(a) Conventional End-to-End AD

Env. Info Sensor Input → E2E Network → Driving Actions

(b) Vision Language Models (VLMs) for AD

Env. Info Sensor Input → VLMs ← Finetune — Database

VLMs → Reasoning Chains → Multi Tasks

(c) Vision-Language-Action (VLA) Models for AD

Env. Info Sensor Input → Multi-Modal Encoder → LLM/VLMs

Action Decoder → Output → Driving Actions