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

Nuo Chen[1*], Lulin Liu[1,2*], Zihao Li[3], Ziyao Zeng[4], Zihao Zhu[1], Wenyan Cong[5], Junyuan Hong[6,7], Yunhao Yang[5], Zhengzhong Tu[1], Yan Wang[8], Boris Ivanovic[8], Marco Pavone[8,9], Zhangyang Wang[5], Yang Zhou[1], Zhiwen Fan[1]

[1]Texas A&M University [2]University of Minnesota
[3]Marquette University [4]Yale University
[5]University of Texas at Austin [6]Massachusetts General Hospital
[7]Harvard Medical School [8]NVIDIA [9]Stanford University

[*]Equal contribution.

Project Website: `https://crash-twin.github.io/`

Abstract. Generative world models hold immense promise as scalable simulators for autonomous systems, particularly for synthesizing rare but 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 cannot reliably quantify whether generated dynamics actually obey the fundamental physical laws required for reliable simulation. Assessing this physical plausibility is inherently difficult due to a lack of physical metrics and the challenge of extracting metric-scale kinematics from uncalibrated video rollouts. To bridge this gap, we introduce **CrashTwin**, a physics-grounded evaluation framework designed to stress-test the physical trustworthiness of world models. CrashTwin couples a diverse dataset of multi-agent collision scenarios, comprising 25K controllable synthetic and 12K in-the-wild real-world collision sequences with a novel calibration-free reconstruction pipeline, enabling the recovery of 3D physical attributes directly from world model rollouts. We propose a diagnostic suite that systematically evaluates three dimensions: spatio-temporal consistency, momentum and kinetic energy conservation, and world-dynamics integrity. Extensive benchmarking of state-of-the-art models reveals a crucial insight: high perceptual quality frequently masks 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. Code and dataset are available at: https://github.com/phai-lab/CrashTwin.

1 Introduction

Generative world models [23, 47, 49, 50, 60] 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, 52]. For these

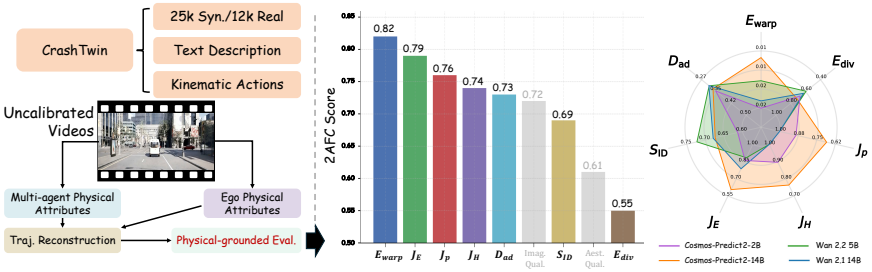

Fig. 1: CrashTwin Benchmark. Left illustrates the content of CrashTwin and the proposed evaluation pipeline that extracts physical attributes from uncalibrated videos to enable physics-grounded evaluation. **Middle** shows the two-alternative forced-choice (2AFC) scores, demonstrating that metrics derived from physical dynamics, indicated by colored bars, align more strongly with human preferences for physical realism compared to conventional visual quality proxies shown in grayscale. **Right** compares representative world models, revealing that both small and large models exhibit notable deficiencies across the proposed physics-grounded criteria.

rollouts to be actionable in downstream data curation, they must not only exhibit high visual fidelity but also strictly adhere to physical laws. However, the current evaluation paradigm predominantly indexes on perceptual quality and semantic alignment [25,44,59]. Consequently, a critical blind spot remains: severe violations of fundamental physics frequently go undetected under standard visual assessments. Furthermore, these conventional visual metrics correlate poorly with human judgments of physical realism, as illustrated in Fig. 1.

Rigorously quantifying this physical plausibility is inherently difficult, particularly in safety-critical scenarios. Existing frameworks often circumvent the underlying mechanics by relying on generalized visual proxies, including distributional similarity metrics such as Fréchet Video Distance [44] and appearance-focused benchmark scores such as the aesthetic and imaging quality [25, 44], or vision-language models (VLMs) as judges [5], which remain brittle under visual ambiguity and are insufficient for precise physical validation [33, 38]. The fundamental bottleneck lies in physical observability: to rigorously verify adherence to physical laws, one must extract metric-scale kinematics, such as precise trajectories, velocities, and angular motion. Yet, recovering these 3D physical attributes from uncalibrated monocular videos is a highly ill-posed problem, requiring the disentanglement of ego-motion, camera intrinsics, and multi-agent dynamics without explicit physical sensors. This difficulty is severely compounded during complex interactions featuring rapid relative motion and heavy occlusion. Furthermore, current video generation benchmarks primarily consist of general-domain or benign driving clips, lacking the targeted, safety-critical data required to expose these specific dynamic failures.

To address these methodological and data bottlenecks, we focus on vehicle-to-vehicle collisions as a representative task for evaluating multi-agent physical interactions. This setting provides a rigorous stress test for generative world models, as physical impacts induce abrupt, coupled state changes whose post-

contact evolution is tightly governed by classical conservation laws [7, 9]. By focusing on these multi-agent interactions, we establish a mathematically grounded environment for quantitative analysis, yielding consistent geometric priors and clear kinematic constraints [61]. Because these specific scenarios are inherently physics-dependent, verifying their structural integrity is a crucial prerequisite before such world models can be trusted as reliable simulators.

