Foundation Models for Automated Vehicles

Varun Vaidhiya(5589918)
Supervisor: Karim El Haloui

March 19, 2024

Contents

List of Figures

List of Tables

List of Abbreviations

AV - Automated Vehicles
FM - Foundation Models
LLM - Large Language Models
ITS - Intellegent Transport System
ViT - Vision Transformers
VFM - Vision Foundation Models
KD - Knowledge Distillation
DiT - Diffusion Transformer
BEV - Birds Eye View
PEFT - Parameter Efficient Fine-Tuning
GPU - Graphical Processing Unit
TPU - Tensor Processing Unit
FPGA - Field Programmable Gate Array
ASIC - Application Specific Integrated Circuit
VQA - Vision Question Answering.
GAN -Generative Adversarial Networks
VAE - Variational Auto Encoders

Abstract

Foundation models are the base model used for generative AI is one of the fast growing and highly researched topics in the last 2 years. While Open AI's Chat GPT uses foundation model in terms of LLM which are trained using transformer architecture for creating text output and other products like Sora and Dall-E use Diffusion Transformer (DiT) to train their Vision Foundation Model for creating video and image output. Similarly, there are lot of research going on to use foundation models for Automated vehicle perception, behavioural planning and prediction as well . The current CNN based approach that has been used for different vision task like object detection, object classification, semantic segmentation , Birds-eye-View (BEV), action recognition , scene understanding and object distance calculation has instrumental for the functioning of the current SAE level 3 Automated vehicle .But to achieve Level 4 and 5 it is important to use much advanced AI techniques . There are lot of research papers related to using Vision Transformer that are trained on smaller dataset instead of using conventional CNN .But there are no enough researches for using it with Large Foundation models due to some limitations .Hence , I will also be working on optimising the algorithm using FM to work efficiently with minimal hardware resources by using both ViT and CNN based hybrid architecture depending on the Operation Design Domain (ODD) and also look into cloud offloading when necessary .All the future Updates of this project can be found on this GitHub repository FMforAD

1 Introduction

Foundation Models [8] [9] which are a large pre-trained model that are trained on vast number of datasets and several billions of parameters. In this Paper I will look into the applications of foundation models and LLMs for autonomous driving from the back-end and the front-end.The back-end utilisation is the set of tasks which include simulation , synthetic data generation and annotation, and the front-end appliance consists of world models, perception , planning/decision making and E2E driving operations.I will also discuss about various techniques for tuning foundation models for our specific tasks as shown in Fig 2

The utilisation of foundation models in AVs is still in its early stages, but research is progressing rapidly. As these models are further optimised for efficiency and real-time performance on edge devices, they have the potential to become a cornerstone in achieving Level 5 autonomy, enabling truly driver less vehicles that navigate complex road environments safely and reliably.Refer to Figure 1 for different levels of AV.

1.1 Foundation Models for AV

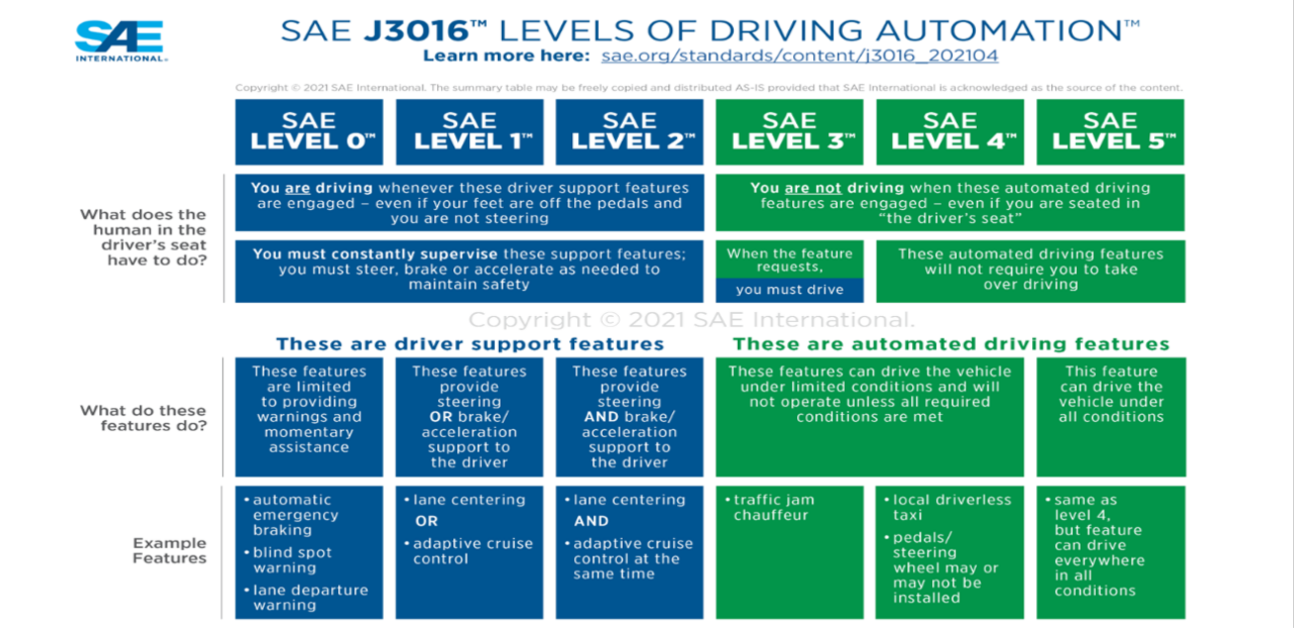
Automated driving is a challenging and long-tailed problem that requires human knowledge, common-sense, and reasoning. Foundation models and LLMs can leverage these abilities to reformulate and improve Automated driving systems. Foundation models when used for Automated vehicles could solve many existing requirements like simulation, world model, data annotation, planning, and end-to-end solutions.

1.1.1 Core Types of Foundation Models

The different types of foundation models crucial to Automated vehicles (AVs) are:

- **Large Language Models (LLMs)**: Help with decision-making, route planning, and communication due to their text understanding and generation abilities.
- **Computer Vision Foundation Models**: Vital for object detection, scene understanding, and mapping.
- **Multi-modal Foundation Models**: Combine information from cameras, LiDAR, text, and other sensors for a comprehensive world view.

Traditionally, AVs have relied heavily on rule-based systems[**?**] and sensor data processing. However, these methods struggle to capture the intricacies and nuances of real-world driving scenarios. Are these traditional AI methods enough to create a SAE level 4 or level 5 (Full self-Driving)[1] vehicle? In this survey I will be exploring all the research that has been done in Foundation Models for doing the



SAE J3016™ LEVELS OF DRIVING AUTOMATION™

Learn more here: sae.org/standards/content/j3016_202104

Copyright © 2021 SAE International. The summary table may be freely copied and distributed AS-IS provided that SAE International is acknowledged as the source of the content.

	SAE LEVEL 0™	SAE LEVEL 1™	SAE LEVEL 2™	SAE LEVEL 3™	SAE LEVEL 4™	SAE LEVEL 5™
What does the human in the driver's seat have to do?	You **are** driving whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering			You **are not** driving when these automated driving features are engaged – even if you are seated in "the driver's seat"		
	You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety			When the feature requests, you must drive	These automated driving features will not require you to take over driving	
	These are driver support features			These are automated driving features		
What do these features do?	These features are limited to providing warnings and momentary assistance	These features provide steering **OR** brake/acceleration support to the driver	These features provide steering **AND** brake/acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met		This feature can drive the vehicle under all conditions
Example Features	• automatic emergency braking • blind spot warning • lane departure warning	• lane centering **OR** • adaptive cruise control	• lane centering **AND** • adaptive cruise control at the same time	• traffic jam chauffeur	• local driverless taxi • pedals/steering wheel may or may not be installed	• same as level 4, but feature can drive everywhere in all conditions

Figure 1: SAE Levels for AV [1].

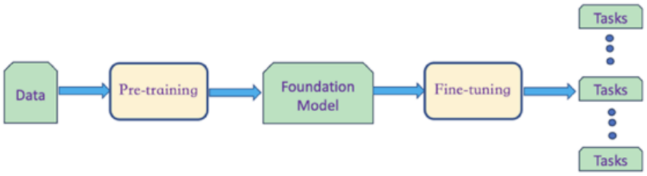


Figure 2: Foundation Model [2].

task mentioned above in an Automated vehicle and make a conclusion on whether using foundation models are beneficial or not ? is it possible to run a foundation model in the current generation Compute hardware or does it need some other techniques to do it ? I will be looking into all this methods and make a decision on when and where to use foundation models . And look into the methods of creating a hybrid architecture .

This paper will delve deeper into the specific applications of foundation models for autonomous vehicles, focusing on object detection, behaviour prediction, route planning and optimisation. We will explore the current research landscape, challenges, and potential solutions for integrating these powerful models into the complex world of autonomous driving.

The landscape changes quickly. Even a paper from a year or two ago might discuss outdated hardware or less computationally intensive foundation models.

2 Background

Runtime systems in autonomous vehicles are the software and hardware components that manage the real-time operation and decision-making of a self-driving car. They work in concert with the pre-developed algorithms and models to do the following tasks:

- **Real-time Perception:**Foundation models can analyse sensor data (cameras, LiDAR) in real-time, leading to robust object detection and tracking (vehicles, pedestrians, traffic signs). This enables the AV to maintain a comprehensive understanding of its surroundings, even in adverse weather or complex traffic patterns.
- **Scene Understanding:**Beyond object detection, foundation models can delve deeper, inferring the overall context of the scene. They can interpret traffic rules, social cues from other drivers,

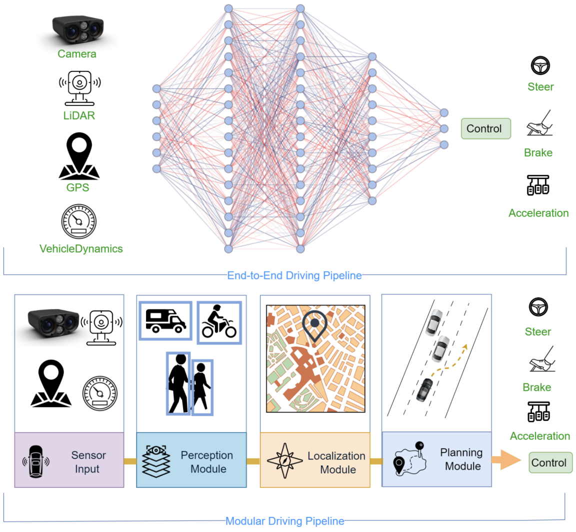


Figure 3: End-End and modular systems.

and potential hazards like road anomalies. This enriched understanding is crucial for safe and informed decision-making.

