VLDrive: Vision-Augmented Lightweight MLLMs for Efficient Language-grounded Autonomous Driving

Ruifei Zhang[1,2], Wei Zhang[4], Xiao Tan[4], Sibei Yang[3], Xiang Wan[2], Xiaonan Luo[5], Guanbin Li[3,6*]
[1] The Chinese University of Hong Kong, Shenzhen [2] Shenzhen Research Institute of Big Data
[3] Sun Yat-sen University [4] Baidu Inc. [5] Guilin University of Electronic Technology
[6] Guangdong Key Laboratory of Big Data Analysis and Processing

Abstract

*Recent advancements in language-grounded autonomous driving have been significantly promoted by the sophisticated cognition and reasoning capabilities of large language models (LLMs). However, current LLM-based approaches encounter critical challenges: (1) Failure analysis reveals that frequent collisions and obstructions, stemming from limitations in visual representations, remain primary obstacles to robust driving performance. (2) The substantial parameters of LLMs pose considerable deployment hurdles. To address these limitations, we introduce VLDrive, a novel approach featuring a lightweight MLLM architecture with enhanced vision components. VLDrive achieves compact visual tokens through innovative strategies, including cycle-consistent dynamic visual pruning and memory-enhanced feature aggregation. Furthermore, we propose a distance-decoupled instruction attention mechanism to improve joint visual-linguistic feature learning, particularly for long-range visual tokens. Extensive experiments conducted in the CARLA simulator demonstrate VLDrive's effectiveness. Notably, VLDrive achieves state-of-the-art driving performance while reducing parameters by 81% (from 7B to 1.3B), yielding substantial driving score improvements of **15.4**%, **16.8**%, and **7.6**% at tiny, short, and long distances, respectively, in closed-loop evaluations. Code is available at* https://github.com/ReaFly/VLDrive.

1. Introduction

Autonomous driving has experienced significant development in recent years, evolving from modular and segregated design to integrated end-to-end networks [10, 15, 26, 29, 46]. However, these networks still rely on specific inputs such as target points or action commands to guide their driving behaviors, significantly constraining the interaction with humans and applicability in real-world scenarios.

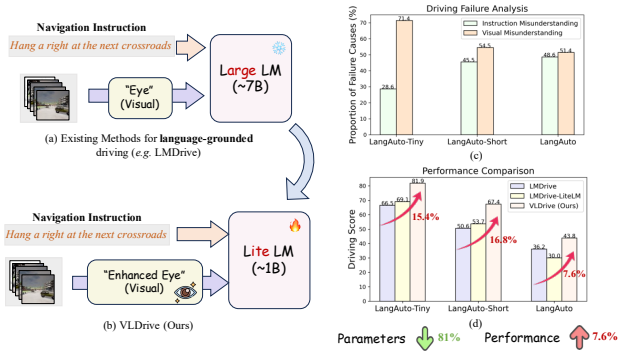

Figure 1. (a) Existing methods for language-grounded driving. (b) Our proposed VLDrive: a novel framework featuring a lightweight MLLM architecture with enhanced vision components. (c) Driving failure analysis of existing methods based on three evaluation runs. (d) Performance comparison between VLDrive and both versions of LMDrive, highlighting our method's superior driving performance with fewer parameters.

The emergence of large language models (LLMs) [18, 33, 49, 50] has catalyzed a revolution in autonomous driving. Impressed by their advanced cognition and logical reasoning capabilities, a multitude of studies have integrated LLMs into autonomous driving systems and achieved noteworthy results [28, 37, 38, 41]. Among them, LMDrive [28] successfully achieves human-friendly **language-grounded** autonomous driving, where vehicle behaviors are **solely** guided by natural language instructions (see Fig. 1(a)), significantly enhancing human-vehicle interaction.

However, despite remarkable advancements, these approaches encounter significant practical deployment challenges, characterized by suboptimal driving performance and excessive computational parameters. To address these issues, we conduct a thorough investigation of driving failure cases in current approaches, coupled with performance analysis across different LLM scales, yielding two critical insights: 1) ***Deficiencies in visual comprehension***

*Corresponding author: liguanbin@mail.sysu.edu.cn

are the predominant cause of driving failures. We analyze the failure cases of existing language-grounded autonomous driving models and identify two primary causes: **instruction misunderstanding**, manifested as the agent deviating from the route; and **visual misunderstanding**, characterized by failures due to collisions or blocks involving vehicles or layouts. The results in Fig. 1(c) reveal that limitations in visual comprehension frequently emerge as a critical factor contributing to driving task failures. 2) ***Driving performance does not scale linearly with LLM's parameter count***. Through comprehensive evaluations across three benchmarks, we compare the full-scale LMDrive (7B parameters) with its lightweight counterpart, LMDrive-LiteLM (1.3B trainable parameters). As demonstrated in Fig. 1(d), LMDrive-LiteLM achieves comparable driving performance despite its significantly reduced parameter count. These findings strongly suggest that massive parameter scales in LLMs may not be a critical prerequisite for achieving superior driving performance. The insights derived from these analyses illuminate the path toward developing a parameter-efficient yet high-performing alternative for language-grounded driving applications.