In this work, we present **CrashTwin**, a comprehensive evaluation framework designed to quantify the physical integrity of world model rollouts. To ensure both rigorous control and real-world relevance, we construct an ever large dataset that contains comprehensive text descriptions, poses indicating agent positions and states, and kinematic action labels, built from 25K synthetic sequences generated via a physically grounded pipeline [11] and 12K diverse in-the-wild traffic incident videos. To overcome the challenge of evaluating uncalibrated world-model rollouts, we develop a calibration-free reconstruction pipeline that leverages vision foundation models to reliably recover physical attributes under complex interactions. Building upon these recovered 3D trajectories, we introduce a standardized diagnostic protocol evaluating three core dimensions: spatio-temporal consistency, momentum and energy conservation, and world-dynamics integrity. Extensive benchmarking demonstrates that our metrics expose physical failure modes largely invisible to appearance-based evaluations, offering an orthogonal diagnostic signal for developing physically trustworthy world models.

Our contributions are threefold:

- We introduce **CrashTwin**, a large-scale benchmark coupling controllable synthetic data with unconstrained real-world collision videos, directly addressing the lack of safety-critical data in current evaluation paradigms.
- We develop a physics-grounded evaluation framework featuring a calibration-free reconstruction pipeline to recover metric-scale physical attributes from monocular videos, resolving the bottleneck of uncalibrated measurement to enable the precise evaluation of spatio-temporal consistency, momentum conservation, energy dissipation, and world dynamics.
- We systematically benchmark representative world models under this protocol, providing quantitative diagnoses of their physical failure modes and demonstrating that our metrics align significantly better with human preference judgments regarding physical realism.

2 Related Works

Video Generation & World Models. Video generation has advanced rapidly with diffusion and autoregressive transformers, enabling high-fidelity text-to-video across diverse domains [20, 41]. In autonomous driving, world models like GAIA-1, DriveDreamer-2, and Epona simulate controllable, multi-view traffic scenes with long-horizon rollouts, while egocentric simulators such as PlayerOne and GEM synthesize first-person videos from user motion or object controls

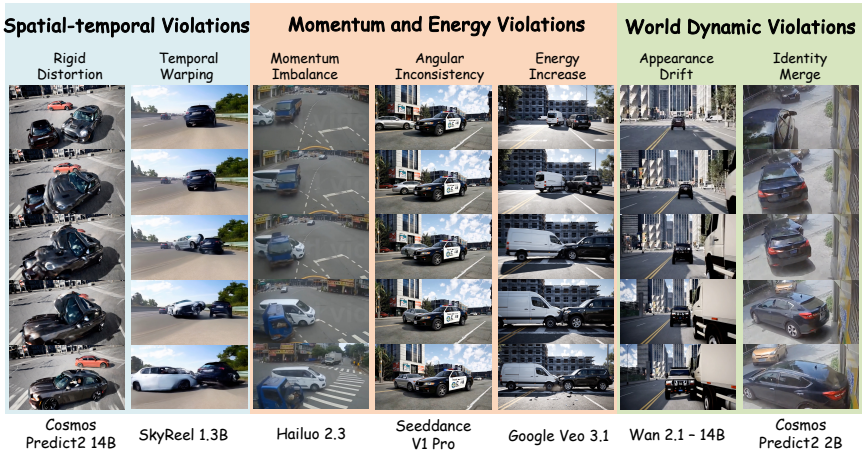

Fig. 2: Failure cases of existing world models revealed by our benchmark. Each region highlights a distinct type of physical breakdown, including violations of spatio-temporal consistency, departures from momentum and energy conservation around impact, and breakdowns of world-dynamics integrity such as unstable identities or appearances. These patterns illustrate the characteristic ways in which current methods deviate from physically consistent behavior.

[21, 23, 25, 43, 57, 58]. Yet appearance fidelity does not guarantee physical validity. Recent crash-oriented video generation models attempt to address safety-critical scenarios but fundamentally sidestep true physical simulation. For instance, DrivingGen [17] is text-driven and does not fix the first-frame, preventing controlled comparisons of physical outcomes. Similarly, AVD2 [30] focuses on extreme ego-centric self-collisions where vital pre- and post-impact kinematic information is effectively destroyed, making their physical fidelity inherently untestable. Furthermore, methods like OAVD [12] and Ctrl-Crash [16] strictly condition generation on comprehensive temporal bounding boxes. By providing the temporal object trajectories as input conditions, these models perform appearance completion within predefined constraints rather than actual physical simulation. Because these specialized models either rely on strong physical priors or generate untestable ego-collisions, they lack the capacity for generalized physical reasoning. Consequently, benchmarking them for physical consistency is inherently meaningless, as the physics are either dictated by the input conditions or completely unrecoverable. Instead, evaluating the physical characteristics of unconstrained world models remains a critical open challenge. Studies show that current models often ignore kinematics and violate momentum during collisions [5, 26, 34]. This has motivated hybrids that embed classical rigid-body simulators and benchmarks that test whether generated trajectories obey fundamental laws rather than only look plausible [5,34,37]. Our CrashTwin benchmark advances this direction by explicitly quantifying spatio temporal consistency, momentum and energy conservation, and instance level structural integrity during complex vehicle collisions.

Physical-grounded Evaluation. Most evaluation in world modeling is still perception-centric, prioritizing appearance fidelity, temporal coherence, and preference alignment over adherence to physical laws and dynamics. Distributional and perceptual scores such as FID, FVD, and embedding- or quality-based measures focus on visual realism instead of law adherence [22, 44], and comprehensive benchmarks like VBench and DEVIL mainly decompose visible quality, including identity consistency, motion smoothness, temporal stability, and dynamics coherence, rather than testing mechanics [25,31]. Recent "physics-aware" work follows three lines: (i) using LLMs or multimodal transformers as commonsense judges of plausibility [5,26,53]; (ii) heuristic proxies enforcing optical-flow or multi-frame depth and pose consistency [35, 51]; and (iii) scripted scenarios that flag violations or probe conservation in simplified collisions [5,26,37]. These approaches are largely indirect or qualitative, avoid directly measuring physical quantities, are prompt- or design-sensitive, and rarely localize errors in space and time, which hampers reproducibility and objective comparison [25,35,37,51]. The limitations are especially severe in collision scenes, where models still show ghosting through obstacles, momentum-inconsistent velocity changes, and invalid contact geometry even when appearance scores are high [5, 26]. This gap motivates physical-law-grounded verification with localized diagnosis. We introduce CrashTwin, which directly tests spatio-temporal consistency, momentum and energy conservation, and world-dynamics integrity in predicted collision events, providing quantitative localized checks that current metrics overlook and supporting progress toward deployable world models for planning and decision-making [18, 57].

