BEYOND VISUAL REALISM: BENCHMARKING GENERATED DRIVING SCENARIOS AS POLICY TESTS

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end autonomous vehicles need to be tested on rare, safety-critical situations that real-world driving data collects too slowly and expensively. Generative models offer a scalable fix—synthesizing unbounded, controllable traffic scenarios from recorded histories—but the community still lacks a standard for whether a generated scenario actually works as a driving test. Existing benchmarks assess visual fidelity, temporal consistency, or traffic plausibility, but rarely ask whether it exposes meaningful policy failures, whereas visual corruption naturally lowers policy performance. We introduce **DriveAudit**, a hierarchical benchmark for evaluating generated driving scenarios as policy tests. The multi-source dataset includes 550 scenarios, combining native multi-camera histories with monocular-derived multi-view inputs constructed through generative view completion. Evaluation proceeds through four coupled levels—observation quality, scenario fidelity, policy response, and quality-aware testing effectiveness. Observation quality assesses visual realism and temporal coherence. Scenario fidelity measures image–motion agreement and traffic plausibility through Trajectory–Visual Consistency (TVC) and Traffic Realism Score (TRS). Policy response use Responsibility-aware PDMS (RPDMS), which gates collisions by event-aligned ego attribution, and the headline Quality-aware Testing Effectiveness Score (QTES) multiplies paired policy degradation by reference-relative quality retention, so a visually corrupted scenario cannot be rewarded merely for breaking the policy. Benchmarking seven scenario-generating baselines with six driving policies reveals a mismatch between visual fidelity and testing effectiveness: models that produce more realistic observations do not consistently yield more effective policy tests. DriveAudit thus enable assessing the generated scenarios through joint analysis of visual realism, traffic fidelity, and policy behavior.

1 INTRODUCTION

Validation of end-to-end autonomous vehicles relies on rare, safety-critical situations that real-world fleets collect too slowly and too expensively. Road recordings provide an empirical foundation, but each clip captures only one realized continuation, leaving the alternative futures hidden in the same driving context (Xu et al., 2022, Zhou et al., 2026). Constructing plausible alternatives is therefore central to extending test coverage beyond the available logs.

Simulation and neural reconstruction provide established foundations for this construction, but both scale unevenly. Engine-based simulators rely on configurable digital assets, whereas neural reconstruction grounds rendering in captured geometry and appearance (Dosovitskiy et al., 2017, Ljungbergh et al., 2024, Zhou et al., 2025). provide a complementary path—reusable visual and motion priors can synthesize diverse appearances and traffic futures around a shared history without per-scene capture or asset construction (Gao et al., 2025, Mei et al., 2024). This scalability is attractive for testing, yet it exposes a gap in how such scenarios are evaluated: a photorealistic video may disagree with the traffic states used to score a policy, and a degraded rendering can break a policy for the wrong reason. Visual realism alone is therefore not evidence that a generated scenario is a credible test.

Recent benchmarks for driving scenario generation have expanded evaluation beyond visual realism. DrivingGen evaluates video quality, temporal consistency, and motion plausibility, while WorldLens

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

further incorporates geometry and downstream driving (Zhou et al., 2026, Liang et al., 2026). However, these benchmarks primarily assess whether a generated world is realistic and coherent, rather than whether the policy degradation induced by that world constitutes a credible driving test. In parallel, driving-policy benchmarks such as NAVSIM provide efficient non-reactive policy evaluation (Dauner et al., 2024), but do not jointly verify generated observations, generated traffic, and the responsibility of resulting failures. This leaves a key gap: a policy may fail because of meaningful traffic interaction or simply because the generated scene is inconsistent or visually degraded. Table 1 summarizes this gap across representative benchmarks.

In this paper, we introduce DriveAudit a hierarchical benchmark that evaluates generated driving scenarios as policy tests rather than as plausible videos. We explicitly disentangle two core scenario-generation paradigms for systematic comparison: *traffic trajectory generation*, which produces diverse future agent behaviors, and *visual rendering*, which synthesizes multi-view observations given fixed recorded traffic trajectories. This decomposition isolates rendering fidelity from traffic plausibility. Our paired non-reactive replay protocol ensures that every driving policy is tested on identical scene anchors. The evaluation comprises four diagnostic levels: *Observation Quality*, *Scenario Fidelity*, *Policy Response*, and *Quality-Aware Testing Effectiveness*. Within this framework, RPDMS attributes collision responsibility to ego or traffic agents to avoid indiscriminate policy penalization, while QTES serves as a robust downstream metric that credits policy degradation only when the generated scenario retains valid visual, geometric, and traffic fidelity.

Our contributions are: (1) A policy-test-oriented benchmark for generated driving scenarios. We formulate scenario generation as a policy-testing problem and connect observation quality, scenario fidelity, and independent policy response within a hierarchical evaluation protocol. **(2) A quality-aware criterion for testing effectiveness.** We introduce QTES, which rewards policy degradation only when the generated scene retains perceptual, geometric, image–state, and traffic quality; together with RPDMS, this prevents generators from gaining credit through visual corruption or unassignable collisions. **(3) A multi-source dataset with paired policy evaluation.** We construct a 550-scenario dataset from native multi-camera and monocular-derived multi-view histories and benchmark seven scenario generators against six independent driving policies, finding that visually realistic observations do not consistently yield more effective policy tests. Our benchmark and evaluation code will be made publicly available.

2 Related Work

2.1 Driving Scenario Generation

Driving scenario generation can be broadly supported by simulation, neural reconstruction, and generative modeling, which emphasize different capabilities for autonomous driving development and testing. Simulation-based approaches provide controllability through explicit assets, traffic logic, and physical rules, while reconstruction-based methods focus on reproducing real-world environments with high geometric and visual fidelity (Dosovitskiy et al., 2017, Ljungbergh et al., 2024, Zhou et al., 2025). Generative models complement these approaches by learning visual and motion priors directly from driving data, enabling diverse and controllable scenario synthesis without explicitly modeling every environment configuration. Representative methods such as MagicDrive and Panacea use structured conditions including maps, camera poses, BEV representations, and 3D objects for multi-view generation (Gao et al., 2024, Wen et al., 2024), while MagicDrive-V2 and DreamForge further improve temporal coherence, resolution, and long-horizon synthesis (Gao et al., 2025, Mei et al., 2024). Recent systems also begin to connect generation with autonomous driving evaluation: DriveArena combines conditional rendering with traffic simulation, and Bench2Drive-R incorporates generated observations into downstream driving-agent testing (Yang et al., 2025, You et al., 2024). These developments indicate a shift from generating visually plausible content toward producing coherent scenarios that preserve traffic semantics and can support policy-level evaluation.

2.2 Driving Scenario Evaluation and Benchmarking

Driving scenario evaluation has progressively expanded from traffic realism and policy safety toward generation fidelity and downstream testing value. Early benchmarks such as SafeBench and WOSAC focus on safety-critical policy evaluation and realistic traffic behavior, respec-

Benchmarks	Input Support		Evaluation Coverage						
	Data source	View	Visual Qual.	Temporal Cons.	Img.–State Align.	Traffic Realism	Policy Eval.	Respons. Score	Testing Effect.
Driving-policy benchmarks									
SafeBench (Xu et al., 2022)	Simulated	SV	—	—	—	—	CL	—	—
Bench2Drive (Jia et al., 2024)	Simulated	MV	—	—	—	—	CL	—	—
NAVSIM (Dauner et al., 2024)	Real	MV	—	—	—	—	NR	Rules	—
Driving scenario generation benchmarks									
WOSAC (Montali et al., 2023)	Real	—	—	—	—	✓	—	—	—
DrivingDojo (Wang et al., 2024)	Real	SV	✓	—	Ego	—	—	—	—
ACT-Bench (Arai et al., 2024)	Real	SV	—	—	Ego	—	—	—	—
DrivingGen (Zhou et al., 2026)	Real	SV	✓	✓	Ego	Ego	—	—	—
Bench2Drive-R (You et al., 2024)	Real	MV	✓	Percep.	Det.	—	OL/CL	Rules	—
WorldLens (Liang et al., 2026)	Real	MV	✓	✓	Det./trk.	Human	OL/CL	—	—
DriveAudit (ours)	Real+ Generated	MV	✓	✓	✓	✓	NR	✓	✓

Table 1: **Comparison of autonomous driving benchmarks by input support and evaluation coverage.** SV/MV denotes single-/multi-view inputs; OL/CL/NR denotes open-/closed-loop and non-reactive evaluation. ✓ denotes reported support and – denotes unreported or inapplicable items.

tively (Xu et al., 2022, Montali et al., 2023). Bench2Drive and NAVSIM further establish structured driving-policy evaluation through interactive or non-reactive protocols, emphasizing safety, comfort, progress, and diverse driving capabilities (Jia et al., 2024, Dauner et al., 2024). With the rapid development of generative driving models, recent benchmarks increasingly assess whether generated observations preserve meaningful driving semantics. ACT-Bench studies action-conditioned consistency, while DrivingGen evaluates distributional quality, temporal consistency, trajectory plausibility, and agent behavior (Arai et al., 2024, Zhou et al., 2026). WorldLens further connects multi-view generation quality with downstream perception, driving behavior, and human evaluation (Liang et al., 2026). Overall, existing benchmarks provide complementary measurements of traffic realism, visual fidelity, and policy performance, but these aspects are typically evaluated separately. DriveAudit instead connects generated-scenario fidelity with independent policy response and testing effectiveness, aiming to determine whether a generated scenario is not only realistic, but also a credible and interpretable autonomous driving test.

3 DriveAudit Benchmark

DriveAudit evaluates the fidelity of generated driving scenarios and their effectiveness in testing independent driving policies. It comprises three components: 1) a generation task combining 2D trajectory generation and 3D scenario rendering, 2) a multi-source dataset of native multi-camera and monocular-derived multi-view histories, and 3) a hierarchical evaluation framework spanning observation quality, scenario fidelity, policy response, and quality-aware testing effectiveness. Recorded and generated trajectories support controlled rendering comparisons and alternative-future evaluation, respectively, with policy responses assessed through shared non-reactive replay. Figure 1 summarizes the benchmark. The generation task is introduced in Sec. 3.1, the benchmark dataset in Sec. 3.2, and the evaluation metrics and scoring protocols in Sec. 3.3.

