Delving Into Multi-Modal Multi-Task Foundation Models for Road Scene Understanding: From Learning Paradigm Perspectives

Sheng Luo, Wei Chen, Wanxin Tian, Rui Liu, Luanxuan Hou, Xiubao Zhang, Haifeng Shen, Ruiqi Wu, Shuyi Geng, Yi Zhou, Ling Shao, *Fellow, IEEE*, Yi Yang, *Member, IEEE*, Bojun Gao, Qun Li, and Guobin Wu

***Abstract*—Foundation models have indeed made a profound impact on various fields, emerging as pivotal components that significantly shape the capabilities of intelligent systems. In the context of intelligent vehicles, leveraging the power of foundation models has proven to be transformative, offering notable advancements in visual understanding. Equipped with multi-modal and multi-task learning capabilities, multi-modal multi-task visual understanding foundation models (MM-VUFMs) effectively process and fuse data from diverse modalities and simultaneously handle various driving-related tasks with powerful adaptability, contributing to a more holistic understanding of the surrounding scene. In this survey, we present a systematic analysis of MM-VUFMs specifically designed for road scenes. Our objective is not only to provide a comprehensive overview of common practices, referring to task-specific models, unified multi-modal models, unified multi-task models, and foundation model prompting techniques, but also to highlight their advanced capabilities in diverse learning paradigms. These paradigms include open-world understanding, efficient transfer for road scenes, continual learning, interactive and generative capability. Moreover, we provide insights into key challenges and future trends, such as closed-loop driving systems, interpretability, low-resource conditions, embodied driving agents, and world models.**

***Index Terms*—Foundation model, multi-modal learning, multi-task learning, road scene, visual understanding.**

I. Introduction

INTELLIGENT vehicles have made great progress by achieving significant advancements in perceiving road scenes, employing various tasks such as object detection, trajectory prediction, and advanced generation methods. The integration of vision-centric (such as camera, Lidar) and vision-beyond (such as text, action) modalities further enhances the ability of autonomous vehicles to perceive the world from diverse dimensionalities.

However, traditional driving-related models are typically designed for specific tasks and utilize common convolution neural networks (CNNs) [70] or Transformer [23], [53], [239] architecture to extract feature maps from a single modality. This fashion often leads to incomplete observation due to complex and unpredictable conditions in real-world road scenes.

To address this problem, researchers have shifted their focus to unified multi-modal and multi-task models, as illustrated in Fig. 1. These models possess multi-task learning capabilities, allowing them to concurrently perform multiple tasks. Moreover, with the integration of multi-modal capabilities (e.g., various visual sensors, text), these models contribute to creating more versatile systems.

The emergence of Foundation Models (FMs) is regarded as a milestone in achieving artificial general intelligence (AGI). Recently, large language models (LLMs) [20], [141], [142], [143], [175], [176], vision language models (VLMs) [44], [113], [114], [201], [228], [247], and large vision models (LVMs) [10], [94], have already drawn significant attention. LLMs require training on a large amount of textual data and considerable computational resources, showing powerful emergent abilities with continuous growth of data scale. With the extensive world knowledge from the pretraining stage, LLMs have been viewed as knowledge bases to solve downstream tasks by prompt engineering. VLMs refer to a series of models designed to bridge the gap between visual information and natural language understanding (e.g., Vision Question Answer(VQA), Captioning), emphasizing on aligning vision modality into language latent space and thereby enhancing the capacity to understanding and reasoning about visual content. Moreover, LVMs aim to achieve powerful strengths in vision-centric tasks without relying on linguistic data. These FMs with exceptional generalization capabilities have achieved great success across various domains.

Motivation of our survey: Among traditional algorithms for road scene understanding, each algorithm is individually designed to solve a specific task, and is usually trained on a single

Manuscript received 12 March 2024; revised 9 May 2024; accepted 20 May 2024. Date of publication 28 May 2024; date of current version 11 July 2025. This work was supported by DiDi GAIA Research Cooperation Initiative under Grant CCF-DiDi GAIA 202304. *(Corresponding author: Yi Zhou.)*

Sheng Luo, Wei Chen, Ruiqi Wu, Shuyi Geng, and Yi Zhou are with the School of Computer Science and Engineering, Southeast University, Nanjing 211102, China, and also with the Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Southeast University, Ministry of Education, Nanjing 211102, China (e-mail: yizhou.szcn@gmail.com).

Wanxin Tian, Rui Liu, Luanxuan Hou, Xiubao Zhang, Haifeng Shen, Yi Yang, Bojun Gao, Qun Li, and Guobin Wu are with the Didi Chuxing, Beijing 100080, China.

Ling Shao is with the UCAS-Terminus AI Lab, University of Chinese Academy of Sciences, Beijing 101408, China.

To facilitate researchers in staying abreast of the latest developments in MM-VUFMs for road scenes, we have established a continuously updated repository at https://github.com/rolsheng/MM-VUFM4DS.git.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2024.3406372.

Digital Object Identifier 10.1109/TIV.2024.3406372

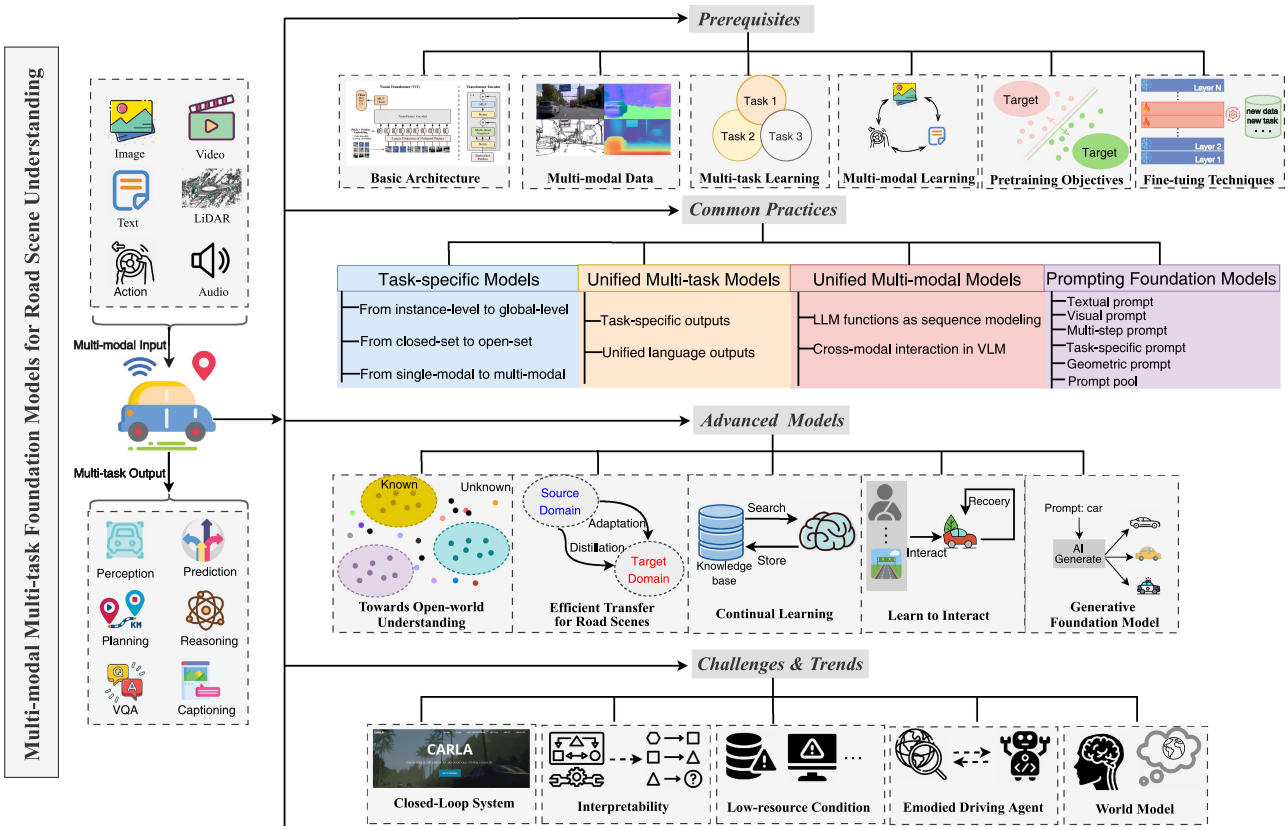

Fig. 1. Overview of our survey at a glance. A multi-modal and multi-task foundation model for road scene understanding is defined as a framework that inputs multi-modal data and outputs multi-task results. In the section of prerequisites, we introduce some basic knowledge in advance before reading the main context. Then, we refer to up-to-date task-specific models, unified multi-task models, unified multi-task models for road scene understanding and prompting foundation models, respectively, in the section of common practice. The section of advanced models is to show strengths in diverse learning paradigms, such as open-world understanding, efficient transfer for road scene, continual learning, interactive and generative capabilities, respectively. Finally, we also list key challenges and promising future trends to address them.

modality. This is beneficial for easy assembly and deployment. However, we argue that these algorithms are inefficient and impractical in real-world scenarios. 1) First, limited knowledge can be learned from a single task, hindering mutual benefit from learning universal knowledge across multiple tasks. 2) Second, the acquisition and processing of multi-modal data have become increasingly feasible due to the advancement of multi-modal learning and the flexibility of Transformer-based architecture, respectively. By integrating these multi-modal data into a unified multi-modal model, we can gain comprehensive understanding of road scene characteristics. 3) Third, the intricacies of road scenes demand systems that possess a profound understanding of the surrounding environment. Motivated by the great success of FMs, the realization of their pivotal role in the visual understanding of road scenes has taken center stage. The great generalization capabilities of MM-VUFMs make them an ideal solution to tackle this issue. Due to the emergence of numerous related research papers and the absence of a review in the current landscape, there is an urgent need for a comprehensive and systematic survey for MM-VUFMs.

Comparisons with related surveys: We investigated previous related surveys to clarify a difference between our survey with them [24], [42], [52], [61], [101], [154], [166], [215], [223], [244]. Some previous surveys [24], [42], [166], [223], [244] cover related datasets, methods, and applications from a perspective of end-to-end autonomous driving (e.g., perception, prediction, and planning), causing a lack of detailed attention to visual understanding. The survey [215] focuses on specific views such as data generation,self-supervised learning, and adaptation, but they have no consideration of multi-modal and multi-task capabilities of FMs. However, our survey not only reviews the latest works of MM-VUFMs from the perspective of multi-modal and multi-task capabilities but also has more emphasis on their advanced strength in diverse learning paradigms, aiming to provide readers with comprehensive awareness and in-depth insights toward this field.

Contributions: In the end, we summarize four contributions of this survey as follows:

- We provide a systematic analysis of up-to-date (until May, 2024) MM-VUFMs for road scenes, covering high-level motivation, common practices, advanced capabilities in diverse learning paradigms, emerging challenges and future trends.
- We classify existing MM-VUFMs into task-specific models, unified multi-task, unified multi-modal models, and foundation model prompting techniques, respectively. We also review datasets for road scenes from multi-modal and multi-task perspectives, and typical evaluation metrics for measuring performance.

- Advanced strengths of MM-VUFMs in diverse learning paradigms are highlighted, including open-world understanding, efficient transfer, low-resource condition, continual learning, interactive and generative capability.
- Emerging key challenges and promising trends are proposed to draw attention to achieve intelligent vehicles.

II. Prerequisites and Roadmap

A. Prerequisites

1) Basic Architectures: Convolutional neural networks (CNNs) [70] are classic architectures for computer vision tasks, opening a new era for computer vision and continuing to maintain their prominence and impact in the era of large-scale foundation models [48], [151], [186]. Despite CNN's long-range dependencies through applying large kernels or recursive gated kernels, the performance is still limited in scaling up in-context learning. Transformer [180] is an attention-based sequence-to-sequence learning architecture originally introduced to extract long-range dependencies from natural language. Unlike previous models that relied on recurrent or convolutional layers, Transformer exclusively employs the self-attention [180] mechanism to weigh different parts of the input sequence, enabling parallelization and improving efficiency in capturing long-range dependencies.

Vision-Language architecture [2], [98], [151]. The success of Transformer in both language and vision domains has spurred the development of vision-language architecture. This architecture leverages Transformers' ability to handle serialized multi-modal data. To be more specific, they often employ mechanisms such as cross-attention, which functions by enabling queries from one modality(e.g., text) to interact with keys and values from another modality(e.g., image), facilitating a context-rich understanding that is essential for tasks requiring joint visual and textual comprehension such as image captioning, visual question answering (VQA).

2) Multi-Modal Data: In road scenes, multi-modal data play a critical role in enhancing the perception, decision-making, and interaction capabilities of intelligent vehicles. These data come from various sources and types, each contributing unique information essential for the comprehensive functioning of road scene understanding [59]. Here we introduce several mainstream multi-modal data used in road scenes in Fig. 2. Based on their data modality, we categorize them into two groups: **vision-centric** multi-modal data which include the most popular used sensor data in the past decades, and **vision-beyond** multi-modal data that include other types of multi-modal data springing up in recent few years.

Vision-centric multi-modal data include images and videos captured by various cameras such as RGB cameras, depth cameras, thermal cameras and lidar. RGB image I can be represented as $I \in \mathbb{R}^{H\times W\times C}$ where H and W stand for the height and width of the image respectively. The symbol C represents the number of channels: 3 for RGB images, and 1 for depth or thermal images. A sequence of images forms a video $V \in \mathbb{R}^{T\times H\times W\times 3}$ where T stands for the number of frames. Depth images measure the distance between the object and the ego-car, and thermal images help detect moving objects with relatively high temperature. Point clouds from the lidar, featuring robustness to various weather conditions, can be represented as $P \in \mathbb{R}^{N\times 3}$ where N denotes the number of points and 3 stands for spatial coordinates for each point.

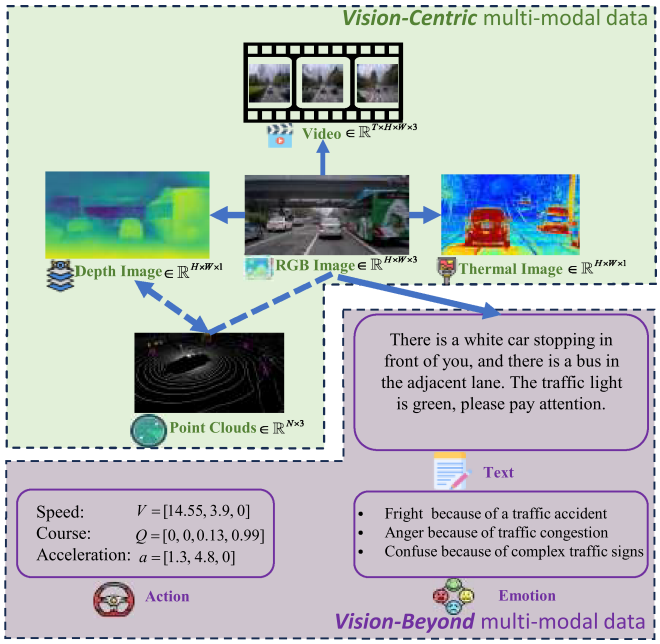

Fig. 2. Common multi-modal data used in road scenes. We divide them into two groups, i.e. **vision-centric** multi-modal data and **vision-beyond** multi-modal data. Solid arrows denote the strong connections between two modalities and dashed arrows denote weak connections. Vision-centric multi-modal data refer to those collected from perception sensors, usually containing detailed visual features, while vision-beyond multi-modal data refer to those springing up recently which contain more semantic and comprehensive information describing the holistic scene.

