AxE: Benchmarking Agent–Environment Generalization of Video World Models via Grounded Events

Haiyu Xu[1,2,*] Yiyang Chen[1,*] Jiayi Zhou[1,2] Jiaming Ji[1,2] Juntao Dai[1,2]
Yaodong Yang[1,†] Boyuan Chen[1,2,3,†]

[1]Peking University [2]Physis [3]Beijing Academy of Artificial Intelligence
[*] Equal contribution [†] Corresponding authors

Abstract

Edge cases are not the long tail; they are the norm. An embodied system deployed across changing environments, bodies, objects, and tasks will routinely encounter combinations that were absent from its training data. General-purpose physical intelligence therefore requires world models that reuse prior knowledge, predict consequences, and adapt with little or no additional training. Yet current evaluations of conditional video generation emphasize visual quality, prompt alignment, or isolated physical plausibility, leaving systematic generalization difficult to measure. We introduce AxE Bench, a benchmark that organizes condition changes along eight factors of the acting agent and of its environment, and pairs every original condition with controlled variants that change one side or both. WBench-plus applies the same structure to an existing multi-turn benchmark, pairing each WBench scenario with variants on the environment side, through its world setting or camera behavior, or on the agent side, through its interaction instructions. Additionally, to score generated futures, we develop an evaluation harness that combines vision-language reasoning with targeted computer-vision evidence, converting each video into temporally localized events and comparing generated events with reference events in both time and semantics. We summarize each model with the AxE Score, event Fidelity weighted by its Stability across original–variant pairs. Across twenty models, generated events match reference events far more often in timing than in content: no model exceeds a Fidelity of 13.01, the highest Stability is 0.55, and the highest AxE Score is 7.10. Together, these components turn world-model generalization from an informal aspiration into a falsifiable evaluation program for real and simulated interactions. Our project website can be found at: AxEBench.github.io

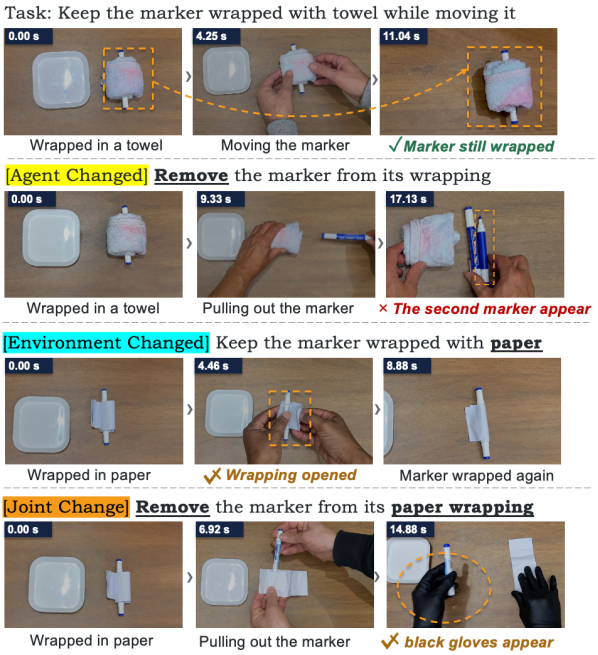


Figure 1. **Success in one condition does not guarantee generalization.** Each row shows three generated frames. The original task is to move a marker while keeping it wrapped in a towel (top). Variants change the action to removing the marker (**Agent Changed**), replace the towel with paper (**Environment Changed**), or combine both changes (**Joint Change**). Despite success in the original condition, the variants exhibit object duplication, unintended unwrapping, and inconsistent agent appearance, respectively.

1. Introduction

Anticipating the consequences of actions is a central goal on the path toward artificial general intelligence, and video generation offers a concrete way to pursue it: given an observation and a description of an action, a model generates a possible continuation of the scene. Since video gener-

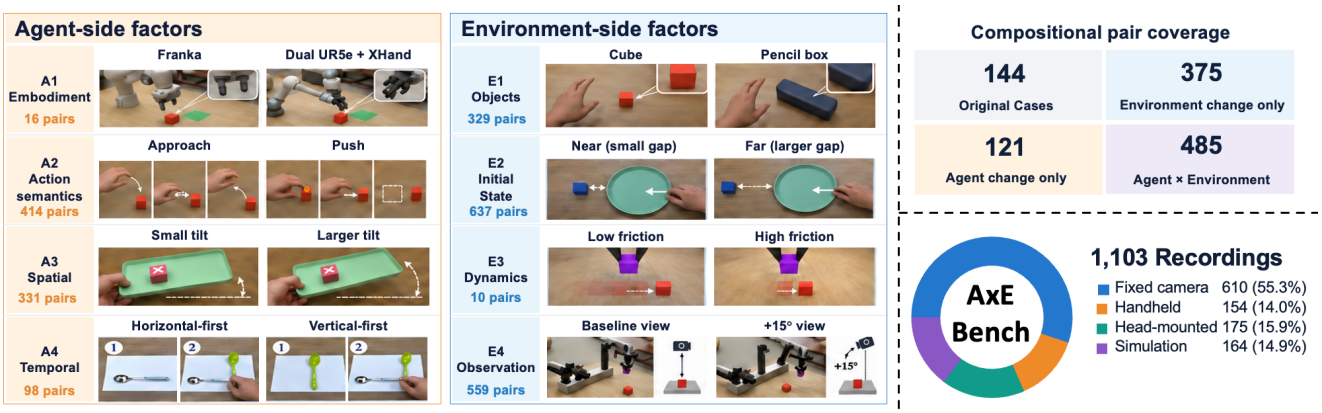


Figure 2. **Data taxonomy and coverage.** We organize changes relative to an original condition into eight agent-side and environment-side factor families; each is shown with an original–variant pair and the number of pairs that change it. The presence of changes on either side defines four regimes: original, agent generalization, environment generalization, and agent–environment composition. The ring breaks the recordings down by capture setup.

ation models were first explored as world simulators [8], later systems have advanced physical fidelity [32], action-conditioned robot simulation [13], and interactive generation [18, 33]. These advances bring video generation closer to embodied intelligence and open-world interaction, where predictions must hold across diverse operating conditions.

