

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

Supplementary Material

A. Appendix Overview

Our appendix encompass author statements, licensing, dataset access, dataset analysis, and the implementation details of benchmark results to ensure reproducibility. Additionally, we offer dataset documentation in adherence to the Datasheet format [4], which covers details such as data distribution, maintenance plan, composition, collection, and other pertinent information.

B. Author Statement

We bear all responsibilities for licensing, distributing, and maintaining our dataset.

C. Licensing

The proposed dataset MapDR is under the CC BY-NC-SA 4.0 license, while the evaluation code is under the Apache License 2.0.

D. Datasheet

D.1. Motivation

For what purpose was the dataset created? Autonomous driving not only requires attention to the vehicle’s trajectory but also to traffic regulations. However, in the online-constructed vectorized HD maps, traffic regulations are often overlooked. Therefore, we propose this dataset to integrate lane-level regulations into the vectorized HD maps. These regulations can serve as navigation data for both human drivers and autonomous vehicles, and are crucial for driving behavior.

D.2. Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes, the dataset is open to public.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? The dataset is available at <https://modelscope.cn/datasets/MIV-XJTU/MapDR>. Code is available at <https://github.com/MIV-XJTU/MapDR>.

D.3. Maintenance

Is there an erratum? No. We will make a statement if there is any error are found in the future, we will release errata on the main web page for the dataset.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Yes, the dataset will be updated as necessary to ensure accuracy, and announcements will be made accordingly. These updates will be posted on the dataset’s webpage on <https://modelscope.cn/datasets/MIV-XJTU/MapDR>.

Will older versions of the dataset continue to be supported/hosted/maintained? Yes, older versions of the dataset will continue to be maintained and hosted.

D.4. Composition

What do the instances that comprise the dataset represent? An instance of the dataset consists of three main parts: a video clip, basic information, and annotation. The video clip comprises at least 30 continuous front-view image frames, with one frame captured every 2 meters to ensure uniform spatial distribution. Basic information of each clip is presented in the form of a JSON file, including the locations of traffic sign, all lane vectors, camera intrinsic parameters, and the camera poses for each frame. Annotation is also organized in JSON format, containing multiple driving rules. Each rule consists of a set of properties in $\{key : value\}$ format, along with the index of each centerline associated. All coordinates are transferred to the ENU coordinate systems, consistent within each segment but distinct between segments. For safety and privacy reasons, reference points are not provided.

How many instances are there in total (of each type, if appropriate)? MapDR is composed of 10,000 newly collected traffic scenes with over 400,000 front-view images, containing more than 18,000 lane-level driving rules.

Are relationships between individual instances made explicit? The frames in a single video clip are continuous in time with a uniform spatial distribution. All video clips are collected among different time periods with consistent capture equipment and vehicles

Are there recommended data splits (e.g., training, development/validation, testing)? We have partitioned the dataset into two distinct splits: training and testing.

Is the dataset self-contained, or does it link to or otherwise rely on external resources? MapDR is totally newly

collected and self-contained. Front-view images are captured and all the vectors are generated by our vectorized algorithm. All driving rules and correspondence are manually annotated.

D.5. Collection Process

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? Based on our HD map annotation scheme and annotation team, we have provided high-quality annotations with the help of experienced annotators and multiple validation stages.

D.6. Use

What (other) tasks could the dataset be used for? MapDR focus on the primary task of integrating driving rules from traffic signs to vectorized HD maps, which can be divided into two distinct sub-tasks: rule extraction and rule-lane correspondence reasoning. Researchers can also adapt to other traffic scene tasks.

E. Dataset Production

E.1. Data Production Pipeline

Data Collection. Search and Retrieval: We use our own database to locate the GPS coordinates of traffic signs, utilizing both text-based and image-based retrieval methods. Route Planning: Our path planning algorithm is employed to design data collection routes. Vehicles equipped with data collection devices gather raw data, including images, camera parameters, and pose information, which are then uploaded to the cloud.

Data Processing: Vectorization. In the cloud, BEV (Bird’s Eye View) perception algorithms are applied to generate vectorized local HD maps. Key point detection and matching algorithms are used to recover the 3D positions of traffic signs.

Rule Extraction. For each set of multiple image frames containing traffic signs, the most representative frame is selected for rule extraction by annotators. Vectorized map results are provided for annotating rule-lane associations. All captured images and the projection of vectorized maps in these images are included as reference material to enhance annotation accuracy.

E.2. Annotation Process

Rule Identification. Annotators identify the number of rules on each traffic sign and group related text information corresponding to each rule.

Annotation Creation. A JSON file is created with eight properties that annotators fill based on their interpretation of the rules.

Vector Association. Each rule is associated with the vector ID corresponding to its location on the vectorized map. Unique IDs are assigned to all vectors.

Quality Assurance. Quality inspection procedures are implemented to ensure the accuracy of annotations. This includes a thorough review and rework process to correct any discrepancies.

F. Analysis of MapDR

Data&Label Composition. MapDR is organized into video clips, with each clip focusing on a single traffic sign. The raw data and annotation are provided as JSON files. We provide the detailed JSON schema of both files. Listing 1 is the JSON schema of data file (data.json). An example is as shown in Listing 2. The 3D spatial location of the traffic sign is provided by 4 points represented as *traffic_board_pose*. Vectors and their types are also provided. Additionally, camera intrinsics and pose for each frame are provided to facilitate vector visualization. Note that all coordinates have been transferred to relative ENU coordinate systems which is consistent within a clip. Listing 3 is the JSON schema of annotation file (label.json). An example is as shown in Listing 4. All pre-defined properties of driving rules are illustrated. The corresponding centerlines of each rule are annotated by the vector index. As mentioned in main submission, spatial location of the symbols and texts which represent the particular rules, referred to as semantic groups, is also provided. Researchers can optionally utilize this information.

