DIVIDE AND PRUNE: GEOMETRY-DRIVEN TOKEN PRUNING FOR EFFICIENT VLA DRIVING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language-Action (VLA) models have emerged as a premier paradigm for embodied AI, enabling scene comprehension and auto-regressive action generation in autonomous driving, yet their deployment in the real-world system is hindered by the excessive number of visual tokens from high-resolution inputs. Previous general visual token pruning approaches primarily exploit text-visual attention, visual attention maps or token similarity, inherently blinding themselves to the non-uniform space and risk-asymmetric nature of driving environments. In this work, we propose **Di**vide and **P**rune (**DiP**), a geometry-driven visual token pruning strategy for VLA driving models. DiP utilizes the depth estimation as spatial priors to divide the visual field into near-field and far-field regions. For near-field tokens, DiP fuses the edge energy with the token informativeness to retain local details critical for obstacle avoidance. For far-field tokens, we propose the depth-modulated diverse token selection to maximize semantic diversity while mitigating the semantic dilution caused by spatial overlap in 2D images. Experiments on the nuScenes and the Bench2Drive benchmarks show that DiP achieves strong performance while delivering an approximately 3.4× inference speedup.

1 INTRODUCTION

Autonomous driving increasingly permeates daily life and shows significant potential to define next-generation transportation Yang et al. (2023); Zhou et al. (2024); Chen et al. (2024a). Prior studies decompose autonomous driving into perception Li et al. (2024b), prediction Liu et al. (2021), and planning modules Caesar et al. (2021), aiming to obtain a globally effective driving system through the optimization of these individual components. Recent advances in multimodal large language models Chen et al. (2024b); Team et al. (2025); Li et al. (2024a) show that end-to-end architectures have achieved remarkable performance in Natural Language Processing and Computer Vision. Motivated by this progress, end-to-end autonomous driving models, particularly the vision-language-action (VLA) models, integrate different driving modules into a unified framework Kim et al. (2025); Chi et al. (2025). This paradigm can simplify the architectural design and bridge the semantic gap inherent in traditional modular pipelines, thereby emerging as a prominent research direction.

Nevertheless, high-resolution images inherently produce a massive number of visual tokens, which significantly increases the computational cost of the end-to-end models. Previous research Bolya et al. (2023) has disclosed substantial redundancy among the visual tokens. Therefore, visual token pruning emerges as an active research area to mitigate the severe inference latency of end-to-end models and facilitate their deployment in the real world. To identify the redundant tokens, previous approaches Zhang et al. (2025c); Jiang et al. (2025) obtain token importance scores using the text-visual attention mechanism. Other methods Zhang et al. (2025a); Yang et al. (2025) rely on visual attention and use the attention maps of the visual encoder to guide token selection. Recent approaches Alvar et al. (2025) also explore token selection strategies based on token similarity, with the goal of preserving a global understanding of the visual input.

While previous works have achieved satisfactory performance in visual question answering tasks, general token pruning methods remain poorly suited for autonomous driving. Specifically, as shown in Appendix A, text prompts in VLA driving models are concise and do not specify target objects, which makes text-visual attention mechanisms ineffective at providing the fine-grained cross-modal guidance. The visual attention-based approaches may suffer from the attention shift in traffic scenes,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

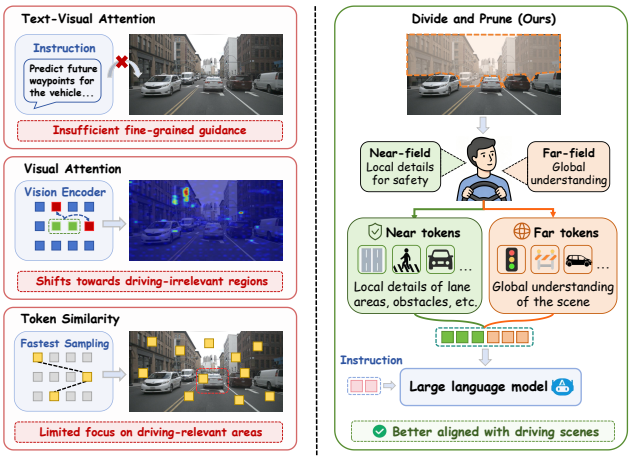


Figure 1: Comparison of prior methods with DiP. Text-visual attention methods struggle to provide fine-grained guidance from concise prompts. Visual attention-based methods risk safety, as the semantic salience may shift toward driving-irrelevant areas. Token similarity-based methods retain scattered tokens, limiting focus on driving-relevant areas. Our DiP employs a spatial divide-and-conquer strategy preserving near-field local details and far-field global understanding, which enables more effective pruning for driving scenes.

where more attention is assigned to semantically salient but driving-irrelevant billboards or buildings, thereby introducing safety risks. Token-similarity methods inherently preserve scattered tokens to maximize diversity, which leaves insufficient focus on driving-related objects under a limited token budget and fails to meet the requirements of safe driving.

To address the above issues, we propose Divide and Prune (DiP), a novel geometry-driven token pruning strategy. Fig. 1 illustrates how DiP differs from prior methods. Motivated by the observation that human drivers focus on obstacle avoidance for nearby objects while relying on coarse-grained understanding for distant objects, DiP implements a spatial division strategy to perform visual token pruning tailored for autonomous driving. In our method, depth information serves as a geometric prior that divides the scene into near and far fields. For the near field, obstacles typically contain richer local details than flat road surfaces. We therefore compute edge energy for near-field tokens and combine it with information capacity to determine their importance scores. As for far tokens, we preserve the token diversity and further modulate the token selection with depth information, which aims to mitigate semantic dilution caused by spatial overlap in the 2D image plane and to improve the global understanding of distant regions.

Overall, the contributions of this work can be summarized as follows:

- Observing spatial informational asymmetry in driving environments, we propose DiP, a novel paradigm utilizing geometric depth priors to divide the spatial representation into near and far fields.
- For near tokens, we fuse the edge energy with information capacity for each token, which aims to preserve local near-field details that are critical for obstacle avoidance.
- For far tokens, we utilize depth information to modulate the selection of diverse tokens, which achieves a better global understanding of driving scenes.
- Our proposed approach, which serves as a token pruning algorithm for end-to-end VLA models, achieves strong performance on the nuScenes and the Bench2Drive benchmarks.