- **Dynamic Route Planning and manoeuvring:**Foundation models can leverage their understanding of traffic conditions and real-time updates to optimise routes, perform on-the-fly re-routing, and even suggest creative manoeuvres in unexpected situations.
- **Human-Machine Interaction:**Natural language processing capabilities of foundation models allow for intuitive interaction with the AV via voice commands or text. Additionally, these models can potentially personalise driving styles based on learned user preferences, enhancing overall comfort and user experience.

The current autonomous driving solutions can be broadly classified into into **modular paradigm** and the **end-to-end system**, as shown in in Fig3.Modular AV systems break down perception, planning, and control tasks into independent modules, while end-to-end [10]approaches train a single neural network to handle all aspects of driving by taking sensor data as input and provides control signal to the actuators as the output.

World modelling involves an AI agent (e.g., a robot, an autonomous vehicle) learning to build a compressed and structured internal representation of its environment. This model goes beyond raw sensory data, capturing the underlying dynamics, rules, and potential future states of the world.A good world model allows the agent to plan effectively, reason abstractly, handle novel situation.

3 Literature Review

3.1 Architectures for Foundation Models

Foundation Models are the pre-trained models that are trained on large data sets of unlabelled data.Though the foundation models can be trained using architectures like CNN , RNN and other architectures they are widely trained using Transformer architecture due to their advantages like Long Range dependency , Parallelisation , Focus on Attention and Encoder Decoder Structure.The first Transformer architecture was introduced by researchers form Google through a paper ”Attention is all you need” [3] in 2017, shown in Fig. 4.

While Transformer are good at Natural Language Processing(NLP)tasks they are not suitable for vision tasks , hence a slightly modified architecture Vision transformers (ViTs),shown in Fig. 5.[4], specifically designed for computer vision tasks, came later, around 2020 and has a huge potential to replace CNN models and led to the development of Vision Foundation Models.

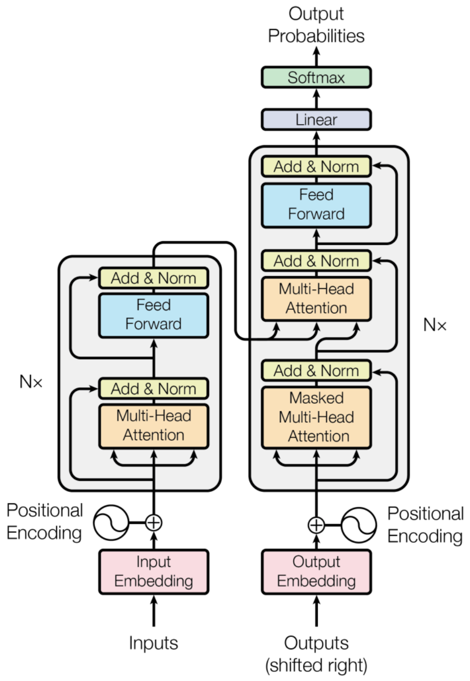


Figure 4: Transformer Architecture [3].

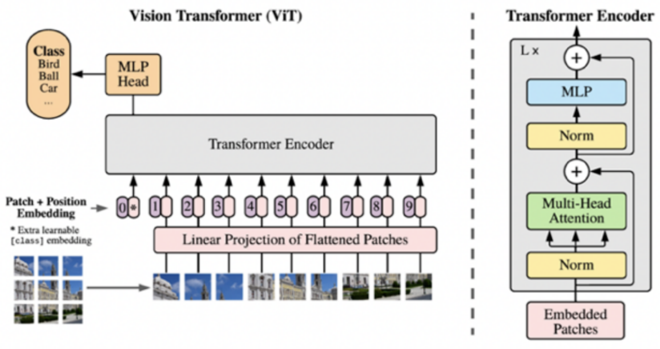


Figure 5: Vision Transformer Architecture [4].

Many recent research has revealed that ViT's perform much better than CNN due to their cross attention mechanism and Global Context Understanding . Some of the perception tasks like segmentation[11] , BEV[12][13] [14] , 2D and 3D object Detection[15][16][17], Scene Understanding[18] and Action Recognition [19], Sensor fusion[20][21] and future prediction [22] are proved to be outperforming the CNN counterparts .But these methods are trained using small amount of labelled dataset and does not use foundation models .

Thus by using the vision transformer architecture and training is with large amount of unlabelled dataset enables us to create much advanced vision foundation models which would unlock whole new possibilities for autonomous vehicle perception and planning tasks .

In this Literature Review I will discuss about the existing foundation models that are suitable for the AV perception and planning tasks and try to improve the performance of it for the specific tasks .Then i will discuss about the tools/frameworks that are available to use those LLM and VFM for creating a application that are specific to our need .Later I will be exploring the various techniques that are available to optimise the large models to run on edge devices like a vehicle .

Image and video generation applications rely on diffusion models in conjunction with transformer architectures [23] to achieve their impressive results.

Later the development of Multi-model architecture has enabled a whole new possibility of creating a Foundation model that would get input in any format and generate output in any format like Text, image, video and audio. But it might take some more time to work as smoothly as the Language models that are recently introduced .

The list of foundation models/LLM's and their application in different modalities that has a potential for Autonomous Vehicle development are given in Table 1

Table 1: Popular Foundation Models/LLM's and General Application Areas

Foundation Model	**General Applications**
GPT-3 [24](and variants)	Text generation, translation, summarisation, chatbots
BERT [25](and variants)	Natural language understanding, question answering, text classification
CLIP [26] (Contrastive Language-Image Pre-training)	Zero-shot Image Classification ,Robust Image Search, Image Generation
DALL-E [27](and variants)	Image generation from text descriptions
AlphaFold (DeepMind)	Made breakthroughs in protein structure prediction, a complex problem within biology.
Stable Diffusion [28]	Image generation, image editing
Codex	Code generation and translation
Flamingo [29], PaLM [30]	Combining text, image, and other modalities for diverse tasks
Claude [31]	Text generation, chatbots (focus on safety and reduced biases)
Midjourney	Image generation from text
Mistral	Open Source Image generation, creative tasks
Gemini [32]	Large language model by Google AI (more details often in research papers)
NExT-GPT	Any-to-Any Multimodality . NExT-GPT can understand and generate different types of data.
Megatron-Turing NLG	Text generation, summarisation, translation and Question answering.

Hugging Face [33] is a Open-Source Model and Dataset Repository that hosts a massive collection of pre-trained machine learning models and datasets across various domains.

The application of large language models in the field of autonomous driving covers a wide range of task types, and has a revolutionary potential . Foundation Models tasks and LLMs in autonomous driving pipelines is shown in the Fig. 6 and 7.

3.2 Perception and Scene Understanding

Vision Foundation Models can be used for this task as these models excel at identifying and classifying objects in complex scenes (cars, pedestrians, traffic signs, road anomalies, etc.). They outperform traditional computer vision techniques in challenging conditions and are more robust to noise and adverse weather conditions.

Object Detection and Tracking: Vision foundation models can excel at identifying vehicles, pedestrians, traffic signs, and other road elements in complex scenes.DINO[34] and DINO V2[35] are the foundation models which are trained using Self Supervised Learning methods . Semantic Segmentation is the process of Understanding the layout of the scene (e.g., road, sidewalk, buildings) is crucial for safe navigation .Vision foundation models excel this tasks one of the example is Segment Anything (SAM)[36] which can differentiate small objects and are more robust to noise and bad weather conditions.

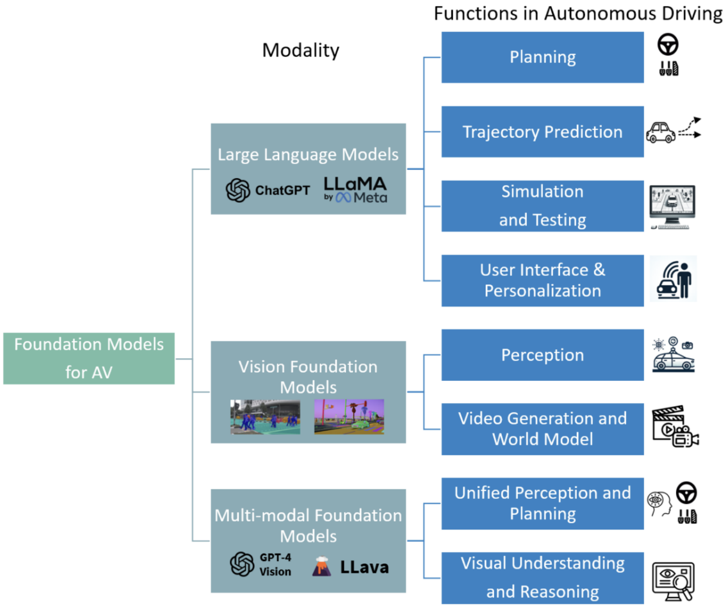

Functions in Autonomous Driving
Modality
Foundation Models for AV
Large Language Models
ChatGPT
LLaMA by Meta
Vision Foundation Models
Multi-modal Foundation Models
GPT-4 Vision
LLava
Planning
Trajectory Prediction
Simulation and Testing
User Interface & Personalization
Perception
Video Generation and World Model
Unified Perception and Planning
Visual Understanding and Reasoning

Figure 6: Foundation Model Tasks [5].

