International Scientific Conference UNITECH 20 26

INTERNATIONAL SCIENTIFIC CONFERENCE
19-21 November 2026
GABROVO

PICS: A NON-COMPENSATORY ASSURANCE PROTOCOL FOR WORLD-MODEL-BASED END-TO-END AUTONOMOUS DRIVING

Jianxun Cui

School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, China; Chongqing Research Institute of HIT, Chongqing, China

**Corresponding author: cuijianxun@hit.edu.cn*

Abstract

Driving world models are increasingly used to forecast scenes, evaluate candidate trajectories, and train end-to-end autonomous-driving policies. Yet visual realism, open-loop accuracy, decision utility, and closed-loop safety are distinct claims that cannot be validated by one score. This paper introduces PICS, a non-compensatory assurance protocol for claim-aligned evaluation. PICS makes three advances. First, it represents an evaluation claim as an auditable contract specifying intended use, scenario distribution, metrics, baselines, thresholds, and failure conditions. Second, it organizes evidence into four gated layers: predictive validity, interventional controllability, consequential utility, and closed-loop safety. Third, it returns pass, fail, or unresolved at each layer and prohibits a strong score at one layer from compensating for missing or failed evidence at another. Mapping CARLA, Bench2Drive, NAVSIM, Vista, WorldModelBench, and ReSim to PICS shows that public resources provide complementary evidence across real-log, non-reactive, reactive, and generative evaluation. A worked trajectory-scoring example and reproducibility dossier translate the protocol into an executable study design. By linking each intended claim to explicit evidence requirements, PICS provides a falsifiable route to proportionate simulation-based assurance and identifies the empirical calibration required before stronger safety claims can be considered.

Keywords: driving world models; end-to-end autonomous driving; closed-loop simulation; safety evaluation; evidence triangulation

Simulation is central to the development of autonomous vehicles because it enables repeatable testing of hazardous and rare situations without exposing road users to experimental risk. CARLA established a widely used open urban driving simulator and demonstrated its value for comparing modular, imitation-learning, and reinforcement-learning approaches [1]. End-to-end autonomous driving (E2E-AD) has since moved from direct sensor-to-control mappings toward planning-oriented systems that learn structured intermediate representations and trajectories. Examples include model-based policy learning in a replayed "world on rails" [2], latent model-based imitation learning in MILE [3], unified perception-prediction-planning in UniAD [4], and vectorized planning in VAD [5].

World models extend this trend by predicting how the scene may evolve under alternative actions. Drive-WM generates controllable multi-view futures and applies image-based rewards to planning [6], while Vista emphasizes generalization, long-horizon consistency, and multiple control interfaces [7]. These capabilities are attractive for end-to-end training and simulation, but they also create a validation problem. A visually plausible video can violate traffic geometry or ignore an ego action. A model can predict logged trajectories accurately while failing after its own actions move the system away from the logged state distribution. Conversely, a

This is an open access article licensed under Creative Commons Attribution 4.0 International
Print ISSN: 1313-230X, Online ISSN: 3033-1404

policy may achieve a high score in one simulator while exploiting benchmark-specific dynamics.

This mismatch is visible in contemporary benchmarks. Bench2Drive provides multi-ability reactive closed-loop evaluation across interactive scenarios [8], whereas NAVSIM trades reactivity for scalable evaluation using real sensor logs and planning-aware proxy metrics [9]. Neither protocol is a substitute for the other. The same distinction applies inside a world model: prediction quality, action controllability, usefulness for choosing a plan, and closed-loop safety are related but separate propositions. Existing benchmarks measure important parts of this chain, but they do not provide a common rule for deciding which combination of evidence is sufficient for a particular claim.

This paper addresses the following question: **What evidence is required before a driving world model can support a claim about end-to-end policy safety or utility?** To answer this question, PICS contributes an architecture-agnostic claim-evidence methodology that specifies the evidence required for end-to-end policy utility and safety claims. It consists of:

First, PICS formalizes a claim contract and separates **P**redictive validity, **I**nterventional controllability, **C**onsequential utility, and closed-loop **S**afety.