Recently, there has been a growing trend to harness a broader spectrum of modalities which are called vision-beyond multi-modal data. In our survey, vision-beyond multi-modal data include text, emotion, and action. Text data provide an overview of the road scene using natural language descriptions. Emotion data are collected and annotated, reflecting the driver's emotions according to the real-time road scene. Action data are numerical data collected from the control system, demonstrating the moving status of the vehicle. The incorporation of these vision-beyond multi-modal data enables systems to gain more comprehensive scene understanding.

3) Multi-Modal Learning: The goal is to train models through the joint learning of multiple representations from diverse data modalities, including image, text, and video, mimicking the human ability to interact with the environment through various senses. Two primary methods for encoding multiple modalities are as follows:

Modal-invariant encoder: These methods map all modalities into a joint embedding space by a shared encoder. They tend to propose a simple and effective architecture to learn the ability to interact across modalities. The success of unified multi-modal learning methods such as Meta-Transformer [236] has inspired methods to learn joint embedding space by utilizing a

unified multi-modal encoder. Meta-Transformer benefits from the versatility of Transformer architecture to learn modal-invariant representations.

Modal-specific encoder: These methods use a modal-specific encoder for each modality. CLIP [151] individually utilizes two encoders for vision and language, achieving strong zero-shot learning toward downstream tasks by aligning visual-semantic space into language space via contrastive pretraining. ImageBind [63] involves more modalities such as video, audio, thermal data. By aligning modal-specific embedding into the image embedding, a joint embedding space for all modalities is obtained.

4) Multi-Task Learning: Multi-task learning (MTL) refers to a paradigm where a model is trained to perform multiple related tasks simultaneously. Instead of training separate models for individual task, MTL aims to leverage shared representations to improve overall performance on all tasks.

Multi-task architectures: Parameter sharing is a key mechanism to encourage the learning of shared information among different tasks. Depending on the extent of parameter sharing within model, multi-task architectures involve two categories: hard-parameter sharing [89] and soft-parameter sharing [120]. Specifically, in hard-parameter sharing architecture, a single set of shared parameters or a shared feature extractor is used for all tasks to save computational resources. The shared parameters are responsible for extracting features that are considered task-agnostic. However, each task has its own task-specific parameters, which are responsible for generating task-specific outputs. Unlike a strictly common set of shared parameters in hard-parameter sharing, the soft-parameter sharing architecture introduces a degree of flexibility by allowing task-specific parameters to be influenced by shared parameters. This architecture aims to strike a balance between leveraging shared knowledge and allowing flexibility for task-specific outputs.

Multi-objective optimization: During joint training, different tasks often correspond to different objectives. Multi-objective optimization aims to find a set of solutions that optimize multiple objectives simultaneously. One common approach involves transforming the multi-objective problem into a single-objective problem using a weighted sum of the objectives [89], [117], which allows to express the relative importance or priority of each task during the optimization process. However, this fashion not only struggles to accurately capture the trade-offs between multiple conflicting objectives where the Pareto front is non-convex or has intricate shapes, but the performance is highly sensitive to the choice of these weights. Gradient-based multi-objective optimization is another valuable approach to address complex problems with conflicting objectives. For example, Pareto-based methods [110], [121], [158] are often employed in gradient-based multi-objective optimization. These methods aim to find solutions along the Pareto front, which represents the set of optimal solutions that cannot be improved in one objective without degrading another, by iteratively updating the model parameters using gradients from multiple objectives.

5) Pretraining Objectives: Foundation models have propelled the fundamental cognitive abilities of intelligent systems. These models are pretrained on extensive datasets primarily through self-supervised or weakly-supervised learning methods. Through the stage of pretraining, the model can acquire the intrinsic structure and patterns from data. We summarized four types of pretraining methods as follows:

Image-only contrastive pretraining: To acquire a robust encoder, image-only contrastive pretraining builds on a specific pretext task where the model learns to compact positive sample pairs and widen negative sample pairs in a unified representation space. Existing methods [27], [28], [30], [69] usually construct pairs of positive and negative samples using various hand-craft augmentation of the same image, such as color jittering, crop and resize operations.

$$L_{image-only} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{exp(x_i^+ \cdot x_{0/}\tau)}{\sum_{j=1,j\neq i}^{N} exp(x_j^- \cdot x_0/\tau)}. \quad (1)$$

Specifically, they use a similar objective function like InfoNCE, as shown in (1), to maximize agreement between positive pairs and minimize agreement between negative pairs. Given a query sample x_0, x^+ and x^- are individually positive and negative samples associated with x_0. N is the total number of samples and $exp(\cdot)$ is similarity function.

Image-Text contrastive pretraining: With the popularity of multi-modal learning, researchers have also explored its potential for aligning image-text representation [11], [22], [83], [147], [151], [161]. The image-text contrastive loss is calculated as (2):

$$L_{image-text} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{exp(x_i \cdot y_i/\tau)}{\sum_{j=1}^{N} exp(x_j \cdot y_i/\tau)}, \quad (2)$$

where x_i and y_i are the normalized embedding of the i-th paired image embedding and text-embedding. τ is the temperature term. N is the total number of samples. Besides, image-text contrastive pretraining with masking strategy [51], [58], [102], [102] have shown more impressive performance than vallina version, reducing the cost of computational resources.

Masked image modeling (MIM): Encoders gain enhanced universality and adaptability through the generation or reconstruction of the input [11], [29], [68], [210], [235], which directs the encoder to learn semantic representations by predicting a percentage of masked regions from the visible ones. The MIM loss is formulated as (3):

$$L_{MIM} = -\mathbb{E}_{t^m\sim D}[\log p(t^m|t^v)], \quad (3)$$

where the goal is to generate masked image tokens t_m given the visible image tokens t_v. According to the predicted target, MIM can be categorized into pixel-based target and feature-based target: Early methods, such as MAE [68] and SimMIM [210], predict the raw pixel values of masked image patches in the input space. While these approaches yield promising results in various downstream tasks, they usually focus on mask strategy and model structure, leading to lower semantic representations [4]. Subsequent methods have started to make predictions in representation space like CAE [29], [235], MaskFeat [194] as an efficient and scalable method for learning semantic representations.

Masked language modeling (MLM): Early masked language models like BERT [1], RoBERTa [118] are trained to predict masked word tokens x_m given a sequence of tokens x_v also

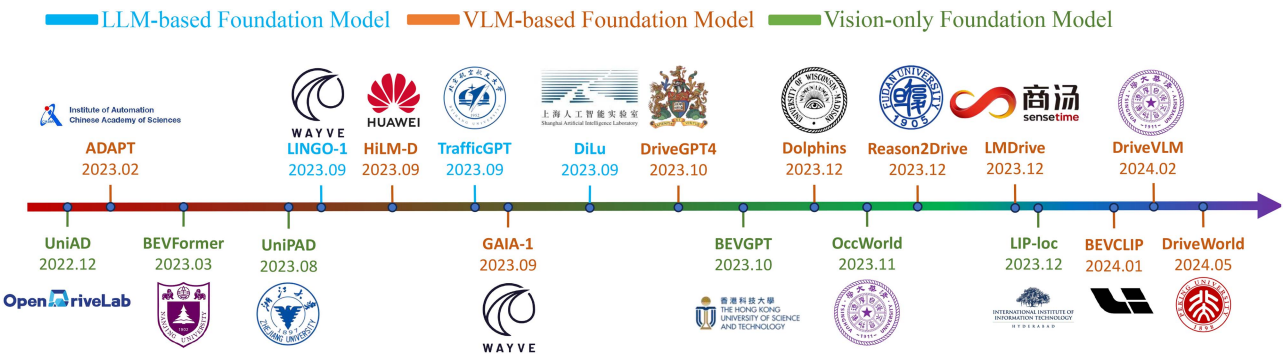

Fig. 3. Roadmap of recent foundation models in driving scenarios. We divide these foundation models into LLFM, VLFM, and LVFM based on the data modality they use. LVFMs are vision-only large-vision foundation models that only take vision-centric data as input. Pretrained on large-scale datasets, these foundation models can act as robust feature representors and facilitate downstream tasks to a great extent. In contrast, LLFMs and VLFMs usually incorporate LLMs or VLMs respectively, leveraging their robust reasoning ability to perform various complicated tasks.

known as "Cloze task", as shown in (4):

$$L_{MLM} = -\mathbb{E}_{x^m \sim D} \left[\log p\left(x^m | x^v\right)\right]. \tag{4}$$

Furthermore, large language models (LLMs) like GPT-3 [20] predict next token given the prefix language tokens in an autoregressive manner. The pretraining objective is employed in Decoder-only LLMs, as shown in (5):

$$L_{LM} = -\mathbb{E}_{x^l \sim D}[\log p(x_l | x_{1:l-1})], \tag{5}$$

where $x_{1:l-1}$ denotes the prefix sequence before l-th token and x_l is the l-th next token prediction.

6) Fine-Tuing Techniques: As foundation models increase in parameter size, they showcase emerging abilities. However, the challenges associated with directly full parameter fine-tuning for new tasks and domains can be a significant concern, including computational resources, complexity, and potential risks such as overfitting. Therefore, to overcome these challenges, we mainly conclude two types of commonly used fine-tune techniques as follows:

Prompt tuning has drawn great attention with the release of GPT-3 [20]. Compared to full parameters fine-tuning, prompt tuning is often more resource-efficient, particularly when computational resources and task-specific data are limited. For example, in-context learning [20], [141] aims to perform a new task based on few-shot examples provided within the context of a prompt, rather than performing gradient updates through explicit fine-tuning on a large labeled dataset. For hard tasks requiring logic reasoning, such as mathematics, chain-of-thought prompting [196] provides a detailed, step-by-step explanation or reasoning within the prompt itself, which guides the model to generate more accurate and logical responses.

Instruction tuning is a simple and straightforward multi-task fine-tuning technique that involves adapting a pretrained foundation model to a specific task by providing it with clear, task-related instructions. During inference, the foundation model can generalize to some unseen tasks by explicitly adding task instructions, especially when the inference tasks are similar to those it was trained on. Moreover, it can also help foundation models act as human assistant whose response are better aligned with human intents. Representative works like InstructGPT [143], Flan-T5 [38].

B. Roadmap of Visual Understanding Foundation Models for Road Scenes

As shown in Fig. 3, we've recently witnessed a large number of foundation models in road scenes. Here, we divide them into mainly three categories: **LLM-based foundation model (LLFM)**, **VLM-based foundation model (VLFM)** and **vision-only large-vision foundation model (LVFM)** based on the data modality used during pretraining phase.

Kim et al. [92] proposed the first framework for generating textual explanations of driving decisions. Despite the novelty of this work, its impact might have been constrained due to reliance on CNN rather than Transformer architecture.

With the prevalence of LLMs in recent years, there emerges numerous foundation models utilizing LLMs (LLFMs) in road scenes. The most popular way to integrate LLMs into the autonomous system is to treat a pretrained LLM as a "brain" that processes structured data from the scene and conducts reasoning. This route prevails after the release of open-source LLMs such as LLaMa [175]. TrafficGPT [233], DiLu [199], and LINGO-1 [193]serialize scene data into structured natural language narrations using task-specific models and feed them into a pretrained LLM for understanding and reasoning. These models just take serialized text data as input, so the performance of these models is highly restricted by the task-specific models, hampering their deployment in real applications.

With the advancements in VLMs, VLFMs seek to train multi-modal encoders for inputs respectively and align them in the manifold space to perform multi-modal understanding tasks in an end-to-end manner. ADAPT [87] trains a unified end-to-end vision-language Transformer to attend across visual and linguistic input, providing narration, reasoning, and control signal prediction ability. However, ADAPT can only answer rigid-form questions and are unable to perform more complicated tasks such as VQA. The breakthrough turn point comes when multi-modal LLMs like LLaVa [114] and BLIP2 [98] release. DriveGPT4 [214], Dolphins [122] and many other works [50],

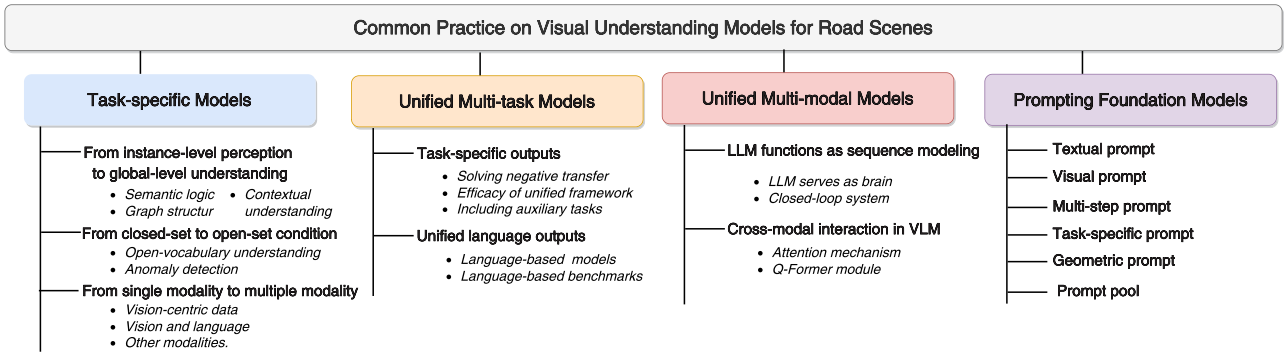

Fig. 4. The overall framework of Section III. We review existing works on visual understanding for road scenes from four perspectives: task-specific models, unified multi-task models, unified multi-modal models, and prompting foundation models, respectively. By organizing the review based on these four perspectives, the framework provides a structured and insightful exploration of the existing literature on visual understanding for road scenes.

[137], [160], [195] fine-tune a visual encoder and align the visual features into the language features via Q-former [98] to leverage the robust reasoning ability of LLM, contributing to a better human-like understanding of the environment. GAIA-1 [73] is a generative world model that synthesizes realistic road scenes from video, text, and action inputs, facilitating a robust model trained on diverse synthetic data and offering fresh opportunities for advancement in intelligent vehicles. Recently, some works [3], [160] have begun to focus on interactive capability in real-world road scenes, which motivates them to design closed-loop systems that are more realistic for deployment.

In contrast, LVFMs only leverage visual inputs such as camera data and lidar for better feature representation and scene construction, thus facilitating downstream vision-centric tasks without dependence on linguistic data. UniAD [75] proposes the first end-to-end framework that incorporates full-stack driving tasks in a unified network. For directly connecting image with lidar in cross-modal localization, LIP-loc [164] applies contrastive pretraining to 2D images and 3D points cloud. UniPAD [218] proposes a self-supervised framework to enhance feature learning using only images. To exploit the strengths of BEV (Bird's-Eye View) map, BEVFormer [103] leverages multi-camera images and spatio-temporal Transformer to construct robust BEV representations. BEVGPT [184] integrates prediction, decision-making, and motion planning taking the BEV images as the only input source. As for vision generative models, OccWorld [240] is a world model that predicts vehicle and scene dynamics in a 3D occupancy-based representation.

III. Common Practices on Visual Understanding Models for Road Scenes

In this section, we provide a comprehensive review of existing works on road scene understanding from four perspectives, as illustrated in Fig. 4. We first introduce task-specific models in Section III-A, unified multi-task models in Section III-B, unified multi-modal models in Section III-C, and prompting foundation models in Section III-D, respectively.

A. Task-specific Models

Task-specific models for road scenes are designed for specific downstream tasks. These models are trained on specific data and architecture to optimize their performance in those particular tasks, making them to be highly proficient in their designated areas. In this subsection, we conclude them from three perspectives: from instance-level perception to global-level understanding, from closed-set to open-set condition, and from single-modality to multi-modality.

