

Driving by the Rules: A Benchmark for Integrating Traffic Sign Regulations into Vectorized HD Map

Xinyuan Chang^{1*} Maixuan Xue^{2*†} Xinran Liu¹ Zheng Pan¹ Xing Wei^{2‡}
¹Amap, Alibaba Group ²Xi'an Jiaotong University
 {changxinyuan.cxy, tom.lxr, panzheng.pan}@alibaba-inc.com
 xmx0809@stu.xjtu.edu.cn, weixing@mail.xjtu.edu.cn

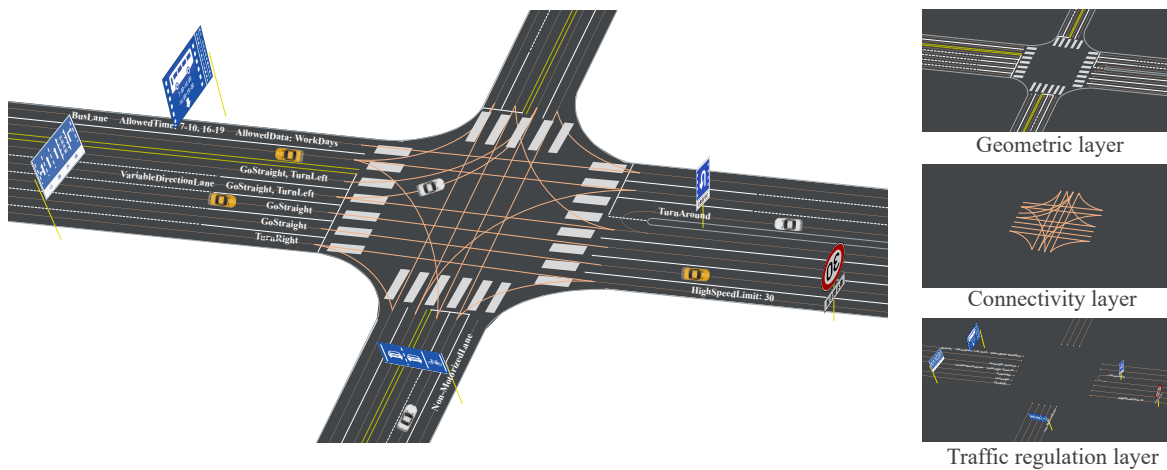


Figure 1. **MapDR Overview and Motivation.** The image on the left depicts the comprehensive results of HD map construction for an intersection scene, whereas the image on the right illustrates the outcomes after partitioning it into three layers. Existing online mapping methods primarily emphasize the construction of the geometric and connectivity layers, neglecting the traffic regulation layer. However, precise comprehension of traffic signs and their correlation with lanes is vital for ensuring the safety of autonomous driving.

Abstract

Ensuring adherence to traffic sign regulations is essential for both human and autonomous vehicle navigation. While current online mapping solutions often prioritize the construction of the geometric and connectivity layers of HD maps, overlooking the construction of the traffic regulation layer within HD maps. Addressing this gap, we introduce **MapDR**, a novel dataset designed for the extraction of **Driving Rules** from traffic signs and their association with vectorized, locally perceived **HD Maps**. MapDR features over 10,000 annotated video clips that capture the intricate correlation between traffic sign regulations and

lanes. Built upon this benchmark and the newly defined task of integrating traffic regulations into online HD maps, we provide modular and end-to-end solutions: **VLE-MEE** and **RuleVLM**, offering a strong baseline for advancing autonomous driving technology. It fills a critical gap in the integration of traffic sign rules, contributing to the development of reliable autonomous driving systems. Code is available at <https://github.com/MIV-XJTU/MapDR>.

1. Introduction

The rise of autonomous vehicles and intelligent transportation systems has emphasized the importance of accurate, reliable navigational data. High-Definition (HD) maps, with their detailed representations of road elements, have become essential to supporting these systems. An HD map typically

*Equal contribution.

†Work done during the internship at Amap, Alibaba Group.

‡Corresponding author.

consists of three primary layers: the geometric layer, connectivity layer, and traffic regulation layer [16, 35], as shown in Fig. 1. The geometric layer provides vector information, such as dividers and lane centerlines; the connectivity layer defines lane relationships to assist in path planning; and the traffic regulation layer encodes rule information (e.g., HOV lanes, bus lanes, speed-limit zones) associated with lanes, ensuring compliant driving.

A limitation of conventional HD maps is their inability to support rapid, real-time updates, making the construction of online HD maps an emerging trend. Existing methods, such as MapTR [30] and TopoMLP [54], address only the geometric and connectivity layers, overlooking the traffic regulation layer. As a result, autonomous driving systems currently still rely on offline maps to obtain traffic regulations, an approach that runs counter to the trend of online HD map construction.

Traffic signs serve as a "visual language" on the road, playing a critical role in defining traffic regulations. Inspired by human driving processes, the creation of traffic regulations can be divided into two stages: 1) understanding the rules conveyed by traffic signs, and 2) determining the lanes to which these rules apply. Although OpenLane-V2 [52] has made strides in associating signs with lanes in online mapping, it has two primary limitations. First, it considers only directional signs, neglecting other types of regulations. Second, its label annotations are limited to sign categories, falling short of the detailed rule descriptions required for autonomous driving. A more structured description, in alignment with HD map standards, is essential for supporting comprehensive autonomous driving applications [16, 35].

