SafeRoute: Enhancing Traffic Scene Understanding via a Unified Deep Learning and Multimodal LLM

Ankit Kumar Shaw[†1] Chandan Kumar Sah[†2] Xiaoli Lian[*2] Arsalan Shahid Baig[2] Tuopu Wen[1]
Kun Jiang[*1] Mengmeng Yang[1] Diange Yang[1] Li Zhang[2]
[1]Tsinghua University [2]Beihang University
[†]Equal Contribution [*]Corresponding author
shawak10@mails.tsinghua.edu.cn, {jiangkun, yangmm_qh}@tsinghua.edu.cn,
ydg@mail.tsinghua.edu.cn, wtp18@tsinghua.org.cn,
{sahchandan98, lianxiaoli, arsalan, lily }@buaa.edu.cn

Abstract

Autonomous vehicles (AVs) require highly reliable traffic sign recognition and robust lane detection to navigate safely in complex and dynamic environments. This paper presents SafeRoute, a unified perception framework that integrates deep learning with instruction-tuned Multimodal Large Language Model (MLLM) for comprehensive road scene understanding. For traffic sign recognition, we benchmark three state-of-the-art architectures, ResNet-50, YOLOv8, and RT-DETR, achieving accuracies of 99.8%, 98.0%, and 96.6% respectively. To address the limitations of traditional vision-only methods in lane detection under adverse conditions (e.g. occlusion, poor lighting, road wear), we introduced a MLLM-based pipeline, fine-tuned via instruction learning without requiring large-scale pretraining. Our approach introduces a novel Multimodal Adapter that fuses CNN-derived spatial features with EVA-CLIP embeddings, enabling fine-grained visual grounding and robustness to occlusion. By integrating these visual tokens into a LLaMA-2 decoder, our system performs semantic-level reasoning and interpretable scene understanding, moving beyond segmentation to structured, language-based lane perception. Quantitatively, SafeRoute achieves a Frame Overall Accuracy (FRM) of 53.87%, Question Overall Accuracy (QNS) of 82.83%, and lane detection accuracies of 99.6% in clear conditions and 93.0% at night. It also demonstrates robust reasoning in adverse conditions, with 88.4% accuracy in rain and 95.6% under lane degradation. Overall, SafeRoute introduces a new paradigm in AV perception by offering a unified, multimodal approach, significantly improving both the robustness and explainability of lane detection in safety-critical scenarios.

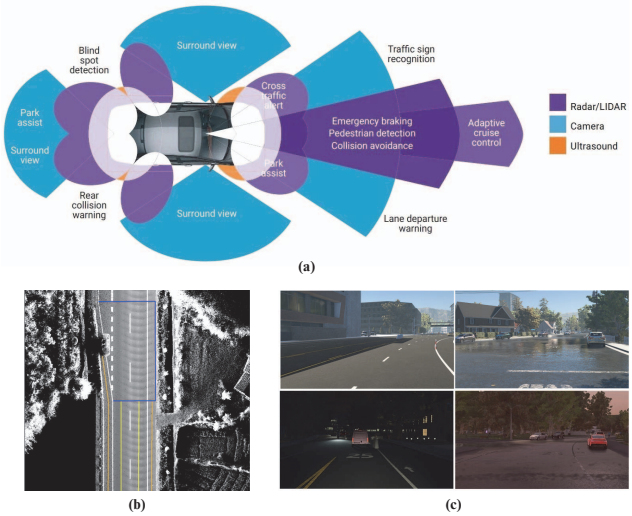

Figure 1. Multimodal LLM-Based Road Elements Understanding. (a) An AV uses multiple sensors (camera, LiDAR, radar, ultrasonic) for perception and navigation. (b) Annotated MapLM dataset with colored poly-lines: Yellow = motorway lanes, Orange = bicycle lanes, Blue = intersections/lane-change zones. (c) Apollo synthetic data representing varied times of day, weather, and road conditions.

1. Introduction

Autonomous vehicles (AVs) rely on precise perception of traffic signs and lane markings to ensure safe and effective navigation in complex, real-world environments [1–3]. With human error contributing to over 90% of traffic incidents [4], perception errors in AV systems can have significant safety implications. However, dynamic road conditions—ranging from occlusions and poor lighting to weather-induced visibility loss—present formidable chal-

lenges to current vision-based systems [5].

Convolutional Neural Networks (CNNs) have achieved strong performance in standard visual conditions. Models such as YOLO [6] and hybrid architectures like Local Vision Transformer (Local-VT) [7] demonstrate state-of-the-art accuracy in traffic sign recognition, while recent variants of YOLOv8 reach up to 97% mAP under varied lighting. Despite these advances, CNN-based systems remain vulnerable in ambiguous or degraded visual scenarios. When lane markings are occluded (e.g., by vehicles), faded, covered by rain or snow, or poorly illuminated, CNNs frequently output low-confidence, missing, or misleading predictions—including false positives that degrade downstream tasks like map updates and trajectory planning.

CNNs, though powerful perceptual models, fundamentally lack the capacity for semantic reasoning, symbolic abstraction, and contextual inference. For instance, they cannot answer questions like *"How many lanes are visible, and why not more?"*, nor can they explain visual uncertainty (e.g., *"The lane is likely hidden due to darkness or occlusion"*). This limitation is critical when AV systems must navigate temporary or ambiguous conditions, such as road construction, inclement weather, or obstructed intersections.

To overcome these limitations, we integrate Multimodal Large Language Models (MLLMs) into the lane detection pipeline. Unlike CNNs, MLLMs jointly interpret visual inputs, semantic context, and environmental cues to perform high-level reasoning. They can infer the continuity of occluded or degraded lane lines based on scene structure, road layout, and vehicle behavior. Moreover, MLLMs produce interpretable, language-based outputs (e.g., *"Lane markings are likely degraded"* or *"This appears to be a T-intersection"*), enhancing transparency and enabling better decision support for both autonomous systems and human operators. These capabilities are especially valuable for refining map updates and ensuring robustness in temporally dynamic road scenarios.

