KineBench: Benchmarking Embodied World Models via IDM-Free Kinematic Grounding

Zeyu Liu[1,2⋆], Zhangzhe Zhu[1,3⋆], Yang Zhang[1,4⋆⋆⋆], Chenyou Fan[1,3]
Chenjia Bai[1,5], and Xuelong Li[1]

[1] Institute of Artificial Intelligence (TeleAI), China Telecom, China
[2] National University of Singapore, Singapore
[3] Fudan University, China
[4] Tsinghua University, China
[5] Shenzhen Research Institute of Northwestern Polytechnical University, China

Abstract. Evaluating the physical consistency of embodied world models (EWMs) is a critical open challenge. While closed-loop evaluation via simulator rollouts offers a more faithful assessment of physical plausibility than open-loop alternatives, existing frameworks almost exclusively rely on Inverse Dynamics Models (IDMs) for action extraction. Due to the intricate mapping from 2D pixel space to 3D kinematic space, the learned IDMs can be brittle to data outside their training distribution, resulting in unreliable action extraction from the generated videos with novel objects and scenarios. This creates an unavoidable attribution ambiguity between world model inaccuracies and extractor errors. To reduce this ambiguity, we present **KineBench**, an IDM-free closed-loop benchmark for EWMs, built upon an explicit kinematic grounding pipeline. Given a generated video, KineBench employs cascaded visual foundation models to directly extract 6D end-effector poses from individual frames, which are then executed in a physics simulator for closed-loop validation. This explicit grounding directly tests the physical feasibility rather than visual plausibility, while remaining sensitive to general physical hallucinations such as gripper vanishing or spatial inconsistency. Beyond execution-based task success, KineBench incorporates two classical 3D kinematic metrics—Spectral Arc Length (SPARC) and the Maruyama Manipulability Index—to characterize trajectory smoothness and kinematic feasibility from a robot-centric perspective. Across the evaluated models and tasks, these metrics exhibit task- and model-dependent associations with physical success rates, suggesting that they provide complementary diagnostic signals for assessing embodied generation quality. Built on 20 diverse manipulation tasks in ManiSkill3, KineBench evaluates EWMs across four progressive suites: basic execution, task transfer, visual out-of-distribution generalization, and complexity-conditioned scaling. Evaluation across frontier models reveals task-complexity-bounded nonlinear scaling in embodied video generation, providing empirical guidance for future data-scaling strategies.

⋆ indicates equal contribution.
⋆⋆ Project Lead
Correspondence to: Chenjia Bai <baicj@chinatelecom.cn>

The code and datasets are available on GitHub at `https://github.com/minecraft-zzz/KineBench` and on Hugging Face at `https://huggingface.co/datasets/Zorkzak/KineBenchDatasets`.

Keywords: Benchmarking Embodied World Models · Closed-Loop Evaluation · 3D Kinematics · Robot-centric Metrics · Physical Plausibility

1 Introduction

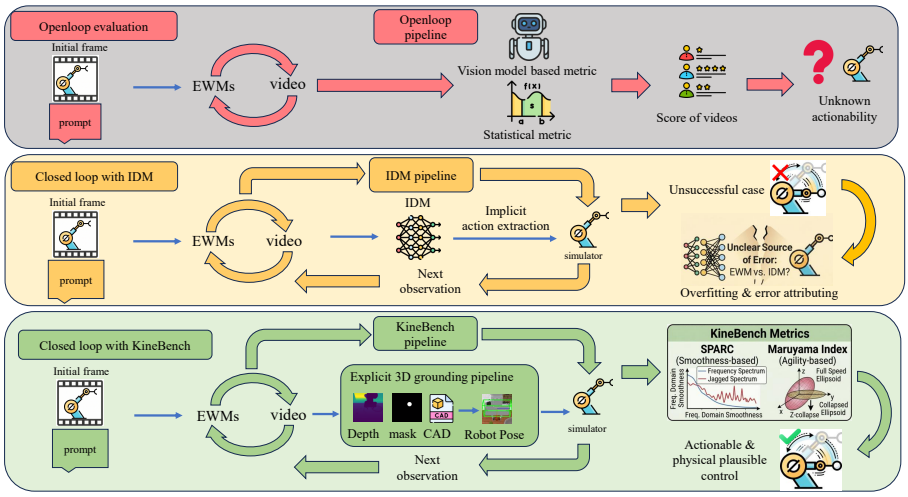


Fig. 1: Comparison of evaluation paradigms for Embodied World Models. KineBench provides an IDM-free, 3D kinematic grounding pipeline for rigorous physical validation, resolving issues of physical actionability and attribution ambiguity.

Constructing Embodied World Models (EWMs) has emerged as a core objective in Embodied AI, aiming to endow agents with a predictive understanding of environmental dynamics [11–14, 22, 23, 41, 42]. Driven by recent advancements in scalable generative architectures, visual foundation models trained on internet-scale data have demonstrated remarkable spatiotemporal predictive capabilities [3, 19, 27, 31, 35]. Consequently, an active line of research seeks to adopt these video generation models directly as predictive priors for physical intelligence and embodied control [7, 40]. These models serve two complementary roles in embodied control: as pre-training backbones for Vision-Language-Action models, where their richer spatiotemporal priors are hypothesized to yield stronger action policies [5, 15, 37, 43]; and as model-based planners that provide predictive rollouts to guide downstream task execution [6]. However, this paradigm shift exposes a critical evaluation issue: *how can we rigorously assess the actionable potential and physical plausibility of these video foundation models, rather than visual generation quality alone?*

Existing video generation benchmarks primarily evaluate pixel-level fidelity and spatiotemporal consistency [16, 30], which fail to capture the physical plausibility and kinematic feasibility required for embodied control. Recent embodied benchmarks have gone further by leveraging vision models [44] and Large Vision-Language Models (VLMs) [8] to assess semantic correctness in an open-loop setting. Beyond open-loop evaluation, a few frameworks attempt to ground generated sequences into physical simulation via simulators, providing a more direct assessment of executability [10, 40]. To close this loop in a model-agnostic manner, these frameworks almost exclusively have to rely on learned Inverse Dynamics Models (IDMs), which is either end-to-end neural mappings or explicit tracking-based approaches, to extract robot actions from generated RGB frames. However, these closed-loop frameworks face a significant methodological bottleneck: how to extract robot actions from generated frames. Consequently, existing closed-loop benchmarks almost exclusively rely on Inverse Dynamics Models (IDMs)—either end-to-end neural mappings or explicit tracking-based approaches—to recover robot actions from RGB frames.