To bridge this gap, we introduce **MapDR**, the first dataset specifically designed for the challenging task of **integrating traffic regulations into existing online HD maps**. MapDR focuses on complex traffic scenes with diverse signage, collecting over 10,000 video clips across three representative cities in China, with each clip containing at least one traffic sign. In addition to vectorized representations of dividers, boundaries, and centerlines, MapDR provides structural annotations of traffic regulations and their associations with lanes, offering a comprehensive foundation for advancing online HD maps. More details of MapDR can be found in Sec. 4.

Based on MapDR, in addition to the end-to-end task definition, we also divide this task into two innovative sub-tasks aimed at bolstering research in this domain: 1) rule extraction from traffic signs, and 2) rule-lane correspondence reasoning. We introduce Vision-Language Encoder (**VLE**) and Map Element Encoder (**MEE**) to handle the interaction of multimodal data, encompassing images, texts, and vectors. Through the concatenation of these two models, integrated traffic regulations can be obtained. Additionally, inspired by recent advancements in Multimodal Large Lan-

Table 1. **Comparison of the existing datasets.** "Sign" and "Lane" denote whether the dataset focus on traffic signs and lanes. Only those annotated with formatted ("Fmt.") rules and the correspondence ("Corr.") between rules and lanes can form driving rules. "Clip" represents whether the data is organized in the form of video clips. "*" denotes that these samples are not newly collected and are built upon the previous dataset.

Dataset	Sign	Lane	Driving Rules		Number of Samples			Year
			Fmt.	Corr.	Image	Clip	Region	
nuScenes [9]		✓			1400K	1K	Worldwide	2019
Argoverse2 [53]		✓			2100K	1K	USA	2021
CTSU [19]	✓				5K	/	China	2021
OpenLane [11]		✓			200K*	1K*	Worldwide	2022
RS10K [20]	✓			✓	10K	/	China	2023
OpenLane-V2 [52]	✓	✓		✓	466K*	2K*	Worldwide	2023
Waymo [47]	✓	✓			390K	2K	USA	2024
MapDR(ours)	✓	✓	✓	✓	400K	10K	China	2024

guage Models (MLLMs), we also explore an end-to-end MLLM solution called **RuleVLM** to solve this task holistically. These approaches establish robust baselines and aim to inspire future research in this area. For detailed descriptions of the proposed tasks and metrics, please refer to Sec. 3. For approaches, please refer to Secs. 5 and 6.

To sum up, our contributions are as follows:

- For the first time, We introduce the novel task of extracting lane-level rules from traffic signs and integrating them into vectorized HD maps, alongside the MapDR dataset and evaluation metrics for benchmarking.
- MapDR comprises over 10,000 video clips and 400,000 front-view images captured across multiple cities in China, covering diverse traffic conditions and including more than 18,000 annotated lane-level rules. All data are newly collected and meticulously validated.
- We propose two modeling approaches, modular and end-to-end, to address the integration of lane-level rules into vectorized HD maps, providing effective baselines for future research.

2. Related Work

2.1. HD Map Construction

HD maps construction have seen significant advancements, with a focus on traffic element perception, including lane detection and traffic sign recognition [8, 9, 17, 18, 22, 46, 53, 58, 60]. The shift towards BEV perception and vectorization for end-to-end HD maps construction has gained traction [9, 11, 53]. Notable works include HDMaNet, which aggregates semantic segmentation results [29], LSS [40] estimates depth to transfer image features to BEV features, while VectorMapNet [36] is the first end-to-end framework for sequential vector point prediction to generate HD maps without post-processing. MapTR [30] and its enhanced

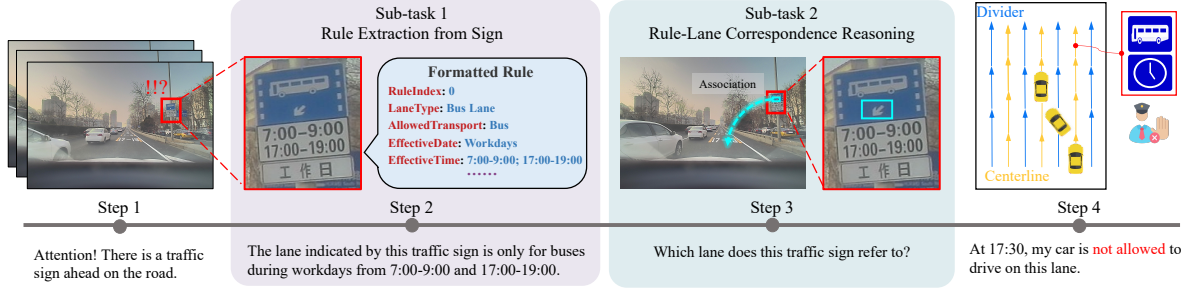


Figure 2. **Overview of the sub-tasks.** Step 1 ~ Step 4 shows a case of driving by the rules. Step 2 and Step 3 demonstrates the specific role of two sub-tasks, respectively.

version, MapTRv2 [31], introduced a unified permutation-equivalent modeling approach and extended it to a general framework supporting centerline learning and 3D map construction. However, these efforts have largely overlooked the integration of traffic sign rules into HD maps.

2.2. Traffic Element Association

Traffic element association aims to link elements like traffic signs with lanes. As demonstrated in Tab. 1, CTSU has initiated internal elements association to describe traffic sign in $\{key : value\}$ form, however lacking both generalization of driving rules from description and lane association [19]. VTKGG [20] propose to utilize a graph model for connectivity but also lacks structured expression of driving rules for motion planning and requires complex integration into HD maps, which is typically expressed in the BEV space. OpenLane-V2 [52] advances traffic sign and lane association in BEV space, but is constrained by single-label classification for traffic sign rather than structured descriptions for fine-grained driving rules. This makes it insufficient for signs with multiple rules, which are complex but common in real scenarios. Recent MLLM-based benchmarks [10, 38, 41, 43, 44] for autonomous driving, such as MAPLM [10], prioritize end-to-end motion planning over precise rule extraction from traffic sign, lacking evaluation for rule reasoning. MapDR addresses this gap by focusing on traffic sign rule extraction and lane association.