From instance-level perception to global-level understanding Unlike the conventional approach of recognizing individual traffic participants[64], [93], [242], modern methods prioritize analyzing the environment through a broader, global-level understanding. For example, semantic logic between instances serves as global-level comprehension. Understanding overall information from traffic signs solely based on instance-level perception can be challenging. Yang et al. [217] introduce a novel traffic sign interpretation task that involves both localizing and recognizing traffic signs, aiming to provide a precise global semantic understanding akin to natural language for road scenes. This approach makes a great innovation by emphasizing the consideration of broader relationships among individual traffic signs and generating global accurate traffic instruction information. Similarly, contextual understanding[152], [226] excels in achieving a comprehensive global perspective. Mask2Anomaly[152] casts traditional road anomaly segmentation task as a context-aware mask classification problem by considering contextual semantics around the anomalies. Compared with pixel-based architecture which densely predicts label for each pixel, mask-based Transformer architecture is more advantageous for anomaly segmentation. This is attributed to the mask-based architecture promoting objectness, facilitating the comprehensive capture of anomalies as complete entities. This approach yields more consistent anomaly scores and diminishes false positives. Graph representation is another kind of global understanding. GP-Graph [8]reconceptualizes the multi-modal pedestrian trajectory prediction task by incorporating both inter- and intra-group relations using graph representations. Unlike existing interaction models treating each pedestrian as a graph node, GP-Graph dynamically segregates pedestrians in crowded settings into individual groups. This innovation enables effective attention to both inter-group and intra-group interactions, significantly enhancing prediction accuracy and facilitating model interpretation. Unlike instance-level perception which focuses on individual objects without exploiting relationships between

objects, global-level understanding lies in its ability to provide a holistic understanding for road scenes, by considering the interactions and dependencies across various objects.

From closed-set to open-set condition: Previous detection methods[66], [162] and segmentation methods [104], [130]for road scenes primarily aimed to achieve high performance within common class sets such as pedestrian and car, often overlooking considerations for underlying unseen objects in open-set scenarios, such as a dog on the highway.

Open-set understanding in autonomous driving refers to long-tailed corner cases detection[99], which holds significant importance in road scenes, particularly for ensuring safety and reliability. Previous method [37] has endeavored to generate corner samples to maximize the coverage with pretrained models, which is cost and impractical. However, a growing trend involves discovering unseen objects through the utilization of external knowledge sources, such as multi-modal data and pretrained vision-language models. As an example, SalienDet [49] solves open-world detection in road scenes via salience maps-based feature enhancement techniques for unknown objects. Likewise, GOOD [76] delves into extracting geometric cues of traffic objects from depth and normal maps within an open-world context, demonstrating that geometric information is a crucial component for open-set understanding. On the other hand, open-vocabulary understanding targets the recognition of arbitrary objects based on textual input by leveraging advancements in vision-language learning. OVTrack [100] extends a closed-set object tracker into an open-vocabulary tracker by leveraging knowledge distillation from a pretrained vision-language model [151]. Additionally, it also mitigates the scarcity of road data through a strategy of hallucination-based generation harnessing the power of the diffusion-based model. Anomaly detection also represents an open-set scenario. Bogdoll et al. [15] delve into the challenge of unexpected situations within road scenes. They explore the application of world models for the demanding anomaly detection task and showcase their seamless integration with current approaches. Having the ability to appropriately recognize inputs that belong to classes not encountered during the training phase, the open-set capability is a crucial aspect for ensuring the safety of intelligent vehicles in road scenes.

From single modality to multiple modalities: Unlike previous methods have the limitation in using only single modality[90], [220], [243], the utilization of multiple modalities proves beneficial in obtaining a comprehensive understanding of road scenes[59], [63], [149].

Vision-centric data. Many methods have shifted their attention to multiple visual sensor data, ranging from camera-only [86], [103], [177], [190] or lidar-only [36], [181], [213], [225], [229], to collaborate vision-centric sensors [9], [35], [168], [209]. These methods for collaborating sensors mainly involve two categories, multi-sensor fusion and multi-sensor calibration. Multi-sensor fusion methods [9], [45], [168], [179] aim to observe road scenes via utilizing complementarity across sensors in a fusion fashion. TransFusion [9] introduces an enhanced LiDAR-Camera fusion approach utilizing a soft-association mechanism for 3D Object Detection. This method is specifically designed to tackle challenges arising from suboptimal image conditions, such as intensive illumination, thereby promoting greater robustness in fusion solutions. During actual driving, multiple sensors on the ego-vehicle may experience positional displacement and misalignment issues. To address these issues, multi-sensor calibration[71], [85], [86], [88], [206] is proposed. Jiang et al.[85] not only use rotational angle noise data augmentation to simulate the poor calibration of multi-modal information but also introduce knowledge distillation strategy to prevent the model from experiencing performance degradation due to noise data augmentation when the model is well-calibrated.

Vision and language. Recent progressions have extended the scope of basic visual modalities from sensor combinations to encompass the fusion of vision and language modalities. Present methodologies[33], [34], [116], [136], [145] focus on further integrating linguistic comprehension into vision-centric tasks. VLPD[116] introduces a novel pedestrian detection approach aimed at overcoming challenges posed by confusion, small scales, and occlusion. This method self-supervises its own detector using segmentation pseudo-labels generated by pretrained vision-language model and learns discriminative pedestrian features through contrastive learning.

Other modalities. For instance, within the realm of risk assessment, driver emotions[14], [172] also serve as valuable information sources. CPSOR-GCN[172] have shifted their attention to exploring multi-modal training strategies on enhancing performance in trajectory prediction. Besides considering the physical features in road scenes, their research proposes a multi-modal prediction framework incorporated with cognitive theory[128], simultaneously considering the environment stimuli, emotion state, and behavior of the driver.

B. Unified Multi-Task Models

Understanding road environments is an essential requirement for ensuring safe autonomous driving. In the past few years, although many methods have been proposed to solve a single task with satisfactory performance, there are still many difficulties in the synergy of various tasks across granularities. Multi-task learning (MTL) aims to jointly train on the unified models through shared parameters across various tasks, enabling parameter efficiency and better performance compared with single-task learning. Recently, researchers have begun to explore integrating various visual understanding tasks in a unified framework [173], [222]. We classify these methods into two categories according to multi-task output formats, as illustrated in Fig. 5.

Task-specific outputs: These methods, usually consisting of one encoder for shared features and task-specific heads, have been adopted in a wide range of tasks such as 3D object detection [91], [179], RGB-X object detection [45].

Solving the negative transfer problem across tasks is the primary concern for MTL. In this trend, VE-Prompt [108] aims to generate task-specific features for different task via the prompting of visual exemplars. This method provides novel insights into the effect of task-specific prototype-based prompts on vehicle detection, lane detection, and drivable area segmentation tasks. It uses CLIP as a visual prompt generator to generate learnable

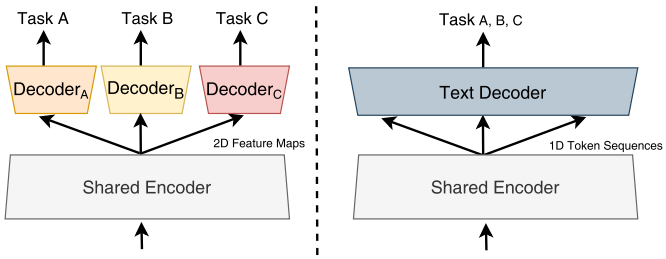

Fig. 5. Unified multi-task models can be categorized based on their outputs into two distinct types. The first type (left) includes models with task-specific outputs, characterized by a shared encoder and individual task-specific heads across all tasks. In this type, the shared encoder processes the input data to produce 2D feature maps, and each task has its dedicated head to generate task-specific output, respectively. Conversely, the second type (right) refers to models with unified language outputs. These models consist of a shared encoder and a unified text decoder to generate texts for all tasks. The shared encoder is responsible for transforming the input data into 1D token sequences, contributing to language-based representations for all tasks.

task-specific prototypes from visual exemplars, and then these task-specific prototypes are attended with features extracted by one shared backbone to acquire high-quality task-specific features for decoders via cross-attention mechanism. In pursuit of effective information propagation across diverse tasks, CML-MOTS [43] introduces a novel multi-task framework designed for video instance segmentation and tracking. This framework incorporates an innovative associative connection across different task heads, facilitating the fusion of outputs from various task heads.

The efficacy of a multi-task model stems from its ability for tasks to mutually benefit from each other. Therefore, real-time unified perception and fast inference are critical considerations. For example, YOLOP [202] extends real-time and lightweight objector YOLOv5 to multi-task model for performing traffic object detection, drivable area segmentation, and lane detection simultaneously. With the same goal of fast inference, YOLOM [183] introduces a real-time and lightweight multi-task model designed for joint object detection, drivable area segmentation, and lane line segmentation tasks. The model enhances generalization without the need for customizable design by incorporating a unified loss function for all segmentation tasks. Additionally, the adoption of a series of convolutional layers as the segmentation head significantly reduces the overall inference time. LiDAR-BEVMTN [134] proposes a LiDAR-only multi-task perception system for collaborative detection, semantics, and motion segmentation. Opting for LiDAR-only perception in Bird's-Eye View (BEV) provides an alternative for fast inference on embedded devices. This consideration arises from the signal sparsity of radar sensors, which operate at long ranges and may exhibit lower quality. Moreover, dense data from camera sensors, which are susceptible to diverse weather conditions, further underscores the appeal of this approach for efficient inference.

Including auxiliary tasks to enhance the performance of the main task is another key advantage of MTL. The concept of world models [96] has been widely discussed recently with the following characteristics of remembering history, learning experience, modeling the world, and predicting the future. For example, OccWorld [240]is a world model in predicting future scenes with the auxiliary ego-car motion trajectory prediction. Through efficiently discretely encoding [178] to obtain high-level semantic representations of past frames and a spatial-temporal autoregressive-based generative transformer to enable interaction world tokens in the same and cross timestamp, task-specific decoders can make a more reasonable prediction than ground truth labels generated by self-supervised methods.

Unified language outputs: While it is a common practice for multi-task models to produce task-specific formats as results, certain limitations persist. Vision-centric tasks often exhibit distinct output formats, varying across different axes such as granularities and the number of classes. A novel approach to address the issue is to use natural language as the unified output. For example, HiLM-D [50]is a LLM-augmented model that it performs risk object localization and intention and suggestion prediction in a natural language manner, without exquisite architecture design. ADAPT [87] jointly predicts the vehicle's action and gives the reasons based on the observation of surrounding scenes in the caption manner. Furthermore, traffic sign interpretation is a new task proposed by [217], which globally detects, and recognizes traffic signs, and gives accurate traffic instruction information in logic-based natural language description.

Newly proposed prompt-based benchmarks provide a convenient way to build unified multi-task models using natural language as unified multi-task outputs. Specifically, NuPrompt [204] expands the Nuscenes dataset [21] to instance-text pair by constructing language descriptions for video clips. Similarly, NuScenes-MQA [78] provides an evaluation of a model's capabilities in sentence generation and VQA annotated in a questions and answers (QA) format. DriveLM [165] also facilitates perception, prediction, and planning with logical reasoning. DRAMA [124] aims to simultaneously detect risk objects and give explanations.

C. Unified Multi-Modal Models

In recent years, the advancements in natural language processing through LLM and the progress in seamlessly integrating visual and linguistic comprehension via VLM, have brought a transformative paradigm shift in road scene understanding. This evolution has given rise to a pursuit of a more comprehensive and unified approach to multi-modal understanding. LLM-based models, particularly multi-modal LLM (MLLM), specialize in language-related generation tasks. Moreover, owing to the robust understanding and reasoning capabilities of LLMs, these models [160], [214] are trained to learn patterns and semantics from a broader spectrum of modalities beyond vision and language, such as control signs of the intelligent vehicles. On the other hand, VLM-based models [122] are designed to comprehend visual and textual data in pairs and bridge the gap between images or videos and the corresponding texts. They emphasize the intricate matching and interaction between vision and language. Comparison of LLM-based and VLM-based multi-modal models is illustrated in Fig. 6. In this subsection, we discuss their specific strengths on contributing to unified multi-modal scene understanding as follows:

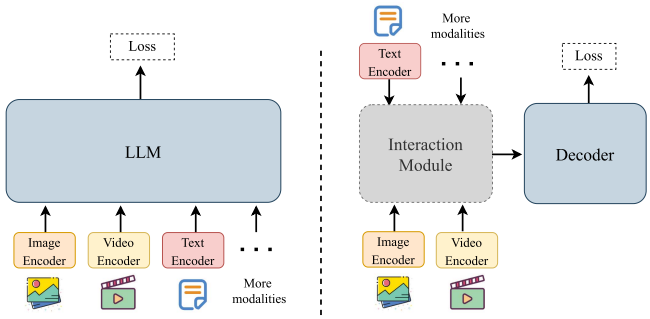

Fig. 6. Comparison of LLM-based (left) and VLM-based (right) unified multi-modal models. The LLM-based model takes LLM as a center place, which transforms multi-modal data into textual tokens that are easily modeled by LLM in the manner of sequence modeling. The VLM-based model emphasizes cross-modal interaction involving fusion, alignment, and matching across multi-modal data.

LLM functions as sequence modeling: In contrast to module design in previous driving systems, introducing LLM as the central brain for various tasks has become a prominent fashion. This kind of methods usually consists of modality-specific encoders for multi-modal input and one unified MLLM as a decoder for various tasks. They focus on transforming the driving environment into serialized textual tokens which can be easily modeled by Transformer architecture. For example, given the poor generalization of previous methods in the open-world environment, GPT-Driver [125] treats LLM as a reliable motion planner for autonomous vehicles by reformulating the motion planning task as a language modeling problem. Specifically, it converts perception and prediction results into language descriptions which are viewed as input prompts for LLM and then decoded to generate ego-movement trajectory in language-format outputs. DriveGPT4 [214] not only utilizes a video tokenizer to encode video frames and a shared text tokenizer to tokenize user query and past control signals, but employs LLM to predict desired answers and future control signals. LanguageMPC [159] harnesses the reasoning capability of LLM to analyze complex road scenes and make high-level textual decisions. These textual decisions are then converted into mathematical representations to seamlessly guide the bottom-level controller for Model Predictive Control. Similarly, some related works have been proposed in open-loop settings(such as Driving-LLM [25]), where the output or behavior of the model is not directly considered or influenced for real-time feedback.

However, these methods discussed above ignore real-time feedback from the environment in an open-loop setting. Closed-loop driving systems are essential for safe and adaptive driving because they allow vehicles to respond to changes in real time. Typically, [60], serving as a catalyst, demonstrates impressive abilities in real-time interaction with the environment, human-like driving with common sense, and memorization to overcome catastrophic forgetting, especially when faced with long-tailed corner cases. LMDrive [160] is a novel language-based closed-loop autonomous driving framework. LMDrive integrates all visual tokens of past timestamps, language tokens tokenized from multi-modal multi-view visual data and language instructions, respectively. Control signs are then predicted until the given instructions are completely conducted. DriverMLM [3] serves as MLLM-based intelligent agents. In addition to the commonly mentioned visual signals, it also employs interfaces for driving rules and user commands to build a comprehensive understanding of road scenes and align the output of MLLM with behavioral planning states predicted by the off-the-shelf behavior planning module. Wang et al. [189] introduce LLM as a behavior planer with safety assurance. The LLM-based driver agent simultaneously ensures safety and reinforces performance in driving tasks, which takes intention prediction, scenario description, behavior state, its own memory of past scenarios, and experience as input, and outputs safety-constrained behavior decisions. Moreover, rendered new observations from the environment are provided and next actions are made based on new observations iteratively. More uniquely, in contrast to directly tokenizing environment observation into language tokens, Agent-Driver [126] designs a tool library to abstract them via function call. The text-based messages are returned, leading to a precise and intensive environment description. Furthermore, cognitive memory for storing common sense and past experience serves as complementary knowledge to enhance the reasoning ability of MLLM to output safe and comfortable actions.

