VEHDYN: A DRIVING WORLD MODEL BENCHMARK FOR VEHICLE DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models are emerging as data engines, action planners, and generative simulators for autonomous driving, but existing benchmarks primarily assess visual fidelity and coarse physical plausibility, providing limited evidence on whether generated driving futures obey realistic vehicle kinematics and dynamics. This limitation is further compounded by the lack of datasets in which vehicle, road, maneuver, and speed conditions are independently controlled, and ground-truth vehicle states are recorded in synchrony with videos. We introduce VehDyn, a driving world model benchmark for vehicle dynamics. VehDyn is built on a CARLA-CarSim co-simulation platform where photorealistic rendering is coupled with a validated multi-body dynamics model, and it contains 10,080 configurations from a full factorial design over five vehicle types, four tire-road friction coefficients, three maneuvers, four target speeds, 14 scenes, and three illuminations, each paired with synchronized position, velocity, and attitude sequences. Built on this dataset, VehDyn introduces a hierarchical evaluation framework that measures trajectory alignment, kinematic consistency, and dynamic consistency, and benchmarks 12 state-of-the-art video world models. We further assess the video quality using two established protocols and correlate it with the VehDyn score. Trajectory-level metrics are nearly saturated, with ten of twelve models within 20% of ground truth, while no model reaches 92% of ground truth on dynamic consistency, and visual-quality metrics are only weakly correlated with vehicle-dynamics fidelity. DrivingWorld achieves the highest VehDyn score, followed by Cosmos 3 Nano and LTX-Video 2.5, and the VehDyn score agrees closely with human judgment. VehDyn provides a systematic foundation for developing driving world models that are physically consistent and visually realistic.

1 INTRODUCTION

A video world model is a generative simulator that predicts how an environment will evolve under different action conditions (Ha & Schmidhuber, 2018), and supports downstream tasks (Zhao et al., 2026). Inspired by this vision, there has been great progress in driving world models: given initial observations and optional conditions such as text prompts, control sequences, or target trajectories, driving world models predict the future evolution of the ego vehicle and its surroundings (Gao et al., 2024). The value of a generated future depends on whether the ego vehicle moves as a real vehicle would under the same commands and road conditions. That motion is governed by tire-road adhesion, inertia, and suspension compliance (Wang et al., 2026b), which bound the achievable acceleration, determine the stopping distance, and couple longitudinal and lateral inputs to the pitch and roll of the body. A generated lane change may follow a plausible planar path while exhibiting a lateral acceleration beyond the available adhesion, and an emergency-braking sequence may appear smooth while producing an incorrect stopping distance, pitch response, or deceleration profile. Such failures are visually unremarkable but physically invalid, and any planner trained or evaluated on such futures inherits the error. Prior studies (Zhang et al., 2025a; Fan et al., 2026) have found visual realism to be largely uncorrelated with physical understanding. Therefore, measuring vehicle-dynamics fidelity in driving video generation represents a prerequisite for deploying driving world models as reliable simulators, especially in safety-critical scenarios.

Existing evaluation of driving world models has progressively moved beyond conventional video quality metrics (Wang et al., 2026a), but vehicle-dynamics fidelity remains largely unmeasured.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of existing driving world model benchmarks and autonomous driving datasets. ✓: supported; △: partial; ✗: not supported.

(a) Comparison of evaluation dimensions in driving world model benchmarks.

Benchmark	Venue & Year	Data	World Model Types		Vehicle Motion Evaluation			Human
			General	Driving	Trajectory	Kinematics	Dynamics	
ACT-Bench	ICLRW '25	nuScenes	✗	✓	✓	✗	✗	✗
WorldSimBench	ICML '25	CARLA	✓	✗	△	✗	✗	✓
DrivingGen	ICLR '26	Real-World	✓	✓	✓	△	✗	✓
WorldLens	CVPR '26	nuScenes	✗	✓	✓	△	△	✓
What-If World	arXiv '26	nuScenes	✓	✗	✗	✗	△	✓
VehDyn (Ours)	–	CARLA-CarSim	✓	✓	✓	✓	✓	✓

(b) Comparison of data controllability in autonomous driving datasets.

Dataset	Venue & Year	Visual Ctrl.	Controlled Dynamics Factors				Dynamics GT	Paired Analysis
			Vehicle	Speed	Maneuver	Friction		
Real-World Datasets								
nuScenes	CVPR '20	△	✗	△	✗	✗	△	✗
nuPlan	CVPRW '21	△	✗	△	△	✗	△	✗
NAVSIM	NeurIPS '24	△	✗	△	△	✗	△	✗
DrivingDojo	NeurIPS '24	△	✗	△	△	✗	△	✗
StyleDrive	AAAI '26	△	✗	△	✗	✗	△	✗
DrivingGen	ICLR '26	△	✗	△	△	✗	△	✗
CARLA-Based Datasets								
CARLA (Leaderboard)	CoRL '17	✓	✗	△	✓	✗	△	✗
Bench2Drive	NeurIPS '24	✓	✗	△	✓	✗	△	✗
Bench2Drive-Speed	arXiv '26	✓	✗	✓	✓	✗	△	△
Bench2Drive-Robust	arXiv '26	✓	✗	△	✓	✗	△	△
Fail2Drive	arXiv '26	✓	✗	△	✓	✗	△	△
VehDyn (Ours)	–	✓	✓	✓	✓	✓	✓	✓

Early work relied primarily on Fréchet inception distance (FID), Fréchet video distance (FVD), and downstream perception or planning performance (Wu et al., 2026; Ying et al., 2026; Shang et al., 2026). ACT-Bench (Arai et al., 2024) introduced action fidelity by recovering the ego trajectory from generated video with monocular visual odometry and comparing it with the commanded trajectory through action-match rate, average displacement error (ADE), and final displacement error (FDE). DrivingGen (Zhou et al., 2026) broadened the scope to distribution realism, driving-specific image quality, temporal consistency, and a trajectory-consistency score that rewards smooth velocity and acceleration. WorldLens (Liang et al., 2026) added reconstruction quality, action following, downstream utility, and human-preference assessment, and What-If World (Cai et al., 2026) tested whether a model's output changes in the direction physics predicts when a single variable in the prompt is altered. Consequently, no prior benchmark compares generated motion with synchronized ground-truth vehicle states in both kinematic and dynamic aspects.