2.3. Vision-Language Models

Vision-Language Models (VLMs) facilitates multimodal applications by learning joint representations of vision and language data. Visual Question Answer (VQA) tasks provide answers to image-related questions [4], while Visual Information Extraction (VIE) tasks extract structured information from visual and textual data [4, 23, 55, 56]. In Autonomous Driving (AD), VLMs are increasingly used for comprehensive traffic scene understanding and decision-making. The field has seen various approaches, including using transformers [51] for joint encoding [23, 25], excelling at multimodal information interaction, and independent en-

coders for different modalities [24, 42] that are proficient in multimodal retrieval. Cross-modal representation methods [26, 59] combine these advantages, and the latest LLM-based research [28, 32–34] has achieved state-of-the-art results in various multimodal tasks. Nowadays, an increasing number of methods are leveraging LLMs to achieve impressive results [12, 37, 45, 57], with works like DriveLLM [13] showing significant potential in AD. However, addressing hallucination [6] remains the most crucial aspect for LLM-based approaches.

3. Task Definition : Integrating Traffic Sign Regulations into HD Maps

The ability to discern rules from traffic signs and to associate them with specific lanes is pivotal for autonomous navigation. Our proposed task focuses on extracting lane-level driving rules denoted as $R = \{r_i\}_{i=1}^m$ and reasoning the correspondence between these rules and the centerlines $L = \{l_i\}_{i=1}^k$ in the local vectorized HD map, where m is the number of rules and k is the number of centerlines. Each rule r_i is a set of pre-defined properties in $\{key : value\}$ pairs, as shown in Fig. 3. The final correspondence forms a bipartite graph $G = (R \cup L, E)$, where $E \subseteq \{0, 1\}^{m \times k}$ and the element E_{ij} in the i -th row and j -th column of matrix E represents the corresponding status between r_i and l_j . In this case, we assume the entire approach as $\mathcal{F}$ and the overall process can be formulated as $G = \mathcal{F}(X, L)$, where $X = \{x_i\}_{i=1}^n$ represents a series of image sequences and n is the number of frames. To independently investigate the capabilities of rule extraction and correspondence reasoning within the overall task, we further divided the overall task into two sub-tasks, defined as follows.

3.1. Rule Extraction from Traffic Sign

As shown in Step 2 of Fig. 2, this sub-task involves extracting multiple rules $R = \{r_i\}_{i=1}^m$ from a series of image sequences $X = \{x_i\}_{i=1}^n$. The rule extraction model, denoted as $\mathcal{M}$, can be expressed as $R = \mathcal{M}(X)$.

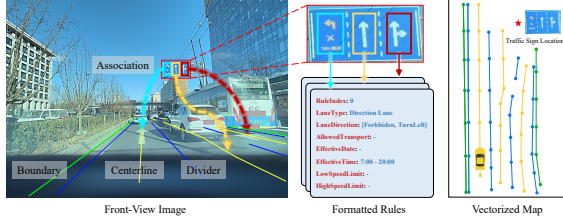


Figure 3. **Visualization of dataset demo.** Multiple lane-level rules of a single traffic sign are annotated in $\{key : value\}$ format. Directed lines indicate the correspondence between rules and particular centerlines.

3.2. Rule-Lane Correspondence Reasoning

As shown in *Step 3* of Fig. 2, the reasoning process establishes the correspondence between centerlines $L = \{l_i\}_{i=1}^k$ and all rules R . We denote the correspondence reasoning model as $\mathcal{T}$, and this process can be described as $E = \mathcal{T}(R, L)$. The final reasoning result forms a bipartite graph $G = (R \cup L, E)$, which means corresponding relationships only exist between rules and centerlines.

4. The MapDR Dataset & Benchmark

We introduce the MapDR dataset, meticulously annotated with traffic sign regulations and their correspondences to lanes, as shown in Fig. 3. The dataset encompasses a diverse range of scenarios, weather conditions, and traffic situations, with over 10,000 traffic scene segments, 18,000 driving rules, and 400,000 images. Traffic signs typically have varying textual descriptions, text layouts, and positions on the road, which add complexity to the task.

The majority of the data originates from Beijing and Shanghai, with additional scenes from Guangzhou. Fig. 4 illustrates the geographic spread and variety of traffic signs. The dataset reflects a natural long-tail distribution, with a prevalence of bus and direction lanes and a scarcity of tidalflow lanes. We primarily focus on traffic signs that indicate lane-level rules, collected from cities with the most complex and diverse traffic scenarios in China, ensuring realistic and representative data. All images have undergone privacy and safety processing to obscure license plates and faces. More comprehensive statistic of dataset and case demonstrations can be found in the supplement.

4.1. Raw Data & Annotation

Raw Data. MapDR is collected from real-world traffic scenes, each scene segment (video clip) captures front-view images within a $100m \times 100m$ area centered on the traffic sign, with a consistent resolution of 1920×1240 . Each clip contains 30 to 60 frames, captured at 1 frame every 2 meters, ensuring consistent spatial intervals. Each video clip focuses on a single traffic sign and provides its position in 3D space.

