WFM-Eval: An Evaluation Framework for Video World Models in Robotic Manipulation

Sahil Khose
UC Irvine
skhose@uci.edu

Mengqi Zhang
UC Irvine

Prithvijit Chattopadhyay
NVIDIA

Judy Hoffman
UC Irvine

Abstract

Video world models can generate synthetic training data for robot manipulation, but standard metrics (FVD, FID) miss the failure modes that determine whether policies trained on that data succeed. VLM-based holistic judges are a natural alternative, but no single VLM reliably predicts task completion: judges exhibit opposing biases that do not cancel under ensembling. We introduce WFM-Eval, a multi-dimensional evaluation framework that decomposes video quality into three axes (task completion, object hallucination, and temporal consistency) with structured object-level diagnostics that bypass holistic judgment. We benchmark five video world models (Cosmos Predict2, Predict2.5, Veo 3.1, HunyuanVideo 1.5, Wan2.2) on GR1 and AgiBot. We find that (1) object hallucination is the dominant model-discriminative failure mode, and (2) model rankings reverse between datasets: the GR1 leader is last on AgiBot and the AgiBot leader is mid-pack on GR1, so single-dataset benchmarks mislead. As supporting evidence, the framework's hallucination ranking aligns with downstream policy success on LIBERO, where Predict2 outperforms Predict2.5 by 8.75 points. We release WFM-Eval as an open toolkit.

1. Introduction

Video world models [1, 2] offer a path to scalable robot learning: rather than collecting expensive teleoperated demonstrations, a world model can *imagine* task execution from a single image and a language instruction, synthesizing video that is then distilled into robot policies via inverse dynamics models [7, 24]. The pipeline has shown early success across embodiments from tabletop manipulators to humanoid robots.

The promise rests on a largely unexamined assumption: that generated videos are faithful enough to teach correct behavior. Standard metrics such as Fréchet Video Distance (FVD) [34], Fréchet Inception Distance (FID) [21], and CLIP-Score [20] measure distributional realism and coarse semantic alignment, but are blind to the failure modes that corrupt downstream policy learning. A generated video may depict the wrong manipulation, hallucinate objects into or out of the scene, or teleport them between frames. These failures have distinct downstream consequences that a single scalar conflates.

A natural alternative is to use a VLM as a holistic judge of task completion. We test this and find it unreliable: judges exhibit opposing failure modes (Qwen3-VL [4] and InternVL3.5-14B [36] over-predict success with high recall and moderate precision, Qwen2.5-VL [5] over-predicts failure, Cosmos-Reason1 [3] sits closer to balanced but underperforms, and even the best single judge Kimi-K2.5 [33] reaches only 69.5% F1) and ensembling does not cancel them. These findings echo known VLM limitations on physical and temporal reasoning [9, 13, 25, 26]. Holistic judgment alone is too coarse to ground a benchmark.

We therefore introduce WFM-Eval, a multi-dimensional evaluation framework inspired by TIDE [8] that decomposes video quality along three complementary axes: whether the instructed *task was completed*, whether *objects are hallucinated* or lost, and whether the video is *temporally consistent*. Each axis targets a distinct mechanism by which generated data can corrupt a downstream policy. We benchmark five video world models (Cosmos Predict2 [30], Predict2.5 [2], Veo 3.1 [15], HunyuanVideo 1.5 [37], Wan2.2 [35]) on GR1 and AgiBot, and validate the framework downstream by training LIBERO [27] policies.

Our contributions are: (1) a structured error taxonomy with three evaluation dimensions and four fine-grained event subtypes, paired with an automated pipeline using VLM grounding for object-level diagnosis; (2) a cross-family empirical study showing that model rankings reverse between datasets, and that domain-specific training trades generalization for in-distribution peak performance; and (3) downstream validation on LIBERO showing the framework's hallucination ranking aligns with policy success.

Accepted at the 1st Workshop on Video World Models, CVPR 2026.
Accepted at the 2nd Workshop on Foundation Models Meet Embodied Agents, CVPR 2026.

2. Related Work

Video World Models for Robotics. World models have evolved from latent-space predictors [16–18] to pixel-space video generators. Recent methods adapt video diffusion models into world models that produce robotic rollout videos for navigation [14, 22] and manipulation [7, 24]. Foundation-scale models such as Cosmos-Predict [1, 2] and Genie2 [31] demonstrate that generated videos paired with inverse dynamics models can yield training data competitive with real teleoperation. General-purpose video generators (Veo [15], HunyuanVideo [37], Wan [35]) are also being adapted to robotic settings, raising the question of whether domain-specific or general-purpose priors transfer better. Whether generated rollouts are task-faithful and physically grounded, rather than merely visually plausible, remains open across both classes.

