Is Your Driving World Model an All-Around Player?

Lingdong Kong[*,1,†], Ao Liang[*,1], Tianyi Yan[*,2], Hongsi Liu[*,3], Yu Yang[*,4], Ziqi Huang[5], Xian Sun[6], Wei Yin[7], Jialong Zuo[8], Yixuan Hu[9], Dekai Zhu[9], Dongyue Lu[1], Youquan Liu[10], Guangfeng Jiang[3], Linfeng Li[1], Xiangtai Li[5], Long Zhuo[11], Lai Xing Ng[12], Benoit R. Cottereau[13], Changxin Gao[8], Liang Pan[11], Wei Tsang Ooi[1,‡], Ziwei Liu[5,‡]

[1]NUS [2]UM [3]USTC [4]ZJU [5]NTU [6]Duke [7]Horizon [8]HUST [9]TUM [10]FDU [11]SH Lab [12]A*STAR [13]CNRS

[*]Equal Contributions [†]Project Lead [‡]Corresponding Authors

Project Page: *Link* **GitHub:** *Link* **HuggingFace:** *Link*

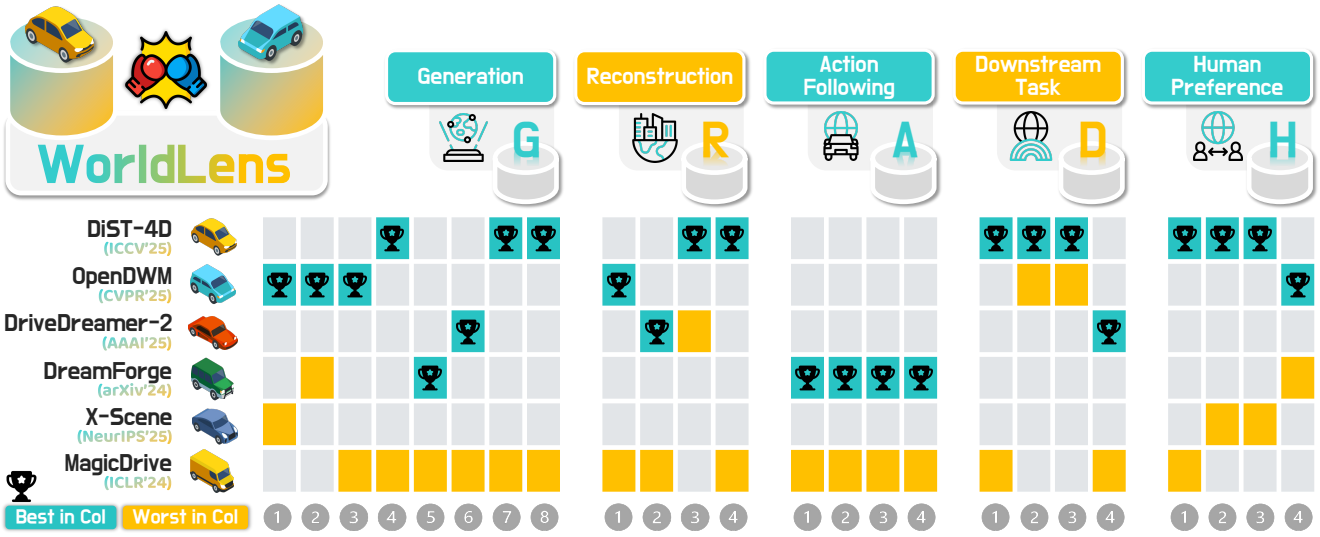


Figure 1. **How do world models perform in the real world?** This work introduces *WorldLens*, a unified benchmark for evaluations on [1]***Generation***, [2]***Reconstruction***, [3]***Action-Following***, [4]***Downstream Task***, and [5]***Human Preference***, across **24 dimensions**. We observe *no single model dominates across all axes*, highlighting the need for balanced progress toward physically realistic world modeling.

