International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 12-No.4 (2026) pp. 17-26
http://www.ijcesen.com

ISSN: 2149-9144

Copyright © IJCESEN

Research Article

From Simulation to Deployment: Practical Design Principles for Physics-Consistent World Models in Autonomous Driving Systems

Tamoghna Das*

Independent Researcher, USA
* **Corresponding Author Email:** tamoghnad55@gmail.com- **ORCID:** 0009-0004-1308-7050

Article Info:

DOI: 10.22399/ijcesen.5533
Received : 11 August 2026
Revised : 15 Octoer 2026
Accepted : 19 Octoer 2026

Keywords

Autonomous Driving Simulation;
World Models;
Physics-Consistent Scene Generation;
Closed-Loop Validation;
Multi-Modal Sensor Synthesis;
Safety-Critical Curriculum;
Annotation Quality

Abstract:

Autonomous driving safety validation depends on the capacity of generative world models to produce high-fidelity, controllable synthetic driving scenarios at scale. The rapid expansion of the field has produced a landscape in which system architects must choose among architectures with substantially different capabilities, trade-offs, and failure modes. This article develops seven evidence-based design principles that distinguish world models capable of measurably improving fleet safety from those that produce perceptually plausible but operationally inert outputs. The principles address structured scene conditioning, multi-modal sensor consistency, novel-trajectory extrapolation, automated quality filtering, closed-loop evaluation priority, safety-curriculum construction, and upstream annotation validation. Drawn from peer-reviewed findings published through 2026, each principle is stated as a verifiable engineering criterion that links architectural choices to observable downstream policy improvements. The analysis reveals that failures in production deployment most commonly trace to causal decoupling between synthesized observations and agent actions, inconsistent geometric relationships across sensor modalities, and benchmark metrics that reward perceptual realism while penalizing behavioral validity. A structured integration checklist translates the seven principles into pass/fail milestones applicable across the major world model architecture families: recurrent latent, video diffusion, occupancy-centric, and hybrid Gaussian splatting systems. The framework is intended to accelerate the transition from experimental world model adoption to systematic deployment in safety-critical autonomous driving pipelines.

1. Introduction

Two years of sustained investment in generative world models have produced an embarrassment of riches for autonomous driving engineers. A dedicated World Model Bench workshop at the 2025 Conference on Computer Vision and Pattern Recognition drew more than sixty paper submissions from one leading research organization alone. Major platform releases, including a world foundation model suite designed for physical AI simulation [8], have extended access to high-fidelity generative simulation far beyond the organizations that originally built these systems. The practical challenge is no longer finding a capable world model; it is selecting the right one and deploying it in a way that actually improves fleet safety.

Offline benchmark scores do not resolve that challenge. A model that achieves state-of-the-art Fréchet Inception Distance on a held-out video dataset may still be unsuitable for closed-loop policy training if it does not translate commanded steering actions into geometrically consistent scene updates. A surround-camera synthesis pipeline that produces photorealistic imagery may corrupt the downstream policy if its accompanying synthetic

LiDAR point clouds are geometrically inconsistent with the same scene. These failures are not predictable from headline perceptual quality metrics, and they have caused production integration projects to stall or regress without obvious diagnostic signals.

Seven evidence-based design principles address this gap. Each derives from documented findings in the research literature spanning 2024 through 2026, and each is stated in a form that generates verifiable engineering milestones. The principles cover structured conditioning [6], [9], multi-modal

sensor consistency [4], [9], novel-trajectory extrapolation [5], [6], automated quality filtering [9], closed-loop evaluation priority [13], [19], safety-curriculum construction [11], [12], and upstream annotation validation [9]. Together they form a coherent integration framework rather than a list of independent tips.

The research question governing this analysis is: what design principles distinguish world models that measurably improve autonomous vehicle safety in production from those that do not? Each section below answers part of that question, and the integration checklist in Section X maps the full answer to a set of pass/fail engineering criteria.

2. Background: world model architectures and the validation shift

A world model, as the term is used throughout this article, is a neural network that learns a compressed dynamic representation of the driving environment and can synthesize future sensor readings on demand [16], [18]. The appeal for autonomous driving is straightforward: rare events such as pedestrian incursions at night or tire failures at highway speed can be generated thousands of times in software, with full physical and perceptual plausibility, rather than waited for on public roads.

Recurrent latent world models, built on Long Short-Term Memory and Gated Recurrent Unit networks, were the first practical approach [18]. They maintain a latent state vector updated at each time step from the previous state and a candidate action. Their computational economy is a genuine advantage, but accumulated prediction error limits reliable simulation to roughly two to five seconds, a horizon too short for meaningful safety assessment of highway-speed maneuvers.