Extending beyond existing vision-language paradigms (e.g., BLIP-2, Flamingo), we present SafeRoute, a lightweight, instruction-tuned MLLM pipeline specifically tailored for structured lane understanding—a domain where prior models have traditionally underperformed. Our key contributions include:

- **Systematically comparison of traffic sign recognition models:** We evaluate ResNet-50, YOLOv8, and RT-DETR across diverse environmental scenarios, quantifying trade-offs between classification accuracy, real-time inference, and robustness to adverse conditions. Our evaluation provides practical guidance for model deployment in AV systems.
- **MLLM-Based Structured Lane Detection via Multimodal Fusion and Contextual Reasoning:** We propose a lightweight, instruction-tuned MLLM framework for structured lane understanding, which integrates a novel multimodal adapter that fuses CNN-derived spatial features with EVA-CLIP embeddings. This fusion enables fine-grained visual grounding and enhances robustness to occlusion, weather degradation, and poor lighting. The system incorporates semantic prompts and annotated training signals, allowing it to contextually reason about road topology, intersections, and merging zones—without requiring large-scale pretraining. Together, these components enable the model to perform lane detection as a semantic, structured perception task rather than a purely pixel-wise operation.
- **Enable interpretable reasoning for safety-critical perception:** By leveraging MLLMs' generative and symbolic reasoning capabilities, our framework produces human-understandable high-level explanations that enhance the interpretability of traffic scene understanding and trustworthiness of AV perception pipelines, especially in ambiguous or degraded visual scenarios.

SafeRoute demonstrates that combining deep learning with multimodal language models provides a scalable, interpretable, and robust solution for real-world autonomous navigation, particularly in the face of uncertain or degraded visual conditions.

2. Related Work

2.1. Traffic Sign Recognition with Deep Learning

Early traffic sign recognition relied on manual feature extraction and traditional classifiers [8]. Deep learning now dominates the field, with CNNs like ResNet-50 achieving over 99% accuracy using residual learning [9]. One-stage detectors such as YOLOF-F improve small sign detection, achieving 77.2% AP on CTSD at 32 FPS [10]. Faster R-CNN [11] offers strong accuracy but is computationally heavy. One-stage detectors like SSD [12] and YOLO series balance speed and accuracy, with recent variants like YOLOv5 optimized for complex weather [13] and ETSR-YOLO enhancing tiny sign detection [14]. Transformer-based models such as DETR [15] eliminate post-processing, while Deformable DETR and DINO improve convergence and small-object performance [10]. Recent work shows transformers can surpass YOLO in real-time accuracy with optimization. Attention-based models like Swin-Transformer [16] and multi-scale attention modules further enhance sign recognition. These advances provide a robust foundation for our traffic sign recognition approach [17].

2.2. Lane Line Detection Techniques

Lane detection has progressed from edge detection and Hough transforms to deep learning-based segmentation

Table 1. Deep Learning Models for Traffic Sign Recognition[22–24].

Model	Approach	Adaptations	Strengths	Weaknesses
ResNet-50	CNN	pre-trained on ImageNet; Residual blocks	99%+ accuracy; Robust	Needs proposals; ROI extraction
YOLOv8	One-stage	Custom anchors; Multiscale; Augmented data	Fast; Very low false alarms	Lower recall; Imbalanced precision
RT-DETR	Transformer	Attention; Multiscale; IoU filtering	High accuracy; Very low NMS	Slow training; Complex tuning

[18]. CNN-based methods reliably segment lanes [19], often integrating post-processing like polynomial curve fitting for smoother lane boundaries [20]. Multi-task learning incorporating temporal information further improves detection stability [5]. Adverse conditions remain a challenge, with lane markings becoming obscured in rain or poor lighting. Deep networks now infer lanes even when markings are degraded by leveraging contextual road cues [21]. State-of-the-art models like LaneNet achieve 96.4% accuracy on TuSimple [3], while SCNN achieves 71.6% F1-score on CULane [20], setting benchmarks for comparison.

2.3. VLMs for Autonomous Driving

Visual-Language Models (VLMs) play a crucial role in driving scene understanding and decision-making [25]. Recent advancements focus on VLM-driven autonomous driving policies, including DiLu [26], DriveGPT4 [27], GPT-Driver [28], HiLM-D [29], DriveMLM [30], and DriveVLM [31], while Talk2BEV and LiDAR-LLM [32] integrate LLMs, VLMs, BEV, and LiDAR for improved perception. LLMs also enhance passenger-vehicle interaction [33] and enable language-guided closed-loop autonomous driving using multi-modal sensor data, as seen in LimSim++ [34] and LMDrive. RAG-Driver improves zero-shot generalization with in-context learning, while Wayve's LINGO-1 [35] introduces an open-loop driving commentator.

3. Methodology

Our framework consists of two main components: a traffic sign recognition module powered by deep learning detectors/classifiers, and an adaptive lane line detection and reasoning module. We describe the models, datasets, and the multimodal large language model integration below.

3.1. Traffic Sign Recognition Module

To develop and evaluate traffic sign recognition models, we compiled data from German Traffic Sign Recognition Benchmark (GTSRB), TT100K, and synthetic sources generated via augmentation (e.g., rotation, contrast, noise) to simulate varied driving conditions. Preprocessing steps included grayscale normalization, histogram equalization, and edge enhancement [38]. Finally, we used polygon-based ground truth annotations for lane segmentation and bounding box labeling for traffic sign detection to ensure robust model training.

