Scalable Offline Metrics for Autonomous Driving

Animikh Aich*, Adwait Kulkarni*, and Eshed Ohn-Bar[1]

***Abstract*— Real-World evaluation of perception-based planning models for robotic systems, such as autonomous vehicles, can be safely and inexpensively conducted offline, i.e., by computing model prediction error over a pre-collected validation dataset with ground-truth annotations. However, extrapolating from offline model performance to online settings remains a challenge. In these settings, seemingly minor errors can compound and result in test-time infractions or collisions. This relationship is understudied, particularly across diverse closed-loop metrics and complex urban maneuvers. In this work, we revisit this undervalued question in policy evaluation through an extensive set of experiments across diverse conditions and metrics. Based on analysis in simulation, we find an even worse correlation between offline and online settings than reported by prior studies, casting doubts on the validity of current evaluation practices and metrics for driving policies. Next, we bridge the gap between offline and online evaluation. We investigate an offline metric based on *epistemic uncertainty*, which aims to capture events that are likely to cause errors in closed-loop settings. The resulting metric achieves over 13% improvement in correlation compared to previous offline metrics. We further validate the generalization of our findings beyond the simulation environment in real-world settings, where even greater gains are observed.**

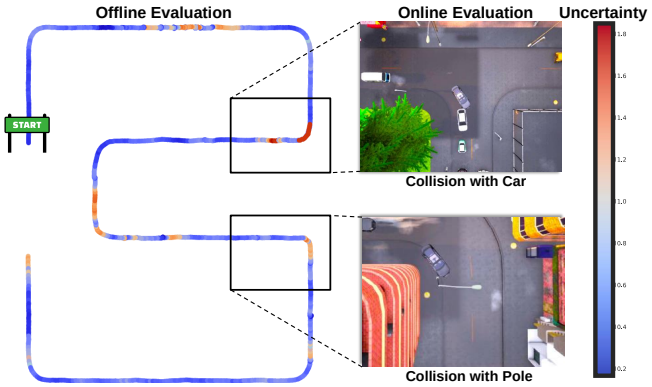

Fig. 1. **Closing the Gap Between Offline and Online Evaluation for Autonomous Driving.** We revisit the relationship between offline and online evaluation of vision-based autonomous driving policies. Our key insight is that model uncertainty-weighted errors (shown on the left) can be used to estimate errors in online settings (shown on the right), such as collisions and traffic infractions. We then devise a simple and scalable metric that can be applied over offline data without requiring high-quality perception information or various surrounding agent prediction models. We validate the effectiveness of the metrics in both simulated and real-world settings across diverse model sampling strategies, types of policy failures, and platforms.

I. Introduction

Can we estimate the online performance of a driving model without actively driving with it? Evaluating driving policy models that map camera input to a driving plan or controller action in real-world conditions is typically expensive and potentially hazardous. During model deployment, traffic infractions and collisions may occur, such as fatal accidents with pedestrians and vehicles [1], [2], Waymo's driving in the wrong lane and veering unsafely [3], and Tesla's Autopilot crashing into a truck on a busy highway [4].

To circumvent such safety-critical evaluations, researchers often use an offline approach. They leverage a computed *expected loss* over a pre-collected, real-world dataset. This method avoids the potentially dangerous and costly quantification of collisions and traffic infractions in live scenarios (Fig. 1). The efficiency in offline evaluation over a pre-collected driving dataset, where the driving model predictions are compared against safe ground-truth targets provided by a human driver (e.g., leveraging an L2 loss [5]–[13]), has recently propelled significant advancements in scalable autonomous driving models [7], [10], [14]–[19]. However, it remains unclear whether current offline evaluation practices adequately reflect actual on-road driving model performance, particularly for decisions that require highly robust and split-second reactions with life-threatening implications.

The reasons for the discrepancy between offline and online model performance are three-fold. First, real-world driving datasets are nearly entirely *uneventful*, with drivers rarely experiencing safety-critical events such as abrupt pedestrian crossings, near-collisions, or even turns and merging. The expected loss over the inherently imbalanced distribution of data can obscure performance under safety-critical events [20], [21]. Second, while automatic data sampling techniques can alleviate data imbalance issues [22]–[26], the standard L2 loss-based metric itself provides a crude measure for the ultimate task of safe and enjoyable driving [27], [28]. For instance, the L2 metric can remain identical among several maneuvers, e.g., merging to the left or right lane on a freeway, yet one decision may be perfectly safe while the other results in a collision, i.e., in the case of an adjacent vehicle in a nearby lane. This observation holds for nearly any scenario in complex urban driving settings, as shown in Fig. 1. Additional open-loop metrics have been devised, such as those based on future collision estimates [5], [7], [9], [10], [21]. However, these metrics induce similar evaluation issues. Their reliance on various naive velocity models prevents them from capturing realistic interactive navigation and negotiation in intricate and dense settings. Moreover, collision metrics lack scalability due to the requirement of perception annotations [29]. Third, online model testing is

[1]Authors are affiliated with the College of Engineering, Boston University, Boston, MA 02215, USA {animikh, adk1361, eohnbar}@bu.edu.
*These authors contributed equally to this work.

done in a *closed-loop manner*, where minor decision errors can propagate and result in unpredictable model behavior over time in testing [30]–[33].

Despite known limitations in offline open-loop evaluation [19], [21], prior work has rarely studied extrapolation from offline to online settings beyond simplistic settings. For instance, Codevilla et al. [28] demonstrated the limited correlation between the two settings but only in the simplest towns and maneuvers in the CARLA simulation [34], i.e., Town01 and Town02, with one online metric (success rate), and partial action values (only steer, excluding throttle and brake ground truth). Therefore, generalization to intricate maneuvers and complex urban scenarios, diverse online metrics, or real-world settings remains understudied despite its broad implications. In this work, we revisit the question of the correlation between offline and online evaluations.

