A Survey for Foundation Models in Autonomous Driving

Haoxiang Gao[1] , Yaqian Li[2] , Kaiwen Long[2] , Ming Yang[3] , Yiqing Shen[4]

[1]Independent Researcher [2]Li Auto Inc. [3]Shanghai Jiao Tong University , [4]Johns Hopkins University

haoxianggao@gmail.com, {liyaqian,longkaiwen}@lixiang.com, mingyang@sjtu.edu.cn, yshen92@jhu.edu

Abstract

The advent of foundation models has revolutionized the fields of natural language processing and computer vision, paving the way for their application in autonomous driving (AD). This survey presents a comprehensive review of more than 40 research papers, demonstrating the role of foundation models in enhancing AD. Large language models contribute to planning and simulation in AD, particularly through their proficiency in reasoning, code generation and translation. In parallel, vision foundation models are increasingly adapted for critical tasks such as 3D object detection and tracking, as well as creating realistic driving scenarios for simulation and testing. Multi-modal foundation models, integrating diverse inputs, exhibit exceptional visual understanding and spatial reasoning, crucial for end-to-end AD. This survey not only provides a structured taxonomy, categorizing foundation models based on their modalities and functionalities within the AD domain but also delves into the methods employed in current research. It identifies the gaps between existing foundation models and cutting-edge AD approaches, thereby charting future research directions and proposing a roadmap for bridging these gaps.

1 Introduction

The integration of deep learning (DL) into autonomous driving (AD) has marked a significant leap in this field, attracting attention from both academic and industrial spheres. AD systems, equipped with cameras and lidars, mimic human-like decision-making processes. These systems are fundamentally composed of three key components: perception, prediction, and planning. Perception, utilizing DL and computer vision algorithms, focuses on object detection and tracking. Prediction forecasts the behavior of traffic agents and their interaction with autonomous vehicles. Planning, which is typically structured hierarchically, involves making strategic driving decisions, calculating optimal trajectories, and executing vehicle control commands. The advent of foundation models, particularly renowned in natural language processing and computer vision, has introduced new dimensions to AD research. These models are distinct due to their training on extensive web-scale datasets and their massive parameter sizes. Given the vast amounts of data generated by autonomous vehicle services and advancements in AI, including NLP and AI-generated content (AIGC), there is a growing curiosity about the potential of foundation models in AD. These models could be instrumental in performing a range of AD tasks, such as object detection, scene understanding, and decision-making, with a level of intelligence akin to human drivers.

Foundation models address several challenges in AD. Conventionally, AD models are trained in a supervised manner, dependent on manually annotated data that often lack diversity, limiting their adaptability. Foundation models, however, show superior generalization capabilities due to their training on diverse, web-scale data. They can potentially replace the complex heuristic rule-based systems in planning with their reasoning capabilities and knowledge derived from extensive pre-training. For example, LLM has reasoning capability and common sense driving knowledge acquired from the pre-training dataset, which can potentially replace heuristic rule-based planning systems, which require complex engineering effort of hand-crafted rules in software codes and debugging on corner cases. Generative models within this domain can create realistic traffic scenarios for simulation, essential for testing safety and reliability in rare or challenging situations. Moreover, foundation models contribute to making AD technology more user-centric, with language models understanding and executing user commands in natural language.

Despite considerable research in applying foundation models to AD, there are notable limitations and gaps in real-world application. Our survey aims to provide a systematic review and propose future research directions. There are two surveys related to foundation models for autonomous driving: LLM4Drive[Yang *et al.*, 2023c] is more focused on large language models. [Huang *et al.*, 2023] has a good breadth of summary of applications of foundation models in autonomous driving, mainly in simulation, data annotation, and planning. We expand upon existing surveys by covering vision foundation models and multi-modal foundation models, analyzing their applications in prediction and perception tasks. This comprehensive approach includes detailed examinations of technical aspects, such as pre-trained models and methods, and identifies future research opportunities. Innova-

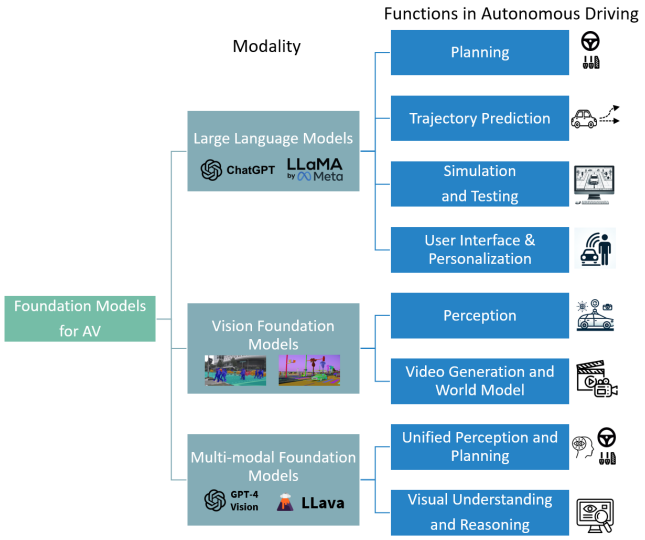

Figure 1: Taxonomy of foundation models for autonomous driving. It delineates the categorization of foundation models according to their modalities, such as Large Language Models, Vision Foundation Models, and Multi-modal Foundation Models, and correlates them with their respective functions in autonomous driving.

tively, we propose a taxonomy categorizing foundation models in AD based on modalities and functions, as shown in Figure 1. The following sections will explore the application of various foundation models, including large language models, vision foundation models, and multi-modal foundation models, in the context of AD.