Abstract

Today's driving world models can generate remarkably realistic dash-cam videos, yet no single model excels universally. Some generate photorealistic textures but violate basic physics; others maintain geometric consistency but fail when subjected to closed-loop planning. This disconnect exposes a critical gap: the field evaluates how real generated worlds appear, but rarely whether they behave realistically. *We introduce* ***WorldLens****, a unified benchmark that measures world-model fidelity across the full spectrum, from pixel quality and 4D geometry to closed-loop driving and human perceptual alignment, through five complementary aspects and 24 standardized dimensions. Our evaluation of six representative models reveals that no existing approach dominates across all axes: texture-rich models violate geometry, geometry-aware models lack behavioral fidelity, and even the strongest performers achieve only 2–3 out of 10 on human realism ratings. To bridge algorithmic metrics with human perception, we further contribute* ***WorldLens-26K****, a 26,808-entry human-annotated preference dataset pairing numerical scores with textual rationales, and* ***WorldLens-Agent****, a vision-language evaluator distilled from these judgments that enables scalable, explainable auto-assessment. Together, the benchmark, dataset, and agent form a unified ecosystem for assessing generated worlds not merely by visual appeal, but by physical and behavioral fidelity.*

1. Introduction

Consider a state-of-the-art driving world model that generates visually compelling multi-view videos. The textures are sharp, the lighting is natural, and the motion appears