Such applications make generalization a central requirement. As illustrated in Fig. 1, success in an original condition can coexist with failures in object consistency, action execution, and agent appearance when the action or environment change. The same interaction may involve a different embodiment, an unfamiliar object, a changed physical setting, or a new viewpoint. This raises a central question for video-based world simulation: how reliably do models generate the appropriate consequences of an interaction when its conditions change, individually or in combination?

Answering this question requires evaluating performance across explicitly related conditions. Existing benchmarks provide important foundations: VBench and EvalCrafter assess video quality and prompt alignment [21, 26], while VideoPhy, Physics-IQ, and WorldModelBench examine physical plausibility and world-modeling capabilities [6, 24, 29]. Our focus is on how these capabilities generalize under controlled changes. Evaluating diverse scenarios alone does not isolate which changes affect performance or how multiple changes interact. We therefore organize evaluation around explicit variations of reference conditions, enabling comparisons of baseline competence, sensitivity to individual factors, and robustness to their composition.

We introduce AxE Bench (Agent × Environment), a benchmark for evaluating the generalization of world-simulation capabilities in conditional video generation. Each condition is described by eight factors, four of the acting agent and four of the environment and how it is perceived, and each original condition is paired with variants that change the agent side, the environment side, or both.

We further apply this framework to extend an existing benchmark to generalization evaluation. Specifically, we construct WBench-plus by augmenting each multi-turn scenario in WBench [41] with variations in world settings and camera behavior on the environment side, and interaction instructions on the agent side.

Evaluating these changes also requires evidence beyond overall visual impressions. A decisive error may occur during a brief contact, emerge after an object becomes occluded, or accumulate across a sequence of events. We therefore develop an evaluation harness that combines computer-vision evidence with vision-language reasoning to assess generated entities, interactions, and event progression. Our contributions are threefold:

- **A benchmark for generalization in video-based world simulation.** We introduce AxE Bench, which structures evaluation conditions into eight atomic factors spanning the agent and environment. As part of the benchmark, we construct WBench-plus (Sec. 3.2), an extension of WBench with paired original and modified multi-turn tasks.
- **An evidence-based evaluation harness.** We introduce WM-PerceptHarness, which combines vision-language reasoning with targeted visual evidence gathering to assess the temporal and semantic fidelity of generated events, with traceable evidence for diagnosis. On top of it, the AxE Score combines event Fidelity with its Stability across original–variant pairs into a single measure of generalization.
- **A systematic analysis of model generalization.** Across twenty models, generated events often fall at the right time but rarely show the right content: Timing is 38.8

to 52.9 for the text-and-image-to-video models, while no model exceeds a Fidelity of 13.01. The scores also do not hold under a change: the highest Stability is 0.55, and the highest AxE Score is 7.10.

2. Related Work

Video Generation Models and World Models A world model is a learned predictive model of an environment that captures its dynamics and predicts future observations or latent states [19]. Through multi-step prediction and imagined rollouts, world models support control and policy learning [19], policy evaluation [25], and environment simulation and interactive generation [9]. Video generation models have become a common way to build such models: they predict future frames from an image and a text prompt [17, 37, 39] and have been proposed as general-purpose simulators of the physical world [8, 30], and additional conditioning on actions or camera motion makes the generated world interactive [20, 33, 35]. These applications require reliable generalization across world settings, interaction demands, and observation conditions, motivating our systematic evaluation under controlled shifts.

World-Model Evaluation *Evaluation of model capabilities:* WorldModelBench [24] evaluates instruction following, physical adherence, and common sense. WorldScore [12] assesses controllability, visual quality, and motion dynamics across video and 3D/4D world generation. WBench [41] evaluates interactive video world models on multi-turn tasks spanning navigation, subject actions, event edits, and perspective switches. *Evaluation of downstream applications:* WorldSimBench [34] combines perceptual evaluation with closed-loop embodied task evaluation. WorldArena [36] jointly evaluates video quality and functional utility in robotic manipulation. *Generalization and model generality:* Kang et al. [22] investigate physical prediction in video diffusion models under in-distribution, out-of-distribution, and combinatorial generalization settings. WorldTest [38] proposes a protocol for evaluating whether agents can use environmental knowledge acquired through reward-free interaction to solve diverse environment-level queries.

3. The AxE Bench

3.1. Agent–Environment Generalization Space

Interaction and model. A condition is an interaction between an agent and its environment, seen through a camera: at step t the camera gives an observation o_t of the environment state, and the agent issues an action a_t that changes it. A video world model G_θ receives the first observation and the action sequence and generates a video:

$$\hat{V} = G_\theta\big(o_1, \mathbf{a}_{1:T}\big) = (o_1, \hat{o}_2, \ldots, \hat{o}_T) \tag{1}$$

Factors. What happens in a condition is determined by two things: the actions that the agent issues, and the environment in which they take effect, that is, the initial state, how the state responds to an action, and how it is observed. A variant can therefore differ from its original condition on the agent side, in the action sequence, or on the environment side, in everything else. We divide each side into four factors (Fig. 2). The four agent-side factors change the action sequence: who issues the actions (**Embodiment**), what they are and what goal they serve (**Action Semantics**), how an action with the same goal is executed, such as its amplitude or direction (**Spatial**), and how the steps are ordered or repeated (**Temporal**). The four environment-side factors leave the actions as they are and change the rest: which task-relevant objects the first state contains (**Objects**), how they are arranged (**Initial State**), the physical conditions under which the state responds to an action, such as friction or mass (**Dynamics**), and the viewpoint and motion of the camera through which the state is observed (**Observation**).