Building on these considerations, in our work, we propose **VLDrive**, a vision-augmented lightweight language model for efficient language-grounded autonomous driving, as showcased in Fig. 1(b). We enhance the model's visual comprehension by drawing inspiration from human driving competencies, focusing on two key aspects: prioritized attention allocation to sparse but critical agents and objects, and temporal trajectory memory for behavioral sequence prediction. These enhancements are achieved through our proposed Cycle-consistent Dynamic Visual Pruning (CCDP) and Memory-enhanced Feature Aggregation (MEFA) mechanisms. Specifically, CCDP framework consists of token sparsification and training-only token reconstruction. The sparsification mechanism enables adaptive selection of critical visual tokens and dynamic adjustment of token count based on driving scenario complexity. The reconstruction process preserves information integrity by explicitly recovering features of pruned tokens, enabling robust visual perception while minimizing computational overhead. Additionally, MEFA incorporates a memory bank of adjacent frames to capture temporal dependencies, providing critical temporal cues for visual feature enhancement.

Furthermore, we introduce a Distance-decoupled Instruction Attention (DDIA) mechanism to enhance instruction comprehension capabilities of lightweight language models. This design stems from our observation that increasing numbers of long-range visual tokens may dilute instruction attention (see Fig. 6), potentially leading to **instruction misunderstanding**. DDIA alleviates this issue and enhances joint visual-linguistic feature learning and alignment, resulting in more accurate trajectory planning. Extensive closed-loop experiments conducted on the CARLA [6] simulation platform demonstrate the effectiveness of our proposed method, surpassing state-of-the-art approaches by **15.4**%, **16.8**% and **7.6**% in driving scores at tiny, short and long distances, respectively.

To summarize, our contributions are as follows:

- We identify that visual comprehension deficiencies constitute the primary bottleneck in language-grounded driving, with performance scaling non-linearly with language model size. This observation motivates our proposed paradigm: a lightweight language model enhanced by vision-augmented strategies, which simultaneously reduces computational overhead and improves driving performance, facilitating practical deployment.
- We propose VLDrive, a lightweight architecture enhanced with vision-centric strategies for language-grounded driving. The framework incorporates CCDP for adaptive visual signal extraction and MEFA for temporal information integration. Additionally, DDIA facilitates visual-language alignment for robust navigation instruction following. These complementary strategies collectively enable more reliable autonomous driving.
- We evaluate our approach through closed-loop simulation experiments on standard language-grounded driving benchmarks using the CARLA platform. VLDrive achieves state-of-the-art performance while using significantly fewer parameters.

2. Related Work

2.1. End-to-End Autonomous Driving

Early end-to-end autonomous driving methods can be mainly divided into two categories: imitation learning [4, 8, 42] and reinforcement learning [2, 16, 32, 47]. Subsequently, benefiting from the robust feature modeling capabilities of the Transformer [36], researchers introduced additional modalities to provide complementary information for autonomous driving [24, 29, 45]. For example, Transfuse [24] uses Transformer to merge features from RGB and LiDAR bird's eye view (BEV) images to enhance the global comprehension of 3D scenes. InterFuse [29] improves the security and interpretability of the model by directly presenting its intermediate features and limiting actions to specified safe sets. In recent years, modular end-to-end planning approaches [1, 10, 26] have gained increased attention, where all components are connected and jointly optimized towards the ultimate goal, further improving the driving performance.

2.2. LLMs in Autonomous Driving

As LLMs continue to evolve and flourish, an increasing number of researchers are incorporating them into au-

tonomous driving tasks, broadening the scope and capabilities of these systems. Among them, a notable strand of research [22, 23, 30] employs LLMs for driving-related visual question answering, providing enhanced interpretability for driving decisions. Another line of work seeks to fully exploit the powerful logical reasoning capabilities of LLMs to fundamentally transform the research paradigm in autonomous driving tasks [5, 7, 20, 21]. Specifically, GPT-Driver [20] feeds LLM with observations and ego-states as language prompts and generates a planned trajectory and corresponding decision-making process in a natural language format. Agent-Driver [21] builds an LLM-powered agent, capitalizing on a tool library, a cognitive memory, and a reasoning engine for human-like autonomous driving. Recently, LLMs have been extended to visual modality successfully, making it possible for LLMs to understand the visual information of autonomous driving. Chen et al. [3] introduces a multimodal LLM to process vectorized numeric modalities data and produce driving-related question answers as well as action decisions. DriveGPT4 [41] curates a visual instruction tuning dataset for interpretable autonomous driving, taming LLM to generate text responses and low-level control signals. Further, based on the CARLA simulation platform, many studies [28, 37, 38] delve deeper and perform closed-loop evaluation driven by LLMs directly. For instance, LMDrive [28] takes multi-modality data and navigation instructions as input, achieving language-grounded driving and enhanced interaction with humans. Despite the impressive performance powered by LLMs, their considerably heavy parameters pose significant deployment challenges. In this work, we demonstrate that a lightweight language model can effectively respond to instructions and execute driving tasks through efficient visual representation extraction and strengthened vision-language alignment.

2.3. Token Reduction

Transformer has achieved substantial performance improvements across a range of tasks, yet it faces criticism for its quadratic computational costs relative to the number of input tokens. To tackle this issue, numerous studies focus on reducing token counts [12, 25, 39, 40]. For instance, DynamicViT [25] adaptively prunes redundant tokens at different Transformer layers to achieve token sparsification. SPVit [12] introduces an attention-based multi-head token selector with a soft pruning strategy. Unlike direct pruning methods, this approach consolidates selected redundant tokens into a single package token, enhancing efficiency while preserving essential information. In the LLM era, token reduction is also a crucial strategy to reduce the heavy computation costs [13, 14, 27, 50]. Most existing works utilize Q-former [13] to aggregate visual features into a fixed number of tokens [13, 43, 50]. LLaMA-VID [14] leverages average pooling followed by a learnable projector to compress the video frame representation. LLaVA-PruMerge [27] proposes an adaptive token selection via outlier detection, paired with a cluster-based merging strategy to supplement the information of pruned tokens to kept ones. Distinct from prior approaches, we present a cycle-consistent dynamic pruning strategy that integrates token reconstruction with token sparsification at the training phase. This design not only enables more representative token selection but also ensures that the retained tokens aggregate fine-grained global visual information.

