FastDriveVLA: Efficient End-to-End Driving via Plug-and-Play Reconstruction-based Token Pruning

Jiajun Cao[1,2], Qizhe Zhang[1], Peidong Jia[1], Xuhui Zhao[1,2], Bo Lan[1,2], Xiaoan Zhang[1,2], Lizhuo[2], Xiaobao Wei[1], Sixiang Chen[1], Liyun Li[2], Xianming Liu[2], Ming Lu[1], Yang Wang[2] ‡, Shanghang Zhang[1] ‡

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2]XPeng Motors

Abstract

Vision-Language-Action (VLA) models have demonstrated significant potential in complex scene understanding and action reasoning, leading to their increasing adoption in end-to-end autonomous driving systems. However, the long visual tokens of VLA models greatly increase computational costs. Current visual token pruning methods in Vision-Language Models (VLM) rely on either visual token similarity or visual-text attention, but both have shown poor performance in autonomous driving scenarios. Given that human drivers concentrate on relevant foreground areas while driving, we assert that retaining visual tokens containing this foreground information is essential for effective decision-making. Inspired by this, we propose **FastDriveVLA**, a novel reconstruction-based vision token pruning framework designed specifically for autonomous driving. FastDriveVLA includes a plug-and-play visual token pruner called ReconPruner, which prioritizes foreground information through MAE-style pixel reconstruction. A novel adversarial foreground-background reconstruction strategy is designed to train ReconPruner for the visual encoder of VLA models. Once trained, ReconPruner can be seamlessly applied to different VLA models with the same visual encoder without retraining. To train ReconPruner, we also introduce a large-scale dataset called nuScenes-FG, consisting of 241K image-mask pairs with annotated foreground regions. Our approach achieves SOTA results on the nuScenes open-loop planning benchmark across different pruning ratios.

Introduction

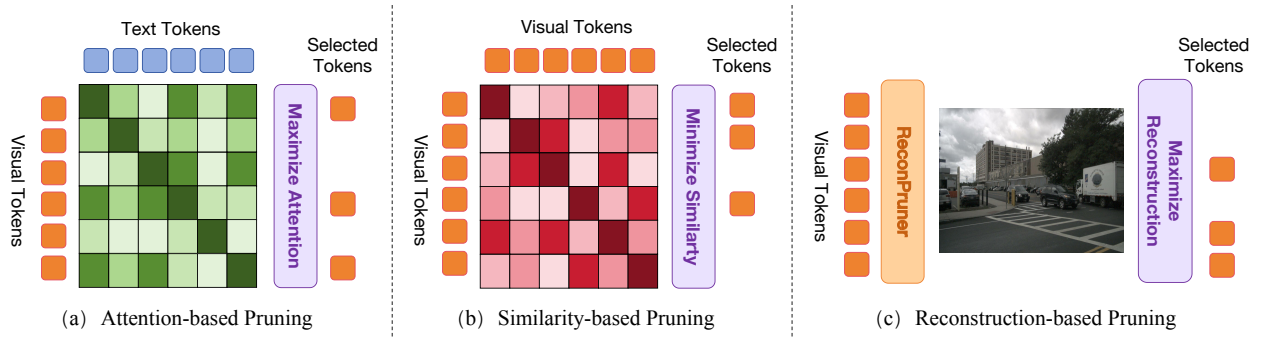
End-to-end autonomous driving (Hu et al. 2023; Prakash, Chitta, and Geiger 2021; Zhang et al. 2021; Jiang et al. 2023) has recently shown remarkable potential, promising to revolutionize future transportation systems. Unlike traditional modular autonomous driving systems—which divide the task into distinct components such as perception (Lang et al. 2019; Li et al. 2024), prediction (Gu, Sun, and Zhao 2021; Liu et al. 2021), and planning (Caesar et al. 2021; Ettinger et al. 2021)—end-to-end approaches learn the entire driving pipeline within a unified framework. This design not only mitigates error propagation between modules but also enhances system simplicity.

‡ Corresponding authors.

Given the remarkable reasoning capabilities demonstrated by Vision-Language Models (VLMs) (Liu et al. 2023; Wang et al. 2024a; Cao et al. 2025) in visual question answering tasks, recent studies have explored their extension to embodied intelligence and autonomous driving by incorporating action generation capabilities. These models, referred to as Vision-Language-Action (VLA) models (Tian et al. 2024; Sima et al. 2024; Wang et al. 2025; Li et al. 2025b), are increasingly adopted in end-to-end autonomous driving systems and have demonstrated superior performance over traditional modular approaches. However, existing visual language models (VLMs) usually convert visual inputs into numerous visual tokens. This approach has also been adopted by visual language attention (VLA) models, leading to considerable computational overhead and increased inference latency. This presents a significant challenge for deploying vehicles in real-world scenarios, where both computational resources and inference speed are severely limited.