Existing benchmarks have introduced different datasets covering diverse and challenging scenarios. Real-world datasets such as nuScenes (Caesar et al., 2020), nuPlan (Caesar et al., 2021), NAVSIM (Dauner et al., 2024), DrivingDojo (Wang et al., 2024c), and StyleDrive (Hao et al., 2026) provided rich traffic diversity and useful pose or trajectory annotations, but their physical conditions are observational rather than experimentally controlled. Simulation-based datasets (Dosovitskiy et al., 2017; Li et al., 2022) offered greater controllability, but they were primarily designed for closed-loop policy evaluation and rely on a game-engine vehicle model that is not intended as a dynamics reference. High-fidelity simulators such as CarSim model tire forces, suspension kinematics, and load transfer with validated multi-body formulations and are the standard tool in vehicle dynamics analysis (Zeng et al., 2023; 2026), but they have not been coupled with photorealistic rendering to construct evaluation data for video world models. In all, current resources (Table 1) offered either rich visual diversity or controlled driving tasks, but not a physics-grounded testbed in which generated videos can be compared with synchronized vehicle-dynamics ground truth under systematically varied conditions. A detailed literature review is provided in Appendix A.1.

To address this gap, we propose **VehDyn** (Figure 1), a comprehensive driving world model benchmark for evaluating vehicle dynamics with diverse data distributions and novel evaluation metrics.

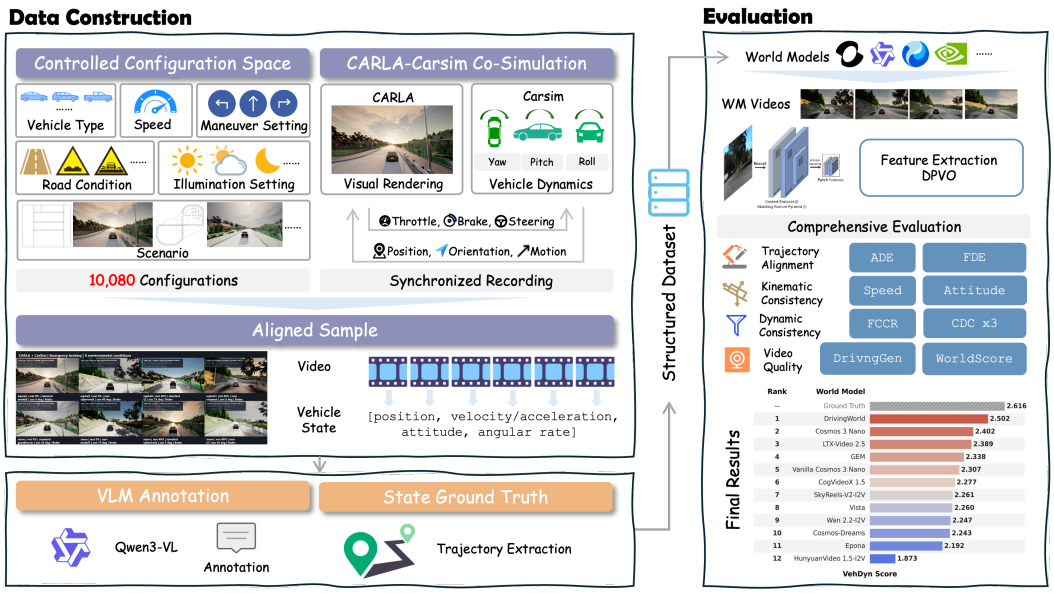


Figure 1: Overview of the proposed VehDyn benchmark. Left: data construction. A full factorial design over vehicle type, target speed, maneuver, road condition, scene, and illumination is instantiated on a CARLA-CarSim co-simulation platform. Right: evaluation. Vehicle states are recovered from the generated videos by DPVO and compared with ground truth at three levels, trajectory alignment, kinematic consistency, and dynamic consistency, and aggregated into the VehDyn score.

VehDyn is built on a CARLA-CarSim co-simulation platform in which CARLA provides visual rendering, while CarSim computes the vehicle response with a high-fidelity multi-body model, with both simulators synchronized on a unified clock. The dataset comprises 10,080 configurations generated from a full factorial design over five vehicle types, four tire-road friction coefficients, three maneuvers, four target speeds, 14 scenes, and three sun angles, each paired with ground-truth position, velocity, and attitude sequences. Built on this dataset, **VehDyn** defines a hierarchical evaluation protocol comprising trajectory alignment, kinematic consistency, and dynamic consistency. We benchmark 12 video world models, additionally score every generated video under the DrivingGen (Zhou et al., 2026) and WorldScore (Duan et al., 2025) protocols, and verify the agreement between the VehDyn score and human judgment. The main contributions are summarized as follows:

- **A physics-grounded dataset**, which includes 10,080 CARLA-CarSim co-simulation configurations generated from a full factorial design over vehicle type, maneuver, speed, friction coefficient, scene, and sun angle, paired with synchronized ground-truth vehicle states.
- **A vehicle-dynamics-aware evaluation protocol**, which extends beyond visual plausibility and planar trajectory alignment to six-degree-of-freedom (6-DoF) kinematic consistency, and condition-aware dynamic consistency measured against ground-truth responses.
- **An extensive benchmark**, which evaluates 12 representative video world models, characterizing the gap between visual fidelity and vehicle-dynamics consistency, and identifies failure modes that visual metrics do not expose.