Cross-modal interaction in VLM: Methods [174] in the VLM-based unified multi-modal models differ from the ones discussed earlier by incorporating an interaction module for multi-modal tokens before processing them with the decoder. Within this interaction module, multi-modal tokens undergo fusion, alignment, or matching processes. Dolphins [122], based on openFlamingo [7] which features in-context learning capabilities, inserts new gated cross-attention layers within a pretrained frozen LLM with the LoRA [74] module. These newly added layers promote stable interaction between vision and language via gate-mechanism. To improve interpretability during the decision-making process, Reason2Drive [137] views the decision-making process as chain-based reasoning task. The algorithm proposes a combination of vision encoder and prior tokenizer, embedding multi-modal data into latent space, respectively. A Q-Former serves as an alignment module followed by an LLM and vision decoder responsible for predicted answers to user questions and perception results. Drive-Anywhere [185] uses multi-modal foundation models to extract patch-level features rather than vector representation for the entire image. To improve generalization to new driving conditions, latent space simulations enhanced language modality is conducted by replacing original visual features with substitute contextual concepts. Talk2BEV [47] utilizes MLLMs to acquire captions for each object in the Bird's Eye View (BEV) maps, as well as descriptions for the overall scene. By aligning this semantic information with the inherent spatial information in BEV maps, metadata is generated, i.e., the LLM-enhanced BEV map. The understanding of the current BEV map is then achieved by prompting the LLM using a Chain of Thought (CoT) approach. PromptTrack [204] incorporates a novel prompt reasoning branch based on current object tracking algorithms to deal with referring object tracking whose objective is to track desired objects according to the user's prompt. Specifically, a cross-modal interaction module is implemented to acquire prompt-aware features

and the refined features are further decoded by the tracking head.

D. Prompting Foundation Models

The emergence of the multi-modal prompts have indeed triggered a paradigm shift for transfer learning. Currently, an emerging trend is to highlight the significance of prompt engineering on transfer learning. For example, Liang, et al. [109] proposes an effective pretrain-adapt-finetune paradigm for unified multi-task learning. In the phase of adapt stage, task-specific concepts are utilized as prompts to maintain consistency between visual features and textual priors, overcoming negative transfer within multiple tasks. Xu, et al. [212] focus on time-to-event analysis to ensure safety in cyber-physical system (CPS), e.g., autonomous driving systems. To alleviate data scarcity, they adopt a novel transfer learning method, namely "pretrain-and-prompt tuning". Specifically, firstly pretraining on large-scale datasets to build fundamental knowledge, and then transfering source knowledge to the target CPS via prompt tuning. In the phase of prompt tuning, it consists of three steps: prompt template designing, answer generation and answer mapping.

Multi-modal prompts play a crucial role in guiding a model's understanding and reasoning capabilities. These prompts leverage diverse types of information or cues, such as images, text, or other modalities, to enhance the model's comprehension of the environment and improve its decision-making processes. We summarize foundation models with various multi-modal prompts as follows:

Textual prompt: In the context of driving-related models, textual promptable models like DriveMLM [3], GPTDriver [125] and DriveGPT4 [214], refer to the integration of natural language where textual instructions or commands are used to direct or interact with the autonomous vehicle (AV). These prompts serve as a form of human-vehicle communication, and can guide the AV's actions or decision-making processes.

Visual prompt: Visual promptable models [46], [47], [200] involve utilizing visual cues to enhance the scenario understanding. The representative algorithms in this line include multi-view vision-prompt [146], cross-modal prompt fusion [249], and learnable visual prompts [108] in various driving scenarios. As the road scene becomes increasingly complex, the plain textual and visual promptable models no longer meet the demands. They demands real-time responses and effective adaption to new scenarios. Novel prompts beyond simple vision and language cues are being explored to enhance autonomous systems and their capabilities.

Multi-step prompt: Existing prompt techniques mentioned above solely activate the powerful reasoning ability of FMs via a single step, where it is sub-optimal to exploit powerful world knowledge from LLM. Typically, CoT embodies the idea that language is not just a collection of individual words or sentences but rather multi-step connected thoughts or concepts. Observation of the lack of fine-grained understanding and reasoning ability in existing VLMs, Dolphin [122] extends generic image datasets to instruction-following datasets with grounded CoT response, enriched with detailed multi-step reasoning. The model is then trained on the enriched image instruction-following datasets to establish more general understanding and reasoning ability. To transfer the ability learned from generic images to the driving domain, the process also refers to finetuning on driving-related video instruction-following datasets. GPT-driver [125] argues that a sequential prompting-reasoning-finetuning strategy is effective for highly encouraging reasoning ability of FMs. This strategy involves a first stage for prompting LLM observed on perception and prediction, a second stage for multi-step CoT reasoning, and the last finetuning stage for aligning LLM's outputs with human driving trajectories. Moreover, OpenAnnotate3D [245] introduces an open-vocabulary auto-labeling pipeline to generate 2D/3D mask and 3D bounding boxes for multi-modal data, which greatly reduce the labor of extensive manual labeling. The algorithms can understand given target description with the CoT reasoning of LLM, and then provide high-quality annotations with multi-modal alignment.

Task-specific prompt: Multi-task learning involves joint training to perform multiple tasks simultaneously, leveraging shared representations to improve overall performance. However, those prompts learned from a specific task might struggle to transfer effectively to entirely different tasks in a multi-task framework. This limitation is mainly due to the specificity and task-oriented nature of the prompts. Consequently, The use of task-specific prompts helps guide the model's learning process for individual tasks, preventing negative transfer within tasks. For example, Liang et al. sequentially proposes two methods [108], [109] to effectively solve gradient conflict in multi-task learning from the perspective of generating task-specific features. They both treat CLIP model as a task-specific prompt generator, but the difference is that task-specific prompts are generated from language concepts and visual exemplar individually. [109] firstly constructs a task-specific sentence concatenated by class names and the combination of textual embeddings that are tokenized from the sentence via the text encoder of CLIP, and some learnable tokens are viewed as task-specific prompts. However, for each task, VE-Prompt [108] firstly crop class-related regions and the image encoder of CLIP is adopted to extract features, and the average of these features is used as task-specific prompts. TaskPrompter [224] is a novel multi-task prompting framework with joint 2D-3D scenarios understanding, referring to 3D vehicle detection, semantic segmentation, and monocular depth estimation. With the help of task prompt embedding, it aims to unify learning of task-specific and task-generic representation as well as cross-task interaction in a compact model capacity, rather than separately learning in specific modules. In this way, the task prompts are mapped into spatial- and channel-prompts to promote interaction patch tokens from input image in the spatial and channel dimensions, and further serve as prompting dense task-specific features and multi-task outputs.

Geometric prompt: Attention mechanisms have been incredibly useful in sequence modeling. However, they do face limitations in quadratic-scaled computational complexity and memory constraints, particularly when dealing with very long sequences. Researchers are actively exploring various methods to mitigate these limitations with the help of a prompt-based attention mechanism. For example, to address the challenge

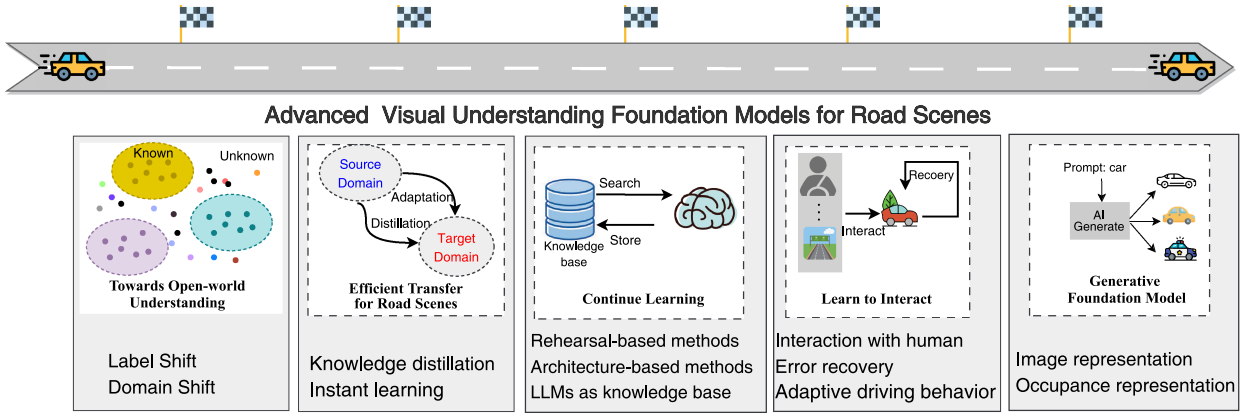

Fig. 7. The overall framework of Section IV. The advanced capabilities of visual understanding foundation models are highlighted from five perspectives: towards open-world understanding, efficient transfer for road scenes, continual learning, learn to interact, and generative foundation models, respectively.

of modeling human state change given long-range series signals, Niu et al. [138] incorporate *window prompt* to enable flexible and efficient attention on a local window. Similarly, DriveGPT4 [214] points out that a vector representation for the entire image significantly causes the loss of spatial information, which is crucial for understanding structural road scenes. Therefore, it innovatively replaces the vector-wise with patch-wise representation extracted by multi-modal FMs. This approach involves constructing attention masks via *anchor prompt*, allowing attention modules to pay attention to specific regions rather than global regions.

Prompt pool: In continual learning, designing the prompt pool that stores previous task-specific prompts is a effective way to overcome catastrophic forgetting, allowing the model to improve its performance over time. I3DOD [107] points out that current incremental 3D detectors fail to model the relationship between object localization and semantic labels and then introduce task-shared prompts maintained by a prompt pool to learn localization-category matching.

In summary, multi-modal prompts play a crucial role in enhancing the capabilities of rich information fusion, contextual understanding, and adaptability to diverse environments by providing a richer, more diverse set of cues.

IV. Advanced Visual Understanding Foundation Models for Road Scenes

In this section, we aim to highlight the advanced strengths of MM-VUFMs on diverse learning paradigms, as illustrated in Fig. 7. These strengths demonstrate the versatility of MM-VUFMs in addressing complex conditions that frequently occur in road scenes.

A. Towards Open-World Understanding

The real-world driving scenario is characterized by its dynamic and complex nature, constituting an open-world environment that challenges autonomous vehicles. Unlike controlled and static settings, the landscape of the road is constantly changing, presenting a multitude of variables and unpredictable elements. Existing foundation models have made great progress towards open-world understanding. We show GPT-4 V's [139], [140] understanding capability on open-world road scenes with both label shift and domain shift in Fig. 8, considering that the access to other models was not released.

Label shift refers to great changes in label distribution between training data and real-world scenarios encountered during deployment. Existing methods often lack the ability to fully understand long-tailed road scenes, such as animals on the road and overturned trucks. These extreme conditions underscore the necessity for further advancements in open-world generalization. In this direction, DriveLM [165] significantly improves generalization to novel objects encountered during inference by introducing a graph structure, where the decision-making process is regarded as a graph-based reasoning task. In this structure, each vertex represents the VQ pair relevant to key objects in the current scene, and logical dependencies between adjacent vertices are reformulated as edges. Unlike previous methods that regard question-answer (QA) pairs as independent individuals, this graph of QA pairs with logic dependencies enables logical reasoning and offers more promising zero-shot generalization. To explore the effect of semantic information on detecting unseen objects, Elhafsi et al. [56] adapt an LLM endowed with contextual understanding and reasoning capabilities to identify semantic anomalies. In the proposed framework, the current environment's observation is translated into a natural language description. This scene description is subsequently incorporated into a prompt template and sent to the LLM, which utilizes contextual reasoning to identify potential semantic anomalies in the described scene. Moreover, existing datasets tailored for road scenes usually fall short of adequately covering unseen cases due to the substantial label cost. This limitation results in a lack of exposure to rare situations and poor generalization to unknown objects. To address this challenge, data-driven methods are emerging. For example, TrafficSim [171] proposes a solution by generating large-scale data through simulation. This simulation is constructed and learned directly from real-world data, incorporating human intentions to ensure a more comprehensive representation of diverse road scenes. Similarly,

Fig. 8. Illustration of GPT-4 V's capability on open-world road scene understanding in both label shift and domain shift scenarios. The example images are from CODA [99]. Given the prompts, right answers in GPT-4 V's response are highlighted. These results reveal that GPT-4 V can identify scene objects and further provide advice for safe driving, underscoring remarkable understanding capability of existing foundation models.

KING [67] proposes a novel gradient-based safety-critical road scene generation in the CARLA simulator.

Domain shift occurs when there is a misalignment between the source domain where the model is trained and the target domain where the model is applied. For road scenes, the domain shift implies that the model may encounter environments, scenarios, or conditions during deployment that differ from what it experienced during training. This can lead to a decrease in performance as the model struggles to generalize effectively to unseen domains. Wen et al. [200]has undergone extensive road experiments on real-world road scenes. Observations indicate that GPT-4 V has demonstrated superior understanding and inference capabilities in dealing with unseen scenarios compared to existing driving agents. DriveAnywhere [185] innovatively alleviates the domain shift problem on unseen scenarios via simulating different scenarios in latent space. Specifically, a series of concepts relevant to road scenes are obtained from the LLM. The textual features of these concepts are computed and used to replace visual features, simulating various scenarios based on the principle of exceeding the similarity threshold. This simulation technique also acts as a data augmentation strategy in the latent space, exposing models to diverse scenarios during training. Ultimately, this process enhances the model's ability to generalize in open-world environments during deployment. Inspired by VLM with rich knowledge, Dolphins [122] is built upon a general vision-language model [7], and leverages the underlying knowledge gained from pretraining. This approach aims to enhance in-context learning to improve generalization across a spectrum of driving-related tasks via few-shot scene images.

In summary, effectively addressing label shift and domain shift is essential for ensuring the robust performance of intelligent vehicles in an open-world environment. By mitigating the challenges posed by shifts in label distribution and domain conditions, the model can enhance their adaptability, generalize effectively, and navigate the dynamic complexities of real-world road scenes with better reliability and safety.

B. Efficient Transfer for Road Scenes

Efficient transfer for road scenes involves adapting generic knowledge learned from natural scenes to road scenes. Due to a great domain gap between natural scenes and road scenes, this means that foundation models(FM) are required to not only distillate foundational knowledge gained from the pretraining stage to achieve good generalization on common objects (e.g., car, pedestrian), but also to be applicable to specific policies which are unique in road scenes. Moreover, instant learning further enhances the adaptability of foundation models. In the dynamic landscape of the road, where conditions can change rapidly, the model continuously updates its understanding. This process allows it to adapt to evolving situations, making informed decisions in the face of dynamic changes on the road. Therefore, we discuss knowledge distillation and instant learning for efficiently transferring foundational knowledge of FMs to new road scenes.

Knowledge distillation refers to the process of transferring foundational knowledge from large and general FMs [94], [142], [151] to small and specialized models. FMs have learned foundational understanding, reasoning, and decision-making capabilities from widespread data and scenes. The small and specialized model is a lightweight model tailored specifically for driving-related tasks. However, directly deploying a large model onto autonomous vehicles might be computationally expensive and impractical due to the constraints in memory, power. To address this, knowledge distillation allows for the compression and transfer of the FM's knowledge into specialized models for downstream tasks.

Distill CLIP knowledge to road scenes. With the remarkable success of multi-modal contrastive pretraining, the recognition of its pivotal role in cross-modal localization tasks has come to the center. LIP-Loc [164] is the first work to solve global visual localization utilizing image-LiDAR contrastive pretraining. It aims to learn a joint embedding space for image and LiDAR using the same contrastive symmetric loss in CLIP and achieve zero-shot pose prediction about the given image within expansive point clouds. LiDAR-based retrieval has also emerged in LidarCLIP [72]. Directly aligning the features of point clouds with text features is a non-trivial solution due to the scarcity of LiDAR-text data. Instead, image features serve as a bridge for connecting text and lidar data. For each training pair of image and lidar, lidar data are first transformed into image planes and then fed to a lidar encoder to fit the features of a frozen CLIP image encoder by maximizing the similarity. In 3D scene understanding, CLIP2Scene [26] leverages CLIP knowledge to solve point cloud semantic segmentation via semantic and spatial-temporal consistency regularization.

