SignEye: Traffic Sign Interpretation from Vehicle First-Person View

Chuang Yang, Xu Han, Tao Han, Yuejiao Su, Junyu Gao Hongyuan Zhang, Yi Wang, and Lap-pui Chau*, *Fellow, IEEE*

***Abstract*—Traffic signs play a key role in assisting autonomous driving systems (ADS) by enabling the assessment of vehicle behavior in compliance with traffic regulations and providing navigation instructions. However, current works are limited to basic sign understanding without considering the egocentric vehicle's spatial position, which fails to support further regulation assessment and direction navigation. Following the above issues, we introduce a new task: traffic sign interpretation from the vehicle's first-person view, referred to as TSI-FPV. Meanwhile, we develop a traffic guidance assistant (TGA) scenario application to re-explore the role of traffic signs in ADS as a complement to popular autonomous technologies (such as obstacle perception). Notably, TGA is not a replacement for electronic map navigation; rather, TGA can be an automatic tool for updating it and complementing it in situations such as offline conditions or temporary sign adjustments. Lastly, a spatial and semantic logic-aware stepwise reasoning pipeline (SignEye) is constructed to achieve the TSI-FPV and TGA, and an application-specific dataset (Traffic-CN) is built. Experiments show that TSI-FPV and TGA are achievable via our SignEye trained on Traffic-CN. The results also demonstrate that the TGA can provide complementary information to ADS beyond existing popular autonomous technologies.**

***Index Terms*—Traffic sign interpretation, autonomous driving, intelligent transportation, egocentric vision.**

I. Introduction

VISION-based autonomous driving requires perceiving the road environment around vehicles to support decision-making regarding driving plans. Traffic signs (including guide panels, symbols, and text), key components of the road scene, contain traffic regulation and navigation information, which helps formulate a driving plan that adheres to road driving criteria. Traditional traffic sign-related methods primarily focus on the basic traffic symbol (such as the speed limitation and no parking symbols) detection and recognition [1], [2].

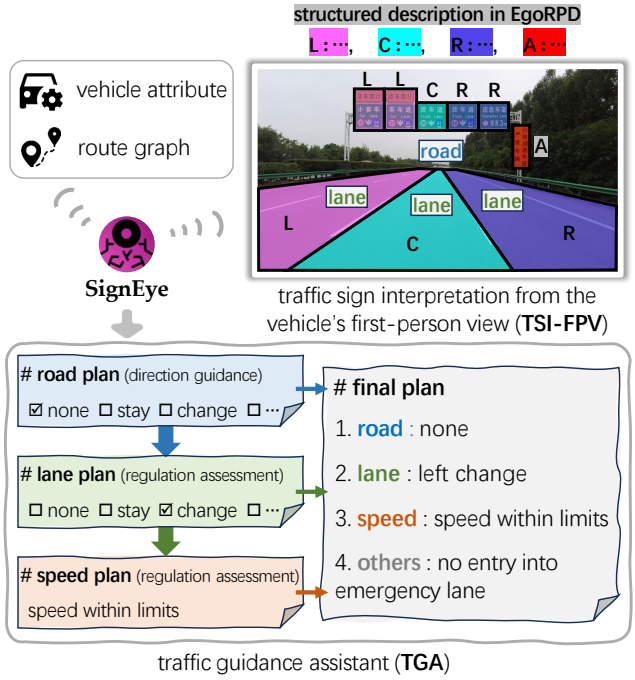


Fig. 1: Illustration of the SignEye, where the TSI-FPV part interprets sign units as structured descriptions and assigns them to different roads and lanes in egocentric relative position definition (EgoRPD), where "C" for current, "L" for left, "R" for right, and "A" for all lanes or roads. The TGA part combines the descriptions with vehicle attributes and a route graph to achieve traffic regulation assessment (e.g., speed limitations, lane restrictions, etc.) and direction navigation.

Although recent approaches [3]–[5] attempt to combine traffic symbols and texts for comprehensive understanding, the lack of consideration for the egocentric vehicle's spatial position makes them hard to establish a connection between sign information and the vehicle. It leads to current methods still struggling to support autonomous driving systems (ADS) for traffic regulation assessment and direction navigation.

Based on the above observations, we propose to interpret traffic signs from the vehicle's first-person view (TSI-FPV) to assist ADS in decision-making. Specifically, as shown in Figure 1, TSI-FPV **firstly** generates a description for each traffic **sign unit** (a sign set, which includes multiple symbols and texts that need to be combined together to interpret complete traffic information) according to the Road Traffic Signs and Markings Criteria. The interpretation process from

The work described in this paper was conducted in the JC STEM Lab of Machine Learning and Computer Vision funded by The Hong Kong Jockey Club Charities Trust, and was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 15215824).

Chuang Yang, Yuejiao Su,Yi Wang, and Lap-pui Chau are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR.

Xu Han, Tao Han, Junyu Gao, and Hongyuan Zhang are with the School of Computer Science, and with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

E-mail: omtcyang@gmail.com, hxu04100@gmail.com, hantao10200@gmail.com, yuejiao.su@connect.polyu.hk, giy3035@gmail.com, hyzhang98@gmail.com, yi-eie.wang@polyu.edu.hk, lap-pui.chau@polyu.edu.hk.

Lap-pui Chau is the corresponding author.

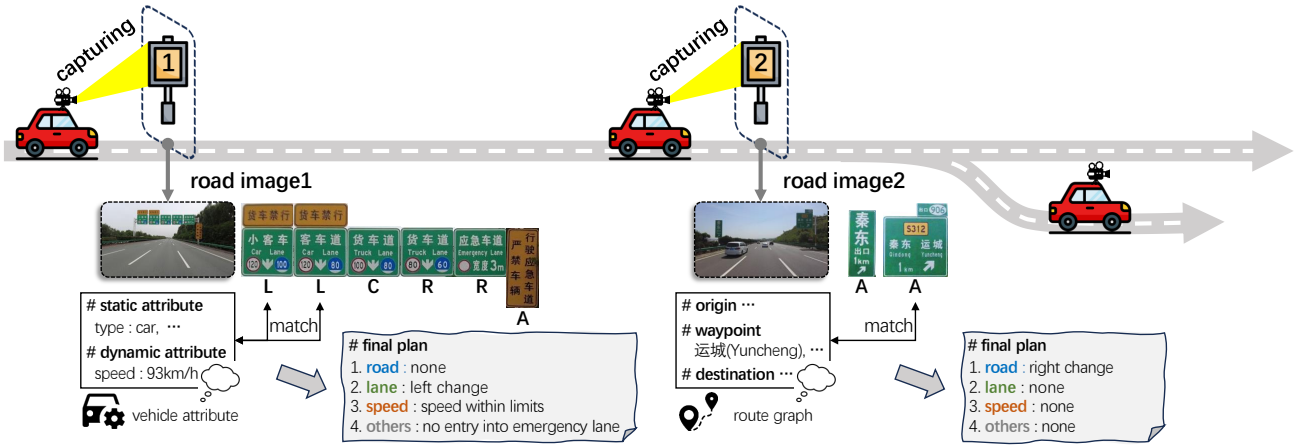


Fig. 2: TGA estimates the structured sign descriptions in EgoRPD the vehicle attribute and route graph for achieving traffic regulation assessment and direction navigation respectively.

a sign unit to a description organizes semantic logic among signs within a unit accurately and naturally. It ensures reliable sign-understanding results. Meanwhile, the description-styled output enjoys intuitional and structured sentence patterns, which makes it easier for ADS or human drivers to take in to further assist in driving plan decision-making. TSI-FPV **then** distinguishes road and lane markings from the vehicle's first-person view. Lanes within one road are labeled as left ('L'), current ('C'), and right ('R'), where the label 'C' indicates the vehicle's located position from an egocentric view, label 'L' and 'R' are defined according to the determined label 'C'. Similarly, roads are labeled in the same way. Compared with the normal absolute strategy (determining the positional order of all lanes or roads from left to right), the introduced egocentric relative position-definition (EgoRPD) strategy only requires considering lanes around the vehicle, which avoids the recognition interference brought by perspective distortion of the distant lane and can provide adequate adjacent lane information for single-step lane change (a recommended practice under Road Traffic Safety Law) decision-making process at the same time. **Lastly**, TSI-FPV distributes the descriptions of all sign units in the image to different roads and lanes based on the analysis of the spatial relationship between units, roads, and lanes and the directional information involved in the descriptions. Notably, on the current road, a description is labeled as 'A' if the corresponding sign unit regulates all lanes instead of a specific one lane.