Numerous efforts have been made to accelerate VLM inference by visual token reduction. Some approaches introduce newly designed multimodal projectors to compress visual tokens (Cha et al. 2024; Li et al. 2025a; Hu et al. 2024; Cai et al. 2024; Zhang et al. 2025b), but these methods require retraining the entire model, making them computationally expensive. Other approaches attempt to remove redundant visual tokens in a plug-and-play manner (Shang et al. 2024; Yang et al. 2025b,a; Dhouib et al. 2025), which can be broadly categorized into attention-based (Chen et al. 2024; Zhang et al. 2024b; Zhao et al. 2025) and similarity-based (Zhang et al. 2024a; Alvar et al. 2025; Zhang et al. 2025a) pruning strategies. Attention-based methods depend significantly on precise text-vision alignment and are particularly vulnerable to irrelevant information in the visual tokens. This issue is further exacerbated in autonomous driving scenarios, where text inputs are typically fixed and concise, offering limited guidance for effective token selection. While similarity-based methods are also ill-suited for autonomous driving, where visual inputs often contain well-defined foreground regions, such as lanes, pedestrians, and vehicles. In such cases, emphasizing token similarity becomes less meaningful, and similarity-based pruning may mistakenly retain background tokens irrelevant to driving.

To address these challenges, we propose **FastDriveVLA**, a novel reconstruction-based vision token pruning frame-



(a) Attention-based Pruning (b) Similarity-based Pruning (c) Reconstruction-based Pruning

Figure 1: Comparison of different visual token pruning strategies.

work tailored for end-to-end autonomous driving VLA models. Fig. 1 illustrates the differences between our visual token pruning strategy and existing methods. Motivated by the observation that human drivers primarily attend to foreground regions—while background areas have minimal influence on driving decisions—we argue that visual tokens encoding foreground information are significantly more valuable for autonomous driving. In contrast, tokens associated with background content are largely redundant. To implement this insight, we propose a plug-and-play visual token pruner named ReconPruner. ReconPruner is trained via MAE-style pixel reconstruction, encouraging it to focus on foreground regions and assign higher saliency scores to visual tokens containing critical foreground information. To prevent the pruner from assigning high saliency scores to all visual tokens, we introduce an adversarial foreground-background reconstruction strategy. This mechanism helps ReconPruner avoid local optima by enforcing discriminative attention between foreground and background areas. During inference, ReconPruner can be seamlessly integrated into various autonomous driving VLA models that share the same vision encoder, without requiring retraining.

To facilitate the training of ReconPruner, we further introduce a large-scale dataset named nuScenes-FG. To construct this dataset, we first define the concept of foreground in autonomous driving scenes, and then leverage Grounded-SAM (Ren et al. 2024) to segment the nuScenes (Caesar et al. 2020) dataset accordingly. This large-scale dataset contains 241k image-mask pairs across six camera views, with segmentation annotations of foreground regions.

Our contributions can be summarized as follows:

- We propose FastDriveVLA, a novel reconstruction-based token pruning framework, which differs from existing attention-based and similarity-based pruning methods.
- We design ReconPruner, a plug-and-play pruner trained via MAE-style pixel reconstruction, and introduce a novel adversarial foreground-background reconstruction strategy to enhance its ability to identify valuable tokens.
- We construct the nuScenes-FG dataset with foreground segmentation annotations for autonomous driving scenarios, comprising a total of 241k image–mask pairs.
- Our method is tailored for end-to-end autonomous driving VLA models and achieves SOTA performance on the nuScenes open-loop planning benchmark.

Related work

End-to-End Autonomous Driving Research in autonomous driving has seen a notable evolution from conventional modular pipelines, which decompose the task into perception, prediction, and planning, towards unified end-to-end learning frameworks. Seminal work like PilotNet demonstrated the feasibility of directly mapping raw pixel inputs to vehicle control commands using deep neural networks (Bojarski et al. 2016). While early behavioral cloning methods demonstrated the promise of end-to-end driving, they suffered from critical issues such as causal confusion and covariate shift. A primary research thrust to mitigate these limitations involved injecting explicit guidance into the learning process; for example, Conditional Imitation Learning (CIL) (Codevilla et al. 2018) incorporated high-level navigational commands to regularize the driving policy. Concurrently, another line of work focused on enhancing model robustness through architectural innovation, with approaches like TransFuser (Prakash, Chitta, and Geiger 2021) leveraging Transformer architectures to effectively fuse multi-modal sensor data. More recently, works such as SOLVE (Wen et al. 2024) and OpenDriveVLA (Zeng et al. 2024) proposed to synergize the direct action generation of end-to-end networks with the power of large vision-language-action architectures to improve both interpretability and performance in complex scenarios.

Driving Vision-Language-Action Models Recently, the integration of large language models (LLMs) has given rise to Vision-Language-Action (VLA) models, which are setting a new frontier in autonomous driving. These models aim to enhance the vehicle's reasoning capabilities, interpretability, and ability to handle long-tail scenarios by grounding driving actions in natural language. DriveGPT4 (Xu et al. 2024) showcased how LLMs can be adapted for motion planning and vehicle control. Building upon this trend, OpenDriveVLA (Zeng et al. 2024) and Impromptu VLA (Sha et al. 2024) are significant contributions that focus on developing open-source, large-scale VLAs specifically for driving. They demonstrate how to train powerful models that can process complex visual scenes and generate fine-grained control actions. The development of such data-hungry models is critically dependent on comprehensive datasets. OmniDrive (Wang et al. 2024b) provides a holistic, vision-language dataset featuring rich annotations

Figure 2: **nuScenes-FG.** It contains 241k foreground segmentation annotations for scenes in the nuScenes dataset.

and counterfactual reasoning scenarios.