2 Large Language Models in AD

2.1 Overview

LLMs, originally transformative in NLP, are now driving innovations in AD. Bidirectional Transformers (BERT) [Devlin *et al.*, 2018] pioneered foundation models in NLP, leveraging Transformer architecture for understanding language semantics. This pre-trained model can be fine-tuned on specific data-sets, and achieve state-of-the-art results in a wide range of tasks. Following this, OpenAI's generative pre-trained transformer (GPT) series [Radford *et al.*, 2018], including GPT-4, demonstrated remarkable NLP capabilities, attributed to training on extensive datasets. Later GPT models, including ChatGPT, GPT-4[Achiam *et al.*, 2023] are trained using billions of parameters and crawled web data with trillions of words, and achieve strong performance on many NLP tasks, including translation, text summarization, question-answering. It also demonstrates one-shot and few-shot reasoning capabilities to learn new skills from the context. More and more researchers have started to apply these reasoning, understanding, and in-context learning capabilities to address challenges in AD.

2.2 Applications in AD

Reasoning and Planning

The decision-making process in AD closely parallels human reasoning, necessitating the interpretation of environmental

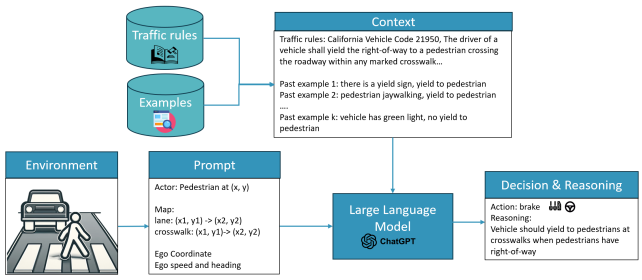

Figure 2: Common pattern of LLM pipelines for autonomous driving, showcasing the integration of textual environment descriptions and LLM reasoning to inform driving decisions.

cues to make safe and comfortable driving decisions. LLMs, through their training on diverse web data, have assimilated common-sense knowledge pertinent to driving, drawing from a plethora of sources including web forums and official government websites. This wealth of information enables LLMs to engage in the nuanced decision-making required for AD. One method for harnessing LLMs in AD involves presenting them with detailed textual descriptions of the driving environment, prompting them to propose driving decisions or control commands. This process, as illustrated in Figure 2, typically encompasses comprehensive prompts detailing agent states, such as coordinates, speed, and past trajectories, the vehicle's state *i.e*, velocity, and acceleration, and map specifics including traffic lights, lane information, and intended route). For enhanced interaction understanding, LLMs can also be directed to provide reasoning along with their responses. For instance, the GPT driver [Mao *et al.*, 2023a] not only recommends vehicle actions but also elucidates the rationale behind these suggestions, significantly enhancing the transparency and explainability of autonomous driving decisions. This approach, exemplified by Driving with LLMs [Chen *et al.*, 2023], enhances the explainability of autonomous driving decisions. Similarly, the "Receive, Reason, and React" approach [Cui *et al.*, 2023] instructs LLM agents to assess lane occupancy and evaluate the safety of potential actions, thereby fostering a deeper comprehension of dynamic driving scenarios. These methods not only leverage LLMs' inherent ability to understand complex scenarios but also employ their reasoning capabilities to simulate human-like decision-making processes. Through the integration of detailed environmental descriptions and strategic prompts, LLMs contribute significantly to the planning and reasoning aspects of AD, offering insights and decisions that mirror human judgment and expertise.

Prediction

Prediction forecasts traffic participants' future trajectories, intents, and possible interactions with the ego vehicle. The common deep learning-based models are based on rasterized or vector images of the traffic scene, which encode spatial information. However, it is still challenging to accurately predict highly interactive scenes, which requires reasoning and semantic information, for example, right-of-ways, vehicles' turning signals, and pedestrians' gestures. The text representation of the scene can provide more semantic context

and better leverage LLM's reasoning capability and common knowledge in the pre-training dataset.

[Keysan *et al.*, 2023] did an early exploration of LLM's power to make trajectory predictions. They convert the scene representation into text prompts, and use BERT[Devlin *et al.*, 2018] model to generate the text encoding, which is finally fused with image encoding to decode trajectory prediction. Their evaluation shows significant improvement compared with baselines only using image encoding or text encoding.