2 VehDyn Dataset

2.1 Data Construction

VehDyn is constructed in three stages (Figure 1). We first build a CARLA-CarSim co-simulation platform in which visual rendering and vehicle dynamics are computed by separate simulators on a shared clock. We then define a controllable scenario space over six factors and instantiate every combination in a full factorial design. For each instance, we synchronously record a video clip, the complete vehicle state sequence, and the scenario configuration, and we finally annotate each clip

with a motion-centric caption. Each configuration contains ten independently generated episodes. The resulting dataset contains 100,800 samples.

Co-simulation platform. CARLA renders road geometry, scene assets, traffic participants, and lighting, while CarSim computes the vehicle response with a validated multi-body model that includes tire force generation, suspension kinematics, and longitudinal and lateral load transfer. At each simulation step, the reference-path controller in CARLA issues throttle, brake, and steering commands; CarSim integrates the vehicle dynamics under the configured vehicle and road parameters and returns the updated position, orientation, and motion states; CARLA then places the vehicle at the returned pose and renders the next frame. CarSim integrates at 2,000 Hz and CARLA renders at 30 Hz in synchronous fixed-step mode, with the CarSim state sampled at every rendered frame, so that each frame corresponds to an exactly known dynamic state. Through this closed-loop interaction, visual observations and high-fidelity vehicle motion states are collected synchronously.

Scenario space. The scenario space comprises six factors (Table 3 in Appendix A.2). Four dynamic factors define the physical conditions to which a world model must respond (vehicle type, target speed, driving maneuver, and tire-road friction coefficient), and the two visual factors provide replicates under which the dynamic response should be invariant (urban scene and sun angle). Vehicle appearance, scene, and sun angle are configured at CARLA initialization, the target speed is imposed by the path-following controller, and driving maneuvers are realized as predefined reference paths or control sequences. Vehicle mass, moment of inertia, and wheelbase are specified in CarSim. In addition, road adhesion conditions are kept semantically consistent: CARLA visually represents the road conditions through road material and surface wetness, and CarSim applies the corresponding tire-road friction coefficient. Route details are given in Figure 6 in Appendix A.2.

Data acquisition. Visual observations are recorded at 30 Hz for 5 s at 1920×1080 resolution from an RGB camera rigidly mounted at the vehicle reference point and moving with the body. This third-person viewpoint is chosen in preference to a front-facing dashcam view for two reasons. First, the vehicle body is visible in the frame, so pitch, yaw, and roll are directly observable in the generated video. Second, because the camera is rigidly attached to the body, its 6-DoF pose differs from the vehicle pose only by a fixed transform, which allows vehicle attitude to be recovered from camera-pose estimates. Vehicle states are recorded from CarSim at every rendered frame and include position, orientation, velocity and rate, control inputs, and progress along the route. All modalities are timestamped on the shared simulation clock.

Scene annotation. Each clip is captioned with Qwen3-VL (Bai et al., 2025) under a prompt that foregrounds visually observable motion rather than static appearance or abstract physical laws. Because the conditioning frame already conveys scene layout, lighting, and vehicle identity, the caption is allocated to the temporal dynamics that a generative model must synthesize: the maneuver, its onset, direction, and extent, the resulting change in vehicle attitude, and the camera's rigid attachment to the vehicle. A manual review pass by three annotators discards captions containing hallucinated or physically ambiguous content. The full prompt is given in Appendix A.2.

2.2 Data Analysis

Dataset composition. Each of the 100,800 samples consists of a 5 s, 150-frame RGB clip, a synchronized ground-truth state sequence, the scenario configuration, and a caption. The dataset totals 140 h of video and 2 TB. Because the design is fully crossed, every level of every factor occurs with every level of the others, and any two samples that differ in a single factor form a matched pair. For each target-speed sweep, friction contrast, or vehicle contrast, the ground-truth change in dynamic response is available directly from the recorded states. Recorded variables, data sources, and acquisition settings are provided in Table 5 in Appendix A.2.

Dynamic characteristics. To verify that the factorial design induces physically meaningful variation, we analyze the ground-truth states across the dataset. Emergency-braking samples characterize longitudinal motion and pitch responses, while single-lane-changing samples characterize lateral motion, yaw, and roll responses. Figure 2 shows the responses across target speeds, vehicle models, and the four road conditions. In the emergency-braking scenario, braking distance increases

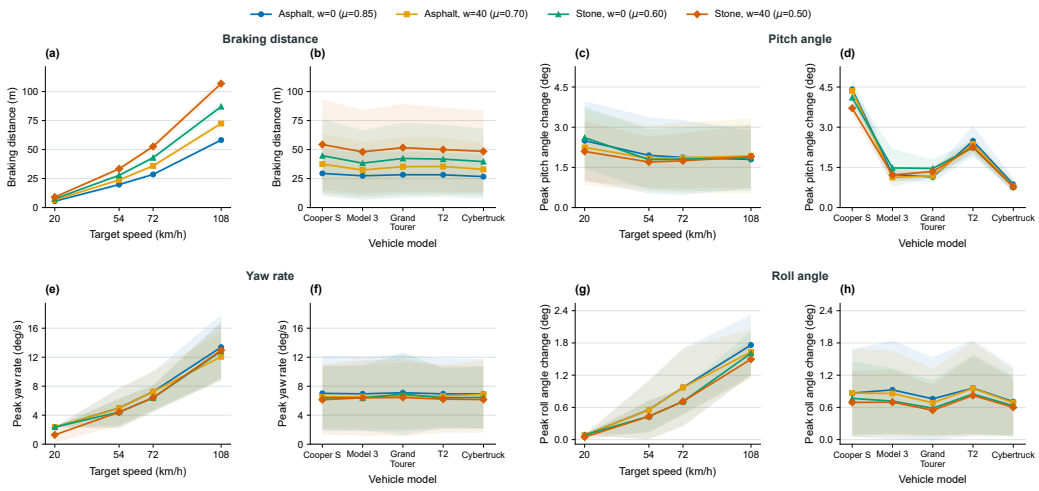


Figure 2: Dynamic characteristics of the VehDyn dataset under different target speeds, road conditions and vehicle types. (a)–(d), emergency-braking condition; (e)–(h), single-lane-changing condition. Curves represent the mean values, and shaded bands denote ± 1 standard deviation across scenes and illumination settings.