Camera intrinsics and poses are provided for each frame, and coordinates for each clip are transformed to distinct ENU systems.

All vectors of local map in the target area are provided as 3D point lists, generated using our algorithm similar to MapTRv2 [31]. Each lane vector has a type, such as divider, centerline, crosswalk, or boundary. For example, the centerline is defined as $L = \{l_i\}_{i=1}^k$, where each vector l_i is composed of multiple 3D points $l_i = [p_1, \dots, p_n]$, and $p_j = (x_j, y_j, z_j)$ represents the coordinates of the current point. The pipeline of dataset production is illustrated in Fig. 5, and detailed data acquisition and annotation procedures can be found in the supplement material.

Formatted Rules. Each video clip may contain multiple lane-level rules, denoted as R . Each rule is expressed by symbols and text on the sign, requiring interpretation. We reference existing HD map design and data specifications [1, 16, 35] to define lane-level driving rules in the JSON format. As shown in Fig. 3, each rule r_i comprises 8 predefined properties in the form of $\{key : value\}$ pairs. This definition schema encompasses most scenarios and supports integration into existing autonomous driving systems. We enclose the symbols and texts denoting each distinct rule on traffic signs with polygons and project them into 3D space as $P_i = [p_1, \dots, p_n]$, where n varies. Researchers can optionally use this information to facilitate rule extraction.

Correspondence between Rules & Lanes. Based on formatted rules R and centerlines L , corresponding centerlines of each rule are annotated as shown in Fig. 3. Therefore correspondence between rules and centerlines can be formed as a bipartite graph $G = (R \cup L, E)$, where $E \subseteq \{0, 1\}^{|R| \times |L|}$ and the positive edges only exist between R and L as demonstrated in Sec. 3.2. Specifically, $E_{ij} = 1$ represents that

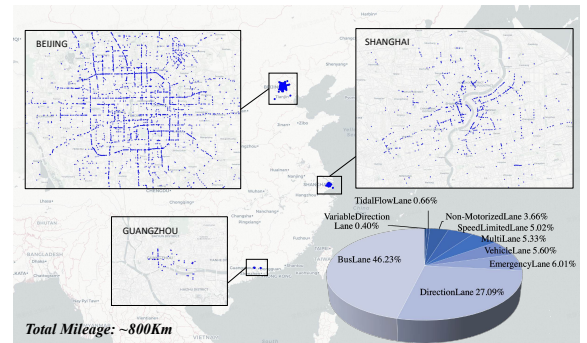


Figure 4. **Geographic location distribution of the collected traffic signs and proportions of various lane types represented in all signs.** The geographic distribution is visualized based on OpenStreetMap [2].

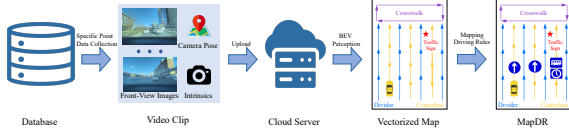


Figure 5. **Pipeline of dataset production.** The location of traffic signs are sampled from existing database then front-view images of each sign are newly collected. Vectorized map is processed in cloud server. Finally formatted rules and correspondence between rules and centerlines are annotated and organized as MapDR.

vehicle driving on the lane with centerline l_j should follow the driving rule r_i .

4.2. Evaluation Metrics

Based on the task defined in Sec. 3, we propose evaluation metrics for entire task as well as for two sub-tasks separately. The F_1 score is adopted as the benchmark for methods rank.

Rule Extraction (R.E.). Given the ground truth R and predicted rules $\hat{R}$, we propose to calculate the *Precision* ($P_{R.E.}$) and *Recall* ($R_{R.E.}$) to evaluate the capability of rules extraction as defined in Eq. (1), where $\hat{r}_i = r_i$ represents all the properties are predicted correctly.

$$P_{R.E.} = \frac{|\hat{R} \cap R|}{|\hat{R}|} \quad R_{R.E.} = \frac{|\hat{R} \cap R|}{|R|} \quad (1)$$

Correspondence Reasoning (C.R.). Given the ground truth of correspondence bipartite graph $G = (R \cup L, E)$ and predicted graph $\hat{G} = (R \cup L, \hat{E})$, we propose to calculate *Precision* ($P_{C.R.}$) and *Recall* ($R_{C.R.}$) of edge set E to evaluate the capability of correspondence reasoning individually. Metrics are defined as Eq. (2).

$$P_{C.R.} = \frac{|\hat{E} \cap E|}{|\hat{E}|} \quad R_{C.R.} = \frac{|\hat{E} \cap E|}{|E|} \quad (2)$$

Overall. As defined in Sec. 3, the final predicted results can be obtained through an end-to-end prediction or a combination of two sequential sub-tasks. Given the predicted rules, correspondence should be reasoned between $\hat{R}$ and L which means the prediction of entire task is $\hat{G} = (\hat{R} \cup L, \hat{E})$ and the ground truth is consistent $G = (R \cup L, E)$. We evaluate *Precision* (P_{all}) and *Recall* (R_{all}) using the sub-graph G^s , where $G^s = \{g_{ij}^s\}_{i=1, j=1}^{m, k}$, $g_{ij}^s = (\{r_i, l_j\}, e_{ij})$. In set of sub-graph G^s , m is the number of rules, and k is the number of centerlines. Furthermore, we propose the F_1 for the final benchmark ranking. Metrics are defined in Eq. (3) and Eq. (4). The F_1 is derived from the calculations of P_{all} and R_{all} . We provide an example of calculating the *Overall* metrics in the supplement material.

$$P_{all} = \frac{|\hat{G}^s \cap G^s|}{|\hat{G}^s|} \quad R_{all} = \frac{|\hat{G}^s \cap G^s|}{|G^s|} \quad (3)$$

$$F_1 = 2 \times \frac{P_{all} \times R_{all}}{P_{all} + R_{all}} \quad (4)$$

5. Modular Approach

In order to conduct a comprehensive analysis of the entire task, a modular approach is implemented in this section to address the two sub-tasks. It is demonstrated that these interconnected modules can effectively achieve the ultimate objective of integrating driving rules into the HD map. To address the multimodal information interaction encompassing images, texts, and vectors, a **Vision-Language Encoder (VLE)** and a **Map Element Encoder (MEE)** are developed. The following sections expound upon their structures, applications, and the experimental results on MapDR.

