A Physics-Grounded Benchmark for Multi-Agent Dynamics in World Models

Nuo Chen[1,*] Lulin Liu[1,2,*] Zihao Li[1] Ziyao Zeng[3] Zihao Zhu[1]
Wenyan Cong[4] Junyuan Hong[4] Yunhao Yang[4] Zhengzhong Tu[1] Yan Wang[5]
Boris Ivanovic[5] Marco Pavone[5,6] Zhangyang Wang[4] Yang Zhou[1] Zhiwen Fan[1]

[1]Texas A&M University [2]University of Minnesota [3]Yale University
[4]University of Texas at Austin [5]NVIDIA [6]Stanford University

[*]Equal contribution

Abstract

*Generative world models hold promise as scalable simulators for autonomous systems, particularly for rare safety-critical multi-agent interactions such as vehicle collisions. However, current evaluation paradigms index heavily on visual fidelity and semantic alignment, leaving a critical blind spot: they rarely quantify whether generated dynamics obey the physical laws required for reliable simulation. To bridge this gap, we introduce **CrashTwin**, a physics-grounded evaluation framework designed to stress-test the physical trustworthiness of world models. CrashTwin combines 25K synthetic sequences and 12K realworld crash sequences with a calibration-free reconstruction pipeline that recovers metric-scale physical attributes from uncalibrated videos. We evaluate spatio-temporal consistency, momentum and energy conservation, and world-dynamics integrity. Benchmarking representative world models reveals that high perceptual quality can mask severe physical violations during complex interactions. By quantitatively exposing these failure modes, CrashTwin provides a vital diagnostic tool for developing physically grounded world models capable of reliable real-world simulation.*

1. Introduction

Generative world models [14, 28–30, 36] have emerged as promising tools for scalable simulation in autonomous driving, offering a pathway to synthesize safety-critical corner cases that are rare in the real world [1, 32]. For these rollouts to be actionable in downstream data curation, they must not only exhibit high visual fidelity but also preserve physically valid motion and interaction dynamics. However, current evaluation protocols predominantly index on perceptual quality, temporal smoothness, and semantic alignment [16, 26, 35]. As shown in Fig. 1, severe violations of

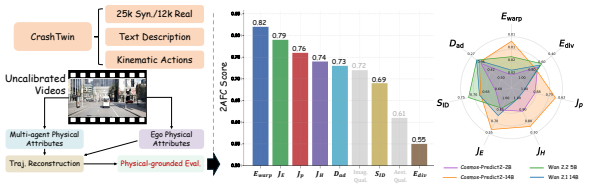

Figure 1. **CrashTwin teaser.** Physics metrics align with 2AFC preferences and expose failures missed by visual-quality proxies.

fundamental physics can remain hidden under standard visual assessments, and conventional visual metrics correlate poorly with human judgments of physical realism.

Rigorously quantifying physical plausibility remains difficult. Existing frameworks often rely on visual proxies, distributional metrics, or VLM judges [3, 13, 16, 17, 26], which are insufficient for precise physical validation. The bottleneck lies in physical observability, since verification requires metric-scale trajectories, velocities, yaw motion, and contact timing. Yet recovering these attributes from uncalibrated monocular videos is highly ill-posed, as ego-motion, camera intrinsics, scene scale, depth, and multi-agent dynamics are entangled. Rapid motion, occlusion, and abrupt post-impact changes further compound this challenge, while current benchmarks lack targeted safety-critical data for exposing such dynamic failures.

We focus on vehicle-to-vehicle collisions as a representative task for evaluating multi-agent interactions. This setting provides a rigorous stress test for generative world models, since physical impacts induce abrupt, coupled state changes whose post-contact evolution is tightly governed by classical conservation laws [4, 5]. Compared with heterogeneous accident types, vehicle-to-vehicle collisions offer more consistent geometry and motion patterns, yielding clearer geometric priors and kinematic constraints [37]. Verifying their structural integrity is crucial before world-model rollouts can be trusted as reliable simulations.