3 Principles of the New Evaluation Protocol

The proposed CrashTwin is a physics grounded framework for analyzing and evaluating crash events. Section 3.1 first presents three dimensions of physical behavior and their associated evaluation metrics. Unlike prior evaluations that focus on image, patch, or pixel level accuracy, CrashTwin instead measures physical quantities and tests physical laws and contact constraints using measurable signals, localizing violations around first contact and reporting their magnitudes through physics-grounded metrics. Next, Section 3.2 introduces a carefully controlled CARLA based data generation pipeline supplemented with a real world crash corpus. Finally, Section 3.3 details the calibration free reconstruction pipeline that recovers the necessary physical attributes from uncalibrated videos to support these evaluations.

3.1 Physical-Grounded Evaluation

We study physics-grounded behavior across three complementary dimensions: spatio-temporal consistency, momentum and energy conservation, and world-dynamics integrity, as illustrated in Fig. 3. Together, these dimensions provide a comprehensive measure of how well generated collisions adhere to physical laws,

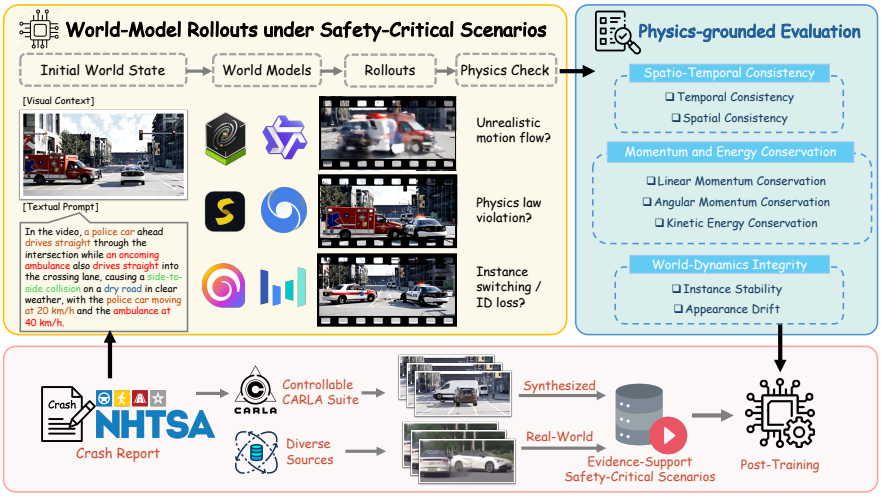

Fig. 3: Overview of the CrashTwin framework. We derive safety-critical collision from national pre-crash statistics and instantiate them in both a controllable CARLA suite and a diverse real-world corpus. Current world models often violate physical principles in these rollouts, motivating our physics-grounded evaluation framework that measures spatio-temporal consistency, momentum and energy conservation, and world-dynamics integrity. Guided by these metrics, post-training improves physical fidelity.

assessing whether motion remains continuous and bounded, impacts satisfy conservation principles, and interacting instances maintain stable, non-penetrating identities throughout the event.

Spatio-Temporal Consistency We aim to evaluate whether generated motion is smooth and plausible, through *temporal coherence*, ensuring continuous motion over time, and *spatial rigidity*, requiring foreground instances keep consistent geometry and scale without unrealistic expansion or compression.

Determining Temporal Consistency. We measure temporal coherence using a *flow warping error*, which quantifies how well pixels propagated by optical flow match their corresponding observations in the next frame:

$$E_{\text{warp}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left| I_t(p) - I_{t+1}(p + F_{t \to t+1}(p)) \right|, \tag{1}$$

where Ω denotes all pixels within the instance mask, I_t denotes the image at frame t, and $F_{t \to t+1}$ is the optical flow from frame t to $t{+}1$.

Evaluating Spatial Consistency. We evaluate spatial rigidity using the *divergence of the velocity field*, which measures local expansion and compression within the foreground region:

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

where $(u, v) = F_{t \to t+1}(p)$ are the flow components. Lower divergence indicates stronger preservation of rigid-body geometry. All computations are performed within instance masks [39] using modern optical-flow estimators [48].

Momentum and Energy Conservation The content generated by world models should preserve momentum and energy within the collision window; otherwise the impact fails to respect conservation laws. We therefore apply three checks: (1) *linear momentum conservation*, (2) *angular momentum conservation* and (3) *kinetic energy conservation*. Together, these indicators assess whether motion transitions around impact remain consistent with vehicle dynamics and collision physics, verifying that generated behaviors follow physically plausible laws under dominant inter-vehicle impulses and negligible external forces.

(1) Linear Momentum Conservation. For each vehicle i with mass m_i and pre-/post-impact velocities v_i^- and v_i^+, the total momentum before and after collision is $p^- = \sum_i m_i v_i^-$ and $p^+ = \sum_i m_i v_i^+$. Their difference $\Delta p = p^+ - p^-$ measures translational imbalance, and the normalized residual

$$J_p = \frac{\|\Delta p\|}{\sum_i m_i \|v_i^-\|} \tag{3}$$

quantifies momentum loss or gain relative to pre-impact motion. The evaluation is performed within a short collision window Δt, during which external forces (e.g., tire friction or aerodynamic drag) contribute negligible impulse compared with inter-vehicle contact. Under this quasi-closed condition, linear momentum can be treated as approximately conserved [40]. Small J_p values indicate physically consistent velocity transitions dominated by a single collision impulse, while larger values imply missing or unbalanced dynamics.

