

xCritic: A Continuous, Scene-Aware Driving Critic

Max Gier¹, Samir Joshi², Duke Lin² and Yunpeng Pan²

Abstract—Benchmarking autonomous driving behavior requires evaluating complex traffic scenarios where driving behavior is inherently context-dependent. Traditional rule-based metrics, such as the Extended Predictive Driver Model Score (EPDMS), offer physical grounding but remain incomplete, context-blind, and misaligned with human preferences. Emerging learning-based evaluators using Vision–Language Models (VLMs) demonstrate promising semantic context awareness. However, existing learning-based approaches typically rely on reference baselines or candidate trajectory pairs, lack a normalized metric directly calibrated to human driving behaviors, and struggle to accommodate scenarios where multiple driving choices are valid. To address these limitations, we propose xCritic, a continuous, scene-conditioned VLM driving critic to quantify driving assertiveness. The model predicts a standardized assertiveness scalar $z \sim \mathcal{N}(0, 1)$ calibrated directly against human baseline driving. Our framework operates in a standalone manner without requiring reference trajectories, while naturally accounting for multimodal driving choices. Furthermore, this normalized scalar provides directional interpretability (distinguishing passive hesitation from over-assertive risk), enables direct auditing of human demonstration, and behavior target setting.

I. INTRODUCTION

While modern autonomy stacks including end-to-end (E2E) approaches have achieved rapid progress [1], evaluating driving behavior in a manner aligned with expert human judgment remains a persistent challenge. Closed-loop simulation platforms like CARLA [2] and nuPlan [3] provide interactive testing, but traditional heuristic metrics remain the standard for benchmarking. However, classical imitation metrics—such as Average Displacement Error (ADE), Final Displacement Error (FDE) or Relative Progress to a reference—suffer from the “single-reference trap,” penalizing valid alternative maneuvers simply because they deviate from a single baseline demonstration [4], [5].

To address the limitations of displacement errors, benchmarks like NAVSIM [6] introduced rule-based evaluation frameworks, primarily the Predictive Driver Model Score (PDMS) and its extended variant, EPDMS [7]. EPDMS evaluates composite performance by combining hard safety checks (e.g., collisions, drivable area compliance) with weighted scores evaluating progress, lane keeping, and comfort. Despite their widespread adoption, rule-based metrics present three core limitations:

- **Incomplete and Context-Blind:** Operating on fixed geometric thresholds, EPDMS penalizes contextually appropriate behaviors [5].

- **Reference Target Dependence:** The progress metric in EPDMS calculates route advancement relative to a target distance (d_{ref}) generated by an external reference planner (e.g., PDM-Closed) or human demonstration:

$$EP = \min \left(1, \frac{d_{\text{ego}}}{d_{\text{ref}}} \right). \quad (1)$$

This design couples evaluation to a secondary reference system, making standalone trajectory scoring impossible when reference targets are unavailable.

- **Rigid Driving Style Assumptions:** Rule-based metrics treat deviations in progress or lane alignment as non-compliant, failing to acknowledge that driving style and multiple distinct trajectory options can be equally reasonable in complex traffic scenarios.

Recent learning-based approaches using Large Vision–Language Models (VLMs) have shown strong promise in integrating visual and semantic context into evaluation. Models such as DriveCritic [5] evaluate trajectory pairs through preference learning, while frameworks like DriveJudge [8] use VLMs to selectively gate rule functions. However, these learning-based methods still present key gaps: they require human reference baselines or paired candidate trajectories, and produce uncalibrated or coarse binary outputs (e.g., relative comparison) rather than a continuous metric calibrated against human driving distributions. Table I situates our proposed critic against these paradigms.

To resolve these challenges, we introduce a Continuous Scene-Conditioned Driving Critic, which we call *xCritic*. We deliberately scope xCritic to a single, context-dependent axis of driving quality: *progress*, expressed as driving *assertiveness*. Safety, comfort, and rule compliance are complementary axes that composite scores such as EPDMS aggregate to a single driving score; xCritic targets the progress component and is intended to be combined with other aspects of driving quality for a full assessment rather than to replace them altogether. Conditioning on Bird’s-Eye-View (BEV) renderings and a textual scene-and-state representation, our model directly predicts a continuous assertiveness scalar z grounded in standard normal human baseline driving style,

$$z \sim \mathcal{N}(0, 1), \quad (2)$$

where $z = 0$ represents expected mean human driver behavior for the given scene context, $z < 0$ indicates conservative or hesitant driving, and $z > 0$ indicates assertive driving.

¹Max Gier is with Gier Ventures, Aachen, Germany. This work was carried out under contract with Zoox.

²Samir Joshi, Duke Lin and Yunpeng Pan are with Zoox, Foster City, CA, USA.

Our proposed framework targets four core contributions:

- **Standalone Reference-Free Evaluation:** Evaluates single trajectories independently without needing counterfactual reference paths or external reference targets (d_{ref}), enabling evaluation on unpaired autonomous vehicle miles (“service miles”).
- **Human-Calibrated Normalized Metric:** Replaces coarse binary pass/fail labels with a standardized scalar directly interpretable relative to human driving distributions across scenarios.
- **Multimodal Style Accommodation:** Evaluates trajectory progress with respect to the scene context rather than rigid baseline trajectories, honoring valid alternative driving maneuvers.
- **Two-Tailed Directional Diagnosis:** Unlike existing progress metrics that only measure one-sided progress deficits, our signed z -score explicitly decouples over-conservative hesitation ($z \ll 0$) from risky over-assertiveness ($z \gg 0$).