Distill SAM knowledge to road scenes. Intermediate representations generated by FMs on the fly provide a rich source of knowledge, contributing to the adaptability and effectiveness of various tasks. In multi-modal 3D object detection, segment anything (SAM) has raised great attention from the research community. Chen et al. propose an extension of SAM [94] to the 3D task using VoxelNeXt [31], a fully sparse 3D object detector. With prompts in the form of points or boxes, the model outputs a combination of a 2D mask and a 3D bounding box. The incorporation of promptable SAM significantly reduces the annotation cost for 3D detection. RoboFusion [169] is a robust framework to suppress the noise in the driving scenes via SAM. To adapt SAM to road scenes with substantial noise, SAM-AD is obtained via mask image modeling pretraining on collective datasets. The SAM-AD is then adopted to extract robust features for the scene image, effectively reducing the impact of noise in road scenes. In unsupervised domain adaptation, it is suboptimal to alleviate the domain gap between the source and target domain by simple alignment. [208] propose to utilize SAM's feature space as a robust supervision, and both explicitly align source and target domain with SAM in 3D scene understanding task.

Distill LLM knowledge to road scenes. LLM pretrained large-scale datasets is regarded as a rich knowledge base. For example, DriveMLM [3] is an LLM-based autonomous driving framework that inputs multi-modal data observed on the surrounding scene, and then employs a MLLM to model behavior planning. Behavior planning involves determining the optimal driving route based on the surrounding environment and giving decision states for vehicle control. Here, the teacher model is an existing AD system such as Apollo [6]. The DriveMLM's linguistic decision outputs are aligned with Apollo's decision state output by its behavioral planning module, transforming them into formats that can be easily processed by MLLMs.

Instant learning addresses the necessity for autonomous driving systems to instantly adapt to previously unseen scenarios. Inspired by the ability of humans to instantly learn new things through just a few examples, the models for road scene understanding should be able to adapt and learn from these new scenes or tasks with few shot exemplars. Moreover, when faced with a new situation, the model can instantly learn to generalize with rich instructions, ensuring the safety of intelligent vehicles in an ever-changing real-world environment. Two effective approaches to instant learning are in-context learning and instruction tuning.

In-context learning, particularly referring to few-shot prompting, involves the rapid adaptation of the perception system to a new environment based on few-shot example pairs, analogously to GPT-3 [20]. This rapid adaption involves scenario adaption and context adaption, simultaneously. For instance, instead of extensive retraining to numerous instances of this scenario, the system can learn and adapt quickly based on a few specific examples when encountering a new scenario on the road. Furthermore, the driving system leverages few-shot examples to make rapid and contextually informed decisions in real time as it faces similar situations. Typically, Dolphins [122] is the pioneering work that harnesses the in-context capability of VLMs [7] to instantly learn and adapt to a series of driving-related tasks.

Extensive experiments on DriveLM [165] have demonstrated instant learning and adaptation capability in prediction, planning, and reasoning tasks through in-context learning.

Instruction tuning has also emerged as an effective technique to enhance generalization abilities of FMs like [44], [143]. To unlock the powerful capabilities of FMs and apply them to solve driving tasks, researchers like [122] have attempted to transform road-related data in input-output tuples into instruction-following data organized in instruction-input-output triplets. These instructions may be task-specific descriptions or requirements annotated by an efficient auto-machine like ChatGPT. Combined with in-context learning, self-instruct tuning is also a simple technique to align the foundational capability of FMs with driving intention.

In conclusion, knowledge distillation facilitates the efficient transfer of foundational knowledge of FMs, while instant learning enables adaptability to new road scenes. Both are crucial components in the development of robust, adaptable, and efficient road scene understanding.

C. Continual Learning

In a real-world application scenario where sustainability is the key concern, we naturally expect the model to behave like humans, adapting to new tasks continually. This gives rise to the study of continual learning, where the pretrained model is required to continually learn a sequence of new tasks without forgetting previously learned ones. Contrary to traditional machine learning models which are predicated on the notion of encapsulating a static data distribution, continual learning is more vulnerable to *catastrophic forgetting* [95], [127], [153], [155], where the acquisition of new knowledge leads to an abrupt erosion of previously learned information.

Rehearsal-based methods are often preferred when applying continual learning methods to task-specific road scene systems. To adapt the pretrained model to various weather conditions, Liang et al. [106] introduce a novel rehearsal strategy in domain-incremental learning, employing a two-stage "Recall" and "Adapt" process. The "Recall" stage serves as a rehearsal mechanism, utilizing self-training with mixed domain data to search and reinforce previously learned domain characteristics, while the "Adapt" stage incrementally introduces new domains, utilizing patch-based adversarial learning to fine-tune the discriminability and generalizability of the model. DISC [133] proposes a model that can incrementally learn to detect things in different weather conditions. It stores the statistical parameters and simply "plug and play" the statistical vectors for the corresponding task into the model in new tasks, without the need for expensive feature banks.

Architecture-based methods: Traditional continual learning methods usually adopt an expandable architecture to accommodate new tasks automatically. DRB [167] presents a novel approach utilizing a neuro-fuzzy architecture called Deep Rule-Based (DRB) system to integrate the malleability of fuzzy logic to adapt to new data continuously. DRB evolves its architecture by comparing incoming data against existing prototypes using a distance metric, adding new rules, or updating existing ones based on predefined thresholds. Concurrently, it modifies its meta-parameters, such as rule shapes and decision thresholds, ensuring that the model continuously refines and adapts its decision-making process to the ever-changing data landscape.

LLMs as knowledge base: Benefiting from the extensive pretraining of LLMs, perception models based on LLMs generally demonstrate the "Emergent ability" [119], which refers to the phenomenon where these models exhibit capabilities not explicitly programmed or anticipated during their training. This ability, along with the transfer ability discussed in IV-B, enables the model to "learn" new prototypes more quickly and accommodate new scenarios more efficiently [200]. Given this kind of ability, in the trend of merging LLMs into autonomous systems, existing systems [40], [41], [60], [189], [199] mainly follow the "store-and-search" manner. In this manner, the model uses an explicit natural language knowledge base to cache past skills and search for useful information when it encounters a similar situation. Fu et al. [60] propose a knowledge base storing the expert advice that deviates from the original pretrained model's knowledge. This expert advice helps the model make more proper and practical decisions just like human drivers do. Next time it encounters the same situation, the model will be reminded of the taught expert knowledge by the in-context learning mechanism [20]. DiLu [199] designs a vector knowledge base taking vectorized scene descriptions as key and searches for related information by cosine similarity. It further designs a reflection module to identify whether the decision is safe, subsequently refining unsafe decisions into safe ones using the knowledge embedded in the LLM and updating these revised decisions into the knowledge base.

We anticipate seeing works in the realm of LLMs or VLMs for road scenes that align with the classical paradigm of continual learning in the future. In fact, within the domain of general-purpose LLMs or VLMs, numerous parameter-efficient fine-tuning (PEFT) methodologies can, to a certain extent, alleviate the issue of catastrophic forgetting, such as LoRA [74] and various Adaptors [231], [232]. We hope that future research will integrate some of the unique aspects of the road scene (corner case [99] and long-tail problem [81]) and propose knowledge-enhanced continualable fine-tuning methods. By tailoring these advanced fine-tuning techniques to the distinctive challenges of the road scene, we can enhance the model's ability to adapt and retain crucial information, ensuring more robust and reliable performance in road-scene understanding applications.

D. Learn to Interact

Although autonomous systems in road scenes are capable of collecting various types of data from the environment, such as images and radar points, it is desirable for them to also consider other environmental factors, including the driver's emotions or instructions. Road scenes are inherently dynamic, with continuous changes in traffic patterns, pedestrian movements, and environmental conditions, necessitating a system that can capture and respond in real-time. The more elements the system takes into account from its environment and the more frequently it interacts with these elements, the closer it approaches the driving

habits of humans and the safer it will be. In this subsection, we will discuss the interaction of autonomous driving systems with their environment, and how these interactions contribute to the development of robust and safe autonomous systems.

Interaction with human: Nobody's going to leave their safety to a black box. There are many works [87], [124] generating additional reasoning explanations for each control/action decision which help users understand the current state of the vehicle and the surrounding environment. To enable continuous conversation, LLM-based models such as DriveGPT4 [214], LMDrive [160]. build vast instruction datasets on road scenes and train the model in an instruction-tuning manner.

Error recovery: Despite the robust reasoning capability of LLMs, LLM-based autonomous systems are still vulnerable to some intricate cases such as a really rare object in the picture, poor feature quality due to the perception sensors, or intrinsic hallucination problem [77]. These challenges are particularly pronounced in road scenes, where unexpected obstacles or conditions can arise at any moment. Dolphins [122] proposes a reflection mechanism to enable the model to correct the mistakes by itself through continuous conversations with users. LMDrive [160] proposes a dataset where notices from humans are included. These occasional notices during driving, such as "Watch out for the red traffic light in front of you" are appended to the corresponding frame tokens and fed into the LLMs for reasoning. These notices, along with navigation instructions from navigation apps, are designed to simulate real-life scenarios where a passenger or a side assistance system communicates with the driver to navigate safely through dynamic and potentially hazardous road environments.

Adaptive driving behavior: In road scenes, the driving behaviors should be taken into consideration for safe driving [198]. The complex interplay between a driver's emotional state and the unpredictable nature of road scenes can significantly influence driving behavior. For example, the ego car should slow down when detecting fright or horror from the driver's expression which may be caused by some potential risks that autonomous systems fail to detect. Sonia et al. [135]introduce a methodology that employs in-vehicle sensors to identify mild cognitive impairment signals in elderly drivers in real-time. Upon detection, the system alerts the driver and offers driving assistance, thereby ensuring safe driving behavior and reducing the risk of accidents amidst the unpredictable dynamics of road scenes. Moreover, the driver's emotions will influence the trajectory prediction system. For example, an angry driver will lead the vehicle trajectory to show a more severe lateral deviation and a more violent longitudinal acceleration [234], which is difficult for traditional models to predict. CPSOR-GCN [172] proposes a trajectory model that integrates both cognitive and physical features and uses stimulus-organism-response theory to model abnormal emotions. The experiment verifies that the incorporation of emotional data helps enhance prediction accuracy.

In conclusion, by enabling sophisticated interactions with human inputs, facilitating error correction through reflective mechanisms, and incorporating adaptive behavior based on emotional and situational awareness, these models contribute to the development of more intuitive and responsive autonomous vehicles. Further research into these areas is essential to address remaining challenges and to fully leverage the capabilities of LLMs or VLMs in improving the performance and trustworthiness of road-scene autonomous systems.

E. Generative Foundation Models

Generative models leverage sophisticated algorithms to model the complex patterns and relationships within visual data. They represent the full probability distribution of all variables, modeling the joint probability distribution of both current and target variables. Consequently, a generative model can simulate or generate the distribution of any variable. They employ various architectures such as generative adversarial networks (GANs), variational autoencoders (VAEs), autoregressive models, or diffusion models to achieve this objective.

In the context of road scene understanding, generative foundation models [16], [65], [73], [84], [96], [131], [148], [170], [187], [238], [241] refers to the ability of the model to imagine and generate a sequence of upcoming scenarios and actions when given the past scenarios and actions, empowering the driving system to adapt to changing environments and making safe actions. Generative world models achieve this predictive capability through leveraging various techniques such as representation abstraction, sequential modeling, auto-regressive prediction, or more advanced architectures. These techniques can all capture temporal-spatial and long-term dependencies across video frames. Modeling the dynamic and structural world is a challenging task, especially easily overlooked details such as traffic lights, so recently many efforts have been made to build a generative world model.

Image representation: Sora [19] recently has attracted great attention due to its ability to follow human instructions to generate high-quality videos with highly spatial and temporal consistency. GAIA-1 [73] as a generative world model, seamlessly integrates video, action, and text inputs, enabling fine-grained control and comprehensive understanding of complex environments. The world model processes discrete tokens from input as sequence modeling and then autoregressively predicts the next scenario tokens. It shows its controllable driving videos by passing the text prompt, such as a minimal change(turn the red light to green). Besides, introducing road structural information as an auxiliary condition is also an effective method for understanding road scenes. DriveDreamer [187] is a real-world-driven generative world model, excelling in controllable video generation and future driving action generation. To improve sampling efficiency, the diffusion-based world model utilizes structural information (such as HDMaps, 3D Boxes) as guidance to efficiently sample in an extensive search space. These methods are learned in 2D representation space such as RGB images and videos. Although RGB images can capture rich information such as context, color, and shape information, they might struggle with spatial structures.

Occupancy representation: OccWorld [240] proposes a novel multi-task world model for joint future scenario generation and ego trajectory prediction in 3D occupancy representation. Unlike predicting a single token at each time like

TABLE I
SUMMARY OF UP-TO-DATE LANGUAGE-BASED MULTI-MODAL MULTI-TASK DATASETS FOR ROAD SCENE UNDERSTANDING

Datasets	Statistics	Scenario	Task				Modality				
			Perception	Prediction	Planning	Understanding	Image	Video	Text	Action	Point Cloud
LaMPilot [125]	4.9k scenes, 7.6 instruction length per frame	simulator	✗	✗	✓	✗	✓	✓	✓	✓	✗
DriveLM-nuScenes [167]	4,871 frames, 91.4 QAs per frame	urban	✓	✓	✓	✓	✓	✗	✓	✗	✗
DriveLM-CARLA [167]	183,373 frames, 20.5 QAs per frame	simulator	✓	✓	✓	✓	✓	✗	✓	✗	✗
DriveGPT4 [216]	28K video, 16K fixed QAs, 12K conversations	urban,rural	✓	✓	✓	✓	✗	✓	✓	✓	✗
Talk2BEV [47]	1k BEV scenarios, 20 QAs per scenarios	urban	✓	✗	✗	✓	✓	✗	✓	✗	✓
Rank2Tell [159]	116 scenarios, 31.95 caption length per scenarios	urban	✓	✗	✗	✓	✓	✓	✓	✓	✓
NuPrompt [206]	34k frames, 1.1 prompts per frame	urban	✓	✓	✗	✗	✓	✗	✓	✗	✗
NuScenes-QA [152]	30k frames, 15.3 QAs per frame	urban	✓	✗	✗	✓	✓	✓	✓	✗	✓
DriveMLM [3]	50k routes, 30 scenarios, 200 trigger points per scenarios	simulator	✓	✓	✓	✓	✓	✗	✓	✓	✓
LMDrive [162]	64K parsed clips, 464K notice instructions	simulator	✓	✓	✓	✓	✓	✓	✓	✓	✓
Reason2Drive [139]	600K video-text pairs	urban,rural	✓	✓	✗	✓	✓	✗	✓	✗	✗
Driving-With-Llm [25]	10k driving scenarios, 16 QAs per scenarios	urban	✓	✓	✗	✓	✗	✗	✓	✗	✗
Refer-KITTI [205]	6.6k frames, 10.7 instances per prompt	urban,rural	✗	✗	✓	✓	✗	✓	✓	✓	✗
NuScenes-MQA [79]	34k scenarios, 41.2 QAs per scenarios	urban	✓	✗	✗	✗	✓	✗	✓	✗	✗
LiDAR-text [222]	420K 3D captioning data, 280K 3D grounding data	simulator	✓	✗	✗	✓	✗	✗	✓	✗	✓
CARLA-NAV [83]	83k frames, 7 command length per frames	simulator	✗	✗	✓	✗	✓	✗	✓	✗	✗
RSUD20K [253]	20K high-resolution, 130K bounding box annotations	urban	✓	✗	✗	✗	✓	✗	✓	✗	✗
DRAMA [126]	18k scenarios, 5.6 text strings per scenarios	urban	✗	✗	✓	✓	✗	✓	✓	✓	✗

The summarized perspectives of these datasets involve considerations such as data statistics, applicable scenarios, tasks, and annotated modalities.