Video diffusion world models represent the current frontier of perceptual quality [15]. Conditioned on structured scene descriptors, these systems synthesize photorealistic multi-camera driving footage with consistent geometry and lighting. DriveDreamer, presented at ECCV 2024, established this lineage by demonstrating real-world-driven video generation conditioned on structured traffic annotations [7]. Building on this foundation, DriveDreamer4D, presented at CVPR 2025, achieved a 32.1 percent relative improvement in Fréchet Inception Distance over a Gaussian-splatting baseline on novel viewpoints by conditioning a diffusion Transformer on high-definition map overlays, three-dimensional bounding boxes, and ego-vehicle trajectory [6]. The deployment constraint for pure diffusion systems is computational: their inference cost exceeds traditional rasterization by two to three orders of magnitude.

Occupancy-centric models sidestep the pixel-level rendering problem by representing the scene as a three-dimensional voxel grid carrying semantic class probabilities, occupancy states, and velocity vectors [3], [11]. UniScene, a CVPR 2025 contribution, jointly generates semantic occupancy, video, and LiDAR through a progressive hierarchical process, producing output that downstream planning software can consume directly without a separate pixel-decoding step [3]. GaussianWorld, also presented at CVPR 2025, pairs three-dimensional Gaussian primitives with a learned dynamics module operating over an occupancy representation and reported a greater than 2% improvement in mean Intersection-over-Union on the nuScenes dataset without increasing inference cost [2].

Hybrid Gaussian splatting systems combine the rendering quality of NeRF-derived methods with real-time performance, making them practical for reinforcement learning applications that require thousands of environmental interactions per training episode [5], [17]. The training-compute budget required for these systems is substantially lower than pure diffusion approaches, which makes them attractive for continuous fleet learning pipelines.

2. Principle 1 — structured scene conditioning

Generative fidelity alone does not produce useful simulation data. A world model that synthesizes visually compelling footage without receiving structured geometric and semantic inputs generates scenes that are internally coherent but externally uncontrollable: the locations of agents, the geometry of lane boundaries, and the planned trajectory of the ego vehicle are all determined by the model's learned priors rather than by the engineer's scenario specification.

The conditioning architecture that has gained widest traction in the literature accepts multiple simultaneous structured signals: a high-definition map encoding lane boundaries, intersection geometry, and road type; three-dimensional bounding boxes with class labels for all surrounding agents; the ego vehicle's intended trajectory as a sequence of waypoints; and control signals specifying the planned velocity profile. Zhao et al. demonstrate that conditioning a diffusion Transformer on this full set of structured geometric inputs maintains photorealistic synthesis quality while preserving scenario-level

controllability, as evidenced by the strong NTA-IoU and FID results achieved by DriveDreamer4D across novel trajectory evaluations [6]. The Cosmos-Drive-Dreams pipeline extends this pattern to eight-camera surround-view generation, where structured map and trajectory conditioning is applied jointly across all camera views to maintain cross-view geometric coherence [9].

The production implication is a shift in workflow. Rather than generating scenarios stochastically and filtering for relevance, engineering teams can specify scenarios that probe known policy weaknesses: a pedestrian appearing at a blind intersection with a lead agent blocking sightlines, or a cut-in maneuver by an aggressive driver on a specific road geometry. Conditioning fidelity is the prerequisite for this workflow, and it must be verified empirically using geometric consistency metrics before integration into the training data pipeline.

3. Principle 2—multi-modal sensor consistency

A policy trained on camera images alone cannot safely control a sensor suite that also includes LiDAR and radar at test time. The geometric relationships between these modalities are tightly constrained by the physics of the sensor placement, the calibration matrices, and the three-dimensional structure of the scene. Synthetic data that satisfies these constraints for camera imagery while violating them for LiDAR introduces systematic calibration errors that the downstream perception stack cannot recover from.

The multi-modal generation framework described in the Cosmos-Drive-Dreams pipeline addresses this requirement by decoding all sensor modalities from a shared latent scene representation [9]. Because camera images, LiDAR point clouds, and radar signatures are derived from the same compressed scene state, their geometric relationships are preserved by construction rather than enforced through post-hoc alignment. Cosmos-Drive-Dreams reports that this shared-representation design enables spatiotemporally consistent generation across all sensor streams [9]. UniScene independently validates this approach by demonstrating that occupancy-centric hierarchical generation produces semantically and geometrically consistent outputs across camera and LiDAR modalities simultaneously [3].

Engineering teams integrating world model outputs into perception training pipelines must verify multi-modal consistency before committing synthetic data to the training set. The verification protocol involves checking that three-dimensional bounding box projections are pixel-accurate across all camera views simultaneously, that LiDAR returns fall within the volume defined by the projected bounding box geometry, and that radar detections are consistent with the velocity vectors encoded in the occupancy representation [13]. Violations of any of these consistency checks indicate a world model that is not suitable for multi-modal policy training regardless of its perceptual quality scores.