(2) Angular Momentum Conservation. We evaluate rotational consistency under the same impulse-window assumption by computing angular momentum about the contact point c. For each vehicle i with yaw moment of inertia $I_{z,i}$, yaw rate ω_i, position r_i, and velocity v_i, the total angular momentum is

$$H_c = \sum_{i=1}^{n} (I_{z,i}\omega_i + m_i(r_i - c) \times v_i)\,, \tag{4a}$$

$$J_H = \frac{\|H_c^+ - H_c^-\|}{\|H_c^-\| + \varepsilon}. \tag{4b}$$

Here $(r_i - c) \times v_i$ denotes the scalar 2D cross product capturing the rotational effect of translational motion around c. Low J_H values indicate that angular momentum transfer is physically consistent with rigid-body impact dynamics, while large values imply unbalanced or nonphysical torque generation. The small constant ε prevents division by zero when pre-impact angular momentum is negligible, ensuring numerical stability.

(3) Kinetic Energy Conservation. We verify that total kinetic energy does not increase during impact, consistent with the dissipative nature of real collisions. The total kinetic energy and its normalized residual are

$$E_k = \tfrac{1}{2} \sum_{i=1}^{n} m_i \|v_i\|^2, \tag{5a}$$

$$J_E = \max\left(0,\ \frac{E_k^+ - E_k^-}{E_k^-}\right). \tag{5b}$$

Here J_E serves as a penalty that activates only when kinetic energy increases; $J_E = 0$ for $E_k^+ \leq E_k^-$ and grows proportionally with any unphysical energy gain.

World-Dynamics Integrity We evaluate world-dynamics integrity to test whether each instance maintains a stable identity and consistent appearance through impact, ensuring coherent behavior at the instance level. Intuitively, a physically plausible collision should preserve the integrity of object identities and avoid sudden visual changes unexplained by the generated dynamics. We examine two complementary dimensions: *Instance Stability*, which measures the temporal persistence of tracked instances, and *Appearance Drift*, which evaluates the smoothness and temporal coherence of instance-level visual features over the trajectory.

Instance Stability. We evaluate instance stability by measuring how consistently an instance retains the same identity throughout the collision. We implement this using the Simpson index: let $c_{k,i}$ denote the number of frames in which instance i is assigned ID k, $N_i = \sum_k c_{k,i}$, and $f_{k,i} = c_{k,i}/N_i$. The score is

$$S_{\mathrm{ID}}^{(i)} = \sum_k f_{k,i}^2 \in [0, 1], \tag{6}$$

where 1 indicates a fully stable identity and lower values reflect fragmentation. Tracking associations are obtained using CenterTrack [62].

Appearance Drift. To capture appearance over time, we use SAM2 to obtain instance masks [39], and then assess instance-level visual coherence by tracking how the CLIP features of each vehicle evolve along its full visible trajectory. For each instance i, let $\{X_{i,t}\}$ denote the sequence of masked crops and let $\hat{z}_{i,t} = \phi(X_{i,t})/\|\phi(X_{i,t})\|_2$ be the L2-normalized CLIP embedding at frame t. Given the ordered embedding path $\{\hat{z}_{i,t_1}, \hat{z}_{i,t_2}, \ldots, \hat{z}_{i,t_T}\}$, we compute the frame-to-frame angular deviation using $\theta_{i,k} = \arccos(\langle \hat{z}_{i,t_k}, \hat{z}_{i,t_{k+1}} \rangle)$ for consecutive frames k. The appearance drift score for instance i is defined as

$$D_{\mathrm{ad}}^{(i)} = \frac{1}{T-1} \sum_{k=1}^{T-1} \theta_{i,k}, \tag{7}$$

with lower values indicating smoother and more temporally stable appearance evolution. The video-level D_{ad} is computed as a frame-count-weighted average over all tracked instances.

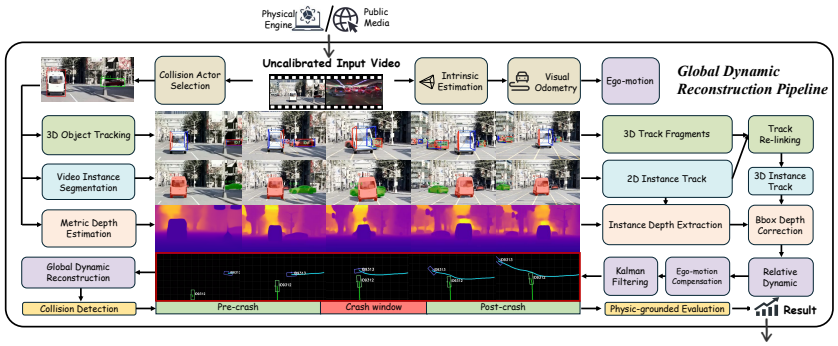

Fig. 4: Global dynamic reconstruction pipeline. The system reconstructs the 3D dynamics of collision participants from uncalibrated accident videos, where camera parameters, scene scale, and ego-motion are not directly available. We combine 3D tracking, video segmentation, metric depth estimation, and visual odometry to obtain coherent per-actor trajectories. Tracking fragments are re-linked to maintain identity continuity, metric depth restores real-world scale, and ego-motion compensation maps relative motion into a global coordinate system. The resulting global trajectories define the pre-crash, crash, and post-crash window and support reliable computation of physical quantities used in our physics-based evaluation.

