

WOD-E2E: Waymo Open Dataset for End-to-End Driving in Challenging Long-tail Scenarios

Runsheng Xu^{*†}, Hubert Lin^{*}, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun
 John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, Ben Sapp
 Mingxing Tan, Jyh-Jing Hwang[†], Dragomir Anguelov

Waymo LLC

Abstract

Vision-based end-to-end (E2E) driving has garnered interest in the research community due to its scalability and synergy with multimodal large language models (MLLMs). However, current E2E driving benchmarks primarily feature nominal scenarios paired with existing open-loop evaluation metrics that fall short in capturing the multimodal nature of driving or effectively evaluating performance in long-tail scenarios. To address these gaps, we introduce the Waymo Open Dataset for End-to-End Driving (WOD-E2E). WOD-E2E contains 4,021 driving segments (approximately 12 hours), specifically curated for challenging long-tail scenarios that are rare in daily life with an occurring frequency of less than 0.03%. Each segment in WOD-E2E includes the high-level routing information, ego states, and 360-degree camera views from 8 surrounding cameras. To evaluate E2E driving performance on these long-tail situations, we propose a novel open-loop evaluation metric: Rater Feedback Score (RFS). Unlike conventional distance-based metrics, RFS measures how closely a predicted trajectory matches rater-annotated trajectory preference labels. WOD-E2E includes rater preference labels for validation, and a separate held out test set is used for the 2025 WOD-E2E Challenge benchmark leaderboard.

1. Introduction

Autonomous driving systems have traditionally followed a modular design approach that decomposes the driving task into distinct sub-tasks [13, 15, 23, 27, 37, 39]. However, recent focus on vision-based end-to-end (E2E) architectures [7, 22, 30, 35, 38] have increased due to their

improved scalability [34] and the natural synergy with multi-modal large language models (MLLMs) [14, 31].

Current real-world E2E driving datasets, such as NAVSIM [8], WOMD [10] and CoVLA [1], predominantly feature *nominal* driving scenarios that do not fully expose systems to the long tail situations. This limits the evaluation of the robustness and generalization ability of E2E driving systems.

We introduce the Waymo Open Dataset for End-to-End Driving (WOD-E2E) which explicitly focuses on long-tail situations, featuring scenarios which occur with a frequency of less than 0.03% (Figure 1). It contains 4,021 challenging driving segments totaling over 12 hours of manual and autonomous driving.

We observe that existing open-loop metrics may be inadequate to evaluate driving in long-tail scenarios. The popular Average Distance Error (ADE) or L2 error metric captures error between a prediction and a single future ground truth trajectory, despite driving behavior being inherently multi-modal, where multiple reasonable future trajectories are possible. Predictive metrics, such as PDMS scores [8], require annotated positions and future trajectories of road agents to calculate collision rates, and thus become impractical when novel or hard-to-detect objects are involved (e.g., the flock of birds in Figure 1). Furthermore, off-road behaviors typically incur high penalties in PDMS, yet in safety-critical scenarios, a vehicle might reasonably deviate off-road to avoid an emergency. To address these limitations, WOD-E2E dataset includes a set of human expert preference ratings on multiple potential trajectories to support a novel open-loop evaluation metric, the Rater Feedback Score (RFS).

We conduct rigorous studies to verify the dataset and RFS. Since its release, WOD-E2E has garnered great interest from the research community, with numerous methods ranked on our public benchmark leaderboard.

^{*}Equal contributions.

[†]Contact emails: Runsheng Xu <runshengxu@waymo.com>, Jyh-Jing Hwang <jyhjh@waymo.com>.

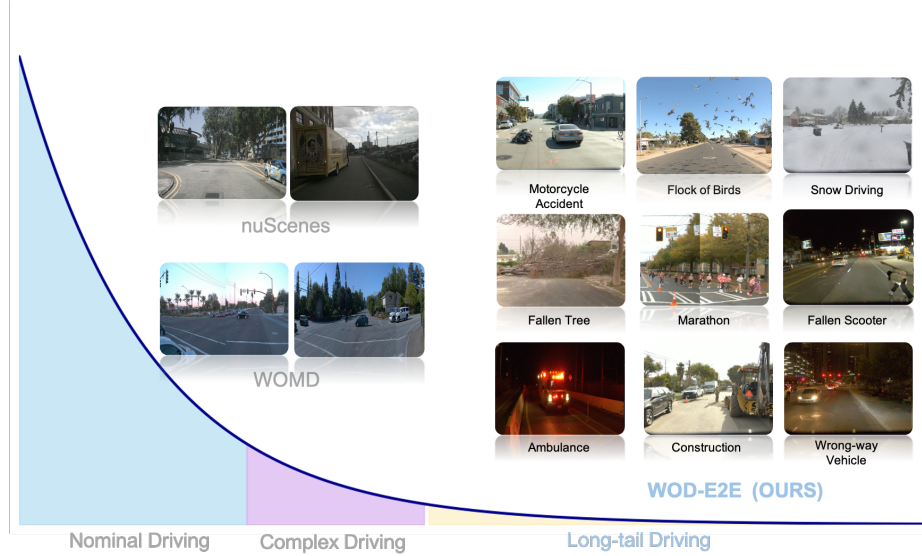


Figure 1. **Long-tail scenario examples from the Waymo Open Dataset for End-to-End Driving (WOD-E2E).** Unlike existing datasets that are commonly used for E2E Driving benchmarking, WOD-E2E dataset has more explicit focus on long-tail scenarios. Our analysis in Section 3.3 shows that WOD-E2E captures the long-tail scenarios with a frequency of less than **0.03%** in daily driving.