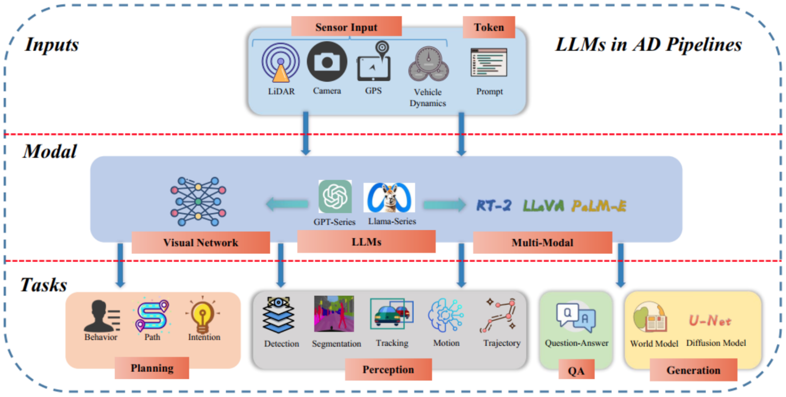


Figure 7: Foundation Model in AD Pipeline.[6]

3.2.1 Enhanced Object Detection and Tracking

Multi-Sensor Fusion: Foundation models (particularly multimodal ones) can effectively combine and align data from cameras, LiDAR, radar, and other sensors. This creates a more robust and detailed representation of the surrounding environment.

3D Scene Reconstruction: Models can create detailed 3D representations of the environment, including depth information, using visual and/or LiDAR data. This improves localisation and mapping abilities for AVs. Understanding Complex Situations: LLMs can reason about traffic rules, social cues, and the context of a scene, enhancing the AV's ability to interpret and react to dynamic situations.

3.2.2 Handling Uncertainty and Adaptability

Dealing with Novel Objects and Situations: Foundation models, due to their training on diverse data, are better at dealing with objects or scenarios not seen before, improving safety in the real world. Quantifying Uncertainty: Models can indicate the confidence level of their predictions, allowing AVs to make informed decisions in ambiguous scenarios. Data-Driven Adaptation: Foundation models can be fine-tuned or updated based on real-world driving data collected by the AV fleet.

3.2.3 Foundation models for Automated vehicle SLAM / mapping and localisation

Semantic Mapping: Foundation models excel at understanding the environment. They can create semantically rich maps, identifying not just geometric elements but also objects (cars, pedestrians, buildings), traffic rules, and road markings. This adds contextual awareness for safer navigation. Prior Knowledge and Map Augmentation: LLMs can store and process vast amounts of geographic knowledge. This information could be integrated with SLAM maps, enhancing localisation and path planning in familiar areas. Object Recognition for Loop Closure: In traditional SLAM, loop closure (recognising a previously visited place) is key for correcting map drift. Foundation models can improve feature matching and place recognition, particularly in visually challenging scenarios. Change Detection: Foundation models can detect changes in an environment over time (e.g., construction zones, new traffic patterns), allowing for updating dynamic elements within maps.

3.3 Foundation Models for decision making and route planning

LLMs show great promise in enhancing Automated vehicle decision-making and route planning. These models can process real-time traffic data, road condition reports, and even social media updates to identify potential bottlenecks, accidents, or unexpected events . They can reason about these factors and suggest alternative routes for optimal travel time and safety . Moreover, LLMs can learn driver preferences and historical patterns, personalising routes in line with individual needs and priorities . This ability to understand complex scenarios and generate human-like solutions positions LLMs as a key component in the intelligent navigation systems of Automated vehicles.

Behavioural Prediction: LLMs can anticipate the actions of other road users, aiding safe interactions with surrounding vehicles and pedestrians. . Route Planning: LLMs, aided by real-time data, can optimise routes, predict delays, and find alternative paths in case of disruptions. . Complex manoeuvres: Foundation models could be used for complex manoeuvres like lane changes, merging, and navigating intersections. For detailed references for Researches on Methods or FM Refer Tables .2 3 4 5 6.

3.3.1 World Model

David Ha and colleagues pioneered the concept of World Models in their seminal 2018 paper titled "Recurrent World Models Facilitate Policy Evolution" [37], laying the foundation for the use of internal representations of the environment in reinforcement learning agents. Generative world models hold significant promise to advance deep learning for robotics applications like self-driving cars. World modelling can help AI models learn general representations of how the world works and how to predict what might happen next. Like how people use mental models to make sense of the world and guide their actions, embodied AI systems could benefit from world models like GAIA-1[38] by wayve . These models could help autonomous vehicles better understand their surroundings, allowing them to efficiently anticipate and plan their driving actions.This research model was developed to enhance and accelerate the training of Wayve's end-to-end AI software for autonomous driving Other research papers related to world models include [39][40][41].

3.4 Natural Language Interaction

Voice Commands: LLMs enable intuitive interaction with the vehicle, enhancing the user experience. Traffic Sign Understanding: LLMs can interpret text on traffic signs or in construction zones for better situation awareness. Personalised Communication: The vehicle can adapt its communication style and information presentation based on the driver's preferences learned by LLMs.

Wayve a UK based Autonomous Vehicle startup recently unveiled another first-of-its-kind AI model, LINGO-1[42], that uses natural language to comment on driving scenes and explain its decision-making. Developments like LINGO-1 and GAIA-1 pave the way to significant advances in the field of autonomy, which can help self-driving cars better understand and predict their surroundings. This could represent a significant leap forward in making their safe deployment a reality. For detailed references for Researches on Methods or FM Refer Tables .2 3 4 5 6.

3.5 Foundation models for Automated vehicle testing

3.5.1 Scenario Generation

Realistic and Diverse Scenarios: Foundation models, particularly LLMs, can generate a vast range of realistic driving scenarios, including challenging edge cases and unexpected events. This allows testing of AVs in situations that may be difficult or dangerous to replicate in the real world. Virtual Environments: Foundation models can create detailed virtual environments, including varying weather conditions, road layouts, and traffic patterns, providing a comprehensive testing ground for AV software.

3.5.2 Augmenting Physical Testing

Stress Testing: Foundation models can identify scenarios most likely to expose weaknesses in AV systems, improving the efficiency of real-world testing. Sensor Simulation: Models can simulate sensor data (camera, LiDAR, etc.) with varying levels of noise and distortion, testing the AV's ability to handle imperfect input.

3.5.3 Behaviour Prediction and Anomaly Detection

Anticipating Other Road Users: LLMs can model the likely behaviours of other vehicles, pedestrians, and cyclists, aiding the AV system in testing its reactions and decision-making in complex interactions. Identifying Anomalies: Foundation models can be trained to detect anomalous situations or deviations from expected driving patterns, potentially flagging unexpected situations during testing.

3.5.4 Explanation and Debugging

Understanding AV Decisions: LLMs can help explain the reasoning behind an AV's actions in specific scenarios, improving transparency and assisting with debugging. Generating Counterfactuals: Foundation models can explore "what-if" situations, potentially revealing the root cause of failures and informing improvements. For detailed references for Researches on Methods or FM Refer Tables .2 3 4 5 6.

3.6 Foundation models for Automated vehicle dataset generation

Generative Artificial Intelligence (Generative AI) has experienced significant advancements, enabling the automatic creation of diverse content types like text, images, videos, and audio in response to user specifications. This technology revolutionises content production efficiency and holds vast application potential. Mature approaches within generative AI include Variational Auto-Encoders (VAE), Generative Adversarial Networks (GAN), Normalising Flows, Energy-Based Models, Generative Models from Physical Processes, Diffusion Models, and Generative Pre-trained Transformers (GPT). These techniques vary in approach but collectively drive innovation in realistic, user-guided content generation.

3.6.1 Synthetic Data Generation

Generative technologies have revolutionized various generation tasks across different modalities, including text, images, videos, and cross-modal data. In text generation, advancements in deep learning, particularly with models like the Transformer architecture and GPT series, have significantly enhanced the quality and diversity of generated text. ChatGPT, a specialized version developed by OpenAI, has been integrated into platforms like Microsoft's Bing, enhancing conversational search experiences by providing detailed and contextually relevant responses.

Image generation tasks leverage neural networks to create new visual content from given information, such as images and text. Techniques such as style transfer, image morphing, and diffusion models have been employed to restore images to their original high-quality state and achieve superior image generation quality.

Video generation involves creating coherent sequences of frames to form videos. Unconditional video generation extracts valuable information from training data to generate new videos, while conditional video generation incorporates additional conditions, such as text or captions, to guide the generation process. Popular methods like Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN) are utilized for conditional video generation.

Cross-modality data generation aims to generate data in one modality based on features from a different modality. This task involves effectively modelling relationships between different modalities to control the generation process. Text-to-image generation, image-to-text generation, and text-to-video generation are common examples. These tasks often rely on architectures like GANs, Transformers, and diffusion models to achieve accurate and contextually relevant generation results.

Overall, the advancements in generative AI have led to significant progress in various generation tasks, offering solutions that range from enhancing conversational interactions to creating realistic visual content across different modalities. These technologies have vast implications for intelligent transportation systems, including simulating environments, creating objects, and generating sensor data for training and testing purposes. For Researches on Methods or FM Refer Tables .2 3 4 5 6.

3.6.2 Data Augmentation

Expanding Existing Datasets: Foundation models can apply transformations to existing data, increasing variety and addressing biases. This includes image rotations, lighting changes, adding noise, etc. Generating Rare Events: Models can create examples of rare but critical scenarios (accidents, unusual weather), which are often underrepresented in real-world datasets.

3.6.3 Domain Adaptation

Bridging Gaps between Simulation and Reality: Foundation models can help adapt synthetically generated data to look more like real-world data, reducing the sim-to-real gap and improving the performance of AVs trained on such datasets. Transfer Learning Across Locations: Models can facilitate knowledge transfer from data collected in one region to data in another, helping AVs generalise to new environments.

3.6.4 Automatic Labelling

Semantic Segmentation Pre-labelling: Vision foundation models can provide preliminary labels for objects and scene elements in images, significantly reducing the time and cost of manual annotation. 3D Bounding Box Pre-labelling: Models can generate initial 3D bounding boxes around objects, aiding in the 3D object detection labelling process.

Fig 8 shows the different applications of Foundation Models in Autonomous Vehicles.