Contributions: We aim to advance safe and reliable evaluation procedures for autonomous driving policies. Our key insights are threefold. *First*, we leverage a more expansive evaluation in simulation, covering multiple towns and closed-loops metrics. Our results reveal even worse correlation patterns than have been previously shown. *Subsequently*, we hypothesize that the poor correlation is due to degraded model performance during instances of uncertainty, i.e., challenging or rare events in the data. We thus analyze an effective metric based on an *uncertainty-weighted error*, and demonstrate its broad effectiveness across settings, models, and error types. While uncertainty-based techniques are standard in robotics and machine learning [35]–[37], to the best of our knowledge, we are the first to quantify their role in estimating online evaluation of driving policies. We observe a significant improvement, e.g., 13% higher correlation in CARLA. *Third*, given challenges in realistic simulation of real-world physics and appearance, we go beyond simplistic simulations to characterize correlation trends in the real world using a small-scale platform, further validating our proposed metric.

II. Related Work

End-to-End Autonomous Driving: With the emergence of deep learning models, autonomous driving research has undergone a significant transformation, shifting from rule-based systems to learning-based approaches. This shift is notably reflected in the proliferation of end-to-end autonomous driving policies utilizing deep learning techniques [6], [8], [14], [17], [38]–[44]. This trend is further exemplified by the CARLA Leaderboard [45], where deep learning-based methods have gained prominence, surpassing modular approaches in autonomous driving performance [46]–[51]. In addition to camera-based RGB images, additional sensors, such as LiDAR, have also been shown to play a key role in improving the safety of driving policies through multi-modal sensor fusion [41], [52], [53].

Issues with Offline Evaluation: The progress of learning-based models is primarily driven by offline evaluation methods. However, the question of their accuracy in assessing driving performance within online settings still remains. Codevilla et al. [28] is one of the first to attempt to study this problem and proposes several novel offline evaluation metrics, including speed-weighted and quantized versions of standard loss metrics. Among the metrics, the NuScenes Detection Score (NDS) [54] is a widely accepted offline metric that has been shown in a recent work by Schreier et al. [29] to provide improved correlation with online metrics like CARLA Driving Score [34]. As shown by Li et al. in [21], a simple baseline model without vision can achieve competitive status on a leaderboard evaluated by offline metrics like NDS, thus proving the limitations of existing offline metrics in predicting driving performance. These recent discussions demand a revisit on whether offline metrics are sufficient in predicting driving performance.

Safety-Critical Scenario Analysis: Safety critical scenarios refer to specific scenarios, such as collisions, where safety is compromised, and the ego vehicle must take preventative or evasive actions to avoid catastrophic outcomes [4], [25], [55], [56]. Current methodologies address this challenge by curating datasets with generated safety-critical scenarios [22], [23], [26], [57]. This approach allows for testing on various generated scenarios to validate if the policy is safe. While we share a similar goal, this is an orthogonal direction to our approach as we focus on designing a useful metric across datasets. However, while generated scenario realism and coverage remain an open question, our proposed metric can be automatically and easily applied at scale to any policy and dataset (i.e., only based on the model and without requiring privileged annotated information).

Off-Policy Evaluation: Offline Policy Evaluation (OPE) in RL [58], as well as offline RL [59], is a related line of research, aiming to minimize the need for direct interaction with the environment to train or evaluate an RL agent. We use an instantiation of OPE and interpret it as a sampling mechanism [60] in our proposed methods. OPE approaches are also known to suffer from high variance in high-dimensional and complex settings. In general, RL and offline RL-based methods have yet to be adapted to complex environments such as CARLA, i.e., moving beyond tabular or simple settings often studied in OPE.

Non-Reactive Simulation Benchmarks: Recent work by Dauner et al. [61] introduces NAVSIM, a data-driven non-reactive simulation framework that bridges offline and online evaluation by unrolling short-horizon simulations on real-world sensor data. It computes safety-critical metrics such as collisions and progress, using bird's-eye-view abstractions. This approach has been shown to outperform traditional displacement errors (e.g., ADE) in closed-loop alignments. While NAVSIM enables scalable benchmarking (e.g., revealing lightweight models like TransFuser [41] can match complex architectures), its non-reactive assumption limits modeling of compounding errors or interactive scenarios. Our work complements NAVSIM by proposing uncertainty-weighted metrics that address rare or uncertain failures with-

out requiring perception annotations, improving scalability in reactive or long-horizon settings.

III. Method

Our goal is to estimate online performance through offline evaluation, i.e., errors computed over a dataset of safely demonstrated driving trajectories. We first formulate our task in Sec. III-A. Then, we define a comprehensive set of offline and online metrics that measure different aspects of model predictions in Sec. III-B.

A. Settings and Data

To validate model performance, we follow standard protocols in vision-based, end-to-end motion planners and open-source simulations. In our settings, we assume access to a collected dataset $\mathscr{D} = \{(\mathbf{o}_i, \mathbf{a}_i)\}_{i=1}^{N}$. Each observation $\mathbf{o}_i = (\mathbf{I}_i, \mathbf{L}_i, c_i, v_i) \in \mathscr{O}$, comprises an image $\mathbf{I}_i \in \mathbb{R}^{N_W \times N_H \times 3}$, a LiDAR point cloud with alpha channel $\mathbf{L}_i \in \mathbb{R}^{N_X \times N_Y \times N_Z \times 1}$, conditional command $c_i \in \mathbb{N}$, and speed measurements $v_i \in \mathbb{R}$, and ground-truth action vector $\mathbf{a}_i \in \mathscr{A}$ executed by the driver. We denote $\pi_\theta : \mathscr{O} \rightarrow \mathscr{A}$ as the policy, a mapping function, parameterized using weights $\theta \in \mathbb{R}^d$, that predicts actions $\hat{\mathbf{a}}_i$ based on input observations $\mathbf{o}_i$. We note that the action may either be a low-level control action, e.g., a 3D vector of steering, throttle, and brake amounts, or a set of future 2D waypoints in the map's view representing an intended trajectory to follow by the vehicle [28], [46]. Vision-based planners today generally employ waypoints as the model output [7], [10], [41], where a low-level controller (e.g., PID) can be used to produce the final control. In contrast, prior work emphasizes direct steer value prediction [28]. Given the widespread use of waypoints-based metrics (defined below), our analysis explores the role of both action definitions and their correlation with online, closed-loop performance.