Meanwhile, to re-explore the role of traffic signs in ADS, a traffic regulation assistant (TGA) is developed in this paper (as visualized at the bottom of Figure 2). TGA introduces vehicle attribute and route graph into the driving plan decision-making process, where the vehicle attribute consists of dynamic (vehicle speed) and static attributes (the type, size, and weight of the vehicle), and the route graph contains all latent waypoints along the travel route from origin to destination. In decision-making, TGA estimates the information between TSI-FPV output and the vehicle attribute and route graph for achieving traffic regulation assessment (e.g., speed limitations, lane restrictions, etc.) and direction navigation (lane change and road change) respectively.

Furthermore, to achieve TGA, a stepwise reasoning pipeline (SignEye) is constructed, which decomposes a complex spatial and semantic logic reasoning task into a series of intuitive problems. With decomposition problems, SignEye analyzes semantic logic among signs more accurately to generate structured descriptions for the TSI-FPV task without a hallucination problem. It also can understand the spatial logic between the descriptions and lane and road more easily for further reasoning the driving plan in the TGA. Meanwhile, compared with the previous three steps (involving detection, classification, and natural language processing)-based traffic sign interpretation methods, SignEye avoids heavy dependence on a complex training process and plenty of difficult-to-obtain intensive symbol and text ground-truth data. It combines these three steps into a single process, which considers sign interpretation as image description and can be directly trained using the description of sparse sign units. Besides, considering the absence of a large available dataset to train SignEye, we build a semi-automatic data engine to generate TSI-FPV and TGA corresponding datasets, namely Traffic-CN, to fulfill the research and evaluation in this field. The contributions of this work can be summarized as follows:

1) The traffic sign interpretation task from the vehicle's first-person view (TSI-FPV) is introduced to provide sign descriptions with egocentric positional information for ADS. Additionally, a TSI-FPV-based traffic guidance assistant (TGA) scenario is developed to re-evaluate the role of traffic signs in supporting ADS decision-making for driving plans.
2) A stepwise reasoning pipeline, called SignEye, is constructed to improve semantic logic analysis among signs and to determine spatial relationships between sign units, lanes, and roads, effectively facilitating TSI-FPV and TGA. Notably, it simplifies the previously multi-step sign interpretation process into a single-step image description, allowing it to be trained directly on sparse descriptions without intensive symbol and text boxes.
3) The TSI-FPV and TGA corresponding dataset (Traffic-

CN) is built. It fulfills the research evaluation and scheme verification in the field of applying traffic signs in ADS. Meanwhile, Traffic-CN will promote the development of the traffic sign community.

The rest of the paper is organized as follows. Section II introduces the related works on traffic sign recognition method and dataset. Section III and Section IV describe the overall framework and Traffic-CN dataset. The experimental results are discussed in Section V. Section VI concludes the paper.

II. Related Work

SignEye is constructed based on sign-related expert models and vision-language models (VLMs). In this section, we review previous methods to provide a clearer understanding of the framework and its advantages.

A. Traffic Sign-Related Method

Previous works on traffic signs primarily pay attention to basic symbol detection and recognition. Initially, deep object detection frameworks [6]–[8] made the task of locating symbols from scene images and recognizing them [1], [9], [10] achievable. However, traffic signs always convey complex instructions via the combination of symbols and texts with different layouts and color styles, which would lead to information omission and even misunderstanding (e.g., the combination of speed limitation symbol and lane information) if relying merely on symbols.

Following the above considerations, recent works [3] attempted to achieve a comprehensive sign understanding. Specifically, Guo *et al.* [3]–[5] detected and recognized symbols and texts by optical character recognition models [11]–[14]. Then symbols and texts were combined according to the analysis of the corresponding layout. Further, the authors proposed a traffic knowledge graph in [4], [15], where roads and lanes were considered in the understanding process. To describe signs more intuitively, Yang *et al.* [5] formulated the traffic knowledge graph as a natural language description. Although these methods achieve a comprehensive sign understanding, the lack of consideration for the egocentric vehicle's spatial position makes them hard to provide adequate information support to ADS.

B. Vision-Language Model

Traffic-sign understanding is formulated as an image-to-text task over VLMs in this paper, which avoids a complex training process and plenty of difficult-to-obtain intensive symbol and text data. Recently, a representative model (CLIP) [16] was designed by OpenAI, which trained two encoders via a contrastive learning strategy for the mutual representation of text and image features. BLIP [17], [18] proposed to unify vision-language understanding and generation for achieving a wide range of vision-language tasks. MiniGPT4 [19] enhanced the model capability to achieve those vision-language tasks by bridging between visual modules and LLMs. LLaVA [20] and LLaVA1.5 [21] had further achieved significant progress by equipping themselves with more advanced LLMs.

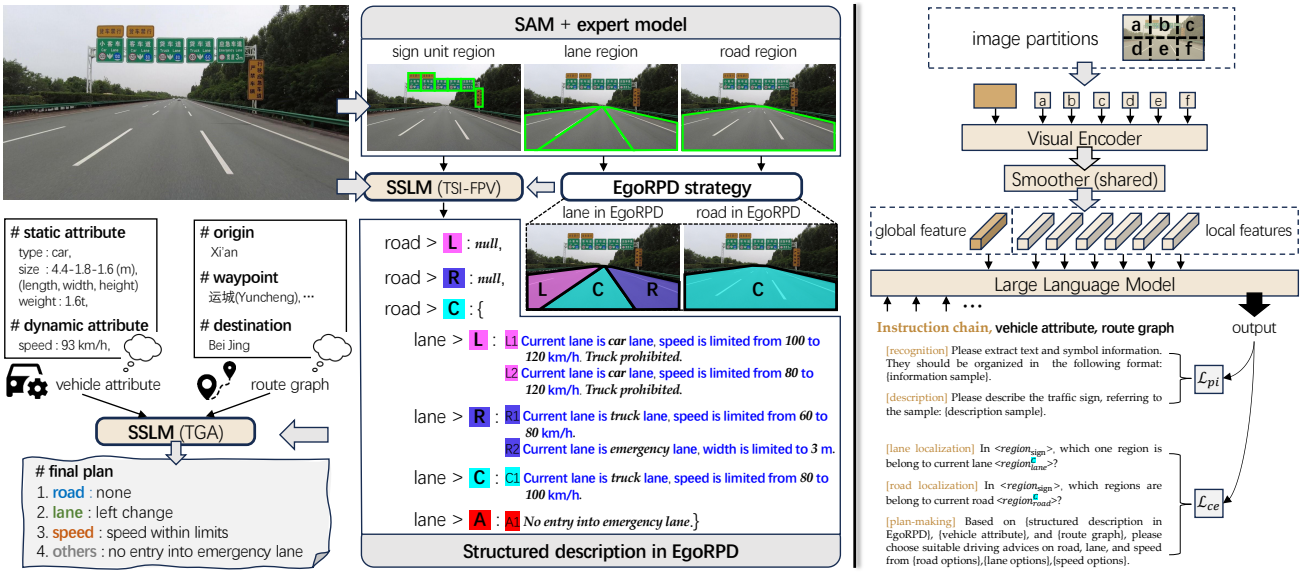
Since input image resolution is an important factor for VLMs improving their capabilities to describe images, a lot of researches concentrate on resolution improvement. Qwen-VL [22] increased the input resolution that supports encoding to 448. Fuyu-8B [23] adopted different-sized images as pre-training supervision signals, which helped OtterHD [24] further improve the image resolution. Monkey [25] ensured high resolution by extracting global features from the original image along with fine local details. To achieve a better description for an image with multiple varied objects, GeoChat [26], RegionGPT [27], and ASM [28], [29] formulated region-level description as an image-level detailed interpretation. However, such general-domain VLMs perform poorly for first-person view traffic sign scenarios, leading to inaccurate or superfluous and irrelevant information when presented with traffic sign-specific queries. Such a behavior emerges due to the unique challenges introduced by traffic sign interpretation of the vehicle's first-person view.

III. Method

In this paper, a stepwise reasoning pipeline (SignEye) is constructed to achieve TSI-FPV and TGA. We will describe the details of SignEye in this section.