3.7 Optimising Foundation Models for Runtime AV Systems

3.8 Challenges of implementing above tasks using FM in AV

Latency and Efficiency: Foundation models often suffer from high latency, and generating detailed driving decisions could cause the Vehicle computer to overload . It takes several seconds for inference according to [43]. Foundation models trained with billions of parameters can consume over 100GB of

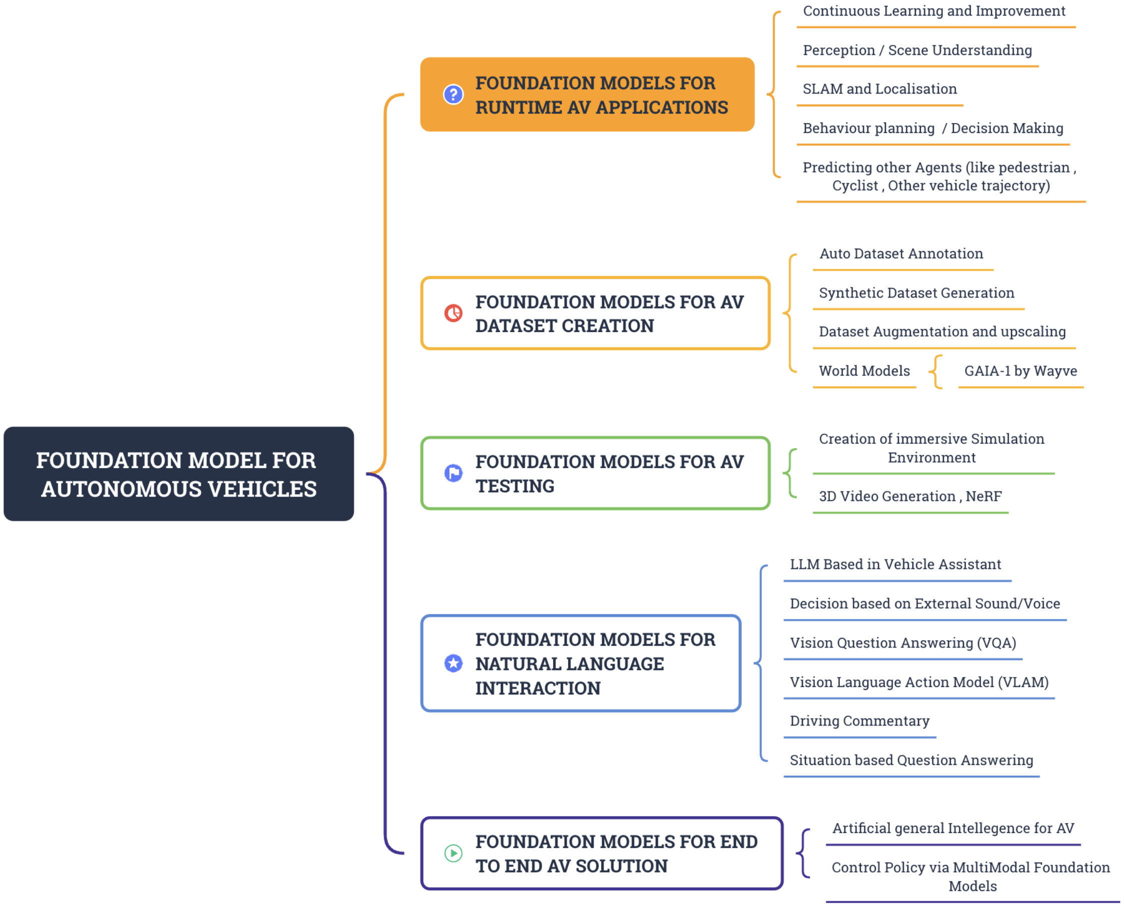

FOUNDATION MODEL FOR AUTONOMOUS VEHICLES
FOUNDATION MODELS FOR RUNTIME AV APPLICATIONS
Continuous Learning and Improvement
Perception / Scene Understanding
SLAM and Localisation
Behaviour planning / Decision Making
Predicting other Agents (like pedestrian , Cyclist , Other vehicle trajectory)
FOUNDATION MODELS FOR AV DATASET CREATION
Auto Dataset Annotation
Synthetic Dataset Generation
Dataset Augmentation and upscaling
World Models
GAIA-1 by Wayve
FOUNDATION MODELS FOR AV TESTING
Creation of immersive Simulation Environment
3D Video Generation , NeRF
FOUNDATION MODELS FOR NATURAL LANGUAGE INTERACTION
LLM Based in Vehicle Assistant
Decision based on External Sound/Voice
Vision Question Answering (VQA)
Vision Language Action Model (VLAM)
Driving Commentary
Situation based Question Answering
FOUNDATION MODELS FOR END TO END AV SOLUTION
Artificial general Intellegence for AV
Control Policy via MultiModal Foundation Models

Figure 8: Foundation Models Use cases in AV

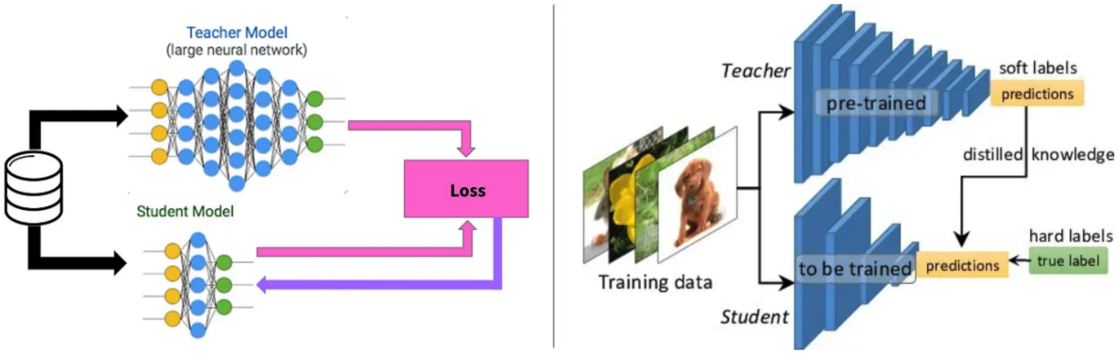


Figure 9: Knowledge Distillation Methodology [7]

memory, which might interfere with other critical modules in autonomous driving vehicles. hence in this paper I will be discussing about techniques such as model compression and knowledge distillation, to make the foundation model more efficient and easier for deployment.

Metric Precision: Classical SLAM demands precise metric measurements (distances, angles) for accurate positioning. Foundation models, while strong in semantic understanding, may not inherently focus on this level of geometric accuracy.

Real-time Constraints: Large foundation models can be computationally expensive. Integrating them seamlessly with the real-time requirements of perception can be a hurdle. Drift and Uncertainty: SLAM algorithms are designed to handle and correct accumulating errors in localisation. Foundation models would need to be adapted to quantify and represent uncertainty effectively within the SLAM framework.

Hybrid Approaches: Combining traditional perception and planning with foundation models is promising. Foundation models could provide semantic information, loop closure enhancements, and prior knowledge.But is there enough Frameworks to implement these methods .

3.8.1 Knowledge Distillation(KD)

The main problems with large foundation models are they need more resources to run , the inference time is more and it is harder to fine-tune and manage them .To fix this methods like Knowledge Distillation can be implemented .Geoffrey Hinton, Oriol Vinyals, and Jeff Dean introduced the concept of knowledge distillation in their paper titled “Distilling the Knowledge in a Neural Network” [44]. Jianping el.al [45] discussed about perspectives of knowledge categories, training schemes, teacher-student architecture, distillation algorithms, performance comparison and applications in detail .

Knowledge distillation refers to the idea of model compression by teaching a smaller network, step by step, exactly what to do using a bigger already trained network. The ‘soft labels’ as shown in Fig 9.refer to the output feature maps by the bigger network after every convolution layer. The smaller network is then trained to learn the exact behaviour of the bigger network by trying to replicate it’s outputs at every level (not just the final loss).

KD works effectively by reducing the number of layers hidden layers or the dimension size of each hidden layers in their student model .In this paper [46] the author discusses a method called DistilBERT which is the distilled version of the BERT foundation model by google . The this student model is obtained by reducing the size of the hidden layer from the main transformer architecture , Also they removed the pooler layer from the original BERT model and done various modifications in token embeddings in order to get a impressive performance in the final model .

3.8.2 Parameter Efficient Fine Tuning

Fine-Tuning is a technique used to improve the performance of Large Foundation Models on specific tasks, Parameter Efficient Finetuning also solve the same problem while addressing the challenges associated with traditional fine-tuning approaches. Here’s a breakdown:

Traditional Fine-Tuning involves taking a pre-trained LLM, and then retraining all or a significant portion of its parameters on a new, smaller dataset specific to a particular task (e.g., question answering, sentiment analysis).This can be computationally expensive and lead to overfitting, especially with limited data. This paper [47] introduces ENGINE, a parameter- and memory-efficient fine-tuning method that integrates large language models (LLMs) and graph neural networks (GNNs) to enhance textual graph modeling. Through a tunable side structure, ENGINE significantly reduces training complexity while achieving superior performance compared to previous methods, further enhanced by caching for faster training and dynamic early exit for faster inference with minimal performance loss.

Instead of retraining the entire model, PEFT [48] focuses on fine-tuning a small number of additional parameters on top of the pre-trained LLM. These new parameters act as an "adapter module" that helps the model adapt to the specific task.Freezing most of the pre-trained parameters in PEFT leverages the vast knowledge already encoded in the pre-trained model while focusing on efficient adaptation for the new task.

3.8.3 Prompt Engineering

Prompt engineering involves crafting clear and informative instructions that guide a large language model (LLM) tasked with controlling an autonomous vehicle. These prompts influence the LLM's interpretation of sensor data and decision-making during navigation. This paper [49]proposes an end-to-end control system for autonomous vehicles using LLMs. It highlights the importance of prompt engineering in defining the desired actions and goals for the LLM based on sensor data.Essentially, well-crafted prompts act as a bridge between the LLM and the complex real-world environment, ensuring the LLM interprets sensor data accurately and makes safe driving decisions.

3.8.4 Quantisation and Pruning

LLM and VFM are very large in size, hence techniques like pruning and quantisation can be used to reduce model size for deployment on edge devices.There are different methods for model pruning and quantisation .All these are discussed in papers [50][51][52][53][54] Also this Github repository [55] provides list of all the recent papers related o Neural Network quantisation .

Quantisation reduces the number of bits used to represent the values within a foundation model. Imagine using fewer digits to store numbers! While this can lead to a slight loss of accuracy, it significantly reduces model size and computational cost, making it easier to deploy on resource-constrained devices.

Pruning focuses on removing unnecessary connections between neurons in the foundation model. Think of it like trimming a tree to make it more efficient. Pruning identifies and removes connections with minimal impact on the model's performance, resulting in a smaller and faster model.

