Masterarbeit

Evaluating Traffic and Scenario Generation using Fidelity and Diversity Metrics

Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik
Lernbasierte Computer Vision
Micha Fauth, micha.fauth@student.uni-tuebingen.de, 2025

Bearbeitungszeitraum: 01. Juni 2025 - 01. Dezember 2025

Betreuer: Kashyap Chitta, Universität Tübingen
Gutachter: Prof. Dr. Andreas Geiger, Universität Tübingen
Zweitgutachter: Prof. Dr. Matthias Hein, Universität Tübingen

Selbstständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Masterarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Micha Fauth (Matrikelnummer 6554279), November 30, 2025

Abstract

Despite recent advances in autonomous driving, ensuring their safety remains challenging and demands extensive testing. Traditional simulations offer a solution but are constrained by human design and heuristics. These limitations have motivated data-driven generative methods to create scalable, realistic, and diverse traffic scenarios that better reflect real-world complexity. However, evaluating the realism of generated scenarios remains a fundamental challenge due to the structured, multimodal, and temporally evolving nature of driving data. Existing evaluation approaches, such as collision metrics, trajectory errors, or distribution-based comparisons, capture only limited aspects of realism and fail to assess multimodality effectively. Benchmark-driven protocols, such as the Waymo Open Sim Agent Challenge, provide partial solutions. To address these limitations, this work introduces an evaluation framework that adapts fidelity and diversity metrics from the image synthesis domain to the driving domain, measuring realism and distributional coverage. We complement this with an interactive visualization tool for qualitative analysis, making the evaluation process more explainable. The framework is validated through comprehensive case studies across multiple rule-based and learning-based generative models for autonomous driving scenes. Our results demonstrate that the adapted metrics provide complementary diagnostic value for agent-level traffic simulation while highlighting limitations in assessing full scene-level scenario generation. We hope our proposed framework provides practical guidance for future research on evaluating traffic and scenario generation.

Acknowledgments

I want to thank Prof. Andreas Geiger for giving me the opportunity to complete my master's thesis within his research group. I am equally grateful to my supervisor, Kashyap Chitta, whose guidance, encouragement, and insights supported me at every stage of this work and shaped both my research skills and personal growth. I also want to thank Maximilian Igl for his valuable input and participation in our weekly project meetings. Furthermore, I would like to thank my mate Long for the successful collaboration on the Waymo Challenges, as well as Daniel and Bernhard for their help and ideas.

On a personal note, I would like to thank my parents for their continuous support throughout my studies. Lastly, I thank Nina for her unconditional love, support, and patience.

Contents

1. Introduction

The ability to generate realistic traffic and driving scenarios is an essential component in the development of autonomous vehicles. Modern autonomous driving systems require extensive testing across a wide range of behaviors, road layouts, and interactive situations to ensure reliability and safety before deployment in the real world. However, collecting large-scale, safety-critical real-world data is expensive and inherently limited in the diversity of rare but important events. Consequently, scalable methods for synthesizing realistic traffic scenarios have become increasingly important. Simulation environments offer a controllable and cost-effective alternative, yet they rely on rule-based or manually crafted heuristics that limit behavioral richness and realism [DRC+17, LBBW+18]. Recent advances in generative modeling, including diffusion models [HJA20] and autoregressive transformers [VSP+17], provide promising solutions by learning from real-world data and producing more diverse and realistic traffic scenes [FLP+23, MML+24, JBC+24, RGG+25]. This progress, however, has amplified the need for rigorous and interpretable evaluation methodologies to assess the quality of generated scenarios.

Unlike images or text, driving scenarios are defined by structured, multimodal, and temporally evolving elements such as agent trajectories, semantic maps, interactions, physical constraints, and traffic rules (see Figure 1.1). Existing evaluation approaches lack standardization and typically rely on heterogeneous metrics such as collision rates [ZLD+23], trajectory reconstruction errors [VLY+23], or feature distribution comparisons [XCIP22, MML+24]. While informative, each such metric captures only a narrow aspect of realism or physical plausibility. Benchmark-oriented protocols such as the Waymo Open Sim Agent Challenge (WOSAC) [MLM+23] measure how likely real-world behavior is by computing the average negative log-likelihood of ground truth observations under the distribution of generated trajectories. However, this approach is also limited in its ability to assess multimodality and exhibits several conceptual constraints.

This thesis addresses these limitations in two complementary ways. First, we adapt fidelity and diversity metrics developed initially for image synthesis [SBL+18, KKL+19] to the domain of traffic and scenario generation. By comparing real and generated samples within an embedding space, these metrics provide a unified and interpretable framework for assessing both realism (fidelity) and distributional coverage (diversity).

Second, we present an interactive analysis tool that complements these metrics with

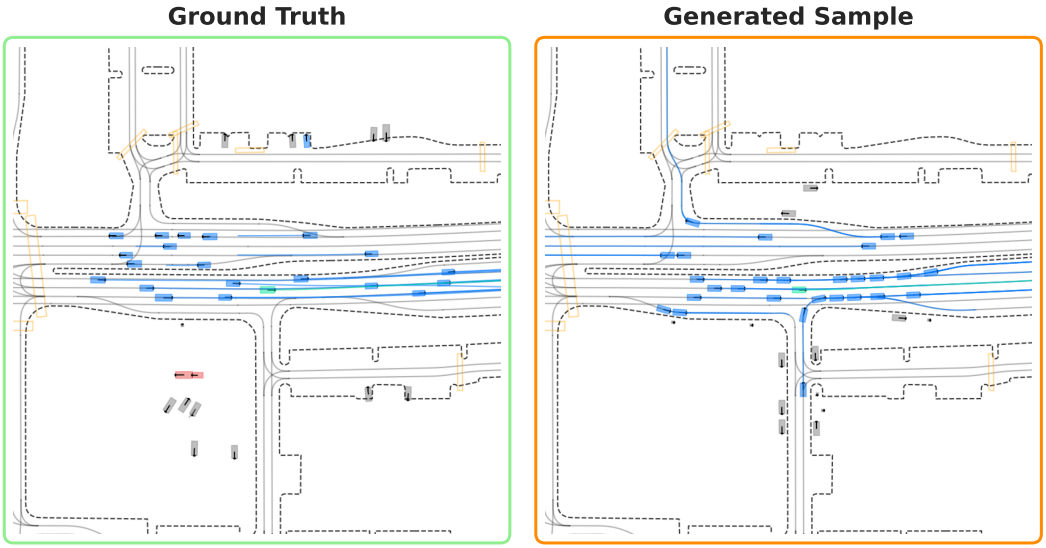


Figure 1.1.: **Example traffic scene in Bird's Eye View (BEV).** Real-world scenario (left) and rule-based generated scenario (right). Although the generated scene appears visually realistic, data-driven evaluation metrics may yield misleading conclusions, as they can be sensitive to artifacts such as annotation noise in the ground truth data.

manifold visualizations, nearest-neighbor inspection, and detailed per-sample comparisons. This tool bridges both quantitative evaluation and qualitative interpretation, making the evaluation process more transparent and explainable.

To demonstrate the utility and sensitivity of our framework, we conduct comprehensive case studies evaluating a range of generative approaches, including rule-based baselines and recent learning-based methods for traffic and scenario generation. Our experiments show that while the metrics provide complementary diagnostic value for traffic simulation, they remain limited in delivering consistent and interpretable evaluations within the more complex scenario generation task.

Thesis Structure: Chapter 2 surveys related research on evaluation approaches for generative models, with a particular emphasis on traffic and scenario generation. Chapter 3 introduces the adapted fidelity and diversity metrics and outlines the main components of our visualization tool for interactive evaluation. Chapter 4 presents a series of experimental case studies that apply and ablate our evaluation framework across multiple traffic and scenario generation settings. Finally, Chapter 5 concludes this thesis by discussing the limitations of current data-driven evaluation approaches and suggesting future research directions.

2. Background and Related Work

This chapter reviews existing literature relevant to this work. Section 2.1 surveys current evaluation methods from other generative domains and provides a detailed discussion about fidelity and diversity metrics. Section 2.2 then shifts to our application domain, providing an overview of current traffic and scenario generation approaches and summarizing commonly used evaluation protocols. Finally, Section 2.3 introduces the mainstream benchmarks for traffic and scenario generation provided by Waymo [MLM+23, Way25].

2.1. Evaluation Methods for Generative Tasks

In general, evaluating generative models remains a fundamentally challenging task and an active area of research [NOU+20, SCH+23, RvBvdS25, RSGA25]. This difficulty is particularly pronounced for implicit models such as generative adversarial networks [GPAM+14] and diffusion models [HJA20], which do not provide explicit likelihood estimates [BPF24]. However, even likelihood-based approaches, such as variational autoencoders [KW14], have been shown to yield unreliable results when comparing log-likelihood values [TvdOB16].

As a result, researchers have increasingly turned to a variety of quantitative proxy metrics, often tailored to the domain and task at hand. The following section briefly surveys the most popular evaluation approaches used in different application areas, with a particular focus on the development of fidelity and diversity metrics.

2.1.1. Other Domains

Different data modalities impose different requirements on the evaluation of generative models. This section summarizes representative evaluation methodologies used in other domains, highlighting how metric design adapts to the properties of text, images, video, audio, and tabular data.

Text Data

In the domain of Natural Language Processing (NLP), traditional reference-based overlap measures such as BLEU [PRWZ02] or ROUGE [Lin04] have long been considered the gold standard for evaluating text translation and summarization.

Even though several studies have shown that these metrics do not correlate well with human judgment, there was no decline in their popularity [SMK20].

With the rapid advancement of Large Language Models (LLMs), language generation tasks have become more sophisticated, and their evaluation has consequently grown more complex. Embedding-based metrics such as BERTScore [ZKW+20] and CLIPScore [HHF+22] leverage pretrained (vision-)language models to compute semantic similarity between generated outputs and reference texts or input images, capturing meaning beyond surface-level word overlap.

Furthermore, evaluation approaches have increasingly shifted from traditional reference-based metrics to reference-free methods [IvDS25]. Even LLMs themselves have recently been employed as evaluators of NLP tasks [ZCS+23], enabling evaluation without human annotations, provided that a pretrained LLM is available.

Image Data

In the image domain, models are judged by how realistic their generated images look. The most popular metrics include the Inception Score (IS) [SGZ+16] and the Fréchet Inception Distance (FID) [HRU+17]. IS is computed based on the inception network [SVI+15] trained on the ImageNet dataset [DDS+09]. Formally, it calculates

$$IS = \exp(\mathbb{E}_{x \sim p_G}[d_{KL}(p_\theta(y \mid x) || p_\theta(y))])$$

where $x \sim p_G$ is a generated sample image, $p_\theta(y|x)$ is the conditional class distribution predicted by the inception network, and $p_\theta(y)$ the marginal distribution over labels for all generated images. Besides the necessity of a labeled data set, IS is weak at guiding model comparison [BS18, LL22, XHY+18].

FID also relies on the inception network, but instead of using its predicted class probabilities, it extracts intermediate feature embeddings for both real and generated images. The metric assumes that the embedded features follow a multivariate Gaussian distribution, which is then fitted to the real and generated samples. It then measures the Fréchet distance's difference between these two distributions, given by

$$FID = ||\mu_r - \mu_g||_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}})$$

where $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_g, \Sigma_g)$ are the fitted Gaussians for the real and generated data, respectively. However, FID is sensitive to both the addition of spurious modes as well as to mode dropping [SBL+18]. Furthermore, several improvements to FID have been proposed. While the Kernel Inception Distance (KID) relaxes the Gaussian assumption [BSAG21], FID_∞ aims to reduce the bias introduced when FID is estimated from a finite number of samples [CF20].

Video and Audio Data

For video generation, evaluation extends beyond visual quality to include temporal coherence across multiple frames. A widely used metric for unconditional video generation is the Fréchet Video Distance (FVD) [UvSK+19] and Kernel Video Distance (KVD) [UvSK+19], which adapts the image-based FID and KID by extracting spatio-temporal features using a 3D convolution network pretrained on a video dataset. Furthermore, CLIP cosine similarities are frequently employed, particularly in text-conditioned video generation tasks [ECA+23].

Similarly, in the audio domain, the Fréchet Audio Distance (FAD) [KZRS19] and the Kernel Audio Distance (KID) [CEL+25] extend FID and KID, respectively. These metrics measure distributional similarity between real and generated audio by comparing deep feature statistics extracted from a pretrained audio embedding network.

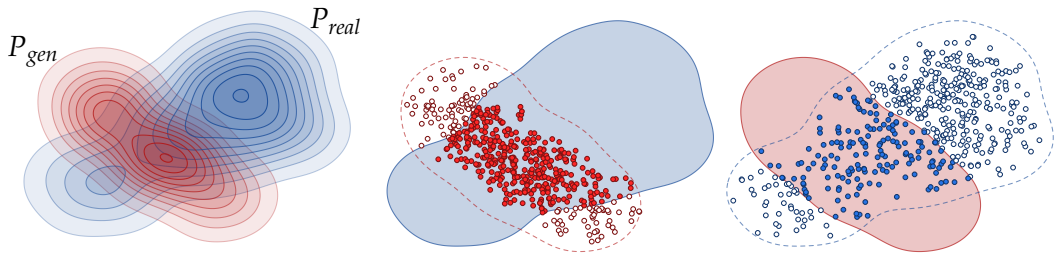
Tabular Data

Generated tabular data is evaluated differently because there is neither visual nor auditory output. Evaluation typically focuses on statistical fidelity and utility [HOMC+25, HEA+23]. Fidelity metrics assess whether the generated data accurately reflect the distributions and relationships present in the real data. For example, [HOMC+25] measures distances between marginal histograms using the Hellinger distance and compares pairwise correlations between variables. Additionally, they treat the problem as a classification task. If a trained classifier can not reliably distinguish between real and generated data, then the generated data has high fidelity.

Utility metrics quantify how useful the generated data is for downstream tasks. A common approach is Train-on-Synthetic Test-on-Real, where a model is trained on the generated data and evaluated on a held-out real test set. The performance drop relative to a model trained and tested on real data indicates the utility of the generated data [RBB+20].

Human Evaluation

Despite the advances in quantitative metrics, human judgment remains crucial for generative models in many domains [TvdOB16, HL08, SMK20]. In video and audio tasks, human participants are often asked to perform pairwise preference tests or quality ranking assessments. Furthermore, new metrics are routinely validated against human judgment [UvSK+19, KZRS19]. Recent studies examine human perception of generated images and find that none of the current quantitative metrics accurately reflect human judgments of realism. This underscores the continued importance of human evaluation as a supplement to quantitative metrics [SCH+23].



(a) Example distributions (b) Precision (c) Recall

Figure 2.1.: **Precision and Recall for distributions** defined by [SBL+18].(a) Given two distributions, one of real data P_{real} (blue) and of generated data with P_{gen} (red), we can define: (b) Precision as the probability that a random data sample from P_{gen} falls within the support of P_{real}. And (c) recall as the probability that a random data sample from P_{real} falls within the support of P_{gen}. The figure is taken from [KKL+19].

2.1.2. Fidelity and Diversity

In the image domain, [SBL+18] point out that one-dimensional metrics such as FID and IS are limited in their diagnostic capability. While these metrics can distinguish between generally good and bad models, they fail to capture specific failure modes of generative models. Moreover, they provide only a dataset-level assessment and do not offer an instance-level metric.

To address these shortcomings, [SBL+18] proposes a precision and recall metric for distribution to assess image generation models. The classic concepts of precision and recall, predominantly known from classification tasks, are illustrated in Figure 2.1. It defines precision as the fraction of generated samples that are realistic, and recall as the extent to which the generated samples cover the real data distribution. Both are computed as expectations over binary set membership, capturing how likely a sample from one distribution lies within the support of the other. The two dimensions of the metric provide an advantage for distinguishing between more failure modes of generative models. Specifically, a model that produces unrealistic samples will exhibit low precision. In contrast, a model that fails to cover the full diversity of the real data distribution will show low recall, even if its generated samples are of high quality (high precision). [SBL+18] formulates its metric by modeling the relative probability densities of the two distributions across a continuum of precision and recall values. This results in a curve-valued metric where the extrema correspond to the classic definitions. However, these extrema are difficult to estimate due to the method's dependence on relative densities and a uniform-density assumption [KKL+19, NOU+20].

Several recent works [KKL+19, NOU+20, AVBSvdS22, CU23, KA23, PK23, FSA25, STT25] have built upon this approach by introducing various extensions of the

precision and recall framework. All metrics transform the original curve-valued formulation into single-valued metrics, facilitating easier interpretation and more practical use [RvBvdS25]. Most of the metrics utilize k-Nearest Neighbor (kNN) methods to estimate the probability density function of the data. [KKL+19], for instance, estimates the underlying support of the real and generated data in feature space by constructing manifolds with the set of hyperspheres around each embedded sample that reaches its kth nearest neighbor. While intuitive and widely adopted, this approach has several limitations, such as it is highly sensitive to outliers and fails to recognize identical distributions, as pointed out by [NOU+20, PK23]. To address these shortcomings, [NOU+20] and [PK23] propose enhanced formulations to be more robust and better capture the relationship between the real and generated data distributions. Other works propose alternative support estimation [AVBSvdS22], and symmetric reformulations [KA23], highlighting the ongoing effort to overcome the limitations of existing metrics.

Some of these metrics have been successfully applied beyond the image domain, including to audio [NOU+20] and tabular data [AVBSvdS22], demonstrating their versatility in evaluating generative models across domains. However, despite these advancements, a recent study by [RvBvdS25] systematically evaluates current fidelity and diversity metrics and finds that none of them pass all basic sanity checks, even in simple distribution settings. The authors further argue that all metrics are bad for absolute model evaluation. Moreover, they highlight that using current fidelity and diversity metrics requires great caution, and practitioners should be aware of their inherent limitations when interpreting results.

Since all these metrics come in pairs, they can be grouped into two categories: fidelity and diversity. While the first category measures the realism of generated samples, the latter measures how well the generated distribution covers the real distribution. This categorization and notation will be used throughout the rest of this work.

2.2. Traffic and Scenario Generation

Shifting focus to our domain, this section reviews the landscape of traffic and scenario generation, summarizing methodological approaches and the corresponding evaluation protocols adopted in recent literature.

2.2.1. Definition and Approaches

Traffic and scenario generation refers to synthesizing realistic driving situations for simulation, which plays a key role in enabling rigorous testing of autonomous vehicle systems under diverse and challenging conditions [DXA+23]. The boundary between traffic generation and scenario generation is often indistinct in the literature, with both terms being used to describe overlapping or closely related tasks [LPF+23, GPZ+25].

In our work, we distinguish traffic simulation from scenario generation based on initial conditions. Thus, traffic simulation assumes predefined initial states, whereas scenario generation involves synthesizing complete scenarios that include initial states. For both tasks, we typically consider the road map and the number of objects given. In this work, we use the terms traffic simulation, sim agent, and traffic generation interchangeably.

Traffic and scenario generation methods can be broadly categorized into two main approaches, namely rule-based and learning-based. In the following, we provide a brief overview of related work for these approaches.

Rule-based

Rule-based methods rely on predefined rules and expert knowledge to generate traffic behaviors through deterministic processes. A widely used rule-based motion model is the Intelligent Driver Model (IDM), which implements vehicle motion using simple car-following rules [THH00]. Simulators such as Carla [DRC+17] and SUMO [LBBW+18] commonly incorporate variants of this model to simulate background traffic.

Similarly, procedural approaches remain popular for scenario generation. Traffic agents are populated in the scene guided by a set of predefined rules and heuristics [DRC+17, LBBW+18]. These methods allow for manual parameter tuning to achieve reasonable agent placements and behaviors. However, purely rule-based methods often struggle to capture the full complexity of human driving, and handcrafted heuristics may miss irregular maneuvers or diverse interactions between vehicles [SRCU21, MML+24].

Learning-based

With recent advances in deep learning, learning-based approaches have become increasingly prominent in traffic and scenario generation. Such methods aim to model the underlying distributions of real-world traffic scenarios and generate novel samples, thereby producing more realistic and diverse driving behaviors. In particular, generative models such as diffusion models [HJA20] and autoregressive transformer architectures [VSP+17] have shown remarkable success in this domain.

For traffic simulation, the current state-of-the-art includes finetuned SMART models, which are autoregressive multi-agent motion forecasting models based on next-token prediction [WFGK24, ZKI+25, ZJC+25]. Other approaches, such as VBD [HZV+24] and SceneDifusser [JBC+24], exemplify diffusion-based motion modeling.

In the context of scenario generation, autoregressive methods typically insert agents into the scene subsequently, conditioning each new agent on the map and initial states of previously placed agents [TWW+21]. Some approaches extend this conditioning

to include also future trajectories [FLP+23, MML+24]. For motion decoding, they commonly use transformer-based trajectory decoders such as MultiPath++ [VHS+21] or Wayformer [NARZ+22].

Additionally, several works adopt hybrid strategies, combining different paradigms for initial state generation and motion modeling. For example, SLEDGE, Scene Diffusion, and Scenario Dreamer all employ diffusion models for generating the initial state scenes, but differ in how they simulate motion, using rule-based [CDG24], reinforcement learning-based [RGG+25], or transformer-based [SGS+24] approaches, respectively.

In this thesis, we primarily adopt rule-based methods to efficiently generate controllable and diverse traffic behaviors with varying performance, which is essential for systematically evaluating metric sensitivity. As a learning-based addition, we incorporate SMART to leverage state-of-the-art motion modeling, thereby complementing rule-based methods with more realistic and data-driven behaviors.

2.2.2. Evaluation Protocols

The evaluation of generative tasks remains an open challenge, as there is yet to be a single metric that can adequately measure all aspects of realism and controllability [TvdOB16]. This also applies to evaluating the quality of generated traffic scenes. To address these challenges, different evaluation protocols have been proposed. The following subsections present representative approaches for assessing traffic and scenario generation.

Traffic Simulation

[MLM+23] point out that evaluating traffic simulations lacks standardization and mainstream benchmarks. We present a subset of their survey in Table 2.1 as a representative summary. Most works measure a form of collision rate for the importance of safety [SRCU21, BYS+21, XCIP22, VLY+23, ZLD+23], but this alone doesn't capture realism, since static agents can simply avoid collisions. To assess realism, researchers additionally compare simulated behavior to logged real-world data using metrics such as distribution matching of vehicle dynamics [XCIP22], off-road rates [XCIP22, SRCU21], spatial coverage [XCIP22, SRCU21], and goal progress [VLY+23].

Furthermore, metrics like Average Displacement Error (ADE) have been used [BYS+21, VLY+23], but they are limited by the inherent multimodality of agent behavior. Synthesized traffic scenes should account for multiple plausible agent behaviors, rather than matching only a single ground-truth observation. Alternative approaches favor point-to-distribution metrics, such as the average negative log likelihood (NLL) of the ground truth trajectories under the distribution of generated trajectories [IP19, ZLD+23]. Building on this approach, Waymo introduced an NLL-based

Evaluation Protocol	ADE or minADE	Offroad Rate	Collision Rate	Instance-Level Distribution Matching	Dataset-Level Distribution Matching	Spatial Coverage or Diversity	Goal Progress or Completion
Trajectron [IP19]	✓			✓			
TrafficSim [SRCU21]	✓	✓	✓			✓	
SimNet [BYS+21]	✓		✓				
BITS [XCIP22]		✓	✓		✓	✓	
Nocturne [VLY+23]	✓		✓				✓
TrafficBots [ZLD+23]	✓		✓	✓			
WOSAC [MLM+23]	✓	✓	✓	✓			

Table 2.1.: **Overview of evaluation protocols for traffic simulation.** Table taken and adapted from [MLM+23].

evaluation protocol within their Waymo Open Sim Agent Challenge (WOSAC), establishing a standardized benchmark for traffic simulation [MLM+23]. By hosting public leaderboards and competitions, Waymo has built a strong research community and positioned this benchmark as a mainstream standard. In-depth details on their evaluation protocol are provided in Section 2.3.1.

