Prune2Drive: A Plug-and-Play Framework for Accelerating Vision-Language Models in Autonomous Driving

Supplementary Material

A. Experiments Details

A.1. Model architectures

A.1.1. DriveMM

DriveMM adapts SigLIP as the vision encoder, which is pre-trained on WebLI with a resolution of 384×384. DriveMM uses a 2-layer MLP as the projector to project the image features into the word embedding space. For the language model, DriveMM selects Llama-3.1 8B, which utilizes a tokenizer with a vocabulary of 128,000 tokens, and the model is trained on sequences of 8,192 tokens.

During training, DriveMM employs a 4-stage curriculum learning approach to train the model progressively across various autonomous driving datasets, including Language-Image Alignment, Single-Image Pre-Training, Multi-Capacity Pre-Training, and Driving Fine-Tuning.

A.1.2. DriveLMM-o1

DriveLMM-o1 is an InternVL2.5-8B model fine-tuned on the DriveLMM-o1 dataset. InternVL2.5 consists of a Vision Transformer image encoder and an LLaMA 2-based language model, enabling strong multimodal understanding. A key feature of InternVL2.5 is dynamic image patching, which enables variable-resolution image processing by dividing the input image into tiles, ensuring that finer details in complex scenes are better captured, which is essential for autonomous driving.

During training, DriveLMM-o1 freezes both the vision encoder and most of the language model's layers, ensuring the model retains its general multimodal knowledge while adapting to domain-specific reasoning. By leveraging LoRA and dynamic image patching, this fine-tuned InternVL2.5 model effectively integrates spatial and textual reasoning, refining its ability to process diverse real-world driving situations while maintaining computational efficiency.

A.2. Evaluation benchmarks

To comprehensively evaluate quantitative performance in autonomous driving scenarios, we adopt the evaluation metrics proposed by both evaluation benchmarks.

A.2.1. DriveLM

- **Accuracy**: calculates the ratio of correctly predicted samples to the total number of samples.
- **BLEU**: measures the similarity between a generated text and one or more reference texts by comparing n-grams in the generated text to those in the reference texts, with higher precision indicating a better match.
- **ROUGE_L**: calculates scores based on the longest common sub-sequence between the model outputs and the reference answers.
- **CIDEr** captures matches between n-grams of different lengths and differentiates the importance of various n-grams through TF-IDF weighting.
- **Chatgpt** is prompted to assign a numerical score between 0 and 100, with higher scores indicative of enhanced prediction accuracy. The prompt we employ is as follows: *Rate my answer based on the correct answer out of 100, with higher scores indicating that the answer is closer to the correct answer, and you should be accurate to single digits like 62, 78, 41, etc..*
- **Match** computes the average IoU of the predicted 2D bounding box and the ground truth 2D bounding box in the detection task.
- **Average Score** computes the weighted sum of Chatgpt Score, Language Score, Match Score, and Accuracy using weights of 0.4, 0.2, 0.2, and 0.2, respectively.

A.2.2. DriveLMM-o1

- **Risk Assessment** examines the model's capacity to prioritize high-risk objects or scenarios, ensuring critical situations receive appropriate urgency in decision-making processes. Accuracy measures the precision of environmental perception by quantifying the correct identification and classification of key objects in the environment.
- **Traffic Rule Adherence** assesses compliance with standard traffic regulations and domain-specific best practices, mirroring real-world operational reliability.
- **Scene Awareness and Object Understanding** jointly evaluate how well the response interprets and reflects critical objects in current scenarios, including their positions, predictions, and actions.
- **Relevance** evaluates alignment between model outputs and scenario-specific requirements relative to ground truth annotations, ensuring contextually appropriate responses aligned with humans.
- **Missing Details** evaluates the extent to which critical information is missing from the response by systematically analyzing perceptual gaps.

Moreover, DriveLMM-o1 employs various complementary metrics to collectively establish a holistic framework for evaluating system reliability, safety margins, and environmental adaptability across diverse driving conditions. DriveLMM-o1 utilizes GPT-4o-mini to complete the eval-

uation steps. The prompt we employ is as follows: *You are an autonomous driving reasoning evaluator. Your task is to assess the alignment, coherence, and quality of reasoning steps in text responses for safety-critical driving scenarios. You will evaluate the model-generated reasoning using the following metrics:*

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: The 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 are 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.

For the overall score, we compute the average of all metric scores. One example of a detailed evaluation metric is listed as follows:


```
{
  "Faithfulness-Step": 6.0,
  "Informativeness-Step": 6.5,
  "Risk Assessment Accuracy": 7.0,
  "Traffic Rule Adherence": 7.5,
  "Object Understanding": 8.0,
  "Repetition-Token": 7.0,
  "Hallucination": 8.5,
  "Semantic Coverage-Step": 7.5,
  "Commonsense Reasoning": 7.0,
  "Missing Step": 8.5,
  "Relevance": 8.5,
  "Missing Details": 7.0,
  "Overall Score": 7.42
}
```


B. Comparison Baselines

We select multiple representative training-free token pruning methods in VLMs for comparison.