This IDM-based protocol, however, introduces a fundamental confounding factor. Due to the intricate mapping from 2D pixel space to 3D kinematic space, IDMs overfit the specific action distributions in their training data and frequently fail when faced with novel trajectories produced by generative models (see Sec. 4.2). This creates an unavoidable attribution ambiguity: when closed-loop execution fails, it is impossible to determine whether the failure stems from the world model generating physically implausible visuals, or simply from the IDM's poor generalization to unseen trajectories.

To reduce this attribution ambiguity, we introduce **KineBench**, an IDM-free closed-loop evaluation framework for EWMs based on explicit kinematic grounding. Rather than inferring actions through an implicit inverse dynamics model, KineBench recovers end-effector motion through a modular geometric pipeline. Figure 1 compares three evaluation paradigms. Given a language instruction and an initial observation, the world model generates a predictive video sequence. KineBench then applies a cascade of visual perception models to estimate the 6D end-effector pose from each generated frame, without requiring any modification to the evaluated world model. The recovered pose sequence is subsequently executed in a physics simulator for closed-loop evaluation. The CAD-constrained pose-tracking procedure also provides a geometric regularization effect: rigid-body constraints can suppress small frame-level jitters and local visual artifacts, while larger inconsistencies, such as gripper disappearance or discontinuous spatial motion, tend to produce unstable or infeasible pose estimates. In this sense, the procedure exhibits a low-pass filtering effect analogous to that of an admittance controller [20], although its reliability remains dependent on the underlying segmentation, depth estimation, and pose-tracking modules.

Beyond execution-based assessment, KineBench provides a **robot-centric** evaluation perspective grounded in 3D kinematics. Rather than assessing generated videos solely through visual appearance, we characterize the physical properties of the generated motions in 3D space. Specifically, KineBench in-

corporates *Spectral Arc Length (SPARC)* to measure trajectory smoothness in the frequency domain and the *Maruyama Manipulability Index* to characterize kinematic dexterity and feasibility under the robot model.

Across the evaluated models and tasks, these metrics exhibit task- and model-dependent associations with simulator execution success, suggesting that trajectory smoothness and kinematic feasibility provide complementary signals for assessing embodied generation quality. To the best of our knowledge, KineBench is the first framework to incorporate these classical robotics metrics into the evaluation of generative video models.

To systematically probe the capability boundaries of EWMs, we design 20 diverse manipulation tasks in ManiSkill3, organized into four evaluation suites of increasing diagnostic depth: basic execution capability, task-level transfer, visual out-of-distribution generalization, and complexity-conditioned scaling.

Across the evaluated frontier models, we observe complexity-conditioned nonlinear scaling in embodied video generation. In contrast to the broadly reported scaling trends of visual and language foundation models [18], the marginal benefits of increasing data and compute diminish as task difficulty rises, providing empirical evidence to inform future scaling strategies for embodied world models.

In summary, our main contributions are as follows:

- **IDM-free closed-loop benchmark.** We present **KineBench**, a closed-loop benchmark for EWMs that connects generated videos to physics simulation through an explicit 6D end-effector pose grounding pipeline. By replacing learned IDM-based action extraction with modular geometric estimation, KineBench reduces the attribution ambiguity between world-model errors and action-extraction errors.
- **Robot-centric kinematic evaluation paradigm.** We incorporate SPARC and the Maruyama Manipulability Index into generative video assessment. Across the evaluated models and tasks, these metrics exhibit task- and model-dependent associations with execution success, providing complementary diagnostic signals for embodied generation quality.
- **Structured four-suite benchmark and scaling analysis.** We design 20 diverse manipulation tasks in ManiSkill3, organized into four evaluation suites, covering basic execution, task transfer, visual OOD generalization, and complexity-conditioned scaling. Experiments across frontier models show that the marginal benefits of increasing data and compute diminish as task difficulty rises, providing empirical evidence to inform future scaling strategies for EWMs.

2 Related Work

2.1 Video Models as Embodied World Models

Large-scale video foundation models have demonstrated remarkable spatiotemporal predictive capabilities, learning rich physical dynamics from internet-scale data [3, 19, 31]. This has catalyzed a novel research paradigm: adopting video

generation models as world models for embodied agents [4, 9, 36]. Under this paradigm, these models serve as predictive priors for robots, autoregressively imagining future physical interaction sequences conditioned on current observations and language instructions. Several works have further explored using such models as pre-training backbones for Vision-Language-Action models [5, 15, 37, 43], hypothesizing that richer physical priors yield stronger downstream action policies. Despite these advances, rigorously evaluating their generation quality and physical plausibility remains an open challenge in the field.

2.2 Benchmarks for Embodied World Models

Open-loop visual benchmarks. Early video benchmarks primarily assess pixel-level fidelity and aesthetic quality [16, 25, 29]. Recent embodied-specific benchmarks have gone further: EWMBench [44] proposes a dedicated framework evaluating visual scene consistency, motion correctness, and semantic alignment for robotic manipulation via VLM-based assessment. RBench [8] introduces a large-scale evaluation covering five task domains and four robot embodiments, benchmarking 25 models with metrics including structural consistency and action completeness. These benchmarks represent significant progress but remain confined to open-loop, pixel-space evaluation, unable to verify physical feasibility.