LG-Traj[Chib and Singh, 2024] leverages pre-trained LLM to generate motion cues for pedestrian trajectory prediction. Given the prompt and pedestrian coordinates, LLM is able to identify pedestrian motion patterns, such as linear motion, curved motion, and standing still. Combined with the Transformer-based motion encoder and social decoder, their method achieved the best results in multiple data sets compared with other state-of-the-art methods.

iMotion-LLM[Felemban *et al.*, 2024] proposes a novel motion prediction method leveraging instructions and captions in text. Given the text instruction inputs and scene encoding projected into text space, LLM decodes the output captions describing the driving decisions and latent tokens to generate multi-modal trajectories. Their model significantly outperforms the baseline model without using LLM. Additionally, they propose a feasibility classification task, which aligns predicted trajectories with feasible instructions and reject infeasible ones to improve safety.

User Interface and Personalization

Autonomous vehicles should be user-friendly and able to follow the instructions from passengers or remote operators. The current Robotaxi remote assistance interface is only used to execute a limited set of pre-defined commands. However, LLM's understanding and interaction capabilities make it possible to let autonomous driving cars to understand human's free-form instruction to better control the autonomous vehicle and satisfy users' personalized requirements. [Cui *et al.*, 2023] explores the LLM-based planner conditioning on personalized commands, e.g. "driving aggressively" or "conservatively", and was able to output actions of various speeds and riskiness. [Yang *et al.*, 2023b] leverages LLM's reasoning abilities, and provides step-by-step rules to decide on response to user commands. The LLM agent is also able to accept or reject user commands based on pre-defined traffic rules and system requirements.

Simulation and Testing

LLM can summarize and extract knowledge from existing text data and generate new content, which can facilitate simulation and testing. The ADEPT system[Wang *et al.*, 2022b] uses GPT to extract key information from NHTSA accident reports using QA approach, and was able to generate diverse scene code used for simulation and testing. TARGET[Deng *et al.*, 2023] system is able to use GPT to translate traffic rules from the natural language to the domain-specific language, which is used for generating testing scenarios. LCTGen[Tan *et al.*, 2023] uses LLM as a powerful interpreter translating user's text query into structured specifications of map lanes and vehicle locations for traffic simulation scenarios.

2.3 Methods and Techniques

Researchers use similar techniques in natural language processing to utilize LLM for autonomous driving tasks, such as prompt engineering, in-context and few-shot learning, and reinforcement learning from human feedback[Ouyang *et al.*, 2022].

Prompt Engineering

Prompt engineering adopts sophisticated designs of input prompts and questions to guide the Large Language Model to generate our desired answers.

Some papers add traffic rules as pre-prompt to make LLM agent law-compliant. Driving with LLMs[Chen *et al.*, 2023] has diving rules covering aspects like traffic light transition and left or right driving side. [Mao *et al.*, 2023b] proposes a module called common-sense module, which stores the rules and instructions for human driving, for example, avoiding collision, and maintaining safety distances.

LanguageMPC[Sha *et al.*, 2023] adopts a top-down decision-making system: given different situations, the vehicle has different possible actions. LLM agent is also instructed to identify important agents in the scenario and output attention, weight, and bias matrices to select from predefined actions.

Memory modules are also introduced in some papers, which store past driving scenarios. At inference time, the relevant examples are retrieved and added as the context in the prompt, and LLM agent can better leverage few-shot learning capabilities and reflect on the most relevant scenarios. DILU[Wen *et al.*, 2023a] proposes a memory module, which stores text descriptions of driving scenarios in the vector database, and the system can retrieve top-k scenarios for few-shot learning. [Mao *et al.*, 2023b] has a two-stage retrieval process: the first stage uses k-nearest-neighbor search to retrieve relevant past examples in the database, and the second stage asks LLM to rank these examples.

More papers built complex systems to manage tasks in the prompt generation, which trigger function calls to other modules or sub-systems to obtain required information for decision-making. [Mao *et al.*, 2023b] has created libraries and function API calls to interact with perception, prediction, and mapping systems so that the LLM can fully leverage all available information. LanguageMPC [Sha *et al.*, 2023] uses LangChain to create tools and interfaces needed by LLM to get relevant vehicles, possible situations, and available actions.

Fine-tuning v.s. In-context Learning

Fine-tuning and in-context learning are both applied to adapt pre-trained models to autonomous driving. Fine-tuning re-trains the model parameters on smaller domain-specific datasets, while in-context learning or few-shot learning leverages LLM's knowledge and reasoning ability to learn from given examples in the input prompt. Most papers are focused on in-context learning, but only a few papers utilize fine-tuning. Researchers have mixed results on which one is the better: [Mao *et al.*, 2023b] compared both approaches and found that few-shot learning is slightly more effective. GPT-Driver [Mao *et al.*, 2023a] has a different conclusion

that using OpenAI fine-tuning performs significantly better than few-shot learning. [Chen *et al.*, 2023] also compared training from scratch and fine-tuning approaches, and found using the pre-trained LLaMA model with LoRA-based fine-tuning can perform better than training from scratch.

Reinforcement Learning and Human Feedback

DILU [Wen *et al.*, 2023a] proposes reflection modules, which store good driving examples and bad driving examples with human corrections to enhance its reasoning capabilities further. In this way, the LLM can learn to reason about what action is safe and unsafe and continuously reflect on a large amount of past driving experiences. Surreal Driver [Jin *et al.*, 2023] interviewed 24 drivers and used their descriptions of driving behavior as chain-of-thought prompts to develop a 'coach agent' module, which can instruct the LLM model to have a human-like driving style. Incorporating Voice Instructions [Wang *et al.*, 2022a] uses instructions from human coaches, and builds a taxonomy of natural language instructions of action, reward, and reasoning, which are used to train a deep reinforcement learning-based autonomous driving agent.