3.1 Generation Task

DriveAudit formulates driving scenario generation as two consecutive tasks: *2D trajectory generation* and *3D scenario rendering*. Given one second of historical observations and traffic states $\mathcal{H}$, together with contextual information $\mathcal{C}$ such as maps, navigation, timestamps, and camera calibration, the trajectory generator predicts future ego and surrounding-agent motion, which is then used to condition multi-view video generation:

$$(\widehat{\mathcal{E}}, \widehat{\mathcal{A}}) = G_{\text{traj}}(\mathcal{H}, \mathcal{C}), \qquad \widehat{\mathcal{V}} = G_{\text{render}}(\mathcal{H}, \mathcal{C}, \widehat{\mathcal{E}}, \widehat{\mathcal{A}}). \tag{1}$$

where $(\widehat{\mathcal{E}}, \widehat{\mathcal{A}})$ denote the generated future trajectories of the camera-carrying ego vehicle and surrounding participants, while $\widehat{\mathcal{V}}$ denotes the synchronized six-view future video. The generated future

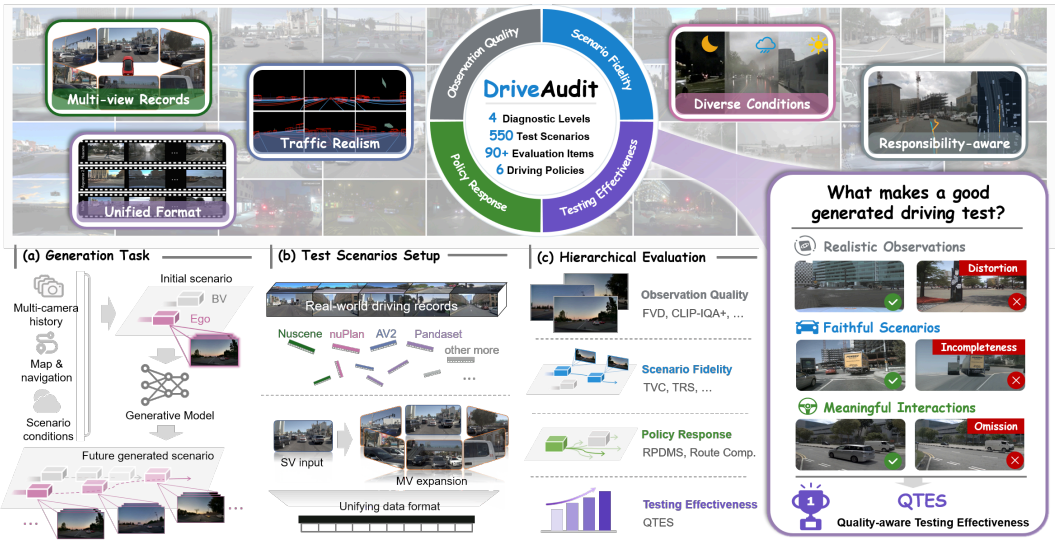

Figure 1: **Overview of DriveAudit: from scenario generation to policy testing.** Multi-source histories support future-traffic and multi-view generation, followed by hierarchical evaluation of scene fidelity, policy response, and testing effectiveness.

is not required to reproduce the recorded continuation, but the rendered observations should remain consistent with the submitted traffic motion.

3.2 Benchmark Dataset

The DriveAudit dataset contains 550 scenario histories from heterogeneous real-world sources, combining native multi-camera recordings with monocular-derived multi-view inputs. We curate the collection for paired evaluation rather than unconstrained clip sampling: source diversity is retained while temporal support, observable interactions, and recording-level dependence are controlled. The full selection protocol is given in Appendix B.3.

Multi-source dataset construction and curation. The native subset contains 500 histories from nuScenes, Argoverse 2, nuPlan, and PandaSet (Caesar et al., 2020, Wilson et al., 2023, Caesar et al., 2021, Xiao et al., 2021), complemented by 50 Nexar monocular histories whose five missing views are synthesized through generative view completion (Liu et al., 2026, Yang et al., 2026). All histories follow a common six-view interface with preserved timestamps, calibration, traffic states, and provenance. All 550 scenarios satisfy the benchmark support requirements and are retained for formal evaluation. Recording-level grouping is used during curation and uncertainty estimation to control duplicated or strongly dependent clips, while the same scene identities and scoring anchors are reused across renderers, motion settings, and driving policies.

Visual characteristics. Figure 2 summarizes image contrast, mean brightness, and temporal visual change. The source distributions overlap while retaining distinct profiles, with the monocular-derived subset extending low-light and high-contrast conditions. These statistics characterize input diversity rather than generation fidelity; detailed source statistics and matched completion controls are provided in Appendices B.1 and B.4.

3.3 Hierarchical Evaluation Metrics

For all evaluated scenario-generation methods, DriveAudit considers four complementary dimensions: *Observation Quality*, *Scenario Fidelity*, *Policy Response*, and *Quality-Aware Testing Effectiveness*. These dimensions progressively evaluate whether generated observations are realistic, whether they faithfully represent the submitted traffic, how independent driving policies respond, and whether the resulting policy degradation constitutes an effective test. Table 2 summarizes the

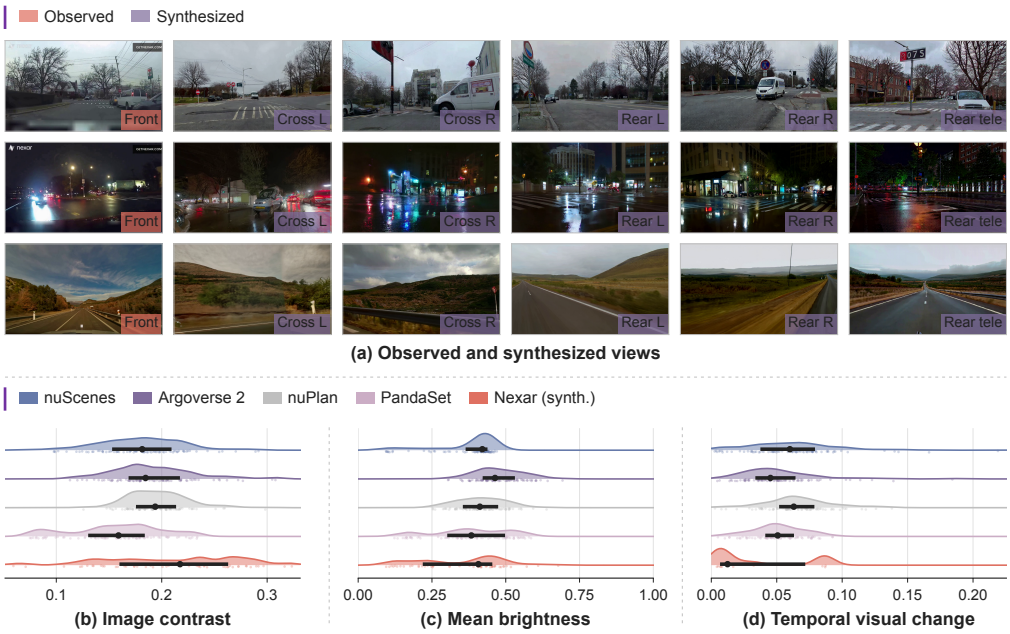

Figure 2: **Multi-source benchmark inputs and visual characteristics.** (a) Native multi-camera and monocular-derived multi-view inputs. (b–d) Source-wise distributions of contrast, brightness, and temporal visual change.

nine headline metrics grouped by these four dimensions. Complete definitions and evaluation protocols are provided in Appendices A and C.

3.3.1 Observation Quality

How realistic and temporally coherent are the generated observations? We first evaluate whether generated videos provide credible visual observations. FVD (Unterthiner et al., 2018) measures distributional discrepancy between real and generated videos, CLIP-IQA+ (Wang et al., 2023) evaluates no-reference perceptual quality, and DINO-T measures temporal feature consistency. CMMD (Jayasumana et al., 2024) is additionally reported as a supplementary CLIP-feature distribution diagnostic. These metrics characterize visual realism, but do not determine whether the rendered observations faithfully realize the traffic states used for downstream evaluation. Detailed computation and aggregation are given in Appendix A.2.

3.3.2 Scenario Fidelity

Do the generated observations and submitted traffic describe the same plausible driving scenario? Visual realism alone is insufficient for test-oriented generation. A photorealistic video may still contain geometry or motion inconsistent with the submitted traffic used for policy evaluation. We therefore evaluate scenario fidelity through **Epi@3**, **TVC**, and **TRS**. Epi@3 measures calibrated cross-view geometric consistency, while TVC evaluates image–state agreement and TRS evaluates the realism of the submitted traffic evolution.

Trajectory–Visual Consistency (TVC). TVC measures whether traffic states are supported by independently extracted RGB evidence. For scenario n, let $S_{nj}^{\text{vis}} \in [0,1]$ denote the available scores for visual support, projected-center agreement, motion-direction agreement, identity persistence, and multi-view support. Let $\mathcal{J}_n$ denote the valid component set. We define:

$$\text{TVC}_n = \frac{1}{|\mathcal{J}_n|} \sum_{j \in \mathcal{J}_n} S_{nj}^{\text{vis}}, \qquad \mathcal{J}_n \neq \varnothing. \tag{2}$$

TVC directly tests whether the visual world observed by a driving policy is consistent with the structured traffic state used by the evaluator. Measured disagreement is retained in the score, while

unavailable evidence is excluded and reported through coverage. Component mappings and support rules are detailed in Appendix A.3.

Traffic Realism Score (TRS). A generated future may represent a plausible alternative to the recorded continuation, making pointwise trajectory error unsuitable as the primary realism criterion. TRS therefore evaluates traffic at the population level by comparing the temporal state and transition distributions of speed, acceleration, angular motion, nearest-object distance, and time-to-collision. These statistics form the distributional similarity term $D_{\mathcal{P}}$, while a reference-relative overlap term penalizes additional geometric conflicts.

$$\mathrm{TRS}_{\mathcal{P}} = D_{\mathcal{P}}\, q^{E^{+}_{\mathcal{P}}/\eta}, \qquad q = 0.55, \quad \eta = 0.01. \tag{3}$$

where $E^{+}_{\mathcal{P}}$ measures the scene-wise excess in rounded-footprint overlap exposure relative to the paired reference. TRS thus combines behavioral similarity with geometric consistency without requiring the generated future to reproduce the recorded trajectory. The overlap term is used as a reproducible geometric diagnostic rather than a physical-collision label. Full feature, temporal, and geometry definitions are provided in Appendix A.3.

3.3.3 Policy Response

How does an independent driving policy respond to the generated scenario? A realistic scenario becomes useful for testing only when it elicits meaningful policy behavior. We therefore evaluate each generated scenario with independent driving policies and report RPDMS together with Route Completion. Route Completion measures episode-level progress along the supported reference route, complementing the local safety and driving-quality measurements captured by RPDMS. Its complete definition is provided in Appendix C.4.

Generated-world non-reactive policy testing. NAVSIM demonstrates efficient policy evaluation by querying a driving agent once from recorded observations and evaluating its fixed plan over a short non-reactive horizon (Dauner et al., 2024). Our setting instead evaluates whether a *generated scenario* can serve as a credible policy test. This requires the synthesized observations to remain consistent with the traffic states used for scoring, while policy responses are examined throughout the generated sequence rather than only at its initial state. We therefore use *repeated-anchor, observation–state-bound replay*, querying each policy at matched timestamps while fixing scene identities, prediction horizons, and scoring anchors across real, recorded-motion, and generated-motion conditions. The protocol remains non-reactive, so policy actions do not alter subsequent observations or traffic, but it enables controlled paired testing without repeatedly invoking the generative renderer for each policy. The complete protocol, motivation, and computational considerations are provided in Appendix C.1.

Responsibility-aware PDMS (RPDMS). Conventional PDM-style scores treat collision as a binary penalty. This is overly coarse for generated stress tests, where a collision may result primarily from the ego policy or from the surrounding generated traffic. Building on the NAVSIM-style aggregation structure (Dauner et al., 2024), RPDMS preserves the drivable-area, TTC-safety, comfort, and progress components while replacing binary collision gating with event-aligned ego responsibility:

$$\mathrm{RPDMS}_i = s_i^{\mathrm{rnc}}\, s_i^{\mathrm{dac}}\, \frac{5s_i^{\mathrm{ttc}} + 2s_i^{\mathrm{comf}} + 5s_i^{\mathrm{prog}}}{12}. \tag{4}$$

For collision-free windows, $s_i^{\mathrm{rnc}} = 1$. For an attributable collision with ego-responsibility estimate ρ_i, $s_i^{\mathrm{rnc}} = 1 - \rho_i$. Collisions without valid attribution retain the conservative zero gate. RPDMS therefore refines collision penalties while leaving the remaining planning-score structure unchanged. Event matching and aggregation are detailed in Appendix A.4.