Second, a non-compensatory decision rule connects each required layer to metrics, scenario splits, baselines, acceptance thresholds, and explicit fail or unresolved outcomes.

Third, a reproducible audit record triangulates real-log, non-reactive, reactive, and generative simulation evidence without collapsing unlike protocols into a misleading global score.

2. THE CLAIM-EVIDENCE GAP

2.1. Different world models support different uses.

Driving world models can be grouped by their operational role rather than only by architecture. A representation model learns temporally informed features for downstream perception or planning. A predictive model forecasts future images, occupancy, or latent states. An action-conditioned model estimates futures under a specified command or trajectory. A policy world model is embedded in training, trajectory scoring, or online planning. These roles overlap, but each requires a different validation target.

A recent decision-making-centric position organizes world-model evidence into an L0--L7 ladder spanning visual plausibility to policy-optimization utility [15]. PICS shares its use-dependent premise but addresses a different decision problem: instead of treating evidence as a hierarchy, it binds each driving claim to required layers and returns pass, fail, or unresolved under a non-compensatory rule. This distinction matters because generative quality does not automatically imply decision quality. Drive-WM [6] and Vista [7] demonstrate strong visual forecasting and controllability, yet safe policy selection also depends on calibrating predicted action differences to actual consequences. WorldModelBench shows that video-quality evaluation can miss instruction-following and physics violations [10], while ReSim exposes unreliability under hazardous or non-expert actions, precisely where policy evaluation requires dependable rollouts [11].

2.2. Evaluation protocols answer different questions.

Open-loop log replay is inexpensive and uses real observations, but it evaluates a policy under the expert's state distribution. Reactive closed-loop simulation evaluates the states induced by the tested policy, including compounding errors and recovery, but introduces simulator bias and stochastic variance. NAVSIM provides an intermediate non-reactive protocol based on real sensor logs and short-horizon simulation-aware metrics [9]. Bench2Drive instead stresses interactive abilities in CARLA and reports route-level outcomes [8].

These protocols should not be placed on a single linear scale. Open-loop displacement error primarily measures imitation fidelity. A planning-aware proxy can test collision, drivable-area compliance, time to collision, comfort, and progress under a fixed environment. A reactive simulator can additionally test interaction and recovery. A generative simulator may increase sensor realism or scenario diversity, but only after its response to interventions has been validated. The general E2E-AD literature has recently emphasized benchmark-consistent claims [12]; PICS narrows this principle to world-model-based training and evaluation and converts it into an operational assurance checklist.

3. THE PICS EVALUATION PROTOCOL

PICS begins with a claim contract

$$\Gamma = (u, \Omega, M, B, \tau, F) \quad (1)$$

where u is the intended use, Ω is the declared scenario distribution, M is the metric set, B is the baseline set, τ contains pre-declared acceptance thresholds, and F contains failure and unresolved conditions. The contract prevents the evaluation target from changing after results are observed.

For a fixed contract, PICS treats evaluation as a vector of tri-state evidence rather than a scalar leaderboard value. Let the state set T contain pass, fail, and unresolved:

$$\vec{E}(\Gamma) = (E_P, E_I, E_C, E_S) \quad (2)$$

where each component records the outcome of its layer under the declared contract. The vector is intentionally not reduced to a weighted sum. A high predictive score cannot compensate for a failed intervention test, and a strong proxy-planning score cannot erase closed-loop collisions. An unresolved result is distinct from failure: it records that required evidence, ground truth, or statistical precision is insufficient for the intended claim.