3.2. Data Collection and Pairing

Recordings. The benchmark contains 1,103 recordings. Real recordings show people manipulating everyday objects and are captured with a fixed camera (610), a handheld moving camera (154), or a head-mounted camera (175). The remaining 164 are simulated robot manipulation, which lets us vary what real capture cannot control cleanly: the robot arm, the friction, and the camera viewpoint, which is rotated by a known angle.

Pairs and evaluation settings. Recordings are organized in task groups, in which one recording is the original condition and the others are variants that keep the task and change designated factors. This yields 981 original–variant pairs over 144 original conditions. A pair that changes several factors is counted under each of them (Fig. 2): initial state (637 pairs), observation (559), and action semantics (414) are the most frequent, while embodiment (16) and dynamics (10) are the rarest. We evaluate two input settings on the same recordings and pairs. In the text-and-image-to-video setting, a model receives the first frame and the text prompt; in the camera-conditioned setting, it additionally receives a camera trajectory, which is the recorded trajectory for the 329 moving-camera recordings and a static one for the 774 fixed-camera recordings, real and simulated.

Extension to multi-turn interaction. To verify that the original–variant design is not tied to our own recordings, we apply it to WBench [41], a multi-turn benchmark that was



Figure 3. **Construction of WBench-plus.** We show representative original–variant pairs selected from the newly constructed WBench-plus cases. The agent-side examples (left) illustrate Change subject appearance (A1), Quantity and Intensity (B1), Add, Remove, and Replace content (B2), and Reorder instructions (B3). The environment-side examples (right) illustrate Change setting (A2), Change rendering (A3), Change camera angle (C1), Change camera path (C2a), and Change view mode (C2b). Dashed outlines mark the original WBench [41] cases or instruction sequences; solid outlines mark their WBench-plus variants.

not built for generalization. Each of its 289 cases becomes an original condition and is paired with variants that apply one designated edit and keep the number of turns (Fig. 3): world setting and view and control change the environment side, and interaction instructions change the agent side. The resulting WBench-plus contains 884 variants (283 world-setting, 283 view-and-control, and 318 instruction variants) that cover all 289 original cases and comprise 3,273 turns. Together with the 1,058 turns in the original cases, the full suite contains 4,331 turns.

4. Grounded Event Evaluation

A generated video can look plausible and still get the interaction wrong, and a single judgment of the whole clip cannot say where. Our harness, WM-PerceptHarness, therefore converts every video into a timeline of events supported by visual evidence and compares the events of a generated video with those of its reference (Fig. 4).

4.1. Evidence-Grounded Event Annotation

An event is a triple of start time, end time, and a text description of what is observed in that interval, with times on the clock of the original video. Annotation proceeds in three steps. First, a vision-language model (VLM) watches the video, drafts a timeline, and lists the claims in its draft that the frames it has seen do not settle, such as whether a hand has fully left the frame or where an event ends; each becomes a review question with a time window. Second, for each question the harness gathers evidence with a segmentation model (SAM 3.1 [27]): it selects the target named in the question at a reference frame, checks the mask, and tracks the target over the requested window, which yields timestamped frames with mask overlays. The segmentation model does not decide what an action means or whether it succeeded. Third, the VLM receives the video, its draft, the questions, and the accepted evidence, and revises event descriptions and boundaries, citing the times that support each change. The same procedure annotates reference recordings and generated videos.

4.2. Event Matching and Scores

Let a reference timeline have N_r events and a generated timeline N_g events. For every pair of a reference event and a generated event we compute the temporal intersection over union (tIoU) of their intervals; pairs with $\mathrm{tIoU} \geq 0.5$ are temporally eligible. A semantic judge then reads the two descriptions of each eligible pair, and decides whether they report the same object, action, direction, and outcome.

We compute two assignments. The *temporal* assignment uses all eligible pairs and matches M_t of them, of which S_t also pass the judge. The *joint* assignment is computed separately, using only pairs that are eligible and pass the semantic judge, and matches M_j pairs. The scores are

$$\begin{aligned} \mathrm{Timing} &= \frac{2M_t}{N_r + N_g}, \qquad \mathrm{Content} = \frac{S_t}{M_t}, \\ \mathrm{Fidelity} &= \frac{2M_j}{N_r + N_g}. \end{aligned} \tag{2}$$

Timing is an F1 score over time-matched events: it measures whether the generated video has events where the reference has them, regardless of what they show. Content is the share of time-matched events whose descriptions also agree. Fidelity is an F1 score over events matched in both time and content. All three are reported $\times 100$.