Scenario Generation

The evaluation of scenario generation methods is characterized by even greater diversity and less standardization than in traffic simulation, due to the additional complexity of generating initial states or lane graphs. Following the approach of [MLM+23] for traffic simulation, we survey recent works on scenario generation and classify their evaluation protocols in Table 2.2.

Evaluation Protocol	ADE or FDE	Offroad Rate	Collision Rate	Instance-Level Distribution Matching	Dataset-Level Distribution Matching	Fidelity and Diversity	Holistic view
SceneGen [TWW+21]				✓	✓		
Scene Diffusion [PWR23]				✓			
SceneControl [LWZ+24]		✓	✓		✓		
SLEDGE [CDG24]		✓	✓	✓	✓	✓	
DriveSceneGen [SGS+24]					✓	✓	
ScenarioDreamer [RGG+25]		✓			✓	✓	
TrafficGen [FLP+23]	✓		✓	✓			
Scenario Diffusion [PGH+23]		✓		✓			
UniGen [MML+24]			✓	✓			
SceneDiffuser [JBC+24]	✓	✓	✓	✓			✓
WOSGC [Way25]		✓	✓	✓			✓

Table 2.2.: **Overview of evaluation protocols for scenario generation.** Approaches are grouped by generative capability: (Top) initial states only, (Middle) initial states with motion and lane-graphs, and (Bottom) initial states with motion.

For methods that focus solely on generating initial states of agents, evaluation is typically performed using distribution matching of key scene features such as position, bounding box size, and velocity [TWW+21, PWR23, LWZ+24]. Two popular metrics for quantifying the difference between real and generated feature distributions in this context are Maximum Mean Discrepancy (MMD) and Jensen-Shannon divergence (JSD). In addition, some works report metrics such as initial

state collision and off-road rates to further assess the physical realism [LWZ+24].

For methods that jointly generate both initial states and agent motions, evaluations are mostly divided by their subtask. While initial state quality is typically assessed as mentioned, motion realism is evaluated using metrics such as dynamic collision rates [MML+24, FLP+23], mean trajectory distance errors [FLP+23], or distribution matching distances over motion features including speed, acceleration, and proximity to other objects or road boundaries [MML+24]. More advanced works, which are capable of generating not only agent positions and motions but also lane graphs, introduce fidelity and diversity metrics to evaluate the realism and variability of their generated road layouts [CDG24, SGS+24, RGG+25].

[JBC+24] presents the first approach to evaluate all aspects of scenario generation using a single holistic metric by adapting the WOSAC evaluation framework. Instead of evaluating individual trajectories, all objects in the scene get aggregated, and the scenario gets assessed as a whole. Building on this, the Waymo Open Scenario Generation Challenge (WOSGC) employs this evaluation protocol for its recently released benchmark on scenario generation.

2.3. Waymo Open Motion Challenge

To advance research in autonomous driving, Waymo, one of the leading companies in this field, hosts annual challenges. These challenges aim to advance state-of-the-art methods in motion prediction, planning, and simulation, while encouraging collaboration and competition within the research community. In the following sections, two recent challenges are described in more detail, with a focus on their task description and evaluation framework. Table 2.3 offers a high-level overview of the tasks and highlights their differences.

Task	Input	Output	Evaluation
Traffic Simulation	- 11 steps state history (all objects) - 11 steps traffic light history - road map	- 32 rollouts - full trajectories (80 steps)	- histograms per object - selected objects
Scenario Generation	- 11 steps state history (ego vehicle) - 11 steps traffic light history - number of objects to simulate - road map	- 32 rollouts - initial states (pose, bounding boxes) - full trajectories (91 steps)	- histograms per scenario - all objects

Table 2.3.: **Overview of traffic simulation and scenario generation.** The table outlines the problem formulation for each task, including inputs, generated outputs, and evaluation approaches defined by Waymo.

2.3.1. Sim Agent

The Waymo Open Sim Agents Challenge was introduced in 2023 as the first public benchmark to address the task of simulating realistic and interactive traffic agents [MLM+23]. WOSAC defines traffic simulation as a conditional generative modeling task under the assumption of a fixed, offboard perception system. The goal is to encourage the design of traffic simulators that can be used to evaluate and train behavior models for autonomous driving. Furthermore, they introduce a data-driven evaluation framework built on the publicly accessible Waymo Open Motion Dataset (WOMD). WOMD is a large-scale motion forecasting dataset containing data mined for interactive behaviors across a diverse set of road layouts from multiple cities [ECC+21]. WOSAC uses WOMD's 9-second scenarios sampled at 10 Hz. The dataset provides about 487,000 training, 44,000 validation, and 45,000 testing scenarios.

Problem Formulation

A traffic scene can be formalized as $S = (C, A)$. $C = \{M, \Gamma\}$ represents the context of the scene, which includes a high-definition (HD) map M and the location and state of traffic lights Γ over a time horizon $T = 91$ with timesteps $t \in \{0, \dots, T-1\}$. The set of agents $A = \{a^i\}, i \in \{1, \dots, N\}$ denotes the N traffic agents present in the scenario, including the ego vehicle. Each agent a^i can be described as its trajectory a^i_t over T timesteps, while a^i_0 is the initial state. At all timesteps, the agent's state is characterized by the center positions (x^i_t, y^i_t, z^i_t), the sizes (w^i, l^i, h^i), heading angle θ^i_t, and velocity v^i_t. In the WOMD there are three different agent types τ ={vehicle, pedestrian, cyclist}. For the Sim Agent Challenge, Waymo defines the task as follows:

Given a scene context C, the past eleven timesteps of all agents' trajectories $a^i_{:10}$, the task is to model the future behavior of $A_{11:}$ in $R = 32$ rollouts over 80 additional timesteps. In other words, the goal is to model the conditional distribution of all agent states:

$$p(A_{11:} \mid C, A_{:10})$$

such that we can sample R scenarios which yield a realistic and diverse traffic behavior. Furthermore, the challenge requires factoring the traffic agents from the ego vehicle, such that the traffic model can be used with new releases of the ego policy. Formally, this can be written as:

$$p(a^{ego}_{11:}, A^{traffic}_{11:} \mid C, A_{:10}) = \prod_{t=11}^{T} \pi_{ego}(a^{ego}_t \mid a^{ego}_{<t}, a^{traffic}_{<t}, C) \cdot p(A^{traffic}_t \mid a^{ego}_{<t}, a^{traffic}_{<t}, C)$$

where $A^{traffic} := A \setminus \{a^{ego}\}$, and $a^{traffic}_t$ refers to the combined states of all agents except for the ego vehicle at time t. Further factorizing the traffic model $p(A^{traffic}_t \mid a^{ego}_{<t}, a^{traffic}_{<t}, C)$ is also allowed but not required. Bounding box sizes are assumed to

remain fixed from the last observed timestep, and any objects appearing after the first 11 timesteps are not considered.

Evaluation Metric

To assess the quality of simulated agents, WOSAC introduced a distribution-matching evaluation framework, based on point-to-distribution distance measuring. The key objective is to measure how closely the generated agent behaviors align with those observed in the real world. Since the true analytic form of the real-world driving distribution is unknown, WOSAC approximates the distribution using WOMD, which provides representative samples of real traffic scenarios [MLM+23].

Specifically, the challenge evaluates submissions on the negative likelihood (NLL) of selected ground truth trajectories under the distribution induced by the same selected and simulated agents. To ensure a consistent evaluation across submissions, predefined histograms are fitted to the trajectories of the 32 generated rollouts. The NNL of the ground truth trajectory samples is then computed under the normalized categorical distribution derived from these histograms. An illustrative example of this computation for a single feature is provided in Figure 2.2. This computation is performed for $M = 10$ behavioral features, as described in Table 2.4, resulting in one component metric per feature. Finally, all component metrics are weighted and aggregated into a single realism score that captures all aspects in one comprehensive metric, the Realism Meta Metric (RMM).

Metric Group	Feature	Weight	Num Bins	Min Val	Max Val	Ind. Timesteps
Kinematic	Linear Speed (LINS)	0.05	10	0.0	25.0	True
	Linear Acceleration (LINA)	0.05	11	-12.0	12.0	True
	Angular Speed (ANGS)	0.05	11	-0.628	0.628	True
	Angular Acceleration (ANGA)	0.05	11	-3.14	3.14	True
Interactive	Collision Indication (COLL)	0.25	2	0	1	False
	Distance to Nearest Object (DTNO)	0.10	10	-5.0	40.0	True
	Time to Collision (TTC)	0.10	10	0.0	5.0	True
Map-based	Off-road Indication (OFFR)	0.25	2	0	1	False
	Traffic Light Violation (TLV)	0.05	2	0	1	False
	Distance to Road Edge (DTRE)	0.05	10	-20.0	40.0	True

Table 2.4.: **Waymo metric configuration.** Overview of the component metrics used in WOSAC and WOSGC evaluations. Each metric belongs to one of three groups and contributes to the overall realism score with an associated weight. The last four columns specify the histogram configurations used to approximate the empirical feature distributions of the simulated trajectories.

Formally, $R = 32$ samples are drawn from the distribution p for each scene. Each sample spans $T = 80$ timesteps and contains A with N agents. Let the generated data be denoted by $\hat{x}(1:N, 1:R, 1:T, 1:D)$ with $D = 7$-dim vector consisting of location (x, y, z), size (w, l, h) and orientation θ. Let the ground truth data be denoted

by $x(1:N', 1:T, 1:D)$. Since the ground truth can contain agents that enter and leave the scene after the initial prefix, a validity mask $v(1:N', 1:T)$ is necessary to exclude ground truth data points from the evaluation. The following describes the mathematical computation of the realism score for one scenario.

Let $F_{j\_sim}(\hat{x}(a,:))$ the set of feature samples of type j derived from $\hat{x}(a, 1:R, 1:T)$ by pooling over T, and R. These samples are used to compute a histogram $p_{ja\_sim}(.)$, specifically, the empirical distribution of F_j for agent a in this scenario.

Similarly, let $F_{j\_gt}(x(a,t))$ denote the value of feature j from the ground truth trajectory of agent a at time t. This applies to all features except for the three dependent-timestep binary features OFFR, COLL, and TLV. For both x and $\hat{x}$, values are aggregated by taking the maximum valid value over T before forming the feature set. In other words, if a trajectory exhibits a collision or goes off-road at any timestep across the rollouts, the corresponding feature is assigned a value of 1.

For independent-timestep features, the negative log likelihood for each ground truth sample gets computed as:

$$NLL(a,t,j) = -\log p_{ja\_sim}(F_{j\_gt}(x(a,t))) \tag{2.1}$$

and for dependent-timestep features:

$$NLL(a,j) = -\log p_{ja\_sim}(F_{j\_gt}(x(a))) \tag{2.2}$$

Then, the j-th component metric for agent a is defined as:

$$m_j(a) = \exp\left(-[\frac{1}{V(a)}]\sum_t^T v(a,t)NLL(a,t,j)\right) \tag{2.3}$$

$$m_j = \frac{1}{N_{eval}}\sum_a^{N_{eval}} m_j(a) \tag{2.4}$$

for independent-timestep features, where $V(a) = \sum_t^T v(a,t)$ denotes the number of valid samples for agent a in the ground truth data. Here, N_{eval} denotes a subset of N, as only a predefined set of agents is evaluated. The number of N_{eval} varies between $\approx$ 2-15 agents per scenario. Since the dependent-timestep features yield a single value per agent and all N_{eval} evaluated agents are valid, the metric is defined as:

$$m_j(a) = \exp\left(-NLL(a,j)\right) \tag{2.5}$$

$$m_j = \frac{1}{N_{eval}}\sum_a^{N_{eval}} m_j(a) \tag{2.6}$$

Finally, all component metrics m_j are weighted and aggregated into the realism meta metric score:

$$RMM = \frac{1}{M}\sum_{j=1}^{M} w_j m_j \tag{2.7}$$

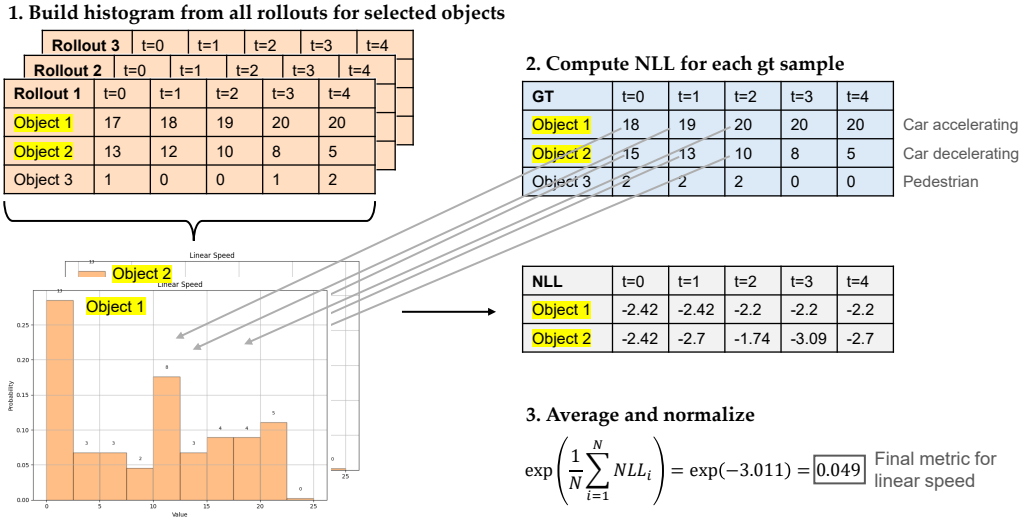


Figure 2.2.: **Example of metric computation for linear speed feature.** (1) For each predefined object (highlighted in yellow), a histogram is created from all feature samples of that object across generated rollouts. (2) The NLLs of the ground-truth samples per object are then computed under these distributions. (3) NLL values are averaged and normalized.

The final leaderboard score is computed as the average of all RMM scores across all scenarios in the test split.

2.3.2. Scenario Generation

In addition to the WOSAC and two other challenges, Waymo introduced the Waymo Open Scenario Generation Challenge (WOSGC) [Way25]. This represents the first publicly available benchmark for evaluating scenario generation methods in a standardized manner. It was released on 1^{st} of April 2025, and participants were given 6 weeks to submit their results. Similar to WOSAC, the challenge uses the WOMD's 9-second scenarios sampled at 10 Hz.

Problem Formulation

The previously described traffic scene formalization continues to apply in this setting. For the WOSGC, however, Waymo defines the task as the following:

Given a scene context C, eleven history timesteps of the ego vehicles' trajectory $a^{ego}_{:11}$ and the number N^{τ} of agents per type to be generated, the task is to synthesize A in $R = 32$ rollouts over $T = 91$ timesteps. Unlike WOSAC, this also includes the initial

states. Formally, the goal is to model the conditional distribution of all agent states:

$$p(A \mid C, a^{ego}_{:11}, N^{\tau})$$

such that R scenarios can be sampled to produce realistic and diverse traffic scenarios.

Evaluation Metric

Due to the lack of comprehensive and standardized evaluation metrics for scenario generation, as described in Section 2.2.2, Waymo employs a slightly modified version of the WOSAC metrics to assess the realism of generated scenarios. In the WOSAC metric, NLLs are computed per agent. Here, they are computed per scenario. Consequently, histograms are constructed by aggregating not only over time and rollouts but also across the agent dimension, reflecting the absence of a one-to-one correspondence between agents in the generated and logged scenarios. Furthermore, rather than evaluating only a predefined subset N^{eval} as in WOSAC, all N agents are included when forming the histograms.

As a result, the set of simulated feature samples of type j can be defined as $F_{j\_sim}(\hat{x}(:))$, derived from $\hat{x}$ by pooling over T, R, and also N. The histogram $p_{j\_sim}(.)$, constructed from $F_{j\_sim}(\hat{x}(:))$, represents then the empirical distribution of F_j across all agents.

Using this, the NLL can be defined as:

$$NLL(a,t,j) = -\log p_{j\_sim}(F_{j\_gt}(x(a,t)) \tag{2.8}$$

and for dependent timestep features:

$$NLL(a,j) = -\log p_{j\_sim}(F_{j\_gt}(x(a)) \tag{2.9}$$

Then the j'th component metric is defined as:

$$m_j = \exp\left(-[\frac{1}{V}]\sum_{t}^{T}\sum_{a}^{N} v(a,t)NLL(a,t,j)\right) \tag{2.10}$$

for independent-timestep features with $V = \sum_t^T \sum_a^N v(a,t)$ as the number of valid samples in the ground truth. And for dependent-timestep features with $V = \sum_a^N v(a)$:

$$m_j = \exp\left(-[\frac{1}{V}]\sum_{a}^{N} v(a)NLL(a,j)\right) \tag{2.11}$$

Finally, all component metrics m_j are weighted and aggregated into the RMM score, as in WOSAC. The challenge leaderboard score is computed as the average of all meta-metric scores across all samples in the test split.

3. Interactive Fidelity and Diversity Assessment

This chapter presents the fidelity and diversity metrics we propose to evaluate traffic and scenario generation. After outlining the theoretical foundations of several metric pairs, we extend them with conditional formulations tailored for an instance-level evaluation. The second part of this chapter describes the main components of our interactive tool for assessing traffic and scenario generation using fidelity and diversity metrics.

3.1. Prerequisites

For assessing traffic and scenario generation, we advocate the use of fidelity and diversity metrics, as they provide a more comprehensive evaluation of generative performance, capturing both realism and the variability across the generated data. In contrast, one-dimensional metrics such as Inception Score and Fréchet Inception Distance offer limited diagnostic power and assume the existence of a pretrained network that captures meaningful domain embeddings. These metrics further provide only a coarse, dataset-level overview, precluding detailed instance-level analysis. Similarly, (V)LLMs are less robust for evaluating traffic scenes [WLAA25] due to the highly domain-specific nature of the data, which is governed by complex temporal and physical dynamics. Furthermore, traffic scenarios comprise structured, multimodal data such as trajectories and semantic maps, which lie beyond the representational capabilities of internet-pretrained models.

Paper	Fidelity Metric	Diversity Metric	Highlights
[KKL+19]	Improved Precision (I-Prec)	Improved Recall (I-Rec)	intuitively and widely adapted
[NOU+20]	Density	Coverage	robust to outliers and computationally efficient
[PK23]	Probabilistic Precision (P-Prec)	Probabilistic Recall (P-Rec)	robust to outliers and probabilistic scoring rule

Table 3.1.: **Fidelity and diversity metrics considered in this work.**

Specifically, we select three distinct pairs of fidelity and diversity for our analysis, summarized in Table 3.1. I-Prec and I-Rec are included due to their wide adoption and recognition in the literature. The remaining two pairs are chosen to address

specific limitations identified in the metric proposed by [KKL+19]. Each pair introduces improvements from different methodological perspectives. The following mathematically describes all metrics in detail and highlights their differences.

3.1.1. Metrics

To formally describe the metrics in detail, we adopt the notation of [RvBvdS25], which offers a unified and standardized formulation for all fidelity and diversity metrics. As already mentioned in Section 2.1.2, all fidelity and diversity metrics are computed using embeddings of the data points. The embedding of a single data point is denoted as ϕ, and the set of all embeddings as Φ. Additionally, the subscripts Φ_r and Φ_g are used to differentiate between real and generated data. The notation $|\Phi|$ represents the number of samples in the dataset Φ.

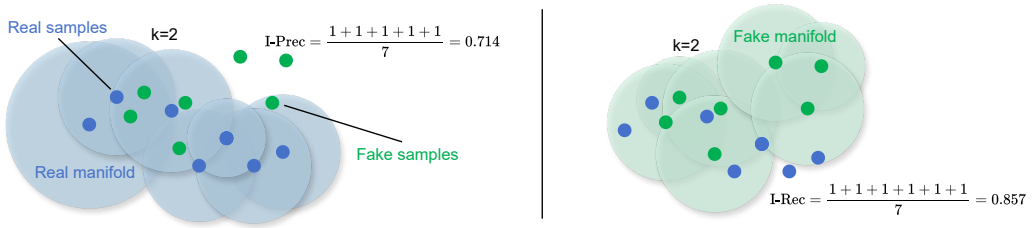


Figure 3.1.: **Improved Precision and Recall.** Real and fake manifolds are approximated by surrounding each sample within a hypersphere defined by its kth-nearest-neighbor distance.

Improved Precision and Recall

The main idea of the I-Prec and I-Rec is approximating the support or manifold S of the real and generated data distribution with a set of hyperspheres around each data point [KKL+19]. The radius r of the hypersphere is defined to be the distance to the k-th nearest neighbor of each point.

$$S(\Phi) = \bigcup_{\phi \in \Phi} B(\phi, NN_k(\phi, \Phi)) \tag{3.1}$$

with $B(\phi, r)$ as a hypersphere with radius r centered at ϕ, and $NN_k(\phi, \Phi)$ as the distance to the k-th nearest neighbor of ϕ in Φ, excluding itself. Next, a binary function is defined to determine whether a given sample ϕ is located within the support:

$$f(\phi, \Phi) = \begin{cases} 1, & \text{if } \|\phi - \phi'\|_2 \leq \|\phi' - NN_k(\phi', \Phi)\|_2 \text{ for at least one } \phi' \in \Phi \\ 0, & \text{otherwise,} \end{cases} \tag{3.2}$$

Finally, I-Prec counts the fraction of generated data points that lie in the support of the real data, while I-Rec counts the fraction of real points that are in the support of the generated data.

$$\text{I-Prec}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\phi_g \in \Phi_g} f(\phi_g, \Phi_r) \tag{3.3}$$

$$\text{I-Rec}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} f(\phi_r, \Phi_g) \tag{3.4}$$

A simple example is illustrated in Figure 3.1. For the parameter k, the authors suggest using the value 3. Since this approach approximates the support with a fixed k for all data points, it implicitly assumes a constant density across the entire distribution. Consequently, the method does not normalize with respect to the hypersphere radii or relative density of samples. [NOU+20, PK23] pointed out that this simplification can be problematic in practice, especially in the presence of outliers. For precision, the kNN method can lead to an overestimation of the true manifold around an outlier. As a result, generating multiple samples around such outliers can artificially inflate the precision score (see (a) in Figure 3.2).

Similarly, for recall, models that produce highly diverse samples (outliers) can overestimate the generated data manifold. Due to the lower sample density, the distances to their k-nearest neighbors increase, leading to larger hyperspheres and an inflated estimate of the generated manifold. Consequently, real samples are more likely to fall within the generated support, thereby falsely increasing recall (compare (c) in Figure 3.2).