3. Method

3.1. Overview

This task aims to predict driving actions by leveraging sequences of multi-sensor data and corresponding human navigation instructions. As shown in Fig. 2, our method follows the common framework of a multi-modal large language model (MLLM) and comprises three main components:

Visual Encoder: Given a sequence of visual data $\{\mathbf{X}_i, i = 1, 2, ..., T\}$, where each frame $\mathbf{X}_i$ includes multi-view camera images and corresponding LiDAR data. A visual encoder processes these multi-modal inputs to produce a unified feature representation $\mathbf{F}_i \in \mathbb{R}^{N \times C}$ for each frame, with N representing the total number of tokens.

Connector: As a bridge to align visual features with the language space, the connector plays a remarkable role in MLLM. To endow our model with adaptive visual perception and long-context modeling capability, as illustrated in Fig. 3, we tailor cycle-consistent dynamic visual pruning (CCDP), consisting of token sparsification and reconstruction, and memory-enhanced feature aggregation (MEFA) for the connector, significantly enhancing the visual representation. We denote the improved features of each frame by the connector as $\mathbf{F}_i^v \in \mathbb{R}^{N_v \times C_t}$, where C_t is the language model's hidden dimension.

Lite Language Model: A lightweight language model, enhanced by our proposed distance-decoupled instruction attention (DDIA), is utilized to process the combination of visual tokens of all frames, $\mathbf{F}_v \in \mathbb{R}^{N_v T \times C_t}$ and corresponding text embedding $\mathbf{F}_t \in \mathbb{R}^{N_t \times C_t}$ tokenized from navigation instructions. A following MLP leverages features from the language model and predicts the future trajectory, which is subsequently converted by PID controllers to produce the lateral steering action and the longitudinal acceleration action. We detail each component in the following subsections.

3.2. CCDP: Token Sparsification

Given the i-th frame's multi-modal features $\mathbf{F}_i$, as in DynamicViT [25], we predict the probabilities for pruning and retaining each token. Specifically, the probabilities are cal-

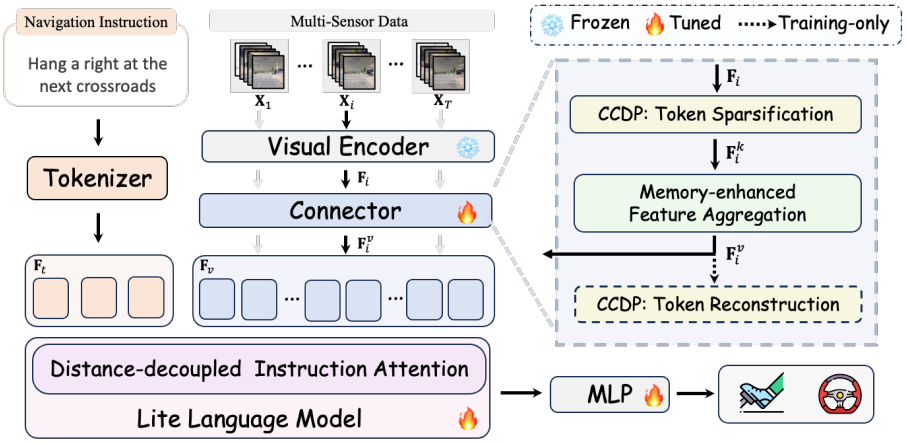

Figure 2. An overview of our proposed VLDrive framework. Given a sequence of visual data, our connector transforms each frame's raw visual features $\mathbf{F}_i$ into sparse yet informative representations $\mathbf{F}_i^v$ through two key components: **CCDP: Token Sparsification** and **Memory-enhanced Feature Aggregation (MEFA)**. Subsequently, a lite language model augmented with **Distance-decoupled Instruction Attention (DDIA)** jointly processes tokenized navigation instructions $\mathbf{F}_t$ and temporal visual features $\mathbf{F}_v = \{\mathbf{F}_i^v \mid i = 1, \ldots, T\}$. The resulting hidden representations are fed into an MLP for trajectory prediction, followed by PID controllers that translate these predictions into concrete driving actions. Additionally, we incorporate **CCDP: Token Reconstruction** as a *training-only* auxiliary task to further strengthen visual information integrity of $\mathbf{F}_i^v$ via explicit token reconstruction.

culated based on the combined global and local features $[\mathbf{G}_i; \mathbf{L}_i]$, which are obtained by MLP$(\cdot)$ operations:

$$\mathbf{L}_i = \text{MLP}(\mathbf{F}_i) \in \mathbb{R}^{N \times \frac{C}{2}} \tag{1}$$

$$\mathbf{G}_i = \text{Avg}(\mathbf{L}_i) \in \mathbb{R}^{1 \times \frac{C}{2}} \tag{2}$$

where Avg$(\cdot)$ is average operation. Another MLP is utilized to predict the probability $\mathbf{S}_i$, and $[\cdot;\cdot]$ denotes concatenation with broadcast mechanism:

$$\mathbf{S}_i = \text{Softmax}(\text{MLP}([\mathbf{G}_i; \mathbf{L}_i])) \in \mathbb{R}^{N \times 2} \tag{3}$$

In our implementation, we leverage Gumbel-Softmax to obtain the binary mask:

$$\mathbf{M}_i = \text{Gumbel-Softmax}(\mathbf{S}_i)_{*,1} \in \{0, 1\}^N \tag{4}$$

where $\mathbf{M}_i$ denotes the token retention masks. As shown in Fig. 3, the kept tokens $\mathbf{F}_i^k$ selected based on $\mathbf{M}_i$ are delivered to the **Memory-enhanced Feature Aggregation** module, which is detailed in the next subsection.