5.1. Architecture

Vision-Language Encoder. Inspired by vision-language frameworks [7, 25–27, 42], we designed a vision-language fusion model named VLE, following [26]. As shown in Fig. 6, VLE uses ViT-b16 [15] as the vision encoder, with the text encoder and multimodal fusion encoder each consisting of L transformer layers [51]. Each layer of the fusion encoder includes a cross-attention module for fusion [26]. In practice, distinct rules are represented by varying numbers of symbols and texts, as shown in the OCR results in Fig. 6. To address the challenge of representing variable-length input as fixed-length features, we introduce a [CLS] token for an entire rule and several [STC] tokens for sentence-level representation. The specific usage of these tokens is detailed in Sec. 5.2. Furthermore, we incorporate inter-instance and intra-instance attention mechanisms [31] to enhance model performance by capturing interactions and independence between and within sentences. In addition to content, layout captures the relative positions of symbols and texts, offering important semantic meaning. To leverage this, we encode the layout using the method from [48] and the relative positions of characters as position embedding following [14]. As shown in VLE in Fig. 6, text embedding, layout embedding, and position embedding together form the input of the text encoder.

Map Element Encoder. Vectors can be represented as sequences of points, similar to words in sentences. Inspired by this, we designed MEE akin to language models [14]. The MEE employs M transformer layers for vector encoding and N cross-attention layers for multimodal fusion. Utilizing the method from [48], points of each vector are embedded as point embedding. To achieve a fixed-length representation,

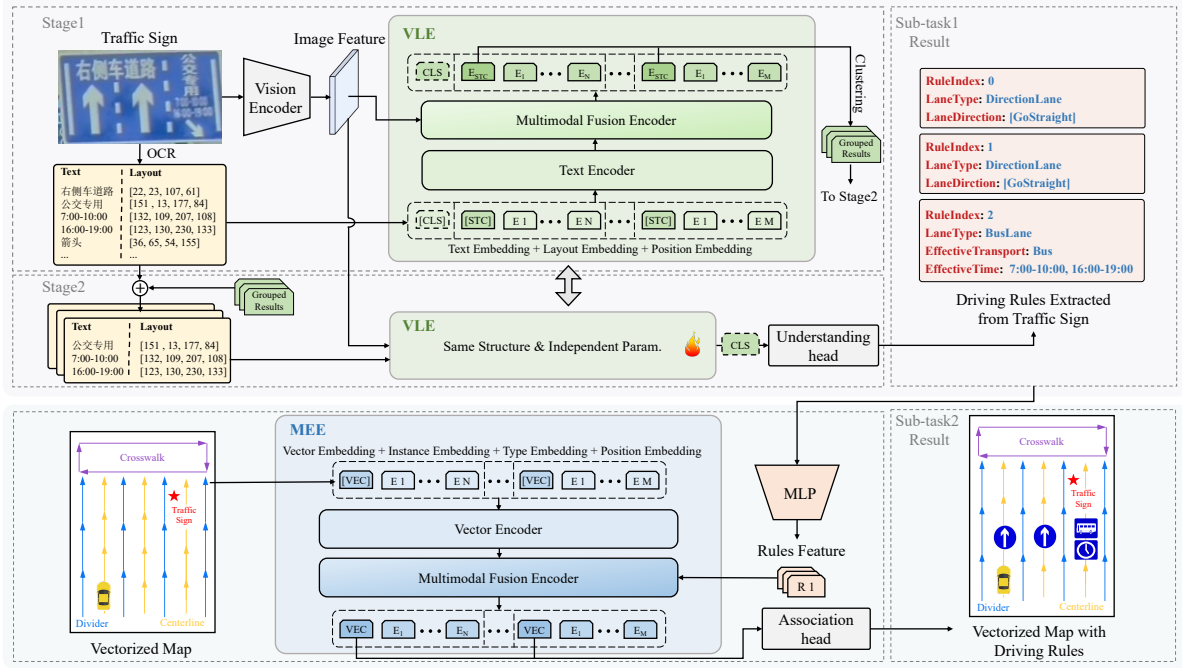


Figure 6. **Overview of the modular approach.** Entire approach can be divided into two main parts: **Rule Extraction from Traffic Sign** (top) and **Rule-Lane Correspondence Reasoning** (bottom). Rule Extraction model consists of two sequential stages with the same structure VLE but unshared parameters, and the training procedure is independent.

we add [VEC] tokens as the first token of each vector, similar to [STC] tokens in the VLE. We also introduce learnable type embedding for vector types, learnable instance embedding to distinguish vector instances, and position embedding from [14] to encode the relative positions of multiple points within a vector. These embedding are aggregated as the input of vector encoder, as shown in Fig. 7. In addition, we employ inter-instance and intra-instance attention mechanisms [31] to prioritize interactions within vectors over interactions between vectors, as depicted in the dashed box on the right side of Fig. 7. The [VEC] token in output serves as fused feature of rules and vectors, enabling the final prediction of their relationships through association head.