4. Principle 3 - novel-trajectory extrapolation

Most world models are trained on forward-facing dashcam sequences with narrow speed and curvature variation [1], [10]. When evaluated on trajectories that deviate substantially from the training distribution, such as U-turns, emergency stops, or split-lane maneuvers, many systems produce geometrically implausible outputs: buildings that slide across the frame, ground planes that tilt, or agent bounding boxes that detach from their visual counterparts. ReconDreamer, presented at CVPR 2025, quantified this failure mode and proposed an online restoration mechanism to address it [5].

ReconDreamer's solution is progressive data incorporation during inference. Rather than completing a novel trajectory in a single forward pass, the model iteratively refines its output by incorporating intermediate synthesized frames into the conditioning context, progressively anchoring the scene geometry as the ego trajectory diverges from the training distribution. Experimental results confirm that ReconDreamer outperforms Street Gaussians with relative improvements of 24.87% in NTA-IoU (Novel Trajectory Agent IoU), 6.72% in NTL-IoU, and 29.97% in FID, where higher NTA-IoU indicates better geometric preservation of agents during novel-trajectory rendering [5]. In the more demanding large-maneuver scenario involving a 6-meter lane shift, ReconDreamer surpasses DriveDreamer4D with PVG by a relative improvement of 195.87% in NTA-IoU, establishing it as the first method to render reliably across multi-lane deviations of that magnitude [5]. DriveDreamer4D complements this finding by demonstrating that structured conditioning on high-definition map overlays and three-dimensional agent bounding boxes maintains geometric plausibility across trajectory deviations that produce complete geometric breakdown in unconditional models [6].

The engineering implication is a mandatory out-of-distribution test suite for any world model deployment. The suite must include trajectories with curvature, speed, and lane position values that fall outside the training distribution by a

statistically significant margin. A model that fails this suite is not suitable for safety-critical scenario generation regardless of its in-distribution benchmark scores.

5. Principle 4 — automated quality filtering with vision-language models

Large-scale synthetic data pipelines generate outputs at rates that make human review impractical. A system producing thousands of synthetic driving sequences per day requires an automated quality gate to detect and discard outputs with physically implausible scene elements, incomplete multi-modal correspondence, or annotation inconsistencies before they corrupt the downstream training set.

The Cosmos-Drive-Dreams pipeline addressed this requirement using a vision-language model as an automated rejection sampler [9]. The quality filter scores each generated sequence against a set of physical plausibility criteria: lane boundary geometry, agent-trajectory consistency, sensor-to-scene correspondence, and annotation completeness. Sequences that fall below threshold on any criterion are discarded without human review. This filtering stage is distinct from the annotation validation stage described in Principle 7; it operates on the generated output rather than on the input conditioning data.

Deployment of a vision-language quality filter requires careful calibration of the rejection threshold. An overly aggressive threshold discards physically plausible sequences that differ from the training distribution in desirable ways, precisely the rare-event scenarios that justify using a world model in the first place. An insufficiently aggressive threshold passes physically implausible sequences that harm downstream policy performance. The calibration process must be performed on a held-out validation set with known ground-truth quality labels, and the threshold must be re-evaluated whenever the generative model is updated [9].

6. Principle 5 — closed-loop evaluation over open-loop surrogates

Open-loop evaluation, in which model predictions are compared against fixed ground-truth trajectories without feedback from the model's own outputs, has been the dominant evaluation paradigm for autonomous driving perception and prediction systems [14], [17]. For world model evaluation, it is demonstrably inadequate. A model evaluated in open loop accumulates displacement error at each time step without the causal feedback that would allow it to correct its trajectory. The result is a systematic underestimate of behavioral errors that only manifest when the model is integrated into a policy training loop.

The distinction between open-loop and closed-loop evaluation is structural, not merely a matter of metric selection. In closed-loop evaluation, the world model receives the agent's actual action at each time step and generates the next scene observation conditioned on that action [13]. Errors made by the agent propagate into the world model's input and affect all subsequent observations. This causal coupling is the defining property of a simulation environment suitable for safety-critical policy training, and Yang et al. demonstrated it systematically in the UniSim closed-loop framework [13].

ACT-Bench provides a standardized closed-loop evaluation protocol that measures action controllability — the degree to which a world model faithfully executes commanded actions — rather than trajectory similarity to a reference policy [19]. Systems that score well on ACT-Bench are those whose closed-loop causal dynamics are sufficiently realistic to enable policies trained in simulation to transfer behavioral patterns to real-world performance. This direct measure of action fidelity is the appropriate complement to perceptual quality scores in any production integration evaluation.