Both quantisation and pruning are crucial for making foundation models more practical for real-world applications, especially when dealing with limited computational power or memory constraints.

3.9 Evaluation Metrics for performance Tracking

Some of the Evaluation Metrics used to measure the performance of the AI models include
1. Accuracy Metrics:

Mean Squared Error (MSE) measures the average squared difference between the predicted output and the ground truth. Lower MSE indicates better performance. **Accuracy** calculates the percentage of correct predictions made by the model. For classification tasks, it's the ratio of correctly classified samples to the total number of samples **Precision** measures the ratio of true positives (correctly identified positive cases) to the total number of identified positive cases and **Recall** measures the ratio of true positives (correctly identified positive cases) to the total number of actual positive cases.
2. Efficiency Metrics:

Latency is used to measures the time taken by the model to process an input and generate an output. Lower latency is desirable for real-time applications on edge devices with limited resources. **Memory Footprint** refers to the amount of memory required by the model to run on the edge

device. Lower memory footprint is essential for devices with limited memory capabilities. **Energy Consumption** measures the amount of energy consumed by the model during operation. Lower energy consumption is crucial for battery-powered edge devices.
3. Interpretability Metrics:

Explainability is a metric that refers to the ability to understand the rationale behind the model's predictions. This is particularly important for LLMs, where the internal workings can be complex. **Fairness** is a metric that assesses whether the model's predictions are biased towards certain groups or data subsets. Fairness is crucial for ethical considerations in real-world applications.

3.10 Frameworks and Tools for building/optimising applications that are usig LLM and VFM

There are various tools and frameworks for building applications using LLM .Some of the popular frameworks that I will be using for the project are mentioned below .

LightLLM [56] is a Python-based LLM (Large Language Model) inference and serving framework, notable for its lightweight design, easy scalability, and high-speed performance.Its focus is reducing computational and memory requirements for deployment on edge devices or resource-constrained settings.

NVIDIA TensorRT LLM[57] is a software development kit (SDK) by NVIDIA for high-performance inference of deep learning models. It includes optimisations specifically tailored for large language models.Its focus is Maximizing execution speed and efficiency of LLMs and VFM's on NVIDIA GPUs.The techniques used are Graph optimisations, kernel fusion, and hardware-aware optimisations. **NVIDIA Faster Transformer**[58] is an optimised implementation of Transformer models (the backbone of most LLMs) by NVIDIA.Its focus is accelerating the core computations within LLMs for faster training and inference on NVIDIA GPUs.Techniques used in this include Low-level code optimisations, efficient use of GPU memory and compute resources. **Microsoft DeepSpeed** [59] is a deep learning optimisation library by Microsoft.Its Focus is Training very large models (LLMs) that wouldn't otherwise fit on a single GPU/node and also improving efficiency in both training and inference.Its techniques include ZeRO-Offload (memory optimisation), model parallelism, gradient checkpointing. **Microsoft Deepspeed MII (Mixture of Industry Experts)**[60], An extension of DeepSpeed specifically designed for Mixture of Experts (MOE) model architectures.Its focus is Scaling LLMs to trillions of parameters by efficiently distributing the model and computations across many GPUs. **VLLM**[61], refers to very large language models or a specific implementation is a another technique

LangChain[62] [63] is A powerful framework for building applications with LLMs.Its focus is Streamlining LLM workflows – data loading, prompt generation, chaining different models. Makes working with LLMs more accessible.

NVIDIA Triton Server is a inference server designed to deploy deep learning models in production environments.Its focus is Managing and serving LLMs or foundation models for real-time applications, handling scaling and requests from multiple clients.

3.11 Existing Foundation models for Autonomous Driving

The papers like [6] , [2] ,[64] ,[65] and [5] gives valuable resources about the latest research in this area and systematically categorise existing works across each process within the proposed framework. Based on the feasibility and time constraints of this project I will be using some of methods that are well suited for implementing in the autonomous vehicles perception and planning tasks as shown in Tables .2 3 4 5 6.

Table 2: GPT Like Tokenisation[2]

Methods	Modalities	Functions	Technologies
Talk to the Vehicle[66]	Depth, semantics, text	Navigation with trajectory ton controller	CNN, LSTM, instruction, Natural Language Encoder, WGN ,Local planner
ADAPT[67]	Videos	E2E decision making for control and action	Motion transformer, visual-language transformer for caption generation
ConBaT[68]	trajectory	Behaviour (action) learning	Casual transformer, tokenizer, world model, control barrier
MTD-GPT[69]	Object trajectory (position and speed), action	Decision making	GPT-2Transformer-based RL, GPT-like sequence modeling, POMDP, policy network, action prediction
BEVGPT[70]	BEV images	Generative pretrained model with prediction,decision making and mothin planning	Pre-training BEV prediction, modified GPT casual transformer, fine-tuning, optimisation-based planning

Table 3: Pretrained Foundation Models[2]

Methods	Modalities	Functions	Technologies
PPGeo[71]	images	Policy pre-training	Posenet, depthnet, policy learning
AD-PT[72]	Point clouds	Unified pre-trained representation	Mean Teacher, pseudo label generator, open set learning
UniPad[73]	Point clouds, images	Pretrained 3D representation learning	MAE-based mask generator, VoxelNet, differentiable neural rendering, SDF network

Table 4: Auto Annotation[2]

Methods	Modalities	Functions	Technologies
Talk2Car[74]	Text, action, object, image, point cloud	Object referral	Amazon Mechanical Turk (AMT), manual caption, GPS+IMU
OpenScene[75]	Point cloud, image, text	Open vocabulary 3D scene understanding	CLIP, segmentation, 2D-3D ensemble feature, zero-shot learning
MSSG[76]	LiDAR, text, image	Multi-modal visual ground	Token fusing, object detector
HILM-D[77]	Videos, text	Video understanding	LLMS, MMLMs, perception, reasoning, prompt, ViT, GradCAM
NuPrompt[78]	Image, text	Language prompt generation	Transformer-based Prompt-Track, LLM(GPT3.5), prompt, manual caption
UP-VL[79]	Image , point cloud	Auto labelling	Open set categories, VLMs, 3-D object detector, tracking
OpenAnnotate[80]	LiDAR, camera	Open vocabulary auto labeling	LLMS, VLMs, prompt, multi-modal spatial alignment

Table 5: LLM / VLM Based Autonomous Driving[2]

Methods	**Modalities**	**Functions**	**Technologies**
Drive-like-a-human[81]	2-D BEV, text	Planning and control	LLM(GPT3.5), perception tool, LLaMA-Adapter v2 prompts
LINGO-1 [82]	Image, text, driving data	Actions and reasoning like VQA	LLM, Vision Language Action Model
Can you text what is happening?[43]	Image, text	Trajectory prediction	LLM (DistilBERT)
Drive as you speak[83]	2-D BEV, map, GNSS, radar, LiDAR, image	Decision making for actions	LLM (chatGPT4), special tool (localisation, perception and monitoring)
DiLu[84]	Text	Decision making to control	LLM(GPT3.5), prompt, recall from memory, CoT, decision decoder
language MPC[85]	Text	Decision making for action commands	LLM(GPT3.5), CoT reasoning
DriveGPT4[49]	Control action, text, image, video	Action interpreting, e2e control, question answering	LLM (LLAMA2), tokeniser, de-tokeniser, visual instruction tuning
GPT-Driver[86]	Text, trajectory	Motion planner for trajectory generation and control	LLM(GPT3.5), prompt, fine tuning, tokeniser, perception prediction
LLMDriver[87]	Text	Driving question answer	LLM (GPT3.5), pretrained model, RL expert , LoRA finetuning
Talk2BEV[88]	BEV, image, LiDAR, text	Augment BEV map with language understanding, reasoning and decision making	LLM, Q-former in BLIP-2, LLAVA (instruction tuning), question answering ,BEV network
DriveLM[89]	2-D BEV, text	Make decision and planning	LLM, perception, prediction and planning (P3), GoT, question answering
Drive-Anywhere[90]	Image, text	E2E driving policies with multi-modal understanding	LLM, BLIP, open set learning, attention mask, ViT, perception, policy net
Agent-Driver[91]	Image, text	E2E cognitive agent for autonomous driving	Tool library, cognitive memory, reasoning engine, self-reflection

Table 6: Different Datasets for using Foundation models in AV [6]

Dataset	**Task**	**Size**	**Annotator**	**Description**
BDD-X[92]	Planning VQA	77 hours, 6970 videos, 8.4M frames, 26228 captions	Human	Ego-vehicle actions description and explanation
HAD[93]	Planning, Perception	30 hours, 5744 videos, 22366 captions	Human	Joint action description for goal oriented advice and attention description for stimulus-driven advice.
Talk2Car[74]	Planning, Perception	15 hours,850 videos of 20s each 30k frames, 11959 captions	Human	Object referral dataset that contains commands written in natural language for self-driving cars
DriveLM[89]	Perception, Prediction, Planning VQA	In Carla, 18k frames and 3.7M QA pairs; In nuScenes, 4.8k frames and 450k QA pairs	Human, Rule-Based	P3 with reasoning logic; Connect the QA pairs in a graph-style structure; Use "What if"-style questions.
DRAMA[94]	VQA	91 hours, 17785 videos, 77639 question, 102830 answering, 17066 captions	Human	Joint risk localization with visual reasoning of driving risks in a free form language descrip
Rank2Tell[95]	Perception, VQA	Several hours, 118 videos of 20s each	Human	Joint important object identification, important object localization ranking, and reasoning.
NuPrompt[78]	Perception	15 hours, 35367 prompts for 3D objects	Human, GPT3.5	Object-centric language prompt set for perception tasks.
NuScenes-QA[96]	VQA	15 hours,Train(24149 scenes, 459941 QA pairs; Test: 6019 scenes, 83337 QA pairs)	Rule-Based	Leverage 3D annotations(object category, position, orientation, relationships information) and designed question templates to construct QA pairs.
Reason2Drive[97]	Perception Prediction VQA	600K video-text pairs	Human, GPT-4	Composed of nuScenes, Waymo and ONCE, with driving instructions.
LingoQA[98]	VQA	419.9k QA pairs, 28k scenarios	Rule-Based, GPT-3.5/4 software Human	Contains reasoning pairs in addition to object presence, description, and localisation.