The diversity of top-performing methods, employing approaches such as MLLMs [21, 29], diffusion models [18], and CNN/ViT with GRU/MLP architectures [24], further underscores the utility of the WOD-E2E dataset in advancing end-to-end autonomous driving research.

Our contribution can be summarized as:

- **WOD-E2E**, a new dataset for benchmarking end-to-end driving. It contains 4,021 challenging driving segments, totaling 12 hours of driving and representing real-world long-tail scenarios occurring with a frequency of less than 0.03% in daily driving.
- **Rater Feedback Score (RFS)**, a novel human-aligned open-loop metric. RFS addresses the limitations of traditional open-loop metrics like ADE and PDMS for long-tailed driving.
- **Experimental analyses** to validate the dataset and RFS metric, and **learnings from top methods** submitted to our public benchmark leaderboard.

Section 2 describes related work, Section 3 covers details about WOD-E2E and RFS, and experiments and learnings are summarized in Section 4.

2. Related Works

2.1. End-to-end autonomous driving research

End-to-end (E2E) autonomous driving aims to overcome modular pipeline complexities by mapping sensor inputs directly to control [14, 34]. Recent foundation-model-based advances fall into three main categories:

Bird’s-Eye-View (BEV) Based E2E Planner:

These methods fuse multi-sensor data into a unified BEV representation for direct planning. UniAD [12] propagates queries from perception to planning tasks, while BEV-Planner [16] learns an explicit planning policy from dense BEV features. By avoiding intermediate perception outputs, these approaches offer improved computational efficiency and spatial coherence.

Multi-modal Large Language Model Based E2E Planner:

Leveraging MLLMs provides E2E systems with reasoning capabilities and world knowledge. DriveGPT4 [38] and DriveVLM [28] utilize LLMs for interpretability and chain-of-thought reasoning. EMMA [14] and its open-source variants [25, 35] process multiple driving tasks within a unified language space, while S4-Driver [34] lifts MLLM vision tokens into 3D space.

Diffusion Based E2E Planner:

Diffusion models excel at capturing the multi-modal nature of driving. DiffusionDrive [18] employs a truncated diffusion policy for real-time performance, while EnDfuser [33] uses ensembles to estimate trajectory uncertainty.

2.2. End-to-end autonomous driving open dataset

A multitude of autonomous driving datasets are available today, supporting a diverse range of driving tasks. Notable examples include Kitti [11], Argoverse [5], Argoverse 2 [32], WOD-Perception [26], and V2V4Real [36]. While these datasets serve various purposes, a spe-

cific subset focuses on end-to-end driving. Among the most prominent open datasets in this category are nuScenes [2], NAVSIM [8], WOMB [10], and CoVLA [1].

nuScenes [2] is primarily designed for perception tasks. While used for E2E research [14, 22, 28], performance is often achieved by simple extrapolation of historical behavior [17, 40], limiting its suitability for complex planning.

NAVSIM [8] is a compact simulation and benchmarking framework built upon a filtered version of nuPlan [3], whose core contribution lies in enabling large-scale real-world evaluation through a non-reactive simulator. Its approach presents two major limitations. First, it relies on filtering existing datasets rather than providing a **raw data collection effort**, which may preclude it from capturing the full, nuanced diversity of real-world long-tail scenarios. Second, the PDMS proposed in NAVSIM—which heavily prioritizes ego progress and comfort, along with time-to-collision (TTC)—may prove insufficient for true safety-critical situations. For instance, TTC is challenging to measure with amorphous obstacles like a flock of birds, as depicted in Figure 1, and the metric of comfort should be secondary to safety when the vehicle must perform an emergency maneuver. Our dataset and evaluation methodology are explicitly designed to overcome these two limitations.

WOMB [10] focuses on agent interaction modeling rather than planning. Crucially, it lacks the full camera images (only embeddings are provided) required for comprehensive vision-based E2E research.

3. WOD-E2E Dataset

3.1. Dataset Overview

This dataset contains 4,021 driving segments mined from real driving logs. Each segment is 20-second long and focused on long-tail scenarios. The dataset is partitioned as: 2,037 segments for training, 479 segments for validation, and the rest 1,505 segments for testing.

3.1.1. Coordinate System

This dataset employs two primary coordinate systems. **Vehicle Coordinates** are centered at the ego vehicle (x -forward, y -left, z -up) and serve as the reference for all trajectory data. **Sensor Frames** are related via extrinsic transformation; for cameras, the x -axis points out from the lens, z -axis upward, with the y/z plane parallel to the image plane (right-handed).

3.1.2. Camera Data

We use eight cameras providing 360° coverage. We provide a single JPEG image per direction, along with

camera intrinsics and extrinsics, enabling the projection of 3D trajectories onto the 10Hz video sequences. Training data spans 20s, while testing data covers 12s of observation with 8s of hidden future data.