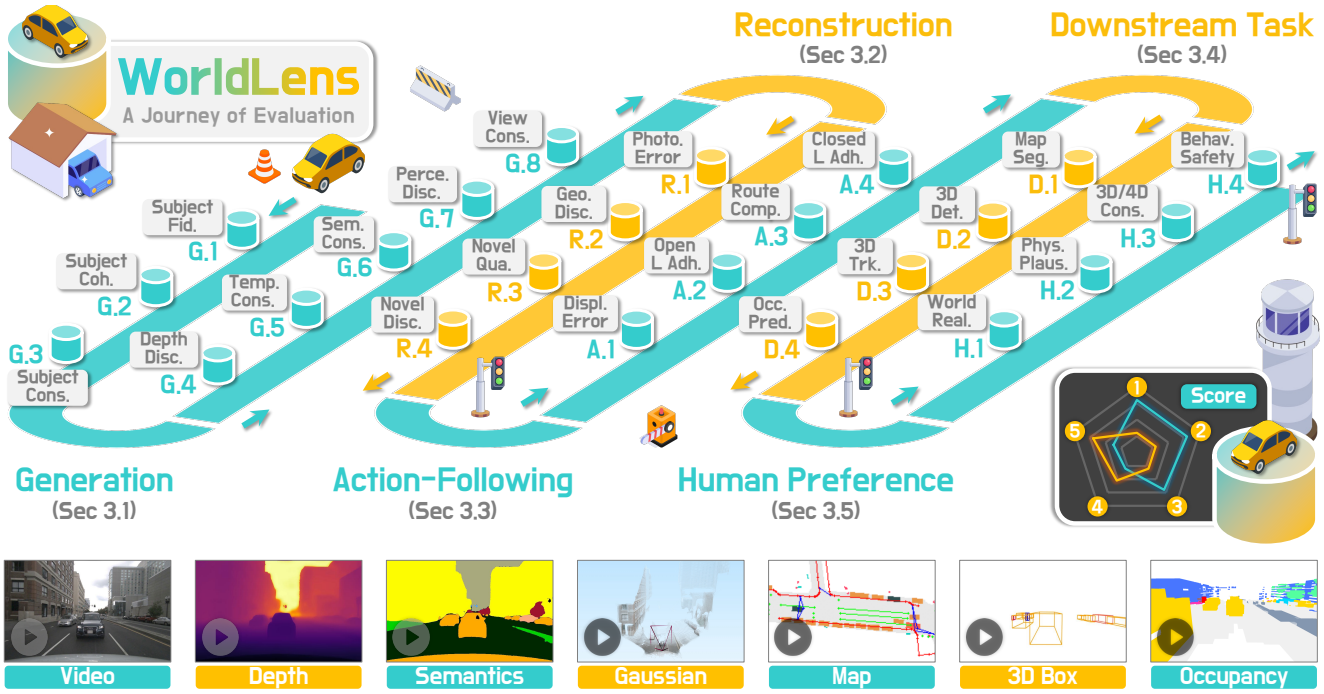


Figure 2. ***WorldLens*** unifies five complementary aspects, namely [1]***Generation***, [2]***Reconstruction***, [3]***Action-Following***, [4]***Downstream Task***, and [5]***Human Preference***, that jointly cover visual, structural, functional, and perceptual quality across 24 interpretable dimensions.

fluid. Yet when a pretrained planner is deployed within this generated world to execute a routine maneuver, it frequently collides or drifts off-road within seconds. This is not an edge case: our experiments show that even the best-performing world models complete fewer than 14% of navigation routes under closed-loop control.

This paradox lies at the heart of the current world-modeling landscape. Recent years have witnessed remarkable progress in generative driving models [2, 22, 27], generating multi-view video sequences that are increasingly difficult to distinguish from real footage. Yet the prevailing evaluation metrics, such as FID, FVD, and LPIPS, were designed for *image quality*, not *world fidelity* [13, 15]. These metrics quantify perceptual similarity but reveal nothing about whether the underlying geometry is coherent, whether the physics are plausible, or whether the generated world can support downstream autonomy tasks [17, 20]. As such, the field has been optimizing for an incomplete objective: worlds that *appear* real but do not *behave* realistically. The absence of a comprehensive evaluation protocol means that progress on one axis (*e.g.*, texture realism) can mask regression on others (*e.g.*, 3D consistency or action controllability), making it difficult to compare models or identify meaningful bottlenecks.

We introduce ***WorldLens***, a benchmark designed to address this gap. Rather than scoring generated videos along a single quality axis, we decompose evaluation along **five complementary axes**: *Generation* (visual realism and temporal stability across eight dimensions), *Reconstruction* (whether videos can be lifted into coherent 4D Gaussian fields), *Action-Following* (whether pretrained planners can operate safely in the generated world), *Downstream Task* (whether synthetic data supports real-world perception models), and *Human Preference* (subjective judgments of realism, physics, and safety from 930+ hours of annotation). Together, these 24 dimensions span the full spectrum from pixel fidelity to functional reliability, as illustrated in Figs. 1 and 2. By evaluating each model through every lens simultaneously, WorldLens makes the trade-offs between competing design choices explicit and quantifiable.

Our evaluation reveals a notable finding: **no existing world model dominates across all axes**. Models that lead in texture quality often violate physics; those with strong geometry fail behaviorally; and even the best performers score only 2–3 out of 10 on human realism ratings. These results suggest that current architectures still treat appearance, geometry, and dynamics as largely independent objectives, a decomposition that prevents holistic world understanding. To enable such evaluation at scale beyond manual annotation, we further contribute **WorldLens-26K**, a dataset of 26,808 human-scored video entries with textual rationales, and **WorldLens-Agent**, a vision-language critic agent distilled from these annotations via LoRA-based SFT on Qwen3-VL-8B [1]. The agent achieves strong zero-shot alignment with human judgments on unseen models, enabling automated, explainable evaluation at scale.

2. The *WorldLens* Benchmark

A world model that generates visually appealing frames does not necessarily *understand* the world it portrays. To bridge this gap between surface-level appearance and deeper understanding, WorldLens structures evaluation along five complementary axes (Fig. 2), each probing a distinct facet of world fidelity, ranging from low-level pixel quality to high-level behavioral realism.