We present **CrashTwin**, a framework for quantifying the

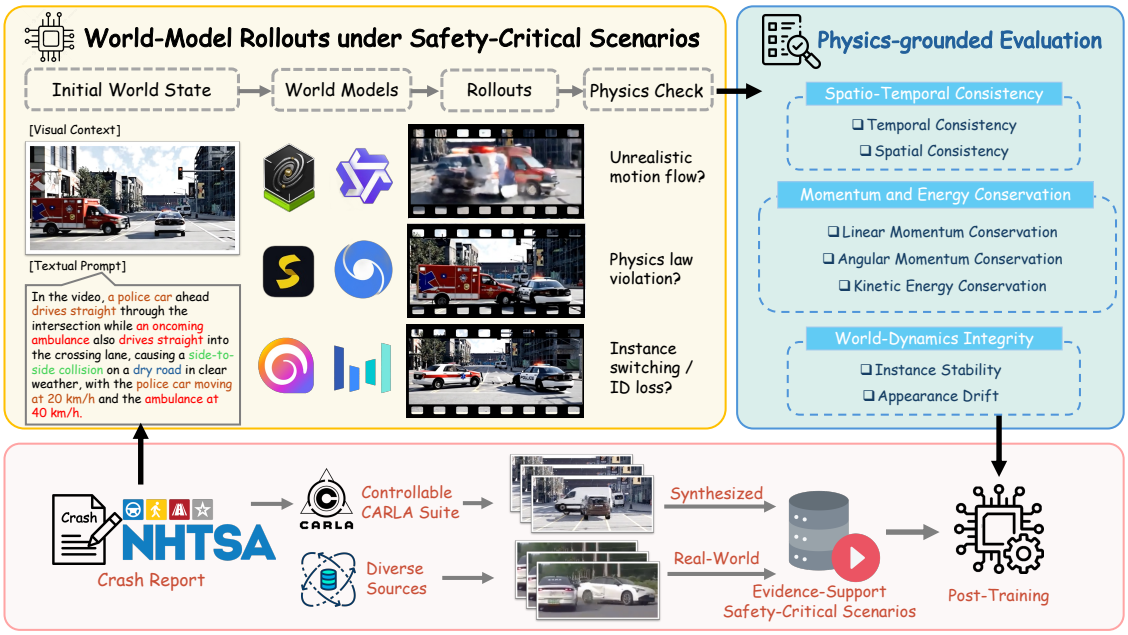

Figure 2. **CrashTwin main overview.** We evaluate collision rollouts by combining controlled crash data, real-world videos, physical reconstruction, and metrics for spatio-temporal consistency, momentum and energy conservation, and world-dynamics integrity.

physical integrity of world-model rollouts. CrashTwin combines 25K synthetic sequences generated through a physically grounded CARLA pipeline [7] with 12K diverse in-the-wild traffic incident videos annotated with text descriptions, poses, and kinematic action labels. To evaluate uncalibrated rollouts, we develop a calibration-free reconstruction pipeline that leverages vision foundation models to recover metric-scale physical attributes under complex interactions. Based on the recovered global trajectories, we introduce a diagnostic protocol over spatio-temporal consistency, momentum and energy conservation, and world-dynamics integrity. Benchmarking representative world models shows that our metrics expose physical failures largely invisible to appearance-based evaluation and align more strongly with human judgments of physical realism.

Our contributions are threefold.

- We introduce **CrashTwin**, a large-scale benchmark coupling controllable synthetic data with unconstrained real-world collision videos for physics-grounded evaluation.
- We develop a physics-grounded evaluation framework with a calibration-free reconstruction pipeline that recovers metric-scale attributes from monocular videos.
- We systematically benchmark representative world models, quantitatively diagnose their physical failure modes, and show that our proposed metrics align more strongly with human judgments of physical realism.

2. Related Work

Video Generation & World Models. Recent video generators and driving world models can synthesize high-fidelity traffic rollouts from text, scene, or motion conditions [11, 12, 14, 23, 25, 33, 34]. Some crash-oriented systems improve visual coverage of rare accidents, but often rely on strong trajectory or layout conditioning, making them closer to appearance completion than unconstrained physical simulation. CrashTwin instead evaluates whether generated collision dynamics are physically plausible when the model itself must produce the post-impact evolution.

Physics-grounded Evaluation. Existing evaluation is still mostly perception-centric, relying on FID/FVD, visual-quality benchmarks, or VLM/LLM judges [3, 13, 16, 17, 19, 26]. Recent physics-aware metrics probe flow, depth, scripted violations, or commonsense physical reasoning [20, 21, 31], but they rarely recover metric-scale quantities or localize failures at the collision window. CrashTwin fills this gap with metric crash dynamics.

3. Methodology

Spatio-temporal consistency. We evaluate whether generated motion remains smooth over time and locally rigid.

Temporal coherence. We measure frame-to-frame consistency by flow warping error within the actor mask:

$$E_{\text{warp}} = |\Omega|^{-1} \sum_{p \in \Omega} |I_t(p) - I_{t+1}(p + F_{t \to t+1}(p))| . \quad (1)$$

Here I_t is the frame at time t, p indexes foreground pixels in Ω, and $F_{t \to t+1}(p)$ is optical flow from t to $t+1$.