2 Related Work

2.1 End-to-End Autonomous Driving Models

End-to-end autonomous driving models have progressively replaced conventional modular pipelines by directly mapping raw sensory inputs to the control or trajectory outputs Prakash et al. (2021); Hu et al. (2023). Early approaches Codevilla et al. (2018) have demonstrated the feasibility of learning driving policies from expert demonstrations, but suffered from poor generalization to the long-tail scenario. To inject high-level semantic priors into the driving policy, recent methods leverage large language models (LLMs) and vision-language models (VLMs) for scene understanding and decision making Shao et al. (2024); Sima et al. (2024); Wang et al. (2025). The vision-language-action (VLA) model Kim et al. (2025) Renz et al. (2025) further unify perception, language reasoning and action generation within an autoregressive framework, which is particularly attractive for driving models where perception and control are tightly coupled. Building on this trend, DriveGPT4 Xu et al. (2024) adapts LLMs for the motion planning. AutoVLA Zhou et al. (2025), DriveMM Huang et al. (2024) and Impromptu-VLA Chi et al. (2025) further scale up VLA training with large-scale driving data for fine-grained control in complex scenes. Nevertheless, the heavy computational cost of the LLM backbones and the dense visual tokens have become a major bottleneck for the real-time deployment.

2.2 Visual Token Reduction

Previous studies Bolya et al. (2023) have already demonstrated that redundant visual tokens bring unnecessary inference latency. A common strategy for acceleration is to reduce the number of visual tokens fed into the LLM. Methods based on text-visual attention, such as SparseVLM Zhang et al. (2025c) and DCP Jiang et al. (2025), score visual tokens by the cross-modal attention and prune those with low relevance. Representative works based on visual attention, such as VisPruner Zhang et al. (2025a) and VisionZip Yang et al. (2025), exploit visual cues from the vision encoder itself to identify salient tokens. DivPrune Alvar et al. (2025) and CDPruner Zhang et al. (2025b), as token similarity-based approaches, aim to select a representative subset of visual tokens through maximizing the inter-token dissimilarity.

Task-specific methods have recently emerged to address the limitations of the generic pruning method under domain-specific tasks. FastDriveVLA Cao et al. (2026) introduces a ReconPruner that prioritizes foreground regions through MAE-style pixel reconstruction for the driving models. Different from the previous approach, our proposed DiP explicitly leverages monocular depth estimation to spatially divide the visual token into near and far groups and designs tailored selection strategies to each group, which achieves a better accuracy-efficiency trade-off.

3 Methodology

3.1 Overview

In VLA models, a vision encoder typically obtains visual features and a projector subsequently maps the features into the word embedding space. Let $h(\cdot)$ denotes the composite function of the vision encoder and the projector, the visual tokens are obtained as $\mathbf{F}_v = h(\mathbf{V})$, where $\mathbf{V}$ represents the visual input. To implement token pruning, each visual token is assigned an importance score. Guided by these scores, we retain the most critical tokens. The textual input is represented by the text tokens $\mathbf{F}_t$, which are concatenated with the final visual tokens to form the input to the LLM $f(\cdot)$.

Figure 2 illustrates the DiP framework. Given the original visual input and visual tokens, the stage 1 employs a lightweight Depth Anything v2 model Yang et al. (2024), denoted as $g(\cdot)$, aim to perform spatial division and obtain near tokens $\mathbf{F}_{near}$ and far tokens $\mathbf{F}_{far}$. For near tokens, the stage 2 computes the edge energy and L2-norm of each token and further applies a proportional Top-K strategy based on the rankings derived from weighted rank fusion. For far tokens, the depth information is converted into Gaussian weights to modulate the semantic diversity of the tokens. Tokens with an infinite distance are added until the number of selected tokens satisfies the budget if necessary. The final selected visual tokens are then combined with the text tokens to form the input, which is fed into the LLM.

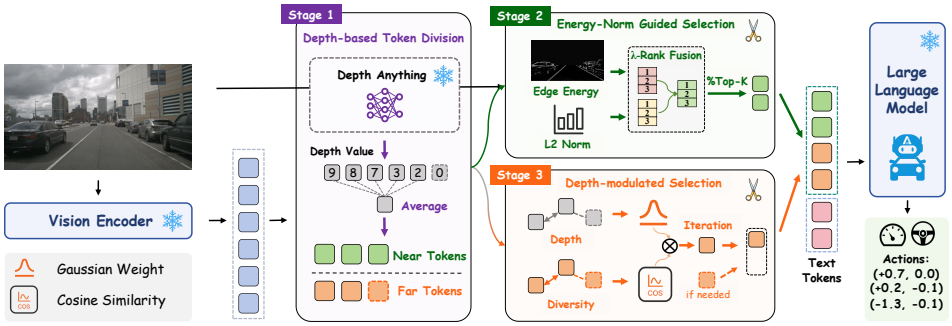


Figure 2: An overview of the DiP framework. In Stage 1, the average depth value for each token is computed using the lightweight Depth Anything model. The tokens are then divided into near and far tokens based on the average of the valid (i.e., non-zero) depth values. For the near tokens, Stage 2 applies a weighted rank fusion to the edge energy values and L2-norm values, which is followed by a proportional Top-K token selection. For the far tokens, Stage 3 retains semantic diversity while incorporating depth information through Gaussian-weighted modulation. Finally, all selected visual and text tokens are processed by the LLM to generate the predicted actions.