GPT-3, they propose a novel spatial-temporal transformer as the world model, which can simultaneously generate multiple tokens each time. A U-shape network is also adopted to generate multi-scale scenario tokens and aggregate them for next token prediction. Structural understanding and easy accessibility of occupancy representation allow this approach to accommodate large-scale training.

It is worth considering how to learn a world model keeping a connection with existing FMs, and finally promote realistic applications in real-world scenarios. Existing generative world models that predict the coming states (including scenarios and actions) conditioned the past states have significant limitations of real-time interaction with the environment, leading to a relatively poor performance and inexplicable generation. Besides, simple action space which merely involves limited levels of steers and speed, is also far from meeting practical needs. Instead, constructing a close-loop generative world model is a promising trend and more in line with reality, where the model keeps real-time interaction with the environment. Furthermore, defining various action spaces[182] and designing a novel search strategy [57], [182] are both useful ways to deal with real-world challenges.

V. DATASETS AND EVALUATION METRICS

A. Datasets

The exploration of language-guided visual understanding for road scenes indeed stands at the forefront of advancements in intelligent vehicles. Pioneering studies in this domain go beyond single-task applications, aiming to create more versatile and adaptable systems through the integration of language guidance across multiple tasks and modalities. As presented in Table I, there is a summary of up-to-date language-based multi-modal multi-task datasets specifically designed for road scene understanding. It's noteworthy that these datasets depart from traditional ones, like CODA [99], KITTI [62], and Cityscapes [39] which typically annotate road scenes solely with geometric labels, such as instance-level bounding boxes and pixel-level masks. The absence of semantic and contextual information in traditional datasets poses a great challenge to achieving intelligent driving systems.

The shift towards language-guided multi-modal multi-task datasets implies a richer and more context-aware annotation approach. Instead of relying solely on geometric annotations, these datasets incorporate language-based descriptions or instructions that guide the understanding and reasoning of the scene image, encompassing multiple tasks and modalities to forge more versatile and adaptable models. Annotations in these datasets are provided in the form of question-answer pairs (QAs), either collected from new scenes like DRAMA [124] or extended on existing datasets like DriveLM [165]. Moreover, there is a trend that QA pairs for different tasks are organized from Chain-of-Thought like Reason2Drive [137], Tree-of-Thought like NuScenes-QA [150] and Graph-of-Thought like DriveLM [165].

B. Evaluation Metrics

Existing MM-VUFMs are applicable to a wide range of tasks, so it is necessary to summarize some representative evaluation metrics. Here, we briefly introduce commonly used evaluation metrics designed to measure the quality and diversity of the data generated by MM-VUFMs from the perspective of modality generation.

For text generation tasks, such as visual question answering (VQA) and image captioning, metrics like BLEU and METEOR are widely adopted, serving to evaluate the quality of the generated texts by comparing them against reference texts. For tasks involving image generation, the Fréchet Inception Distance

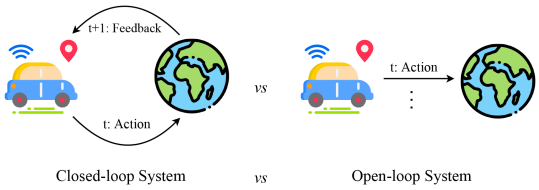

Fig. 9. Closed-loop system vs Open-loop system. In the closed-loop system, it maintains real-time interaction with the surrounding environment. This involves executing actions to modify the environment at the current time t and receiving feedback from the environment at the next time $t+1$. In the open-loop system, actions are undertaken at each time, but no feedback is received from the environment.

(FID) is a metric used to assess the quality and diversity of generated images. It quantifies the quality of generated images by examining the distance between the feature distributions of generated images and real images. Similarly, the Frechet Video Distance (FVD) is employed to evaluate video generation tasks, analogous to FID for images. FVD evaluates the quality and diversity of generated videos by measuring the discrepancy in feature distributions between the synthetic videos and real videos. For action generation tasks, such as simulating driving actions and steering angle, typical metrics are adopted like L1 and L2 errors.

Besides, for specific vision tasks in the context of perception, prediction, we use specialized metrics such as mean average precision (mAP) for object detection, mean intersection over union (mIoU) for semantic segmentation, average end point error (ADE) for trajectory prediction, and etc.

These metrics can be used individually or jointly to comprehensively evaluate diverse MM-VUFMs.

VI. Open Challenges & Future Trends

Within the promising landscape of road scene understanding, numerous challenges persist. Identifying and addressing these challenges are crucial to ensure efficient and reliable autonomous driving systems. In this section, we delve into open challenges shared by existing approaches and also highlight future trends that reveal insights toward solving these challenges.

A. Closed-Loop vs Open-Loop Driving System

The concepts of closed-loop system and open-loop system play a crucial role in control systems. An open-loop autonomous driving system operates without real-time feedback from the environment. They execute predefined programs or instructions without paying attention to external changes. In contrast, their behavior or reaction in the closed-loop system depends on the real-time feedback observed from sensors and the environment. Compared to the open-loop system, the closed-loop system is generally considered more reliable and safer for real-world road scenes, as illustrated in Fig. 9.

Among visual understanding foundation models for road scenes, there is a promising trend from single-modal to multi-modal input. Coupled with the language modeling ability of LLM, existing approaches mostly use LLM as a unified multi-modal interface by tokenizing multi-modal input into language tokens and making next token prediction for understanding and reasoning. However, these approaches operate in an open-loop setting. More specifically, these approaches are usually trained and evaluated on assumed benchmarks such as commonly used nuScenes [21] with already predefined properties. This pipeline brings two primary challenges: 1) restricted exposure to real-world data and 2) failure to incorporate feedback signals from a dynamic environment. Recently, researchers have started to explore closed-loop driving systems. LMDrive [160] is the first closed-loop driving framework that takes language instruction and multi-sensor data as input and output control actions, based on real-time observation from real-world road scenes, but there is still limited performance to achieve fully closed-loop driving systems.

To bridge the gap between unrealistic open-loop and the reliable closed-loop driving agent, driving simulator is a promising research trend. It serves as a customized and flexible platform to expose proposed approaches to 3D driving scenes and mimic closed-loop evaluation with real-time feedback. CARLA [54], known as a commonly used open-source simulator for autonomous driving, has become a cornerstone in the development and evaluation of autonomous driving systems. But existing simulators [54], [221] still have problems of lacking fidelity and limited real-world reappearance. These simulated traffic participants like pedestrians have significant differences in appearance compared with real traffic participants, potentially impacting the performance of algorithms trained solely on simulated data. Besides, real sensors face various environmental factors (like various weather conditions and occlusions) that might not be fully captured in simulations and the simulators struggle to reproduce large-scale environments (like complex intersections) with the same level of detail and accuracy. Therefore, developing a simulator that can simulate high-quality driving scenes is a promising direction.

B. Interpretability

Interpretability in road scene understanding is indispensable, as it ensures safety and accountability, and enables effective debugging along with continuous improvement. Interpretable foundation models show not only what they respond, but also the reasons behind these responses.

In this direction, existing algorithms [50], [87], [124] have investigated the use of FMs (LLMs, VLMs) to explain their output in natural language. However, we contend that directly incorporating FMs, which are used as the text generation head, into the algorithm framework cannot ensure interpretability. Although excelling at language tasks, they operate as black boxes due to their internal reasoning process remains unclear. Therefore, Chain-of-Thought (CoT) has been proposed to visualize the multi-step reasoning of LLMs to simulate the human thinking process. By depicting how LLMs process inputs and visualize outputs, this technique tries to unmask internal thought for enhancing interpretability.

It is also greatly crucial to make reasonable actions for end-to-end autonomous driving systems. Although they often employ

simpler architecture than modular design and adopt global optimization objectives, it is difficult to explain. Some post-hoc techniques can be adapted to enhance the interpretability of driving models such as saliency maps [50], [79], decoding intermediate outputs into results. The output of interpretable tasks such as object detection [35], [149], depth estimation [35], [80] is another effective way to enhance interpretability. Recently, there is also an emerging solution that transforms driving tasks into visual question answering to explain the decision-making process [5], [165].

C. Low-Resource Condition

Low-resource condition usually exists in road scenes where there is a great challenge of limited high-quality data, insufficient computing resources and memory capacity on an embedded intelligent vehicle. Early methods have attempted to deal with this challenge via model quantization [13], [211], memory-efficiency cache technique [12], robust post-processing step [156] and edge-cloud cooperation [105], respectively. They significantly reduce latency time and save memory capacity under the constraint of low-resource condition, ensuring safe and efficient driving.

In the new era of foundation models, although their generalization capability allows easy adaptation to various domains and tasks, they struggle to generalize well under some low-resource conditions due to limited data, fine-grained differences, and highly specialized domain, as reported in [237]. These characteristics also exist in road scenes where large-scale high-quality data are difficult to collect, and there is a very subtle appearance difference between traffic objects. Obviously, simply reducing model parameters or improving model efficiency like above methods [12], [13], [105], [156], [211] fail to work well. Many researchers have investigated many learning methods to overcome these new challenges, e.g., zero-shot learning [216], [227], few-shot learning [115], [129]. They aim to achieve good performance with only few training exemplars or even without any training exemplars. To differentiate fine-grained details across objects with different semantics and similar appearances, support examples or templates in few-shot learning [188] seamlessly benefit distinguishing them. FOMO [248] provides a novel perspective to use natural language predicted by LLM to describe fine-grained differences to address the challenge of open-world detection in real-world scenarios. Generative methods are also effective ways to alleviate data scarcity via generating diverse and unlimited data, which make controllable and human-friendly condition generation a reality. These methods are mainly based on diffusion models [111], [216], road scene simulation [32], [197], [205], [230], generative gaussian splatting [207], respectively. Moreover, some parameter-efficiency techniques [97], [163], [192] are proposed to adapt foundation model to downstream scenarios with only finetuning a few trainable parameters. Although great progress have been made, performance gain is very limited, still existing a long way to go before foundation models achieve good generalization under the low-resource condition.

D. Embodied Driving Agent

Although visual understanding FMs can easily incorporate sensor data, model the realistic world, and complete driving-related tasks according to human instruction, these models are pretrained on given the input data collected and produced by humans without actively solving unknown tasks and polishing their own behaviors. In contrast, we argue that transforming visual understanding FMs into embodied driving agents is a promising trend to achieve DriveAGI by endowing them with embodied reasoning capacity.

Recent advancements in LLM and multi-modal learning have brought the concept of embodied intelligence into practical application, such as robotics. PaLM-E [55], RT series [17], [18], [144] have begun to explore generic robotic agents with embodied reasoning capabilities such as mobile manipulation and task/motion planning. However, having similar embodied reasoning capabilities in driving systems encountered obstacles. Unlike splitting a large task into smaller ones and then progressively tackling them as they typically do in robotic tasks, driving outdoors is a highly coupled and complex task that cannot be step-by-step solved by simply task decomposition strategy. Moreover, the uncertainty associated with driving behavior poses a significant challenge. In this direction, a feasible solution is to be in line with end-to-end driving framework based on reinforcement learning. To imitate human driving behavior, the system collects multi-sensor data, maps them into low-level actions with the global optimization objective, and calibrates or stimulates their next actions with the feedback path. Surprisingly, ELM [246] has recently taken a solid step towards achieving embodied driving agent.

E. World Model

The world model [132], [191] refers to the model that makes reasonable predictions given past scenarios and actions, which can not only generate controlled, diverse, and scalable samples for training robust driving systems, but also promote them to perfectly generalize real-world situations.

Modeling complex driving scenarios is a great challenge due to small yet important details such as traffic lights, which are easily overlooked. Existing methods such as GAIA-1 [73] and DriveDreamer [187] harness the remarkable modeling capability of the diffusion model but still face the challenge of sample inefficiency, leading to slow convergence and a great requirement for computing resources. Considering that they only model sequences with lengths of thousands, LWM [112] proposes the RingAttention technique to extend the sequence length to millions, enabling broader capability in understanding the world. In addition, we contend that the simple action space (such as speed and steering) cannot be sufficient to meet the real-world driving requirement since human drivers can make fine-grained controls when facing unpredictable tasks. In embodied reinforcement learning, Voyager [182] is designed to progressively solve open-ended tasks in Minecraft Game using a high-dimensional action space in code rather than low-level motion commands. These action programs are not generated at one time but require multiple iterations to mend until being

verified for successfully completing the task. Similar inspiration could be implemented in the driving world model.

VII. Conclusion

In this survey, we provide an overview of multi-modal multi-task visual understanding foundation models (MM-VUFMs) for road scenes. We systematically review the extensive literature focusing on task-specific models, unified multi-modal models, unified multi-task models and foundation model prompting techniques. Besides, we highlight their advanced strengths in diverse learning paradigms, involving open-world understanding, efficient transfer for road scenes, continual learning, interactive and generative capability. Finally, we also point out key challenges and future trends to push the boundaries of foundation models for road scene understanding. In the new era of foundation models, we hope that this survey can provide researchers with comprehensive awareness and inspire future research in this domain.

References

[1] J. Devlin, M.-W. ChangK. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[2] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 23716–23736.

[3] W. Wang et al., "DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving," 2023, *arXiv:2312.09245*.

[4] M. Assran et al., "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15619–15629.

[5] S. Atakishiyev, M. Salameh, H. Babiker, and R. Goebel, "Explaining autonomous driving actions with visual question answering,'' in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst.*, 2023, pp. 1207–1214.

[6] "Apollo auto. Baidu," 2019. [Online]. Available: https://github.com/ApolloAuto/apollo

[7] A. Awadalla et al., "Openflamingo: An open-source framework for training large autoregressive vision-language models," 2023, *arXiv:2308.01390*.

[8] I. Bae, J.-H. Park, and H.-G. Jeon, "Learning pedestrian group representations for multi-modal trajectory prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 270–289.

[9] X. Bai et al., "Transfusion: Robust lidar-camera fusion for 3D object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1090–1099.

[10] Y. Bai et al., "Sequential modeling enables scalable learning for large vision models," 2023, *arXiv:2312.00785*.

[11] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Representations*, 2022.

[12] M. G. Bechtel, E. Mcellhiney, M. Kim, and H. Yun, "DeepPicar: A low-cost deep neural network-based autonomous car," in *Proc. IEEE 24th Int. Conf. Embedded Real-Time Comput. Syst. Appl.*, 2018, pp. 11–21.

[13] I. Ben-Yair, G. B. Shalom, M. Eliasof, and E. Treister, "Quantized convolutional neural networks through the lens of partial differential equations," *Res. Math. Sci.*, vol. 9, no. 4, 2022, Art. no. 58.

[14] X. Zou, D. B. Logan, and H. L. Vu, "Modeling public acceptance of private autonomous vehicles: Value of time and motion sickness viewpoints," *Transp. Res. Part C: Emerg. Technol.*, vol. 137, 2022, Art. no. 103548.

[15] D. Bogdoll, L. Bosch, T. Joseph, H. Gremmelmaier, Y. Yang, and J. M. Z¨ollner, "Exploring the potential of world models for anomaly detection in autonomous driving," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2023, pp. 488–495.

[16] D. Bogdoll, Y. Yang, and J. M. Zöllner, "MUVO: A multimodal generative world model for autonomous driving with geometric representations," 2023, *arXiv:2311.11762*.

[17] B. Zitkovich et al., "RT-2: Vision-language-action models transfer web knowledge to robotic control,'' in *Proc. 7th Conf. Robot Learn.*, 2023, vol. 229, pp. 2165–2183.