Beyond trajectory scoring, this framework can act as an automated dataset auditor on human driving logs themselves, filtering out-of-distribution (OOD) human errors or interesting long-tail human driving behaviors.

II. RELATED WORK

A. Context-Blind Driving Evaluation

Imitation-based driving evaluation relies on geometric displacement metrics such as Average Displacement Error (ADE) and Final Displacement Error (FDE) [3], [9]. These imitation metrics measure distance against a single recorded human trajectory and are mostly applied to open-loop evaluation of learning-based planners. However, because urban driving is inherently multimodal, an autonomous agent can execute multiple valid, safe maneuvers that deviate geometrically from the recorded path. Imitation metrics treat all such deviations as errors, creating misalignment with actual driving quality [4]. Waymo’s long-tail benchmark [9] extends this approach by providing three trajectory candidates which are graded individually by human labelers. The similarity to those candidate trajectories is used to derive a final driving score.

To move beyond displacement errors, modern benchmarks like NAVSIM [6] and Pseudo-Simulation [7] established rule-based metrics, primarily PDMS and EPDMS. EPDMS combines hard safety penalties—No At-fault Collisions (NC), Drivable Area Compliance (DAC), Driving Direction Compliance (DDC)—with continuous scores for progress, lane keeping, and comfort.

While rule-based metrics offer spatial precision and physical grounding, they remain context-blind. Evaluations on the NAVSIM benchmark show that expert human drivers frequently receive low EPDMS scores due to context-appropriate nudges or cautious slowdowns [5]. Furthermore, EPDMS progress evaluation is explicitly tied to an external reference planner target (d_{ref}), preventing standalone trajectory evaluation and making it impossible to evaluate human

driving quality directly; style-aware variants swap this target for a human demonstration [10] but remain reference-bound.

B. Context-Aware Driving Evaluation

Large Vision–Language Models have demonstrated strong scene understanding and reasoning for autonomous driving [11]–[13]. Beyond autonomous driving, learning-based architectures including Vision–Language Models (VLMs) have been explored as general-purpose behavior critics and reward models for embodied agents [14], [15]. Existing context-aware driving evaluators broadly fall into three paradigms:

a) Learned Drive-Style (STAGE): To capture driving style in context, STAGE [16] learns a continuous, monotonic aggressiveness scalar from pairwise preferences. However, its supervision relies on hand-designed kinematic rules rather than semantic scene understanding, and the authors note cross-scenario comparability as an open limitation—a gap our population-calibrated z -score framework explicitly addresses.

b) Preference-Based Comparison (DriveCritic): DriveCritic [5] frames evaluation as pairwise trajectory preference selection ($A > B$). By fine-tuning a VLM on curated ambiguous driving pairs using Supervised Fine-Tuning (SFT) and Reinforcement Learning from Verifiable Rewards (RLVR via GRPO / DAPO [17], [18]), DriveCritic achieves high agreement with expert human annotators. However, preference models only provide relative rankings between two specific candidate trajectories; they do not produce an absolute, standalone scalar score for an isolated trajectory.

c) Rule-Gated Metric Invocation (DriveJudge): DriveJudge [8] addresses VLM spatial imprecision by introducing an agentic tool-invocation framework. A VLM analyzes scene context to dynamically enable or disable deterministic EPDMS rule functions (e.g., disabling Lane Keeping during an obstacle nudge while enforcing collision checks). While DriveJudge improves context sensitivity, it inherits EPDMS’s reliance on d_{ref} for progress evaluation and evaluates driving primarily through selectively composing the metric terms of EPDMS.

C. Positioning Against Existing Evaluators

Table I systematically compares evaluation paradigms across the dimensions most relevant to context-aware benchmarking. Preference VLMs and rule-gated VLMs both add semantic context awareness, but neither produces a normalized standalone scalar calibrated to human driving distributions, and both remain dependent on candidate pairs or external reference targets. Our continuous critic is the only approach that is simultaneously context-aware, multimodal, normalized to $z \sim \mathcal{N}(0, 1)$, directionally interpretable, fully standalone, and capable of auditing human baselines.

III. PROBLEM FORMULATION & PRELIMINARIES

A. Continuous Assertiveness Formulation

We address the limitations above by formulating driving evaluation as mapping a single candidate trajectory τ_{eval} and

TABLE I
COMPARISON OF DRIVING EVALUATION PARADIGMS.

Evaluation Paradigm	Context Awareness	Multimodal Styles	Normalized Continuous Score	Directional Diagnosis	Standalone (No d_{ref})
Imitation-Based [4]	Low	Low	No	No	No
Rule-Based [6], [7]	Low	Low	Yes (uncal.)	Partial	No (d_{ref})
Driving Style [16]	Low	High	No (uncal.)	Partial	Yes
Preference VLMs [5]	High	High	No (pairwise)	Partial	No (needs pair)
Rule-Gated VLMs [8]	High	High	No	Partial	No (d_{ref})
Ours (Continuous Critic)	High	High	Yes ($\mathcal{N}(0, 1)$)	Yes (two-tailed)	Yes (standalone)

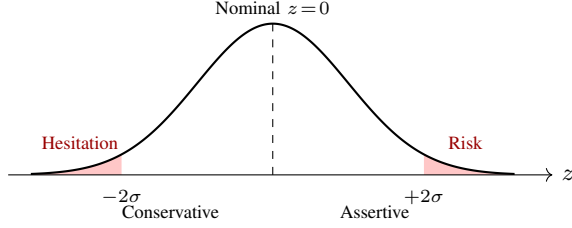


Fig. 1. The scene-conditioned assertiveness score $z \sim \mathcal{N}(0, 1)$. Nominal human driving sits at $z = 0$; the conservative tail ($z < 0$) ranges from cautious yielding to over-conservative hesitation ($z < -2$), and the assertive tail ($z > 0$) ranges from proactive navigation to over-assertive risk ($z > +2$).

its scene context X directly to a continuous assertiveness scalar z :

$$f_{\theta}(X, \tau_{\text{eval}}) \rightarrow z \in \mathbb{R}. \quad (3)$$

We ground z in a standard normal distribution representing context-conditioned human baseline driving, $z \sim \mathcal{N}(0, 1)$, so that z represents standard-deviation units (σ) from expected mean human driving behavior for scene context X (Figure 1).