Visual Token Pruning Existing VLMs convert visual inputs into a large number of tokens, leading to significant computational overhead and inference latency. Many studies have explored visual token pruning as a plug-and-play approach to improve the inference efficiency (Shang et al. 2024; Yang et al. 2025b; Zhang et al. 2025a; Chen et al. 2025; Ma et al. 2025), which can be broadly categorized based on their pruning criteria. The first category, attention-based methods, such as FastV (Chen et al. 2024) and SparseVLM (Zhang et al. 2024b), assesses the importance of visual tokens using attention scores from text tokens, which heavily rely on the correlation between user instructions and input images. However, in the driving tasks, where instructions are typically fixed and concise, this correlation is insufficient to guide effective token selection. The second category, similarity-based methods, like VisPruner (Zhang et al. 2024a) and DivPrune (Alvar et al. 2025), removes redundancy by selecting a diverse subset of visual tokens. Nevertheless, in the driving scenarios, input images often contain well-defined foreground regions, and excessive retention of background tokens irrelevant to the driving task can degrade performance under constrained computational budgets.

Methodology

nuScenes-FG Dataset

Inspired by human driving behavior, we first define the foreground regions in autonomous driving scenarios as areas that include humans, roads, vehicles, traffic signs (including traffic lights), and traffic barriers (such as obstacles located on or adjacent to the roadway). In contrast, other regions—such as buildings, the sky, and roadside trees—have little to no impact on human driving decisions, even when they are completely occluded.

The nuScenes (Caesar et al. 2020) dataset includes 3D bounding box annotations for humans and vehicles, yet this representation inherently captures extraneous background elements due to the coarse nature of axis-aligned bounding volumes. Although a subsequent map expansion package with 11 semantic layers is available, these annotations still fail to comprehensively cover all relevant regions. To address this, we employ Grounded-SAM (Ren et al. 2024) to generate consistent and fine-grained foreground segmentation annotations across nuScenes scenes. The resulting nuScenes-FG dataset comprises 241k image–mask pairs from six camera views, with examples shown in Fig. 2.

ReconPruner: Reconstruction-based Pruner

We propose a novel and lightweight plug-and-play pruner named ReconPruner, which is trained via a pixel-level reconstruction. The architecture of ReconPruner consists of a PrunerLayer and a Scorer, as illustrated in Fig. 3. The PrunerLayer is implemented as a single decoder layer of Qwen2.5-VL-3B (Bai et al. 2025). The Scorer is implemented as a single-layer feedforward network with a weight shape of $\mathbb{R}^{D\times 1}$, where D denotes the hidden state dimension. Overall, ReconPruner is highly lightweight, with a total size of only 0.07B parameters.

During training and inference, we introduce a learnable query token $Q \in \mathbb{R}^{1\times D}$ to capture the saliency of the visual tokens in the foreground. The query token Q and the visual tokens $V \in \mathbb{R}^{N\times D}$ are jointly fed into the PrunerLayer, producing $Q^* \in \mathbb{R}^{1\times D}$ and $V^* \in \mathbb{R}^{N\times D}$, where N denotes the number of visual tokens. The process is as follows:

$$[Q^*, V^*] = PrunerLayer([Q, V]), \tag{1}$$

The fused tokens are obtained by computing the Hadamard product between V^* and Q^*, which are subsequently fed into the Scorer to assign saliency scores $S \in \mathbb{R}^{N\times 1}$ to visual tokens, as computed below:

$$S = Scorer(V^* \odot Q^*). \tag{2}$$

Since our primary objective is to enable ReconPruner to effectively identify and select visual tokens that contain meaningful foreground information, we draw inspiration from prior masked image modeling (MIM) approaches (He et al. 2022; Xie et al. 2022) and design a MAE-style pixel reconstruction strategy. During training, we select the subset of visual tokens with the highest saliency scores as predicted by ReconPruner and use them for masked foreground reconstruction. The reconstruction loss computed on this subset serves as a supervisory signal, encouraging ReconPruner to assign higher saliency scores to visual tokens that genuinely correspond to foreground content.

Adversarial Foreground-Background Reconstruction Strategy

However, relying solely on foreground reconstruction can lead to a degenerate solution where ReconPruner takes a shortcut by indiscriminately assigning high saliency scores to all visual tokens, thus boosting the reconstruction performance. To address this issue, we draw inspiration from Generative Adversarial Networks (GANs) (Goodfellow et al.