7. Principle 6 —safety-curriculum construction with counterfactual scenarios

A world model that generates only statistically typical driving scenarios provides limited value for safety validation. The scenarios most relevant to safety assessment, those involving rare agent behaviors, unusual environmental conditions, or edge-case roadway configurations, are precisely those most underrepresented in real-world driving logs [1], [10], [20]. Safety-curriculum construction is the principled approach to synthesizing these scenarios at the scale required for policy hardening.

Tan et al. addressed city-scale safety curriculum generation in SceneDiffuser++ through a joint generative world model that synthesizes the trajectories of all agents in a scene simultaneously, ensuring that the resulting traffic configurations are physically and behaviorally coherent at the scenario level rather than the agent level [11]. Its CitySim task evaluation demonstrated capability to simulate full driving journeys from point A to point B rather than isolated interaction clips, producing a scenario distribution that covers the

long tail of driving situations inadequately represented in logged data [11]. The behavioral plausibility of synthesized agent interactions was verified against real-world traffic flow statistics.

Reinforcement learning-based approaches offer an additional and complementary route to safety-critical curriculum construction. Yan et al. identify a fundamental optimistic bias in world models trained exclusively on safe expert data: when conditioned on an unsafe trajectory, such models hallucinate an implausibly safe future rather than forecasting the impending failure [12]. AD-R1 addresses this through Counterfactual Synthesis, a data synthesis pipeline that systematically generates a curriculum of plausible collision and off-road events, transforming the world model from a passive scene completer into one that remains faithful to the causal link between actions and outcomes [12].

Before scaling to city-level or adversarial curriculum generation, it is advisable to validate closed-loop integration on a small representative scenario set to confirm that the world model's causal dynamics are suitable for policy training [13].

Expanding data generation volume before confirming the causal coupling between world model outputs and agent actions risks scaling a fundamentally miscalibrated training loop, a failure mode that produces large synthetic datasets of limited policy utility.

8. Principle 7 — upstream annotation validation

The conditioning quality of a world model is bounded by the quality of the annotations used to construct its structured conditioning inputs. High-definition map errors, mislabeled agent bounding boxes, and incorrect ego-trajectory labels propagate directly into the synthesized output: a world model conditioned on a structurally correct but semantically mislabeled scene will generate a structurally consistent but physically implausible synthetic sequence [9].

Upstream annotation validation is therefore a prerequisite for any deployment that relies on structured conditioning.

The Cosmos-Drive-Dreams pipeline deploys annotation generation and quality assessment as a pre-generation filter [9]. Cosmos-Drive annotation models generate bounding boxes, verify label consistency across camera views, assess trajectory plausibility given the surrounding road geometry, and check map-to-image alignment. This pre-generation annotation filter reduces annotation-induced geometric inconsistencies in the resulting synthetic dataset, establishing a measurable quality floor for the conditioning inputs before generation begins [9].

Annotation validation must be treated as a continuous process rather than a one-time data audit. As the real-world fleet accumulates new driving logs in novel environments and conditions, the annotation pipeline that processes those logs for use as world model conditioning input must maintain consistent quality standards. Periodic re-calibration of the annotation validator against human-reviewed ground truth is necessary to detect and correct systematic annotation errors before they propagate into the world model training set.

9. Integration checklist and common failure modes

The seven principles above reduce to a set of concrete engineering criteria that can be applied to any world model integration project. The checklist below provides pass/fail gates structured around the sequence in which architectural decisions are typically made: from conditioning architecture through evaluation protocol to data pipeline quality assurance.

Three failure modes recur most frequently in integration projects that do not apply these checks. Causal decoupling occurs when a world model generates future scenes without incorporating the agent's most recent action into the scene state update [12], [13].

The model produces visually coherent sequences, but the policy that trained on them learns to ignore the causal relationship between its outputs and scene evolution, producing plans that are well-calibrated to the synthesized environment but brittle in the real world. Cross-modal contamination occurs when camera, LiDAR, and radar outputs are generated from independent models without a shared scene representation [9].

Geometric inconsistencies between modalities introduce systematic errors that the downstream perception stack interprets as calibration drift, degrading detection accuracy on real-world data. Benchmark overfitting occurs when the world model is optimized for perceptual quality metrics on a specific held-out dataset, producing a generative model that achieves strong Fréchet Inception Distance scores but fails to generalize to novel road geometries or unusual weather conditions encountered in fleet deployment [5], [6].