Mathematically, the z -score partitions trajectory behaviors into distinct continuous regimes:

- **Nominal Human Driving** ($z \approx 0$): Trajectories that match normal human speed profiles, yielding behaviors, and path choices under the current road topology and traffic density.
- **Conservative Tail** ($z < 0$): *Cautious Progression* ($-1.0 \leq z < 0$) captures contextually appropriate deceleration (e.g., slowing for an incoming merge or yielding to pedestrians), while *Over-Conservative Hesitation* ($z < -2.0$) captures unnecessary stopping, lingering at green traffic lights, or unwarranted passivity that obstructs traffic flow.
- **Assertive Tail** ($z > 0$): *Proactive Navigation* ($0 < z \leq 1.0$) captures efficient progress, decisive merging, or minor lateral nudges around parked obstacles, while *Over-Assertive Risk* ($z > +2.0$) captures excessive speed or assertive cut-offs with potential safety impact.

B. Multimodal Input Representation

The critic evaluates a candidate future trajectory τ_{eval} by conditioning on a multimodal historical scene representation X . Let a scenario consist of a history of $h = 15$ observed ticks (3 seconds) and a candidate future horizon of $f = 20$

ticks (4 seconds) sampled at 5 Hz ($\Delta t = 0.2$ s). At each historical tick t , the scene observation X_t combines a visual BEV semantic map, a textual description of surrounding agents and scene properties, and the ego-vehicle kinematic state $s_t = (x, y, v, \delta, \psi)$ comprising 2D position, speed, steering angle, and yaw angle.

We summarize the full historical observation sequence $X = \{X_1, \dots, X_h\}$ into a compact context embedding $c = \text{EMB}_{\phi}(X)$ using our backbone. This summary vector c encapsulates the scene and historical ego behavior. Given this context vector c , xCritic models the conditional density $p_{\theta}(\tau_{\text{eval}} | c)$ of expert human driving over the evaluation horizon. The assertiveness of any candidate trajectory τ_{eval} is subsequently characterized by where it falls within this learned distribution.

IV. CONTINUOUS DRIVING CRITIC (METHODOLOGY)

xCritic is composed of (1) frozen vision-language encoders, (2) a latent world model that produces a scene summary, (3) a flow-matching trajectory decoder that models expert actions, and (4) a conditional variational assertiveness encoder.

A. Inputs, World Model, and Trajectory Decoder

Each history tick is encoded by two frozen, pretrained backbones: a Vision Transformer (ViT) [19] image encoder and a multilingual General Text Embeddings (GTE) text encoder [20] for the structured scene description. We take the ViT class token and the text token as per-tick visual and textual features. Freezing the encoders keeps the perception representation stable and confines learning to the world and scoring modules.

A JEPA-style [21], [22] latent world model fuses the per-tick visual and textual features and predicts tick-wise latent future world states. Based on the observed past and the latent future predictions a scene summary vector $c \in \mathbb{R}^{768}$ is derived. Conditioned on the world model’s outputs, a DiT-style [23] trajectory decoder models the distribution of expert future driving, $p_{\theta}(\tau | c)$, with conditional flow matching [24].

B. Conditional Variational Assertiveness Encoder

To obtain an explicit, interpretable assertiveness axis we incorporate a conditional variational autoencoder

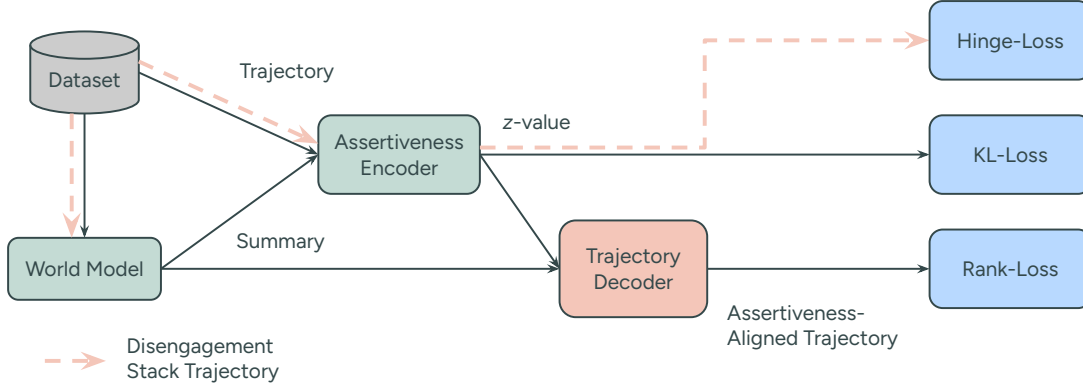


Fig. 2. The assertiveness encoder (CVAE) and its training signals. The encoder maps the ground-truth trajectory and the world-model summary to a scalar style latent (the z -value of human assertiveness); conditioning the trajectory decoder on this latent yields an assertiveness-aligned trajectory. The trajectory decoder is optional during inference but helpful for introspection. Three losses shape the latent: $KL/population$ calibration to $\mathcal{N}(0, 1)$, a monotonic *rank* loss on decoded progress, and a *hinge* that pushes resimulated disengagement negative trajectories (dashed path) toward the appropriate tail ($\pm 2\sigma$).