Distribution of MapDR. Fig. 1 illustrates the diverse metadata distribution in the MapDR dataset. Upper depicts the distribution of the time period for data collection, primarily from 07 : 00 AM to 18 : 00 PM, indicating that the dataset was mainly collected during daytime. The lower displays the majority of clips containing between 30 and 45 frames.

Auxiliary Evaluation Results. We conducted separate evaluations on all traffic signs of different lane types in MapDR. As shown in Tab. 1, the results indicate that the prediction difficulty varies among different categories of traffic signs.

Potential negative societal impacts. To minimize negative societal impact, we have applied obfuscation techniques

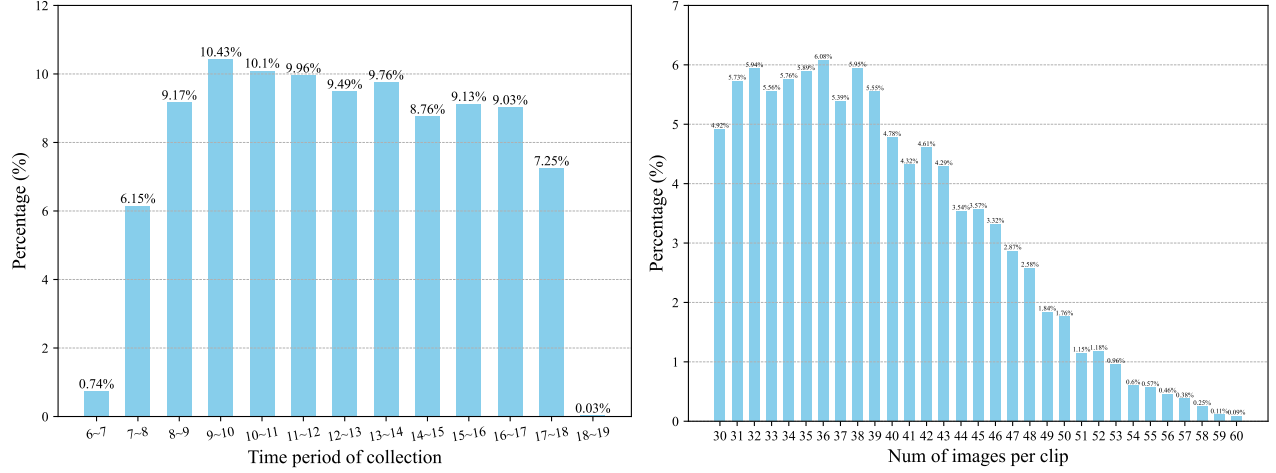


Figure 1. Metadata distribution of MapDR.

to license plate numbers, facial features, and other personally identifiable information in our dataset. Additionally, sensitive geographical locations have been excluded, and coordinates in the ENU coordinate system have been provided without reference points to safeguard privacy. However, considering the potential inaccuracies and deviation of data distribution, the model may have misinterpretations and biases during the learning process. If such models are used on public roads, it could pose safety issues. Therefore, we recommend thorough testing of models before deploying to any autonomous driving system.

G. Visualization of MapDR

Fig. 4 visualizes driving rules for different lane types in the dataset, including BEV and front-view images, as well as formatted driving rules. The red pentagram in the BEV image marks the position of the traffic sign. The front-view image displays the lane vectors and manually annotated semantic groups, with driving rules organized as sets of $\{key : value\}$ pairs. Fig. 6 shows diverse types of traffic signs collected at different times, locations, and weather conditions, demonstrating rich inter-class differences and intra-class diversity,

Table 1. Evaluation results of all traffic signs with different lane types in MapDR. The results are all based on proposed modular method, and the split of dataset remains unchanged.

Metric	BusLane	DirectionLane	EmergencyLane	VariableDirectionLane
$P_{R.E.}(\%)$	73.44	78.44	92.20	71.42
$R_{R.E.}(\%)$	71.98	77.36	91.03	57.14
$P_{C.R.}(\%)$	73.34	82.12	92.85	71.42
$R_{C.R.}(\%)$	76.76	87.03	91.00	85.71

Metric	NonMotorizedLane	VehicleLane	TidalFlowLane	MultiLane	SpeedLimitedLane
$P_{R.E.}(\%)$	80.00	88.88	0	82.09	60.34
$R_{R.E.}(\%)$	72.00	74.41	0	82.56	53.85
$P_{C.R.}(\%)$	85.41	61.90	0	81.33	88.15
$R_{C.R.}(\%)$	83.67	72.22	0	83.94	97.10

highlighting the complexity of the MapDR dataset.

H. Example for Evaluation Metric

We provide an example of metric calculation as Fig. 2 shown, illustrating the evaluation process. Given the ground truth G with 5 rule nodes and 8 centerline nodes while 6 edges between them, we assume that the algorithm has predicted \hat{G} with 6 rules and 5 edges, the metric calculation process is detailed as below.

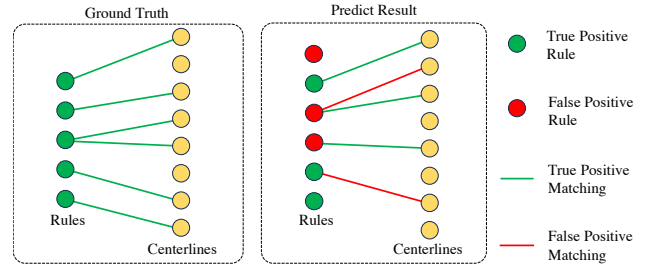


Figure 2. Illustration example for Evaluation Metrics.

