Artificial Intelligence

Generative AI, Agentic AI, and Physical AI

1141AI09
MBA, IM, NTPU (M5276) (Fall 2025)
Tue 2, 3, 4 (9:10-12:00) (B3F17)

https://meet.google.com/paj-zhhj-mya

Min-Yuh Day, Ph.D,
Professor and Director
Institute of Information Management, National Taipei University
https://web.ntpu.edu.tw/~myday

2025-12-02

Syllabus

Week Date Subject/Topics

1 2025/09/09 Introduction to Artificial Intelligence

2 2025/09/16 Artificial Intelligence and Intelligent Agents;
Problem Solving

3 2025/09/23 Knowledge, Reasoning and Knowledge Representation;
Uncertain Knowledge and Reasoning

4 2025/09/30 Case Study on Artificial Intelligence I

5 2025/10/07 Machine Learning: Supervised and Unsupervised Learning;
The Theory of Learning and Ensemble Learning

Syllabus

Week Date Subject/Topics

6 2025/10/14 NVIDIA Fundamentals of Deep Learning I:
Deep Learning; Neural Networks

7 2025/10/21 NVIDIA Fundamentals of Deep Learning II:
Convolutional Neural Networks;
Data Augmentation and Deployment

8 2025/10/28 Self-Learning

9 2025/11/04 Midterm Project Report

10 2025/11/11 NVIDIA Fundamentals of Deep Learning III:
Pre-trained Models; Natural Language Processing

Syllabus

Week Date Subject/Topics

11 2025/11/18 Case Study on Artificial Intelligence II

12 2025/11/25 Computer Vision and Robotics

13 2025/12/02 Generative AI, Agentic AI, and Physical AI

14 2025/12/09 Philosophy and Ethics of AI and the Future of AI

15 2025/12/16 Final Project Report I

16 2025/12/23 Final Project Report II

Generative AI,
Agentic AI,
and Physical AI

Outline

- Generative AI
- Agentic AI
- Physical AI (Robotics)