3.3. Memory-enhanced Feature Aggregation

Temporal reasoning is integral to our everyday driving scenarios. Human drivers can deduce the likely future direction or intentions of an approaching vehicle by analyzing its historical trajectory, and subsequently tailor their driving actions. To attain autonomous driving that matches the safety and reliability of human driving, we explicitly inject the temporal context into the current visual tokens. Specifically, for the current time step i, we introduce a memory bank storing the adjacent Z frames from the current one, denoted as $\mathbf{B}_i = [\mathbf{F}_{i-Z}, \mathbf{F}_{i-Z+1}, ..., \mathbf{F}_{i-1}]$, $\mathbf{B}_i \in \mathbb{R}^{Z \times N \times C}$. Temporal encoding $\mathbf{TE}_i$ is derived by an MLP utilizing the historical average data and current frame information:

$$\mathbf{B}_i^{avg} = \text{Avg}(\mathbf{B}_i) \in \mathbb{R}^{N \times C} \tag{5}$$

$$\mathbf{TE}_i = \text{MLP}([\mathbf{F}_i; \mathbf{B}_i^{avg}]) \tag{6}$$

Afterwards, a query transformer (Q-former) [13] is introduced to aggregate the features of current frames enhanced by temporary encoding:

$$\mathbf{F}_i^v = \text{Q-Former}(\mathbf{F}_i^k, \mathbf{F}_i + \mathbf{TE}_i) \tag{7}$$

In contrast to the original Q-former, we implement two key modifications: 1) Instead of using learnable queries, we utilize the retained visual tokens $\mathbf{F}_i^k$ as queries to probe and aggregate features. These retained tokens serve as significant anchors within the current frame, allowing for more effective feature aggregation. 2) We introduce a temporary embedding $\mathbf{TE}_i$ to the original visual features $\mathbf{F}_i$, which enhances the model's ability to focus on temporary-salient moving objects.

Finally, the generated $\mathbf{F}_i^v$ serves as part of the aggregated visual tokens across all frames, $\mathbf{F}_v$, which is then fed into the subsequent language model for trajectory prediction based on the corresponding navigation instructions. Additionally, we incorporate **CCDP: Token Reconstruction** as a *training-only* auxiliary task to further strengthen visual information integrity via explicit token reconstruction.

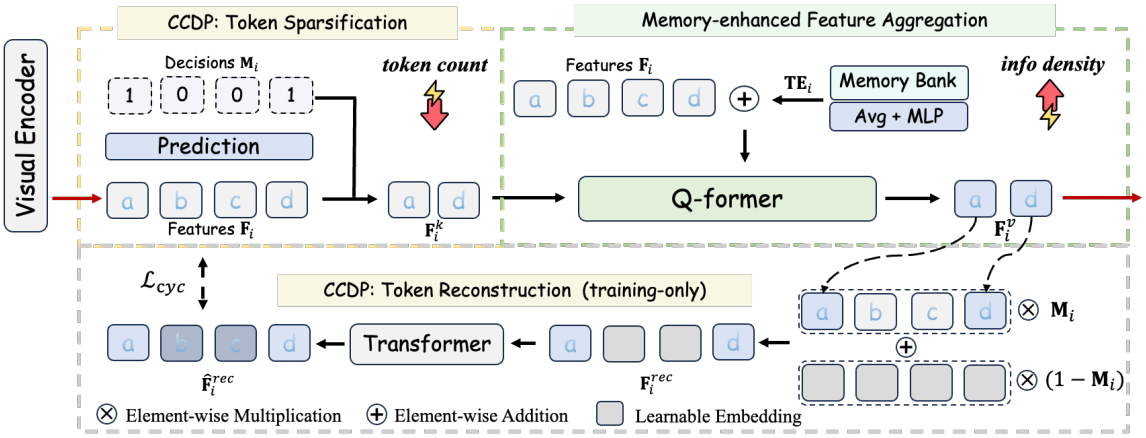

Figure 3. A detailed illustration of our proposed connector. **CCDP: Token Sparsification** and **Memory-enhanced Feature Aggregation** are proposed to reduce token count while enhancing information density. **CCDP: Token Reconstruction** serves as a training-only task that further ensures the information integrity of the retained tokens. Red arrows (→) indicate the input and output paths of our connector.

