Validate the Dream Before You Trust Its Verdict: Admissibility for World-Model Simulators

Christian Oefinger[1], Finn Rasmus Schäfer[1], Korbinian Moller[1], Mattia Piccinini[1], and Johannes Betz[1]
[1]Autonomous Vehicle Systems Lab
Technical University of Munich, Garching b. München, Germany
Email: christian.oefinger@tum.de

***Abstract*—Across robotics, World Models (WMs) are increasingly used to evaluate action policies by simulating the consequences of actions in an imagined world, and returning a success or safety verdict. Yet a verdict is only as trustworthy as the WM that produced it, and the WM itself needs to be certified. In video-generation WMs, fidelity metrics such as Fréchet Video Distance (FVD) reward visual realism, but ignore whether the world responds correctly to the policy's actions, including those unseen in training. Classical simulation-based validation assumes a trusted simulator evaluating an untrusted policy, whereas generative WMs are themselves unverified learned artifacts. Hence, we argue that any WM used as a test oracle must first be accredited before its verdicts can serve as evidence. Building on credibility practices from safety-critical simulation, including Verification, Validation & Accreditation (VV&A), Safety of the Intended Functionality (SOTIF), and scenario-based testing standards, we define an admissibility ladder (L0–L4) that a WM must climb before its closed-loop verdicts are accepted as assurance evidence. Our framework is embodiment-agnostic, and is instantiated in autonomous driving (AD), where assurance methods for traditional simulation are most mature. Applied to two driving WMs, the lower rungs reveal a reversal: the model that ranks higher on visual generation quality (L0) ranks lower on action-following (L1–L2), so visual fidelity does not predict the action-robustness a closed-loop verdict depends on.**

I. Introduction

WMs are generative models that learn an internal representation of an environment's dynamics, letting them predict how the world responds to an agent's actions [15, 16]. Recent video-generation models turn this prediction into high-fidelity, controllable simulation [12, 18], with the potential to reshape how robotic systems are developed and tested. Across robotics, the role of WMs is shifting from *imagining* plausible futures to serving as closed-loop simulators that *test* the action policies that act within them [17]: A WM rolls out the consequences of a policy in an imagined world, and returns a verdict on success or safety. Thus, WMs are used as *test oracles* in several robotic embodiments. WMs serve for instance as (i) a testbed for *manipulation* policies [31], (ii) a benchmark for world-model planning on *legged robots* [39], or (iii) a generative simulator for closed-loop evaluation in *autonomous driving* [43]. However, the reliability of WMs is typically left unverified, treating their verdicts as *evidence* of real-world behavior, without any basis to assess admissibility.

Trusting the verdict assumes a high-fidelity imagined world can substitute for reality. For video-generation WMs, which this paper targets rather than reconstruction-based simulators,

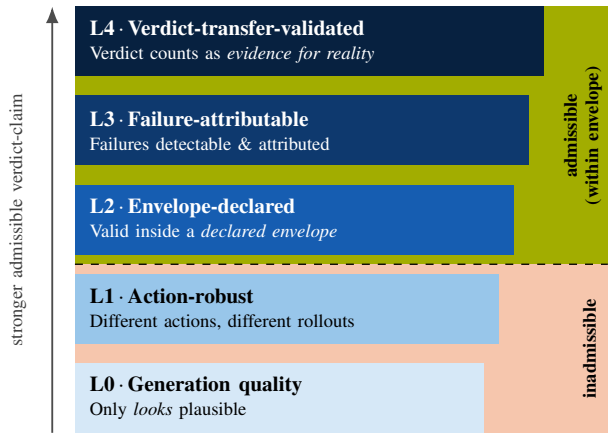

Fig. 1. **The levels-of-admissibility ladder (L0–L4).** A generative WM used as a closed-loop test oracle earns the right to have its verdict counted as assurance evidence only by climbing the ladder. Verdict *validity* first appears at L2. L0–L1 support no admissibility claim, and all guarantees hold only within the declared operating envelope. Table I lists the evidence required at each level.

visual quality alone does not determine task success [9, 45]. Distribution-level video scores, such as the FVD [38], reward perceptual and temporal realism while ignoring physical plausibility and the world's realistic reactions to the policy's actions. This disconnect has been demonstrated across embodiments. For legged robots, the best video WMs reach only 0.341 (of a maximum of 1.0) on Target-Bench's overall path-planning score [39] despite generating visually convincing footage, and an analogous gap holds in AD, where open-loop trajectory metrics only partly track closed-loop outcomes [8, 41]. Hence, a world can look right and still judge wrong.

The reason lies in what a verdict implicitly trusts. Classical simulation-based safety validation relies on trusted simulators, engineered and validated against physics [37], and produces verdicts about action policies under test [7]. This relationship is inverted for a generative WM. The simulator, trusted by construction in the classical setting, becomes the component whose reliability is in question, since it is learned and has no ground-truth physics to validate against. Action policies tested within WMs still require validation. What is new is not that the tested policy needs validation, but that the world judging it must now be validated too, before its verdict can count. Unlike classical simulators, existing WMs carry neither a measurable error nor a certification.

This vulnerability is currently unaddressed. The established methodologies for licensing safety-critical simulators, VV&A [6, 32], SOTIF [21], and scenario-based testing [20], were all built for classical, physics-based simulators and presuppose an independent ground truth against which the simulator is validated. None has been adapted to a generative WM that is itself the unverified artifact, leaving no principled criterion for when its verdicts may be trusted.

This paper proposes that missing criterion. We adopt simulation VV&A, SOTIF, and scenario-based testing [6, 20, 21, 22, 32] methods to form a new *admissibility standard*: criteria, organized as a levels-of-admissibility ladder (L0–L4), that a generative WM must meet before its closed-loop verdicts are admissible as evidence (Figure 1).