3.2 CrashTwin Data Collection and Composition

To evaluate collision understanding of world models under both controllable synthetic settings and in-the-wild conditions, we construct CrashTwin by combining a CARLA-based crash generator with a curated real-world crash corpus.

Collision Data Collection Pipeline. Our collision data collection pipeline includes both simulated and real-world sources. In simulation, we reproduce seven high-frequency intersection collision types based on national pre-crash statistics [2]. Using CARLA, built on Unreal Engine 5, CrashTwin simulates scalable, high-fidelity scenes where vehicles approach from diverse directions and speeds. Each case is parameterized by initial position, velocity, and orientation, with stochastic perturbations that vary timing and contact geometry. The simulator enforces rigid-body dynamics with frictional contact and records full logs of vehicle states, 3D bounding boxes, segmentation, depth, and optical flow, enabling reproducible, physics-based evaluation. For real-world data, we collect diverse crash videos from public online sources and clean the raw data to obtain crash videos in varied environments such as intersections, highways, and rural roads. Each video is temporally segmented around the impact moment and annotated with detailed captions describing vehicle motion and scene context. Together, these simulated and real-world streams form a comprehensive and physically interpretable foundation for benchmarking collision understanding.

Data Composition and Statistics. Our final collection includes approximately 25K synthetic and 12K real-world crash sequences. The synthetic subset covers 7 representative intersection collision configurations derived from national

Table 1: Accident video dataset comparison. We summarize prior datasets by release year, scale (#clips and #frames), whether object bounding boxes are provided (Bboxes), whether tracklets are provided (Tracklet), whether text descriptions are provided (TT), and whether the data are real or synthetic (R/S).

Dataset	Year	#Clips	#Frames	Bboxes	Tracklet	TT	R/S
DAD [8]	2016	1,750	175K		✓		R
A3D [55]	2019	3,757	208K			✓	R
GTACrash [29]	2019	7,720	–			✓	S
VIENA^2 [4]	2019	15,000	2.25M			✓	S
CTA [56]	2020	1,935	–			✓	R
CCD [6]	2021	1,381	75K		✓	✓	R
TRA [32]	2022	560	–			✓	R
DADA-2000 [13]	2022	2,000	658K			✓	R
DoTA [54]	2022	5,586	732K	partial		✓	R
ROL [27]	2023	1,000	100K			✓	R
DeepAccident [46]	2023	–	57K			✓	S
CTAD [36]	2023	1,100	–			✓	S
MM-AU [12]	2023	11,217	2.19M	✓		✓	R
Ours	**2026**	**38,349**	**7.04M**	✓	✓	✓	**Both**

Bboxes object bounding boxes. **Tracklet** short-term bounding-box trajectories with identity linkage. **TT** text descriptions. **R/S** real or synthetic.

pre-crash typologies, spanning diverse approach angles, impact geometries, and environmental conditions such as weather and illumination. Each synthetic event is recorded at 30 Hz with full state logs, segmentation, depth, and optical flow, ensuring frame-level physical observability. The real-world subset contains 12K curated crash clips sourced from online videos, featuring varied environments including urban intersections, highways, and rural roads. Each clip is temporally segmented around the impact frame and captioned to describe vehicle motion, collision timing, and scene context. Together, these two complementary sources provide a balanced corpus that combines physically controlled synthetic data for quantitative evaluation with visually rich real-world data for assessing realism and generalization. We provide further details and a few data samples in the supplementary materials. A detailed comparison with prior accident video datasets is summarized in Tab. 1.

3.3 Global Dynamic Reconstruction Pipeline.

To consistently evaluate both generated crash content and web-scale real videos, which are typically uncalibrated, we build a calibration-free pipeline that reconstructs the 3D dynamics of the collision participants from raw video, as illustrated in Fig. 4. Conventional 2D/3D detection and tracking pipelines often fail under the large viewpoint changes, motion blur, and occlusions present in crash footage, making them unreliable for physics analysis. Our pipeline instead integrates multiple vision foundation models which remains robust under rapid motion and provides metric trajectories for subsequent physics-based evaluation.

Actor Selection. For each collision, we must first determine the pair of vehicles that will collide, using the initial frame of the sequence. Automatic identification

Table 2: 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. 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 [10]	0.0227	0.6103	0.9566	0.9628	0.9457	0.6660	0.3592
Wan 2.1-14B [45]	0.0179	0.6320	0.8235	0.8494	0.7864	0.6760	0.3117
Wan 2.2-5B [45]	0.0145	**0.5959**	0.8899	0.8975	0.8649	**0.7254**	**0.3109**
Cosmos-Predict2-2B [3]	0.0240	0.6748	0.8890	0.8954	0.8590	0.6129	0.3462
Cosmos-Predict2-14B [3]	**0.0117**	0.7180	**0.6828**	**0.7629**	**0.6047**	0.6737	0.3327
Proprietary Models							
Google Veo 3.1 [15]	**0.0097**	0.6202	0.7743	0.8011	0.7460	**0.7232**	0.3075
Hailuo 2.3 [19]	0.0151	0.6143	**0.7664**	**0.7812**	0.7285	0.6203	**0.2968**
Seedance V1 Pro [14]	0.0166	**0.6130**	0.7725	0.7837	**0.7209**	0.6007	0.3002

of collision participants is not yet reliable for current vision-language models, especially under occlusion and rapid motion. We therefore use simulator ground-truth identities in synthetic clips, and annotate collision actors manually in real crash videos. These identities are released with the test set to ensure consistent actor selection across all evaluation methods.

Tracking Fragments. We obtain initial 3D track fragments for the selected actors using a 3D object tracker. These fragments provide coarse estimates of the underlying dynamics but often break under strong acceleration, occlusion, and rapid directional changes that are typical of collision events.

