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

Supplementary Material

1. Data Mining and Scenario Details

1.1. Long-Tail Scenario Clusters

The WOD-E2E dataset is categorized into the following 11 challenging long-tail clusters:

- **Construction:** Scenarios involving construction zones.
- **Intersection:** Scenarios with complex interactions at intersections.
- **Pedestrians:** Scenarios involving interactions with pedestrians.
- **Cyclists:** Scenarios involving interactions with cyclists.
- **Multi-Lane Maneuvers:** Scenarios where the ego vehicle must change lanes on multi-lane roads.
- **Single-Lane Maneuvers:** Scenarios where the ego vehicle must take actions on single-lane roads.
- **Cut-ins:** Scenarios where other on-road agents cut into the ego vehicle's lane.
- **Foreign Object Debris (FOD):** Scenarios with rare objects such as animals or furniture, or abnormal road conditions.
- **Special Vehicles:** Scenarios involving emergency vehicles.
- **Spotlight:** Manually selected challenging scenarios found via targeted database searches.
- **Others:** Scenarios that do not belong to any of the above clusters.

1.2. Mining Criteria for Long-Tail Scenarios

The automated and manual mining strategy uses the following rule-based criteria derived from rich auto-labels to identify segments for each long-tail scenario category, as table 1 demonstrated.

2. Qualitative Rater Feedback Score (RFS) Validation

This section provides additional qualitative examples to validate the RFS metric, illustrating model behavior under different scoring conditions (Figure 1).

2.1. Scores within the Trust Region (Figure 1a)

In these examples, the model's predicted trajectory (blue) aligns closely with the highest-rated human trajectory, resulting in a flat (un-decayed) RFS score.

- **Left:** The optimal trajectory follows a slow-moving construction vehicle carefully. The model's prediction aligns perfectly (Score 10.0, RFS 10.0).
- **Center:** In a complex urban intersection with a cable car and a turning vehicle, the preferred trajectory is to proceed carefully (Score 8.0). The model's prediction is well-aligned, yielding RFS 8.0.
- **Right:** The best action is to safely nudge right to pass a bus. The model's prediction accurately follows this optimal behavior, yielding RFS 10.0.

2.2. Decayed Scores outside the Trust Region (Figure 1b)

In these cases, the predicted trajectory deviates slightly from the preferred path but maintains a safe general maneuver. Since the predictions fall outside the mathematically defined trust region, the final scores are exponentially decayed.

- **Left:** In snowy conditions, the prediction executes a required left-turn but at a slightly higher velocity than the human-labeled path, causing the score to decay.
- **Center:** An avoidance maneuver is required due to an oncoming motorcycle. The prediction executes a similar lateral swerve but maintains a smaller (more conservative) lateral distance to the lane edge, resulting in a decayed score.
- **Right:** The model predicts a speed significantly slower than the optimal (Score 10.0) path when navigating around a cyclist, leading to a decayed score.

2.3. Floor Scores for Predictions Far from Rater-Specified Trajectories (Figure 1c)

Floored scores (RFS=4) are assigned when the model's prediction deviates fundamentally from the set of acceptable behaviors defined by human raters.

- **Left:** Labeled paths include lane-following and a lane-change maneuver. The model proceeds at a high velocity in the unrated region *between* the two maneuvers, receiving the floor score.
- **Center:** While all labeled trajectories indicate a left turn, the model erroneously turns right, resulting in the floor score.
- **Right:** The labeled paths execute a right turn. The model proceeds straight, diverging completely from the specified maneuvers and receiving the floor score.

Table 1. Mining criteria for each long-tail scenario category.