B. Metrics

Given a loss $\mathscr{L}$ and a per-sample scalar weight $\alpha_i \in \mathbb{R}$, an offline metric can be defined as an expectation wrt. a dataset:

$$\mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathscr{D}}[\alpha_i \mathscr{L}(\mathbf{a}_i, \pi_\theta(\mathbf{o}_i))] \tag{1}$$

A choice of setting weight scalars to $\alpha_i = 1$ and setting an L1 loss, i.e., $\mathscr{L} = \|\mathbf{a}_i - \hat{\mathbf{a}}_i\|_1$, gives the mean absolute error (MAE) metric. Similarly, mean squared error (MSE) is given by using the L2 loss [7], [27]. In principle, a large number of different offline metrics may be derived for a given dataset, yet only a few have been shown to produce a correlation with online performance, particularly for policy's overall success rate along a route known as the *QCE* and *TRE* (based on Codevilla et al. [28]), defined next for clarity. For clarity, we summarize all studied metrics in Table I.

Quantized Classification Error (QCE): QCE quantizes the predictions such that only larger errors from the ground truth are considered, defined via a threshold. QCE is given by:

$$\mathscr{L}_{QCE} = (1 - \delta(Q(\mathbf{a}_i, \sigma), Q(\hat{\mathbf{a}}_i, \sigma))) \tag{2}$$

Where δ is the Kronecker delta function and $Q(x)$ is given by:

$$Q(x) = \begin{cases} -1 & \text{if } x < -\sigma \\ 0 & \text{if } -\sigma \leq x < \sigma \\ 1 & \text{if } x > \sigma \end{cases} \tag{3}$$

To ensure meaningful comparison, we employ the same threshold settings provided in Codevilla et al. [28].

Thresholded Relative Error (TRE): TRE builds on QCE by making use of an adaptive threshold proportional to the ground truth steering angle. It penalizes errors during small ground truth action values, i.e., when the steering angle is low. TRE is given by:

$$\mathscr{L}_{TRE} = H(\|\hat{\mathbf{a}}_i - \mathbf{a}_i\| - \lambda \|\mathbf{a}_i\|) \tag{4}$$

where H is the Heaviside step function. We set $\lambda = 0.1$ consistently with [28] and $\sigma = 0.5$ as the midpoint of our steering range. We note that we set $\alpha_i = 1$ in general for computing errors such as MSE, QCE, and TRE. However, we hypothesize that online errors occur in scenarios where the model may be uncertain about its outputs. We thus propose to leverage the weights α_i to proportionally weigh such instances, as defined next.

Uncertainty-Weighted Error (UWE): Codevilla et al. [28] were motivated by the intuition that errors at higher ego speeds have a greater impact on driving performance. They studied a speed-weighted metric where the weights in Eq. 1 are set to the speed, i.e., $\alpha_i = v_i$. However, this did not result in a better correlation with the route's success rate. Instead, we propose to leverage a model-dependent, *uncertainty-based weight*. We employ monte carlo (MC) dropout [62] to efficiently estimate epistemic uncertainty. This involves enabling dropout during the prediction phase and performing K forward passes through the network for each input sample i. We then calculate the variance u_i of these predictions to estimate the model's uncertainty [63]. Then, u_i is raised to γ to scale the variance's importance, and a weighted sum is applied to various offline metrics, each with optimized weights β_t to compute the UWE:

$$\text{UWE} = \sum_{t=1}^{n} \beta_t \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathscr{D}}[u_i^\gamma \mathscr{L}_t(\mathbf{a}_i, \pi_\theta, (\mathbf{o}_i))], s.t. \\ u_i = \text{var}\left(\{\pi_{\theta_k}(\mathbf{o}_i)\}_{k=1}^{K}\right) \tag{5}$$

where n is the number of offline metrics and $\mathscr{L}_t$ is different offline metric. A weighted sum of the different uncertainty-based weighted offline metrics is able to capture various edge cases from each offline metric. While uncertainty has been applied to improve and regularize a range of robot perception and learning tasks [37], [64]–[66], we are the first to measure its utility in the context of offline and online policy evaluation. The MC Dropout-based uncertainty provides a simple and efficient weight that can be applied to any prior offline metric, yet we find its combinations with the basic errors, i.e., MAE and MSE, sufficient.

TABLE I

SUMMARY OF OFFLINE AND ONLINE METRICS. AN OVERVIEW OF METRIC TERMS. THE LIST IS NOT EXHAUSTIVE BUT COMPARES THE MOST COMMONLY EMPLOYED ONLINE METRICS AND THE BASELINE OFFLINE METRICS.