ResNet-50 Classifier: We employed a pretrained ResNet-50 CNN [36], originally trained on ImageNet, for 43-class traffic sign classification using GTSRB as the primary dataset, supplemented with additional images under varied conditions to enhance robustness. Training em-

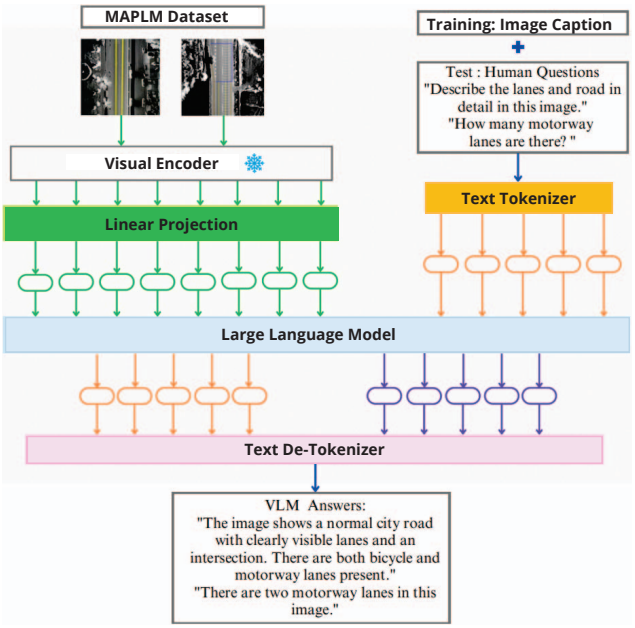

Figure 2. Framework for Multimodal LLM-based Road Elements Understanding. The system processes multimodal inputs, including annotated 3D LiDAR BEV point clouds and textual annotations. The EVA visual encoder extracts features from the point clouds, which are then integrated with text annotations and task-specific prompts. These combined inputs are processed by the LLAMA-2 LLM to generate coherent and contextually relevant responses.

Table 2. Comparison of ResNet_50 with previous works. The table presents accuracy, precision, recall, F1-score, and training time across different ResNet architectures.

Model Source	ResNet Architectures	Accuracy	Precision	Recall	F1-score	Training Time (s)
[36]	ResNet 50	99.67%	0.9968	0.9967	0.9967	1109.65
	ResNet 50V2	99.78%	0.9978	0.9978	0.9978	1016.82
	ReNet 101	99.51%	0.9944	0.995	0.9947	1893.66
	ReNet 101V2	99.74%	0.9975	0.9974	0.9974	1893.8
	ResNet 152	99.49%	0.9948	0.9949	0.9949	2632.62
	ResNet 152V2	99.76%	0.9976	0.9976	0.9976	2617.94
[37]	ResNet 50	91.43%	-	-	-	-
[36]	ResNet 50	96.41%	0.9416	0.9594	0.9464	-
	ResNet 50 + SVM	97.55%	0.9751	0.9676	0.9713	-
Ours	ResNet_50	99.80%	1.00	1.00	1.00	5412.34

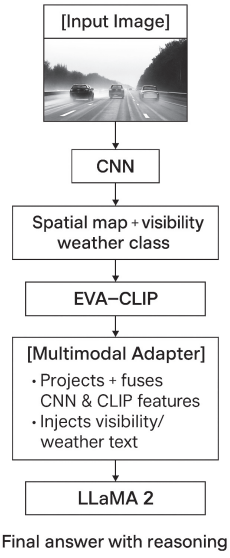

Figure 3. Multimodal-Adapter based Perception Pipeline for Visibility-Aware Lane Reasoning. The input image is processed by a ResNet-50 CNN to extract spatial features and infer visibility-aware weather conditions. EVA-CLIP provides high-level visual embeddings, which are fused with CNN-derived features via a multimodal adapter. This adapter projects and aligns the modalities, while injecting contextual cues (e.g., visibility class, rain severity) as auxiliary text. The fused representation is passed to an instruction-tuned LLaMA-2 model, enabling reasoning-driven outputs that explain lane visibility under adverse conditions.

ployed Adam optimization [39], cross-entropy loss, and early stopping. ResNet-50 achieved strong baseline accuracy and robustness under varied conditions without task-specific fine-tuning.

YOLOv8 Detector: For full-frame traffic sign detection, we utilized YOLOv8, a one-stage object detector trained end-to-end to output bounding boxes and class labels. Custom anchor box priors were used to align with common traffic sign aspect ratios, and mosaic data augmentation was applied to improve generalization across different scales and occlusions. The training dataset integrated real-world driving images (Dashcam datasets, TT100K) with synthetic scenes (Figure 1 (c)) to cover diverse environmental conditions. YOLOv8 was optimized using a combination of localization loss (complete IoU loss [40]) and classification loss. Its multi-scale feature maps enable real-time detection of small signs, as demonstrated in Figure 6, where it successfully detects traffic signs in challenging road conditions.

RT-DETR Transformer Detector: We adapted the Real-Time Detection Transformer (RT-DETR) model for traffic sign detection. Unlike conventional object detectors, RT-DETR employs a transformer encoder-decoder architecture, eliminating the need for region proposals and non-maximum suppression. Table 3 compares its performance with other DETR-based architectures, showcasing its competitive mAP50 score. To enhance detection of small signs, an attention-based feature fusion module was incorporated, inspired by [41]. The model produces a fixed set of predictions, matched with ground truth boxes using the Hungarian algorithm during training [42], optimizing a joint loss function (L1 and GIoU for box regression, focal loss for classification). Although RT-DETR requires significantly more training epochs than YOLOv8, it excels in detecting signs in cluttered scenes due to its global self-attention mechanism. However, its computational cost remains high, running at 10 FPS in our setup, reflecting its emphasis on contextual reasoning over real-time efficiency.