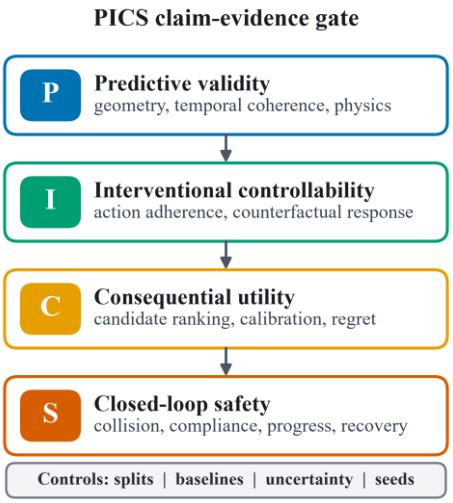


Fig. 1. *PICS non-compensatory claim-evidence gate. Each intended use activates a required prefix of P, I, C, and S; a claim passes only when every required layer passes.*

3.1. P — Predictive validity.

The first layer asks whether the model predicts plausible future observations or latent states over the intended horizon. Evaluation should combine perceptual, geometric, temporal, and physical criteria. Perceptual metrics may include FID or FVD for comparison with previous work, but they are insufficient alone. Geometry-aware checks should measure lane-boundary stability, object scale and depth consistency, ego-motion consistency, and multi-view coherence. Temporal checks should identify identity switches, object disappearance, and accumulated drift. Physics-oriented checks should test collision geometry, traffic-light transitions, kinematic feasibility, and conservation-like constraints where applicable.

The scenario split must separate ordinary validation clips from rare events, adverse weather, unfamiliar locations, and long-horizon rollouts. Results should be reported as degradation curves over horizon and shift severity rather than as one average. A deterministic replay or kinematic predictor provides a useful lower-complexity baseline; a general video generator provides a perceptual baseline. Failure at this layer limits claims to representation learning or qualitative generation.

3.2. I — Interventional controllability.

The second layer asks whether changing an action or command produces the correct change in the predicted world. It requires paired counterfactual tests with the same initial scene and controlled variations in steering, speed, route command, or candidate trajectory. Recommended measures include action adherence, pose error relative to the commanded trajectory, counterfactual separability, response monotonicity, and invariance of scene factors that should not change.

A critical test is support outside the expert-action distribution. Hazardous braking, delayed response, abrupt steering, and recovery maneuvers should be included within physically valid limits. This directly addresses the expert-data limitation highlighted by ReSim [11]. A model that generates attractive videos but has weak action sensitivity cannot be used as a simulator or trajectory evaluator; it can only support unconditional forecasting or data visualization.

3.3. C — Consequential utility.

The third layer asks whether the world model helps select or improve a driving action. The unit of evaluation is a set of candidate actions with observable downstream outcomes. The model should rank safer and more effective candidates above unsafe alternatives. Suitable measures include pairwise ranking accuracy, top-k regret, reward calibration, collision-risk calibration, and improvement over a policy-only or evaluator-only baseline.

This layer isolates utility from visual quality. For example, a low-resolution BEV or latent model may be more useful for trajectory ranking than a photorealistic model if it preserves geometry and interaction. Conversely, a visually strong model may assign similar futures to meaningfully different actions. The evaluation should hold the candidate generator fixed when comparing evaluators and hold the evaluator fixed when comparing candidate generators. Learning to Drive from a World Model provides an example of using learned simulation for on-policy end-to-end training [13], while World4Drive uses predicted latent futures to evaluate multi-modal trajectories [14]. These designs motivate the layer, but their reported benchmark results remain protocol-specific.

3.4. S — Closed-loop safety and robustness.

The final layer asks whether a policy trained, scored, or simulated with the world model behaves safely when its actions alter subsequent observations. Required outcomes include at-fault collisions, time to collision, route completion, red-light and stop-sign violations, off-road duration, comfort, intervention frequency, and recovery after perturbation. Results should be disaggregated by scenario family rather than reported only as a driving score.