A. Overall Pipeline of SignEye

Existing representative VLMs (such as BLIP2 [18], MiniGPT-4 [19], LLaVA [20], Qwen-VL [22], Yi-VL [30], DeepSeek-VL [31], Intern-VL [32], MiniCPM [33], Monkey [25]) focus on pre-training their models on large datasets and aim to achieve a strong ability on general image description. However, the generated description from them is either too long or too short for the introduced TSI-FPV task, which leads to information redundancy or omission and further interferes with the final decision-making about the driving plan. Besides, the understanding of the latent spatial and semantic logic among signs and the position information around the vehicle's first-person view is important for TGA. Specifically, it is found that different traffic signs always correspond to different lanes and roads, which means the determination of the current vehicle's lane and road locations is essential for interpreting traffic sign information to provide support for automatic driving decision-making. Based on the above observation, we start from the first-person view of the driving vehicle to identify the corresponding lane and road locations and build a connection between the location and the traffic sign through the proposed EgoRPD strategy by explicitly modeling the positional information of both entities and the geometric relationship between them. This helps our method achieve a strong ability to interpret traffic signs when combined with the corresponding location information. However, existing VLMs take the whole image as input and interpret traffic signs directly without analyzing the relationship between the vehicle location and traffic signs, which makes it hard for VLMs to identify the correct sign that matches the vehicle's location and leads to information deficiency in automatic driving decisions. As a result, even when these models are well-trained for describing general images, they struggle to accurately transfer



(a) the stepwise reasoning pipeline (**SignEye**) for TSI-FPV and TGA

(b) spatial and semantic logic-aware model (**SSLM**)

Fig. 3: Overall pipeline of SignEye. It takes a road image from the vehicle's first-person view, and object (sign, lane, and road) regions as input to generate structured sign descriptions in EgoRPD and combines them with vehicle attributes and route graphs to achieve the TGA scenario, automatically.

traffic sign units to structured descriptions from the first-person view. This limitation affects the models to effectively assist in driving plan decision-making.

To this end, our SignEye takes a road image from the vehicle's first-person view, and object (sign, lane, and road) regions as input to generate sign unit descriptions and combines them with vehicle attributes and route graphs to achieve the TGA scenario, automatically. It can assist ADS as a complement to current popular autonomous technologies (such as obstacle perception [34]–[36]) to provide extra traffic-sign-based traffic regulation assessment and direction navigation assistant beyond current popular technologies. Specifically, as shown in Figure 3(a), SignEye adopts several advanced models as components and follows multiple steps to achieve TSI-FPV and TGA:

Region proposal generation. SAM [37] and expert models (i.e., SignDet [5] and UFLD [38]) are adopted for generating the region proposals of roads, sign units, and lanes, respectively. SAM, a strong segmenter, is finetuned through the road data in RS10K [4] to provide road masks in the pipeline. Considering there are lots of adhesive samples among sign units and lanes while the point or box prompt of each instance is expensive to obtain, the SignDet [5] and UFLD [38] are chosen to replace SAM for the responsibility to extract the regions of sign units and lanes in the inference process, where SignDet is trained on our Traffic-CN dataset and UFLD is optimized on both training and test data of Tusimple [39].

Structured and unstructured sign description. Different from formulating traffic sign unit as a directional combination [3] or traffic knowledge graph [4], we describe it via the structured natural language (as shown in Table I) according to the Road Traffic Signs and Markings Criteria. The language description-styled output enjoys intuitional and structured sentence patterns, which makes it easier for the model to take in and further assist in driving plan decision-making for TGA. Importantly, structured descriptions are the key to generating plenty of TGA samples to help our model learn to accomplish traffic regulation assessment and direction navigation (details can be referred to in Section IV). Compared with description generation based on text information [5], our spatial and semantic logic-aware model (SSLM) combines vision and text features to describe a sign unit, which helps focus on the spatial and semantic logic simultaneously for understanding the traffic symbol and text more effectively. Specifically, SSLM decomposes the describing process into two steps: (1) extracting traffic symbol and text information from the given traffic sign unit and determining the corresponding description structure from Table I based on information; (2) describing the sign unit following the determined description structure. The two-step process ensures correct spatial semantic logic between different symbols and texts, which either helps prevent the model from generating a redundant description or omitting important information. The details of SSLM can be referred to the following Section III-B.

In addition to the structured descriptions outlined above, there are numerous unstructured descriptions pertaining to special circumstances. These descriptions often deviate from conventional traffic regulations and, if disregarded by drivers, can result in traffic accidents. We illustrate several representative examples in Table I. For instance, the emergency lane is designated exclusively for emergency stops, emergency vehicles, safety buffers, and accident avoidance, and is strictly off-limits to regular vehicles except in emergency situations (as depicted in the first row of the bottom sub-table in

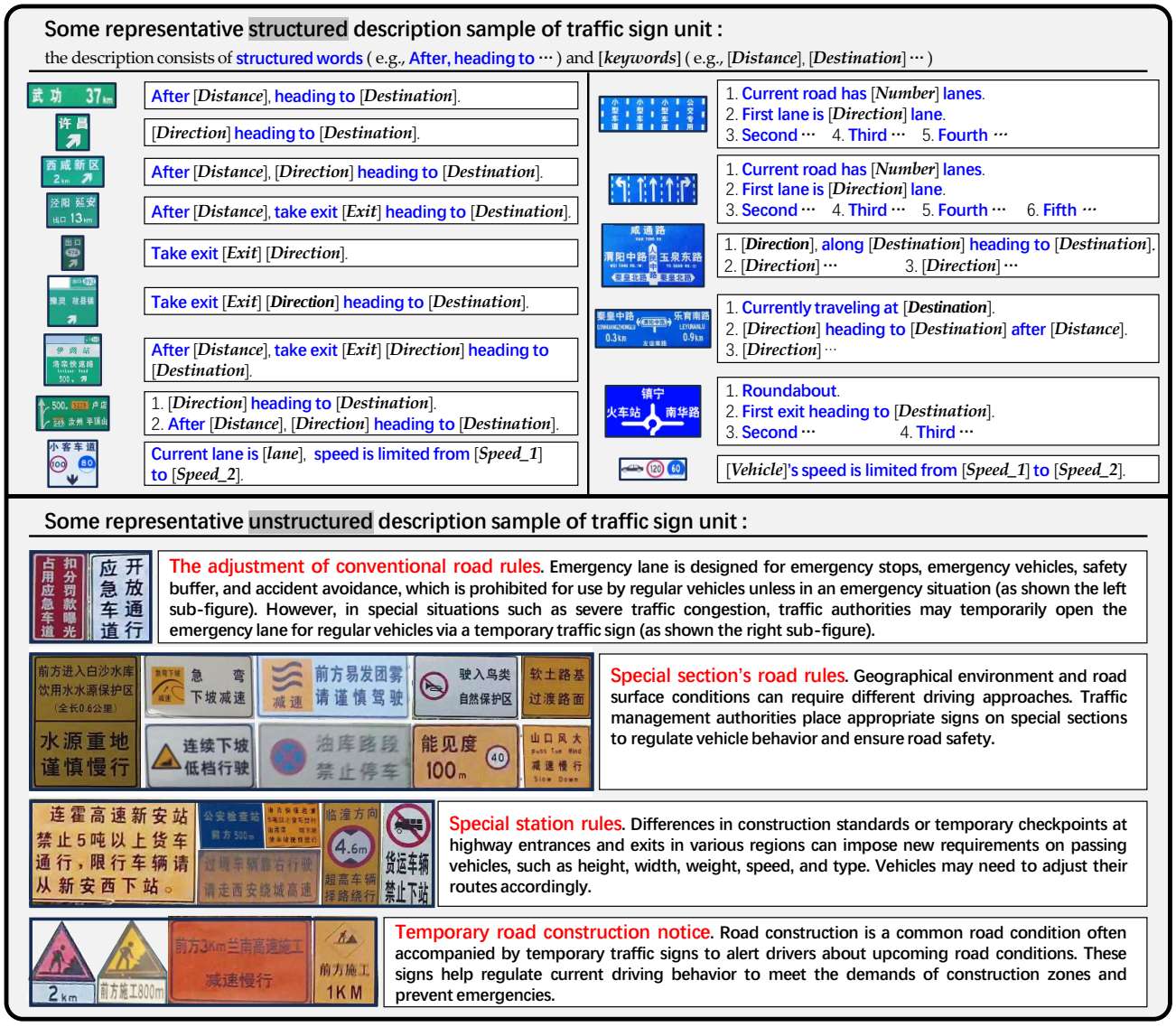


TABLE I: Illustration of structured and unstructured sign descriptions.

Table I). However, under exceptional circumstances such as severe traffic congestion, traffic authorities may temporarily permit regular vehicles to use the emergency lane, indicated by temporary traffic signs. Additionally, road construction is a frequent road condition that is typically accompanied by temporary traffic signs to warn drivers of upcoming changes. These signs are crucial for adjusting driving behavior to accommodate the requirements of construction zones and to prevent potential emergencies (as shown in the last row of the bottom sub-table in Table I).

