$ policy optimization, exploitability, distribution shift, and uncertainty become important.
- If $U =$ synthetic data, the main question is whether generated rollouts improve downstream learning.
- If $U =$ representation, the emphasis shifts toward abstraction, probing, and planning relevance rather than decoding quality.

This is why a single use-independent world-model score is difficult to defend. Different tasks and uses make different demands, a point that the value-equivalence literature makes precise: model fidelity is only well defined relative to a set of policies and value functions [21, 22].

2.6 Open-loop, closed-loop, intervention, and identifiability

Several distinctions are central for the rest of the paper.

Open-loop vs. closed-loop. In open-loop evaluation, the model predicts future observations from a ground-truth history. In closed-loop evaluation, the policy acts on model-generated histories and the model must remain coherent under its own predictions. Many world models look strong open-loop and degrade sharply closed-loop.

Observational vs. interventional prediction. A model can predict likely futures under the behavior-policy data distribution without correctly modeling the consequences of *alternative* actions. For decision-making, the relevant object is

$$\mathbb{P}_{\mathcal{E}}(o_{t+1:t+H} \mid h_t, do(a_{t:t+H-1})),$$

not merely

$$\mathbb{P}_{\mathcal{E}}(o_{t+1:t+H} \mid h_t).$$

We use $do(\cdot)$ in the interventional sense (Pearl's second rung): we ask about the distribution of futures when actions are *set* externally, holding the history fixed. We deliberately avoid the stronger word "counterfactual" in its technical sense, which would additionally require abduction over a fixed realization of exogenous noise; the evaluations we propose are interventional, and where the term "counterfactual" appears (including in cited titles) we mean the interventional, action-branching notion.

Identifiability, and its connection to pipeline entanglement. Estimating $\mathbb{P}_{\mathcal{E}}(o_{t+1:t+H} \mid h_t, do(a_{t:t+H-1}))$ from logged data is not automatic. It generally requires (i) no unobserved confounding between the data-collection policy and the dynamics, so that conditioning on the recorded history is sufficient to block back-door paths, and (ii) adequate action coverage, so that the relevant action branches are supported in the data or can be actively queried. In fully instrumented simulators and many tabletop robot setups where the logged history captures the controller's information state and actions are recorded without hidden side-channels, the interventional and observational conditionals coincide and the quantity is identified by conditioning. This favorable case is, however, less common than the framing of the recent literature suggests. In the increasingly popular regime of learning world models from *human* video (e.g., generalist video world models and executable-video pipelines), the "action" is not recorded but *inferred* by an inverse-dynamics or retargeting model, and the recorded history is not the executed controller's information state. There the interventional target is generally *not* identified even with environment resets, because the video-to-action map is itself a learned, error-prone component. This is the same phenomenon we elsewhere call pipeline entanglement: when executability or action recovery depends on inverse dynamics, retargeting, or success checkers [31, 16, 55], an interventional claim is a claim about the whole pipeline, not about the world model in isolation. In partially observed or open-world settings these assumptions can likewise fail, and interventional error must then be measured through active intervention (resets and executed action branches) rather than inferred from passive logs. Our benchmark protocol (Section 6) is designed so that the strongest interventional claims are made only where such interventions are actually performed and the action interface is directly controlled rather than inferred.

Outcome fidelity. For models intended to inform policy choice or policy improvement, preserving task-relevant outcomes (e.g., rewards, success predicates, progress variables, and constraints) is often more important than preserving every pixel. This is, informally, the value-equivalence intuition restated at the level of evaluation [44, 21].

Representation sufficiency. A model may fail to reconstruct visual detail yet still preserve the abstract state structure needed for planning. Conversely, a model may reconstruct visually plausible futures while failing to encode the causal distinctions that matter for action selection [53].

3 What Is Being Evaluated?

This section is intentionally inventory-like. Tables 2–5 enumerate recent works, grouped by what they most directly evaluate (the levels L0–L7 are described in the next section). The goal is solely to separate distinct evaluation cultures that are often collapsed under the single phrase "world-model evaluation."

Level-assignment rule. To keep the tables auditable, we use a single mechanical rule: a paper is listed at a level if and only if it reports at least one *quantitative* metric whose target matches that level's core question (Table 7), regardless of how much emphasis the paper places on it. We do not weight by prominence and do not exercise a separate "dominant evidence" judgment; the level set in the last column is therefore intended to be reconstructable from each paper's reported metrics. Purely qualitative demonstrations (e.g., "we show controllability in Figure X" without a number) are noted in the text but not counted as occupying a level.

3.1 Benchmark and diagnostic literature

Table 2: Benchmark and diagnostic literature. The last column lists the main levels of the L0–L7 ladder occupied by the benchmark.

Work	Claimed object	What is actually evaluated	Representative metrics or outputs	Main levels
EVA-Bench [10]	embodied future-video anticipation	Offline embodied video prediction from current visual context and language; action description, next-step prediction, how-to generation, finish-thinking, and future-video quality	BLEU, METEOR, ROUGE-L, CIDEr, SPICE, CLIPScore, GPT-4o score; SC, BC, MS, FVD, GCE; EVA-Score	L1–L3
WorldSimBench [48]	video generation models as world simulators	Explicit perceptual evaluation in open embodied, driving, and manipulation settings; implicit closed-loop evaluation in Minecraft, CARLA, and CALVIN	Human Preference Evaluator scores; route completion, infractions, driving score, resource collection, CALVIN task success	L0–L3, L7
EWMBench [67]	embodied robot-video benchmark	Scene consistency, trajectory correctness, dynamics, semantics, diversity, and logical consistency in robot manipulation videos	SceneC, HSD, nDTW, DYN, Diversity, BLEU, CLIP score, logical-error penalty	L1–L3
DreamGen Bench [28]	benchmark for robot video world models used in policy learning	Instruction following and physics alignment across RoboCasa and GR1 object/behavior/environment generalization settings	GPT-4o, Qwen2.5-VL, human evaluation, VideoCon-Physics; IF, PA	L2–L3
WoW-World-Eval [14]	embodied world-model Turing test	Video quality, instruction understanding, physical law, planning DAG quality, replay executability, OOD generalization, and human deception	FVD/PSNR/SSIM/DINO/DreamSim; sequence match and execution quality; trajectory L2/DTW/FD; replay success; deceive-human ratio	L0–L4, L7
RBench [11]	robot-oriented video generation benchmark	Task correctness and embodiment-specific plausibility for common manipulation, long-horizon planning, collaboration, spatial relations, and visual reasoning	VLM/LLM judging of PSS, TAC, RSS; programmatic motion amplitude and smoothness	L0–L3
WorldArena [50]	unified benchmark for perception and functional utility	Open-loop video quality plus world model as data engine, policy evaluator, and action planner	16 video metrics; EWMScore; policy performance gain; correlation with simulator; planner success; human evaluation	L0–L4, L6, L7
RoboWM-Bench [31]	benchmark for world models in robotic manipulation	Whether generated human-hand or robot videos can be converted into executable actions that complete tasks in simulation	Task-level and step-level success; real-to-sim consistency; video-to-action replay reliability	L4, L7
WorldScore [13]	world-generation benchmark	Static and dynamic world generation under layout and camera control	Camera control, object control, content alignment, 3D consistency, photo consistency, motion accuracy, motion smoothness, WorldScore	L0–L3
WorldModelBench [36]	benchmark for judging video generators as world models	Instruction following, commonsense, and physics adherence of generated videos across domains	Human labels and trained VLM judge; instruction-following level, frame-wise/temporal quality, physics sub-scores, ELO	L2–L3
WorldPrediction [6]	high-level world modeling and procedural planning benchmark	Multiple-choice action or action-sequence selection from initial and final states	Single-step world-modeling accuracy; multi-step procedural-planning accuracy	L4, L7
AutumnBench / WorldTest [57]	environment-level query benchmark after exploration	Masked frame prediction, change detection, and planning in text-based grid-world POMDPs	MFP accuracy, change-detection score, planning success, aggregate score	L1, L4, L7

Work	Claimed object	What is actually evaluated	Representative metrics or outputs	Main levels
PBench [46]	Physical-AI image-to-video benchmark	Domain-specific physical and commonsense QA plus generic video quality	Domain score via Qwen2.5-VL QA; VBench-style quality score; overall score	L0–L3
MVP [34]	shortcut-resistant physical video QA	Minimal-pair physical understanding across human-object, robot-object, intuitive-physics, and temporal-reasoning settings	Paired minimal-pair accuracy	L3
IntPhys 2 [4]	intuitive-physics benchmark	Possible vs. impossible events in complex synthetic scenes	Overall and difficulty-split accuracy; pairwise and single-video evaluation	L3
CausalVQA [15]	causal reasoning benchmark for video models	Counterfactual, hypothetical, anticipation, planning, and descriptive reasoning on real egocentric videos	Paired accuracy, unpaired accuracy, reasoning accuracy, difficulty splits, human baseline	L3

The benchmark literature has moved beyond pure FID/FVD. Newer suites ask about instruction following, physical plausibility, trajectory correctness, executability, and policy-centric use [48, 50, 31]. Even so, many benchmark suites place most of their evaluative weight on *artifact quality* (e.g., the generated video, description, or QA answer) rather than on whether the model supports reliable policy evaluation or policy improvement. In objective-mismatch terms, these suites measure predictive or perceptual quality, which is known to be an unreliable proxy for control utility [35].

3.2 Environment-model and policy-optimization literature

Table 3: Environment-model and policy-optimization literature. These papers are closer to the original decision-making use of world models, but they vary in how directly they evaluate that use.