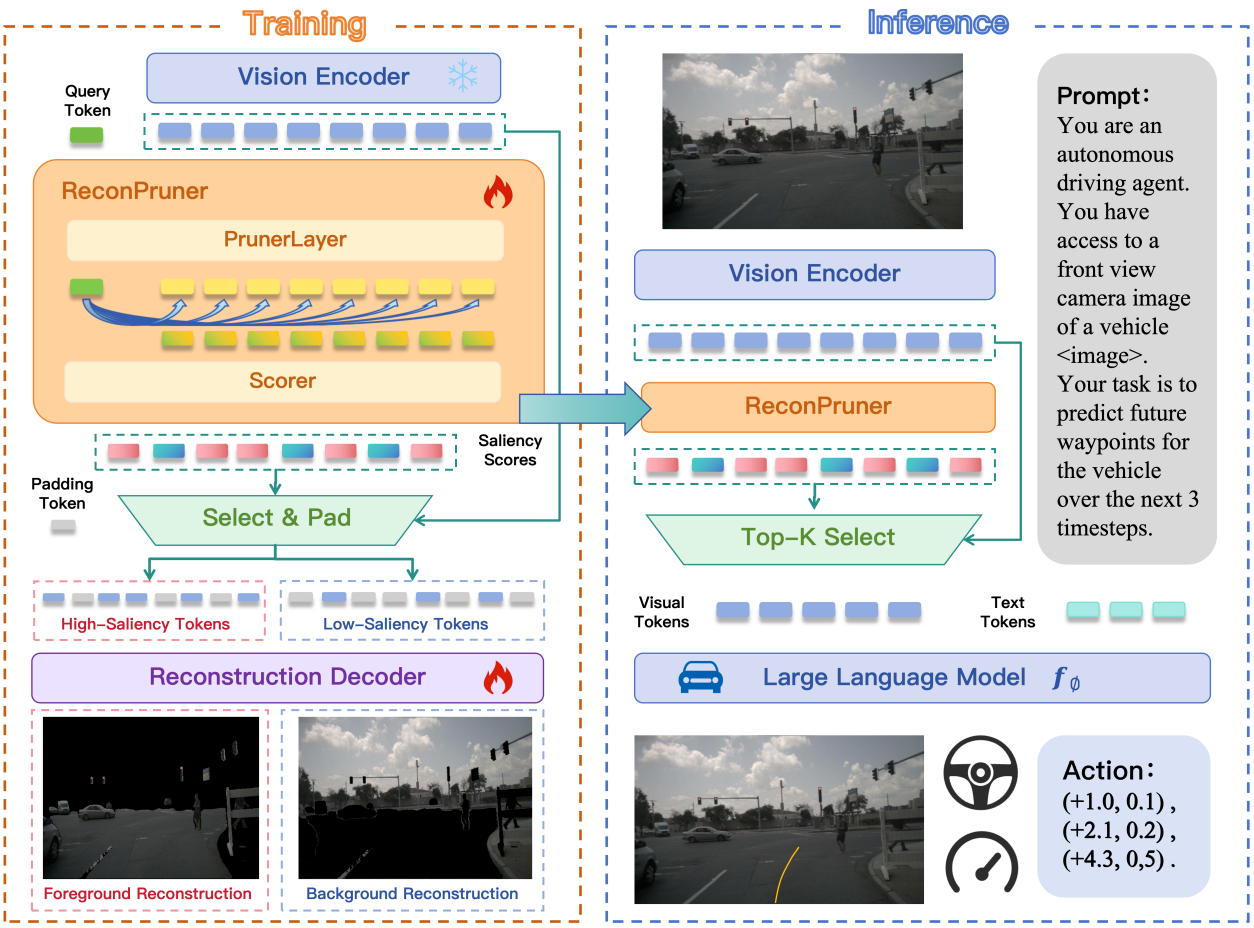


Figure 3: **FastDriveVLA** framework. During training, a novel adversarial foreground-background reconstruction strategy is proposed to enhance ReconPruner's ability to perceive foreground visual tokens. During inference, ReconPruner can be directly integrated into autonomous driving VLA models for token pruning.

2020) and propose an adversarial foreground-background reconstruction strategy. Specifically, ReconPruner is additionally required to reconstruct the background regions using the visual tokens that receive low saliency scores. By imposing this complementary background reconstruction constraint, the model is effectively discouraged from assigning uniformly high saliency scores, thereby promoting a more precise and discriminative scoring of visual tokens. This adversarial setup enhances ReconPruner's ability to differentiate foreground tokens from background ones, resulting in improved token selection performance.

The overall training strategy proceeds as follows:

We first generate a binary mask $M \in \{0,1\}^N$ based on the saliency scores S predicted by ReconPruner, where each element M_i is set to 1 if the corresponding saliency score $S_i > 0$, and 0 otherwise, as defined below:

$$M_i = \begin{cases} 1, & \text{if } S_i > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{for } i = 1, 2, \ldots, N, \tag{3}$$

where M_i and S_i denote the i-th element of M and S, respectively. However, since M is non-differentiable, directly masking visual tokens V with mask M would block the gradient flow during backpropagation. To address this issue, we adopt the Straight-Through Estimator (STE) (Bengio, Léonard, and Courville 2013) technique, which applies a discrete mask in the forward pass while using a continuous approximation in the backward pass to enable gradient propagation. This operation is defined as follows:

$$\tilde{M} = S + \text{stop\_grad}(M - S), \tag{4}$$

where $\tilde{M} \in \{0,1\}^N$ denotes the gradient-friendly approximation of the binary mask.