The standard is embodiment-agnostic and instantiated in AD, where the assurance machinery is most mature and provides the richest source of adaptation. This paper makes four contributions. (i) We characterize the *trust inversion* and make precise why a generative WM, unlike a classical simulator, is itself the unverified artifact under test, and therefore why existing simulator-validation methods do not license its verdicts. (ii) We formalize the *action-coverage gap* as an off-policy evaluation problem, identifying the actions outside the training behavior policy as the regime where a WM's verdict is unsupported. (iii) We construct an *admissibility standard*: a five-level ladder (L0–L4) that, by adapting simulation VV&A, SOTIF, and scenario-based testing to the generative setting, specifies the evidence a WM must supply at each level before its closed-loop verdicts count as assurance evidence. (iv) We instantiate the lower rungs on two driving WMs (Vista, Epona), measuring L0, L1, and the L2 horizon, and demonstrate empirically that generation quality and action-robustness decouple (Appendix A).

II. Related Work

The use of learned WMs to test policies, rather than to plan an agent's actions within them, is a recent but rapidly growing practice. In general robotics, WorldGym shows that in-simulation policy rankings can correlate with real-world outcomes [31], and World-in-World proposes task success inside imagined rollouts as a general-purpose benchmark signal [45]. In AD, DriveArena and Vista already use generative rollouts to score driving behavior [12, 43]. The practice of treating such verdicts as evidence is spreading faster than the principled criteria for trusting them.

Several recent works argue that visual-realism metrics are insufficient to characterize a WM's action-conditioned predictions, pointing to three independent lines of evidence: (i) Chen and Zhu identify visual conflation and dynamical hallucinations as failure modes that generation scores cannot capture [9]. (ii) Action-controllability benchmarks show that models with high visual fidelity can still fail to respond correctly to ego-action inputs [4]. (iii) A recent survey frames the gap between generation-quality metrics and closed-loop planning as a central evaluation problem [44], because fidelity measures the plausibility of the output distribution rather than whether the world reacts correctly to the policy's chosen actions. We take this growing body of work as a premise rather than a contribution. Our aim is to specify what must be demonstrated *instead*.

Simulator trustworthiness is addressed in safety-critical engineering through VV&A, which is a structured process for establishing whether a simulator's outputs are sufficiently representative for their intended use [6]. In robotics, the reality gap between simulated and real-world dynamics has been formalized and surveyed across robotics domains [1]. In AD, credibility assessment defines structured evidence requirements for scenario-based virtual testing [37], verification, validation, and uncertainty quantification provide structured procedures for bounding predictive uncertainty [32], and simulator fidelity has been shown to shape which failures a safety-validation process uncovers [5]. The most recent standards-body treatment flags WMs as emerging simulation technology yet defines no admissibility criteria for their verdicts [19]. Every existing procedure assumes a classical, physics-based simulator. None has been applied to a generative WM.

The work most closely premised on our concern is Geng et al. [14], who verify end-to-end controllers against a *deterministic* WM via reachability analysis. They share the same premise (a WM's outputs must be validated before verdicts count) but take a complementary scope, deterministic and reachability-bounded rather than generative and VV&A-grounded. Chen and Zhu [9] argue that physical grounding is a necessary property of useful simulators. We extend this by specifying when grounding is sufficient for a verdict to count as assurance evidence. Neither line of work addresses when a generative WM's closed-loop verdict is admissible as assurance evidence.

III. An Admissibility Standard

A. The Credibility Gap

Like a generative WM, a classical simulator can roll out the consequences of a policy. Its verdict is credible because its components, such as sensor models or scenario logic, are explicitly engineered and can be validated against physics or recorded data [37]. Additionally, structured procedures such as VV&A [6], credibility assessment for scenario-based virtual testing [37], and verification, validation, and uncertainty quantification [32] provide principled means of assessing whether their outputs are sufficiently representative for their intended use. A sim-to-real gap remains, but it is *measurable*: it can be quantified against real-world reference data [1], bounded, and reported. This measurability is the evidence for the verdict's admissibility. A generative WM bypasses this traditional pipeline by learning the simulator directly from data. Its sim-to-real gap is immeasurable, since it synthesizes futures for which no recording exists to validate against. This is the trust inversion of Section I.

A subtler reason a WM may mislead its own verdict lies in how it is trained. Most generative world models learn from logged interaction data, where actions were chosen by the data-collecting behavior policy, such as a human

driver, rather than by diverse interventions spanning the action space [12, 18]. This yields a model of the observational conditional $p(o' \mid o, a)$, with o and o' the current and next observations and a an action, learned under the behavior policy's action distribution. This model is reliable only on the (o, a) pairs the behavior policy actually exercised. A policy under test queries it at the actions *it* selects, including those the behavior policy rarely took or that fall outside the training distribution, where the estimate is subject to *extrapolation error* [11] and therefore unvalidated. This is the off-policy evaluation problem [30], which we call the *action-coverage gap*: the estimates hold only where the target and behavior policies overlap. Recent evidence confirms it is real. Action-conditioned reliability depends on training-data diversity [47] and the majority of policy actions are underrepresented in action-labeled training data [26].

This problem is not equally severe for all neural simulators. Reconstruction-based simulators that fit a scene representation such as NeRF [29] or 3DGS [23] to a recorded capture reduce the sim-to-real gap largely to appearance fidelity, which is partially measurable against the source recording [25]. The trust problem is qualitatively different for *fully generative* WMs that synthesize novel futures from a learned prior. Their dynamics oracle is the unverified artifact, with no recording to validate against. This paper targets that class for which no accreditation procedure exists yet. What remains is to specify what such a WM must demonstrate to earn the trust a classical simulator holds by construction.

B. Adapting Existing Certification Frameworks

Classical accreditation procedures presuppose a physics-based simulator and therefore do not apply directly to a learned one. Two of their underlying principles, however, are independent of that assumption, and we adapt them to the generative WM setting to derive criteria for when a verdict may be admitted as evidence. The resulting standard is prescriptive rather than empirical. It specifies what a WM must demonstrate and which evidence it must provide.