3.2 Spatial Division via Depth Priors

A lightweight Depth Anything v2 model is first introduced to provide dense depth cues, with only approximately 0.02B parameters. Note that the output of the Depth Anything v2 model is a relative depth map, where larger numerical values indicate closer proximity, with values ranging from 0 to 10. Given a visual input $\mathbf{V}$, we obtain the pixel-level depth map $\mathbf{Z} = g(\mathbf{V})$. To align the depth information with patch-level visual tokens, we partition the image into N patches following the same patch layout as the vision encoder. We compute the average depth of each patch aiming to obtain the patch-level depth representation $\mathbf{D} = [d_1, \ldots, d_N]^\top \in \mathbb{R}^{N\times 1}$, which is formulated as:

$$d_i = \frac{1}{|\mathcal{P}_i|} \sum_{(x,y)\in\mathcal{P}_i} \mathbf{Z}_{x,y},\ i = 1, \ldots, N, \tag{1}$$

where $\mathcal{P}_i$ denotes the set of pixel coordinates of each image patch, i means each index of the image patches and x, y denotes the pixel coordinates.

The Depth Anything model inherently assigns a depth value of 0 to objects at infinite distance, such as the sky, which indicates that the depth is too far to be estimated accurately. Based on this observation, we regard nonzero depth values as valid depth values $\mathcal{V} = \{i \mid d_i > 0\}$. To further separate near and far tokens, we utilize the mean of the valid depth values as the threshold τ. Tokens whose depth values are no less than the threshold are assigned to the near token set $\mathbf{F}_{near}$, while the remaining tokens are classified into the far token set $\mathbf{F}_{far}$, which can be represented as follows:

$$\begin{aligned} \mathbf{F}_{near} &= \{f_i \mid d_i \geq \tau\}, \\ \mathbf{F}_{far} &= \{f_i \mid d_i < \tau\}, \end{aligned} \tag{2}$$

where f_i denotes each visual token.

3.3 Energy-Norm Guided Token Selection

For objects in the near field, the objective is to avoid obstacles and recognize lane signs, thereby supporting safe driving. Compared with flat road surfaces, obstacles and lane signs typically contain richer edge and texture details. Building on this observation, we apply the lightweight and efficient Sobel operator Sobel & Feldman (1968) to the visual input $\mathbf{V}$ to obtain the gradient map. Given a token f_i within the near-token set $\mathbf{F}_{near}$, we can similarly compute the average gradient value of each image patch associated with f_i as its edge energy, which can be expressed as:

$$\mathbf{E}_i = \frac{1}{|\Omega_i|} \sum_{(x,y)\in\Omega_i} \text{Sobel}(\mathbf{V})(x, y), \tag{3}$$

where Ω_i represents the set of pixel coordinates in the image patch associated with the token f_i and $\text{Sobel}(\cdot)$ denotes the Sobel operator.

To further select tokens with high information capacity, the L2-norm of each near token $||f_i||_2$ is incorporated as an auxiliary criterion. Due to the discrepancy in scale between the edge energy and L2-norm, we rank the edge energy and L2-norm of each token and apply a weighted rank fusion approach, aiming to identify tokens that can both represent obstacles or lane signs in the near field and carry rich information. The above can be formulated as:

$$\mathbf{R}_i = \text{Rank}(\mathbf{E}_i) + \lambda \cdot \text{Rank}(\|f_i\|_2), \tag{4}$$

where $\mathbf{R}_i$ represents the fused rank of f_i and λ denotes the trade-off parameter.

Building on the computed token statistics, we next focus on the allocation budget k_1 of the near tokens. To preserve the global proportion of the near tokens, we determine the allocation budget through a proportional Top-K strategy. Specifically, we first compute the ratio p between the number of near tokens and the total number of tokens. Given a final retention budget of n tokens, the budget for near tokens is defined as $k_1 = \lceil p \cdot n \rceil$, which dictates the selection of the Top-K highest-ranking tokens:

$$\begin{aligned} \mathcal{I}_{near} &= \text{TopK}(\mathbf{R}, k_1), \\ \mathcal{T}_{near} &= \{f_i \mid i \in \mathcal{I}_{near}\}, \end{aligned} \tag{5}$$

where $\mathcal{T}_{near}$ denotes the selected near tokens and $\mathcal{I}_{near}$ denotes the corresponding indices.

3.4 Depth-modulated Diverse Token Selection

Considering that the objective of the far tokens $\mathbf{F}_{far}$ is to capture the global context, a straightforward approach is to employ the farthest point sampling (FPS) Qi et al. (2017). Suppose two far tokens as f_i and f_j, token-level FPS enhances diversity by comparing the cosine similarity $s_{i,j}$ between the tokens. Nevertheless, in 2D driving images, distant traffic lights may spatially overlap with the background buildings or sky, which dilutes the semantic distinctiveness of their tokens and can result in erroneous pruning.

To address the above issue, we propose a depth-modulated diverse token selection strategy based on FPS. Specifically, our method is performed only on tokens that belong to the valid depth values set $\mathcal{V}$, to mitigate the influence of the objects at infinite distance. To mitigate the semantic dilution that occurs when objects at varying depths overlap in the 2D space, we further introduce a depth-based Gaussian weight to adjust the similarity between any two distant tokens f_i and f_j, which can be expressed as follows:

$$w_{i,j} = \exp\left(-\frac{(d_i - d_j)^2}{2\sigma^2}\right), \tag{6}$$

where d_i and d_j represents the depth values of f_i and f_j, $\exp$ denotes the exponential function and σ is a hyperparameter.

We furthermore select the first centroid point by maximizing the feature dissimilarity to achieve a deterministic and representative initialization. Moreover, the cosine similarity between tokens is linearly rescaled to $[0, 1]$ and a depth-based Gaussian weight is then applied to reweight the token-level similarities, yielding the final depth-modulated semantic distinctiveness, which can be represented as:

$$\delta_{i,j} = 1 - \hat{s}_{i,j} \cdot w_{i,j}, \tag{7}$$

where $\hat{s}_{i,j}$ denotes the scaled cosine similarity.

Following the FPS, we perform iterative updates on the minimum distance array $\mathbf{A}_i$ for all candidate tokens to the sampled set $\mathcal{I}_{far}$. The next centroid is selected as the point furthest from the current sampled set, which can be expressed as:

$$\mathbf{A}_i = \min_{c \in \mathcal{I}_{far}} \delta_{i,c}, \; i^\star = \arg\max_{i \in \mathcal{V}} \mathbf{A}_i. \tag{8}$$

The allocation budget k_2 for the far token is calculated by subtracting the number of retained near tokens k_1 from the final retention budget. To maintain stability, when the number of far tokens with valid depth values is less than k_2, we can randomly sample tokens with invalid depth values to fulfill the budget and obtain the final sampled set $\mathcal{I}_{far}$. The selected far tokens are therefore given by $\mathcal{T}_{far} = \{f_i \mid i \in \mathcal{I}_{far}\}$. The overall algorithm flow is detailed in Appendix B.