Generation: How Realistic Does It Appear? We decompose visual quality into eight dimensions organized around three questions. *Are individual objects realistic?* Subject Fidelity (G.1, ↑) scores cropped objects via class-specific classifiers [8]; Subject Coherence (G.2, ↑) and Consistency (G.3, ↑) track identity stability across frames using ReID [11, 38] and DINO [4] features. *Is the scene temporally and geometrically stable?* Depth Discrepancy (G.4, ↓) measures frame-to-frame depth smoothness [34]; Temporal (G.5, ↑) and Semantic (G.6, ↑) Consistency assess global stability in CLIP [23] and SegFormer [32] spaces, respectively. *Is the video perceptually convincing as a whole?* Perceptual Discrepancy (G.7, ↓) computes FVD on I3D features [5, 31], and Cross-View Consistency (G.8, ↑) evaluates multi-camera alignment via LoFTR [29].

Reconstruction: Can a Coherent 3D World Be Recovered? This aspect serves as the definitive test for geometric coherence. We lift each generated sequence into a 4D Gaussian field [16] and re-render it under both training and novel camera poses. Models generating sharp 2D frames often exhibit severe degradation under this protocol, generating geometric "floaters" and depth artifacts that reveal how loosely current architectures couple temporal frames. We measure Photometric Error (R.1, ↓) via LPIPS/PSNR/SSIM at training poses, Geometric Discrepancy (R.2, ↓) via depth comparison using Grounded-SAM 2 [24, 25], and both Novel-View Quality (R.3, ↑) and Novel-View Discrepancy (R.4, ↓) under unseen viewpoints.

Action-Following: Can a Planner Operate Reliably Within It? This is arguably the most consequential aspect: if a pretrained planner [12, 14] cannot make safe decisions within the generated world, the model's practical value for autonomous driving remains limited. We measure Displacement Error (A.1, ↓) between predicted and real trajectories, Open-Loop Adherence (A.2, ↑) via the PDMS score [6], Route Completion (A.3, ↑) in closed-loop simulation, and Closed-Loop Adherence (A.4, ↑) via the Arena Driving Score [35]. As we show in Sec. 3, the disparity between open- and closed-loop results is substantial and highly informative.

Downstream Task: Is the Synthetic Data Practically Useful? We assess whether generated videos can support 3D perception pipelines trained on real data. Specifically, we evaluate Map Segmentation (D.1, ↑) via BEVFusion [19], 3D Object Detection (D.2, ↑) via NDS [3], 3D Tracking (D.3, ↑) via AMOTA [7], and Occupancy Prediction (D.4, ↑) via SparseOcc RayIoU [30]. Notably, even models with strong perceptual quality can degrade detection accuracy by 30–50%.

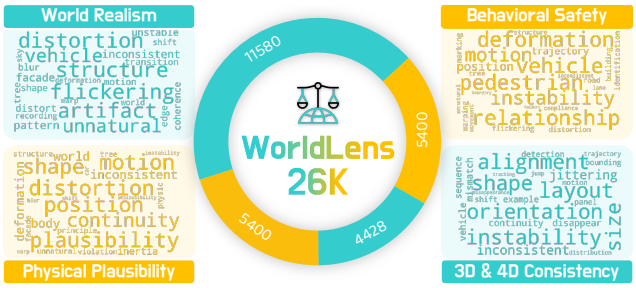


Figure 3. Statistics and word clouds of **WorldLens-26K**. Frequent keywords align with target criteria, confirming that annotators maintain consistent, dimension-specific reasoning.

Human Preference: How Do Human Observers Judge Quality? Quantitative metrics can miss perceptual artifacts that human observers identify readily. Ten annotators in two independent groups rated generated videos on a 1–10 scale across four perceptual dimensions: World Realism (H.1), Physical Plausibility (H.2), Behavioral Safety (H.4), and 3D & 4D Consistency (H.3). Each annotator reviewed four time-synchronized modalities, including the generated video alongside its semantic segmentation mask, estimated depth map, and 3D bounding box overlay, facilitating holistic judgment across visual, structural, and behavioral aspects. Each annotation required ∼128 seconds on average, totaling over **930** hours. Divergent ratings between groups were re-evaluated to ensure inter-annotator consistency.