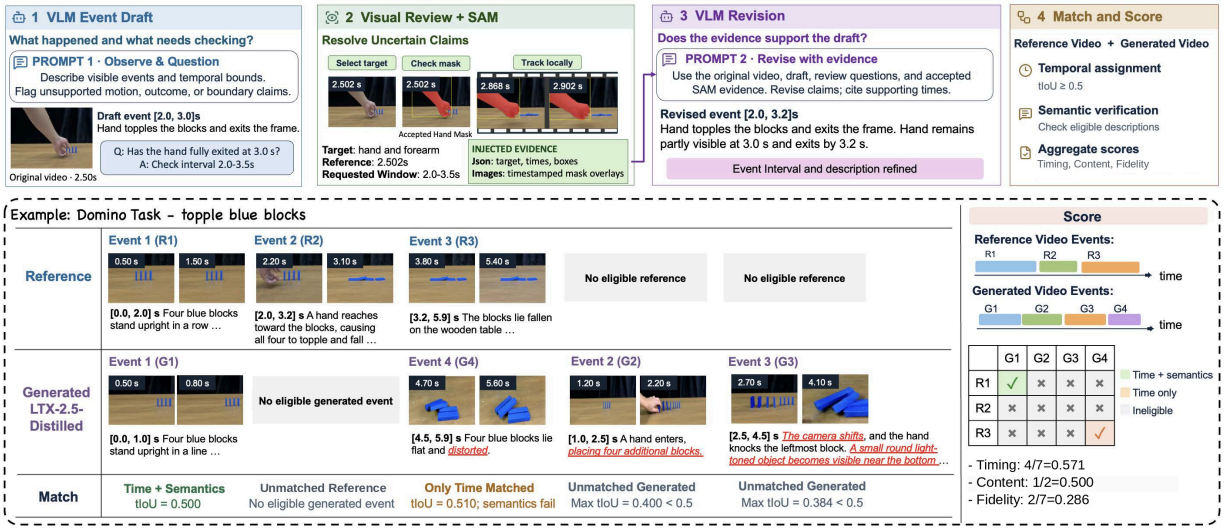


Figure 4. **Grounded event evaluation.** A vision-language model converts each video into timestamped event descriptions in three steps: it drafts the events and flags claims that need checking, gathers question-driven SAM evidence (target selection, mask check, local tracking) for those claims, and revises the events with the accepted evidence. Reference and generated events are then compared. Bottom: a domino task with three reference and four generated events, its match matrix, and the resulting scores.

Paired comparison. To measure the effect of a change, each variant is compared only with its own original condition. Let $f(x)$ be the Fidelity of a single video x, and let (v_i, o_i), $i = 1, \ldots, n$, be the pairs of a variant v_i and its original condition o_i. The *Stability* of a model is

$$\text{Stability} = 1 - \frac{\sum_i \left| f(v_i) - f(o_i) \right|}{\sum_i \left(f(v_i) + f(o_i) \right)}, \tag{3}$$

the share of the paired score that is reached on both sides. It is one when no score changes within a pair and zero when every pair scores on one side only, whichever side that is. The *AxE Score* is the product of the two,

$$\text{AxE Score} = \text{Fidelity} \times \text{Stability}, \tag{4}$$

the score of a model weighted by how well it holds under a change.

Complementary metrics. Event scores do not measure visual quality, so we also report PSNR against the reference recording, Dynamic Degree and Motion Smoothness [21], Semantic Adherence and Physical Commonsense on a 1–5 scale [7], the camera motion error CamMC [43] for camera-conditioned models, and the GeCo structure error [15] relative to the real recording of the same task.

4.3. Agreement with Human Judgments

We validate the semantic judge and Fidelity against blinded human annotations (1,360 annotations across 1,060 items).

Semantic judge. On 578 event pairs, the judge achieves 87% agreement with human judgments (Cohen's $\kappa = 0.72$).
Fidelity. Human annotators compare two generated videos against a reference, selecting the closer one or a tie. Across nine models, rankings by Fidelity strongly correlate with human win-rate rankings (Spearman $\rho = 0.90$).

5. Experiments

5.1. Models and Evaluation Setup

We evaluate twenty video generation models on our recorded benchmark. Thirteen take a first frame and a text prompt: nine open models, MiniMax H3 [28], LTX-2.5 (distilled) [17], VPP2 [16], Cosmos 3 Super and Nano [31], Wan2.2 TI2V-5B [37], HunyuanVideo 1.5 [39], Cosmos-Predict2.5-2B [30], and GigaWorld-0 [14], and four commercial models, Wan 3.0 [4], Kling 4.0 [23], Seedance 2.5 [11], and HappyHorse 1.1 [3]. Seven further models additionally take a camera trajectory: Alaya-World [2], Matrix-Game-2.0 [20], Wan2.2 Fun Camera [5], SANA-WM [44], LingBot-World [35], Evoke [40], and Evoke Turbo [1].

The vision-language model that drafts and revises the events and the semantic judge are both Doubao-Seed-2.1-Pro [10]; visual evidence comes from SAM 3.1 [27]. The judge runs at temperature zero.

5.2. Main Results

Reproducing the reference events is hard for every model. The four commercial models score highest, with a

Table 1. Main results on the 1,103 recordings. Event scores (Sec. 4.2, $\times 100$) are pooled over all recordings; Stability (Eq. (3)) is computed on the 981 pairs. The complementary metrics are averaged over the same recordings. MS: Motion Smoothness; SA: Semantic Adherence; PC: Physical Commonsense; DD: Dynamic Degree. CamMC is computed on the 329 moving-camera recordings only. Rows are sorted by Fidelity; the first four text-and-image-to-video models are commercial. The best value of each column within each of the two groups is in bold.