We then utilize the approximated mask $\tilde{M}$ to retain the high-saliency visual tokens and replace the low-saliency ones with padding tokens (typically zeros) to obtain the foreground visual tokens $V_{fore} \in \mathbb{R}^{N \times D}$. Similarly, we invert $\tilde{M}$ to obtain the background visual tokens $V_{back} \in \mathbb{R}^{N \times D}$. This process is formulated as follows:

$$V_{fore} = \tilde{M} \odot V, \quad V_{back} = (1 - \tilde{M}) \odot V. \tag{5}$$

The reconstruction decoder D consists of six Qwen2.5-VL-3B (Bai et al. 2025) decoder layers and a feedforward reconstruction head. We feed both V_{fore} and V_{back} into

the reconstruction decoder D to obtain the reconstructed foreground $I_{fore}^{pred} \in \mathbb{R}^{3\times H\times W}$ and background $I_{back}^{pred} \in \mathbb{R}^{3\times H\times W}$, which can be formulated as follows:

$$I_{fore}^{pred} = D(V_{fore}), \quad I_{back}^{pred} = D(V_{back}). \tag{6}$$

Training Loss

In order to leverage both pixel-level accuracy and perceptual consistency, we formulate the reconstruction loss as a weighted combination of the Mean Squared Error (MSE) and the Structural Similarity Index Measure (SSIM) loss (Wang et al. 2004), as defined below:

$$\begin{aligned}\mathcal{L}_{fore} &= \lambda \left(1 - \text{SSIM}(I_{fore}^{gt}, I_{fore}^{pred})\right) \\ &\quad + (1-\lambda)\,\text{MSE}(I_{fore}^{gt}, I_{fore}^{pred}), \\ \mathcal{L}_{back} &= \lambda \left(1 - \text{SSIM}(I_{back}^{gt}, I_{back}^{pred})\right) \\ &\quad + (1-\lambda)\,\text{MSE}(I_{back}^{gt}, I_{back}^{pred}),\end{aligned} \tag{7}$$

where I_{fore}^{gt} and I_{back}^{gt} denote the masked foreground and background images, respectively, and we set $\lambda = 0.2$.

The overall training loss is defined as follows:

$$\mathcal{L}_{all} = \alpha\mathcal{L}_{fore} + (1-\alpha)\mathcal{L}_{back}, \tag{8}$$

where we set $\alpha = 0.5$.

Pruning During Inference

During inference, ReconPruner assigns saliency scores S to a sequence of N visual tokens. Given a target pruning ratio $p \in [0, 1]$, we apply a Top-K selection strategy to retain the top $K = \lfloor N\cdot(1-p)\rfloor$ visual tokens with the highest saliency scores, which can be formulated as:

$$V_{\text{select}} = \{v_i \mid i \in \mathcal{I}\}, \quad \mathcal{I} = \text{TopK}(S, K). \tag{9}$$

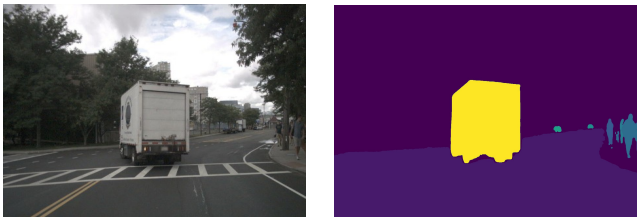
To ensure that the retained visual tokens preserve their original spatial semantics, we also retain their corresponding position embeddings. The selected visual tokens $V_{\text{select}} \in \mathbb{R}^{K\times D}$ and the text tokens $T \in \mathbb{R}^{L\times D}$ are then jointly fed into the large language model f_ϕ to predict the final action, which can be formulated as:

$$Action = f_\phi([V_{\text{select}}, T]). \tag{10}$$

Experiments

Experimental Settings

Models. We adopt Impromptu-VLA (Chi et al. 2025), the current state-of-the-art end-to-end VLA model for autonomous driving, as the base model for visual token pruning. It is built upon the Qwen2.5-VL (Bai et al. 2025) architecture. The encoder of Impromptu-VLA remains frozen during its original training process, making its parameters and architecture identical to those of Qwen2.5-VL. Since the reconstruction task is non-causal by nature, we replace the causal attention mechanism with full attention in both the ReconPruner and reconstruction decoder.



(a) Input Image (b) Foreground Segmentation

(c) Foreground Reconstruction (d) Background Reconstruction

Figure 4: Visualization of reconstruction.

Datasets and Metrics. We evaluate our method on the nuScenes (Caesar et al. 2020) dataset, a large-scale benchmark specifically designed for autonomous driving in urban environments. It consists of 1,000 driving scenes, each lasting approximately 20 seconds. For testing, we follow the official evaluation protocol of Impromptu-VLA and use a total of 6,019 test samples. Following prior work (Wang et al. 2025), we evaluate the performance of open-loop planning using three metrics: trajectory prediction **L2** error, **Collision Rate**, and **Intersection Rate** with the road boundary.

Baselines. For comparison, we adopt FastV (Chen et al. 2024) and SparseVLM (Zhang et al. 2024b) as attention-based baselines, and DivPrune (Alvar et al. 2025) and VisPruner (Zhang et al. 2024a) as similarity-based baselines.

Training. We train FastDriveVLA with a learning rate of 2e-5 using cosine scheduler. The training runs for a total of 10 epochs and takes only 3 hours on two H800 GPUs.