Construction	**Intersection**
• Driving route changes due to road closures from a construction zone. • Uniformed pedestrians directing traffic. • Abnormal road surface conditions due to construction.	• Unprotected maneuvers with limited visibility or heavy traffic interactions. • Complex interactions at stop sign intersections. • Interactions with other traffic-violating agents at traffic light intersections. • Interactions with rails and cable cars at intersections.
Pedestrians	**Cyclists**
• Pedestrians crossing with low visibility due to occlusion or weather. • Emergent behavior required to avoid collisions with pedestrians exhibiting unexpected behaviors. • Pedestrians performing unsafe maneuvers specific to the autonomous vehicle.	• Cyclists losing control nearby. • Interactions with a group of cyclists.
Cut-ins	**Foreign Object Debris (FOD)**
• Oncoming agent cuts across the ego vehicle's trajectory. • An agent in a neighboring lane cuts across the ego vehicle's lane aggressively.	• Interactions with animals on road. • Debris that can causes damage on the ADV's path (e.g., large box, glass, metal debris). • Abnormal road condition (e.g., flooded road, fire on the roadside, severely degraded road).
Multi-lane Maneuvers	**Single Lane Maneuvers**
• Nudge maneuvers to overtake blocked agents in the current lane. • Lane merging maneuvers on freeway. • Other agents in the other lane get too close to ADV that could cause hazards.	• Overtake maneuvers in narrow single lane roads. • Interactions with open-door vehicle in a narrow single lane road.
Special Vehicles	**Spotlight**
• Emergency vehicles blocking road due to accidents or construction. • Pull-over required due to the emergency vehicles.	• Leveraging Gemini to search over the database to find scenarios containing certain long-tail objects.

2.4. RFS vs Closed-Loop Evaluation

We conduct a "reactive log-replay" experiment using the Waymax simulator. To isolate the quality of the plans scored by RFS, we fix the ego-vehicle to RFS-rated trajectories while allowing other agents to react dynamically, and measure alignment between RFS and closed-loop metrics (collision, off-road, progression, wrong way, off-route, *etc.*). RFS demonstrates a statistically significant alignment (p-value < 0.001) with closed-loop performance (see table below). RFS also vastly outperforms ADE (82% vs 51%) in top-1 accuracy, which measures how often the metric correctly identifies the same optimal trajectory as the closed-loop metrics. This confirms RFS is a reliable proxy for closed-loop safety. Comparing RFS to PDMS: PDMS is ill-suited for long-tail scenarios due to its need for precise annotations, which is infeasible for amorphous hazards (*e.g.*, flocks of birds). It also incorrectly penalizes context-dependent emergent behaviors, such as driving off-road to avoid collisions. In contrast, RFS leverages human judgment to correctly contextualize these necessary maneuvers where rigid rule-based metrics fail.

Metric	**Stat. Sig. (p-value)**	**Top-1 Acc.**
ADE (Baseline)	0.015	0.51
RFS (Ours)	**<0.001**	**0.82**

3. Detailed Trajectory Scoring Rubric

The Rater Feedback Score (RFS) is derived from a rigorous manual grading process applied to three selected candidate trajectories (optimal, plausible, and suboptimal). The full scenario visualization lasts **20** seconds

and includes comprehensive data (mapping, camera, agent annotations).

Grading Criteria (5 Dimensions)

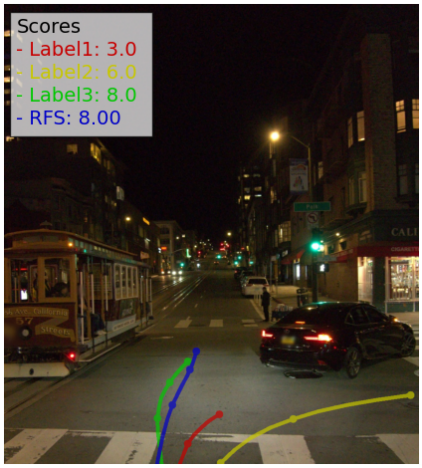
Raters score trajectories based on five distinct dimensions:

- **Safety:** Whether the trajectory results in collisions, near-misses, or other unsafe conditions.
- **Legality:** Whether the trajectory complies with all traffic laws and regulations, including proper behavior around emergency vehicles.
- **Reaction Time:** Whether the autonomous vehicle's actions within the trajectory are timely in response to unfolding events.
- **Braking Necessity:** Whether the trajectory includes unnecessary, sudden, or overly conservative braking.
- **Efficiency:** Whether the trajectory demonstrates efficient progress, avoiding unnecessary lane changes, hesitations, or over-reactions to distant or irrelevant agents.

