IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Sce2DriveX: A Generalized MLLM Framework for Scene-to-Drive Learning

Rui Zhao, *Member, IEEE*, Qirui Yuan, Jinyu Li, Haofeng Hu, Yun Li, *Graduate Student Member, IEEE*, Zhenhai Gao, and Fei Gao

***Abstract*—End-to-end autonomous driving, which directly maps raw sensor inputs to low-level vehicle controls, is an crucial part of Embodied AI. Despite successes in applying Multimodal Large Language Models (MLLMs) for high-level traffic scene semantic understanding, it remains challenging to effectively translate these conceptual semantics understandings into low-level motion control commands and achieve cross-scene driving generalization and consensus. We propose Sce2DriveX, a human-like chain-of-thought (CoT) driving reasoning MLLM framework, designed to achieve progressive learning from multi-view scene understanding to behavior analysis, motion planning, and vehicle control driving process. Sce2DriveX utilizes multimodal joint learning of local scene videos and global Bird's Eye View (BEV) maps to deeply understand long-range spatiotemporal relationships and road topology, enhancing its 3D dynamic/static scene perception and reasoning capabilities and achieving cross-scene generalization. Meanwhile, it reconstructs the implicit cognitive chain inherent in human driving, further enhancing the consensus between autonomous driving and human thought. To improve model performance, we construct the first comprehensive Visual Question Answering (VQA) driving instruction dataset, which tailored for 3D spatial understanding and long-axis task reasoning, and introduce a task-oriented three-stage training pipeline to support supervised fine-tuning. Extensive experiments demonstrate that Sce2DriveX achieves state-of-the-art performance across tasks from scene understanding to end-to-end driving, as well as robust generalization in handling diverse driving scenes on the CARLA Bench2Drive benchmark.**

***Index Terms*—Multimodal large language, autonomous driving, multimodal joint learning, visual question answering.**

I. Introduction

EMBODIED artificial intelligence (AI) empowers intelligent agents, such as autonomous driving (AD) models, with the capability to perceive, reason, and interact with the real world. However, a key challenge lies in achieving generalization and consensus within AD frameworks [1], [2], [3]. On one hand, AD systems struggle to generalize across complex, dynamic, and heterogeneous traffic scenarios involving diverse weather, road layouts, traffic semantics, and participant behaviors. On the other hand, AD strategies often diverge from human cognitive processes, making their decision-making opaque. These issues arise from the gap between high-level scene understanding and low-level motion control [4], [5]. *Developing a human-like, generalized framework capable of all-weather, all-scene perception, reasoning, and interpretable strategy mapping has thus become a central research pursuit.*

Recently, the rapid development of Multimodal Large Language Models (MLLMs) [6], [7], [8] have shown great promise in vision-language tasks and offer a potential solution to generalization and consensus challenges in AD. Leveraging extensive cross-modal pretraining, MLLMs possess excellent common reasoning and generalization capabilities, enabling better cross-scene understanding and adaptability. They also possess strong textual and cognitive abilities, allowing them to generate human-aligned driving thoughts and translate implicit reasoning into interpretable natural language. However, AD involves spatiotemporal continuity and dynamic global coordination, which current MLLM-based methods struggle to capture due to their reliance on single-frame front-view images [9], [10]. This limits their understanding of road semantics and temporal context. Moreover, most approaches focus narrowly on mapping scene elements to low-level control signals [4], [5], failing to exploit the MLLMs' advantages for generalized cognitive reasoning and deviates from human driving thought.

Beyond model architecture, well-matched datasets are crucial for effective training and performance. Although many datasets follow the VQA format [11], [12], [13], they fail to capture the full complexity of AD due to the gap between traffic scenes and typical VQA content. Bridging this gap requires multimodal perceptual data to model complex environments and object dynamics across frames. Furthermore, most VQA datasets focus on single driving tasks and provide only simple yes/no or limited multiple-choice answers [13], lacking the scale and diversity required for comprehensive reasoning.

To address these gaps, this work proposes Sce2DriveX framework, as shown in Fig. 1 (left), which employs modal encoders to emergently align visual representations of multi-view scene videos and BEV map images into a unified visual feature

Received 13 May 2025; accepted 3 October 2025. Date of publication 15 October 2025; date of current version 28 October 2025. This article was recommended for publication by Associate Editor A. Thakur and Editor A. Banerjee upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 52572475 and in part by the Science and Technology Development Project of Jilin Province under Grant 20250102130JC. *(Corresponding author: Fei Gao.)*

Rui Zhao, Qirui Yuan, Jinyu Li, and Haofeng Hu are with the School of Automotive Engineering, Jilin University, Changchun 130025, China (e-mail: rzhao@jlu.edu.cn; yuanqr23@mails.jlu.edu.cn; lijy1522@mails.jlu.edu.cn; huhf24@mails.jlu.edu.cn).

Yun Li is with the School of Information and Science Technology, The University of Tokyo, Tokyo 113-8654, Japan (e-mail: li-yun@g.ecc.u-tokyo.ac.jp).

Zhenhai Gao and Fei Gao are with the School of Automotive Engineering and the National Key Laboratory of Automotive Chassis Integration and Bionics, Jilin University, Changchun 130025, China (e-mail: gaozh@jlu.edu.cn; gaofei123284123@jlu.edu.cn).