3.1.3. Routing Information

A high-level command (enum: `{GO_STRAIGHT, GO_LEFT, GO_RIGHT}`) serves as a routing input [4, 12]. These commands specify the expected driving direction at decision points and are determined by comparing the vehicle’s 10s future driven route. Commands do not refer to micro-maneuvers (e.g. lane changes) or provide speed profile information.

3.1.4. Ego Status

Each segment includes ego vehicle status information. This comprises the ego vehicle’s **past 4s trajectory** (waypoints at 4Hz in vehicle coordinates) along with aligned velocity and acceleration data. Additionally, the ground truth **future 5s trajectory** is provided in the same format and frequency, available exclusively for the training and validation sets.

3.1.5. Labels

Scenario Cluster: Each segment is tagged with one of 11 scenario types, which will be explained in detail in the following sections.

Rater Feedback Labels: To capture the diversity of acceptable driving decisions during critical events, this dataset includes rater feedback labels. At specific moments within each driving segment, expert labelers rate three distinct 5-second future trajectories on a scale of 0 to 10, where 0 indicates the worst driving and 10 the best. Importantly, we ensure that at least one of the rater-specified trajectories receives a score higher than 6. This label is provided only for the validation set. Details on the creation of these labels will be provided in a subsequent section.

3.2. Quantitative Rariness Comparison

In this section, we quantitatively compare the **rarity** of WOD-E2E against other popular E2E driving datasets. To achieve a standardized, impartial rarity assessment, we utilized a large language model, Gemini 2.5 Pro [6], to score the test set of each dataset. The model was provided with the front camera sequences and a detailed scoring prompt outlining four tiers of rarity based on complexity, risk, and long-tail factors. The prompt required the output to be a JSON object containing the **rarity_score** that is ranged from 0-100, identified **rare_factors**, and a **reasoning** trace for maximum transparency.

After scoring each scene, we plot the comparative rarity distribution in Figure 2 (left). This curve is

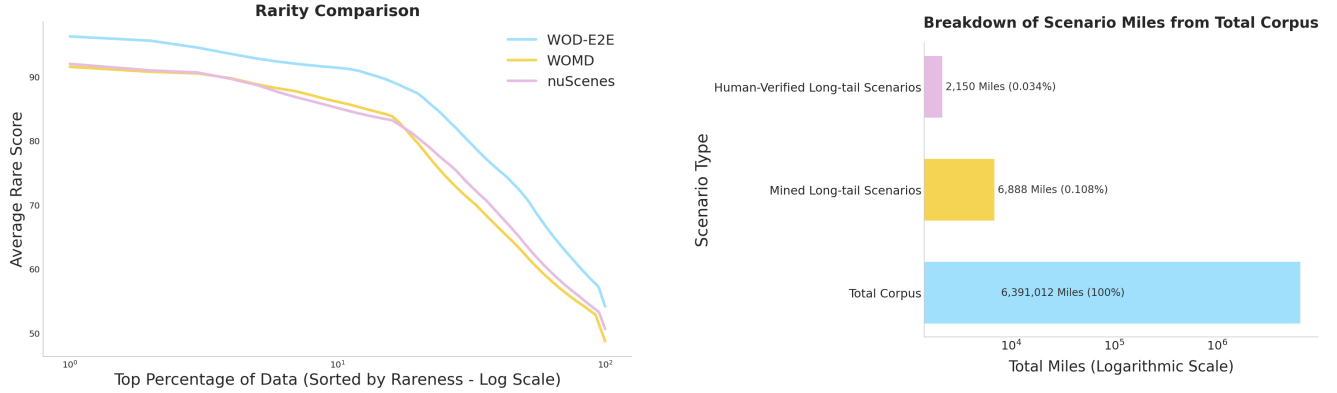


Figure 2. **Left:** Rarity comparison of driving datasets. This figure shows the average rarity score for the top percentage of data in each dataset, highlighting the distribution of rare events in WOD-E2E. **Right:** Proportion of mined long-tail scenarios (**0.03%**) from the total driving corpus (**6.4 million miles**).

generated by ranking all scenes by their rarity score (high to low) and plotting the average rarity score for all scenes up to that percentage of the dataset.

We can see that the WOD-E2E curve is the highest across all percentiles, indicating a consistently high density of rare scenarios relative to other datasets. Looking just at the most extreme 10% of each dataset, WOD-E2E reaches the highest score at around 93.

3.3. Long-tail Data Mining

3.3.1. Mining Strategy

We have access to a very large database containing diverse, real-world driving logs that span millions of miles. The vast majority of this data, however, consists of nominal scenarios. To effectively extract only the long-tail scenarios, we developed an efficient mining strategy that combines rule-based heuristics and MLLMs. Firstly, we categorized all driving logs into 11 different categories, including Construction, Intersection, Pedestrians, Cyclists, Multi- and Single-Lane Maneuvers, Cut-ins, Foreign Object Debris, Special Vehicles, Spotlight, and Others. The category definition and detailed mining criteria for each cluster can be found in the supplementary materials. These criteria are made possible by the rich auto-labels available in our dataset, including 3D detection, mapping, tracking, and prediction, which provide the necessary heuristics for our mining process.