	Event scores			Generalization		Complementary metrics						
Model	Fidelity ↑	Timing ↑	Content ↑	Stability ↑	AxE Score ↑	PSNR ↑	MS ↑	SA ↑	PC ↑	CamMC ↓	DD	GeCo
Text-and-image-to-video												
Wan 3.0	**13.01**	49.71	**26.16**	**0.55**	**7.10**	17.17	0.995	2.94	**4.04**	–	0.637	1.043
HappyHorse 1.1	12.55	**52.92**	23.72	0.54	6.79	15.44	0.990	2.94	3.97	–	0.792	1.411
Kling 4.0	12.26	49.92	24.56	0.49	5.98	**18.08**	**0.996**	2.93	3.99	–	0.545	0.790
Seedance 2.5	12.22	52.26	23.38	0.42	5.12	17.06	**0.996**	2.93	4.01	–	0.579	0.994
MiniMax H3	10.06	52.04	19.30	0.39	3.97	15.86	0.994	2.94	3.97	–	0.652	1.069
LTX-2.5-Distilled	5.59	41.06	13.61	0.35	1.93	15.67	0.992	2.96	3.86	–	0.762	1.411
VPP2	2.98	45.02	6.57	0.18	0.53	14.84	0.978	2.99	3.50	–	0.931	1.179
Cosmos3-Super	2.82	43.31	6.52	0.25	0.72	15.41	0.992	2.99	3.66	–	0.896	1.748
Wan2.2-TI2V-5B	2.78	38.75	7.18	0.30	0.84	15.87	0.995	2.96	3.85	–	0.581	1.280
HunyuanVideo-1.5	1.73	42.39	4.09	0.18	0.31	14.91	0.995	2.97	3.82	–	0.821	0.756
Cosmos-Predict2.5-2B	1.57	44.75	3.52	0.37	0.59	13.21	0.993	3.00	3.82	–	0.844	1.691
GigaWorld-0	1.49	40.95	3.61	0.19	0.28	16.01	0.991	**3.01**	3.46	–	0.855	1.013
Cosmos3-Nano	1.48	42.19	3.52	0.03	0.04	14.30	0.990	3.00	3.61	–	0.937	1.885
Camera-conditioned image-to-video												
AlayaWorld	**4.45**	32.32	**13.78**	0.24	**1.09**	**17.64**	**0.996**	2.62	4.26	1.116	0.240	0.380
Wan2.2-Fun-Camera	2.69	**46.05**	5.81	0.23	0.61	13.18	0.992	2.97	3.86	1.080	0.830	1.437
Matrix-Game-2.0	2.66	31.03	8.58	0.25	0.67	15.33	0.993	2.44	**4.45**	1.151	0.290	1.401
SANA-WM	2.65	45.88	5.73	0.15	0.39	14.36	0.985	2.94	3.46	0.966	0.802	1.556
LingBot-World	0.99	41.24	2.41	0.02	0.02	13.24	0.982	2.92	3.88	1.212	0.762	2.689
Evoke	0.84	44.57	1.89	0.42	0.35	14.71	0.993	**3.01**	3.80	0.898	0.866	0.960
Evoke-Turbo	0.68	42.98	1.54	0.00	0.00	14.73	0.993	3.00	3.76	**0.850**	0.869	1.100

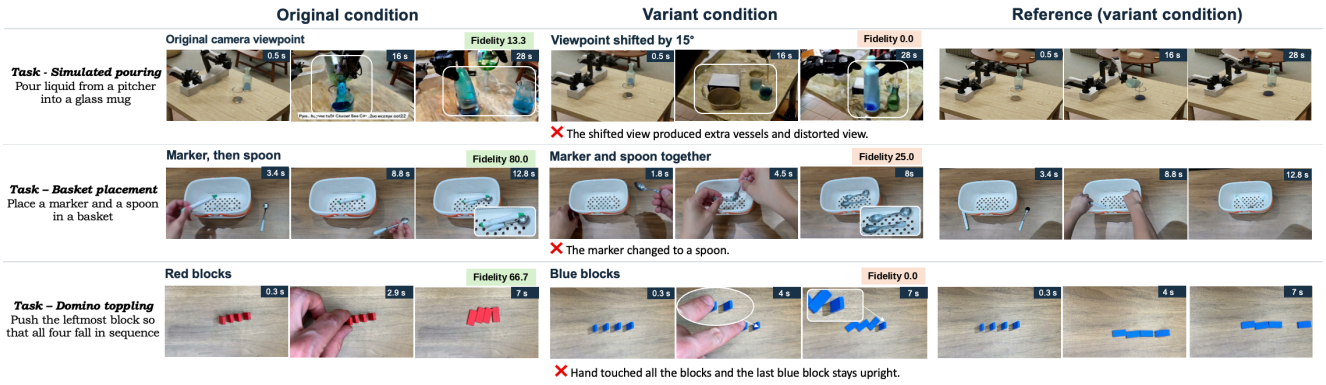


Figure 5. **Typical examples.** Each row shows key frames of the generated video for the original condition (left) and the variant condition (middle), with their Fidelity, and the reference recording of the variant condition (right). Changed factor by row: observation (simulated camera rotated by 15°; LTX-2.5), action semantics (marker and spoon placed together instead of in sequence; LTX-2.5), initial state (block color; MiniMax H3).

Fidelity of 12.22 to 13.01 (Table 1). Among the open models, the best is 10.06 (MiniMax H3), the second 5.59 (LTX-2.5), and the other seven are below 3.0. Models that additionally take a camera trajectory do not score higher: the best reaches 4.45 (AlayaWorld), below the two best open models, and Evoke Turbo has both the lowest camera motion error (0.850) and the lowest Fidelity (0.68). Timing lies between 38.8 and 52.9 for the thirteen text-and-image-to-video models, but Content is at most 26.2: generated events fall at roughly the times of the reference events, and their content usually differs. Once Stability is taken into account, the gap widens further: the highest AxE Score is 7.10 (Wan