Relinking and Correction. To obtain coherent trajectories, fragmented 3D segments are relinked using video instance segmentation, which provides more reliable identity continuity under rapid motion. Metric depth is then applied to refine geometry and restore real-world scale, yielding stable relative 3D dynamics in the camera coordinate frame.

Global Dynamics. Relative trajectories are transformed into a global coordinate system by composing them with ego motion estimated through visual odometry. A Kalman filter is then applied to smooth positions and orientations, producing stable global dynamics for both actors.

Collision Detection. Given the reliable global trajectories, we monitor the distance between the two actors and mark the first moment it falls within a safety threshold, which defines the start of the collision window. A short pre-crash, crash, and post-crash segment is then extracted and passed to our physics-based evaluation. See more details in our supplementary materials.

Table 3: Reconstruction accuracy on simulated crashes. Since no publicly accessible models exist for uncalibrated collision scenarios, we incrementally add components to a basic tracking baseline. We report mean absolute trajectory error (ATE) in 3D for both scene-level and instance-level dynamics.

Model Configuration	Scene-level ATE (m) ↓		Instance-level ATE (m) ↓	
	$SE(3)$	$Sim(3)$	$SE(3)$	$Sim(3)$
Basic 3D Tracking [62]	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**

Benchmark Protocol. For all evaluations, we estimate camera intrinsics with MapAnything [28], obtain ego motion using DROID-SLAM [42] combined with metric-scale correction from Metric3D V2 [24], generate initial 3D object tracks with CenterTrack [62], and use SAM2 [39] to produce instance masks for relinking and refinement. These components together provide the 3D trajectories required by the physics-based metrics in Section 3.1. Collision timing is defined as the first frame in which the inter-vehicle distance falls below a predefined safety threshold. For clips where no contact occurs, collision-based momentum and energy metrics can become numerically unstable, so we assign a fixed penalty value in all non-collision cases.

4 Experiments

4.1 Datasets

To study how current world models perform on physically realistic collisions and how to reduce the gap between generated and real crash videos, we benchmark them on the proposed CrashTwin test split under our evaluation protocol. As summarized in Tab. 1, prior accident video datasets are restricted in scale and confined to a single data source, lacking both the volume and cross-domain diversity necessary for unified evaluation and model adaptation. To bridge this gap, CrashTwin contains 38K crash events, including both 25.6K synthetic and 12.6K real-world sequences, and is divided into a training split for model development and a held-out test split for evaluation.

We construct a comprehensive evaluation set, *CrashTwin-Eval*, comprising 300 synthetic and 44 real-world videos. Each clip provides two annotated collision actors in the first frame and a textual description of the impact configuration. To accommodate closed-source models with high inference cost or limited access quotas, we further define a smaller evaluation partition, *CrashTwin-Eval-Mini*, which contains randomly sampled 100 synthetic and 16 real-world videos. Both subsets are sampled from the CrashTwin test split and are used for quantitative benchmarking, while the remaining data serve as training samples. We report the physics-grounded metrics defined in Section 3.1.

Table 4: Effect of post training. Metrics are grouped into three physical families. Lower is better unless marked with ↑.

	Spatio-temporal Consistency		Momentum & Energy Conservation			World-dynamics Integrity	
Model Variant	Warp Error	Flow Div.	Momentum Residual	Angular Residual	Energy Gain	Instance Stability	App. Drift
	$E_{\text{warp}} \downarrow$	$E_{\text{div}} \downarrow$	$J_p \downarrow$	$J_H \downarrow$	$J_E \downarrow$	$S_{\text{ID}} \uparrow$	$D_{\text{ad}} \downarrow$
Base (w/o Post-training)	0.0240	0.6748	0.8890	0.8954	0.8590	0.6129	0.3462
Base (w/ Post-training)	0.0085	0.6279	0.6479	0.7296	0.5534	0.7746	0.2953
Ground Truth Reference	0.0075	0.6549	0.3089	0.4620	0.2502	0.8010	0.2819

4.2 Leaderboard Comparison.

We benchmark representative world models, including open-source SkyReel [10], Wan 2.1 [45], and Cosmos-Predict2 [3], on the CrashTwin-Eval subset, as well as proprietary models Veo 3.1 [15], Hailuo [19], and Seedance [14], on the CrashTwin-Eval-Mini subset. Tab. 2 shows that momentum and energy residuals remain noticeably high across all methods, indicating that current models still struggle to exchange momentum and energy in a physically consistent way, even when some show strong spatio-temporal and world-dynamics consistency.

As illustrated by the Hailuo 2.3 example in Fig. 2, the scene remains temporally stable with consistent identities, yet the blue truck suddenly aligns with the sedan after impact, causing its lateral momentum to vanish. In the Seedance V1 Pro case, the police vehicle preserves appearance and identity throughout the sequence, but undergoes an implausible yaw reversal after contact, revealing a clear angular-momentum inconsistency. The Veo 3.1 example further shows two vehicles nearly coming to rest at impact before accelerating again, producing an unphysical increase in kinetic energy despite otherwise coherent frames. This gap indicates that maintaining spatio-temporal smoothness and stable object identities is easier for current world models than producing physically correct momentum and energy exchanges. It highlights the necessity of physics-grounded evaluation to reveal failures that perceptual metrics cannot capture.

Effect of Post-Training. We fine-tune pretrained models on our CrashTwin training split using physical consistency objectives that regularize flow-based temporal coherence, spatial rigidity, and short-window momentum and energy residuals. Post-training brings clear gains across all physical dimensions. As seen in Tab. 4, most spatio-temporal and world-dynamics metrics move closer to the ground-truth reference, and momentum consistency also improves. This indicates that post-training improves spatio-temporal consistency and world-dynamics integrity more than momentum and energy conservation, reinforcing that spatio-temporal smoothness and stable identities are easier to achieve than physically correct momentum and energy exchange. A representative qualitative example in Fig. 5 further shows that post-training restores lateral momentum transfer and stabilizes instance appearance.