The first principle is fidelity-sufficiency. VV&A certifies a simulator not by maximizing realism but by establishing that its fidelity is sufficient for a declared intended use [6]. When the intended use is policy evaluation, the link between fidelity and verdict is consequential. Below a sufficient fidelity, the simulator's outcomes cease to be representative of real-world behavior, so what fails is the evaluation itself, not merely its numerical accuracy [33]. Adapting this principle to a generative WM changes *which* fidelity must be certified. A WM is trained and scored on visual and distributional realism, quantities that are marginal over actions and therefore say nothing about how the world reacts to any particular one [26, 45]. A closed-loop verdict depends entirely on those reactions. We therefore take the object of certification to be the model's *action-conditioned* fidelity: whether, for the actions the policy under test actually selects, the imagined world evolves as it would in the real environment. Since the WM offers no single real-world reference against which this fidelity can be validated at once (Section III-A), sufficiency cannot be certified in one step. It must be established as a sequence of progressively stronger, separately verifiable claims, which the ladder of Section III-C makes explicit.

The second principle is the explicit declaration of operating limits. SOTIF and scenario-based testing standards require a system to bound the conditions under which its behavior is claimed [22], and to treat conditions beyond that boundary as a recognized source of risk [21]. Adapting this principle changes what defines the boundary. A classical system declares an operating domain specified by design. A generative WM has none, because the region in which its action-conditioned fidelity can hold is fixed by what it was trained on, making the boundary statistical rather than specified. Establishing where that boundary lies, declaring it, and detecting departures from it is the role of the ladder's envelope level (Section III-C), where we make the requirement precise.

C. The Admissibility Ladder

Each rung of the ladder (Figure 1) licenses a stronger verdict on the policy under test and repurposes an existing diagnostic instrument as an admissibility gate. No rung may be reached without satisfying the one below. Table I lists the evidence required at each level.

At **L0 (Generation quality)**, the simulator is assessed solely on visual and temporal fidelity, typically via Fréchet Video Distance (FVD) [38]. L0 is a minimum-viability gate. A model that cannot render a coherent, temporally stable world cannot host a policy test at all, which is why L0 forms the foundation of the admissibility ladder. Passing this gate is a binary admission, not a quality ranking, and on its own it supports no admissibility claim. A high fidelity score does not predict the action-following measured at the levels above (Appendix B1 instantiates this rung for two driving WMs). At L0 the simulated world is not required to respond correctly to the policy's actions, the property on which a closed-loop verdict depends. Content-debiased and successor fidelity measures correct FVD's known biases and estimator flaws [13, 27], yet they remain L0 instruments.

L1 (Action-robust) requires that the imagined world reacts to the policy's actions, and semantically different actions produce correspondingly different rollouts. This establishes that the simulator is responsive to the policy under test rather than replaying an action-agnostic future. Action-controllability benchmarks supply the instruments at this level. Arai et al. [4] measure whether a rollout follows a commanded action or trajectory, and action-replay analyses probe whether the response degrades under forced trajectory replay [35]. Clearing L1 certifies *differentiation*, not *validity*. The rollouts vary systematically with the action, but their correspondence to real-world dynamics remains unestablished and is assessed only at higher levels.

L2 (Envelope-declared) adds validity over a bounded region of operation. Within a declared set of conditions, the simulator's reactions can be checked against real-world behavior, for instance, against measured physical dynamics,

TABLE I

REQUIRED EVIDENCE PER ADMISSIBILITY LEVEL (L0–L4). ADMISSIBILITY LEVELS FOR GENERATIVE WORLD-MODEL SIMULATORS USED AS CLOSED-LOOP TEST ORACLES.

Level	Admissibility claim	Required evidence
L0 *Generation quality*	None: the world only looks plausible and need not respond to actions.	Visual and temporal fidelity only [13, 27, 38].
L1 *Action-robust*	The world responds to the policy's actions, and semantically different actions yield correspondingly different rollouts.	Rollouts vary systematically with the commanded action [4, 35].
L2 *Envelope-declared*	The verdict is valid only within a declared training envelope and rollout horizon, with out-of-distribution detection and refusal outside.	Declared training envelope, bounded-drift horizon, and out-of-distribution detection/refusal [3, 10, 20, 21, 24, 28].
L3 *Failure-attributable*	Failures at and beyond the boundary are detectable and attributable to simulator vs. policy. Out-of-envelope verdicts remain inadmissible.	Out-of-distribution failure signatures, plus an attribution protocol separating simulator contributions from policy contributions [36].
L4 *Verdict-transfer-validated*	Within the envelope, good-in-sim predicts good-in-reality and the verdict counts as assurance evidence. Outside, L3 keeps failures transparent.	Measured in-sim $\leftrightarrow$ real correlation within the envelope, with L3 attribution as the out-of-envelope companion [31].

and are required to hold. Outside that region, the simulator's outputs are undefined and do not carry an admissibility claim. Unlike a classical simulator with an engineered operating domain (in automated driving, an operational design domain [22]), a generative WM is bounded by its *training envelope*, the distribution of conditions it was exposed to during training. The relevant boundary is therefore statistical rather than specified by design. L2 requires the simulator to declare this envelope and to detect and refuse inputs that fall outside it, mirroring how scenario-based safety and SOTIF bound a system to a declared operating region and treat its violation as a hazard [21, 22]. L2 additionally requires the simulator to bound the rollout horizon over which its reactions remain valid (Appendix B3 instantiates the horizon component for two driving WMs). The operative mechanism is out-of-distribution detection, because silently crossing the boundary is what renders a verdict inadmissible. Viable instruments include conformal prediction, which is distribution-free, gives coverage guarantees under exchangeability, and can be adapted to flag out-of-distribution inputs [3, 28], and ensemble-rollout disagreement, an epistemic-uncertainty heuristic [10, 24].

L3 (Failure-attributable) concerns behavior at and beyond the envelope boundary. It requires that failures be detectable and attributable. The framework must distinguish failures induced by the simulator from failures of the policy under test. Verdicts outside the declared envelope remain inadmissible, but at L3 the failures that occur there are transparent rather than silent. The required evidence is a set of out-of-distribution failure signatures together with an attribution protocol that separates simulator contributions from policy contributions at the boundary. The supporting instrument is calibration against real failure data, such as field disengagement and criticality records [36].