Digital Object Identifier 10.1109/LRA.2025.3621971

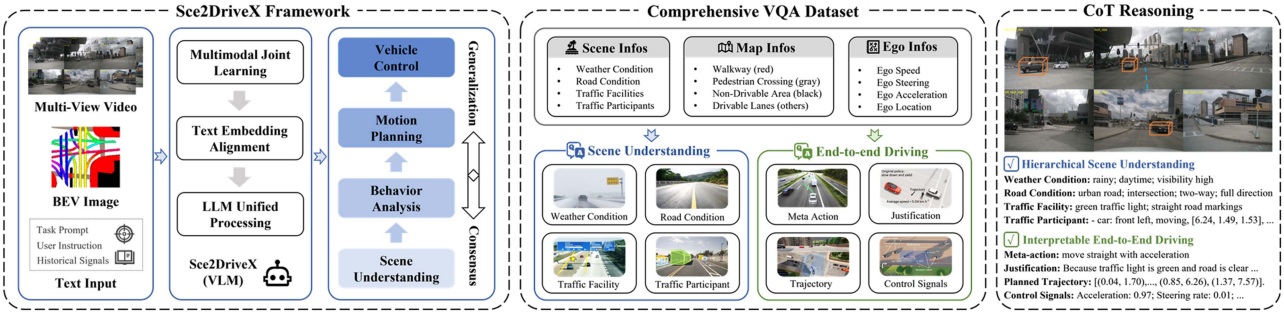

Fig. 1. Overall pipeline. This work aims to achieve cross-scene driving generalization and consensus, including two key contributions: 1) A novel MLLM framework named Sce2DriveX; 2) A comprehensive VQA driving instruction dataset.

space, then mapped to a text embedding space through a shared projection and processed by the Large Language Model (LLM) backbone to generate natural language responses, including scene understanding, behavior analysis, motion planning, and vehicle control. This multimodal joint learning of local scenes and global maps endows model a deep understanding of long-range spatiotemporal relationships and road topology features, extending its capabilities to 3D dynamic/static scene perception and reasoning, thereby achieving cross-scene driving generalization and consensus. Meanwhile, it reconstructs the implicit cognitive chain inherent in human driving, further enhancing the consensus between AD and human thought. To support model training, the first comprehensive VQA driving instruction dataset for 3D spatial understanding and long-axis task reasoning is constructed, as shown in Fig. 1 (right). It emphasizes hierarchical scene comprehension and interpretable end-to-end driving in multimodal, multi-view, multi-frame settings. A task-oriented three-stage training pipeline further guides model fine-tuning. The main contributions can be summarized as follows:

- Sce2DriveX, a human-like CoT driving reasoning MLLM framework, aimed to achieve progressive learning from multi-view scene understanding to behavior analysis, motion planning, and vehicle control driving process.
- The first comprehensive VQA driving instruction dataset for 3D spatial understanding and long-axis task reasoning, and a task-oriented three-stage training pipeline, which are used to support Sce2DriveX training and enhance its perception-reasoning capabilities.
- Extensive experiments demonstrate that Sce2DriveX achieves state-of-the-art performance across tasks from scene understanding and interpretable end-to-end driving, exhibiting robust generalization.

II. Related Works

A. Multimodal Large Language Models

With LLMs showing powerful comprehension capabilities in natural language processing (NLP), extending these capabilities to multimodal domains is a natural progression. Benefiting from scalable Transformer architecture and web-scale multimodal data, MLLMs efficiently handle visual tasks. To align text queries with visual signals, existing studies focus on modality latent space fusion. BLIP2 [6] use gated attention and Q-Former to align visual features with LLM embeddings, while LLaVA [7] applies MLPs to combine vision module with LLM. Additionally, Video-LLaVA [8] extends modality interaction to video and image, integrating visual encoders with LLMs for joint processing.

B. MLLMs in Autonomous Driving

MLLMs have potential to understand traffic scenes, optimize driving decisions, and improve human-vehicle interactions. Unlike traditional perception systems, MLLMs offer a new paradigm, leveraging their inherent few-shot learning capabilities to learn from extensive multimodal data, providing richer supervision. PromptTrack [14] integrates cross-modal features into language prompts for 3D detection and tracking. Talk2BEV [15] fuses BEV images with language prompts for driving audiovisual integration. For end-to-end driving, MLLMs enhance interpretability and trust. DriveGPT4 [4] pioneers to map sensor data and instructions to control signals and text responses. RAG-Driver [5] uses retrieval-augmented MLLM to generate driving behavior justifications and control. Senna [16] further uses MLLM to generate meta-action, guiding the end-to-end model to predict motion. However, existing works fall short in aligning MLLMs with the implicit cognitive chain of human driving, limiting cross-scene driving generalization and human-consensus driving capabilities.

C. Visual Question Answering Datasets

To enable efficient MLLM training, large-scale VQA datasets have become a key research focus. Existing datasets include image-based [11] and video-based ones [12]. In ImageQA, early methods fused CNN-extracted image features with encoded questions for answer generation, while Transformer-based models now yield stronger results. VideoQA focuses on temporal context modeling, using attention mechanisms to capture spatial-temporal relations. 3D QA [13] extends this to 3D scenes, requiring understanding of object geometry and spatial relations. Yet, challenges persist in handling complex traffic scenes with multimodal, multi-view, and multi-frame contexts, and comprehensive VQA driving datasets remain lacking.

III. VQA Driving Instruction Dataset

To support Sce2DriveX training, we construct the comprehensive VQA driving instruction dataset for 3D spatial

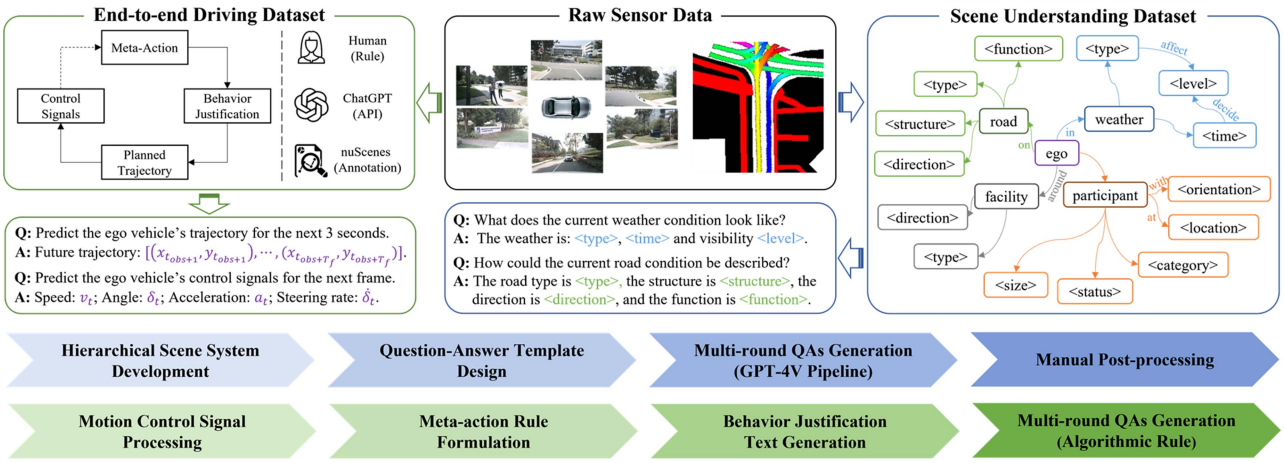

Fig. 2. Dataset construction. Based on the raw nuScenes dataset, our comprehensive VQA driving instruction dataset includes two subsets: 1) Hierarchical Scene Understanding Dataset; 2) Interpretable End-to-End Driving Dataset.

