OmniDrive-R1: Reinforcement-driven Interleaved Multi-modal Chain-of-Thought for Trustworthy Vision-Language Autonomous Driving

Supplementary Material

A. Evaluation Metrics Details

In this section, we provide a detailed explanation of the evaluation metrics used in our experiments, specifically for the DriveLMM-o1 [20] and SURDS [17] benchmarks.

A.1. DriveLMM-o1 Benchmark Metrics

Following the evaluation protocol established in previous works [20, 32], we employ a dual-metric system consisting of **Overall Reasoning Score** and **Multiple Choice Quality (MCQ)**. To ensure fair comparison and reproducibility, we strictly adhere to the DriveLMM-o1 evaluation protocol. Specifically, we utilize **GPT-4o-mini** as the automated evaluator. The exact system prompt and the comprehensive scoring criteria employed in our evaluation are presented below:

1. **Faithfulness-Step (1-10):** Measures how well the model's reasoning steps align with the ground truth.
   - **9-10:** All steps correctly match or closely reflect the reference.
   - **7-8:** Most steps align, with minor deviations.
   - **5-6:** Some steps align, but several are incorrect or missing.
   - **3-4:** Few steps align; most are inaccurate or missing.
   - **1-2:** Majority of steps are incorrect.
2. **Informativeness-Step (1-10):** Measures completeness of reasoning.
   - **9-10:** Captures almost all critical information.
   - **7-8:** Covers most key points, with minor omissions.
   - **5-6:** Missing significant details.
   - **3-4:** Only partial reasoning present.
   - **1-2:** Poor extraction of relevant reasoning.
3. **Risk Assessment Accuracy (1-10):** Evaluates if the model correctly prioritizes high-risk objects or scenarios.
   - **9-10:** Correctly identifies and prioritizes key dangers.
   - **7-8:** Mostly accurate, with minor misprioritizations.
   - **5-6:** Some important risks are overlooked.
   - **3-4:** Significant misjudgments in risk prioritization.
   - **1-2:** Misidentifies key risks or misses them entirely.
4. **Traffic Rule Adherence (1-10):** Evaluates whether the response follows traffic laws and driving best practices.
   - **9-10:** Fully compliant with legal and safe driving practices.
   - **7-8:** Minor deviations, but mostly correct.
   - **5-6:** Some inaccuracies in legal/safe driving recommendations.
   - **3-4:** Several rule violations or unsafe suggestions.
   - **1-2:** Promotes highly unsafe driving behavior.
5. **Scene Awareness & Object Understanding (1-10):** Measures how well the response interprets objects, their positions, and actions.
   - **9-10:** Clearly understands all relevant objects and their relationships.
   - **7-8:** Minor misinterpretations but mostly correct.
   - **5-6:** Some key objects misunderstood or ignored.
   - **3-4:** Many errors in object recognition and reasoning.
   - **1-2:** Misidentifies or ignores key objects.
6. **Repetition-Token (1-10):** Identifies unnecessary repetition in reasoning.
   - **9-10:** No redundancy, very concise.
   - **7-8:** Minor repetition but still clear.
   - **5-6:** Noticeable redundancy.
   - **3-4:** Frequent repetition that disrupts reasoning.
   - **1-2:** Excessive redundancy, making reasoning unclear.
7. **Hallucination (1-10):** Detects irrelevant or invented reasoning steps not aligned with ground truth.
   - **9-10:** No hallucinations, all reasoning is grounded.
   - **7-8:** One or two minor hallucinations.
   - **5-6:** Some fabricated details.
   - **3-4:** Frequent hallucinations.
   - **1-2:** Majority of reasoning is hallucinated.
8. **Semantic Coverage-Step (1-10):** Checks if the response fully covers the critical reasoning elements.
   - **9-10:** Nearly complete semantic coverage.
   - **7-8:** Good coverage, some minor omissions.
   - **5-6:** Partial coverage with key gaps.
   - **3-4:** Major gaps in reasoning.
   - **1-2:** Very poor semantic coverage.
9. **Commonsense Reasoning (1-10):** Assesses the use of intuitive driving logic in reasoning.
   - **9-10:** Displays strong commonsense understanding.
   - **7-8:** Mostly correct, with minor gaps.
   - **5-6:** Some commonsense errors.
   - **3-4:** Frequent commonsense mistakes.
   - **1-2:** Lacks basic driving commonsense.
10. **Missing Step (1-10):** Evaluates if any necessary reasoning steps are missing.
    - **9-10:** No critical steps missing.
    - **7-8:** Minor missing steps, but answer is mostly intact.
    - **5-6:** Some important steps missing.