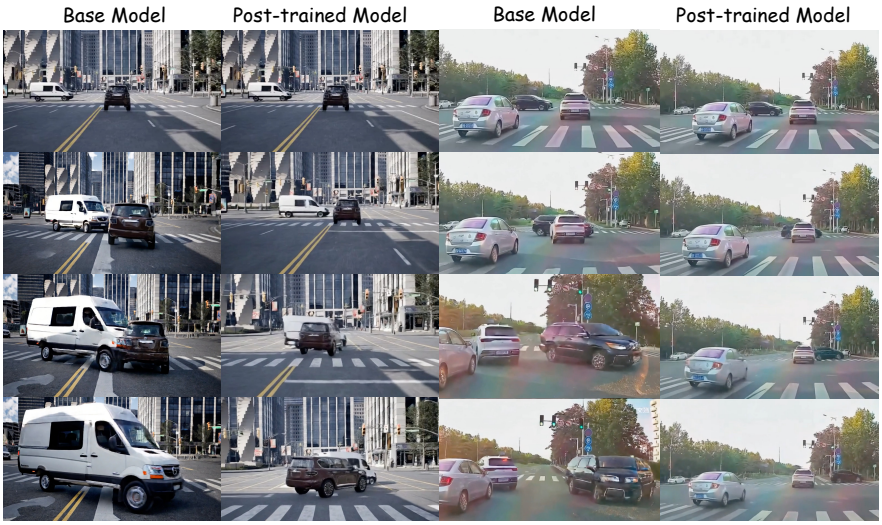

Fig. 5: Qualitative comparison before and after physics post-training. *Left:* The base model lets the van pass through the other vehicle without transferring lateral momentum, while the post-trained model produces a realistic rightward push. *Right:* In a mild real crash, the base model shows appearance distortion and deformation, whereas the post-trained model preserves stable geometry and identity.

4.3 Effectiveness of the Proposed Reconstruction

As no publicly accessible model can recover metric-scale dynamics from uncalibrated severe-collision videos, we build a basic 3D tracking baseline [62] and incrementally add our components. Tab. 3 shows large drift under heavy occlusion for the baseline, which is steadily reduced by instance relinking, Kalman filtering, and metric depth correction. We report $SE(3)$ and $Sim(3)$ errors to reflect accuracy with fixed scale and up to a global similarity transform. The final instance-level error reaches a sub-meter regime, supporting that our pipeline recovers stable kinematics for physics-grounded evaluation. Additional qualitative results are included in the supplementary materials.

5 Conclusions

This paper has introduced a physics-based benchmark for world models in crash scenarios, enabling a fine-grained examination of how generative rollouts behave under real physical constraints. We have demonstrated that our benchmark reveals systematic weaknesses in current models, particularly in momentum exchange, energy transfer, and post-impact dynamics, even when their visual predictions appear coherent. This is achieved by decomposing crash dynamics into interpretable physical metrics and by providing a pipeline that highlights where and how physical violations arise around high-force events. Our study shows

that post-training reduces violations more in spatio-temporal consistency and world-dynamics integrity than in momentum and energy exchange, motivating stronger physics priors and more reliable dynamic modeling. Altogether, CrashTwin serves as a diagnostic tool that combines data curation, physical attribute estimation, and physics-grounded metrics to bring physical fidelity to the forefront of world-model evaluation, paving the way for future models that are not only visually plausible but physically trustworthy.

Acknowledgements. This work was supported in part by GPU hardware and cloud credits provided to the Texas A&M University PHAI Lab and TACO Lab through the NVIDIA Academic Grant Program. We also acknowledge computational resources provided by the Texas A&M VISION GPU Cluster. Z. Fan was supported in part by research gifts from Meta and the US AI Glasses Impact Grants program. J. Hong is supported by NAIRR250526. Z. Wang is supported by DARPA ANSR (RTX CW2231110), DARPA TIAMAT (HR0011-24-9-0431), ARL StAmant (W911NF-23-S-0001), as well as the NSF AI Institute for Foundations of Machine Learning (IFML).

References

1. Waymo safety report. Tech. rep., Waymo LLC (Dec 2021), `https://storage.googleapis.com/waymo-uploads/files/documents/safety/2021-12-waymo-safety-report.pdf`, accessed 2026-03-04
2. Administration, N.H.T.S., et al.: Pre-crash scenario typology for crash avoidance research. DOT HS **810**(767), 4 (2007)
3. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
4. Aliakbarian, M.S., Saleh, F.S., Salzmann, M., Fernando, B., Petersson, L., Andersson, L.: Viena: A driving anticipation dataset. In: Asian Conference on Computer Vision. pp. 449–466. Springer (2018)
5. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024)
6. Bao, W., Yu, Q., Kong, Y.: Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2682–2690 (2020)
7. Brach, R.M., Brach, R.M.: A review of impact models for vehicle collision (1987)
8. Chan, F.H., Chen, Y.T., Xiang, Y., Sun, M.: Anticipating accidents in dashcam videos. In: Asian conference on computer vision. pp. 136–153. Springer (2016)
9. Chatterjee, A.: Rigid body collisions: some general considerations, new collision laws, and some experimental data. Cornell University (1997)
10. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
11. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)