Reactive testing should include at least one standardized suite such as CARLA/Bench2Drive, with simulator version, maps, traffic-manager settings, controller, sensor configuration, route files, and random seeds disclosed. Each scenario should be repeated across seeds, and uncertainty intervals should be reported. NAVSIM or another real-log proxy can complement this evidence by testing real sensor inputs at scale [9], but it cannot establish interaction with reactive agents. A claim of improved closed-loop safety therefore requires direct layer-S evidence; layers P–C only explain why such an improvement may occur.

3.5. Claim-to-evidence decision rule.

The intended use determines the required layer set (R(u)). A forecasting claim requires P. A controllable-simulation claim requires P and I. A trajectory-evaluation or training-utility claim requires P, I, and C. A safety or robust-driving claim requires all four layers, with S tested directly. The non-compensatory gate accepts a claim if and only if

$$E_\ell = \text{pass} \forall \ell \in R(u) \quad (3)$$

"Pass" must be defined before evaluation using baseline-relative margins, uncertainty

intervals, and application-specific safety constraints. Failure of any required layer rejects the stated claim; an unresolved required layer defers it. PICS does not prescribe universal thresholds because acceptable risk depends on the operational design domain, simulator fidelity, and deployment context.

3.6. Audit record and falsifiability.

Each layer must produce an evidence card containing the claim, dataset or scenario identifiers, model and policy versions, metric implementation, baseline, threshold, uncertainty procedure, observed outcome, and failure examples. The card must also record what evidence would overturn the result. For P, this may be a horizon or shift severity at which geometry or temporal coherence fails. For I, it may be an action pair that the model cannot distinguish. For C, it may be a systematic candidate-ranking inversion. For S, it may be a scenario family with excess at-fault collisions or failed recovery.

This requirement makes the protocol falsifiable and audit-ready. A paper cannot upgrade a forecasting claim to a safety claim by adding qualitative rollouts after the fact; it must open a new claim contract with the additional required layers. Likewise, negative findings remain informative because the failed layer localizes whether the bottleneck lies in prediction, control response, decision calibration, or reactive policy behavior.

4. BENCHMARK MAPPING AND WORKED USE

Table 1 summarizes how representative public resources contribute evidence. The mapping is functional rather than a ranking. CARLA enables interactive sensor-level simulation [1]. Bench2Drive standardizes multi-ability closed-loop scenarios [8]. NAVSIM supplies a scalable planning-aware proxy on real logs [9]. WorldModelBench contributes physics- and instruction-oriented tests for generative models [10]. Vista supplies a representative action-controllable visual world model [7], while ReSim explicitly extends action support toward non-expert behavior [11]. No resource covers every layer and every real-world validity concern.

Table 1. Representative evidence coverage in the PICS protocol.

Resource	Strongest evidence contribution	Limitation if used alone
CARLA [1]	Reactive closed-loop behavior and recovery	Synthetic sensor and behavior-model gap
Bench2Drive [8]	Standardized interactive abilities and route outcomes	Inherits CARLA distribution and stochasticity
NAVSIM [9]	Scalable real-log planning proxy	Background agents are non-reactive
WorldModelBench [10]	Physics and instruction adherence of generated futures	Does not establish driving-policy safety
Vista [7]	Long-horizon generation and multi-modal action control	Utility and safety remain benchmark-dependent
ReSim [11]	Response under expert and non-expert actions	Generative and reward-model biases may be coupled

A worked evaluation of a trajectory-scoring world model proceeds as follows. First, freeze a claim contract for "trajectory-ranking utility" and construct paired initial states with nominal, near-boundary, and unsafe candidates. Second, verify P on real validation logs and shifted subsets, reporting degradation over horizon. Third, verify I with paired interventions and measure whether trajectory changes produce the expected ego and agent responses. Fourth, test C by comparing predicted candidate ordering with simulator-observed outcomes and report calibration, ranking error, and regret against a policy-only baseline. Fifth, open a separate safety contract, integrate the scorer with a fixed E2E policy, and run S in repeated reactive scenarios. If C improves but S does not, PICS returns offline ranking utility under the tested distribution and rejects the stronger safety claim. This outcome is a localized scientific result rather than an ambiguous global-score change.