Density and Coverage

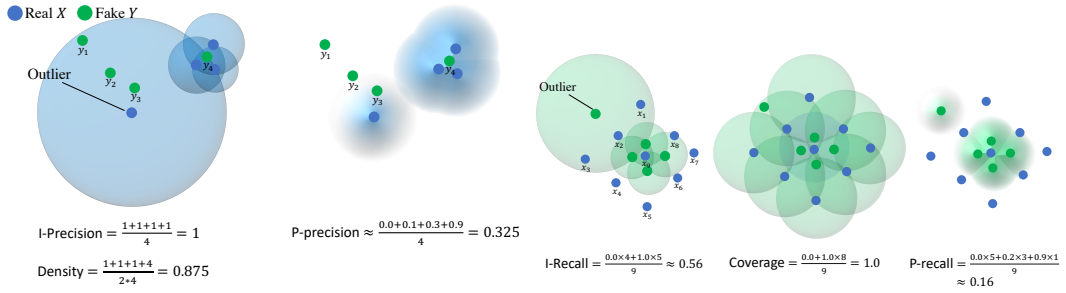
To overcome the weaknesses of the I-Prec and P-Rec, [NOU+20] propose the density and coverage metrics. Density is defined as

$$\text{density} = \frac{1}{k|\Phi_g|} \sum_{\phi_g \in \Phi_g} \sum_{\phi_r \in \Phi_r} 1_{\phi_g \in B(\phi_r, NN_k(\phi_r, \Phi_r))} \tag{3.5}$$

and counts, for each generated data point ϕ_g, the number of real data hyperspheres that include ϕ_g. The result is then normalized by both the sum over all generated data and the neighborhood size k.

$$\text{coverage} = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} 1_{\exists \phi_g \in \Phi_g \,:\, \phi_g \in B(\phi_r, NN_k(\phi_r, \Phi_r))} \tag{3.6}$$

Coverage counts the fraction of real data points that have at least one generated data point in their hypersphere. This means coverage uses the approximate manifold of



(a) I-Precision / Density (b) P-precision (c) I-Recall (d) Coverage (e) P-recall

Figure 3.2.: **Robustness to outliers.** Each metric pair applies its own scoring rule and handles outliers differently. I-Prec and density (a) overestimate outlier regions, whereas P-Prec (b) normalizes neighborhood sizes and applies a probabilistic scoring rule, yielding greater robustness. Similar effects are illustrated for the corresponding diversity metrics (c-e). The figure is taken from [PK23] and $k = 2$.

the real data distribution. The authors claim that the metric is more robust to outliers because outliers in the real data distribution are less common, and generated outliers do not influence it. While coverage is bounded between 0 and 1, the density can exceed 1 if the density of generated data points is high near a real data point.

However, despite the improvements, [PK23] find that density is still vulnerable to outliers, which can be seen in (a) of Figure 3.2. Coverage on the other side is fully independent of the distances between generated samples, as it only measures how many real hyperspheres are occupied by generated samples. This introduces a limitation in sensitivity to the relative diversity between real and generated distributions. For example, if the real data distribution is relatively sparse, both highly diverse and more dense generated distributions may yield similar coverage scores (see (d) in Figure 3.2 as an example for a more dense generated distribution).

Probabilistic Precision and Recall

Unlike Density and Coverage, P-Prec and P-Rec address the outlier-robustness problem by approximating the supports probabilistically [PK23]. Instead of assigning a binary value to indicate whether a data point lies within the approximate support, they estimate the probability that the point belongs to the support. The authors define a probabilistic scoring rule (PSR), which estimates the probability of ϕ to be in the support of Φ as,

$$PSR_{\Phi}(\phi') = 1 - \prod_{\phi \in \Phi} (1 - f(\phi, \phi', R_{\Phi})) \tag{3.7}$$

with $f(\phi, \phi', R_\Phi)$ being a simple estimate of the probability that ϕ' is in the support around ϕ:

$$f(\phi, \phi', R) = \begin{cases} 1 - \frac{\|\phi - \phi'\|_2}{R}, & \text{if } \|\phi - \phi'\|_2 \leq R \\ 0, & \text{otherwise,} \end{cases} \tag{3.8}$$

where R_Φ denotes the average k-th nearest neighbor distance within Φ, scaled by the parameter a.

$$R_\Phi = \frac{a}{|\Phi|} \sum_{\phi \in \Phi} NN_k(\phi, \Phi) \tag{3.9}$$

In the end, all probability estimates are averaged to produce the overall metric.

$$\text{P-Prec}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\phi_g \in \Phi_g} PSR_{\Phi_r}(\phi_g) \tag{3.10}$$

$$\text{P-Rec}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} PSR_{\Phi_g}(\phi_r) \tag{3.11}$$

The authors set the parameters to default values of $a = 1.2$ and $k = 4$. The improvements of this metric over prior ones in their failure cases are shown in Figure 3.2 (b) and (e), for precision and recall, respectively. However, it still fails to satisfy several sanity checks for clear upper and lower bounds, as studied by [RvBvdS25].

3.1.2. Conditional Fidelity and Diversity

In the domain of traffic and scenario generation, the generative task is conditioned on specific contexts, such as road layouts, the number of objects, or initial states. Consequently, evaluation metrics should account for these conditioning factors to accurately assess model performance and the quality of the generated data. The Waymo metric achieves this by computing the log-likelihood of the ground truth with respect to their corresponding generated trajectories or scenarios, and then averaging the result across the dataset. This provides instance-level diagnostics and enables a more detailed and interpretable analysis of model performance.

To obtain similar capabilities for fidelity and diversity metrics, we extend their traditional definitions to a conditional setting. Since all fidelity and diversity metrics depend on the underlying manifold, and thus on the distribution of the data, we continue to construct the manifold using all data points. This preserves the global geometric structure of the support, and nearest neighbors are ideally from similar instances. However, when testing membership for the conditional metrics, a data

point is compared only to points in its corresponding counterpart set, rather than to the entire dataset. For I-Prec and I-Rec we can redefine the Equations 3.3 and 3.4 as:

$$\text{con-I-Prec}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\Phi_g^{(s)} \in \Phi_g} \sum_{\phi_g \in \Phi_g^{(s)}} f(\phi_g, \Phi_r^{(s)}) \tag{3.12}$$

$$\text{con-I-Rec}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\Phi_r^{(s)} \in \Phi_r} \sum_{\phi_r \in \Phi_r^{(s)}} f(\phi_r, \Phi_g^{(s)}) \tag{3.13}$$

$$= \frac{1}{|\Phi_r|} \sum_{\phi_r^{(s)} \in \Phi_r} f(\phi_r^{(s)}, \Phi_g^{(s)}) \quad \text{since } |\Phi_r^{(s)}| = 1 \text{ in our case} \tag{3.14}$$

Here, the superscript (s) denotes the data points which belong to instance s. In our context, this corresponds either to a set of trajectories or a set of scenarios generated from multiple rollouts. Since each instance has only one ground truth counterpart, the size of the set $\Phi_r^{(s)}$ is equal to one when computing the metrics.

Similarly, for P-Prec and P-Rec, we can express the conditioned forms of Equations 3.10 and 3.11 as follows:

$$\text{con-P-Prec}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\Phi_g^{(s)} \in \Phi_g} \sum_{\phi_g \in \Phi_g^{(s)}} conPSR_{\Phi_r^{(s)}}(\phi_g) \tag{3.15}$$

$$\text{con-P-Rec}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\Phi_r^{(s)} \in \Phi_r} \sum_{\phi_r \in \Phi_r^{(s)}} conPSR_{\Phi_g^{(s)}}(\phi_r) \tag{3.16}$$

$$= \frac{1}{|\Phi_r|} \sum_{\phi_r^{(s)} \in \Phi_r} conPSR_{\Phi_g^{(s)}}(\phi_r^{(s)}) \quad \text{since } |\Phi_r^{(s)}| = 1 \text{ in our case} \tag{3.17}$$

while keeping R_Φ on global scale, we compute the PSR on the instance level:

$$conPSR_{\Phi^{(s)}}(\phi') = 1 - \prod_{\phi \in \Phi^{(s)}} (1 - f(\phi, \phi', R_\Phi)) \tag{3.18}$$

For the conditioning of density and coverage, the Equation 3.5 can be expressed as:

$$\text{con-density} = \frac{1}{k|\Phi_g|} \sum_{\Phi_g^{(s)} \in \Phi_g} \sum_{\phi_g \in \Phi_g^{(s)}} 1_{\phi_g \in B(\phi_r^{(s)}, NN_k(\phi_r^{(s)}, \Phi_r))} \tag{3.19}$$

where we consider only one ground truth sample per instance, while summing over all generated samples from instance s that fall within the neighborhood of the corresponding real data point $\phi_r^{(s)}$. And for Equation 3.6:

$$\text{con-coverage} = \frac{1}{|\Phi_r|} \sum_{\phi_r^{(s)} \in \Phi_r} 1_{\exists \phi_g \in \Phi_g^{(s)} \,:\, \phi_g \in B(\phi_r^{(s)}, NN_k(\phi_r^{(s)}, \Phi_r))} \tag{3.20}$$

All these conditional formulations allow us to retain the global geometric structure of the data distribution while evaluating generated samples specifically against their corresponding ground truth instance. As a result, it provides a more detailed assessment of fidelity and diversity in a context-aware and instance-conditioned setting.

3.2. Interactive Components

This section describes the core components of our visualization tool for evaluating generative tasks using fidelity and diversity metrics. Although the primary focus of this work lies on assessing traffic and scenario generation, the tool can be expanded to other domains where human evaluation remains crucial. Its purpose is to provide researchers with an interactive tool for exploring and analyzing generated samples. To this end, the tool integrates dimensionality reduction methods to visualize data manifolds and provides dashboard-like overviews of evaluation metrics. We also include capabilities for detailed per-sample and per-method analysis. Together, these components allow both a high-level understanding of model performance and a deeper qualitative inspection of specific samples.

3.2.1. Manifold Visualization

Figure 3.3.: **Dashboard view.** The interface provides an interactive manifold visualization of real and generated distributions, complemented by an overview of relevant evaluation metrics. Users can explore the underlying data structure, inspect nearest neighbors, and examine correspondences between ground-truth and generated samples.

To provide users with an intuitive way of exploring their data, we employ dimensionality reduction methods to project high-dimensional data samples into

two-dimensional (2D) or three-dimensional (3D) spaces. This allows for visual inspection of similarities and differences between real and generated data distributions.

For dimensionality reduction, we consider two complementary nonlinear techniques. The first is t-distributed Stochastic Neighbor Embedding (t-SNE) [vdMH08], which constructs a low-dimensional embedding by minimizing the Kullback–Leibler divergence between high- and low-dimensional pairwise similarity distributions, thereby emphasizing the preservation of local neighborhood structure. In this work, we use the *sklearn* implementation with a perplexity of 20 and random initialization, while keeping all other parameters at their defaults.

The second option is Uniform Manifold Approximation and Projection (UMAP) [MHM20]. UMAP models the high-dimensional data manifold as a fuzzy topological structure and optimizes a low-dimensional representation by minimizing a corresponding cross-entropy objective. Compared to t-SNE, UMAP tends to retain more global structure while still preserving local relationships. We employ the official *UMAP* implementation with default parameters.

Figure 3.3 shows the visualization of the embeddings produced by the dimensionality reduction methods within the frontend. The interface allows users to zoom, pan, and inspect individual data points through hover interactions. Hovering over a data point reveals more information, such as the three nearest neighbors in feature space (highlighted in green) and the corresponding counterpart(s) from the other distribution (highlighted with increased size and a red outline). This interaction visualizes the conditional correspondence between ground-truth and generated samples associated with the same underlying trajectory or scenario. Further details such as the distance to the counterpart, per-sample metric scores, and identifiers are also displayed. Selecting a data point opens a detailed neighbor comparison view, allowing for fine-grained inspection of the sample and its local neighborhood.

3.2.2. Nearest Neighbor Analysis

Since k-nearest neighbors are already computed for the fidelity and diversity metrics, our tool leverages this information to support neighbor-based inspection. This analysis provides users with insights into how samples relate to their closest points and how they differ from each other.

Depending on the application domain, users can explore both the original data and their feature representations, if applicable. In our case, the original data can be visualized as traffic scenes rendered as a 2D Graphics Interchange Format (GIF) file. Since we summarize trajectories and scenarios into handcrafted feature histograms, we can also illustrate them. This dual perspective can be valuable for assessing whether the feature space adequately captures meaningful similarities. Shortcomings in the representation can be revealed by samples that may appear close in feature space but differ substantially in raw form. On the other hand, high similarity across

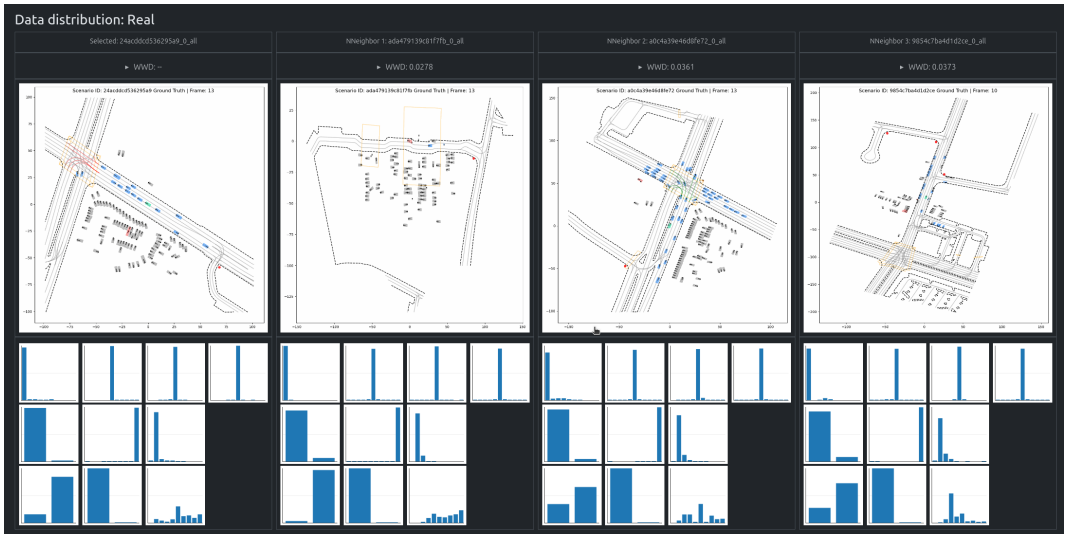


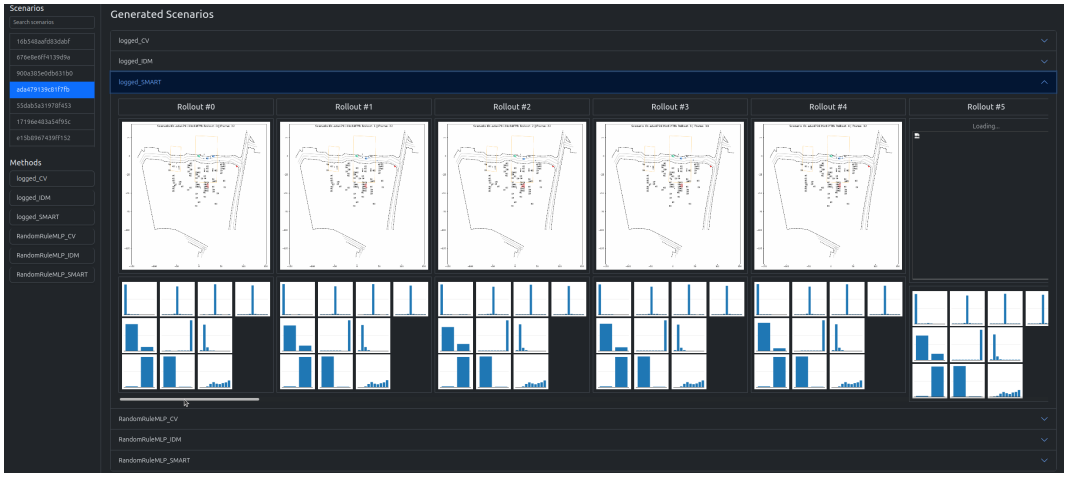
Figure 3.4.: **Nearest Neighbor Analysis.** Our tool allows users to inspect how individual samples relate to their closest neighbors in feature space and in the original data, visualized as GIFs. In this example, a traffic scenario with many off-road parked cars is highlighted, showing that the nearest neighbors share similar scene characteristics.

both spaces provides confidence that the features align with human perception. An example analysis is shown in Figure 3.4. A scenario with many off-road parked cars was selected, as indicated by the high bar on the right side of the first histogram in the last row. Looking at the raw data from the nearest neighbors, their scenes look similar: numerous off-road parked vehicles, along with fewer on-road vehicles either approaching a traffic light or driving at low speed.

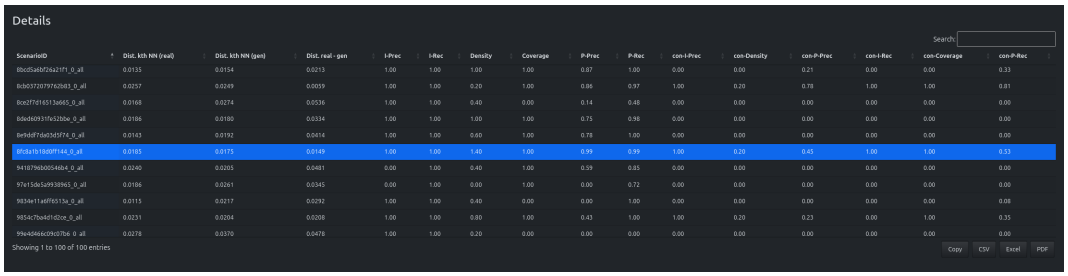
3.2.3. Detailed Method and Sample Comparison

A central part of developing generative models is to compare different methods and determine which performs best for a given task. However, evaluating generative models remains challenging due to missing ground truth data and a lack of reliable metrics [NOU+20, SCH+23, RvBvdS25, RSGA25]. To address this, our tool supports both quantitative and qualitative comparison across methods and generated samples.

Qualitatively, our tool provides sample-level comparisons, allowing generated outputs to be directly inspected alongside their corresponding ground truth scenes. This facilitates visual assessment and can aid data curation for downstream tasks requiring specific scenario characteristics. Selecting a real data point reveals the associated generated samples from all methods, as illustrated in Figure 3.5 (b). Combined with the manifold visualization introduced in Section 3.2.1, users obtain an overview of the distributions produced by different approaches, where deviations



(a)



(b)

Figure 3.5.: **(a) Per-method comparison.** Users can visually inspect generated samples alongside their corresponding ground truth scenes across different methods. **(b) Per-sample quantitative results.** A table presents detailed metrics for each sample, including distances to counterparts, hypersphere radii, and individual fidelity and diversity scores.

from the ground-truth manifold indicate poorer performance.

Quantitatively, the tool aggregates various fidelity and diversity metrics for each method and displays them in the dashboard view (Figure 3.3). We additionally report the Waymo Realism Metric to complement these scores. A detailed per-sample table provides further information such as hypersphere radii, distances to counterparts, and individual metric values (Figure 3.5 (a)).

By combining both qualitative and quantitative perspectives, the tool allows researchers to balance objective metric scores with subjective human judgment, which remains essential in generative domains [TvdOB16, HL08].

4. Case Studies: Traffic and Scenario Generation

This chapter presents case studies demonstrating how fidelity and diversity metrics can be practically applied to evaluate traffic and scenario generation tasks. In Section 4.1, we introduce all the methods evaluated in this study. Section 4.2 delineates the experimental design and details the series of experiments conducted as part of our case studies. Finally, Section 4.3 reports the quantitative and qualitative results for both traffic and scenario generation tasks, incorporating our new evaluation metrics. For further insights, we provide supplementary figures in the Appendix A.

4.1. Methods

The first section of this chapter provides a comprehensive overview of the methods employed in our experiments. To ensure modularity, extensibility, and reproducibility, we develop a toolkit named SHED (Synthesizing Environments for Driving). SHED integrates different methods for scenario and traffic generation within a unified workflow, as illustrated in Figure 4.1. In the first stage, initial states are generated and stored as intermediate files. This enables the reuse of generated states across subsequent methods, facilitating efficient experimentation and reducing computational redundancy. The first stage is particularly relevant for scenario generation, where the use of historical trajectory data is not allowed. The second stage involves simulating the motion of all traffic participants. For rule-based approaches, we utilize the Waymax simulator [GFL+23] to execute closed-loop simulations. The toolkit supports the integration of learning-based methods for both traffic simulation and initial state generation, enabling future extensions beyond our baselines. Finally, simulated trajectories are serialized and persisted as pickle files, enabling downstream evaluation and visualization workflows.

4.1.1. Traffic Simulation

Traffic simulation focuses on realistic motion forecasting of traffic participants, as detailed in Section 2.3.1. We compare seven motion forecasting approaches with varying complexity and realism. Method variations enable systematic assessment of evaluation metric sensitivity. The following subsections briefly describe the methods in more detail.

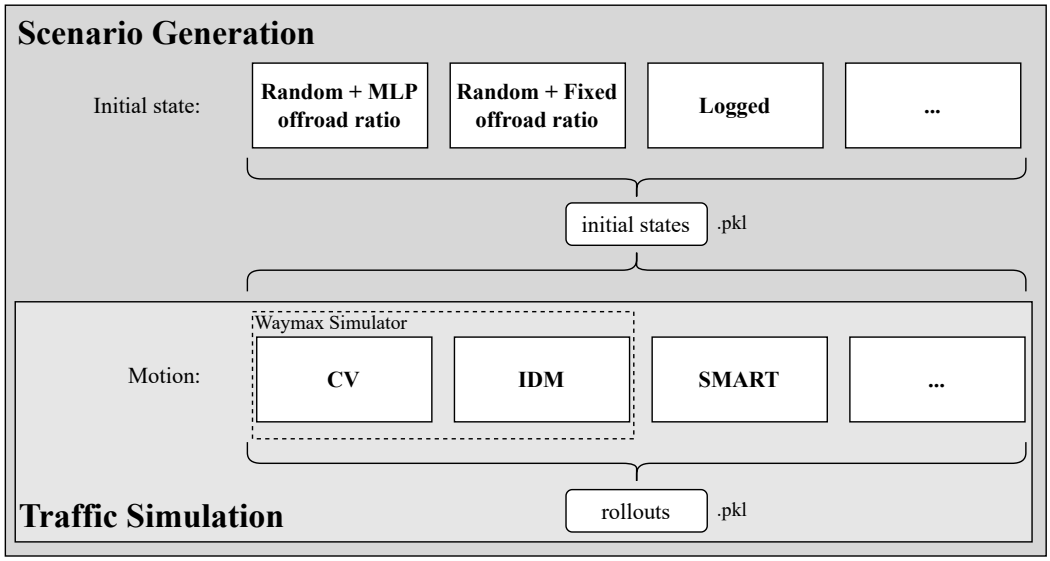


Figure 4.1.: **SHED.** Comprehensive overview of methods integrated within SHED for scenario and traffic generation. SHED can be easily extended with additional methods.

Constant Velocity

The **logged_CV** baseline implements a Constant Velocity (CV) assumption. The model computes each agent's velocity based on its last two logged states from the historical trajectory and extrapolates future positions in the direction of the current heading. We add Gaussian noise to the velocity to promote trajectory variability across rollouts. A slight modification, denoted as **logged_CV_slow**, reduces the extrapolated velocity to half of the logged speed. We expect fewer collisions and off-road violations, though the driving behavior remains unrealistic.

Intelligent Driver Model

The Intelligent Driver Model (IDM) [THH00] is a deterministic, rule-based driving model that computes a vehicle's longitudinal acceleration based on its distance and relative velocity to the preceding vehicle. On free roads, it allows the car to accelerate toward a predefined desired speed. We evaluate three IDM variants using the built-in Waymax implementation [GFL+23].