First, for the **Rule Extraction from Traffic Sign** sub-task, the ground truth has 5 rules, while the algorithm predicted 6 rules, of which 3 are correct (green circles) and 3 are incorrect (red circles). Then the precision ($P_{R.E.}$) and recall ($R_{R.E.}$) are calculated as Eq. (1):

$$P_{R.E.} = \frac{|\hat{R} \cap R|}{|\hat{R}|} = \frac{3}{6} \quad R_{R.E.} = \frac{|\hat{R} \cap R|}{|R|} = \frac{3}{5} \quad (1)$$

Next, for the **Rule-Lane Correspondence Reasoning** task, there are 6 association results in the ground truth, but

the algorithm predicted 5, with 3 being correct (green lines) and 2 being incorrect (red lines). Then, the precision ($P_{C.R.}$) and recall ($R_{C.R.}$) are calculated as Eq. (2):

$$P_{C.R.} = \frac{|\hat{E} \cap E|}{|\hat{E}|} = \frac{3}{5} \quad R_{C.R.} = \frac{|\hat{E} \cap E|}{|E|} = \frac{3}{6} \quad (2)$$

Finally, considering the entire task, in the ground truth, a total of 6 lanes are assigned driving rules. The model predicted driving rules for 5 lanes, with correct predictions for both the association relationship and driving rules for only 1 lane. Therefore, the precision (P_{all}) and recall (R_{all}) for the entire task are calculated as Eq. (3):

$$P_{all} = \frac{|\hat{G}^s \cap G^s|}{|\hat{G}^s|} = \frac{1}{5} \quad R_{all} = \frac{|\hat{G}^s \cap G^s|}{|G^s|} = \frac{1}{6} \quad (3)$$

I. Implementation Details

All experiments utilizing the modular approach are conducted on 8 NVIDIA V100 16G GPUs, whereas the end-to-end approach experiments are performed on 8 NVIDIA RTX A6000 48G GPUs. We utilize pre-trained weights of DeiT [10] and BERT [3] to initialize the modular model in our experiments. Assets of DeiT and BERT are licensed under the Apache-2.0 license. Pre-trained weights of Qwen-VL-Chat [2] are employed to initialize the end-to-end model and the weights are under Tongyi Qianwen license. Additionally, we have adopted ALBEF [5] and Qwen-VL as our code base, which are available under the BSD 3-Clause and Tongyi Qianwen license respectively.

I.1. Heuristic approach

We design the heuristic method based on OCR character matching and nearest lane association. Specifically, we first perform OCR detection for the sign images then predict the values corresponding to different properties in the driving rule based on the presence of specific text or symbols in the OCR detection results. For example, if there is a "bus" symbol in the OCR result the "LaneType" property will be predicted as "BusLane", meanwhile the "AllowedTransport" property will be predicted as "Bus". If a text line contains purely numeric text similar to time or speed limits, its format is used to determine whether it represents the value of "EffectiveTime", "HighSpeedLimit" and "LowSpeedLimit". For C.R., we calculate the shortest distance between each centerline in the local vectorized HD map and the sign coordinates, selecting the nearest distance as the corresponding centerline associated with the rule.

The experiment result of heuristic method indicates that relying solely on OCR and heuristics is insufficient for this complex task, which requires a more sophisticated approach integrating image features, OCR results, and layout analysis. We agree that while this method offers some insight, however it lacks long-term research value.

I.2. Vision-Language Encoder (VLE)

Hyperparameters and Configurations. We conduct $lr = 1e - 4$, $warmup_lr = 1e - 5$, $decay_rate = 1$, $weight_decay = 0.02$, $embedding_dim = 768$, $momentum = 0.995$, $alpha = 0.4$, $attention_heads = 12$, and $batch_size = 32$ for all experiments. We initialize vision encoder with pre-trained weight of DeiT [10], text encoder and fusion encoder with the first 6 layers and last 6 layers of BERT [3], respectively. The fine-tuning epoch is set to 50. Input image is resized to 256×256 . The maximum number of tokens for input in the text encoder is 1000. *RandomAugment* is used, with hyperparameters $N = 2$, $M = 7$, and it includes the following data augmentations: "Identity", "AutoContrast", "Equalize", "Brightness", "Sharpness".

Clustering head. We calculate the cosine similarity between the [STC] tokens to determine if they represent the same rule. The training procedure is supervised by *Contrastive Loss*. The positive margin is set to 0.7, and the negative margin is set to 0.3.

Understanding head. For properties in each rule, we prefer to classify their value into pre-defined classes. Specifically, for "RuleIndex", "LaneType", "AllowedTransport", "EffectiveDate" we employ linear layer to perform classification with *Cross-Entropy Loss*. For "LaneDirection", this property is predicted by a multi-label classification that direction is defined as a combination of multi-choice from ["None", "Forbidden", "GoStraight", "TurnLeft", "TurnRight", "TurnAround"]. The training loss is *Binary Cross-Entropy Loss*. Additionally, properties of "EffectiveTime", "LowSpeedLimit" and "HighSpeedLimit" are formed as *string*. In practice, we classify the [STC] token to determine whether the OCR text is time or speed and use the original OCR text as the predicted value of these three properties.

I.3. Map Element Encoder (MEE)

Hyperparameters and Configurations. We conduct $lr = 1e - 4$, $warmup_lr = 1e - 5$, $decay_rate = 1$, $weight_decay = 0.02$, $embedding_dim = 768$, $momentum = 0.995$, $alpha = 0.4$, $attention_heads = 12$, and $batch_size = 48$ for all experiments. We train MEE from scratch, the training epoch is set to 120. The maximum number of tokens for input in the vector encoder is 1000. The formatted rule is mapped to a 768-dimensional

vector by an MLP. Specifically, each property in the rule is mapped to a 768-dimensional vector (except for "*EffectiveTime*", "*LowSpeedLimit*" and "*HighSpeedLimit*"), and the position of the traffic sign is also mapped to a 768-dimensional vector through a position encoding method (as described in the main submission), and finally, all these vectors are added together to obtain the final feature of the rule. In MEE, there are a total of four types of embeddings: vector embedding, position embedding, type Embedding, and instance embedding. The encoding method for vector embedding and position Embedding is detailed in the main submission. For type embedding, as there are 5 types in total, we initialize it using *nn.Embedding*, with the hyperparameters *num_embeddings* = 5 and *embedding_dim* = 768. Similarly, we also use *nn.Embedding* to initialize the instance embedding, with the *num_embeddings* = 120 and *embedding_dim* = 768, meaning it can support a maximum of 120 vectors. It is important to note that since the instance embedding is only used to distinguish different vectors, we shuffle the order of these embeddings at each iteration. After the multimodal fusion encoder of MEE, we further incorporate an *nn.Linear* to map the 768-dimensional features to 256, which is then connected to the association head.