(CVAE) [25]. The assertiveness encoder maps the ground-truth trajectory and the scene summary to a *scalar* latent distribution $(\mu, \log \sigma^2)$ (Figure 2). A reparameterized sample is passed through a style embedder and added to the scene summary, so the trajectory decoder generates an assertiveness-conditioned trajectory

$$\hat{\tau} \sim p_{\theta}(\tau | c, z), \quad z = \mu + \sigma \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1). \quad (4)$$

The style latent is regularized so that z maps directly to a percentile of the human-driver population. Rather than the per-example Kullback–Leibler (KL) divergence, we match the batch statistics of the latent to $\mathcal{N}(0, 1)$: the population mean is pushed to zero and the total population variance $\text{Var}[\mu] + \mathbb{E}[\sigma^2]$ is pushed to one, with a soft penalty keeping per-example variances well-conditioned. A monotonic ranking loss enforces that decoded progress grows with z : for sorted samples $z_1 < \dots < z_n$ the realized progress must be monotonically increasing up to a margin. Figure 2 illustrates these training signals, including how disengagement negative examples are pushed toward the tails (e.g. $\pm 2\sigma$) depending on whether they under- or over-progressed relative to the human. Disengagement events supply out-of-distribution negative driving examples that a safety operator logged. Those events are resimulated starting from the time of disengagement and used during training and evaluation. The negative hinge relocates the assertiveness mean of a resimulated disengagement toward $\pm 2\sigma$. The sign is set by whether the simulated behavior under- or over-progressed compared to the human driver’s trajectory post-disengagement.

C. Joint Training

The modules are trained jointly. The total loss combines the world-model prediction loss, the flow-matching loss, the CVAE population and rank losses, a weak hinge that nudges disengagement negatives toward the assertiveness tails, and auxiliary consistency terms.

D. Two-Tailed Directional Diagnosis & Standalone Scoring

The critic assigns a z -value to any (scene, trajectory) pair. Given the past, the world model produces the summary c , and the trajectory is placed on the calibrated assertiveness axis of Section IV-B. Because the axis is population-calibrated to $\mathcal{N}(0, 1)$, z maps monotonically to an assertiveness percentile $\Phi(z) \in [0, 1]$ via the standard normal cumulative distribution function Φ ; we report the equivalent but more precise z -value throughout. $z \approx 0$ is human-like, $z \ll 0$ indicates under-progression (cautious), and $z \gg 0$ over-progression (aggressive). Importantly, z is a property of a *trajectory in a scene*: the same critic scores any trajectory standalone, regardless of who produced it, requiring no reference planner target d_{ref} or paired counterfactual.

The key to evaluation is *which* trajectory we score. For each scenario we can apply the critic to two trajectories—one from the recorded on-road log (LOG) and one from simulation (SIM):

- z_{LOG} (the LOG score): the assertiveness of the **recorded expert human driver**, taken from the on-road log of the scenario.
- z_{SIM} (the SIM score): the assertiveness of the **autonomous vehicle (AV) stack** re-simulated on the exact same scenario.

Scoring z_{SIM} alone (the standalone mode) asks whether the stack looks unusually cautious or assertive in absolute terms. Scoring z_{LOG} alone carries little information about stack failures, but helps evaluate the driving style and behavior of the human baseline driver. When a paired reference is available, the two scores can additionally be compared on the shared axis: the recorded human thus may serve as a *per-scenario baseline of the assertiveness the scenario demanded*, against which the stack is judged.

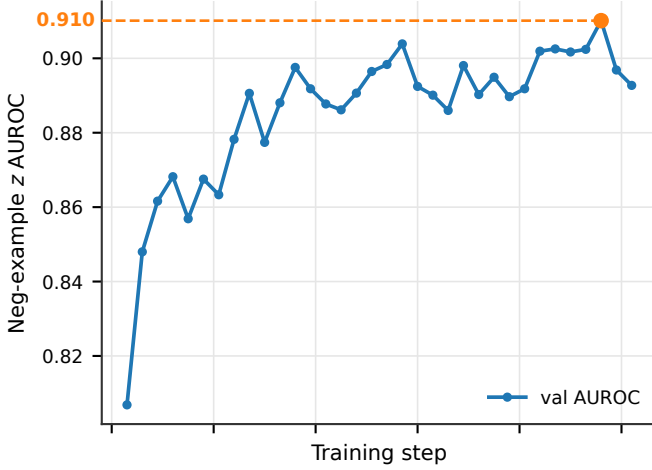


Fig. 3. Standalone, cross-scenario anomaly separation on the validation set during training. The threshold-free rank AUROC with which the single-trajectory magnitude $|z|$ distinguishes disengagement negatives from normal driving—pooled across thousands of different scenarios—rises to 0.91. No classifier or AUROC objective is optimized.

V. EXPERIMENTS & RESULTS

A. Experimental Setup & Datasets

We evaluate on a large fleet corpus of roughly 1M 10-second scenarios spanning nominal human driving, challenging long-tail maneuvers, and disengagement events. Each scenario is *paired*: we score both the recorded expert human log (LOG, z_{LOG}) and the autonomous stack re-simulated on the identical scenario (SIM, z_{SIM}) on the shared, human-calibrated assertiveness axis $z \sim \mathcal{N}(0, 1)$.