The **logged_IDM** variant follows logged lateral trajectory while computing longitudinal motion via IDM dynamics. To ensure compliance with traffic signals, we extend the default Waymax implementation by overriding the acceleration output whenever a vehicle approaches a red traffic light. For model configuration, we use the default IDM parameters from Table 4.1, with additive noise (e.g. ±1.0 to

Parameter	Default Value	Description
desired_vel	v_{adv_lane}	Desired velocity. Either adv lane speed limit or $v_{adv\_lane} = 30$ *mph* if speed limit is not available.
min_spacing	2.0 m	Minimum spacing to the leading vehicle.
safe_time_headway	2.0 s	Desired time headway to leading vehicle.
max_accel	2.0 m/s^2	Maximum acceleration of vehicle.
max_decel	4.0 m/s^2	Maximum deceleration of vehicle.
delta	4.0	Acceleration exponent.

Table 4.1.: **Default IDM parameters.**

max_accel) to promote stochasticity across rollouts.

Secondly, we define **logged_IDM_aggressive**, which represents a more assertive driving behavior. By increasing the *desired_vel* with a factor of 1.5, higher speeds are expected. Furthermore, we adjust the *min_spacing* and the acceleration/deceleration parameters to less conservative values, allowing vehicles to follow more closely and respond more abruptly to traffic dynamics.

The **IDM** variant modifies the default Waymax IDM behavior such that, instead of following logged trajectories, agents follow randomly generated paths along road centerlines, starting from their last known history state. All Off-road vehicles remain static throughout the simulation. This approach is particularly important for tasks where logged trajectories are unavailable, such as in the scenario generation task (see Section 4.1.2).

In all IDM-based simulations, pedestrians maintain constant velocity and heading, while cyclists employ IDM with reduced desired velocity (10 *mph*) to reflect typical motion patterns.

SMART

The **SMART** (Scalable Multi-Agent Real-Time Motion Generation) model [WFGK24] represents a state-of-the-art learning-based approach for multi-agent motion forecasting in traffic simulation. It serves as the base architecture for all top-performing models in its domain and ranked 1st on the WOMD leaderboard in 2024. SMART formulates motion forecasting as a next-token prediction task, in which a classifier iteratively predicts discrete trajectory tokens. This is achieved by transforming both the vectorized map features and agent trajectories into a unified sequence of tokens, enabling efficient autoregressive trajectory generation.

Several extensions were introduced during the Waymo Sim Agent Challenge 2025, including closed-loop [ZKI+25] and reinforcement-learning [PSS25] fine-tuned variants, as well as improved tokenization approaches [ZJC+25]. However, most of these models have not been made publicly available. To ensure reproducibility, we employ the original SMART model without additional modifications. We use the lightweight

SMART-nano variant, which was included in [ZKI+25]. This model comprises approximately 1.2 million parameters, making it well-suited to accommodate GPU memory constraints while preserving predictive capabilities. Training comprised 30 epochs on 4 NVIDIA 2080ti GPUs using default hyperparameters with minor adjustments specific to the *nano* configuration.

We define **SMART_earlyStop** as an additional variant, corresponding to an earlier checkpoint obtained at approximately 80% of the full training process.

4.1.2. Scenario Generation

Unlike traffic simulation, scenario generation does not assume prior knowledge of initial agent states or trajectory history. Therefore, a primary challenge is accurately predicting the initial scene composition. In reality, a scene can vary considerably depending on factors such as time of day, day of the week, holidays, or weather conditions.

For the scenario generation challenge described in Section 2.3.2, accurate initial state prediction is essential. Particularly regarding the off-road/on-road ratio, which the evaluation metric is highly sensitive to. Since this ratio is assumed to remain stable across the 91-timestep horizon with appropriate motion models, precise initial states are critical for strong performance.

Since this work prioritizes evaluation methodology over method development, we leverage privileged information such as the logged initial states and trajectories. This approach provides a reliable reference baseline representing "perfect" scene prediction. We adopt the naming convention ***initialStateMethod_MotionModel***, yielding three privileged scenario generation methods: **logged_CV**, **logged_IDM**, and **logged_SMART**.

Random Initial State Generator

To avoid exclusive reliance on privileged information, we implement a rule-based approach that generates initial agent states through two primary tasks: specifying the off-road/on-road ratio and positioning the objects.

Off-road Rate Regressor: Given the prevalence of off-road parked vehicles and pedestrians in the dataset, accurately predicting off-road object counts is critical for overall performance. We explore using both a constant off-road ratio (e.g., 40% off-road) and a 3-Layer Multi-Layer Perceptron (MLP) regressor trained on 9,000 randomly sampled training scenarios, taking object count, lane count, and average speed limit as inputs. The MLP achieves a mean absolute error (MAE) of ~ 5.4.

Position Initializer: Objects are positioned under simplified assumptions: all pedestrians are placed off-road, cyclists on bike lanes or on-road, and vehicles are distributed to satisfy the off-road ratio. The ego vehicle anchors placement, followed by

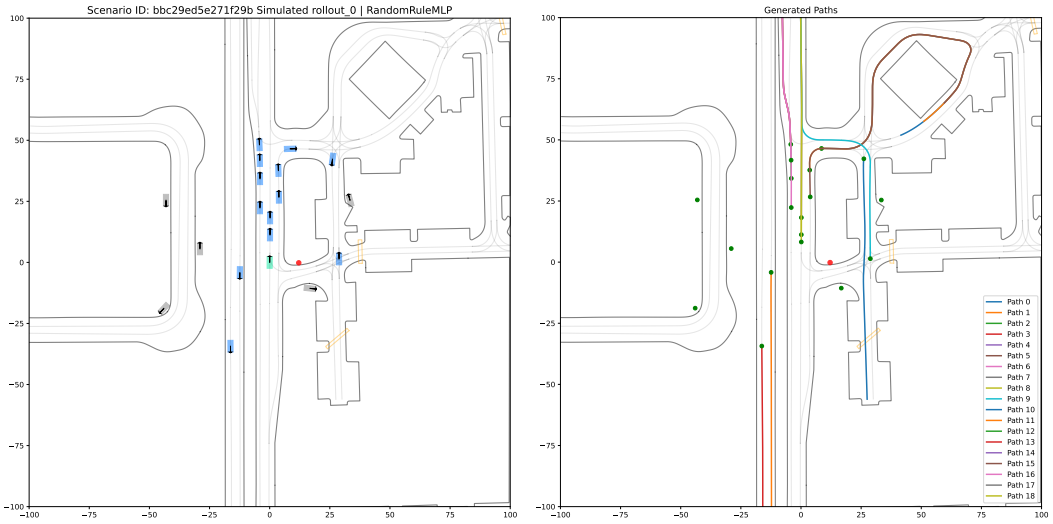


Figure 4.2.: **Rule-based scenario generation approach.** RandomRuleMLP position initializer output (left) and the corresponding generated paths for all on-road vehicles (right) when using IDM. Vehicles colored gray represent off-road objects, blue indicates on-road objects, and the ego is shown in green.

on-road objects randomly positioned on lane centerlines within the ego vehicle's proximity. To avoid initial state collisions, we apply rejection sampling with spatial conflict checks against previously placed objects. A sampled location is rejected if the distance between the centers of two objects is within

$$d = \frac{\max(l_1, w_1)}{2} + \frac{\max(l_2, w_2)}{2} \tag{4.1}$$

where l_i, w_i are the length and width of object i. Off-road objects are distributed near road edges with minimal offset, replicating typical parked vehicle and pedestrian positioning. To maintain computational efficiency, the number of placement attempts per object is limited to 10. Unsuccessful placements result in static off-road positioning with a progressively increasing road-edge offset. Initial velocities default to ego speed for vehicles and 10 *mph* for cyclists. Figure 4.2 (left) illustrates a fully populated scene from the **RandomRuleMLP** initializer.

Path Generation for IDM

To ensure compatibility with the IDM motion model, it requires path information for lateral control. Since off-road objects are static and only on-road objects follow IDM dynamics, paths are generated exclusively for on-road agents. Following prior work [KGT+24, CDG24], fixed-length paths are sampled along lane centerlines from initial positions. For the ego vehicle, which follows the logged trajectory for the first

11 timesteps, we identify the closest point on a lane centerline after this interval. The subsequent path is constructed from this point onward along the centerline. Figure 4.2 (right) illustrates representative pathways.

Reverse Token Matching for SMART

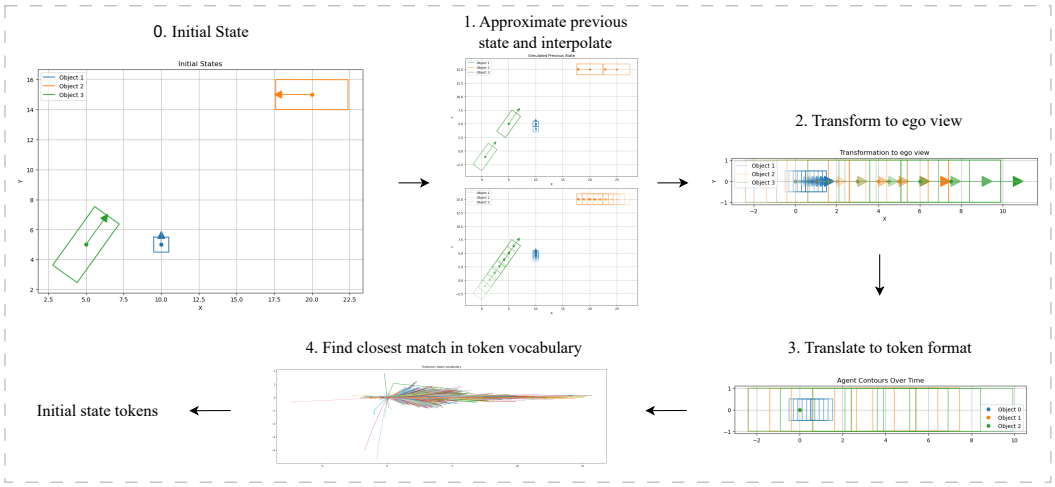


Figure 4.3.: **Reverse token matching strategy.** (1) We estimate a physically plausible "previous" state for each initialized object using a constant velocity and heading model. (2-3) The estimated state is then converted into the SMART token representation and (4) matched against all available trajectory tokens to select the closest one.

SMART is originally trained for traffic simulation tasks that assume at least one trajectory token from the motion history. To ensure compatibility with our random position initializer, we introduce a mechanism that aligns each initialized state with a corresponding "previous" trajectory token. Inspired by [WJZ+25], we apply a reverse token matching strategy, illustrated in Figure 4.3. Given the initial states, we estimate a preceding state using a constant velocity and heading model. These states are then transformed to ego-centric coordinates and translated into SMART token representation (contours of the bounding boxes). In the end, we match this token against all trajectory tokens in SMART's vocabulary to identify the closest match. Subsequent trajectory generation proceeds autoregressively from the selected initial token conditioned on scenario context and prior tokens.

During preliminary analysis, we ablate several non-logged methods. This ablation study is described in Section 5.1.1 and guided the selection of methods for our experiments. Along with the logged methods, we therefore focus on **RandomRuleMLP_CV**, **RandomRuleMLP_IDM**, and **RandomRuleMLP_SMART**.

4.2. Experiments

Given the additional complexity and aggregation inherent in scenario generation, we first conduct all experiments on the traffic simulation task. This provides a more controlled and better-understood environment for evaluation. To systematically assess the influence of individual configuration factors, we perform ablation studies on the embedding, metrics, rollouts, and data size, varying a single experimental dimension at a time. Building on the insights gained from these experiments, we subsequently apply the fidelity and diversity metrics to the scenario generation setting in Section 4.2.5.

4.2.1. Embeddings and Distances

In the first experiment, we investigate different types of embeddings and distance metrics for approximating the support of the real and generated data distributions. Since no unbiased off-the-shelf encoder exists for complex multimodal traffic scenes, we employ Waymo's handcrafted features, which are described in Section 2.3.1. This feature space comprises kinematic, map-based, and interactive features, capturing essential driving behavior characteristics and critical safety indicators, such as collision and off-road indications.

Waymo characterizes trajectories via ten handcrafted per-timestep features, aggregated across time and rollouts. In our case, we aggregate only temporally and treat each rollout as an independent sample. The subsequent subsection details two approaches for aggregating these features.

Histogram-based Embeddings

The first variant aggregates the features with the same histograms defined by Waymo and concatenates all histogram bins into a high-dimensional feature vector. To reflect feature importance, we apply feature-wise weighting:

$$\texttt{histogramBins\_eucl\_weighted}(x_i, x_j) = \sqrt{\sum_{m=1}^{M} w^{(m)} \sum_{k=1}^{b^{(m)}} \frac{1}{b^{(m)}} (h_{i,k}^{(m)} - h_{j,k}^{(m)})^2} \tag{4.2}$$

where $w^{(m)}$ is the Waymo weight for feature m equally distributed across its $b^{(m)}$ bins, and $x_i = [h_i^{(1)} || h_i^{(2)} || ... || h_i^{(M)}] \in \mathbb{R}^d$ the vector of the concatenated histogram bins.

While compatible with the Euclidean assumptions of the fidelity and diversity metrics, this approach neglects the underlying structure of the categorical distributions. In particular, it treats all histogram bins as independent dimensions, even though adjacent bins within the same feature are semantically related and inherently ordered.

To address this limitation, we also experiment with the p-Wasserstein Distance (WD) as an alternative similarity measure between histogram-based features. WD is a distance metric that quantifies the dissimilarity between two probability distributions that are defined on a metric space [Kan39].

We formalize the distance between two trajectories in the categorical distribution space of their feature histograms as follows:

$$\texttt{histogram\_wasserstein\_weighted}(x_i, x_j) = \sum_{m=1}^{M} w^{(m)} W_1\left(h_i^{(m)}, h_j^{(m)}\right) \tag{4.3}$$

where W_1 denotes the 1-WD between the normalized histograms $h_i^{(m)}$ and $h_j^{(m)}$ of feature m. The overall trajectory distance is obtained by a weighted sum of the per-feature Wasserstein distances, using the feature weights specified by Waymo. This formulation penalizes larger shifts between distant bins, yielding more semantically meaningful trajectory similarity.

However, this distance definition departs from the Euclidean space assumption on which all fidelity and diversity metrics are based. Since the Wasserstein metric induces a different geometry that does not directly correspond to hyperspheres or Euclidean neighborhoods, the resulting fidelity and diversity values should be interpreted as heuristic measures rather than exact counterparts of the original definitions.

Simple Stats Embeddings

As an alternative to Waymo histograms, we extract summary statistics. Specifically, for each trajectory and for each of the M features, we extract the minimum, mean, and maximum values over the time domain. Concatenating and normalizing these values yields lightweight embeddings. We compute the distance between two trajectories i and j as:

$$\texttt{MeanMinMax\_eucl\_weighted}(x_i, x_j) = \sqrt{\sum_{m=1}^{M} w^{(m)} \sum_{k=1}^{3} \frac{1}{3} (x_{i,k}^{(m)} - x_{j,k}^{(m)})^2} \tag{4.4}$$

$$\texttt{MinMax\_eucl\_weighted}(x_i, x_j) = \sqrt{\sum_{m=1}^{M} w^{(m)} \sum_{k=1}^{2} \frac{1}{2} (x_{i,k}^{(m)} - x_{j,k}^{(m)})^2} \tag{4.5}$$

where $x_{i,k}^{(m)}$ is the k-th statistic (mean, min, or max) of feature m for trajectory i. These embeddings remain within Euclidean space, ensuring direct compatibility with standard fidelity and diversity metrics while preserving key behavioral information.

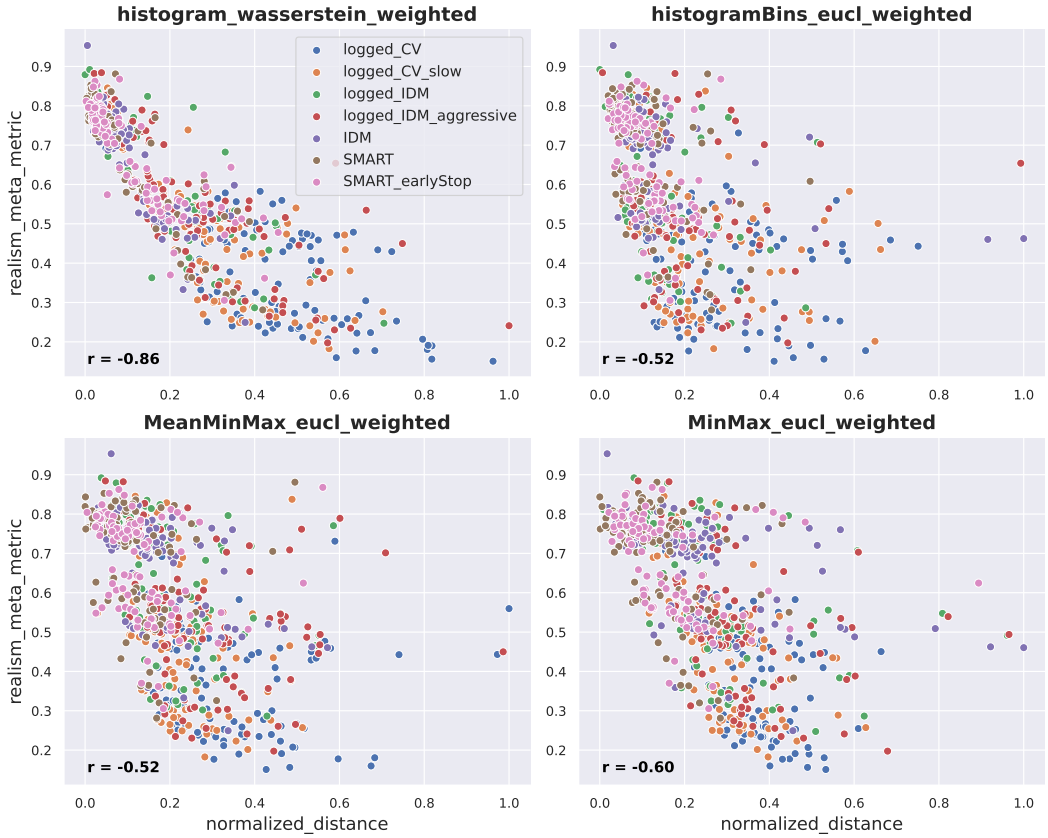


Figure 4.4.: **Correlation between normalized distance and the Waymo RMM score** ($n = 100$). We observe the most negative Pearson correlation r for the `histogram_wasserstein_weighted` embedding across methods, as it captures the underlying structure of the feature histograms most effectively.

Results

We first compare embedding types by correlating their computed distances with the Waymo RMM score, examining whether distance measures themselves constitute evaluation metrics. For 100 randomly sampled scenarios per method, we compute the average distance between generated and ground-truth trajectories at the scenario level and normalize it to $[0, 1]$.

As Figure 4.4 shows, `histogram_wasserstein_weighted` embedding exhibits the strongest negative correlation with the RMM score, indicating that higher realism corresponds to lower distances. The Euclidean variants `histogramBins_eucl_weighted` and `MeanMinMax_eucl_weighted` achieve similar correlations.

In contrast, the `MinMax_eucl_weighted` variant performs slightly better, likely be-

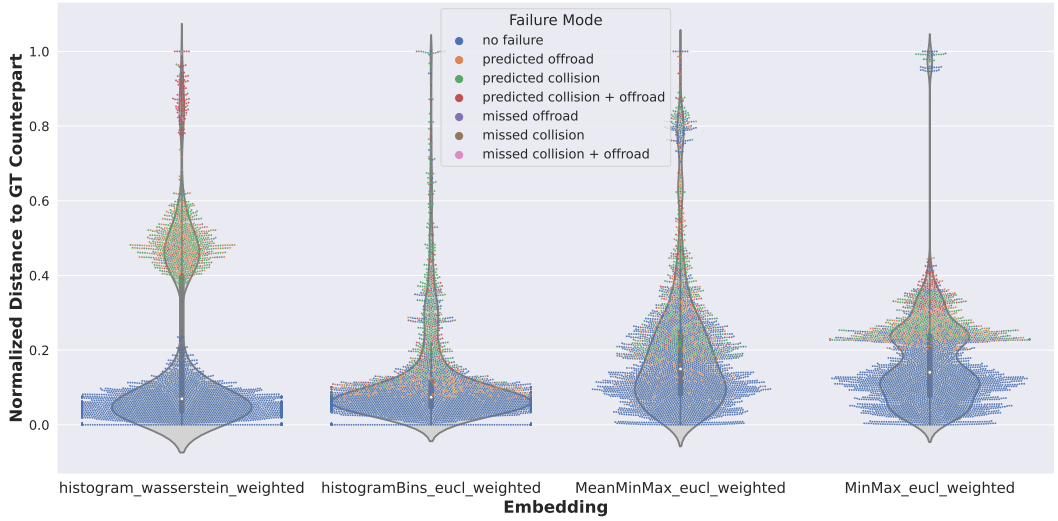


Figure 4.5.: **Distance distributions to corresponding ground-truth trajectories.** Generated trajectory samples are color-coded by mode-failure type. A clearer separation between these categories indicates a stronger embedding sensitivity to behavioral deviations.

cause it treats binary off-road and collision indicators as such rather than incorporating a "mean" dimension.

Figure 4.5 further analyzes distance distributions between generated and ground truth trajectories across all methods. We overlay samples, classified by the most relevant mode-failure types (off-road, collision, or both). For example, an orange point represents a generated trajectory that goes off-road while its ground truth counterpart remains on-road. Across embeddings, failure-free trajectories exhibit the lowest distances, while mode failures display higher distances. However, separation clarity varies substantially across embedding types.

The `histogram_wasserstein_weighted` representation yields the clearest separation between non-failure (blue), failure in one dimension (orange and green), and failure in two dimensions (red). This indicates that the embedding is more sensitive to semantic changes in feature distributions. In contrast, the Euclidean-based embeddings show greater overlap. `MinMax_eucl_weighted` preserves separation better than `MeanMinMax_eucl_weighted`, which exhibits minimal category distinction, likely due to the additional "mean" dimension. Similarly, `histogramBins_eucl_weighted` struggles to distinguish between single- and two-dimensional failures and tends to assign smaller distances to off-road than to collision failures. A detailed per-method analysis is provided in the Appendix A.1.

Qualitative nearest-neighbor analysis using our visualization tool (Section 3.2) confirms that `histogram_wasserstein_weighted` produces the most semantically

meaningful neighborhoods.

Overall, these insights support the earlier findings that the Wasserstein embedding not only correlates most strongly with the Waymo realism metric but also provides the most discriminative distance distribution across mode failure types, reinforcing its value as a meaningful trajectory similarity measure. Among Euclidean-based embeddings, `MinMax_eucl_weighted` achieves the most meaningful results.

4.2.2. Fidelity and Diversity Metrics

In the second experiment, we investigate the behavior of the fidelity and diversity metrics introduced in Section 3.1.1. We randomly sample 200 scenarios from the validation split, each containing, on average, five trajectories selected for evaluation. We use the same set of trajectories to establish the support for both the generated and real trajectory distributions. To maintain data symmetry, we consider only one rollout per generated trajectory in this experiment.

The parameter k is chosen according to the author's recommendations: $k = 3$ for I-Prec/I-Rec [KKL+19] and $k = 4$ for P-Prec/P-Rec [PK23]. For coverage and density, k is derived using the expected coverage formula proposed by the authors, which accounts for the number of the generated and real samples [NOU+20], resulting in $k = 5$.