Offline	
Metric	**Definition**
Steer MSE	MSE of predicted steering and expert steering.
SW Steer MSE	Speed weighted MSE of predicted steering angle and expert steering angle.
UW Steer MSE	Uncertainty weighted MSE of predicted steering angle and expert steering angle.
Steer MAE	MAE of predicted steering angle and expert steering angle.
SW Steer MAE	Speed weighted MAE of predicted steering angle and expert steering angle.
UW Steer MAE	Uncertainty weighted MAE of predicted steering angle and expert steering angle.
Throttle MAE	MAE of predicted throttle value and expert throttle value.
Waypoint MAE	MAE of predicted waypoints and expert waypoints.
Waypoint FDE	Final displacement error of final predicted waypoints and expert final waypoints.
TRE	Thresholded relative error.
QCE	Quantized classification error.
FDE	Error between final predicted displacement and expert final displacement.
Action MSE	MSE of combined throttle and steering.
Action MAE	MAE of combined throttle and steering.
UW Action MSE	Uncertainty weighted MSE of combined throttle and steering.
UW Action MAE	Uncertainty weighted MAE of combined throttle and steering.
PDM Score	Composite metric evaluating collisions, progress, and comfort in non-reactive simulations.
Online	
Metric	**Definition**
Outside Route Lane	The number of times the wheels of the car cross the lane of the assigned route.
Route Deviation	More than 30-meter deviation from route assigned.
Route Timeout	Takes too long to complete route assigned.
Vehicle Blocked	Vehicle does not take an action for more than 180 seconds.
Driving Score	Product of route completion and the infractions penalty.
Success	Vehicle completes the assigned route, denoted by 0 (incomplete) or 1 (complete).
Route Completion	Percentage of the route distance completed by vehicle.
Infractions	Total number of infractions occurred in one route.
Collisions (All)	Number of times vehicle did not stop when it was supposed to.
Collisions (Vehicle)	Number of times the vehicle collided with another vehicle.
Collisions (Environment)	Number of times the vehicle collided with its surroundings.
Red Light Violation	Number of times the vehicle did not stop in the presence of a red light.
Stop Infraction	Number of times the vehicle did not stop in the presence of a stop sign.

Online Metrics: Online metrics convey information about the driving performance in on-policy test time settings. We follow metrics defined in CARLA Leaderboard [41], [45], including the rate of various traffic incidences, such as Collisions and Stop Infractions. The main metric in CARLA is the Driving Score (DS). DS is the average of the product of the Route Completion (% from the planned route) score R_i and an Infraction penalty P_i across N routes:

$$\text{DS} = \frac{1}{N}\sum_{i=1}^{N} R_i P_i \tag{6}$$

We emphasize that prior work has only looked into Success Rate (SR) as the main online metric [28]. However, SR does not consider route completion (e.g., 95% completion score still means zero success) nor infraction types along the route. Instead, DS provides a holistic evaluation of driving behavior, promoting safer and more efficient driving practices. In this paper, we will study the correlation of offline metrics with various online metrics, with a particular focus on the DS.

IV. EXPERIMENTS

We first describe our experimental setup and model training and testing procedure. We then leverage the setup to perform comprehensive experiments in simulation over various towns, ambient settings, and traffic infractions using the CARLA simulation [34]. However, while most online performance studies for autonomous driving are performed in simulation, this may bias results, i.e., due to appearance and physics artifacts. To ensure our findings generalize to real-world models, we thus also incorporate analysis using a small-scale platform in the real-world.

A. Experimental Setup

Simulation Benchmark: We follow the standard Longest6 testing setup [9], [41]. Longest6 benchmark comprises 36 routes with an average route length of 1.5km, which goes beyond the prior study on CARLA's Town02 by Codevilla [28]. We leverage a privileged rule-based expert to generate a large training and validation dataset, where trained policies are evaluated over unseen weather and routes in training.

Real-World Benchmark: Online evaluation of driving policies is generally done in simulation [10], [41], [67]. To

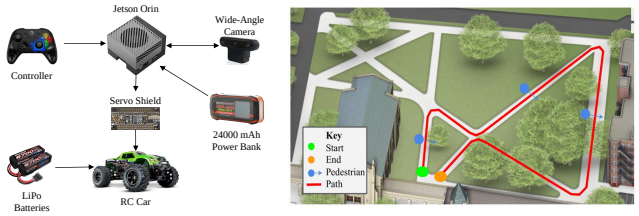

Fig. 2. **Real-World Vehicle Platform (Left) and Bird's Eye View of an Example Evaluation Route (Right).** The vehicle is a hacked off-the-shelf RC car (Traxxas XMaxx). It is equipped with a Jetson Orin with camera and is controlled by a joystick controller for data collection and evaluation.

ensure our findings generalize to the real-world, we utilize a safe, small-scale platform. This platform is based on a commercial RC car (Traxxas X-Maxx [68]). We set up a Jetson Orin and a wide-angle camera to capture observations and control the steering and speed of the car using custom scripts and a joystick controller. We note that we do not leverage LiDAR in our real-world platform. We collect a training set of 32,356 frames on various sidewalks, times of day, and weather. We evaluate on an uneven road that has three right turns, two left turns, and three pedestrians crossing (example routes are shown in Fig. 2). The platform can travel at speeds of up to 50MPH. Online evaluations are safely done with an emergency software stop. In addition to the long route evaluation, we also perform targeted evaluations with clearly defined scenarios over traffic lights, vehicles, and object crossings.

Model Training Strategies: In the simulation, our baseline model is TransFuser [9], which provides a state-of-the-art conditional imitation learning model. The model is a multi-modal policy with two inputs: RGB image and LiDAR. The features of these modalities are fused together by a transformer-based architecture with a cross-modal attention mechanism to predict output actions. To ensure our findings are applicable to a broad range of models, we train a large set of model variations by modifying the model input (e.g., RGB only vs. RGB and LiDAR) and backbone (two ConvNext [69], small and tiny, and nine RegNet [70], from RegNetY-200MF to RegNetY-8.0GF). We train each model (input and backbone combination) for 40 epochs while also saving the checkpoints every 10 epochs to sample additional model variations. Applying test-time dropouts provides another means of sampling models and covering the space of all possible policies. Moreover, we sample models using test-time dropout for values between 10% to 30% with increments of 10%. The Pearson correlation coefficient is then computed over all sampled models. Results are presented across 82 models, each varying in backbones, epochs, inputs, and test-time dropout rates. For real-world models, in the interest of optimizing for real-time performance, we target smaller and simpler architectures motivated by prior works from Codevilla et al. [12], Hawke et al. [71] and Bojarski et al. [38] with a baseline of a five-layer CNN as well as variants of RegNetY as feature extractor backbones. In the real-world experiment, we train an end-to-end policy model solely using RGB input, as it provides an easy and scalable implementation for researchers to reproduce and build upon in the future (i.e., without LiDAR).