5. REPRODUCIBLE IMPLEMENTATION

PICS requires a compact evidence dossier for every experiment. It should include the

frozen claim contract; model and checkpoint identifiers; training-data sources and licenses; observation, action, and prediction spaces; simulator and benchmark versions; scenario taxonomy; random seeds; controller and safety-wrapper settings; metric implementations; hardware; and failed runs. World-model and policy checkpoints must be versioned separately so that a policy change is not misattributed to simulator improvement. Every reported claim should be traceable to the exact evidence cards that determine its gate outcome.

Scenario sampling should be stratified by maneuver, weather, traffic density, road topology, interaction type, and shift severity. Nominal and safety-critical subsets must both be retained. To avoid leakage, routes, towns, and temporal clips used to tune world-model thresholds should not appear in the final claim test. Comparisons should use the same sensor inputs, candidate actions, controllers, and evaluation routes wherever possible.

Statistical reporting should prioritize effect sizes and uncertainty. Route-level or scenario-level paired differences are more informative than unpaired averages when methods are evaluated on the same cases. Closed-loop results should use repeated seeds and bootstrap or other appropriate confidence intervals, with sensitivity analysis over simulator stochasticity and threshold choices. Failures should be categorized into prediction error, control non-compliance, ranking error, policy error, simulator artifact, and ambiguous ground truth. This taxonomy makes negative results actionable and prevents every collision from being attributed to the end-to-end policy.

6. DISCUSSION AND LIMITATIONS

PICS is an assurance-oriented research protocol rather than a vehicle certification standard. Its value is to make the evidential content of a simulation claim explicit and contestable. Real-world validity additionally depends on the operational design domain, sensor degradation, human behavior, and simulator-to-reality transfer. The framework also avoids treating a photorealistic world model as inherently preferable. The representation should be selected according to the claim: image-space models support sensor-level analysis, BEV and occupancy models emphasize geometry, and latent models may provide efficient policy learning while requiring stronger interpretability checks.

The main unresolved issue is threshold calibration. Metric thresholds should be linked to downstream outcomes through cross-protocol studies and, where feasible, controlled real-world tests. Three threats to validity require explicit treatment. Construct validity is threatened when a convenient proxy is interpreted as safety. Internal validity is threatened by coupled bias when a generated world and a learned reward share training data or features. External validity is threatened by simulator-specific behavior and incomplete rare-event coverage. Independent simulators, rule-based safety checks, real-log evaluation, and sampled human review reduce these risks but do not eliminate them. PICS therefore treats its evidence dossier as a versioned assurance argument that must be reopened when the model, policy, threshold, simulator, or operational domain changes.

7. CONCLUSION

Driving world models connect prediction, simulation, and end-to-end decision making, but these functions require different evidence. PICS contributes a claim contract, a four-layer tri-state evidence vector, a non-compensatory acceptance gate, and an auditable dossier for reproducible evaluation. The protocol makes clear when visual forecasting supports only prediction, when action conditioning supports controllable simulation, when calibrated ranking supports decision utility, and when reactive testing supports a bounded safety claim. By combining real-log, non-reactive, reactive, and generative evidence without merging unlike scores, PICS turns benchmark results into proportionate,

falsifiable claims. The next empirical step is to calibrate layer-specific thresholds against cross-benchmark and controlled real-world outcomes.

ACKNOWLEDGMENTS

This work was supported by the Chongqing Natural Science Foundation Innovation and Development Joint Fund (Changan Automobile) under Grant CSTB2024NSCQ-LZX0157.

REFERENCE

[1] Dosovitskiy A, Ros G, Codevilla F, Lopez A, Koltun V. CARLA: An open urban driving simulator. Proceedings of the 1st Annual Conference on Robot Learning, PMLR 78, 2017, p. 1-16. Available at: https://proceedings.mlr.press/v78/dosovitskiy17a.html (accessed 22 August 2026).