Throughout, our reference baseline is the *progress (distance) ratio* similar to the EPDMS-style progress scoring. In contrast to (1) we clamp the ratio to $[0, 2.0]$ to allow capturing overprogressing beyond the human’s baseline progress.

$$r = \min \left(2, \frac{d_{\text{ego}}}{d_{\text{ref}}} \right). \quad (5)$$

Our evaluation addresses four key questions: (i) Applied to a *single* trajectory with no baseline, does the calibrated axis encode a genuine notion of abnormal driving that stays comparable *across* scenarios (Sec. V-B)? (ii) Under a fixed triage budget, does xCritic surface undesirable progression more precisely than the progress ratio (Sec. V-C)? (iii) *Why* does a single reference ratio fail where a scene-conditioned axis succeeds (Sec. V-F)? (iv) Does scoring LOG and SIM on one axis yield an actionable driving improvement target for development and driving style enhancements (Sec. V-E)?

B. $|z|$ Is a Standalone, Cross-Scenario Anomaly Score

Our primary finding is that the calibrated axis grades a *single* trajectory, with no reference, on a scale that remains comparable across scenarios. The axis is shaped by only three signals (Sec. IV-B): population calibration to $\mathcal{N}(0, 1)$, a monotonic rank loss on decoded progress, and a deliberately coarse hinge $\mathcal{L}_{\text{neg}}$ that relocates a flagged disengagement’s latent mean toward $\pm 2\sigma$ (the sign set by whether it under- or over-progressed). Critically, we *never* train a binary quality

TABLE II

PROGRESS-EVENT TRIAGE UNDER A FIXED BUDGET OF 1,000 CANDIDATES PER DIRECTION AND MINER. xCRITIC IS MORE PRECISE ON BOTH TAILS, AND THE COMBINED MINER IS THE MOST PRECISE.

Tail	Miner	Precision
Conservative	Progress ratio	0.447
Conservative	xCritic	0.521
Conservative	xCritic + ratio gate	0.573
Assertive	Progress ratio	0.167
Assertive	xCritic	0.272
Assertive	xCritic + ratio gate	0.302

classifier, never optimize the negative-vs-normal AUROC, and give negatives no per-tick or discriminative label—the hinge only nudges a mean, and only for disengagements with an available progress sign. As a diagnostic we monitor the rank AUROC with which the latent magnitude $|z|$ separates disengagement negatives from normal driving on the validation set; Figure 3 shows it climbing throughout training to 0.91.

This number is a deliberately hard test of our global normalization claim. The AUROC is a threshold-free rank statistic—the probability that a randomly drawn disengagement carries a larger $|z|$ than a randomly drawn normal sample—and we compute it *across scenarios*, pooling negatives and normals from thousands of different scenarios and asking whether one global ordering on $|z|$ separates them. This is precisely the comparison that preference critics such as DriveCritic [5] do not attempt: they grade *pairs* of trajectories *within a single scenario*, where the shared context cancels and only a relative, in-scene ordering is required. A cross-scenario ranking affords no such cancellation: a $|z| = 2$ in an easy scene must denote the same degree of abnormality as a $|z| = 2$ in a hard one, or the pooled ordering decays toward chance. That the AUROC for $|z|$ reaches 0.91 across scenes is therefore direct evidence that the per-scene calibration transfers to a shared population scale—the same discrimination task DriveCritic solves in the easier within-scene, paired setting, xCritic solves in the harder standalone, cross-scenario setting.

Two properties follow. First, because $|z|$ is computed from a single trajectory graded individually, the score is *reference-free*: it needs neither a paired human counterpart nor an external target d_{ref} , which is what makes it applicable to unpaired autonomous service driving. Second, the value is directly operational: at 0.91, drawing one disengagement and one normal sample at random, the disengagement is ranked the more abnormal 91% of the time, with no threshold chosen.

C. Progress-Event Triage

The standalone axis also sharpens the paired, operational task of mining undesirable out-of-distribution driving events under a fixed *budget*: a reviewer inspects a bounded number of scenarios, so the objective is to maximize genuine findings per batch, not recall. Under an equal budget of

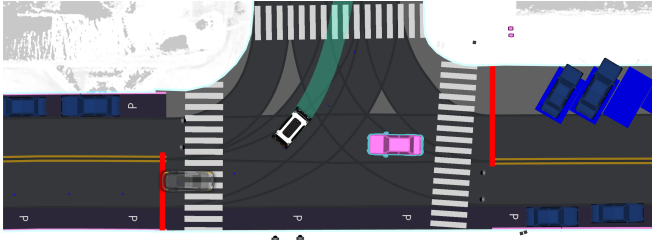


Fig. 4. Over-progression example (assertive tail). The stack (white) advances into a closing gap more aggressively than the human baseline supports for this scene, landing in the upper tail ($z \gg 0$) of the same two-sided score.

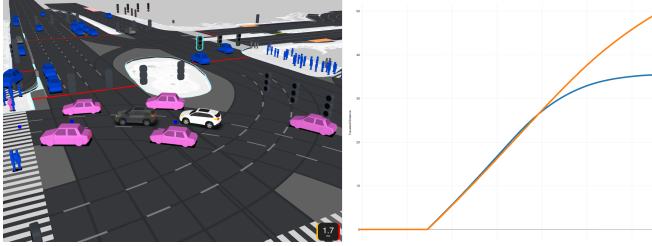


Fig. 5. Endpoint bias of the progress ratio. *Left*: the stack (white) comes to a full stop mid-intersection while the recorded human proceeds. *Right*: traveled distance vs. time for the human (orange) and the stack (blue); the curves diverge as the stack halts and its distance plateaus, yet the endpoint ratio remains $r > 0.7$, masking the stall. xCritc scores the same behavior at $z < -2$, correctly flagging severe traffic disruption the ratio misses.