3.5 PRUNING-BASED INFERENCE

With the pruned near tokens $\mathcal{T}_{near}$ and far tokens $\mathcal{T}_{far}$, we merge them according to their original spatial positions in the image sequence. We further concatenate all selected visual tokens with the text tokens $\mathbf{F}_t$ to construct the multimodal input to the LLM $f(\cdot)$ to generate the actions, which can be expressed as:

$$Actions = f([\mathcal{T}_{far}, \mathcal{T}_{near}],\ \mathbf{F}_t), \tag{9}$$

where $[\cdot]$ denotes the merge operation. Our DiP pruning algorithm enables the LLM to achieve higher inference efficiency while maintaining fairly satisfactory performance.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. We conduct the open-loop evaluation of our method on nuScenes Caesar et al. (2020), a large-scale autonomous driving benchmark containing 1,000 urban driving scenes, each lasting approximately 20 seconds. Following the evaluation protocol of prior work Cao et al. (2026), we use 6019 test samples in total with trajectory-prediction L2 error, Collision Rate and Intersection Rate as metrics. Closed-loop evaluation is assessed on the Bench2Drive benchmark Jia et al. (2024), which is based on the CARLA simulator and encompasses 220 routes across Town01 through Town15 under diverse weather conditions. Following the official evaluation metrics, we employ driving score, success rate, efficiency and comfortness to evaluate the performance of our method. Detailed descriptions of the metrics and the comparison baselines can be found in Appendix C and Appendix D.

Implementation Details. The baseline model for the open-loop evaluation is Impromptu-VLA Chi et al. (2025), a leading end-to-end autonomous driving model built on Qwen2.5-VL-3B Bai et al. (2025). For the closed-loop evaluation, we select SimLingo Renz et al. (2025) as the base model, which is mainly based on the InternVL-2 Gao et al. (2024) architecture. To reduce additional computational overhead, we employ the ONNX implementation of the small variant of Depth Anything v2 Yang et al. (2024), a well-established and representative monocular depth estimation model. All the experiments are conducted on NVIDIA RTX 4090 GPUs.

4.2 MAIN RESULTS

In this section, we perform comparison experiments across various datasets and benchmarks to evaluate our approach.

NuScenes Benchmark Performance. As shown in Table 1, with an original visual token count of 3249, our proposed method achieves the highest average inference performance across the pruning rates of 25%, 50% and 75%. Notably, at a 25% pruning rate, our approach outperforms the base model by about 0.4% and 2.4% in L2 and Intersection metrics, which is consistent with FastDriveVLA, demonstrating the necessity of pruning methods tailored to driving scenarios and the superior efficiency of our spatial divide-and-conquer strategy. At a 50% pruning rate, all the evaluated methods experience a pronounced performance decline in the Intersection Rate. We attribute this degradation to the substantial noise introduced by sparsifying irrelevant information under this pruning ratio. While alternative approaches suffer performance decreases ranging from 4.9% to 5.5%, our approach can successfully retain a performance level of 96.0%. At a pruning ratio of 75%, our method shows stronger robustness than other methods across all metrics, indicating that it can further filter noisy information while focusing more on driving-relevant objects for safe driving.

Bench2Drive Benchmark Results. As shown in Table 2, our method retains most of the driving capability of the vanilla model while preserving 25% of the visual tokens. Specifically, with a 75% reduction in visual tokens, our method achieves a Driving Score of 81.64, a Success Rate of 60.00% and an Efficiency score of 245.16, corresponding to 96.0%, 89.2% and 94.6% of the respective full-token performance. Notably, our method retains a clear performance advantage over existing full-token approaches despite operating with only one quarter of their visual-token budget. These results show that our pruning strategy can effectively remove redundant visual tokens while preserving the task-relevant information required for strong closed-loop driving performance, offering a favorable performance-efficiency trade-off under an aggressive token budget.

Table 1: Comparison results of different pruning methods on Impromptu-VLA.

Methods	L2 (cm) ↓					Collision (%) ↓					Intersection (%) ↓				
	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.
Upper Bound, All 3249 tokens (100%)															
Impromptu-VLA	13.97	28.38	53.13	31.83	100%	0.00	0.13	0.60	0.24	100%	0.53	2.34	5.52	2.80	100%
Retain 2436 Tokens (↓ 25%)															
SparseVLM	14.09	28.72	53.74	32.18	98.9%	0.00	0.17	0.67	0.28	86.9%	0.51	2.41	5.52	2.81	99.4%
VisPruner	14.02	28.50	53.44	31.99	99.5%	0.00	0.17	**0.61**	0.26	93.6%	0.51	2.40	5.51	2.81	99.6%
DivPrune	14.17	28.83	53.72	32.24	98.7%	0.00	0.17	0.73	0.30	81.1%	0.50	2.47	5.61	2.86	97.8%
FastDriveVLA	13.99	28.36	53.04	31.80	100.1%	0.00	0.15	0.63	0.26	93.6%	0.53	2.36	**5.42**	2.77	101.0%
DiP (Ours)	**13.88**	**28.28**	**52.97**	**31.71**	**100.4%**	**0.00**	**0.13**	0.63	**0.25**	**96.1%**	**0.50**	**2.26**	5.43	**2.73**	**102.4%**
Retain 1624 Tokens (↓ 50%)															
SparseVLM	14.24	28.97	54.17	32.46	98.0%	0.00	0.18	0.73	0.30	80.2%	0.53	2.62	5.73	2.96	94.5%
VisPruner	14.16	28.77	53.82	32.25	98.7%	0.00	0.17	0.65	0.27	89.0%	0.52	2.54	5.78	2.95	94.9%
DivPrune	14.20	28.98	54.12	32.43	98.1%	0.00	0.20	0.78	0.33	74.5%	**0.50**	2.63	5.72	2.95	94.8%
FastDriveVLA	14.08	28.65	53.57	32.10	99.1%	0.00	**0.15**	0.60	0.25	97.3%	0.55	**2.49**	5.78	2.94	95.1%
DiP (Ours)	**13.92**	**28.42**	**53.22**	**31.85**	**99.9%**	**0.00**	0.17	**0.58**	**0.25**	**97.3%**	0.61	2.51	**5.62**	**2.91**	**96.0%**
Retain 812 Tokens (↓ 75%)															
SparseVLM	14.58	29.47	54.81	32.95	96.6%	0.00	0.21	0.75	0.32	76.0%	0.57	2.58	5.74	2.96	94.4%
VisPruner	14.42	29.38	54.52	32.77	97.1%	0.00	0.19	0.73	0.31	79.3%	**0.52**	2.57	5.72	2.94	95.2%
DivPrune	14.50	29.46	54.57	32.84	96.9%	0.00	0.20	0.76	0.32	76.0%	0.55	2.54	5.70	2.93	95.4%
FastDriveVLA	14.28	29.18	54.46	32.64	97.5%	0.00	0.18	0.70	0.29	83.0%	0.55	2.50	5.68	2.91	96.1%
DiP (Ours)	**13.99**	**28.54**	**53.38**	**31.97**	**99.6%**	**0.00**	**0.17**	**0.68**	**0.28**	**85.9%**	0.55	**2.41**	**5.67**	**2.88**	**97.2%**