L4 (Verdict-transfer-validated) is the highest rung and the only one resting on *external*, empirical rather than internal evidence. Inside the declared envelope, it requires a demonstrated correlation between in-simulation and real-world outcomes, so that good in simulation predicts good in reality, and the verdict counts as assurance evidence. Outside the envelope, L3's attribution capability ensures that failures remain transparent rather than hidden. The required evidence is a measured correlation between in-simulation and real outcomes within the envelope, complemented outside it by L3's failure attribution. The instrument is a real-to-simulation policy-performance correlation, as reported for manipulation by WorldGym [31].

IV. DISCUSSION AND CONCLUSION

As generative WMs are adopted to test policies, the robotics community increasingly needs to agree on when a WM's verdict can be trusted as evidence. The admissibility ladder is our proposal for structuring that decision. Rather than a mandatory standard, the ladder offers a structured vocabulary for that conversation. By naming the levels a WM can demonstrably clear, and the evidence required at each, the ladder makes the implicit assumptions behind in-simulation verdicts explicit and auditable. Concretely, we suggest that developers proposing generative WMs as closed-loop test oracles support their fidelity scores and task metrics with a credibility argument: which rungs does the system clear, and on what evidence?

Three directions remain open. First, the levels are derived from first principles and by analogy with classical simulation validation. Their application to existing generative WMs is still in progress, as is the test of whether the levels (L0–L4) are jointly sufficient, mutually exclusive, and correctly ordered. Appendix A takes a first step in this direction, instantiating L0, L1, and the horizon component of L2 on two driving WMs. Second, embodiment-agnostic L2 protocols that define a training envelope and reliably detect out-of-distribution inputs do not yet exist. Developing and benchmarking such detectors across manipulation, locomotion, and navigation is required to make L2 operational. Third, L3 attribution requires real-world failure data, which is scarce and domain-specific. Building calibrated failure datasets together with an attribution protocol that separates simulator from policy contributions is the concrete prerequisite for L3. We present the ladder as an initial proposal, to be refined as the community applies it across embodiments and tasks.

References

[1] Elie Aljalbout, Jiaxu Xing, Angel Romero, Iretiayo Akinola, Caelan Reed Garrett, Eric Heiden, Abhishek Gupta, Tucker Hermans, Yashraj Narang, Dieter Fox, Davide Scaramuzza, and Fabio Ramos. The Reality Gap in Robotics: Challenges, Solutions, and Best Practices. *Annual Review of Control, Robotics, and Autonomous Systems*, 9, dec 2025. ISSN 2573-5144. doi: 10.1146/annurev-control-031924-100130.

[2] American Association of State Highway and Transportation Officials. *A Policy on Geometric Design of Highways and Streets: The Green Book*. American Association of State Highway and Transportation Officials, Washington, DC, 7th edition, 2018. ISBN 978-1-5231-1958-5.

[3] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, 2021. URL https://arxiv.org/abs/2107.07511.

[4] Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. ACT-Bench: Towards Action Controllable World Models for Autonomous Driving, 2024. URL https://arxiv.org/abs/2412.05337.

[5] Ali Baheri. Exploring the Role of Simulator Fidelity in the Safety Validation of Learning-Enabled Autonomous Systems. *AI Magazine*, 44(4):453–459, 2023. ISSN 2371-9621. doi: 10.1002/aaai.12141.

[6] Osman Balci. Verification Validation and Accreditation of Simulation Models. In *Proceedings of the 29th Conference on Winter Simulation*, WSC '97, pages 135–141, Atlanta, GA, USA, dec 1997. IEEE Computer Society. ISBN 978-0-7803-4278-1. doi: 10.1145/268437.268462. URL https://dl.acm.org/doi/10.1145/268437.268462.

[7] Jared J. Beard and Ali Baheri. Black-Box Safety Validation of Autonomous Systems: A Multi-Fidelity Reinforcement Learning Approach, mar 2022. URL https://arxiv.org/abs/2203.03451.

[8] Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-Simulation for Autonomous Driving, aug 2025. URL https://arxiv.org/abs/2506.04218.

[9] Zhikang Chen and Tingting Zhu. From Generative Engines to Actionable Simulators: The Imperative of Physical Grounding in World Models, 2026. URL https://arxiv.org/abs/2601.15533.

[10] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep Reinforcement Learning in a Handful of Trials Using Probabilistic Dynamics Models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, volume 31, Montréal, Canada, 2018. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3de568f8597b94bda53149c7d7f5958c-Paper.pdf.

[11] Scott Fujimoto, David Meger, and Doina Precup. Off-Policy Deep Reinforcement Learning Without Exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2052–2062, Long Beach, CA, USA, June 2019. PMLR.

[12] 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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 37*, volume 37, Vancouver, BC, Canada, oct 2024. Curran Associates, Inc. doi: 10.48550/arXiv.2405.17398. URL https://papers.nips.cc/paper_files/paper/2024/file/a6a066fb44f2fe0d36cf740c873b8890-Paper-Conference.pdf.

[13] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the Content Bias in Fréchet Video Distance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7277–7288, Seattle, WA, USA, jun 2024. IEEE. doi: 10.1109/CVPR52733.2024.00695. URL https://ieeexplore.ieee.org/document/10655312.

[14] Yuang Geng, Zhuoyang Zhou, Zhongzheng Zhang, Siyuan Pan, Hoang-Dung Tran, and Ivan Ruchkin. Deterministic World Model for Closed-loop Verification of End-to-End Vision-based Controller, dec 2025. URL https://arxiv.org/abs/2512.08991.

[15] David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, volume 31, pages 2450–2462, Red Hook, NY, USA, 2018. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf.

[16] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Control Tasks through World Models. *Nature*, 640(8059):647–653, apr 2025. doi: 10.1038/s41586-025-08744-2. URL https://www.nature.com/articles/s41586-025-08744-2.

[17] Bohan Hou, Gen Li, Jindou Jia, Tuo An, Xinying Guo, Sicong Leng, Haoran Geng, Yanjie Ze, Tatsuya Harada, Philip Torr, Oier Mees, Marc Pollefeys, Zhuang Liu, Jiajun Wu, Pieter Abbeel, Jitendra Malik, Yilun Du, and Jianfei Yang. World Model for Robot Learning: A Comprehensive Survey, apr 2026. URL http://arxiv.org/abs/2605.00080.