Physics-grounded open-loop benchmarks. To measure physical plausibility without physical simulation, existing approaches formulate physics-grounded prompt suites and employ multimodal evaluators to decouple semantic adherence from physical commonsense [2, 24]. Advanced paradigms bypass external judges by leveraging surrogate likelihoods from diffusion processes or extracting verifiable optical flow proxies to enforce Newtonian kinematics via reward constraints [21, 39]. However, these metrics still measure visual plausibility rather than actual physical feasibility.

Closed-loop execution benchmarks. A few benchmarks ground generated sequences into physical simulation for closed-loop evaluation [10, 26, 40]. World-in-World [40] introduces a unified closed-loop interface and demonstrates that visual quality does not reliably correlate with task success—underscoring the necessity of execution-based evaluation. WoW-World-Eval [10] further proposes an IDM-based Turing Test to assess whether generated videos can fool a dynamics model trained on real trajectories. However, all existing closed-loop benchmarks rely on IDMs for action extraction, inheriting the generalization limitations we identify in Sec. 4.2. KineBench is the first to entirely eliminate this dependency via explicit kinematic grounding.

2.3 Evaluation Metrics: From Pixels to 3D Kinematics

Early video quality metrics measure frame-level pixel distribution distances or structural similarities [29, 33]. Subsequently, metrics based on 2D point tracking or optical flow were introduced to capture temporal consistency [16]. For real-world robotic manipulation, however, movement smoothness and kinematic dexterity are the core indicators of robust control. Spectral Arc Length (SPARC) [1]

has been widely validated in clinical and robotic settings as a more robust smoothness metric than traditional jerk-based measures, as it is insensitive to high-frequency noise. The Yoshikawa and Maruyama Manipulability Index [38] evaluates a robot's ability to avoid kinematic singularities and exert multi-directional forces across configurations. To the best of our knowledge, KineBench is the first work to introduce these classical 3D kinematic metrics into the evaluation of generative video models.

3 Methodology

To evaluate the actionable capabilities and physical consistency of EWMs, we introduce **KineBench**, an IDM-free closed-loop evaluation framework comprising three components: (1) an explicit kinematic grounding pipeline that extracts 6D end-effector poses directly from generated frames; (2) a robot-centric kinematic evaluation system grounded in 3D metrics; and (3) a four-suite benchmark of 20 diverse manipulation tasks in ManiSkill3.

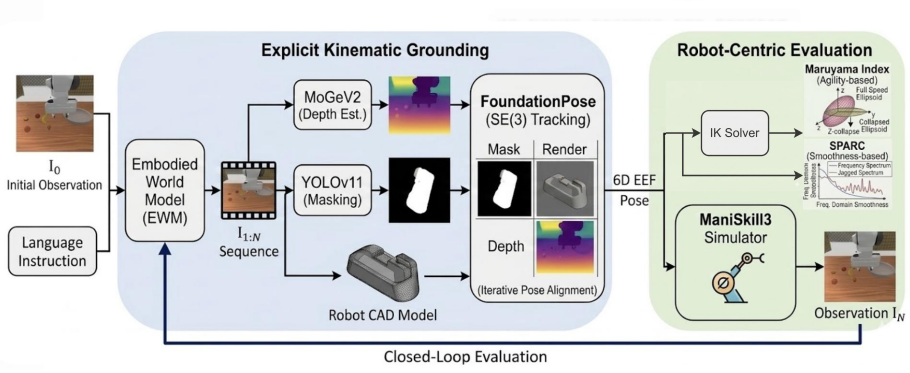


Fig. 2: KineBench evaluation pipeline

3.1 Explicit Kinematic Grounding Pipeline

The inherent generalization limitations of IDMs introduce significant confounding factors into closed-loop evaluation, as discussed in Sec. 1. To reduce this bottleneck, KineBench constructs a pipeline integrating 2D instance segmentation, monocular depth estimation (MoGeV2 [32]), and 6D pose tracking (FoundationPose [34]). The overall KineBench pipeline is illustrated in Figure 2

– **Masking and Depth Recovery.** Prior to 6D pose extraction, accurate 2D localization and 3D spatial grounding are required. We employ a fine-tuned YOLO model [28] to segment the end-effector mask. Since the exact CAD models of the gripper are readily available from manufacturers, no manual annotation is required. We then apply a fine-tuned MoGeV2 within the simulation domain to produce high-precision absolute metric depth maps.

– **CAD-based Pose Matching and Geometric Filtering.** Given the segmentation mask, CAD model, and depth map, we apply FoundationPose to extract the 6D pose of the gripper. Its core mechanism follows a Render-and-Compare paradigm: the pose within $SE(3)$ space is iteratively refined by minimizing the photometric and geometric residuals between rendered CAD templates and the generated observation frames.

Crucially, while generated videos inevitably exhibit high-frequency temporal jitter or local non-rigid pixel deformations, FoundationPose enforces strict alignment with rigid-body constraints in 3D space. Consequently, this end-effector-centric approach naturally absorbs negligible non-physical pixel noise while remaining acutely sensitive to genuine physical hallucinations, e.g., gripper vanishing or extreme spatial inconsistencies that fundamentally violate rigid-body mechanics. The extracted 6D poses are then injected into ManiSkill3, which executes the actions, renders the next observation, and initiates the subsequent generation-and-rollout cycle.

3.2 Robot-Centric Kinematic Evaluation

Traditional video generation metrics, such as PSNR and FVD [29], operate exclusively within the 2D pixel domain, making them inadequate for assessing the physical plausibility of robotic motion. Because KineBench explicitly recovers 3D end-effector poses, we can shift the evaluation paradigm from visual appearance to kinematic fidelity. To this end, we incorporate two classical 3D kinematic metrics to measure trajectory smoothness and spatial dexterity, establishing a direct standard for evaluating an embodied world model's physical understanding.