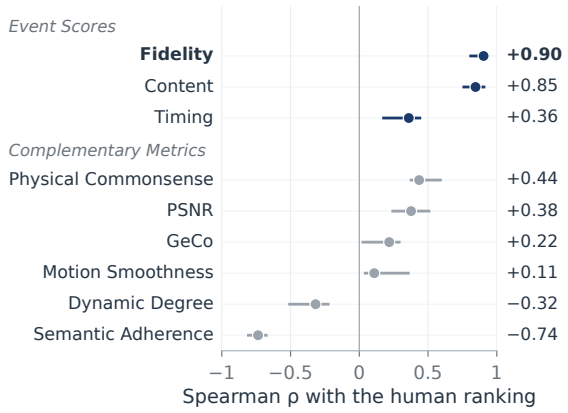


Figure 6. **Agreement of each metric with human preference.** Spearman ρ between the ranking of the nine open text-and-image-to-video models by each metric and their ranking by human win rate (Sec. 4.3). Bars are 95% intervals from resampling the human comparisons. GeCo is counted as lower-is-better, all other metrics as higher-is-better.

3.0), and Sec. 5.3 examines why. In Fig. 5, a marker turns into a second spoon, a rotated view gains vessels that were never in the scene, and a hand touches every block instead of pushing the first one.

Complementary metrics separate the models less and do not follow human preference. These metrics describe how a video looks and moves, not whether it shows the instructed interaction. Some barely separate the models: across the thirteen text-and-image-to-video models, Motion Smoothness lies between 0.978 and 0.996 and Semantic Adherence between 2.93 and 3.01, while Fidelity ranges from 1.48 to 13.01. Where they do rank the models, the ranking does not match human preference (Fig. 6): Fidelity follows the human ranking ($\rho = 0.90$), but no complementary metric exceeds 0.44, and Semantic Adherence runs against it. PSNR and GeCo favor videos that move little: among the text-and-image-to-video models, the three with the highest PSNR are among the four with the lowest Dynamic Degree.

5.3. Generalization under Paired Changes

We compare each variant with its own original condition as in Sec. 4.2, on the same 981 pairs for all models.

Pooled scores separate three kinds of behavior. Figure 7 gives, for each model, the Fidelity over all original videos of the 981 pairs and over all their variant videos. We read the two scores together. A model with high Fidelity on both sides and a small gap between them generates the instructed events and keeps doing so when the condition changes. A model with high Fidelity but a large gap is capable but not robust: Seedance 2.5 has the largest gap of all twenty models (3.2 points), although its Fidelity is within one point of the other commercial models, and it also has the lowest stability of the four (0.42, Table 1). A model with low Fidelity on both sides lacks the ability in the first place, and its small gap is not evidence of robustness: for the models at the bottom of Fig. 7, both scores are close to zero. The gap is therefore informative only together with the level, which is what the AxE Score combines.

Table 2. **Results on the multi-turn WBench-plus cases.** Rows are sorted by AxE Score within each of the three groups; the first four text-and-image-to-video models are commercial. The best value of each column within each of the three groups is in bold.

Model	Fidelity ↑	Timing ↑	Content ↑	Stability ↑	AxE Score ↑
Text-and-image-to-video					
Kling 4.0	**27.80**	41.81	**65.89**	**0.62**	**17.18**
Seedance 2.5	27.02	**44.02**	60.91	**0.62**	16.64
Wan 3.0	24.70	42.90	57.07	0.60	14.70
HappyHorse 1.1	21.68	38.21	56.59	0.55	11.96
MiniMax H3	**17.55**	43.22	**40.32**	**0.53**	**9.39**
LTX-2.5-Distilled	17.14	43.73	38.28	0.52	8.84
Cosmos3-Super	15.63	40.82	38.16	**0.53**	8.34
GigaWorld-0	15.15	**47.79**	31.57	0.47	7.18
HunyuanVideo-1.5	14.15	38.20	36.95	0.46	6.44
VPP2	13.30	39.36	33.65	0.43	5.67
Wan2.2-TI2V-5B	7.38	42.92	16.37	0.46	3.43
Cosmos3-Nano	8.17	34.54	23.37	0.37	3.03
Cosmos-Predict2.5-2B	7.08	44.46	15.30	0.37	2.63
EchoWM	5.51	26.91	20.48	0.31	1.70
Camera-conditioned image-to-video					
Wan2.2-Fun-Camera	**34.35**	59.41	57.57	0.59	**20.14**
SANA-WM	28.09	**64.47**	43.10	**0.67**	18.94
EchoWM Flash	26.77	42.57	**62.44**	0.60	15.93
Matrix-Game-2.0	24.52	59.35	41.32	0.64	15.75
LingBot-World	21.44	43.25	49.59	0.61	13.17
Evoke	12.82	26.89	45.49	0.46	5.91
Evoke-Turbo	11.91	26.37	44.44	0.45	5.41
AlayaWorld	9.13	35.65	25.62	0.37	3.38

Within pairs, even the strongest models hold about half of their score. Stability (Eq. (3)) measures whether a model reproduces both conditions of the same pair, and Fig. 8 multiplies it onto Fidelity. It ranges from 0.39 (MiniMax H3) to 0.55 (Wan 3.0) for the five highest models and is 0.03 for Cosmos 3 Nano. The AxE Score of Wan 3.0 is 7.10 against a Fidelity of 13.01. Among the text-and-image-to-video models, the six highest models keep their order; below them, VPP2 falls from seventh to tenth and Wan2.2 TI2V-5B rises from ninth to seventh. Among the camera-conditioned models, the highest AxE Score is 1.09 (AlayaWorld). For models that score on few pairs, Stability is unreliable (0.42 for Evoke), and we do not interpret it.