Structured description in EgoRPD. Previous sign-related methods did not establish a connection between signs and the egocentric vehicle's position, which results in those methods could not provide effective sign information for ADS. Following the above issues, we introduce the EgoRPD strategy from the vehicle's egocentric view to determining the vehicle's corresponding lane and road positions, which helps connect egocentric vehicles and related signs to support ADS achieve traffic regulation assessment and direction navigation for the current vehicle. Importantly, the EgoRPD strategy only requires considering lanes around the vehicle, which helps avoid the recognition interference brought by perspective distortion of the distant lane and can provide adequate adjacent lane information for single-step lane change process.

Specifically, the strategy firstly takes lane and road regions as input and assigns them to left ('L'), current ('C'), right ('R'), and all ('A') parts, respectively (referred to Algorithm 1 r1–r6). It then determines whether a sign unit is a lane-level or road-level traffic guidance according to the [***keywords***] of the corresponding structured description from SSLM (as the introduction in Section III-A). Next, SSLM assigns all units to the corresponding road and lane according to their relative position analysis (Algorithm 1 r7–r15). In this stage, SSLM formulates the assignment process as a single-choice question and a multi-choice question for lane and road, respectively, because only one sign unit corresponds to one lane, and there may be multiple units related to one road. It takes the image features, and the coordinate-style sign regions, lane

Algorithm 1 EgoRPD Strategy

Require: Points of lane or road lines $\mathbf{S}$ and sign boxes $\mathbf{B}$;
Ensure: Sign boxes $\mathbf{B}_e$ in EgoRPD;
for $i \leftarrow 1$ to $(N_S - 1)$ **do** *//N_S is line number*
  $\theta_{i-1}, \theta_i \leftarrow$ ARCTAN(POLYFIT($\mathbf{S}[i-1]$;$\mathbf{S}[i]$))
  if $\theta_{i-1} <= 90$ and $\theta_i <= 90$ **then** break
  end if
end for
$\mathbf{S}_e^L, \mathbf{S}_e^C, \mathbf{S}_e^R \leftarrow \mathbf{S}[:i], \mathbf{S}[i-1:i+1], \mathbf{S}[i:]$
for $j \leftarrow 0$ to $(N_B - 1)$ **do** *//N_B is box number*
  $d^L \leftarrow |\mathbf{B}_j^{x_{mid}} - \mathbf{S}_{e,\frac{2H}{3}}^{L,x_{mid}}|$
  $d^C \leftarrow |\mathbf{B}_{j,\frac{2H}{3}}^{x_{mid}} - \mathbf{S}_{e,\frac{2H}{3}}^{C,x_{mid}}|$
  $d^R \leftarrow |\mathbf{B}_{j,\frac{2H}{3}}^{x_{mid}} - \mathbf{S}_{e,\frac{2H}{3}}^{R,x_{mid}}|$ *//$\cdot^{x_{mid}}_{\frac{2H}{3}}$ is the x-axis midpoint of $\mathbf{S}$ two points that y equals $\frac{2}{3}$ image height H*
  if $d^L =$MIN(d^L, d^C, d^R) **then** $\mathbf{B}_e^L \leftarrow \mathbf{B}_j$
  else if $d^C =$MIN(d^L, d^C, d^R) **then** $\mathbf{B}_e^C \leftarrow \mathbf{B}_j$
  else if $d^R =$MIN(d^L, d^C, d^R) **then** $\mathbf{B}_e^R \leftarrow \mathbf{B}_j$
  end if *//$\cdot_e^C$ is the box assigned to current road/lane*
end for

TABLE II: List of options for different-level questions.

Level	Option
road	none; stay; left change; turn left; right change; turn right; exit;
lane	none; stay; left change; right change;
speed	none; speed within limits; speeding; driving too slowly;
other	none; excessive vehicle height; excessive vehicle width; excessive vehicle weight; [description];

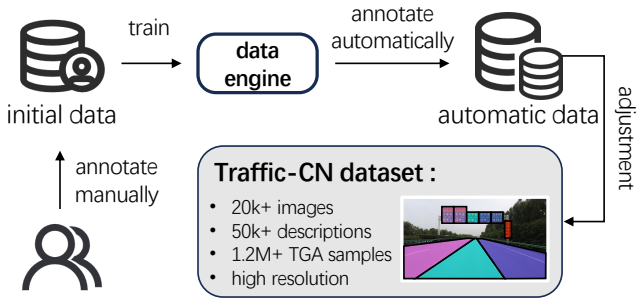


Fig. 4: Workflow of the data engine for constructing the Traffic-CN dataset.

and road regions with 'C' mark as input to achieve the the unit assignment (as shown in Figure 3(b) lane and road localization) and output the structured description in EgoRPD. Notably, units belonging to the left or right lane/road can be determined via the position in the X-axis relative to the unit with 'C' mark. The trainable data for this assignment stage is generated via our data engine based on Algorithm 1.

Traffic guidance assistant. Based on the structured description in EgoRPD, we develop the TGA to re-explore the role of traffic signs in ADS by achieving traffic regulation assessment and direction navigation for the vehicle in the first-person view. In the traffic regulation assessment aspect, vehicle attributes are introduced to simulate the vehicle's driving state, which consists of static (type, size, and weight of vehicle) and dynamic (vehicle speed) attributes. For direction navigation, a route graph is pre-defined and contains plenty of latent waypoints from the origin and destination. With the context information of description, vehicle attributes, and route graph, SSLM formulates the plan generation as a single-choice question. For example, if the current lane limits vehicle speed from 90 to 120 km/h, the SSLM will choose the option of driving too slowly when the speed is 60 km/h. Notably, the 'other' plan is mainly responsible for estimating limitations on the vehicle's height, width, and weight. Our model will output the corresponding description directly if there is no information related to the vehicle's height, width, and weight. The pre-defined options for various questions are presented in Table II, where [description] indicates that the sign descriptions are directly outputted, as they are unrelated to lane or road changing scenarios.

B. SSLM Architecture

Different from previous sign understanding works [4], [5], [15], SSLM transfers road images to structured descriptions in EgoRPD, which is essential to achieve TGA for supporting ADS to achieve traffic regulation assessment and direction navigation. SSLM is shown in Figure 3(b), which consists of a visual encoder, smoother, and a LLM-based decoder.

Visual encoder. Specifically, given an input road image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$, we divide it into 2×3 slices for support handling the high-resolution (i.e., 1344×1344 resolution) images, where H and W are the height and width of the input image. These slices then are fed into the visual encoder (the pre-trained SigLIP-ViT [40] is employed as our visual encoder) for embedding them as the corresponding features $\boldsymbol{I}_v \in \mathbb{R}^{N \times D}$, where N and D are the number of image slices and the feature hidden dimension, respectively.

Smoother. Considering the large hidden dimension D (D is set as 1096 in the pre-trained model) of the feature $\boldsymbol{I}_v$ from the visual encoder leads to a high token count, which results in extensive GPU memory consumption in the training and inference process, we insert a Smoother after visual encoder to reduce the hidden dimension of $\boldsymbol{I}_v$ from D to M ($M << D$). The smoothed visual features $\boldsymbol{I}_s \in \mathbb{R}^{N \times M}$ allows us to train and evaluate the model on RTX 4090 of 24 GB memory and achieves competitive results.

LLM-based decoder. To decode the combination features of vision and text into the natural language description, a strong open source LLM (QWen2 [41]) is adopted as our decoder. It takes a sequence of smoothed visual features $\boldsymbol{I}_s$ and pre-designed instruction tokens $\boldsymbol{T}_{ins}$ as input, generating task-specific answers.

C. SSLM Training

We employ a strategy that involves initializing the visual encoder and LLM-based decoder with pre-trained weights proposed in SigLIP [40] and QWen2 [41] and fine-tuning specific segments with the Low-Rank Adaptation (LoRA [42]) method for SSLM. In the fine-tuning process, considering even simple descriptions (e.g., examples in Table I), VLMs

often struggle to generate a structured description due to the hallucination problem. Meanwhile, VLMs hard to propose suitable advice for the TGA that is dependent on spatial relative position deeply. To this end, we follow the stepwise pipeline to achieve multi-task instruction tuning (as shown in Figure 3(b)).