Table 1. *World Model Architecture Families: Approximate Deployment Trade-offs*

Architecture Family	Spatial Consistency	Multi-Modal	RL Suitability	Inferenc e Cost	Sim Horizon
Recurrent latent	Low	Limited	High	Low	2–5 s
Video diffusion	High	Partial	Moderate	Very High	Not specified
Occupancy-centric	High	High	High	Moderate	Not specified
Gaussian splatting hybrid	Very High	High	High	Low-Moderat e	Not specified

[Note: horizon and consistency ratings are indicative summaries derived from cited papers; precise figures vary by implementation.]

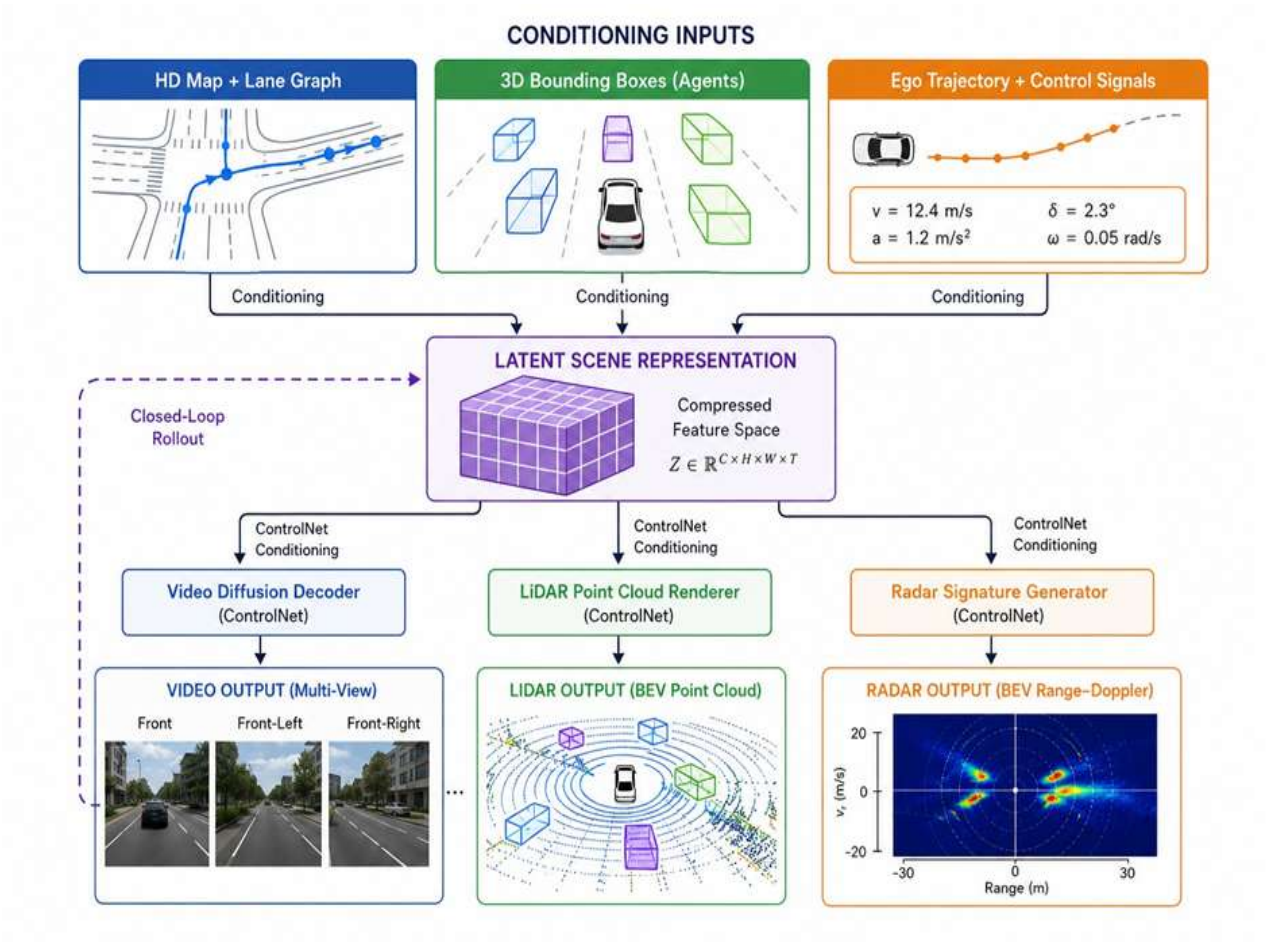


Figure. 1*: Multi-modal world model architecture — structured scene conditioning inputs flow through a shared latent scene representation, with modality-specific decoding branches producing sensor-consistent camera, LiDAR, and radar outputs for closed-loop rollout.*