- **3-4:** Many critical reasoning gaps.
- **1-2:** Response is highly incomplete.

11. **Relevance (1-10):** Measures how well the response is specific to the given scenario and ground truth.
    - **9-10:** Highly specific and directly relevant to the driving scenario. Captures critical elements precisely, with no unnecessary generalization.
    - **7-8:** Mostly relevant, but some minor parts may be overly generic or slightly off-focus.
    - **5-6:** Somewhat relevant but lacks precision; response contains vague or general reasoning without clear scenario-based details.
    - **3-4:** Mostly generic or off-topic reasoning, with significant irrelevant content.
    - **1-2:** Largely irrelevant, missing key aspects of the scenario and failing to align with the ground truth.
12. **Missing Details (1-10):** Evaluates the extent to which critical information is missing from the response, impacting the reasoning quality.
    - **9-10:** No significant details are missing; response is comprehensive and complete.
    - **7-8:** Covers most important details, with minor omissions that do not severely impact reasoning.
    - **5-6:** Some essential details are missing, affecting the completeness of reasoning.
    - **3-4:** Many critical reasoning steps or contextual details are absent, making the response incomplete.
    - **1-2:** Response is highly lacking in necessary details, leaving major gaps in understanding.

A.2. SURDS Benchmark Metrics

The SURDS benchmark [17] is designed to rigorously evaluate the fine-grained spatial reasoning capabilities of VLMs. It comprises six distinct tasks divided into single-object and multi-object categories. We report the **accuracy (%)** for each task, utilizing specific calculation protocols based on the output type.

Single-object Spatial Reasoning:

- **Yaw (Orientation):** Evaluates the model's ability to estimate the heading angle of a specific object. The metric adopts a classification accuracy based on normalized textual matching.
- **Pixel (Localization):** Tests the model's capability to ground a textual description to a specific 2D region. Unlike standard bounding box IoU, SURDS employs a centerness score to evaluate the precision of the predicted point $P = (x, y)$ relative to the ground truth bounding box $B = \{x_{min}, y_{min}, x_{max}, y_{max}\}$.
  If P falls outside B, the score is 0. If inside, we compute the distances to the four borders (l, r, t, b) and the center-aligned ratios:

$$\mathcal{R}_{lr} = \frac{\min(l, r)}{\max(l, r)}, \quad \mathcal{R}_{tb} = \frac{\min(t, b)}{\max(t, b)} \tag{7}$$

  The score for a sample is defined as the geometric mean of these ratios: $S_{pixel} = \sqrt{\mathcal{R}_{lr} \cdot \mathcal{R}_{tb}}$.
- **Depth:** Assesses the model's ability to estimate the absolute distance or depth range. Similar to Yaw, this is evaluated via textual classification accuracy.

Multi-object Relational Reasoning:

- **Dis (Distance):** Evaluates the understanding of pairwise distances (e.g., determining if Object A is closer than Object B).
- **L/R (Left/Right):** Tests the ability to resolve lateral spatial relationships between two objects.
- **F/B (Front/Back):** Measures the comprehension of longitudinal spatial relationships in 3D space.

Metric for Classification Tasks: For the five tasks excluding Pixel (i.e., Yaw, Depth, Dis, L/R, F/B), we employ a normalized exact match accuracy. To ensure robust evaluation, we define a normalization function $\mathcal{N}(\cdot)$ which removes punctuation, articles (e.g., "a", "the"), and extra whitespace, converting text to lowercase. A prediction A_{pred} is considered correct if and only if:

$$\mathcal{N}(A_{pred}) = \mathcal{N}(A_{gt}) \tag{8}$$

The **Overall Score** for SURDS is calculated as the arithmetic mean of the accuracy scores across these six tasks, providing a holistic view of the VLM's spatial intelligence.