For the SSLM (TSI-FPV) (the extraction and description tasks in Figure 3(a)) optimization, we formulate the process as maximizing a permutation-invariant likelihood function $\mathcal{L}_{pi}$ to estimate the difference between the input answer and the corresponding ground truth, which avoids the interference brought by the arrangement sequence of [***keywords***]. Specifically, as shown in Table I, there always more than one keywords within the same [***keywords***] position (e.g., **heading to** [***Destination1, Destination2, ...***]). The sign unit description is correct no matter how the relative position between ***Destination1*** and ***Destination2*** distribute. The permutation-invariant likelihood function is formulated as:

$$\mathcal{L}_{pi} = \min(\log P(\boldsymbol{T}_{ans}^{k} | \boldsymbol{I}_s, \boldsymbol{T}_{ins}; \Theta)), \tag{1}$$

$$k = 1, 2, ..., M! \times \prod_{j}^{M} V_j!, \tag{2}$$

where $\boldsymbol{T}_{ins}$ and Θ are the input instruction tokens and the trainable parameters. P is the conditional probability. M represents the number of sub-sentences, and V denotes the number of [***keywords***] in each sub-sentence. $\boldsymbol{T}_{ans}^{k}$ denotes the k th styled answer tokens, which is generated through the rearrangement of sub-sentences and [***keywords***] sequence of the given answer tokens $\boldsymbol{T}_{ans}$.

For the SSLM (TGA) (the localization and plan-making tasks in Figure 3(a)) optimization, we estimate the prediction via cross entropy function directly.

IV. Traffic-CN Dataset

Images. Traffic-CN collects 20k+ images with high resolution (1920×1080 pixels), of which 18K are used for training and 2K for testing. They are captured from the vehicle's first-person view through the DJI OSMO Pocket2 camera. The dataset involves some popular Chinese provinces (including Shaanxi, Henan, Shanxi, Hebei, Tianjin, and Beijing), typical road scenes (such as highways, urban roads, urban streets, and rural roads), and various challenged optical environments (e.g., overexposure, rain and fog blur, shadows, occlusion, and motion blur).

Data engine. To fulfill the research of the TSI-FPV and TGA, we build a data engine (Figure 4) to enable the label collection of the Traffic-CN dataset. For **TSI-FPV data**, the data engine follows three stages to generate structured descriptions: (1) a totally manual labeling stage for a small amount of initial data; (2) a model-assisted automatic annotation stage, where SignEye is trained on initial data to predict automatic annotations; (3) a manual adjustment stage, where partial automatic data is corrected manually and used for training model again with initial data together. In the end, repeat stage (3) for labeling all image sign units. Our data engine produced 50k+ structured descriptions for sign units, 85% of which were generated fully automatically. **TGA data** consists of localization and plan-making data (as shown in Figure 3(b)). The former is generated according to the spatial distribution among sign, lane, and road regions through Algorithm 1. The plan-making data is generated by combining structured descriptions in EgoRPD from the localization task with vehicle attributes and the route graph, where the description is decomposed into smaller traffic instructions following its structure, and vehicle attributes and waypoints are matched with the instruction for generating suitable advice from options. Based on the images, the introduced data engine generate 1.2M+ TGA data automatically. Notably, the Traffic-CN dataset comprises 20,000 images, each sourced from a video averaging 1.5 to 4 seconds in duration, with consistent traffic sign information throughout. The single-frame data supports our TSI-FPV research, while the video data offers opportunities for further exploration.

V. Experiments

A. Implementation Details

We initialize our model using the well-trained SigLIP-ViT [40] and QWen2 [41]. During the training process, we refine the parameters by employing LoRA [42] for low-rank adaptation, with the rank r set to 64. We utilize the AdamW [45] optimizer along with a cosine learning rate scheduler to train SSLM over 5 epochs using a batch size of 16 across eight NVIDIA 4090 GPUs, with a total training time of approximately 18 hours. For the sign detection module, shrink-mask expanding strategy-based detector [11] is adopted to achieve efficient sign detecting task. It takes ResNet-18 [46] as backbone and trained on a unified text detection datasets (SynthText [47] and ICDAR2019 [48]). Data augmentation is used in our work including the following three strategies: (1) random horizontal flipping, (2) random scaling and cropping, (3) random rotating. In the training stage, Adam [49] is employed with an initial learning rate of 1e-6, and the learning rate is reduced by the "poly" learning rate strategy that proposed in [50], in which the initial rate is multiplied by $\left(1 - \frac{iter}{\max\_iter}\right)^{0.9}$. For the lane detection module, we take the well-trained UFLD [38] to determine the lane and road locations. The input image is uniformly resized with aspect ratio preservation to a target dimension of 1056 in the following all experiments.

B. Main Results

Traffic Sign Interpretation (TSI) Task. This task is introduced to generate accurate and structured descriptions, aiming to facilitate the model's comprehension of traffic instruction information derived from sign units. The quality of these descriptions plays a critical role in determining the effectiveness of Traffic Guidance and Assistance (TGA) scenario applications. To evaluate the performance of our proposed approach, we conduct experiments on the Traffic-CN dataset, a comprehensive benchmark tailored for traffic-related visual-language understanding. The experimental results, as summarized in Table III, demonstrate that the instruction chain of extraction and description (illustrated in Figure 3(b)) employed by SignEye significantly outperforms general Vision-Language

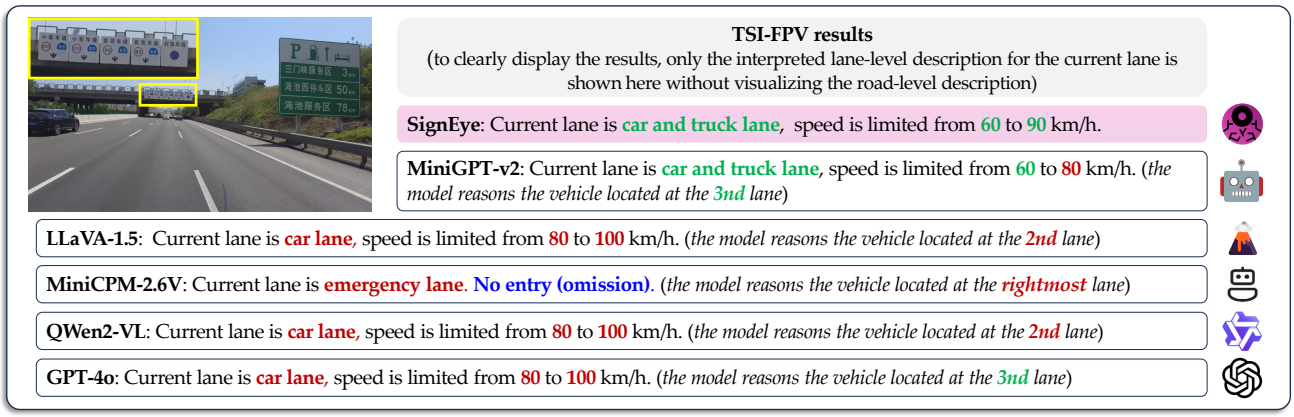


Fig. 5: Visualization of the egocentric vehicle corresponding description in the TSI-FPV task.

TABLE III: Comparisons with existing popular VLM baselines on TSI task.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MiniGPT-v2 [43]	56.1	41.6	32.9	27.5	33.9	55.7	1.97
LLaVA-1.5 [21]	73.1	64.6	58.7	53.9	45.9	71.0	4.20
MiniCPM-2.6 [33]	69.4	60.6	54.0	48.6	45.6	68.6	4.05
QWen2-VL [44]	66.2	57.8	51.2	45.4	47.1	70.1	4.24
SignEye (Ours)	**74.8** (1.7↑)	**67.1** (2.5↑)	**61.1** (2.4↑)	**56.2** (2.3↑)	**49.3** (2.2↑)	**72.7** (1.7↑)	**4.70** (0.46↑)

Models (VLMs) that attempt to describe sign units directly. This improvement can be attributed to the structured and systematic approach adopted by SignEye, which ensures a more precise interpretation of traffic sign information, thereby enhancing the overall reliability and applicability of the model in real-world TGA scenarios. The comparative analysis underscores the importance of specialized methodologies over generic approaches in handling domain-specific tasks such as traffic sign understanding.