Spatial rigidity. We measure local expansion and compres-

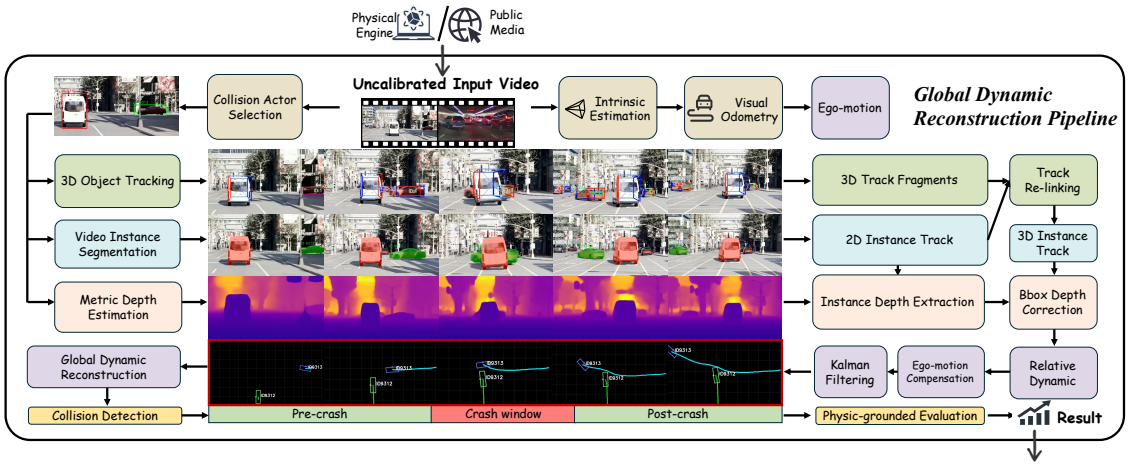

Figure 3. **Global dynamic reconstruction pipeline.** The system reconstructs 3D collision dynamics from uncalibrated accident videos by combining 3D tracking, segmentation, metric depth, and visual odometry. Re-linked tracks, metric scale recovery, and ego-motion compensation produce global per-actor trajectories, which define pre-crash, crash, and post-crash windows for physics-based evaluation.

sion by the divergence of the flow field:

$$E_{\text{div}} = |\Omega|^{-1} \sum_{p \in \Omega} |\nabla_x u(p) + \nabla_y v(p)| \,. \tag{2}$$

Here $F_{t \to t+1} = (u, v)$, with u and v denoting horizontal and vertical flow components. Lower E_{warp} and E_{div} indicate stronger temporal and geometric consistency.

Momentum and energy conservation. For each vehicle i, with mass m_i, position r_i, velocity v_i, yaw rate ω_i, and yaw inertia $I_{z,i}$, we compare pre-impact and post-impact states in a short collision window centered at first contact.
Linear momentum. We compute a normalized residual:

$$J_p = \frac{\left\| \sum_i m_i v_i^+ - \sum_i m_i v_i^- \right\|}{\sum_i m_i \| v_i^- \|}. \tag{3}$$

Here v_i^- and v_i^+ are velocities before and after impact, and the denominator normalizes by incoming momentum.
Angular momentum. We compute an angular residual:

$$J_H = \frac{\| H_c^+ - H_c^- \|}{\left\| H_c^- \right\| + \varepsilon}, \quad H_c = \sum_i (I_{z,i}\omega_i + m_i (r_i - c) \times v_i). \tag{4}$$

Here H_c sums yaw and translational angular momentum around contact point c, and ε avoids zero denominators in low-angular-momentum cases.
Kinetic energy. The kinetic-energy gain penalty is:

$$J_E = \max\left(0, \frac{E_k^+ - E_k^-}{E_k^-}\right), \quad E_k = \tfrac{1}{2} \sum_i m_i \| v_i \|^2. \tag{5}$$

Here E_k^- and E_k^+ are total kinetic energies before and after impact. The one-sided penalty allows dissipative loss but flags impossible energy creation. Small residuals indicate plausible impulse exchange, while large values expose missing momentum transfer, nonphysical torque, or impossible energy gain.