Work	Primary use claim	What is actually evaluated	Representative metrics or outputs	Main levels
WHALE [69]	generalizable embodied decision world model	Value estimation, video fidelity, uncertainty estimation, and downstream offline policy optimization under generalization shift	Value-estimation quality, video fidelity, uncertainty quality, policy optimization gains	L4–L7
ACEM / GALILEO [8]	counterfactual environment-model learning	Counterfactual prediction, off-policy evaluation, offline RL, and online decision making under behavior-policy bias	Counterfactual prediction accuracy, OPE quality, policy-improvement performance	L4, L6, L7
ADM-v2 [41]	full-horizon dynamics model for offline learning	Full-horizon rollout quality via off-policy evaluation and offline RL	OPE reliability, full-horizon rollout performance, offline RL returns	L4, L6, L7
PCM [7]	policy-conditioned environment model	Value estimation, policy selection, and MPC under policy-distribution shift	Value-gap reduction, policy-evaluation quality, MPC performance	L4, L6, L7
DayDreamer [59]	world model for real-robot RL	End-to-end policy learning on real robots; no standalone world-model benchmark	Policy success, sample efficiency, robot learning time	L7
UniSim [64]	interactive real-world simulator	Video-generation quality plus downstream policy learning and synthetic-video-based captioning	FID, FVD, IS, CLIP; Language Table success; CIDEr	L0–L1, L7
DiWA [5]	diffusion-policy adaptation with world models	Reward classifier quality, imagined PPO updates, and downstream policy adaptation	Precision/Recall for reward; policy success; sample efficiency	L5, L7
World4RL [32]	diffusion world model for policy refinement	Open-loop video quality plus manipulation policy refinement in simulation and real robots	FID, FVD, LPIPS; success rate; interaction cost	L0–L1, L7
VLA-RFT [37]	world-simulator RL fine-tuning for VLAs	Image prediction quality and downstream success/robustness on LIBERO	MSE, PSNR, SSIM, LPIPS; success rate; perturbation success	L1, L7
ProphRL [68]	future-video world model with VLM reward	Video prediction, optical-flow correctness, reward precision/recall, and policy success	PSNR, SSIM, tSSIM, flow EPE/cosine; RM precision/recall/FPR; success	L1, L5, L7
WMPO [73]	world-model-based policy optimization for VLA models	Policy optimization outcome, successful-trajectory length, and continual-learning performance	Success rate, successful trajectory length, lifelong-learning success	L7
World-Env [60]	simulated post-training with a world model as virtual environment	Video quality, reward-model quality, and policy success in simulation and real robots	FID, FVD, PSNR, SSIM, LPIPS; RM accuracy/precision/recall/F1; success	L1, L5, L7
World-Gymnast [51]	VLA RL inside a video world model	Mainly downstream policy performance under imagined RL	Success rate on tabletop and real-robot tasks	L7
RISE [62]	self-improving robot policy with compositional world model	Multi-view world-model quality, progress-value modeling, and real-robot policy success	PSNR, SSIM, LPIPS, FVD, EPE; success and sub-step scores	L1, L5, L7

Work	Primary use claim	What is actually evaluated	Representative metrics or outputs	Main levels
VLAW [23]	iterative co-improvement of VLA and world model	Video quality, interaction-event correctness, reward quality, and policy success	PSNR, SSIM, LPIPS, FID, FVD; TP/FN/TN/FP for interaction and reward; success	L1, L4, L5, L7
GigaBrain-0.5M [19]	VLA trained with world-model-based RL	Process reward/value prediction quality and downstream real-robot success	Inference time, MAE, MSE, RMSE, Kendall's tau; success rate	L5, L7
WoVR [33]	reliable simulator for VLA post-training	Long-horizon video quality, speed, and downstream policy success	LPIPS, FID, FVD, FloLPIPS, FPS; success rate	L1, L7
World-VLA-Loop [42]	closed-loop co-training of world model and VLA policy	World-model image quality, reward accuracy, and final success	SSIM, PSNR, LPIPS, MSE; reward accuracy; success rate	L1, L5, L7
PlayWorld [66]	world model learned from autonomous play	Video quality, progress-reward modeling, failure-mode alignment, and policy success	LPIPS, SSIM, PSNR, MSE; RM accuracy; failure-mode alignment; success	L1, L5, L7
VLA-MBPO [70]	practical model-based RL for VLA models	Image prediction, reward prediction, inference time, rollout-length ablation, and downstream success	LPIPS, PSNR, SSIM; reward ACC/F1; inference time; success	L1, L5, L7

This family lies closer to the decision-making end of the spectrum because it evaluates what policies achieve when the model is used for imagination, fine-tuning, or planning. Many papers in this family combine lower-level video diagnostics with final policy success, which leaves the contribution of the world model itself only partially isolated. The clearest exceptions are the papers that make counterfactual generalization, full-horizon OPE, policy shift, or uncertainty an explicit part of the evaluation, such as ACEM, PCM, ADM-v2, and WHALE [8, 7, 41, 69]; we note these as exemplars of the evaluation *style* we advocate, not as the only or best instances, and several policy-evaluation papers discussed below adopt a comparably decision-aligned style.

3.3 Policy evaluation, executability, and synthetic-data literature

Table 4: Policy-evaluation, executability, and synthetic-data literature. These works move from artifact quality toward decision utility, often through pipeline-level metrics.

Work	Primary use claim	What is actually evaluated	Representative metrics or outputs	Main levels
WorldGym [49]	world model as environment for policy evaluation	Qualitative rollout fidelity, qualitative action controllability, and correlation between model-evaluated and real policy success	Pearson correlation; relative policy ranking	L4, L6
Vid2World [27]	interactive world model from video diffusion	Video prediction quality and policy evaluation on robotic manipulation	FVD, FID, SSIM, PSNR, LPIPS, DreamSim; simulated vs. real success	L1, L6
Scalable Policy Evaluation [52]	video world models as policy evaluators	Video quality plus policy-value correlation and ranking fidelity	PSNR, SSIM, FVD, latent L2; Pearson correlation; MMRV	L1, L6
Gemini/Veo Simulator [18]	world simulator for policy evaluation and safety	Nominal evaluation, OOD ranking, and safety red-teaming in a video simulator	Pearson correlation, MMRV, qualitative safety findings	L6
dWorldEval [38]	robotic policy evaluation via diffusion world model	Action controllability, round-trip consistency, and policy-ranking fidelity	Δ-LPIPS, round-trip LPIPS, Pearson correlation, MMRV	L4, L6
IRASim [72]	fine-grained world model for manipulation	Video prediction, qualitative flexible controllability, policy evaluation, and model-based planning	PSNR, SSIM, latent L2, FID, FVD; Pearson correlation; planning success	L1, L4, L6, L7
DreamDojo [17]	generalist robot world model	Video prediction, human judgment of physics/action following, policy evaluation, and planning	PSNR, SSIM, LPIPS; human physics/action-following; success, Pearson, MMRV	L1, L3, L6, L7
Kinema4D [61]	4D kinematic world model	RGB rollout quality, geometry quality, and policy-evaluation quality	PSNR, SSIM, latent L2, FID, FVD, LPIPS; Chamfer, F-score, temporal F-score; success-gap to ground truth	L1, L3, L6
Persistent Robot World Models [3]	stabilized multi-step rollouts for policy evaluation	Per-camera rollout quality, masked task-relevant metrics, human preference, and policy ranking	SSIM, PSNR, LPIPS; temporal curves; masked metrics; 2AFC/ELO; Pearson, MMRV	L1, L6
DreMa [2]	synthetic imagination for imitation learning	Object final-position accuracy and downstream imitation-learning gains	Object final-position error; policy success	L4, L7
DreamGen [28]	synthetic robot-video data engine	VLM/human instruction following and physics alignment plus downstream policy learning and generalization	IF, PA; downstream policy success	L2, L3, L7

Work	Primary use claim	What is actually evaluated	Representative metrics or outputs	Main levels
Ctrl-World [24]	controllable generative world model	Video quality, qualitative action controllability, policy evaluation correlation, and policy improvement with synthetic successful trajectories	PSNR, SSIM, LPIPS, FID, FVD; evaluation correlation; success rate	L1, L4, L6, L7
RoboMaster [16]	embodied action planning from generated videos	Bridge video quality, trajectory fidelity, user preference, action-planning success, and IDM action quality	FVD, PSNR, SSIM, TrajError, user preference, planning success	L1, L4, L7
GigaWorld-0 [20]	world-model-based data engine	Physical-AI/world-generation quality, filtering score, and action recovery from generated data	PBench, DreamGen Bench, quality filtering, IDM action recovery	L0–L4
RoboVIP [54]	multi-view video augmentation for manipulation	Multi-view generation quality and downstream policy gains	FID, FVD, LPIPS, MV-Mat; success rate	L1, L7
Interactive World Simulator [56]	interactive simulator for policy training and evaluation	Video prediction, speed/stability, simulator-collected training data, and evaluation correlation	MSE, LPIPS, FID, PSNR, SSIM, UIQI, FVD; FPS/stability; success; correlation	L1, L6, L7
VLP [12]	video language planning	Human judgment of long-horizon video-plan completion and downstream execution quality	Human plan-completion rate; task reward/completion	L2, L7
Dreamitate [39]	video generation for visuomotor policy learning	Real-robot policy success using generated videos as supervision or guidance	Success rate	L7
RoboDreamer [71]	compositional world model for robot imagination	Video generation quality, human task completion judgment, and execution success	FVD; human completion judgment; RLBench success	L0, L2, L7
RoboEnvision [63]	long-horizon robot video generation	Long-horizon video quality and downstream policy-model task success	LPIPS, SSIM, PSNR, FVD, CLIP score; success rate	L1, L2, L7
Genie Envisioner [40]	unified platform for manipulation	Real-robot action-model success and EWMBench-style scene/motion/semantic world-model evaluation	Success rate; SceneC, HSD, nDTW, DYN, Diversity, BLEU, CLIP, Logics	L1, L2, L3, L7
EVA (model) [55]	executable video world model via IDM rewards	Human judgment of kinematic plausibility, interaction plausibility, instruction adherence, and execution success in sim and real robots	Human ratings; simulator success; real-robot success; IDM task success	L2, L3, L7

This family is central for a decision-making-centric paper because it contains some of the strongest current attempts to evaluate world models as *functional* objects rather than as generators. The move is away from asking only "Does the video look plausible?" and toward asking "Does the model rank policies correctly?", "Can it support planning?", or "Can generated futures be executed?" Many of these are still *pipeline metrics*: executability depends on inverse dynamics or retargeting; synthetic-data value depends on the downstream learner; policy-evaluation correlation depends on reward checkers and the policy set being ranked.

3.4 Latent and foundation-model literature

Table 5: Latent and foundation-model literature. "repr." denotes cross-cutting state-abstraction or representation diagnostics that do not map cleanly to a single rung of the ladder but that matter chiefly because they support higher-level decision claims.