[18] 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, 2023. URL https://arxiv.org/

abs/2309.17080.

[19] IEEE SA Autonomous Driving Working Group. Simulation, Testing, Verification, and Validation (STV2) of Autonomous Driving. White Paper, IEEE, New York, NY, USA, oct 2024. URL https://ieeexplore.ieee.org/document/10709902.

[20] International Organization for Standardization (ISO). Road vehicles - Test scenarios for automated driving systems - Scenario based safety evaluation framework, nov 2022.

[21] International Organization for Standardization (ISO). Road Vehicles - Safety of the intended functionality, jun 2022.

[22] International Organization for Standardization (ISO). Road Vehicles - Test scenarios for automated driving systems - Specification for operational design domain, aug 2023.

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (TOG)*, 42(4):139:1–139:14, jul 2023. ISSN 0730-0301. doi: 10.1145/3592433. URL https://dl.acm.org/doi/10.1145/3592433.

[24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, volume 30, Long Beach, CA, USA, 2017. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

[25] Carl Lindström, Georg Hess, Adam Lilja, Maryam Fatemi, Lars Hammarstrand, Christoffer Petersson, and Lennart Svensson. Are NeRFs Ready for Autonomous Driving? Towards Closing the Real-to-Simulation Gap. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4461–4471, Seattle, WA, USA, jun 2024. IEEE. doi: 10.1109/CVPRW63382.2024.00449.

[26] Yuejiang Liu, Fan Feng, Lingjing Kong, Weifeng Lu, Jinzhou Tang, Kun Zhang, Kevin Murphy, Chelsea Finn, and Yilun Du. World Action Verifier: Self-Improving World Models via Forward-Inverse Asymmetry, may 2026. URL https://arxiv.org/abs/2604.01985.

[27] Ge Ya Luo, Gian Mario Favero, Zhi Hao Luo, Alexia Jolicoeur-Martineau, and Christopher Pal. Beyond FVD: Enhanced Evaluation Metrics for Video Generation Quality, oct 2024. URL https://arxiv.org/abs/2410.05203v2.

[28] Rachel Luo, Shengjia Zhao, Jonathan Kuck, Boris Ivanovic, Silvio Savarese, Edward Schmerling, and Marco Pavone. Sample-Efficient Safety Assurances Using Conformal Prediction. *The International Journal of Robotics Research*, 43(9):1409–1424, aug 2024. ISSN 0278-3649. doi: 10.1177/02783649231221580. URL https://doi.org/10.1177/02783649231221580.

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65 (1):99–106, dec 2021. ISSN 0001-0782. doi: 10.1145/3503250. URL https://dl.acm.org/doi/10.1145/3503250.

[30] Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann. URL https://dl.acm.org/doi/10.5555/645529.658134.

[31] Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World Model as An Environment for Policy Evaluation, may 2025. URL https://arxiv.org/abs/2506.00613.

[32] Stefan Riedmaier, Benedikt Danquah, Bernhard Schick, and Frank Diermeyer. Unified Framework and Survey for Model Verification, Validation and Uncertainty Quantification. *Archives of Computational Methods in Engineering*, 28:2655–2688, jun 2021. ISSN 1886-1784. doi: 10.1007/s11831-020-09473-7.

[33] Simon Sagmeister, Panagiotis Kounatidis, Sven Goblirsch, and Markus Lienkamp. Analyzing the Impact of Simulation Fidelity on the Evaluation of Autonomous Driving Motion Control. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 230–237, Jeju Island, Republic of Korea, jun 2024. IEEE. doi: 10.1109/IV55156.2024.10588858. URL https://ieeexplore.ieee.org/document/10588858.

[34] Hiroaki Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, feb 1978. ISSN 0096-3518. doi: 10.1109/TASSP.1978.1163055.

[35] Hunter Schofield, Mohammed Elmahgiubi, Kasra Rezaee, and Jinjun Shan. Beyond Simulation: Benchmarking World Models for Planning and Causality in Autonomous Driving. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1308–1314, Atlanta, GA, USA, may 2025. IEEE. doi: 10.1109/ICRA55743.2025.11127709. URL https://doi.org/10.1109/ICRA55743.2025.11127709.

[36] Marvin Seegert, Christian Oefinger, Korbinian Moller, Christoph Bank, and Johannes Betz. Disengagement Analysis and Field Tests of a Prototypical Open-Source Level 4 Autonomous Driving System, 2026. URL https://arxiv.org/abs/2603.21926.

[37] Christoph Stadler, Francesco Montanari, Wojciech Baron, Christoph Sippl, and Anatoli Djanatliev. A Credibility Assessment Approach for Scenario-Based Virtual Testing of Automated Driving Functions. *IEEE Open Journal of Intelligent Transportation Systems*, 3:45–60, 2022. ISSN 2687-7813. doi: 10.1109/OJITS.2022.3140493.

[38] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges, 2018. URL https://arxiv.org/abs/1812.01717.

[39] Dingrui Wang, Zhihao Liang, Hongyuan Ye, Zhexiao Sun, Zhaowei Lu, Yuchen Zhang, Yuyu Zhao, Yuan Gao, Marvin Seegert, Finn Schäfer, Haotong Qin, Wei Li, Luigi Palmieri, Felix Jahncke, Mattia Piccinini, and Johannes Betz. Target-Bench: Can Video World Models Achieve Mapless Path Planning with Semantic Targets?, 2025. URL https://arxiv.org/abs/2511.17792.

[40] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, Vancouver, BC, Canada, jun 2023. IEEE. doi: 10.1109/CVPR52729.2023.01398.

[41] Yiru Wang, Anqing Jiang, Shuo Wang, Yuwen Heng, Hai Yang, Yang Chen, and Hao Sun. Do Open-Loop Metrics Predict Closed-Loop Driving? A Cross-Benchmark Correlation Study of NAVSIM and Bench2Drive, apr 2026. URL https://arxiv.org/abs/2605.00066.

[42] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. In Joaquin Vanschoren and Serena Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)*, Red Hook, NY, USA, 2021. Curran Associates, Inc.