[18] A. Brohan et al., "RT-1: Robotics transformer for real-world control at scale,'' in *Proc. Robot.: Sci. Syst.*, 2023.

[19] T. Brooks et al., "Video generation models as world simulators," 2024.

[20] T. Brown et al., "Language models are few-shot learners," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.

[21] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631.

[22] L. Cao et al., "Less is more: Removing text-regions improves clip training efficiency and robustness," 2023, *arXiv:2305.05095*.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[24] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," 2023, *arXiv:2306.16927*.

[25] L. Chen et al., "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving," 2023, *arXiv:2310.01957*.

[26] R. Chen et al., "Clip2scene: Towards label-efficient 3D scene understanding by clip," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7020–7030.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A. simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[28] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 22243–22255.

[29] X. Chen et al., "Context autoencoder for self-supervised representation learning," *Int. J. Comput. Vis.*, vol. 132, pp. 208–223, 2024.

[30] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[31] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia, "VoxelNext: Fully sparse voxelnet for 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21674–21683.

[32] Y. Chen et al., "S-Nerf: Autonomous driving simulation via neural reconstruction and generation," 2024, *arXiv:2402.02112*.

[33] S. Cheng, G.-P. Ji, P. Qin, D.-P. Fan, B. Zhou, and P. Xu, "Large model based referring camouflaged object detection," 2023, *arXiv:2311.17122*.

[34] W. Cheng, J. Yin, W. Li, R. Yang, and J. Shen, "Language-guided 3D object detection in point cloud for autonomous driving," 2023, *arXiv:2305.15765*.

[35] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "TransFuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12878–12895, Nov. 2023.

[36] D. Choi, W. Cho, K. Kim, and J. Choo, "idet3d: Towards efficient interactive object detection for lidar point clouds," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 1335–1343.

[37] G. Chou, Y. E. Sahin, L. Yang, K. J. Rutledge, P. Nilsson, and N. Ozay, "Using control synthesis to generate corner cases: A case study on autonomous driving," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 37, no. 11, pp. 2906–2917, Nov. 2018.

[38] H. W. Chung et al., "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–53, 2024.

[39] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[40] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops*, 2024, pp. 902–909.

[41] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles," *IEEE Intell. Transp. Syst. Mag.*, vol. 16, no. 4, pp. 81–94, Jul.-Aug. 2024.

[42] C. Cui et al., "A survey on multimodal large language models for autonomous driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 958–979.

[43] Y. Cui, C. Han, and D. Liu, "Collaborative multi-task learning for multi-object tracking and segmentation," *ACM J. Auton. Transp. Syst.*, vol. 1, no. 2, pp. 1–23, Apr. 2024.

[44] W. Dai et al., "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36.

[45] S. A. Deevi, C. Lee, L. Gan, S. Nagesh, G. Pandey, and S.-J. Chung, "RGB-X object detection via scene-specific fusion modules," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 7366–7375.
[46] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M.-F. Moens, "Talk2Car: Taking control of your self- driving car," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf.Natural Lang. Process.*, 2019, pp. 2088–2098.
[47] V. Dewangan et al., "Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving," 2023, *arXiv:2310.02251*.
[48] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 74–92.
[49] N. Ding, C. Zhang, and A. Eskandarian, "SalienDet: A saliency-based feature enhancement algorithm for object detection for autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 2624–2635, Jan. 2024.
[50] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "HiLM-D: Towards high-resolution understanding in multimodal large language models for autonomous driving," 2023, *arXiv:2309.05186*.
[51] X. Dong et al., "MaskCLIP: Masked self-distillation advances contrastive language-image pretraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10995–11005.
[52] X. Dong and M. L. Cappuccio, "Applications of computer vision in autonomous vehicles: Methods, challenges and future directions," 2023, *arXiv:2311.09093*.
[53] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[54] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, 2017, pp. 1–16.
[55] D. Driess et al., "PaLM-E: An embodied multimodal language model," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 8469–8488.
[56] A. Elhafsi, R. Sinha, C. Agia, E. Schmerling, I. A. D. Nesnas, and M. Pavone, "Semantic anomaly detection with large language models," *Auton. Robots*, vol. 47, no. 8, pp. 1035–1055, 2023.
[57] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," in *Proc. Int. Conf. Learn. Representations*, 2019.
[58] Y. Fang et al., "EVA: Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19358–19369.
[59] D. Feng et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.
[60] D. Fu et al., "Drive like a human: Rethinking autonomous driving with large language models," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 910–919.
[61] H. Gao, Y. Li, K. Long, M. Yang, and Y. Shen, "A. survey for foundation models in autonomous driving," 2024, *arXiv:2402.01105*.
[62] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
[63] R. Girdhar et al., "ImageBind: One embedding space to bind them all," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15180–15190.
[64] D. Guo, D.-P. Fan, T. Lu, C. Sakaridis, and L. Van Gool, "Vanishing-point-guided video semantic segmentation of driving scenes," 2024, *arXiv:2401.15261*.
[65] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
[66] L. V. Haar, T. Elvira, L. Newcomb, and O. Ochoa, "Measuring the impact of scene level objects on object detection: Towards quantitative explanations of detection decisions," 2024, *arXiv:2401.10790*.
[67] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger, "King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 335–352.
[68] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
[69] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[71] Q. Herau et al., "SOAC: Spatio-temporal overlap-aware multi-sensor calibration using neural radiance fields," 2023, *arXiv:2311.15803*.
[72] G. Hess, A. Tonderski, C. Petersson, K. Åström, and L. Svensson, "LidarCLIP or: How I learned to talk to point clouds," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 7438–7447.
[73] A. Hu et al., "GAIA-1: A generative world model for autonomous driving," 2023, *arXiv:2309.17080*.
[74] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022.
[75] Y. Hu et al., "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17853–17862.
[76] H. Huang, A. Geiger, and D. Zhang, "GOOD: Exploring geometric cues for detecting objects in an open world," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.
[77] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," 2023, *arXiv:2311.05232*.
[78] Y. Inoue, Y. Yada, K. Tanahashi, and Y. Yamaguchi, "NuScenes-MQA: Integrated evaluation of captions and QA for autonomous driving datasets using markup annotations," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 930–938.
[79] K. Ishihara, A. Kanervisto, J. Miura, and V. Hautamaki, "Multi-task learning with attention for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2902–2911.
[80] B. Jaeger, K. Chitta, and A. Geiger, "Hidden biases of end-to-end driving models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8240–8249.
[81] A. Jain, L. Del Pero, H. Grimmett, and P. Ondruska, "Autonomy 2.0: Why is self-driving always 5 years away?," 2021, *arXiv:2107.08142*.
[82] K. Jain, V. Chhangani, A. Tiwari, K. M. Krishna, and V. Gandhi, "Ground then navigate: Language-guided navigation in dynamic scenes," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 4113–4120.
[83] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.
[84] F. Jia et al., "Adriver-I: A general world model for autonomous driving," 2023, *arXiv:2311.13549*.
[85] F. Jiang et al., "Revisiting multi-modal 3D semantic segmentation in real-world autonomous driving," 2023, *arXiv:2310.08826*.
[86] Y. Jiao et al., "Instance-aware multi-camera 3d object detection with structural priors mining and self-boosting learning," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 2598–2606.
[87] B. Jin et al., "Adapt: Action-aware driving caption transformer," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 7554–7561.
[88] T. Kanai, I. Vasiljevic, V. Guizilini, A. Gaidon, and R. Ambrus, "Robust self-supervised extrinsic self-calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 1932–1939.
[89] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7482–7491.
[90] A. H. Khan, S. T. Raza Rizvi, and A. Dengel, "Real-time traffic object detection for autonomous driving," 2024, *arXiv:2402.00128*.
[91] A. Khoche, L. P. Sánchez, N. Batool, S. S. Mansouri, and P. Jensfelt, "Fully sparse long range 3D object detection using range experts and multimodal virtual points," 2023, *arXiv:2310.04800*.
[92] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 563–578.
[93] G. Kinoshita and K. Nishino, "Camera height doesn't change: Unsupervised monocular scale-aware road-scene depth estimation," 2023, *arXiv:2312.04530*.
[94] A. Kirillov et al., "Segment anything,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.
[95] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.
[96] Y. LeCun, "A path towards autonomous machine intelligence version 0.9.2, 2022-06-27," *Open Rev.*, vol. 62, no. 1, 2022.
[97] H. Lee, M. Jeong, S.-Y. Yun, and K.-E. Kim, "Bayesian multi-task transfer learning for soft prompt tuning," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP 2023*, 2023, pp. 4942–4958.

[98] 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. 19730–1974.

[99] K. Li et al., "Coda: A real-world road corner case dataset for object detection in autonomous driving," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 406–423.

[100] S. Li, T. Fischer, L. Ke, H. Ding, M. Danelljan, and F. Yu, "OVTrack: Open-vocabulary multiple object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5567–5577.

[101] X. Li et al., "Towards knowledge-driven autonomous driving," 2023, *arXiv:2312.04316*.

[102] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23390–23400.

[103] Z. Li et al., "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–18.

[104] Z. Li, H. Lin, Z. Wang, H. Li, M. Yu, and J. Wang, "A. ground segmentation method based on point cloud map for unstructured roads," 2023, *arXiv:2309.08164*.

[105] S. Liang et al., "Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 25345–25360, Dec. 2022.

[106] W. Liang, L. Gan, P. Wang, and W. Meng, "Brain-inspired domain-incremental adaptive detection for autonomous driving," *Front. Neurorobot.*, vol. 16, 2022, Art. no. 916808.

[107] W. Liang, G. Sun, C. Liu, J. Dong, and K. Wang, "I3DOD: Towards incremental 3D object detection via prompting," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 5738–5743.

[108] X. Liang, M. Niu, J. Han, H. Xu, C. Xu, and X. Liang, "Visual exemplar driven task-prompting for unified perception in autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9611–9621.

[109] X. Liang, Y. Wu, J. Han, H. Xu, C. Xu, and X. Liang, "Effective adaptation in multi-task co-training for unified autonomous driving," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 19645–19658.

[110] X. Lin, H.-L. Zhen, Z. Li, Q.-F. Zhang, and S. Kwong, "Pareto multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 12037–12047.

[111] H. Ling, S. W. Kim, A. Torralba, S. Fidler, and K. Kreis, "Align your Gaussians: Text-to-4D with dynamic 3D Gaussians and composed diffusion models," 2023, *arXiv:2312.13763*.

[112] H. Liu, W. Yan, M. Zaharia, and P. Abbeel, "World model on million-length video and language with ringattention," 2024, *arXiv:2402.08268*.

[113] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023, *arXiv:2310.03744*.

[114] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, 2023

[115] J. Liu, X. Dong, S. Zhao, and J. Shen, "Generalized few-shot 3D object detection of lidar point cloud for autonomous driving," 2023, *arXiv:2302.03914*.

[116] M. Liu, J. Jiang, C. Zhu, and Xu-Cheng Yin, "VLPD: Context-aware pedestrian detection via vision-language semantic self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6662–6671.

[117] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1871–1880.

[118] Y. Liu et al., "Roberta: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[119] S. Lu, I. Bigoulaeva, R. Sachdeva, H. T. Madabushi, and I. Gurevych, "Are emergent abilities in large language models just in-context learning?" 2023, *arXiv:2309.01809*.

[120] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1930–1939.

[121] P. Ma, T. Du, and W. Matusik, "Efficient continuous Pareto exploration in multi-task learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6522–6531.

[122] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, "Dolphins: Multimodal language model for driving," 2023, *arXiv:2312.00438*.

[123] Y. Ma et al., "Lampilot: An open benchmark dataset for autonomous driving with language model programs," 2023, *arXiv:2312.04372*.

[124] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "DRAMA: Joint risk localization and captioning in driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 1043–1052.

[125] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "GPT-driver: Learning to drive with GPT," 2023, *arXiv:2310.01415*.

[126] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A. language agent for autonomous driving," 2023, *arXiv:2311.10813*.

[127] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychol. Learn. Motivation*, vol. 24, pp. 109–165, 1989.

[128] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. Cambridge, MA, USA: MIT Press, 1974.

[129] J. Mei, J. Zhou, and Y. Hu, "Few-shot 3d LiDAR semantic segmentation for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 9324–9330.

[130] E. Milli, Ö. Erkent, and A. E. Ylmaz, "Multi-modal multi-task (3MT) road segmentation," *IEEE Robot. Automat. Lett.*, vol. 8, no. 9, pp. 5408–5415, Sep. 2023.

[131] C. Min, D. Zhao, L. Xiao, Y. Nie, and B. Dai, "Uniworld: Autonomous driving pre-training via world models," 2023, *arXiv:2308.07234*.

[132] C. Min et al., "Driveworld: 4D pre-trained scene understanding via world models for autonomous driving," 2024, *arXiv:2405.04390*.

[133] M J. Mirza, M. Masana, H. Possegger, and H. Bischof, "An efficient domain-incremental learning approach to drive in all weather conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3001–3011.

[134] S. Mohapatra, S. Yogamani, V. R. Kumar, S. Milz, H. Gotzig, and P. Mäder, "LIDAR-BEVMTN: Real-time lidar bird's-eye view multi-task perception network for autonomous driving," 2023, *arXiv:2307.08850*.

[135] S. Moshfeghi et al., "In-vehicle sensing and data analysis for older drivers with mild cognitive impairment," in *Proc. IEEE 20th Int. Conf. Smart Communities: Improving Qual. Life Using AI, Robot. IoT*, 2023, pp. 140–145.

[136] M. Najibi et al., "Unsupervised 3D perception with 2D vision-language distillation for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8602–8612.

[137] M. Nie et al., "Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving," 2023, *arXiv:2312.03661*.

[138] M. Niu, Z.K. Zheng, K. Akash, and T. Misu, "Beyond empirical windowing: An attention-based approach for trust prediction in autonomous vehicles," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 5615–5619.

[139] OpenAI, "GPT-4v(ision) system card," 2023. [Online]. Available: https://openai.com/research/gpt-4v-system-card

[140] OpenAI, "GPT-4v(ision) technical work and authors," 2023. [Online]. Available: https://openai.com/contributions/gpt-4v

[141] OpenAI, "GPT-4 technical report," 2023.

[142] TB OpenAI, "ChatGPT: Optimizing language models for dialogue," OpenAI, 2022.

[143] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 27730–27744.

[144] Q. Vuong et al., "Open X-embodiment: Robotic learning datasets and RT-X models," in *Proc. Towards Generalist Robots: Learn. Paradigms Scalable Skill Acquisition @ CoRL2023*, 2023.

[145] S. Park, H. Kim, and Y. M. Ro, "Incorporating language-driven appearance knowledge units with visual cues in pedestrian detection," 2023, *arXiv:2311.01025*.

[146] H. Peng, B. Li, B. Zhang, X. Chen, T. Chen, and H. Zhu, "Multi-view vision-prompt fusion network: Can 2D pre-trained model boost 3D point cloud data-scarce learning?," 2023, *arXiv:2304.10224*.

[147] H. Pham et al., "Combined scaling for zero-shot transfer learning," *Neurocomputing*, vol. 555, 2023, Art. no. 126658.

[148] R. P. K. Poudel, H. Pandya, S. Liwicki, and R. Cipolla, "Recore: Regularized contrastive representation learning of world model," 2023, *arXiv:2312.09056*.

[149] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7077–7087.

[150] 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.*, 2024, vol. 38, pp. 4542–4550.

[151] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[152] S. N. Rai, F. Cermelli, D. Fontanel, C. Masone, and B. Caputo, "Unmasking anomalies in road-scene segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4037–4046.

[153] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2001–2010.

[154] H. Ren, H. Gao, H. Chen, and G. Liu, "A survey of autonomous driving scenarios and scenario databases," in *Proc. IEEE 9th Int. Conf. Dependable Syst. Appl.*, 2022, pp. 754–762.