3.4. CCDP: Token Reconstruction

Through memory-enhanced feature aggregation, the generated $\mathbf{F}_i^v$ ideally encapsulates all visual features of the current frame. To further enhance the connector's ability to select critical visual tokens and aggregate complete information, we expect to reconstruct features of those pruned tokens based on $\mathbf{F}_i^v$. Firstly, we add the enhanced tokens $\mathbf{F}_i^v$ into the original features $\mathbf{F}_i$ according to their original token positions, forming a new feature vector denoted as $\langle \text{MLP}(\mathbf{F}_i^v), \mathbf{F}_i \rangle$, where an MLP function is used to align the dimensions of $\mathbf{F}_i^v$ and $\mathbf{F}_i$. Secondly, drawing inspiration from the Masked Autoencoder [9], a learnable embedding $\mathbf{e} \in \mathbb{R}^{N \times C}$ is introduced as a foundation for reconstructing the pruned tokens. Finally, we construct the input feature vector to be reconstructed through the following formula:

$$\mathbf{F}_i^{rec} = \langle \text{MLP}(\mathbf{F}_i^v), \mathbf{F}_i \rangle \cdot \mathbf{M}_i + \mathbf{e} \cdot (1 - \mathbf{M}_i) \quad (8)$$

In this way, we incorporate the decision $\mathbf{M}_i$ into the computational graph of feature reconstruction, enabling the reconstruction loss to influence pruning decisions and compelling the model to retain the most critical tokens. Afterwards, a 4-layer Transformer block takes as input $\mathbf{F}_i^{rec}$ and predicts the reconstructed results $\hat{\mathbf{F}}_i^{rec}$. A cycle consistency constraint (Eq. 11) is imposed to ensure that the retained tokens $\mathbf{F}_i^v$ preserve the integral information of the current video frame.

3.5. Distance-decoupled Instruction Attention

We introduce a **lightweight language model** for efficient autonomous driving, reducing both the number of parameters and computational resource requirements. Furthermore, distance-decoupled instruction attention (DDIA) is tailored for our lite language model, significantly alleviating the attention dilution for navigation instructions in the long-context driving scenarios. Specifically, as shown in Fig. 4, our proposed DDIA diverges from the traditional vanilla attention in two key aspects:

1): Causal-attention → Self-attention: Rather than adhering to the original causal setting, our self-attention **within instructional segments** allows each text token to access ample contextual information, thereby producing a more comprehensive representation of text features.

2): Distance-dependent → Distance-decoupled: In contemporary research on LLM, position encoding (e.g. RoPE [31]) is recognized as a critical ingredient to explicitly inject positional information into input tokens and enhance the model's contextual modeling capabilities. However, position encodings may introduce challenges, particularly in long-range driving scenarios. As historical visual data accumulates, the distance between current visual tokens and instructional tokens increases. The feature disparity introduced by position encoding may negatively impact the attention sensitivity of visual tokens to instruction ones, resulting in trajectory predictions that deviate from the specified instructions. To address this problem, inspired by [19], we maintain the distance-dependent attention within the unimodal text or visual tokens, but make the cross-modal attention between instructions and visual tokens unaffected by the position encoding. Specifically, with $\mathcal{V}$ and $\mathcal{I}$ denoting the sets of visual and instruction tokens, taking the j-th position as an example, our DDIA is defined as:

$$\begin{aligned} \text{DDIA}(\mathbf{Q},\mathbf{K},\mathbf{V})_j &= \frac{\sum_{k_i \in \mathcal{I}} sim(R(\mathbf{q}_j), R(\mathbf{k}_i))\mathbf{v}_i}{\sum_{k_i \in \mathcal{I}} sim(R(\mathbf{q}_j), R(\mathbf{k}_i))}, \text{if } \mathbf{q}_j \in \mathcal{I}, \\ &= \frac{\sum_{k_i \in \mathcal{I}} sim(\mathbf{q}_j, \mathbf{k}_i)\mathbf{v}_i + \sum_{k_i \in \{\mathcal{V}|<j\}} sim(R(\mathbf{q}_j), R(\mathbf{k}_i))\mathbf{v}_i}{\sum_{k_i \in \mathcal{I}} sim(\mathbf{q}_j, \mathbf{k}_i) + \sum_{k_i \in \{\mathcal{V}|<j\}} sim(R(\mathbf{q}_j), R(\mathbf{k}_i))}, \text{if } \mathbf{q}_j \in \mathcal{V} \end{aligned} \quad (9)$$

where $R(\cdot)$ denotes RoPE position embedding,

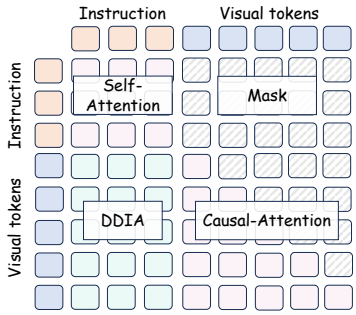

Figure 4. A detailed illustration of our proposed Distance-decoupled Instruction Attention (DDIA).

$sim(\mathbf{q}_j, \mathbf{k}_i) = exp(\mathbf{q}_j^T \mathbf{k}_i / \sqrt{d})$ and $\{\mathcal{V} \mid < j\}$ represents the causal subset within visual tokens.