[43] Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, Liang He, Yong Liu, Botian Shi, and Yu Qiao. DriveArena: A Closed-loop Generative Simulation Platform for Autonomous Driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu, HI, USA, oct 2025. IEEE. doi: 10.1109/ICCV51701.2025.02500. URL https://ieeexplore.ieee.org/document/11446071.

[44] Rongxiang Zeng and Yongqi Dong. Latent World Models for Automated Driving: A Unified Taxonomy, Evaluation Framework, and Open Challenges, mar 2026. URL https://arxiv.org/abs/2603.09086v1.

[45] Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-World: World Models in a Closed-Loop World, oct 2025. URL https://arxiv.org/abs/2510.18135.

[46] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive Diffusion World Model for Autonomous Driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 27220–27230, Honolulu, HI, USA, oct 2025. IEEE. doi: 10.1109/ICCV51701.2025.02527. URL https://ieeexplore.ieee.org/document/11445740.

[47] Zhanguang Zhang, Zhiyuan Li, Behnam Rahmati, Rui Heng Yang, Yintao Ma, Amir Rasouli, Sajjad Pakdamansavoji, Yangzheng Wu, Lingfeng Zhang, Tongtong Cao, Feng Wen, Xinyu Wang, Xingyue Quan, and Yingxue Zhang. Do World Action Models Generalize Better than VLAs? A Robustness Study, apr 2026. URL https://arxiv.org/abs/2603.22078.

[48] Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven L. Waslander. DrivingGen: A Comprehensive Benchmark for Generative Video World Models in Autonomous Driving, mar 2026. URL https://arxiv.org/abs/2601.01528.

APPENDIX

This appendix takes a first step toward applying the admissibility ladder to existing generative WM. We instantiate the lower rungs of the ladder (L0–L2) as concrete, reproducible measurements and apply them to two open video-generation WMs, placing each on the highest rung its measured evidence supports. The intent is illustrative rather than a benchmark: a worked example showing that the proposed levels can be operationalized and that they discriminate between models, not a ranking of systems or a general-purpose evaluation protocol.

A. Experimental Scope

Instantiating the lower rungs in AD requires generative WMs that can be both run and driven. The model must be open-weight, so that its rollouts can be regenerated and scored. Furthermore, it must be action-conditioned, since only a model steered by an external control input can serve as a closed-loop oracle for the policy under test. Two driving WMs meet these criteria: Vista [12] and Epona [46]. Both accept an injected ego trajectory and differ in generation paradigm: Vista builds on clip-level video diffusion, and Epona on autoregressive diffusion that predicts one frame at a time. We evaluate both on the nuScenes-based ACT-Bench split [4], which supplies the action templates and the pretrained action-recognition instrument that the lower rungs rely on.

We use two models, chosen for three reasons. First, they come from successive years (Vista from 2024 and Epona from 2025), so a difference in placement reflects progress between them. Second, Vista has already been evaluated by ACT-Bench [4], which provides an external score ($30.7\,\%$) that we can use to validate our pipeline, whereas Epona has not been evaluated, so its placement is a new result. Third, and most importantly, showing that the rungs are independent rather than redundant requires at least two models whose rankings can disagree.

B. Measurement Protocol

We operationalize each rung as one or more reproducible metrics, each paired with an explicit decision rule that maps a measured value to a placement. In keeping with the ladder, we do not introduce new instruments but repurpose existing ones, and a model is placed at the highest rung whose evidence clears its decision rule. L0 draws on video-distribution metrics. L1 and L2 instead share a single instrument from ACT-Bench [4]: the ACT-Estimator, a pretrained model that infers which maneuver a generated rollout executed and recovers its trajectory. We reuse this estimator rather than build our own recovery method, since ACT-Bench already trains and validates it, which keeps the L1 and L2 measurements independent of our implementation. All metrics are computed on a fixed stratified subsample of the ACT-Bench split: 400 nuScenes clips drawn evenly across ACT-Bench's eight commandable maneuver categories, each paired with its commanded action template. Checkpoints, seeds, and thresholds are pinned for reproducibility.

1) L0: Generation Quality: L0 measures the visual and temporal fidelity of the generated video against real driving footage, the property that existing fidelity metrics target. We report three Fréchet distances between generated and held-out real nuScenes clips. The FVD [38] compares I3D video features and is the field standard. The content-debiased Fréchet Video Distance (CD-FVD) [13] replaces these with VideoMAE-v2 features [40], which respond more to temporal and motion artifacts than to per-frame appearance. The Fréchet Trajectory Distance (FTD) [48] instead compares the distribution of generated ego trajectories, embedded with a pretrained motion encoder, and so captures trajectory realism rather than pixel realism. Because Fréchet distances depend on sample size, all three are computed at a matched number of clips, so that the two models are directly comparable. L0 carries no admissibility claim: any functioning generator clears it, and we report it only as the foil against which the higher rungs are read.

Both models clear L0 (Table II). Vista attains the better pixel- and motion-level scores (FVD 151.3 and CD-FVD 51.6, against 159.4 and 86.1 for Epona), whereas Epona attains the better trajectory-distribution score (FTD 2.59 against 2.72). Visual fidelity therefore already separates from trajectory fidelity within L0, a split we return to in Section C.

TABLE II
MEASURED LADDER METRICS FOR VISTA AND EPONA (L0–L2).
PERFORMANCE COMPARISON OF VISTA AND EPONA ACROSS THE ADMISSIBILITY-LADDER METRICS. h^* IS THE MAXIMUM ADMISSIBLE ROLLOUT HORIZON, THE LARGEST EVALUATED HORIZON AT WHICH ADE STAYS WITHIN THE $1.8\,$M BAND (HALF A $3.6\,$M US LANE [2]; SEE FIG. 4).