Video Generation Evaluation. Early approaches relied on distribution-based metrics (FVD [34], FID [21]) and semantic metrics (CLIP-Score [20]), which provide limited insight into real-world faithfulness. Multi-dimensional benchmarks such as VBench [23] and EvalCrafter [28] broaden evaluation using specialist models, while follow-up efforts [6, 29] leverage VLMs to reason about physical law violations. Embodied benchmarks [12, 38, 39] tailor evaluation to robot world models but assess individual dimensions in isolation, and few report whether their metrics align with downstream policy performance. WFM-Eval addresses this gap with a robotics-aware taxonomy that decomposes generation quality into task incompletion, object hallucination, and physically implausible dynamics, and validates the framework against policy success on LIBERO.

VLMs as Evaluators. VLMs have been increasingly used as automated judges for video quality [19], yet physics-centered benchmarks [9, 25] report near-chance VLM performance on object permanence, continuity, and solidity. Complementary work on causal reasoning and hallucination [13, 26] further exposes weaknesses in the capabilities required for task-level evaluation. Our cross-judge analysis (§4.2) confirms these limitations in the robotics setting: opposing biases across judges do not cancel under ensembling. WFM-Eval complements VLM scoring with structured, per-object hallucination and temporal consistency diagnostics.

3. WFM-Eval Framework

3.1. Error Taxonomy

We define a three-level error taxonomy. Each dimension targets a distinct failure mode that can corrupt downstream policy learning.

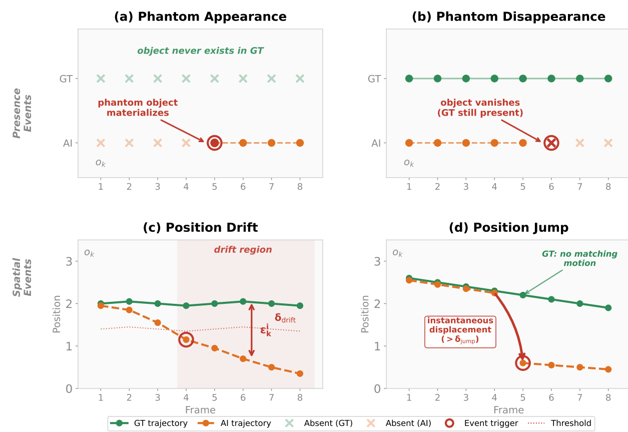

Figure 1. **Schematic of the four error event types in WFM-Eval.** Top row: presence events, (a) phantom appearance and (b) phantom disappearance. Bottom row: spatial events, (c) position drift and (d) frame jump.

Dimension 1: Task Completion ($\mathcal{E}_{\text{task}}$). Does the generated video depict successful execution of the instructed manipulation? A visually plausible video that teaches incorrect behavior is the highest-level failure.

Dimension 2: Object Hallucination ($\mathcal{E}_{\text{hall}}$). Do objects in the generated video match those in the GT initial frame? Policies trained on hallucinated objects may reach for targets that do not exist. We distinguish *Phantom Appearance* (E2a), where an object is detected with no GT counterpart, and *Phantom Disappearance* (E2b), where a GT object ceases to be detected.

Dimension 3: Temporal Consistency ($\mathcal{E}_{\text{temp}}$). Is the generated video temporally coherent? We distinguish *Position Drift* (E3a), where an object's position diverges from GT beyond δ_{drift}, and *Frame Jump* (E3b), where instantaneous displacement exceeds δ_{jump} with no corresponding GT motion. Figure 1 illustrates all four types.

3.2. Evaluation Pipeline

Let $\hat{V} = \{\hat{f}_1, \ldots, \hat{f}_M\}$ denote the generated video, $V^* = \{f_1^*, \ldots, f_N^*\}$ the GT reference, T the language instruction, and $\mathcal{O}$ the set of tracked objects with $K = |\mathcal{O}|$. WFM-Eval runs two parallel tracks.

Track A: Task Completion. A VLM judges $f_{\text{VLM}} : (\hat{V}, T) \to \{0, 1\}$ where 1 indicates success. Predictions are compared against human annotations to compute accuracy, precision, recall, and F1.