3.6. Objective Function

The full training objective for i-th frame consists of two parts, formulated as follows:

$$\mathcal{L} = \mathcal{L}_{way} + \mathcal{L}_{cyc}, \tag{10}$$

here, $\mathcal{L}_{way}$ supervises trajectory prediction by minimizing the L_1 norm between predicted and ground-truth paths [28], while $\mathcal{L}_{cyc}$ facilitates our model to select the most discriminative tokens and aggregate more representative visual features, which is composed of two adversarial losses:

$$\mathcal{L}_{cyc} = \lambda_1 \mathcal{L}_{prun} + \lambda_2 \mathcal{L}_{rec} \tag{11}$$

$$\mathcal{L}_{prun} = (R - \frac{1}{N} \sum_{j=1}^{N} \mathbf{M}_{ij})^2 \tag{12}$$

$$\mathcal{L}_{rec} = \sum_{j=1}^{N} (1 - \mathbf{M}_{ij})(\mathbf{F}_{ij} - \hat{\mathbf{F}}_{ij}^{rec})^2 \tag{13}$$

$\mathcal{L}_{prun}$ restricts the ratio of kept tokens to our pre-defined value R, and $\mathcal{L}_{rec}$ reconstructs pruned tokens from kept ones after memory-enhanced feature aggregation. λ_1 and λ_2 balance the weights of two loss items. $\mathbf{M}_{ij}$, $\mathbf{F}_{ij}$ and $\hat{\mathbf{F}}_{ij}^{rec}$ denote the j-th element of $\mathbf{M}_i$, $\mathbf{F}_i$ and $\hat{\mathbf{F}}_i^{rec}$, respectively.

4. Experiments

4.1. Experiment Settings

Datasets. We train our model on the official language-driven autonomous driving dataset [28], which includes 64K instruction-following data clips collected across 8 towns on CARLA [6] simulation environment. Each clip not only equips multi-sensor input data (multi-view camera images and LiDAR data), but also introduces aligned navigation instructions in the natural language format.

Implementation Details. We adopt the visual encoder in LMDrive [28], producing 106 visual tokens (N=106) and kept it fixed during training. Two configurations of lightweight language models are integrated into our framework: LLaMA* (a 4-layer lite version of LLaMA [34]) and TinyLLaMA [44]. The hype parameters R determining the ratio of kept tokens and Z denoting the number of stored adjacent frames in the memory bank are set to 0.3 and 10, respectively. λ_1 and λ_2 in the loss function (Eq. 11) are configured to 10 and 1. More details are provided in the *Supplementary Material*.

Evaluation Metrics. As in LMDrive [28], we evaluate our model on three benchmarks, categorized by route length: LangAuto (>500m), LangAuto-Short (150m–500m) and LangAuto-Tiny (<150m) and employ route completion (RC), infraction score (IS), and driving score (DS) to assess close-looped driving performance. Specifically, route completion measures the extent to which a specified route is completed. Infraction score quantifies driving safety, starting from an ideal 1.0 base score and decrementing proportionally in the event of collisions or traffic rule violations. Driving Score is calculated as the product of RC and IS, offering a comprehensive evaluation of overall performance.

4.2. Comparisons with LLM-based Agents

We compare our VLDrive with LMDrive, a pioneering LLM-driven, closed-loop autonomous driving method, across several distance benchmarks. Table 1 presents the comparison results on the LangAuto benchmark. VLDrive has approximately 81% fewer parameters yet demonstrates superior driving performance compared to LMDrive, which is equipped with various LLM configurations. Specifically, VLDrive-LLaMA* achieves driving scores of 41.7% and route completion rates of 52.6%, outperforming LMDrive with LLaVA-v1.5 by 5.5% and 6.1%. This trend continues on the LangAuto-Short and LangAuto-Tiny benchmarks, as shown in Table 2, where VLDrive-LLaMA* attains driving scores of 63.6% and 78.4%, surpassing LMDrive by 13% and 11.9%. VLDrive-TinyLLaMA achieves new state-of-the-art performance, boosting the DS score to 43.8%, 67.4%, and 81.9% on the LangAuto, LangAuto-Short, and LangAuto-Tiny benchmarks, respectively.

4.3. Ablation Study

To investigate the impact of various components of our method, we conduct comprehensive ablation studies on the standard LangAuto benchmark. To reduce time costs and enhance training efficiency, we randomly extract approximately 25% of the training data from each town, forming a mini-training set for all ablation experiments. We present the metrics from 3 evaluation runs.

Table 1. Performance comparison of our method with LMDrive incorporated various LLM backbones on the LangAuto benchmark. We present the metrics from 3 evaluation runs.

Method	LLM	Params (Total)	LangAuto		
			DS ↑	RC ↑	IS ↑
	Large Language Model				
LMDrive	LLaMA [34]	7B	31.3±1.5	37.1±1.6	0.82±0.01
	LLaMA2 [35]	7B	32.8±2.1	40.1±2.2	0.81±0.02
	Vicuna [48]	7B	33.5±1.9	39.3±1.9	0.83±0.02
	Vicuna-v1.5 [48]	7B	34.0±3.8	39.0±3.3	**0.85±0.06**
	LLaVA-v1.5 [17]	7B	36.2±2.3	46.5±4.3	0.81±0.03
	Lightweight Language Model				
	Phi-2 [11]	3B	22.3±0.1	28.9±1.0	0.82±0.02
	TinyLLaMA [44]	1.3B	25.2±2.3	38.6±3.7	0.71±0.02
Ours	LLaMA*	1.3B	41.7±1.8	52.6±2.5	0.81±0.02
	TinyLLaMA [44]	1.3B	**43.8±2.4**	**54.5±3.0**	0.84±0.02

Table 2. Performance comparison of our method with LMDrive incorporated various LLM backbones on the LangAuto-Short and LangAuto-Tiny benchmarks. We present the metrics from 3 evaluation runs.