Rung	Metric	Vista (2024)	Epona (2025)
L0	FVD ↓	**151.3**	159.4
	CD-FVD ↓	**51.6**	86.1
	FTD ↓	2.72	**2.59**
L1	IEC ↑	0.33	**0.54**
	ADE ↓ (m)	4.56	**2.35**
	DTW ↓	97.3	**41.5**
	Success ↑	0.08	**0.28**
L2	h^* ↑ (s)	1.6	**3.2**
	drift ↓ (m/s)	1.06	**0.53**

2) L1: Action-Robustness: L1 asks whether a rollout follows the action it was commanded, rather than replaying a future that ignores the policy. For each clip, we feed the commanded action template to the model, generate a rollout, and apply the ACT-Estimator to recover the executed maneuver and trajectory. The recovered values are then compared with the command by four metrics. Instruction-Execution Consistency (IEC) [4] is the fraction of rollouts whose executed maneuver matches the commanded one across the eight commandable maneuver categories,[1] where chance is $1/8$. Average Displacement Error (ADE) and Final Displacement

[1]ACT-Bench's estimator classifies nine maneuvers, but only eight are commandable: no action template instructs the ego to remain *stopped*, so it can never appear as a commanded maneuver. IEC is scored over these eight (chance $1/8$); the estimator-reconstruction check below retains all nine to match the estimator's classification task.

Error (FDE) are the mean and final Euclidean distance between the commanded and recovered trajectories [42]. Dynamic Time Warping (DTW) measures the same mismatch under an elastic time alignment, so that a correctly shaped but delayed maneuver is not over-penalized [34]. Success Rate (SR) is the fraction of rollouts whose FDE falls below $3\,\mathrm{m}$, the default threshold of the `success_rate` metric in DrivingGen's [48] reference implementation. This mirrors the complementary miss-rate convention of motion-forecasting benchmarks, which threshold the endpoint error at a fixed distance [42] ($2\,\mathrm{m}$ in Argoverse). We adopt it rather than define our own. A model clears L1 if its IEC lies significantly above chance, which shows the rollout tracks the commanded maneuver rather than ignoring it. Clearing L1 certifies responsiveness, not that the resulting dynamics are physically correct, which the higher rungs address.

We validate the instrument before trusting any generated-video score. On real nuScenes clips the ACT-Estimator recovers trajectories to a mean ADE of $0.77\,\mathrm{m}$, and a reconstruction of ACT-Bench's rule-based maneuver classifier reproduces the estimator's maneuver predictions on $93.4\,\%$ of the clips it labels, close to the $94.0\,\%$ ACT-Bench reports. Scoring Vista's ACT-Bench-released rollouts, our pipeline reproduces the published score (IEC $30.7\,\%$). We then apply the identical scoring to Epona rollouts that we generate with an adapter mapping each template to Epona's action format, as ACT-Bench did not evaluate Epona. Because Epona's action format includes a heading the templates lack, the adapter synthesizes it as the path tangent. On real ego trajectories, this reproduces recorded yaw to a mean of $0.4°$, with the turning categories no worse than the straight ones, so the synthesis does not distort Epona's maneuver results. The same pipeline we validate against Vista's external score thus produces Epona's first reported number.

Both models clear L1 (Table II): their IEC of 0.33 (Vista) and 0.54 (Epona) lies well above the $1/8$ chance level, so each follows the commanded maneuver rather than replaying a fixed future. Epona is the more action-robust on every metric, with a lower ADE ($2.35\,\mathrm{m}$ against $4.56\,\mathrm{m}$). Figure 2 shows the full per-clip distributions. Epona's shifted lower and tighter and correspondingly has lower DTW and higher SR. The per-category breakdown (Figure 3) locates the difference: Epona reaches near-perfect IEC on the curve and constant-speed categories, whereas Vista is weaker and asymmetric between left and right curves. Both models follow the start and stop transitions poorly, a shared limit at the boundary between rest and motion.

3) L2: Action-Following Horizon: L2 asks whether a model's reactions remain valid over a bounded region of operation, and requires it to declare that envelope rather than be trusted everywhere. We instantiate neither part of L2's validity core: we neither check reactions against the measured physical dynamics L2 requires as its real-world reference, nor detect and refuse inputs that fall outside the envelope. What we measure instead is the one component an action-conditioned rollout makes directly observable, the horizon over which action-following stays accurate. Because it reuses the L1 metric, comparing the recovered trajectory against the commanded one rather than against measured dynamics, it bounds the horizon over which L1 correspondence holds, not validity against reality. Reusing the L1 rollouts, we truncate each recovered trajectory to its first h seconds, recompute each clip's ADE, and average over all 400 clips, giving a drift curve $\mathrm{ADE}(h)$ (Figure 4). The admissible horizon h^* is the largest evaluated h at which the mean $\mathrm{ADE}(h)$ stays within $1.8\,\mathrm{m}$, half a $3.6\,\mathrm{m}$ (12 ft) US lane width [2], so that the rollout's ego path remains within its own lane. We read h^* off the sampled horizons without interpolating. Reporting h^* declares the horizon envelope, and a larger h^* is a longer window over which the model's action-following can be trusted.

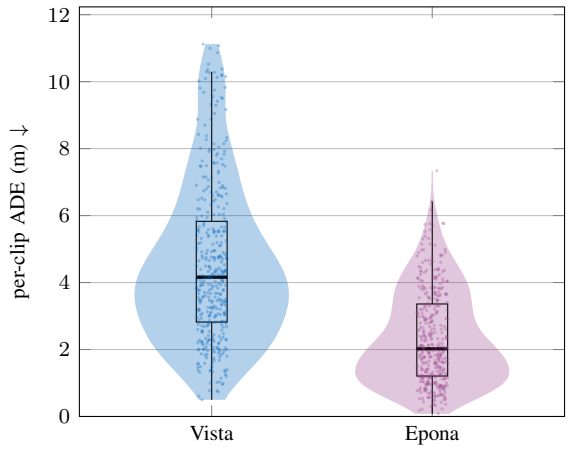

Fig. 2. **Per-clip action-following error (ADE, L1).** Distribution of per-clip ADE (m) for Vista and Epona ($n = 400$ each), shown as violin plots with overlaid box plots (median, interquartile range, and whiskers).