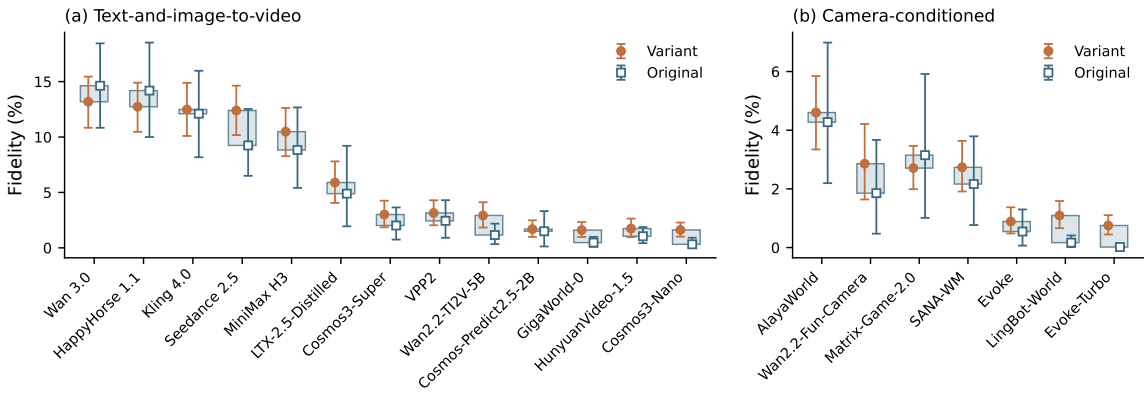


Figure 7. **Fidelity on the original and on the variant conditions.** For each model, the two markers give the Fidelity over all original videos of the 981 pairs and over all their variant videos, each computed from pooled event counts; bars are 95% intervals from resampling task families. Models are ordered by their mean Fidelity over the pairs. A model that generalizes scores high on both sides with a small difference between them; a small difference at a low score does not indicate this. (a) Text-and-image-to-video; (b) camera-conditioned, on a different axis range.

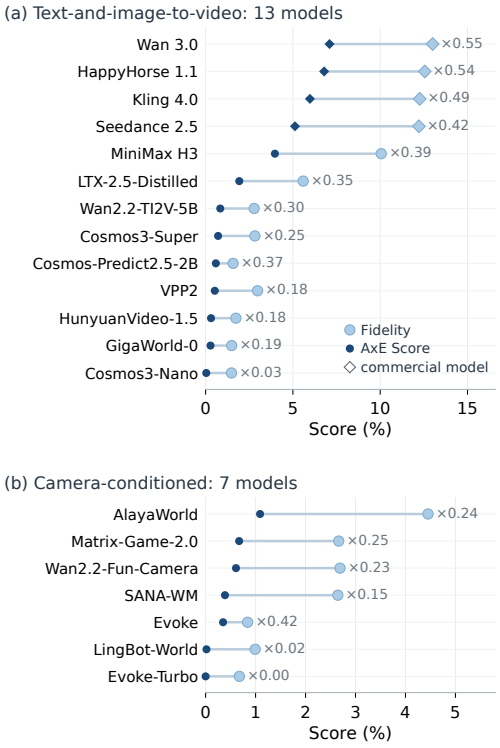


Figure 8. **Fidelity and AxE Score.** Light markers: Fidelity (Table 1). Dark markers: the AxE Score, Fidelity times Stability (Eq. (3)); the Stability value is printed beside each row. Models are ordered by AxE Score; diamonds mark the commercial models. For scores close to zero, Stability rests on few pairs. The axis runs from 0 to 15 in (a) and from 0 to 5 in (b).

5.4. Multi-Turn Generalization on WBench-plus

For the multi-turn evaluation, Table 2 reports 20 models from the main benchmark together with two additional models, EchoWM and its four-step autoregressive variant, EchoWM Flash [42]. In total, we report 22 models on WBench-plus, expanding our generalization design and event evaluation.

Event evaluation extends to tasks without ground-truth videos. WBench-plus also demonstrates how the event scores can be applied when ground-truth videos are unavailable. Each condition's turn instructions specify the required behavior, and its turn windows provide the reference intervals. Generated videos are converted into event timelines and evaluated against these instruction-derived references using the temporal and semantic matching of Sec. 4.2. Timing measures alignment with the reference intervals, Content assesses whether time-matched events fulfill the instructions, and Fidelity requires both. This provides a common evaluation structure for text-and-image-to-video and camera-conditioned models, while allowing each modified task to be scored against its own requirements.

The generalization design reveals sensitivity beyond our own recordings. Applying the agent–environment variation design to WBench reveals substantial sensitivity in an existing multi-turn benchmark. Across changes to world settings, camera behavior, and interaction instructions, the highest Stability is only 0.62 among the commercial models, 0.53 among the open text-and-image-to-video models, and 0.67 among the camera-conditioned models. Even the most stable models therefore show substantial differences

between paired conditions. These results show that the original–variant design exposes limitations that evaluation on individual conditions alone would leave unmeasured, extending the analysis of generalization from our recorded interactions to WBench's multi-turn tasks.

6. Conclusion

We introduced AxE Bench, which pairs each original condition with variants that change the acting agent, the environment, or both along eight factors, and WM-PerceptHarness, which scores a generated video by the events it contains. On 1,103 recordings and 981 pairs, twenty models place events at roughly the right times but rarely with the right content: no model exceeds a Fidelity of 13.01. The scores also do not hold under a change. Stability is at most 0.55, so even the strongest model reaches only about half of its paired score on both conditions of a pair, and the highest AxE Score is 7.10. Applied to WBench, the same design gives the same picture for multi-turn tasks, with a Stability of at most 0.67. Fidelity follows human preference across models ($\rho = 0.90$), while no complementary metric exceeds 0.44. A deployed world model meets changed conditions routinely; current models do not yet carry what they get right in one condition over to a related one. These findings motivate evaluating event fidelity and robustness jointly when assessing progress in video world models.