FastV proposes a straightforward solution by leveraging the attention map in the second layer of LLM. It prunes tokens with the lowest visual-text attention score after layer 2 to achieve training-free token pruning.

SparseVLM adopts a multi-stage token pruning strategy. It mainly investigates the role of instruction tokens in vision-language attention mechanisms. It demonstrates that not all text tokens contribute equally to visual token pruning—only those highly relevant to the visual content are critical. To address this, the method first identifies the most vision-aligned text tokens as 'raters' and leverages their attention patterns to guide visual token pruning.

DART challenges attention-based token pruning by emphasizing redundancy reduction over token importance. It first randomly chooses pivot tokens and then selects visual tokens with minimal cosine similarity to chosen sets, thereby preventing duplication of visual information.

PACT achieves efficient visual token pruning by pruning irrelevant tokens and merging visually redundant ones at an early layer of the language model. It also proposes a novel clustering algorithm, called Distance Bounded Density Peak Clustering, which efficiently clusters visual tokens while constraining the distances between elements within a cluster by a predefined threshold.

C. Additional Visualizations

C.1. Comparison with baselines

We present additional visualizations comparing the results of retained visual tokens in Figure 1. It can be clearly observed that our method effectively considers the view importance while retaining informative visual tokens in aggressive pruning ratios, which is crucial for downstream autonomous driving tasks in VLMs.

C.2. Failure analysis

T-FPS may under-sample safety-critical objects when they occupy large image regions with uniform color, as areas sharing similar image features lead to fewer retained tokens. Fig. 2 illustrates two failure cases: an orange bus and three white vehicles are sparsely represented despite their safety relevance.

D. Additional Implementation Details

D.1. Evaluation Details

We introduce our evaluation details of both benchmarks in this section. For the DriveLMM-o1 benchmark, we use the official evaluation script to evaluate the full test dataset with GPT-4o-mini. For the DriveLM benchmark, following the DriveLM challenge repo, we upload our evaluation result to the official leaderboard and obtain detailed evaluation metrics after submission to the test server in Huggingface.

D.2. Pruning ratio optimization settings

We list our hyperparameters during the process of pruning ratio optimization below in Table 1:

Table 1. Hyperparameters in pruning ratio optimization.

Hyperparameter term	Value
Initial pruning ratio	0.9
Pruning ratio upperbound	1
Pruning ratio lowerbound	0.01
Dataset size	500
Max iteration	100
Penalty scale	-0.05
Reward scale	0.5

D.3. Hyperparameter λ sensitivity.

Tab. 2 shows that similar λ values yield comparable accuracy, demonstrating the robustness of our optimization method.

λ	0.04	**0.05**	0.06
Accuracy ↑	58.2	**58.3**	58.1

Table 2. Sensitivity of λ.