Evaluation on the nuScenes

We evaluate and compare our method against both attention-based (FastV & SparseVLM) and similarity-based (VisPruner & DivPrune) baselines on the open-loop nuScenes benchmark. The input image resolution is set to 1596×1596, resulting in a total of 3249 visual tokens. We consider three pruning ratios of visual tokens: 25%, 50%, and 75%. We avoid using more aggressive pruning ratios, as driving is a safety-critical task that prioritizes maintaining high model performance over maximizing computational efficiency, in contrast to general visual question answering tasks.

As shown in Tab. 1, when pruning 25% of the visual tokens, our method outperforms all baseline methods across all metrics. Notably, our approach even surpasses the original unpruned model in terms of L2 and Intersection metrics, with improvements of 0.1% and 1.0%, respectively. This encouraging result supports our hypothesis that focusing on foreground-relevant visual tokens is key to autonomous driving. When pruning 50% of the visual tokens, we observe

Methods	L2 (cm) ↓					Collision (%) ↓					Intersection (%) ↓				
	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.
Input size 1596 × 1596, *3249 tokens* (**100%**)															
Impromptu-VLA (NeurIPS25)	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%**)															
FastV (ECCV25)	14.23	28.85	53.80	32.29	98.6%	0.00	0.18	0.74	0.31	79.3%	0.52	2.44	5.65	2.87	97.4%
SparseVLM (ICML25)	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 (ICCV25)	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 (CVPR25)	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 (Ours)	**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%**
Retain 1624 Tokens (↓ **50%**)															
FastV (ECCV25)	14.29	29.14	54.33	32.59	97.7%	0.00	0.20	0.79	0.33	73.7%	0.52	2.67	5.77	2.99	93.6%
SparseVLM (ICML25)	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 (ICCV25)	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 (CVPR25)	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 (Ours)	**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%**
Retain 812 Tokens (↓ **75%**)															
FastV (ECCV25)	14.63	29.54	54.97	33.05	96.3%	0.00	0.21	0.79	0.33	73.0%	0.58	2.63	5.76	2.99	93.5%
SparseVLM (ICML25)	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 (ICCV25)	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 (CVPR25)	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 (Ours)	**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%**

Table 1: **Performance comparison of different pruning methods on Impromptu-VLA.** Input images are of resolution 1596×1596, resulting in 3249 visual tokens. Here, **Rel.** represents the average percentage of performance maintained, and the underlined values indicate improvements over the original unpruned model.

(a) Input Image (b) FastV

(c) Divprune (d) FastDriveVLA (Ours)

Figure 5: Visual comparison of visual tokens retained by different visual token pruning methods.

an interesting phenomenon: most methods exhibit a better Collision performance compared to the 25% pruning setting. Similarly, at a 75% pruning ratio, some methods even achieve higher Intersection performance than at 50%. However, this performance improvement with increasing pruning ratios is not observed under the L2 metric. We attribute this to the relatively small absolute values of Collision and Intersection metrics, making them more susceptible to noise.

Overall, our method consistently outperforms existing approaches across all pruning ratios. Notably, pruning 50% of the visual tokens achieves a more balanced performance across all metrics. Therefore, we recommend this pruning ratio for practical deployment in autonomous driving.

Ablation Study

As shown in Tab. 2, we separately investigate the contributions of pixel reconstruction and the adversarial foreground-background reconstruction strategy to our method. When we replace pixel reconstruction with foreground mask prediction, we observe performance degradation across all metrics. We attribute this to the fact that the mask prediction objective merely distinguishes between foreground and background regions, assigning equal importance to all tokens within the foreground. This fails to emphasize the more complex and critical objects. Moreover, when the adversarial foreground-background reconstruction strategy is removed and only pixel reconstruction is performed on the foreground region, pruning performance deteriorates significantly. This is because the ReconPruner lacks the ability to further distinguish between foreground and background content without adversarial supervision.

Pruning with Foreground Masks

To achieve reconstruction-based visual token pruning, a straightforward approach is to resize the foreground mask to match the spatial resolution of the visual tokens and prune the tokens at the corresponding positions. However, this approach encounters two major challenges: (1) the foreground

Pixel Reconstruction	AFBR Strategy	L2 (cm) ↓					Collision (%) ↓					Intersection (%) ↓				
		1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.
✓	✗	14.11	28.82	53.78	32.24	98.7%	0.00	0.18	0.70	0.29	83.0%	0.59	2.55	5.82	2.99	93.6%
✗	✓	14.14	28.76	53.66	32.19	98.9%	0.00	0.17	0.67	0.28	86.9%	0.58	2.59	5.84	3.00	93.1%
✓	✓	**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%**

Table 2: Ablation study on pixel reconstruction and adversarial foreground-background reconstruction (AFBR) strategy.

Methods	L2 (cm) ↓					Collision (%) ↓					Intersection (%) ↓				
	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.	1s	2s	3s	Avg.	Rel.
Gt-mask+Text-attn	**14.07**	28.71	53.70	32.16	99.0%	0.00	0.16	0.63	0.26	92.4%	**0.53**	2.50	5.82	2.95	94.8%
Text-attn	14.15	29.01	53.89	32.35	98.4%	0.00	0.19	0.72	0.30	80.2%	0.60	2.63	5.85	3.03	92.4%
FastDriveVLA (Ours)	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%**

Table 3: Comparison of visual token pruning with ground-truth foreground masks.