3.3.2. Case Study

To validate the effectiveness of our mining strategy, we conducted a case study on a recent set of driving logs that includes a total of 6,391,012 miles. After applying our automated mining strategy, we found that only 6,888 miles (0.1%) of the data fit our criteria for long-tail scenarios. This initial result shows that our

strategy is highly effective at isolating rare, challenging events from a massive volume of nominal driving data, as demonstrated in the right figure of Figure 2.

Moreover, to ensure the highest quality of our dataset, we perform a subsequent round of human filtering. This manual review process, which has a conversion rate of 30%, further refined the mined data by removing non-long-tail scenarios. This final filtering step reduced the overall portion of long-tail scenarios to an even rarer 0.03%, highlighting the significant infrequency of these critical events in real-world driving.

3.3.3. Data Analysis

As shown in Figure 3, we analyze the dataset’s distributions across three key dimensions: city locations, scenario clusters, and driving behaviors.

City Distribution. The top-left subfigure shows the geographical distribution of the dataset across different cities. For confidential reasons, all city names have been anonymized. The data is predominantly sourced from cities L, K, and J, and the remaining cities, which are only present in the test set, contribute a smaller but more diverse set of scenarios, which is crucial for evaluating model generalization.

Scenario Clusters. The bottom-left subfigure provides a clear overview of our dataset’s composition by problem cluster and road type.

- We first analyze the distribution of long-tail scenarios by their problem clusters. The clusters for Intersections, Foreign Object Debris, and Pedestrians account for the largest share of the dataset. This highlights our focus on a variety of complex and safety-critical events, including intricate interactions at intersections, challenging scenes for the perception module, and high-risk encounters with pedestrians.
- Our dataset contains three major road types: Local

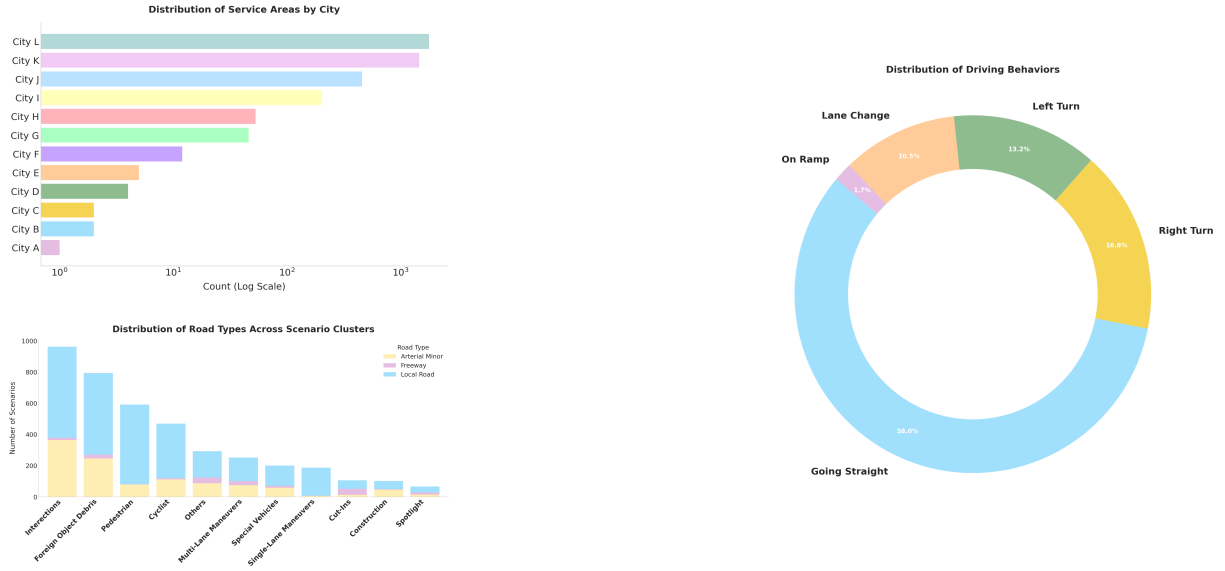


Figure 3. Comprehensive data distribution analysis. This figure illustrates the key characteristics of the WOD-E2E dataset across three critical dimensions. **Top Left:** Distribution of service areas by city. **Bottom Left:** Distribution of scenario clusters and their breakdowns by road type. **Right:** Distribution of driving behaviors.

Road, Arterial Minor, and Freeway. The Freeway road type is most prominent in the Cut-ins cluster, which is a particularly safety-critical event at high speeds. It is also notably present in the Interactions cluster. This is because these scenarios specifically capture interactions at freeway entrances and exits, such as making a right turn to enter an on-ramp.

Driving Behavior Distribution. The right subfigure shows the distribution of driving behaviors. We have a variety of diverse behaviors, including moving straight, lane changes, left turns, right turns, and on-ramp maneuvers. The majority of behavior is moving straight, which includes typical lane-following, but also hard braking and swerving for emergency situations. Turning behaviors at intersections, including left and right turns, make up approximately 30% of the data, with roughly equal proportions. Additionally, lane changes account for 10.3% of the scenarios, which usually involve collision or obstacle avoidance. Finally, a small portion of the data (1.7%) is dedicated to on-ramp behaviors, which are often challenging to tackle due to the interaction of merging vehicles at high speeds.