nonlinearly with target speed, more markedly under lower-friction conditions, and increases as friction decreases for every vehicle model, while differences among models at fixed conditions remain small. Peak pitch-angle changes, by contrast, vary more across vehicle models than across target speeds, potentially reflecting differences in mass distribution and suspension characteristics. In the single-lane-changing scenario, yaw-angle changes are relatively similar across conditions, whereas yaw rates are more sensitive to speed, making them more suitable for characterizing turning dynamics. Both the peak yaw rate and the peak roll-angle changes increase with target speed. Yaw rate is nearly identical across vehicle models, while roll-angle changes exhibit more pronounced differences among models. The full dynamic characteristics analysis is provided in Appendix A.2.

3 VehDyn Evaluation

VehDyn evaluates a generated video by recovering the ego-vehicle state sequence from the video and comparing it with the synchronized ground-truth states at three levels (Figure 1): (1) trajectory alignment, which measures planar position; (2) kinematic consistency, which measures translational and rotational motion states; and (3) dynamic consistency, which measures whether the generated motion is physically feasible and whether it responds to changed physical conditions as the real vehicle does. Per-metric implementation details are provided in Appendix A.3.

3.1 Feature Extraction

We employ deep patch visual odometry (DPVO) (Teed et al., 2023) to estimate a sequence of 6-DoF relative camera poses $\{\hat{\mathbf{T}}_t\}_{t=1}^{T}$ from the video frames and the known camera intrinsics. DPVO tracks sparse image patches through recurrent correspondence updates and jointly refines camera poses by differentiable bundle adjustment. Because the camera is rigidly attached to the vehicle body, the body pose is obtained from the camera pose by the fixed extrinsic transform, so that roll, pitch, and yaw of the vehicle follow directly from the estimated camera rotation. A comparison with trajectory-reconstruction methods adopted by other benchmarks is provided in Appendix A.4.

3.2 Metric Design

Trajectory alignment. Planar position is compared using ADE and FDE, computed over all maneuvers. Lower ADE and FDE indicate better trajectory alignment.

Kinematic consistency. Two trajectories with similar planar paths may differ substantially in velocity and attitude. Therefore, we compare both the translational states (longitudinal and lateral velocity, v_x and v_y), and the rotational states (roll, pitch, yaw) by root mean square error (RMSE) over the corresponding state sequence. Since RMSE penalizes large deviations more heavily than mean absolute error, it reflects not only overall accuracy, but also severe local inconsistencies. All kinematic-consistency metrics are lower-is-better.

Dynamic consistency. Trajectory and kinematic errors quantify agreement with one recorded outcome but do not test physical feasibility or causal sensitivity. We employ the friction-circle constraint rate (FCCR), which uses the adhesion bound on the resultant planar acceleration and reports the fraction of active-maneuver frames in which the recovered acceleration satisfies the bound for the friction coefficient of the sample's road condition, within a fixed tolerance. Then we construct samples under different physical conditions and perform cross-instance evaluation. We evaluate condition-aware dynamic consistency (CDC) from three aspects: response direction consistency ($\mathrm{CDC}_{\mathrm{Dir}}$) measures whether the model correctly predicts the direction of vehicle-state changes induced by variations in physical conditions, response magnitude consistency ($\mathrm{CDC}_{\mathrm{Mag}}$) evaluates the accuracy with which the model captures the strength of such response changes, and overall trend consistency ($\mathrm{CDC}_{\mathrm{Trend}}$) measures whether, as a continuous physical condition varies across multiple levels, the model response follows the same variation pattern as the real vehicle.

VehDyn score. Each metric m is first mapped to a normalized score $s_m \in [0, 1]$ with higher-is-better orientation. Metrics are then averaged within each level and the level scores are summed:

$$\mathrm{VehDyn} = \sum_{\ell \in \mathcal{L}} \underbrace{\frac{1}{|M_\ell|} \sum_{m \in M_\ell} \tilde{s}_m}_{S_\ell}, \tag{1}$$

where $\mathcal{L}$ indexes the three levels, M_ℓ is the set of metrics in level ℓ, and S_ℓ is the level score. Averaging within M_ℓ makes a level score independent of how many metrics instantiate it. Summing over $\mathcal{L}$ gives each level equal weight, so that the relative importance of trajectory, kinematics, and dynamics is an explicit design choice. Scores are computed per maneuver family and macro-averaged with equal weight. The normalization process is described in Appendix A.3.

3.3 Complementary Visual Evaluation

We also score every generated video under two established protocols without modification, i.e., DrivingGen (Zhou et al., 2026) and WorldScore (Duan et al., 2025) to quantify the gap between visual and dynamic scores.

DrivingGen. We adopt Distribution (FVD) and Temporal Consistency (Video Consistency and Trajectory Consistency). FVD measures clip-level distributional realism between inflated 3D ConvNet (Carreira & Zisserman, 2017) feature distributions of generated and ground-truth videos. Video Consistency measures frame-to-frame DINOv3 (Siméoni et al., 2025) feature similarity after motion-adaptive temporal subsampling with an optical-flow model (Wang et al., 2024b). Trajectory Consistency computes how stable a trajectory's velocity and acceleration are over time.

WorldScore. We adopt Quality (3D Consistency, Photometric Consistency, Style Consistency, and Subjective Quality). 3D Consistency uses DROID-SLAM (Teed & Deng, 2021) to compute the reprojection error between co-visible pixels in consecutive frames. Photometric Consistency measures the average end-point error (AEPE) of optical flow between consecutive frames. Style Consistency measures Gram-matrix style drift (Gatys et al., 2015). Subjective Quality uses the arithmetic mean of CLIP-Aesthetic (Schuhmann, 2022) and CLIP-IQA+ (Wang et al., 2023).