B. Results in Simulation

Given the more extensive benchmark and model sampling strategies, i.e., with dropout over model weights for broader coverage of the performance spectrum, we revisit the analysis of Codevilla et al. [28] (only analyzed steering-based metrics, while *fixing other action values based on the ground-truth*). We also leverage the Driving Score, as it provides a holistic measure for driving quality, i.e., beyond Success Rate, as well as finer-grained online metrics.

Revisiting Correlation in Simulation: As shown in Fig. 3, we show an overall poor correlation (both for Driving Score and Success Rate, isolated in Fig. 4) for standard offline metrics. Specifically, QCE and TRE, which are designed for higher correlation with the online Success Rate metric, show poor correlation (0.22 and 0.56, respectively) to more holistic online performance metrics such as the Driving Score. As shown in Fig. 3, QCE and TRE even perform worse than FDE (waypoint-based final displacement error), which is a standard offline metric for high-performing systems today. We hypothesize that the reason these were successful is due to the highly simplistic test settings in Codevilla et al. [28] (Town02, not considering collisions or infractions). However, we observe that our proposed UWE achieves the highest correlation within this challenging benchmark, 0.69, 0.08 higher than prior offline metrics. We only show selected metrics, as the remaining analyzed offline metrics performed with worse correlation. We study nine online metrics to get a complete understanding of the strengths and weaknesses of these metrics and identify key aspects that help choose the most promising metric. Our UWE demonstrates a stronger correlation with all online metrics and significantly outperforms other common offline metric baselines, as shown in Fig. 4. We also compare our UWE metric with the PDM (Predictive Driver Model) score from NAVSIM [61], as shown in Fig. 3, using the publicly available code. As shown in Fig. 3, the PDM score is lower in our more realistic CARLA simulation compared to its performance on nuPlan. Specifically, in settings with more complex physics and reactivity, UWE's correlation exhibits significantly higher fidelity than the PDM score. We note that this is likely because the PDM score makes assumptions about agent reactivity and environmental information (i.e., *being overly accommodating*). In contrast, UWE does not require such assumptions. As such, our metric provides a simple yet effective and scalable alternative potentially applicable to a broader range of situations with various agent dynamics.

C. Generalization to Real-World Settings

We use two distinct real-world settings for evaluation: targeted and naturalistic driving settings. Physics in the real-world such as momentum, friction, and other dynamic-affecting constructs are simplified in simulation. Therefore, evaluations in simulation are not always indicative of the expectations of real-world evaluations.

