

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, Yi Yang, 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, embodied driving agents, and world models. 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>.

Index Terms—Foundation Model, Visual Understanding, Multi-modal Learning, Multi-task Learning, Road Scene.

I. INTRODUCTION

INTELLIGENT vehicles have made great progress by achieving significant advancements in perceiving road scenes, employing various tasks such as object detection, tracking, segmentation, trajectory prediction, and advanced generation methods. The integration of vision-centric (such as camera, Lidar) and vision-beyond (such as language, action) modalities further enhances the ability of autonomous vehicles to perceive the world from diverse dimensionalities. However, traditional driving-relevant models are typically designed for specific tasks and utilize common convolution neural networks (CNNs) or Transformer architecture to extract feature maps

S. Luo, W. Chen, R. Wu, S. Geng and Y. Zhou are with the School of Computer Science and Engineering, Southeast University, Nanjing, China, and Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China. Corresponding author: Yi Zhou (yizhou.szc@gmail.com)

W. Tian, R. Liu, L. Hou, X. Zhang, H. Shen, Y. Yang, B. Gao, Q. Li and G. Wu are with DiDi Chuxing, Beijing, China.

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

from a single modality. This fashion often leads to incomplete observation due to complex and unpredictable conditions in real-world road scenes and poses a significant risk to achieving reliable and safe driving. 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 such as recognizing and localizing traffic participants on the road. This approach enables the models to learn shared representations, facilitating mutual performance gains across tasks. Moreover, with the integration of multi-modal capabilities (e.g., various sensors, vision, and language), these models contribute to creating more versatile and human-compatible systems. These systems excel in processing data from multiple modalities, leveraging complementary information to build a comprehensive understanding of the surrounding environment.

The emergence of Foundation Models (FMs) is regarded as a milestone in achieving artificial general intelligence (AGI). Recently, large language models (LLMs) [16], [129]–[131], [160], [161], vision language models (VLMs) [39], [106], [107], [182], [204], [219], and large vision models (LVMs) [9], [89], 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, well-trained 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 aligning vision modality into language latent space. Their key focus lies in aligning visual information with language understanding, thereby enhancing the capacity to comprehend and reason about visual content. This alignment contributes to more intuitive and interactive human-machine interactions. 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, following a modular design pattern, and is usually trained on a single 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. Early works have shown that a unified model trained by performing multiple tasks can achieve better performance than that of each specific task. 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 a more comprehensive understanding of road scene characteristics. 3) Third, the intricacies of road scenes demand systems that possess a profound understanding of the surrounding environment and the ability to make accurate judgments. Therefore, motivated by the great success of FMs, the realization of their pivotal role in the visual understanding of driving scenarios has taken center stage. The outstanding generalization capabilities of MM-VUFMs make them an ideal solution to tackle this challenge. 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 the MM-VUFMs.

Comparisons with related surveys. We investigated previous related surveys to clarify a difference between our survey with them [20], [36], [48], [57], [96], [143], [153], [192], [200], [217]. Some previous surveys [20], [36], [153], [200], [217] 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 [192] 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 (up until February 2024) multi-modal multi-task visual understanding foundation models (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.
- Advanced strengths of MM-VUFMs in diverse paradigms are also highlighted, including open-world understanding, efficient transfer learning, 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) [66]. CNNs are classic and fundamental architectures for computer vision tasks, primarily known for their ability to extract local features using convolutional layers. These layers apply filters to capture local patterns and textures, efficiently encoding spatial hierarchies and features in visual data. CNNs open up a new era for computer vision and continue to maintain their prominence and impact in the era of large-scale foundation models [43], [140], [172]. Despite CNN's long-range dependencies through applying large kernels or recursive gated kernels, when it comes to scaling up ability, there still exists a performance gap between CNN and Transformer architecture.

Transformer [165]. Transformer is an attention-based sequence-to-sequence learning architecture originally introduced to extract long-range dependencies from natural language. Unlike previous sequence-to-sequence models that relied on recurrent or convolutional layers, Transformer exclusively employs the self-attention [165] mechanism which allows the model to weigh different parts of the input sequence differently during processing, enabling parallelization and improving efficiency in capturing long-range dependencies. Transformer has formed the foundation for major advancements in natural language processing (NLP) and, more recently, computer vision.

Vision Transformer(ViT) [49]. ViT marks a significant extension of the Transformer architecture from the area of natural language processing to computer vision. ViT divides an image into fixed-size patches which are linearly embedded, augmented with positional embeddings to maintain spatial information, treated as sequences, and fed into the Transformer architecture. The self-attention mechanism enables ViT more adept at capturing global features across the image compared to CNN, making it suitable for classic computer vision tasks [19], [49], [213].

Vision-Language architecture. 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, vision-language architectures 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

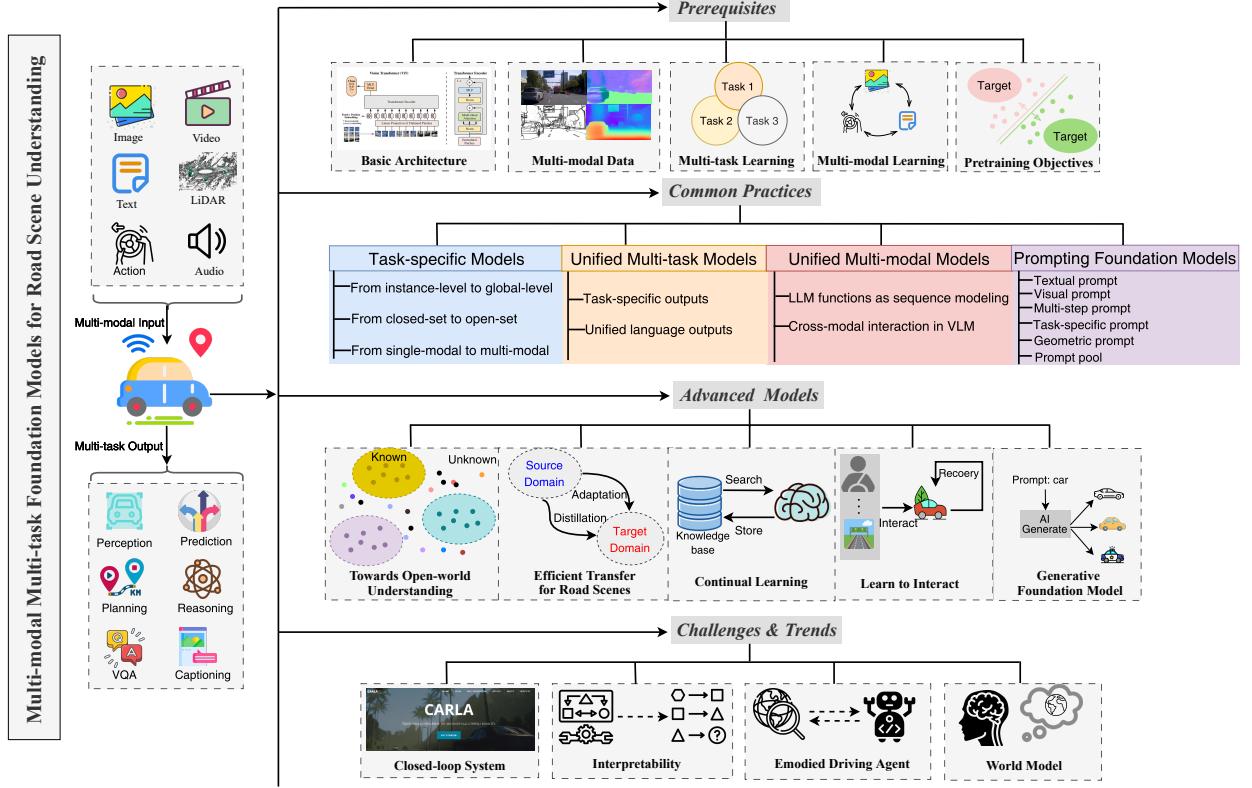


Fig. 1. Overview of our survey at a glance. A multi-modal and multi-task foundation model for road scene understanding is defined as a framework that inputs multi-modal data and outputs multi-task results. In the section of prerequisites, we introduce some basic knowledge in advance before reading the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

functioning of road scene understanding [55]. 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.

Camera data include images and videos captured by various onboard cameras such as RGB cameras, depth cameras, and thermal cameras. Raw image data 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. RGB images present a naturalistic perspective of the scene and furnish abundant semantic and textured information, helping understand the scene. A sequence of RGB 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 from the object to the car, helping in localization, and thermal images help detect moving objects (e.g. humans and vehicles) with relatively high temperature.