3.2. Our MLLM Datasets and Annotations

We utilize the MAPLM [43] dataset for road element understanding, comprising 10,775 training and 1,500 test Point Cloud Bird's-Eye View (BEV) images collected from diverse environments, including highways and urban roads. This dataset also provides detailed geometric lane information and attributes, enabling a comprehensive analysis of road structures. To enhance model performance, we introduce a novel preprocessing strategy: annotating BEV point cloud images with color-coded poly-lines, as illustrated in Figure 1 with the help of the given geometrical coordinates for each road elements i.e. motorway lanes are marked in yellow, bicycle lanes in orange, and intersections/lane-change zones in blue, aiding in lane detection and reasoning. As seen, each color represents a distinct road feature, offering explicit visual cues to improve lane feature recognition and cross-section interpretation. For lane detection under adverse conditions, we employ the Apollo Synthetic Dataset, which consists of 10,000 training and 2,200 test RGB images. Since task-specific labels are unavailable, we manually annotate this dataset in a zero-shot setting to ensure reliable ground truth generation.

General Caption Template for Lane Elements Understanding (MAPLM Dataset): The scene contains a *[scene type]* with *[data quality]* data quality. It includes *[total number of lanes]* lanes, specifically a *[type of lane]* lane extending from *[start coordinate]* to *[end coordinate]*, and a *[type of lane]* lane spanning from *[start coordinate]* to *[end coordinate]*. Additionally, a *[type of cross-section]* is present at the intersection, defined by vertices *[vertex 1]*, *[vertex 2]*, *[vertex 3]*, and *[vertex 4]*.

General Caption Template for Adaptive Lane Detection (Apollo Synthetic Dataset): 1. Lane lines are fully visible. 2. Lane lines are *[partially visible / invisible]* due to *[specific reason]*.

This structured annotation methodology enhances the dataset's utility for evaluating lane detection and road scene understanding across both standard and challenging envi-

ronmental conditions.

We also employ targeted prompts such as "*Describe the lanes and road elements in detail*", "*How many lanes are there?*", "*Are the lanes visible? If not, why?*", and "*Are there intersections or cross-sections?*" (see Supplementary). These task-specific inputs improve interpretability, refine feature extraction, and reduce irrelevant token processing—enabling the model to focus on semantically relevant aspects of the scene, particularly under challenging conditions.

3.3. Our MLLM Framework Architecture

As illustrated in Figure 2 and 3, we design a lightweight yet effective multimodal architecture inspired by MiniGPT-v2 [44]. The framework leverages an EVA encoder for visual feature extraction—chosen for its strong generalization and high-resolution capacity (448×448)—alongside a LLaMA-2-7B decoder as the language model backbone.

3.3.1. Input Processing

Multimodal Observations for Scene Understanding. Given observations $O = \{X_{pc}, X_{ann}\}$, consisting of BEV (Bird's Eye View) point cloud representations and textual annotations, along with a natural language query X_q, the model predicts an output answer $\hat{Y}$ using a multimodal function F_θ:

$$\hat{Y} = F_\theta(O, X_q) \tag{1}$$

Here, X_{pc} denotes the BEV image from the point cloud, X_{ann} includes contextual descriptions or captions of road elements, and X_q represents the task-specific query. $\hat{Y}$ is the predicted autoregressive answer.

3.3.2. Encoding

Point Cloud BEV Tokenization & Feature Extraction. BEV images are tokenized using a frozen EVA encoder ϕ_{bev} to extract patch-based visual embeddings:

$$Z_{bev} = W_{bev} \cdot k_{bev}(\phi_{bev}(\text{BEV}(X_{pc}))), \quad Z_{bev} \in \mathbb{R}^{d \times k_{bev}/4} \tag{2}$$

where k_{bev} is the EVA backbone, and $W_{bev} \in \mathbb{R}^{d \times 4d_{bev}}$ projects the extracted BEV features into the language token space. These tokens serve as the vision-language bridge for downstream LLM processing.

3.3.3. Linear Projection into LLM Layer

To reduce computational load and input sequence length, we concatenate every four adjacent BEV tokens and linearly project them into the LLaMA-2 language embedding space:

$$Z'_{bev} = W'_{bev} \cdot \text{Concat}_4(Z_{bev}), \quad Z'_{bev} \in \mathbb{R}^{d \times k_{bev}/4} \tag{3}$$

Here, W'_{bev} acts as a learnable projection layer. We apply Low-Rank Adaptation (LoRA) with rank $r = 64$ to fine-tune key subcomponents (e.g., query/value matrices) of the LLM efficiently.

3.3.4. Multimodal Question Answering Output

The final answer $\hat{Y}$ is generated autoregressively by the LLM, conditioned on both vision and text inputs:

$$P(\hat{Y}|Z_{bev}, X_{ann}, X_q) = \prod_{i=1}^{L} P(y_i|Z_{bev}, X_{ann}, X_q, < i, \hat{Y}_{<i}; \theta) \tag{4}$$

This formulation models conditional token probabilities over generated answers, allowing the LLM to reason over visual semantics, language cues, and road context holistically.

3.3.5. Multimodal Fusion Strategy

To enrich the visual representation and improve robustness in complex scenarios (e.g., occlusion, rain), we propose a novel multimodal fusion mechanism illustrated in figure 3. Let:

$Z_v^{\text{clip}} \in \mathbb{R}^{T_v \times d}$ (Visual tokens from EVA-CLIP)

$F_{\text{cnn}} \in \mathbb{R}^{H'W' \times d'}$ (Flattened CNN feature map)

We project CNN-derived features to match the EVA token dimension using a learnable MLP:

$$F_{\text{cnn}}^{\text{proj}} = \text{MLP}(F_{\text{cnn}})$$

We then concatenate the two modalities and apply a fusion operator ψ_{fuse}:

$$Z_{\text{fused}} = \psi_{\text{fuse}}\left(\left[Z_v^{\text{clip}}; F_{\text{cnn}}^{\text{proj}}\right]\right)$$

The fused representation is aligned to the LLaMA-2 input embedding space using another MLP:

$$Z_{\text{LLM-input}} = \text{MLP}_{\text{align}}(Z_{\text{fused}}) \in \mathbb{R}^{T \times d_{\text{llm}}}$$

This vision-language structured reasoning pipeline redefines lane detection from a purely pixel-wise classification task to one of semantic, multimodal interpretation—introducing a new paradigm for robust and explainable AV perception.