Work	Primary use claim	What is actually evaluated	Representative metrics or outputs	Main levels
Cosmos Predict 2.5 [47]	diffusion-based world foundation model	Physical-AI domain competence, generic video quality, multi-view robot geometry, action-conditioned robot prediction, and DreamGen-Bench performance	Domain score, quality score, TransErr, RotErr, Sampson error, PSNR, SSIM, latent L2, FVD	L0–L4
ABot-PhysWorld [9]	interactive world foundation model for manipulation	PBench/EZS-Bench world-generation quality and action-to-video trajectory consistency	Domain/Robot score, VBench-style quality metrics, trajectory consistency	L0–L4
EA-WM [65]	event-aware generative world model	Interaction quality, trajectory accuracy, depth accuracy, perspectivity, instruction following, semantic alignment, action following, and action recoverability	VLM interaction and perspective scores; trajectory NDTW; depth accuracy; KVAF translation/rotation/gripper error	L2–L4

Work	Primary use claim	What is actually evaluated	Representative metrics or outputs	Main levels
V-JEPA 2 [1]	latent predictive world model for understanding, prediction, and planning	Qualitative decoded rollouts, robot planning, action anticipation, classification, and video QA	Goal distance, success rate, planning time, Top-1 accuracy, Recall@5, VQA accuracy	L7 + repr.
V-JEPA 2.1 [45]	dense-feature latent predictive model	Robot planning, navigation, dense prediction, action anticipation, classification, and video QA	Success rate, ATE/RTE, RMSE, mIoU, J&F, Recall@5, Top-1 accuracy	L7 + repr.
LeWorldModel [43]	latent JEPA-style world model from pixels	Latent-space MPC planning, latent physical probing, and detection of physically implausible latent trajectories	Planning success, MSE, Pearson correlation, anomaly separation	L5, L7 + repr.
Implicit World Model Evaluation [53]	formal evaluation of the state-transition structure learned by generative models	Whether the model merges histories that correspond to the same true state and separates histories with different future possibilities	Sequence compression and sequence distinction	repr.

These papers make two points. First, a world model need not be a pixel generator: V-JEPA 2, V-JEPA 2.1, and LeWorldModel are evaluated mainly through planning, probing, and downstream utility [1, 45, 43]. This echoes, informally, the value-equivalence intuition that a model can be useful by capturing the value-relevant subspace without modeling pixels [21], though these latent models typically do not define an explicit reward/value set against which equivalence is formally stated. Second, strong next-step or surface-level prediction can be misleading about state abstraction: Vafa et al. show that a model may appear strong under standard probes while still failing to recover coherent transition structure [53]. A decision world model may ultimately depend more on the right *state-transition abstraction* than on the right pixels.

3.5 What is actually being evaluated?

Taken together, Tables 2–5 suggest that the literature is not evaluating a single object.

First, there are multiple evaluation cultures. Benchmark papers such as EVA-Bench, EWMBench, WorldModelBench, PBench, RBench, and WorldScore mainly evaluate the *generated artifact*: realism, semantics, or physics [10, 67, 36, 46, 11, 13]. Policy-evaluation papers such as WorldGym, Scalable Policy Evaluation, dWorldEval, and DreamDojo evaluate the model as a *surrogate evaluator of policies* [49, 52, 38, 17]. Optimization papers evaluate the model as part of an *end-to-end control-improvement pipeline* [59, 32, 37, 73]. Latent papers evaluate the model as a *planning-relevant representation* [1, 45, 43]. Synthetic-data and executability papers evaluate the model as a *data engine or executable planner* [28, 31, 16, 55].

Second, the distribution of evidence is skewed toward lower levels, read directly off the survey. Tables 3 and 4 together list roughly forty environment-model, optimization, policy-evaluation, executability, and synthetic-data works, the subset of the survey that most plausibly carries decision-making claims. Reading the recorded metrics under the rule of Section 3, three patterns stand out. (i) Open-loop reconstruction (L1) and final policy success (L7) are the two most common rungs, and a large fraction of works report *only* this pair as their world-model evidence, leaving the model's specific contribution entangled with the optimizer, reward model, and data pipeline. (ii) Fixed-policy ranking agreement (L6) appears in only about a dozen works, almost all of them in the dedicated policy-evaluation cluster (e.g., WorldGym, Scalable Policy Evaluation, dWorldEval, DreamDojo, Persistent, Ctrl-World, IRASim). (iii) Most strikingly, we find essentially no work in these tables that reports an *exploitability gap*, a measured discrepancy between model-predicted value and real value for policies or action sequences *optimized against the model* (the XGap of Section 5). A handful report related diagnostics, such as a success-gap to ground truth for fixed rollouts (Kinema4D) or uncertainty quality (WHALE), but the specific failure mode that model-based optimization is known to provoke is, by the evidence of our own survey, almost never measured. This absence is the concrete form of the claim that the field has re-encountered objective mismatch without adopting the diagnostics developed for it. We present these counts as a reading of the metrics recorded in our tables rather than as a full audit of every paper's appendix; the qualitative conclusion is robust to small recounting differences.

Third, the main fault line is not simply "video vs. RL." The deeper distinction is between evaluating *the generated artifact* and evaluating *the decisions enabled by the model*. Many VLA/RL papers report policy success but diagnose the world model mainly through L1-style reconstruction metrics. Many benchmark papers add semantics and physics but remain artifact-centered. Policy-evaluation and executability papers move closer to decision utility, but often through black-box or pipeline-level metrics.

Fourth, pixels are neither necessary nor sufficient, and the divergence is observed, not merely possible. V-JEPA 2, V-JEPA 2.1, and LeWorldModel show that a model can be useful for planning without photorealistic reconstruction [1, 45, 43]. The converse, high visual quality with low functional utility, is exactly what the two benchmarks designed to separate these axes report. WorldArena pairs sixteen video-quality metrics with functional measures (data-engine gain, policy-evaluation correlation, planner success) and finds that perceptual ranking and functional ranking of world models do not coincide [50]; RoboWM-Bench similarly evaluates whether generated videos can be *executed* into task success and reports that video quality is not a reliable predictor of executability [31]. We emphasize that these are published findings on real model sets, not a thought experiment: the artifact-vs-utility divergence the objective-mismatch literature predicts is already being measured in the generative-world-model setting, wherever a benchmark bothers to report both axes. The illustrative example in Section 5 is included only to make the mechanism intuitive, not to establish that the divergence occurs.

3.6 The world-model claim

Different papers use the same term, "world model," while implicitly making different claims:

1. **Future-video claim:** the model predicts plausible or realistic future observations.
2. **Policy-evaluation claim:** the model can estimate or rank the performance of candidate policies.
3. **Policy-optimization claim:** using the model inside a planner, optimizer, or RL loop improves policies in the real environment.
4. **Planning/executability claim:** the model can help choose or recover action sequences that succeed in the real environment.
5. **Synthetic-data claim:** model-generated samples improve downstream learning.
6. **Representation claim:** the model's latent state preserves task-relevant transition structure sufficient for prediction and control.

These claims are not interchangeable. Evidence for one of them is not automatically evidence for another.

This observation can be made concrete in the recent literature. Benchmarks such as PBench, WorldModelBench, WorldScore, and RBench are informative if the claim is embodied world generation or physical-video quality, but they do not by themselves establish policy-evaluation or policy-optimization utility [46, 36, 13, 11]. Many RL/VLA papers such as World4RL, VLA-RFT, World-Env, RISE, WoVR, and VLA-MBPO evaluate final policy success, which is important, but often diagnose the world model itself mainly through L1-style image or video metrics [32, 37, 60, 62, 33, 70]. Policy-evaluation papers such as WorldGym, Scalable Policy Evaluation, Vid2World, dWorldEval, DreamDojo, and Persistent Robot World Models are closer to the decision-making use, yet they mostly evaluate fixed-policy ranking rather than optimizer interaction or exploitability [49, 52, 27, 38, 17, 3]. Executability papers such as RoboWM-Bench, RoboMaster, and EVA test a stronger bridge from video to control, but their performance also depends on inverse dynamics, retargeting, simulators, and success checkers [31, 16, 55]. Latent papers show the converse point: planning utility can exist without high-quality pixels [1, 43, 53].

The conclusion is not that the literature has been "doing it wrong." Rather, the field has been evaluating several different things under one name, and has often re-discovered the objective-mismatch problem without using the existing decision-aware vocabulary to address it. The next section organizes the evaluation targets explicitly.

4 World-Model Evaluation Ladder

The survey suggests an L0–L7 ladder of evaluation targets. The ladder is *descriptive*, because it summarizes what the literature already measures, and partly *normative*, because the levels answer increasingly strong questions about whether a model supports embodied decision-making. The point is not that lower levels are useless; it is that lower levels generally answer weaker questions about decision use.

The ladder spans several axes, not one. We caution at the outset that "ladder" is a simplification. The L0–L7 ordering actually moves along several partly independent axes: artifact-quality versus decision-utility (L0–L3 vs. L4–L7), observational versus interventional (L0–L3 vs. L4 onward), open-loop versus closed-loop (L1 vs. L6–L7), and fixed-policy versus optimization-induced (L6 vs. L7). These axes are correlated but not collinear. We retain the linear presentation only because it tracks, roughly, the strength of the decision-making question being asked; it should be read as an evidential hierarchy, not as a single scalar scale, and the level assignments in Tables 2–5 are sets of levels precisely because most papers occupy several axes at once.

Claim type	What stronger supporting evidence would typically include	Common weaker evidence sometimes substituted	Why the substitution can be misleading
Future-video / world-generation claim	L0–L3 evidence: visual plausibility, logged-future fidelity, semantic alignment, physical plausibility	None; these may be appropriate if this is the actual claim	The problem begins only when this evidence is later treated as if it also established policy evaluation or optimization utility.
Policy-evaluation claim	Closed-loop rollouts of fixed policies with value/ranking agreement, ideally supported by L4–L5 diagnostics	FVD/PSNR, human preference, or instruction-following scores alone	A model can generate plausible videos while misranking policies or misestimating success.
Policy-optimization claim	Real or trusted-sim policy lift under fixed budgets, plus exploitability tests and L4–L6 diagnostics	Open-loop image/video metrics, or a single downstream success number without decomposition	Final success entangles the world model with reward models, filters, optimizers, and data curation; open-loop metrics do not test interventional usefulness.
Planning / executability claim	Task success under model-based planning or action recovery, ideally with action-intervention diagnostics	Instruction-following videos or human preference for plausible plans	A plausible-looking plan can still be dynamically wrong or non-executable.
Synthetic-data claim	Controlled downstream learning gains under matched training budgets and learners	Video aesthetics, FVD, or VLM preference over generated data	Visually clean data need not contain the right task-relevant variation for learning.
Representation claim	Planning utility, physical probes, or state-abstraction diagnostics such as sequence compression/distinction	Decoder quality or linear probes alone	A model can reconstruct well or support shallow probing while still lacking the correct transition structure.

Table 6: A recurring issue in the literature is claim/evidence mismatch: lower-level evidence is sometimes taken to support stronger world-model claims than it can justify. This is the objective-mismatch problem [35] restated at the level of evaluation.

The levels are not mutually exclusive, and they are not strictly monotone in practice. A latent model might score modestly on L0 or L1 and still be useful at L7; a video model might score highly at L0–L3 and still be weak at L6. The ladder is therefore best read as a hierarchy of increasingly direct evidence for decision-making claims, not as a single linear scorecard.