Lidar measures the accurate surrounding 3D points and is robust to various illumination conditions compared to a video camera, helping reconstruct the entire environment and obtain holistic perception. Raw lidar point data can be represented

as $P \in \mathbb{R}^{N \times 3}$ where N denotes the number of points and 3 stands for 3 values for each point, i.e., the coordinates x, y, z .

Recently, with advancements in LLMs and VLMs, there has been a growing trend to harness a broader spectrum of data from other modalities that extends beyond vision-centric inputs. In our survey, these vision-beyond multi-modal data include text data, emotional data, and action data. Text data provide an overview understanding of the scene using natural language descriptions. Emotion data are collected from the camera and annotated by annotators, 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 corporation of these vision-beyond multi-modal data enables the autonomous system to gain a more comprehensive human-like understanding of the holistic road scenes and make the system more interpretable to common drivers.

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

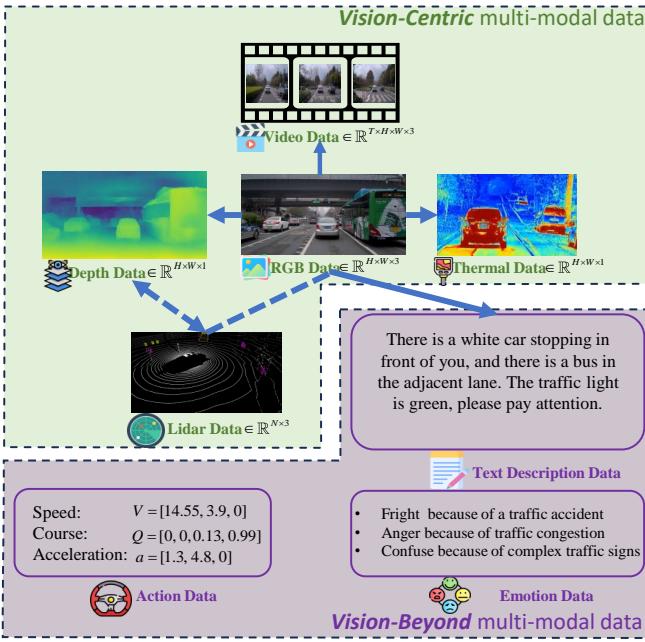


Fig. 2. Common multi-modal data used in road scenes. We divide them into two groups, i.e. **vision-centric** multi-modal data and **vision-beyond** multi-modal data. Solid arrows denote the strong connections between two modalities and dashed arrows denote weak connections. Vision-centric multi-modal 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.

the ability to interact across modalities. The success of unified multi-modal learning methods such as Meta-Transformer [211] 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 [140] 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 [59] 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 each individual task, a multi-task learning approach aims to leverage shared representations and features 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 for handling various tasks, multi-task architectures involve two categories: hard-parameter sharing [84] and soft-parameter sharing [112]. Specifically, in hard-parameter sharing architecture, a single set of shared parameters or a

shared feature extractor is used for all tasks. The shared parameters are responsible for extracting features that are considered task-agnostic. However, each task has its own task-specific heads or layers that follow the shared parameters. These task-specific heads are responsible for generating task-specific outputs. The primary objective is to reduce the risk of overfitting on individual tasks by sharing parameters in the main body of the model. 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 [84], [109], which allows to express the relative importance or priority of each task during the optimization process. Weighting strategy can be further categorized as follows: Static weighting involves adjusting weights before the start of training based on task frequency and difficulty. For example, assign higher weights to less frequently occurring tasks or more challenging instances to give them more emphasis during optimization. Dynamic weighting involves adjusting weights dynamically during the optimization process based on real-time task performance, task gradient. For example, adaptively increase the weight of tasks with poorer performance or divergence of objectives via a adaptive task balancing module. 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 a valuable approach to address complex problems with conflicting objectives. For example, Pareto-based methods [105], [113], [146] 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 recently propelled the fundamental cognitive abilities of intelligent agents. These models, characterized by their largeness and generality, are pretrained on extensive datasets primarily through self-supervised or weakly-supervised learning methods. They exhibit versatility by being easily adapted to various downstream tasks, serving as an excellent starting point for model initialization. Contrastive and generative pretraining both refer to providing a good model initialization using large-scale multi-modal data (such as image-text pairs). Through the stage of pretraining, the model can acquire the intrinsic structure and patterns from multi-modal data. The learned knowledge is easily transferred to downstream tasks. As fol-

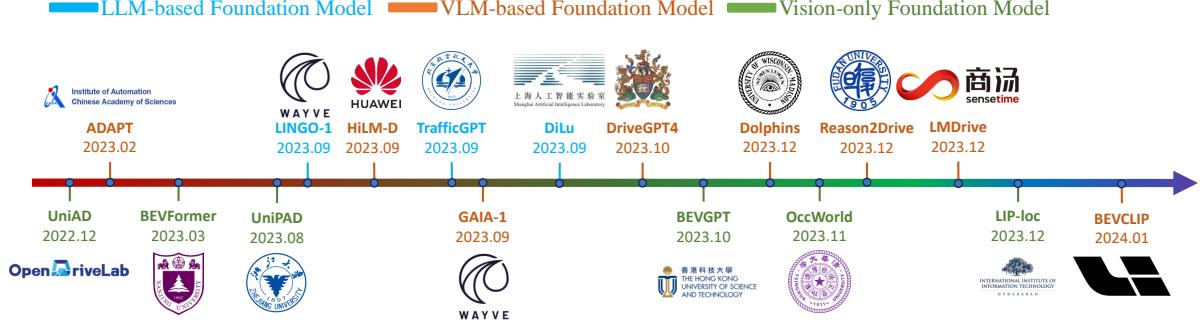


Fig. 3. Roadmap of recent foundation models in driving scenarios. We divide these foundation models into LLFM, VLFM, and LVFM based on the data modality they use. LVFMs are vision-only large-vision foundation models that only take vision-centric data as input. Pretrained on large-scale datasets, these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

lows, we summarized two types of pretraining methods:

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 [23], [24], [26], [65] 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 Eq. 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 [10], [18], [78], [136], [140], [149]. The image-text contrastive loss is calculated as Eq. 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 [47], [54], [97], [98] have shown more impressive performance than vallina version, reducing the cost of computation resources and time.

Masked image modeling (MIM). Encoders gain enhanced universality and adaptability through the generation or reconstruction of the input [10], [25], [64], [188], [210], which directs the encoder to learn semantic representations by randomly masking a percentage of image regions and predicting the masked regions from the visible ones. The MIM loss is formulated as Eq. 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 [64] and SimMIM [188], 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 [3]. Subsequent methods have started to emphasize making predictions in the encoded representation space like CAE [25], [210], MaskFeat [177] as an efficient and scalable method for learning semantic representations.

Masked language modeling (MLM). Early masked language models like BERT [1], RoBERTa [110] are trained to predict masked word tokens x_m given a sequence of tokens x_v also known as "Cloze task", as shown in Eq. 4:

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

Furthermore, large language models (LLMs) like GPT-3 [16] predict next token given the prefix language tokens in an autoregressive manner. The pretraining objective is employed in Decoder-only LLMs, as shown in Eq. 5:

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

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

B. Roadmap of Visual Understanding Foundation Models for Road Scenes

As shown in Fig. 3, in recent two years, we've 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 they use during pretraining phase.

Kim et al. [87] 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 [160]. TrafficGPT [208], DiLu [179], and LINGO-1 [176] 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 [82] 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 [107] and BLIP2 [93] release. DriveGPT4 [191], Dolphins [114] and many other works [45], [127], [148], [178] finetune a visual encoder and align the visual features into the language features via Q-former [93] to leverage the robust reasoning ability of LLM, contributing to a better human-like understanding of the environment. GAIA-1 [69] 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 [2], [148] 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 [71] 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 [151] applies contrastive pretraining to 2D images and 3D points cloud. UniPAD [195] proposes a self-supervised framework to enhance feature learning using only images. To exploit the strengths of BEV (Bird's-Eye View) map, BEVFormer [99] leverages multi-camera images and spatio-temporal Transformer to construct robust BEV representations. BEVGPT [169] integrates prediction, decision-making, and motion planning taking the BEV images as the only input source. As for vision generative models, OccWorld [214] 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 Subsec. III-A, unified multi-task models in Subsec. III-B, unified multi-modal models in Subsec. III-C, and prompting foundation models in Subsec. III-D, respectively.