Table 2: Comparison results on the Bench2Drive closed-loop benchmark.

Method	Tokens Retained	Driving Score ↑	Success Rate (%) ↑	Efficiency ↑	Comfortness ↑
Vanilla	100%	85.07	67.27	259.23	33.67
DriveAdapter	100%	64.22	33.08	70.22	16.01
Orion	100%	77.74	54.62	151.48	17.38
AutoVLA	100%	78.84	57.73	146.93	**39.33**
DiP (Ours)	25%	**81.64**	**60.00**	**245.16**	21.56

4.3 Ablation Study

Different Depth Estimators. In this subsection we replace the depth estimator with an earlier work MiDaS Birkl et al. (2023) to further validate the effectiveness of our proposed method. As shown in Table 3, the results indicate that the model has a slight performance degradation when compared to the our configuration. We attribute this variation to the lower accuracy of the earlier depth estimation model, which affects the downstream spatial partitioning and token sampling processes. The achieved performance, however, remains at a high level. This robustness further confirms the overall effectiveness of the proposed depth-based spatial division strategy.

Table 3: Ablation study of different depth estimators

Methods	L2 (cm) ↓					Collision (%) ↓					Intersection (%) ↓				
	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.
Vanilla	13.97	28.38	53.13	31.83	100%	0.00	0.13	0.60	0.24	100%	0.53	2.34	5.52	2.80	100%
MiDaS	13.90	28.33	53.10	31.78	100.2%	0.00	0.13	0.65	0.26	93.6%	0.53	2.29	5.58	2.80	99.9%
DiP (Ours)	**13.88**	**28.28**	**52.97**	**31.71**	**100.4%**	**0.00**	**0.13**	**0.63**	**0.25**	**96.1%**	**0.50**	**2.26**	**5.43**	**2.73**	**102.4%**

Sensitivity Analysis of λ. This subsection presents a sensitivity analysis of the hyperparameter λ to further evaluate its effect on model performance. As shown in Figure 3, when λ increases from 0 to 1, the Collision Rate only changes slightly, which we attribute to the fact that both edge energy and L2-norm contribute to the obstacle perception. The Intersection Rate initially decreases, followed by a subsequent increase. Our analysis suggests that when λ is small, the objective is dominated by edge energy, which introduces certain redundancy and limits the perception of the spatial structure of the road. Conversely, a large λ value

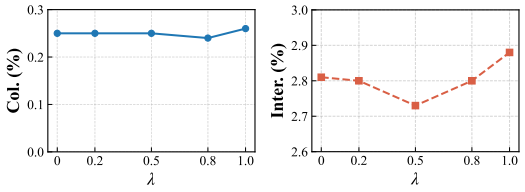


Figure 3: Sensitivity analysis of λ. Here, Col. and Inter. denote the average collision rate and average intersection rate, respectively.

may diminish the influence of the edge energy and thereby reduces the fine-grained perception of the scenes. Finally, we set λ to 0.5 to achieve a satisfactory balance between Collision Rate and Intersection Rate.

Sensitivity Analysis of σ. This subsection conducts a sensitivity analysis of the parameter σ to evaluate its potential impact on model performance, as illustrated in Figure 4. We define the baseline version, denoted as Base., which excludes the depth modulation process. As the value of σ increases, both Collision Rate and Intersection Rate decrease, which validates the effectiveness of the proposed depth-modulated reweighting method. However, an excessively large σ reduces the influence of depth reweighting and the small depth weighting coefficients may act as noisy perturbations to the token diversity, which reduces the model's global understanding of the far field and thus degrading its performance. Therefore, we set the σ value to 0.3 to achieve a good trade-off between the strength of depth reweighting and model performance.

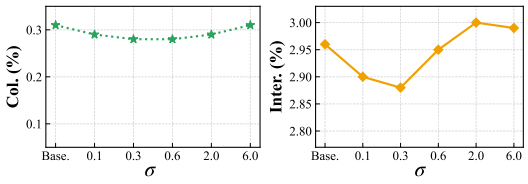


Figure 4: Sensitivity analysis of σ. Here, Col. abbreviates the average collision rate and Inter. abbreviates the average intersection rate.