3.12 Hardware for training and running/testing Foundation Models

While it is important to optimise the software for the final performance of the product it is also equally important to choose the best hardware for running the foundation Model .According to Moore's law [99] the performance of the hardware increases every year with the increase in number of transistors at a particular rate ,It is always better to choose the latest hardware in the market in order to get the better results .

For training the large foundation models the most popular hardware that are used are Nvidia H100, A100 and and Google's custom TPU . Models like BLOOM , CLIP , DINO , PaLM , GPT-3 are trained by using these Processors in large clustres.

Companies like Google and Meta AI possess large-scale, custom-designed AI infrastructure. They might use specialized TPUs in addition to traditional GPUs.The list of potential hardware that could be used for running realtime FM applications is hive in the table 7

3.12.1 Important Considerations when Choosing Edge Hardware

- Estimate the FLOPs (floating-point operations per second), memory usage, and latency requirements of the final foundation models for different tasks (perception, planning, etc.).
- For battery-powered vehicles it is crucial to Look for specs like TDP (Thermal Design Power) or benchmark results that include power consumption.
- Ease of development, available libraries (TensorFlow, PyTorch support), tools for quantisation/optimisation matter.
- Hardware can range from budget-friendly to very expensive.Hence it is important to Consider the cost-performance trade-off for the project's scope.
- Some hardware platforms have automotive certifications, making the integration process easier and potentially ensuring robust operation within the vehicle environment.

Hardware acceleration refers to utilizing specialized hardware components alongside traditional CPUs to improve the performance of running large foundation models (LLMs) in autonomous vehicles (AVs). These LLMs require significant computational resources for tasks like real-time scene understanding and decision-making.

Hardware acceleration tackles these challenges in several ways. Firstly, it offloads computationally intensive tasks from the CPU to dedicated hardware accelerators, allowing the CPU to focus on other crucial AV operations. Examples include dedicated chips for tasks like matrix multiplication, a cornerstone of many LLM computations.

Secondly, hardware accelerators are often designed for specific tasks related to LLMs, leading to significant speedups compared to general-purpose CPUs. This translates to faster processing of sensor data and real-time decision-making for the AV. Finally, specialized hardware can be more energy-efficient than CPUs for specific tasks, reducing overall power consumption and extending the operational range of the AV.

Hardware acceleration with components like GPUs, AI accelerators, and FPGAs[100] is becoming increasingly important for deploying LLMs in resource-constrained environments like autonomous vehicles.

3.13 Limitations of Foundation Models

Hallucination [101] is a big problem of LLMs to avoid, which refers to a situation where the model generates content that is not based on factual or accurate information . Hallucination can occur when the model produces output that includes details, facts, or claims that are fictional, misleading, or entirely fabricated, rather than providing reliable and truthful information. Hallucination can be unintentional and may result from various factors, including biases in the training data, the model's lack of access to real-time or up-to-date information, or the inherent limitations of the model in comprehending and generating contextually accurate responses. Explainability [102] refers to the ability to explain or present the behavior of models in human-understandable terms. Improving the explainability of LLMs is crucial. With that, end users are able to understand the capabilities, limitations, and potential

Table 7: Different types of Edge Hardware for running a foundation model in real-time

Hardware Type	Example Products	Key Specifications	Advantages
GPUs (Edge-Focused)	NVIDIA Jetson Series (Nano, Xavier NX, AGX Orin)	CUDA Cores (for parallel processing), Tensor Cores (for AI workloads), Memory, Power Consumption	Wide software support, powerful for deep learning, growing focus on AV applications
AI Accelerators	Google Coral Edge TPU, Intel Movidius, Gyrfalcon Lightspeeur	Specialised neural network computation units, Power Efficiency (often measured in TOPS/Watt)	Optimised for ML inference, lower power than GPUs, potential for very low latency
System-on-Chips (SoCs)	Qualcomm Snapdragon Ride, Mobileye EyeQ Series	Integrate CPU, GPU, AI accelerators, may have AV-specific features	Purpose-built for AVs, balanced performance and efficiency, software toolkits
FPGAs	Xilinx Automotive-grade FPGAs	Customisable logic, very low latency potential, Power Efficiency	Flexibility to tailor hardware to model, but steeper learning curve for programming
Consumer Grade PCs	Custom PC with RTX 4080 NVIDIA Gaming GPU and an 8-core CPU	CUDA Cores, Overclocking, High Memory and Cores	Cheap, Easy to get, Customisable

flaws of LLMs. Besides, explainability acts as a debugging aid to quickly advance model performance on downstream tasks. From the application view, LLMs can handle high-level reasoning tasks such as question answering and commonsense reasoning. Understanding exclusive abilities of LLMs in in context learning and chain-of-thought prompting, as well as the phenomenon of hallucination, are indispensable to explaining and improving models.

4 Discussions

4.1 Research Gaps

This study focuses on improving the computational efficiency and edge deployment of foundation models within autonomous vehicles (AVs). Based on the above Literature Review The following key research gaps and challenges are addressed:

- **Balancing Accuracy and Efficiency:** Tailoring existing and developing novel compression, pruning, and knowledge distillation techniques are needed to reduce the size and computational demands of foundation models while minimising accuracy loss on safety-critical AV tasks.

- **Real-time Edge Performance and Efficiency:** Thorough benchmarking of AI accelerators within realistic AV workloads is essential to identify optimal hardware choices. This underscores the need for hardware-aware model optimisation tools and exploration of neuromorphic computing's potential to balance efficiency and performance.

- **Optimising Energy Consumption:** The development of energy-centric model architectures, training techniques, power profiling frameworks, and intelligent energy management systems are crucial for the extended operation of battery-powered AVs.

- **Intelligent and Robust Offloading:** Research into context-aware decision-making algorithms is needed to determine the optimal balance between edge and cloud execution. This includes ensuring robust operation under intermittent connectivity and investigating secure and privacy-preserving offloading protocols.

- **Addressing Data Requirements and Bias Mitigation:** Effective strategies for large-scale, unbiased data collection and labeling, along with the development of debiasing techniques for foundation models trained on driving data, are required to promote equitable performance across diverse scenarios. Exploring the potential of synthetic data generated with foundation models is promising to address data scarcity and bias issues.

- **Challenges of End-to-End Learning:** Improving the interpretability and debugging capabilities of end-to-end AV systems powered by foundation models is paramount for safety and explainability. Furthermore, seamless integration techniques that allow for graceful interaction between foundation models and traditional AV components warrant further investigation.
- **Need for Benchmarks:** The lack of standardized benchmarks for AV-specific edge deployment of foundation models hinders progress. Efforts towards comprehensive benchmark suites are essential.

4.2 Research Questions

In order to implement the mentioned tasks in an Automated vehicles we have to fix certain things first.Hence based on the above research gaps I have derived some of the important questions that needs to be solved first .

- RQ1: Is it feasible to develop a custom foundation model specifically tailored for object detection and planning in autonomous vehicles, and what are the key dataset, hardware, and edge compatibility requirements for such a model?
- RQ2: How can we optimise foundation models for object detection and planning tasks, ensuring real-time performance, low latency, and energy efficiency on edge devices in autonomous vehicles? Which compression, distillation, hardware acceleration, or novel techniques are most effective?
- RQ3: Can we design adaptive offloading strategies for object detection and planning tasks, intelligently partitioning foundation model execution between edge and cloud based on factors like scene complexity, object density, connectivity, and safety-criticality?
- RQ4: How can foundation models, used for object detection and planning, be effectively integrated with traditional methods in autonomous vehicles? Could this hybrid approach offer flexibility for different operational design domains (ODDs) or safety requirements?

4.3 Research Objective

Foundation models has various applications in Autonomous vehicles as mentioned in the literature review. Even though there are lot of methods and researched conducted to optimise the foundation model to run on edge device there is no single perfect solution as the performance of the processor increases every year with increasing transistor count. Any solution that is proposed this year may be completely irrelevant next year due to the drastic change in hardware performance or a new inventions of an AI architecture . Hence optimisation is a never ending problem that keeps on requiring improvements every year . In this project I am going to work on optimising the foundation model algorithm by implementing various methods that I mentioned above and by choosing the best hardware for running the model.

This research project aims to improve the efficiency and real-time performance of foundation models for object detection and planning in autonomous vehicles. The following specific objectives guide the investigation:

1. **Develop and Benchmark Novel Optimisation Techniques:**
    - Employ advanced quantization, pruning, and neural architecture search tailored to object detection and planning tasks in AVs.
    - Investigate knowledge distillation to create smaller models optimized for edge deployment, preserving accuracy on route planning and object detection.
    - Establish benchmarks reflecting real-world AV workloads for comprehensive evaluation of optimization techniques, measuring speed, memory, energy, and accuracy.

2. **Optimize Hardware-Software Co-Design:**

    - Benchmark diverse AI accelerators (TPUs, GPUs, potentially neuromorphic chips) on realistic foundation model workloads for AVs, considering power consumption, performance, and programmability.
    - Develop hardware-aware model optimization techniques, leveraging accelerator-specific features.
    - Explore the potential of neuromorphic architectures for long-term efficiency gains in object detection and planning tasks.
    - Design Adaptive and Robust Offloading Strategies:
    - Develop intelligent offloading algorithms that dynamically decide execution between edge and cloud, based on network conditions, task complexity, latency, and hardware capabilities.
    - Ensure seamless operation under intermittent connectivity, including graceful fallback to edge-only execution.
    - Investigate secure offloading techniques to protect user privacy when sensitive data is involved.

3. **Methodology** The project will employ a combination of theoretical analysis, algorithm development, and thorough experimentation. Key steps include:

    - Project Setup: Selection of technology stack, dataset acquisition (real-world and potentially synthetic), and choice of a pre-trained foundation model.
    - Hardware Exploration: Profiling of the baseline model, benchmarking of AI accelerators.
    - Model Optimization: Task-specific pruning and quantization, potentially in collaboration with hardware-specific optimization experts.
    - Knowledge Distillation: Design of a smaller student model and creation of customized loss functions for knowledge transfer.
    - Integration and Testing: Hardware-optimized deployment of the student model, evaluation on a realistic AV dataset with rigorous performance tracking.