3.4. Data Labeling

The mined data is sent to our data labeling pipeline, which consists of three major steps: critical moment selection, trajectory sampling, and trajectory scoring.

3.4.1. Critical Moment Selection

The **critical moment** is the specific frame where a safety-critical event emerges, requiring a critical driv-

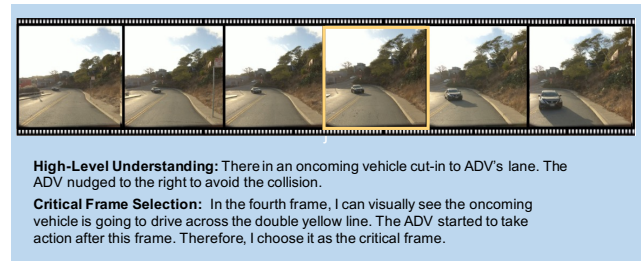


Figure 4. An illustration of how a critical frame is selected. The human raters first scan through the video for high-level understanding, and then select the critical frame, which is the earliest moment when a critical event is visually apparent in the camera images. Finally, the raters also document the rationales for the critical frame selection.

ing decision (e.g., slowing down, nudging, or yielding). Raters select this moment via a three-step process (Figure 4): 1) establish the high-level understanding of the necessary maneuver; 2) precisely identify the earliest frame where the event is visually apparent and the ego-action starts (to avoid historical motion bias); and 3) document the rationale for selection consistency.

3.4.2. Trajectory Sampling

Trajectory sampling generates a diverse set of motion plans for human review. We use an existing motion planning model (e.g. Wayformer [20]) to produce an initial set of up to **64 diverse trajectories**, using inputs like perception detections and predicted road agent

behaviors. This set is refined through a two-step process involving automated filtering into decision-based "buckets" (e.g., velocity, lane changes) and subsequent human-guided selection. Labelers then choose **three final trajectories** from these candidates for ranking and reasoning, ensuring the labeled data spans optimal to suboptimal behaviors.

3.4.3. Trajectory Scoring

Scored trajectories are generated and reviewed to establish accurate ground truth boundaries. Using a visualization tool, human raters review three candidate trajectories (optimal, plausible, and suboptimal) plotted over a **20** second scenario, including comprehensive data and agent annotations. The primary grading is based on five distinct dimensions: **Safety, Legality, Reaction Time, Braking Necessity, and Efficiency**. Trajectories are scored on a deductive **0** to **10** scale, where an initial score of **10** is reduced by **2** points for major infractions (Safety, Legality, Reaction Time) and **1** point for minor infractions (Braking Necessity, Efficiency). The resulting diverse score distribution (Figure 5) provides strong training boundaries. Further details on the specific scoring rubric are provided in Appendix.

3.5. Rater Feedback Score

The Rater Feedback Score (RFS) is a metric designed to evaluate the quality of a model's predicted trajectory with the reference of multiple human-annotated trajectories. The WOD-E2E dataset includes 3 reference trajectories generated by human raters, each assigned a score s_{rater} in $[0, 10]$.

The RFS is designed to see how much the model's prediction is aligned with three rated trajectories by considering trust regions, as illustrated in Figure 5. A *trust region* is defined around each rater trajectory at evaluation times t in $\{3, 5\}$ seconds. This region represents the rectangular space within specified longitudinal and lateral distance thresholds from the rater trajectory at a given time t .

The base thresholds follows WOMD [10], and they are set as $\bar{\tau}_{\text{lat}} = 1.0, \bar{\tau}_{\text{lng}} = 4.0$ at $t = 3$ and $\bar{\tau}_{\text{lat}} = 1.8, \bar{\tau}_{\text{lng}} = 7.2$ at $t = 5$, where the longitudinal threshold $\bar{\tau}_{\text{lng}}$ is always set to be 4 times larger than the lateral threshold $\bar{\tau}_{\text{lat}}$. These base thresholds are scaled based on the initial speed v (m/s) of the rater trajectory. The scaling function is a piece-wise linear function of v :

$$\text{scale}(v) = \begin{cases} 0.5, & v < 1.4, \\ 0.5 + 0.5 \times \frac{v-1.4}{11-1.4}, & 1.4 \leq v < 11, \\ 1, & v \geq 11. \end{cases}$$

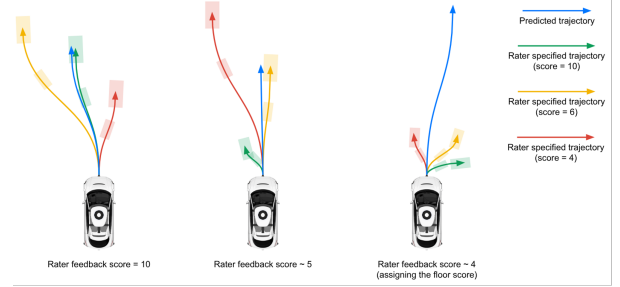


Figure 5. **Rater Feedback Score Mechanism.** This figure illustrates how the RFS evaluates a model's Predicted Trajectory (Blue) against three human-rated reference trajectories. The predicted score is based on the highest-rated reference trajectory it aligns with within the defined trust region.