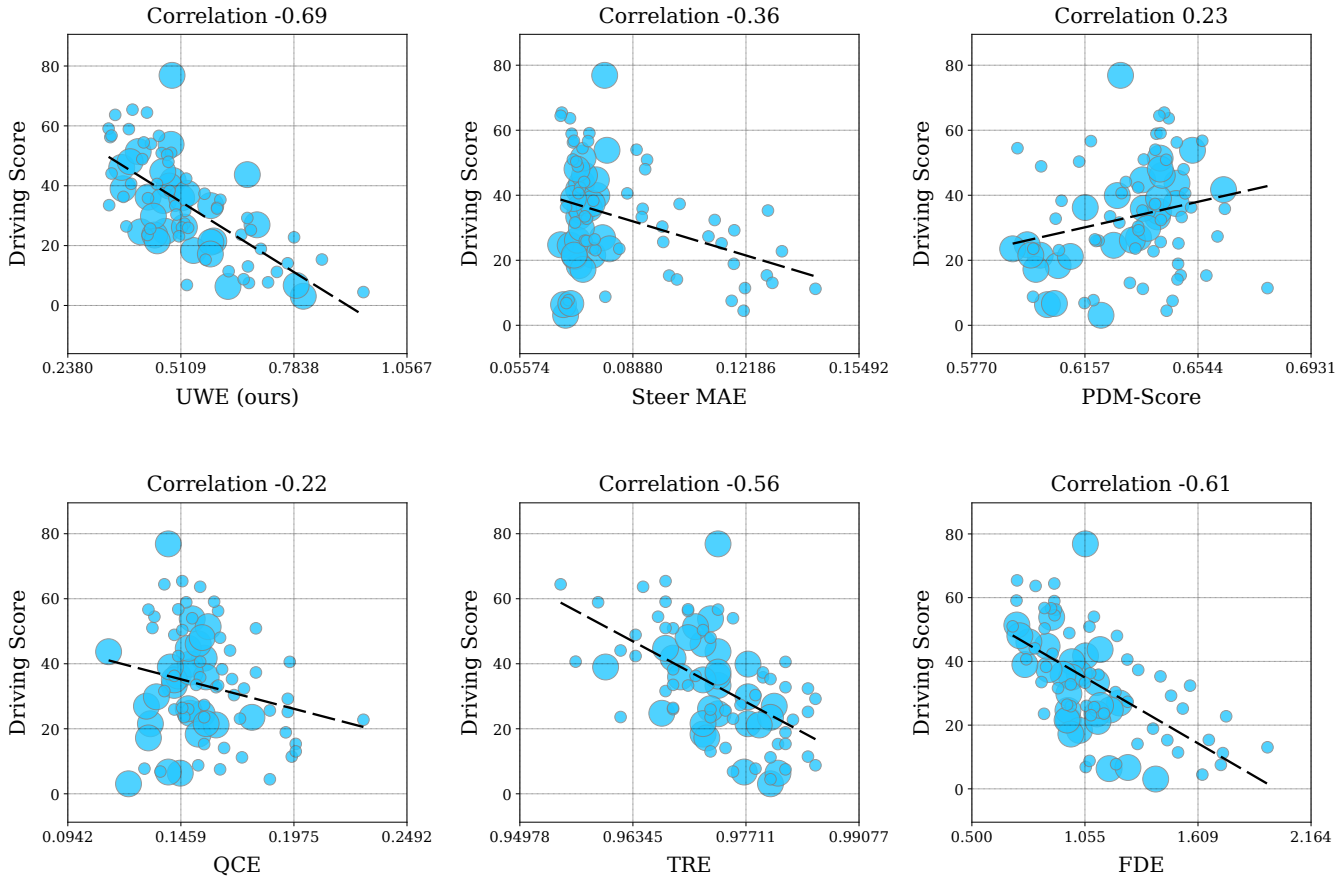

Fig. 3. **Driving Score Correlation Analysis in Simulation.** The plot shows updated correlation for offline metrics, including TRE, QCE and PDM, the most successful reported metrics by prior research [28], [61], as well as widely employed metrics such as waypoint FDE (Final waypoint Displacement Error). Given our updated evaluation setup with complex traffic scenarios and diverse models, correlations are low overall, besides the introduced uncertainty-weighted (UW) version. Each disc shows one model sampled from a certain epoch, backbone, input (with or without LiDAR), and a test-time dropout rate used to obtain coverage of models with varying performances (see further explanation in Sec. IV-A). The radius of each marker is proportional to the percent of test-time dropout used (this sampling is independent of our proposed metric).

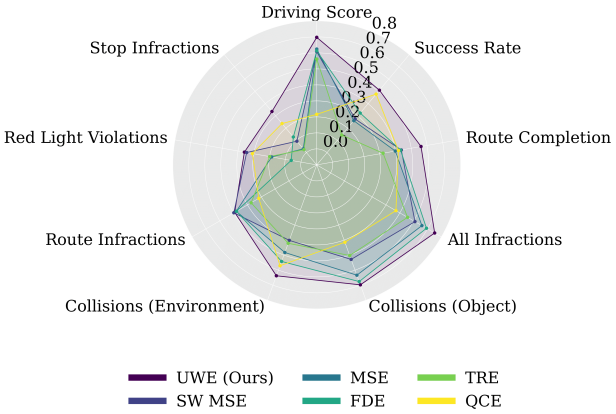

Fig. 4. **Correlation Analysis in Simulation.** Best overall performance is shown by UWE (the proposed uncertainty-weighted error).

Uncertainty Metrics in the Real-World: Fig. 5 shows the effect of the UW action metric on the real-world evaluations. We compare our UW metric for both the full action or only steering against the baseline, as well as the standard TRE and QCE metrics. Due to complexity of computing some metrics in real-world conditions, we focus on the primary five online metrics. Consistent with our findings in the simulation, Fig. 5 shows that in our targeted scenario setting, the UW action metric has the highest correlation across all metrics in comparison to the baseline. This is further validated by the naturalistic driving setting where the UW action MAE shows a competitive correlation when compared to baseline steer-only or full action MAE for most of the online metrics.

Targeted Setting: We conduct experiments in a controlled targeted setting where the brightness and hue of the lighting are static and the floor is leveled. We deploy an Agile-X LIMO for data capturing and inference. Our targeted setting experiments include dense events such as stoplights and crossing automated mechatronic toys (robots, dinosaurs, etc). In Fig. 5, our UW action MAE outperforms other metrics in all correlations with the online metrics.

Naturalistic Driving Setting: Uneven surfaces such as brick roads or potholes are common in naturalistic driving settings but are not yet simulated by CARLA. However, these have major impacts on driving performance in the real-world. Challenges of naturalistic driving settings include varying weather conditions and unexpected obstacles such as road debris. Hence, we mitigate a few of these challenges

Fig. 5. **Correlation Analysis in the Real-World.** Results depict targeted scenarios evaluation (left) in the real-world (short routes over a traffic light or pedestrian scenario) and naturalistic longer routes (right). We find consistent trends with simulation-based results, i.e., UW action correlates well all online metrics when compared to the TRE, QCE, and baseline steer MAE and action MAE in targeted scenarios. Similarly, in longer routes, we show competitive correlation, i.e., compared to the baseline steer and action MAE, while outperforming TRE and QCE.

using hardware such as deploying a hacked suspension-based RC car and utilizing an automatic white-balancing Android phone camera. Additionally, we chose a pedestrian walkway plaza to run naturalistic driving evaluations as it is rich with obstacles, uneven surfaces, and diverse weather conditions. The complete setup is seen in Fig. 2. In Fig. 5, we see that UW action MAE once again performs exceptionally well at indicating good online-driving performance, outperforming the other metrics when correlated with Success Rate and coming in second in correlation with Driving Score. We note that action MAE has a higher correlation to Driving Score than our UW action MAE, and that can be attributed to the greater variation between online and offline evaluation in naturalistic driving testing environments.

D. UWE Ablation

We further investigate the reliability of the uncertainty estimates obtained via Monte Carlo Dropout. When we replace the dropout mechanism with a model ensemble and observe an increase in correlation from 0.69 to 0.80 in the Pearson coefficient. While this aligns with prior literature [37], dropout-based inference remains more efficient for larger models. To assess metric sensitivity, we recompute the correlation shown in Fig. 3. For this test, we use uncertainty weights derived from a single model, keeping weights fixed rather than adapting them per model. Specifically, we compute the overall correlation with the Driving Score using weights from a model trained for 10 epochs (achieving a 0.67 Pearson coefficient) and 40 epochs (Pearson coefficient of 0.73). These results indicate that UWE is not highly sensitive to model selection. While we evaluate UWE across models and deployment settings, including real-world applications on a physical platform, each metric introduces trade-offs. For instance, UWE does not require HD Maps or the meticulous offline simulation design used in NAVSIM [61].