Methods	Token	FLOPs (T)	Prefill Time (ms/token)	Decode Time (ms/token)
Impromptu-VLA	3249	38.2	187	23
FastV	812	4.1 (×9.3)	49 (×3.8)	21 (×1.2)
SparseVLM	812	4.2 (×9.1)	55 (×3.4)	19 (×1.1)
VisPruner	812	**3.6 (×10.6)**	**43 (×4.3)**	**18 (×1.3)**
Divprune	812	**3.6 (×10.6)**	**43 (×4.3)**	**18 (×1.3)**
FastDriveVLA (Ours)	812	5.1 (×7.5)	51 (×3.7)	**18 (×1.3)**

Table 4: Efficiency analysis of different pruning methods.

mask provides only binary cues and lacks the capacity to quantify the saliency of individual visual tokens, making it unsuitable for ranking and pruning at arbitrary ratios; and (2) the spatial alignment between the foreground mask and visual tokens is often inaccurate — prior work (Darcet et al. 2023) has shown that the positions of visual tokens generated by vision encoders frequently exhibit spatial misalignment with the original image patches.

To compare with the pruning method based on foreground masks, we use text attention to estimate the saliency of visual tokens and prioritize those located within the foreground mask region. We also compare this with a baseline that prunes solely based on text attention. As shown in Tab. 3, we find that pruning guided by foreground masks achieves a clear performance improvement over text-attention-only pruning, indicating that foreground visual tokens are indeed more informative. However, our method remains more efficient, as it addresses the spatial misalignment issue of foreground visual tokens. Moreover, generating foreground masks using Grounded-SAM (Ren et al. 2024) typically takes around 3 seconds per image, which incurs a prohibitive time cost for practical deployment.

Efficiency Analysis

To demonstrate the efficiency of FastDriveVLA, we conduct a efficiency analysis against other pruning methods in terms of FLOPs and CUDA latency. As shown in Tab. 4, when the number of visual tokens is reduced from 3249 to 812, FastDriveVLA achieves nearly a 7.5× reduction in FLOPs. In terms of CUDA latency, FastDriveVLA reduces the prefill and decode time by 3.7× and 1.3×, respectively, significantly enhancing real-world inference efficiency. Although our method introduces a parameterized pruner, which results in slightly higher FLOPs compared to some non-parametric approaches, its lightweight design still achieves lower CUDA latency than some of them.

Qualitative Visualization

To validate the effectiveness of our reconstruction-based pruning method, we present qualitative visualizations of foreground and background reconstructions. As shown in Fig. 4, ReconPruner effectively preserves tokens related to foreground objects while distinguishing background regions, enabling high-quality reconstruction and demonstrating its ability to retain essential visual information with reduced token redundancy.

We further visualize the visual tokens selected by FastV (attention-based) and Divprune (similarity-based), alongside our method. As shown in Fig. 5, our approach better preserves the lane area and effectively attends to lane signs and vehicles. In contrast, FastV tends to overlook vehicles, while Divprune retains a greater number of more scattered tokens but demonstrates limited focus on the lane area.

Conclusion

We propose a novel reconstruction-based visual token pruning framework, FastDriveVLA, which is more suitable for autonomous driving tasks with clearly defined foregrounds compared to traditional attention-based and similarity-based pruning methods. We train the plug-and-play ReconPruner through MAE-style pixel reconstruction and enhance its foreground perception capability with a novel adversarial foreground-background reconstruction strategy. Additionally, we have constructed a large-scale autonomous driving scene dataset annotated with foreground segmentation masks, which can be widely utilized for future autonomous driving research. Overall, our work not only establishes a new paradigm for efficient visual token pruning in autonomous driving VLA models but also provides valuable insights into task-specific pruning strategies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62476011) and the Beijing Natural Science Foundation (L252060).

References

Alvar, S. R.; Singh, G.; Akbari, M.; and Zhang, Y. 2025. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9392–9401.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bengio, Y.; Léonard, N.; and Courville, A. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Nagy, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Caesar, H.; Kabzan, J.; Tan, K. S.; Fong, W. K.; Wolff, E.; Lang, A.; Fletcher, L.; Beijbom, O.; and Omari, S. 2021. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*.

Cai, M.; Yang, J.; Gao, J.; and Lee, Y. J. 2024. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*.

Cao, J.; Zhang, Y.; Huang, T.; Lu, M.; Zhang, Q.; An, R.; Ma, N.; and Zhang, S. 2025. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19846–19856.

Cha, J.; Kang, W.; Mun, J.; and Roh, B. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13817–13827.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35. Springer.

Chen, Y.; Bai, X.; Wang, Z.; Bai, C.; Dai, Y.; Lu, M.; and Zhang, S. 2025. StreamKV: Streaming Video Question-Answering with Segment-based KV Cache Retrieval and Compression. arXiv:2511.07278.