5.2. Implementation

We utilize VLE and MEE as backbones to integrate multiple modalities and address these two sub-tasks. The specific procedures are detailed as follows:

Rule Extraction from Traffic sign. To clarify the objectives of model, we first *cluster symbols and texts into groups*. As shown in the upper part of Fig. 6, the VLE is used to encode OCR results and images. By calculating the cosine similarity between [STC] tokens, different symbols and texts are clustered into groups. This process is supervised by contrastive loss during training. Next, using grouped OCR results as text input and maintaining the VLE structure, we

extract lane-level rules. We employ multiple linear layers as understanding head for the [CLS] token to predict the corresponding value for each property of the rules. This facilitates to express all rules as $\{key : value\}$ pairs.

Rule-Lane Correspondence Reasoning. MEE is designed for vector encoding and interaction with rules. Each formatted rule is mapped to an embedding through MLP and

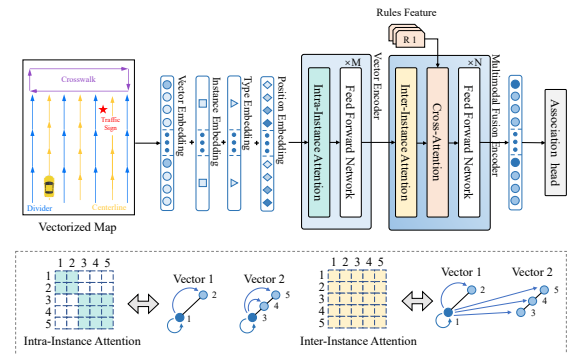


Figure 7. **Structure of MEE.** MEE serves as correspondence reasoning model. Learnable embeddings are introduced within input to enhance the representing capacity of vector types. inter & intra-instance attention mechanisms facilitate to capture the relationships and independence of individual vectors.

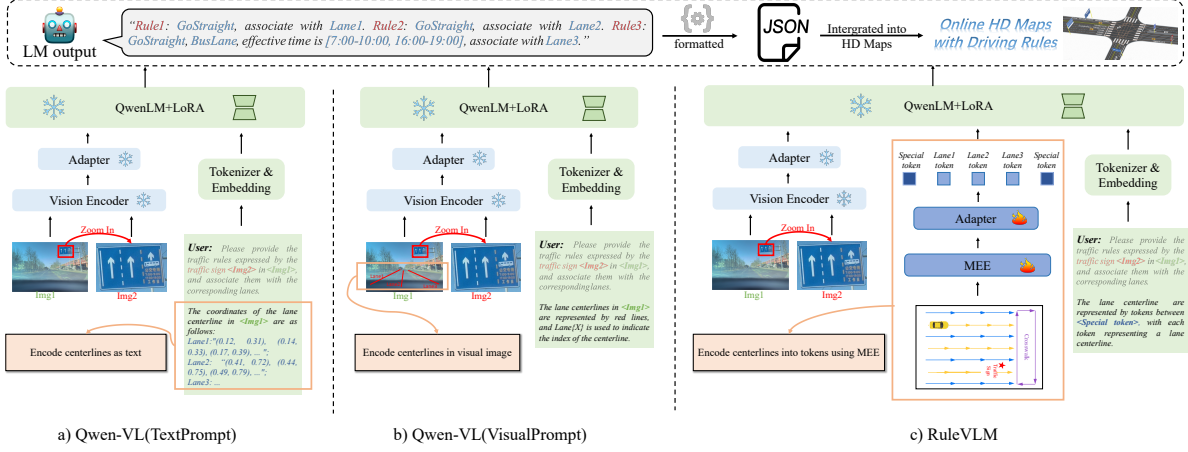


Figure 8. **Overview of end-to-end approaches.** Based on different vectorized HD maps encoding methods, approaches can be categorized into three types.

fused with vector features in the fusion encoder, as shown in the lower part of Fig. 6. We add a binary classification head after each [VEC] token to determine the relationship between the current centerline and rule. Details can be found in supplement.

5.3. Experiment

Setups. The dataset is split into *train* and *test* sets in the ratio of 9 : 1. $L = 6$ in VLE and $M = 2, N = 2$ in MEE. Input images are resized to 256×256 and the feature dimension is 768 with consistent 12 attention heads. We initialize VLE with pre-trained weights of DeiT [50] and BERT [14] while MEE is trained from scratch. The training procedure runs 50 and 120 epochs for VLE and MEE, respectively. More details can be found in the supplement.

Results. We use a heuristic method based on OCR character matching and nearest lane matching as the modular baseline. Meanwhile we make modifications to ALBEF [26] and BERT [14] to adapt them to our task. As shown in Tab. 2, the heuristic method performs poorly in both R.E. and C.R. tasks, underscoring the challenging nature of the MapDR dataset. ALBEF-BERT method significantly enhances performance in the R.E. sub-task however lacking adequate adaptation to vector information hampers its performance on the C.R.. Tab. 3 indicates the attention mechanisms significantly improve $R_{R.E.}$, while layout of text brings marginal improvement. For C.R., the attention mechanisms is the critical part that promotes the effectiveness of MEE. Instance embedding slightly improves $P_{C.R.}$ and $R_{C.R.}$, while type embedding significantly enhances both, indicating that vector types help the model establish rule-lane correspondence. The separate evaluation results of all lane types and the strategies of heuristic method can be found in the supplement.