The final thresholds at $t = 3, 5$ are determined by

$$\tau_{\text{lng}} = \text{scale}(v) \times \bar{\tau}_{\text{lng}}, \tau_{\text{lat}} = \text{scale}(v) \times \bar{\tau}_{\text{lat}}.$$

For distance errors Δ_{lng} (longitudinal) and Δ_{lat} (lateral) and the final thresholds, the score from each rater feedback trajectory is defined by

$$s_{\text{rater}} \times 0.1^{\max\left\{\max\left\{\frac{\Delta_{\text{lng}}}{\tau_{\text{lng}}}, \frac{\Delta_{\text{lat}}}{\tau_{\text{lat}}}\right\} - 1, 0\right\}}.$$

Intuitively, we assign either the flat score s_{rater} , if a predicted trajectory is within the trust region, or the score exponentially decayed from s_{rater} . Then, the final score is determined by choosing the maximum score over all rater specified trajectories, followed by averaging over $t = 3, 5$ and flooring with 4.

4. Experimental Results

4.1. Baseline Model Setup

We implemented a highly simplified version of EMMA [14], which we call NaiveEMMA, as our baseline model. NaiveEMMA is finetuned directly from Gemini Flash [6] and is trained exclusively on the released WOD-E2E training split. The model consumes a combined image from all eight cameras at the current timestep, concatenated into a single 768×768 resolution image. It also takes in 3 seconds of past ego-status history and the high-level routing input. It does not use past camera frames. Note that NaiveEMMA omits several advanced components of the original EMMA model described in [14], specifically generalist task training mixtures, Chain-of-Thought reasoning, and any test-time scaling methods.

			# of Dataset	Training Schema	Model Arch.	Model Param.
Methods	RFS↑	ADE↓	Training	Strategy	Model Setting	
MLP-based						
Swin-Trajectory	7.543	2.814	1	SFT	Swin Transformer	36M
Diffusion-based						
DiffusionLTF	7.717	2.977	4	SFT	DiffusionDrive	60M
UniPlan	7.779	2.986	2	SFT	DiffusionDrive	60M
MLLM-based						
Baseline	7.528	3.018	1	SFT	Gemini1 Nano	3B
AutoVLA	7.556	2.958	3	SFT+RL	Qwen2.5	3B
HMVLM	7.736	3.071	1	SFT	Qwen2.5	3B
Poutine	7.986	2.741	2	SFT+RL	Qwen 2.5	3B

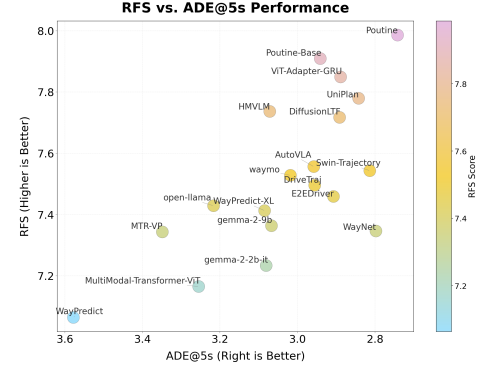


Table 1. WOD-E2E leaderboard submission results. **Left:** We summarize the results and configurations of selected representative methods among 3 categorical methodology (MLP-based, Diffusion-based, and MLLM-based). **Right:** We plot RFS vs ADE using 19 submissions. We only observe a mild positive correlation between RFS and ADE.

4.2. RFS Metric Validation

4.2.1. Quantitative Validation

We train several models based on NaiveEMMA and evaluate RFS on an internal test split. This test split contains long-tailed scenarios similar to the WOD-E2E test split. This experiment controls for several factors that are expected to improve model quality in long-tail settings: exposure to long-tailed scenarios via the WOD-E2E training split, multi-camera inputs to reason about surroundings, and test-time scaling to handle scenario ambiguities. RFS aligns with these intuitions, assigning higher scores to models that utilize more of these features (Table 2).

Model	RFS
Baseline	7.14
+ WOD-E2E finetuning	7.22
+ multi-camera inputs	7.30
+ test-time scaling (multi sampling)	7.39

Table 2. RFS assigns higher scores to models that are better-equipped to handle long-tailed scenarios. Evaluation is performed on an internal test split.

4.2.2. Qualitative Validation

Figure 6 validates RFS across three distinct conditions (additional examples provided in Appendix). First, **within the trust region** (Fig. 6a), the model correctly identifies a slow-moving construction vehicle and follows safely, aligning with the best-rated trajectory for a perfect RFS of 10.0. Second, for predictions **slightly outside the trust region** (Fig. 6b), the model follows the left-turn maneuver but at a slightly higher velocity than the labeled trajectory, causing the score to decay. Finally, for predictions **far from rater trajectories** (Fig. 6c), the model proceeds at high velocity in an

unrated region between valid lane-following and lane-change maneuvers; this significant deviation correctly triggers the floor score ($RFS = 4$).

4.3. Benchmark Models

Since the release of WOD-E2E, we have received a many submissions to our leaderboard. These models can be broadly divided into three categories: MLLM-based, Diffusion-based, and MLP-based models. The following section details the methods that have released a detailed report, as shown in Table 1.

Swin-Trajectory [24] is a lightweight, MLP-based model. It uses a Swin Transformer [19] to extract image features from three front cameras and a simple MLP to directly predict waypoints. The model is lightweight and achieves a slightly better RFS (7.543) than the baseline.