Spatial Division Strategies. We compare three lightweight approaches in order to obtain the spatial division threshold: K-Means Ikotun et al. (2023), Otsu Otsu (1979) and average (Avg). As shown in Table 4, Avg method achieves the best average performance across all the metrics. We believe K-Means is sensitive to cluster initialization, where random initialization possibly lead to suboptimal division thresholds and therefore limit the model performance. Otsu assumes a bimodal distribution for the near and far fields, whereas depth values tend to exhibit continuous variations or multi-level distributions. This discrepancy can also produce suboptimal thresholds that impair the model performance. It is worth noting that our adopted Avg method achieve considerable trade-off between the prediction performance and the computational cost, demonstrating the effectiveness of the proposed spatial divide-and-conquer approach.

Table 4: Ablation study of the spatial division strategy

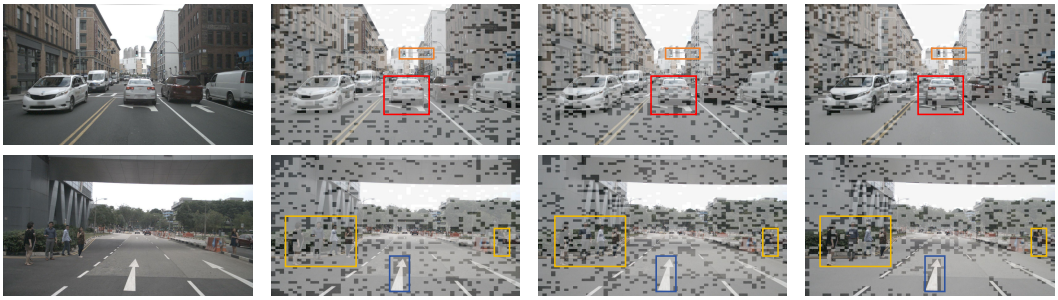
Methods	L2 (cm) ↓					Collision (%) ↓					Intersection (%) ↓				
	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.
Vanilla	13.97	28.38	53.13	31.83	100%	0.00	0.13	0.60	0.24	100%	0.53	2.34	5.52	2.80	100%
K-Means	13.90	28.34	53.10	31.78	100.1%	0.00	0.15	**0.61**	0.25	96.1%	0.56	2.28	5.57	2.80	99.8%
Otsu	13.88	28.28	52.98	31.71	100.4%	0.00	0.13	0.63	0.25	96.1%	0.53	2.36	5.70	2.86	99.4%
Avg (Ours)	**13.88**	**28.28**	**52.97**	**31.71**	**100.4%**	**0.00**	**0.13**	0.63	**0.25**	**96.1%**	**0.50**	**2.26**	**5.43**	**2.73**	**102.4%**

4.4 Efficiency Analysis.

To validate the efficiency of the proposed DiP, we conduct efficiency experiments, as illustrated in Table 5. Since visual token pruning mainly accelerates the pefill stage by reducing the number of visual tokens fed into the LLM, we report the prefill FLOPs and prefill time of DiP and recent advanced pruning methods. It is worth noting that the experimental results represent the full-pipeline cost, which already includes the overhead of the pruning algorithm. The results show that, compared with the vanilla model, DiP reduces FLOPs by 68.8% and achieves about a 3.4× speedup. While our approach does not yield the highest speedup when compared with non-parametric methods, it maintains a favorable trade-off between inference latency and model performance for safety-critical autonomous driving. Furthermore, acceleration techniques tailored to depth estimation models still remain underexplored, suggesting that our method holds potential for further optimization.

Table 5: Efficiency analysis across different pruning methods, where Δ represents the reduction ratio.

Methods	Token	FLOPs (T)↓	Δ	Prefill Time↓ (ms)	Δ
Vanilla	3249	24.7	-	185	-
VisPruner	812	8.1	↓ 67.21%	62	2.98×
DivPrune	812	**7.6**	↓ **69.23%**	**43**	**4.30**×
FastDriveVLA	812	8.2	↓ 66.80%	49	3.78×
DiP (Ours)	812	7.7	↓ 68.83%	55	3.36×



(a) Input Image (b) VisPruner (c) DivPrune (d) DiP (Ours)

Figure 5: Qualitative comparison of our method and prior pruning methods in the driving scene. Colored boxes highlight critical driving objects, including vehicles, traffic lights, pedestrians and lane areas. The results show that our method allocates more focus to driving-related objects within a limited token budget.

4.5 Qualitative Analysis.

Figure 5 shows the visualization results of token selection for recent baseline methods and our DiP. Visual attention-based method VisPruner tends to retain tokens that concentrate on buildings, billboards and other regions that are less relevant to driving, rather than on critical elements such as vehicles, pedestrians and lane areas, indicating that such approaches fail to satisfy the safety priority of autonomous driving. The pruning results of the token similarity-based DivPrune contain numerous scattered tokens, which can inherently restrict its focus on driving-related objects such as lane signs. As highlighted by the orange boxes in the figure, DivPrune also omits the tokens associated with the traffic light, which introduces severe safety hazards. In contrast, our method allocates more tokens to regions related to obstacles and lane areas under a limited token budget, which improves inference efficiency while supporting safer driving decisions.

5 Conclusion

In this work, we propose DiP, a novel visual token pruning method tailored for auto-regressive VLA driving models. We propose a spatial divide-and-conquer strategy that divides the space into near and far fields with depth information as explicit spatial-cognitive priors. To ensure better obstacle avoidance, we retain the near tokens based on the edge energy and the information capacity. For far-field tokens, DiP improves scene understanding by maximizing token diversity while using depth information to modulate the token selection process. Experiments on the nuScenes benchmark validate the effectiveness of our method. Overall, Our DiP serves as a domain-specific token pruning strategy for autonomous driving and also provides new insights into visual token pruning.

AI use statement

In this work, we used generative AI tools for assisting in writing proofs, interpreting results, providing feedback on research methodology or experiments, implementing methods and translating text. We did not use generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, or support qualitative and thematic data analysis. Generating synthetic data sets and cleaning or reformating datasets were not applicable to this work. Additionally, we used generative AI tools to edit software code, draft parts of the paper, and edit the research paper to improve readability. We have reviewed all AI-assisted work. The LLM-generated code was verified and tested for correctness by two authors. We manually checked the rationality and readability of the LLM-generated content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

This paper proposes a novel token pruning method for vision-language-action (VLA) models to improve their inference efficiency. The foundational models and datasets used in this work originate from open-source and widely cited resources, ensuring that this aspect does not impede experimental reproducibility. To further guarantee reproducibility, we provide comprehensive implementation details in Section 4.1 of the main text and in Appendix B.