L0 (Visual plausibility) asks whether the output *looks* like a plausible image or video. This is the dominant level in video-generation evaluations, where metrics such as FID, FVD, aesthetics, image quality, and human preference are natural first checks [64, 13, 46, 47, 9]. L0 is useful because obvious visual failures often indicate model collapse, temporal artifacts, or poor conditioning, and it matters when humans inspect rollouts or when generated data must pass a basic realism filter. But L0 is only indirect evidence of decision usefulness: a model can attain good L0 by producing smooth videos, copying backgrounds, or generating semantically plausible but action-insensitive futures. L0 therefore diagnoses surface quality, not whether the model gets the *consequences of chosen actions* right.

L1 (Logged-future prediction) asks whether a predicted future matches a held-out future from the behavior distribution. This is the standard open-loop world-model setting of MSE, PSNR, SSIM, LPIPS, DreamSim, latent L2, temporal SSIM, or optical-flow accuracy [32, 37, 68, 72, 61, 3]. L1 is a useful sanity check: a model that cannot predict held-out futures at all is less likely to support higher-level use. However, L1 remains observational and is exactly the quantity the objective-mismatch literature found to be weakly correlated with control performance [35]: it measures whether the model matches *what happened*, not *what would happen under different actions or policies*, and in stochastic settings it can penalize plausible alternative futures. A model can be mediocre at L1 but useful for planning if it preserves reward-relevant structure; conversely, it can be strong at L1 yet fail under policy shift.

L2 (Semantic alignment) evaluates whether the rollout matches the instruction, task, objects, and scene semantics. This level appears in EVA-Bench, DreamGen Bench, RBench, WorldArena, WoW-World-Eval, and EA-WM [10, 28, 11, 50, 14, 65], via CLIPScore, caption overlap, VLM/LLM judges, and task-completion labels. L2 is often important for language-conditioned systems, because a model that misunderstands the task is unlikely to support meaningful

Level	Name	Core question	Typical metrics	Representative works
L0	Visual plausibility	Does the output look like a realistic image or video?	FID, FVD, aesthetics, image quality, human preference, VBench-style scores	UniSim, WorldScore, PBench, Cosmos Predict 2.5, ABot-PhysWorld
L1	Logged-future prediction	Does the predicted future match held-out trajectories from the behavior distribution?	MSE, PSNR, SSIM, LPIPS, DreamSim, latent L2, tSSIM, optical-flow error	World4RL, VLA-RFT, ProphRL, IRASim, Kinema4D, Persistent
L2	Semantic alignment	Does the rollout match the instruction, task, and scene semantics?	CLIPScore, caption metrics, VLM/LLM judges, task-completion labels	EVA-Bench, DreamGen Bench, RBench, WorldArena, EA-WM
L3	Physical plausibility	Does the rollout obey intuitive physical and geometric constraints?	Physics QA, VLM physics scores, object permanence, depth, contact, trajectory, 3D consistency	WorldModelBench, PBench, MVP, IntPhys 2, CausalVQA, WorldArena
L4	Action controllability / interventional fidelity	Does changing actions produce the correct task-relevant changes?	Action-effect tests, trajectory error, Δ-LPIPS, round-trip consistency, interventional OPE diagnostics	ACEM, PCM, WHALE, dWorldEval, WorldGym, IRASim, RoboMaster
L5	Reward, value, and outcome fidelity	Does the model predict success, reward, progress, or value accurately enough for decision making?	Reward accuracy, precision/recall/F1, calibration, value error, Kendall's tau, progress ranking	DiWA, ProphRL, RISE, VLAW, GigaBrain, PlayWorld, VLA-MBPO
L6	Policy evaluation and ranking	Does model-based evaluation agree with real/simulator policy performance?	Pearson/Spearman correlation, pairwise ranking accuracy, MMRV, calibration	WorldGym, Vid2World, Scalable Policy Evaluation, dWorldEval, DreamDojo
L7	Policy optimization / planning utility	Does using the model improve decisions?	Policy lift, optimization regret, sample efficiency, planning success, safe improvement, exploitability gap	DayDreamer, ADM-v2, WMPO, DreamGen, VLP, EVA, V-JEPA 2

Table 7: The L0–L7 ladder. Levels are ordered by the strength of the question they answer about world-model usefulness for embodied decision making, but they span several orthogonal axes (artifact/decision, observational/interventional, open-loop/closed-loop, fixed-policy/optimization).

downstream use. Still, L2 remains artifact-centered: VLM judges and semantic similarity scores can reward plausible narratives rather than faithful dynamics, and a video may appear to follow the instruction while encoding the wrong object contact, force direction, or success state.

L3 (Physical plausibility) asks whether the rollout obeys intuitive physics and geometric consistency: object permanence, continuity, gravity, non-penetration, contact, depth, or trajectory coherence. This level appears in WorldModelBench, PBench, WorldArena, MVP, IntPhys 2, CausalVQA, and EA-WM [36, 46, 50, 34, 4, 15, 65]. Relative to L0 and L2, L3 is a meaningful advance because it penalizes physically impossible or causally incoherent artifacts. Yet L3 is still not enough for decision use unless it is tied to interventions: many physical benchmarks ask whether a video looks physically plausible or whether a model answers a physical question correctly, which does not directly test whether the model predicts the consequences of *the agent's* chosen actions in the regions of state-action space that matter for policy optimization.

L4 (Action controllability and interventional fidelity) asks whether changing the action causes the *correct task-relevant change*. This is explicit in ACEM, PCM, WHALE, dWorldEval, WorldGym, IRASim, and RoboMaster [8, 7, 69, 38, 49, 72, 16]. Typical evidence includes action-effect tests, end-effector or object trajectory accuracy, Δ-LPIPS, round-trip consistency, or interventional OPE diagnostics. L4 is the first level that clearly distinguishes a decision world model from a future-video prior, and for embodied decision-making it is, in our view, a strong candidate for the minimal genuinely interventional requirement. If a model appears insensitive to actions or produces the wrong action-dependent changes, then strong L0–L3 scores provide limited reassurance about decision usefulness.

L5 (Reward, value, and outcome fidelity) asks whether the model predicts success, reward, progress, constraint violation, or value accurately enough for decision making. This is explicit in DiWA, ProphRL, RISE, VLAW, GigaBrain, PlayWorld, and VLA-MBPO [5, 68, 62, 23, 19, 66, 70], via reward accuracy, precision/recall/F1, success-probability calibration, value error, or Kendall's tau for progress ranking. L5 matters because a visually imperfect model can still be useful if it preserves the variables that determine reward, whereas a photorealistic simulator that hallucinates success can be risky for policy optimization. This is, informally, the value-equivalence intuition made measurable, and a reason to treat reward and value as first-class evaluation targets rather than downstream afterthoughts [21].

L6 (Policy evaluation and ranking) asks whether model-based evaluation agrees with real or simulator policy performance. This is the focus of WorldGym, Vid2World, Scalable Policy Evaluation, dWorldEval, DreamDojo, Gemini/Veo, and Persistent Robot World Models [49, 27, 52, 38, 17, 18, 3], via Pearson or Spearman correlation, pairwise ranking accuracy, and MMRV. This is one of the stronger *fixed-policy* criteria in the current literature: it directly tests whether the model preserves the ordering of policies, which is often more decision-relevant than image similarity. But L6 is still weaker than L7, because an optimizer can drive the policy into parts of the space that were never tested by the fixed policy set.

L7 (Policy optimization and planning utility) asks the pragmatic question: does using the model improve decisions? This includes model-based planning, model-based RL, executable video planning, and synthetic-data-driven gains. DayDreamer, ADM-v2, WMPO, DreamGen, VLP, EVA, and V-JEPA 2 are representative [59, 41, 73, 28, 12, 55, 1]. L7 provides the most direct evidence for world models claimed for embodied decision-making. At the same time, L7 is *entangled*: it depends not only on the world model, but also on the reward model, the optimizer, the rollout horizon, uncertainty handling, and the surrounding data pipeline. This is why L7 is usually easier to interpret when paired with L4–L6 decompositions rather than treated as a standalone number.

The ladder suggests a useful distinction. **L0–L3 are best interpreted as diagnostic levels**: they evaluate the generated artifact and are useful and often necessary, especially for video-based interfaces. **L4 can be viewed as an interventional threshold**: it asks whether the model responds correctly to actions. **L5–L7 provide the most direct evidence of decision use**: they test whether the model preserves outcomes, ranks policies, and improves decisions. State-abstraction and latent-representation diagnostics such as [53] cut across the ladder rather than forming an extra rung; they matter chiefly because they help explain success or failure at L4–L7. For decision-making claims, lower-level success is not a reliable substitute for higher-level evidence, even though lower-level metrics remain useful in practice and may correlate with higher-level performance in some domains.

5 A Decision-Making-Centric Evaluation Framework

The ladder clarifies what kinds of evidence exist. We now turn from description to proposal. The points below are recommendations for *models whose stated purpose is embodied decision-making*, not universal requirements that every world-model paper must satisfy.

A declared decision contract clarifies evaluation. There is no single use-independent world-model score; the value-equivalence principle makes this precise, since model fidelity is only defined relative to a set of policies and value functions [21, 22]. Evaluation becomes easier to interpret once a *decision contract* is declared:

> *This model is a world model for task family $\mathcal{T}$, policy class Π, action interface $\mathcal{A}$, horizon H, and decision use U.*

Without $\mathcal{T}$, Π, H, and U, it is difficult to know whether FVD, VLM-based physics QA, policy ranking, or final task success is the most relevant primary metric. This explains a large part of the survey: many benchmark papers are reasonable once interpreted as L0–L3 contracts, and the difficulty arises only when those results are generalized to stronger L6–L7 claims without additional evidence.

Interventional action fidelity is often the first distinguishing requirement. A passive video predictor estimates likely futures under the data distribution. A decision world model is more compelling when it can answer interventional queries, distinguishing

$$\mathbb{P}_{\mathcal{E}}(o_{t+1:t+H} \mid h_t) \quad \text{from} \quad \mathbb{P}_{\mathcal{E}}(o_{t+1:t+H} \mid h_t, do(a_{t:t+H-1})).$$

The second object is the one needed for planning and policy optimization, and (per Section 2.6) is identifiable from logs only under no-unobserved-confounding and coverage assumptions, and only when the action interface is directly controlled rather than inferred from generated video; otherwise it must be measured through executed interventions. For a task-relevant feature map Φ, define the interventional prediction error

$$\mathrm{IPE}_{H,\Phi}(\widehat{\mathcal{E}}) = \mathbb{E}_{(h,a_{0:H-1})\sim\mathcal{Q}}\left[d_{\Phi}\left(\Phi(\widehat{\tau}_{1:H}), \Phi(\tau_{1:H})\right)\right],$$

where $\widehat{\tau} \sim \widehat{\mathcal{E}}(\cdot \mid h, a_{0:H-1})$ and $\tau \sim \mathcal{E}(\cdot \mid h, do(a_{0:H-1}))$. The feature map Φ may include object poses, end-effector states, contact events, success predicates, safety constraints, or latent task variables, depending on the domain. A complementary action-effect metric compares two actions from the same history:

$$\Delta_{\mathcal{E}} = d_{\Phi}\left(\Phi(\tau^{a}), \Phi(\tau^{a'})\right), \qquad \Delta_{\widehat{\mathcal{E}}} = d_{\Phi}\left(\Phi(\widehat{\tau}^{a}), \Phi(\widehat{\tau}^{a'})\right).$$

A model that preserves the magnitude and ordering of action-induced differences offers stronger evidence of decision relevance than one that merely predicts plausible futures.