From Human Judgment to Scalable Evaluation.

The annotation effort leads to **WorldLens-26K**, a dataset of 26,808 scoring records, each coupling a numerical rating with a free-text rationale that articulates *why* a particular score was assigned. As in Fig. 3, word-cloud of these rationales reveals strong topical alignment with their respective dimensions (*e.g.*, "shape" and "reflection" for realism, "motion" and "safety" for behavioral assessment), confirming that annotators systematically attend to the intended perceptual criteria. Building on this paired quantitative–qualitative supervision, we train **WorldLens-Agent**, a vision-language evaluator distilled from human preferences through LoRA-based SFT on Qwen3-VL [1]. The agent simultaneously predicts dimension-specific scores and produces free-text explanations that mirror human reasoning patterns, demonstrating strong zero-shot generalization to previously unseen models (Fig. 7) and offering a practical pathway toward scalable, interpretable auto-evaluation without ongoing manual labeling.

Table 1. Summary of benchmarking results of state-of-the-art driving world models for **Generation** and **Reconstruction** in WorldLens.

Model	Venue	Aspect: Generation								Aspect: Reconstruction			
		Subject Fid.↑ G.1	Subject Cohe.↑ G.2	Subject Cons.↑ G.3	Depth Disc.↓ G.4	Temp. Cons.↑ G.5	Sem. Cons.↑ G.6	Percept. Disc.↓ G.7	View Cons.↑ G.8	Photo. Error↓ R.1	Geo. Disc.↓ R.2	Novel Qual.↑ R.3	Novel Disc.↓ R.4
MagicDrive [9]	ICLR'23	28.49	75.95	65.22%	24.19	74.44%	80.63%	222.00	185.77	0.140	0.115	39.82%	427.30
DreamForge [21]	arXiv'24	31.99	75.12	76.40%	19.27	**79.82%**	84.99%	189.76	194.99	0.097	0.105	41.23%	347.70
DriveDreamer-2 [37]	AAAI'25	27.38	78.97	74.49%	17.73	79.51%	**85.91%**	127.07	302.83	0.093	**0.073**	36.10%	259.91
OpenDWM [28]	CVPR'25	**36.30**	**83.13**	**78.33%**	18.17	79.63%	84.08%	90.42	211.18	**0.065**	0.088	39.54%	287.73
DiST-4D [10]	ICCV'25	30.32	79.36	74.69%	**17.71**	77.76%	84.32%	**58.08**	**389.78**	0.066	0.080	**43.09%**	**192.39**
$\mathcal{X}$-Scene [36]	NeurIPS'25	27.17	77.22	74.37%	20.50	79.41%	83.80%	179.74	201.00	0.098	0.096	38.04%	365.71
Empirical Max	-	60.22	83.25	93.66%	14.27	93.24%	86.39%	-	570.75	0.056	-	45.69%	-

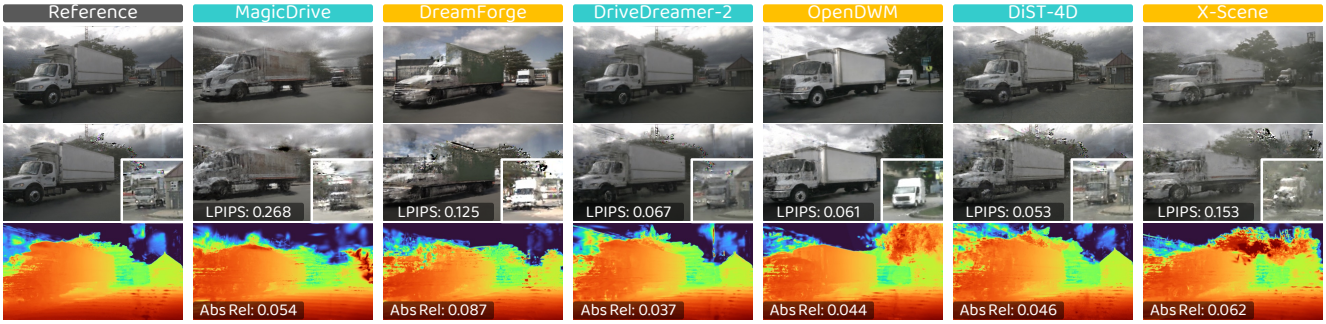


Figure 4. **4D reconstruction** from generated videos. Rows: [1]generated frame, [2]novel-view rendering at a lateral offset, [3]depth map.