References

[1] Alaya Lab. Evoke-Turbo. Hugging Face model card, https://huggingface.co/AlayaLab/Evoke-Turbo, 2026.

[2] AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. AlayaWorld: Long-horizon and playable video world generation. *arXiv preprint arXiv:2607.06291*, 2026.

[3] Alibaba Cloud. HappyHorse gets stronger motion expressiveness, higher generation consistency, and enhanced visual quality. Alibaba Cloud Blog, https://www.alibabacloud.com/blog/happyhorse-gets-stronger-motion-expressiveness-higher-generation-consistency-and-enhanced-visual-quality_603293, 2026.

[4] Alibaba Cloud. Alibaba unveils Wan3.0 with twice as long video outputs from a richer variety of inputs. Alibaba Cloud Blog, https://www.alibabacloud.com/blog/alibaba-unveils-wan3-0-with-twice-as-long-video-outputs-from-a-richer-variety-of-inputs_603439, 2026.

[5] Alibaba PAI. Wan2.2-Fun-A14B-Control-Camera. Hugging Face model card, https://huggingface.co/alibaba-pai/Wan2.2-Fun-A14B-Control-Camera, 2025.

[6] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, 2025.

[7] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, 2026.

[8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI, https://openai.com/index/video-generation-models-as-world-simulators/, 2024.

[9] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4603–4623. PMLR, 2024.

[10] ByteDance Seed. Doubao-Seed-2.1-Pro. Volcano Engine Ark, https://www.volcengine.com/product/ark, 2026. Model snapshot doubao-seed-2-1-pro-260628.

[11] ByteDance Seed. One-take creation, flexible referencing: Introducing Seedance 2.5. ByteDance Seed Blog, https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5, 2026.

[12] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 27713–27724, 2025.

[13] Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Naveed 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. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 33340–33366. PMLR, 2026.

[14] 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. *arXiv preprint arXiv:2511.19861*, 2025.

[15] Leslie Gu, Junhwa Hur, Charles Herrmann, Fangneng Zhan, Todd Zickler, Deqing Sun, and Hanspeter Pfister. GeCo: Evaluating geometric consistency for video generation via motion and structure. *arXiv preprint arXiv:2512.22274*, 2026.

[16] Yanjiang Guo, Haodong Yan, Zhide Zhong, Zhongru Zhang, Qingyuan Yang, Qingzhou Lu, Xiaoyu Chen, Yen-Jen Wang, Shuying Deng, Chenghan Yang, Puzhen Yuan, Chenxin Liu, Tun Ban, Xiang Zhu, Yichen Liu, Kun Feng, Haoang Li, and Jianyu Chen. Video prediction policy 2: Predict better, act better. *arXiv preprint arXiv:2610.10270*, 2026.

[17] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

[18] Raia Hadsell, Diego Rivas, and Hang Qi. Genie 4: A new generation of world simulation. Google DeepMind Blog, https://deepmind.google/blog/genie-4-a-new-generation-of-world-simulation/, 2026.

[19] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

[20] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

[21] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, 2024.

[22] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 28991–29017. PMLR, 2025.

[23] Kling AI. Meet all-new Kling 4.0! Kling AI Release Notes, https://kling.ai/release-note/release-notes/Kling_4, 2026.

[24] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2025.

[25] Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. WorldEval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025.

[26] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22139–22149, 2024.

[27] Meta AI. SAM 3.1: Faster and more accessible real-time video detection and tracking with multiplexing and global reasoning. Meta AI Blog, https://ai.meta.com/blog/segment-anything-model-3/, 2026.

[28] MiniMax. MiniMax-H3. Hugging Face model card, https://huggingface.co/MiniMaxAI/MiniMax-H3, 2026.

[29] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 948–958, 2026.

[30] NVIDIA et al. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

[31] NVIDIA et al. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026.

[32] OpenAI. Sora 2 is here. https://openai.com/index/sora-2/, 2025.

[33] Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. Google DeepMind Blog, https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/, 2025.

[34] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, and Ruimao Zhang. WorldSimBench: Towards video generation models as world simulators. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 50338–50362. PMLR, 2025.

[35] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

[36] 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. *arXiv preprint arXiv:2602.08971*, 2026.

[37] Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[38] Archana Warrier, Dat Nguyen, Michelangelo Naim, Moksh Jain, Yichao Liang, Karen Schroeder, Cambridge Yang, Joshua B. Tenenbaum, Sebastian Josef Vollmer, Kevin Ellis, and Zenna Tavares. Benchmarking world-model learning with environment-level queries. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 133215–133246. PMLR, 2026.

[39] Bing Wu, Chang Zou, Changlin Li, et al. HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

[40] Yuanyang Yin, Gongxuan Wang, Yifan Zhan, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-EVOKE: From linear-scaling supervision to endless world. *arXiv preprint arXiv:2608.13546*, 2026.

[41] Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

[42] Songchun Zhang, Yaowei Li, Junhao Zhuang, Weiyang Jin, Haoyu Wang, Xin Lu, Yilang Sun, Shiyi Zhang, Haoran Li, Xiaoxiao Ma, Yuming Li, Yijun Liu, Yaofeng Su, Yanwen Ma, Haoyu Wu, Zihan Su, Yue Ma, Lvmin Zhang, Haoyang Huang, Zeyue Xue, Anyi Rao, and Nan Duan. EchoWM: Open and enterable omnimodal world models. *arXiv preprint arXiv:2608.23189*, 2026.

[43] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. CamI2V: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

[44] Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026.