4 Experiments

4.1 Experimental Setup

We evaluate 12 open-source video world models in three categories: five general world models (Wan 2.2-I2V (Wan et al., 2025), LTX-Video 2.5 (HaCohen et al., 2025), SkyReels-V2-I2V (Chen et al.,

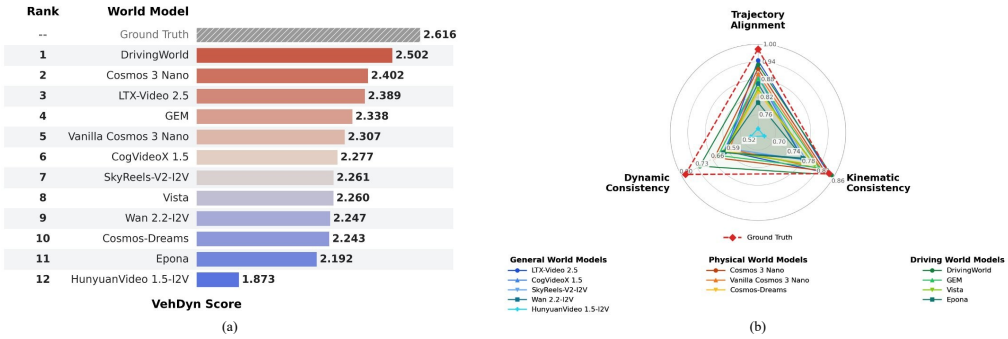


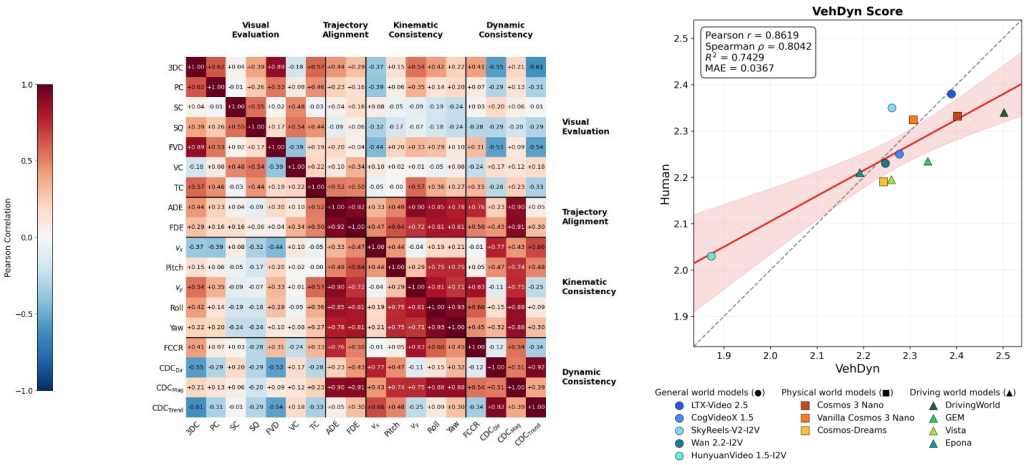
Figure 3: Overall results of the proposed VehDyn benchmark. (a) VehDyn score of the 12 evaluated models ordered by rank, with the ground-truth video scored through the same pipeline as the reference ceiling (dashed line). (b) Normalized level scores of the 12 evaluated models and ground truth.

Table 2: Detailed quantitative results of the VehDyn benchmark for 12 representative video world models. Top: VehDyn metrics. Rank is by VehDyn score. Bottom: Visual evaluation under the WorldScore and DrivingGen benchmarks. Rank is by the sum of the normalized level scores. Best results are in red region, second best in orange region, and third best in blue region. Models fall into three categories: general world models, physical world models, and driving world models.

VehDyn Metrics	Trajectory Alignment		Kinematic Consistency					Dynamic Consistency				Rank
Models	All		Emergency-Braking		Single-Lane-Changing			All	All			
	ADE ↓	FDE ↓	v_x ↓	Pitch ↓	v_y ↓	Roll ↓	Yaw ↓	FCCR ↑	CDC_{Dir} ↑	CDC_{Mag} ↑	CDC_{Trend} ↑	
Ground Truth	0.514	0.814	0.897	2.522	0.697	0.602	3.885	0.900	0.722	0.713	0.803	–
General World Models												
LTX-Video 2.5	**1.665**	**2.310**	3.217	**2.077**	**0.932**	**0.558**	3.330	0.686	0.563	0.610	0.539	3
CogVideoX 1.5	2.804	5.368	5.683	3.018	1.270	0.971	3.543	**0.738**	0.582	0.557	0.565	6
SkyReels-V2-I2V	2.853	4.757	7.243	2.415	1.127	0.828	3.291	0.569	0.577	0.584	0.590	7
Wan 2.2-I2V	3.620	5.902	9.063	2.161	1.552	0.658	3.283	0.595	0.571	0.602	0.643	9
HunyuanVideo 1.5-I2V	6.826	14.437	7.471	3.283	4.403	1.872	4.334	0.488	0.528	0.436	0.485	12
Physical World Models												
Cosmos 3 Nano	2.432	3.517	1.866	2.173	1.847	0.772	3.556	0.591	0.646	0.617	0.746	2
Vanilla Cosmos 3 Nano	2.736	4.476	4.415	2.203	1.288	0.590	3.335	0.601	0.581	0.599	0.543	5
Cosmos-Dreams	3.699	8.599	5.959	2.158	1.740	0.588	3.325	0.675	0.566	0.572	0.576	10
Driving World Models												
DrivingWorld	1.914	3.214	**1.418**	2.141	1.313	0.741	**3.250**	0.685	**0.708**	**0.664**	0.817	1
GEM	3.110	5.347	3.166	2.398	1.971	0.802	3.280	0.549	0.624	0.591	0.769	4
Vista	4.176	6.663	3.653	2.282	2.328	0.869	3.476	0.533	0.604	0.568	0.706	8
Epona	5.775	7.805	2.474	2.127	5.269	1.275	3.624	0.394	0.672	0.528	**0.844**	11