[2] Chen D, Koltun V, Krähenbühl P. Learning to drive from a world on rails. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. Available at: https://openaccess.thecvf.com/content/ICCV2021/html/Chen_Learning_To_Drive_From_a_World_on_Rails_ICCV_2021_paper.html (accessed 22 August 2026).

[3] Hu A, Corrado G, Griffiths N, et al. Model-based imitation learning for urban driving. Advances in Neural Information Processing Systems 35, 2022. Available at: https://proceedings.neurips.cc/paper_files/paper/2022/hash/827cb489449ea216e4a257c47e407d18-Abstract-Conference.html (accessed 22 August 2026).

[4] Hu Y, Yang J, Chen L, et al. Planning-oriented autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, p. 17853-17862. Available at: https://openaccess.thecvf.com/content/CVPR2023/html/Hu_Planning-Oriented_Autonomous_Driving_CVPR_2023_paper.html (accessed 22 August 2026).

[5] Jiang B, Chen S, Xu Q, et al. VAD: Vectorized scene representation for efficient autonomous driving. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, p. 8340-8350. Available at: https://openaccess.thecvf.com/content/ICCV2023/html/Jiang_VAD_Vectorized_Scene_Representation_for_Efficient_Autonomous_Driving_ICCV_2023_paper.html (accessed 22 August 2026).

[6] Wang Y, He J, Fan L, Li H, Chen Y, Zhang Z. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, p. 14749-14759. Available at: https://openaccess.thecvf.com/content/CVPR2024/html/Wang_Driving_into_the_Future_Multiview_Visual_Forecasting_and_Planning_with_CVPR_2024_paper.html (accessed 22 August 2026).

[7] Gao S, Yang J, Chen L, et al. Vista: A generalizable driving world model with high fidelity and versatile controllability. Advances in Neural Information Processing Systems 37, 2024. DOI: 10.52202/079017-2906.

[8] Jia X, Yang Z, Li Q, Zhang Z, Yan J. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. Advances in Neural Information Processing Systems 37, 2024. DOI: 10.52202/079017-0025.

[9] Dauner D, Hallgarten M, Li T, et al. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. Advances in Neural Information Processing Systems 37, 2024. DOI: 10.52202/079017-0902.

[10] Li D, Fang Y, Chen Y, et al. WorldModelBench: Judging video generation models as world models. Advances in Neural Information Processing Systems 38, 2025. DOI: 10.52202/085713-1834.

[11] Yang J, Chitta K, Gao S, et al. ReSim: Reliable world simulation for autonomous driving. Advances in Neural Information Processing Systems 38, 2025. DOI: 10.52202/085713-5591.

[12] Guan Y, Liu X, Rao B, et al. Planning-oriented end-to-end autonomous driving: Architectures, evaluation, and emerging paradigms. arXiv:2608.20111, 2026. Available at: https://arxiv.org/abs/2608.20111 (accessed 22 August 2026).

[13] Goff M, Hogan G, Hotz G, et al. Learning to drive from a world model. Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition Workshops, 2025, p. 1989-1998. Available at: https://openaccess.thecvf.com/content/CVPR2025W/DDADS/html/Goff_Learning_to_Drive_from_a_World_Model_CVPRW_2025_paper.html (accessed 22 August 2026).

[14] Zheng Y, Yang P, Xing Z, et al. World4Drive: End-to-end autonomous driving via intention-aware physical latent world model. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, p. 28632-28642. Available at: https://openaccess.thecvf.com/content/ICCV2025/html/Zheng_World4Drive_End-to-End_Autonomous_Driving_via_Intention-aware_Physical_Latent_World_Model_ICCV_2025_paper.html (accessed 22 August 2026).

[15] Yu Y, Zhang S, Sheng Y, et al. How should world models be evaluated for embodied decision-making? A decision-making-centric position. arXiv:2606.15032, 2026. Available at: https://arxiv.org/abs/2606.15032 (accessed 22 August 2026).