Table 2. *Action Fidelity vs. Perceptual Quality: Five Representative Systems*

System	Perceptual Quality Metric	Action Fidelity	CL Supported	Primary Application	Ref.
DriveDrea mer4D	FID: 32.1% / 46.4% / 16.3% rel. improvement (vs. PVG / S3G / Def-GS); NTA-IoU: 22.6% / 43.5% / 15.6%	Moderate	Partial	Novel viewpoint synthesis	[6]
Gaussian World	>2% mIoU gain (nuScenes)	High	Yes	RL policy training	[2]

UniScene	Competitive (occupancy/video/LiDAR)	High	Yes	E2E planning	[3]
ReconDreamer	NTA-IoU: +24.87%; NTL-IoU: +6.72%; FID: +29.97% (vs. Street Gaussians); NTA-IoU: +195.87% (vs. DD4D+PVG, large maneuver)	High	Yes	Novel trajectory eval	[5]
SceneDiffuser++	N/A (trajectory/behavior focus)	Very High	Yes	City-scale simulation	[11]

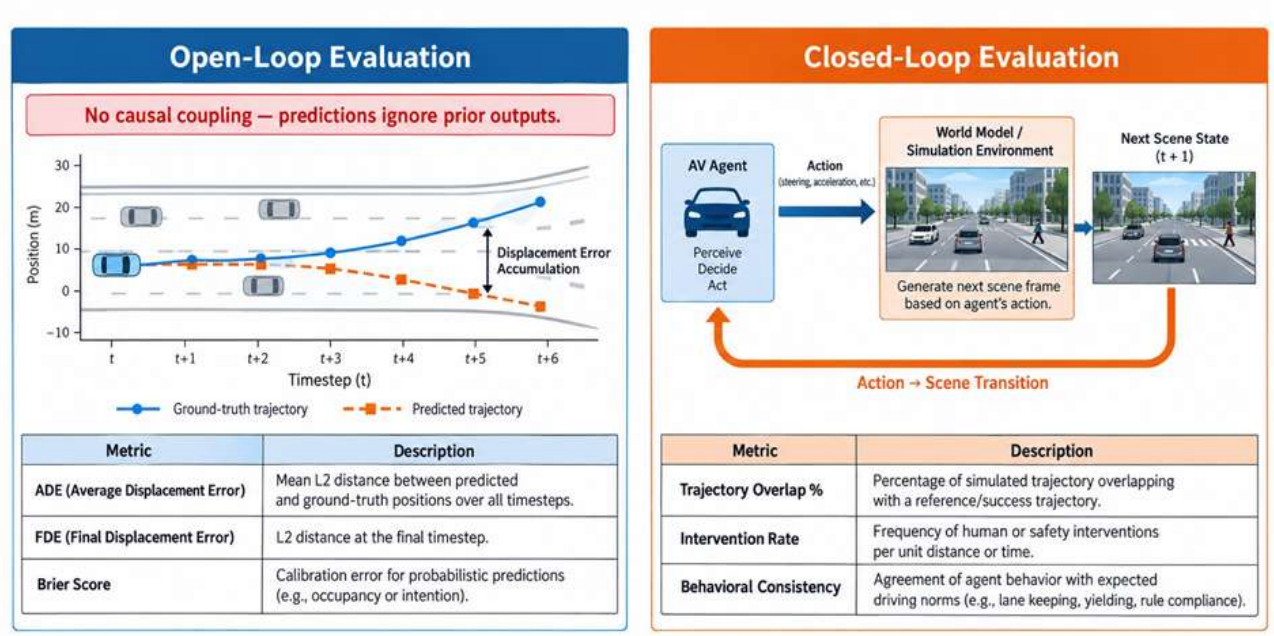


***Figure 2**: Comparison of open-loop and closed-loop evaluation paradigms — open-loop metrics accumulate displacement error without causal coupling, while closed-loop evaluation feeds agent decisions back into world model state transitions, enabling behaviorally valid safety assessment.*

***Table 3.** Integration Checklist: Seven Principles to Engineering Gates*

#	Principle	Pass Criterion	Failure Indicator	Ref.
1	Structured conditioning	Synthesized scenes match engineer-specified agent positions and lane geometry within acceptable tolerance	Agent positions or lane geometry drift measurably from conditioning input	[6], [9]
2	Multi-modal consistency	LiDAR returns fall within volume defined by camera-projected bounding box geometry; cross-view projections are pixel-accurate	Cross-modal geometric misalignment detected under standard sensor geometry	[3], [9], [13]
3	Novel-trajectory extrapolation	NTA-IoU on out-of-distribution trajectories remains within acceptable margin of in-distribution baseline; large-maneuver (≥6m lane shift) rendering does not degrade to artifact-dominated output	Scene geometry collapse or NTA-IoU degradation on curvature or lane deviation beyond training distribution	[5], [6]