3.3.4 Quality-Aware Testing Effectiveness

Does the generated scenario expose policy weakness while remaining a credible test? A low policy score alone does not establish testing value because degraded observations or inconsistent traffic can also cause policy failure. QTES therefore measures policy degradation relative to the paired recorded reference while accounting for retained scenario quality.

Models	Observation Quality			Scenario Fidelity			Policy Response		Testing Eff.
	FVD↓	CLIP-IQA+↑	DINO-T↑	Epi@3↑	**TVC↑**	**TRS↑**	**RPDMS↑**	Route Comp.↑	**QTES↑**
Real video	0.000	0.619	0.792	0.379	0.584	1.000	0.788	0.931	50.000
Fixed recorded trajectories + rendering									
MagicDrive-V2	216.224	0.435	0.740	0.179	0.469	1.000	0.764	0.906	39.868
OpenDWM	**93.767**	0.537	0.733	**0.225**	**0.573**	1.000	0.774	0.916	**44.631**
Panacea	278.398	0.581	0.729	0.163	0.408	1.000	0.770	**0.922**	39.637
DreamForge	279.270	0.573	0.704	0.169	0.534	1.000	**0.782**	0.920	41.573
BaseDreamer	615.745	**0.662**	0.674	0.144	0.448	1.000	0.777	0.921	38.765
HorizonDrive	176.006	0.593	**0.840**	0.117	0.551	1.000	0.777	0.920	40.507
MagicDrive	372.394	0.534	0.684	0.136	0.383	1.000	0.721	0.887	37.839
Generated trajectories + rendering									
MagicDrive-V2	217.146	0.437	0.744	0.179	0.447	0.782	0.718	0.870	39.509
OpenDWM	**93.383**	0.531	0.730	**0.222**	**0.543**	0.782	0.724	0.878	**44.079**
Panacea	280.866	0.581	0.728	0.164	0.374	0.782	0.717	**0.881**	39.146
DreamForge	272.880	0.573	0.705	0.168	0.495	0.782	**0.729**	**0.881**	41.311
BaseDreamer	630.836	**0.661**	0.677	0.143	0.410	0.782	0.721	0.880	38.522
HorizonDrive	178.708	0.593	**0.842**	0.116	0.519	0.782	0.725	0.880	40.108
MagicDrive	378.423	0.530	0.681	0.132	0.363	0.782	0.670	0.855	37.282

Table 2: **Cross-generator comparison from visual fidelity to policy-testing effectiveness.** Best and second-best generator results are highlighted in **coral** and blue, respectively.

Quality-aware Testing Effectiveness Score (QTES). QTES measures whether a generated scenario exposes policy weakness for the right reason. A useful test should both degrade policy performance relative to the recorded reference and retain sufficient scene fidelity for that degradation to remain interpretable. We therefore treat scenario quality as a multiplicative gate rather than an additive bonus: a large policy drop cannot compensate for severely degraded observations or implausible traffic. For each scenario, TE_n measures paired policy degradation relative to the recorded reference, with $\mathrm{TE}_n = 0.5$ denoting reference-equivalent response. The local retention factor B_n summarizes CLIP-IQA+, DINO-T, Epi@3, and TVC, while $\mathrm{TRS}_{\mathcal{P}}$ measures population-level traffic realism. The final score is:

$$\mathrm{QTES}_{\mathcal{P}} = 100\,\mathrm{TRS}_{\mathcal{P}}^{1/4}\,\frac{1}{|\mathcal{P}|}\sum_{n\in\mathcal{P}} V_n\,\mathrm{TE}_n\,B_n. \tag{5}$$

where V_n enforces the benchmark validity requirements. The product therefore rewards stronger policy degradation only when the corresponding generated scene remains visually, geometrically, and behaviorally credible. A valid reference self-comparison gives $\mathrm{TE}_n = 0.5$, $B_n = 1$, and $\mathrm{TRS}_{\mathcal{P}} = 1$, yielding the interpretable anchor $\mathrm{QTES} = 50$. Detailed definitions are provided in Appendices A.5.1, A.5.2, and A.5.3.

4 Experiments

In this section, we present experiments addressing two questions: (1) How closely does visual realism align with quality-aware testing effectiveness? (2) What additional insights do traffic-fidelity and responsibility-aware measurements provide when interpreting generated tests? Table 2 compares seven generators using six independent driving policies under matched recorded-motion and generated-motion settings. Model descriptions and replay configurations are provided in Appendices C.2 and C.3.

4.1 Alignment Between Visual Realism and Testing Effectiveness

Visual quality and scene fidelity provide complementary evidence. Table 2 shows a broad correspondence between observation quality and QTES, with OpenDWM performing strongly on both. The ordering is not preserved across all models, however: HorizonDrive improves several observation-quality metrics over DreamForge, while DreamForge attains higher QTES and stronger cross-view consistency; BaseDreamer similarly leads perceptual quality without leading the overall comparison. Thus, visual realism remains important, but scene-level consistency further differentiates whether the rendered views describe a coherent driving situation.

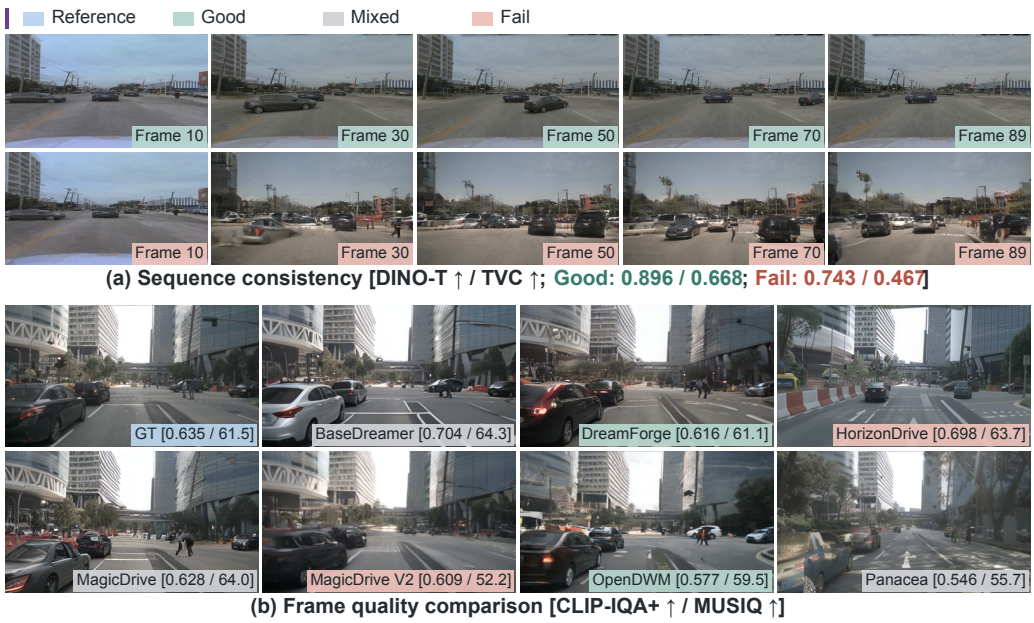

Figure 3: **Qualitative comparison of temporal consistency and perceptual quality.** (a) DINO-T/TVC sequence comparison. (b) CLIP-IQA+/MUSIQ rendering comparison.

Alternative traffic further differentiates policy response. Within a fixed renderer, replacing the recorded continuation with generated future traffic changes RPDMS and route completion while observation-quality metrics remain comparatively stable. The paired settings therefore separate rendering under fixed traffic from the additional policy response induced by alternative future interactions. This distinction motivates the quality-aware analysis below.

Policy responses are architecture dependent. Figure 4 further decomposes the policy-level results across the six ego policies on the same paired scene–condition grid. The effect of generated traffic is not uniform across driving architectures: the shifts in RPDMS and collision rate vary substantially across policies, while route completion remains comparatively stable for most models. This heterogeneity supports evaluating generated scenarios with a policy panel rather than relying on a single driving model. Additional scene-level policy responses are provided in Appendix D.3.

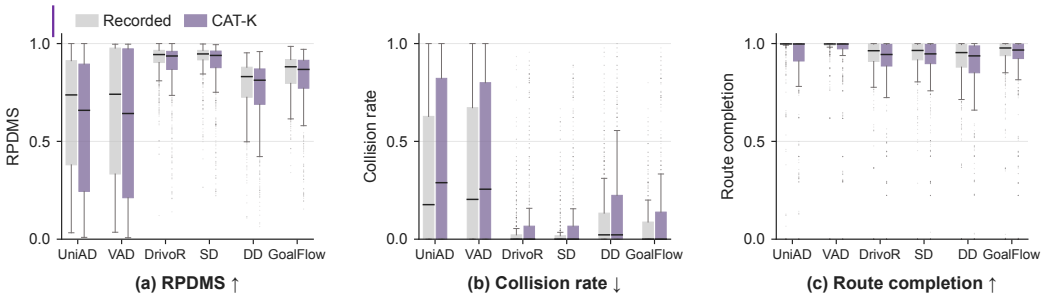

Figure 4: **Policy-specific responses under recorded and generated traffic.** Distributions of RPDMS, collision rate, and route completion for six ego policies evaluated on the same paired scene–condition grid. CAT-K denotes the shared generated-traffic condition.

4.2 Role of Quality Retention in Interpreting Policy Degradation

Stronger policy degradation does not necessarily imply a more effective test. Generated motion produces larger paired-response increments, yet QTES does not increase accordingly because the

score evaluates policy degradation together with retained scene quality. The same principle explains differences across renderers: all rows within a motion setting share the same traffic-level TRS, but differ in retained observation/scene quality and policy response. MagicDrive, for example, produces the largest generated-motion response increment while remaining lowest in QTES, whereas OpenDWM achieves the highest QTES without the largest response increment. Testing effectiveness therefore reflects the balance between credible scene retention and policy degradation, rather than degradation alone. Figure 3 further illustrates the complementary roles of temporal consistency and perceptual quality, with additional QTES decomposition and six-view cases provided in Appendices A.5.4 and D.4.

4.3 Diagnosing Traffic Inconsistency and Attributing Collisions

Scene consistency requires visual–state and traffic-level checks. In the fixed-RGB control, the video and DrivoR predictions remain unchanged while only background states supplied to the evaluator are perturbed. The resulting RPDMS changes show that evaluation-state mismatch can alter the outcome independently of policy behavior; TVC responds on average, although not uniformly across cases.

We design experiments to test both explicit geometric inconsistency and sustained traffic realism as presented in Figure 5(a–b). Under controlled geometric corruption, the distributional component remains close to the reference while TRS decreases markedly, showing the added sensitivity of the geometric term. Applied to CAT-K, SMART, and VBD over longer horizons, TRS further reveals method-dependent differences that are less apparent in short-duration comparisons.

Responsibility-aware scoring differentiates collision penalties. Figure 5(c) shows that binary gating assigns similar collision penalties to other-dominant and ego-dominant events, whereas responsibility-aware gating reduces the former while retaining larger penalties for the latter. Because the remaining planning components are fixed, the comparison isolates the effect of responsibility-aware collision weighting. Independent attribution validation and cross-source ordering results are provided in Appendix A.6.