– **Spectral Arc Length (SPARC).** Traditional movement smoothness metrics rely on high-order temporal derivatives of position signals, which disproportionately amplify the minute high-frequency pose noise in video generation, causing the metrics invalid. SPARC resolves this issue by shifting the analytical domain to the frequency domain. It is defined as the arc length of the normalized Fourier magnitude spectrum:

$$\text{SPARC} \triangleq -\int_0^{\omega_c} \left[\left(\frac{1}{\omega_c} \right)^2 + \left(\frac{d\hat{V}(\omega)}{d\omega} \right)^2 \right]^{\frac{1}{2}} d\omega$$

where $\hat{V}(\omega)$ is the normalized velocity magnitude spectrum, and ω_c is the adaptive cutoff frequency. Physically, smooth actions adhering to the dynamic principle of minimum energy concentrate their energy in low frequencies, resulting in a flat spectrum curve with a SPARC value approaching 0; conversely, if the model generates hallucinations such as stuttering or teleportation, high-frequency harmonics increase, the curve becomes convoluted, and the SPARC value significantly decreases. In this context, SPARC emerges as an exceptional standard for scrutinizing the temporal consistency of world models.

– **Maruyama Manipulability Index.** Smoothness alone is insufficient to guarantee manipulation success; robots must persistently avoid entering singular configurations. The manipulability index $w = \sqrt{\det(J(q)J^T(q))}$, based on the determinant of the Jacobian matrix $J(q)$, characterizes the volume of the end-effector's velocity ellipsoid. By inputting the video-generated 6D poses into an inverse kinematics (IK) solver, if the poses generated by the world model exceed the physical joint constraints of the robot, the manipulability index will sharply decline. Therefore, the integral of manipulability over the entire trajectory directly reflects the world model's implicit cognition of the robot's kinematic boundaries.

3.3 Four-Suite Benchmark Design

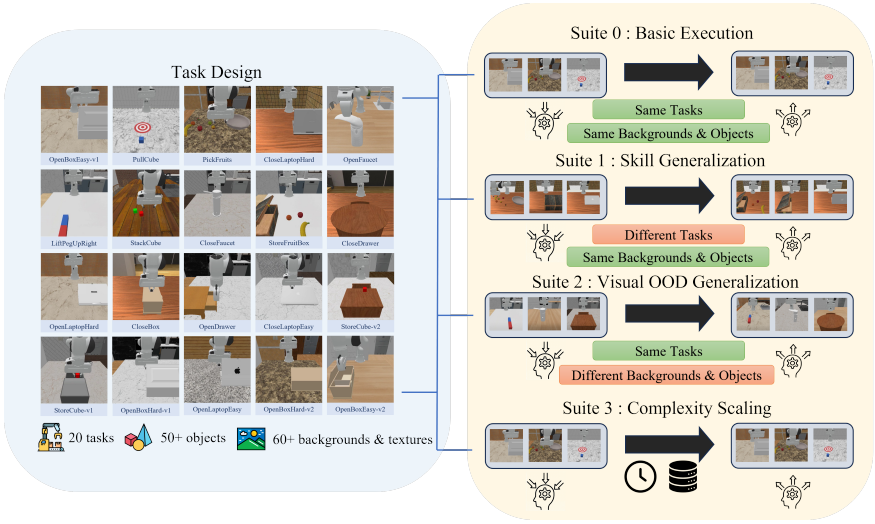


Fig. 3: Overview of the KineBench Task and Benchmark Design

To systematically probe the capability boundaries of EWMs, we design 20 diverse manipulation tasks in ManiSkill3, spanning a wide spectrum of physical complexities from primitive rigid-body interactions (e.g., pick-and-place) to long-horizon articulated object manipulation[6]. To isolate specific axes of generalization, these tasks are organized into four evaluation suites:

Suite 0: Basic Execution. The training and evaluation sets are strictly identically distributed (I.I.D.), encompassing all tasks, object instances, and backgrounds.

[6] Part of the assets used in our tasks are derived from the HumanoidGen project [17]. We thank the HumanoidGen team for their valuable contributions.

This suite establishes the fundamental kinematic execution baseline, quantifying a model's capacity to reliably reproduce demonstrated dynamics without domain shift.

Suite 1: Task Transfer. To test whether models digest physical causality rather than merely memorizing trajectory distributions, this suite evaluates zero-shot transfer to conceptually related but disjoint tasks. For instance, an EWM trained on `CloseBox`, `OpenBoxEasy-v1`, and `CloseLaptopEasy` is evaluated on unseen spatial recombinations like `OpenBoxEasy-v2` (novel box orientations) or harder variants like `CloseLaptopHard`. This suite probes the model's capacity for instruction understanding, spatial extrapolation and physical causality reasoning.

Suite 2: Visual OOD Generalization. This suite decouples visual representation robustness from physical reasoning. The training set is restricted to 50% of the randomized asset library (object instances and background textures), while evaluation is conducted entirely on the disjoint remaining 50%. This strictly out-of-distribution (OOD) visual setting tests whether kinematic generation catastrophically degrades under novel textures, geometries, and lighting conditions.

Suite 3: Complexity Scaling. Following the I.I.D. task distribution of Suite 0, we conduct controlled scaling experiments by systematically varying the training data volume and compute budget, with training duration used as a proxy for compute. This suite examines how intrinsic task complexity, such as degrees of freedom and contact dynamics, affects the performance gains obtained from increased data and compute, thereby revealing complexity-conditioned empirical scaling trends for EWMs.

