ViT是计算机视觉和自然语言处理两个领域的融合成果。它使用transformer架构来处理图像数据，这种架构原本是用于处理自然语言的。

ViT的主要思想是将图像分割成固定大小的块（patches），再将块转换成序列形式（自然语言中就是词语序列）。ViT就是利用transformer的编码器来处理这些序列的。

此外ViT引入了visual token视觉标记的概念，它通过在输入的图像块中添加一些特定的位置编码信息，使模型可以处理不同位置和内容的图像信息。

Swin Transformer

**核心创新点**

* **引入滑动窗口机制**：Swin Transformer 提出了一种包含滑窗操作的层级设计，使用层级式的滑动窗口在窗口内计算注意力，将图片切分成若干个窗口，自注意力的计算在局部的非重叠窗口内进行。这种方式一方面减少了计算的复杂度，使复杂度与图像大小呈线性关系，另一方面，由于采用非重叠窗口，自注意计算时不同 query 会共享同样的 key 集合，从而对硬件友好，更实用1.
* **具有层级结构**：借鉴了卷积神经网络的层次化构建方法，其网络结构具有多个 Stage，每个 Stage 都会缩小输入特征的分辨率，同时通道数翻倍，从而能够提取多尺度的特征信息，适用于多种 CV 任务，如目标检测、实例分割等 。这种层级结构有助于更好地理解图像的不同层次的语义信息，与传统的 Vision Transformer 相比，更符合人类视觉系统对图像的感知方式1

**Combining sensing technology and big language modeling: a convergence path from perception to intelligent decision making**

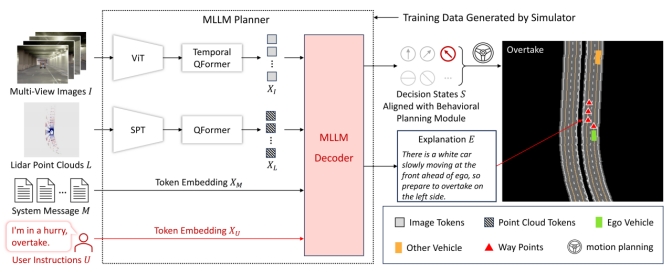
Wang Jibo

Artificial Intelligence Institute ,School of Energy and Mining Engineering,China University of Mining and Technology-Beijing,Beijing 100083,China;

**Abstract:** With the rapid development of Artificial Intelligence (AI), Large Language Models (LLMs) have become an important tool for information processing and decision support. Meanwhile, advances in sensing technology have provided efficient means for perception and data acquisition, especially in the fields of autonomous driving, smart cities, and robotics, where the combination of sensors and AI has begun to change the traditional decision-making process. This review paper examines the current state of the art in the fusion of sensing technology and big language models, discusses how the advantages of both can be utilized in the process of data acquisition, processing, fusion, and perception-decision integration to achieve intelligent systems, focuses on the challenges and potentials of this fusion, and looks at future research directions.

**Keywords:** sensing technology, big language modeling, deep learning, data fusion, autonomous driving, smart cities, robotics

1. Xu Z, Zhang Y, Xie E, et al. Drivegpt4: Interpretable end-to-end autonomous driving via large language model[J]. IEEE Robotics and Automation Letters, 2024.
2. Wang W, Xie J, Hu C Y, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving[J]. arXiv preprint arXiv:2312.09245, 2023.
3. Li Z, Xia L, Shi L, et al. OpenCity: Open Spatio-Temporal Foundation Models for Traffic Prediction[J]. arXiv preprint arXiv:2408.10269, 2024.
4. Abbasian Dehkordi S, Farajzadeh K, Rezazadeh J, et al. A survey on data aggregation techniques in IoT sensor networks[J]. Wireless Networks, 2020, 26(2): 1243-1263.
5. Santos P M, Rodrigues J G P, Cruz S B, et al. PortoLivingLab: An IoT-based sensing platform for smart cities[J]. IEEE Internet of Things Journal, 2018, 5(2): 523-532.



**图1 基于多模态大语言模型（MLLM）的自动驾驶系统架构**

**Fig. 1 Automated driving system architecture based on multimodal large language model (MLLM)**

图1便是一个基于大模型的自动驾驶架构，此架构与传感技术结合形成一个自动驾驶系统。