A. Preliminary

A.1. Multimodal LLMs

LLMs have attracted considerable attention for their remarkable capabilities across various linguistic tasks. This wave of interest has paved the way for the development of recent Multimodal Large Language Models (MLLMs), which are built upon strong pre-trained LLMs and consume both visual and text tokens by integrating pre-trained vision encoders. Modality alignment models like CLIP [132], which significantly narrow the semantic gap between language and visual information, lay the foundation of MLLMs. Early efforts such as Flamingo [2] and BLIP-2 [92] further enhance cross-modal alignment with intensive image-text datasets. GPT-4o [73], Claude 3 [4], Gemini [150] are popular closed-source MLLMs that can support multimodal inputs from the users. Several open-source MLLMs have also been developed. LLaVA [104] is trained end-to-end using instruction-tuning with language-only GPT-4 generated multimodal instruction-following data. Qwen-VL [10] augments MLLMs with visual grounding capabilities, allowing detecting objects in accordance with user prompts. InstructBLIP [38] utilizes a Q-Former [92] to extract instruction-aware visual features from the output embeddings of a frozen image encoder. These visual features are then fed as soft prompt input to a frozen LLM. MiniGPT-4 [204] combines a frozen visual encoder with a frozen LLM using a single trainable projection layer. This allows for efficient integration of visual and textual information while minimizing the number of trainable parameters. Other notable open-source examples include CogVLM [162], InternLM-XComposer [39], etc. These advancements highlight the diverse and expanding landscape of MLLMs, which has remarkably facilitated the application and iteration of MLLMs.

A.2. AI Alignment

AI alignment [77] has become a major focus in the machine learning community as AI models grow increasingly capable. With the potential emergence of superintelligence on the horizon, ensuring that AI models reflect human values is of critical importance. AI alignment seeks to guide AI systems to behave according to human intentions. As AI systems increase in capability, the risks from misalignment also grow [207].

Broadly, AI alignment can be divided into two directions: forward alignment and backward alignment. Forward alignment focuses on aligning the AI model with human values through alignment training. This involves techniques such as learning from feedback, which includes preference modeling [7, 133, 174] and reinforcement learning from human feedback (RLHF) [124], and adapting to distribution shifts (or spurious correlations) [168, 186] to ensure robust alignment across varying contexts. The goal is to shape AI systems during their development to act according to human intentions and ethical considerations. Backward alignment, on the other hand, focuses on evaluating AI systems after their development to ensure continued alignment with human values. This involves assurance techniques, such as safety evaluations [191], interpretability methods [141], and human values verification [140], as well as governance practices aimed at managing AI risks. These processes aim to refine alignment by identifying and addressing misalignment risks throughout the AI system's lifecycle, including post-deployment scenarios.

A.3. State Space Models (SSM) and Mamba

Modern state space models (SSMs), derived from the classical state space one [82], are commonly considered linear time-invariant systems that map an one-dimensional sequence $x(t) \rightarrow y(t) \in \mathbb{R}$ through a hidden state $h(t) \in \mathbb{R}^N$. Subsequently, the seminal work of [55] introduced the structured state space model (S4), which further employed discrete versions of the aforementioned continuous system to be appropriate for deep learning algorithms and modifies the $\mathcal{A}$ matrix by low-rank correction [54]. A recent work, called Mamba [53], further improves S4 with a selection mechanism and a hardware parallel scanning algorithm to solve Transformer's quadratic complexity calculation problem [154] in long sequence modeling and Transformer's inability to model data outside the attention window. Mamba's superior scaling performance shows that it has emerged as a powerful long-sequence modeling approach with linear complexity in terms of input size, shedding light on efficient modeling of both local and global dependencies.

Despite being devised to address Natural Language Processing (NLP) tasks, Mamba also shows effective performance in vision fields [156, 173, 182]. Vision Mamba (Vim) [205] improves Mamba with bidirectional SSM scanning mode for data-dependent global visual context modeling and position embedding for location-aware visual understanding. VMamba [108] introduces a 2D selective scan mode to bridge 1D array scanning and 2D plane traversal, enabling the extension of selective SSMs to process vision data. In addition, many works further extend Mamba in multi-modal fusion [97, 155, 171, 189]. Cobra [200] first extends Mamba to multi-modal large language models with linear computational complexity. VL-Mamba [130] exploits SSMs in solving multimodal learning tasks, providing a novel framework option for multimodal large language models other than transformer-based architectures. These studies highlight the promising potential of Mamba across various multimodal scenarios, demonstrating its capacity to enhance autonomous driving, particularly by improving system speed.