Track B: Object Hallucination & Temporal Consistency. For each object o_k at frame t, Molmo [11] returns a point detection $d_k^t = (x_k^t, y_k^t) \in \mathbb{R}^2$ or $\varnothing$ (absent). Biases in this single grounding model propagate to all downstream metrics, partially mitigated by step (2).

(1) Frame alignment. When generated and GT videos differ in length, we map each generated frame index to its

Table 1. **Dataset overview.** Six datasets with diverse robot manipulation scenarios.

Dataset	Samples	Robot	Description
AgiBot [10]	210	Bimanual	Kitchen manipulation
RigVid [32]	40	Tabletop	Rigid object handling
GR1-100 [24]	92	Humanoid	Pick-and-place
EVAL-175-Obj [24]	50	Humanoid	Object variations
EVAL-175-Beh [24]	47	Humanoid	Behavior variations
EVAL-175-Env [24]	29	Humanoid	Environment variation
Total	**468**		

nearest GT counterpart via linear interpolation.

(2) Temporal filtering. We apply a voting filter (μ=3 detections within window W=5) and a velocity filter (rejecting physically implausible motion).

(3) Per-object event detection. Phantom events are triggered when presence transitions in the generated video have no corresponding GT transition within tolerance τ. Spatial events are flagged when positional error exceeds δ_{drift} or inter-frame displacement exceeds δ_{jump} without matching GT motion.

(4) Aggregate metrics. Let $\mathcal{E}$ denote detected events with counts $n_{\text{pa}}, n_{\text{pd}}, n_{\text{dr}}, n_{\text{jp}}$. We define:

$$\text{OHR} = \frac{n_{\text{pa}}+n_{\text{pd}}}{M}, \quad \text{PAR} = \frac{n_{\text{dr}}+n_{\text{jp}}}{M}, \quad \text{TCS} = 1 - \frac{n_{\text{jp}}}{MK}. \tag{1}$$

$$\text{HSS} = \min\Big(1, \frac{\sum_{e\in\mathcal{E}} w(e)}{MK}\Big), \tag{2}$$

with weights $w(e)$: phantom appearance = 10, disappearance = 8, jump = 5, drift = 3, ranked by relative downstream harm. Per-axis metrics (OHR, PAR) are reported alongside HSS so conclusions do not rest on the exact weighting. Lower HSS is better; HSS=0 indicates no hallucinations.

4. Experiments

4.1. Setup

Datasets. We evaluate on six robotic manipulation datasets (Tab. 1) spanning diverse embodiments, tasks, and environments: AgiBot [10] (210 bimanual kitchen manipulation clips), RigVid [32] (40 rigid object clips), GR1-100 [24] (92 humanoid pick-and-place clips), and three EVAL-175 splits [24] covering object (50), behavior (47), and environment (29) variations.

Models. We evaluate two robotics-specific video world models, Cosmos Predict2 (2B) [30] and Predict2.5 (2B) [2], against three general-purpose video generators: Veo 3.1 [15], HunyuanVideo 1.5 (13B) [37], and Wan2.2 (14B) [35]. Predict2 is evaluated across four resolution×framerate combinations (480p/720p × 10/16 fps); other models at their available configurations.

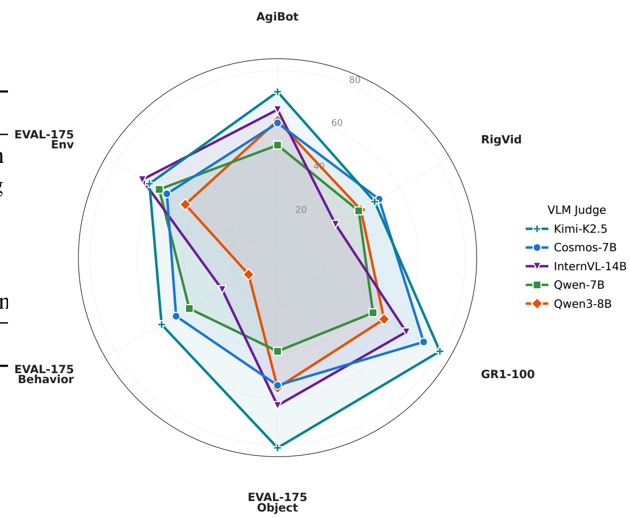

Figure 2. **VLM judge F1 by dataset.** Kimi-K2.5 (teal) leads on four of six datasets, trails Cosmos-Reason1-7B by ∼2 F1 points on RigVid, and trails InternVL3.5-14B by ∼3 F1 points on EVAL-175-Env; among the remaining judges no single model dominates uniformly, with each exhibiting distinct dataset-specific biases.