6. End-to-End Approach

Recently, there has been notable progress in the development of Multimodal Large Language Models (MLLMs) [3, 5, 39, 49], demonstrating significant capabilities in processing multimodal data and acquiring implicit knowledge of modality alignment. Their large-scale pre-training endows them with exceptional understanding and generative abilities, thereby enabling the comprehensive execution of the task proposed by MapDR in an end-to-end manner. This section aims to address the overall task using Qwen-VL [5] and to investigate three distinct vector encoding methods. Moreover, we introduce an innovative **RuleVLM** model based on MEE, designed to process inputs from three modalities (images, texts, and vectors) and to produce driving rules, which yield favorable results on MapDR. Further elaboration on these concepts is presented in subsequent sections.

6.1. Architecture

As shown in Fig. 8, we classify MLLM methods into three types: Qwen-VL(TextPrompt), Qwen-VL(VisualPrompt), and RuleVLM, based on different vector encoding methods. These methods utilize the framework of Qwen-VL [5], with ViT [15] serving as the image encoder. Qwen-VL(TextPrompt) encodes centerline coordinates in the PV image as text, inputs them into QwenLM; Qwen-VL(VisualPrompt) visualizes the centerline and its index in the PV image and inputs it as Visual Prompt into QwenLM, while RuleVLM uses MEE with the cross-attention layer removed to encode vectorized centerline results and aligns it with LLM through an adapter. The goal of all these methods is to output a serialized set of driving rules in text form, decode it through a json decoder to restore formatted data, and then integrate it into HD maps for further applications.

Table 2. **Evaluation of the overall task.** The heuristic method and the Qwen-VL(TextPrompt) method serve as the baselines for the modular and end-to-end approach, respectively. "—" denotes end-to-end approach is not suitable to independent evaluations of C.R. because these approaches do not utilize ground truth of rules for correspondence reasoning independently.

Model	Type	R.E.		C.R.		Overall		
		$P_{R.E.}(\%)$	$R_{R.E.}(\%)$	$P_{C.R.}(\%)$	$R_{C.R.}(\%)$	$P_{all}(\%)$	$R_{all}(\%)$	$F1\ Score$
Heuristic	Modular	18.01	11.51	33.05	17.99	5.01	2.73	0.035
ALBEF-BERT		75.78	57.56	4.14	17.25	0.24	0.78	0.003
VLE-MEE		76.67	74.54	78.05	82.16	63.35	67.37	0.653
Qwen-VL(TextPrompt)	End-to-End	42.21	41.09	—	—	8.39	8.17	0.083
Qwen-VL(VisualPrompt)		89.29	89.50	—	—	39.14	39.23	0.392
RuleVLM		89.28	89.44	—	—	64.16	64.28	0.642

Table 3. **Evaluation of sub-tasks.** "Attn." indicates intra & inter-instance attention mechanisms. "Layout" refers to the text layout applied in VLE. "In.E." and "Ty.E." denotes instance and type embedding in MEE, respectively.

VLE		$P_{R.E.}(\%)$	$R_{R.E.}(\%)$
Attn.	Layout		
$\times$	$\times$	75.78	57.56
$\checkmark$	$\times$	76.86	71.75
$\checkmark$	$\checkmark$	76.67	74.54

MEE			$P_{C.R.}(\%)$	$R_{C.R.}(\%)$
Attn.	In.E.	Ty.E.		
$\times$	$\times$	$\times$	4.14	17.25
$\checkmark$	$\times$	$\times$	68.91	71.39
$\checkmark$	$\checkmark$	$\times$	69.68	72.76
$\checkmark$	$\checkmark$	$\checkmark$	78.05	82.16

6.2. Implementation

The JSON-formatted driving rules in MapDR are serialized into text and used as the answers in the QA format for training and evaluation. For all methods based on QwenVL, the pre-trained weights of Qwen-VL-Chat (9.6B) are used for initialization. For method using MEE, the trained weights of MEE from the modular approach are utilized for its initialization. During the supervised fine-tuning, we perform LoRA [21] fine-tuning of the QwenLM, while keeping the parameters of ViT frozen. For RuleVLM, we employ a linear layer as the adapter between QwenLM and MEE, while the parameters of MEE and the adapter remains trainable. We use special tokens to separate the centerline tokens extracted by MEE from the visual and text tokens. Additionally, to prevent overfitting, we shuffle the order of centerlines in the local vectorized map during training.

6.3. Experiment.

Setups. All experiments are conducted on QA formatted MapDR dataset with 20 epochs of training on NVIDIA RTX

A6000. The dimension of hidden states remains 4096 while r and α of LoRA configuration are set 64 and 16 respectively. Meanwhile training batch size is 6 and the max training token length remains 2048.

Results. We conducted zero-shot evaluations on a subset of MapDR using existing MLLMs [3, 5, 39, 49] as qualitative investigation, but we encounter challenges in achieving stable structured outputs, further details are provided in the supplement. Additionally, as shown in the Tab. 2, Qwen-VL(TextPrompt) method serves as a end-to-end baseline do not perform better on both the R.E. and overall evaluation. Textual representations of centerline coordinates are challenging for MLLMs to perform spatial reasoning and also resulted in excessive length of sequence, which affected $P_{R.E.}$ and $R_{R.E.}$. Qwen-VL(VisualPrompt) visualizes the centerline within the image, resulting in reduced text sequence length, and demonstrates notable enhancements in the R.E. task compared to the baseline. However, the overall metrics are suboptimal, suggesting that image-based spatial reasoning remains an exceedingly challenging endeavor. The RuleVLM method that we propose introduces MEE for the independent encoding of vectorized HD maps information. This facilitates a more direct extraction of centerline features, resulting in significant enhancements in the overall metrics.