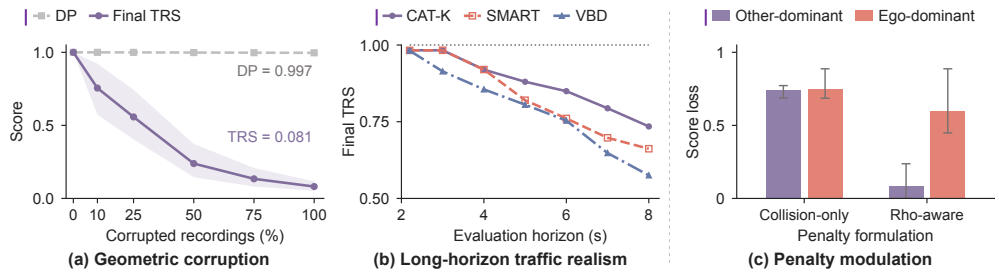

Figure 5: **Diagnostic validation of traffic realism and responsibility-aware policy scoring.** (a) Distributional similarity and TRS under increasing geometric corruption. (b) Long-horizon TRS for CAT-K, SMART, and VBD. (c) Collision-induced score loss for other-dominant and ego-dominant events under collision-only and responsibility-aware gating.

5 Conclusion

We introduce DriveAudit a benchmark for evaluating generated driving scenarios as policy tests through observation quality, scenario fidelity, policy response, and quality-aware testing effectiveness. The experiments show that visual realism and policy degradation are informative but should be interpreted together with retained scene fidelity. QTES captures this joint objective, while TVC, TRS, and responsibility-aware scoring provide diagnostics for image–state consistency, traffic realism, and collision attribution. The resulting framework evaluates not only whether a generated scenario looks realistic or challenges a policy, but whether it remains credible as a test. Nevertheless, the current evaluation is non-reactive and depends on the declared policy panel and metric configuration, as extending to reactive simulation and broader policy families remains future work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we document the generation task and evaluation framework of DriveAudit in Section 3. Section 3.2 and Appendix B describe the dataset sources, curation protocol, provenance tracking, and monocular-to-multi-view construction. Section 3.3 and Appendix A provide the metric definitions, validity criteria, aggregation rules, and validation experiments. Appendix C specifies the paired non-reactive replay protocol, computational considerations, evaluation configuration, model suite, and sensor compatibility settings. Run-level configuration and provenance are recorded as described in Appendices B.2 and C.2. Additional numerical results, policy-specific diagnostics, and qualitative examples are provided in Appendix D. We will make the benchmark and evaluation code publicly available.

ETHICS STATEMENT

DriveAudit evaluates generated driving scenarios using existing recordings and synthesized observations in an offline, non-reactive setting. Source provenance is retained, and synthesized camera views are explicitly distinguished from recorded observations. Any release of source-derived data will be subject to the original usage and redistribution terms and appropriate privacy review. The benchmark supports autonomous driving research, but its results do not establish safety for real-world deployment. Responsibility scores represent model-based collision attribution and should not be interpreted as determinations of legal liability.

AI USE STATEMENT

AI tools were used to assist with manuscript drafting and revision, including language editing, organization, and LaTeX preparation. The authors reviewed all AI-assisted text for accuracy and consistency with the methods and reported results, and take full responsibility for the final manuscript, including its claims, results, and conclusions.

REFERENCES

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

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

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

Rui Chen, Zehuan Wu, Yichen Liu, Yuxin Guo, Jingcheng Ni, Haifeng Xia, and Siyu Xia. Unimlvg: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. *arXiv preprint arXiv:2412.04842*, 2024.

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

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

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations*, volume 2024, pp. 22841–22860, 2024.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28135–28144. IEEE, 2025.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862. IEEE, 2023.

Zhiyu Huang, Zixu Zhang, Ameya Vaidya, Yuxiao Chen, Jaime Fernández Fisac, and Chen Lv. Versatile behavior diffusion for generalized traffic agent simulation. *IEEE Transactions on Intelligent Transportation Systems*, 2026.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315. IEEE, 2024.

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

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8306–8316. IEEE, 2023.

Ellington Kirby, Alexandre Boulch, Yihong Xu, Yuan Yin, Gilles Puy, Éloi Zablocki, Andrei Bursuc, Spyros Gidaris, Renaud Marlet, Florent Bartoccioni, et al. Driving on registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32058–32069, 2026.

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

Lulin Liu, Nuo Chen, Yan Wang, Bangya Liu, Wenyan Cong, Hezhen Hu, Boris Ivanovic, Hao Wang, Ziyao Zeng, Xinyu Gong, et al. Openlongtail: Generative scaling of long-tail driving data. *arXiv preprint arXiv:2607.09655*, 2026.

William Ljungbergh, Adam Tonderski, Joakim Johnander, Holger Caesar, Kalle Åström, Michael Felsberg, and Christoffer Petersson. Neuroncap: Photorealistic closed-loop safety testing for autonomous driving. In *European Conference on Computer Vision*, pp. 161–177. Springer, 2024.

Jianbiao Mei, Tao Hu, Xuemeng Yang, Licheng Wen, Yu Yang, Tiantian Wei, Yukai Ma, Min Dou, Botian Shi, and Yong Liu. Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003*, 2024.

Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nicholas Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, et al. The waymo open sim agents challenge. *Advances in Neural Information Processing Systems*, 36:59151–59171, 2023.

Jingcheng Ni, Yuxin Guo, Yichen Liu, Rui Chen, Lewei Lu, and Zehuan Wu. Maskgwm: A generalizable driving world model with video mask reconstruction. *arXiv preprint arXiv:2502.11663*, 2025.

Wenchao Sun, Xuewu Lin, Keyu Chen, Zixiang Pei, Xiang Li, Yining Shi, and Sifa Zheng. Sparsedrivev2: Scoring is all you need for end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 446–463. Springer, 2026.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

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

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6902–6912, 2024.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. Smart: Scalable multi-agent real-time motion generation via next-token prediction. *Advances in Neural Information Processing Systems*, 37: 114048–114071, 2024.

Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pp. 3095–3101. IEEE, 2021.

Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, and Wei Yin. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1602–1611. IEEE, 2025.

Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. *Advances in Neural Information Processing Systems*, 35:25667–25682, 2022.

Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, et al. Drivearena: A closed-loop generative simulation platform for autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26933–26943. IEEE, 2025.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Junqi You, Xiaosong Jia, Zhiyuan Zhang, Yutao Zhu, and Junchi Yan. Bench2drive-r: Turning real world data into reactive closed-loop autonomous driving benchmark by generative model. *arXiv preprint arXiv:2412.09647*, 2024.

Conglang Zhang, Yifan Zhan, Qingjie Wang, Zhanpeng Ouyang, Yu Li, Zihao Yang, Xiaoyang Guo, Weiqiang Ren, Qian Zhang, Zhen Dong, et al. Horizondrive: Self-corrective autoregressive world model for long-horizon driving simulation. *arXiv preprint arXiv:2605.11596*, 2026.

Zhejun Zhang, Peter Karkus, Maximilian Igl, Wenhao Ding, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Closed-loop supervised fine-tuning of tokenized traffic models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5422–5432. IEEE, 2025.

Hongyu Zhou, Longzhong Lin, Jiabao Wang, Yichong Lu, Dongfeng Bai, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven Waslander. Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 103502–103524, 2026.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Diffusiondrivev2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.

APPENDIX

This appendix provides the notation, metric definitions, dataset construction, replay protocol, model configurations, and supplementary results used by the benchmark. The same paired evaluation units and separation between scene fidelity and policy response are retained throughout.

APPENDIX CONTENTS

A Metric Definitions and Validation **14**
- A.1 Notation and Evaluation Units . . . 14
- A.2 Observation Quality Metrics . . . 15
- A.3 Scenario Fidelity . . . 16
- A.4 Policy Response: RPDMS . . . 17
- A.5 Quality-Aware Testing Effectiveness . . . 17
- A.6 Core-Metric Validation Experiments . . . 18

B Dataset Composition and Curation **20**
- B.1 Supplementary Dataset Statistics . . . 20
- B.2 Scenario Interface and Provenance . . . 20
- B.3 Evaluation-Oriented Scenario Curation . . . 20
- B.4 Monocular-to-Multi-View Construction . . . 21

C Replay and Evaluation Protocol **21**
- C.1 Repeated-Anchor Paired Non-Reactive Replay . . . 21
- C.2 Evaluation Configuration . . . 22
- C.3 Evaluated Model Suite . . . 22
- C.4 Route Completion and Episode Summary . . . 24

D Supplementary Numerical Results **24**
- D.1 Supplementary Renderer Diagnostics . . . 24
- D.2 Agent-Specific Responses . . . 25
- D.3 Supplementary Policy-Response Cases . . . 25
- D.4 Supplementary Qualitative Cases . . . 27

A METRIC DEFINITIONS AND VALIDATION

A.1 NOTATION AND EVALUATION UNITS

We first make the evaluation units explicit. Let $n \in \mathcal{P}$ index a source scenario in the declared comparison population $\mathcal{P}$, $c \in \mathcal{C}_n$ a camera, t a physical exposure time, $a \in \mathcal{A}$ an ego policy, and $t_k \in \mathcal{T}_n$ a scoring anchor. A condition $\gamma \in \{\mathrm{real}, \mathrm{rec}, \mathrm{gen}\}$ denotes real video, recorded-motion rendering, or generated-motion rendering. Each condition provides a synchronized observation stream $\mathcal{O}_n^{(\gamma)}$ and a traffic-state stream $\mathcal{X}_n^{(\gamma)}$. The policy output at an anchor is a finite-horizon ego plan $\widehat{\tau}_{n,a,k}^{(\gamma)}$.

All headline metrics are defined on one of three evidence populations. *Observation-level* metrics operate on valid image frames or clips. *Scenario-level* metrics operate on paired scene evidence, while *policy-level* metrics operate on valid scenario–policy–anchor tuples. Missing measurements are not treated as poor measurements. Availability masks and denominators are retained explicitly, and population-level statistics are recomputed on the declared support.

A.2 Observation Quality Metrics

Observation Quality measures the visual distribution, perceptual quality, and temporal stability of the rendered future. Unless otherwise stated, all methods use identical camera subsets, exposure sampling, preprocessing, and frozen feature extractors.

Fréchet Video Distance (FVD). Let $\{\mathbf{x}_i\}_{i=1}^{N_r}$ and $\{\mathbf{y}_j\}_{j=1}^{N_g}$ denote clip-level I3D activations for the recorded and generated populations, respectively. Their empirical means and covariance matrices are $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ and $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. The distributional discrepancy is:

$$\mathrm{FVD} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \mathrm{Tr}\Big(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{1/2}\Big). \tag{6}$$

The final evaluator follows the reference FVD convention. We use 16-frame clips with stride 2, resize them to 224×224, and extract the 400-dimensional Kinetics-400 I3D logits before softmax. No feature normalization is applied before Equation 6. FVD is computed once from the complete activation pools rather than averaged over scene-wise Fréchet distances. The checkpoint, feature dimension, clip manifest, and sample counts are versioned with every run. By construction, $\mathrm{FVD} \in [0, \infty)$ and lower values indicate closer recorded and generated feature distributions. $\mathrm{FVD} = 0$ corresponds to identical empirical Gaussian feature statistics. FVD has no universal upper bound, and its absolute scale depends on the feature extractor, preprocessing, clip sampling, and evaluation population. We therefore compare only scores computed under the same protocol. Under our fixed protocol, the generator results in Table 2 span 93.383–630.836 across the two motion settings.