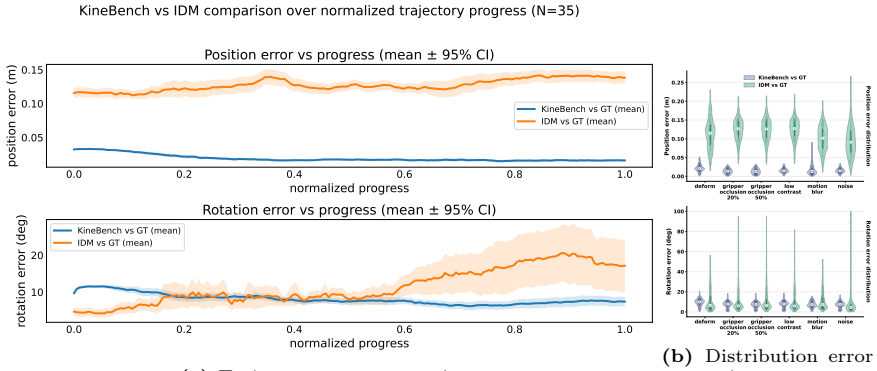
4 Experiments

4.1 Experimental Setup

Evaluation Environments and Task Taxonomy We systematically evaluate the models using the ManiSkill3 physics simulator. As detailed in Figure 3, we meticulously design 20 diverse manipulation tasks encompassing a wide spectrum of physical challenges: basic reaching motions, rigid body grasping (with and without obstacle interference), multi-axis articulated object manipulations (e.g., hinges and sliders oriented along the X, Y, and Z axes), and long-horizon compositional tasks. To rigorously assess generalization, each task incorporates highly randomized object instances and background textures, and is further stratified into different difficulty levels.

Hardware Constraints and Model Pool Constrained by a realistic computing budget of four NVIDIA A100 GPUs, we benchmark a representative suite of state-of-the-art video foundation models, including Wan 2.1 (1.3B), Wan 2.2 (5B), CogVideoX (2B), as well as LoRA-finetuned variants (CogVideoX 5B-LoRA and Wan 2.2 5B-LoRA).

Implementation Details of the KineBench Pipeline For explicit action extraction, we utilize a fine-tuned YOLOv11 for 2D mask extraction, a two-stage fine-tuned MoGeV2 which amplifies the loss on fine-grained geometric details on the second stage for high-fidelity metric depth estimation, and FoundationPose for zero-shot 6D pose tracking. This decoupled extraction-and-rollout pipeline ensures a highly efficient and stable closed-loop evaluation.



(a) Trajectory error comparison

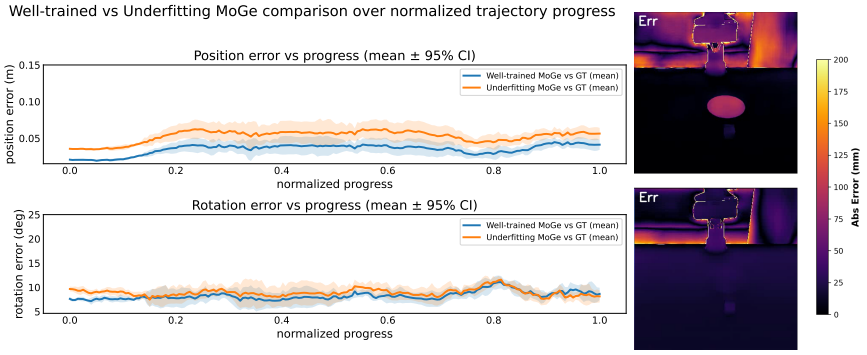
(b) Distribution error comparison

Fig. 4: Generalization and robustness comparison between KineBench and the IDM baseline. (a) Trajectory-wise translational and rotational errors on seven unseen test tasks. The IDM is trained on all trajectories from the ten training tasks in Suite 1, with 100 trajectories per task. For evaluation, we randomly sample five trajectories from each unseen task and compare the recovered end-effector poses with simulator ground-truth poses. **(b)** Distributions of pose-estimation errors under controlled visual perturbations, including gripper deformation, increasing levels of occlusion, reduced contrast, motion blur, and image noise. Lower values indicate more accurate pose recovery.

4.2 Generalization and Robustness Analysis of the KineBench Pipeline

A central design choice of KineBench is to replace implicit IDM-based action inference with an explicit and modular geometric grounding pipeline. Although the pipeline still contains learned perception components, including segmentation and depth estimation, its intermediate outputs can be independently inspected and ablated. Our goal is therefore not to treat the extracted poses as ground truth, but to examine whether explicit geometric grounding provides a more stable and diagnosable interface than an IDM when evaluating out-of-distribution generated videos. We conduct controlled analyses covering task-level generalization, robustness to visual degradations, and sensitivity to depth estimation.

Generalization and Robustness of Action Extraction We compare KineBench with Dino3DFlowIDM [10] on simulator-rendered trajectories from tasks excluded from IDM training. Both methods receive the same visual observations, and their recovered end-effector poses are evaluated against simulator ground truth. We further introduce controlled perturbations that approximate common generated-video artifacts, including gripper deformation, varying levels of occlusion, reduced contrast, motion blur, and image noise. As shown in Figures 4a and 4b, across all conditions, the KineBench pipeline achieves substantially lower position error than the IDM baseline. Although rotation mean error is slightly higher in some cases, its distribution is much more concentrated, with far fewer long-tail failures. These results support the reliability of our pipeline and indicate that it provides a more robust evaluation interface than IDM.



(a) Trajectory error comparison **(b)** Depth error heatmap

Fig. 5: Comparison of pose and depth estimation errors under different MoGe depth quality levels. (a) Pose errors between trajectories reconstructed from MoGe-estimated depths and the ground-truth poses. (b) Heatmap visualization of depth errors between MoGe-estimated depths and the corresponding ground-truth depths. Despite the difference in depth estimation quality, the resulting pose errors remain very similar, demonstrating the robustness of the proposed pipeline.