1,000 candidates per direction, we compare two miners: the **progress ratio** takes the tails of $r < 0.5$ conservative, $r > 1.5$ assertive), while **xCritc** ranks the scenarios directly from both tails. Expert reviewers confirm each surfaced candidate; precision is confirmed true positives over the reviewed budget.

xCritc converts the same budget into materially more confirmed findings (Table II): on the conservative tail, precision rises $0.447 \rightarrow 0.521$ (+17% relative); on the harder over-progression tail it rises $0.167 \rightarrow 0.272$ (a $1.6\times$, +63% gain). It also recovers failures a fixed threshold cannot: 78 confirmed conservative true positives lie *above* the $r < 0.5$ cutoff, where raising the threshold only admits more false positives.

Crucially, in this paired setting a baseline *is* available, so the two signals compose rather than compete: gating the critic’s candidates with the progress ratio tails ($r < 0.5$ conservative, $r > 1.5$ assertive) raises the precision of the surfaced batch beyond either signal alone, to 0.573 on the conservative tail and 0.302 on the assertive tail (Table II)—making xCritc a drop-in enhancement to existing ratio-based triage, not merely a replacement.

D. Qualitative Two-Tailed Examples

Figures 5 and 4 show representative failures the critic surfaces, one per tail, making the two-sided score concrete.

Figure 5 shows an instructive under-progression case ($z \ll 0$) that also exposes the *endpoint bias* of the progress ratio. On a protected left turn (green arrow), the stack

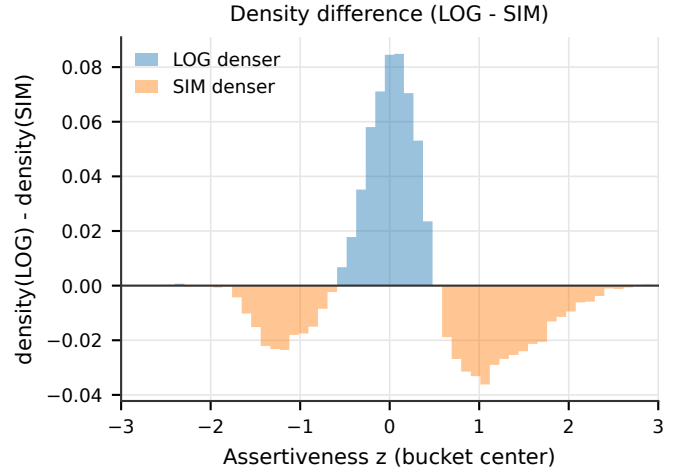


Fig. 6. Per-bucket assertiveness density difference (LOG — SIM) over the full $\sim 1\text{M}$ -scenario corpus. The human log concentrates near $z \approx 0$ (blue excess), while the autonomous stack carries excess mass in both out-of-distribution tails (orange). Closing this tail gap is a direct, directional “drive more human-like” target.

(white) comes to a *full stop* mid-intersection while the recorded human proceeds smoothly. The traveled-distance curves (right) track together early, before the stack’s traveled distance plateaus as it halts while the human’s distance keeps climbing. Because the ratio is built from endpoint distances it depends heavily on where the window begins and ends—our pipeline already evaluates it at the tick where log and sim start to diverge—yet even so the end-of-scenario ratio stays $r > 0.7$, so the stall looks acceptable and is masked. xCritc, scoring the fixed-horizon behavior against the scene, assigns $z < -2$ and flags the stop as severely conservative. This directly illustrates the endpoint bias of ratio-based metrics: a ratio built from start/end distance can hide a mid-scenario failure whenever total displacement recovers, which is exactly why a ratio-based miner suffers from limited recall (Sec. V-C).

In Figure 4, the stack over-progresses ($z \gg 0$) into a closing gap, landing in the upper tail of the same axis. Both failure modes are captured by one continuous score and become unambiguous once referenced against the human baseline.

E. The Human–Sim Gap as a Concrete Development Target

Scoring both LOG and SIM on the same calibrated axis turns evaluation into an actionable objective. Figure 6 plots, per z -bucket over the full corpus, the density difference $\text{density}(\text{LOG}) - \text{density}(\text{SIM})$. The human log is tightly concentrated near $z \approx 0$ (the calibration target), whereas the stack exhibits systematic mass in both out-of-distribution tails—over-conservative ($z < 0$) and over-assertive ($z > 0$). This makes the human-likeness gap of the stack’s driving behavior a measurable goal of shrinking the tail excess. Because the axis is calibrated, the gap is quantifiable bucket-by-bucket and is *directional*—something an uncalibrated score or a pass/fail label cannot provide, and something a single progress ratio cannot represent either.

F. Ratio-Conditioned Audit: Exposing Single-Reference Bias

The preceding results show the xCritic’s ability to evaluate driving behavior and surface undesirable behaviors on a per scenario level; subsequently we show distributional insights exposing a structural flaw of any single-reference metric. For this, we condition the LOG and SIM assertiveness distributions on the progress ratio itself.

Figure 7 restricts scenarios to strong under-progression ($r < 0.3$) as measured by the progress ratio. The stack is indeed genuinely conservative (SIM posterior mean $z = -1.42$), whereas the recorded human sits near the population center, tending to slightly increased assertiveness (LOG mean $z = +0.32$). The xCritic separates the scene-appropriate human behavior from the truly timid stack.