CLIP-IQA+. Let $\mathcal{F}_{\mathcal{P}}$ be the set of valid future frame tuples (n, c, t) and let $h_{\mathrm{IQA}}(\cdot)$ denote the frozen CLIP-IQA+ predictor. The population score is the frame-equal mean:

$$\text{CLIP-IQA+}_{\mathcal{P}} = \frac{1}{|\mathcal{F}_{\mathcal{P}}|} \sum_{(n,c,t)\in\mathcal{F}_{\mathcal{P}}} h_{\mathrm{IQA}}(I_{nct}). \tag{7}$$

This metric is no-reference and therefore measures perceptual image quality rather than agreement with the paired recorded future.

CLIP Maximum Mean Discrepancy (CMMD). For L2-normalized CLIP embeddings $X = \{\mathbf{x}_i\}_{i=1}^{N_r}$ and $Y = \{\mathbf{y}_j\}_{j=1}^{N_g}$, we retain CMMD as a supplementary distributional diagnostic using the biased Gaussian-kernel estimator:

$$\widehat{\mathrm{CMMD}} = \alpha \left[\frac{1}{N_r^2} \sum_{i,i'} k_\sigma(\mathbf{x}_i, \mathbf{x}_{i'}) + \frac{1}{N_g^2} \sum_{j,j'} k_\sigma(\mathbf{y}_j, \mathbf{y}_{j'}) - \frac{2}{N_r N_g} \sum_{i,j} k_\sigma(\mathbf{x}_i, \mathbf{y}_j) \right], \tag{8}$$
$$k_\sigma(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{2\sigma^2} \right).$$

We use OpenAI CLIP ViT-L/14@336px features, $\sigma = 10$, and $\alpha = 1000$. The same feature pools and preprocessing are used for all renderers.

DINO temporal consistency (DINO-T). Temporal consistency is measured over physical rather than frame-index intervals. Let $\mathcal{P}_\Delta$ contain valid same-camera pairs whose exposure-time separation matches target interval Δ, and let $\mathbf{z}_{c,t}$ be the corresponding frozen DINO representation. We define:

$$\mathrm{DINO}_\Delta = \frac{1}{|\mathcal{P}_\Delta|} \sum_{(c,t,t')\in\mathcal{P}_\Delta} \frac{\mathbf{z}_{c,t}^\top \mathbf{z}_{c,t'}}{\|\mathbf{z}_{c,t}\|_2 \|\mathbf{z}_{c,t'}\|_2}, \ \text{DINO-T} = \frac{1}{3} \sum_{\Delta\in\{0.3,1.0,2.5\}\mathrm{s}} \mathrm{DINO}_\Delta. \tag{9}$$

For each requested interval, the nearest later exposure is retained only when the absolute timing error is at most 0.08 s. Duplicate timestamps are removed before pairing. Because high feature similarity can also result from weak motion, DINO-T is interpreted jointly with the remaining observation and scenario-fidelity evidence.

A.3 SCENARIO FIDELITY

Scenario Fidelity asks whether the rendered observations and the submitted traffic describe the same coherent world. We separate cross-view geometry, local image–state agreement, and population-level traffic realism because these measurements operate on different evidence.

Cross-view geometry (Epi@3). Let $\mathcal{Q}$ denote valid camera-pair–frame units and $\mathcal{M}_q$ the calibrated correspondences for unit q. For correspondence $p \in \mathcal{M}_q$, let e_{qp} be the epipolar residual in pixels after applying the resized-camera calibration. The inlier rate is:

$$\mathrm{Epi@3} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left[\frac{1}{|\mathcal{M}_q|} \sum_{p \in \mathcal{M}_q} \mathbf{1}\{e_{qp} \le 3\,\mathrm{px}\} \right]. \tag{10}$$

Correspondences are first normalized within each camera-pair–frame unit, after which valid units receive equal weight. Unsupported units remain unavailable rather than being assigned perfect consistency.

Trajectory–Visual Consistency (TVC). For scenario n, TVC combines five independent evidence families: persistent visual support, projected-center agreement, image-motion direction agreement, identity persistence, and multi-view support. Their normalized component scores are:

$$\begin{aligned} S_{n,\mathrm{sup}} &= u_n, & S_{n,\mathrm{ctr}} &= 1 - \mathrm{clip}_{[0,1]}(e_n^{\mathrm{ctr}}), \\ S_{n,\mathrm{mot}} &= \frac{a_n^{\mathrm{mot}} + 1}{2}, & S_{n,\mathrm{id}} &= p_n^{\mathrm{id}}, \qquad S_{n,\mathrm{mv}} = r_n^{\mathrm{mv}}. \end{aligned} \tag{11}$$

Let $\mathcal{J} = \{\mathrm{sup}, \mathrm{ctr}, \mathrm{mot}, \mathrm{id}, \mathrm{mv}\}$, $m_{nj} \in \{0, 1\}$ indicate whether component j is supported, and $w_j = 1/5$. The scenario score and its evidence coverage are:

$$\mathrm{TVC}_n = \frac{\sum_{j \in \mathcal{J}} w_j m_{nj} S_{nj}}{\sum_{j \in \mathcal{J}} w_j m_{nj}}, \ \mathrm{Cov}_n^{\mathrm{TVC}} = \frac{\sum_{j \in \mathcal{J}} w_j m_{nj}}{\sum_{j \in \mathcal{J}} w_j}. \tag{12}$$

Equation 12 averages only observable evidence. Measured disagreement remains a low component score rather than being masked as missing. The registered full-profile threshold is $\mathrm{Cov}_n^{\mathrm{TVC}} \ge 0.8$.

Traffic Realism Score (TRS). A generated future is not required to reproduce the single recorded continuation. TRS therefore compares the generated and recorded traffic distributions over physical state and transition statistics. Let

$$\mathcal{F} = \{v, a, \omega, \dot{\omega}, d_{\mathrm{near}}, \mathrm{TTC}\}$$

contain linear speed, linear acceleration, angular speed, angular acceleration, nearest-object distance, and time-to-collision. For feature $f \in \mathcal{F}$, the state-distribution similarity and transition similarity are:

$$\begin{aligned} S_f^{\mathrm{state}} &= 1 - \frac{\mathrm{JSD}\left(P_f^{\mathrm{gt}} \| P_f^{\mathrm{gen}}\right)}{\log 2}, \\ S_f^{\mathrm{trans}} &= \frac{1}{|\mathcal{D}|} \sum_{\delta \in \mathcal{D}} \left[1 - \frac{\mathrm{JSD}\left(P_{f,\delta}^{\mathrm{gt}} \| P_{f,\delta}^{\mathrm{gen}}\right)}{\log 2} \right], \ \mathcal{D} = \{0.5, 1.0, 2.0\}\,\mathrm{s}. \end{aligned} \tag{13}$$

We combine state and transition evidence as $S_f = \sqrt{S_f^{\mathrm{state}} S_f^{\mathrm{trans}}}$. Using relative feature weights $(1, 1, 1, 1, 2, 2)$ in the order above, the distributional traffic term is:

$$D_{\mathcal{P}} = \exp\left(\frac{\sum_{f \in \mathcal{F}} w_f \log S_f}{\sum_{f \in \mathcal{F}} w_f} \right). \tag{14}$$

Distribution matching alone can hide newly introduced geometric conflicts. Let r_n^{gt} and r_n^{gen} denote actor-time exposure to rounded-footprint overlap in the paired recorded and generated traffic. We therefore retain only the positive scene-wise excess:

$$E_{\mathcal{P}}^{+} = \frac{1}{|\mathcal{P}|} \sum_{n \in \mathcal{P}} \max\left(r_n^{\mathrm{gen}} - r_n^{\mathrm{gt}}, 0\right), \qquad \mathrm{TRS}_{\mathcal{P}} = D_{\mathcal{P}} q^{E_{\mathcal{P}}^{+}/\eta}, \quad q = 0.55,\ \eta = 0.01. \tag{15}$$

The overlap test uses the fixed rounded-footprint geometry with corner-rounding factor $c = 0.7$. The geometric term is a reproducible consistency diagnostic, not a calibrated physical-collision probability.

A.4 Policy Response: RPDMS

Policy Response is evaluated on planning windows rather than on the submitted background trajectories. For eligible window i, let s_i^{dac}, s_i^{ttc}, s_i^{comf}, and s_i^{prog} denote drivable-area compliance, TTC safety, comfort, and progress, respectively. Holding these components fixed, the non-collision planning factor is:

$$B_i = s_i^{\mathrm{dac}} \frac{5s_i^{\mathrm{ttc}} + 2s_i^{\mathrm{comf}} + 5s_i^{\mathrm{prog}}}{12}, \qquad B_i \in [0,1]. \tag{16}$$

Conventional binary collision gating assigns the same zero factor to every supported collision. RPDMS instead conditions the gate on event-aligned ego attribution. Let $C_i \in \{0,1\}$ be the collision indicator, $\mathcal{I}_{\mathrm{attr}}$ the set of attributable collision windows, and $\rho_i \in [0,1]$ the estimated ego-responsibility ratio. We define:

$$\mathrm{RNC}_i = \begin{cases} 1-\rho_i, & i \in \mathcal{I}_{\mathrm{attr}}, \\ 1-C_i, & i \notin \mathcal{I}_{\mathrm{attr}}, \end{cases} \qquad \mathrm{RPDMS}_i = \mathrm{RNC}_i B_i. \tag{17}$$

Thus, collision-free windows remain unchanged, attributable collisions receive a continuous penalty, and unsupported attribution falls back to the conservative binary gate. For $i \in \mathcal{I}_{\mathrm{attr}}$,

$$\frac{\partial\, \mathrm{RPDMS}_i}{\partial \rho_i} = -B_i \le 0, \tag{18}$$

which makes the score monotone in ego responsibility while preserving all non-collision components. Window-level products are first aggregated within each scenario–policy pair and then across the fixed policy panel. Responsibility confidence is retained separately and is not multiplied into ρ_i.

A.5 Quality-Aware Testing Effectiveness

QTES formalizes the test-oriented objective used in the main paper. Policy degradation contributes only when the generated scenario retains enough quality to remain a credible test. The score therefore separates paired policy degradation, local quality retention, population-level traffic realism, and validity.