understanding and long-axis task reasoning. As shown in Fig. 2, it includes two subsets: 1) Hierarchical Scene Understanding Dataset; 2) Interpretable End-to-End Driving Dataset.

A. Hierarchical Scene Understanding Dataset

The hierarchical scene understanding dataset is generated via a scalable, fully automated pipeline, enabling hierarchical and structured descriptions of traffic scenes with weather condition, road condition, traffic facilities, and traffic participants. As shown in Fig. 2 (right), the construction involves hierarchical scene system development, QA template design, multi-round QA generation, and manual post-processing.

1) Hierarchical Scene System Development: To enhance the model's recognition of long-tail scenes, hierarchical scene system is developed, covering four scene elements: $E := \{E_{weather}, E_{road}, E_{facility}, E_{participant}\}$. Each scene element is defined multiple attributes. Specifically, weather condition $E_{weather}$ is defined by three attributes: $< type >$, $< time >$, and $< level >$, reflecting driving difficulty. Road condition E_{road} is defined by four attributes: $< type >$, $< structure >$, $< direction >$, and $< function >$, determining driving behavior. Traffic facility $E_{facility}$ is categorized into traffic signs and road markings, with their inherent attributes: $< type >$ and $< direction >$, regulating driving behavior. Traffic participant $E_{participant}$ within a specific detection range is defined by five attributes: $< category >$, $< status >$, $< size >$, $< location >$, and $< orientation >$, influencing driving states. The scene graph is also constructed to visualize the hierarchical scene system. As shown in Fig. 2, it serves as a structured enhancement of the visual scene: central node denotes the ego vehicle, intermediate layer nodes denote four scene elements, and outermost nodes denote their attributes. These nodes are connected by relational edges capturing actions (verbs) or spatial relations (prepositions).

2) QA Template Design: Four question templates are manually designed based on scene elements: $Q := \{Q_{weather}, Q_{road}, Q_{facility}, Q_{participant}\}$. To avoid overfitting to fixed patterns, each template includes multiple synonymous expressions per question option. This is achieved by collecting grammatical rules and synonym patterns to generate grammatically correct and semantically equivalent variants. Four matched answer templates are manually designed based on scene graph: $A := \{A_{weather}, A_{road}, A_{facility}, A_{participant}\}$, incorporating parameterized attributes. Scene graph's nodes and relational edges are converted into [element-relation-attribute] triplets and then transformed into fixed text. Future work may further enrich them to enhance the dataset diversity.

3) Multi-Round QA Generation and Manual Post-Processing: A fully automated pipeline with ChatGPT is used to generate multi-round QAs, using multi-view scene video frames and BEV map images as visual input, and QA templates as context, covering situational awareness, object recognition, and long-term tracking. Specifically, ChatGPT randomly selects a question option via an optimal sampling strategy, then combines visual information to produce answers using predefined templates. A known challenge with ChatGPT is hallucination tendency. To address this, manual post-processing is applied to remove inappropriate QAs, correct errors, and add missing options.

B. Interpretable End-to-End Driving Dataset

The interpretable end-to-end driving dataset is integrated via intent-cognition-oriented algorithm, enabling sequential and transparent descriptions of driving process with meta-action, behavior justification, planned trajectory, and control signals. As shown in Fig. 2 (left), the construction involves motion control signal processing, meta-action rule formulation, and behavior justification text generation.

1) Motion Control Signal Processing: The motion control signals $\mathcal{S}$ include motion trajectory $\{(x_t, y_t)\}_{t=t_{obs}-T_h}^{T_f}$ and multi-type control signals $\{a_t, \dot{\delta}_t, v_{t+1}, \delta_{t+1}\}_{t=t_{obs}-T_h}^{t_{obs}}$, where a is acceleration, $\dot{\delta}$ is steering rate, v is speed and δ is steering angle. The nuScenes annotations provide raw motion control signals, these signals are parsed into structured JSON entries

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

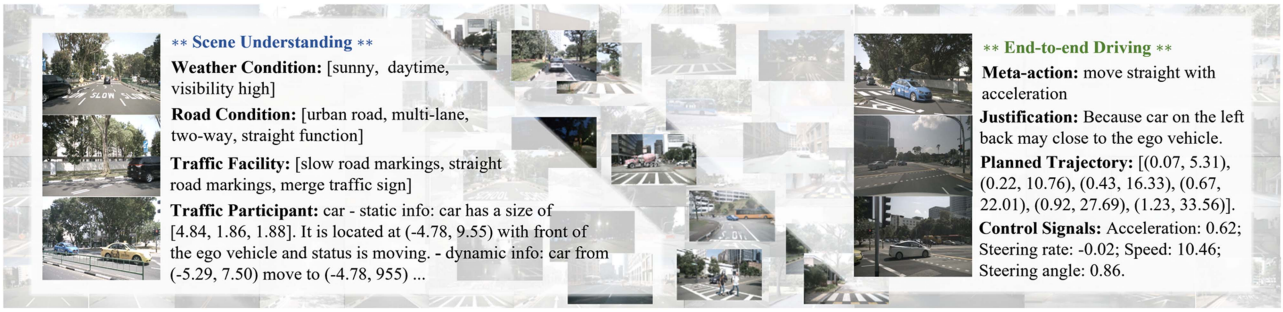

Fig. 3. Qualitative demonstration of the comprehensive VQA driving instruction dataset.

