



xCritic: A Continuous Scene-Aware Driving Critic

Max Gier, 2026-10-07



Autonomous Driving Evaluation

Developing autonomous driving software requires thorough behavior evaluation. Common benchmarks rely on:

- **Imitation-Based Metrics (e.g. Displacement Errors)**
- **Rule-Based & Heuristic Metrics (e.g. EPDMS)**

Novel approaches achieve context-awareness through learned driving critics.

xCritic focuses on the **Progress Dimension** to evaluate the right amount of **Assertiveness** in a given scenario.

Subscore	w_m	Range
No at-fault Coll. (NC)	-	$\{0, \frac{1}{2}, 1\}$
Drivable Area Compl. (DAC)	-	$\{0, 1\}$
Driving Direction Compl. (DDC)	-	$\{0, \frac{1}{2}, 1\}$
Traffic Light Compl. (TLC)	-	$\{0, 1\}$
Ego Progress (EP)	5	$[0, 1]$
Time to Collision (TTC)	3	$\{0, 1\}$
Lane Keeping (LK)	2	$\{0, 1\}$
History Comfort (HC)	2	$\{0, 1\}$
Extended Comfort (EC)	2	$\{0, 1\}$

Table 1: **EPDMS**. Subscores, weights, and ranges.

Flaws of Hand-Crafted Progress Evaluation

Incomplete & Context-Blind

Standard benchmarks (EPDMS, PDMS) rely on fixed geometric thresholds and composite scores. They penalize contextually appropriate maneuvers treating safe flexibility as non-compliance.

Reference Target Coupling

EPDMS computes progress relative to an external reference planner with a single baseline: $EP = \min(1, (d_{ego} / d_{ref}))$



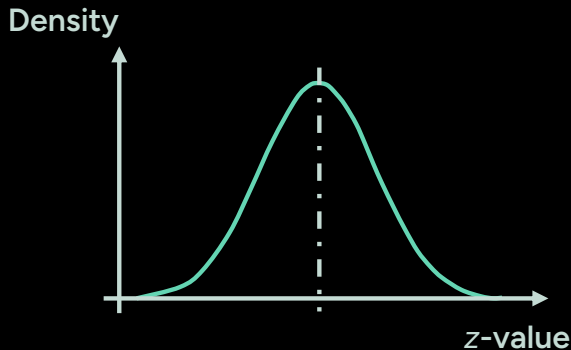
xCritic: Continuous Scene-Aware Driving Critic

Main Contributions



Standalone Scoring

Evaluates single trajectories without reference targets (d_{ref}) or paired counterfactual paths, enabling scoring of unpaired service miles.



Human Calibration

Standardized scalar $z \sim N(0, 1)$ directly translates to human population percentiles $\Phi(z)$, making scores globally comparable across scenes.

Multimodal Styles

Judges progress relative to semantic scene context, honoring valid alternative driving choices instead of penalizing divergence from one human log.



Two-Tailed Diagnosis

Explicitly decouples over-conservative hesitation ($z \ll 0$) from risky over-assertiveness ($z \gg 0$), unlike one-sided progress ratios.

xCritic: Continuous Scene-Aware Driving Critic

Main Contributions

Evaluation Paradigm	Context Awareness	Multimodal Styles	Normalized Continuous Score	Directional Diagnosis	Standalone (No d_ref)
Imitation Metrics	Low	Low	No	No	No
Rule-Based	Low	Low	Yes	Partial	No
Driving Style	Low	High	No	Partial	Yes
Preference VLMs	High	High	No	Partial	No
Rule-Gated VLMs	High	High	No	Partial	No
xCritic (Ours)	High	High	Yes ($z \sim N(0, 1)$)	Yes (two-tailed)	Yes (fully standalone)