7. Conclusion

In this work, we defined the task of integrating traffic regulations into vectorized HD maps to facilitate the construction of the traffic regulation layer within online HD maps. We constructed the MapDR dataset, which includes over 10,000 video clips, more than 400,000 images, and at least 18,000 driving rules, covering a total mileage of approximately 800Km. Additionally, we proposed the modular approach MEE-VLE and the end-to-end approach RuleVLM, establishing effective baselines for further research. We hope that MapDR can serve as a starting point, and that more researchers can contribute to expanding this dataset or proposing even better solutions.

References

- [1] Navigation data standard. <https://nds-association.org/>. 4
- [2] Openstreetmap. <https://github.com/openmaptiles/openmaptiles>. 4
- [3] Anthropic. Claude-3. <https://www.anthropic.com/news/claude-3-family>, 2024. 7, 8
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015. 3
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 7, 8
- [6] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 3
- [7] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *NeurIPS*, 2022. 5
- [8] Karsten Behrendt, Libor Novak, and Rami Botros. A deep learning approach to traffic lights: Detection, tracking, and classification. In *ICRA*, 2017. 2
- [9] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2
- [10] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James Rehg, and Chao Zheng. Maplm: A real-world large-scale vision-language dataset for map and traffic scene understanding. In *CVPR*, 2024. 3
- [11] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, and Junchi Yan. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *ECCV*, 2022. 2
- [12] Long Chen, Oleg Sinavski, Jan Hünemann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *ICRA*, 2024. 3
- [13] Yaodong Cui, Shucheng Huang, Jiaming Zhong, Zhenan Liu, Yutong Wang, Chen Sun, Bai Li, Xiao Wang, and Amir Khajepour. Drivellm: Charting the path toward full autonomous driving with large language models. *IEEE TIV*, 2024. 3
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 5, 6, 7
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5, 7
- [16] Gamal Elghazaly, Raphaël Frank, Scott Harvey, and Stefan Saffo. High-definition maps: Comprehensive survey, challenges, and future perspectives. *IEEE Open Journal of Intelligent Transportation Systems*, 2023. 2, 4
- [17] Andreas Fregin, Julian Müller, Ulrich Krebel, and Klaus Dietmayer. The driveu traffic light dataset: Introduction and comparison with existing datasets. In *ICRA*, 2018. 2
- [18] Shuo Gu, Yigong Zhang, Jinhui Tang, Jian Yang, and Hui Kong. Road detection through CRF based lidar-camera fusion. In *ICRA*, 2019. 2
- [19] Yunfei Guo, Wei Feng, Fei Yin, Tao Xue, Shuqi Mei, and Cheng-Lin Liu. Learning to understand traffic signs. In *ACMMM*, 2021. 2, 3
- [20] Yunfei Guo, Fei Yin, Xiao-Hui Li, Xudong Yan, Tao Xue, Shuqi Mei, and Cheng-Lin Liu. Visual traffic knowledge graph generation from scene images. In *ICCV*, 2023. 2, 3
- [21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 8
- [22] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE TPAMI*, 2020. 2
- [23] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document AI with unified text and image masking. In *ACMMM*, 2022. 3
- [24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3
- [25] Wonjae Kim, Bokyoung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 3, 5
- [26] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 3, 5, 7
- [27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 5
- [28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3
- [29] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online HD map construction and evaluation framework. In *ICRA*, 2022. 2
- [30] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr:

- Structured modeling and learning for online vectorized HD map construction. In *ICLR*, 2023. 2
- [31] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *arXiv preprint arXiv:2308.05736*, 2023. 3, 4, 5, 6
- [32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3
- [33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- [34] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next>, 2024. 3
- [35] Rong Liu, Jinling Wang, and Bingqi Zhang. High definition map for automated driving: Overview and analysis. *Journal of Navigation*, 2020. 2, 4
- [36] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized HD map learning. In *ICML*, 2023. 2
- [37] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with GPT. In *NeurIPS Workshop*, 2023. 3
- [38] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoît Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and Oleg Sinavski. Lingoa: Video question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*, 2023. 3
- [39] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. 7, 8
- [40] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020. 2
- [41] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *AAAI*, 2024. 3
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 5
- [43] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel J. Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *WACV*, 2024. 3
- [44] 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
- [45] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *ECCV*, 2024. 3
- [46] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 2012. 2
- [47] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2
- [48] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020. 5
- [49] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, and et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2024. 7, 8
- [50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 7
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 5
- [52] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, Feng Wen, Hang Xu, Ping Luo, Junchi Yan, Wei Zhang, and Hongyang Li. Openlane-v2: A topology reasoning benchmark for unified 3d HD mapping. In *NeurIPS*, 2023. 2, 3
- [53] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS*, 2021. 2
- [54] Dongming Wu, Jiahao Chang, Fan Jia, Yingfei Liu, Tiancai Wang, and Jianbing Shen. Topomlp: A simple yet strong pipeline for driving topology reasoning. In *ICLR*, 2024. 2
- [55] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD*, 2020. 3
- [56] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *ACL*, 2021. 3
- [57] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics Autom. Lett.*, 2024. 3
- [58] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Dar-

- rell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. [2](#)
- [59] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. [3](#)
- [60] Zhe Zhu, Dun Liang, Song-Hai Zhang, Xiaolei Huang, Baoli Li, and Shi-Min Hu. Traffic-sign detection and classification in the wild. In *CVPR*, 2016. [2](#)