Association head. We perform binary classification on [VEC] tokens to determine whether the vector is corresponding to the input rule. The training procedure is supervised with *Binary Cross-Entropy Loss*.

I.4. Analysis of Evaluation Error

We conduct multiple experiments on proposed modular approach with various random seed, and the experimental results are shown in Fig. 3. We repeated all experiments 5 times with various seed which are depicted in different colors. We uniformly sampled 100 points within the range of 0 to 1 as the binary classification threshold for association head in correspondence reasoning procedure, and then calculate the P_{all} and R_{all} for each threshold. The mean fitted line is shown in black, demonstrating the stability of our method. Specifically, we calculated the standard deviation of all evaluation metrics at a fixed threshold among different random seeds. For rule extraction sub-task, the standard deviation of $P_{R.E.}$ and $R_{R.E.}$ are 0.32 and 0.38. In the rule-lane correspondence reasoning sub-task the standard deviations are 0.07 and 0.38 for $P_{C.R.}$ and $R_{C.R.}$. Overall, the standard deviations of P_{all} , R_{all} and AP are 0.18 0.10 and 1.07, respectively.

J. Qualitative results of MLLM

We qualitatively evaluated the zero-shot performance of existing MLLMs on the two subtasks of **Rule Extraction**

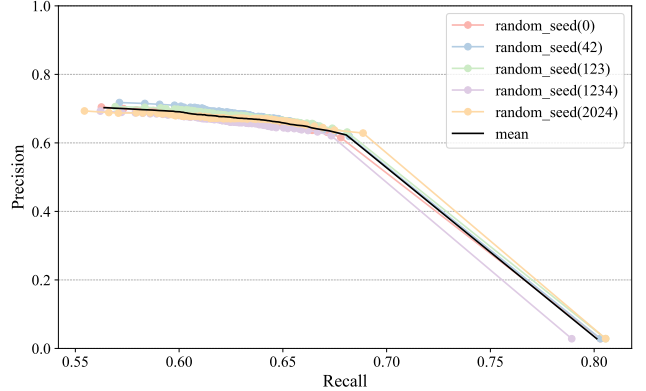


Figure 3. Overall P-R curves with various random seeds.

and **Correspondence Reasoning** using a subset of MapDR, which consists of 20 randomly sampled examples for traffic signs among all lane types, totaling 180 cases. Annotators subjectively assessed the correctness of MLLM outputs. Since MLLMs cannot provide confidence scores for their predictions, we could not use a threshold to calculate precision and recall metrics. Therefore, we evaluated $Acc_{R.E.}$ and $Acc_{C.R.}$ denotes the accuracy of R.E. and C.R. on the board-level. The evaluation reflect whether the model can interpret all the rules within a traffic sign and associated with correct centerlines, as shown in Tab. 2.

Table 2. **Zero-shot accuracy on the subset of MapDR.** MLLMs are subjectively evaluated by annotators, so the results only approximately reflect their capacity.

Model	$Acc_{R.E.}(\%)$	$Acc_{C.R.}(\%)$
Qwen-VL Max [2]	44.4	20.6
Gemini Pro [9]	31.1	6.1
Claude3 Opus [1]	4.4	1.1
GPT-4V [6]	3.3	1.7

All existing MLLMs are evaluated without SFT, clearing former memories before each prompt to avoid contextual influence. This experiment primarily aims to qualitatively analyze the zero-shot capacity of MLLMs in traffic scene understanding, rather than a rigorous quantitative comparison. Overall, the results highlight the necessity of this task and dataset.

As all the traffic signs and rules are from China, described in Chinese, we utilized a Chinese prompt. In Fig. 7, we present our input, including the image and prompt, along with the results generated by MLLMs. Our prompt can be translated as: "What is the meaning of the traffic sign in the red box? In this picture, the red lines represent the lane centerlines, which centerline or centerlines are related to the traffic sign in the red box?". The use of a Chinese prompt

may also contribute to Qwen-VL’s better performance, as it originates from Alibaba, a Chinese company, and its training process involved more Chinese text compared to other models [2].

Additionally, we referenced [8] to mark the red boxes and red lines in the images as visual prompts for the signs of interest and the centerlines of the lanes, which is convenient but may not be the most effective method and may also limit the performance of MLLMs. Furthermore, according to [7], we can learn that apart from the Qwen-VL model, other models such as GPT-4V have weak capabilities in Chinese OCR, so this possibly limit their cognitive performance. Overall, despite MLLMs’ zero-shot performance not achieving remarkable results, they possess significant potential. We believe that with further prompt optimization, the implementation of SFT, and other methods, larger models will undoubtedly achieve improved results in the future.

```

json_schema = {
  "$schema": "http://json-schema.org/draft-07/schema#",
  "type": "object",
  "properties": {
    "traffic_board_pose": {
      "type": "array",
      "minItems": 4,
      "maxItems": 4,
      "items": {
        "type": "array",
        "minItems": 3,
        "maxItems": 3,
        "items": {
          "type": "number"
        }
      }
    },
    "vector": {
      "type": "object",
      "additionalProperties": {
        "type": "object",
        "properties": {
          "type": {
            "type": "string",
            "enum": ["0", "1", "2", "3", "4"]
          },
          "vec_geo": {
            "type": "array",
            "items": {
              "type": "array",
              "minItems": 3,
              "maxItems": 3,
              "items": {
                "type": "number"
              }
            }
          }
        }
      },
      "required": ["type", "vec_geo"],
      "additionalProperties": false
    },
    "camera_intrinsic_matrix": {
      "type": "array",
      "minItems": 3,
      "maxItems": 3,
      "items": {
        "type": "array",
        "minItems": 3,
        "maxItems": 3,
        "items": {
          "type": "number"
        }
      }
    },
    "camera_pose": {
      "type": "object",
      "additionalProperties": {
        "type": "object",
        "properties": {
          "tvec_enu": {
            "type": "array",
            "minItems": 3,
            "maxItems": 3,
            "items": {
              "type": "number"
            }
          },
          "rvec_enu": {
            "type": "array",
            "minItems": 4,
            "maxItems": 4,
            "items": {
              "type": "number"
            }
          }
        }
      },
      "additionalProperties": false
    }
  },
  "required": ["traffic_board_pose", "vector", "camera_intrinsic_matrix", "camera_pose"],
  "additionalProperties": false
}