Policy-induced distribution shift is usually worth including. Behavior-policy data and target-policy rollouts generally come from different state-action distributions. Let $d^{\pi}_{\mathcal{E}}(s, a)$ denote the discounted occupancy measure of policy π in the real environment, and $d^{\pi}_{\widehat{\mathcal{E}}}$ the analogous occupancy under the model. A model may have low error under $d^{\mu}_{\mathcal{E}}$, where μ is the behavior policy, but high error under $d^{\pi}_{\mathcal{E}}$, where π is the target or optimized policy. This is why policy-conditioned models, counterfactual environment-model learning, and full-horizon dynamics models are relevant here [8, 7, 41]. A decision-centric benchmark is more informative when it includes target policies that differ from the data-collection policies, ideally including policies produced by the model-based optimizer itself.

For many decision uses, full-horizon outcome fidelity is more informative than short-horizon reconstruction. One-step or short-horizon accuracy can be misleading: small local errors may compound, while visually large errors may be irrelevant to reward. For many decision uses, it is therefore more informative to evaluate the world model in terms of full-horizon task-relevant outcomes,

$$\widehat{J}_{\widehat{\mathcal{E}}}(\pi) = \mathbb{E}_{\widehat{\tau} \sim \widehat{\mathcal{E}}, \pi}\left[\sum_{t=0}^{H-1} \gamma^{t} \widehat{r}_{t}\right],$$

and, for a policy set Π_{eval}, the full-horizon value error

$$\mathrm{FVE}(\widehat{\mathcal{E}}, \Pi_{\text{eval}}) = \frac{1}{|\Pi_{\text{eval}}|} \sum_{\pi \in \Pi_{\text{eval}}} \left| \widehat{J}_{\widehat{\mathcal{E}}}(\pi) - J_{\mathcal{E}}(\pi) \right|.$$

For sparse-success tasks, success-probability calibration is also useful:

$$\widehat{p}_{\widehat{\mathcal{E}}}(\pi) = \mathbb{P}_{\widehat{\tau} \sim \widehat{\mathcal{E}}, \pi}\left[\widehat{\tau} \text{ succeeds}\right], \qquad p_{\mathcal{E}}(\pi) = \mathbb{P}_{\tau \sim \mathcal{E}, \pi}\left[\tau \text{ succeeds}\right].$$

When the model is used to compare or optimize policies, it is often more important that it preserve what the policy is trying to optimize than that it preserve every visual detail.

Closed-loop rollout and occupancy fidelity can be more informative than teacher-forced prediction. A world model used by a policy is often more meaningfully evaluated *closed-loop*: the policy acts on model-generated histories, not only on teacher-forced ground-truth prefixes. One useful target is occupancy mismatch,

$$D_{\mathrm{occ}}(\widehat{\mathcal{E}}, \pi) = \sum_{t=0}^{H} \gamma^{t} D\left(d^{\pi}_{\mathcal{E},t}, d^{\pi}_{\widehat{\mathcal{E}},t}\right),$$

where D may be MMD, Wasserstein distance, KL divergence, total variation, or a task-specific discrepancy in feature space. A world model that remains stable only while conditioned on ground-truth context may still be useful for some applications, but it offers weaker evidence as a closed-loop simulator.

Policy ranking can often be measured directly. For policy evaluation, exact value may matter less than choosing the better policy. For $\Pi_{\text{eval}} = \{\pi_1, \ldots, \pi_n\}$, pairwise ranking accuracy is

$$\mathrm{PRA} = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{1}\left[\left(\widehat{J}_{\widehat{\mathcal{E}}}(\pi_i) - \widehat{J}_{\widehat{\mathcal{E}}}(\pi_j)\right)\left(J_{\mathcal{E}}(\pi_i) - J_{\mathcal{E}}(\pi_j)\right) > 0\right],$$

and mean maximum rank violation is

$$\mathrm{MMRV} = \frac{1}{n} \sum_{i=1}^{n} \max_{j : J_{\mathcal{E}}(\pi_i) > J_{\mathcal{E}}(\pi_j)} \left[\mathrm{rank}_{\widehat{\mathcal{E}}}(\pi_i) - \mathrm{rank}_{\widehat{\mathcal{E}}}(\pi_j)\right]_{+}.$$

These are the kinds of quantities emphasized in the emerging policy-evaluation literature [49, 52, 38]; we aggregate rather than originate them.

Optimization utility is easier to interpret alongside exploitability tests. If a fixed optimizer $\mathcal{A}$ uses the world model to produce $\widehat{\pi}_{\widehat{\mathcal{E}}} = \mathcal{A}(\widehat{\mathcal{E}}, \mathcal{D}, B)$, then the primary system-level quantity is policy lift,

$$\text{Lift} = J_{\mathcal{E}}(\widehat{\pi}_{\widehat{\mathcal{E}}}) - J_{\mathcal{E}}(\pi_{\text{base}}),$$

and, when an oracle or strong reference $\pi^\star$ is available, optimization regret $\text{OptRegret} = J_{\mathcal{E}}(\pi^\star) - J_{\mathcal{E}}(\widehat{\pi}_{\widehat{\mathcal{E}}})$. A complementary exploitability metric is

$$\text{XGap} = \widehat{J}_{\widehat{\mathcal{E}}}(\widehat{\pi}_{\widehat{\mathcal{E}}}) - J_{\mathcal{E}}(\widehat{\pi}_{\widehat{\mathcal{E}}}).$$

A large positive exploitability gap indicates that the optimizer found trajectories that look good to the model but fail in the environment. This is the model-exploitation failure mode long noted in model-based RL [29, 35]; we restate it here only as an evaluation quantity, because, as the count in Section 3.6 shows, the recent generative-world-model literature essentially never reports it.

Uncertainty and abstention can affect evaluation. A world model used for optimization is easier to trust if it can indicate when it is unreliable. Let $u_{\widehat{\mathcal{E}}}(h, a_{0:H-1})$ denote an uncertainty score. A useful uncertainty measure is ideally calibrated with respect to outcome or value error:

$$\mathbb{P}\left(\left|\widehat{J}_{\widehat{\mathcal{E}}}(\pi) - J_{\mathcal{E}}(\pi)\right| \le \epsilon \;\middle|\; u_{\widehat{\mathcal{E}}}(\pi) \le \alpha\right) \approx 1 - \delta.$$

In practice, this can be reported through risk-coverage curves, error-vs.-uncertainty correlation, abstention performance, pessimistic-planning performance, Brier score, and ECE. This matters because offline or model-based optimizers may actively seek overconfident errors; WHALE's emphasis on generalization and uncertainty is therefore relevant for decision-centric evaluation rather than a side issue [69].

Estimation caveats. The quantities above are targets, not free measurements, and several are statistically demanding. $J_{\mathcal{E}}(\pi)$ on real robots requires many rollouts and carries high variance; occupancy divergences such as MMD or Wasserstein over high-dimensional trajectory distributions are hard to estimate and sensitive to the feature map Φ; and ranking metrics such as PRA and MMRV are unstable for small or poorly separated policy sets. We therefore recommend reporting these metrics with confidence intervals over tasks, seeds, and policy sets (Step 7 of the protocol), preferring low-dimensional task-relevant features inside Φ where possible, and treating a single point estimate of any of these quantities with caution. None of these caveats is unique to our proposal, but they bear directly on whether a reported decision-utility number is trustworthy.

An illustrative example of level divergence. The divergence between artifact quality and decision utility is already *measured* in practice (Section 3.6); the following stylized example is included only to make the mechanism intuitive, and its numbers are hypothetical and not drawn from any specific paper. Consider three hypothetical manipulation world models evaluated on a pick-and-place family. Model A is a high-capacity video diffusion model with excellent FVD and VLM physics scores (strong L0–L3) but which, on closed-loop rollouts, largely reproduces a smooth "default" grasp regardless of the commanded gripper action; its action-effect agreement (L4) and policy-ranking correlation (L6) are near chance. Model B has visibly worse FVD and occasional texture artifacts (weaker L0–L1) but correctly predicts whether a commanded grasp width contacts the object and whether the lift succeeds; it attains high L4 action-effect agreement, high L5 success calibration, and high L6 ranking correlation. Model C is a latent JEPA-style model with no decoder, so L0–L1 are undefined or poor, yet it supports accurate latent MPC and ranks policies well (high L4–L7). Under an artifact-weighted aggregate, Model A wins; under a decision-weighted profile, Models B and C win. The point is the structure, not the numbers: artifact quality and decision utility can be *anti-correlated* across models, exactly as the objective-mismatch literature predicts [35] and as WorldArena and RoboWM-Bench already report on real model sets [50, 31].

We recommend not letting lower levels compensate for higher-level failure. For decision-making claims, we recommend caution when aggregating lower- and higher-level metrics into a single score. A model should not rank highly as a decision world model merely because it has good FVD or VLM-based physical plausibility if it performs poorly at action controllability, reward fidelity, or policy ranking. This reflects not hostility to lower-level metrics, but the observation, documented above, that they can dominate an aggregate while answering a different question.

6 A Benchmark Protocol

The framework describes *what* may be worth measuring. This section turns it into an operational protocol, presented as a *modular template* rather than a requirement that every benchmark release in every domain include every component at full scale. Because the most decision-relevant components are also the least feasible on real hardware, we conclude the section with a minimal feasible variant and a frank statement of what it can and cannot establish.

Step 0: Declare the world-model contract

We recommend that each submission declare the decision contract in Table 8, making explicit what the model claims to support.