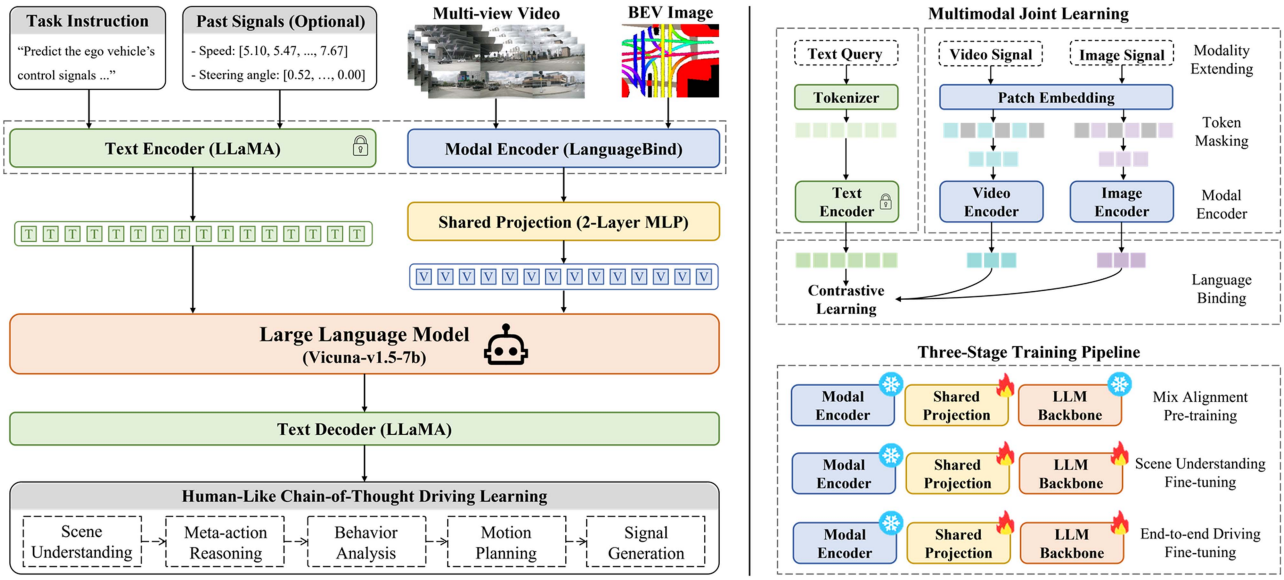

Fig. 4. Model architecture. Sce2DriveX uses modal encoders to emergently align the visual representations of multi-view scene videos and BEV map images into a unified visual feature space, which are then mapped to the LLM backbone through a shared projection.

and undergo classification processing. Specifically, historical trajectory and control signals as known information, with task background text, are integrated into system prompt. Additionally, future trajectory and control signals as ground truth labels, are filled into predefined answer templates.

2) Meta-Action Rule Formulation: The ego vehicle's meta-action $\mathcal{A}$ is defined as a combination of lateral speed level μ^{sp_x}, longitudinal speed level μ^{sp_y}, and steering level μ^{st}: $\mathcal{A} := (\mu^{sp\_x} * \mu^{st} * \mu^{sp\_y})$. Based on the predefined threshold space $\Omega = [\varepsilon^{a\_x}_{min}, \varepsilon^{a\_x}_{max}, \varepsilon^{a\_y}_{min}, \varepsilon^{a\_y}_{max}, \varepsilon^{v}, \varepsilon^{\Delta x}, \varepsilon^{\Delta \theta}]$, μ^{sp_x} includes "$rapidly$", "$slightly$" and "$moderately$", is determined by $a^{x}_{t_{obs}}$, and μ^{sp_y} includes "$(quick)\ acceleration$", "$(quick)deceleration$" and "$constantspeed$", is determined by $a^{y}_{t_{obs}}$. μ^{st} follows layered judgment principle. First level, judging $v_{t_{obs}}$ and $\Delta x_{t_{obs}}$ to determine "$idle$" and "$move\ straight$" respectively. Second level, traversing future timestep that satisfies the straight-moving, and calculating $\Delta\theta^{*}_{t_{obs}}$ and $\Delta x^{*}_{t_{obs}}$. Third level, judging $\Delta\theta^{*}_{t_{obs}}$ and sign of $\Delta x^{*}_{t_{obs}}$ to determine "$turn\ left/right$" and "$shift\ to\ left/right$". Through the above method, 64 meta-action types are generated, comprehensively simulating the vehicle's behavioral patterns in diverse scenarios.

3) Behavior Justification Text Generation: The behavior justification text $\mathcal{T}$ provides an analysis of the ego vehicle's short-term driving strategy, improving the interpretability of decision-making. Leveraging scene QAs and meta-actions as contextual inputs, ChatGPT's API automatically generates these justifications. Compared with manual annotations, this method yields more diverse and accurate descriptions, capturing potential traffic and social factors more comprehensively.

In summary, this work builds a large-scale VQA driving instruction dataset for hierarchical scene understanding and interpretable end-to-end driving, encompassing 22,697 scenes and 158.9 K QAs. Visual examples are shown in Fig. 3.

IV. Methodology

A. Sce2DriveX Framework

This work develops Sce2DriveX, a human-like CoT driving reasoning MLLM framework. As shown in Fig. 4, Sce2DriveX consists of four components: 1) Modal Encoder, including video

encoder $f_{\mathrm{V\_E}}$ and image encoder $f_{\mathrm{I\_E}}$, initialized from OpenCLIP; 2) Shared Projection f_{P}, using two-layer MLP; 3) LLM Backbone f_{LLM}, employing Vicuna-v1.5-7b; 4) Text Encoder $f_{\mathrm{T\_E}}$ and Text Decoder $f_{\mathrm{T\_D}}$, provided by LLaMA.