Generative AI, Agentic AI, Physical AI

Physical AI
Self-driving cars
General robotics

Agentic AI
Coding assistants
Customer service
Patient care

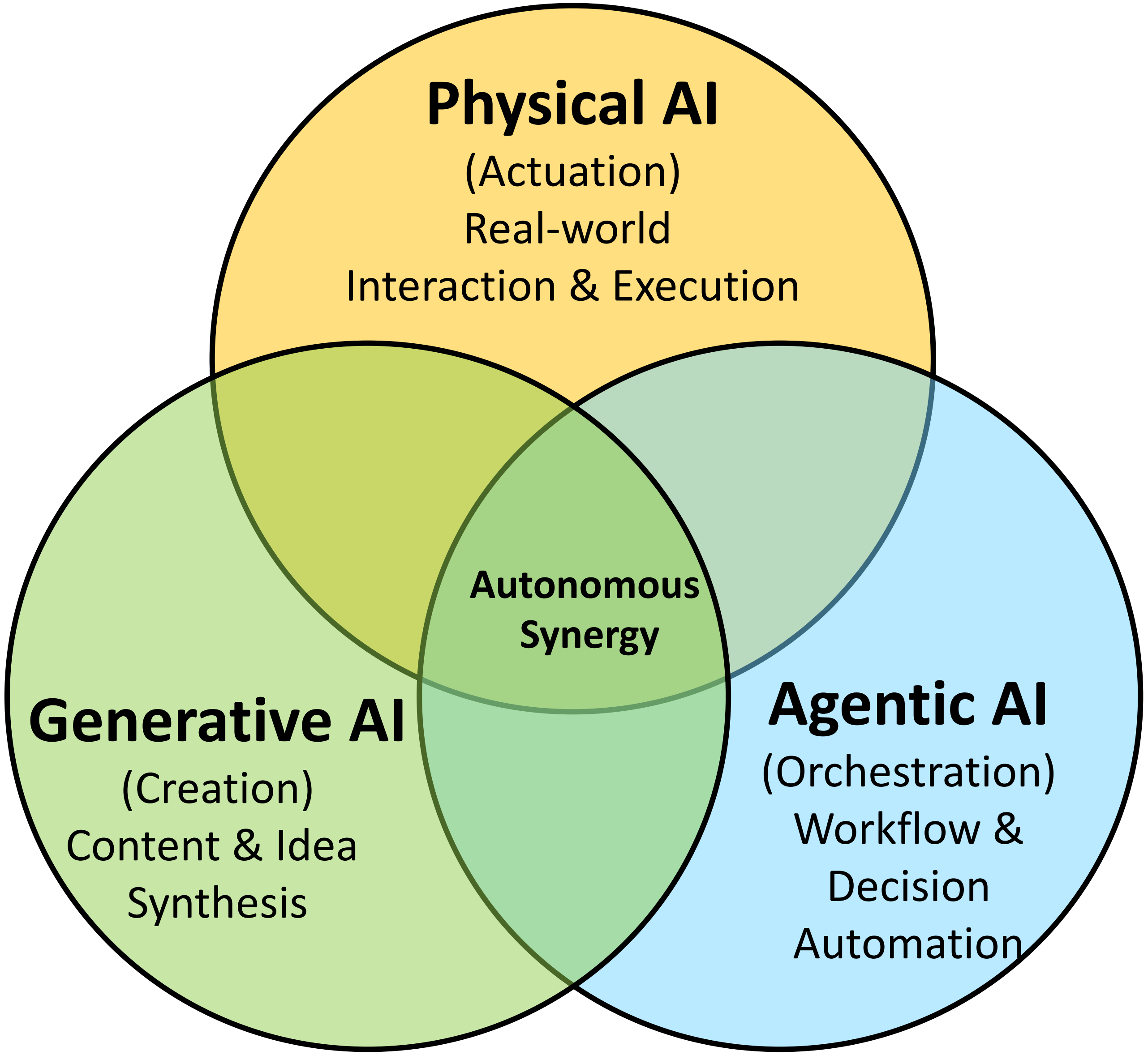
Generative AI
Digital marketing
Content creation

Perception AI
Speech recognition
Deep recommender systems
Medical imaging

2012 AlexNet
Deep learning breakthrough

Source: NVIDIA (2025), GTC March 2025 Keynote with NVIDIA CEO Jensen Huang, https://www.youtube.com/watch?v=_waPvOwL9Z8