Model Inputs & Outputs

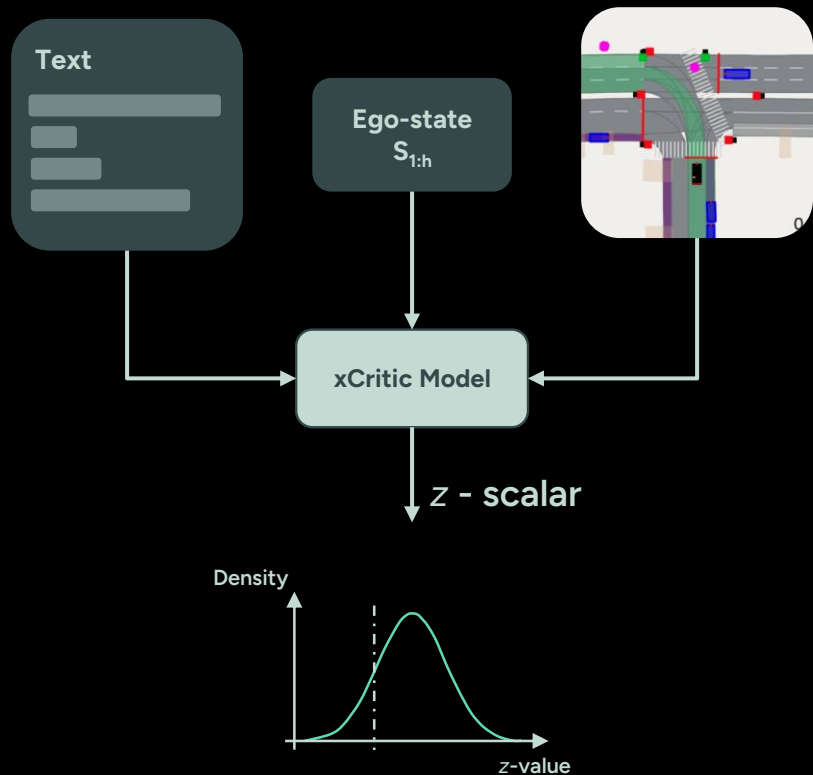
Model Inputs

The model receives a textual scene description, bird's-eye view rendering and ego-vehicle states over a fixed time horizon

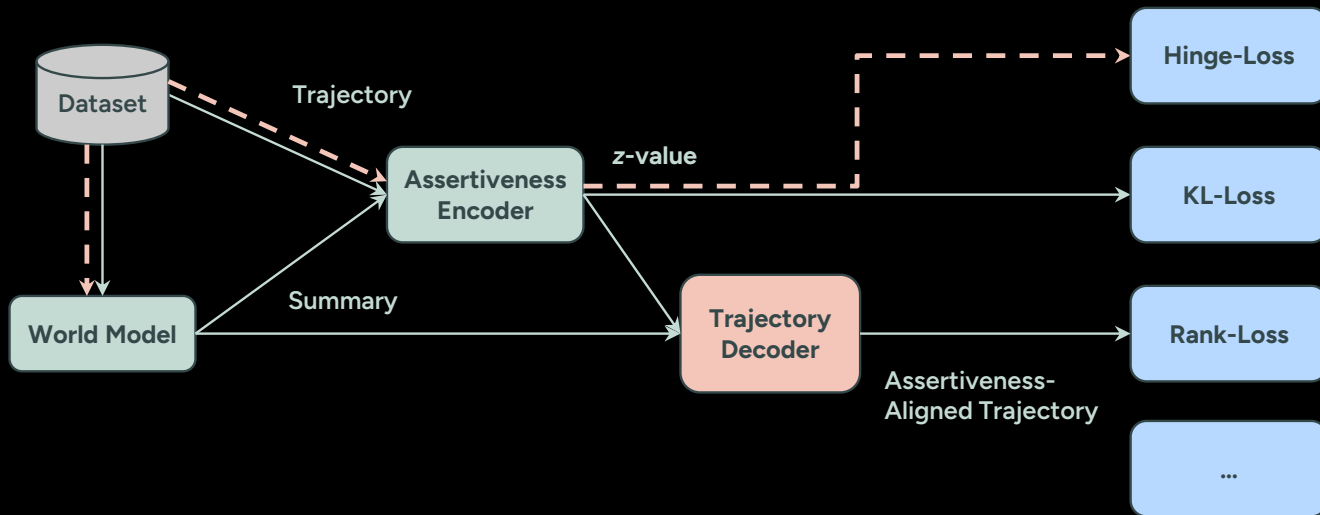
Model Outputs

Outputs a scalar z value as assertiveness prediction over the next 4 seconds

- $z \approx 0$ (average human assertiveness)
- $z \gg 0$ (over-assertive behavior)
- $z \ll 0$ (over-conservative behavior)