Figure 8 restricts to strong over-progression ($r > 1.7$), which the ratio flags as over-assertive. The picture inverts: the *stack is centered and human-like* (SIM mean $z = +0.24$), while the *human baseline is overly conservative* (LOG mean $z = -0.76$). The large ratio is an artifact of a timid human reference, not of stack over-assertiveness. This is the mechanism behind the low assertive precision of the progress ratio (Table II): human drivers do not always maximize progress [5], so a ratio computed against them manufactures phantom over-progression whenever the human under-progressed. Scoring the trajectory against the *scene* rather than the human path disambiguates those behavioral differences and lets us audit the human baseline itself.

This same $r > 1.7$ regime is our clearest evidence of multimodal style accommodation. A conservative human log is an inherently valid human decision; measured against it, FDE or the progress ratio report a large divergence for the (faster) stack, yet xCritic scores a high fraction of these scenarios as entirely human-normal ($z \approx 0$). The critic thus honors a valid alternative maneuver instead of penalizing it as a deviation from one reference.

None of these distinctions—scene-appropriate versus timid at low ratio, human-normal versus conservative at high ratio, or a standalone anomaly score—are available from a single progress ratio value. They require a continuous, scene-conditioned, human-calibrated axis on which any trajectory, human or autonomous, can be evaluated individually and compared.

VI. DISCUSSION

Scope and integration with full driving-quality metrics.

xCritic is deliberately confined to the progress/assertiveness axis of driving quality. Rule-based composite scores such as EPDMS remain strong on the axes they were designed for—no at-fault collisions and safety, drivable-area and direction compliance, and comfort—but are context-blind precisely on progress, which is where a scene-conditioned, human-calibrated axis adds the most. A complete evaluation of driving quality will therefore *combine* xCritic’s assertiveness signal with these safety and comfort metrics. An important direction for future work is to characterize how these axes correlate: in particular, whether over-assertiveness ($z \gg 0$) is predictive of safety-relevant outcomes such as near-collisions

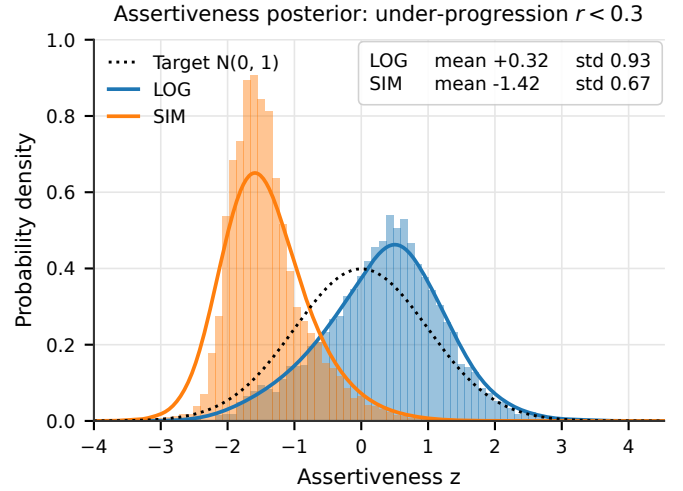


Fig. 7. Assertiveness posteriors conditioned on strong under-progression ($r < 0.3$). The autonomous stack is genuinely conservative (SIM mean $z = -1.42$), but the recorded human is mostly centered (LOG mean $z = +0.32$).

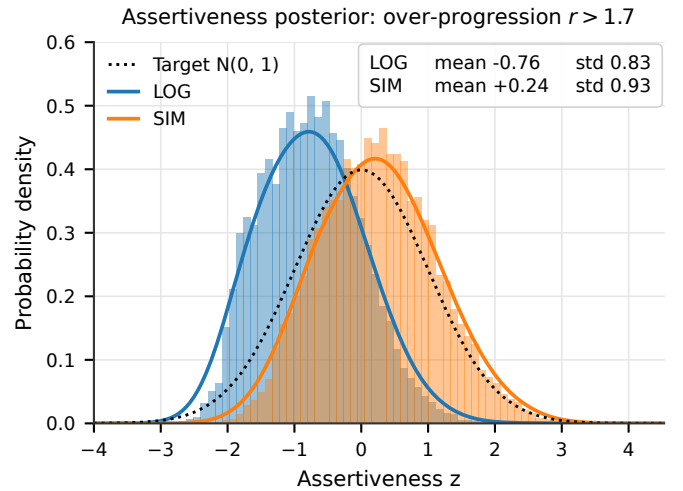


Fig. 8. Assertiveness posteriors conditioned on strong over-progression ($r > 1.7$). Here the stack is human-normal (SIM mean $z = +0.24$) while the human baseline is conservative (LOG mean $z = -0.76$): the large ratio reflects a timid human reference, not stack over-assertiveness—explaining the progress ratio’s poor assertive precision.

or hard interactions. If so, the assertiveness axis could serve as an early, standalone indicator of at-fault risk—flagging behavior that today’s rule-based safety checks only register once a violation has already occurred.

VII. CONCLUSION, LIMITATIONS, & OUTLOOK

A. Conclusion

We presented xCritic, a continuous, scene-conditioned VLM driving critic that scores trajectory assertiveness by combining a latent world model, a conditional variational assertiveness encoder, and a flow-matching trajectory decoder. We set out to (a) improve heuristic progress metrics with a scene-conditioned, calibrated, two-sided assertiveness score $z \sim \mathcal{N}(0, 1)$; (b) learn it from expert driving and disengagements; (c) detect out-of-distribution behavior more

precisely than the progress ratio; and (d) turn evaluation into an actionable, human-referenced target. We reached these goals: on a $\sim 1\text{M}$ -scenario fleet corpus the single-trajectory magnitude $|z|$ is a standalone, cross-scenario anomaly detector (0.91 AUROC); under a fixed triage budget the critic surfaces confirmed undesirable behavior more precisely than the progress ratio baseline on both tails (Table II), where the availability of a paired baseline further lets the critic and ratio signals be combined; and scoring LOG and SIM on one axis exposes a directional human-sim gap that a single reference ratio structurally cannot represent. These results concern the progress/assertiveness axis specifically, which we position as complementary to the safety and comfort components of existing composite scores.