Field	Suggested declaration
Task family	Environments, embodiments, observation modalities, task definitions, reward/success specification.
Policy class	BC, VLA, diffusion policy, MPC planner, RL policy, human teleoperator, or other deployment policy.
Action interface	Joint control, end-effector commands, action chunks, trajectories, language actions, latent actions.
Decision use	Prediction, policy evaluation, planning, policy optimization, safety testing, synthetic data, or representation.
Task-relevant feature map Φ	The variables used to judge action effects and outcomes: object poses, contacts, safety predicates, progress variables, latent task state, etc.
Horizon	One-step, short-horizon, full-episode, or planning horizon.
Deployment regime	Offline, online, simulator, real robot, real-to-sim, or mixed.
Allowed supervision	Pixels, states, rewards, language, videos, actions, demonstrations, human labels, VLM labels.
Uncertainty interface	Ensemble variance, confidence, likelihood, pessimistic bound, or none.

Table 8: A world-model claim becomes easier to interpret once a decision contract is declared.

Step 1: Construct policy and intervention splits

A decision-centric benchmark is more informative if it does not evaluate only behavior-policy rollouts. One useful decomposition is into four policy sets:

1. Π_{beh}: the behavior policies that generated the training data;
2. Π_{anchor}: fixed anchor policies spanning weak, medium, and strong performance;
3. Π_{shift}: target policies that induce distribution shift relative to the data;
4. Π_{opt}: policies produced by optimizing inside the world model.

The benchmark may also include an intervention set $\mathcal{Q}$ of matched histories and action branches. In simulation, this can be done exactly by resetting to the same state. On real robots, one may approximate it with controlled tabletop resets, real-to-sim reconstruction, matched trajectory segments, or branchable tasks; as noted in Section 2.6, only the actually-executed branches with a directly controlled action interface license strong interventional claims.

Step 2: Report L0–L3 diagnostics as diagnostics

Open-loop diagnostic reporting remains useful, especially for video-based models (L0: realism/aesthetics/human preference; L1: MSE, PSNR, SSIM, LPIPS, DreamSim, latent L2, temporal metrics; L2: instruction-following, caption similarity, VLM task completion; L3: object permanence, non-penetration, contact, depth, 3D consistency, physical QA). These metrics are useful for debugging failure modes and understanding interfaces, but for a decision benchmark they are best reported as auxiliary diagnostics rather than as the primary score.

Step 3: Evaluate interventional action fidelity

For each $(h, a_{0:H-1}) \in \mathcal{Q}$, evaluate interventional prediction error $\mathrm{IPE}_{H,\Phi}$ and action-effect agreement. Useful components include matched-action branching from the same history; single-dimension action sweeps where meaningful; round-trip or reversibility tests when applicable; robot/object trajectory fidelity for action-conditioned models; and contact/event prediction for manipulation. This step is the operational realization of L4. If a model performs poorly here, we would hesitate to treat it as strong evidence of a decision world model, regardless of its L0–L3 scores.

Step 4: Evaluate closed-loop fixed-policy rollouts

For policies in $\Pi_{\mathrm{anchor}} \cup \Pi_{\mathrm{shift}}$, roll them out both in the real environment and inside the world model, reporting closed-loop rollout fidelity or occupancy mismatch; full-horizon value error; success-probability calibration; reward/progress

prediction; and Pearson/Spearman correlation, pairwise ranking accuracy, and MMRV. This step operationalizes L5 and L6 and is most informative when the rollouts are *closed-loop*; teacher-forced prefix conditioning alone is often insufficient.

Step 5: Evaluate policy optimization under a fixed budget

A benchmark for decision world models may include a model-based optimization challenge. Fix an optimizer $\mathcal{A}$, a data regime, a compute budget, and, if applicable, an interaction budget; depending on the declared use, $\mathcal{A}$ may be MPC, CEM, imagined RL, synthetic-data filtering, or another planner/optimizer. Evaluate the optimized policy in the real environment or a trusted simulator, reporting policy lift relative to a fixed baseline; optimization regret relative to a strong reference when available; sample efficiency and compute cost; and safe-improvement probability and constraint-violation rate. This is the operational realization of L7 and remains an end-to-end system metric.

Step 6: Adversarial exploitability and uncertainty

A stronger decision benchmark may also test the model under *adversarial use*: search for action sequences or policies that maximize predicted value under the model, subject to action constraints and safety filters, execute a safe subset in the real environment or a trusted simulator, and report the exploitability gap and failure rate. If the model provides uncertainty, evaluate whether abstaining on high-uncertainty rollouts improves calibration and optimization safety. A model with calibrated abstention may be more useful than one with superficially better raw prediction but miscalibrated confidence.

Step 7: Hidden tasks, held-out policies, and statistical reporting

To reduce benchmark overfitting, keep some tasks or objects hidden until final evaluation; evaluate on held-out policy families, not just small perturbations of the same policy class; report confidence intervals over tasks, seeds, initial states, and policy sets; and publish per-task metrics, not only averages. The estimation caveats of Section 5 make this step a substantive requirement rather than a formality.

Step 8: Report a decision-utility profile, not only a scalar

We recommend reporting a profile $(S_0, \ldots, S_7)$ rather than only a single averaged number. If a leaderboard requires a scalar, one option is to use gates,

$$S_{\mathrm{DC}} = G_4 G_5 G_6 \left(w_4 S_4 + w_5 S_5 + w_6 S_6 + w_7 S_7 \right),$$

where $G_4, G_5, G_6 \in \{0, 1\}$ are pass/fail gates for action controllability, outcome fidelity, and policy-evaluation validity. The multiplicative gating encodes one qualitative judgment, i.e., a model failing the interventional, outcome, or ranking tests should not be rescued by artifact quality, and nothing more; the form, thresholds, and choice of gating levels are not derived from theory and are domain-dependent. To make this concrete rather than a placeholder, we give one defended instantiation for tabletop manipulation with sparse success: set $G_4 = \mathbf{1}[\text{action-effect ordering accuracy} \geq 0.7]$, $G_5 = \mathbf{1}[\text{success-probability ECE} \leq 0.15]$, $G_6 = \mathbf{1}[\text{pairwise ranking accuracy} \geq 0.75]$, and weights $(w_4, w_5, w_6, w_7) = (0.2, 0.2, 0.2, 0.4)$ emphasizing optimization utility; the gate thresholds correspond to "clearly better than chance" on action ordering and ranking and "usable" calibration. These specific numbers are illustrative and should be pre-registered per domain to prevent post-hoc tuning. We recommend the profile as primary and the gated scalar only where a single number is unavoidable, and in all cases L0–L3 should not compensate for failure at L4–L7.

A minimal feasible variant for real robots, and what it does not establish

Steps 3–6 are easiest in simulation, where resets and large policy sets are cheap. On real hardware, exact interventions and many rollouts are expensive, and a full protocol may be impractical. As a concrete minimum that a real-robot world-model paper could report *today*, we suggest: (i) a small set of reset-matched action branches on a handful of tabletop tasks, scoring action-effect agreement on a low-dimensional Φ such as end-effector and object pose (a partial L4); (ii) closed-loop success calibration and ranking for three to five anchor policies of clearly different strength, with confidence intervals over a modest number of trials (a partial L5/L6); and (iii) at least a qualitative exploitability probe, reporting whether the optimizer's top model-scored trajectories actually succeed when executed (a partial XGap).

We are deliberately frank about the limits of this minimum. It is genuinely weaker than the full protocol, and on real hardware the decisive evidence, i.e., large-scale L7 policy lift and a quantitative adversarial exploitability search,

Question	What to report
What is the claimed use?	Prediction, policy evaluation, planning, policy optimization, safety testing, synthetic data, or representation.
What ladder levels are actually evaluated?	Explicitly state whether the paper provides evidence at L0, L1, . . . , L7.
What is the action interface, and is it controlled or inferred?	Low-level control, joint action, end-effector command, action chunk, trajectory, language action, or latent action; and whether actions are directly commanded or recovered by an inverse-dynamics/retargeting model.
What policy class is being evaluated or improved?	BC, VLA, diffusion policy, MPC planner, RL policy, human teleoperator, or optimized policy.
What distribution shift is tested?	Behavior-to-target, task, object, environment, embodiment, visual, or optimization-induced shift.
What horizon is evaluated?	One-step, short-horizon, full-episode, or planning horizon, including degradation curves when applicable.
What interventional data is used?	Reset-matched branches, simulator interventions, action sweeps, round-trip tests, adversarial policies, or none.
Are rewards and outcomes evaluated directly?	Reward accuracy, value error, success calibration, progress ranking, safety-constraint prediction.
Does model-based policy evaluation match reality?	Pearson/Spearman correlation, pairwise ranking accuracy, MMRV, calibration, confidence intervals.
Does model-based optimization improve reality?	Policy lift, optimization regret, sample efficiency, exploitability gap, safe-improvement probability.
Can the model be exploited?	Predicted-vs.-real gap for policies or action sequences optimized against the model.
Is uncertainty calibrated?	Risk-coverage curves, error-vs.-uncertainty correlation, abstention performance, pessimistic-planning results.

Table 9: A suggested reporting card for decision-making-centric world-model evaluation.

remains largely out of reach; in that sense, our strong recommendations are partly aspirational for the real-robot setting. We nonetheless argue the minimum is a meaningful advance over current practice for two reasons. First, by the count in Section 3.6, even items (i) and (iii) are almost never reported in the surveyed real-robot papers, which typically pair L1 reconstruction with a single end-to-end success number; adding a partial L4 and a qualitative XGap therefore supplies decision-aligned evidence that is currently missing rather than merely duplicating what the best policy-evaluation papers already do. Second, the qualitative exploitability probe in (iii) is, to our knowledge, not standard even in the strongest current policy-evaluation work (WorldGym, dWorldEval), which focuses on fixed-policy ranking; it directly targets the optimization-induced failure mode that fixed-policy L6 cannot detect. Where simulation or a trusted digital twin is available, we strongly encourage promoting (iii) to the full quantitative XGap of Step 6.

The world-model evaluation card

We recommend that every paper claiming a decision world model include an evaluation card such as Table 9.

7 Discussion and Conclusion

7.1 Why this distinction matters

If the community ranks world models primarily by L0–L3, it may optimize for the wrong target: visually convincing videos that remain unreliable for decision-making. This is the objective-mismatch problem at the level of community-wide evaluation incentives [35]. In robotics, autonomous driving, and embodied agents, this is not a cosmetic problem: a misleading world model can contribute to unsafe policy updates, misrank candidate policies, or create false confidence about robustness under distribution shift.

We do not argue that video metrics are useless. They are useful for debugging, visualization, synthetic-data filtering, and human interpretability, and they are often necessary for video-based interfaces. The concern is treating them as final evidence of decision usefulness. The same applies to VLM judges: they are useful because they provide scalable semantic and physical evaluation [48, 36, 46, 50], but they should not be treated as ground truth for policy utility, since a VLM may reward plausible-looking outcomes while missing subtle action-relevant errors. VLM-based judgments are easier to interpret when validated against executable outcomes, policy rankings, and real or trusted-sim performance.