Method	LLM	Params (Total)	LangAuto-Short			LangAuto-Tiny		
			DS ↑	RC ↑	IS ↑	DS ↑	RC ↑	IS ↑
	Large Language Model							
LMDrive	LLaMA [34]	7B	42.8±7.2	49.1±8.5	0.87±0.03	52.2±5.3	57.8±8.0	0.91±0.05
	LLaMA2 [35]	7B	44.8±6.2	53.5±5.5	0.84±0.02	56.1±4.1	64.2±4.7	0.87±0.04
	Vicuna [48]	7B	45.3±4.9	54.3±3.9	0.83±0.03	55.5±3.9	63.1±4.2	0.88±0.04
	Vicuna-v1.5 [48]	7B	47.0±4.3	56.5±2.4	0.83±0.04	59.0±2.6	69.9±2.3	0.84±0.02
	LLaVA-v1.5 [17]	7B	50.6±1.7	60.0±3.4	0.84±0.04	66.5±3.6	77.9±2.3	0.85±0.02
	Lightweight Language Model							
	Phi-2 [11]	3B	48.0±3.6	55.2±4.1	**0.87±0.01**	56.3±4.9	68.2±3.9	0.82±0.04
	TinyLLaMA [44]	1.3B	46.2±3.4	59.7±3.4	0.79±0.02	64.1±1.2	75.0±1.4	0.86 ±0.01
Ours	LLaMA*	1.3B	63.6±4.1	77.6±4.1	0.81±0.06	78.4±4.6	**85.6±2.4**	0.91±0.04
	TinyLLaMA [44]	1.3B	**67.4±2.0**	**78.1±2.3**	0.85±0.02	**81.9±0.9**	85.5±0.6	**0.94±0.02**

Effectiveness of Each Individual Component. As shown in Table 3, we begin our ablation study from the baseline, which utilizes LLaMA* as the trainable language model and Q-Former [13] with learnable queries as the connector. Our method improves the baseline by: (1) introducing CCDP to extract critical sparse tokens, (2) replacing the vanilla Q-Former with our proposed MEFA module, and (3) substituting the causal attention mechanism with our DDIA module. The results in Table 3 validate the contribution of each component, with their synergistic integration leading to substantial improvements in driving performance.

Effectiveness of CCDP. We compare our CCDP approach with other token reduction strategies, including: 1) Structural pooling, a token reduction technique based on average pooling operations. 2) Dynamic pruning without token reconstruction. We maintain a consistent token retention ratio across all methods. As shown in Table 4, benefiting from the cycle-consistency constraints, CCDP is able to extract the most critical visual information and consequently achieves the highest driving scores. Fig. 5 further illustrates the positive correlation between reconstruction loss and trajectory prediction loss on the open-looped validation set, with a pearson correlation coefficient of 0.65.

Table 3. Ablation studies of each individual component.

Method	DS ↑	RC ↑	IS ↑
Baseline	30.0±0.8	46.2±3.5	0.69±0.03
Ours	39.8±1.7	49.1±1.8	0.83±0.02
w/o CCDP	35.5±3.8	45.3±3.5	0.78±0.02
w/o MEFA	35.9±0.8	46.9±1.3	0.81±0.02
w/o DDIA	36.3±1.3	47.0±1.0	0.81±0.02

Table 4. Ablation studies of various token reduction strategies.

Method	DS ↑	RC ↑	IS ↑
Structural Pooling	33.6±2.4	45.8±4.3	0.78±0.06
Dynamic Pruning	36.3±2.6	48.0±2.3	0.76±0.03
CCDP (Ours)	39.8±1.7	49.1±1.8	0.83±0.02

Table 5. Ablation studies of different token retention ratios.

Retention Ratio (R)	DS ↑	RC ↑	IS ↑
5%	33.4±0.4	45.1±3.2	0.77±0.03
10%	35.6±1.7	47.3±1.0	0.79±0.02
30%	39.8±1.7	49.1±1.8	0.83±0.02
50%	39.8±1.2	51.4±3.7	0.79±0.04

Table 6. Ablation of memory bank with different capacity.

Capacity (Z)	DS ↑	RC ↑	IS ↑
5	37.1±1.9	49.1±0.5	0.80±0.01
10	39.8±1.7	49.1±1.8	0.83±0.02
20	38.3±3.0	49.9±2.6	0.81±0.02

Analysis of Token Sparsification. Table 5 demonstrates that, retaining 30% of the visual tokens captures critical clues for autonomous driving. Further increment in the number of tokens produces no significant performance gains in driving scores.

Effectiveness of MEFA. As illustrated in Table 6, incorporating a memory bank provides essential temporal information that aids in predicting the future trajectories and behaviors of the ego-car and other objects, thereby enhancing the reliability of autonomous driving. Furthermore, our experiments indicate that a capacity of 10 adjacent frames yields the most significant performance improvements.

Effectiveness of DDIA. Fig. 6 shows the visual comparison of attention maps generated by vanilla causal attention and DDIA, where brighter colors indicate higher values. As highlighted in the red dashed box, DDIA enhances attention to instruction tokens, enabling visual tokens to better integrate instructional information and thus make more accurate driving decisions.

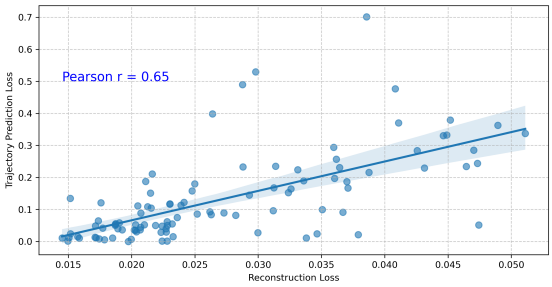

Figure 5. Correlation analysis between reconstruction and trajectory prediction losses, revealing a significant positive relationship (Pearson's r = 0.65).