Other Benchmarks	WorldScore				DrivingGen			Rank
Models	Quality				Distribution	Temporal Consistency		
	3D Consistency ↑	Photometric Consistency ↑	Style Consistency ↑	Subjective Quality ↑	FVD ↓	Video Consistency ↑	Trajectory Consistency ↑	
Ground Truth	79.32	19.88	94.71	42.84	30.1	0.9102	0.4469	–
General World Models								
SkyReels-V2-I2V	71.90	**60.36**	88.78	40.85	277.2	0.9094	**0.5130**	1
LTX-Video 2.5	**75.51**	27.36	94.46	29.99	**224.0**	0.9156	0.4709	2
Wan 2.2-I2V	66.37	18.58	95.55	38.71	243.8	0.9206	0.3936	4
HunyuanVideo 1.5-I2V	58.39	16.01	94.50	33.43	316.0	0.8771	0.3726	6
CogVideoX 1.5	58.33	0.00	94.86	31.90	559.2	0.9030	0.4752	7
Physical World Models								
Cosmos 3 Nano	67.61	9.92	96.17	39.40	286.8	0.8742	0.4744	3
Vanilla Cosmos 3 Nano	64.45	7.52	95.62	**40.92**	372.3	0.9125	0.4374	5
Cosmos-Dreams	66.23	12.51	76.28	5.78	233.8	0.7747	0.3981	8
Driving World Models								
DrivingWorld	50.79	14.81	**99.32**	10.08	506.9	0.8982	0.3637	9
Epona	48.94	0.72	96.05	33.00	599.4	0.9126	0.3701	10
Vista	49.16	5.96	86.14	22.13	772.2	**0.9281**	0.4475	11
GEM	54.73	7.39	72.92	16.62	599.8	0.8958	0.4138	12

2025), HunyuanVideo 1.5-I2V (Wu et al., 2025), and CogVideoX 1.5 (Yang et al., 2025)), three physical world models (Cosmos 3 Nano (Agarwal et al., 2026), Vanilla Cosmos 3 Nano (Agarwal et al., 2026), and Cosmos-Dreams (Basant et al., 2026)), and four driving world models (Vista (Gao et al., 2024), DrivingWorld (Hu et al., 2026), GEM (Hassan et al., 2025), and Epona (Zhang et al., 2025b)). Their architectures, conditioning, and inference costs are summarized in Table 6 in Ap-



(a) (b)

Figure 4: Cross-dimension correlation analysis. (a) Pearson correlation among the 18 metrics, with black lines separating the visual, trajectory, kinematic, and dynamic groups. (b) Pearson correlation between VehDyn score and human evaluation.

pendix A.3. All models except Vanilla Cosmos 3 Nano, which is evaluated zero-shot to isolate the effect of post-training, are post-trained on the VehDyn dataset following their official implementations on eight A100 80 GB GPUs. Episodes are split sequence-wise into training, validation, and test sets at an 80%/10%/10% ratio. General and physical world models predict future frames from the initial observation and the caption, while driving world models from the initial observation and the trajectories/poses. Full quantitative and qualitative results are given in Appendices A.4 and A.5.

4.2 Quantitative Results

Per-Dimension Analysis. Overall VehDyn results are reported in Figure 3 and Table 2, with per-level normalized scores in Tables 14–17 in Appendix A.4. DrivingWorld achieves the strongest performance among the evaluated models, with a VehDyn score of 2.502 (95.6% of Ground Truth), followed by Cosmos 3 Nano (2.402), LTX-Video 2.5 (2.389), and GEM (2.338), with Hunyuan-Video 1.5-I2V ranking last (1.873). Post-training helps Cosmos 3 Nano exceed its zero-shot base by 0.095, with the gain concentrated in dynamic consistency. *Trajectory alignment* is nearly saturated: ten of the twelve models fall within 20% of Ground Truth. The top three are: LTX-Video 2.5 at 0.945, DrivingWorld at 0.931, and Cosmos 3 Nano at 0.917. *Kinematic consistency* is comparatively insensitive to model choice at the level score, where DrivingWorld (0.853) exceeds Ground Truth (0.847), but the per-state errors separate the models: longitudinal velocity spans an order of magnitude (DrivingWorld 1.418 m/s vs. Wan 2.2-I2V 9.063 m/s), whereas yaw varies only between 0.571 and 0.682. *Dynamic consistency* shows that DrivingWorld retains 91.6% of Ground Truth, Cosmos 3 Nano and GEM retain 82.9% and 80.7%, and every other model retains 78% or less. CogVideoX 1.5 attains the highest friction-circle compliance (0.738) but ranks sixth on dynamics, and LTX-Video 2.5, second in compliance (0.686), ranks eighth, indicating that feasible motion does not imply a correct response to changed friction or speed. Conversely, Epona has the highest trend consistency (0.844) and the second-highest direction consistency (0.672) but the lowest compliance (0.394), so its responses have the correct ordering while exceeding the adhesion limit. Only DrivingWorld is strong on all four sub-metrics, which accounts for its overall rank.

Cross-Dimension Analysis. Figure 4 reports Pearson correlations among the metrics and the agreement between VehDyn score and human judgments. Visual evaluation under the WorldScore and DrivingGen protocols is only weakly related to vehicle-dynamics fidelity. Following the human-alignment protocol (Huang et al., 2024), 50 annotators scored trajectory alignment, kinematic consistency, and dynamic consistency on a [0, 1] scale. The VehDyn score exhibits strong agreement with human evaluation, achieving Pearson $r = 0.8619$ and Spearman $\rho = 0.8042$. More details about human alignment are provided in Appendix A.6.