TSI-FPV based TGA scenario application. To re-explore the role of traffic signs in ADS, we develop a TGA scenario (as visualized at the bottom of Figure 1). Considering the deep dependency of TGA on egocentric vehicle position information, we introduce the EgoRPD strategy to build a connection between vehicle and sign units for further proposing the TSI-FPV task based on previous TSI to support TGA. In Figure 5, we first visualize the model's ability to match the egocentric vehicle and the corresponding sign units. It can be observed that general VLMs struggle to accurately locate the vehicle's position and connect the egocentric view of the vehicle to the corresponding sign units, even though humans can easily interpret these spatial relationships. While GPT-4o and MiniGPT-v2 correctly identify the lane position, it still finds it challenging to associate the position with the corresponding correct lane-level sign unit. Unlike these general VLMs, which determine all lanes sequentially from left to right before identifying the vehicle's lane, SignEye determines the current lane by analyzing the geometry of lane lines and connects the egocentric view with the corresponding sign unit based on their spatial relationship (see Algorithm 1) rather than by analyzing the absolute spatial location of each instance globally, which brings significant improvements for our model in making lane-level plan (referred to Section V-C).

SignEye distinguishes itself from existing approaches by explicitly modeling the positional relationships between vehicles, lanes, roads, and traffic signs through its innovative EgoRPD strategy. This capability is particularly crucial for the TSI-FPV task and TGA applications. In contrast, current general Vision-Language Models (VLMs) and VLM-based end-to-end systems, including DriveLM (the most widely adopted open-source end-to-end system), lack the ability to analyze these specific spatial correspondence relationships (as shown in IV). This limitation explains why even when fine-tuned on our dataset, existing end-to-end systems without EgoRPD and instruction chain mechanisms perform comparably to general-purpose VLMs such as Qwen-VL and LLaVA, as demonstrated by our experimental results.

To verify the importance of TSI-FPV for TGA, we then show the performance results in Table IV, where SignEye achieves remarkable proficiency in plan accuracy, surpassing the nearest competing general VLMs without descriptions in EgoRPD 7.3% at least. Importantly, our model achieves 9.7% superiority when we drop the #1 option that does not require the position information but only sign units. These results highlight the effectiveness of the connection between the egocentric vehicle and the corresponding sign descriptions provided by TSI-FPV and the stepwise reasoning pipeline. The effectiveness of SignEye for analyzing the complex spatial relationship among the egocentric vehicle, signs, lanes, and roads is also demonstrated.

Visualization results on TGA. Based on the quantitative experimental results before, our SignEye, leveraging the ad-

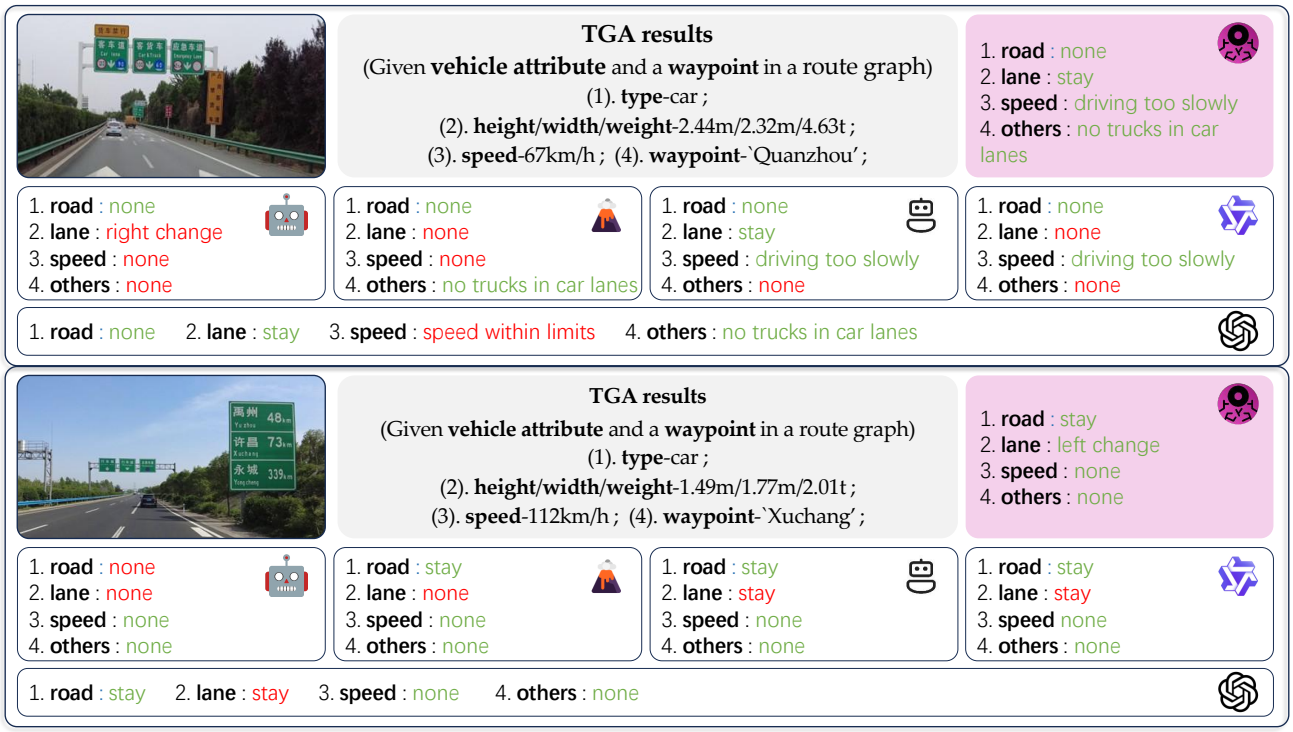


Fig. 6: Visualization of the egocentric vehicle corresponding description in the TGA application.

TABLE IV: Accuracy results on TGA scenario application. $O_{all}^{drop\#1}$ is the all overall accuracy without #1 option.

Model	Road	Lane	Speed	Other	O_{all}	$O_{all}^{drop\#1}$
MiniGPT-v2 [43]	42.6	37.4	62.2	49.5	47.9	34.0
LLaVA-1.5 [21]	65.8	47.2	67.7	72.2	63.2	54.8
MiniCPM-2.6 [33]	79.7	83.0	80.5	82.1	79.7	75.9
QWen2-VL [44]	78.8	63.8	70.5	72.6	71.5	65.2
DriveLM [51]	70.6	66.2	67.9	62.0	66.7	62.5
SignEye (Ours)	**88.8** (9.1↑)	**86.3** (3.3↑)	**85.6** (5.1↑)	**87.4** (5.3↑)	**87.0** (7.3↑)	**85.6** (9.7↑)

vantages of the EgoRPD strategy, achieves superior performance in the TGA scenario, particularly in lane and speed planning. Here, we present some representative visualizations of TGA to highlight the differences between SignEye and existing strong general VLMs in their final planning decisions.

Specifically, as shown in Figure 6, the input image illustrates a typical road scene for lane and speed planning, providing specific vehicle type and speed limit information for various lanes. The model can make accurate lane and speed plans if it locates the lane position, connects the lane to the corresponding sign descriptions, and matches the egocentric vehicle status with the appropriate lane. However, MiniGPT-v2, LLaVA-1.5, and QWen2-VL, even when fine-tuned on our TGA data, struggle to complete all these steps, leading to incorrect lane planning. Interestingly, they might identify the correct lane position but fail to connect it accurately with the corresponding sign descriptions. For instance, Qwen2-VL makes an incorrect lane plan but a correct speed plan.

Additionally, Figure 6 depicts another typical road scene for road and lane planning, focusing on specific lane attributes and destination information rather than vehicle attributes. This makes it challenging for strong general VLMs like MiniGPT-v2 and LLaVA to understand the relationship between the given vehicle attributes and lane attributes. Models such as MiniCPM-2.6, QWen2-VL, and GPT4o consistently identify the vehicle as being in the center lane, which results in incorrect lane planning due to this misjudgment of lane position.

The visualization results further demonstrate the effectiveness of our SignEye model and explain why existing strong general VLMs struggle to achieve excellent results in the TSI-FPV-based TGA scenario.

C. Ablation Studies

We conduct thorough experiments to validate the effectiveness of the designed stepwise reasoning pipeline and permutation-invariant loss.

Effectiveness of the TSI-FPV on TGA. As shown in Figure 3, the vehicle's relative position information is important for TGA in making a driving plan, especially for lane and speed plans. We visualize the plan accuracy of road, lane,

TABLE V: Ablation study on EgoRPD strategy-based TSI-FPV for TGA. #n means the nth option, and O_r, O_l, O_s and O_o are the overall accuracy of all options of road, lane, speed, and other respectively. O_{all} is the overall accuracy of all options of road, lane, speed, and other. $O_{all}^{drop\#1}$ is the all overall accuracy without #1 option. 'w/o EgoRPD' denotes TSI, and 'with EgoRPD' means TSI-FPV.