DiffusionLTF and **UniPlan** are both Diffusion-based models built on the DiffusionDrive [18] architecture. Their primary difference lies in the training datasets used: DiffusionLTF utilizes WOD-E2E CARLA [9], NAVSIM [8], and WOD-Perception [26], whereas UniPlan is trained on WOD-E2E and nuPlan [3]. They achieve comparable performance, with RFS scores of 7.717 and 7.779, respectively.

Poutine [21], **HMVLM** [29], and **AutoVLA** [41] are all MLLM-based models that use Qwen2.5 as their backbone. They share a similar problem formulation, taking camera images and ego states as input modalities and outputting future waypoints as text. Additionally, all three models utilize Chain-of-Thought (CoT) reasoning before generating a trajectory. Despite these similarities, their results show a significant performance gap, with AutoVLA achieving an RFS of 7.556, HMVLM 7.736, and Poutine 7.986. The primary differences among these methods stem from:

- **Training data sources:** AutoVLA uses a com-

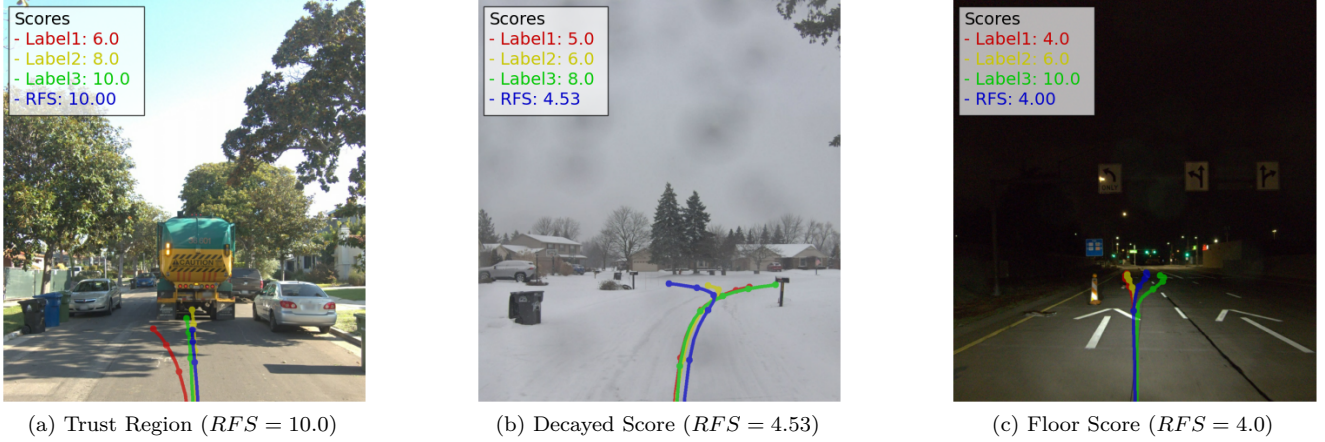


Figure 6. Qualitative RFS validation. (a) The prediction aligns with the rater’s optimal trajectory. (b) The prediction mimics the maneuver but deviates in velocity, causing score decay. (c) The prediction deviates significantly from valid regions, resulting in the floor score.

bination of WOD-E2E nuPlan, and nuScenes. In contrast, HMVLM is trained exclusively on WOD-E2E whereas Poutine uses a blend of WOD-E2E and the CoVLA dataset.

- **CoT captioning style** These three models employ different methods for generating reasoning captions and use distinct prompt templates.
- **RL training:** HMVLM does not include any post-training reinforcement learning. AutoVLA incorporates GPRO with ADE as the reward, whereas Poutine uses GPRO with RFS as the reward.

4.4. Discussion of the Results

From the results of these benchmark models, below we discuss important research questions in E2E Driving.

Q1: Is extra data source with large data distribution gap helpful for E2E Driving?

It depends. MLLM-based models (*e.g.* Poutine and AutoVLA) show obvious performance gains from supplementary data, while Diffusion-based models (*e.g.* UniPlan and DiffusionLTF) show only minor improvements. We hypothesize that the CoT reasoning in MLLMs allows them to **leverage diverse world knowledge** and abstract structures, mitigating visual or geometric distribution shifts. In contrast, diffusion-based models, relying more directly on pixel-level prediction, are more susceptible to performance degradation from combining visually disparate sources.

Q2: Does a better ADE always lead to a better RFS?

No. As shown in Figure 1 (Right), ADE and RFS exhibit only a mild positive correlation, confirming the limitations of ADE for safety-critical, multi-modal driving. We observe clear divergence: for example, WayNet has strong ADE but low RFS, whereas HMVLM has higher RFS despite a worse ADE. This highlights RFS’s

necessity for evaluating performance in long-tail scenarios.

Q3: Is RL effective in E2E Driving?

Yes, but only if the reward aligns with the target metric. Both Poutine and AutoVLA use RL post-training; however, Poutine, which uses RFS as its reward signal, achieved significantly greater gains than AutoVLA, which uses ADE. This confirms that selecting a reward function directly aligned with evaluation goals (long-tail safety) is critical for effective E2E optimization.