3.4. Task Description

We propose a two-stage Multimodal LLM-based framework for adaptive lane detection and reasoning in adverse driving conditions.

3.4.1. Stage 1: Instruction Tuning for Road Element Understanding

In the first stage, we instruction-tune a Multimodal LLM to recognize structured road elements—such as lane types, intersections, and merging areas—that are typically ignored by standard VLMs like MiniGPT-v2. Using the MAPLM dataset (polyline-annotated BEV images with textual descriptions and structured prompts), the model is trained to describe and reason about road layouts. Performance is evaluated using the MAPLM-QA benchmark, verifying the model's ability to interpret complex traffic scenes.

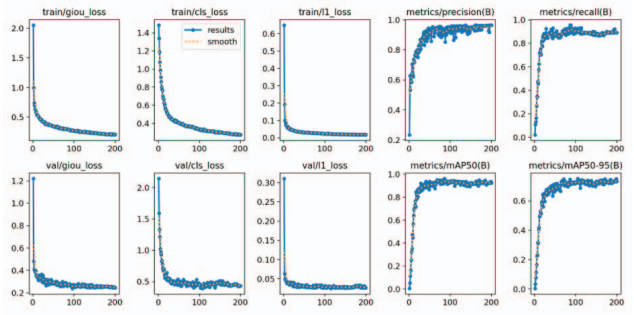

Figure 4. RT-DETR Training Performance: Loss and Validation curves over epochs. The transformer-based model takes longer to converge than YOLO, but eventually reaches a stable high accuracy.

Table 3. RT-DETR comparison. The table compares the performance of various DETR-based architectures for traffic sign detection, specifically focusing on the mean Average Precision at 50% intersection over union (mAP50).

Model Source	DETR Architectures	mAP50
[15]	DINO-DETR	77%
[45]	Deformable DETR	96.52%
[46]	DETR	95.10%
Ours	RT-DETR	96.60%

3.4.2. Stage 2: Adaptive Lane Detection under Adverse Conditions

Building on the instruction-tuned model, the second stage targets robustness under real-world challenges such as rain, night-time, and lane degradation across urban, rural, and highway scenarios. Fine-tuning is performed on the Apollo Synthetic dataset, comprising RGB images with varied weather and lighting. The enhanced model accurately detects lanes and explains visibility failures (e.g., occlusion, wear), supporting safer navigation in difficult conditions.

3.4.3. Model Training and Optimization

We adopt EVA as the frozen visual encoder and LLaMA-2-7B as the LLM decoder. For low-level spatial reasoning, we use a ResNet-50 CNN backbone to extract visibility-aware spatial maps and weather condition classes. A linear projection layer aligns CNN and EVA-CLIP features, which are then fused via a multimodal adapter. LoRA (rank $r = 64$) is applied to the LLM's query and value matrices for efficient fine-tuning. Training is conducted on 8 NVIDIA RTX 3090 GPUs, with dataset splits of 10,775/1,500 (MAPLM) and 10,000/2,200 (Apollo Synthetic). Input images are resized to 448×448 pixels. Optimization employs AdamW with a cosine learning rate scheduler (max LR $1e^{-5}$, warmup ratio 0.05), and Cross-Entropy Loss is used for both lane detection and reasoning tasks. The model is trained for 20 epochs (Stage 1) and 10 epochs (Stage 2), yielding a robust and interpretable lane detection system for real-world autonomous driving.

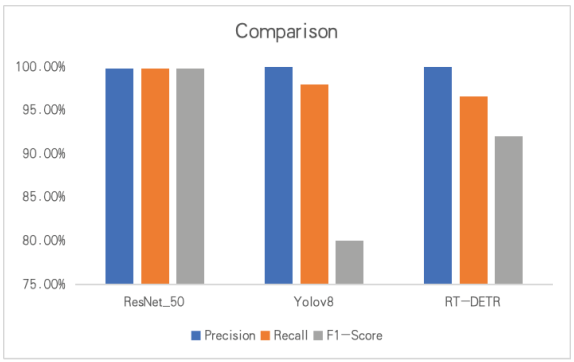

Figure 5. Performance Comparison of ResNet-50, YOLOv8, and RT-DETR in Terms of Precision, Recall, and F1-Score.

Figure 6. YOLOv8-Based Traffic Sign Detection results on real-world scenes. The model detects multiple signs (bounding boxes) in diverse conditions, demonstrating high recall and precision.

Table 4. YOLO v8 comparison. The table compares YOLO-based architectures for traffic sign recognition, focusing on precision, recall, F1-scores, and training times across different studies.

Model Source	YOLO Architectures	Precision	Recall	F1-Score	Training Time (s)
[47]	Yolo v5	98.40%	89.30%	-	-
[48]	GRFS Yolo v8	91.30%	93.00%	89%	-
Ours	Yolo v8	100.00%	98.00%	80.00%	2700

4. Results and Discussion

We evaluate our system on multiple datasets and scenarios to assess both traffic sign recognition and lane detection performance. All experiments were carried out on a machine using Nvidia RTX 3090 GPUs. We compare against baseline methods from the literature where possible and analyze

Table 5. Traffic sign recognition results for different models. ResNet-50's Precision/Recall are effectively classification accuracy on the sign crops. YOLOv8 and RT-DETR results are for full-frame detection (Prec/Rec here treat each correctly detected sign as true positive, etc.). Training times were measured on the same hardware.