```

Listing 1. Json schema of data file.

```

{
  "traffic_board_pose": [
    [6250.741478919514, -23002.897461687568, -51.60124124214053],
    [6250.767766343895, -23002.852551855587, -53.601367057301104],
    [6247.90629957122, -23005.522309921853, -53.698920409195125],
    [6247.880012146425, -23005.5672197543, -51.69879459403455]
  ],
  "vector": {
    "0": {
      "type": "2",
      "vec_geo": [
        [6222.740794670596, -22977.551953653423, -59.28851334284991],
        [6224.65054626556, -22979.753116989126, -59.31985123641789],
        [6229.777790947785, -22985.886256590424, -59.40054347272962],
        [6237.236963539255, -22995.08138003234, -59.51233040448278],
        [6242.709547414123, -23002.134314719562, -59.58363144751638],
        [6247.894389983971, -23008.135111707456, -59.648408086039126],
        [6253.242476279292, -23014.058069147195, -59.700414426624775],
        [6258.56982873722, -23020.026259167204, -59.72872495371848]
      ]
    }
  },
  "camera_intrinsic_matrix": [
    [904.9299114165748, 0.0, 949.2163397703193],
    [0.0, 904.9866120329268, 623.7475554790544],
    [0.0, 0.0, 1.0]
  ],
  "camera_pose": {
    "1710907374739989000": {
      "tvec_enu": [6217.6643413086995, -22963.182929283157, -57.714795432053506],
      "rvec_enu": [-0.2097012215148481, 0.6478309996572192, -0.6804515437189796, 0.2707879063036554]
    }
  }
}

```

Listing 2. Example of data file.

```

json_schema = {
    "$schema": "http://json-schema.org/draft-07/schema#",
    "type": "object",
    "additionalProperties": {
        "type": "object",
        "properties": {
            "attr_info": {
                "type": "object",
                "properties": {
                    "LaneType": {
                        "type": "string",
                        "enum": ["DirectionLane", "BusLane", "EmergencyLane", "MultiLane", "Non-MotorizedLane", "SpeedLimitedLane", "TidalFlowLane", "VariableDirectionLane", "VehicleLane"]},
                    "RuleIndex": {
                        "type": "string",
                        "enum": ["None", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10"]},
                    "LaneDirection": {
                        "type": "array",
                        "items": {
                            "type": "string",
                            "enum": ["GoStraight", "TurnLeft", "TurnRight", "TurnAround", "Forbidden", "None"]},
                        "minItems": 1,
                        "maxItems": 5},
                    "AllowedTransport": {
                        "type": "string",
                        "enum": ["None", "Vehicle", "Non-Motor", "Truck"]},
                    "EffectiveDate": {
                        "type": "string",
                        "enum": ["None", "WorkDays"]},
                    "EffectiveTime": {
                        "oneOf": [
                            {
                                "type": "string",
                                "enum": ["None"]},
                            {
                                "type": "string",
                                "pattern": "^(?([01]?[0-9]|2[0-3]):[0-5][0-9]$)}"],
                                "LowSpeedLimit": {
                                    "oneOf": [
                                        {
                                            "type": "string",
                                            "enum": ["None"]},
                                        {
                                            "type": "string",
                                            "pattern": "^[0-9]+$"}],
                                            "HighSpeedLimit": {
                                                "oneOf": [
                                                    {
                                                        "type": "string",
                                                        "enum": ["None"]},
                                                    {
                                                        "type": "string",
                                                        "pattern": "^[0-9]+$"}]]},
                                                "required": ["LaneType", "RuleIndex", "LaneDirection", "EffectiveTime", "AllowedTransport", "EffectiveDate", "LowSpeedLimit", "HighSpeedLimit"],
                                                "additionalProperties": false},
                                                "centerline": {
                                                    "type": "array",
                                                    "items": {
                                                        "type": "number"}},
                                                "semantic_polygon": {
                                                    "type": "array",
                                                    "minItems": 3,
                                                    "maxItems": 3,
                                                    "items": {
                                                        "type": "array",
                                                        "minItems": 3,
                                                        "maxItems": 3,
                                                        "items": {
                                                            "type": "number" }}}},
                                                "required": ["attr_info", "centerline", "semantic_polygon"],
                                                "additionalProperties": false }
}