References

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9392–9401, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *Proceedings of the International Conference on Learning Representations*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

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.

Jiajun Cao, Qizhe Zhang, Peidong Jia, Xuhui Zhao, Bo Lan, Xiaoan Zhang, Xiaobao Wei, Sixiang Chen, Liyun Li, Xianming Liu, et al. Fastdrivevla: Efficient end-to-end driving via plug-and-play reconstruction-based token pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 4, pp. 2571–2579, 2026.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, Leichen Wang, Xingtao Hu, Hao Sun, Hang Zhao, and Hao Zhao. Impromptu vla: Open weights and open data for driving vision-language-action models. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 38, 2025.

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 4693–4700, 2018.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24823–24834, 2025.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):32, 2024.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

Zhijian Huang, Chengjie Feng, Feng Yan, Baihui Xiao, Zequn Jie, Yujie Zhong, Xiaodan Liang, and Lin Ma. DriveMM: All-in-one large multimodal model for autonomous driving. *arXiv preprint arXiv:2412.07689*, 2024.

Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7919–7929, 2023.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Proceedings of the Advances in Neural Information Processing Systems*, 37:819–844, 2024.

Lei Jiang, Zixun Zhang, Yuting Zeng, Chunzhao Xie, Tongxuan Liu, Zhen Li, Lechao Cheng, and Xiaohua Xu. Dcp: Dual-cue pruning for efficient large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21202–21215, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 2679–2713, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024b.

Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7577–7586, 2021.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7077–7087, 2021.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Proceedings of the Advances in neural information processing systems*, 30, 2017.

Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11993–12003, 2025.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. LMDrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15120–15130, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *Proceedings of the European Conference on Computer Vision*, pp. 256–274, 2024.

Irwin Sobel and Gary Feldman. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, 1968:271–272, 1968.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22442–22452, 2025.

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. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. LLM4Drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*, 2023.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20857–20867, 2025a.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond attention: Toward conditional diversity-based visual token pruning for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29662–29672, 2025b.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 74840–74857, 2025c.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, 2024.

Zewei Zhou, Tianhui Cai, Seth 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. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 38, pp. 27920–27956, 2025.

Prompt Design for Autonomous Driving

You are an autonomous driving agent. You have access to a front view camera image of a vehicle <image>. Your task is to do your best to **predict future waypoints for the vehicle over the next 3 timesteps**, given the vehicle's intent inferred from the images. Provided are the previous ego vehicle status recorded over the last 0.5 seconds (at 0.5-second intervals). This includes the x and y coordinates of the ego vehicle. Positive x means forward direction while positive y means leftwards. The data is presented in the format [x, y]:.(t-0.5s) [-2.78, -0.02], Acceleration: X 0.85, Y -1.56 m/s^2, Velocity: 5.35 m/s, Steering angle: -1.31 (positive: left turn, negative: right turn), (t-0.0s) [0.0, 0.0], Acceleration: X 0.27, Y -1.04 m/s^2, Velocity: 5.53 m/s, Steering angle: -0.63 (positive: left turn, negative: right turn)

Figure 6: Prompt design for driving VLA models.

APPENDIX

A PROMPTING DETAILS

Figure 6 illustrates the prompt design tailored for the autonomous driving VLA model. We adopt the prompt configuration established by Impromptu-VLA Chi et al. (2025), which primarily consists of a role-playing component as a driving agent, task instructions for trajectory prediction and the ego-vehicle status. Such a prompt design evidently lacks specific driving-related objects, which makes it hard to provide fine-grained text-visual guidance.

B ALGORITHMIC DETAILS OF OUR METHOD

To illustrate the proposed method more clearly, particularly the depth-modulated diverse token selection process, we provide pseudocode for the algorithm in Algorithm 1.

C DETAILS OF THE EVALUATION METRICS

L2 metric evaluates the 2D Euclidean distance between the predicted trajectory points of the model and the ground-truth trajectory across the future prediction horizon, such as the sampling points at 1 second, 2 seconds, and 3 seconds, and then averages the values.

Collision Rate measures the probability that the 2D bounding box of the ego vehicle spatially overlaps with the bounding boxes of other dynamic obstacles, such as surrounding vehicles and pedestrians. A sample is categorized as a collision if the ego vehicle intersects with any obstacle at least once at any future time step.

Intersection Rate computes the probability that the future planned trajectory predicted by the model exceeds the road boundaries. With the rasterized road boundaries from the high-definition map, the metric detects whether the predicted trajectory line or the vehicle polygon of the ego vehicle intersects with these rasterized road boundaries, aiming to evaluate the capability of the model to understand map topological boundaries.

Driving Score records the score for route i as $DS_i = RC_i \times IS_i$, where RC_i denotes the route completion rate and IS_i represents the infraction penalty factor. The factor IS_i starts at 1.0 and