3. Key Findings

We evaluate six representative models: MagicDrive [9], DreamForge [21], DriveDreamer-2 [37], OpenDWM [28], DiST-4D [10], and $\mathcal{X}$-Scene [36], across all five aspects. Rather than enumerating each metric individually, we organize the results around the most salient insights. Comprehensive details are available in the main paper [18].

No model is an all-rounder. The radar chart in Fig. 1 summarizes the landscape concisely: every model exhibits notable weaknesses. OpenDWM [28] leads in subject fidelity and coherence owing to multi-dataset training, yet DiST-4D [10] dominates in perceptual discrepancy, cross-view consistency, and all reconstruction metrics (Tab. 1). DriveDreamer-2 [37] achieves the strongest semantic consistency and geometric discrepancy, but falls short in subject-level realism. All models remain substantially below the Empirical Max, indicating that the frontier of driving world generation is far from saturated.

Perceptual quality $\neq$ functional utility. Perhaps the most significant finding is the disconnect between perceptual quality and downstream utility. OpenDWM achieves the strongest subject fidelity (36.30) and the lowest photometric reconstruction error (0.065 LPIPS), yet it scores only 21.96% on 3D object detection and 6.90% on tracking, roughly 30–50% below DiST-4D. This reveals that large-scale multi-domain training can enhance visual diversity while simultaneously hindering alignment with task-specific data distributions. In contrast, DiST-4D's geometry-aware RGB-D generation trades some visual fidelity for improved downstream performance (35.55% map segmentation, 33.22% detection, 15.30% tracking).

2D fidelity does not guarantee 4D coherence. Reconstruction serves as the definitive test, and most models exhibit significant degradation. When generated videos are lifted into 4D Gaussian fields and re-rendered from novel viewpoints, MagicDrive [9] generates dense floaters and over 2× higher photometric error compared to OpenDWM (Fig. 4). DreamForge [21] exhibits similar artifacts. Only DiST-4D maintains clean, stable geometry under lateral camera shifts, with its decoupled spatiotemporal diffusion and explicit depth supervision cutting photometric and geometric error by ∼55%. This reveals a fundamental limitation: most diffusion-based architectures couple temporal frames too loosely to achieve the spatial consistency required for faithful 3D reconstruction.

Open-loop performance masks closed-loop failure. All evaluated models achieve reasonable open-loop PDMS scores (71%–79%), indicating that planners can extract meaningful cues from generated frames in isolation. However, under closed-loop control, where planning decisions feed back into the simulation, route completion rates drop sharply to 6–14%. The best-performing model, RLGF [33], completes only 13.51% of routes. Frequent collisions and off-road departures confirm that photorealistic appearance alone cannot sustain safe navigation; the generated worlds lack the causal and physical consistency required for sustained autonomous control.

Human perception correlates with geometric consistency. Human preference scores remain notably low: averages