1) Multimodal Joint Learning: Given text instruction $\mathbf{X_T}$, BPE tokenizer is first used to segment the words into relatively common subwords, each corresponding to a unique logit. Then, these logits are encoded using the text encoder $f_{\mathrm{T\_E}}$:

$$\mathbf{H_T} = f_{\mathrm{T\_E}}(BPE(\mathbf{X_T})) \tag{1}$$

where $\mathbf{H_T} \in \mathbb{R}^{M_{txt} \times L}$ denotes the text tokens, M_{txt} is the number of text tokens, and L is the LLM's feature dimension. Given multi-view video $\mathbf{X_V} \in \mathbb{R}^{T \times H \times W \times C}$ and BEV image $\mathbf{X_I} \in \mathbb{R}^{H \times W \times C}$, where T is the number of video frames, (H, W) is original image's resolution, and C is the channels, patch masking method is adopted. Some patches are selected by encoder mask $\mathbb{M}_e$ and segmented to alleviate the issue of excessive token numbers. Specifically, $\mathbf{X_V}$ and $\mathbf{X_I}$ are first converted into corresponding patches $\mathbf{P_V} \in \mathbb{R}^{T \times N_p \times C}$ and $\mathbf{P_I} \in \mathbb{R}^{N_p \times C}$ through patch embedding with non-overlapping filters, where $N_p = \frac{H \times W}{B^2}$ is the number of patches, and B is each patch's size. Then, positional embeddings are applied to the visible token, divided by $\mathbb{M}_e$, forming the combined $\mathbf{S_V}$ and image sequence $\mathbf{S_I}$:

$$\mathbf{S_V} = \{\mathbf{P_V} + \mathbf{Q}_i\}_{i \in \mathbb{M}_e}, \quad \mathbf{S_I} = \{\mathbf{P_I} + \mathbf{Q}_j\}_{j \in \mathbb{M}_e} \tag{2}$$

where $\mathbf{Q}$ denotes a series of learnable positional tokens. Finally, video sequence $\mathbf{S_V}$ and image sequence $\mathbf{S_I}$ are encoded by video encoder $f_{\mathrm{V\_E}}$ and image encoder $f_{\mathrm{I\_E}}$ respectively:

$$\mathbf{H_V} = f_{\mathrm{V\_E}}(\mathbf{S_V}), \quad \mathbf{H_I} = f_{\mathrm{I\_E}}(\mathbf{S_I}) \tag{3}$$

where $\mathbf{H_V} \in \mathbb{R}^{T \times M_{vid} \times V}$ denotes the video tokens, $\mathbf{H_I} \in \mathbb{R}^{M_{img} \times V}$ denotes the image tokens, M_{vid} is the number of video tokens, M_{img} is the number of image tokens, and V is the unified visual feature dimension. Notably, to achieve multimodal semantic alignment, modal encoding approach of LanguageBind [17] is adopted, which uses text as the bridge between different modalities. Through contrastive learning principles, other modalities are bound to the text modality and emergently aligned to the unified visual feature space.

2) LLM Backbone Unified Processing: This work maps multimodal tokens into the text embedding space to create unified visual representation, which is then combined with tokenized text instruction and fed into the LLM backbone for response generation. Specifically, shared projection f_{P} is first used to map the video tokens $\mathbf{H_V}$ and image tokens $\mathbf{H_I}$:

$$\mathbf{H_L} = f_{\mathrm{P}}(\mathbf{H_V}, \mathbf{H_I}) \tag{4}$$

where $\mathbf{H_L} \in \mathbb{R}^{T \times M_{vsl} \times L}$ denotes the unified visual tokens, sharing the same feature dimension as text tokens $\mathbf{H_T}$, M_{vsl} is the number of visual tokens. Next, $\mathbf{H_L}$ are combined with $\mathbf{H_T}$ and fed into the LLM backbone f_{LLM} to generated corresponding predicted tokens, which are finally decoded into natural language responses $\mathbf{Z}$ by the text decoder $f_{\mathrm{T\_D}}$:

$$\mathbf{Z} = f_{\mathrm{T\_D}}(f_{\mathrm{LLM}}(\mathbf{H_L} \oplus \mathbf{H_T})) \tag{5}$$

where $\oplus$ denotes the concatenation, $\mathbf{Z}$ includes scene understanding $\mathbf{Z_{sce}}$, meta-action reasoning $\mathbf{Z_{act}}$, behavior analysis $\mathbf{Z_{bev}}$, motion planning $\mathbf{Z_{mot}}$, and control signal generation $\mathbf{Z_{sig}}$.

B. Training Pipeline

To enhance the Sce2DriveX's perception-reasoning capabilities, this work introduces a task-oriented three-stage training pipeline: 1) Mixed Alignment Pre-training; 2) Scene Understanding Fine-tuning; and 3) End-to-End Driving Fine-tuning.

1) Mixed Alignment Pre-Training: This stage aligns image/video features with the LLM backbone. Sce2DriveX is pretrained on CC3M and WebVid-10 M dataset, covering various topics beyond a domain. During this stage, the video encoder, image encoder, and LLM backbone weights are frozen, with only the shared projection trained.

2) Scene Understanding Fine-Tuning: This stage improves the model's 3D spatial perception capabilities for hierarchical scene understanding task. Sce2DriveX is fine-tuned on the hierarchical scene understanding dataset. Considering that the model's responses are in natural language, text cross-entropy loss is used to supervise its reasoning process:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{l} p_\theta(\mathbf{Z}^{[i]} \mid \mathbf{X_V}, \mathbf{X_I}, \mathbf{X_T}, \mathbf{Z}^{[1:i-1]}) \tag{6}$$

where θ denotes the trainable parameters, and l denotes the length of the response $\mathbf{Z}$. During this stage, the video encoder and image encoder weights are frozen, with the LLM backbone and shared projection trained.

3) End-to-End Driving Fine-Tuning: This stage further strengthens the model's long-horizon reasoning for interpretable end-to-end driving, using the same training strategy as the scene understanding fine-tuning stage.