V. Conclusion

Offline evaluation offers several benefits for researchers and engineers, such as safety, scalability, and reduced cost. In this work, we seek to better characterize the correlation between offline and online evaluation. We do so through a series of experiments in both simulation and real-world. To ensure the broad impact of our findings, i.e., across diverse model types and conditions, we further incorporate diverse model sampling strategies in order. An uncertainty-weighted metric is shown to improve correlation with various online aspects of autonomous driving policies, e.g., over standard offline metrics used today. However, our study is a preliminary step toward efficient and scalable evaluation procedures. While our real-world platform (RC car) validates UWE's feasibility, scaling to full-sized vehicles in complex urban environments remains future work. We pursue a finer-grained evaluation of various online metrics suitable for complex urban driving, predicting subtle errors can be difficult. There is a need to devise scalable and efficient metrics that can account for more subtle violations. We plan to study this in the future using our physical platform. Although we evaluate using a field study in the real-world, additional larger-scale evaluations, e.g., with full-scale vehicles and diverse controlled test routes, should be pursued to validate our framework and metrics further in the future.

Acknowledgments: We thank the Red Hat Collaboratory (award #2024-01-RH07, #2025-01-RH04) and the National Science Foundation (IIS-2152077) for supporting this research.

References

[1] "The New York Times. Self-driving uber car kills pedestrian in arizona, where robots roam," 2018. 1
[2] Government-Fleet, "California DMV removes cruise's driverless vehicle testing permits," 2023. 1
[3] The Last Driver License Holder, "Waymo Confused Behind a Trailer with a Tree," 2024. 1
[4] P. Ghorai *et al.*, "A causation analysis of autonomous vehicle crashes," *ITSM*, 2024. 1, 2
[5] P. Hu, A. Huang, J. Dolan, D. Held, and D. Ramanan, "Safe local motion planning with self-supervised freespace forecasting," in *CVPR*, 2021. 1
[6] A. Behl, K. Chitta, A. Prakash, E. Ohn-Bar, and A. Geiger, "Label efficient visual abstractions for autonomous driving," in *IROS*, 2020. 1, 2
[7] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *CVPR*, 2023. 1, 3
[8] J. Zhang, Z. Huang, and E. Ohn-Bar, "Coaching a teachable student," in *CVPR*, 2023. 1, 2
[9] B. Jaeger, K. Chitta, and A. Geiger, "Hidden biases of end-to-end driving models," *ICCV*, 2023. 1, 4, 5
[10] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "VAD: Vectorized scene representation for efficient autonomous driving," *ICCV*, 2023. 1, 3, 4
[11] D. Wang, C. Devin, Q.-Z. Cai, P. Krähenbühl, and T. Darrell, "Monocular plan view networks for autonomous driving," in *IROS*, 2019. 1
[12] F. Codevilla, M. Miiller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *ICRA*, 2018. 1, 5
[13] R. Zhu, P. Huang, E. Ohn-Bar, and V. Saligrama, "Learning to drive anywhere," *CoRL*, 2023. 1
[14] M. Bansal, A. Krizhevsky, and A. Ogale, "ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst," in *RSS*, 2019. 1, 2
[15] P. Wu, L. Chen, H. Li, X. Jia, J. Yan, and Y. Qiao, "Policy pre-training for end-to-end autonomous driving via self-supervised geometric modeling," *ICLR*, 2023. 1
[16] D. Dauner, M. Hallgarten, A. Geiger, and K. Chitta, "Parting with misconceptions about learning-based vehicle motion planning," *arXiv preprint arXiv:2306.07962*, 2023. 1

[17] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *ECCV*, 2022. 1, 2

[18] J. Zhang, Z. Huang, A. Ray, and E. Ohn-Bar, "Feedback-guided autonomous driving," in *CVPR*, 2024. 1

[19] C. Zhang, R. Guo, W. Zeng, Y. Xiong, B. Dai, R. Hu, M. Ren, and R. Urtasun, "Rethinking closed-loop training for autonomous driving," in *ECCV*, 2022. 1, 2

[20] J. Tu, S. Suo, C. Zhang, K. Wong, and R. Urtasun, "Towards scalable coverage-based testing of autonomous vehicles," in *CoRL*, 2023. 1

[21] Z. Li, Z. Yu, S. Lan, J. Li, J. Kautz, T. Lu, and J. M. Alvarez, "Is ego status all you need for open-loop end-to-end autonomous driving?" *arXiv preprint arXiv:2312.03031*, 2023. 1, 2

[22] E. Thorn, S. C. Kimmel, M. Chaka, B. A. Hamilton *et al.*, "A framework for automated driving system testable cases and scenarios," Department of Transportation, Tech. Rep., 2018. 1, 2

[23] T. Menzel, G. Bagschik, and M. Maurer, "Scenarios for development, test and validation of automated vehicles," in *IV*, 2018. 1, 2

[24] M. O'Kelly, A. Sinha, H. Namkoong, R. Tedrake, and J. C. Duchi, "Scalable end-to-end autonomous vehicle testing via rare-event simulation," *NeurIPS*, 2018. 1

[25] H. Weber, J. Bock, J. Klimke, C. Roesener, J. Hiller, R. Krajewski, A. Zlocki, and L. Eckstein, "A framework for definition of logical scenarios for safety assurance of automated driving," *Traffic Injury Prevention*, 2019. 1, 2