A. Task-specific Models

Task-specific models for road scenes are designed and trained for specific downstream tasks. These models are trained with specific data and tailored architectures 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 [60], [88], [215], 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 traffic instructions on road based solely on a single traffic sign can be challenging. Yang et al. [193] 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 [141], [203] excels in achieving a comprehensive global perspective. Mask2Anomaly [141] 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 [7] 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

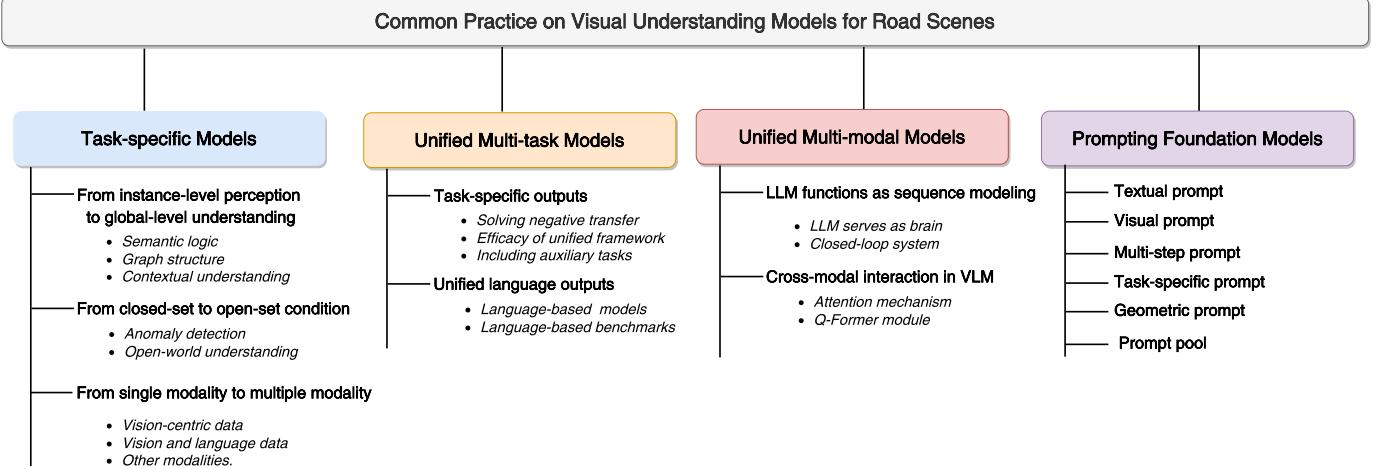


Fig. 4. The overall framework of Sec. III. We review existing works on visual understanding for road scenes from four perspectives: task-specific models, unified multi-task models, unified multi-modal models, and prompting foundation models, respectively. By organizing the review based on these four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

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 [62], [150] and segmentation methods [100], [121] 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 [94], which holds significant importance in road scenes, particularly for ensuring safety and reliability. Previous method [32] 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, SalientDet [44] solves open-world detection in road scenes via salience maps-based feature enhancement techniques for unknown objects. Likewise, GOOD [72] 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 learning targets the recognition of arbitrary objects based on textual input by leveraging advancements in vision-language learning. OVTrack [95] extends a closed-set object tracker into an open-vocabulary tracker by leveraging knowledge distillation from a pretrained vision-language model [140]. 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. Bogdol et al. [12] 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 recog-

nize 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 [85], [197], [216], the utilization of multiple modalities proves beneficial in obtaining a comprehensive understanding of road scenes [55], [59], [138].

Vision-centric data. Many methods have shifted their attention to multiple visual sensor data, ranging from camera-only [81], [99], [162], [175] or lidar-only [31], [166], [190], [202], [205], to collaborate vision-centric sensors [8], [30], [155], [187]. These methods for collaborating sensors mainly involve two categories, multi-sensor fusion and multi-sensor calibration. Multi-sensor fusion methods [8], [40], [155], [164] aim to observe road scenes via utilizing complementarity across sensors in a fusion fashion. TransFusion [8] 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 [67], [80], [81], [83], [186] is proposed. Jiang et al. [80] 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 [28], [29], [108], [126], [133] focus on further integrating linguistic comprehension into vision-centric tasks. VLPD [108] introduces a novel pedestrian detection

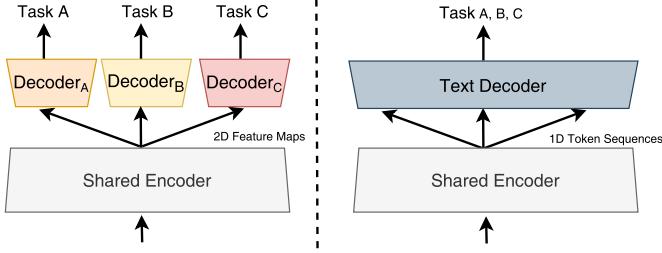


Fig. 5. Unified multi-task models can be categorized based on their outputs into two distinct types. The first type (left) includes models with task-specific outputs, characterized by a shared encoder and individual task-specific heads across all tasks. In this architecture, the shared encoder processes 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.

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 [11], [158] also serve as valuable information sources. CPSOR-GCN [158] 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 [120], 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 [159], [199]. 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 [86], [164], RGB-X object detection [40], and traffic sign interpretation [193].

Solving the negative transfer problem across tasks is the primary concern for MTL. In this trend, VE-Prompt [103] aims to generate task-specific features for different task heads 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 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 [38] introduces a novel multi-task framework designed for video instance segmentation and tracking. This framework incorporates an innovative associative connection across different task heads. The associative connection facilitates the fusion of outputs from various task heads. For instance, bounding boxes are repurposed for the mask head, leading to more accurate segmentation results.

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 [183] 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 [168] 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 [124] 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 [91] has been widely discussed recently with the following characteristics of remembering history, learning experience, modeling the world, and predicting the future. For example, OccWorld [214] is a recent world model in predicting future scenes with the auxiliary ego-car motion trajectory prediction. This method reveals the potential of modeling on 3D occupancy representation to capture fine-grained structural understanding, rather than conventional bounding boxes and segmentation maps. Through efficiently discretely encoding [163] to get high-level semantic representations of past frames and spatial-temporal autoregressive-based generative transformer to enable interaction world tokens in the same and cross timestamp, task-specific decoders can make more reasonable prediction than ground truth labels generated by self-supervised methods, providing a promising trend for interpretable end-to-end autonomous driving without additional supervision.

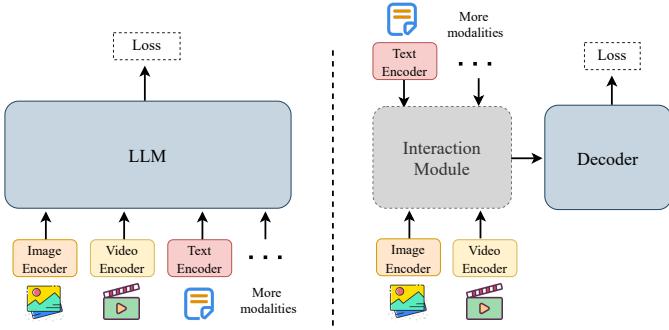


Fig. 6. Comparison of LLM-based (left) and VLM-based (right) unified multi-modal models. The LLM-based model takes LLM as a center place, which transforms multi-modal data into textual tokens that are easily modeled by LLM in the manner of sequence modeling. The VLM-based model emphasizes cross-modal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

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.

In generic vision, VisionLLM [171] has pioneered a solution to address fragmented output formats by mapping visual data into text tokens. Likewise in related driving tasks, a novel approach in road scenes involves adopting natural language as the unified output. For example, HILM-D [46] 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 [82] 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 [194], 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 [185] expands the Nuscenes dataset [17] to instance-text pair by constructing language descriptions for video clips. Similarly, NuScenes-MQA [37] provides an evaluation of a model’s capabilities in sentence generation and VQA annotated in a questions and answers (QA) format. DriveLM [152] also facilitates perception, prediction, and planning with logical reasoning. DRAMA [116] 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 about 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. While both LLM-based and VLM-based

models exhibit powerful multi-modal understanding in the context of road scenes, there exist nuanced differences in their functionalities. 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 [148], [191] 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 [114] 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 from two perspectives: LLM functions as sequence modeling and cross-modal interaction in VLM.