V. Experiments

A. Dataset Analysis

Our dataset has notable advantages: 1) Visual data is multimodal, including multi-view videos and BEV images, requiring the extraction of spatial and temporal information. 2) Dataset is large-scale, containing 22,697 diverse visual scenes and 158.9 K QAs. 3) QAs are automatically collected, saving significant human effort and time costs. 4) Focus on outdoor scenarios, increasing the authenticity and complexity of dataset. Furthermore, Statistical analysis of the dataset is performed. Fig. 5 illustrates the distribution of meta-actions (top 20), validating the consistency and balance of the dataset.

B. Autonomous Driving Task Evaluation

1) Experimental Setup: Images are uniformly cropped to 224×224, videos are sampled 8 frames evenly and performed same processing. Each batch includes a mix of images and videos. During the pre-training, model is trained for 1 epoch with a batch size of 128, distributed across 6 A100 GPUs. During the fine-tuning, AdamW optimizer and cosine learning

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

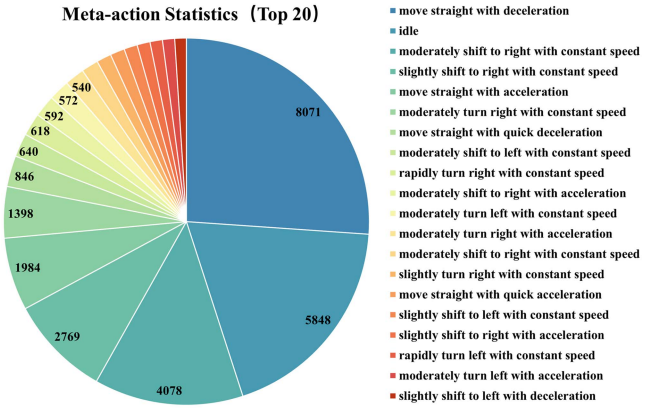

Fig. 5. Distributions of meta-actions (top20).

TABLE I
QUANTITATIVE RESULTS IN HIERARCHICAL SCENE UNDERSTANDING

Method	Ele.	M	B4	R	C	Acc.
Qwen-VL [22]	$E_{wea.}$	61.22	6.94	86.65	698.05	89.16
	$E_{roa.}$	73.81	7.19	84.88	721.83	78.24
	$E_{fac.}$	73.81	7.67	91.93	789.95	90.64
	$E_{part.}$	37.45	4.44	57.23	564.46	64.45
	Avg.	61.57	6.56	80.17	693.57	80.62
LLaVANeXT [23]	$E_{wea.}$	60.88	6.94	86.63	697.69	89.24
	$E_{roa.}$	73.62	7.21	85.46	723.04	80.20
	$E_{fac.}$	73.62	7.58	91.17	787.40	89.72
	$E_{part.}$	41.44	4.64	65.55	584.89	71.24
	Avg.	62.39	6.59	82.20	698.26	82.60
Sce2DriveX	$E_{wea.}$	**60.89**	**7.01**	**87.75**	**713.68**	**89.15**
	$E_{roa.}$	**75.70**	**7.71**	**87.64**	**757.63**	**87.56**
	$E_{fac.}$	**75.70**	**7.84**	**92.52**	**808.03**	**93.71**
	$E_{part.}$	**58.06**	**6.30**	**76.97**	**621.21**	**81.03**
	Avg.	**67.59**	**7.22**	**86.22**	**725.14**	**87.86**

rate scheduler are used, with an initial learning rate of 2e-5, warm-up ratio of 0.03, and gradient accumulation step of 2. Scene understanding fine-tuning trains for 1 epoch, while end-to-end driving fine-tuning trains for 3 epochs, both completed on 8 L20 GPUs, with a batch size of 4 per GPU.

2) Hierarchical Scene Understanding Evaluation: All evaluation metrics for hierarchical scene understanding task are reported, including NLP metrics and accuracy (Acc.), where NLP metrics consist of METEOR (M) [18], BLEU-4 (B4) [19], ROUGE (R) [20], and CIDEr (C) [21]. Table I presents the comparison results between Sce2DriveX and open-source VLM baselines (Qwen-VL [22] and LLaVANeXT [23]) in hierarchical scene understanding, where Qwen-VL inputs multi-view images and BEV map, while LLaVANeXT only inputs multi-view video. Sce2DriveX achieves the highest scores and accuracy, especially in the road condition and traffic participants, significantly outperforming the baseline models, benefiting from the multimodal joint learning of local scene videos and global BEV maps.

3) Interpretable End-to-End Driving Evaluation: All evaluation metrics are reported. For motion planning, L2 error (m) and collision rate ($\%$) are used to evaluate the planned trajectory

TABLE II
COMPARISON RESULTS IN MOTION PLANNING

Method	L2(m)↓				Collision(%)↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
NMP [1]	-	-	2.31	-	-	-	1.92	-
ST-P3 [24]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
FF [25]	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
VAD [2]	0.17	0.34	0.60	0.37	0.04	0.27	0.67	0.33
UniAD [3]	0.48	0.96	1.65	1.03	0.02	0.17	0.71	0.31
GPT-Driver [9]	0.27	0.74	1.52	0.84	0.07	0.15	1.10	0.44
OmniDrive [26]	0.40	0.80	1.32	0.84	0.04	0.46	0.78	0.43
DriveVLM [10]	0.18	0.34	0.68	0.40	0.10	0.22	0.45	0.27
Sce2DriveX	**0.15**	**0.33**	**0.59**	**0.36**	**0.00**	**0.12**	**0.58**	**0.23**

within a 3-second horizon. For meta-action reasoning, weighted accuracy αAcc ($\%$) is used, with steering level weighted at 0.7 and speed level weighted at 0.3. For behavior analysis, four NLP metrics and GPT score [4] are used. For control signal generation, including the next-step speed (m/s) and steering angle ($^\circ$), as well as the current-frame acceleration (m/s^2) and steering rate ($^\circ/s$), RMSE is used.