[26] R. Lee, O. J. Mengshoel, A. Saksena, R. W. Gardner, D. Genin, J. Silbermann, M. Owen, and M. J. Kochenderfer, "Adaptive stress testing: Finding likely failure events with reinforcement learning," *Journal of Artificial Intelligence Research*, 2020. 1, 2

[27] L. L. Li, B. Yang, M. Liang, W. Zeng, M. Ren, S. Segal, and R. Urtasun, "End-to-end contextual perception and prediction with interaction transformer," in *IROS*, 2020. 1, 3

[28] F. Codevilla, A. M. Lopez, V. Koltun, and A. Dosovitskiy, "On offline evaluation of vision-based driving models," in *ECCV*, 2018. 1, 2, 3, 4, 5, 6

[29] T. Schreier, K. Renz, A. Geiger, and K. Chitta, "On offline evaluation of 3d object detection for autonomous driving," in *ICCV*, 2023. 1, 2

[30] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *AISTATS*, 2011. 2

[31] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, 2018. 2

[32] A. Prakash, A. Behl, E. Ohn-Bar, K. Chitta, and A. Geiger, "Exploring data aggregation in policy learning for vision-based urban autonomous driving," in *CVPR*, 2020. 2

[33] M. Laskey, J. Lee, R. Fox, A. Dragan, and K. Goldberg, "Dart: Noise injection for robust imitation learning," in *CoRL*, 2017. 2

[34] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *CoRL*, 2017. 2, 4

[35] A. Filos, P. Tigas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" in *ICML*, 2020. 2

[36] K. Stachowicz and S. Levine, "RACER: Epistemic risk-sensitive rl enables fast driving with fewer crashes," *arXiv preprint arXiv:2405.04714*, 2024. 2

[37] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *NeurIPS*, 2019. 2, 3, 7

[38] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016. 2, 5

[39] L. Lai, E. Ohn-Bar, S. Arora, and J. S. K. Yi, "Uncertainty-guided never-ending learning to drive," in *CVPR*, 2024. 2

[40] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *CVPR*, 2017. 2

[41] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *PAMI*, 2023. 2, 3, 4

[42] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *NeurIPS*, 1989. 2

[43] A. Cui, S. Casas, A. Sadat, R. Liao, and R. Urtasun, "Lookout: Diverse multi-future prediction and planning for self-driving," in *ICCV*, 2021. 2

[44] E. Ohn-Bar, A. Prakash, A. Behl, K. Chitta, and A. Geiger, "Learning situational driving," in *CVPR*, 2020. 2

[45] "Carla autonomous driving leaderboard," *https://leaderboard.carla.org/*, 2023. 2, 4

[46] M. Müller, A. Dosovitskiy, B. Ghanem, and V. Koltun, "Driving policy transfer via modularity and abstraction," in *CoRL*, 2018. 2, 3

[47] J. Tang, L. Shaoshan, S. Pei, S. Zuckerman, L. Chen, W. Shi, and J.-L. Gaudiot, "Teaching autonomous driving using a modular and integrated approach," in *COMPSAC*, 2018. 2

[48] L. Chen, L. Platinsky, S. Speichert, B. Osiński, O. Scheel, Y. Ye, H. Grimmett, L. Del Pero, and P. Ondruska, "What data do we need for training an av motion planner?" in *ICRA*, 2021. 2

[49] D. González, J. Pérez, V. Milanés, and F. Nashashibi, "A review of motion planning techniques for automated vehicles," *T-ITS*, 2015. 2

[50] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, "Learning to drive in a day," in *ICRA*, 2019. 2

[51] W. Xu, Q. Wang, and J. M. Dolan, "Autonomous vehicle motion planning via recurrent spline optimization," in *ICRA*, 2021. 2

[52] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *CoRL*, 2023. 2

[53] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *CVPR*, 2021. 2

[54] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020. 2

[55] J. Wang, A. Pun, J. Tu, S. Manivasagam, A. Sadat, S. Casas, M. Ren, and R. Urtasun, "Advsim: Generating safety-critical scenarios for self-driving vehicles," in *CVPR*, 2021. 2

[56] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger, "King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients," *ECCV*, 2022. 2

[57] L. Lai, Z. Shangguan, J. Zhang, and E. Ohn-Bar, "XVO: Generalized visual odometry via cross-modal self-training," in *ICCV*, 2023. 2

[58] K. Uehara, Shi, "A review of off-policy evaluation in reinforcement learning," *arXiv preprint arXiv:2212.06355*, 2022. 2

[59] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020. 2

[60] C. Shi, R. Wan, V. Chernozhukov, and R. Song, "Deeply-debiased off-policy interval estimation," in *international conference on machine learning*, 2021. 2

[61] 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 *NeurIPS*, 2024. 2, 5, 6, 7

[62] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016. 3

[63] L. Lai, Z. Yin, and E. Ohn-Bar, "ZeroVO: Visual odometry with minimal assumptions," in *CVPR*, 2025. 3

[64] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *CVPR*, 2018. 3

[65] M. Henaff, A. Canziani, and Y. LeCun, "Model-predictive policy learning with uncertainty regularization for driving in dense traffic," *ICLR*, 2019. 3

[66] Y. Wang, J. Peng, and Z. Zhang, "Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation," in *ICCV*, 2021. 3

[67] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *CoRL*, 2020. 4

[68] "Traxxas X-Maxx RC Vehicle, https://traxxas.com/products/landing/x-maxx/." 5

[69] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *CVPR*, 2022. 5

[70] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *CVPR*, 2020. 5

[71] J. Hawke, R. Shen, C. Gurau, S. Sharma, D. Reda, N. Nikolov, P. Mazur, S. Micklethwaite, N. Griffiths, A. Shah *et al.*, "Urban driving with conditional imitation learning," in *ICRA*, 2020. 5