We first analyze the unconditional setting, which evaluates realism and diversity over the full data distribution. We then examine the conditional setting, which operates at the instance level, analogous to the Waymo metric.

Results

In Figure 4.6, the unconditional results are shown for all methods and embeddings. For I-Prec/I-Rec (left), the SMART variants attain the highest scores, indicating strong fidelity and diversity, whereas IDM variants lie in a lower but relatively balanced range. By contrast, the CV variants exhibit similar precision but very low recall. This suggests they primarily capture simple, straight-line behaviors, which are commonly present in the ground truth data, yet fail to capture the diversity of the real trajectory distribution. Moreover, the ground truth data includes outlier trajectories with collision and off-road indications arising from sensor and labeling noise. Consequently, colliding and off-road samples generated by the model can fall within the sparse neighborhoods of these outliers, artificially inflating the precision scores.

For density and coverage (middle), the trends remain similar, except IDM variants are more widely dispersed, and CV variants achieve lower fidelity scores. This aligns with the definition of density, which counts how many generated samples fall within each real sample's neighborhood and normalizes the count by k. This scoring rule

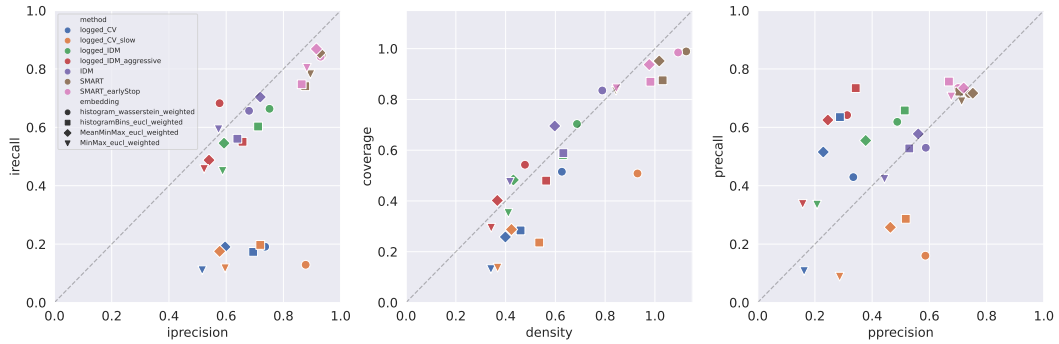


Figure 4.6.: **Results of the unconditional metric experiment.** Each panel visualizes a distinct metric pair, illustrating how the evaluated methods trade off realism and diversity when measured over the full trajectory distributions. We identify the `MinMax_eucl_weighted` embedding as yielding the most interpretable results across metric pairs. The dotted line represents a perfectly balanced score between realism and diversity.

accounts for local concentration by superposing the hyperspheres rather than taking their union. The reduced density of the CV variants suggests that the generated samples occupy sparse, outlier neighborhoods, failing to represent the core of the real distribution adequately. Additionally, we observe that SMART variants achieve density scores exceeding 1.0, a permissible outcome given the metrics' unbounded nature.

The probabilistic metrics (right) distribute the methods irregularly across embeddings. Among all embeddings, `MinMax_eucl_weighted` exhibits the highest sensitivity across methods and yields the most interpretable pattern: IDM variants show high diversity but low realism, CV variants exhibit both low diversity and low precision, and SMART variants achieve a high, well-balanced realism and diversity score. In general, the probabilistic metrics behave more conservatively, likely due to their normalized NN-hyperspheres sizes and probabilistic membership scoring.

In Figure 4.7, we report the conditional results as defined in Section 3.1.2. This setting aims to closely mirror the Waymo metric by checking whether each generated trajectory lies within the neighborhood of its exact ground truth counterpart. As expected in the single-rollout setting, all conditional metrics are noticeably lower than their unconditional parts. Falling in the counterpart's neighborhood is inherently more difficult, especially in dense areas where NN-hyperspheres are small. Improved and probabilistic variants yield relatively balanced fidelity and diversity, whereas conditional density remains low due to normalization by $k > 1$, making it ill-suited for single-sample evaluation.

Across methods, we observe consistently higher scores for histogram-based embeddings compared to our simple statistical embeddings. Moreover, we argue that

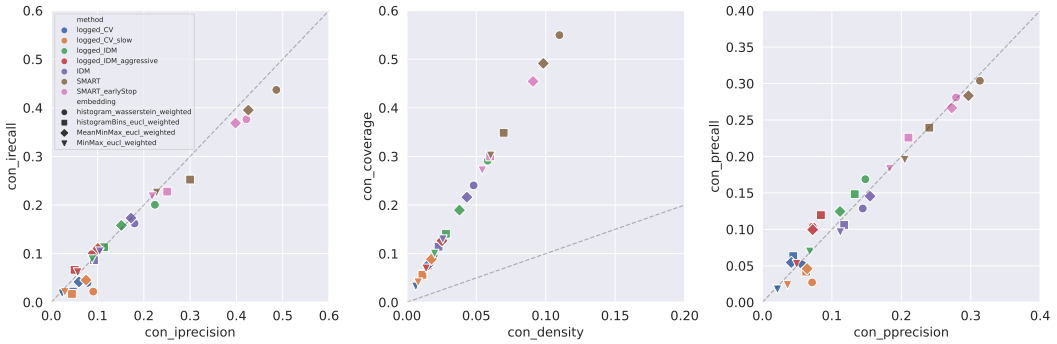


Figure 4.7.: **Results of the conditional metric experiment.** Each generated trajectory is assessed with respect to the neighborhood of its ground-truth counterpart and vice versa. All metrics show the inherently conservative nature of conditional metrics in the single-sample setting, making them less informative. The dotted line represents a perfectly balanced score between realism and diversity.

conditional metrics are informative when multiple reference samples per instance are available.

4.2.3. Number of Rollouts

As shown in the previous experiment, conditional metrics can be overly pessimistic when only a single rollout is considered. Therefore, we believe increasing the number of rollouts better probes a model's support and improves the reliability of instance-level fidelity and diversity estimates. However, aggregating all rollouts into a single histogram, as done in the RMM metric, blurs differences between individual rollouts and prevents detailed diagnosis.

Conversely, treating each rollout as an individual sample from the generated trajectory distribution can improve diagnostic power. However, this approach asymmetrically increases the number of generated samples, making a fixed k inappropriate for reliable support estimation of the generated data. We therefore consider $n \in \{1, 5, 16\}$ rollouts and explore different scalings of k, reusing the 200 scenarios from previous experiments. For clarity, we report results only for the `MinMax_eucl_weighted` embedding, with additional results deferred to the Appendix A.1.

For the improved and probabilistic metric variants, we define three strategies for scaling k when estimating the support of the generated data:

$$\text{fixK} = k \tag{4.6}$$

$$\text{nrolloutK} = \max\{k, n\} \tag{4.7}$$

$$\text{nxrolloutK} = n \cdot k \tag{4.8}$$

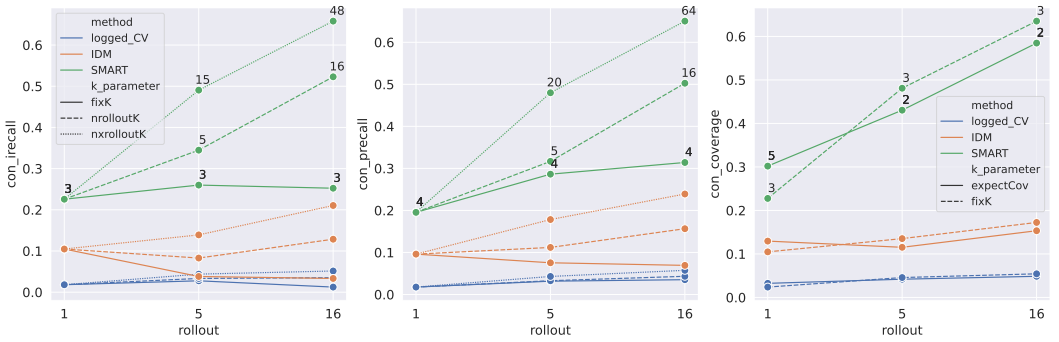


Figure 4.8.: **Results of the rollout experiment**. Scaling k to match the number of rollouts n yields the most stable and reliable diversity estimation across metrics. The figure presents results obtained using the `MinMax_eucl_weighted` embedding for three representative methods and highlights the value of k for each scaling strategy and n.

where n is the number of rollouts and k the default parameter specified by the original metric definition.

For coverage, the generated support is not approximated, and k is not explicitly scaled. Instead, following the original formulation, k is derived from the expected coverage under identical distributions, which depends on the number of data samples [NOU+20]. This results in a $k = 5$ for a single rollout, and $k = 2$ for 5 and 16 rollouts, reflecting faster coverage saturation with more generated samples.

Because rollout count primarily affects diversity metrics, we only report diversity results for three representative methods in Figure 4.8. Overall, diversity estimates become more reliable as the number of rollouts increases. For conditional irecall (left) and precall (middle), all methods exhibit a consistent increase when moving from a single to multiple rollouts with appropriately scaled k, indicating that larger rollout sets capture a broader portion of the model's support. The effect is most pronounced for the IDM and SMART models under the *nxrolloutK* scaling. This suggests that scaling k proportionally to the number of rollouts increases the likelihood that a real sample lies within at least one neighborhood of its corresponding generated trajectories.

In contrast, the *fixK* strategy tends to underestimate diversity at higher rollout counts, as it ignores the asymmetric increase in the number of generated samples. As n grows, the generated distribution becomes denser, causing the neighborhoods to shrink and making them harder to hit. As a result, diversity scores decline, contrary to the expected increase in diversity with more varied rollouts.

The *nrolloutK* strategy offers a compromise, yielding moderate improvements without overestimating. For weaker methods, scaling k with n has a limited impact, which is desirable, as enlarging neighborhoods for poorly distributed rollouts could otherwise

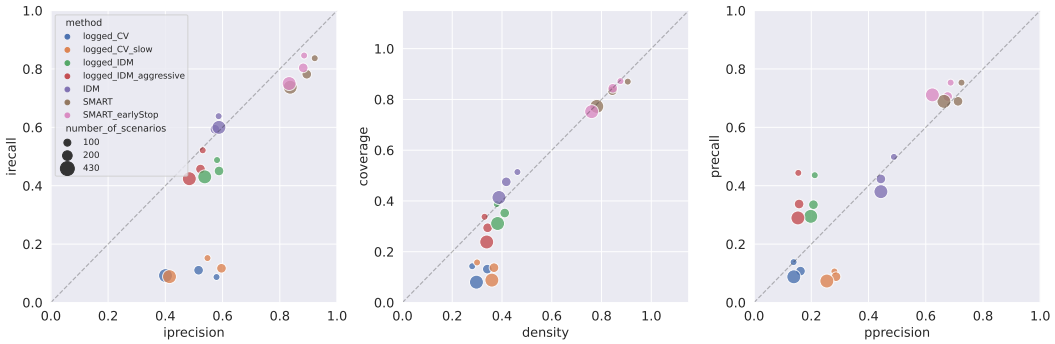


Figure 4.9.: **Results of the datasize experiment**. All metrics are affected by the number of data samples, but the overall effect remains moderate. Notably, weaker methods exhibit a larger spread than stronger ones. We report unconditional metric results for the `MinMax_eucl_weighted` embedding with one sample per trajectory.

artificially increase the likelihood of covering real data points.

For conditional coverage (right), differences between the fixed k and the varying k based on the expected coverage formula are less pronounced. At higher rollout counts, the fix k tends to yield slightly higher scores, since the expected coverage approach selects a smaller k as coverage saturates more quickly.

In summary, increasing the number of rollouts improves conditional diversity estimation by mitigating the strong dependence on a single rollout. Among the evaluated scaling strategies, *nrolloutK* provides the most stable and interpretable diversity estimates across methods. When using coverage, we suggest using the expected coverage approach to determine k. For a more comprehensive view, see Appendix A.1.

4.2.4. Datasize

In this final traffic simulation experiment, we examine how dataset size affects evaluation. As noted by [RvBvdS25], all fidelity and diversity metrics are affected by the number of samples used to estimate the underlying support of the data. Consequently, the dataset size should be investigated in every evaluation to avoid misleading conclusions.

To isolate this effect, we fix the setting to one rollout per trajectory and vary the number of scenarios. Specifically, we compare our baseline, consisting of 200 scenarios, with smaller (100) and larger (430) scenario sets. Note that each scenario contains, on average, 5 trajectories for evaluation, thereby scaling the effective sample size by a factor of ~ 5. We use the default k values for all metrics and report results for the `MinMax_eucl_weighted` embedding in Figure 4.9.

Reducing data size consistently inflates metric scores in most cases. This is expected given how the metrics are defined with their k nearest neighborhoods. Sparser data increases kNN radii, enlarging hyperspheres and making overlaps between real and generated samples more likely. However, the effect is moderate, typically within ~ 0.1 absolute metric points.

We observe similar patterns for the remaining embedding types (see Appendix A.1). Across embeddings, weaker methods (e.g., CV variants) exhibit a larger spread under varying data sizes compared to stronger methods such as SMART. This indicates that methods prone to distribution shifts and outliers are more sensitive to the neighborhood inflation effects induced by sparser data.

To summarize the empirical findings from the preceding experiments, we distill our observations into the following guideline checklist:

1. Do you require a distance metric to evaluate trajectories in feature space?
    - Yes: Use histogram embedding combined with a Wasserstein distance
2. Do you want to evaluate using fidelity and diversity metrics?
    - Yes: Use simple Euclidean-based embeddings such as `MinMax_eucl_weighted`
3. Is robustness to outliers important?
    - Yes: Avoid I-Prec and I-Rec
4. Do you want to evaluate multiple rollouts individually?
    - Yes: Use conditional diversity metrics with appropriately scaled k
5. Do you require non-binary instance-level scores?
    - Yes: Use P-Prec and P-Rec

4.2.5. Application to Scenario Generation

We extend the proposed evaluation framework to the setting of scenario generation. Scenario generation substantially increases the complexity of both the generation task itself and its evaluation, as the assessment must consider the entire scenario as a coherent whole, rather than one-to-one trajectories. Additionally, the absence of initial states significantly broadens the range of plausible scenarios, further complicating the evaluation process.

Consequently, applying existing fidelity and diversity metrics requires careful adaptation in terms of data representation and aggregation. We outline two approaches for this adaptation.

Scenario-based

Inspired by the Waymo challenge, this approach aggregates features over time and all scene objects, producing a single embedding per scenario rollout. The underlying data distribution thus represents entire scenarios rather than individual trajectories. Here, the `MinMax` embedding is less informative, as it lacks average scene-level statistics, especially for critical collision and off-road rates.

For clarity, we report only probabilistic precision and recall, as they demonstrated stable behavior in prior traffic generation experiments. Using 200 scenarios, results for the `MeanMinMax_eucl_weighted` embedding are shown in Figure 4.10.

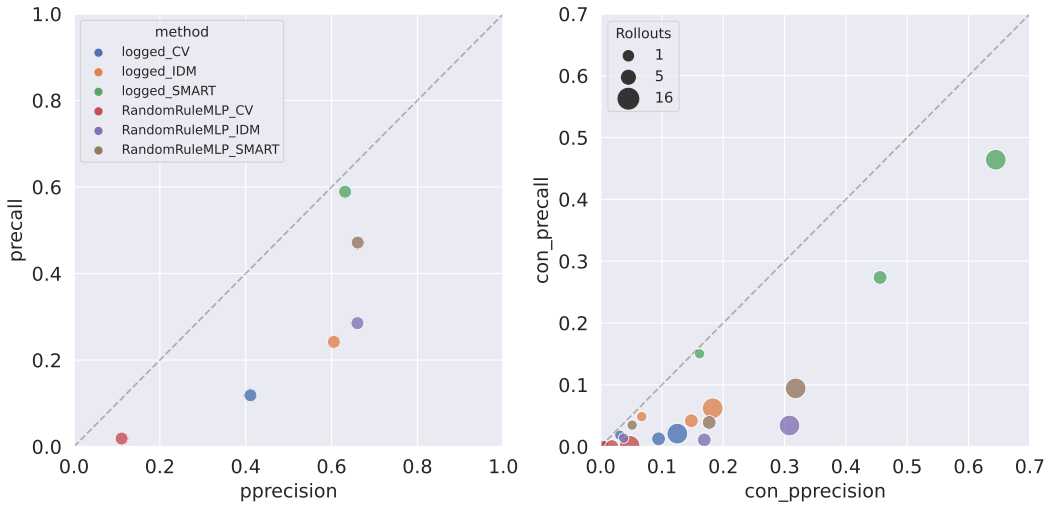


Figure 4.10.: **Evaluation results using aggregated scenario embeddings.** Probabilistic precision and recall for unconditional (left) and conditional (right) metrics using the `MeanMinMax_eucl_weighted` embedding. Low recall across methods indicates insufficient diversity capture, while single ground truth conditioning limits the reliability of conditional precision. Logged methods mostly outperform random initialization.

Unconditional metrics (left) show that logged initialization outperforms random initialization for CV and SMART methods, while IDM exhibits the opposite trend. Across methods, recall scores remain lower than precision, indicating that methods generate scenarios that resemble average ground truth characteristics but fail to capture distributional diversity. Only logged_SMART achieves a balanced trade-off.

Conditional metrics (right) provide limited insight due to a single ground-truth sample, yielding uninformative precision. Furthermore, we observe low recall sensitivity across methods, with SMART being the only method to achieve a substantially high score.

While this approach provides a holistic view of the scene, it also exhibits conceptual

limitations. Specifically, averaging all objects into simple statistics loses detailed characteristics of individual trajectories. Consequently, different combinations of trajectories can produce similar aggregated statistics without necessarily corresponding to realistic or diverse agent behaviors within the scenario.

Trajectory-based

To address the limitations introduced by averaging over objects, we propose an alternative approach that considers all trajectories within each scenario individually. In this formulation, a scenario is represented as the set of all trajectories in the scene. Conceptually, this aligns with the traffic generation setting, but it considers all trajectories rather than a predefined subset. Each scenario contains, on average, 40-50 trajectories.

In the conditional setting, fidelity and diversity metrics assess whether generated trajectories correspond to ground truth trajectories from the counterpart scenario. This yields multiple ground truth references per conditioning instance, enabling more meaningful conditional fidelity metrics. Each generated trajectory is evaluated based on whether it lies within the neighborhood of at least one real trajectory from the corresponding ground truth set. This set-based comparison addresses the limitation of one-to-one mapping in scenario generation.

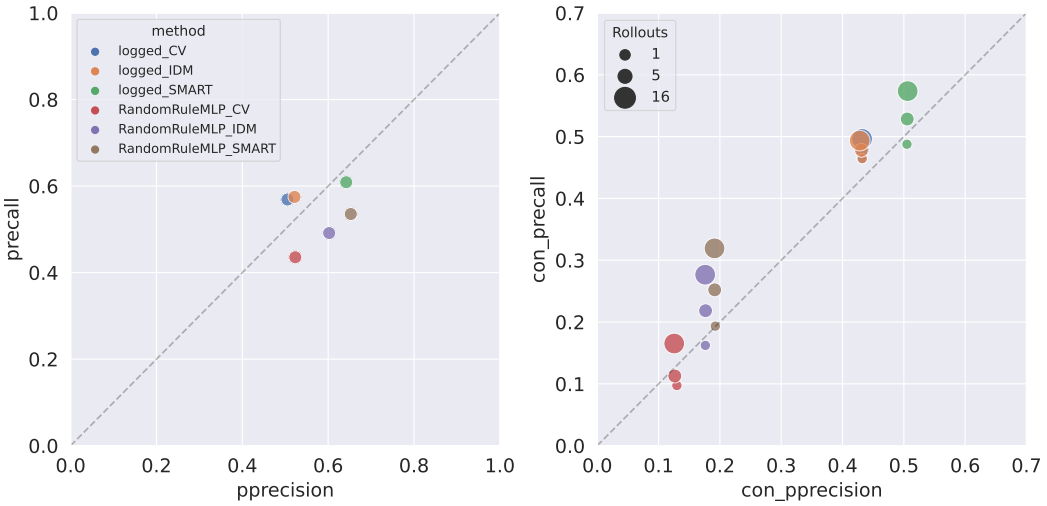


Figure 4.11.: **Evaluation results using all trajectories.** The unconditional metrics (left) exhibit limited sensitivity due to heterogeneous trajectories with numerous trivial matches (e.g., static background vehicles). The conditional setting (right) clearly distinguishes between random and logged initialization, emphasizing initial state prediction. Precision remains stable while recall improves with increased rollouts per trajectory set. Results are reported for the `MeanMinMax_eucl_weighted` embedding.

We evaluate the same 200 scenarios using the `MeanMinMax_eucl_weighted` embedding to be comparable with the previous approach. Figure 4.11 presents both unconditional (left) and conditional (right) probabilistic, while more details can be found in Appendix A.1.

Both unconditional recall and precision exhibit low sensitivity. This limitation is likely attributable to the composition of the ground truth distribution, which encompasses a broad spectrum of trajectories, including undesirable characteristics such as collisions, off-road driving, and parked vehicles. Consequently, poorly performing models attain artificially favorable metric scores. The realism of scenarios is mainly driven by a limited set of trajectories, while static background objects (e.g., parked vehicles) facilitate trivial matching and dominate the overall scenes. As a result, the unconditional approach lacks sufficient discrimination capacity for quality assessment, since evaluation is driven by many trivial rather than a few critical trajectory characteristics. This limitation is further compounded by the simplistic and imperfect embedding, which may not sufficiently capture variations in trajectory quality.

Conditional metrics demonstrate greater sensitivity across methods due to scenario-level conditioning. Logged methods clearly separate from randomly initialized approaches, emphasizing initialization quality over motion modeling. Notably, logged_CV achieves scores comparable to logged_IDM, indicating insufficient discrimination when evaluating trajectories with undesirable characteristics against diverse (ground truth) sets. We believe this effect is further exacerbated by the numerous off-road-parked vehicles in many scenarios, which constitute a trivial modeling task and can achieve high similarity across logged methods.

Building on these observations, our practical guidelines can be extended with the following recommendations specific to scenario generation:

6. For scenario generation, do you want to evaluate aggregated scenarios?
    - Yes: Use unconditional metrics with scenario-based embeddings (e.g., `MeanMinMax_eucl_weighted`).
    - Limitation: Object-level aggregation obscures trajectory-level distinctions
7. Do you require conditional fidelity and diversity metrics for scenario generation?
    - Yes: Use the trajectory-based approach
    - Limitation: Using all trajectories of the scenes includes numerous non-critical but simple instances (e.g., static background vehicles), thereby introducing noise and diluting the evaluation

4.3. Results

This section presents a comprehensive analysis of both quantitative metrics and representative qualitative results. Specifically, we evaluate traffic and scenario generation following the practical guidance delineated in our proposed checklist, thereby demonstrating its applicability and utility.

4.3.1. Quantitative Results

Traffic Simulation

Table 4.2 summarizes the quantitative results for the traffic simulation task, reporting both Waymo's Realism Meta Metric (RMM) and our proposed set of probabilistic fidelity and diversity metrics. Assuming RMM serves as a meaningful baseline, we compute Kendall's τ rank correlation between RMM and our metrics. With $\tau = 0.714$, the unconditional precision and recall metrics rank the evaluated methods comparably to the RMM.