Model	Precision	Recall	F1-Score	mAP@0.5	Training Time (s)
ResNet-50 (classifier)	99.80%	99.79%	99.79%	- (classification)	5,400
YOLOv8 (detector)	100.0%	98.0%	80.0%	97.5%	2,700
RT-DETR (detector)	100.0%	96.6%	90.0%	96.8%	28,200

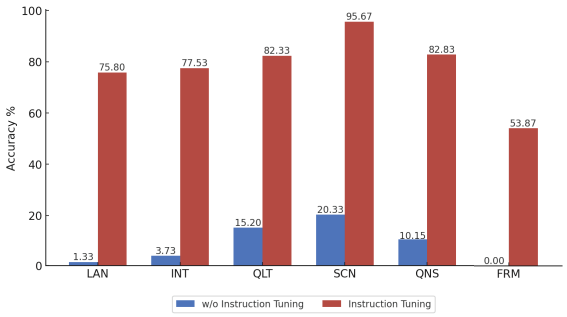

Figure 7. Comparative Evaluation of MLLM Model on QA Tasks before and after Instruction Tuning

strengths and weaknesses of each component. Key quantitative results are summarized in Tables 2, 3, 4, 5, 6, 7 and S1 and qualitative examples are shown in Figures 4, 5, 6, 7, 8 and 9 .

4.1. Traffic Sign Recognition Performance

For sign classification, assuming accurate localization, our ResNet-50 model achieved 100% test accuracy, 100%precision, and 100% recall} on GTSRB, effectively reaching state-of-the-art performance [51]. This improvement over previous ResNet results (Table 2) is attributed to augmented training data and improved understanding of m-centric road elements. The longer training time was due to a larger dataset and extended epochs for convergence. For full-image sign detection, Table 4 compares ResNet-50 (with ideal region proposals), YOLOv8, and RT-DETR. ResNet-50's metrics apply only to correctly classified sign crops (hence mAP is not applicable). YOLOv8 achieved 100% precision and 98.0% recall, yielding an F1-score of 0.80 due to occasional over-prediction of bounding boxes on one sign (counted as lowering the harmonic mean despite high raw precision/recall). RT-DETR attained 100% precision and 96.6% recall, giving a higher F1 of 0.90. YOLOv8 led in mAP@0.5 with 97.5%, slightly outperforming RT-DETR's 96.8%. Both models surpassed prior YOLO-based results, such as an improved YOLOv5 with 89% F1 [52] and YOLOv4-tiny with 91% precision, 93% recall [13].Figure 6 qualitatively shows that YOLOv8 demonstrated high real-time efficiency, excelling in detecting clear, visible signs but occasionally missing very small or motion-blurred ones.

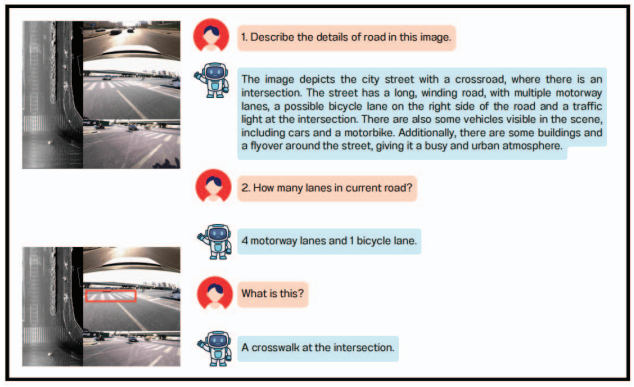

Figure 8. Multimodal Large Language Model (MLLM)-Based Urban Road Scene Analysis: Multi-View Detection of Lanes and Crosswalks.

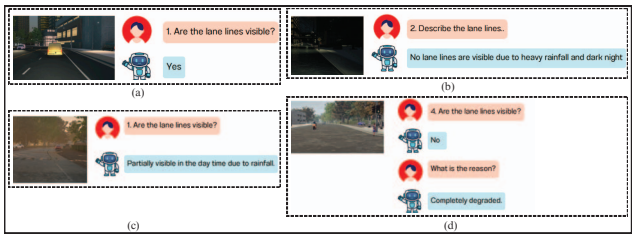

Figure 9. Qualitative Analysis of Adaptive Lane Line Detection and Reasoning. (a) Clear Visibility in Urban Environment. (b) Adverse Weather Conditions. (c) Lane Line Degradation. (d) Partial Visibility Due to Rainfall.

RT-DETR, leveraging transformer-based global reasoning, performed better in complex cases, detecting tiny or occluded signs that YOLOv8 missed (Figure 4). However, its 10 FPS inference speed is a limitation. For practical deployment, YOLOv8 offers the best speed-accuracy tradeoff, while RT-DETR is better suited for high-accuracy applications like HD map updates. Performance comparisons are illustrated in Figure 5.

4.2. MLLM based Lane Detection Performance

Our MLLM based framework achieves state-of-the-art performance in road scene understanding, significantly outperforming GPT-4V, LLaVA, and MAPLM. As shown in Table 6, our model attains the highest Frame-Overall Accuracy (FRM) at 53.87%, surpassing MAPLM (49.07%), LLaVA (38.13%), and GPT-4V (18.75%) for LLAMA-2-7b as backbone for all the compared models, demonstrating superior ability to integrate multimodal cues for comprehensive scene understanding. Additionally, it achieves the best Question-Overall Accuracy (QNS) at 82.83%, outperforming MAPLM (81.65%), LLaVA (76.08%), and GPT-4V (57.81%), proving its dominance in context-aware reasoning and structured road element recognition. Qualitative evaluations in Figure 8 further validate these results,

Table 6. Comparison of our Instruction Tuned MLLM with GPT-4V and State-of-the-Art MLLMs on MAPLM-QA Dataset.