Scope. Track A uses all 468 samples across the six datasets. Cross-family hallucination analysis (§4.3) is run on GR1 and AgiBot, the two datasets for which all five model families have generated rollouts.

VLM judges and detectors. Task completion judges include Qwen2.5-VL-7B [5], Qwen3-VL-8B [4], InternVL3.5-8B/14B [36], Cosmos-Reason1-7B [3], and Kimi-K2.5 [33]. Molmo-7B [11] serves as the object detector. Human binary annotations provide ground truth for all 468 samples.

4.2. Task completion (Track A)

No VLM reliably judges task completion (Fig. 2). The best single judge (Kimi-K2.5) achieves 74.4% accuracy and 69.5% F1 (micro-averaged), with weak, non-significant correlation to human completion rates (Pearson r=0.287, p=0.248). Models fail in distinct, consistent ways: Qwen3-VL and InternVL3.5-14B over-predict success (high recall, moderate precision), Qwen2.5-VL over-predicts failure, and Cosmos-Reason1 lands closer to balanced but still below the threshold for benchmark use. Ensembling does not resolve these opposing biases. These findings echo known VLM limitations on physical and temporal reasoning [9, 13, 25, 26], motivating Track B's structured per-object diagnostics.

4.3. Cross-family object hallucination

We compare five model families on GR1 (n=92) and AgiBot (n=210): the robotics-specific Cosmos Predict2 and Predict2.5, and three general-purpose generators (Veo 3.1,

Table 2. **Cross-family hallucination on GR1 and AgiBot.** Each model at its best configuration per dataset; lower HSS is better. Rank is by HSS within each dataset; 1st and 5th highlighted. Predict2 collapses from rank 1 on GR1 to rank 5 on AgiBot, while Hunyuan rises 3→1, so single-dataset benchmarks mislead. P2.5 = Cosmos Predict2.5; Hunyuan = HunyuanVideo 1.5.

	GR1				AgiBot			
Model	**Rank**	HSS	OHR	PAR	**Rank**	HSS	OHR	PAR
Predict2	**1**	**.396**	.108	.067	**5**	.699	.341	.128
Veo 3.1	2	.436	.171	.050	2	.630	.274	.131
Hunyuan	3	.482	.200	.048	**1**	**.596**	.256	.128
P2.5	4	.501	.210	.049	4	.682	.322	.128
Wan2.2	**5**	.518	.207	.058	3	.661	.301	.145

HunyuanVideo 1.5, Wan2.2). Each family is reported at its best configuration per dataset.

Rankings reverse across datasets. Tab. 2 shows model rankings on GR1 against AgiBot. Predict2 leads on GR1 (HSS=0.396) but drops to last on AgiBot (0.699). Hunyuan follows the opposite trajectory, rising from rank 3 to rank 1 (0.482→0.596). Veo 3.1 is the only family that holds rank across both datasets. The same model can therefore be best-in-class or worst-in-class depending on the embodiment, and single-dataset benchmarks mislead.

Domain training trades generalization for peak performance. Predict2's HSS degrades by 81% from GR1 to AgiBot, against 19% for Hunyuan and 44% for Veo 3.1. The robotics-specific prior helps inside its training distribution and hurts outside it. Parameter count does not explain the pattern: 2B Predict2 beats 14B Wan2.2 by 32% on GR1, while 13B Hunyuan beats 2B Predict2 by 17% on AgiBot.

Predict2.5 regresses on GR1. At matched configuration (720p, 16 fps), Predict2.5 raises HSS by +0.095 over Predict2 (0.406→0.501) and OHR by +0.102 (10.9%→21.0%). Position Anomaly Rate decreases (0.066→0.049): Predict2.5 invents more objects but places them more stably. A 52% increase in phantom appearances per frame-object drives the regression, and the same ranking shows up downstream (§4.4).

Bimodal severity hides in the mean. Figure 3 shows per-clip severity distributions. On AgiBot the two distributions overlap tightly (KS p=0.817). On GR1 they diverge: Predict2.5 introduces a spike at severity 1.0 absent under Predict2, with 56 of 92 clips scoring worse under Predict2.5 and the regressions concentrated among clips Predict2 handles cleanly (severity < 0.3). A small subset of catastrophic failures coexists with broad parity, hidden by the mean.