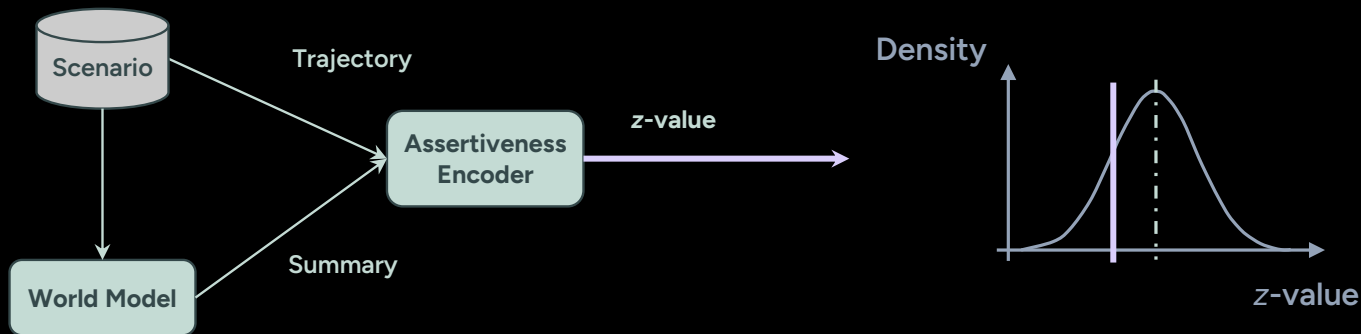
Architecture & Joint Training



— —> Disengagement
Stack Trajectory

Conditional Variational Assertiveness Encoder (CVAE): Maps trajectory and summary to latent assertiveness $z = \mu + \sigma\epsilon$, shaped by batch population loss to $N(0,1)$, monotonic rank loss, and negative disengagement hinge loss pushing failures to $\pm 2\sigma$.

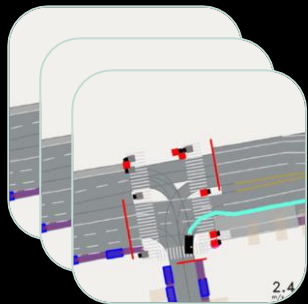
Architecture & Joint Training



Conditional Variational Assertiveness Encoder (CVAE): Maps trajectory and summary to latent assertiveness $z = \mu + \sigma\epsilon$, shaped by batch population loss to $N(0,1)$, monotonic rank loss, and negative disengagement hinge loss pushing failures to $\pm 2\sigma$.