[155] A. A. Rusu et al., "Progressive neural networks," 2016, *arXiv:1606.04671*.

[156] A. Sabater, L. Montesano, and A. C. Murillo, "Robust and efficient post-processing for video object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 10536–10542.

[157] E. Sachdeva et al., "Rank2Tell: A multimodal driving dataset for joint importance ranking and reasoning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 7513–7522.

[158] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proc. Int.Conf. Adv. Neural Inf. Process. Syst.*, 2018, pp. 525–536.

[159] H. Sha et al., "LanguageMPC: Large language models as decision makers for autonomous driving," 2023, *arXiv:2310.03026*.

[160] H. Shao, Y. Hu, L. Wang, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," 2023, *arXiv:2312. 07488*.

[161] S. Shen et al., "K-lite: Learning transferable visual models with external knowledge," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, pp. 15558–15573.

[162] H. Shi et al., "Cobev: Elevating roadside 3D object detection with depth and height complementarity," 2023, *arXiv:2310.02815*.

[163] Hayati Shirley Anugrah et al., "Chain-of-instructions: Compositional instruction tuning on large language models," 2024, *arXiv:2402.11532*.

[164] S. Shubodh, M. Omama, H. Zaidi, U. S. Parihar, and M. Krishna, "Lip-Loc: Lidar image pretraining for cross-modal localization," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 948–957.

[165] C. Sima et al., "DriveLM: Driving with graph visual question answering," 2023, *arXiv:2312.14150*.

[166] A. Singh, "End-to-end autonomous driving using deep learning: A systematic review," 2023, *arXiv:2311.18636*.

[167] E. Soares, P. Angelov, B. Costa, and M. Castro, "Actively semi-supervised deep rule-based classifier applied to adverse driving scenarios," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[168] Z. Song et al., "VoxelNextFusion: A simple, unified and effective voxel fusion framework for multi-modal 3D object detection," 2024, *arXiv:2401.02702*.

[169] Z. Song et al., "RoboFusion: Towards robust multi-modal 3D obiect detection via SAM," *Proc. Int. Joint Conf. Artif. Intell.*, 2024.

[170] J. Su, S. Gu, Y. Duan, X. Chen, and J. Luo, "Text2Street: Controllable text-to-image generation for street views," 2024, *arXiv:2402.04504*.

[171] S. Suo, S. Regalado, S. Casas, and R. Urtasun, "TrafficSim: Learning to simulate realistic multi-agent behaviors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10400–10409.

[172] L. Tang, Y. Li, J. Yuan, A. Fu, and J. Sun, "CPSOR-GCN: A vehicle trajectory prediction method powered by emotion and cognitive theory," 2023, *arXiv:2311.08086*.

[173] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Veh. Symp.*, 2018, pp. 1013–1020.

[174] X. Tian et al., "DriveVLM: The convergence of autonomous driving and large vision-language models," 2024, *arXiv:2402.12289*.

[175] H. Touvron et al., "Llama: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[176] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[177] D. Unger, N. Gosala, V. R. Kumar, S. Borse, A. Valada, and S. Yogamani, "Multi-camera bird's eye view perception for autonomous driving," 2023, *arXiv:2309.09080*.

[178] A. Van Den Oord et al., "Neural discrete representation learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6309–6318.

[179] M. R. Van Geerenstein, F. Ruppel, K. Dietmayer, and D. M. Gavrila, "Multimodal object query initialization for 3D object detection," 2023, *arXiv:2310.10353*.

[180] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[181] K. Viswanath, P. Jiang, S. PB, and S. Saripalli, "Off-road lidar intensity based semantic segmentation," 2024, *arXiv:2401.01439*.

[182] G. Wang et al., "Voyager: An open-ended embodied agent with large language models," in *Proc. NeurIPS Found. Models Decis. Mak. Workshop*, 2023.

[183] J. Wang, Q. M. J. Wu, and N. Zhang, "You only look at once for real-time generic multi-task," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 12625–12637, Sep. 2024.

[184] P. Wang et al., "BevGPT: Generative pre-trained large model for autonomous driving prediction, decision-making, and planning," 2023, *arXiv:2310.10357*.

[185] T.-H. Wang et al., "Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models," in *Proc. 1st Workshop Out–Distrib. Generalization Robot. CoRL*, 2023.

[186] W. Wang et al., "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14408–14419.

[187] X. Wang, Z. Zhu, G. Huang, X. Chen, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," 2023, *arXiv:2309.09777*.

[188] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," 2020, *arXiv:2003.06957*.>

[189] Y. Wang et al., "Empowering autonomous driving with large language models: A safety perspective," 2023, *arXiv:2312.00812*.

[190] Y. Wang, Y. Chen, X. Liao, L. Fan, and Z. Zhang, "PanoOcc: Unified occupancy representation for camera-based 3D panoptic segmentation," 2023, *arXiv:2306.10013*.

[191] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024.

[192] Y. Wang, L. Cheng, C. Fang, D. Zhang, M. Duan, and M. Wang, "Revisiting the power of prompt for visual tuning," 2024, *arXiv:2402.02382*.

[193] Wayve, "Lingo-1: Exploring natural language for autonomous driving," 2023. [Online]. Available: https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/

[194] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14668–14678.

[195] D. Wei et al., "Bev-clip: Multi-modal Bev retrieval methodology for complex scene in autonomous driving," 2024, *arXiv:2401.01065*.

[196] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, pp. 24824–24837, 2022.

[197] Y. Wei et al., "Editable scene simulation for autonomous driving via collaborative LLM-agents,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024.

[198] L. Weiwei, H. Wenxuan, J. Wei, L. Lanxin, G. Lingping, and L. Yong, "Learning to model diverse driving behaviors in highly interactive autonomous driving scenarios with multi-agent reinforcement learning," 2024, *arXiv:2402.13481*.

[199] L. Wen et al., "DiLu: A knowledge-driven approach to autonomous driving with large language models,'' in *Proc.12th Int. Conf. Learn. Representations*, 2024.

[200] L. Wen et al., "On the road with GPT-4V (ision): Early explorations of visual-language model on autonomous driving," 2023, *arXiv:2311.05332*.

[201] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual chatGPT: Talking, drawing and editing with visual foundation models," 2023, *arXiv:2303.04671*.

[202] D. Wu et al., "Yolop: You only look once for panoptic driving perception," *Mach. Intell. Res.*, vol. 19, no. 6, pp. 550–562, 2022.

[203] D. Wu, W. Han, T. Wang, X. Dong, X. Zhang, and J. Shen, "Referring multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14633–14642.

[204] D. Wu, W. Han, T. Wang, Y. Liu, X. Zhang, and J. Shen, "Language prompt for autonomous driving," 2023, *arXiv:2309.04379*.

[205] Z. Wu et al., "Mars: An instance-aware, modular and realistic simulator for autonomous driving," in *Proc. CAAI Int. Conf. Artif. Intell.*, 2023, pp. 3–15.

[206] Y. Xiao, Y. Li, C. Meng, X. Li, and Y. Zhang, "CalibFormer: A transformer-based automatic lidar-camera calibration network," 2023, *arXiv:2311.15241*.

[207] X. Zhou et al., "DrivingGaussian: Composite Gaussian splatting for surrounding dynamic autonomous driving scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024.
[208] P. Xidong et al., "Learning to adapt sam for segmenting cross-domain point clouds," 2023, *arXiv:2310.08820*.
[209] Y. Xie et al., "SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17545–17556.
[210] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9653–9663.
[211] J. Xu, Y. Nie, P. Wang, and A. M. López, "Training a binary weight object detector by knowledge transfer for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 2379–2384.
[212] Q. Xu, T. Yue, S. Ali, and M. Arratibel, "Pretrain, prompt, and transfer: Evolving digital twins for time-to-event analysis in cyber-physical systems," 2023, *arXiv:2310.00032*.
[213] X. Xu, L. Kong, H. Shuai, and Q. Liu, "FRNet: Frustum-range networks for scalable lidar segmentation," 2023, *arXiv:2312.04484*.
[214] Z. Xu et al., "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," 2023, *arXiv:2310.01412*.
[215] X. Yan et al., "Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities," 2024, *arXiv:2401.08045*.
[216] B. Yang et al., "Diffusion-es: Gradient-free planning with diffusion for autonomous driving and zero-shot instruction following," 2024, *arXiv:2402.06559*.
[217] C. Yang et al., "Traffic sign interpretation in real road scene," 2023, *arXiv:2311.10793*.
[218] H. Yang et al., "Unipad: A universal pre-training paradigm for autonomous driving," 2023, *arXiv:2310.08370*.
[219] S. Yang et al., "Lidar-LLM: Exploring the potential of large language models for 3D lidar understanding," 2023, *arXiv:2312.14074*.
[220] Y. Yang, L. Fan, and Z. Zhang, "MixSup: Mixed-grained supervision for label-efficient LiDAR-based 3D object detection," in *Proc. 12th Int. Conf. Learn. Representations*, 2024.
[221] Z. Yang et al., "UniSim: A neural closed-loop sensor simulator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1389–1399.
[222] Z. Yang, Y. Zhang, J. Yu, J. Cai, and J. Luo, "End-to-end multi-modal multi-task vehicle control for self-driving cars with visual perceptions," in *Proc. IEEE 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2289–2294.
[223] Z. Yang, X. Jia, H. Li, and J. Yan, "A survey of large language models for autonomous driving," 2023, *arXiv:2311.01043*.
[224] H. Ye and D. Xu, "Taskprompter: Spatial-channel multi-task prompting for dense scene understanding," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.
[225] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11784–11793.
[226] Y. Zang, W. Li, J. Han, K. Zhou, and C. C. Loy, "Contextual object detection with multimodal large language models," 2023, *arXiv:2305.18279*.
[227] B. Zhang et al., "ReSimAD: Zero-shot 3D domain transfer for autonomous driving with source reconstruction and target simulation," in *Proc. 12th Int. Conf. Learn. Representations*, 2024.
[228] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," 2023, *arXiv:2306.02858*.
[229] H. Zhang et al., "OpenSight: A simple open-vocabulary framework for lidar-based object detection," 2023, *arXiv:2312.08876*.
[230] J. Zhang, F. Zhang, S. Kuang, and L. Zhang, "NeRF-LiDAR: Generating realistic LiDAR point clouds with neural radiance fields," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 7178–718.
[231] Q. Zhang et al., "Adaptive budget allocation for parameter-efficient fine-tuning," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.
[232] R. Zhang et al., "LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention," in *Proc.12th Int. Conf. Learn. Representations*, 2024.
[233] S. Zhang, D. Fu, Z. Zhang, B. Yu, and P. Cai, "TrafficGPT: Viewing, processing and interacting with traffic foundation models," 2023, *arXiv:2309.06719*.
[234] T. Zhang, A. H. S. Chan, Y. Ba, and W. Zhang, "Situational driving anger, driving performance and allocation of visual attention," *Transp. Res. Part F: Traffic Psychol. Behav.*, vol. 42, pp. 376–388, 2016.
[235] X. Zhang et al., "CAE v2: Context autoencoder with CLIP latent alignment," *Trans. Mach. Learn. Res.*, 2023.
[236] Y. Zhang et al., "Meta-transformer: A unified framework for multimodal learning," 2023, *arXiv:2307.10802*.
[237] Y. Zhang, H. Doughty, and C. G. M. Snoek, "Low-resource vision challenges for foundation models," 2024, *arXiv:2401.04716*.
[238] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "TrafficBots: Towards world models for autonomous driving simulation and motion prediction," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 1522–1529.
[239] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
[240] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3D occupancy world model for autonomous driving," 2023, *arXiv:2311.16038*.
[241] W. Zheng, R. Song, X. Guo, and L. Chen, "Genad: Generative end-to-end autonomous driving," 2024, *arXiv:2402.11502*.
[242] H. Zhou, J. Chang, T. Lu, and H. Zhou, "3D lane detection from front or surround-view using joint-modeling & matching," 2024, *arXiv:2401.08036*.
[243] S. Zhou et al., "LiDAR-PTQ: Post-training quantization for point cloud 3D object detection," in *Proc.12th Int. Conf. Learn. Representations*, 2024.
[244] X. Zhou, M. Liu, B. L. Zagar, E. Yurtsever, and A. C. Knoll, "Vision language models in autonomous driving and intelligent transportation systems," 2023, *arXiv:2310.14414*.
[245] Y. Zhou, L. Cai, X. Cheng, Z. Gan, X. Xue, and W. Ding, "Openannotate3D: Open-vocabulary auto-labeling system for multi-modal 3D data," 2023, *arXiv:2310.13398*.
[246] Y. Zhou et al., "Embodied understanding of driving scenarios," 2024, *arXiv:2403.04593*.
[247] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *Proc. 12th Int. Conf. Learn. Representations*, 2024.
[248] O. Zohar, A. Lozano, S. Goel, S. Yeung, and K.-C. Wang, "Open world object detection in the era of foundation models," 2023, *arXiv:2312.05745*.
[249] S. Zou, X. Huang, and X. Shen, "Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 5994–6003.
[250] H. Zunair, S. Khan, and A. Ben Hamza, "Rsud20k: A dataset for road scene understanding in autonomous driving," 2024, *arXiv:2401.07322*.

Sheng Luo iscurrently working toward the master's degree with Southeast University, Nanjing, China. His research interests include road scene understanding and computer vision.

Wei Chen iscurrently working toward the master's degree with Southeast University, Nanjing, China. His research interests include continual learning and computer vision.

Wanxin Tian received the master's degree in electronics and information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2021. He is currently a Senior Algorithm Engineer with DiDi, Beijing. His research interests include computer vision, large language models, and World Model.

Rui Liu is currently working toward the master's degree with the Harbin Institute of Technology, Harbin, China. Her research interests include visual generation and computer vision.

Luanxuan Hou is currently a Senior Algorithm Engineer with Voager, DiDi Chuxing, Beijing, China. He has authored or coauthored papers in CVPR, ICPR, and ICME. He was a Reviewer of NeurIPS, ICML, and ICLR. His research interests include computer vision and diffusion model.

Xiubao Zhang is currently a Senior Expert Algorithm Engineer with DiDi Company, Beijing, China. His research interests include computer vision, pattern recognition, and machine learning.

Haifeng Shen received the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications, Beijing, China, in 2006. He is currently a Principal Algorithm Engineer with DiDi, Beijing. His research interests include computer vision, speech recognition, time sequence prediction, and large language modeling.

Ruiqi Wu is currently working toward the master's degree with Southeast University, Nanjing, China. Her research interests include multitask learning and computer vision.

Shuyi Geng is currently working toward the master's degree with Southeast University, Nanjing, China. Her research interests include continual learning and computer vision.

Yi Zhou is currently an Associate Professor with the School of Computer Science and Engineering, Southeast University, Nanjing, China. His research interests include computer vision, pattern recognition, and machine learning.

Ling Shao (Fellow, IEEE) is the Founding CEO and Chief Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include computer vision, deep learning/machine learning, and image/video processing. He is an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and several other journals. He is a Fellow of the IAPR.

Yi Yang (Member, IEEE) is the Director of Technology Innovation Group with DiDi, Beijing, China. His research interests include artificial intelligence, infrastructure, and cloud native.

Bojun Gao is currently a Research Project Manager with DiDi, Beijing, China. His research interests include artificial intelligence and educational technology.

Qun Li is currently the Head of DiDi Research Outreach and leading DiDi's efforts to work together with academic institutions worldwide, including research collaboration, talent cultivation, and academic exchange. She was in the organizing committee of many workshops and tutorials. Her research interests include in the intersection of artificial intelligence, autonomous driving, and computer vision.

Guobin Wu is currently the Director of Technology Ecology and Development Department with DiDi, Beijing, China. His research interests include artificial intelligence and computer vision.