Table II compares Sce2DriveX with baseline methods in motion planning on the nuScenes dataset. Compared to small model-based and LM-based approaches, Sce2DriveX achieves the lowest L2 error (0.36 m) and collision rate (0.23 $\%$), confirming its strong planning performance. To further assess interpretability, the VQA driving instruction dataset (Section III) is used to evaluate recent MLLM-based methods—DriveGPT4 and RAG-Driver. As shown in Table III, Sce2DriveX attains the highest weighted accuracy of 94.29 $\%$, indicating precise reasoning of future meta-actions. In behavior analysis, it also surpasses others in NLP metrics and GPT score, improving decision interpretability. Moreover, its lower RMSE demonstrates robust control signal prediction.

C. Ablation Study

Ablation study is conducted to evaluate the contributions of the proposed VQA driving dataset and the multimodal joint learning strategy. The specific settings are as follows: 1) using in-context learning method (ICL); 2) modifying the CoT QAs, including omitting the scene QAs (w/o Scene QAs) and omitting the meta-action and explanation QAs (w/o Action QAs); 3) modifying the visual inputs, including changing to front-view scene video (w/o Multi-View), removing the multi-view video-only using a single multi-view image (w/o Video Input) and removing the BEV map (w/o BEV Input). Table IV provides the quantitative results of the ablation study. When any part of the VQA dataset or model visual input is altered, Sce2DriveX's performance declines, validating the effectiveness of the proposed framework design.

D. Real-Time Analysis

To evaluate the real-time performance of Sce2DriveX, test are conducted on an L20 (48 GB) GPU. Table V shows the parameter count (M), time cost (s), and FLOPs (T) for each

TABLE III
COMPARISON RESULTS IN META-ACTION REASONING, BEHAVIOR ANALYSIS AND CONTROL SIGNAL GENERATION

Method	αAcc(%)↑	M↑	B4↑	R↑	C↑	GPT↑	RMSE acc(m/s^2)↓	RMSE rat($^\circ/s$)↓	RMSE spd(m/s)↓	RMSE ang($^\circ$)↓
DriveGPT4 [4]	90.86	18.75	5.18	31.89	130.60	89.89	0.388	0.032	0.096	0.511
RAG-Driver [5]	93.06	19.52	6.13	32.99	131.26	90.78	0.271	0.025	0.089	0.449
Sce2DriveX	**94.29**	**20.19**	**6.69**	**33.21**	**131.99**	**91.11**	**0.241**	**0.021**	**0.081**	**0.427**

TABLE IV
ABLATION STUDY ON THE MODULE DESIGN

Design	αAcc(%)	M	B4	R	C	GPT	L2(m)	RMSE acc(m/s^2)	RMSE rat($^\circ/s$)	RMSE spd(m/s)	RMSE ang($^\circ$)
ICL	81.46	12.69	2.11	23.45	101.36	75.56	1.01	0.401	0.051	0.101	0.814
w/o Scene QAs	86.35	17.11	3.95	28.31	120.48	79.99	0.74	0.282	0.038	0.089	0.675
w/o Action QAs	-	-	-	-	-	-	0.38	0.246	0.025	0.084	0.515
w/o Multi-View	91.56	19.17	5.95	31.74	124.29	83.14	0.51	0.250	0.026	0.084	0.594
w/o Video Input	88.17	19.09	5.37	30.02	122.41	81.05	0.52	0.254	0.028	0.085	0.613
w/o BEV Input	94.00	19.19	6.28	31.95	126.76	86.25	0.39	0.243	0.022	0.083	0.486
Sce2DriveX	**94.29**	**20.19**	**6.69**	**33.21**	**131.99**	**91.11**	**0.36**	**0.241**	**0.021**	**0.081**	**0.427**

TABLE V
RUNTIME AND OCCUPANCY OF SCE2DRIVEX'S MODULE

Module	Params (M)	Time (s)	FLOPs (T)
Image Encoder	302.92	0.007	0.078
Video Encoder	403.73	0.013	0.830
LLM Backbone	6628.32	0.569	7.695
		1.524	20.439
		0.869	11.583

module of the model. Image encoder (302.92 M) has an encoding time and TFLOPs of 0.007 s and 0.078, and video encoder (403.73 M) has an encoding time and TFLOPs of 0.013 s and 0.830. LLM backbone (6628.32 M) is the primary computational bottleneck, time cost and TFLOPs for three rounds of QA are showed, where motion planning QA has the longest inference time, reaching 1.524 s, mainly due to the large number of tokens relative to the trajectory in numeric form. Overall, the total inference time for LLM backbone is 2.982 s, with total TFLOPs is 39.717. Lightweight techniques like knowledge distillation and bit quantization can reduce Sce2DriveX's inference latency and enhance its deployment.

E. Qualitative Demonstration

Fig. 6 presents visualized cases from the VQA driving instruction dataset to validate Sce2DriveX's outstanding performance. For clarity, only the current full-view frame and BEV map are displayed. These cases provide compelling evidence of Sce2DriveX's intrinsic capabilities: when confronted with complex outdoor driving scenes, Sce2DriveX can perform progressive reasoning learning from hierarchical scene understanding to interpretable end-to-end driving.

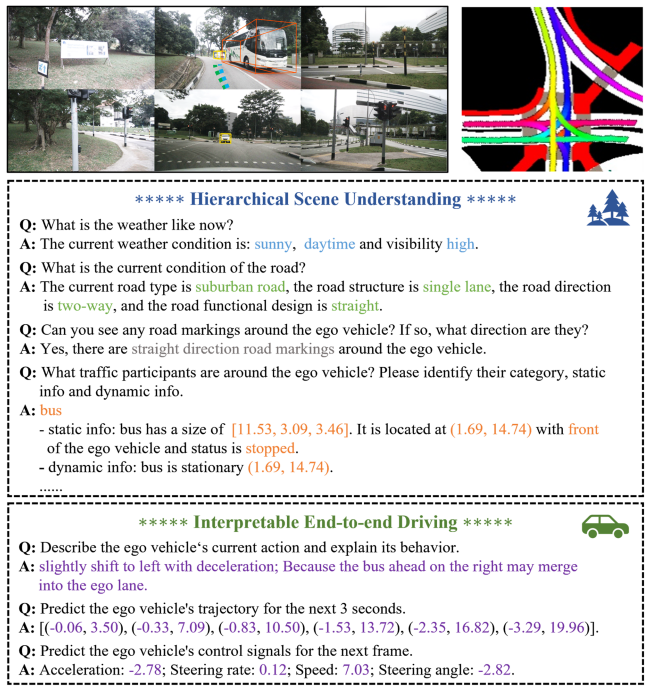

Fig. 6. Qualitative demonstrations of Sce2DriveX on the VQA driving instruction dataset.