LLM functions as sequence modeling. In contrast to module design in previous driving systems, introducing LLM as the central brain has become a prominent fashion. This kind of method 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 [117] 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 [191] not only leverages a video tokenizer to encode video frames and a shared text tokenizer to tokenize user query and past control signals but employs MLLM to predict desired answers and future control signals. LanguageMPC [147] harnesses the reasoning capability of LLM to analyze complex road scenes and make high-level textual decisions. Then these textual decisions are 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 [21]), 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 unexpected obstacles and changes in road scenes. Typically, [56], 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 [148] 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 [2] serves as MLLM-based intelligent agents. In addition to the commonly mentioned visual signals, DriverMLM 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. [174] introduce LLM as a behavior planer with safety assurance. The LLM-based driver agent simultaneously ensures safety and reinforcing 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 (referring to the outputs of perception, prediction, mapping, and occupancy module) into language tokens, Agent-Driver [118] 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 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 [114] based on a foundational VLM, named openFlamingo [6], featuring in-context learning capabilities, inserts new gated cross-attention layers within a pretrained frozen LLM with the LoRA [70] module. These newly added layers promote stable interaction between vision and language via gate-mechanism. To improve interpretability during the decision-making process, Reason2Drive [127] views the decision-making process as chain-based reasoning task. It also provides insights into reasoning abilities of VLM, by solving the limitation of VLM in leveraging object-level perception prior. 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 [170] 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 [42] 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 [185] incorporates a novel prompt reasoning branch based on current object tracking algorithms to deal with referring object tracking whose goal is to track desired objects according to multi-view multi-frame images input and user 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. The traditional paradigm, referring to the "pretrain-then-finetune", involves two stages of pretraining a model on large-scale datasets to learn general representations, and then finetuning the pre-trained model on specific tasks or domains. Currently, an emerging trend is to highlight the significance of prompt engineering on transfer learning. For example, Liang, et al. [104] 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. [189] 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 transferring 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 [2], GPTDriver [117] and DriveGPT4 [191], 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 [41], [42], [180] involve utilizing visual cues to enhance the scenario understanding. The representative algorithms in this line include multi-view vision-prompt [134], cross-modal prompt fusion [220], and learnable visual prompts [103] 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 [114] 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 [117] 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 [218] introduces an open-vocabulary auto-labeling pipeline to generate 2D mask, 3D mask, and 3D bounding boxes for multi-modal data (vision and point cloud 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 [103], [104] 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. [104] 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 [103] 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 [201] 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 of modeling human state change given long-range series signals, Niu et al. [128] incorporate *window prompt* to enable flexible and efficient attention on a local window. Similarly, DriveGPT4 [191] 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 perception system to adapt and improve its performance over time. I3DOD [102] 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. Specifically, after training on the current task, the prompts used will be updated and stored in the prompt pool, allowing the model to leverage them for establishing relationships with previous classes when tackling the next task.

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 paradigms by providing concrete examples and specific algorithms. 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 road landscape is constantly changing, presenting a multitude of variables and unpredictable elements. The dynamics of traffic flow, diverse road conditions, and unexpected events result in the complexity of

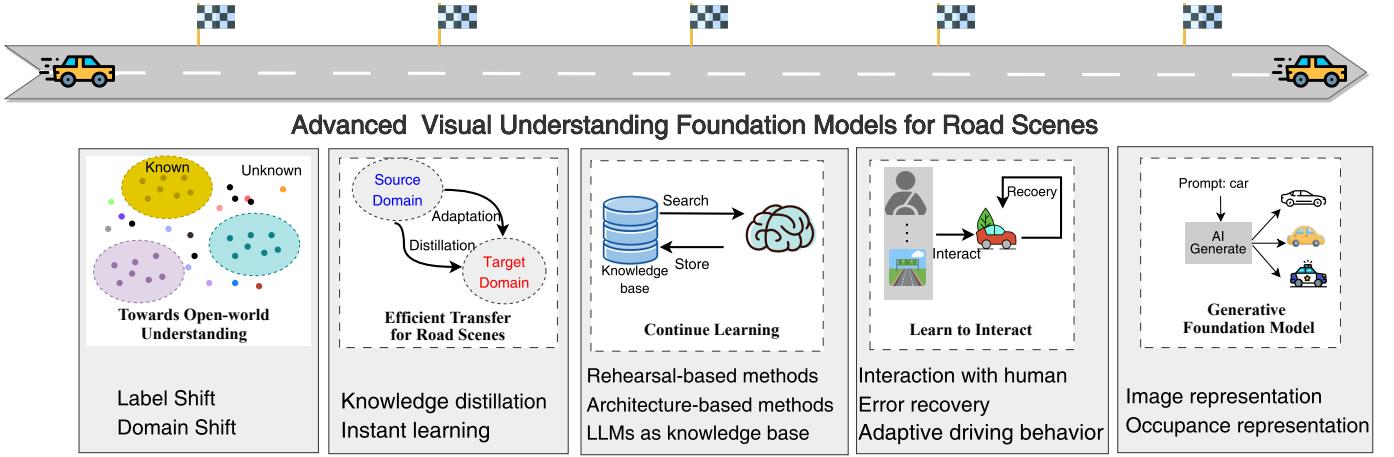


Fig. 7. The overall framework of Sec. IV. The advanced capabilities of visual understanding foundation models are highlighted from five perspectives: towards open-world understanding, efficient transfer for road scenes, continual learning, learn to interact, and generative foundation models, respectively.

the open-world environment. Existing foundation models have made great progress towards open-world understanding. The Fig. 8 illustrates GPT-4V's [180] understanding capability on open-world road scenes with both label shift and domain shift.

Label shift refers to great changes in label distribution between the training data and real-world scenarios encountered during deployment. Existing methods often lack the capability for a holistic understanding of long-tailed road scenes, encompassing special scenarios, such as animals on the highway and overturned trucks. These extreme conditions underscore the necessity for further advancements in open-world generalization. In this direction, DriveLM [152] 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. [52] 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. In addition, existing datasets tailored for road scenes usually fall short of adequately covering unusual cases due to the substantial cost associated with capturing such scenarios. 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 [157] 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, KING [63] 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, 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. GPT-4V [180] has undergone extensive road experiments in real-world road scenes. Observations indicate that GPT-4V has demonstrated superior understanding and inference capabilities in dealing with unseen scenarios compared to existing driving agents. DriveAnywhere [170] 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 [114] is built upon a general vision-language model [6], 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.



Fig. 8. Illustration of GPT-4V's capability on open-world road scene understanding in both label shift and domain shift scenarios. The example images are from CODA [94]. Given the prompts, right answers in GPT-4V's response are highlighted. These results reveal that GPT-4V can identify scene objects and further provide advice for safe driving, underscoring remarkable understanding capability of existing foundation models.

B. Efficient Transfer for Road Scenes

Efficiently transferring to road scenes involves adapting generic knowledge gained from pretraining to road scenes. Due to a great domain gap between natural scenes and road scenes, this process demands a robust framework that can extrapolate relevant information from the known natural scenes and then apply it intelligently to unseen road scenes. This means that the model is required to not only utilize 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. To achieve efficient knowledge transfer, FMs, which are equipped with adaptive capabilities, are essential for identifying commonalities and specialties, enabling a smoother knowledge distillation between road scenes and other scenes.

These models not only recognize patterns in the existing road data but also discern underlying conditions that can be translated into diverse 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 [89], [130], [140] 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-relevant tasks. However, directly deploying a large model onto autonomous vehicles might be computationally expensive and impractical due to the constraints in memory, power, or processing capabilities. To address this, knowledge distillation allows for the compression and transfer of the teacher model’s knowledge into the student model.

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 [151] 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 [68]. 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 [22] 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 [89] to the 3D task using VoxelNeXt [27], 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 [156] 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. [135] 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 [2] 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 [5]. 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 [16]. 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 [114] is the pioneering work that harnesses the in-context capability of VLMs [6] to instantly learn and adapt to a series of driving-related tasks. Extensive experiments on DriveLM [152] 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 the generalization abilities of FMs like [92], [107], [219]. To unlock the powerful capabilities of FMs and apply them to solve driving tasks, researchers like [114] have attempted to transform road-related benchmarks and datasets via input-output tuples into instruction-following data organized in the form of 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 in the same way. 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* [119], where the acquisition of new knowledge leads to an abrupt erosion of previously learned information.

In the field of continual learning, many works are tackling the catastrophic forgetting problem. Rehearsal-based methods, such as ICaRL [142] use a subset of old data or synthetic samples to replay old knowledge. Regularization-based techniques, such as EWC [90], preserve crucial weights to maintain past learning. Architecture-based approaches like progressive neural networks [144], add new modules for new tasks while preserving existing ones. These methods, addressing the plasticity-stability dilemma, enable neural networks to acquire new capabilities without losing previously learned skills.