2.4 Limitations and Future Directions

Hallucination and Harmfulness

Hallucination is a big challenge in LLM, and state-of-the-art large language models can still produce misleading and false information. Most methods proposed in existing papers still require parsing driving actions from LLM's response. When given an unseen scenario, the LLM model can still produce unhelpful or wrong driving decisions. Autonomous driving is a safety-critical application, which has much higher reliability and safety requirements than chat-bots. According to evaluation result [Mao *et al.*, 2023a], the LLM model for autonomous driving has a 0.44% collision rate, higher than other methods. [Chen *et al.*, 2023] proposes a method to reduce hallucination by asking questions without enough information to make decisions, and instructs LLM to answer "I don't know". The pre-trained LLM may also include harmful content, for example, aggressive driving and speeding. More human-in-the-loop training and alignment can reduce hallucinations and harmful driving decisions.

Latency and Efficiency

Large language models often suffer from high latency, and generating detailed driving decisions can exhaust the latency budget of limited compute resources in the car. It takes several seconds for inference according to [Jin *et al.*, 2023]. LLMs with billions of parameters can consume over 100GB of memory, which might interfere with other critical modules in autonomous driving vehicles. More research needs to be done in this domain, such as model compression and knowledge distillation, to make LLM more efficient and easier for deployment.

Dependency on Perception System

Despite the supreme reasoning capability of LLM, the environment description still depends on the upstream perception module. The driving decisions can go wrong and cause critical accidents with minor errors in environmental inputs. For example, [Mao *et al.*, 2023b] shows failure cases when upstream heading data has errors. LLM also needs to be better adapted to perception models and make better decisions when there are errors and uncertainty.

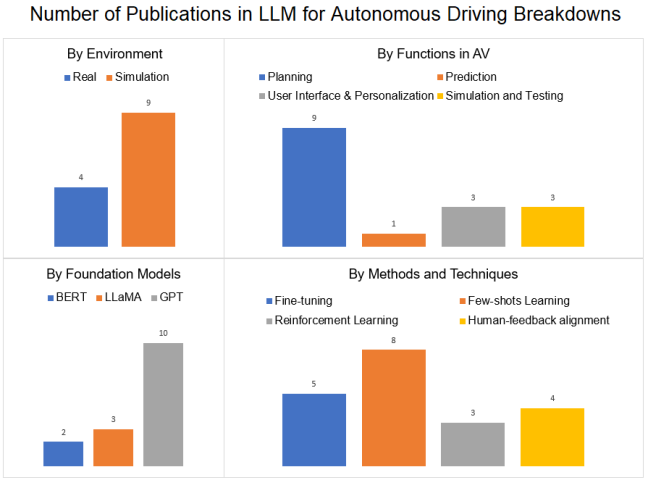

Figure 3: Summary of publications in LLM for autonomous driving.

Sim to Real Gap

Most of the research is done in simulated environments, and driving scenarios are much simpler than real-world environments. A lot of engineering and human detailed annotation efforts are needed for prompt engineering to cover all scenarios in the real world, for example, the model knows how to yield to humans, but is probably not good at handling interaction with small animals.

2.5 Summary

The publications in LLM are summarized in Figure 3. We propose more fine-grained classifications by environments(real or sim), functions in autonomous driving, foundation models, and techniques used in the research.

3 Vision Foundation Model

Vision foundation models have achieved great success in multiple computer vision tasks, such as object detection and segmentation. DINO [Caron *et al.*, 2021] uses vision-transformer architecture, and is trained in a self-supervised manner, predicting global image features given local image patches. DINOV2[Oquab *et al.*, 2023] scales the training with one billion parameters and a diversely curated dataset of 1.2 billion images and achieves state-of-the-art results in multiple tasks. Segment-anything model[Kirillov *et al.*, 2023] is a foundation model for image-segmentation. The model is trained with different types of prompts (points, boxes, or texts) to generate segmentation masks. Trained with billions of segmentation masks in the dataset, the model shows zero-shot transfer capability to segment new objects given the appropriate prompt.

Diffusion model[Sohl-Dickstein *et al.*, 2015] is a generative foundation model widely used for image generation. The

diffusion model iteratively adds noise to the image and applies a reverse diffusion process to restore the image. To generate the image, we can sample from the learned distribution and restore highly realistic images from random noises. Stable-Diffusion[Rombach *et al.*, 2022] model uses VAE[Kingma and Welling, 2013] to encode images to latent representation and use UNet[Ronneberger *et al.*, 2015] to decode from latent variable to pixel-wise images. It also has an optional text encoder and applies the cross-attention mechanism to generate images conditional on prompts (text description or other images). DALL-E[Ramesh *et al.*, 2021] model was trained with billions of image and text pairs and uses stable diffusion to generate high-fidelity images and creative arts following human instructions.