```

Listing 3. Json schema of label file.

```

{
  "0": {
    "attr_info": {
      "LaneType": "DirectionLane",
      "RuleIndex": "1",
      "LaneDirection": ["GoStraight", "TurnLeft"],
      "EffectiveTime": "None",
      "AllowedTransport": "None",
      "EffectiveDate": "None",
      "LowSpeedLimit": "None",
      "HighSpeedLimit": "None"
    },
    "centerline": [17],
    "semantic_polygon": [
      [6250.473053530053, -23003.147903473426, -51.91421646422327],
      [6250.387053162556, -23003.22814210385, -53.56106227565867],
      [6249.308139461227, -23004.234772194584, -53.48654436563898],
      [6249.381109470012, -23004.166690932405, -51.82106907669865]
    ]
  },
  "1": {
    "attr_info": {
      "LaneType": "DirectionLane",
      "RuleIndex": "2",
      "LaneDirection": ["GoStraight"],
      "EffectiveTime": "None",
      "AllowedTransport": "None",
      "EffectiveDate": "None",
      "LowSpeedLimit": "None",
      "HighSpeedLimit": "None"
    },
    "centerline": [16],
    "semantic_polygon": [
      [6249.081411219644, -23004.446310402054, -53.45673720163109 ],
      [6249.21171480676, -23004.324736719598, -51.76890653968486 ],
      [6248.1406193206585, -23005.324072389387, -51.694388629665156],
      [6248.0546189531615, -23005.404311019807, -53.37476750060943 ]
    ]
  }
}

```

Listing 4. Example of label file.




Figure 4. Visualization of MapDR.



Figure 5. Visualization of MapDR.



Figure 6. Visualization of traffic signs.




图中红色框里的交通标牌表达了什么含义？这幅图中，用红色的线来表达车道线，和这个红色框中的交通标牌相关的车道线是哪一条或者哪几条？

红色框内的交通标志是一个公交车道指示牌。它表明在指定的时间段内（7:00-9:00 和 16:00-19:00），该车道被划分为公交车专用道。在这张图片中，有三条红色线标记的道路。根据交通标志的内容，与之相关的车道应该是**最右侧的一条红色线路**，因为通常情况下，公交车道会设置在道路的右侧。然而，请注意实际路况可能因地区和具体情况而异，因此在驾驶时应遵循当地的实际交通标志和规则。

图中红色框里的交通标志表示：公交车上下客时间为**早7点至早9点，晚16点至晚19点**。
与该交通标志相关的车道线是**虚线和实线组成的车道分界线**。

图中红色框里的交通标志表示前方道路要分流，有岔路口。与这个交通标志对应的红色车道线包括**中间两条直行的车道线**，以及右侧分流出去的斜向的**车道线**。这些车道线的位置和走向与交通标志提示的分流信息相吻合。

红色框中的交通标志是蓝底白字的指路牌，通常用于指示方向或者提供路名、地点信息。由于图片的分辨率不足，我**无法读取上面的具体文字内容**。
至于车道线，红色线被用来表示车道的划分。与红色框中标牌相关的车道线应该是最靠近该标志的那些车道线，特别是直接位于标志下方的那条车道。通常，这样的标志会指示下方车道的行驶方向，比如直行、左转、右转等。这里的红色线标记了三条车道，而**最右侧的这条车道很可能是直接与交通标志关联的车道**，给予驾驶者即将到来的路口或方向选择的信息。



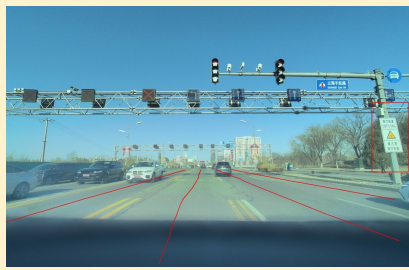
What is the meaning of the traffic sign in the red box? In this picture, the red lines represent the lane centerlines, which centerline or centerlines are related to the traffic sign in the red box?

The traffic sign in the red box is a bus lane indication. It indicates that during specified times (7:00-9:00 and 16:00-19:00), the lane is designated as a bus lane. In this picture, there are three road markings in red. Based on the content of the traffic sign, the associated lane should be the **rightmost red lane**, as bus lanes are typically positioned on the right side of the road. However, please note that actual road conditions may vary by location and specific circumstances, so drivers should adhere to local traffic signs and regulations while driving.

The traffic sign in the red box in the picture indicates that the bus boarding and alighting times are from 7:00 am to 9:00 am and from 4:00 pm to 7:00 pm. The **lane markings related to this traffic sign are composed of dashed and solid lines that delineate the lanes**.

The traffic sign in the red box in the picture indicates that the road ahead will bifurcate, and there is a fork in the road. The red lane markings corresponding to this traffic sign include the **two middle straight lanes and the diagonal lane on the right for diverging**. The positioning and direction of these lane markings align with the indication of the bifurcation provided by the traffic sign.

The traffic sign in the red box has a blue background with white lettering and is typically used to indicate direction or provide information about road names and locations. Due to the limited resolution of the image, I am **unable to read the specific text on it**. As for the lane markings, the red lines are used to delineate the lanes. The lane markings related to the sign in the red box should be those closest to the sign, particularly the lane directly below the sign. Typically, such signs indicate the direction of travel for the lane below, such as straight, left turn, right turn, etc. Here, the red lines mark three lanes, and the lane on the far right is **likely directly associated with the traffic sign**, providing drivers with information about upcoming intersections or direction choices.



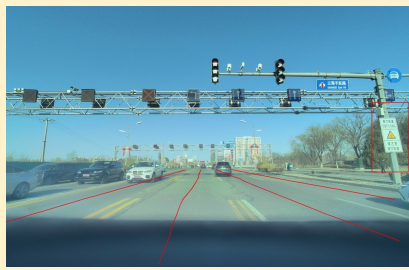
图中红色框里的交通标牌表达了什么含义？这幅图中，用红色的线来表达车道线，和这个红色框中的交通标牌相关的车道线是哪一条或者哪几条？

这张图片显示的是一个城市道路场景。在红色框内的交通标志上写着“**潮汐车道，请注意**”，并配有一个指向左侧的箭头。这意味着这条车道是一条潮汐车道，其方向会根据交通流量的需求进行调整。
关于红色线条的问题，由于**图像中没有明确指出哪些车道与该标志相关**，我们无法确定具体是哪一条或哪几条车道。但是通常情况下，在潮汐车道设置的地方，会有相应的指示牌和地面标记来引导驾驶员如何使用这些可变方向的车道。如果需要更准确的信息，可能需要查看其他角度的照片或者现场实地观察。