Method	RMM-1	P-Prec	P-Rec	con-P-Rec
logged_CV	0.3714 (7)	0.1622 (6)	0.1075 (6)	0.0433 (7)
logged_CV_slow	0.4663 (6)	0.2864 (4)	0.0884 (7)	0.0486 (6)
logged_IDM	0.6316 (4)	0.2074 (5)	0.3350 (5)	0.0984 (4)
logged_IDM_aggressive	0.5505 (5)	0.1573 (7)	0.3373 (4)	0.0768 (5)
IDM	0.6694 (3)	0.4432 (3)	0.4232 (3)	0.1565 (3)
SMART	**0.7060** (1)	**0.7128** (1)	0.6891 (2)	0.5025 (2)
SMART_earlyStop	0.6826 (2)	0.6765 (2)	**0.7056** (1)	**0.5042** (1)
τ rank correlation		0.7143	0.7143	**0.9048**

Table 4.2.: **Quantitative results for traffic simulation.** We report the results using the `MinMax_eucl_weighted` configuration across 200 scenarios. RMM-1 denotes the RMM score computed with a single rollout (RMM-16 yields the same ranking and is therefore omitted). A high rank correlation is observed between con-P-Rec and RMM, despite the increasing sensitivity across methods.

Several interesting patterns emerge. First, *SMART* achieves the highest precision, whereas *SMART_earlyStop* attains the highest recall. This suggests that SMART produces more precise trajectories but may overfit in training, thereby failing to cover the full diversity of real trajectories in testing. Stopping the training earlier appears to improve generalization and recall, but at the cost of precision.

Second, reducing velocity in constant-velocity models improves fidelity but not diversity by mitigating collisions and off-road deviations in simple straight-line following scenarios. A similar pattern is observable for *logged_IDM* and *logged_IDM_aggressive*.

Third, switching from CV to IDM in the logged cases increases recall but not precision. While RMM provides limited diagnostic insight, our metrics show that reactive models enhance trajectory diversity by enabling turning, acceleration, and braking. Yet IDM variants remain inferior to learning-based approaches in capturing complex real-world driving behavior.

It is important to note that both evaluation approaches aggregate information differently. The RMM score operates at the instance level, computing a likelihood per trajectory and averaging across scenarios and datasets. The unconditional fidelity and diversity metrics operate at the dataset level, treating all trajectories as samples from one real and one generated distribution. Therefore, any comparison between the two has limitations and should be interpreted with caution.

Since we interpret RMM as an instance-level recall-like measure, conditional recall provides the most appropriate comparison. With the highest rank correlation for this configuration, conditional recall shows strong sensitivity between learning-based and rule-based methods but weaker sensitivity within each group.

Scenario Generation

		Trajectory-based			Scenario-based		
Method	**RMM-16**	**con-P-Prec**	**con-P-Rec**	**RMM-1**	**P-Prec**	**P-Rec**	**con-P-Rec**
logged_CV	0.6097 (5)	0.4318 (2)	0.4963 (2)	0.6035 (5)	0.4110 (5)	0.1186 (5)	0.0210 (5)
logged_IDM	0.6668 (2)	0.4280 (3)	0.4935 (3)	**0.6668** (1)	0.6053 (4)	0.2420 (4)	0.0621 (3)
logged_SMART	**0.6722** (1)	**0.5062** (1)	**0.5736** (1)	0.6639 (2)	0.6315 (3)	**0.5890** (1)	**0.4641** (1)
RandomRuleMLP_CV	0.5397 (6)	0.1256 (6)	0.1653 (6)	0.5348 (6)	0.1109 (6)	0.0186 (6)	0.0013 (6)
RandomRuleMLP_IDM	0.6353 (3)	0.1760 (5)	0.2766 (5)	0.6287 (3)	0.6603 (2)	0.2853 (3)	0.0342 (4)
RandomRuleMLP_SMART	0.6342 (4)	0.1913 (4)	0.3192 (4)	0.6227 (4)	**0.6609** (1)	0.4716 (2)	0.0942 (2)
τ rank correlation		0.4667	0.4667		0.2	0.4667	**0.7333**

Table 4.3.: **Quantitative results for scenario generation.** Rank correlations are generally low across both approaches, with the exception of conditional recall under scenario-based embeddings. Compared to RMM, the trajectory-based metrics exhibit stronger emphasis on initial state prediction relative to motion quality. The observed weak correlations and ambiguous results suggest that trajectory and scenario-based approaches alone are insufficient for comprehensive scenario generation evaluation.

In the context of scenario generation, the trajectory-based and scenario-based metrics do not exhibit the strong correlations observed in the traffic simulation setting. This suggests that scenario generation is inherently more challenging, and fidelity and diversity metrics alone are insufficient for reliable evaluation. Consequently, establishing comprehensive and robust evaluation methodologies for scenario generation remains an open research challenge.

Experiments are conducted with the same 200 scenarios: one rollout in the unconditional case and 16 rollouts in the conditional case, with k appropriately scaled. We

report results in Table 4.3 for both approaches using the `MeanMinMax_eucl_weighted` embedding type.

RMM heavily depends on the correct off-road/on-road ratios, which benefit reactive methods (e.g., *logged_IDM*, *logged_SMART*) that preserve initial distributions. In contrast, *logged_CV* starts with the correct ratio but gradually drifts during rollout due to collisions and off-road deviations in any scenario that departs from simple straight lane following.

Both fidelity and diversity metric variants exhibit low rank correlation with the RMM score. A few potential reasons for this were already discussed in Section 4.2.5. However, all metrics consistently identify *RandomRuleMLP_CV* as the worst-performing method.

4.3.2. Qualitative Results

For qualitative results, we display four representative scenarios for both traffic simulation and scenario generation tasks in Figure 4.12 and Figure 4.13, respectively. The top row shows the ground truth scenarios. Objects labeled as off-road are shown in gray, while on-road objects appear in blue. The ego vehicle is depicted in green, and the predefined agents used for evaluation in traffic simulation are shown in yellow. Collisions between objects are highlighted in red. Furthermore, we visualize object motion through a dedicated trajectory line. A few method variants were omitted for presentation clarity, but can be found in Appendix A.2.

While CV variants often produce visually unrealistic behaviors (e.g., going off-road or causing collisions), the IDM and SMART variants yield trajectories that more closely match the ground truth. However, simple straight-line scenarios (e.g., Scenario 4) can be modeled reasonably well by CV. The IDM variants, though not perfectly capturing real-world dynamics, exhibit smoother, less noisy behavior due to their deterministic, rule-based formulation. SMART, trained on real-world data, exhibits greater variability while retaining the noise present in the training data. The performance degradation observed between *logged_IDM* and *IDM* can be examined through visual inspection of the first scenario in Figure 4.12. Whereas *IDM* follows randomly generated paths along centerlines, *logged_IDM* transitions to a constant velocity model once the end of the logged trajectory is reached. This results in driving off-road and unreactive driving behavior.

Figure 4.12.: **Qualitative results for traffic simulation.** Note: Method *logged_CV_slow* and *logged_IDM_aggressive* are moved to the Appendix A.2

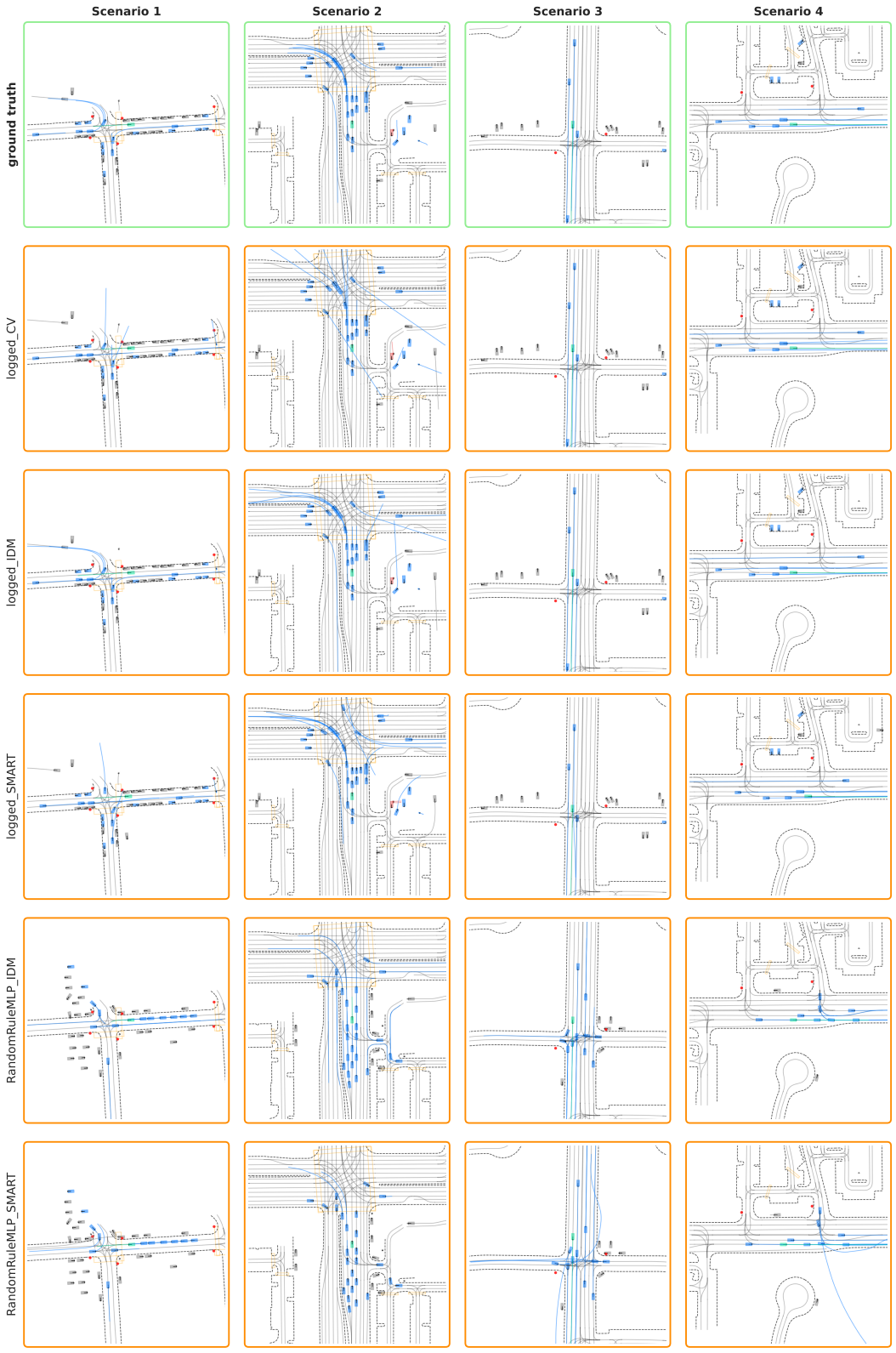


Figure 4.13.: **Qualitative results for scenario generation.** Note: The least promising method *RandomMLP_CV* is moved to the Appendix A.2

5. Discussion

This chapter provides a final discussion of the work presented in this thesis. We first outline the preliminary analysis conducted during our participation in the Waymo Open Scenario Generation Challenge (WOSGC), which motivated our subsequent studies on evaluation metrics. We then examine the key limitations and implications of this work by critically assessing both the Waymo benchmark and the proposed fidelity and diversity metrics. Furthermore, we discuss promising directions for future research and enumerate open questions. Finally, we conclude with a concise summary of our findings.

5.1. Waymo Open Scenario Generation Challenge

In this section, we describe our approach to the WOSGC, which served as a preliminary analysis of scenario generation methods conducted prior to our study of evaluation metrics. We subsequently report the official challenge results, showing that our lightweight rule-based system achieves competitive performance despite its simplicity.

5.1.1. Ablation Study

During our participation in the scenario generation challenge, we ablate several non-logged configurations of our methods on a subset of 430 samples from the validation set. For efficient evaluation, all experiments are conducted using a single rollout per scenario. However, we observe a slight but consistent improvement in the metrics when evaluating 32 rollouts. The corresponding results are summarized in Table 5.1.

As a baseline, we begin with the Constant Velocity (CV) model using a fixed initial distribution of 40% off-road and 60% on-road objects (A1). This configuration exhibits poor kinematic performance due to its simplistic motion model. Moreover, it leads to off-road trajectories and collisions. Introducing more realistic longitudinal control via the Intelligent Driver Model (IDM) (A2), aligned with lateral control by following lane centerlines, substantially improves kinematic scores. Further enhancing IDM with traffic light awareness leads to additional gains (A3), also in the map-based metrics. Next, we replace the fixed off-road ratio with a learned multi layer perceptron (MLP)-based regressor (B1) that dynamically predicts the off-road ratio based on

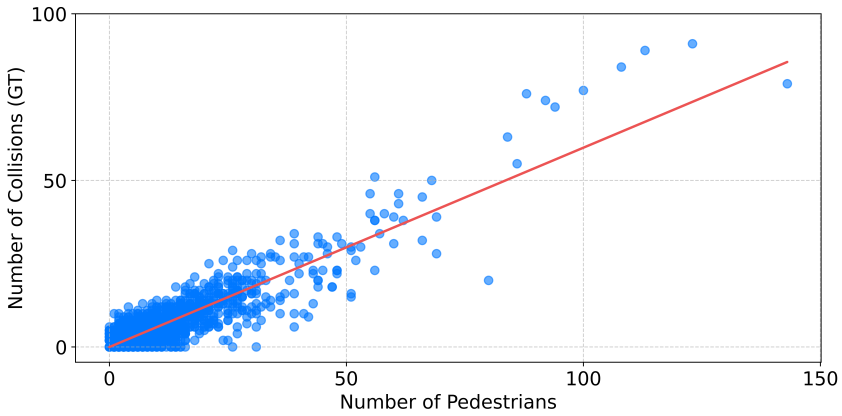


Figure 5.1.: **Ground truth pedestrian-collision correlation.** Scenarios with more pedestrians tend to exhibit more collisions, reflecting variability from sensor noise and labeling errors.

scene context. This modification significantly improves the map-based metrics by enabling adaptive initialization across diverse traffic scenes.

We also discover a strong correlation between the number of pedestrians and the number of collisions in the ground truth, as illustrated in Figure 5.1. This motivates the implementation of a simple forced collision mechanism (A4, B2, C1). Specifically, when the number of pedestrians exceeds four, we enforce collisions for every second pedestrian to reflect the likelihood of collisions seen in the real-world data.

Notably, although SMART achieves the most realistic kinematic score (C1), it does not outperform IDM-based configuration in our setup. This is likely due to

ID	Position	Lat. Con.	Long. Con.	Interactive	Map-based	Kinematic	META
A1	40/60	CV	CV	0.719	0.541	0.275	0.568
A2	40/60	Paths	IDM	0.705	0.530	0.484	0.599
A3	40/60	Paths	IDM w/ Lights	0.705	0.553	0.491	0.609
A4	40/60 w/ Col.	Paths	IDM w/ Lights	**0.723**	0.554	0.491	0.617
B1	MLP	Paths	IDM w/ Lights	0.705	**0.573**	0.490	0.616
B2	MLP w/ Col.	Paths	IDM w/ Lights	**0.723**	0.572	0.490	**0.624**
C1*	MLP w/ Col.	SMART	SMART	0.704	0.571	**0.510**	0.619

Table 5.1.: **Ablation study.** Each row represents a different configuration of our method, varying in initial position handling, lateral control (Lat. Con.) and longitudinal control (Lon. Con.). META shows the aggregated performance score as a weighted sum of the interactive, map-based, and kinematic metrics. * C1 was implemented and evaluated after the challenge deadline.

SMART's dependence on a meaningful trajectory history for accurate prediction. Our simplified approximation of the previous trajectory token fails to fully capture the necessary motion context. Furthermore, the rule-based positioning of objects by RandomRuleMLP results in out-of-distribution initial states that challenge the generalization capacity of learning-based models like SMART.

5.1.2. Results

The official results of WOSGC 2025 are summarized in Table 5.2. For our final submission we employed the *RandomRuleMLP_IDM* method with forced collisions and traffic light consideration (B2). Despite relying on a lightweight rule-based system with practical heuristics, we achieved 3rd place, outperforming both a 4.6M-parameter learning-based model and the Waymo baseline. It should be noted, however, that only four teams ultimately published their results on the leaderboard, which may reflect limited engagement with the task or the substantial computational resources required to participate.

Method Name	RMM ↑	Kinematic Metrics ↑	Interactive Metrics ↑	Map-based Metrics ↑	# of Parameters
SimFormer [WJZ+25]	0.6623	0.5416	0.7417	0.6293	7M
UniTSG [XGL+25]	0.6604	0.5415	0.7378	0.6288	7M
SHRED [ours]	**0.6185**	**0.4815**	**0.7197**	**0.5668**	<7100
infgen-full-large [PLZ25]	0.6030	0.5044	0.6774	0.5638	4.6M
Waymo-Baseline	0.4928	0.3299	0.5696	0.4872	1M

Table 5.2.: **Results.** Performance comparison on WOMD test split at the end of the challenge period. Our rule-based approach ranked third overall.

5.2. Limitations and Issues

This section outlines the main limitations we encountered in both the Waymo metric and our proposed evaluation framework. By highlighting these issues, we aim to clarify how current evaluation approaches fall short and where future improvements are needed.

5.2.1. Waymo Benchmark

Over the course of WOSGC and this work, several limitations of the Waymo Realism Meta Metric (RMM) became apparent. While RMM offers valuable feature-level diagnostics, it does not provide insight into the underlying causes. For instance, it can reveal that kinematic features closely follow the ground-truth distribution while interactive features such as collisions exhibit low likelihood, but it offers no indication of why these discrepancies occur. Consequently, the metric captures realism only along a single dimension, without distinguishing between different failure modes regarding fidelity and diversity.

Inconsistent upper bounds: Due to the structure of the evaluation framework, each trajectory has a distinct maximum attainable negative log-likelihood. The optimal score is achieved when the generated feature distribution matches the ground-truth feature distribution. However, trajectories whose ground-truth features are approximately uniform have inherently lower achievable likelihoods than those with concentrated, low-entropy distributions. As a result, simple trajectories (e.g., straight, constant-velocity motion) yield higher potential scores, whereas complex trajectories with more dispersed feature distributions yield lower ones. This introduces uneven weighting across trajectories and makes score interpretation difficult without knowing their respective upper bounds.

Labeling and sensor noise: A major limitation stems from labeling inaccuracies and sensor noise in the ground-truth data. As illustrated on the left side of Figure 5.2, the logged data contain spurious collisions and numerous parked vehicles labeled as on-road. Because the metric assumes the logged trajectories represent ideal behavior, it rewards models for reproducing these artifacts. Consequently, agents that avoid unsafe or implausible behavior may receive lower likelihoods, while agents that replicate erroneous logged patterns may obtain higher scores. Similar issues have been reported in recent works [CRJ+25, WWY+25]. Overall, the metric primarily measures conformity to noisy data rather than safety, human likeness, or the validity of alternative plausible behaviors. In fact, we were able to exploit this flaw in the scenario generation challenge by intentionally inducing collisions in scenes with many pedestrians.

Handcrafted feature space: The metric relies on handcrafted, computationally intensive features and pre-defined feature weights, with a strong emphasis on collision and off-road indicators. As mentioned, these features are particularly affected by noise in the logged data. Extending the trajectory-level formulation to scenario-level evaluation by averaging features over all objects implicitly assumes that a scenario can be characterized using the same feature space as an individual trajectory. This aggregation leads to substantial information loss and encourages models to reproduce noisy off-road and collision ratios rather than evaluating the realism of individual agent behaviors.

In general, we believe that instance-level distribution matching becomes problematic when only a single ground truth sample is available. In practice, scenes at the same location can vary considerably depending on contextual factors such as day of week or time of day. Metrics that evaluate against a single logged scene, therefore, cannot accommodate the natural diversity of plausible scenes that may occur at that location.

Quick Fixes

To partially mitigate the effects of labeling and sensor noise, we propose two straightforward adjustments to the metric and dataset preprocessing.

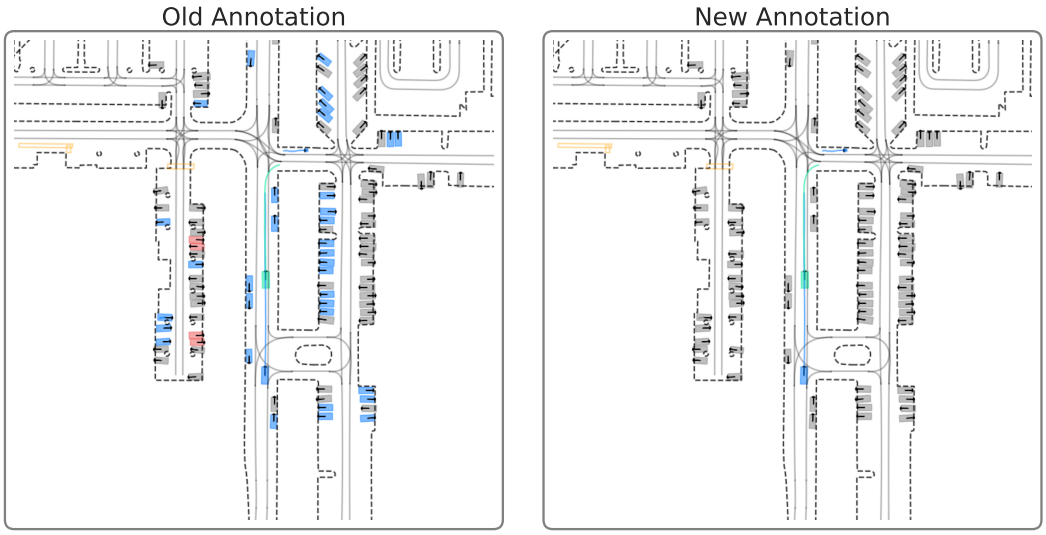


Figure 5.2.: **Comparison of labeling strategies.** The old off-road indication (left), based solely on road-edge information, produces inconsistent and noisy labels for stationary parked cars. Our new off-road indication (right) incorporates vehicle motion and a small margin to the road edges, resulting in more reliable and consistent labeling. Colliding objects (red) are removed from the scenario. Gray denotes off-road objects, blue indicates on-road objects, and green represents the ego vehicle.

First, to address the off-road labeling issue, we propose a more semantically meaningful definition of off-road indication by considering not only the road edge geometry but also the agent's position over time. Specifically, if a vehicle remains static and sufficiently close to the road edge, we classify it as off-road. This refinement enables a more consistent off-road classification for parked vehicles, both at the roadside and within parking areas, which illustrates the right part of Figure 5.2. To avoid incorrectly labeling vehicles that are temporarily stopped (e.g., at a traffic light), the margin around the road edge is deliberately kept small.

Second, to handle ground-truth collisions arising from sensor or labeling noise, we recommend either discarding scenarios that contain such collisions or removing the agents involved as illustrated in Figure 5.2 (right). This prevents learning-based models from reproducing noisy or erroneous behaviors and encourages the generation of more realistic and safety-aligned scenarios. Importantly, these adjustments must also be applied to the test evaluation sets. Otherwise, models would still be implicitly rewarded for replicating noisy labels, undermining the intended improvements.

5.2.2. Fidelity and Diversity Metrics

Fidelity and diversity metrics were originally introduced to address the shortcomings of single-value performance measures by disentangling how realistic the generated data is from how well its variability matches the real distribution. While this decomposition is intuitively appealing, it does not fully resolve the fundamental challenges of evaluating generative models, especially when applied to complex domains such as traffic and scenario generation.