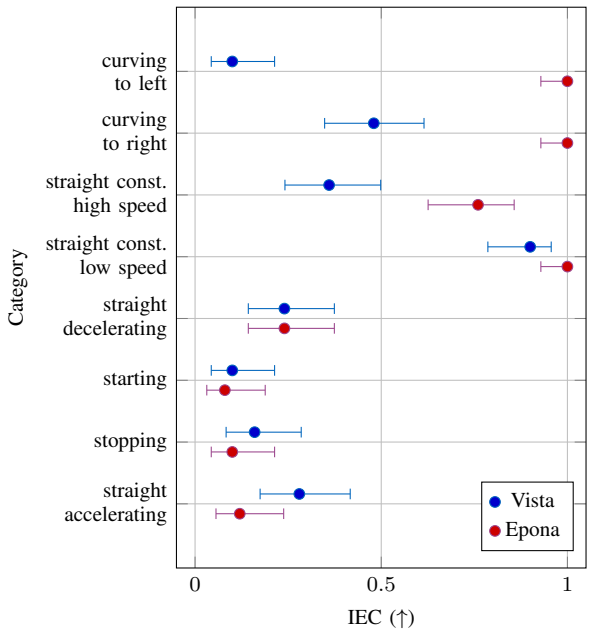

Fig. 3. **Action-following consistency by category (IEC, L1).** Per-category instruction-execution consistency (IEC) for Vista and Epona, with $95\,\%$ Wilson confidence intervals (50 clips per category).

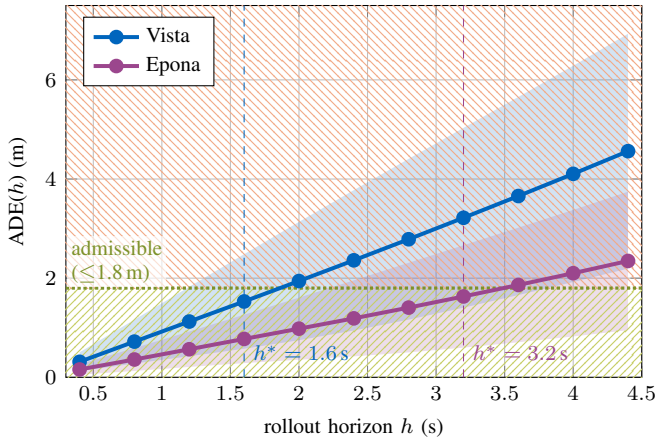

Fig. 4. **Action-following drift over the rollout horizon (ADE, L2).** Average displacement error (ADE) as a function of rollout horizon h for Vista and Epona. The admissible band marks ADE at or below $1.8\,\mathrm{m}$, half a $3.6\,\mathrm{m}$ US lane width [2] (ego path stays within its own lane); dashed lines indicate the maximum admissible horizon h^*, the largest evaluated horizon at which ADE remains within the band. Solid curves show the mean ADE over the 400 clips; shaded regions span ± 1 standard deviation across clips.

Both curves grow roughly linearly but at different rates (Figure 4, Table II). Vista drifts at about $1.06\,\mathrm{m\,s^{-1}}$ and leaves the band at $h^* = 1.6\,\mathrm{s}$, whereas Epona drifts at about $0.53\,\mathrm{m\,s^{-1}}$ and stays admissible until $h^* = 3.2\,\mathrm{s}$, roughly twice as long. The ordering does not depend on the threshold: at a more permissive $3.0\,\mathrm{m}$ band Vista reaches $2.8\,\mathrm{s}$ while Epona never leaves the band over the measured horizon. Epona therefore supports action-following claims over a substantially longer window, the envelope within which its L1 verdicts would hold.

C. Decoupling of Generation Quality and Action-Robustness

Taken one rung at a time, the preceding measurements each report a single ordering of the two models. Read together, they reveal a relation that no single rung shows: the ranking by generation quality is the reverse of the ranking by action-robustness. Vista leads on the L0 visual-fidelity scores (FVD and CD-FVD), whereas Epona leads on every L1 metric and sustains the longer L2 horizon (Table II). Were the two rungs expressions of one underlying quality, the same model would lead on both. The reversal shows that they are not. A credibility argument based on visual fidelity would therefore select Vista, while the action-robustness evidence would select Epona.

For a generative WM used as a test oracle, this reversal is decisive: a model's L0 visual fidelity does not predict its L1 action-following, the property a closed-loop verdict depends on. Visual fidelity, therefore, cannot, on its own, accredit a WM as a test oracle. A single model pair does not establish how often the two properties diverge, but one clear divergence is enough to show that they can, and hence that separating the rungs is necessary rather than redundant.

D. Limitations and Threats to Validity

Several assumptions enter the measurement and bound the strength of these results. All L1 and L2 metrics inherit the ACT-Estimator's error: on real clips it recovers trajectories to a mean ADE of $0.77\,\mathrm{m}$ and classifies maneuvers at about $94\,\%$ accuracy, so displacement differences near that resolution are not meaningful and the reported IEC carries the estimator's misclassification rate as a noise floor. The maneuver categories themselves are defined by ACT-Bench's rule-based classifier, several of whose thresholds are underspecified. The class boundaries are therefore not crisp, which plausibly contributes to the low IEC of both models on the transition maneuvers.

The comparison is not fully symmetric. Vista is scored on the rollouts released with ACT-Bench, whereas Epona is generated through our adapter. The scorer is identical and Vista reproduces its published IEC exactly, but the two models are not conditioned through the same pipeline. Fidelity is reported at a matched sample size, since FVD depends on it. The values support the within-study comparison rather than the absolute numbers reported elsewhere, and all metrics use a 400-clip stratified subsample rather than the full split. The two models also differ in architecture, training data, and scale simultaneously, so although their rankings decouple (Section C), we do not attribute the decoupling to any single cause.

Finally, the instantiation reaches L0, L1, and the horizon component of L2. These are reachable with compute and an existing instrument. However, L2 out-of-distribution detection and the failure-attribution and verdict-transfer evidence of L3 and L4 (Section III) are not relaxed by computation but require data, methods, or a real-world anchor that the model cannot supply. The upper bound of the instantiation is therefore evidentiary rather than computational, and falls within L2.