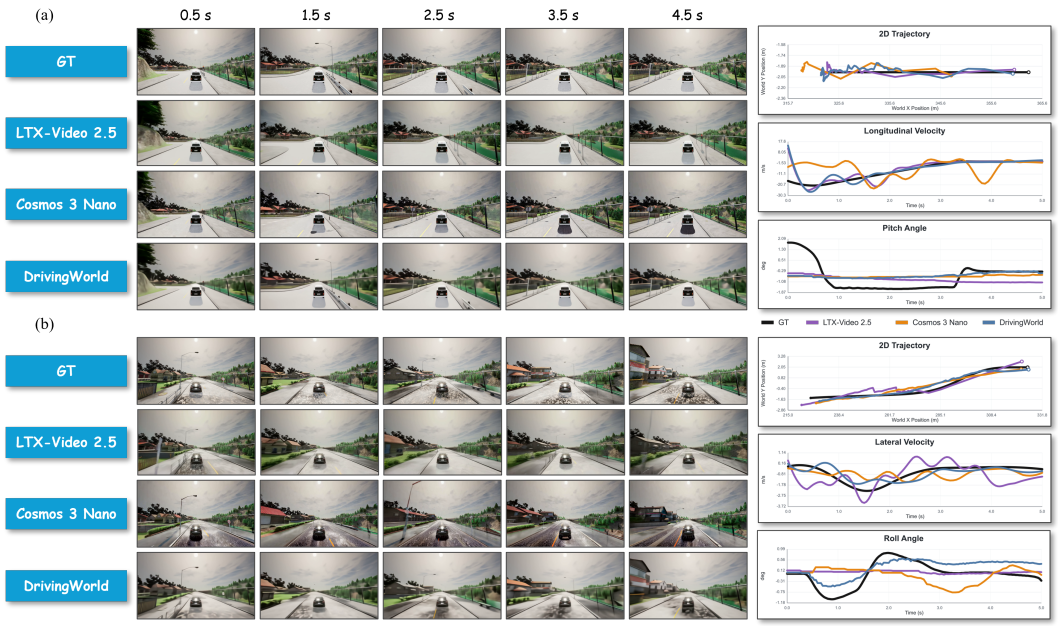


Figure 5: Qualitative comparison of the top-three models with ground truth. (a) Emergency-braking scenario. (b) Single-lane-changing scenario. Frames are shown at 0.5 s to 4.5 s. The right-hand panels plot the trajectory, velocity, and attitude.

4.3 Qualitative Results

Figure 5 compares the three highest-ranked models with ground truth under identical scenario configurations. In the emergency-braking scenario, all three models successfully completed the maneuver but exhibited distinct dynamic behaviors. LTX-Video 2.5 produced temporally coherent visual transitions and a relatively smooth deceleration, but failed to capture the transient pitch response. Cosmos 3 Nano preserved richer scene details, but its deceleration showed noticeable deviations and fluctuations. DrivingWorld, despite lower visual fidelity, reproduced the ground-truth motion more accurately, with both its velocity and trajectory closely following the ground truth. In the single-lane-changing scenario, all three models also complete the maneuver. LTX-Video 2.5 and Cosmos 3 Nano retain better visual appearance while exhibiting larger motion discrepancies, whereas DrivingWorld sacrifices some visual details but achieves better consistency with ground-truth states.

5 Conclusion

We introduce VehDyn, a driving world model benchmark that evaluates whether generated driving futures obey vehicle kinematics and dynamics. VehDyn couples CARLA visual rendering with CarSim vehicle dynamics to produce 10,080 configurations from a full factorial design over six factors, each paired with synchronized ground-truth vehicle states. A hierarchical protocol evaluates trajectory alignment, kinematic consistency, and dynamic consistency. The resulting VehDyn score agrees closely with human judgment. Across 12 video world models, DrivingWorld achieves the highest VehDyn score, followed by Cosmos 3 Nano and LTX-Video 2.5. The results reveal a systematic gap between plausible and physically consistent generation: trajectory-level metrics are nearly saturated, kinematic errors of the best models approach the extraction floor, and no model reaches 92% of ground truth on dynamic consistency. Visual-quality metrics are only weakly related to these outcomes. This work has limitations. VehDyn is simulation-based, evaluates a single ego vehicle without traffic interaction, and relies on monocular state recovery whose error bounds the final score. Future work will extend the benchmark to multi-agent interaction, multi-view observation, and closed-loop planning, and will validate the protocol against real-world driving scenarios.

AI USE STATEMENT

In this work, we used generative AI tools for grammar check and language improvement. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The proposed *dataset* and *evaluation metrics* will be released when the paper is accepted.

REFERENCES

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.

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. Act-bench: Towards action controllable world models for autonomous driving. *arXiv preprint arXiv:2412.05337*, 2024.

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.

Aarti Basant, Amlan Kar, Despoina Paschalidou, Fangyin Wei, Francesco Ferroni, Guillermo Garcia Cobo, Haithem Turki, Huan Ling, Jaewoo Seo, James Lucas, et al. Nvidia omnidreams: Real-time generative world model for closed-loop autonomous vehicle simulation. *arXiv preprint arXiv:2606.03159*, 2026.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11618–11628. IEEE, 2020.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

Kunlin Cai, Rui Song, Jinghuai Zhang, Kaiyuan Zhang, Pranav Bodapati, Alicia Yu, Fnu Suya, Mohammad Rostami, Jiaqi Ma, and Yuan Tian. What-if world: A causal benchmark for general world models in embodied scenarios. *arXiv preprint arXiv:2605.27589*, 2026.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

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.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning*, pp. 1–16. PMLR, 2017.

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*, pp. 27713–27724, 2025.

Chun-Kai Fan, Xiaowei Chi, Xiaozhu Ju, Hao Li, Yong Bao, Yu-Kai Wang, Lizhang Chen, Zhiyuan Jiang, Kuangzhi Ge, Ying Li, et al. Wow, wo, val! a comprehensive embodied world model evaluation turing test. *arXiv preprint arXiv:2601.04137*, 2026.