4. **Expected Outcomes**

    This research is expected to yield:

    - Optimized Foundation Models: Foundation models tailored for object detection and planning in AVs, demonstrating improved efficiency on edge devices.
    - Hardware Insights: A comprehensive understanding of AI accelerator performance on AV-relevant workloads, aiding hardware selection.
    - Adaptive Offloading: Robust offloading strategies for intelligent resource utilization, ensuring safety and performance even under connectivity limitations.

5. **Significance**

    This work contributes to advancing the deployment of foundation models in resource-constrained AV environments, addressing challenges in computation, energy consumption, and real-time operation. Success in these areas paves the way towards more reliable and efficient autonomous driving systems.

6. **Important Considerations**

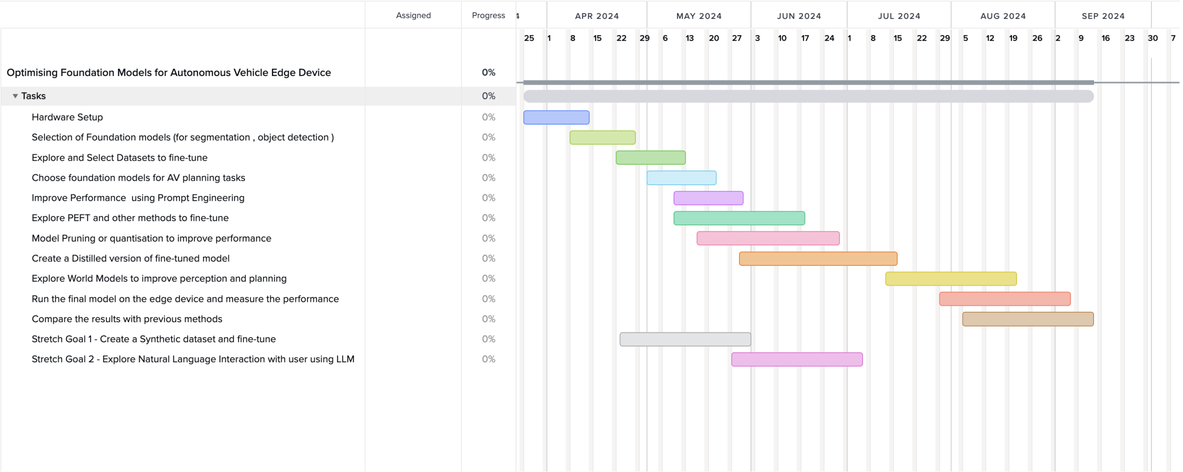
    - Datasets: Involves using realistic AV datasets (or simulators) that reflect the complexity of real-world scenarios.
    - Evaluation Metrics: Going beyond accuracy – will be including latency, FLOPs, and power consumption as key metrics.

5 Project Timeline

The tasks are listed in chronological order, starting with ”Hardware Setup” and ending with ”Compare the FM results with SOTA methods”. The Gantt chart shows that the project is expected to take approximately six months to complete, starting on March 25, 2024 and ending on September 12, 2024.

Here is a brief description of the tasks involved in the project:

1. Hardware Setup
2. Selection of Foundation models (for segmentation, object detection, tracking, BEV)
3. Select suitable Datasets to fine-tune based on Literature Review.
4. Choose foundation models for AV planning tasks
5. Improve Performance using Prompt Engineering
6. Use best PEFT method and other methods to fine-tune
7. Model Pruning or quantisation to improve performance
8. Create a Distilled version of fine-tuned model
9. Explore World Models to improve perception and planning.
10. Run the final model on the edge device and measure the performance
11. Compare the results with SOTA methods
12. Stretch Goal 1 Create a Synthetic dataset for fine-tuning
13. Stretch Goal 2- Explore Natural Language Interaction and VQA with the user using VFM.



	Assigned	Progress
Optimising Foundation Models for Autonomous Vehicle Edge Device		0%
▾ Tasks		0%
Hardware Setup		0%
Selection of Foundation models (for segmentation , object detection)		0%
Explore and Select Datasets to fine-tune		0%
Choose foundation models for AV planning tasks		0%
Improve Performance using Prompt Engineering		0%
Explore PEFT and other methods to fine-tune		0%
Model Pruning or quantisation to improve performance		0%
Create a Distilled version of fine-tuned model		0%
Explore World Models to improve perception and planning		0%
Run the final model on the edge device and measure the performance		0%
Compare the results with previous methods		0%
Stretch Goal 1 - Create a Synthetic dataset and fine-tune		0%
Stretch Goal 2 - Explore Natural Language Interaction with user using LLM		0%

Figure 10: Gantt-Chart for Project Timeline

6 Conclusion

Foundation models are poised to revolutionise autonomous driving, accelerating the push towards full Level 5 autonomy. Research trends focus on addressing core challenges like safety guarantees, real-time efficiency, and handling the vast complexity of real-world driving scenarios. Improving model explainability and developing rigorous verification methods are crucial for ensuring the trustworthiness of foundation model decisions on the road. Additionally, advancements in hardware-software co-design and optimisation techniques will enable seamless integration of these models into edge devices. The future will likely see hybrid approaches where foundation models work alongside traditional rule-based systems, leveraging their complementary strengths. By overcoming these challenges, foundation models hold the key to unlocking the full potential of Level 5 autonomous vehicles, promising safer, more efficient, and intelligent transportation systems.As research on foundation models for autonomous driving progresses, we can anticipate significant strides in making AVs safer, more intelligent, and capable of navigating the complexities of the real world. This technology holds the key to transforming the transportation landscape and redefining the relationship between humans and vehicles.

References

[1] "J3016_202104: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles - SAE International."

[2] Y. Huang, Y. Chen, Z. Li, and S. Member, "Applications of Large Scale Foundation Models for Autonomous Driving," 11 2023.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, 6 2017.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR 2021 - 9th International Conference on Learning Representations*, 10 2020.

[5] H. Gao, Y. Li, K. Long, M. Yang, and Y. Shen, "A Survey for Foundation Models in Autonomous Driving," 2 2024.

[6] Z. Yang, X. Jia, H. Li, and J. Yan, "LLM4Drive: A Survey of Large Language Models for Autonomous Driving,"

[7] "Knowledge Distillation : Simplified — by Prakhar Ganesh — Towards Data Science."

[8] J. Schneider, C. Meske, and P. Kuss, "Foundation Models: A New Paradigm for Artificial Intelligence," *Business and Information Systems Engineering*, pp. 1–11, 1 2024.

[9] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the Opportunities and Risks of Foundation Models," 8 2021.

[10] T.-H. Wang, A. Maalouf, W. Xiao, Y. Ban, A. Amini, G. Rosman, S. Karaman, D. Rus, and S. Jiao Tong University, "Drive Anywhere: Generalizable End-to-end Autonomous Driving with Multi-modal Foundation Models," 10 2023.

[11] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,"

[12] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13669 LNCS, pp. 1–18, 3 2022.

[13] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "BEVSegFormer: Bird's Eye View Semantic Segmentation From Arbitrary Camera Rigs,"

[14] H. Zhou, Z. Ge, Z. Li, and X. Zhang, "MatrixVT: Efficient Multi-Camera to BEV Transformation for 3D Perception," 11 2022.

[15] W. Chen, Y. Li, Z. Tian, and F. Zhang, "2D and 3D object detection algorithms from images: A Survey," *Array*, vol. 19, p. 100305, 9 2023.

[16] T. Guan, J. Wang, S. Lan, R. Chandra, Z. Wu, L. Davis, and D. Manocha, "M3DETR: Multi-representation, Multi-scale, Mutual-relation 3D Object Detection with Transformers," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2293–2303, IEEE, 1 2022.

[17] Z. Zhou, D. Ye, W. Chen, Y. Xie, Y. Wang, P. Wang, and H. Foroosh, "LiDARFormer: A Unified Transformer-based Multi-task Network for LiDAR Perception," 3 2023.

[18] Q. Zhou, X. Li, L. He, Y. Yang, G. Cheng, Y. Tong, L. Ma, and D. Tao, "TransVOD: End-to-End Video Object Detection With Spatial-Temporal Transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 7853–7869, 6 2023.

[19] Y. Zhang, J. Li, N. Jiang, G. Wu, H. Zhang, Z. Shi, Z. Liu, Z. Wu, and X. Liu, "Temporal Transformer Networks With Self-Supervision for Action Recognition," *IEEE Internet of Things Journal*, vol. 10, pp. 12999–13011, 7 2023.

[20] A. Singh, "Transformer-Based Sensor Fusion for Autonomous Driving: A Survey," 2 2023.

[21] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C. L. Tai, "TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 1080–1089, 2022.

[22] K. Chen, H. Zhu, D. Tang, and K. Zheng, "Future pedestrian location prediction in first-person videos for autonomous vehicles and social robots," *Image and Vision Computing*, vol. 134, p. 104671, 6 2023.

[23] W. Peebles, U. C. Berkeley, and S. Xie, "Scalable Diffusion Models with Transformers," pp. 4172–4182, 12 2022.

[24] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 2020-December, 5 2020.

[25] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, 10 2018.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of Machine Learning Research*, vol. 139, pp. 8748–8763, 2 2021.

[27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen OpenAI, "Hierarchical Text-Conditional Image Generation with CLIP Latents,"

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 10674–10685, 12 2021.

[29] J. B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. C. T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a Visual Language Model for Few-Shot Learning," *Advances in Neural Information Processing Systems*, vol. 35, 4 2022.

[30] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "PaLM: Scaling Language Modeling with Pathways," 4 2022.

[31] "Home \ Anthropic."

[32] "Google AI."

[33] "Hugging Face – The AI community building the future.."

[34] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9630–9640, 4 2021.

[35] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning Robust Visual Features without Supervision," 4 2023.

[36] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything,"

[37] D. Ha and J. Urgen Schmidhuber, "World Models,"

[38] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, G. Corrado, and W. R. Ai, "GAIA-1: A Generative World Model for Autonomous Driving," 9 2023.

[39] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the Future: Multiview Visual Forecasting and Planning with World Model for Autonomous Driving," 11 2023.

[40] D. Bogdoll, Y. Yang, and J. M. Zöllner, "MUVO: A Multimodal Generative World Model for Autonomous Driving with Geometric Representations," 11 2023.