Lack of ground truth multimodality: Similar to the limitation of the Waymo metric, the reliance on a single ground-truth instance limits the utility of fidelity and diversity metrics. Using unconditional definitions allows comparison at the dataset level but overly generalizes and obscures scene-specific structure. Conversely, conditional adaptations become overly sensitive to individual trajectories, particularly in the case of conditional fidelity. Furthermore, the effects of introducing conditioning within the metric formulation remain largely unexplored, and it is unclear whether these variations yield meaningful results. In the context of scenario generation, both unconditional and conditional metrics often fail to align with qualitative assessments, producing scores that are difficult to interpret. This mismatch underscores the inherent challenges of evaluating complex generative tasks such as scenario generation.

Handcrafted embedding space: Since there is no ImageNet-like, method-agnostic encoder for autonomous driving scenarios, we rely on the handcrafted Waymo feature space in this work. This representation is computationally expensive and may not fully capture the complexity of real-world driving behaviors. Because fidelity and diversity metrics depend heavily on the quality and expressiveness of the underlying feature space, their reliability is inherently limited by these handcrafted features and by our chosen aggregation setting.

General flaws: Recent work highlights that all existing fidelity and diversity metrics exhibit fundamental shortcomings in one way or another [RvBvdS25]. While most of the metrics have clearly defined lower and upper bounds in theory, such bounds are often absent in practice, even in simple toy settings. Moreover, all metrics are affected by the dataset size, which we also observed in our experiments. Despite these limitations, the authors argue that imperfect metrics can still provide valuable insights, and "ideal" metrics may not even exist at all. Nevertheless, it is essential for practitioners to recognize these limitations and consider their impact on experimental conclusions.

Finally, introducing additional parameters, such as the choice of k, dataset size, and embedding space, increases the complexity of the evaluation pipeline and makes the resulting metric increasingly opaque, potentially even more than the Waymo metric it seeks to replace.

5.3. Further Research Directions

While recent work highlights the need for improved metrics in generative tasks more generally [RvBvdS25], our findings underscore that this need is also critical in the domain of traffic and scenario generation. Current evaluation frameworks remain limited in their ability to assess fidelity, diversity, and structural realism of generated scenes. Additionally, they still rely on handcrafted features and manually tuned parameters.

Building on our fidelity and diversity framework, future work could address the limitations of single ground truth references by clustering semantically similar trajectories or scenarios. Rather than conditioning on a single reference or none at all, conditioning can be performed on a representative set of reference samples. This approach would allow conditional fidelity metrics to account for multiple plausible outcomes and provide more meaningful assessments of the generated data. Furthermore, conditional fidelity and diversity metrics should be evaluated across other domains to assess their general applicability and to determine whether they provide meaningful insights beyond traffic and scenario generation.

An additional challenge for distributional evaluation methods is accurately estimating the underlying data support. Replacing handcrafted feature spaces with learning-based embeddings may alleviate some of these issues and provide a more meaningful representation. However, reliable support estimation remains difficult in high-dimensional, multimodal datasets such as those encountered in traffic and scenario generation. Advances in robust support estimation, as demonstrated in other domains [KJKY24], could significantly strengthen the foundations of these evaluation frameworks.

Finally, in this work, metric evaluation was conducted primarily using simple and clearly differentiated methods on a limited subset of the data. However, for an evaluation metric to be suitable for benchmarking, it must not only correctly rank substantially different approaches but also exhibit sufficient sensitivity when comparing methods with similar performance. Future studies should therefore include a broader set of closely matched methods to more rigorously assess the metric's discriminative capability. Additionally, evaluations should be performed on larger set of data to enable a more rigorous investigation of how varying dataset sizes affect the stability and reliability of the metrics.

5.4. Conclusion

This thesis presents a series of case studies on assessing traffic and scenario generation using fidelity and diversity metrics. To the best of our knowledge, this work constitutes the first application of these metrics to traffic and scenario generation. While the metrics provided valuable insights for traffic generation, their interpreta-

tion proved considerably more challenging for scenario generation, reflecting the additional complexity inherent in this task. More broadly, the study highlights a fundamental challenge: evaluating evaluation metrics is itself a non-trivial problem.

To support researchers in this area, we developed a visualization tool for fidelity and diversity assessments. Quantitative metrics alone are insufficient to guarantee the realism or usefulness of generated data. Our tool highlights the ongoing importance of qualitative inspection in traffic and scenario generation. We believe that visual analysis remains essential, as it helps identify failure modes that purely metric-based evaluations may overlook. Furthermore, the tool is broadly applicable to other generative domains that employ fidelity and diversity metrics.

We also identified several limitations in both the widely adopted Waymo metric and our proposed evaluation methodology, including unreliable upper bounds and dependence on handcrafted feature representations. These limitations, however, point toward valuable directions for future work. We argue that, within the domain of traffic and scenario generation, greater emphasis should be placed on developing robust, meaningful, and domain-aware evaluation metrics, rather than solely advancing generative models.

In conclusion, this thesis offers a novel perspective on the assessment of traffic and scenario generation tasks. The insights gained through our studies underscore the importance of developing more meaningful and comprehensive evaluation methodologies. We hope this work serves as a valuable contribution to ongoing research in this field. Moreover, we believe that the tool introduced in this thesis offers practical use for researchers. It not only supports the development of new evaluation metrics but also facilitates a deeper understanding of existing methods by enabling clear visualization of their failure modes.

Bibliography

[AVBSvdS22] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306. PMLR, 17–23 Jul 2022.

[BPF24] Eyal Betzalel, Coby Penso, and Ethan Fetaya. Evaluation metrics for generative models: An empirical study. *Machine Learning and Knowledge Extraction*, 6(3):1531–1544, 2024.

[BS18] Shane Barratt and Rishi Sharma. A note on the inception score, 2018.

[BSAG21] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021.

[BYS+21] Luca Bergamini, Yawei Ye, Oliver Scheel, Long Chen, Chih Hu, Luca Del Pero, Blazej Osinski, Hugo Grimmett, and Peter Ondruska. Simnet: Learning reactive self-driving simulations from real-world observations, 2021.

[CDG24] Kashyap Chitta, Daniel Dauner, and Andreas Geiger. Sledge: Synthesizing driving environments with generative models and rule-based traffic, 2024.

[CEL+25] Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon. Kad: No more fad! an effective and efficient evaluation metric for audio generation, 2025.

[CF20] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them, 2020.

[CRJ+25] Wei-Jer Chang, Akshay Rangesh, Kevin Joseph, Matthew Strong, Masayoshi Tomizuka, Yihan Hu, and Wei Zhan. Spacer: Self-play anchoring with centralized reference models, 2025.

[CU23] Fasil Cheema and Ruth Urner. Precision recall cover: A method for assessing generative models. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International*

Conference on Artificial Intelligence and Statistics, volume 206 of *Proceedings of Machine Learning Research*, pages 6571–6594. PMLR, 25–27 Apr 2023.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[DRC+17] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator, 2017.

[DXA+23] Wenhao Ding, Chejian Xu, Mansur Arief, Haohong Lin, Bo Li, and Ding Zhao. A survey on safety-critical driving scenario generation – a methodological perspective, 2023.

[ECA+23] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023.

[ECC+21] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Y ang, Aur'elien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719, October 2021.

[FLP+23] Lan Feng, Quanyi Li, Zhenghao Peng, Shuhan Tan, and Bolei Zhou. Trafficgen: Learning to generate diverse and realistic traffic scenarios. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3567–3575. IEEE, 2023.

[FSA25] Alexis Fox, Samarth Swarup, and Abhijin Adiga. A unifying information-theoretic perspective on evaluating generative models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:16630–16638, 04 2025.

[GFL+23] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John D. Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research, 2023.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David

Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3, 06 2014.

[GPZ+25] Yuan Gao, Mattia Piccinini, Yuchen Zhang, Dingrui Wang, Korbinian Moller, Roberto Brusnicki, Baha Zarrouki, Alessio Gambi, Jan Frederik Totz, Kai Storms, Steven Peters, Andrea Stocco, Bassam Alrifaee, Marco Pavone, and Johannes Betz. Foundation models in autonomous driving: A survey on scenario generation and scenario analysis, 2025.

[HEA+23] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods of information in medicine*, 62, 01 2023.

[HHF+22] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[HL08] Yi Hu and Philipos C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

[HOMC+25] Mikel Hernandez, Pablo Osorio Marulanda, Mikel Catalina, Lorea Loinaz, Gorka Epelde, and Naiara Aginako. Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees. *Frontiers in Digital Health*, 7:1576290, 04 2025.

[HRU+17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 12 2017.

[HZV+24] Zhiyu Huang, Zixu Zhang, Ameya Vaidya, Yuxiao Chen, Chen Lv, and Jaime Fernández Fisac. Versatile behavior diffusion for generalized traffic agent simulation, 2024.

[IP19] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs, 2019.

[IvDS25] Takumi Ito, Kees van Deemter, and Jun Suzuki. Reference-free evaluation metrics for text generation: A survey, 2025.

[JBC+24] Chiyu Max Jiang, Yijing Bai, Andre Cornman, Christopher Davis, Xiukun Huang, Hong Jeon, Sakshum Kulshrestha, John Lambert, Shuangyu Li, Xuanyu Zhou, Carlos Fuertes, Chang Yuan, Mingxing

Tan, Yin Zhou, and Dragomir Anguelov. Scenediffuser: Efficient and controllable driving simulation initialization and rollout, 2024.

[KA23] Mahyar Khayatkhoei and Wael Abdalmageed. Emergent asymmetry of precision and recall for measuring fidelity and diversity of generative models in high dimensions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 16326–16343. PMLR, 2023.

[Kan39] L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6(4):363–422, 1939.

[KGT+24] Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, and Holger Caesar. Towards learning-based planning:the nuplan benchmark for real-world autonomous driving, 2024.

[KJKY24] Pum Jun Kim, Yoojin Jang, Jisu Kim, and Jaejun Yoo. Toppr: Robust support estimation approach for evaluating fidelity and diversity in generative models, 2024.

[KKL+19] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *33rd Conference on Neural Information Processing Systems*, Advances in Neural Information Processing Systems, United States, 2019. Neural Information Processing Systems Foundation.

[KW14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

[KZRS19] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019.

[LBBW+18] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wiessner. Microscopic traffic simulation using sumo. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2575–2582, 2018.

[Lin04] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[LL22] Junghyuk Lee and Jong-Seok Lee. Trend: Truncated generalized normal density estimation of inception embeddings for gan evaluation, 2022.

[LPF+23] Quanyi Li, Zhenghao Peng, Lan Feng, Zhizheng Liu, Chenda Duan, Wenjie Mo, and Bolei Zhou. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling, 2023.

[LWZ+24] Jack Lu, Kelvin Wong, Chris Zhang, Simon Suo, and Raquel Urtasun. Scenecontrol: Diffusion for controllable traffic scene generation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[MHM20] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[MLM+23] Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nick Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, Brandyn White, and Dragomir Anguelov. The waymo open sim agents challenge, 2023.

[MML+24] Reza Mahjourian, Rongbing Mu, Valerii Likhosherstov, Paul Mougin, Xiukun Huang, Joao Messias, and Shimon Whiteson. Unigen: Unified modeling of initial agent states and trajectories for generating autonomous driving scenarios, 2024.

[NARZ+22] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S. Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple efficient attention networks, 2022.

[NOU+20] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models, 2020.

[PGH+23] Ethan Pronovost, Meghana Reddy Ganesina, Noureldin Hendy, Zeyu Wang, Andres Morales, Kai Wang, and Nicholas Roy. Scenario diffusion: Controllable driving scenario generation with diffusion, 2023.

[PK23] Dogyun Park and Suhyun Kim. Probabilistic precision and recall towards reliable evaluation of generative models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 20042–20052. IEEE, October 2023.

[PLZ25] Zhenghao Peng, Yuxin Liu, and Bolei Zhou. Infgen: Scenario generation as next token group prediction, 2025.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. Bleu: a method for automatic evaluation of machine translation. 10 2002.

[PSS25] Muleilan Pei, Shaoshuai Shi, and Shaojie Shen. Advancing multi-agent traffic simulation via r1-style reinforcement fine-tuning, 2025.

[PWR23] Ethan Pronovost, Kai Wang, and Nick Roy. Generating driving scenes with diffusion, 2023.

[RBB+20] Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, and Gorka Epelde. Reliability of supervised machine learning using synthetic data in healthcare: A model to preserve privacy for data sharing (preprint). *JMIR Medical Informatics*, 8, 03 2020.

[RGG+25] Luke Rowe, Roger Girgis, Anthony Gosselin, Liam Paull, Christopher Pal, and Felix Heide. Scenario dreamer: Vectorized latent diffusion for generating driving simulation environments. In *CVPR*, 2025.

[RSGA25] Lyle Regenwetter, Akash Srivastava, Dan Gutfreund, and Faez Ahmed. Beyond statistical similarity: Rethinking metrics for deep generative models in engineering design, 2025.

[RvBvdS25] Ossi Räisä, Boris van Breugel, and Mihaela van der Schaar. Position: All current generative fidelity and diversity metrics are flawed, 2025.

[SBL+18] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall, 2018.

[SCH+23] George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models, 2023.

[SGS+24] Shuo Sun, Zekai Gu, Tianchen Sun, Jiawei Sun, Chengran Yuan, Yuhang Han, Dongen Li, and Marcelo H. Ang Jr. Drivescenegen: Generating diverse and realistic driving scenarios from scratch, 2024.

[SGZ+16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[SMK20] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A survey of evaluation metrics used for nlg systems, 2020.

[SRCU21] Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors, 2021.

[STT25] Nicolas Salvy, Hugues Talbot, and Bertrand Thirion. Enhanced generative model evaluation with clipped density and coverage, 2025.

[SVI+15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2015.

[THH00] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested

traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2), 2000.

[TvdOB16] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models, 2016.

[TWW+21] Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Scenegen: Learning to generate realistic traffic scenes, 2021.

[UvSK+19] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *DGS@ICLR*, 2019.

[vdMH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[VHS+21] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S. Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, and Benjamin Sapp. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction, 2021.

[VLY+23] Eugene Vinitsky, Nathan Lichtlé, Xiaomeng Yang, Brandon Amos, and Jakob Foerster. Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world, 2023.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Way25] Waymo. Scenario generation 2025 waymo open dataset. `https://waymo.com/open/challenges/2025/scenario-generation/`, 2025.

[WFGK24] Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. Smart: Scalable multi-agent real-time motion generation via next-token prediction, 2024.

[WJZ+25] Sen Wang, Xu Jianrong, Xiaoyong Zhang, Fangqiao Hu, Zhijun Huang, JiaChen Luo, Kechen Zhu, Jiaxiang Zhu, Yong Zhou, and Zhenwu Chen. Simformer: 1st place in the waymo open scenario generation challenge 2025, 2025.

[WLAA25] Jiahui Wu, Chengjie Lu, Aitor Arrieta, and Shaukat Ali. A tool for benchmarking large language models' robustness in assessing the realism of driving scenarios, 2025.

[WWY+25] Mingyi Wang, Jingke Wang, Tengju Ye, Junbo Chen, and Kaicheng Yu. Do llm modules generalize? a study on motion generation for autonomous driving, 2025.

[XCIP22] Danfei Xu, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Bits: Bi-level imitation for traffic simulation, 2022.

[XGL+25] Jianrong Xu, Baicang Guo, Xingchen Lui, Wei Hong, Liangliang Li, Chenyun Xi, Yewei Shi, Peng Wang, and Ruohai Di. Unitsg: Unified modeling token scenarios generating for autonomous driving task, 2025.

[XHY+18] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks, 2018.

[ZCS+23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[ZJC+25] Zhiyuan Zhang, Xiaosong Jia, Guanyu Chen, Qifeng Li, and Junchi Yan. Trajtok: Technical report for 2025 waymo open sim agents challenge, 2025.

[ZKI+25] Zhejun Zhang, Peter Karkus, Maximilian Igl, Wenhao Ding, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Closed-loop supervised fine-tuning of tokenized traffic models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[ZKW+20] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[ZLD+23] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. Trafficbots: Towards world models for autonomous driving simulation and motion prediction, 2023.

A. Appendix

A.1. Supplementary Details for Experiments

Embeddings and Distances

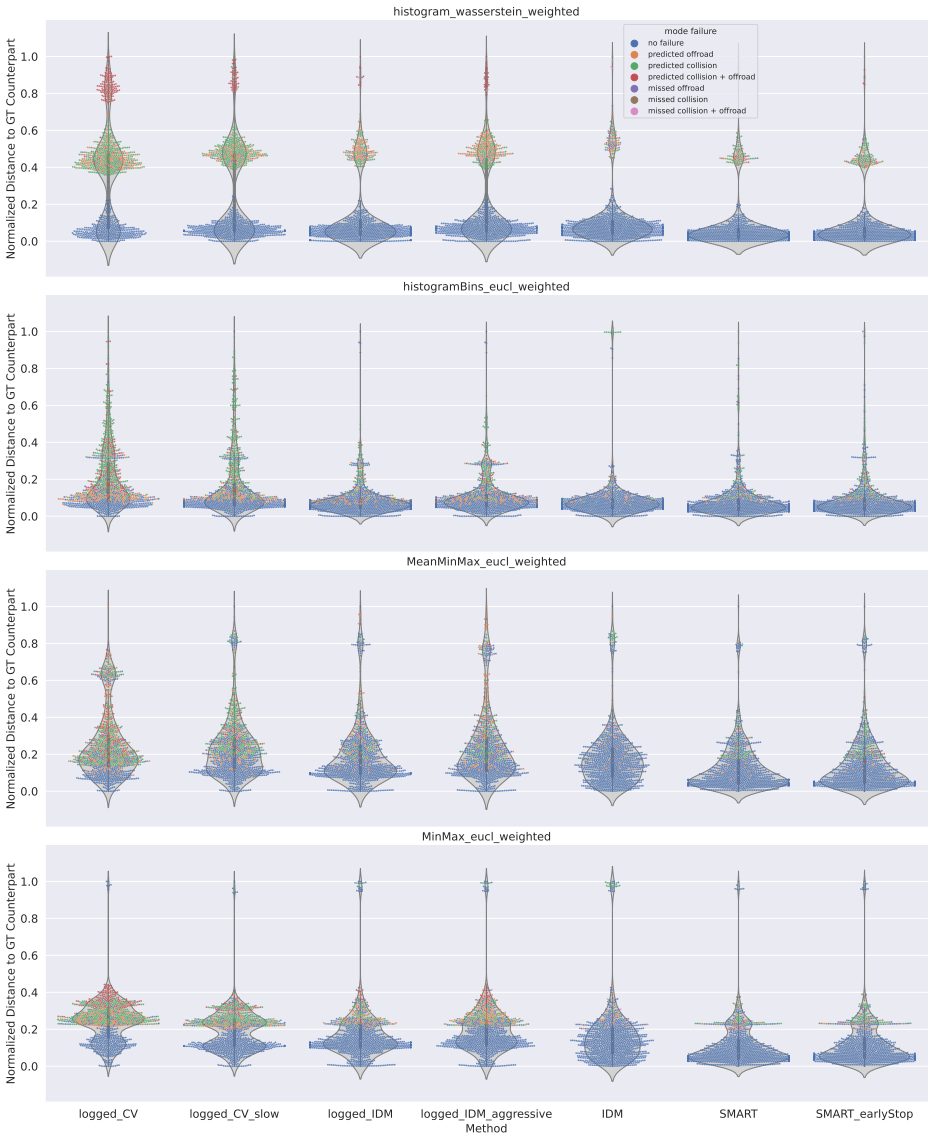


Figure A.1.: **Distance distributions.** For each embedding type (rows), we show the distribution of normalized distances between generated trajectories and their ground-truth counterparts. Notably, `Histogram_wasserstein_weighted` demonstrates the most distinct separation between failure modes, with greater distances assigned to single failure and even larger values observed for double failure modes across all methods. The second best separation is achieved by `MinMax_euclidean_weighted`.

Rollouts

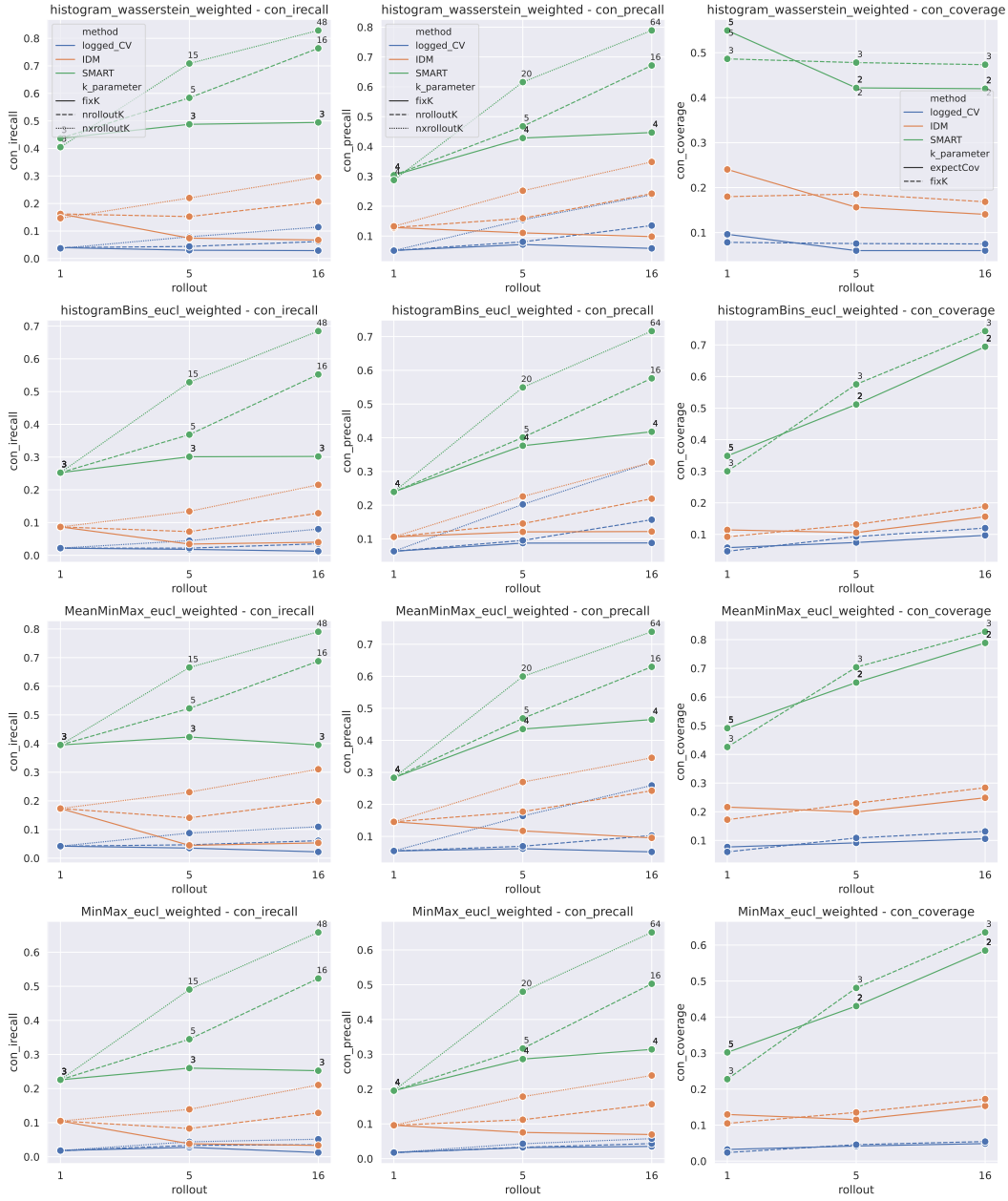


Figure A.2.: **Increasing number of rollouts under different k-scaling strategies.** Across all embedding types (rows), a consistent improvement in diversity metrics is observed when k is scaled proportionally to the number of rollouts. Conversely, maintaining a fixed k value leads to undesirable behavior, as recall scores decline despite the increased diversity induced by additional rollouts.