B. Limitations & Future Work

The main limitation of our current approach is triage precision. Even with the combined critic-plus-gate miner, roughly 43% of conservative and 70% of assertive candidates are not confirmed as genuine failures (Table II), which we want to improve further. Part of this gap is limited information: xCritic deliberately conditions on BEV renderings and a structured scene description rather than raw camera input—which also keeps it applicable in closed-loop simulation—so the model’s information is limited to the quality of the scene representation and the encoders. The model also emits only a scalar score: a reasoning trace explaining *why* a scenario was scored as an outlier is not available yet. The complementary goal of a *baseline-free* evaluator is only partially achieved: the standalone $|z|$ anomaly score is already strong (AUROC 0.91 on disengagement negatives), but paired evaluation against the recorded human remains the more precise triage signal, indicating that a per-scenario human reference still carries signal that absolute scoring cannot fully recover. Finally, scoring is more expensive than a heuristic metric, which opens scaling concerns applying to all model-based approaches.

Future work targets these gaps directly: closing the standalone-paired precision gap, producing reasoning traces alongside the score, and incorporating the critic into a more holistic driving evaluation system alongside other driving evaluation dimensions such as safety and comfort.

REFERENCES

- [1] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, “End-to-end autonomous driving: Challenges and frontiers,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10 164–10 183, 2024.
- [2] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “CARLA: An open urban driving simulator,” in *Conf. Robot Learn. (CoRL)*, 2017, pp. 1–16.
- [3] N. Karnchanachari *et al.*, “Towards learning-based planning: The nuPlan benchmark for real-world autonomous driving,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024, pp. 629–636.
- [4] J.-T. Zhai, Z. Feng, J. Du, Y. Mao, J.-J. Liu, Z. Tan, Y. Zhang, X. Ye, and J. Wang, “Rethinking the open-loop evaluation of end-to-end autonomous driving in nuScenes,” 2023, arXiv:2305.10430.
- [5] J. Song, Z. Li, S. Lan, X. Sun, N. Chang, M. Shen, J. Chen, K. A. Skinner, and J. M. Alvarez, “DriveCritic: Towards context-aware, human-aligned evaluation for autonomous driving with vision-language models,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2026.
- [6] D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, A. Geiger, and K. Chitta, “NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024, pp. 28 706–28 719.
- [7] W. Cao, M. Hallgarten, T. Li, D. Dauner, X. Gu, C. Wang, Y. Miron, M. Aiello, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, A. Geiger, and K. Chitta, “Pseudo-simulation for autonomous driving,” in *Conf. Robot Learn. (CoRL)*, 2025, pp. 4709–4722.
- [8] X. Sun, K. Xie, J. Schmalfuss, D. Paschalidou, X. Zhang, S. Fidler, K. Chitta, and J. M. Alvarez, “DriveJudge: Rethinking autonomous driving evaluation with vision-language models,” 2026, arXiv:2606.17362.
- [9] R. Xu, H. Lin, W. Jeon, H. Feng, Y. Zou, L. Sun, J. Gorman, K. Tolstaya, S. Tang, B. White, B. Sapp, M. Tan, J.-J. Hwang, and D. Anguelov, “WOD-E2E: Waymo open dataset for end-to-end driving in challenging long-tail scenarios,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2026, pp. 3709–3718.
- [10] R. Hao, B. Jing, H. Yu, and Z. Nie, “StyleDrive: Towards driving-style aware benchmarking of end-to-end autonomous driving,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 6, 2026, pp. 4627–4635.
- [11] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, “DriveVLM: The convergence of autonomous driving and large vision-language models,” in *Conf. Robot Learn. (CoRL)*, 2024.
- [12] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, “DriveLM: Driving with graph visual question answering,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2024, pp. 256–274.
- [13] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, “OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 22 442–22 452.
- [14] L. Guan, Y. Zhou, D. Liu, Y. Zha, H. Ben Amor, and S. Kambhampati, “Task success is not enough: Investigating the use of video-language models as behavior critics for catching undesirable agent behaviors,” 2024, arXiv:2402.04210.
- [15] T. Lee, A. Wagenmaker, K. Pertsch, P. Liang, S. Levine, and C. Finn, “RoboReward: General-purpose vision-language reward models for robotics,” 2026, arXiv:2601.00675.
- [16] Z. Liu, X. Liu, Y. Zhang, and P. Huang, “STAGE: Style-controllable action generation for personalized autonomous driving,” *IEEE Robot. Autom. Lett.*, vol. 11, no. 2, pp. 2130–2137, 2026.
- [17] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, *et al.*, “DeepSeekMath: Pushing the limits of mathematical reasoning in open language models,” 2024, arXiv:2402.03300.
- [18] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, *et al.*, “DAPO: An open-source LLM reinforcement learning system at scale,” 2025, arXiv:2503.14476.
- [19] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [20] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, M. Zhang, W. Li, and M. Zhang, “mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP): Industry Track*, 2024, pp. 1393–1412.
- [21] Y. LeCun, “A path towards autonomous machine intelligence,” Open-Review, 2022.
- [22] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, “Revisiting feature prediction for learning visual representations from video,” 2024, arXiv:2404.08471.
- [23] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.
- [24] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [25] K. Sohn, H. Lee, and X. Yan, “Learning structured output representation using deep conditional generative models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015.