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. [101] 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 [123] 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 [154] 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” [111], 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 [180]. This is akin to human learning through conversation, where each question and instruction helps narrow down the topic and clarify understanding. Given this kind of ability, in the trend of merging LLMs into autonomous systems, existing systems [34], [35], [56], [174], [179] 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. [56] 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 [16]. DiLu [179] 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.

In conclusion, compared to traditional task-specific models’ continual learning which mainly addresses the issue of “forgetting” during the fine-tuning stage, these LLMs-based autonomous systems emphasize continuously acquiring new skills during inference and do not strictly evaluate their performance on old tasks like the traditional ones. Moreover, unlike humans, LLMs don’t actually “learn” or “remember” during the conversation [181]. What might appear as “remembering” is the model’s ability to use its immediate context window effectively, i.e., in-context learning [16]. When a user provides additional information or clarification in a series of prompts, the model incorporates this new information within its immediate context to generate a response [16]. This distinction recognizes the nuanced way these models leverage their vast pretrained knowledge and reasoning skills to adapt to new road scenes, reflecting a more general and interpretable adaptive form of learning, rather than the parameter-driven paradigm in traditional continual learning models.

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 [70] and various Adaptors [206], [207]. We hope that future research will integrate some of the unique aspects of the road scene (corner case [94] and long-tail problem [76]) and propose knowledge-enhanced continualizable 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 and everyone wants to figure out how autonomous driving systems make decisions so that passengers find it safe to ride in the car. There are many works [82], [116] 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 [191], LMDrive [148]. 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 [73]. These challenges are particularly pronounced in road scenes, where unexpected obstacles or conditions can arise at any moment. Dolphins [114] proposes a reflection mechanism to enable the model to correct the mistakes by itself through continuous conversations with users. LMDrive [148] 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 driver's emotion should be taken into consideration for safe driving. 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. [125] 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 [209], which is difficult for traditional models to predict. CPSOR-GCN [158] 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 the 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, structures, 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, known as generative world model [13], [61], [69], [79], [91], [122], [137], [173], [212], 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. GAIA-1 [69] 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 [173] 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

Datasets	Statistics	Scenario	Task					Modality			
			Perception	Prediction	Planning	Understanding	Image	Video	Text	Action	Point Cloud
LaMPIlot [115]	4.9k scenes, 7.6 instruction length per frame	simulator	✗	✗	✓	✗	✓	✓	✓	✓	✗
DriveLM-nuScenes [152]	4,871 frames, 91.4 QAs per frame	urban	✓	✓	✓	✓	✓	✗	✓	✗	✗
DriveLM-CARLA [152]	183,373 frames, 20.5 QAs per frame	simulator	✓	✓	✓	✓	✓	✗	✓	✗	✗
DriveGPT4 [191]	28K video, 16K fixed QAs, 12K conversations	urban,rural	✓	✓	✓	✓	✗	✓	✓	✓	✗
Talk2BEV [42]	1k BEV scenarios, 20 QAs per scenarios	urban	✓	✗	✗	✓	✓	✗	✓	✗	✓
Rank2Tell [145]	116 scenarios, 31.95 caption length per scenarios	urban	✓	✗	✗	✓	✓	✓	✓	✓	✓
NuPrompt [185]	34k frames, 1.1 prompts per frame	urban	✓	✓	✗	✗	✓	✗	✓	✗	✗
NuScenes-QA [139]	30k frames, 15.3 QAs per frame	urban	✓	✗	✗	✓	✓	✓	✓	✗	✓
DriveMLM [2]	50k routes, 30 scenarios, 200 trigger points per scenarios	simulator	✓	✓	✓	✓	✓	✗	✓	✓	✓
LMDdrive [148]	64K parsed clips, 464K notice instructions	simulator	✓	✓	✓	✓	✓	✓	✓	✓	✓
Reason2Drive [127]	600K video-text pairs	urban,rural	✓	✓	✗	✓	✓	✗	✓	✗	✗
Driving-With-Llm [21]	10k driving scenarios, 16 QAs per scenarios	urban	✓	✓	✗	✓	✗	✗	✓	✗	✗
Refer-KITTI [184]	6.6k frames, 10.7 instances per prompt	urban,rural	✗	✗	✓	✓	✗	✓	✓	✓	✗
NuScenes-MQA [37]	34k scenarios, 41.2 QAs per scenarios	urban	✓	✗	✗	✗	✓	✗	✓	✗	✗
LiDAR-text [196]	420K 3D captioning data, 280K 3D grounding data	simulator	✓	✗	✗	✓	✗	✗	✓	✗	✓
CARLA-NAV [77]	83k frames, 7 command length per frames	simulator	✗	✗	✓	✗	✓	✗	✓	✗	✗
DRAMA [116]	18k scenarios, 5.6 text strings per scenarios	urban	✗	✗	✓	✓	✗	✓	✓	✓	✗

TABLE I

SUMMARY OF UP-TO-DATE LANGUAGE-BASED MULTI-MODAL MULTI-TASK DATASETS FOR ROAD SCENE UNDERSTANDING. THE SUMMARIZED PERSPECTIVES OF THESE DATASETS INVOLVE CONSIDERATIONS SUCH AS DATA STATISTICS, APPLICABLE SCENARIOS, TASKS, AND ANNOTATED MODALITIES.

context, color, and shape information, they might struggle with spatial structures.

Occupancy representation. OccWorld [214] 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 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 [167] and designing a novel search strategy [53], [167] are both useful ways to deal with real-world challenges.

V. 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. V, 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 [94], KITTI [58], and Cityscapes [33] 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 [116] or extended on existing datasets like DriveLM [152]. Moreover, there is a trend that QA pairs for different tasks are organized from Chain-of-Thought like Reason2Drive [127], Tree-of-Thought like NuScenes-QA [139] and Graph-of-Thought like DriveLM [152].

VI. OPEN CHALLENGES & FUTURE TRENDS

Within the promising landscape of road scene understanding, numerous challenges persist. Identifying and addressing these challenges are crucial steps toward ensuring a safe, efficient, and reliable autonomous driving system. In this section, we delve into open challenges shared by existing

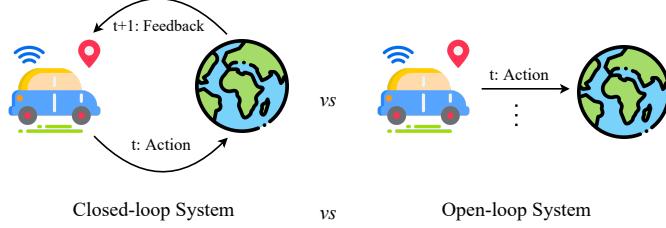


Fig. 9. Closed-loop system vs Open-loop system. In the closed-loop system, it maintains real-time interaction with the surrounding environment. This involves executing actions to modify the environment at the current time t and receiving feedback from the environment at the next time $t + 1$. In the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

approaches and also highlight future trends that unveil insights towards solving these challenges.

A. Closed-loop vs Open-loop Driving Systems

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 [17] 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 [148] 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 the unrealistic open-loop and the reliable closed-loop driving agent, open-source 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 [50], 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 [50], [198] 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 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 a high-quality driving scenario is a promising direction in the future.

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 [46], [82], [116] 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 [46], [74], decoding intermediate outputs into results. The output of interpretable tasks such as object detection [30], [138], depth estimation [30], [75] 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 [4], [152].

C. 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 the goal of DriveAGI by endowing them with embodied reasoning capacity. It emphasizes the importance of *perception-action loop*, which involves observing, modeling the world, taking action, and actively adjusting their driving behavior.

Recent advancements in LLM and multi-modal learning have brought the concept of embodied intelligence into practical application, such as robotics. PaLM-E [51], RT series [14], [15], [132] 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. Notably, an embodied driving agent can be seen as a specialized form of a robot agent. Borrowing technology from the field of robotics is also indeed a great direction for developing embodied driving agents.

D. World Model

The world model 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 [69] and DriveDreamer [173] 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. Besides, 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 [167] 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.

ACKNOWLEDGEMENT

This work was supported by DiDi GAIA Research Cooperation Initiative.