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.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Simon Gerstenecker, Andreas Geiger, and Katrin Renz. Fail2drive: Benchmarking closed-loop driving generalization. *arXiv preprint arXiv:2604.08535*, 2026.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems*, 31, 2018.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2025.

Ruiyang Hao, Bowen Jing, Haibao Yu, and Zaiqing Nie. Styledrive: Towards driving-style aware benchmarking of end-to-end autonomous driving. *AAAI Conference on Artificial Intelligence*, 40 (6):4627–4635, 2026.

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*, pp. 22404–22415, 2025.

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.

Xiaotao Hu, Mingkai Jia, Xiaoyang Guo, Qian Zhang, Xiao-xiao Long, and Wei Yin. Drivingworld: Constructing world model for autonomous driving via video gpt. In *International Conference on Pattern Recognition*, pp. 276–291. Springer, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818. IEEE, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems*, 38, 2026.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3461–3475, 2022.

Ao Liang, Lingdong Kong, Tianyi Yan, Hongsi Liu, Yu Yang, Ziqi Huang, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, et al. Worldlens: Full-spectrum evaluations of driving world models in real world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36385–36399, 2026.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Finn Rasmus Schäfer, Yuan Gao, Dingrui Wang, Thomas Stauner, Stephan Günnemann, Mattia Piccinini, Sebastian Schmidt, and Johannes Betz. Egodyn-bench: Evaluating ego-motion understanding in vision-centric foundation models for autonomous driving. *arXiv preprint arXiv:2604.22851*, 2026.

Christoph Schuhmann. Clip+ mlp aesthetic score predictor. *Clip+ Mlp Aesthetic Score Predictor*, 6:S1, 2022.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *AI Open*, 7:208–226, 2026.

Yuqian Shao, Xiaosong Jia, Langechuan Liu, and Junchi Yan. Can users specify driving speed? bench2drive-speed: Benchmark and baselines for desired-speed conditioned autonomous driving. *arXiv preprint arXiv:2603.25672*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Xinglong Sun, Kevin Xie, Jenny Schmalfuss, Despoina Paschalidou, Xiuming Zhang, Sanja Fidler, Kashyap Chitta, and Jose M Alvarez. Drivejudge: Rethinking autonomous driving evaluation with vision-language models. *arXiv preprint arXiv:2606.17362*, 2026.

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.

Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. *AAAI Conference on Artificial Intelligence*, 37(2):2555–2563, 2023.

Tianyi Wang, Jiazhou Chen, Yiming Xu, Xiangyu Li, Tianyi Zeng, Chih-Hsien Chou, Ning Lu, Liang Peng, Junfeng Jiao, and Christian Claudel. Robophys-3d: A comprehensive embodied world model evaluation via 3d reconstruction. *arXiv preprint arXiv:2608.28718*, 2026a.

Tianyi Wang, Tianyi Zeng, Zimo Zeng, Feiyang Zhang, Yujin Wang, Xiangyu Li, Yiming Xu, Sikai Chen, Junfeng Jiao, Christian Claudel, et al. Dbpnet: Damper characteristics-based bayesian physics-informed neural network for wheel load estimation. *arXiv preprint arXiv:2605.24860*, 2026b.

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*, pp. 55–72. Springer, 2024a.

Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pp. 36–54. Springer, 2024b.

Yuqi Wang, Ke Cheng, Jiawei He, Qitai Wang, Hengchen Dai, Yuntao Chen, Fei Xia, and Zhaoxiang Zhang. Drivingdojo dataset: Advancing interactive and knowledge-enriched driving world model. *Advances in Neural Information Processing Systems*, 37:13020–13034, 2024c.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759. IEEE, 2024d.

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.

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, Jing Tang, Lei Sun, Jiahong Wu, Xiangxiang Chu, Zeming Liu, and Kaiqi Huang. Omni-worldbench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14662–14672. IEEE, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

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

Tianyi Zeng, Tianyi Wang, Liyang Yu, Zeyu Liu, Haotian Chen, and Xinbo Chen. Wheel load estimation and anti-roll bar control using suspension analysis with neural network. In *IFToMM World Congress on Mechanism and Machine Science*, pp. 332–342. Springer, 2023.

Tianyi Zeng, Tianyi Wang, Zimo Zeng, Feiyang Zhang, Jiseop Byeon, Yujin Wang, Yajie Zou, Yangyang Wang, Junfeng Jiao, Christian Claudel, et al. Damper-b-pinn: damper characteristics-based bayesian physics-informed neural network for vehicle state estimation. In *IEEE Intelligent Vehicles Symposium*, pp. 220–227. IEEE, 2026.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, et al. World-in-world: World models in a closed-loop world. *arXiv preprint arXiv:2510.18135*, 2025a.

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 *IEEE/CVF International Conference on Computer Vision*, pp. 27220–27230. IEEE, 2025b.

Zhiyuan Zhang, Zhenghao Jin, Yanlun Peng, Xianda Guo, Haoran Liu, Shaofeng Zhang, Xingjun Ma, Zuxuan Wu, Junchi Yan, Xiaosong Jia, et al. Bench2drive-robust: Benchmarking closed-loop autonomous driving under deployment perturbations. *arXiv preprint arXiv:2605.18059*, 2026.

Changyuan Zhao, Guangyuan Liu, Ruichen Zhang, Yinqiu Liu, Jiacheng Wang, Jiawen Kang, Dusit Niyato, Zan Li, Xuemin Shen, Zhu Han, et al. Edge general intelligence through world models, large language models, and agentic ai: Fundamentals, solutions, and challenges. *IEEE Transactions on Cognitive Communications and Networking*, 2026.

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.

Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven L. Waslander. Driv-inggen: A comprehensive benchmark for generative video world models in autonomous driving. In *International Conference on Learning Representations*, 2026.