A.5.1 Paired Policy-Degradation Term

For source scenario n and policy a, let $R_{n,a}^{\mathrm{gt}}$ and $R_{n,a}^{\mathrm{gen}}$ be paired mean RPDMS values under the recorded and generated conditions. The pair fixes scenario identity, checkpoint, scoring anchors, prediction horizon, and camera-only input protocol. With numerical floor $\epsilon_{\mathrm{TE}} > 0$,

$$\mathrm{TE}_{n,a} = \begin{cases} \frac{1}{2} + \frac{1}{2}\,\mathrm{clip}_{[-1,1]}\left(\dfrac{R_{n,a}^{\mathrm{gt}} - R_{n,a}^{\mathrm{gen}}}{\max\{R_{n,a}^{\mathrm{gt}}, \epsilon_{\mathrm{TE}}\}}\right), & R_{n,a}^{\mathrm{gt}} \ge \epsilon_{\mathrm{TE}}, \\ \frac{1}{2}, & \max\{R_{n,a}^{\mathrm{gt}}, R_{n,a}^{\mathrm{gen}}\} < \epsilon_{\mathrm{TE}}, \\ 0, & R_{n,a}^{\mathrm{gt}} < \epsilon_{\mathrm{TE}} \le R_{n,a}^{\mathrm{gen}}. \end{cases} \tag{19}$$

The neutral value $\mathrm{TE}_{n,a} = 0.5$ denotes reference-equivalent response. Values above 0.5 correspond to degradation of the tested policy under the generated condition. Scenario-level degradation is the equal-policy mean:

$$\mathrm{TE}_n = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathrm{TE}_{n,a}. \tag{20}$$

A.5.2 Reference-Relative Quality Retention

Let $\mathcal{J}_Q = \{\mathrm{iqa}, \mathrm{dino}, \mathrm{epi}, \mathrm{tvc}\}$ denote the paired local-quality components. For each component j,

$$q_{nj}^{\mathrm{rel}} = \mathrm{clip}_{[0,1]}\left(\frac{q_{nj}^{\mathrm{gen}}}{\max\{q_{nj}^{\mathrm{gt}}, \epsilon_j\}}\right). \tag{21}$$

The ratio is capped at one because this term measures retention rather than improvement over the reference. The local quality factor is the weighted geometric mean:

$$B_n = \left(\prod_{j \in \{\text{iqa,dino,epi}\}} q_{nj}^{\text{rel}} \right)^{1/6} \left(q_{n,\text{tvc}}^{\text{rel}} \right)^{1/4}. \tag{22}$$

The three image-derived terms jointly contribute one half of the total quality exponent, TVC contributes one quarter, and the population-level factor $\text{TRS}_{\mathcal{P}}^{1/4}$ contributes the remaining quarter. FVD and CMMD are excluded because they are population-level distribution distances rather than paired scenario-level retention terms.

A.5.3 Aggregation and Decomposition

Let $V_n \in \{0, 1\}$ indicate that the paired observations, local quality components, and all policies required by the formal protocol are valid. The benchmark-level score is:

$$\text{QTES}_{\mathcal{P}} = 100\, \text{TRS}_{\mathcal{P}}^{1/4} \frac{1}{|\mathcal{P}|} \sum_{n \in \mathcal{P}} V_n\, \text{TE}_n\, B_n. \tag{23}$$

A valid self-comparison satisfies $\text{TE}_n = 0.5$, $B_n = 1$, and $\text{TRS}_{\mathcal{P}} = 1$, yielding the reference anchor $\text{QTES} = 50$. Methods are compared only under the same population, policy panel, support contract, and metric configuration.

For interpretation, Equation 23 admits the exact decomposition:

$$Q_{\text{neutral}} = \frac{50\, \text{TRS}_{\mathcal{P}}^{1/4}}{|\mathcal{P}|} \sum_{n \in \mathcal{P}} V_n B_n,\ \Delta_{\text{resp}} \quad = \frac{100\, \text{TRS}_{\mathcal{P}}^{1/4}}{|\mathcal{P}|} \sum_{n \in \mathcal{P}} V_n B_n \left(\text{TE}_n - \frac{1}{2} \right). \tag{24}$$

such that $\text{QTES}_{\mathcal{P}} = Q_{\text{neutral}} + \Delta_{\text{resp}}$. The decomposition preserves the scenario-wise products in the original score and is used only for diagnosis. It is not a second ranking criterion. Uncertainty resamples independent source-recording groups and recomputes the complete population statistic.

	Recorded trajectories		CAT-K trajectories	
Model	Δ_{resp}	QTES↑	Δ_{resp}	QTES↑
Real video	0.000	50.000	–	–
MagicDrive-V2	+1.065	39.868	+3.403	39.509
OpenDWM	+0.655	**44.631**	+3.417	**44.079**
Panacea	+0.899	39.637	+3.407	39.146
DreamForge	-0.019	41.573	+2.871	41.311
BaseDreamer	+0.063	38.765	+2.899	38.522
HorizonDrive	+0.781	40.507	+3.296	40.108
MagicDrive	+1.609	37.839	+3.768	37.282

Table 3: **Paired-response contribution to QTES.** The signed increment isolates the policy-response contribution while retaining the quality and validity factors used by QTES. Best and second-best QTES results are highlighted in **coral** and blue, respectively.

A.5.4 Supplementary QTES Decomposition

Figure 6 shows the decomposition in Equation 24. The neutral component retains measured quality, traffic realism, and validity at the reference-equivalent policy response. The signed increment isolates the contribution of paired policy degradation. This figure is diagnostic and does not define a second ranking criterion.

A.6 Core-Metric Validation Experiments

Figure 5 contains the completed validation studies used in the main paper. The geometric-corruption intervention fixes the underlying comparison population while increasing rounded-footprint overlap, isolating the geometric term from distributional behavior matching. The horizon experiment applies

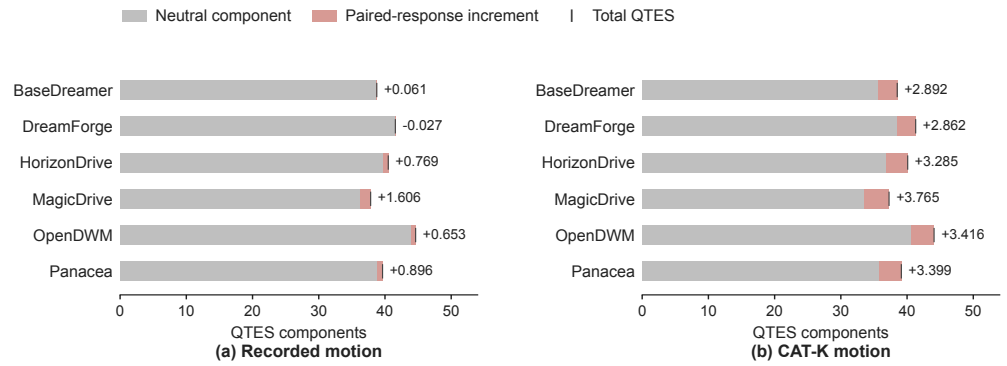

Figure 6: **QTES decomposition under recorded and CAT-K motion.** Each bar separates the neutral quality-retention component from the signed paired-response increment. The marker denotes the resulting total QTES.

the same TRS definition to CAT-K, SMART, and VBD (Zhang et al., 2025, Wu et al., 2024, Huang et al., 2026), allowing long-horizon traffic realism to be compared without changing the metric itself.

For RPDMS, the analytical properties in Equations 17–18 are checked under matched planning windows. In the completed gate comparison, the binary planning score is 0.763 and the responsibility-aware score is 0.788. Attribution coverage is 0.993 with a 0.007 fallback rate. Because the non-collision factor B_i is held fixed, the observed difference isolates the responsibility-aware collision gate rather than a change in comfort, progress, or TTC scoring.

These checks establish implementation consistency and diagnostic sensitivity. They do not establish legal liability, calibrated collision probability, or closed-loop safety. These interpretations remain outside the benchmark scope.

Figure 7 provides complementary responsibility diagnostics across event types and data sources. The separation and ordering panels test whether the continuous responsibility estimate distinguishes other-dominant from ego-dominant events, while the cross-source panel verifies that the TRS overlap response is not confined to a single dataset.

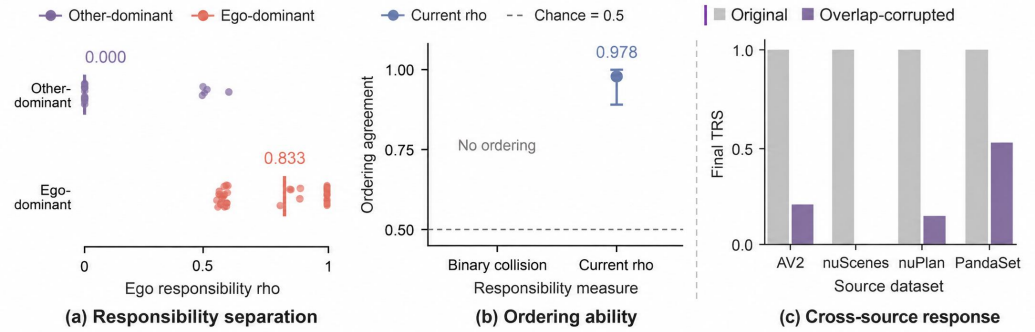

Figure 7: **Supplementary responsibility and cross-source validation.** (a) Responsibility separation between other-dominant and ego-dominant events. (b) Pairwise ordering ability of binary collision gating and the continuous responsibility estimate. (c) TRS response to overlap corruption across source datasets.

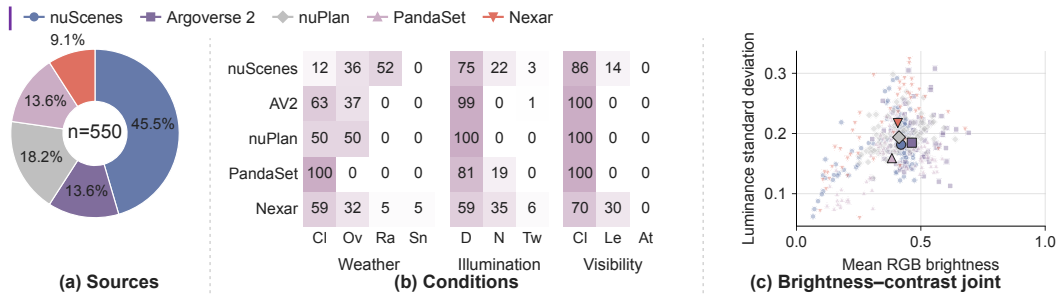

Figure 8: **Dataset composition and visual characteristics.** (a) Source proportions. (b) Weather, illumination, and visibility profiles. (c) Brightness–contrast distributions, with larger points denoting source medians.

B Dataset Composition and Curation

B.1 Supplementary Dataset Statistics

The 550 histories comprise 250 nuScenes, 75 AV2, 100 nuPlan, 75 PandaSet, and 50 Nexar histories. Figure 8 summarizes source composition, weather/illumination/visibility profiles, and the joint brightness–contrast distribution. These statistics describe the selected benchmark population rather than the prevalence of conditions in the underlying public datasets.

Unknown condition labels are normalized only for visualization, and the plots do not recover missing annotations or establish synthesis quality. The Nexar subset contributes a larger share of low-light and high-contrast histories, while the source distributions retain substantial overlap. Temporal visual change is likewise descriptive because it is affected by camera motion, exposure, frame spacing, and synthesis variation.

B.2 Scenario Interface and Provenance

Each scenario package separates participant inputs, submitted futures, and organizer-only references. Historical images preserve camera identifiers and exposure timestamps. Historical states preserve ego motion, actor identities, boxes, and validity, while calibration records intrinsics, transforms, map support, and resize/crop updates. Generated futures contain synchronized camera outputs and submitted ego/background states. Recorded future images and states are used only for paired organizer-side diagnostics. Every evaluation record stores the configuration, input hashes, raw measurements, status, and source provenance for deduplication and audit.