Generative AI, Agentic AI, Physical AI

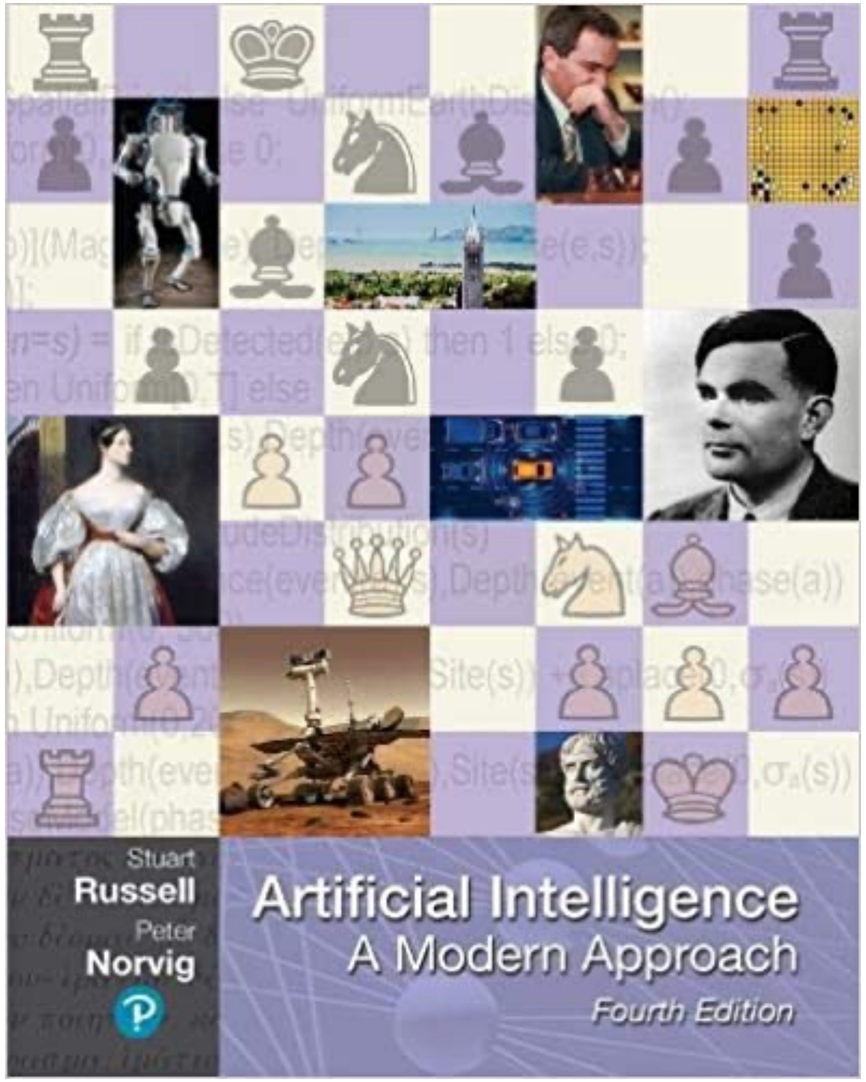


New Economic Paradigm Shift: From Creation to Execution

Stuart Russell and Peter Norvig (2020),

Artificial Intelligence: A Modern Approach,

4th Edition, Pearson



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

https://www.amazon.com/Artificial-Intelligence-A-Modern-Approach/dp/0134610997/

Artificial Intelligence: A Modern Approach

1. Artificial Intelligence
2. Problem Solving
3. Knowledge and Reasoning
4. Uncertain Knowledge and Reasoning
5. Machine Learning
6. Communicating, Perceiving, and Acting
7. Philosophy and Ethics of AI

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Artificial Intelligence: Communicating, perceiving, and acting

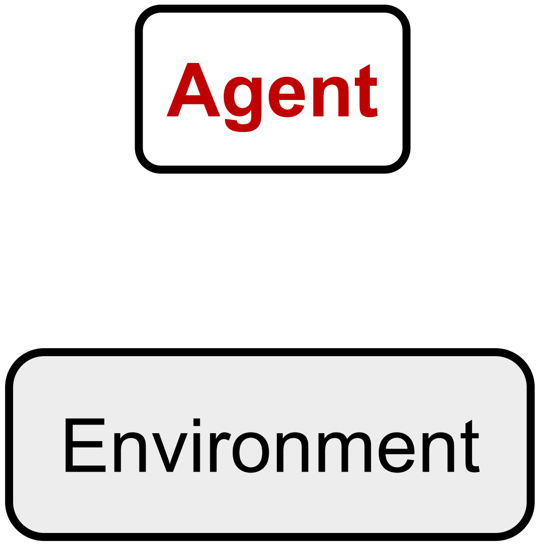
Artificial Intelligence:

6. Communicating, Perceiving, and Acting

- Natural Language Processing
- Deep Learning for Natural Language Processing
- Computer Vision
- Robotics