REFERENCES

- [1] Shivaji Alaparthi and Manit Mishra. Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey. *arXiv preprint arXiv:2007.01127*, 2020.
- [2] Anonymous. Drivelm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *Under Review*, 2023.
- [3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.
- [4] Shahin Atakishiyev, Mohammad Salameh, Housam Babiker, and Randy Goebel. Explaining autonomous driving actions with visual question answering. *arXiv preprint arXiv:2307.10408*, 2023.
- [5] Apollo auto. Baidu. <https://github.com/ApolloAuto/apollo>, 2019.
- [6] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- [7] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Learning pedestrian group representations for multi-modal trajectory prediction. In *European Conference on Computer Vision*, pages 270–289. Springer, 2022.
- [8] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022.
- [9] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023.
- [10] H Bao, L Dong, and F Wei. Beit: Bert pre-training of image transformers. *arxiv* 2021. *arXiv preprint arXiv:2106.08254*.
- [11] Neeraja Bhide, Nanami Hashimoto, Kazimierz Dokurno, Chris Van der Hoorn, Sascha Hoogendoorn-Lanser, and Sina Nordhoff. Defining, measuring, and modeling passenger’s in-vehicle experience and acceptance of automated vehicles. *arXiv preprint arXiv:2309.10596*, 2023.
- [12] Daniel Bogdolla, Lukas Bosch, Tim Joseph, Helen Gremmelmaier, Yitian Yang, and J Marius Zöllner. Exploring the potential of world models for anomaly detection in autonomous driving. *arXiv preprint arXiv:2308.05701*, 2023.
- [13] Daniel Bogdolla, Yitian Yang, and J Marius Zöllner. Muvo: A multi-modal generative world model for autonomous driving with geometric representations. *arXiv preprint arXiv:2311.11762*, 2023.
- [14] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
- [15] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.
- [16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

- [17] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.
- [18] Liangliang Cao, Bowen Zhang, Chen Chen, Yinfei Yang, Xianzhi Du, Wencong Zhang, Zhiyun Lu, and Yantao Zheng. Less is more: Removing text-regions improves clip training efficiency and robustness. *arXiv preprint arXiv:2305.05095*, 2023.
- [19] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [20] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*, 2023.
- [21] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with ilms: Fusing object-level vector modality for explainable autonomous driving. *arXiv preprint arXiv:2310.01957*, 2023.
- [22] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuanan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.
- [23] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [24] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.
- [25] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, pages 1–16, 2023.
- [26] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [27] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelenext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [28] Shupeng Cheng, Ge-Peng Ji, Pengda Qin, Deng-Ping Fan, Bowen Zhou, and Peng Xu. Large model based referring camouflaged object detection. *arXiv preprint arXiv:2311.17122*, 2023.
- [29] Wenhao Cheng, Junbo Yin, Wei Li, Ruigang Yang, and Jianbing Shen. Language-guided 3d object detection in point cloud for autonomous driving. *arXiv preprint arXiv:2305.15765*, 2023.
- [30] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [31] Dongmin Choi, Wonwoo Cho, Kangyeol Kim, and Jaegul Choo. idet3d: Towards efficient interactive object detection for lidar point clouds. *arXiv preprint arXiv:2312.15449*, 2023.
- [32] Glen Chou, Yunus Emre Sahin, Liren Yang, Kwesi J. Rutledge, Petter Nilsson, and Necmiye Ozay. Using control synthesis to generate corner cases: A case study on autonomous driving. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(11):2906–2917, November 2018.
- [33] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [34] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. *arXiv preprint arXiv:2309.10228*, 2023.
- [35] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Receive, reason, and react: Drive as you say with large language models in autonomous vehicles. *arXiv preprint arXiv:2310.08034*, 2023.
- [36] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. *arXiv preprint arXiv:2311.12320*, 2023.
- [37] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. *arXiv preprint arXiv:2311.12320*, 2023.
- [38] Yiming Cui, Cheng Han, and Dongfang Liu. Cml-mots: Collaborative multi-task learning for multi-object tracking and segmentation. *arXiv preprint arXiv:2311.00987*, 2023.
- [39] W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructclip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2, 2023.
- [40] Sri Aditya Deevi, Connor Lee, Lu Gan, Sushruth Nagesh, Gaurav Pandey, and Soon-Jo Chung. Rgb-x object detection via scene-specific fusion modules. *arXiv preprint arXiv:2310.19372*, 2023.
- [41] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838*, 2019.
- [42] Vikrant Dewangan, Tushar Choudhary, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K Madhava Krishna. Talk2bev: Language-enhanced bird’s-eye view maps for autonomous driving. *arXiv preprint arXiv:2310.02251*, 2023.
- [43] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European Conference on Computer Vision*, pages 74–92. Springer, 2022.
- [44] Ning Ding, Ce Zhang, and Azim Eskandarian. Saliendet: A saliency-based feature enhancement algorithm for object detection for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2023.
- [45] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *arXiv preprint arXiv:2309.05186*, 2023.
- [46] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *arXiv preprint arXiv:2309.05186*, 2023.
- [47] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023.
- [48] Xingshuai Dong and Massimiliano L Cappuccio. Applications of computer vision in autonomous vehicles: Methods, challenges and future directions. *arXiv preprint arXiv:2311.09093*, 2023.
- [49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [50] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.
- [51] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- [52] Amine Elhafsi, Rohan Sinha, Christopher Agia, Edward Schmerling, Issa AD Nesnas, and Marco Pavone. Semantic anomaly detection with large language models. *Autonomous Robots*, 47(8):1035–1055, 2023.
- [53] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.
- [54] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinfeng Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.
- [55] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudio Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

- [56] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. *arXiv preprint arXiv:2307.07162*, 2023.
- [57] Haoxiang Gao, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*, 2024.
- [58] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [59] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.
- [60] Diandian Guo, Deng-Ping Fan, Tongyu Lu, Christos Sakaridis, and Luc Van Gool. Vanishing-point-guided video semantic segmentation of driving scenes. *arXiv preprint arXiv:2401.15261*, 2024.
- [61] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.
- [62] Lynn Vonder Haar, Timothy Elvira, Luke Newcomb, and Omar Ochoa. Measuring the impact of scene level objects on object detection: Towards quantitative explanations of detection decisions. *arXiv preprint arXiv:2401.10790*, 2024.
- [63] Niklas Hanselmann, Katrin Renz, Kashyap Chitta, Apratim Bhattacharya, and Andreas Geiger. King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In *European Conference on Computer Vision*, pages 335–352. Springer, 2022.
- [64] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [65] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [67] Quentin Herau, Nathan Piasco, Moussab Bennehar, Luis Roldão, Dzmitry Tsishkou, Cyrille Mignot, Pascal Vasseur, and Cédric Demonceaux. Soac: Spatio-temporal overlap-aware multi-sensor calibration using neural radiance fields. *arXiv preprint arXiv:2311.15803*, 2023.
- [68] Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, and Lennart Svensson. Lidarclip or: How i learned to talk to point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7438–7447, 2024.
- [69] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.
- [70] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [71] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhui Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.
- [72] Haiwen Huang, Andreas Geiger, and Dan Zhang. Good: Exploring geometric cues for detecting objects in an open world. *arXiv preprint arXiv:2212.11720*, 2022.
- [73] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.
- [74] Keishi Ishihara, Anssi Kanervisto, Jun Miura, and Ville Hautamaki. Multi-task learning with attention for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2902–2911, 2021.
- [75] Bernhard Jaeger, Kashyap Chitta, and Andreas Geiger. Hidden biases of end-to-end driving models. *arXiv preprint arXiv:2306.07957*, 2023.
- [76] Ashesh Jain, Luca Del Pero, Hugo Grimmett, and Peter Ondruska. Autonomy 2.0: Why is self-driving always 5 years away? *arXiv preprint arXiv:2107.08142*, 2021.
- [77] Kanishk Jain, Varun Chhangani, Amogh Tiwari, K Madhava Krishna, and Vineet Gandhi. Ground then navigate: Language-guided navigation in dynamic scenes. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4113–4120. IEEE, 2023.
- [78] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- [79] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.
- [80] Feng Jiang, Chaoping Tu, Gang Zhang, Jun Li, Hanqing Huang, Junyu Lin, Di Feng, and Jian Pu. Revisiting multi-modal 3d semantic segmentation in real-world autonomous driving. *arXiv preprint arXiv:2310.08826*, 2023.
- [81] Yang Jiao, Zequn Jie, Shaoxiang Chen, Lechao Cheng, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Instance-aware multi-camera 3d object detection with structural priors mining and self-boosting learning. *arXiv preprint arXiv:2312.08004*, 2023.
- [82] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. Adapt: Action-aware driving caption transformer. *arXiv preprint arXiv:2302.00673*, 2023.
- [83] Takayuki Kanai, Igor Vasiljevic, Vitor Guizilini, Adrién Gaidon, and Rares Ambrus. Robust self-supervised extrinsic self-calibration. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1932–1939. IEEE, 2023.
- [84] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.
- [85] Abdul Hannan Khan, Syed Tahseen Raza Rizvi, and Andreas Dengel. Real-time traffic object detection for autonomous driving, 2024.
- [86] Ajinkya Khoche, Laura Pereira Sánchez, Nazre Batool, Sina Sharif Mansouri, and Patric Jensfelt. Fully sparse long range 3d object detection using range experts and multimodal virtual points. *arXiv preprint arXiv:2310.04800*, 2023.
- [87] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018.
- [88] Genki Kinoshita and Ko Nishino. Camera height doesn't change: Unsupervised monocular scale-aware road-scene depth estimation. *arXiv preprint arXiv:2312.04530*, 2023.
- [89] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
- [90] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [91] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.
- [92] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.
- [93] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [94] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022.
- [95] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5567–5577, 2023.
- [96] Xin Li, Yeqi Bai, Pinlong Cai, Licheng Wen, Daocheng Fu, Bo Zhang, Xuemeng Yang, Xinyu Cai, Tao Ma, Jianfei Guo, et al. Towards knowledge-driven autonomous driving. *arXiv preprint arXiv:2312.04316*, 2023.