B.3 Evaluation-Oriented Scenario Curation

The benchmark uses the complete 550-scenario collection for formal evaluation. Scenario curation is therefore performed during dataset construction rather than by selecting a smaller evaluation subset. Let $\mathcal{D}$ denote the candidate history pool and $i \in \mathcal{D}$ a candidate associated with source $s(i)$ and recording group $g(i)$. Each candidate is required to satisfy four support conditions: synchronized multi-view/calibration support, temporal completeness, persistent and observable traffic participants, and sufficient local interaction evidence. We define:

$$e_i = I_{\mathrm{mv}}(i) I_{\mathrm{time}}(i) I_{\mathrm{actor}}(i) I_{\mathrm{int}}(i), \qquad e_i \in \{0, 1\}. \tag{25}$$

Only candidates satisfying $e_i = 1$ enter the released benchmark collection. Recording identifiers are retained to prevent duplicated or strongly overlapping clips from being treated as independent evidence during dataset construction and uncertainty estimation. The resulting evaluation population is:

$$\mathcal{P} = \{i \in \mathcal{D} : e_i = 1\}, \qquad |\mathcal{P}| = 550, \tag{26}$$

with 250 nuScenes, 100 nuPlan, 75 Argoverse 2, 75 PandaSet, and 50 Nexar histories. Once the collection is frozen, the same scenario identities and temporal anchors are reused across renderers, motion settings, and ego policies. The curation protocol therefore controls data validity, source diversity, and recording-level dependence without introducing a smaller evaluation subset.

Algorithm 1 Evaluation-oriented construction of the benchmark population

Require: candidate histories $\mathcal{D}$
Ensure: benchmark population $\mathcal{P}$
$\mathcal{P} \leftarrow \varnothing$
for all $i \in \mathcal{D}$ **do**
  verify multi-view, temporal, actor, and interaction support
  assign source and recording-group identifiers
  if $e_i = 1$ and i satisfies the recording-level duplication checks **then**
    add i to $\mathcal{P}$
  end if
end for
freeze the 550 scenario identities and their evaluation anchors
reuse the same $\mathcal{P}$ across renderers, motion settings, and ego policies
return $\mathcal{P}$

B.4 Monocular-to-Multi-View Construction

For the 50 Nexar histories, the observed front view is retained and five missing views are synthesized with a fixed view-completion model (Liu et al., 2026, Yang et al., 2026). Each synthesized camera retains target calibration, timestamps, checkpoint identity, and source-clip provenance. All pose, depth, caption, and view-synthesis preprocessing is truncated at the history boundary. Synthesized views are never promoted to independent ground truth. View completion is fixed before future-generator comparison, so every renderer receives the same completed history.

Matched completion control. On a supported native multi-view subset, the front-view history is retained while the remaining historical views are reconstructed by the same completion model and compared with their held-out recorded counterparts. The traffic candidate pool, future timestamps, renderer, and tested policies are fixed across the pair. This control separates completion-related changes from between-source variation without treating the synthesized views as real reference observations.

C Replay and Evaluation Protocol

C.1 Repeated-Anchor Paired Non-Reactive Replay

NAVSIM performs non-reactive policy evaluation from an initial recorded observation (Dauner et al., 2024). Generated-world evaluation additionally requires the policy observation to correspond to the traffic state used for scoring and the policy response to be measured at multiple stages of the generated future. We therefore evaluate each policy at a fixed set of temporal anchors while keeping the generated observation–state stream unchanged.

Let condition $\gamma \in \{\mathrm{real}, \mathrm{rec}, \mathrm{gen}\}$ provide synchronized observations $O_n^{(\gamma)}$ and traffic states $X_n^{(\gamma)}$ for scenario n. At scoring anchor t_k, policy a receives only causal observations and permitted context:

$$\widehat{\tau}_{n,a,k}^{(\gamma)} = \pi_a\Big(O_{n,\le t_k}^{(\gamma)}, C_{n,\le t_k}^{(\gamma)}\Big). \tag{27}$$

The predicted plan is transformed to the common scene frame and evaluated against the traffic state from the same condition:

$$m_{n,a,k}^{(\gamma)} = \mathcal{M}\Big(\widehat{\tau}_{n,a,k}^{(\gamma)}, X_{n,t_k:t_k+H}^{(\gamma)}, G_n^{(\gamma)}\Big), \tag{28}$$

where $G_n^{(\gamma)}$ denotes evaluator-side map and context support. This binding prevents a plan inferred from generated observations from being scored against an unrelated traffic continuation.

The replay stream is fixed before policy execution:

$$\Big(O_{n,t_{k+1}}^{(\gamma)}, X_{n,t_{k+1}}^{(\gamma)}\Big) = \mathcal{R}_n^{(\gamma)}(t_{k+1}),\ \mathcal{R}_n^{(\gamma)} \text{ is independent of } \widehat{\tau}_{n,a,k}^{(\gamma)}. \tag{29}$$

Policy outputs are therefore scored but are not written back into later observations or traffic states. Paired comparisons keep the scenario, policy, anchor, horizon, sensor protocol, and metric configuration fixed while changing only γ. Repeated anchors provide temporal coverage without turning the replay into a closed-loop rollout.

Computational considerations. Closed-loop generative simulation requires policy inference, traffic update, rendering, and observation feedback at each control step. Current multi-view generation is already expensive. Generating 500 scenarios in our pipeline takes more than 30 hours on eight NVIDIA L40 GPUs. A DriveArena-style closed loop (Yang et al., 2025) also requires repeated synchronization and data exchange among the driving agent, traffic simulator, and renderer. Once future observations depend on policy actions, the rendered sequence becomes policy specific and cannot be reused across the six-policy panel. Our protocol instead generates each condition once and reuses the same observation–state stream for all policies and anchors.

Algorithm 2 Repeated-anchor paired non-reactive policy testing

Require: scenario condition $(O^{(\gamma)}, X^{(\gamma)})$, anchors $\mathcal{T}$, policies $\mathcal{A}$, horizon H
Ensure: response records $\mathcal{R}$
for all policy $a \in \mathcal{A}$ **do**
  reset policy state for the current scenario
  for all $t_k \in \mathcal{T}$ **do**
    construct the causal policy input at t_k
    infer $\widehat{\tau}_{a,k}^{(\gamma)}$ and transform it to the common scene frame
    evaluate against $X_{t_k:t_k+H}^{(\gamma)}$
    store key $(\text{scene}, a, t_k, H, \gamma)$
    advance to the next fixed anchor without modifying the replay stream
  end for
end for
join supported conditions by the common scene–policy–anchor key
return $\mathcal{R}$

C.2 Evaluation Configuration

Recorded-motion rendering fixes the recorded ego and background trajectories, whereas generated-motion rendering uses one shared alternative-future pool across renderers. Historical inputs, scenario identities, traffic candidates, and scoring timestamps are matched across the paired conditions. The main generated-motion renderer panel uses CAT-K as the shared traffic backend. SMART and VBD are evaluated separately in the long-horizon traffic-realism analysis.

All ego policies predict three-second plans at 0.5-second offsets under the same camera-only replay contract. DiffusionDriveV2 and GoalFlow use explicit zero/null-LiDAR compatibility tensors in both reference and generated conditions, and no real LiDAR enters the formal comparison. Every run records the traffic-generator, renderer, and ego-policy checkpoints together with camera ordering, calibration, image transformations, observation and trajectory timestamps, coordinate conventions, feature checkpoints, sample counts, collision geometry, and support masks. Independent source recordings form the resampling unit.

C.3 Evaluated Model Suite

This subsection summarizes the traffic-motion generators, seven visual renderers, and six ego policies used in the benchmark. The renderer suite consists of MagicDrive-V2, OpenDWM, Panacea, DreamForge, BaseDreamer, HorizonDrive, and MagicDrive. All models are adapted only through the common benchmark interfaces, while their learned parameters remain frozen unless explicitly stated otherwise.

C.3.1 TRAFFIC MOTION GENERATORS

CAT-K. CAT-K is the default generated-traffic backend for the paired renderer comparison. It performs closed-loop supervised fine-tuning of tokenized traffic models through Closest-Among-Top-K rollouts, mitigating the covariate shift that arises when an open-loop behavior-cloning policy is recursively rolled out in simulation (Zhang et al., 2025). In our benchmark, one frozen CAT-K candidate pool is shared across renderers so that visual generation, rather than traffic selection, is the controlled variable.

SMART. SMART represents vectorized maps and multi-agent trajectories as discrete motion tokens and applies a decoder-only Transformer to autoregressively predict the next token in the joint traffic sequence (Wu et al., 2024). Its tokenized formulation provides a scalable learned traffic prior and serves as an independent behavior-generation baseline in our long-horizon TRS analysis.

VBD. Versatile Behavior Diffusion (VBD) is a diffusion-based multi-agent simulator designed to generate scene-consistent and controllable traffic interactions (Huang et al., 2026). Unlike next-token traffic models, VBD samples joint behavior through iterative denoising and supports inference-time guidance. We include it in the horizon study to test whether TRS differentiates traffic generators with substantially different generative mechanisms.

C.3.2 SCENE RENDERERS

MagicDrive. MagicDrive is a latent-diffusion renderer for controllable street-view generation, conditioning the synthesis process on camera geometry and structured 3D scene information such as BEV maps and object boxes (Gao et al., 2024). We use it as an established multi-view diffusion baseline and map its native camera outputs to the benchmark's synchronized six-view interface.

MagicDrive-V2. MagicDrive-V2 is a high-resolution long-video generation model for autonomous driving with structured control (Gao et al., 2025). We evaluate the frozen released configuration under the same historical input, traffic trajectory, camera layout, and temporal protocol as the remaining renderers.

Panacea. Panacea targets panoramic and controllable driving-video generation with explicit emphasis on temporal and cross-view consistency (Wen et al., 2024). Its structured BEV conditioning and multi-view temporal modeling make it a representative controllable video-generation baseline for evaluating whether high perceptual quality also preserves scene-level testing evidence.

DreamForge. DreamForge is a motion-aware autoregressive framework for multi-view driving-video generation (Mei et al., 2024). It propagates generated context over time to support longer sequences while modeling inter-view and temporal dependencies. In our benchmark it is evaluated as a renderer under both recorded and generated traffic trajectories.

OpenDWM. OpenDWM is an open-source driving-world-model baseline used here for structured multi-view future rendering (Chen et al., 2024, Ni et al., 2025). We retain the released model and inference configuration and adapt only its input/output packaging to the common benchmark format. Its visual and downstream policy effects are then evaluated under the same traffic conditions as the remaining renderers.

BaseDreamer. BaseDreamer is the baseline driving-scene renderer released with DriveArena (Yang et al., 2025). It provides a complementary generative stack to the diffusion and autoregressive renderers above. All evaluations use the same frozen traffic trajectories, historical inputs, and camera protocol.

HorizonDrive. HorizonDrive is a self-corrective autoregressive world model for long-horizon driving simulation (Zhang et al., 2026). Its scheduled rollout recovery and teacher-rollout distillation are designed to reduce autoregressive drift over extended generation horizons. We include it to test whether long-horizon stabilization also translates into stronger observation quality, scene fidelity, and policy-testing effectiveness.

C.3.3 EGO DRIVING POLICIES

UniAD. UniAD is a planning-oriented end-to-end driving framework that unifies perception, prediction, and planning tasks through shared query-based representations (Hu et al., 2023). Its multi-stage scene understanding and planning stack provides a representative full-stack camera-based policy for evaluating how generated observations affect a tightly coupled autonomous driving pipeline.

VAD. VAD represents agents and map elements in a fully vectorized form and uses these instance-level structures as explicit constraints for end-to-end planning (Jiang et al., 2023). Its representation differs substantially from UniAD's multi-task query hierarchy, providing architectural diversity in the policy panel.