#	Principle	Pass Criterion	Failure Indicator	Ref.
4	Automated quality filtering	Vision-language filter calibrated on labeled hold-out set; threshold validated and documented before pipeline deployment	Physically implausible sequences detected in training pipeline after filtering step	[9]
5	Closed-loop evaluation	World model passes action controllability evaluation confirming commanded actions produce corresponding scene state changes	Open-loop perceptual quality improves while closed-loop policy performance regresses	[13], [19]
6	Safety curriculum	Synthesized rare events cover the relevant known failure taxonomy; policy performance on held-out edge cases does not regress after curriculum training	Policy regression on unseen scenarios following world model curriculum training	[11], [12]
7	Annotation validation	Annotation generation and quality gate deployed as pre-generation step; calibration maintained against human-reviewed ground truth	Annotation errors in conditioning inputs detected post-generation	[9]

11. Discussion

The seven principles developed above converge on a common architectural requirement: shared latent scene representation. Every failure mode documented in the literature traces, in some form, to a breakdown in the coupling between what the world model knows about the current scene and what it generates as the next sensor observation [9], [13]. Structured conditioning, multi-modal consistency, novel-trajectory stability, and closed-loop causal coupling are all properties of the latent scene representation, not of the individual sensor decoders. This convergence suggests that the most productive direction for future world model research is not more powerful rendering pipelines but more expressive and geometrically faithful scene encoders [2], [3].

The computational cost gap between diffusion-based world models and traditional simulation rasterizers remains the primary deployment barrier for production applications [15]. Approaches that decouple the scene representation from the pixel rendering pipeline, occupancy-centric and Gaussian splatting hybrid systems in particular, show the most promise for bridging this gap within a training-compute budget that is compatible with continuous fleet learning [2], [3], [5]. The emergence of large-parameter foundation models fine-tuned specifically for autonomous driving annotation and multi-view synthesis suggests that the industry is converging on a layered deployment architecture: a large foundation model for scene representation and a suite of task-specific modules for conditioning, generation, and evaluation [8], [9]. Regulatory and safety certification implications of world model-based training data are not yet fully resolved. Standards bodies have begun to address synthetic data in the context of autonomous driving safety cases, but specific guidance on the quality criteria that world model-generated data must meet to contribute to a safety argument remains nascent. The integration checklist developed in Section X provides a technical framework that is compatible with emerging safety case requirements, but formal certification guidance will require additional work from the research community and standards bodies together.

12. Conclusion

The seven principles presented in this article translate a complex and rapidly evolving research landscape into a coherent set of engineering decisions for teams deploying world models in production autonomous driving pipelines. The analysis confirms that perceptual quality metrics are insufficient predictors of operational utility: closed-loop causal coupling, multi-modal geometric consistency, and structured scene controllability are the properties that determine whether a world model improves fleet safety or merely generates visually impressive simulation outputs.

Structured conditioning anchors scenario generation to engineer specifications [6], [9]. Multi-modal consistency ensures that synthetic sensor streams share the geometric relationships required for downstream perception and planning [3], [9], [13]. Novel-trajectory evaluation exposes geometric failures that in-distribution benchmarks conceal, as demonstrated by NTA-IoU improvements of 24.87% over Street Gaussians and 195.87% over DriveDreamer4D with PVG during large-maneuver rendering [5]. Automated quality filtering provides a scalable gate against physically implausible outputs [9]. Closed-loop evaluation with action

controllability benchmarks provides a direct measure of synthetic data utility for policy training [13], [19]. Safety-curriculum construction using city-scale generative world models and counterfactual synthesis fills the rare-event gaps that logged data cannot provide [11], [12]. Upstream annotation validation prevents conditioning errors from propagating into synthesized outputs at scale [9].

Together these principles constitute an engineering discipline for world model deployment that complements and partially supersedes the research discipline of world model design. Future work should focus on three open challenges: standardized closed-loop benchmarks that enable cross-architecture comparison on safety-relevant scenario sets, formal certification frameworks that incorporate world model-generated synthetic data into safety cases for autonomous driving systems, and efficient scene representation architectures that eliminate the computational gap between diffusion-quality synthesis and real-time simulation requirements.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 11621–11631. [Online]. Available: https://www.nuscenes.org/

[2] S. Zuo, W. Zheng, Y. Huang, J. Zhou, and J. Lu, "GaussianWorld: Gaussian world model for streaming 3D occupancy prediction," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 6772–6781. [Online]. Available: https://arxiv.org/abs/2412.10373