EgoRPD	Road							O_r	Lane				O_l
	#1	#2	#3	#4	#5	#6	#7	/	#1	#2	#3	#4	/
w/o	90.4	85.7	62.5	70.0	82.1	75.0	71.9	86.5	90.1	87.8	53.3	46.4	82.6
with	90.2	88.7	76.5	66.7	81.8	70.2	88.9	**88.8** (2.3↑)	89.8	88.4	83.3(**30.0**↑)	67.9(**21.5**↑)	**86.3** (3.7↑)

EgoRPD	Speed				O_s	Other				O_o	O_{all}	$O_{all}^{drop\#1}$
	#1	#2	#3	#4	/	#1	#2	#3	#4	/	/	/
w/o	87.4	67.2	78.2	65.5	81.5	88.5	81.3	80.1	87.3	85.3	84.0	80.3
with	88.3	82.5	84.4	77.2	**85.6** (4.1↑)	89.2	80.0	84.4	90.8	**87.4** (2.1↑)	**87.0** (3.0↑)	**85.6** (5.3↑)

speed, and other in Table V to show the effectiveness of the EgoRPD strategy-based TSI-FPV task on TGA.

Specifically, for all kinds of plans, SignEye achieves superior accuracy in #1 option. It is mainly because #1 corresponds to the 'none' option (referred to II), which means the model can choose the correct option by recognizing whether signs occurred in images and no need for position information. Instead, for those relative position-sensitive lane change options (e.g., lane #3 and #4), the model has to determine the lane-level sign description corresponding to the egocentric vehicle's lane position first. Then, it finds out the adjacent left and right lane-level sign descriptions. In the end, SignEye makes a lane plan according to vehicle attributes and descriptions. It can be observed that our model achieves 30% and 21% improvements on the lane change options compared with the model that is without the description in EgoRPD and has to analyze the complex spatial distribution based on vision features only. A similar conclusion can be given for making a speed plan because there are plenty of situations in which the speed description occurs on lane-level sign units. We also show the overall accuracy of all options of road, lane, speed, and other, the EgoRPD strategy can bring 3.0% and 5.3% improvements when with and without considering the 'none' option. The results demonstrate the essential of EgoRPD strategy to TGA, especially for position-sensitive plans.

Furthermore, traffic signs in different countries generally follow a relatively unified style. For example, common symbols (e.g., speed limits, no entry) adhere to the same design standards, and common panels (combinations of symbols and texts) share the same semantic logic. The biggest difference between Chinese-style traffic signs and those of other countries is the language. This challenge of recognizing different languages can be alleviated by the pre-trained large language model in our SSLM. Therefore, our model, trained on the Traffic-CN dataset, can be applied to traffic scenes in other countries (such as United States, England, and Japan).

Effectiveness of Permutation-Invariant Loss on TSI. As detailed in Table I and Section 2, the prediction description for samples containing multiple sub-sentences and destinations is designed to align with multiple order-adjusted ground truths. This approach accounts for the inherent variability in how traffic sign information can be interpreted and expressed. To address this, the proposed $\mathcal{L}_{pi}$ (prediction-optimization loss) is specifically designed to enhance model efficiency by minimizing the discrepancy between the predicted description and the most semantically aligned ground truth among the available options. This optimization strategy not only ensures a more robust training process but also leads to significant performance improvements in the Traffic Sign Interpretation (TSI) task. Notably, as the number of [***keywords***] in the description increases, the model demonstrates enhanced accuracy and reliability, as evidenced by the experimental results presented in Table VI. This improvement highlights the effectiveness of $\mathcal{L}_{pi}$ in capturing the nuanced relationships between prediction descriptions and their corresponding ground truths, particularly in complex scenarios involving multi-faceted traffic sign information. The findings underscore the importance of tailored loss functions in advancing the capabilities of models for domain-specific tasks such as TSI.

Speed Analysis. Except for the accuracy performance, we further analyze the interpretation speed and computational resources of our framework in the inference process. As shown in the Table VII, the sign detection module and the lane detection module only take about 13% and 17% of the total time consumption, which is consistent with the module's GFLOPs and Params. For the VLM-based SSLM, although it consumes more computational resources compared with sign detection and lane detection modules, the whole framework still enjoys a competitive inference speed (1.12 FPS).

VI. Discussion and Conclusion

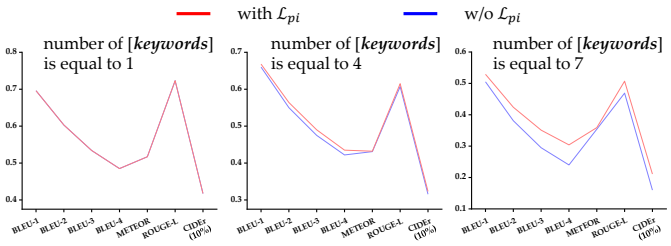
A. Discussion

TSI-FPV *vs.* Traffic Sign Understanding. Traffic signs are inherently associated with specific roads or lanes, making it crucial to accurately identify signs relevant to the egocentric vehicle based on their spatial positioning. Unlike conventional methods that overlook positional information, TSI-FPV (Traffic Sign Interpretation for First-Person View) equips Autonomous Driving Systems (ADS) with the capability to prioritize and apply only those signs pertinent to the egocentric driving path. By filtering out irrelevant signs in input road scene images, TSI-FPV ensures that the driving plans are guided solely by appropriate traffic instructions, thereby minimizing potential interference and enhancing the overall safety and efficiency of autonomous navigation.

Structured Description *vs.* General Description. Unlike general description tasks (e.g., image captioning, VQA), SignEye generates structured descriptions (see Table I) without

TABLE VI: Performance gain brought by $\mathcal{L}_{pi}$ to the TSI task. $\mathrm{num_k}$ denotes the number of [***keywords***] in a ground truth description. 'B-1', 'B-2', 'B-3', and 'B-4' are 'BLEU-1', 'BLEU-2', 'BLEU-3', and 'BLEU-4'. 'R-L' denotes 'ROUGE-L' metric.

$\mathrm{num_k}$	B-1	B-2	B-3	B-4	METEOR	R-L	CIDEr
1	**0.0**↑	**0.0**↑	**0.0**↑	**0.0**↑	**0.0**↑	**0.0**↑	**0.0**↑
4	**0.8**↑	**1.4**↑	**1.5**↑	**1.3**↑	**0.1**↑	**0.8**↑	**0.08**↑
7	**2.4**↑	**4.3**↑	**5.6**↑	**6.4**↑	**0.5**↑	**3.8**↑	**0.51**↑



redundancy or omissions, which offers two advantages: (1) it enables VLMs to convey key information directly to ADS or drivers more effectively, and (2) it allows our data engine to perform regularized analysis of key information in sign units for generating large-scale TGA data.

TGA *vs.* Electric Map Navigation. Both of them are responsible for direction navigation and traffic regulation assessment. However, ***they are not substitutive but complementary***. TGA enjoys the superiority of dynamic and offline relative to electric map navigation, which makes them complementary: (1) TGA can provide newly added or changed sign information for ADS to handle the situations of road construction or rule's temporary adjustment; (2) the TSI-FPV of TGA can maintenance of the sign information for the electric map as an automatic tool; (3) TGA complements the electric map under weak signal or offline.

B. Limitation Analysis

The above experiments verify the superior performance of the proposed pipeline in the tasks of TSI-FPV and TGA. Although our method can handle challenging visual conditions (such as low light, rain, fog, blur, and shadows) due to its strong image recognition capability brought by the well-pretrained VLMs, it fails to detect multiple signs independently when facing severe occlusion problems, where some traffic signs are completely missing visual information. This means our method struggles to perceive occluded traffic signs during the detection process and cannot provide useful visual information to the subsequent recognition module (SSLM). Therefore, finding an effective solution to these problems will be our future work.

C. Conclusion

In this paper, we introduce TSI-FPV to interpret sign units as structured descriptions in EgoRPD. Building on it, we develop TGA to re-explore the role of traffic signs in assisting ADS. Additionally, the Traffic-CN dataset, created through our data engine, supports research and evaluation of TSI-FPV and TGA, encouraging broader participation in sign-related ADS research. Experiments show that TSI-FPV and TGA are feasible, and the SignEye achieves superior performance over general VLMs. In future work, we will focus on integrating TGA with mainstream ADS technologies (e.g., occupancy networks) to enhance sign-related support for driving plan-decisions. Notably, traffic signs in different countries generally follow a relatively unified style. For example, common symbols (e.g., speed limits, no entry) adhere to the same design standards, and common panels (combinations of symbols and texts) share the same semantic logic. The biggest difference between Chinese-style traffic signs and those of other countries is the language. This challenge of recognizing different languages can be alleviated by the pre-trained large language model in our SSLM. Therefore, our model, trained on the Traffic-CN dataset, can be applied to traffic scenes in other countries (such as United States, England, and Japan).