There is growing interest in the application of vision foundation models in autonomous driving, mainly for 3D perception and video generation tasks.

3.1 Perception

SAM3D[Zhang *et al.*, 2023a] applies SAM(Segment-anything model) to 3D object detection in autonomous driving. Lidar point clouds are projected to BEV(bird-eye-view) images, and it uses 32x32 mesh grids to generate point prompts to detect masks for foreground objects. It leverages the SAM model's zero-shot transfer capability to generate segmentation masks and 2D boxes. Then it uses vertical attributes of those lidar points inside 2D boxes to generate 3D boxes. However, the Waymo Open Dataset evaluation shows the average-precision metrics are still far from existing state-of-the-art 3D object detection models. They observed that SAM trained foundation model can not handle those sparse and noisy points very well, and often results in false negatives for distant objects.

SAM is applied to domain adaptation for 3D segmentation tasks, leveraging the SAM model's feature space which contains more semantic information and generalization capability. [Peng *et al.*, 2023] proposes SAM-guided feature alignment, learning unified representation of 3D point cloud features from different domains. It uses the SAM feature extractor to generate the camera image's feature embedding and projects 3D point clouds into camera images to obtain SAM features. The training process optimizes the alignment loss so that 3D features from different domains have a unified representation in SAM's feature space. This approach achieves state-of-the-art performance in 3D segmentation in multiple domain switching datasets, e.g. different cities, weathers, and lidar devices.

SAM and Grounding-DINO[Liu *et al.*, 2023b] are used to create a unified segmentation and tracking framework leveraging temporal consistency between video frames[Cheng *et al.*, 2023]. Grounding-DINO is an open-set object detector that takes input from text descriptions of objects and outputs the corresponding bounding boxes. Given the text prompts of object classes related to autonomous driving, it can detect objects in video frames and generate bounding boxes of vehicles and pedestrians. SAM model further takes these boxes as prompts and generates segmentation masks for detected objects. The resulting masks of objects are then passed to the downstream tracker, which compares the masks from continuous frames to determine if there are new objects.

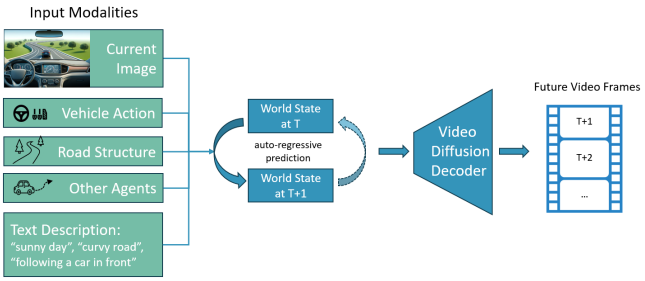

Figure 4: Video generation pipeline for AD.

3.2 Video Generation and World Model

The foundation models, especially generative models and world models can generate realistic virtual driving scenes, which can be used for autonomous driving simulation. Many researchers have started to apply diffusion models to autonomous driving for realistic scene generation. The video generation problem is often formulated as a world model: given the current world state, conditioning on environment input, the model predicts the next world state and uses diffusion to decode highly realistic driving scenes.

GAIA-1[Hu *et al.*, 2023] is developed by Wayve to generate realistic driving videos. The world model uses camera images, text descriptions, and vehicle control signals as input tokens and predicts the next frame. The paper uses pre-trained DINO[Caron *et al.*, 2021] model's embedding and cosine similarity loss to distill more semantic knowledge to image token embedding. They use the video diffusion model[Ho *et al.*, 2022] to decode high-fidelity driving scenes from the predicted image token. There are two separate tasks to train the diffusion model: image generation and video generation. The image generation task helps the decoder generate high-quality images, while the video generation task uses temporal attention to generate temporally consistent video frames. The generated video follows high-level real-world constraints and has realistic scene dynamics, such as the object's location, interactions, traffic rules, and road structures. The video also shows diversity and creativity, which have realistic possible outcomes conditioned on different text descriptions and the ego vehicle's action.

DriveDreamer[Wang *et al.*, 2023b] also uses the world model and diffusion model to generate video for autonomous driving. In addition to images, text descriptions, and vehicle actions, the model also uses more structural traffic information as input, such as HDMap and object 3D boxes, so that the model can better understand higher-level structural constraints of traffic scenes. The model training has two stages: the first stage is video generation using the diffusion model conditioned on structured traffic information. It was built on a pre-trained Stable-Diffusion model[Rombach *et al.*, 2022] with parameters frozen. In the second stage, the model is trained with both future video prediction tasks and action prediction tasks to better learn future prediction and interactions between objects.

[Zhang *et al.*, 2023c] built a point cloud-based world model

that achieves SOTA performance in point cloud forecasting tasks. They propose a VQVAE-like[Oord *et al.*, 2017] tokenizer to represent 3D point clouds as latent BEV tokens and use discrete diffusion to forecast future point clouds given the past BEV tokens and ego vehicle's actions tokens.