5. Conclusion

We introduced the Waymo Open Dataset for End-to-End Driving (WOD-E2E), a benchmark for challenging long-tail scenarios for end-to-end driving, addressing a gap in existing benchmarks that primarily focus on nominal driving. Our dataset provides 4,021 driving segments totaling 12 hours, focusing on rare events that occur with a frequency of less than 0.03%. We also introduced a novel human-aligned Rater Feedback Score (RFS), leveraging expert preference labels to overcome the limitations of traditional metrics like ADE in complex situations. Our dataset and metric are validated with a complete set of analyses, and community benchmark models highlight the promise of MLLMs with RFS-aligned RL rewards for E2E driving. We adopted an open-loop setup for WOD-E2E due to the prohibitive computational cost of realistic sensor simulation. Moving from open-loop to closed-loop continues to be an important limitation to overcome. However, WOD-E2E advances the state-of-the-art for open-loop E2E benchmarks, and its scenarios are applicable for testing the generalizability of high-fidelity simulators.

References

- [1] Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1933–1943. IEEE, 2025. 1, 3
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 3
- [3] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 3, 7
- [4] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3
- [5] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 2
- [6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3, 6
- [7] Can Cui, Yupeng Zhou, Juntong Peng, Sung-Yeon Park, Zichong Yang, Prashanth Sankaranarayanan, Jiaru Zhang, Ruqi Zhang, and Ziran Wang. Vilad: A large vision language diffusion framework for end-to-end autonomous driving. *arXiv preprint arXiv:2508.12603*, 2025. 1
- [8] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 3, 7
- [9] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017. 7
- [10] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, 2021. 1, 3, 6
- [11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 2013. 2
- [12] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, 2023. 2, 3
- [13] Jyh-Jing Hwang, Henrik Kretzschmar, Joshua Manela, Sean Rafferty, Nicholas Armstrong-Crews, Tiffany Chen, and Dragomir Anguelov. Cramnet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In *ECCV*, 2022. 1
- [14] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. EMMA: End-to-end multimodal model for autonomous driving. *Transactions on Machine Learning Research*, 2025. 1, 2, 3, 6
- [15] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1
- [16] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahao Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *CVPR*, 2024. 2
- [17] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahao Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14864–14873, 2024. 3
- [18] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2, 7
- [19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 7
- [20] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *ICRA*, 2023. 5
- [21] Christopher Pal, Liam Paull, Roger Girgis, Luke Rowe, and Rodrigue de Schaetzen. Poutine: Vision-language-trajectory pre-training and reinforcement learning post-training enable robust end-to-end autonomous driving, 2025. 2, 7
- [22] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for

- autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 1, 3
- [23] Chenbin Pan, Burhaneddin Yaman, Senem Velipasalar, and Liu Ren. Clip-bevformer: Enhancing multi-view image-based bev detector with ground truth flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15216–15225, 2024. 1
- [24] Sungjin Park, Gwangik Shin, Jaeha Song, Sumin Lee, Hyukju Shon, Byounggun Park, Jinhee Na, Hawook Jeong, and Soonmin Hwang. Swin-trajectory: Technical report for 2025 waymo vision-based end-to-end driving challenge. 2, 7
- [25] Zhijie Qiao, Haowei Li, Zhong Cao, and Henry X Liu. Lightemma: Lightweight end-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2505.00284*, 2025. 2
- [26] Pei Sun, Henrik Kretschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2, 7
- [27] Pei Sun, Mingxing Tan, Weiyue Wang, Chenxi Liu, Fei Xia, Zhaoqi Leng, and Dragomir Anguelov. Swformer: Sparse window transformer for 3d object detection in point clouds. In *ECCV*, 2022. 1
- [28] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *CoRL*, 2024. 2, 3
- [29] Daming Wang, Yuhao Song, Zijian He, Kangliang Chen, Xing Pan, Lu Deng, and Weihao Gu. Hmvlm: Multistage reasoning-enhanced vision-language model for long-tailed driving scenarios. *arXiv preprint arXiv:2506.05883*, 2025. 2, 7
- [30] Hang Wang, Xin Ye, Feng Tao, Chenbin Pan, Abhirup Mallik, Burhaneddin Yaman, Liu Ren, and Junshan Zhang. AdaWM: Adaptive world model based planning for autonomous driving. In *The Thirteenth International Conference on Learning Representations*, 2025. 1
- [31] Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025. 1
- [32] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaese-model Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 2
- [33] Florian Wintel, Sigmund H Høeg, Gabriel Kiss, and Frank Lindseth. Using diffusion ensembles to estimate uncertainty for end-to-end autonomous driving. *arXiv preprint arXiv:2506.00560*, 2025. 2
- [34] Yichen Xie, Runsheng Xu, Tong He, Jyh-Jing Hwang, Katie Luo, Jingwei Ji, Hubert Lin, Letian Chen, Yiren Lu, Zhaoqi Leng, et al. S4-driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 1, 2
- [35] Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. Openemma: Open-source multimodal model for end-to-end autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025. 1, 2
- [36] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 2
- [37] Runsheng Xu, Chia-Ju Chen, Zhengzhong Tu, and Ming-Hsuan Yang. V2x-vitv2: Improved vision transformers for vehicle-to-everything cooperative perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1
- [38] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *RA-L*, 2024. 1, 2
- [39] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 2020. 1
- [40] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023. 3
- [41] Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025. 7