7.2 Common objections and scope conditions

Objection 1: not every world model is for control. We agree. Some systems are best understood as future-video predictors, world generators, or representation learners. Our argument is conditional: when the stated use is embodied decision-making, action-, outcome-, and policy-level evidence becomes especially informative.

Objection 2: lower-level metrics may correlate with higher-level utility. This can happen, and in some domains better L0–L3 performance may be a good proxy for L6–L7 performance. Our objection is not to using proxies when they are empirically validated; it is to assuming such correlations hold a priori. The objective-mismatch evidence [35], and the fact that the two benchmarks reporting both axes find the rankings diverge [50, 31], are precisely reasons to validate rather than assume the correlation.

Objection 3: full interventional evaluation is impractical. This is true, especially on real robots, which is why we provide the minimal feasible variant in Section 6 and state plainly what it cannot establish. Exact reset-matched branches may be available only in simulation; partial approximations may still be useful in real-world domains.

Objection 4: the thesis risks being tautological. Readers might note that "models for decision-making should be evaluated on decision-making" is close to trivially true. The non-trivial, falsifiable content is empirical: that artifact-quality metrics frequently diverge from, and can invert, decision-utility orderings, so that the substitution is not merely incomplete but actively misleading. This is supported in the classical setting [35] and, increasingly, in the generative-world-model setting wherever a benchmark reports both axes [50, 31]; our Section 3.6 count further shows the diagnostics that would test it are largely absent, which is itself an actionable finding.

7.3 Limitations and open problems

Resettable interventional data is hard. Exact interventional evaluation requires resetting the environment to the same state and executing different actions. This is easy in simulators, difficult on real robots, and sometimes impossible in open-world settings. Real-to-sim reconstruction, branchable tabletop tasks, and matched trajectory segments are useful approximations but not perfect, and (Section 2.6) do not identify the interventional target when the action interface is inferred rather than controlled.

Task-relevant feature maps are domain dependent. The feature map Φ cannot be universal: manipulation needs object poses and contacts; driving needs lane position and safety constraints; navigation needs map and goal progress; games need latent state variables. This is a consequence of taking intended use seriously, mirroring the policy/function dependence in the value-equivalence principle [21].

Metrics carry estimation error. As discussed in Section 5, several proposed quantities are statistically demanding, and a benchmark built on noisy estimates of $J_{\mathcal{E}}(\pi)$ or occupancy divergence could itself mislead. Estimator design and variance reduction for these quantities are open problems.

Our field-level claims are read from a survey, not a full audit. The counts in Section 3.6 are read from the metrics recorded in Tables 3–4, not from an exhaustive re-implementation of every paper. We expect the qualitative conclusions (L1/L7 dominance, sparse L6, near-absent XGap) to be robust, but a systematic, criteria-locked meta-analysis is future work and would strengthen the diagnosis further.

Optimization benchmarks can themselves be gamed. If the optimization challenge is public and static, methods may overfit it. Hidden tasks, hidden policies, held-out embodiments, and adversarial exploitability tests are therefore valuable.

Latent models need special treatment. Latent predictive models should not be penalized for lacking photorealistic decoders. They are often better judged by planning, probing, outcome prediction, and state-abstraction diagnostics in their own representational space.

Safety requires worst-case evaluation. Average policy lift is not enough for safety-critical domains; constraint-violation rates, uncertainty calibration, adversarial stress tests, and worst-case failures all matter.

7.4 Conclusion

The central question is not "Can the model generate a realistic future video?" but "In what sense does the model support better decisions?" For embodied decision-making, the most informative evidence comes from whether the model preserves the interventional, long-horizon, reward-relevant structure needed for policy evaluation, planning, and policy optimization. This is the modern, generative-world-model expression of a lesson already established in model-based RL through the objective-mismatch and decision-aware model learning literature [35, 44, 21]; our contribution is the survey that shows the lesson has been forgotten in practice, the count that shows which diagnostics are missing, and the protocol that makes them reportable. Visual fidelity, semantic alignment, and physical plausibility remain valuable diagnostics, but they do not by themselves settle the stronger claim.

The survey suggests that the field has been evaluating several different objects under one name. The L0–L7 hierarchy helps separate these objects, and the proposed framework and protocol make explicit what kinds of evidence are most directly relevant to stronger decision-making claims. Our position can be stated as follows:

> *For models whose stated purpose is embodied decision-making, the strongest evidence for the label "world model" is that they enable reliable interventional evaluation and, in favorable cases, improvement of policies under intervention and distribution shift. Other evaluations remain useful, but they play a more auxiliary role for that particular claim.*

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62495090 and 62495093, the Natural Science Foundation of Jiangsu under Grants BK20243039, the "111 Center" (No. B26023), and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118). GPT-5.4 was employed to improve the language of this paper.

References

[1] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985. arXiv:2506.09985.

[2] Leonardo Barcellona, Andrii Zadaianchuk, Davide Allegro, Samuele Papa, Stefano Ghidoni, and Efstratios Gavves. Dream to manipulate: Compositional world models empowering robot imitation learning with imagination. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.14957.

[3] Jai Bardhan, Patrik Drozdik, Josef Sivic, and Vladimir Petrik. Persistent robot world models: Stabilizing multi-step rollouts via reinforcement learning, 2026. URL https://arxiv.org/abs/2603.25685. arXiv:2603.25685.

[4] Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849. arXiv:2506.09849.

[5] Akshay L. Chandra, Iman Nematollahi, Chenguang Huang, Tim Welschehold, Wolfram Burgard, and Abhinav Valada. DiWA: Diffusion policy adaptation with world models, 2025. URL https://arxiv.org/abs/2508.03645. arXiv:2508.03645.

[6] Delong Chen, Willy Chung, Yejin Bang, Ziwei Ji, and Pascale Fung. WorldPrediction: A benchmark for high-level world modeling and long-horizon procedural planning. In *ICML World Models Workshop*, 2025. URL https://openreview.net/forum?id=3GuGN0bacr.

[7] Ruifeng Chen, Xiong-Hui Chen, Yihao Sun, Siyuan Xiao, Minhui Li, and Yang Yu. Policy-conditioned environment models are more generalizable. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=g9mYBdooPA.

[8] Xiong-Hui Chen, Yang Yu, Zheng-Mao Zhu, Zhihua Yu, Zhenjun Chen, Chenghe Wang, Yinan Wu, Hongqiu Wu, Rong-Jun Qin, Ruijin Ding, and Fangsheng Huang. Adversarial counterfactual environment model learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=rHAX0LRwk8.

[9] Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Haoyun Liu, Tong Lin, Shuang Zeng, Junjin Xiao, Xinyuan Chang, Feng Xiong, Xing Wei, Zhiheng Ma, and Mu Xu. ABot-PhysWorld: Interactive world foundation model for robotic manipulation with physics alignment, 2026. URL https://arxiv.org/abs/2603.23376. arXiv:2603.23376.

[10] Xiaowei Chi, Hengyuan Zhang, Chun-Kai Fan, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chi min Chan, Wei Xue, Wenhan Luo, Shanghang Zhang, and Yike Guo. EVA: An embodied world model for future video anticipation, 2024. URL https://arxiv.org/abs/2410.15461. arXiv:2410.15461.

[11] Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruoqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking video generation model for the embodied world, 2026. URL https://arxiv.org/abs/2601.15282. arXiv:2601.15282.

[12] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.10625.

[13] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Fei-Fei Li, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2504.00983.

[14] Chun-Kai Fan, Xiaowei Chi, Xiaozhu Ju, Hao Li, Yong Bao, Yu-Kai Wang, Lizhang Chen, Zhiyuan Jiang, Kuangzhi Ge, Ying Li, Weishi Mi, Qingpo Wuwu, Peidong Jia, Yulin Luo, Kevin Zhang, Zhiyuan Qin, Yong Dai, Sirui Han, Yike Guo, Shanghang Zhang, and Jian Tang. Wow, wo, val! a comprehensive embodied world model evaluation turing test, 2026. URL https://arxiv.org/abs/2601.04137. arXiv:2601.04137.

[15] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. CausalVQA: A physically grounded causal reasoning benchmark for video models, 2025. URL https://arxiv.org/abs/2506.09943. arXiv:2506.09943.

[16] Xiao Fu, Xintao Wang, Xian Liu, Jianhong Bai, Runsen Xu, Pengfei Wan, Di Zhang, and Dahua Lin. Learning video generation for robotic manipulation with collaborative trajectory control. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0eDwYtp8n1.

[17] Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi Fan. DreamDojo: A generalist robot world model from large-scale human videos, 2026. URL https://arxiv.org/abs/2602.06949. arXiv:2602.06949.

[18] Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, Fangchen Liu, Anirudha Majumdar, Andrew Marmon, Carolina Parada, Yulia Rubanova, Dhruv Shah, Vikas Sindhwani, Jie Tan, Fei Xia, Ted Xiao, Sherry Yang, Wenhao Yu, and Allan Zhou. Evaluating gemini robotics policies in a veo world simulator, 2025. URL https://arxiv.org/abs/2512.10675. arXiv:2512.10675.

[19] GigaBrain Team, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Jie Li, Jindi Lv, Jingyu Liu, Lv Feng, Mingming Yu, Peng Li, Qiuping Deng, Tianze Liu, Xinyu Zhou, Xinze Chen, Xiaofeng Wang, Yang Wang, Yifan Li, Yifei Nie, Yilong Li, Yukun Zhou, Yun Ye, Zhichao Liu, and Zheng Zhu. GigaBrain-0.5M*: A VLA that learns from world model-based reinforcement learning, 2026. URL https://arxiv.org/abs/2602.12099. arXiv:2602.12099.

[20] GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, Qiuping Deng, Siting Wang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yankai Wang, Yu Cao, Yifan Chang, Yuan Xu, Yun Ye, Yang Wang, Yukun Zhou, Zhengyuan Zhang, Zhehao Dong, and Zheng Zhu. GigaWorld-0: World models as data engine to empower embodied AI, 2025. URL https://arxiv.org/abs/2511.19861. arXiv:2511.19861.

[21] Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3bb585ea00014b0e3ebe4c6dd165a358-Abstract.html. arXiv:2011.03506.

[22] Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/00ac8ed3b4327bdd4ebbebcb2ba10a00-Abstract.html.

[23] Yanjiang Guo, Tony Lee, Lucy Xiaoyang Shi, Jianyu Chen, Percy Liang, and Chelsea Finn. VLAW: Iterative co-improvement of vision-language-action policy and world model, 2026. URL https://arxiv.org/abs/2602.12063. arXiv:2602.12063.

[24] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.10125.