图中红色框里的交通标志表达了：“**该路段为潮汐车道**，在高峰时段，该车道可能只允许一个方向通行，请注意观察指示牌。”
这幅图中，用红色的线来表达车道线，和这个红色框中的交通标志相关的车道线是**实线和虚线组成的车道分界线**。

红色框中的交通标志表示该车道为专用车道，只允许公交车、出租车等特定车辆通行。
与该标志相关的车道线有两条：1. 最左侧的红色实线，表示该专用车道的左边界。2. 专用车道右侧的红色虚线，表示其他车辆可以跨越这条虚线进入专用车道，但要让已在专用车道内行驶的车辆优先通过。
所以综合起来，这个交通标志结合车道线的设置，明确规定了该车道为公交车、出租车等车辆的专用通道，其他社会车辆需要避让，且只能在必要时临时进入该车道行驶。

红色框内的交通标志是一个**右转箭头**，这表示该车道是专用于向右转弯的车道。红色线所标识的车道线显示了几个车道，而与右转弯箭头交通标志相关的车道线是靠图片右侧的，或者说道路**最右边的那一条车道线**。这条车道是为向右方向行驶的车辆准备的。



What is the meaning of the traffic sign in the red box? In this picture, the red lines represent the lane centerlines, which centerline or centerlines are related to the traffic sign in the red box?


The image depicts an urban road scene. The traffic sign within the red box reads “**HOV Lane, Watch for Changes**” with an arrow pointing to the left. This signifies that the lane is a high-occupancy vehicle (HOV) lane, and its direction may be adjusted based on traffic flow requirements.
Regarding the issue of the red lines, since the **image does not clearly indicate which lanes are associated with this sign**, we cannot determine the specific lane or lanes. However, typically in locations with HOV lanes, there are corresponding signs and pavement markings to guide drivers on how to use these variable-direction lanes. For more precise information, it may be necessary to examine photos from different angles or conduct an on-site observation.

The traffic sign within the red box in the picture indicates: “**This section is an HOV lane**. During peak hours, this lane may only allow travel in one direction. Please pay attention to the signs.”
In this image, red lines are used to represent lane markings, and the **lane markings related to the traffic sign within the red box consist of solid and dashed lines, forming lane dividers**.

The traffic sign in the red box indicates that the lane is a dedicated lane, **allowing only specific vehicles such as buses and taxis to travel**.
The lane markings related to this sign are as follows: 1. The leftmost red solid line serves as the boundary for the dedicated lane. 2. The red dashed line to the right of the dedicated lane allows other vehicles to cross into the dedicated lane, but they must yield to vehicles already traveling in the dedicated lane.
Overall, the traffic sign, in conjunction with the lane markings, clearly specifies the lane as a dedicated route for buses, taxis, and other authorized vehicles, with other vehicles required to yield and only temporarily enter the lane when necessary.

The traffic sign within the red box is a **right turn arrow**, indicating that the lane is dedicated for making right turns. The red lines denote multiple lanes, and the lane marking **associated with the right turn arrow traffic sign is the one on the right side of the image**, or the lane furthest to the right on the road. This lane is intended for vehicles traveling in

Figure 7. Prompts and answers for MLLMs.




图中红色框里的交通标表达了什么含义？这幅图中，用红色的线来表达车道线，和这个红色框中的交通标相关的车道线是哪一条或者哪几条？

红色框内的交通标志是一个自行车标志，表示这是一个**非机动车道**。在图像中，可以看到**一条红色线条标记的道路**，这条道路就是与该交通标志相关的车道，它表明这是供骑自行车或其他非机动车使用的专用通道。

图中红色框里的交通标志是**自行车道**，与自行车相关的车道线是图中**两条最左侧的红色线**。

红色框中的交通标志表示**禁止直行，只能右转**。与该标志相关的红色车道线有两条：
1. 最右侧的车道线，指引车辆只能右转进入右侧道路。
2. 中间的虚线，允许其他车道的车辆变道进入右转车道。
所以该标志结合车道线的设置，明确告知司机该车道只能用于右转，禁止直行通过路口。车辆若要右转，可以提前变道进入这条右转专用车道。

红色框中的交通标志是一个蓝底圆形的标志，里面有一个白色的箭头向上，表示**只能直行**。这个标志告诉驾驶员在当前的路口或者车道上不允许转弯，只能继续直行。
图中用红色线标示的是道路上的车道线，用以指导车辆行驶。与红色框中直行的交通标志直接相关的车道线应该包括了直行车道上的红色线条，即图中中央位置指向交通标志方向的那些线条。这表明直行的车辆应当沿着这些车道线行驶。



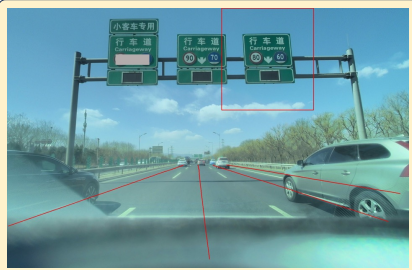
What is the meaning of the traffic sign in the red box? In this picture, the red lines represent the lane centerlines, which centerline or centerlines are related to the traffic sign in the red box?

The traffic sign within the red box is a bicycle symbol, indicating that it is a **non-motorized vehicle lane**. In the image, **there is a road marked with a red line**, which is the lane associated with this traffic sign. It indicates that this is a dedicated lane for bicycles and other non-motorized vehicles.

The traffic sign within the red box in the picture is a **bicycle lane**, and the **lane markings related to the bicycle lane are the two leftmost red lines in the image**.