- [97] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.
- [98] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.
- [99] Zhiqi Li, Wenhui Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022.
- [100] Zixuan Li, Haiying Lin, Zhangyu Wang, Huazhi Li, Miao Yu, and Jie Wang. A ground segmentation method based on point cloud map for unstructured roads. *arXiv preprint arXiv:2309.08164*, 2023.
- [101] Weihao Liang, Lu Gan, Pengfei Wang, and Wei Meng. Brain-inspired domain-incremental adaptive detection for autonomous driving. *Frontiers in neurorobotics*, 16:916808, 2022.
- [102] Wenqi Liang, Gan Sun, Chenxi Liu, Jiahua Dong, and Kangru Wang. I3dod: Towards incremental 3d object detection via prompting. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5738–5743. IEEE, 2023.
- [103] Xiwen Liang, Minzhe Niu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Visual exemplar driven task-prompting for unified perception in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9611–9621, 2023.
- [104] Xiwen Liang, Yangxin Wu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Effective adaptation in multi-task co-training for unified autonomous driving. *Advances in Neural Information Processing Systems*, 35:19645–19658, 2022.
- [105] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019.
- [106] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [107] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [108] Mengyi Liu, Jie Jiang, Chao Zhu, and Xu-Cheng Yin. Vlpd: Context-aware pedestrian detection via vision-language semantic self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6662–6671, 2023.
- [109] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.
- [110] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [111] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? *arXiv preprint arXiv:2309.01809*, 2023.
- [112] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939, 2018.
- [113] Pingchuan Ma, Tao Du, and Wojciech Matusik. Efficient continuous pareto exploration in multi-task learning. In *International Conference on Machine Learning*, pages 6522–6531. PMLR, 2020.
- [114] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. *arXiv preprint arXiv:2312.00438*, 2023.
- [115] Yunsheng Ma, Can Cui, Xu Cao, Wenqian Ye, Peiran Liu, Juanwu Lu, Amr Abdelraouf, Rohit Gupta, Kyungtae Han, Aniket Bera, et al. Lampilot: An open benchmark dataset for autonomous driving with language model programs. *arXiv preprint arXiv:2312.04372*, 2023.
- [116] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1043–1052, 2023.
- [117] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023.
- [118] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*, 2023.
- [119] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.
- [120] Albert Mehrabian and James A Russell. *An approach to environmental psychology*. the MIT Press, 1974.
- [121] Erkan Milli, Özgür Erkent, and Asım Egemen Yılmaz. Multi-modal multi-task (3mt) road segmentation. *IEEE Robotics and Automation Letters*, 8(9):5408–5415, September 2023.
- [122] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Uniworld: Autonomous driving pre-training via world models. *arXiv preprint arXiv:2308.07234*, 2023.
- [123] M Jehanzeb Mirza, Marc Masana, Horst Possegger, and Horst Bischof. An efficient domain-incremental learning approach to drive in all weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2022.
- [124] Sambit Mohapatra, Senthil Yogamani, Varun Ravi Kumar, Stefan Milz, Heinrich Gotzig, and Patrick Mäder. Lidar-bevmt: Real-time lidar bird's-eye view multi-task perception network for autonomous driving. *arXiv preprint arXiv:2307.08850*, 2023.
- [125] Sonia Moshfeghi, Muhammad Tanveer Jan, Joshua Conniff, Seyedeh Gol Ara Ghoreishi, Jinwoo Jang, Borko Furht, Kwangsoo Yang, Monica Rosselli, David Newman, Ruth Tappen, et al. In-vehicle sensing and data analysis for older drivers with mild cognitive impairment. In *2023 IEEE 20th International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 140–145. IEEE, 2023.
- [126] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8602–8612, 2023.
- [127] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. *arXiv preprint arXiv:2312.03661*, 2023.
- [128] Minxue Niu, Zhaobo Zheng, Kumar Akash, and Teruhisa Misu. Beyond empirical windowing: An attention-based approach for trust prediction in autonomous vehicles. Dec 2023.
- [129] OpenAI OpenAI. Gpt-4 technical report. 2023.
- [130] TB OpenAI. Chatgpt: Optimizing language models for dialogue. openai, 2022.
- [131] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [132] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.
- [133] Sungjune Park, Hyunjoo Kim, and Yong Man Ro. Incorporating language-driven appearance knowledge units with visual cues in pedestrian detection. *arXiv preprint arXiv:2311.01025*, 2023.
- [134] Haoyang Peng, Baopu Li, Bo Zhang, Xin Chen, Tao Chen, and Hongyuan Zhu. Multi-view vision-prompt fusion network: Can 2d pre-trained model boost 3d point cloud data-scarce learning? *arXiv preprint arXiv:2304.10224*, 2023.
- [135] Xidong Peng, Runnan Chen, Feng Qiao, Lingdong Kong, Youquan Liu, Tai Wang, Xinge Zhu, and Yuexin Ma. Learning to adapt sam for segmenting cross-domain point clouds, 2023.
- [136] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.
- [137] Rudra PK Poudel, Harit Pandya, Stephan Liwicki, and Roberto Cipolla. Recore: Regularized contrastive representation learning of word model. *arXiv preprint arXiv:2312.09056*, 2023.
- [138] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021.