World-dynamics integrity. We assess identity persistence and appearance drift during impact.
Instance stability. We measure identity persistence with the Simpson index:

$$S_{\text{ID}}^{(i)} = \sum_k f_{k,i}^2, \quad f_{k,i} = c_{k,i} / \sum_j c_{j,i}. \tag{6}$$

Here $c_{k,i}$ counts frames where physical instance i receives track ID k, and $f_{k,i}$ is the corresponding frame fraction.
Appearance drift. We measure visual drift from masked CLIP embeddings along each vehicle trajectory:

$$D_{\text{ad}}^{(i)} = (T-1)^{-1} \sum_{k=1}^{T-1} \arccos\left(\langle \hat{z}_{i,t_k}, \hat{z}_{i,t_{k+1}} \rangle \right). \tag{7}$$

Here $\hat{z}_{i,t_k}$ is the normalized CLIP embedding of instance i at sampled time t_k, and T is the number of samples. Higher S_{ID} and lower D_{ad} indicate more stable world dynamics.

Reconstruction pipeline. In Figure 3, we summarize how we convert an uncalibrated monocular rollout into actor trajectories, scale, yaw, and contact timing before applying the metrics above.
Actor selection. We first choose the two incident vehicles, using simulator identities for synthetic clips and real-video annotations for curated crashes.
Tracking fragments. Starting from the uncalibrated video, we estimate camera intrinsics with MapAnything [18], ego motion with DROID-SLAM [24], metric scale with Metric3D V2 [15], initial 3D tracks with CenterTrack [38], and instance masks with SAM2 [22].
Relinking and correction. SAM2 masks relink broken fragments through occlusion, while instance-depth correction restores scale for metric velocity estimates.

Table 1. **CrashTwin Evaluation Leaderboard.** Open-source models are evaluated on CrashTwin-Eval, while proprietary models are evaluated on CrashTwin-Eval-Mini due to API and rate-limit constraints. Metrics are grouped into three physical families: (A) spatio-temporal consistency, (B) momentum and energy conservation, and (C) world-dynamics integrity. Lower is better unless marked with ↑.

	Spatio-temporal Consistency		Momentum & Energy Conservation			World-dynamics Integrity	
Models	Warp Error	Flow Divergence	Momentum Residual	Angular Residual	Energy Gain	Instance Stability	App. Drift
	$E_{\text{warp}} \downarrow$	$E_{\text{div}} \downarrow$	$J_p \downarrow$	$J_H \downarrow$	$J_E \downarrow$	$S_{\text{ID}} \uparrow$	$D_{\text{ad}} \downarrow$
Open-Source Models							
SkyReel-1.3B [6]	0.0227	0.6103	0.9566	0.9628	0.9457	0.6660	0.3592
Wan 2.1-14B [27]	0.0179	0.6320	0.8235	0.8494	0.7864	0.6760	0.3117
Wan 2.2-5B [27]	0.0145	**0.5959**	0.8899	0.8975	0.8649	**0.7254**	**0.3109**
Cosmos-Predict2-2B [2]	0.0240	0.6748	0.8890	0.8954	0.8590	0.6129	0.3462
Cosmos-Predict2-14B [2]	**0.0117**	0.7180	**0.6828**	**0.7629**	**0.6047**	0.6737	0.3327
Proprietary Models							
Google Veo 3.1 [9]	**0.0097**	0.6202	0.7743	0.8011	0.7460	**0.7232**	0.3075
Hailuo 2.3 [10]	0.0151	0.6143	**0.7664**	**0.7812**	0.7285	0.6203	**0.2968**
Seedance V1 Pro [8]	0.0166	**0.6130**	0.7725	0.7837	**0.7209**	0.6007	0.3002

Global dynamics. Ego-motion compensation maps relative actor motion into a global frame, and Kalman filtering smooths positions and yaw.

Collision detection. We identify the first frame where inter-vehicle distance falls below a threshold, then extract pre-crash, crash, and post-crash windows for the metrics above.

4. Experiments

Datasets and protocol. CrashTwin contains 38K crash events: 25.6K controllable CARLA sequences and 12.6K curated real-world clips. Synthetic data cover seven NHTSA-derived intersection crash types with randomized speeds, approach directions, lighting, and weather, and provide 3D boxes, vehicle states, segmentation, depth, and optical flow at 30 Hz. Real clips are cropped around impact from public dashcam and traffic-camera footage and annotated with scene context, participants, motion, and first-frame collision actors. CrashTwin-Eval uses 300 synthetic and 44 real held-out videos; CrashTwin-Eval-Mini uses 100 synthetic and 16 real videos for proprietary models. We benchmark SkyReel-1.3B [6], Wan [27], and Cosmos-Predict2 [2] on CrashTwin-Eval, and Veo 3.1 [9], Hailuo [10], and Seedance [8] on CrashTwin-Eval-Mini. All rollouts use the same reconstruction pipeline, 20 FPS scoring grid, and matched pre-/post-impact windows.