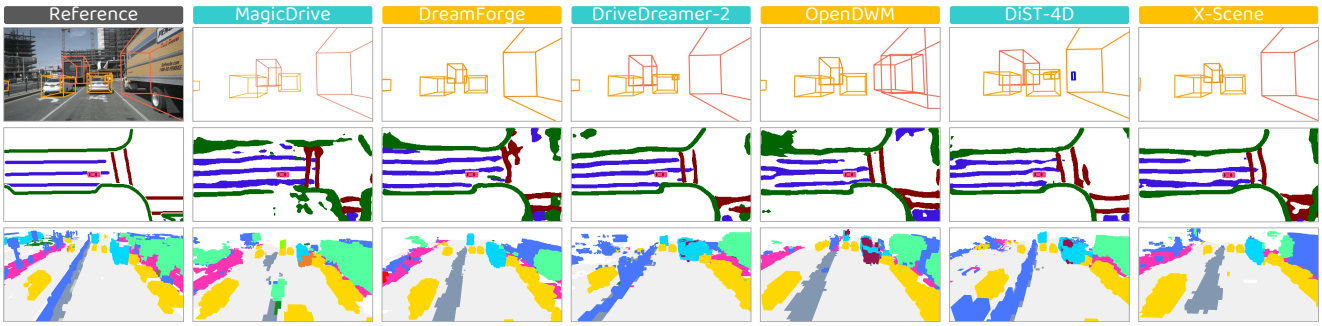
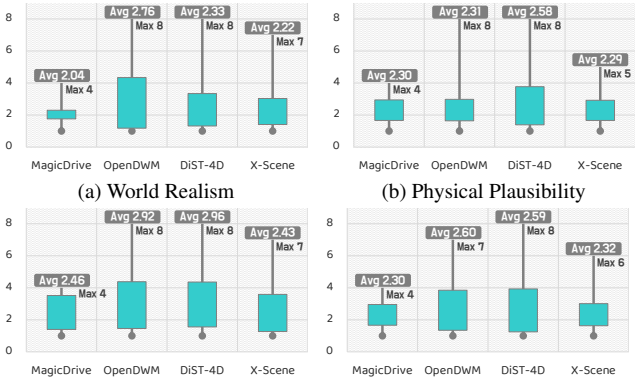


Figure 5. **Downstream task** qualitative results. Rows: [1]3D detection, [2]BEV map segmentation, and [3]semantic occupancy prediction.



(a) World Realism (b) Physical Plausibility

(c) 3D & 4D Consistency (d) Behavioral Safety

Figure 6. **Human Preference** alignments. Max, median, and average scores for each model across four perceptual dimensions. All scores remain modest (2–3 out of 10), with geometric consistency strongly correlated with perceived realism.

of 2–3 out of 10 across all dimensions (Fig. 6), underscoring that existing world models still fall well short of human-perceived realism. However, a meaningful pattern emerges: models with stronger geometric consistency (*e.g.*, DiST-4D) receive higher world realism and physical plausibility ratings, while perceptually appealing but geometrically unstable models (*e.g.*, MagicDrive) score lowest. The strong correlation between consistency and perceived realism indicates that human observers are inherently sensitive to structural coherence, which underscores the importance of geometry-aware evaluation and training.

WorldLens-Agent generalizes to unseen models. To validate the automated evaluator, we perform zero-shot assessment on Gen3C [26] videos, which were never seen during training. As in Fig. 7, the agent's ratings closely track human annotations, and its generated rationales faithfully reflect annotator reasoning, confirming reliable score-level agreement and coherent interpretive alignment. This validates the feasibility of encoding human perceptual standards into a scalable, automated evaluation pipeline.

Design Principles for Future World Models.

These findings converge on four actionable principles: *(1)* Treat geometry as a first-class training objective, as explicit depth supervision consistently improves both reconstruction and downstream perception. *(2)* Optimize jointly for appearance and structure, since current pipelines that decouple texture from geometry incur significant degradation in 4D consistency. *(3)* Evaluate under closed-loop conditions, not solely open-loop, because the open/closed-loop disparity is the most reliable indicator of world-model maturity. *(4)* Benchmark comprehensively, as single-axis metrics systematically overlook the trade-offs that characterize the current landscape.

Figure 7. Zero-shot evaluations by **WorldLens-Agent** on unseen videos (from Gen3C [26]), exhibiting strong alignment with human scores and reasoning.