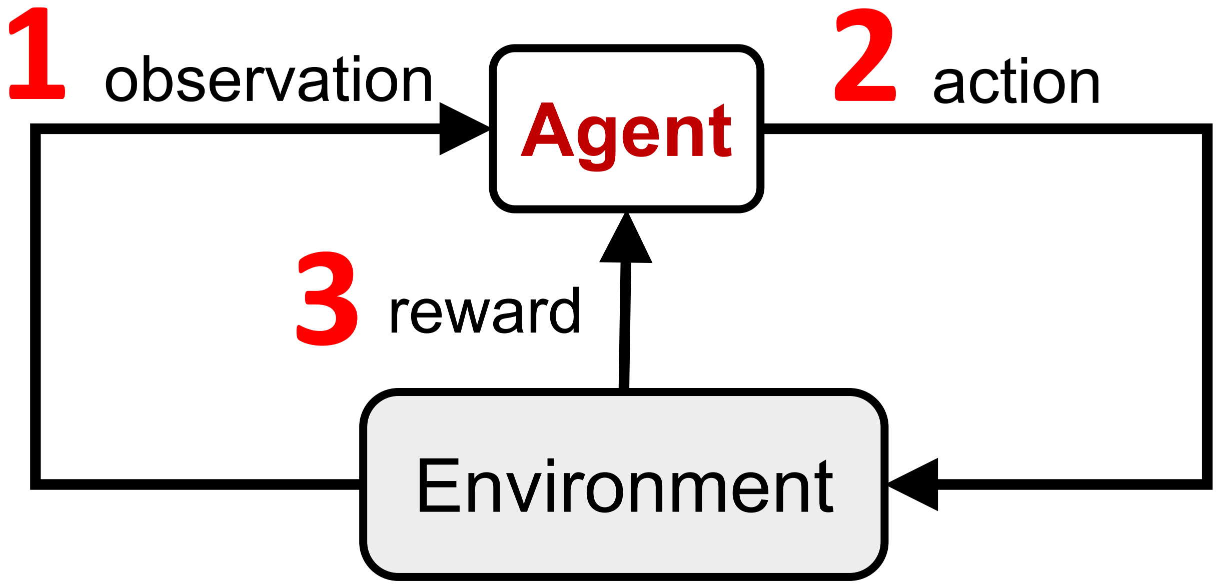
Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Reinforcement Learning (DL)



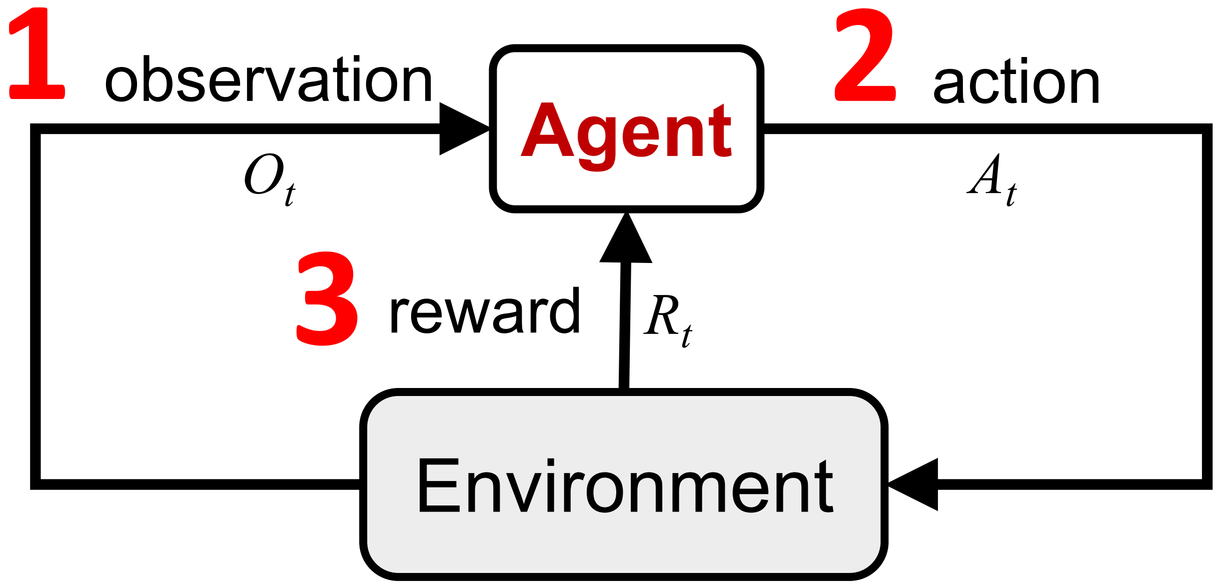
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