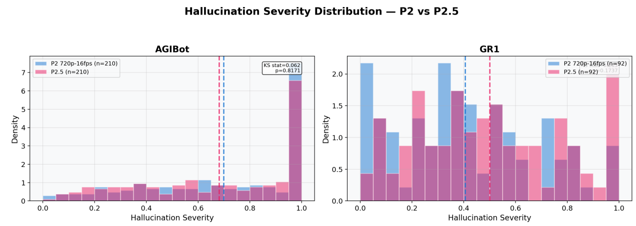

Figure 3. **Hallucination severity distribution: P2 vs. P2.5.** Histograms for AgiBot (*left*) and GR1 (*right*). On AgiBot the distributions nearly coincide (KS p=0.817). On GR1, Predict2 carries more mass in the clean range while Predict2.5 introduces a spike at severity 1.0, indicating catastrophic failures on a subset of clips.

4.4. Downstream validation on LIBERO

To check whether WFM-Eval's verdicts track downstream utility, we evaluate Predict2 and Predict2.5 on LIBERO [27] across its spatial, object, goal, and long-horizon suites.

Setup. Each model is fine-tuned on in-domain LIBERO data and adapted for action prediction following [24]. A single multitask policy is trained over the four suites with 50 demonstrations per task and evaluated under the standard protocol (3 seeds × 50 episodes).

Results. Predict2 outperforms Predict2.5 on every suite. The gap is largest on LIBERO-Long (89.6% vs. 69.4%), and Predict2 averages **95.45%** success against **86.70%** for Predict2.5.

Validation. The downstream ranking matches WFM-Eval on the discriminative axes: at matched configuration (720p, 16 fps) on GR1, Predict2 has lower HSS (0.406 vs. 0.501) and lower OHR (10.9% vs. 21.0%). The 52% increase in phantom appearances behind Predict2.5's regression (§4.3) is consistent with the policy drop: a generator that invents objects produces training rollouts that mislead the policy. This is a single-pair test, not a correlation study, but it ties the framework's diagnostics to end-task performance.

5. Conclusion

We presented WFM-Eval, an evaluation framework that decomposes video quality into task completion, object hallucination, and temporal consistency with fine-grained event subtypes, providing per-axis diagnostics that aggregate metrics cannot. Benchmarking five model families on GR1 and AgiBot, we find that (1) object hallucination is the dominant discriminative failure mode, with Predict2.5's regression on GR1 driven by a 52% increase in phantom appearances; (2) rankings reverse across datasets (Predict2 leads on GR1 but trails on AgiBot; Hunyuan does the opposite), so single-dataset benchmarks mislead; and (3) the hallucination ranking transfers to LIBERO, where Predict2 outperforms Predict2.5 by 8.75 points. We release WFM-Eval as an open toolkit.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1, 2

[2] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025. 1, 2, 3

[3] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. 1, 3

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1, 3

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 1, 3

[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. *arXiv preprint arXiv:2406.03520*, 2024. 2

[7] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. 1, 2

[8] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision*, pages 558–573. Springer, 2020. 1

[9] Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025. 1, 2, 3

[10] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025. 3

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 2, 3

[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*, pages 27713–27724, 2025. 2

[13] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025. 1, 2, 3

[14] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024. 2

[15] Google DeepMind. Veo 3.1. `https://deepmind.google/models/veo/`, 2025. 1, 2, 3

[16] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018. 2

[17] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[18] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. 2

[19] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, 2024. 2

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021. 1, 2

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 2

[22] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: a generative world model for autonomous driving. *URL https://arxiv. org/abs/2309.17080*, 3, 2023. 2

[23] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2

[24] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking gen-

eralization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025. 1, 2, 3, 4

[25] Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*, 2023. 1, 2, 3

[26] Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Fuxiao Liu, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *arXiv preprint arXiv:2505.01481*, 2025. 1, 2, 3

[27] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. 1, 4

[28] 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*, pages 22139–22149, 2024. 2

[29] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 2

[30] NVIDIA Research. Cosmos-predict2. https://research.nvidia.com/labs/dir/cosmos-predict2/, 2025. 1, 3

[31] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model, 2024. 2

[32] Shivansh Patel, Shraddhaa Mohan, Hanlin Mai, Unnat Jain, Svetlana Lazebnik, and Yunzhu Li. Robotic manipulation by imitating generated videos without physical demonstrations. *arXiv preprint arXiv:2507.00990*, 2025. 3

[33] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng,

Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026. 1, 3

[34] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. *International Conference on Learning Representations Workshop*, 2019. 1, 2

[35] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2, 3

[36] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 3

[37] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025. 1, 2, 3

[38] Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025. 2

[39] Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai. *arXiv preprint arXiv:2512.01989*, 2025. 2