F. Generalization Testing

To assess the robust generalization of Sce2DriveX, cross-dataset testing is conducted. Specifically, 3,000 corner cases with distinct styles from the CARLA Bench2Drive benchmark [27] are selected. Under a zero-shot transfer protocol, Sce2DriveX achieved 90.11% weighted accuracy (meta-action reasoning), along with 0.42m L2 error and 0.28% collision rate (motion planning). Additionally, RMSE for speed, steering angle,

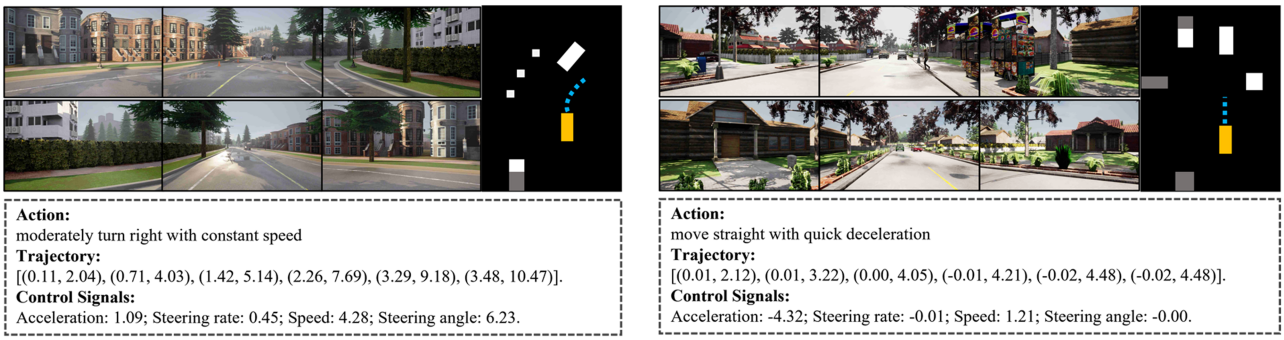

Fig. 7. Visualization results of generalization testing (corner cases from the CARLA Bench2Drive benchmark).

acceleration, and steering rate are 0.308 m/s, 0.031°, $0.087 m/s^2$, and $0.455°/s$, respectively. Fig. 7 further visualizes the qualitative generalization results.

VI. Conclusion

This paper presents Sce2DriveX, a framework that enables progressive reasoning from hierarchical scene understanding to interpretable end-to-end driving. By jointly learning from multimodal local scenes and global maps, it captures long-range spatiotemporal and road-topology features, facilitating cross-scene driving generalization. Furthermore, we construct the first comprehensive VQA driving instruction dataset for 3D spatial understanding and long-horizon reasoning, and design a three-stage task-oriented supervised fine-tuning pipeline. Experimental results demonstrate that Sce2DriveX achieves strong performance in scene understanding, meta-action reasoning, behavior analysis, motion planning, and control signal generation. Future work will focus on enhancing real-time efficiency through soft-label distillation and bit-level quantization.

References

[1] W. Zeng et al., "End-to-end interpretable neural motion planner," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8660–8669.
[2] B. Jiang et al., "VAD: Vectorized scene representation for efficient autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8340–8350.
[3] Y. Hu et al., "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17853–17862.
[4] Z. Xu et al., "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robot. Automat. Lett.*, vol. 9, no. 10, pp. 8186–8193, Oct. 2024.
[5] J. Yuan et al., "RAG-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," 2024, *arXiv:2402.10828*.
[6] 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 *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.
[7] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26296–26306.
[8] B. Lin et al., "Video-LLaVa: Learning united visual representation by alignment before projection," 2023, *arXiv:2311.10122*.
[9] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "GPT-driver: Learning to drive with GPT," 2023, *arXiv:2310.01415*.
[10] X. Tian et al., "DriveVLM: The convergence of autonomous driving and large vision-language models," 2024, *arXiv:2402.12289*.
[11] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *Proc. Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2018, pp. 1–10.
[12] Z. Yu et al., "ActivityNet-QA: A dataset for understanding complex Web videos via question answering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 9127–9134.
[13] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-QA: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4542–4550.
[14] D. Wu et al., "Language prompt for autonomous driving," in *Proc. AAAIConf. Artif. Intell.*, vol. 39, no. 8, 2025, pp. 8359–8367.
[15] T. Choudhary et al., "Talk2BEV: Language-enhanced bird's-eye view maps for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 16345–16352.
[16] B. Jiang et al., "Senna: Bridging large vision-language models and end-to-end autonomous driving," 2024, *arXiv:2410.22313*.
[17] B. Zhu et al., "LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment," 2023, *arXiv:2310.01852*.
[18] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[20] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text summarization Branches Out*, 2004, pp. 74–81.
[21] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2015, pp. 4566–4575.
[22] J. Bai et al., "Qwen-Vl: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023. [Online].Available: https://arxiv.org/abs/2308.12966
[23] B. Li et al., "LLaVA-Onevision: Easy visual task transfer," 2024. [Online]. Available: https://arxiv.org/abs/2408.03326
[24] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 533–549.
[25] P. Hu, A. Huang, J. Dolan, D. Held, and D. Ramanan, "Safe local motion planning with self-supervised freespace forecasting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12732–12741.
[26] S. Wang et al., "Omnidrive: A holistic LLM-agent framework for autonomous driving with 3D perception, reasoning and planning," *CoRR*, 2024.
[27] X. Jia, Z. Yang, Q. Li, Z. Zhang, and J. Yan, "Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 819–844, 2024.