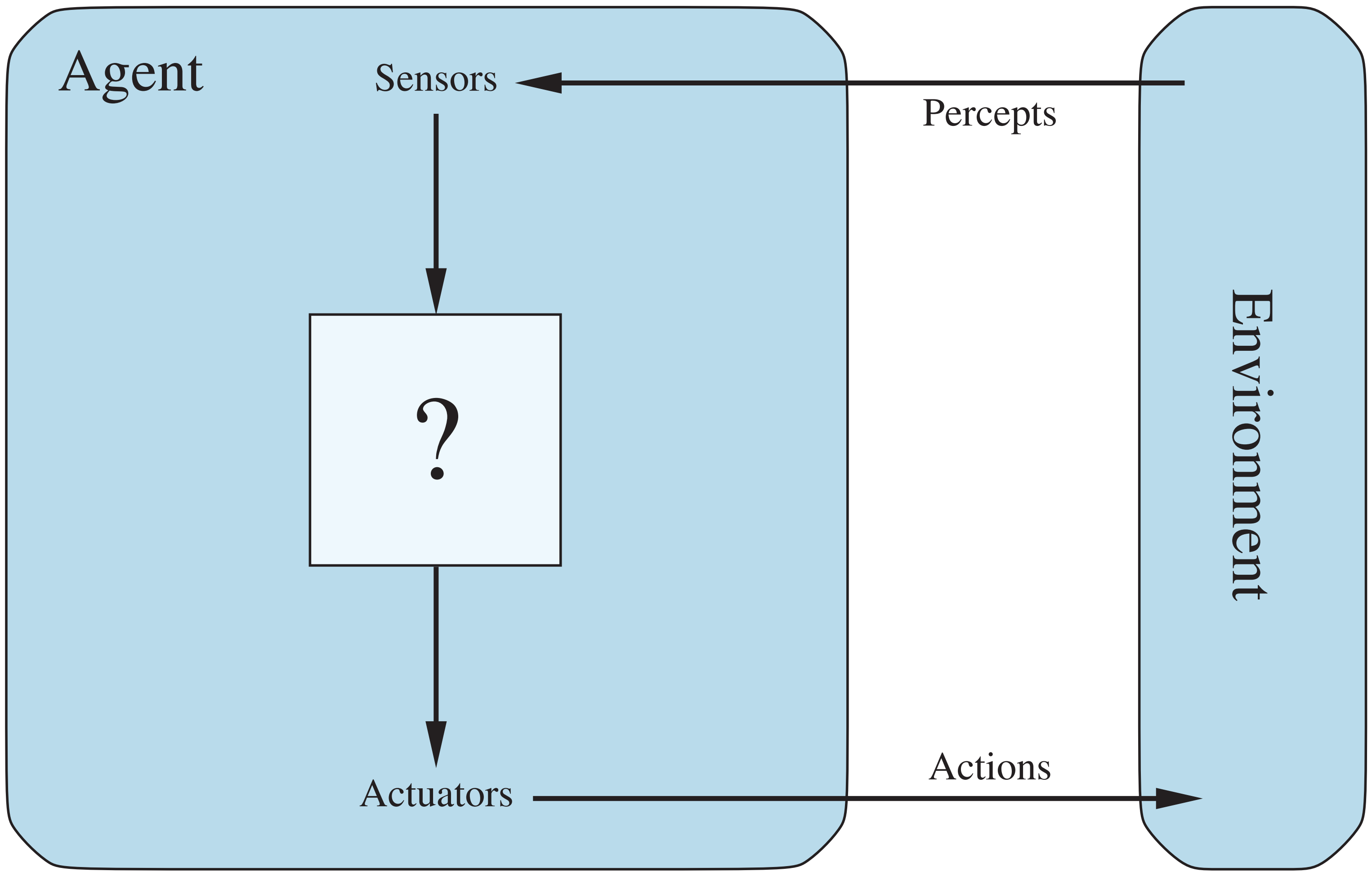
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Agents interact with environments through sensors and actuators