3.3 Limitations and Future Directions

The current state-of-the-art foundation model like SAM doesn't have good enough zero-shot transfer ability for 3D autonomous driving perception tasks, such as object detection, and segmentation. Autonomous driving perception relies on multiple cameras, lidars, and sensor fusions to obtain the highest accuracy object detection result, which is much different from image datasets randomly collected from the web. The scale of current public datasets for autonomous driving perception tasks is still not large enough to train a foundational model and cover all possible long-tail scenarios. Despite the limitation, the existing 2D vision foundation models can serve as useful feature extractors for knowledge distillation, which helps models better incorporate semantic information. In the domain of video generation and forecasting tasks, we have already seen promising progress leveraging existing diffusion models for video generation and point cloud forecasting, which can be further applied to creating high-fidelity scenarios for autonomous driving simulation and testing.

4 Multi-modal Foundation Models

Multi-modal foundation models benefit more by taking input data from multiple modalities, e.g. sounds, images, and video, to perform more complex tasks, e.g. generating text from images, analyzing and reasoning with visual inputs.

One of the most well-known multi-modal foundation models is CLIP[Radford *et al.*, 2021]. The model is pre-trained using the contrastive pre-training method. The inputs are noisy images and text pairs, and the model is trained to predict if the given image and text are a correct pair. The model is trained to maximize the cosine similarity of embedding from the image encoder and text encoder. The CLIP model shows zero-shot transfer ability for other computer vision tasks, such as image classification, and predicting the correct text description of the class without supervised training.

Multi-modal foundation models, like LLaVA[Liu *et al.*, 2023a], LISA[Lai *et al.*, 2023], and CogVLM[Wang *et al.*, 2023a] can be used for the general-purpose visual AI agent, which demonstrates superior performance in vision tasks, such as object segmentation, detection, localization, and spatial reasoning. Video-LLaMA[Zhang *et al.*, 2023b] can further perceive video and audio data, which may help autonomous vehicles better understand the world from temporal images and audio sequences.

Multi-modal foundation model is also used for robot learning, which leverages the robot's action as a new modality to create more general-purpose agents that can perform real-world tasks. DeepMind proposed a vision-language-action model[Brohan *et al.*, 2023] trained on text and images from the web and learned to output control commands to complete real-world object manipulation tasks.

Transferring general knowledge from large-scale pre-training datasets to autonomous driving, the multi-modal foundation models can be used for object detection, visual understanding, and spatial reasoning, which enables more powerful applications in autonomous driving.

4.1 Visual Understanding and Reasoning

Traditional object detection or classification models are not enough for autonomous driving, because we need better semantic understanding and visual reasoning of the scene, for example, identifying risky objects, and understanding the intents of traffic participants. Most of the existing deep learning-based prediction and planning models are dark-box models, which have poor explainability and debuggability when accidents or discomfort events happen. With the help of the multi-modal foundation models, we can generate explanations and the reasoning process of the model to better investigate the issues.

To further improve the perception system, HiLM-D[Ding *et al.*, 2023] utilizes multi-modal foundation models for ROLISP(Risk Object Localization and Intention and Suggestion Prediction). It uses natural language to identify risky objects from camera images and provide suggestions on the ego vehicle's actions. To overcome the drawback of missing small objects, it proposes a pipeline with both high-resolution and low-resolution branches. The low-resolution reasoning branch is used to understand high-level information and identify risk objects from continuous video frames; The high-resolution perception branch enables further refinement of object detection and localization quality. Their model backbone uses the pre-trained visual encoder and LLM weights following BLIP2[Li *et al.*, 2023].

Talk2BEV[Dewangan *et al.*, 2023] proposes an innovative bird's-eye view (BEV) representation of the scene fusing both visual and semantic information. The pipeline first generates the BEV map from image and lidar data and uses general-purpose visual-language foundation models to add more detailed text descriptions of cropped images of objects. The JSON text representation of the BEV map is then passed to general-purpose LLM to perform Visual QA, which covers spatial and visual reasoning tasks. The result shows a good understanding of detailed instance attributes and also higher-level intent of objects, and the ability to provide free-formed advice on the ego vehicle's actions.

LiDAR-LLM[Yang *et al.*, 2023a] uses a novel approach that combines point cloud data with the advanced reasoning abilities of Large Language Models to interpret real-world 3D environments and achieves excellent performance in 3D captioning, grounding, and QA tasks. The model employs a unique three-stage training and a View-Aware Transformer(VAT) to align 3D data with text embedding, enhancing spatial comprehension. Their examples show the model can understand the traffic scenes and provide suggestions for autonomous driving planning tasks.

[Atakishiyev *et al.*, 2023] focus on the the explainability of vehicle's actions using a visual QA approach. They collected driving videos in simulated environments from 5 different action categories(like going straight and turning left) and used manually labeled explanations of actions to train

the model. The model was able to explain the driving decision based on road geometry and clearance of obstacles. They find it promising to apply state-of-the-art multi-modal foundation models to generate structured explanations of vehicle actions.