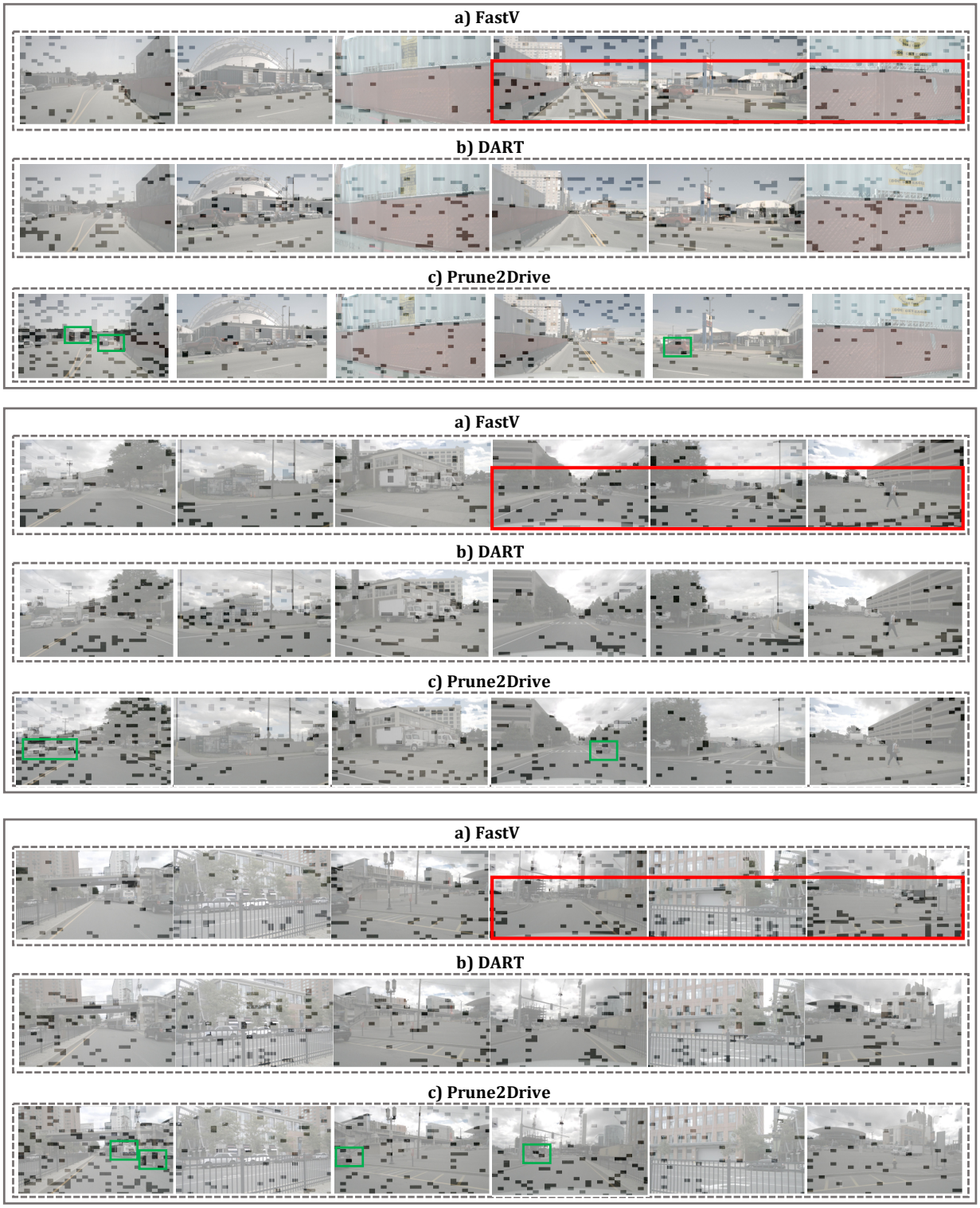

Figure 1. **Qualitative results of selected visual tokens.** We compare selected visual tokens by FastV, DART, and Prune2Drive. FastV shows position bias (red boxes), retaining mostly posterior tokens, DART neglects view importance, while our Prune2Drive (green boxes) captures critical objects through view-importance and diversity-aware selection.

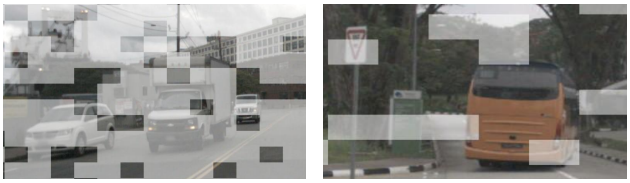

Figure 2. Qualitative failure cases analysis of T-FPS

D.4. Regular VLM benchmarks

GQA. GQA is structured around three core components: scene graphs, questions, and images. It includes not only the images themselves but also detailed spatial features and object-level attributes. The questions are crafted to assess a model's ability to comprehend visual scenes and perform reasoning tasks based on the image content.

MME. The MME benchmark is designed to rigorously evaluate a model's perceptual and cognitive abilities through 14 subtasks. It employs carefully constructed instruction-answer pairs and concise instructions to minimize data leakage and ensure fair evaluation. This setup provides a robust measure of a model's performance across various tasks.

POPE. POPE is tailored to assess object hallucination. It presents a series of binary questions about the presence of objects in images, using accuracy, recall, precision, and F1 score as metrics. This approach offers a precise evaluation of hallucination levels under different sampling strategies.

VQA V2. VQA V2 challenges models with open-ended questions based on 265,016 images depicting a variety of real-world scenes. Each question is accompanied by 10 human-annotated answers, enabling a thorough assessment of a model's ability to accurately interpret and respond to visual queries.

VizWiz. VizWiz is a visual benchmark designed to assist visually impaired individuals. It contains real-world images captured by blind users, paired with questions they ask about the images. The dataset includes 20,523 training, 4,319 validation, and 8,000 test image-question pairs, with each question accompanied by 10 human-annotated answers. VizWiz challenges models to answer questions accurately or determine if a question is answerable, focusing on practical visual understanding and accessibility.

D.5. Video AD benchmark OmniDrive

OmniDrive is a large-scale, multi-modal dataset curated for autonomous driving VLMs. With 374,329 training and 72,184 test samples derived from nuScenes, it supports both multi-view image and video inputs. This makes it particularly suitable for tasks requiring temporal reasoning and holistic scene understanding. Its evaluation protocol utilizes rule-based language metrics for fine-grained, word-level assessment.

E. Future Works

A key limitation of this work is the lack of closed-loop evaluation. Our experiments, confined to offline AD datasets, preclude the assessment of the model's performance in dynamic, interactive environments. This fundamentally limits our understanding of its real-world applicability and long-term behavior.