12. Fang, J., Li, L.l., Zhou, J., Xiao, J., Yu, H., Lv, C., Xue, J., Chua, T.S.: Abductive ego-view accident video understanding for safe driving perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22030–22040 (2024)
13. Fang, J., Yan, D., Qiao, J., Xue, J., Yu, H.: Dada: Driver attention prediction in driving accident scenarios. IEEE transactions on intelligent transportation systems **23**(6), 4959–4971 (2021)
14. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
15. Google DeepMind: Veo 3.1. Technical report, Google DeepMind (2026), `https://blog.google/innovation-and-ai/technology/ai/veo-3-1-ingredients-to-video/`, released: January 13, 2026. Accessed: March 5, 2026
16. Gosselin, A., Luo, G.Y., Lara, L., Golemo, F., Nowrouzezahrai, D., Paull, L., Jolicoeur-Martineau, A., Pal, C.: Ctrl-crash: Controllable diffusion for realistic car crashes. arXiv preprint arXiv:2506.00227 (2025)
17. Guo, Z., Zhou, Y., Gou, C.: Drivinggen: Efficient safety-critical driving video generation with latent diffusion models. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
18. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels. In: International conference on machine learning. pp. 2555–2565. PMLR (2019)
19. HailuoAI: Hailuo. `https://hailuoai.video/` (2024), accessed: 2025-02-24
20. Han, H., Li, S., Chen, J., Yuan, Y., Wu, Y., Deng, Y., Leong, C.T., Du, H., Fu, J., Li, Y., et al.: Video-bench: Human-aligned video generation benchmark. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18858–18868 (2025)
21. Hassan, M., Stapf, S., Rahimi, A., Rezende, P., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., 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. pp. 22404–22415 (2025)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
23. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
24. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: 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)
25. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
26. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective, 2024. URL https://arxiv. org/abs/2411.02385 **2**, 36

27. Karim, M.M., Yin, Z., Qin, R.: An attention-guided multistream feature fusion network for early localization of risky traffic agents in driving videos. IEEE Transactions on Intelligent Vehicles **9**(1), 1792–1803 (2023)
28. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
29. Kim, H., Lee, K., Hwang, G., Suh, C.: Crash to not crash: Learn to identify dangerous vehicles using a simulator. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 978–985 (2019)
30. Li, C., Zhou, K., Liu, T., Wang, Y., Zhuang, M., Gao, H.a., Jin, B., Zhao, H.: Avd2: Accident video diffusion for accident video description. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 13289–13296. IEEE (2025)
31. Liao, M., Ye, Q., Zuo, W., Wan, F., Wang, T., Zhao, Y., Wang, J., Zhang, X., et al.: Evaluation of text-to-video generation models: A dynamics perspective. Advances in Neural Information Processing Systems **37**, 109790–109816 (2024)
32. Liu, C., Li, Z., Chang, F., Li, S., Xie, J.: Temporal shift and spatial attention-based two-stream network for traffic risk assessment. IEEE Transactions on Intelligent Transportation Systems **23**(8), 12518–12530 (2021)
33. Liu, M., Zhang, W.: Is your video language model a reliable judge? arXiv preprint arXiv:2503.05977 (2025)
34. Liu, S., Ren, Z., Gupta, S., Wang, S.: Physgen: Rigid-body physics-grounded image-to-video generation. In: European Conference on Computer Vision. pp. 360–378. Springer (2024)
35. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22139–22149 (2024)
36. Luo, H., Wang, F.: A simulation-based framework for urban traffic accident detection. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
37. Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. arXiv preprint arXiv:2410.05363 (2024)
38. Puyin, L., Xiang, T., Mao, E., Wei, S., Chen, X., Masood, A., Fei-Fei, L., Adeli, E.: Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models. arXiv preprint arXiv:2512.19526 (2025)
39. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
40. Stronge, W.J.: Impact mechanics. Cambridge university press (2018)
41. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8406–8416 (2025)
42. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in neural information processing systems **34**, 16558–16569 (2021)
43. Tu, Y., Luo, H., Chen, X., Bai, X., Wang, F., Zhao, H.: Playerone: Egocentric world simulator. arXiv preprint arXiv:2506.09995 (2025)

44. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
45. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
46. Wang, T., Kim, S., Wenxuan, J., Xie, E., Ge, C., Chen, J., Li, Z., Luo, P.: Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5599–5606 (2024)
47. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
48. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: European Conference on Computer Vision. pp. 36–54. Springer (2024)
49. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: 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. pp. 14749–14759 (2024)
50. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6902–6912 (2024)
51. Wu, J.Z., Fang, G., Wu, H., Wang, X., Ge, Y., Cun, X., Zhang, D.J., Liu, J.W., Gu, Y., Zhao, R., et al.: Towards a better metric for text-to-video generation. arXiv preprint arXiv:2401.07781 (2024)
52. Xu, R., Lin, H., Jeon, W., Feng, H., Zou, Y., Sun, L., Gorman, J., Tolstaya, E., Tang, S., White, B., et al.: Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. arXiv preprint arXiv:2510.26125 (2025)
53. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18826–18836 (2025)
54. Yao, Y., Wang, X., Xu, M., Pu, Z., Wang, Y., Atkins, E., Crandall, D.J.: Dota: Unsupervised detection of traffic anomaly in driving videos. IEEE transactions on pattern analysis and machine intelligence **45**(1), 444–459 (2022)
55. Yao, Y., Xu, M., Wang, Y., Crandall, D.J., Atkins, E.M.: Unsupervised traffic accident detection in first-person videos. In: 2019 IEEE/RSJ International conference on intelligent robots and systems (IROS). pp. 273–280. IEEE (2019)
56. You, T., Han, B.: Traffic accident benchmark for causality recognition. In: European Conference on Computer Vision. pp. 540–556. Springer (2020)
57. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., et al.: Epona: Autoregressive diffusion world model for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27220–27230 (2025)
58. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10412–10420 (2025)
59. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)

60. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)
61. Zhou, J., Peng, H., Lu, J.: Collision model for vehicle motion prediction after light impacts. Vehicle System Dynamics **46**(S1), 3–15 (2008)
62. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European conference on computer vision. pp. 474–490. Springer (2020)