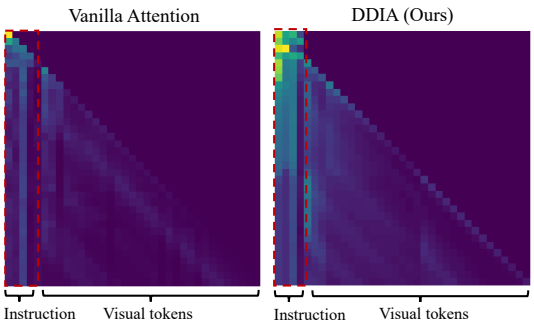

Figure 6. Visual comparison of attention maps generated by the language model equipped with vanilla causal attention and DDIA. We aggregate the attention weights within the instruction segment into the first four tokens for clearer visualization.

5. Conclusion

In this work, we introduce VLDrive, a vision-enhanced, lightweight model for language-grounded autonomous driving. To facilitate the practical deployment of autonomous vehicles, we streamline the heavy LLM into a more compact form while augmenting its visual modeling and vision-language alignment. Specifically, we employ cycle-consistent visual dynamic pruning to efficiently capture the most salient visual information for each frame, and incorporate a memory-enhanced feature aggregation that enriches critical temporal information to help the model comprehend historical trajectories and infer future movements. Additionally, a distance-decoupled instruction attention strategy is tailored for our lightweight language model, alleviating the attention dilution to navigation instructions and boosting the model's instruction-following capability. Extensive experiments demonstrate that our model, despite having significantly fewer parameters, achieves superior and robust driving performance across various benchmarks.

Acknowledgments

This work is supported in part by the National Key R&D Program of China (2024YFB3908503), in part by the National Natural Science Foundation of China (62322608), in part by the Shenzhen Longgang District Science and Technology Innovation Special Fund (No. LGKCYLWS2023018), in part by the Futian Healthcare Research Project (No.FTWS002), and in part by the Shenzhen Medical Research Fund (No. C2401036). This work is also sponsored by CIE-Tencent Robotics X Rhino-Bird Focused Research Program.

References

[1] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14403–14412, 2021. 2

[2] Raphael Chekroun, Marin Toromanoff, Sascha Hornauer, and Fabien Moutarde. Gri: General reinforced imitation and its application to vision-based autonomous driving. *Robotics*, 12(5):127, 2023. 2

[3] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. *arXiv preprint arXiv:2310.01957*, 2023. 3

[4] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9329–9338, 2019. 2

[5] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 902–909, 2024. 3

[6] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2, 6

[7] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 910–919, 2024. 3

[8] Jeffrey Hawke, Richard Shen, Corina Gurau, Siddharth Sharma, Daniele Reda, Nikolay Nikolov, Przemysław Mazur, Sean Micklethwaite, Nicolas Griffiths, Amar Shah, et al. Urban driving with conditional imitation learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 251–257. IEEE, 2020. 2

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5

[10] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1, 2

[11] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023. 7

[12] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European conference on computer vision*, pages 620–640. Springer, 2022. 3

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3, 4, 7

[14] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 3

[15] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020. 1

[16] Xiaodan Liang, Tairui Wang, Luona Yang, and Eric Xing. Cirl: Controllable imitative reinforcement learning for vision-based self-driving. In *Proceedings of the European conference on computer vision (ECCV)*, pages 584–599, 2018. 2

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 7

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[19] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*, 2023. 5

[20] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 3

[21] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*, 2023. 3

[22] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Video question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*, 2023. 3

[23] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for au-

tonomous driving. *arXiv preprint arXiv:2312.03661*, 2023. 3

[24] 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*, pages 7077–7087, 2021. 2

[25] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 3

[26] Abbas Sadat, Sergio Casas, Mengye Ren, Xinyu Wu, Pranaab Dhawan, and Raquel Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 414–430. Springer, 2020. 1, 2

[27] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. 3

[28] Hao Shao, Yuxuan Hu, Letian Wang, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488*, 2023. 1, 3, 6

[29] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, pages 726–737. PMLR, 2023. 1, 2

[30] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 3

[31] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[32] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7153–7162, 2020. 2

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 6, 7

[35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 7

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[37] Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. *arXiv preprint arXiv:2403.16996*, 2024. 1, 3

[38] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 1, 3

[39] Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. Joint token pruning and squeezing towards more aggressive compression of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2092–2101, 2023. 3

[40] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2964–2972, 2022. 3

[41] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023. 1, 3

[42] Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 2

[43] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 3

[44] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024. 6, 7

[45] Ruifei Zhang, Xiangru Lin, Wei Zhang, Jincheng Lu, Xuekuan Wang, Xiao Tan, Yingying Li, Errui Ding, Jingdong Wang, and Guanbin Li. Interactive 3d object detection with prompts. In *European Conference on Computer Vision*, pages 140–157. Springer, 2024. 2

[46] Wei Zhang, Jiaming Li, Meng Xia, Xu Gao, Xiao Tan, Yifeng Shi, Zhenhua Huang, and Guanbin Li. Offsetnet: Towards efficient multiple object tracking, detection, and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[47] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15222–15232, 2021. 2

[48] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 7

[49] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[50] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 3