Method	Additional Learning	Modality		Metrics (↑)					
		Img	PC	LAN	INT	QLT	SCN	QNS	FRM
MiniGPT-v2[44]	None	-	-	1.33	3.73	15.20	20.33	10.15	0
GPT-4V[49]	0-shot	✓	✓	56.25	62.53	43.75	68.73	57.81	18.75
LLaVA[50]	P+IT+LoRA	✓	✓	64.33	65.27	81.60	90.94	76.08	38.13
MAPLM[43]	P+IT	✓	✓	59.67	75.87	77.47	92.53	76.38	36.27
MAPLM[43]	P+IT+LoRA	✓	✓	72.33	56.40	78.67	93.53	81.65	49.07
Ours	IT+LoRA	×	✓	75.80	77.53	82.33	95.67	82.83	53.87

Metrics: LAN (Lane Counting), INT (Intersection Recognition), QLT (Point Cloud Quality), SCN (Scene Understanding), FRM (Frame-Overall Accuracy: 1 if all LAN, INT, QLT, and SCN are correct; else 0), QNS (Question-Overall Accuracy). Training paradigms: P (Pretraining), IT (Instruction Tuning), LoRA (Low-Rank Adaptation), PC (Point Cloud). Note: All methods utilized LLAMA-2-7b as their LLM backbone.

Table 7. Lane Detection Accuracy under Adverse Conditions Compared to SCNN.

Condition	Total Images	Correct Detections	Correct Reasoning	Accuracy (%)	SCNN Baseline (%)
Visible Lane Lines (Daytime)	500	498	-	99.6*	95.0
Visible Lane Lines (Nighttime)	500	465	-	93.0*	85.0
Partially Visible	200	162	-	81.0*	65.0
Invisible Lane Lines (Rain)	500	500	442	88.4#	70.0
Invisible Lane Lines (Degradation)	500	500	478	95.6#	80.0

The symbol (*) indicates vision-only detection accuracy, while (#) represents accuracy achieved with the reasoning module. SCNN values are sourced from the CULane benchmark. The runtime cost on a single RTX 3090, batch=1, 448×448, SafeRoute (Resnet-50+EVA-Clip+Llama-2 7B) runs at 3.8 FPS 263 ms using 14.6 GB. Under the same setup, SCNN achieves 22 FPS 45.5 ms, 1.8 GB.

with our model precisely identifying lane types, intersections, road markings, and environmental context in both urban and rural settings, a capability that baseline models fail to achieve as illustrated in Figure 7. The integration of instruction tuning and LoRA optimization significantly enhances the model's robustness and spatial reasoning, making it the most effective Vision-Language Model (VLM) for HD map updates and autonomous driving applications. For adaptive lane detection under adverse conditions, our model outperforms SCNN and other state-of-the-art approaches, achieving 99.6% accuracy for daytime lane detection and 93.0% at night, surpassing SCNN's 95.0% and 85.0%, respectively, as seen in Table 7. In partially visible lane scenarios, our model maintains 81.0% accuracy, exceeding SCNN's 65.0%, proving its resilience against occlusions and degraded visibility. Notably, in challenging conditions where lane markings are invisible, the vision-language reasoning module enables 88.4% accuracy in heavy rain and 95.6% for degraded lanes, outperforming SCNN's 70.0% and 80.0%, respectively. As shown in Figure 9, our model not only detects lane lines in extreme environments but also infers missing or degraded lanes by leveraging contextual reasoning, a capability beyond traditional segmentation-based methods. By integrating multimodal vision-language reasoning, our model ensures higher reliability, adaptability, and interpretability, making it more robust than conventional CNN-based methods. These results confirm that our LLM-augmented framework achieves state-of-the-art performance in both road scene understanding and adaptive lane detection under adverse conditions, setting a new benchmark for autonomous vehicle perception.

5. Conclusion

This work advances autonomous vehicle perception by integrating deep learning models with a Multimodal LLM for robust traffic sign recognition and lane detection. Through comparative evaluation, we show that YOLOv8 offers the best trade-off between speed and accuracy for real-time sign detection. Our proposed novel MLLM-Adapter based framework enhances lane detection under adverse conditions by leveraging multimodal reasoning and semantic-level understanding. Fine-tuned on structured datasets, the model effectively interprets complex road layouts, supports HD map updates, and generalizes well across urban and rural scenes. These contributions move toward safer, more interpretable AV systems capable of robust perception under real-world variability.

However, challenges remain particularly in poor visibility, complex lighting, and overlapping environmental factors (as shown in supplementary material sections S1.6 and S1.7. Future work will address these through advanced augmentation, large-scale end-to-end training, multisensor fusion, and further optimization and usage of other LLM and CNN backbones to improve reasoning and real-time decision-making.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (U22A20104, 52472449), the Beijing Natural Science Foundation (L231008, L243008), and the Beijing Municipal Science and Technology Commission (Z241100003524013, Z241100003524009). This work was also sponsored by the Tsinghua University-DiDi Joint Research Center, Chinese Government Scholarship and Software Engineering Institute of Beihang University.

References

[1] Noor Jannah Zakaria, Mohd Ibrahim Shapiai, and Nurbaiti Abd Ghani. Lane detection in autonomous vehicles: A systematic review. *IEEE access*, 11:3729–3765, 2023. 1

[2] Miguel Angel Sotelo, Francisco Javier Rodriguez, Luis Magdalena, Luis Miguel Bergasa, and Luciano Boquete. A color vision-based lane tracking system for autonomous driving on unmarked roads. *Autonomous Robots*, 16(1):95–116, 2004.