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

AI Acting Humanly: The Turing Test Approach

(Alan Turing, 1950)

- Knowledge Representation
- Automated Reasoning
- Machine Learning (ML)
  - Deep Learning (DL)
- Computer Vision (Image, Video)
- Natural Language Processing (NLP)
- Robotics

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

4 Approaches of AI

2. Thinking Humanly: The Cognitive Modeling Approach	3. Thinking Rationally: The "Laws of Thought" Approach
1. Acting Humanly: The Turing Test Approach (1950)	4. Acting Rationally: The Rational Agent Approach

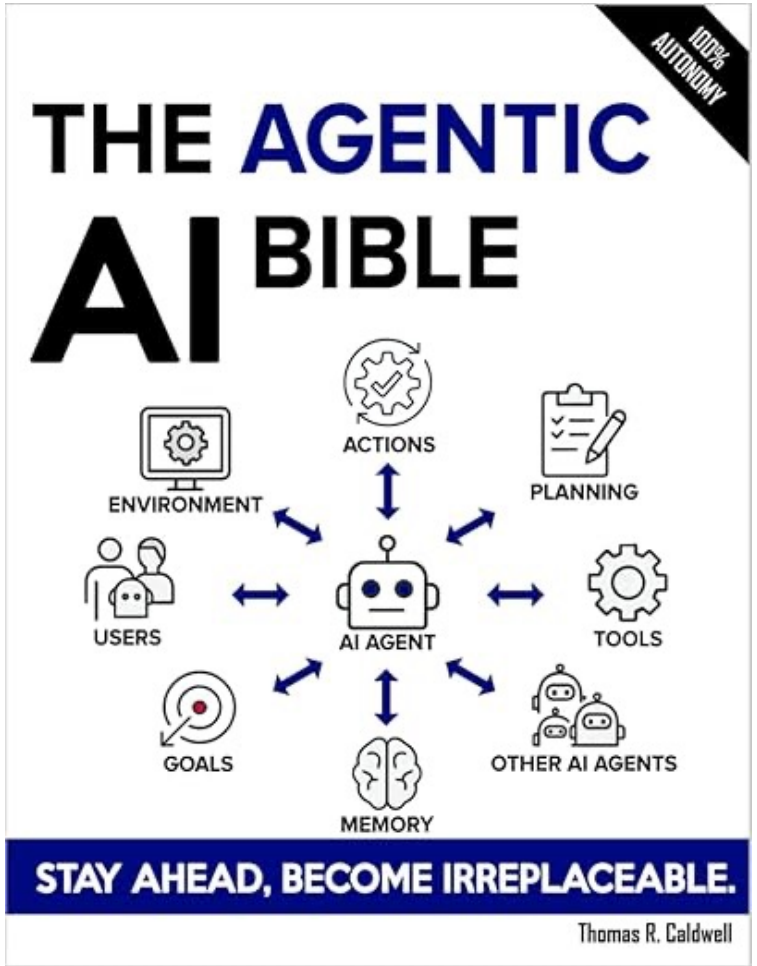
Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Thomas R. Caldwell (2025),

The Agentic AI Bible:

The Complete and Up-to-Date Guide to Design, Build, and Scale Goal-Driven, LLM-Powered Agents that Think, Execute and Evolve,