- [139] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*, 2023.
- [140] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [141] Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, and Barbara Caputo. Unmasking anomalies in road-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4037–4046, 2023.
- [142] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- [143] Hongping Ren, Hui Gao, He Chen, and Guangzhen Liu. A survey of autonomous driving scenarios and scenario databases. In *2022 9th International Conference on Dependable Systems and Their Applications (DSA)*, pages 754–762. IEEE, 2022.
- [144] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
- [145] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7513–7522, 2024.
- [146] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.
- [147] Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempe: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*, 2023.
- [148] Hao Shao, Yuxuan Hu, Letian Wang, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488*, 2023.
- [149] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, et al. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35:15558–15573, 2022.
- [150] Hao Shi, Chengshan Pang, Jiaming Zhang, Kailun Yang, Yuhao Wu, Huajian Ni, Yining Lin, Rainer Stiefelhagen, and Kaiwei Wang. Cobev: Elevating roadside 3d object detection with depth and height complementarity. *arXiv preprint arXiv:2310.02815*, 2023.
- [151] Sai Shubodh, Mohammad Osama, Husain Zaidi, Udit Singh Parihar, and Madhava Krishna. Lip-loc: Lidar image pretraining for cross-modal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–957, 2024.
- [152] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengan Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023.
- [153] Apoorv Singh. End-to-end autonomous driving using deep learning: A systematic review. *arXiv preprint arXiv:2311.18636*, 2023.
- [154] Eduardo Soares, Plamen Angelov, Bruno Costa, and Marcos Castro. Actively semi-supervised deep rule-based classifier applied to adverse driving scenarios. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.
- [155] Ziying Song, Guoxin Zhang, Jun Xie, Lin Liu, Caiyan Jia, Shaoqing Xu, and Zhepeng Wang. Voxelnextfusion: A simple, unified and effective voxel fusion framework for multi-modal 3d object detection. *arXiv preprint arXiv:2401.02702*, 2024.
- [156] Ziying Song, Guoxing Zhang, Lin Liu, Lei Yang, Shaoqing Xu, Caiyan Jia, Feiyang Jia, and Li Wang. Robofusion: Towards robust multi-modal 3d object detection via sam. *arXiv preprint arXiv:2401.03907*, 2024.
- [157] Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10400–10409, 2021.
- [158] Lanyue Tang, Y Li, Jinghui Yuan, A Fu, and J Sun. Cpsor-gcn: A vehicle trajectory prediction method powered by emotion and cognitive theory. *arXiv preprint arXiv:2311.08086*, 2023.
- [159] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 1013–1020. IEEE, 2018.
- [160] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [161] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almalhairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [162] David Unger, Nikhil Gosala, Varun Ravi Kumar, Shubhankar Borse, Abhinav Valada, and Senthil Yogamani. Multi-camera bird’s eye view perception for autonomous driving. *arXiv preprint arXiv:2309.09080*, 2023.
- [163] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [164] Mathijs R van Geerenstein, Felicia Ruppel, Klaus Dietmayer, and Dariu M Gavrila. Multimodal object query initialization for 3d object detection. *arXiv preprint arXiv:2310.10353*, 2023.
- [165] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [166] Kasi Viswanath, Peng Jiang, Sujit PB, and Srikanth Saripalli. Off-road lidar intensity based semantic segmentation. *arXiv preprint arXiv:2401.01439*, 2024.
- [167] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.
- [168] Jiayuan Wang, QM Wu, and Ning Zhang. You only look at once for real-time and generic multi-task. *arXiv preprint arXiv:2310.01641*, 2023.
- [169] Pengjin Wang, Meixin Zhu, Hongliang Lu, Hui Zhong, Xianda Chen, Shaojie Shen, Xuesong Wang, and Yinhai Wang. Bevgpt: Generative pre-trained large model for autonomous driving prediction, decision-making, and planning. *arXiv preprint arXiv:2310.10357*, 2023.
- [170] Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multimodal foundation models. *arXiv preprint arXiv:2310.17642*, 2023.
- [171] Wenhui Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023.
- [172] Wenhui Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023.
- [173] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.
- [174] Yixuan Wang, Ruochen Jiao, Chengtian Lang, Sinong Simon Zhan, Chao Huang, Zhaoran Wang, Zhuoran Yang, and Qi Zhu. Empowering autonomous driving with large language models: A safety perspective. *arXiv preprint arXiv:2312.00812*, 2023.
- [175] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panocc: Unified occupancy representation for camera-based 3d panoptic segmentation. *arXiv preprint arXiv:2306.10013*, 2023.
- [176] Wayve. Lingo-1: Exploring natural language for autonomous driving. <https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/>, 2023.
- [177] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.
- [178] Dafeng Wei, Tian Gao, Zhengyu Jia, Changwei Cai, Chengkai Hou, Peng Jia, Fu Liu, Kun Zhan, Jingchen Fan, Yixing Zhao, et al. Bevclip: Multi-modal bev retrieval methodology for complex scene in autonomous driving. *arXiv preprint arXiv:2401.01065*, 2024.
- [179] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-

- driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023.
- [180] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*, 2023.
- [181] What Makes In-Context Learning Work. Rethinking the role of demonstrations: What makes in-context learning work?
- [182] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.
- [183] Dong Wu, Man-Wen Liao, Wei-Tian Zhang, Xing-Gang Wang, Xiang Bai, Wen-Qing Cheng, and Wen-Yu Liu. Yolop: You only look once for panoptic driving perception. *Machine Intelligence Research*, 19(6):550–562, 2022.
- [184] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14633–14642, 2023.
- [185] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *arXiv preprint arXiv:2309.04379*, 2023.
- [186] Yuxuan Xiao, Yao Li, Chengzhen Meng, Xingchen Li, and Yanyong Zhang. Calibformer: A transformer-based automatic lidar-camera calibration network. *arXiv preprint arXiv:2311.15241*, 2023.
- [187] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. *arXiv preprint arXiv:2304.14340*, 2023.
- [188] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.
- [189] Qinghua Xu, Tao Yue, Shaukat Ali, and Maite Arratibel. Pretrain, prompt, and transfer: Evolving digital twins for time-to-event analysis in cyber-physical systems. *arXiv preprint arXiv:2310.00032*, 2023.
- [190] Xiang Xu, Lingdong Kong, Hui Shuai, and Qingshan Liu. Frnet: Frustum-range networks for scalable lidar segmentation. *arXiv preprint arXiv:2312.04484*, 2023.
- [191] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023.
- [192] Xu Yan, Haiming Zhang, Yingjie Cai, Jingming Guo, Weichao Qiu, Bin Gao, Kaiqiang Zhou, Yue Zhao, Huan Jin, Jiantao Gao, et al. Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities. *arXiv preprint arXiv:2401.08045*, 2024.
- [193] Chuang Yang, Kai Zhuang, Mulin Chen, Haozhao Ma, Xu Han, Tao Han, Changxing Guo, Han Han, Bingxuan Zhao, and Qi Wang. Traffic sign interpretation in real road scene. *arXiv preprint arXiv:2311.10793*, 2023.
- [194] Chuang Yang, Kai Zhuang, Mulin Chen, Haozhao Ma, Xu Han, Tao Han, Changxing Guo, Han Han, Bingxuan Zhao, and Qi Wang. Traffic sign interpretation in real road scene. *arXiv preprint arXiv:2311.10793*, 2023.
- [195] Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, et al. Unipad: A universal pre-training paradigm for autonomous driving. *arXiv preprint arXiv:2310.08370*, 2023.
- [196] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023.
- [197] Yuxue Yang, Lue Fan, and Zhaoxiang Zhang. Mixsup: Mixed-grained supervision for label-efficient lidar-based 3d object detection. *arXiv preprint arXiv:2401.16305*, 2024.
- [198] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023.
- [199] Zhengyuan Yang, Yixuan Zhang, Jerry Yu, Junjie Cai, and Jiebo Luo. End-to-end multi-modal multi-task vehicle control for self-driving cars with visual perceptions. In *2018 24th international conference on pattern recognition (ICPR)*, pages 2289–2294. IEEE, 2018.
- [200] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*, 2023.
- [201] Hanrong Ye and Dan Xu. Taskprompter: Spatial-channel multi-task prompting for dense scene understanding. In *ICLR*, 2023.
- [202] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.
- [203] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *arXiv preprint arXiv:2305.18279*, 2023.
- [204] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- [205] Hu Zhang, Jianhua Xu, Tao Tang, Haiyang Sun, Xin Yu, Zi Huang, and Kaicheng Yu. Opensight: A simple open-vocabulary framework for lidar-based object detection. *arXiv preprint arXiv:2312.08876*, 2023.
- [206] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.
- [207] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.
- [208] Siyao Zhang, Daocheng Fu, Zhao Zhang, Bin Yu, and Pinlong Cai. Trafficgpt: Viewing, processing and interacting with traffic foundation models. *arXiv preprint arXiv:2309.06719*, 2023.
- [209] Tingru Zhang, Alan HS Chan, Yutao Ba, and Wei Zhang. Situational driving anger, driving performance and allocation of visual attention. *Transportation research part F: traffic psychology and behaviour*, 42:376–388, 2016.
- [210] Xinyu Zhang, Jiahui Chen, Junkun Yuan, Qiang Chen, Jian Wang, Xiaodi Wang, Shumin Han, Xiaokang Chen, Jimin Pi, Kun Yao, et al. Cae v2: Context autoencoder with clip target. *arXiv preprint arXiv:2211.09799*, 2022.
- [211] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.
- [212] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. TrafficBots: Towards world models for autonomous driving simulation and motion prediction. In *International Conference on Robotics and Automation (ICRA)*, 2023.
- [213] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.
- [214] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023.
- [215] Haibin Zhou, Jun Chang, Tao Lu, and Huabing Zhou. 3d lane detection from front or surround-view using joint-modeling & matching. *arXiv preprint arXiv:2401.08036*, 2024.
- [216] Sifan Zhou, Liang Li, Xinyu Zhang, Bo Zhang, Shipeng Bai, Miao Sun, Ziyu Zhao, Xiaobo Lu, and Xiangxiang Chu. Lidar-ptq: Post-training quantization for point cloud 3d object detection. *arXiv preprint arXiv:2401.15865*, 2024.
- [217] Xingcheng Zhou, Mingyu Liu, Bare Luka Zagar, Ekim Yurtsever, and Alois C Knoll. Vision language models in autonomous driving and intelligent transportation systems. *arXiv preprint arXiv:2310.14414*, 2023.
- [218] Yijie Zhou, Likun Cai, Xianhui Cheng, Zhongxue Gan, Xiangyang Xue, and Wenchao Ding. Openannotate3d: Open-vocabulary auto-labeling system for multi-modal 3d data. *arXiv preprint arXiv:2310.13398*, 2023.
- [219] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [220] Shihao Zou, Xianying Huang, and Xudong Shen. Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5994–6003, 2023.