[3] Ze Wang, Weiqiang Ren, and Qiang Qiu. Lanenet: Real-time lane detection networks for autonomous driving. *arXiv preprint arXiv:1807.01726*, 2018. 1, 3

[4] Wenshuo Wang, Ding Zhao, Wei Han, and Junqiang Xi. A learning-based approach for lane departure warning systems with a personalized driver model. *IEEE Transactions on Vehicular Technology*, 67(10):9145–9157, 2018. 1

[5] Der-Hau Lee and Jinn-Liang Liu. End-to-end deep learning of lane detection and path prediction for real-time autonomous driving. *Signal, Image and Video Processing*, 17(1):199–205, 2023. 2, 3

[6] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[7] Ali Farzipour, Omid Nejati Manzari, and Shahriar B. Shokouhi. Traffic sign recognition using local vision transformer. In *Proc. 13th Int. Conf. on Computer and Knowledge Engineering (ICCKE)*, 2023. 2

[8] Yassmina Saadna and Ali Behloul. An overview of traffic sign detection and classification methods. *International journal of multimedia information retrieval*, 6:193–210, 2017. 2

[9] Anamika Dhillon and Gyanendra K Verma. Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, 9(2):85–112, 2020. 2

[10] Lili Zhang, Kang Yang, Yucheng Han, Jing Li, and Wei. Tsd-detr: A lightweight real-time detection transformer of traffic sign detection for long-range perception of autonomous driving. *Eng. Appl. Artif. Intell.*, 139(PA), January 2025. 2

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 2

[12] XiongFei Liu and Fan Xiong. A real-time traffic sign detection model based on improved yolov3. In *IOP Conference Series: Materials Science and Engineering*, volume 787, page 012034. IOP Publishing, 2020. 2

[13] Jie Shen, Hengsong Liao, and Li Zheng. A lightweight method for small scale traffic sign detection based on yolov4-tiny. *Multimedia Tools and Applications*, 83(40):88387–88409, 2024. 2, 7

[14] Feng Liu, Yurong Qian, Hua Li, Yongqiang Wang, and Hao Zhang. Caffnet: channel attention and feature fusion network for multi-target traffic sign detection. *International journal of pattern recognition and artificial intelligence*, 35(07):2152008, 2021. 2

[15] Jiaao Xia, Meijuan Li, Weikang Liu, and Xuebo Chen. Dsra-detr: An improved detr for multiscale traffic sign detection. *Sustainability*, 15(14):10862, 2023. 2, 6

[16] Shi Luo, Chenghang Wu, and Lingen Li. Detection and recognition of obscured traffic signs during vehicle movement. *Ieee Access*, 11:122516–122525, 2023. 2

[17] Ankit Kumar Shaw, Kun Jiang, Tuopu Wen, Chandan Kumar Sah, Yining Shi, Mengmeng Yang, Diange Yang, and Xiaoli Lian. Cleanmap: Distilling multimodal llms for confidence-driven crowdsourced hd map updates. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3798–3807, 2025. 2

[18] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37(3):362–386, 2020. 3

[19] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 286–291. IEEE, 2018. 3

[20] Der-Hau Lee, Kuan-Lin Chen, Kuan-Han Liou, Chang-Lun Liu, and Jinn-Liang Liu. Deep learning and control algorithms of direct perception for autonomous driving. *Applied Intelligence*, 51(1):237–247, 2021. 3

[21] Dev S Kochhar, Hong Zhao, Paul Watta, and Yi Murphey. Robust prediction of lane departure based on driver physiological signals. Technical report, SAE Technical Paper, 2016. 3

[22] Anamika Dhillon and Gyanendra K Verma. Convolutional neural network: A review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, 9(2):85–112, 2020. 3

[23] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[24] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 3

[25] Yunsheng Ma, Can Cui, Xu Cao, Wenqian Ye, et al. Lampilot: An open benchmark dataset for autonomous driving with language model programs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[26] Licheng Wen, Daocheng Fu, Xin Li, et al. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023. 3

[27] Zhenhua Xu, Yujia Zhang, Enze Xie, et al. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023. 3

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

[29] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *arXiv preprint arXiv:2309.05186*, 2023. 3

[30] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 3

[31] Xiaoyu Tian, Junru Gu, Bailin Li, et al. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 3

[32] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. LiDAR-LLM: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023. 3

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

[34] Daocheng Fu, Wenjie Lei, et al. Limsim++: A closed-loop platform for deploying multimodal llms in autonomous driving. *arXiv preprint arXiv:2402.01246*, 2024. 3

[35] Wayve. Lingo-1: Exploring natural language for autonomous driving, 2023. 3

[36] Arunima Jaiswal, Nitin Sachdeva, et al. Empirical analysis of traffic sign recognition using resnet architectures. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 280–285. IEEE, 2023. 3

[37] Bhavjot Singh, Sanskar Rajak, Piyush Mehra, Saksham Azad, and Deeksha Kumari. Deep neural networks for traffic sign recognition systems. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, volume 6, pages 472–477. IEEE, 2023. 3

[38] Simran Somal. Image enhancement using local and global histogram equalization technique and their comparison. In *First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019*, pages 739–753. Springer, 2020. 3

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. 4

[40] JunYoung Gwak Amir Sadeghian Ian Reid Silvio Savarese Hamid Rezatofighi, Nathan Tsoi. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019. 4

[41] Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 4

[42] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4

[43] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21819–21830, 2024. 4, 8

[44] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 5, 8

[45] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13039–13048, 2021. 6

[46] Osama Mustafa, Khizer Ali, Anam Bibi, Imran Siddiqi, and Momina Moetesum. Object detection for vehicle dashcams using transformers. *arXiv preprint arXiv:2408.15809*, 2024. 6

[47] Omran Nacir, Maraoui Amna, Werda Imen, and Belgacem Hamdi. Yolo v5 for traffic sign recognition and detection using transfer learning. In *2022 IEEE International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, volume 4, pages 1–4. IEEE, 2022. 6

[48] Guobo Xie, Zhijun Xu, Zhiyi Lin, Xingming Liao, and Teng Zhou. Grfs-yolov8: an efficient traffic sign detection algorithm based on multiscale features and enhanced path aggregation. *Signal, Image and Video Processing*, 18(6):5519–5534, 2024. 6

[49] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 8

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 8

[51] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19702–19712, 2023. 7

[52] Junfan Wang, Yi Chen, Zhekang Dong, and Mingyu Gao. Improved yolov5 network for real-time multi-scale traffic sign detection. *Neural Computing and Applications*, 35(10):7853–7865, 2023. 7