TABLE VII: Time consumption analysis of the proposed SignEye framework. "Size" means that the longer side size of each testing image. "SD" and "LD" are the sign and lane detection modules in our framework.

Size	GLOPs			Params (M)		
	SD	LD	SSLM	SD	LD	SSLM
1056	67.02	49.74	2453.97	12.11	219.03	7000
Size	Time comsuption (s)			FPS		
	SD	LD	SSLM			
1056	0.12	0.15	0.62	1.12		

REFERENCES

[1] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2110–2118.

[2] A. De La Escalera, L. E. Moreno, M. A. Salichs, and J. M. Armingol, "Road traffic sign detection and classification," *IEEE transactions on industrial electronics*, vol. 44, no. 6, pp. 848–859, 1997.

[3] Y. Guo, W. Feng, F. Yin, T. Xue, S. Mei, and C.-L. Liu, "Learning to understand traffic signs," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2076–2084.

[4] Y. Guo, F. Yin, X.-h. Li, X. Yan, T. Xue, S. Mei, and C.-L. Liu, "Visual traffic knowledge graph generation from scene images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 604–21 613.

[5] C. Yang, K. Zhuang, M. Chen, H. Ma, X. Han, T. Han, C. Guo, H. Han, B. Zhao, and Q. Wang, "Traffic sign interpretation via natural language description," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[9] D. Tabernik and D. Skočaj, "Deep learning for large-scale traffic-sign detection and recognition," *IEEE transactions on intelligent transportation systems*, vol. 21, no. 4, pp. 1427–1440, 2019.

[10] Y. Yang, H. Luo, H. Xu, and F. Wu, "Towards real-time traffic sign detection and classification," *IEEE Transactions on Intelligent transportation systems*, vol. 17, no. 7, pp. 2022–2031, 2015.

[11] C. Yang, M. Chen, Z. Xiong, Y. Yuan, and Q. Wang, "Cm-net: Concentric mask based arbitrary-shaped text detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2864–2877, 2022.

[12] M. Ye, J. Zhang, S. Zhao, J. Liu, T. Liu, B. Du, and D. Tao, "Deepsolo: Let transformer decoder with explicit points solo for text spotting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 348–19 357.

[13] C. Yang, X. Han, T. Han, H. Han, B. Zhao, and Q. Wang, "Edge approximation text detector," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[14] C. Yang, M. Chen, Y. Yuan, and Q. Wang, "Text growing on leaf," *IEEE Transactions on Multimedia*, vol. 25, pp. 9029–9043, 2023.

[15] Y. Guo, W. Feng, F. Yin, and C.-L. Liu, "Signparser: An end-to-end framework for traffic sign understanding," *International Journal of Computer Vision*, vol. 132, no. 3, pp. 805–821, 2024.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[17] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[18] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[19] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[20] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[21] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

[22] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[23] R. Bavishi, E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşırlar, "Introducing our multimodal models," 2023. [Online]. Available: https://www.adept.ai/blog/fuyu-8b

[24] B. Li, P. Zhang, J. Yang, Y. Zhang, F. Pu, and Z. Liu, "Otterhd: A high-resolution multi-modality model," *arXiv preprint arXiv:2311.04219*, 2023.

[25] Z. Li, B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, and X. Bai, "Monkey: Image resolution and text label are important things for large multi-modal models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 763–26 773.

[26] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "Geochat: Grounded large vision-language model for remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 831–27 840.

[27] Q. Guo, S. De Mello, H. Yin, W. Byeon, K. C. Cheung, Y. Yu, P. Luo, and S. Liu, "Regiongpt: Towards region understanding vision language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 796–13 806.

[28] W. Wang, M. Shi, Q. Li, W. Wang, Z. Huang, L. Xing, Z. Chen, H. Li, X. Zhu, Z. Cao *et al.*, "The all-seeing project: Towards panoptic visual recognition and understanding of the open world," *arXiv preprint arXiv:2308.01907*, 2023.

[29] W. Wang, Y. Ren, H. Luo, T. Li, C. Yan, Z. Chen, W. Wang, Q. Li, L. Lu, X. Zhu *et al.*, "The all-seeing project v2: Towards general relation comprehension of the open world," *arXiv preprint arXiv:2402.19474*, 2024.

[30] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang *et al.*, "Yi: Open foundation models by 01. ai," *arXiv preprint arXiv:2403.04652*, 2024.

[31] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, Y. Sun *et al.*, "Deepseek-vl: towards real-world vision-language understanding," *arXiv preprint arXiv:2403.05525*, 2024.

[32] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites," *arXiv preprint arXiv:2404.16821*, 2024.

[33] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

[34] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[35] Q. Ma, X. Tan, Y. Qu, L. Ma, Z. Zhang, and Y. Xie, "Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 936–19 945.

[36] Y. Huang, W. Zheng, B. Zhang, J. Zhou, and J. Lu, "Selfocc: Self-supervised vision-based 3d occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 946–19 956.

[37] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[38] Z. Qin, H. Wang, and X. Li, "Ultra fast structure-aware deep lane detection," in *The European Conference on Computer Vision (ECCV)*, 2020.

[39] Tusimple. https://github.com/TuSimple/tusimple-benchmark. [Online]. Available: https://github.com/TuSimple/tusimple-benchmark

[40] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.

[41] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, and Z. Fan, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[42] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[43] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

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

[45] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2315–2324.

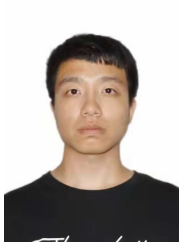
[48] C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding *et al.*, "Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1571–1576.

[49] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[50] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 325–341.

[51] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," in *European Conference on Computer Vision*. Springer, 2024, pp. 256–274.

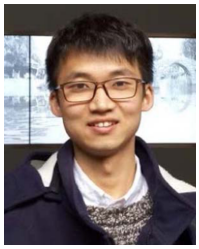
Chuang Yang received the Ph.D. degree in the School of Computer Science and School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China, in 2024. He is currently a Postdoc Fellow with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong. His research interests include sign-driven VLN and AIGC for remote sensing.



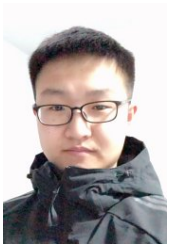
Xu Han received the B.E. degree in information and computing sciences from Northeast Agricultural University, Harbin, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Computer Science and School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and text detection.

Tao Han received a B.E. degree in transportation equipment and control engineering and an M.S. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2019 and 2022. He is currently pursuing a Ph.D. degree in computer science and engineering at the Hong Kong University of Science and Technology. His research interests include computer vision, ai4science, and AIGC.

Yuejiao Su received the B.Sc. and M.Sc. degrees in computer science and engineering from the Northwestern Polytechnical University, China in 2020 and 2023 respectively. She is currently a Ph.D. degree candidate with The Hong Kong Polytechnic University, China. Her research interests include egocentric analysis, image segmentation, and embodied AI.



Junyu Gao (Member, IEEE) received the B.E. degree and the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an, China, in 2015 and 2021, respectively. He is currently an Associate Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include computer vision and pattern recognition.



Hongyuan Zhang received the B.E. degree in software engineering from Xidian University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an.

Yi Wang (Member, IEEE) received the B.Eng. degree in electronic information engineering and the M.Eng. degree in information and signal processing from the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China, in 2013 and 2016, respectively, and the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2021. He is currently a Research Assistant Professor with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong. His research interests include image/video processing, computer vision, intelligent transport systems, and digital forensics.

Lap-Pui Chau (M'15-SM'15) (Fellow, IEEE) received the Ph.D. degree from The Hong Kong Polytechnic University in 1997. He was with the School of Electrical and Electronic Engineering, Nanyang Technological University, from 1997 to 2022. He is currently a Professor with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. His current research interests include image and video analytics and autonomous driving. He was the Chair of the Technical Committee on Circuits and Systems for Communications of the IEEE Circuits and Systems Society from 2010 to 2012. He was the general chair and the program chair of some international conferences. Besides, he served as an associate editor for several IEEE journals and a Distinguished Lecturer for IEEE BTS.