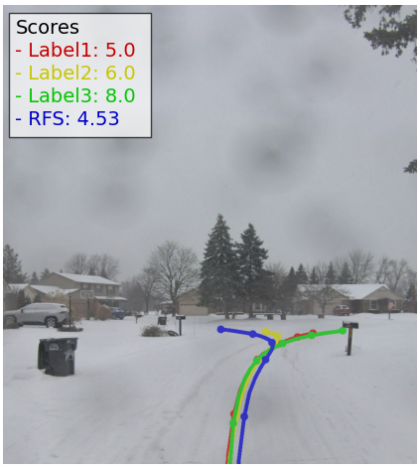
Scoring Mechanism

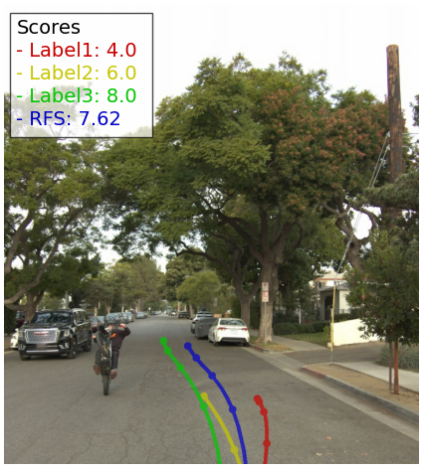
Trajectories are scored on a scale from **0** (worst) to **10** (perfect). The process is deductive:

1. Each trajectory is initialized with a base score of **10 points**.
2. Points are then deducted based on violations of the grading criteria:
   - **Major infractions:** A deduction of **2 points** is applied for violations related to Safety, Reaction Time, or Legality.
   - **Minor infractions:** A deduction of **1 point** is applied for violations related to Braking Necessity or Efficiency.
3. Penalties are cumulative, and raters may apply additional discretionary deductions to reflect the severity of combined faults.

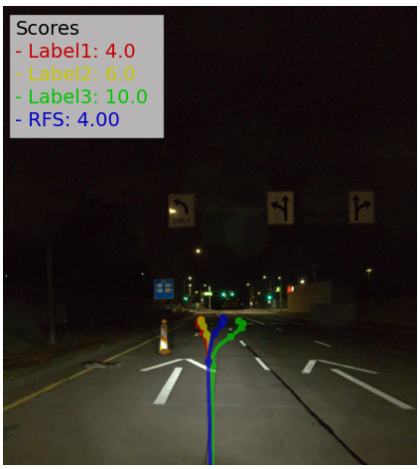


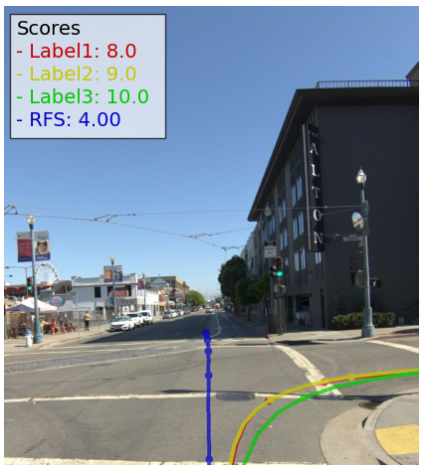
(a) The model predicted future trajectory (blue) aligns well with one of the rater specified trajectories. The corresponding flat scores are assigned as the predictions fall within the trust region.





(b) The model predicted future trajectory (blue) deviates from rater specified trajectories. Since the predictions fall outside the trust regions, final scores are exponentially decayed.





(c) Floored scores (RFS=4) are assigned because predictions are far from any of the rater-specified trajectories.

Figure 1. Visualization for the RFS metric in 3 different conditions. **Top:** The model predictions fall within the trust region. **Middle:** The model predictions fall slightly outside the trust region. **Bottom:** The model predictions are far from any of the rater-specified trajectories.