[3] B. Li, J. Guo, H. Liu, Y. Zou, Y. Ding, X. Chen, H. Zhu, F. Tan, C. Zhang, T. Wang, et al., "UniScene: Unified occupancy-centric driving scene generation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 11971–11981. [Online]. Available: https://arxiv.org/abs/2412.05435

[4] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao, and X. Wang, "DriveDreamer-2:LLM-enhanced world models for diverse driving video generation," in Proc. AAAI Conf. Artif.Intell., vol. 39, 2025, pp. 10412–10420. [Online]. Available: https://arxiv.org/abs/2403.06845

[5] C. Ni, G. Zhao, X. Wang, Z. Zhu, W. Qin, G. Huang, C. Liu, Y. Chen, Y. Wang, X. Zhang, Y. Zhan, K. Zhan, P. Jia, X. Lang, X. Wang, and W. Mei, "ReconDreamer: Crafting world models for driving scene reconstruction via online restoration," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 1559–1569. [Online]. Available: https://arxiv.org/abs/2411.19548

[6] G. Zhao, C. Ni, X. Wang, Z. Zhu, X. Zhang, Y. Wang, G. Huang, X. Chen, B. Wang, Y. Zhang, W. Mei, and X. Wang, "DriveDreamer4D: World models are effective data machines for 4D driving scene representation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025,pp. 12015–12026. [Online]. Available: https://arxiv.org/abs/2410.13571

[7] W. Zhao, J. Zheng, X. He, T. Zhou, and C. C. Loy, "DriveDreamer: Towards real-world-driven world models for autonomous driving," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2024. [Online].Available: https://arxiv.org/abs/2309.09777

[8] NVIDIA Research, "Cosmos: World foundation models for physical AI development," 2025. [Online]. Available: https://research.nvidia.com/labs/toronto-ai/cosmos-av/

[9] X. Ren, Y. Lu, T. Cao, R. Gao, S. Huang, A. Sabour, T. Shen, T. Pfaff, J. Z. Wu, R. Chen, S. W. Kim, J. Gao, L. Leal-Taixé, M. Chen, S. Fidler, and H. Ling, "Cosmos-Drive-Dreams: Scalable synthetic driving data generation with world foundation models," arXiv preprint arXiv:2506.09042, 2025. [Online]. Available: https://arxiv.org/abs/2506.09042

[10] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai,B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in

Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 2446–2454. [Online]. Available: https://waymo.com/open/

[11] S. Tan, J. Lambert, H. Jeon, S. Kulshrestha, Y. Bai, J. Luo, D. Anguelov, M. Tan, and C. M. Jiang, "SceneDiffuser++: City-scale traffic simulation via a generative world model," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 1570–1580. [Online].Available: https://arxiv.org/abs/2506.21976

[12] T. Yan, T. Tang, X. Gui, Y. Li, J. Zheng, W. Huang, L. Kong, W. Han, X. Zhou, X. Zhang, et al., "AD-R1: Closed-loop reinforcement learning for end-to-end autonomous driving with impartial world models," arXiv preprint arXiv:2511.20325, 2025. [Online]. Available: https://arxiv.org/abs/2511.20325

[13] Z. Yang, Y. Chen, J. Wang, S. Manivasagam, W.-C. Ma, A. J. Yang, and R. Urtasun, "UniSim: A neural closed-loop sensor simulator," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 1389–1399. [Online]. Available: https://arxiv.org/abs/2308.01898

[14] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "GPT-Driver: Learning to drive with GPT," arXiv preprint arXiv:2310.01415, 2023. [Online]. Available: https://arxiv.org/abs/2310.01415

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. Advances in Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 6840–6851.

[16] A. Hu, L. Russell, H. Hudson, J. Ziegler, S. Banerjee, S. James, M. Bertram, B. Mitchinson, C. Sherlock, A. Bhatt, K. Manakul, A. Cass, J. Sherlock, V. Dobrescu, S. Leask, W. Jones, M. Hind,J. Hawke, and A. Sherlock, "GAIA-1: A generative world model for autonomous driving," arXiv preprint arXiv:2309.17080, 2023. [Online]. Available: https://arxiv.org/abs/2309.17080

[17] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia,
Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 17853–17862.

[18] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in Proc. Int. Conf. Learn. Represent. (ICLR), 2020. [Online]. Available: https://arxiv.org/abs/1912.01603

[19] Hidehisa Arai et al., “ACT-BENCH: Towards Action Controllable World Models for Autonomous Driving,” [Online]. URL : https://www.alphaxiv.org/pdf/2412.05337

[20] S. Suo, S. Regalado, S. Casas, and R. Urtasun, "TrafficSim: Learning to simulate realisticmulti-agent behaviors," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021,pp. 10973–10982.