Ablation of Depth Estimation and Pose Tracking To isolate the effect of upstream depth estimation, we further compare three action-extraction settings: FoundationPose with simulator ground-truth depth, FoundationPose with MoGeV2-estimated depth, and Dino3DFlowIDM. Using ground-truth depth yields approximately centimeter-level translational errors, while replacing it with MoGeV2 increases the error to roughly 1.5–3 cm. Figure 5 further examines sensitivity to depth-estimation quality. In comparison, the IDM exhibits errors on the order of ten centimeters on the evaluated unseen trajectories. The remaining rotational error of the MoGeV2-based pipeline is approximately ten degrees, indicating that orientation estimation remains an important limitation for precision manipulation.

Table 1: Closed-loop success rates across evaluation suites. All values are reported as percentages. "Seen" and "Unseen" correspond to the training and validation visual assets in Suite 2, respectively. A dash indicates that the model configuration was not evaluated in the corresponding suite.

Suite	Main model comparison							Wan2.1-1.3B scaling configurations					
	Wan2.2 5B 7.5k	**CogVideoX 2B** 7.5k	**Wan2.1 1.3B** 7.5k	**Wan2.2 5B LoRA** 7.5k	**CogVideoX 5B LoRA** 7.5k	**Wan2.6** API	**Hailuo-V2** API	**1.5k** steps	**4.5k** steps	**15k** steps	**Scale 10** 7.5k	**Scale 25** 7.5k	**Scale 50** 7.5k
Suite 0	**56.32**	46.17	43.96	19.00	18.33	14.08	6.56	–	–	–	–	–	–
Suite 1	11.90	**57.78**	20.00	16.67	3.81	–	–	–	–	–	–	–	–
Suite 2 (Seen)	**58.00**	56.11	57.19	25.33	–	–	–	–	–	–	–	–	–
Suite 2 (Unseen)	**55.50**	51.67	52.83	26.33	–	–	–	–	–	–	–	–	–
Suite 3	–	–	43.96	–	–	–	–	44.83	47.88	**73.33**	53.67	52.00	51.17

This ablation suggests that most of the translational error introduced by the KineBench pipeline can be localized to the depth-estimation stage, while the CAD-constrained pose tracker remains comparatively stable. More importantly, the modular structure makes these error sources directly measurable, unlike an end-to-end IDM whose failures cannot be easily separated into perception and action-inference components.

Overall, the results indicate that KineBench provides a more robust and interpretable action-extraction interface under the evaluated distribution shifts and visual perturbations. However, its estimates remain dependent on segmentation, depth recovery, and sufficient visibility of the end-effector; accordingly, we view the pipeline as reducing IDM-specific attribution ambiguity rather than providing error-free pose recovery.

4.3 Main Results: Physical Execution and Generalization

We rigorously evaluate the models across three dimensions: base capacity (S0), cross-task transfer (S1), and visual OOD robustness (S2). Table 1 presents a performance comparison across evaluated models.

Base Physical Execution (Suite 0) . The results show that larger models excel on complex tasks demanding fine-grained manipulation or long-horizon sequencing (e.g., OpenBoxHard-v1, StoreCube-v2). However, severe performance drops on contact-rich tasks (e.g., StackCube, PickFruits) reveal that while video models capture kinematic priors, modeling complex friction and collision dynamics remains an open challenge. Under the matched Wan2.2-5B setting, full-parameter fine-tuning achieves higher performance than LoRA in S0, indicating that updating the full model may be beneficial for acquiring task-specific execution dynamics. However, this advantage does not hold consistently across the other evaluation suites.

Task and Visual Generalization (Suites 1 & 2) In Suite 2, models remain relatively robust on simple tasks such as CloseBox, but degrade substantially on

tasks requiring precise affordance localization under visual shifts. For example, Wan 2.2 drops from 60.0% to 30.0% on OpenBoxHard when evaluated on unseen assets. This suggests that current models may rely on appearance-specific features rather than learning transferable 3D geometric affordances.

Suite 1 further shows that cross-task transfer remains challenging. Notably, the LoRA-tuned model outperforms its fully fine-tuned counterpart, suggesting a trade-off between task-specific execution and transferability. One possible explanation is that LoRA preserves more pretrained visual and motion priors, whereas full-parameter fine-tuning specializes more strongly to the training tasks. We treat this explanation as a hypothesis rather than a causal conclusion.

To complement visual evaluation, we analyze the extracted 6D trajectories using SPARC and the Maruyama Manipulability Index. Their task- and model-dependent associations with execution success indicate that they provide complementary diagnostic signals for trajectory smoothness and kinematic feasibility.

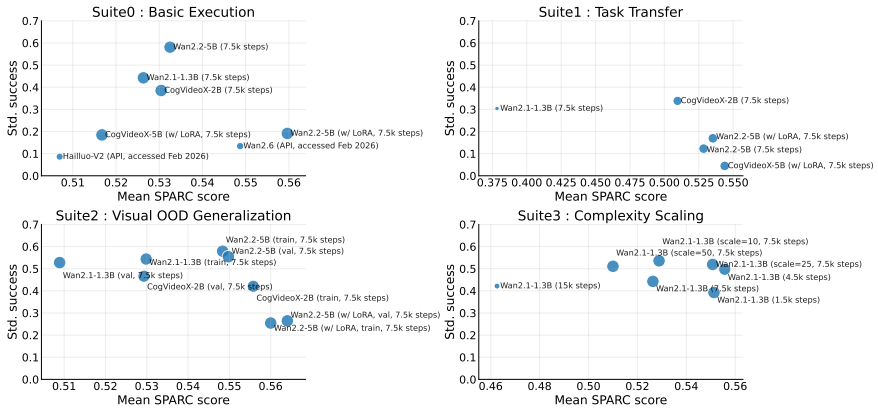


Fig. 6: Relationship between SPARC score and standardized success rate across models. Each point represents one model within a benchmark suite. The x-axis shows the task-balanced mean SPARC score, and the y-axis shows the standardized success rate.