[41] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "TrafficBots: Towards World Models for Autonomous Driving Simulation and Motion Prediction," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2023-May, pp. 1522–1529, 3 2023.

[42] "LINGO-1: Exploring Natural Language for Autonomous Driving - Wayve."

[43] Y. Jin, X. Shen, H. Peng, X. Liu, J. Qin, J. Li, J. Xie, P. Gao, G. Zhou, and J. Gong, "SurrealDriver: Designing Generative Driver Agent Simulation Framework in Urban Contexts based on Large Language Model," *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03â•fi05, 2018, Woodstock, NY*, vol. 1, 9 2023.

[44] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” 3 2015.

[45] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 6 2020.

[46] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” 10 2019.

[47] Y. Zhu, Y. Wang, H. Shi, and S. Tang, “Efficient Tuning and Inference for Large Language Models on Textual Graphs,” 1 2024.

[48] R. K. Mahabadi, S. R. Deepmind, M. Dehghani, G. Brain, and J. Henderson, “Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks,”

[49] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, “DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model,” 10 2023.

[50] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, “Pruning and Quantization for Deep Neural Network Acceleration: A Survey,”

[51] M. Sun, Z. Liu, A. Bair, and J. Zico Kolter, “A SIMPLE AND EFFECTIVE PRUNING APPROACH FOR LARGE LANGUAGE MODELS,”

[52] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, “A Survey on Model Compression for Large Language Models,” 8 2023.

[53] S. Bai, J. Chen, X. Shen, Y. Qian, and Y. Liu, “Unified Data-Free Compression: Pruning and Quantization without Fine-Tuning,”

[54] Z. Lyu, T. Yu, F. Pan, Y. Zhang, J. Luo, D. Zhang, Y. Chen, B. Zhang, and G. Li, “A survey of model compression strategies for object detection,” *Multimedia Tools and Applications 2023*, pp. 1–72, 11 2023.

[55] “Zhen-Dong/Awesome-Quantization-Papers: List of papers related to neural network quantization in recent AI conferences and journals..”

[56] “ModelTC/lightllm: LightLLM is a Python-based LLM (Large Language Model) inference and serving framework, notable for its lightweight design, easy scalability, and high-speed performance..”

[57] “NVIDIA/TensorRT-LLM: TensorRT-LLM provides users with an easy-to-use Python API to define Large Language Models (LLMs) and build TensorRT engines that contain state-of-the-art optimizations to perform inference efficiently on NVIDIA GPUs. TensorRT-LLM also contains components to create Python and C++ runtimes that execute those TensorRT engines..”

[58] “NVIDIA/FasterTransformer: Transformer related optimization, including BERT, GPT.”

[59] “microsoft/DeepSpeed: DeepSpeed is a deep learning optimization library that makes distributed training and inference easy, efficient, and effective..”

[60] “microsoft/DeepSpeed-MII: MII makes low-latency and high-throughput inference possible, powered by DeepSpeed..”

[61] “vllm-project/vllm: A high-throughput and memory-efficient inference and serving engine for LLMs.”

[62] “LangChain.”

[63] “langchain 0.1.12 — LangChain 0.1.12.”

[64] X. Yan, H. Zhang, Y. Cai, J. Guo, W. Qiu, B. Gao, K. Zhou, Y. Zhao, H. Jin, J. Gao, Z. Li, L. Jiang, W. Zhang, H. Zhang, D. Dai, and B. Liu, “Forging Vision Foundation Models for Autonomous Driving: Challenges, Methodologies, and Opportunities,” 1 2024.

[65] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundational Models Defining a New Era in Vision: A Survey and Outlook," 7 2023.

[66] b. N. Sriram N, T. Maniar, J. Kalyanasundaram, V. Gandhi, M. Krishna, S. N. N, B. Bhowmick, and K. Madhava Krishna, "Talk to the Vehicle: Language Conditioned Autonomous Navigation of Self Driving Cars," 2019.

[67] B. Jin, X. Liu, Y. Zheng, P. Li, H. Zhao, T. Zhang, Y. Zheng, G. Zhou, and J. Liu, "ADAPT: Action-aware Driving Caption Transformer," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2023-May, pp. 7554–7561, 2 2023.

[68] Y. Meng, S. Vemprala, R. Bonatti, C. Fan, and A. Kapoor, "ConBaT: Control Barrier Transformer for Safe Policy Learning," 3 2023.

[69] J. Liu, P. Hang, X. qi, J. Wang, and J. Sun, "MTD-GPT: A Multi-Task Decision-Making GPT Model for Autonomous Driving at Unsignalized Intersections," pp. 5154–5161, 7 2023.

[70] P. Wang, M. Zhu, H. Lu, H. Zhong, X. Chen, S. Shen, X. Wang, and Y. Wang, "BEVGPT: Generative Pre-trained Large Model for Autonomous Driving Prediction, Decision-Making, and Planning," 10 2023.

[71] P. Wu, L. Chen, H. Li, X. Jia, J. Yan, and Y. Qiao, "Policy Pre-training for Autonomous Driving via Self-supervised Geometric Modeling," 1 2023.

[72] J. Yuan, B. Zhang, X. Yan, T. Chen, B. Shi, Y. Li, and Y. Qiao, "AD-PT: Autonomous Driving Pre-Training with Large-scale Point Cloud Dataset," 6 2023.

[73] H. Yang, S. Zhang, D. Huang, X. Wu, H. Zhu, T. He, S. Tang, H. Zhao, Q. Qiu, B. Lin, X. He, and W. Ouyang, "UniPAD: A Universal Pre-training Paradigm for Autonomous Driving," 10 2023.

[74] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M.-F. Moens, "Talk2Car: Taking Control of Your Self-Driving Car," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 2088–2098, 9 2019.

[75] S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "OpenScene: 3D Scene Understanding with Open Vocabularies," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June, pp. 815–824, 11 2022.

[76] W. Cheng, J. Yin, W. Li, I. Shanghai, C. R. Yang, and C. J. Shen, "Language-Guided 3D Object Detection in Point Cloud for Autonomous Driving," *Proceedings of (Preprint)*, vol. 1, 5 2023.

[77] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "HiLM-D: Towards High-Resolution Understanding in Multimodal Large Language Models for Autonomous Driving," 9 2023.

[78] D. Wu, W. Han, T. Wang, Y. Liu, X. Zhang, and J. Shen, "Language Prompt for Autonomous Driving," 9 2023.

[79] M. Najibi, J. Ji, Y. Zhou, C. R. Qi, X. Yan, S. Ettinger, D. Anguelov, and W. Llc, "Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving," pp. 8568–8578, 9 2023.

[80] Y. Zhou, L. Cai, X. Cheng, Z. Gan, X. Xue, and W. Ding, "OpenAnnotate3D: Open-Vocabulary Auto-Labeling System for Multi-modal 3D Data," 10 2023.

[81] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, Y. Qiao, and S. A. Lab, "Drive Like a Human: Rethinking Autonomous Driving with Large Language Models," 7 2023.

[82] "LINGO-1: Exploring Natural Language for Autonomous Driving - Wayve."

[83] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Drive as You Speak: Enabling Human-Like Interaction with Large Language Models in Autonomous Vehicles," 9 2023.

[84] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models," 9 2023.

[85] H. Sha, Y. Mu, Y. Jiang, G. Zhan, L. Chen, C. Xu, P. Luo, S. E. Li, M. Tomizuka, W. Zhan, and M. Ding, "LanguageMPC: Large Language Models as Decision Makers for Autonomous Driving," 10 2023.

[86] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "GPT-Driver: Learning to Drive with GPT," 10 2023.

[87] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, J. Shotton, and W. R. Ai, "Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving," 10 2023.

[88] T. Choudhary, V. Dewangan, S. Chandhok, S. Priyadarshan, A. Jain, A. K. Singh, S. Srivastava, K. Murthy Jatavallabhula, K. Madhava Krishna, and I. Hyderabad, "Talk2BEV: Language-enhanced Bird's-eye View Maps for Autonomous Driving," 10 2023.

[89] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li, "DriveLM: Driving with Graph Visual Question Answering," 12 2023.

[90] T.-H. Wang, A. Maalouf, W. Xiao, Y. Ban, A. Amini, G. Rosman, S. Karaman, D. Rus, and S. Jiao Tong University, "Drive Anywhere: Generalizable End-to-end Autonomous Driving with Multi-modal Foundation Models," 10 2023.

[91] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A Language Agent for Autonomous Driving," 11 2023.

[92] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual Explanations for Self-Driving Vehicles," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11206 LNCS, pp. 577–593, 7 2018.

[93] J. Kim, T. Misu, Y. T. Chen, A. Tawari, and J. Canny, "Grounding Human-to-Vehicle Advice for Self-driving Vehicles," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 10583–10591, 11 2019.

[94] S. Malla, C. Choi, I. Dwivedi, J. Hee Choi, and J. Li, "DRAMA: Joint Risk Localization and Captioning in Driving," *Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023*, pp. 1043–1052, 9 2022.

[95] E. Sachdeva, N. Agarwal, S. Chundi, S. Roelofs, J. Li, M. Kochenderfer, C. Choi, and B. Dariush, "Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning,"

[96] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "NuScenes-QA: A Multi-modal Visual Question Answering Benchmark for Autonomous Driving Scenario," 5 2023.

[97] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2Drive: Towards Interpretable and Chain-based Reasoning for Autonomous Driving," 12 2023.

[98] A.-M. Marcu, L. Chen, J. Hünermann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton, and O. Sinavski, "LingoQA: Video Question Answering for Autonomous Driving," 12 2023.

[99] R. R. Schaller, "Moore's law: past, present, and future," *IEEE Spectrum*, vol. 34, no. 6, 1997.

[100] M. Sun, Z. Li, A. Lu, Y. Li, S. E. Chang, X. Ma, X. Lin, and Z. Fang, "FILM-QNN: Efficient FPGA Acceleration of Deep Neural Networks with Intra-Layer, Mixed-Precision Quantization," *FPGA 2022 - Proceedings of the 2022 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 134–145, 2 2022.

[101] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. Tuan Luu, W. Bi, F. Shi, S. Shi, and T. A. lab, "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," 9 2023.

[102] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for Large Language Models: A Survey," *ACM Transactions on Intelligent Systems and Technology*, 9 2023.