DrivoR. DrivoR builds on pretrained vision transformers and compresses multi-camera image features into camera-aware register tokens, which are consumed by lightweight trajectory-generation and trajectory-scoring decoders (Kirby et al., 2026). The scorer predicts interpretable planning sub-scores, making DrivoR a compact vision-centric policy with a different representation and selection mechanism from BEV-heavy baselines.

SparseDriveV2. SparseDriveV2 formulates multimodal planning around efficient scoring of a dense trajectory vocabulary (Sun et al., 2026). Its design increases candidate density while retaining tractable scoring, providing a strong candidate-selection policy for testing whether generated scenes affect trajectory ranking even when the underlying plan set is highly expressive.

DiffusionDriveV2. DiffusionDriveV2 uses truncated diffusion for multimodal trajectory generation and introduces reinforcement-learning constraints to suppress low-quality modes while preserving trajectory diversity (Zou et al., 2025). In our camera-only compatibility setting, unavailable LiDAR inputs are replaced by fixed zero/null tensors in both paired conditions so that the sensor adaptation itself does not vary across comparisons.

GoalFlow. GoalFlow generates multimodal plans with goal-conditioned flow matching and scores candidate goals and trajectories against the current scene (Xing et al., 2025). The flow-matching formulation provides a generative planning mechanism distinct from diffusion-based and vocabulary-scoring policies. As with DiffusionDriveV2, the same zero/null-LiDAR compatibility interface is used in both paired conditions.

C.4 ROUTE COMPLETION AND EPISODE SUMMARY

Let s_0 be the initial coordinate on the supported reference route, s_{end} the route endpoint, and $\mathcal{Q}$ the set of valid route projections of committed replay samples before termination. Distance-based route completion is

$$\mathrm{RC}_{\text{ep}} = \mathrm{clip}_{[0,1]}\left(\frac{\max_{q\in\mathcal{Q}} s_q - s_0}{s_{\text{end}} - s_0}\right), \qquad s_{\text{end}} > s_0, \quad \mathcal{Q} \neq \varnothing. \tag{30}$$

Replay episodes commit declared prefixes of predicted plans and stop at registered terminal conditions. Invalid route projections do not advance progress. Because the observation stream remains fixed and the ego action is not written back to the simulator, RC_{ep} is interpreted as a replay progress proxy rather than completion of a dynamically continuous closed-loop mission.

D SUPPLEMENTARY NUMERICAL RESULTS

D.1 SUPPLEMENTARY RENDERER DIAGNOSTICS

The main table already reports the headline Observation Quality, Scenario Fidelity, Policy Response, and QTES results. We retain three complementary diagnostics here to avoid repeating the full generator grid. FID provides an additional image-distribution measure, collision-window rate provides a direct policy-risk diagnostic, and plan jitter measures temporal planning stability. Recorded and CAT-K settings are shown in paired columns.

	Observation Quality		Policy Response			
	FID↓		CR_{win}↓		Jit↓	
Model	Record	CAT-K	Record	CAT-K	Record	CAT-K
Real video	–	–	0.135	–	2.621	–
MagicDrive-V2	31.235	31.797	0.164	0.214	2.808	2.838
OpenDWM	18.921	19.628	0.144	0.206	2.754	2.752
Panacea	**17.008**	**17.081**	0.153	0.211	2.793	2.808
DreamForge	27.294	27.742	**0.141**	**0.202**	**2.685**	2.712
BaseDreamer	36.130	36.328	0.148	0.206	2.728	**2.751**
HorizonDrive	29.711	30.388	0.152	**0.202**	2.707	2.754
MagicDrive	39.842	41.201	0.197	0.247	2.934	2.967

Table 4: **Supplementary diagnostics under recorded and CAT-K traffic.** Best and second-best generator results are highlighted in **coral** and blue, respectively.

The supplementary trends are consistent with the main analysis but serve a different purpose. Panacea obtains the lowest FID in both motion settings, whereas the main-table FVD and QTES are led by OpenDWM. This disagreement reinforces the distinction between a single visual-distribution metric and the broader test-oriented evaluation. The policy diagnostics also change with traffic condition, but Figure 4 already decomposes these effects across individual ego policies. We therefore avoid duplicating another full agent-by-renderer table in the appendix.

D.2 Agent-Specific Responses

The six-policy decomposition in Figure 4 emphasizes paired changes between recorded and generated traffic. Table 5 complements that figure by reporting the intrinsic safety, planning, and execution profiles of the six ego policies under the common camera-only replay protocol.

	Safety		Planning Behavior				Execution	
Ego policy	CR_{win} ↓	TTC_p ↑	DAC↑	Prog.↑	Comf.↑	Jit.↓	ISR↑	Lat.↓
UniAD	0.302	2.598	0.952	**0.960**	0.633	4.196	1.000	1.380
VAD	0.337	2.632	0.950	0.941	**1.000**	4.868	1.000	1.137
DrivoR	0.026	**2.967**	**0.999**	0.868	0.960	1.565	1.000	0.312
SparseDriveV2	**0.016**	2.960	**0.999**	0.874	0.998	1.707	1.000	**0.061**
DiffusionDriveV2	0.078	2.947	0.992	0.831	0.465	1.855	1.000	0.095
GoalFlow	0.051	2.954	0.991	0.869	0.602	**1.536**	1.000	0.646

Table 5: **Ego-policy diagnostics under the common replay protocol.** CR_{win} denotes collision-window rate. TTC_p, DAC, Prog., Comf., Jit., ISR, and Lat. denote plan-level time-to-collision, drivable-area compliance, progress, comfort, plan jitter, inference success rate, and latency, respectively. Best and second-best results are highlighted in **coral** and blue. ISR is not ranked because all policies attain the same value.

D.3 Supplementary Policy-Response Cases

Figures 9–12 provide four additional scene-level examples of the policy differences summarized in Figure 4. Each case shows synchronized six-view observations together with trajectories predicted by the six ego policies under the same scene condition. The examples further illustrate the architecture-dependent policy responses observed in the quantitative results.

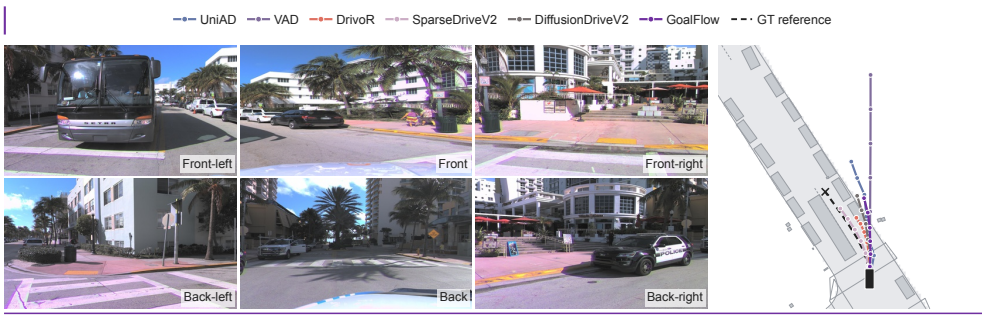

Figure 9: **Supplementary policy-response case (a).** Six-view observations and trajectories predicted by UniAD, VAD, DrivoR, SparseDriveV2, DiffusionDriveV2, and GoalFlow. The GT trajectory is shown as reference.

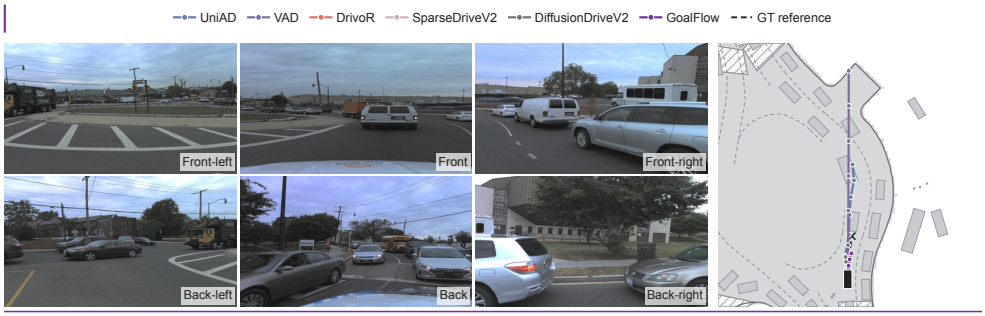

Figure 10: **Supplementary policy-response case (b).** Six-view observations and policy-specific trajectories under the same scene condition.

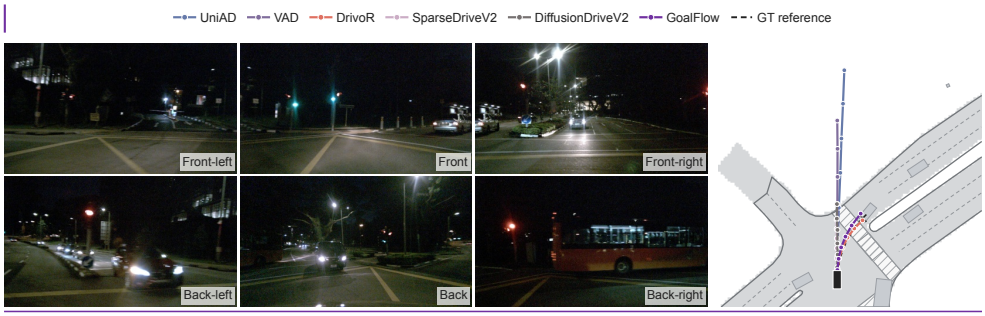

Figure 11: **Supplementary policy-response case (c).** Six-view observations and policy-specific trajectories under the same scene condition.

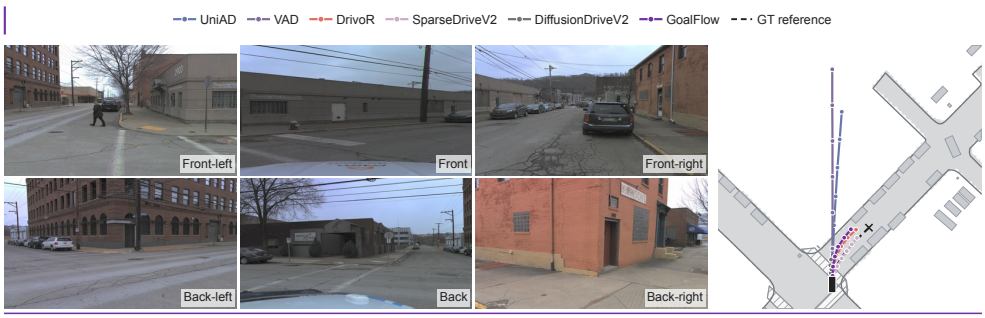

Figure 12: **Supplementary policy-response case (d).** Six-view observations and policy-specific trajectories under the same scene condition.

D.4 Supplementary Qualitative Cases

Figures 13 and 14 provide additional six-view temporal comparisons on nuScenes and nuPlan. The nuScenes case compares OpenDWM and BaseDreamer with the recorded reference, while the nuPlan case compares HorizonDrive and MagicDrive-V2. Across five future timestamps, the examples show how temporal drift and cross-view inconsistency develop over the generated sequence, complementing the DINO-T and TVC results.



Figure 13: **Supplementary six-view temporal comparison on nuScenes.** Reference, OpenDWM, and BaseDreamer outputs are shown at five future timestamps across six camera views.



Figure 14: **Supplementary six-view temporal comparison on nuPlan.** Reference, HorizonDrive, and MagicDrive-V2 outputs are shown at five future timestamps across six camera views.