4. Conclusion

WorldLens reveals a critical reality: the driving world models widely recognized for visual realism remain far from functional deployment. No model dominates across all

evaluation axes, human ratings average only 2–3 out of 10, and closed-loop control degrades almost universally. These deficiencies are not incremental; they point to fundamental gaps in how generative worlds are *constructed* and *evaluated*. By unifying 24 evaluation dimensions, we provide the community with a standardized, scalable, and human-aligned protocol for measuring progress toward worlds that are physically coherent, functionally reliable, and suitable for autonomous navigation. We anticipate that this benchmark will accelerate the transition from appearance-driven generation to behavior-grounded world modeling.

References

[1] Shuai Bai et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Philip J. Ball et al. Genie 3: A new frontier for world models, 2025.

[3] Holger Caesar et al. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020.

[4] Mathilde Caron et al. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021.

[5] Joao Carreira et al. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017.

[6] Daniel Dauner et al. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *NeurIPS*, pages 28706–28719, 2024.

[7] Shuxiao Ding et al. ADA-Track: End-to-end multi-camera 3D multi-object tracking with alternating detection and association. In *CVPR*, pages 15184–15194, 2024.

[8] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[9] Ruiyuan Gao et al. MagicDrive: Street view generation with diverse 3D geometry control. In *ICLR*, 2023.

[10] Jiazhe Guo et al. DiST-4D: Disentangled spatiotemporal diffusion with metric depth for 4D driving scene generation. In *ICCV*, pages 27231–27241, 2025.

[11] Shuting He et al. TransReID: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021.

[12] Yihan Hu et al. Planning-oriented autonomous driving. In *CVPR*, pages 17853–17862, 2023.

[13] Ziqi Huang et al. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

[14] Bo Jiang et al. VAD: Vectorized scene representation for efficient autonomous driving. In *ICCV*, 2023.

[15] Junjie Ke et al. MUSIQ: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021.

[16] Bernhard Kerbl et al. 3D Gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):1–14, 2023.

[17] Lingdong Kong et al. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.

[18] Ao Liang et al. WorldLens: Full-spectrum evaluations of driving world models in real world. In *CVPR*, 2026.

[19] Zhijian Liu et al. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, pages 2774–2781, 2023.

[20] Jinghui Lu et al. OneVL: One-step latent reasoning and planning with vision-language explanation. *arXiv preprint arXiv:2604.18486*, 2026.

[21] Jianbiao Mei et al. DreamForge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003*, 2024.

[22] Jack Parker-Holder et al. Genie 2: A large-scale foundation world model, 2024.

[23] Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[24] Nikhila Ravi et al. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

[25] Tianhe Ren et al. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[26] Xuanchi Ren et al. Gen3C: 3D-informed world-consistent video generation with precise camera control. In *CVPR*, pages 6121–6132, 2025.

[27] Lloyd Russell et al. GAIA-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

[28] SenseTime-FVG. Open Driving World Models (OpenDWM). https://github.com/SenseTime-FVG/OpenDWM, 2025.

[29] Jiaming Sun et al. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, pages 8922–8931, 2021.

[30] Pin Tang et al. SparseOCC: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *CVPR*, pages 15035–15044, 2024.

[31] Thomas Unterthiner et al. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[32] Enze Xie et al. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, pages 12077–12090, 2021.

[33] Tianyi Yan et al. RLGF: Reinforcement learning with geometric feedback for autonomous driving video generation. In *NeurIPS*, 2025.

[34] Lihe Yang et al. Depth anything v2. In *NeurIPS*, pages 21875–21911, 2024.

[35] Xuemeng Yang et al. DriveArena: A closed-loop generative simulation platform for autonomous driving. In *ICCV*, pages 26933–26943, 2025.

[36] Yu Yang et al. X-Scene: Large-scale driving scene generation with high fidelity and flexible controllability. In *NeurIPS*, 2025.

[37] Guosheng Zhao et al. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation. In *AAAI*, pages 10412–10420, 2025.

[38] Jialong Zuo et al. Cross-video identity correlating for person re-identification pre-training. *NeurIPS*, 37, 2024.