Chi, H.; Gao, H.-a.; Liu, Z.; Liu, J.; Liu, C.; Li, J.; Yang, K.; Yu, Y.; Wang, Z.; Li, W.; et al. 2025. Impromptu VLA: Open Weights and Open Data for Driving Vision-Language-Action Models. *arXiv preprint arXiv:2505.23757*.

Codevilla, F.; Müller, M.; López, A.; Koltun, V.; and Dosovitskiy, A. 2018. Imitation learning for autonomous driving in urban environments. In *Proceedings of the IEEE international conference on robotics and automation (ICRA)*.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.

Dhouib, M.; Buscaldi, D.; Vanier, S.; and Shabou, A. 2025. Pact: Pruning and clustering-based token reduction for faster visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14582–14592.

Ettinger, S.; Cheng, S.; Caine, B.; Liu, C.; Zhao, H.; Pradhan, S.; Chai, Y.; Sapp, B.; Qi, C. R.; Zhou, Y.; et al. 2021. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9710–9719.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gu, J.; Sun, C.; and Zhao, H. 2021. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15303–15312.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hu, W.; Dou, Z.-Y.; Li, L.; Kamath, A.; Peng, N.; and Chang, K.-W. 2024. Matryoshka query transformer for large vision-language models. *Advances in Neural Information Processing Systems*, 37: 50168–50188.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17853–17862.

Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8350.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Li, W.; Yuan, Y.; Liu, J.; Tang, D.; Wang, S.; Qin, J.; Zhu, J.; and Zhang, L. 2025a. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 1–19.

Li, Y.; Wei, X.; Chi, X.; Li, Y.; Zhao, Z.; Wang, H.; Ma, N.; Lu, M.; and Zhang, S. 2025b. Manipdreamer3d: Synthesizing plausible robotic manipulation video with occupancy-aware 3d trajectory. *arXiv preprint arXiv:2509.05314*.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Yu, Q.; and Dai, J. 2024. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Y.; Zhang, J.; Fang, L.; Jiang, Q.; and Zhou, B. 2021. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7577–7586.

Ma, J.; Zhang, Q.; Lu, M.; Wang, Z.; Zhou, Q.; Song, J.; and Zhang, S. 2025. MMG-Vid: Maximizing Marginal Gains at Segment-level and Token-level for Efficient Video LLMs. *arXiv preprint arXiv:2508.21044*.

Prakash, A.; Chitta, K.; and Geiger, A. 2021. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7077–7087.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Sha, H.; Zhuo, T. Y.; Richards, L. E.; Morin, S.; Karkus, P.; Adrien Taylor, A. T. B. W. W.; Bewley, A.; Shotton, J.; and Kanazawa, A. 2024. Impromptu VLA: Open Weights and Open Data for Driving Vision-Language-Action Models. arXiv:2405.11623.

Shang, Y.; Cai, M.; Xu, B.; Lee, Y. J.; and Yan, Y. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, 256–274. Springer.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; and Alvarez, J. M. 2025. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22442–22452.

Wang, Y.; cheng, Z.; Yao, Y.; and zhang, W. 2024b. OmniDrive: A Holistic Vision-Language Dataset for Autonomous Driving with Counterfactual Reasoning. arXiv:2405.00835.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wen, L.; chen, Y.; Li, Y.; Zhang, L.; Cui, C.; Zong, G.; Liu, J.; Li, R.; Yue, X.; and Qiao, Y. 2024. SOLVE: Synergy of Language-Vision and End-to-End Networks for Autonomous Driving. arXiv:2405.02213.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9653–9663.

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Zhang, L.; Qiao, Y.; Li, Z.; and Liu, S. 2024. DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model. arXiv:2310.05739.

Yang, C.; Sui, Y.; Xiao, J.; Huang, L.; Gong, Y.; Li, C.; Yan, J.; Bai, Y.; Sadayappan, P.; Hu, X.; et al. 2025a. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19803–19813.

Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2025b. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19792–19802.

Zeng, W.; Wong, K.; Sun, J.; Tang, B.; Huer, F.; Girish, B. J.; Lu, F.; Chen, Y.; Xu, R.; Li, H.; Qiao, Y.; and Liu, S. 2024. OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model. arXiv:2405.12212.

Zhang, Q.; Cheng, A.; Lu, M.; Zhang, R.; Zhuo, Z.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2024a. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. *arXiv preprint arXiv:2412.01818*.

Zhang, Q.; Liu, M.; Li, L.; Lu, M.; Zhang, Y.; Pan, J.; She, Q.; and Zhang, S. 2025a. Beyond Attention or Similarity: Maximizing Conditional Diversity for Token Pruning in MLLMs. *arXiv preprint arXiv:2506.10967*.

Zhang, S.; Fang, Q.; Yang, Z.; and Feng, Y. 2025b. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.

Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; et al. 2024b. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

Zhang, Z.; Liniger, A.; Dai, D.; Yu, F.; and Van Gool, L. 2021. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15222–15232.

Zhao, S.; Wang, Z.; Juefei-Xu, F.; Xia, X.; Liu, M.; Wang, X.; Liang, M.; Zhang, N.; Metaxas, D. N.; and Yu, L. 2025. Accelerating Multimodal Large Language Models by Searching Optimal Vision Token Reduction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29869–29879.