Leaderboard observations. Table 1 reveals a consistent separation between visual coherence and physical correctness. Hailuo and Seedance obtain competitive drift scores but substantial conservation residuals, while Cosmos-Predict2-14B improves several conservation metrics without uniformly dominating world-dynamics measures. Thus E_{warp} and E_{div} diagnose unstable local geometry, whereas J_p, J_H, and J_E expose impulse-transfer, yaw-motion, and post-impact energy inconsistencies hidden in visually smooth videos.

Reconstruction validation. We validate reconstruction on simulated crashes with metric trajectories. A 3D tracker drifts under occlusion and abrupt post-impact motion; each added component in Tab. 2 reduces scene- and instance-level ATE, showing that relinking, filtering, and depth correction are required for reliable physical measurements.

Table 2. **Reconstruction accuracy on simulated crashes.** Mean absolute trajectory error (ATE); lower is better.

Configuration	**Scene ATE (m)** ↓		**Inst. ATE (m)** ↓	
	$SE(3)$	$Sim(3)$	$SE(3)$	$Sim(3)$
Basic 3D Tracking	11.71	9.38	3.89	1.98
+ Instance Relinking	5.96	5.10	3.73	1.89
+ Kalman Filtering	5.61	4.68	3.29	1.47
+ Metric Depth Correction	**5.48**	**3.70**	**2.63**	**0.91**

Human alignment. The two-alternative forced-choice (2AFC) study in Fig. 1 shows that annotators prefer rollouts with more plausible post-impact dynamics, even under similar visual quality. This suggests that our physics metrics capture dynamic validity beyond perceptual realism, including non-rigid flow, momentum imbalance, energy gain, identity switching, and drift.

5. Conclusion

We introduced CrashTwin, a physics-grounded benchmark for evaluating the physical plausibility of world models in multi-agent interaction scenarios. By coupling controlled synthetic crashes, curated real-world videos, calibration-free reconstruction, and direct physical metrics, CrashTwin tests whether generated rollouts respect impact constraints rather than merely look plausible. Our experiments show that current models still fail in spatio-temporal consistency, Momentum and energy conservation, and World dynamics integrity. CrashTwin therefore provides a diagnostic tool for data curation and world-model evaluation, motivating future world models that are both visually realistic and physically trustworthy.

References

[1] Waymo safety report. Technical report, Waymo LLC, 2021. Accessed 2026-03-04. 1

[2] 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. 4

[3] 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. 1, 2

[4] Raymond M Brach and R Matthew Brach. A review of impact models for vehicle collision. 1987. 1

[5] Anindya Chatterjee. *Rigid body collisions: some general considerations, new collision laws, and some experimental data*. Cornell University, 1997. 1

[6] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 4

[7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2

[8] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 4

[9] Google DeepMind. Veo 3.1. Technical report, Google DeepMind, 2026. Released: January 13, 2026. Accessed: March 5, 2026. 4

[10] HailuoAI. Hailuo. https://hailuoai.video/, 2024. Accessed: 2025-02-24. 4

[11] Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18858–18868, 2025. 2

[12] Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22404–22415, 2025. 2

[13] 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

[14] 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. *arXiv preprint arXiv:2309.17080*, 2023. 1, 2

[15] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. 3

[16] 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. 1, 2

[17] 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, 2024. *URL https://arxiv. org/abs/2411.02385*, 2:36. 1, 2

[18] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 3

[19] Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, Xinyu Zhang, et al. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2024. 2

[20] 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

[21] 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

[22] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[23] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8406–8416, 2025. 2

[24] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 3

[25] Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and

Hengshuang Zhao. Playerone: Egocentric world simulator. *arXiv preprint arXiv:2506.09995*, 2025. 2

[26] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1, 2

[27] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4

[28] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer, 2024. 1

[29] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024.

[30] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6902–6912, 2024. 1

[31] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, et al. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024. 2

[32] Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Ekaterina Tolstaya, Sarah Tang, Brandyn White, et al. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint arXiv:2510.26125*, 2025. 1

[33] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27220–27230, 2025. 2

[34] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10412–10420, 2025. 2

[35] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 1

[36] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer, 2024. 1

[37] Jing Zhou, Huei Peng, and Jianbo Lu. Collision model for vehicle motion prediction after light impacts. *Vehicle System Dynamics*, 46(S1):3–15, 2008. 1

[38] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020. 3