The traffic sign within the red box indicates **a prohibition on proceeding straight and only allows for right turns**. There are two red lane markings related to this sign:
1. The rightmost lane marking instructs vehicles to make a right turn into the right lane.
2. The dashed line in the middle allows vehicles from other lanes to change lanes into the right-turn lane.
Therefore, this sign, in conjunction with the lane markings, clearly informs drivers that the lane is only for right turns and prohibits proceeding straight through the intersection. If a vehicle needs to make a right turn, it can change lanes in advance and enter this dedicated right-turn lane.

The traffic sign in the red box is a blue circular sign with a white arrow pointing upwards, indicating that **only straight-ahead movement is allowed**. This sign informs drivers that at the upcoming intersection or on the lane, turning is not permitted, and vehicles must continue straight.
The red lines in the image indicate the lane markings on the road, guiding the movement of vehicles. **The lane markings directly related to the straight-ahead traffic sign in the red box should include the red lines in the straight-ahead lane**, i.e., those in the central position of the image pointing in the direction of the traffic sign. This indicates that vehicles proceeding straight should follow these lane markings.



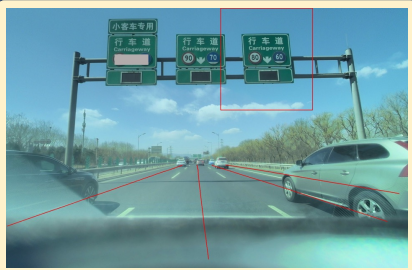
图中红色框里的交通标表达了什么含义？这幅图中，用红色的线来表达车道线，和这个红色框中的交通标相关的车道线是哪一条或者哪几条？

这张图片显示的是一个高速公路场景。红色框内的交通标志表明了**不同车道的速度限制**，最左边的绿色标志表示“小客车专用”，中间的标志写着“90”和“70”，右边的标志写着“80”和“60”。这些数字代表了该车道允许的最大速度。在图像中，红色线条标记了**三条车道**。根据交通标志的位置和方向，可以推断出与红色框内交通标志相关的车道应该是中间和右侧的两条车道，因为左侧的标志只适用于小客车，而中间和右侧的标志则分别对应着不同的速度限制。所以，如果车辆行驶在这两条车道上，则需要遵守相应的速度限制。

图中红色框里的交通标志表示的是：右侧两条车道是行车道，**限速分别为70km/h和60km/h，与之相关的车道线是红色虚线**。

图中红色框里的交通标志是**行车道选择标志**。这种标志通常用来预告前方路口处可供选择的行车道方向。
与该标志相关的车道线是中间和右侧两条红色实线之间的那条车道线。该车道对应标志中的右转和直行箭头，表示这个车道可用于右转或直行通过前方路口。

图中红色框中的交通标志是指示标志，用以指导驾驶员所处的车道可以前往的目的地以及该车道的**最低与最高限速**。右侧两个框内的标志显示的是最高车速为90公里/小时，最低车速为70公里/小时；左侧的框内的标志显示的是最高车速为60公里/小时，最低车速为40公里/小时。同时标志上方有方向箭头，分别指示前方道路的走向。与这些标志相关的车道线是与交通标志上方箭头方向一致的车道。红色的线标示了这些车道。在图片中我看不清具体车道，但一般来说，右侧的标志对应的车道线会是图片中标记红线右侧的车道，而左侧的车道则会对应于左侧的标志中的信息。



What is the meaning of the traffic sign in the red box? In this picture, the red lines represent the lane centerlines, which centerline or centerlines are related to the traffic sign in the red box?

The image depicts a highway scene. The traffic signs within the red box indicate **the speed limits for different lanes**: the green sign on the far left indicates “passenger cars only”, the middle sign shows “90” and “70”, and the right sign shows “80” and “60”. These numbers represent the maximum speeds allowed for each lane. In the image, **red lines mark three lanes**. Based on the position and direction of the traffic signs, it can be inferred that the lanes related to the traffic signs within the red box should be the middle and right lanes. This is because the left sign applies only to passenger cars, while the middle and right signs correspond to different speed limits. Therefore, vehicles traveling in these two lanes should adhere to the respective speed limits.

The traffic sign within the red box indicates that the right two lanes are travel lanes with speed limits of 70km/h and 60km/h, respectively. **The related lane markings are the red dashed lines**.

The traffic sign within the red box is a **lane selection sign**. This type of sign is typically used to indicate the available choices of travel lane direction at upcoming intersections. **The lane markings related to this sign are the lane between the middle and right red solid lines**. This lane corresponds to the right-turn and straight-ahead arrow in the sign, indicating that this lane can be used for either right turns or for proceeding straight through the upcoming intersection.

The traffic signs within the red boxes are directional signs that indicate the destination the driver can reach from the lane and the **minimum and maximum speed for that lane**. The sign in the two boxes on the right indicate a maximum speed of 90 kilometers per hour and a minimum speed of 70 kilometers per hour. The sign in the left box indicates a maximum speed of 60 kilometers per hour and a minimum speed of 40 kilometers per hour. Above the signs, there are directional arrows indicating the direction of the road ahead. **The lane markings related to these signs are the lanes aligned with the arrows above the signs**. The red lines indicate these lanes. In the picture, it's not clear which specific lanes correspond to the signs, but generally, the lane corresponding to the sign on the right will be to the right of the red lines marked in the picture, and the left lane will correspond to the information in the sign on the left.

Figure 8. Prompts and answers of MLLMs.

References

- [1] Anthropic. Claude-3. <https://www.anthropic.com/news/claude-3-family>, 2024. 5
- [2] 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. 4, 5, 6
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 4
- [4] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021. 1
- [5] 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. 4
- [6] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. 5
- [7] Miao Rang, Zhenni Bi, Chuanjian Liu, Yunhe Wang, and Kai Han. Large ocr model: An empirical study of scaling law for ocr. *arXiv preprint arXiv:2401.00028*, 2023. 6
- [8] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*, 2023. 6
- [9] 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. 5
- [10] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 4