Loss on Model Output

Batch of Expert
Human Driving



KL-Loss enforces
 $z_{batch} \sim N(0, 1)$

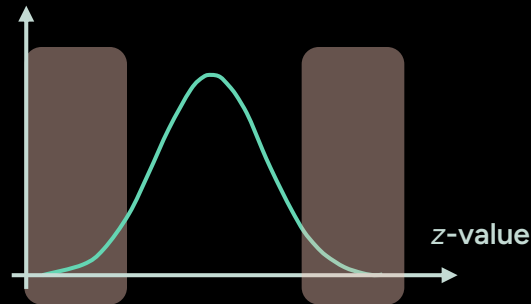
Resimulated
Disengagements



Post-Disengagement
Expert Trajectory

Post-Disengagement
Sim Trajectory

Density



Hinge-Loss enforces

$z_{disengagement} < -2$ or

$z_{disengagement} > +2$

Cross-Scenario Separation

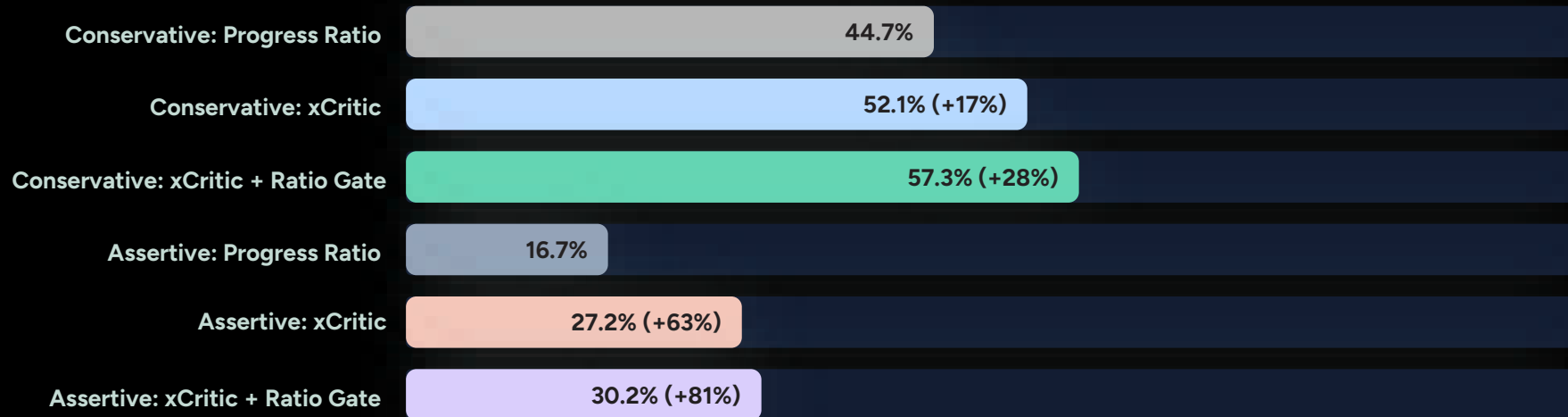
0.91

Disengagement AUROC

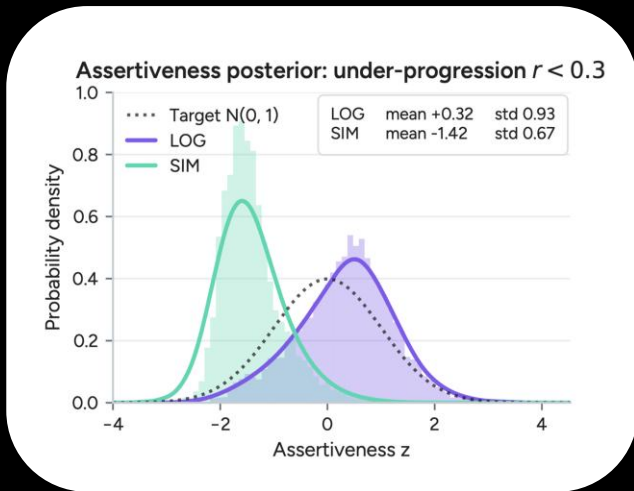
Global Standalone Discrimination

The single-trajectory latent magnitude $|z|$ separates real-world disengagements from human driving with a 0.91 AUROC across thousands of heterogeneous scenarios.

Undesirable Driving Detection - Precision

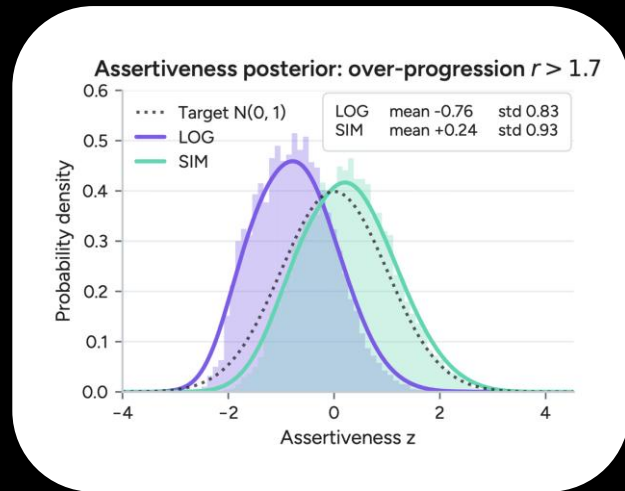


Human and Sim Assertiveness Distribution



Distributional Behavior Insights

Looking at distance ratio < 0.3 shows clear conservativeness for the simulation.



Human Baseline Bias

Human baseline does not always maximize progress for high-progress ratios [1]. xCritik allows auditing the human baseline itself.

xCritic: Continuous Scene-Aware Driving Critic

Distributional Difference

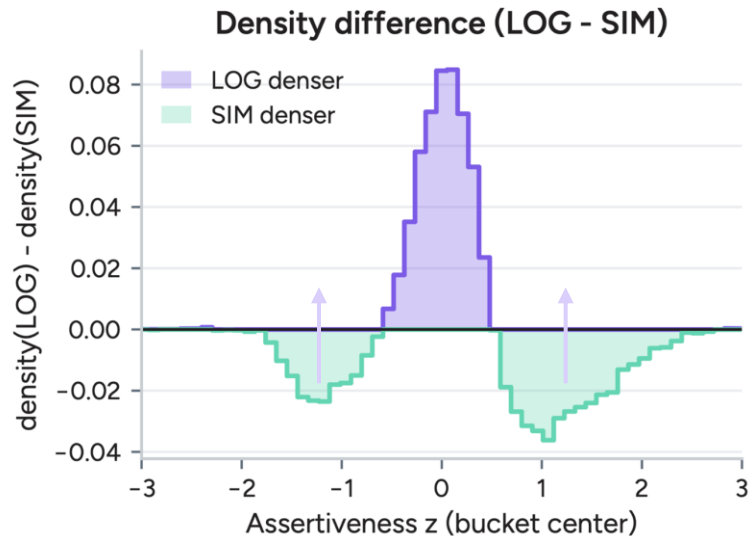
Identify Out-of-Distribution Behavior

We compare the LOG (human) and SIM distribution across roughly 1 million scenarios of 10s length.

Difference of distributions helps identify out of distribution groups compared to the baseline.

Behavioral Target Setting

The distributional difference can be used to set behavioral targets towards more human-like driving.



xCritic: Continuous Scene-Aware Driving Critic

Limitations & Future Work

Precision

Improve the precision and close the precision gap without baseline assertiveness

Reasoning

Improve introspection e.g. through reasoning traces

Complete Driving Evaluation

Incorporate xCritic's assertiveness signal into a more holistic evaluation system

Take a look at our Paper about this work!

xCritic: A Continuous, Scene-Aware Driving Critic

Max Gier¹, Samir Joshi², Duke Lin² and Yungpeng 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 quantifies composite performance by combining hard safety

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

$$EP = \min \left(1, \frac{d_{epo}}{d_{det}} \right). \quad (1)$$

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.



Join the Journey

zoox.com/careers





ZOOX



Thank You!