Algorithm 1 Proposed Depth-modulated Diverse Token Selection

Require: Far-field tokens $\mathbf{F}_{far}$; valid depth index set $\mathcal{V}$; token depth values $\mathbf{D}$; far token allocation budget k_2; hyperparameter σ
Ensure: Selected far token set $\mathcal{T}_{far}$
Initialize the sampled index set with the cosine similarity matrix $\mathbf{S}$:
$i_1 \leftarrow \arg\max_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} (1 - s_{i,j})$
$\mathcal{I}_{far} \leftarrow \{i_1\}$
Compute the linearly rescaled cosine similarity matrix:
$\hat{s}_{i,j} \leftarrow \frac{s_{i,j}+1}{2}, \quad \forall i, j \in \mathcal{V}$
Compute the depth-modulated diversity matrix:
$w_{i,j} \leftarrow \exp\left(-\frac{(d_i - d_j)^2}{2\sigma^2}\right), \quad \forall i, j \in \mathcal{V}$
$\delta_{i,j} \leftarrow 1 - \hat{s}_{i,j} w_{i,j}, \quad \forall i, j \in \mathcal{V}$
Initialize minimum distance array:
$\mathbf{A}_i \leftarrow \infty, \quad \forall i \in \mathcal{V}$
$c \leftarrow i_1$
while $|\mathcal{I}_{far}| < k_2$ **and** $|\mathcal{I}_{far}| < |\mathcal{V}|$ **do**
  Update the minimum depth-modulated diversity:
  $\mathbf{A} \leftarrow \min(\mathbf{A}, \boldsymbol{\delta}_{:,c})$
  Select the furthest candidate token:
  $i^\star \leftarrow \arg\max_{i \in \mathcal{V} \setminus \mathcal{I}_{far}} \mathbf{A}_i$
  $\mathcal{I}_{far} \leftarrow \mathcal{I}_{far} \cup \{i^\star\}$
  $c \leftarrow i^\star$
end while
if $|\mathcal{I}_{far}| < k_2$ **then**
  Obtain zero-depth token index set:
  $\mathcal{V}' \leftarrow \{1, 2, \ldots, N\} \setminus \mathcal{V}$
  Randomly sample $k_2 - |\mathcal{I}_{far}|$ indices without replacement from $\mathcal{V}'$ and append them to $\mathcal{I}_{far}$
end if
$\mathcal{T}_{far} \leftarrow \{f_i \mid i \in \mathcal{I}_{far}\}$
return $\mathcal{T}_{far}$

is multiplied by a penalty coefficient p_j for each occurrence of infraction j. Multiple occurrences of the same infraction exponentiate p_j. Primary penalties include colliding with pedestrians (0.50), vehicles (0.60), or static objects (0.65), running red lights (0.70), ignoring stop signs (0.80), failing to yield to emergency vehicles (0.70), and failing to maintain minimum speed (up to 0.70). Off-road driving proportionally reduces route completion. A route terminates immediately upon deviating over 30 meters, remaining blocked for over 180 simulated seconds, encountering a communication timeout exceeding 60 seconds, or exceeding the overall time limit.

Success Rate represents the percentage of routes completed without any infractions relative to the total number of routes. Consequently, a route is deemed successful only when the vehicle reaches the destination without committing any infractions, such as collisions, running red lights or ignoring stop signs.

Efficiency is defined as the ratio of the ego vehicle speed to the speed of surrounding traffic participants, which measures whether the model completes driving tasks at higher speeds. A higher ratio indicates that the ego vehicle travels faster relative to surrounding vehicles, thereby increasing the difficulty of the driving task. This metric prevents the model from reducing task difficulty by driving at extremely low speeds.

Comfortness evaluates the ego vehicle across the route using jerk magnitude, lateral acceleration, longitudinal acceleration, yaw angular acceleration, longitudinal jerk, and yaw rate. A driving execution achieves comfort success if route-wide means of these measurements fall within specified thresholds: maximum absolute jerk magnitude of 8.37 m/s^3, maximum absolute lateral acceleration of 4.89 m/s^2, longitudinal acceleration between -4.05 m/s^2 and 2.40 m/s^2, maximum absolute yaw angular acceleration of 1.93 rad/s^2, maximum absolute longitudinal jerk of 4.13 m/s^3, and maximum absolute yaw rate of 0.95 rad/s. Higher scores reflect smoother and more comfortable driving.

D Details of the Comparison Baselines

SparseVLM Zhang et al. (2025c) designs a text-guided visual token pruning framework in order to accelerate the VLM inference. It utilizes the cross-attention to isolate visually relevant text tokens and further designates them as raters. With the self-attention logits between these raters and the visual sequence, the model accurately computes a significance score for each visual patch. This text-aware guidance ensures the adaptive pruning of tokens irrelevant to the specific linguistic query.

VisPruner Zhang et al. (2025a) proposes a token pruning framework guided by visual attention cues. To be specific, the framework initially extracts a subset of highly informative tokens with the attention scores generated within the visual encoder. To ensure comprehensive scene representation, it also supplements the selection with background-preserving tokens by eliminating redundant tokens which exhibits high cosine similarity within the remaining token set.

DivPrune Alvar et al. (2025) formulates visual token pruning as a max-min diversity problem to systematically reduce redundancy in the LLMs. It iteratively selects a subset of visual tokens by maximizing the minimum cosine distance between them. DivPrune aims to ensure the retained tokens are highly diverse, maintaining robust performance under extreme pruning ratios without requiring fine-tuning or calibration.

FastDriveVLA Cao et al. (2026) proposes the ReconPruner, a parametric visual token pruner tailored for end-to-end autonomous driving VLA models. Being trained via masked autoencoder-style pixel reconstruction, it dynamically assigns saliency scores to emphasize predefined foreground objects crucial for driving planning. To prevent degenerate uniform scoring across spatial regions, an adversarial foreground-background reconstruction strategy is proposed, which facilitates highly discriminative token selection during the inference.

DriveAdapter Jia et al. (2023) adopts a decoupled architecture separating perception from planning. In this framework, the student model learns bird's-eye-view (BEV) perception representations solely from raw sensor data to generate privileged inputs required by the teacher, while the frozen teacher model focuses exclusively on trajectory planning. To address distribution bias between models, DriveAdapter integrates learnable adapter modules into each hierarchical level of the teacher and applies layer-wise supervision via feature alignment losses.

Orion Fu et al. (2025) first employs QT-Former to extract current scene features from multi-view images, aggregating long-term temporal context using historical queries alongside a memory bank. Scene tokens, historical tokens, and user instructions are fed into a large language model for a planning token. Finally, a VAE-based generative planner aligns the language model's semantic reasoning space with the trajectory action space in latent space, producing multimodal trajectories conditioned on the planning token.

AutoVLA Zhou et al. (2025) utilizes supervised fine-tuning on mixed trajectory and chain-of-thought data distilled from large vision-language models, endowing the model with fast thinking mode and slow thinking mode. It then applies group relative policy optimization for reinforcement fine-tuning, guided by a reward function combining driving performance metrics. Taking multi-view multi-frame images, navigation instructions, and ego-vehicle states as input, the model directly predicts reasoning text alongside action tokens.