4.2 Unified Perception and Planning

[Wen *et al.*, 2023b] performed an early exploration of GPT-4Vision[Achiam *et al.*, 2023]'s application in perception and planning tasks, and evaluated its capabilities in several scenarios. It shows that GPT-4Vision can understand weather, traffic signs, and traffic lights and identify traffic participants in the scene. It can also provide more detailed semantic descriptions of these objects, such as vehicle rear lights, intents like U-turn, and detailed vehicle types(e.g. cement mixer truck, trailer, and SUV). It also shows the foundation model's potential for understanding point cloud data, GPT-4V can identify vehicles from point cloud contours projected in BEV images. They also evaluated the model's performance on planning tasks. Given the traffic scenario, GPT4-V is asked to describe its observation and decision on the vehicle's action. The results show good interaction with other traffic participants and compliance with the traffic rules and common sense, e.g. following the car at a safety distance, yielding to cyclists at a crosswalk, remaining stopped until the light turns green. It can even handle some long-tail scenarios very well, such as the gated parking lot.

Instruction tuning is used to better adapt general-purpose multi-modal foundation models to autonomous driving tasks. DriveGPT4[Xu *et al.*, 2023] created an instruction-following dataset, where ChatGPT, YOLOV8[Reis *et al.*, 2023] and ground truth vehicle control signals from BBD-X dataset[Kim *et al.*, 2018] are used to generate question and answers about common objects detections, spatial relations, traffic light signals, the ego vehicle's actions. Following LLaVA, it used the pre-trained CLIP[Radford *et al.*, 2021] encoder and LLM weights and fine-tuned the model with their instruction-following dataset specifically designed for autonomous driving. They were able to build an end-to-end interpretable autonomous driving system, which is able to have a good understanding of the surrounding environment and make decisions on vehicle actions with jurisdictions and lower-level control commands.

4.3 Limitations and Future Directions

The multi-modal foundation models show capability for spatial and visual reasoning, which is required by autonomous driving tasks. Compared to traditional object detection, classification model trained on the closed-set dataset, the visual reasoning capability and free-formed text description can provide more abundant semantic information, which can solve many long-tail detection problems, such as classification of special vehicles, and understanding of hand signals from the police officers and traffic controllers. The multi-modal foundation models have good generalization capability and can handle some challenging long-tail scenarios very well using common sense, like stopping at a gate with controlled access. Further leveraging its reasoning capability for planning tasks, the vision-language models can be used for unified perception planning and end-to-end autonomous driving.

There are still limitations of multi-foundation models in autonomous driving. [Wen *et al.*, 2023b] shows the GPT-4V model still suffers from hallucination and generates unclear responses or false answers in several examples. The model also shows incompetence in utilizing multi-view cameras and lidar data for accurate 3D object detections and localization, because the pre-training dataset only contains 2D images from the web. More domain-specific fine-tuning or pre-training is required to train multi-modal foundation models to better understand point cloud data and sensor fusion to achieve comparable performance of the state-of-the-art perception system.

5 Conclusion and Future Directions

We have summarized and categorized recent papers applying foundation models to autonomous driving. We build a new taxonomy based on modality and functions in autonomous driving. We have detailed discussions on methods and techniques for adapting foundation models to autonomous driving, e.g. in-context learning, fine-tuning, reinforcement learning, and visual instruction tuning. We also analyze the limitations of foundation models in autonomous driving, e.g. hallucination, latency, and efficiency as well as the domain gap in the dataset, and thereby propose the following research directions:

- Domain-specific pre-training or fine-tuning on autonomous driving dataset
- Reinforcement Learning, and Human-in-the-loop alignment to improve safety and reduce hallucinations
- Adaptation of 2D foundation models to 3D, e.g. language guided sensor fusion, fine-tuning, or few-shot learning on the 3D dataset
- Latency and memory optimization, model compression, and knowledge distillation for deployment of foundation models to vehicles

We also notice that the dataset is one of the biggest obstacles in the future development of foundation models in autonomous driving. The existing open-sourced dataset[Li *et al.*, 2024] for autonomous driving at the scale of 1000 hours, is far less than pre-training datasets used for state-of-the-art LLMs. The web dataset used for existing foundation models doesn't leverage all modalities required by autonomous driving, such as lidar and surround cameras. The web data domain is also quite different from the real driving scenes.

We propose the longer-term future road map in Figure 5. In the first stage, we can collect a large-scale 2D dataset that can cover all data distribution, diversity, and complexity of driving scenes in the real-world environment for pre-training or fine-tuning. Most vehicles can be equipped with front cameras to collect the data in different cities, at various times of the day. In the second stage, we can use smaller but higher-quality 3D datasets with lidar to improve the foundation model's 3D perception and reasoning, for example, we can use existing state-of-the-art 3D object detection models as teachers to fine-tune the foundation model. Finally, we can

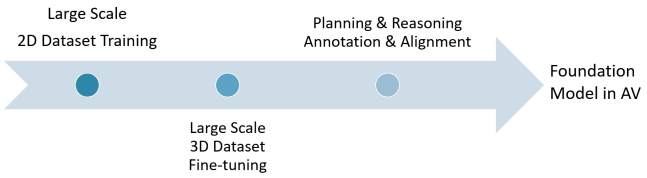

Figure 5: Road map of foundation models in AD.

leverage human driving examples or annotations in planning and reasoning for alignment, reaching the utmost safety goal of autonomous driving.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Atakishiyev *et al.*, 2023] Shahin Atakishiyev, Mohammad Salameh, et al. Explaining autonomous driving actions with visual question answering. *arXiv preprint arXiv:2307.10408*, 2023.