[25] David Ha and Jürgen Schmidhuber. World models, 2018. URL https://arxiv.org/abs/1803.10122. arXiv:1803.10122.

[26] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2023. URL https://arxiv.org/abs/2301.04104. arXiv:2301.04104.

[27] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pFyzqbUiF9.

[28] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DreamGen: Unlocking generalization in robot learning through neural trajectories. In *Conference on Robot Learning*, 2025. URL https://arxiv.org/abs/2505.12705.

[29] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1906.08253.

[30] Chengxing Jia, Fuxiang Zhang, Tian Xu, Jing-Cheng Pang, Zongzhang Zhang, and Yang Yu. Model gradient: unified model and policy learning in model-based reinforcement learning. *Frontiers of Computer Science*, 18:184339, 2024.

[31] Feng Jiang, Yang Chen, Kyle Xu, Yuchen Liu, Haifeng Wang, Zhenhao Shen, Jasper Lu, Shengze Huang, Yuanfei Wang, Chen Xie, and Ruihai Wu. RoboWM-Bench: A benchmark for evaluating world models in robotic manipulation, 2026. URL https://arxiv.org/abs/2604.19092. arXiv:2604.19092.

[32] Zhennan Jiang, Kai Liu, Yuxin Qin, Shuai Tian, Yupeng Zheng, Mingcai Zhou, Chao Yu, Haoran Li, and Dongbin Zhao. World4RL: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation, 2025. URL https://arxiv.org/abs/2509.19080. arXiv:2509.19080.

[33] Zhennan Jiang, Shangqing Zhou, Yutong Jiang, Zefang Huang, Mingjie Wei, Yuhui Chen, Tianxing Zhou, Zhen Guo, Hao Lin, Quanlu Zhang, Yu Wang, Haoran Li, Chao Yu, and Dongbin Zhao. WoVR: World models as reliable simulators for post-training VLA policies with RL, 2026. URL https://arxiv.org/abs/2602.13977. arXiv:2602.13977.

[34] Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-QA benchmark for physical understanding via minimal video pairs. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2506.09987.

[35] Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control (L4DC)*, volume 120 of *Proceedings of Machine Learning Research*, pages 761–770, 2020. URL https://proceedings.mlr.press/v120/lambert20a.html. arXiv:2002.04523.

[36] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models, 2025. URL https://arxiv.org/abs/2502.20694. arXiv:2502.20694.

[37] Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, and Weihua Su. VLA-RFT: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators, 2025. URL https://arxiv.org/abs/2510.00406. arXiv:2510.00406.

[38] Yaxuan Li, Zhongyi Zhou, Yefei Chen, Yaokai Xue, and Yichen Zhu. dWorldEval: Scalable robotic policy evaluation via discrete diffusion world model, 2026. URL https://arxiv.org/abs/2604.22152. arXiv:2604.22152.

[39] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. In *Conference on Robot Learning*, 2024. URL https://arxiv.org/abs/2406.16862.

[40] Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie envisioner: A unified world foundation platform for robotic manipulation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fHLtSxDFKC.

[41] Haoxin Lin, Siyuan Xiao, Yi-Chen Li, Zhilong Zhang, Yihao Sun, Chengxing Jia, and Yang Yu. ADM-v2: Pursuing full-horizon roll-out in dynamics models for offline policy learning and evaluation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ICbXEwqpga.

[42] Xiaokang Liu, Zechen Bai, Hai Ci, Kevin Yuchen Ma, and Mike Zheng Shou. World-VLA-Loop: Closed-loop learning of video world model and VLA policy, 2026. URL https://arxiv.org/abs/2602.06508. arXiv:2602.06508.

[43] Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312. arXiv:2603.19312.

[44] Amir massoud Farahmand, André Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pages 1486–1494, 2017. URL https://proceedings.mlr.press/v54/farahmand17a.html.

[45] Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning, 2026. URL https://arxiv.org/abs/2603.14482. arXiv:2603.14482.

[46] NVIDIA. PBench: A physical AI benchmark for world models, 2025. URL https://research.nvidia.com/labs/cosmos-lab/pbench/.

[47] NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jiaojiao Fan, Linxi Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, Jacob Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezanali, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shuran Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne Tchapmi, Wei-Cheng Tseng, Jibin Varghese, Andrew Z. Wang, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Jiashu Xu, Dinghao Yang, Xiaodong Yang, Haotian Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World simulation with video foundation models for physical AI. Technical report, NVIDIA, 2025. URL https://research.nvidia.com/labs/dir/cosmos-predict2.5/. Technical report; project page branded as Cosmos-Predict2.5.

[48] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. WorldSimBench: Towards video generation models as world simulators, 2024. URL https://arxiv.org/abs/2410.18072. arXiv:2410.18072.

[49] Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation, 2025. URL https://arxiv.org/abs/2506.00613. arXiv:2506.00613.

[50] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong

Tian, Tat-Seng Chua, Wenwu Zhu, and Yong Li. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models, 2026. URL https://arxiv.org/abs/2602.08971. arXiv:2602.08971.

[51] Ansh Kumar Sharma, Yixiang Sun, Ninghao Lu, Yunzhe Zhang, Jiarao Liu, and Sherry Yang. World-Gymnast: Training robots with reinforcement learning in a world model, 2026. URL https://arxiv.org/abs/2602.02454. arXiv:2602.02454.

[52] Wei-Cheng Tseng, Jinwei Gu, Qinsheng Zhang, Hanzi Mao, Ming-Yu Liu, Florian Shkurti, and Lin Yen-Chen. Scalable policy evaluation with video world models, 2025. URL https://arxiv.org/abs/2511.11520. arXiv:2511.11520.

[53] Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model, 2024. URL https://arxiv.org/abs/2406.03689. arXiv:2406.03689.

[54] Boyang Wang, Haoran Zhang, Shujie Zhang, Jinkun Hao, Mingda Jia, Qi Lv, Yucheng Mao, Zhaoyang Lyu, Jia Zeng, Xudong Xu, and Jiangmiao Pang. RoboVIP: Multi-view video generation with visual identity prompting augments robot manipulation, 2026. URL https://arxiv.org/abs/2601.05241. arXiv:2601.05241.

[55] Ruixiang Wang, Qingming Liu, Yueci Deng, Guiliang Liu, Zhen Liu, and Kui Jia. EVA: Aligning video world models with executable robot actions via inverse dynamics rewards, 2026. URL https://arxiv.org/abs/2603.17808. arXiv:2603.17808.

[56] Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive world simulator for robot policy training and evaluation, 2026. URL https://arxiv.org/abs/2603.08546. arXiv:2603.08546.

[57] Archana Warrier, Dat Nguyen, Michelangelo Naim, Moksh Jain, Yichao Liang, Karen Schroeder, Cambridge Yang, Joshua B. Tenenbaum, Sebastian Vollmer, Kevin Ellis, and Zenna Tavares. Benchmarking world-model learning with environment-level queries, 2025. URL https://arxiv.org/abs/2510.19788. arXiv:2510.19788.

[58] Ran Wei, Nathan Lambert, Anthony McDonald, Alfredo Garcia, and Roberto Calandra. A unified view on solving objective mismatch in model-based reinforcement learning. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=tQVZgvXhZb. arXiv:2310.06253.

[59] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. DayDreamer: World models for physical robot learning. In *Conference on Robot Learning*, 2023. URL https://arxiv.org/abs/2206.14176.

[60] Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. World-Env: Leveraging world model as a virtual environment for VLA post-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2509.24948.

[61] Mutian Xu, Tianbao Zhang, Tianqi Liu, Zhaoxi Chen, Xiaoguang Han, and Ziwei Liu. Kinema4D: Kinematic 4d world modeling for spatiotemporal embodied simulation, 2026. URL https://arxiv.org/abs/2603.16669. arXiv:2603.16669.

[62] Jiazhi Yang, Kunyang Lin, Jinwei Li, Wencong Zhang, Tianwei Lin, Longyan Wu, Zhizhong Su, Hao Zhao, Ya-Qin Zhang, Li Chen, Ping Luo, Xiangyu Yue, and Hongyang Li. RISE: Self-improving robot policy with compositional world model, 2026. URL https://arxiv.org/abs/2602.11075. arXiv:2602.11075.

[63] Liudi Yang, Yang Bai, George Eskandar, Fengyi Shen, Mohammad Altillawi, Dong Chen, Soumajit Majumder, Ziyuan Liu, Gitta Kutyniok, and Abhinav Valada. RoboEnvision: A long-horizon video generation model for multi-task robot manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2025. URL https://arxiv.org/abs/2506.22007.

[64] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06114.

[65] Zhaoyang Yang, Yurun Jin, Lizhe Qi, Cong Huang, and Kai Chen. EA-WM: Event-aware generative world model with structured kinematic-to-visual action fields, 2026. URL https://arxiv.org/abs/2605.06192. arXiv:2605.06192.

[66] Tenny Yin, Zhiting Mei, Zhonghe Zheng, Miyu Yamane, David Wang, Jade Sceats, Samuel M. Bateman, Lihan Zha, Apurva Badithela, Ola Shorinwa, and Anirudha Majumdar. PlayWorld: Learning robot world models from autonomous play, 2026. URL https://arxiv.org/abs/2603.09030. arXiv:2603.09030.

[67] Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. EWMBench: Evaluating scene, motion, and semantic quality in embodied world models, 2025. URL https://arxiv.org/abs/2505.09694. arXiv:2505.09694.

[68] Jiahui Zhang, Ze Huang, Chun Gu, Zipei Ma, and Li Zhang. ProphRL: Reinforcing action policies by prophesying, 2025. URL https://arxiv.org/abs/2511.20633. arXiv:2511.20633.

[69] Zhilong Zhang, Ruifeng Chen, Junyin Ye, Yihao Sun, Pengyuan Wang, Jingcheng Pang, Kaiyuan Li, Tianshuo Liu, Haoxin Lin, Yang Yu, and Zhi-Hua Zhou. WHALE: Towards generalizable and scalable world models for embodied decision-making, 2024. URL https://arxiv.org/abs/2411.05619. arXiv:2411.05619.

[70] Zhilong Zhang, Haoxiang Ren, Yihao Sun, Yifei Sheng, Haonan Wang, Haoxin Lin, Zhichao Wu, Pierre-Luc Bacon, and Yang Yu. Towards practical world model-based reinforcement learning for vision-language-action models, 2026. URL https://openreview.net/forum?id=gB1yFEd106. ICLR 2026 World Models Workshop.

[71] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2404.12377.

[72] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9844, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_IRASim_A_Fine-Grained_World_Model_for_Robot_Manipulation_ICCV_2025_paper.html.

[73] Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. WMPO: World model-based policy optimization for vision-language-action models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qE2FyvRvuF.