SPARC: Evaluating Model Motion Fluency SPARC's predictive power is highly conditioned on a model's generation capabilities. Figure 6 shows that for zero-shot and under-optimized models, this metric strongly correlates with success rates, effectively filtering catastrophic physical hallucinations caused by poor motion fluency. Conversely, fully fine-tuned models generate kinematically smooth trajectories that saturate SPARC scores; their failures arise primarily from semantic misalignments rather than local jitter. Therefore, SPARC serves as a rigorous, specialized metric for evaluating fine-grained temporal fluency in physical actions.

Maruyama Manipulability: Evaluating Embodiment Awareness While SPARC captures temporal dynamics, the Maruyama Manipulability Index evaluates spatial constraints and embodiment awareness. Fully fine-tuned models effectively internalize the 7-DoF robot's structural boundaries, consistently maintaining low manipulability costs (Figure 7). Conversely, zero-shot closed-source and capacity-limited LoRA models lack this familiarity. Although generating visually plausible motions, their inherent stochasticity frequently pushes the robot toward singularities or kinematically unreachable points, leading to execution failures. Consequently, this index serves as a crucial metric for embodiment familiarity, exposing the risks of unconstrained generative randomness in world models.

Together, these two metrics establish a comprehensive 3D evaluation paradigm: SPARC scrutinizes the dynamic motion details, while Manipulability bounds the spatial positioning and embodiment constraints, offering profound insights into why generated videos succeed or fail in the physical world.

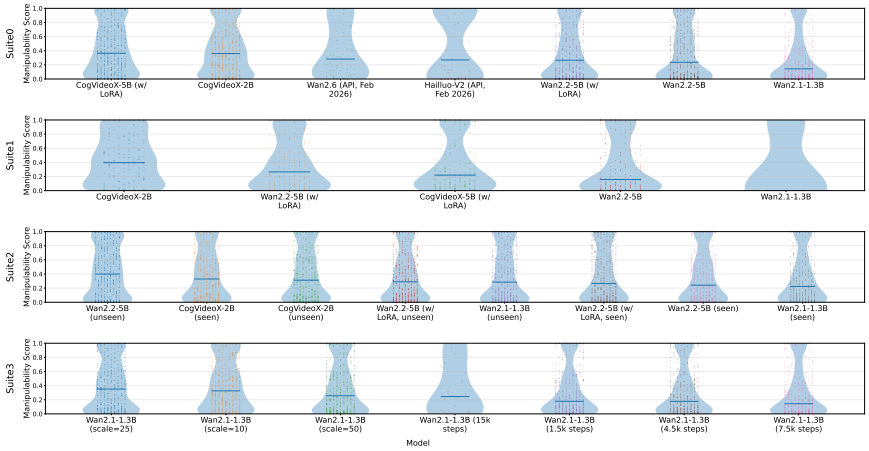


Fig. 7: Manipulation score across different suites. The figure presents violin plots of the Manip score across different suites, illustrating the distribution of manipulation performance across models. Lower values indicate better manipulation quality, corresponding to more accurate and stable execution of the generated trajectories. The Manip score is computed from trajectory costs obtained via PyRoki and normalized using a robust min–max scaling based on the 10th–90th percentiles.

4.4 Complexity-Conditioned Scaling Behaviors

In S3 of Table 1, we conduct a controlled scaling analysis using the Wan 2.1 1.3B architecture by varying the number of expert trajectories from 10 to 100 and the number of optimization steps from 1,500 to 15,000. Rather than exhibiting uniformly monotonic improvements, the results show task-dependent and nonlinear scaling trends.

For the relatively simple OpenBoxEasy task, increasing the number of homogeneous training trajectories from 10 to 100 reduces the success rate from 100% to 83.33%, while additional optimization steps also provide limited or negative gains. This suggests that increasing data and training duration does not necessarily improve performance when the task distribution is narrow and the model has already learned the dominant motion patterns. One possible explanation is that prolonged training on highly similar trajectories leads to overspecialization, although our experiments do not directly verify this mechanism.

In contrast, for the more complex StoreCube-v2 task, increasing the optimization budget improves the success rate from 13.34% to 69.52%, indicating that tasks with more diverse motion and interaction requirements may benefit from longer training. Overall, these results suggest that the marginal benefits of increasing data and compute depend on task complexity and trajectory diversity. Broader experiments across architectures, data regimes, and collection strategies are needed to determine how generally these trends hold.

5 Discussion and Limitations

While our benchmark evaluates several complementary aspects of embodied world models, the expert trajectories utilized in our benchmark are generated via motion planning. Future iterations will integrate human teleoperation data to enrich the action distribution. Although our 3D kinematic metrics effectively capture low-level physical coherence, future work will explore integrating them with existing 2D pixel-based evaluations, enabling a more holistic diagnosis of embodied world models.

6 Conclusion

In conclusion, we present KineBench, an IDM-free closed-loop benchmark for evaluating Embodied World Models through explicit 3D kinematic grounding. By incorporating classical robotic kinematic metrics, KineBench provides an executor-centric perspective on how selected properties of generated motions relate to downstream physical execution. Across four evaluation suites, our experiments reveal task-complexity-dependent nonlinear scaling trends among the evaluated video models. Overall, KineBench offers a transparent evaluation framework and empirical observations that may inform future research on physically grounded and actionable world models.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (Grant No.2024YFE0210900), the National Natural Science Foundation of China (Grant No.62306242), the Young Elite Scientists Sponsorship Program by CAST (Grant No. 2024QNRC001), and the Yangfan Project of the Shanghai (Grant No.23YF11462200).