Data size

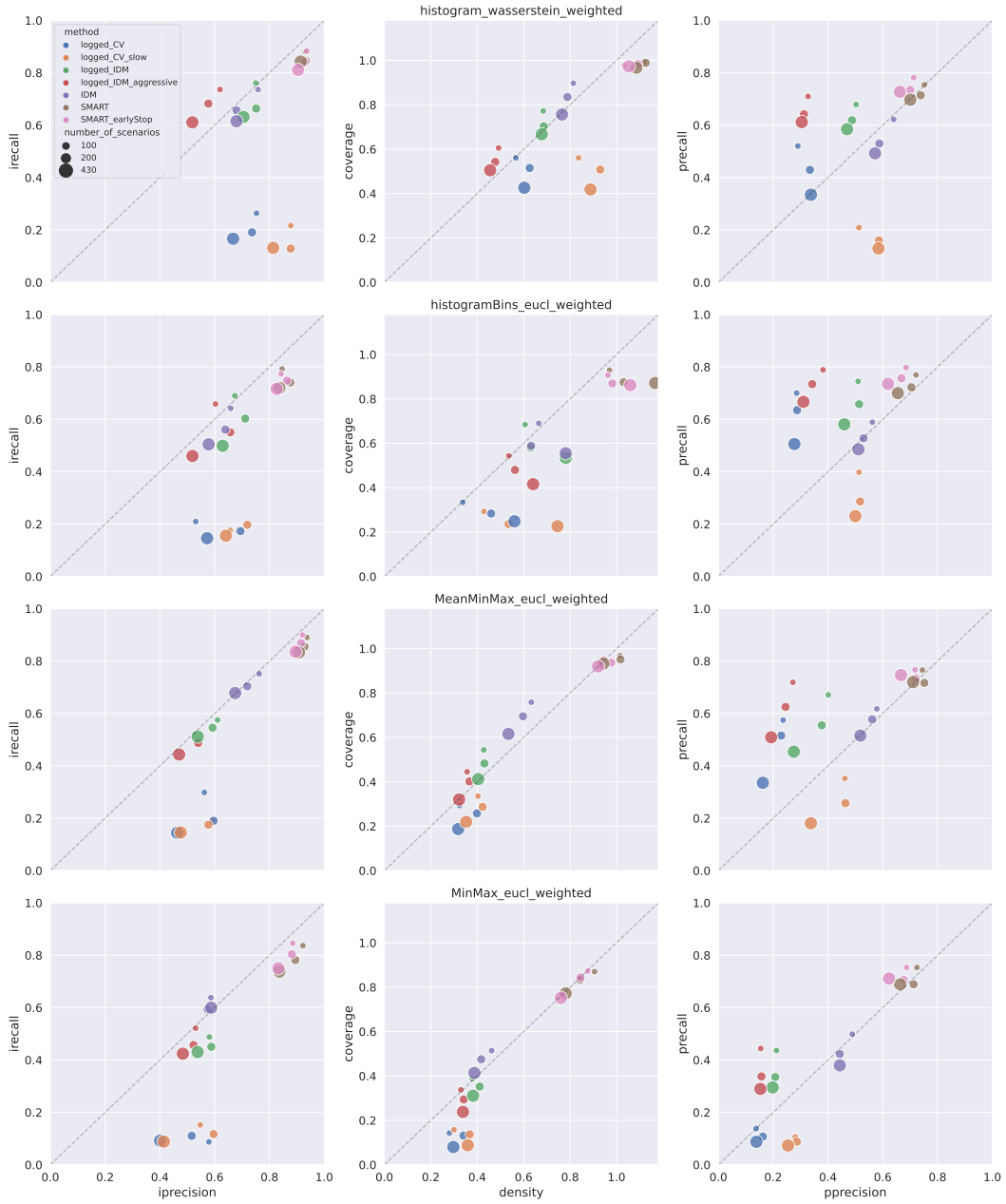


Figure A.3.: **Data size effect.** We examine how different data sizes affect each embedding type (rows) and each metric pair (columns). Across all configurations, the results demonstrate a consistent dependence on data size. Most of the time, the metrics exhibit elevated scores for sparser datasets and diminished scores as data density increases.

Application to Scenario Generation

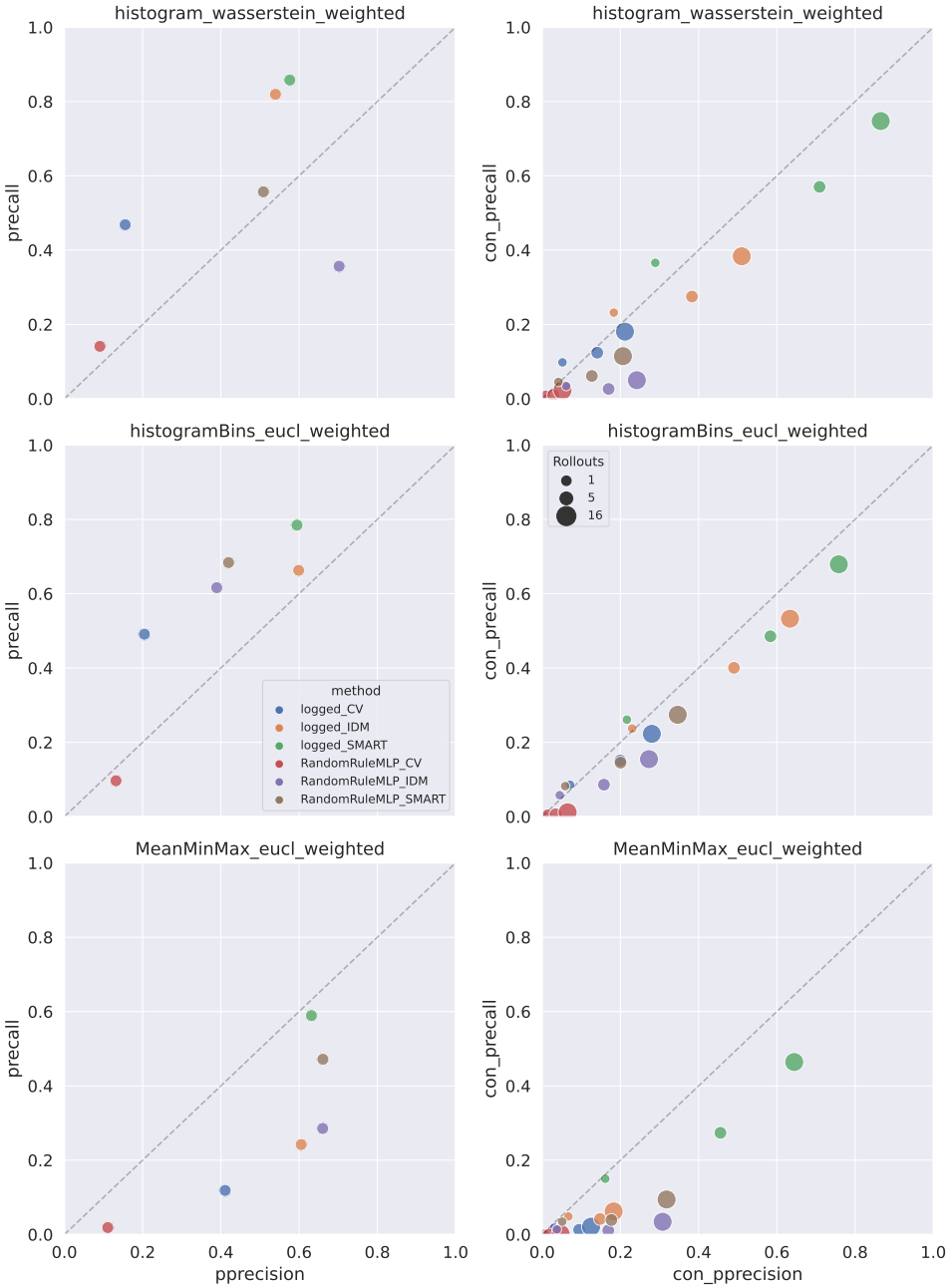


Figure A.4.: **Scenario-based evaluation of scenario generation.** In the unconditional case, histogram-based embeddings tend to emphasize higher recall, while simple statistical embeddings favor precision. As in traffic simulation, conditional fidelity metrics provide little information across embeddings because only a single ground-truth sample is available.

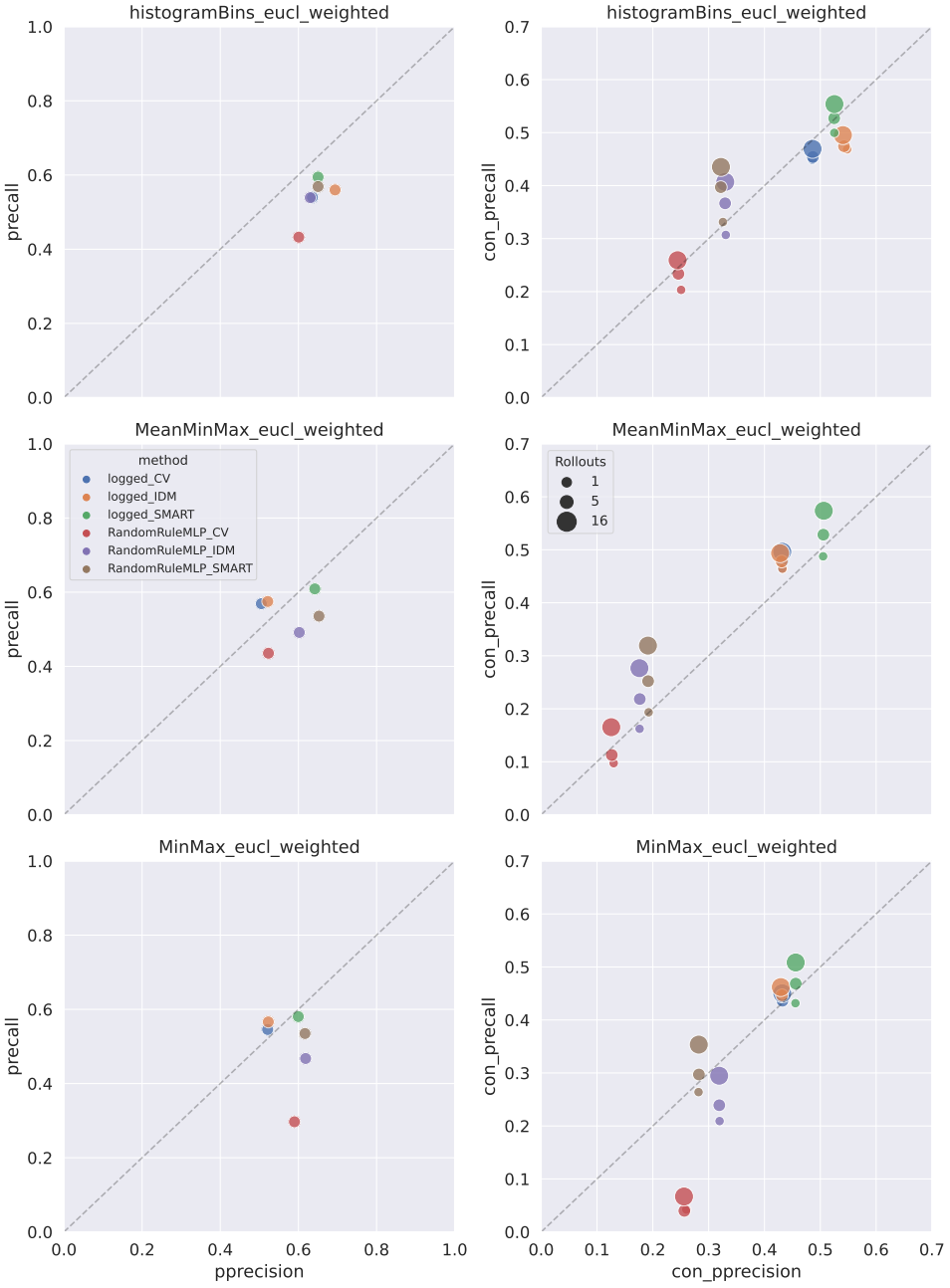


Figure A.5.: **Trajectory-based evaluation of scenario generation.** Supplementary results demonstrating low sensitivity across all embedding types in the unconditional setting, and largely consistent trends observed in the conditional case.

A.2. Supplementary Details for Results

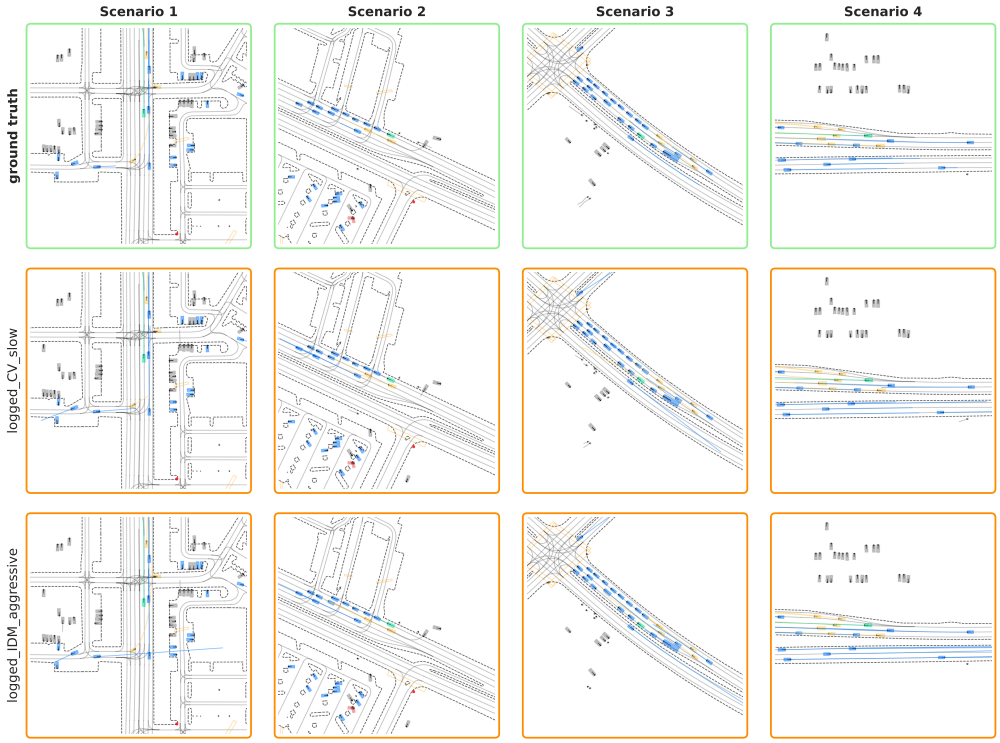


Figure A.6.: **Supplementary qualitative results for traffic simulation.** For completeness, we report the same scenarios for the methods excluded from Figure 4.12.

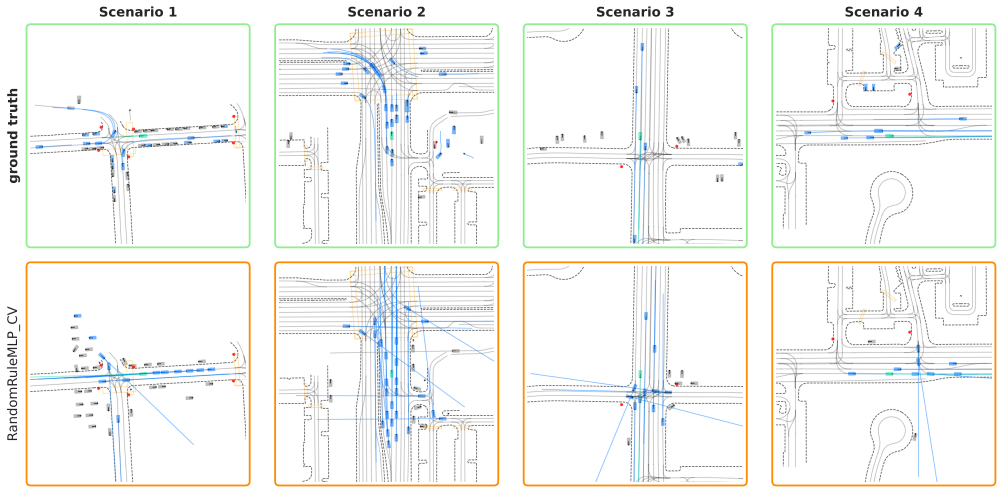


Figure A.7.: **Supplementary qualitative results for scenario generation.** For completeness, we report the same scenarios for the method excluded from Figure 4.13.

Erklärung

Laut Beschlüssen der Prüfungsausschüsse Bioinformatik, Informatik, Informatik Lehramt, Kognitionswissenschaft, Machine Learning, Medieninformatik und Medizininformatik der Universität Tübingen vom 05.02.2025. Gültig für Abschlussarbeiten (B.Sc./M.Sc./B.Ed./M.Ed.) in den zugehörigen Fächern. Bei Studienarbeiten und Hausarbeiten bitte nach Maßgabe des/der jeweiligen Prüfers/Prüferin.

1. Allgemeine Erklärungen

Hiermit erkläre ich:

- Ich habe die vorgelegte Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt.
- Ich habe alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet.
- Die Arbeit war weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens.
- Falls ich ein elektronisches Exemplar und eines oder mehrere gedruckte und gebundene Exemplare eingereicht habe (z.B., weil der/die Prüfer/in(nen) dies wünschen): Das elektronisch eingereichte Exemplar stimmt exakt mit dem bzw. den von mir eingereichten gedruckten und gebundenen Exemplar(en) überein.

2. Erklärung bezüglich Veröffentlichungen

Eine Veröffentlichung ist häufig ein Qualitätsmerkmal (z.B. bei Veröffentlichung in Fachzeitschrift, Konferenz, Preprint, etc.). Sie muss aber korrekt angegeben werden. Bitte kreuzen Sie die für Ihre Arbeit zutreffende Variante an:

- [x] Die Arbeit wurde bisher weder vollständig noch in Teilen veröffentlicht.
- [ ] Die Arbeit wurde in Teilen oder vollständig schon veröffentlicht. Hierfür findet sich im Anhang eine vollständige Tabelle mit bibliographischen Angaben.

3. Nutzung von Methoden der künstlichen Intelligenz (KI, z.B. chatGPT, DeepL, etc.)

Die Nutzung von KI kann sinnvoll sein. Sie muss aber korrekt angegeben werden und kann die Schwerpunkte bei der Bewertung der Arbeit beeinflussen. Bitte kreuzen Sie alle für Ihre Arbeit zutreffenden Varianten an und beachten Sie, dass die Varianten 3.4 - 3.6 eine vorherige Absprache mit dem/der Betreuer/in voraussetzen:

- [ ] 3.1. Keine Nutzung: Ich habe zur Erstellung meiner Arbeit keine KI benutzt.
- [x] 3.2. Korrektur Rechtschreibung & Grammatik: Ich habe KI für Korrekturen der Rechtschreibung und Grammatik genutzt, ohne dass es dabei zu inhaltlich relevanter Textgeneration oder Übersetzungen kam. Das heißt, ich habe von mir verfasste Texte in derselben Sprache korrigieren lassen. Es handelt sich um rein sprachliche Korrekturen, sodass die von mir ursprünglich intendierte Bedeutung nicht wesentlich verändert oder erweitert wurde. Im Zweifelsfall habe ich mich mit meinem/r Betreuer/in besprochen. Alle genutzten Programme mit Versionsnummer sind im Anhang meiner Arbeit in einer Tabelle aufgelistet.

☒ 3.3. Unterstützung bei der Softwareentwicklung: Ich habe KI als Unterstützung beim Schreiben von Code in der Softwareentwicklung genutzt. Es handelt sich hierbei lediglich um Unterstützung und nicht um die automatische Generierung von größeren Programm-Teilen. Im Zweifelsfall habe ich mich mit meinem/r Betreuer/in besprochen. Alle genutzten Programme mit Versionsnummer sind im Anhang meiner Arbeit in einer Tabelle aufgelistet.

☐ 3.4. Übersetzung: Ich habe *nach vorheriger Absprache und mit Erlaubnis meines/r Betreuer/in* KI zur Übersetzung von mir in einer anderen Sprache geschriebenen Texte genutzt. Jede derartige Übersetzung ist im laufenden Text gekennzeichnet und der Anhang meiner Arbeit enthält eine Tabelle mit einem vollständigen Nachweis aller übersetzten Textstellen und der verwendeten Programme mit Versionsnummer.

☐ 3.5. Code-Generierung: Ich habe *nach vorheriger Absprache und mit Erlaubnis meines/r Betreuer/in* KI zur Erzeugung von Code in der Softwareentwicklung genutzt. Der Anhang meiner Arbeit enthält eine Tabelle mit einem vollständigen Nachweis aller derartigen Nutzungen, der verwendeten Programme mit Versionsnummer und der verwendeten Prompts.

☐ 3.6. Text-Generierung: Ich habe *nach vorheriger Absprache und mit Erlaubnis meines/r Betreuer/in* KI zur Erzeugung von Text in meiner Arbeit genutzt. Jede derartige Verwendung von KI ist im laufenden Text gekennzeichnet und der Anhang meiner Arbeit enthält eine Tabelle mit einem vollständigen Nachweis aller derartigen Nutzungen, der verwendeten Programme mit Versionsnummer und der verwendeten Prompts.

Falls ich in irgendeiner Form KI genutzt haben (siehe oben), dann erkläre ich:

Mir ist bewusst, dass ich die Verantwortung trage, falls es durch die Verwendung von KI zu fehlerhaften Inhalten, zu Verstößen gegen das Datenschutzrecht, Urheberrecht oder zu wissenschaftlichem Fehlverhalten (z.B. Plagiaten) kommt.

4. Abschluss und Unterschrift(en)

Mir ist bekannt, dass ein Verstoß gegen diese Erklärung prüfungsrechtliche Konsequenzen haben und insbesondere dazu führen kann, dass die Prüfungsleistung mit „nicht ausreichend“ bzw. die Studienleistung mit „nicht bestanden“ bewertet wird und bei mehrfachem oder schwerwiegendem Täuschungsversuch eine Exmatrikulation erfolgen bzw. ein Verfahren zur Entziehung eines eventuell verliehenen akademischen Titels eingeleitet werden kann.

Micha Fauth	Tübingen, 30.11.2025	M. Fauth
Vorname, Nachname Student/in	Ort, Datum	Unterschrift

Die Punkte 3.4 - 3.6 erfordern eine Zustimmung des/r Betreuer/in. Sollten Sie einen dieser Punkte angekreuzt haben, dann sollte der/die Betreuer/in bitte hier unterschreiben:

Ich habe der oben genannten Nutzung von KI zur Erstellung der Arbeit zugestimmt.

Vorname, Nachname Betreuer/in	Ort, Datum	Unterschrift

Generative AI tool usage

The following table lists the AI tools used in this thesis according to the marked conditions mentioned above:

Used generative AI tools
ChatGPT-5, released in August 2025
Claude 4.5 Sonnet released in September 2025
Grammarly AI tool, which was available during November 2025

Table A.1.: Used generative AI tools