[Brohan *et al.*, 2023] 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.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[Chen *et al.*, 2023] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving, 2023.

[Cheng *et al.*, 2023] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything, 2023.

[Chib and Singh, 2024] Pranav Singh Chib and Pravendra Singh. Lg-traj: Llm guided pedestrian trajectory prediction. *arXiv preprint arXiv:2403.08032*, 2024.

[Cui *et al.*, 2023] 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, 2023.

[Deng *et al.*, 2023] Yao Deng, Jiaohong Yao, Zhi Tu, Xi Zheng, Mengshi Zhang, and Tianyi Zhang. Target: Automated scenario generation from traffic rules for testing autonomous vehicles, 2023.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dewangan *et al.*, 2023] Vikrant Dewangan, Tushar Choudhary, et al. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. *arXiv preprint arXiv:2310.02251*, 2023.

[Ding *et al.*, 2023] 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, 2023.

[Felemban *et al.*, 2024] Abdulwahab Felemban, Eslam Mohamed Bakr, Xiaoqian Shen, Jian Ding, Abduallah Mohamed, and Mohamed Elhoseiny. imotion-llm: Motion prediction instruction tuning. *arXiv preprint arXiv:2406.06211*, 2024.

[Ho *et al.*, 2022] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022.

[Hu *et al.*, 2023] Anthony Hu, Lloyd Russell, et al. Gaia-1: A generative world model for autonomous driving, 2023.

[Huang *et al.*, 2023] Yu Huang, Yue Chen, et al. Applications of large scale foundation models for autonomous driving, 2023.

[Jin *et al.*, 2023] Ye Jin, Xiaoxi Shen, Huiling Peng, Xiaoan Liu, Jingli Qin, Jiayang Li, Jintao Xie, Peizhong Gao, Guyue Zhou, and Jiangtao Gong. Surrealdriver: Designing generative driver agent simulation framework in urban contexts based on large language model, 2023.

[Keysan *et al.*, 2023] Ali Keysan, Andreas Look, Eitan Kosman, Gonca Gürsun, Jörg Wagner, Yu Yao, and Barbara Rakitsch. Can you text what is happening? integrating pre-trained language encoders into trajectory prediction models for autonomous driving, 2023.

[Kim *et al.*, 2018] 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.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[Lai *et al.*, 2023] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[Li *et al.*, 2024] Hongyang Li, Yang Li, et al. Open-sourced data ecosystem in autonomous driving: the present and future, 2024.

[Liu *et al.*, 2023a] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[Liu *et al.*, 2023b] Shilong Liu, Zhaoyang Zeng, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[Mao *et al.*, 2023a] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt, 2023.

[Mao *et al.*, 2023b] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving, 2023.

[Oord *et al.*, 2017] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Ouyang *et al.*, 2022] 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.

[Peng *et al.*, 2023] Xidong Peng, Runnan Chen, et al. Learning to adapt sam for segmenting cross-domain point clouds, 2023.

[Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[Reis *et al.*, 2023] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8, 2023.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[Sha *et al.*, 2023] Hao Sha, Yao Mu, et al. Languagempc: Large language models as decision makers for autonomous driving, 2023.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[Tan *et al.*, 2023] Shuhan Tan, Boris Ivanovic, Xinshuo Weng, Marco Pavone, and Philipp Kraehenbuehl. Language conditioned traffic generation, 2023.

[Wang *et al.*, 2022a] Mingze Wang, Ziyang Zhang, and Grace Hui Yang. Incorporating voice instructions in model-based reinforcement learning for self-driving cars, 2022.

[Wang *et al.*, 2022b] Sen Wang, Zhuheng Sheng, et al. Adept: A testing platform for simulated autonomous driving. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–4, 2022.

[Wang *et al.*, 2023a] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[Wang *et al.*, 2023b] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving, 2023.

[Wen *et al.*, 2023a] Licheng Wen, Daocheng Fu, et al. Dilu: A knowledge-driven approach to autonomous driving with large language models, 2023.

[Wen *et al.*, 2023b] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi Bai, Xinyu Cai, Min Dou, Shuanglu Hu, Botian Shi, and Yu Qiao. On the road with gpt-4v(ision): Early explorations of visual-language model on autonomous driving, 2023.

[Xu *et al.*, 2023] 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.

[Yang *et al.*, 2023a] 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.

[Yang *et al.*, 2023b] Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. Human-centric autonomous systems with llms for user command reasoning, 2023.

[Yang *et al.*, 2023c] Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving, 2023.

[Zhang *et al.*, 2023a] Dingyuan Zhang, Dingkang Liang, et al. Sam3d: Zero-shot 3d object detection via segment anything model, 2023.

[Zhang *et al.*, 2023b] 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.

[Zhang *et al.*, 2023c] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Learning unsupervised world models for autonomous driving via discrete diffusion, 2023.