References

1. Balasubramanian, S., Melendez-Calderon, A., Burdet, E.: A robust and sensitive metric for quantifying movement smoothness. IEEE Rransactions On Biomedical Engineering **59**(8), 2126–2136 (2011)
2. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. In: International Conference on Learning Representations (ICLR) (2025)
3. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), `https://openai.com/research/video-generation-models-as-world-simulators`
4. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024)
5. Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025)
6. Chen, B., Zhang, T., Geng, H., Song, K., Zhang, C., Li, P., Freeman, W.T., Malik, J., Abbeel, P., Tedrake, R., et al.: Large video planner enables generalizable robot control. arXiv preprint arXiv:2512.15840 (2025)
7. Chi, X., Jia, P., Fan, C.K., Ju, X., Mi, W., Zhang, K., Qin, Z., Tian, W., Ge, K., Li, H., et al.: Wow: Towards a world omniscient world model through embodied interaction. arXiv preprint arXiv:2509.22642 (2025)
8. Deng, Y., Pan, Z., Zhang, H., Li, X., Hu, R., Ding, Y., Zou, Y., Zeng, Y., Zhou, D.: Rethinking video generation model for the embodied world. arXiv preprint arXiv:2601.15282 (2026)
9. Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., Abbeel, P.: Learning universal policies via text-guided video generation. Advances in neural information processing systems **36**, 9156–9172 (2023)
10. Fan, C.K., Chi, X., Ju, X., et al.: Wow, wo, val! a comprehensive embodied world model evaluation turing test. arXiv preprint arXiv:2601.04137 (2026)
11. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: Advances in Neural Information Processing Systems 31, pp. 2451–2463. Curran Associates, Inc. (2018), `https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution`, `https://worldmodels.github.io`
12. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 (2019)
13. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering atari with discrete world models. arXiv preprint arXiv:2010.02193 (2020)
14. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse control tasks through world models. Nature pp. 1–7 (2025)
15. Hu, Y., Guo, Y., Wang, P., Chen, X., Wang, Y.J., Zhang, J., Sreenath, K., Lu, C., Chen, J.: Video prediction policy: A generalist robot policy with predictive visual representations. arXiv preprint arXiv:2412.14803 (2024)
16. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)

17. Jing, Z., Yang, S., Ao, J., Xiao, T., Jiang, Y.G., Bai, C.: Humanoidgen: Data generation for bimanual dexterous manipulation via llm reasoning. Advances in Neural Information Processing Systems **38**, 156210–156256 (2026)
18. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
19. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
20. Landi, C.T., Ferraguti, F., Sabattini, L., Secchi, C., Fantuzzi, C.: Admittance control parameter adaptation for physical human-robot interaction. In: 2017 IEEE international conference on robotics and automation (ICRA). pp. 2911–2916. IEEE (2017)
21. Le, M.Q., Zhu, Y., Kalogeiton, V., Samaras, D.: What about gravity in video generation? post-training newton's laws with verifiable rewards. arXiv preprint arXiv:2512.00425 (2025)
22. LeCun, Y., et al.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1), 1–62 (2022)
23. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)
24. Meng, F., et al.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In: International Conference on Machine Learning (ICML) (2025)
25. Qiao, Q., et al.: Vadb: A large-scale video aesthetic database with professional and multi-dimensional annotations. arXiv preprint arXiv:2510.25238 (2025)
26. Qin, I., et al.: Worldsimbench: Towards video generation models as world simulators. arXiv preprint arXiv:2410.18072 (2024)
27. Seedance, T., Chen, H., Chen, S., Chen, X., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Cheng, T., Cheng, X., et al.: Seedance 1.5 pro: A native audio-visual joint generation foundation model. arXiv preprint arXiv:2512.13507 (2025)
28. Team, U.: Ultralytics yolo evolution: An overview of yolo26, yolo11, yolov8, and yolov5. arXiv preprint arXiv:2510.09653 (2025)
29. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
30. Upadhyay, U., et al.: How close are world models to the physical world? arXiv preprint arXiv:2601.21282 (2026)
31. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
32. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details (2025), https://arxiv.org/abs/2507.02546
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
34. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 17868–17879 (2024)

35. Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., et al.: Hunyuanvideo 1.5 technical report. arXiv preprint arXiv:2511.18870 (2025)
36. Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., Abbeel, P.: Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114 (2023)
37. Ye, S., et al.: World action models are zero-shot policies. arXiv preprint arXiv:2602.15922 (2026)
38. Yoshikawa, T.: Manipulability of robotic mechanisms. The international journal of Robotics Research **4**(2), 3–9 (1985)
39. Yuan, J., et al.: Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. arXiv preprint arXiv:2510.11512 (2025)
40. Zhang, J., Jiang, M., Dai, N., Lu, T., Uzunoglu, A., Zhang, S., Wei, Y., Wang, J., Patel, V.M., Liang, P.P., Khashabi, D., Peng, C., Chellappa, R., Shu, T., Yuille, A., Du, Y., Chen, J.: Wow!: World models in a closed-loop world. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=yDmb7xAfeb
41. Zhang, Y., Bai, C., Zhao, B., Yan, J., Li, X., Li, X.: Decentralized transformers with centralized aggregation are sample-efficient multi-agent world models. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=xT8BEgXmVc
42. Zhang, Y., Li, X., Ye, J., Qiu, S., Qu, D., Li, X., Zhang, C., Bai, C.: Revisiting multi-agent world modeling from a diffusion-inspired perspective. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=rRxFIOoEeF
43. Zheng, Y., et al.: Videovla: Video generation as generalizable visual planning for robotic manipulation. arXiv preprint arXiv:2512.06963 (2025)
44. Zhu, Y., et al.: Ewmbench: A comprehensive evaluation benchmark for embodied world models. arXiv preprint arXiv:2506.18241 (2025)