Independently published



Source: https://www.amazon.com/Agentic-Bible-Up-Date-Goal-Driven/dp/B0FL21R86Q

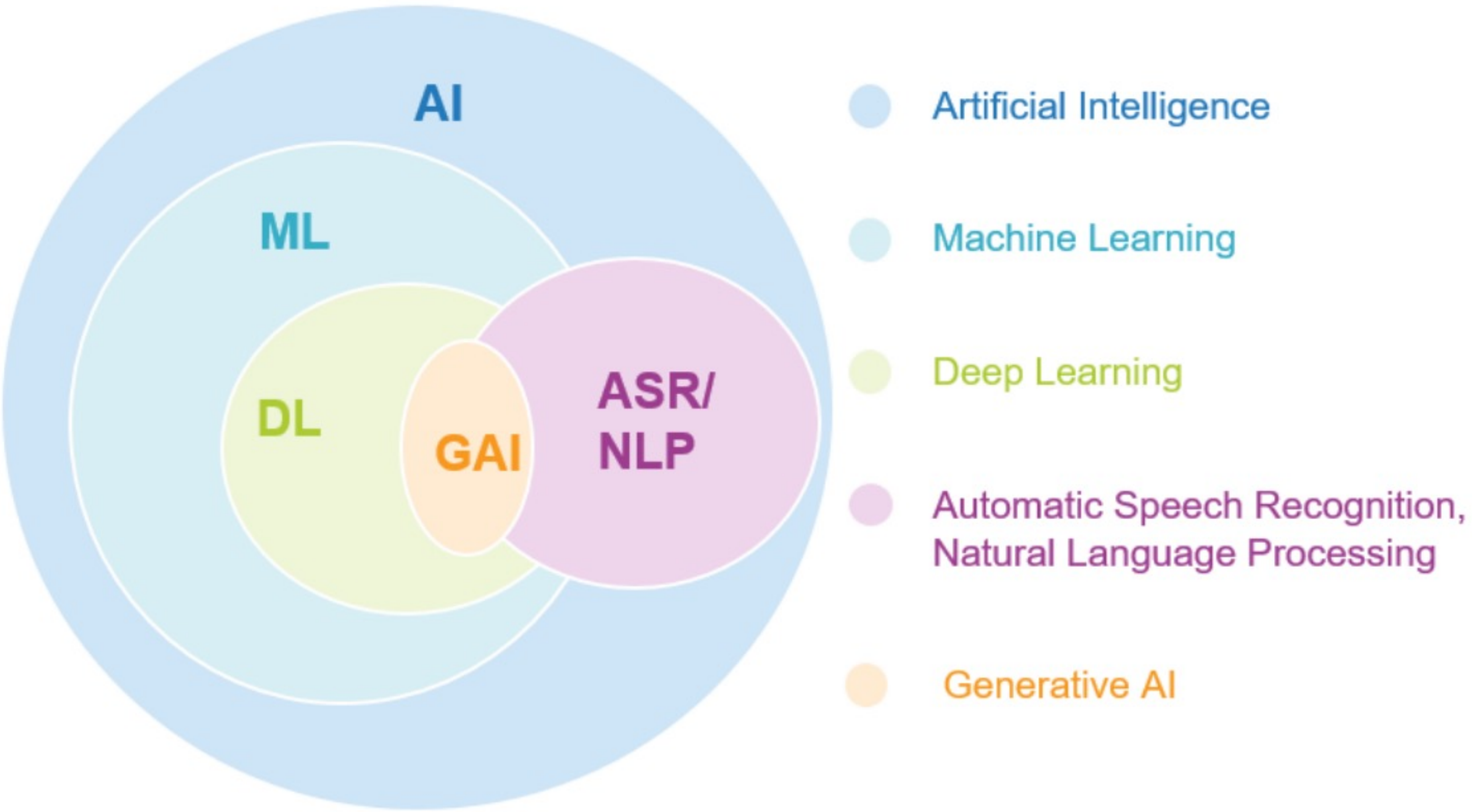
Generative AI-Driven ESG Report Generation Technology

Industrial Technology Research Institute (ITRI),
Fintech and Green Finance Center (FGFC, NTPU),
NTPU-113A513E01, 2024/03/01~2024/12/31

Innovative Agentic AI Technology for Autonomous ESG Report Generation

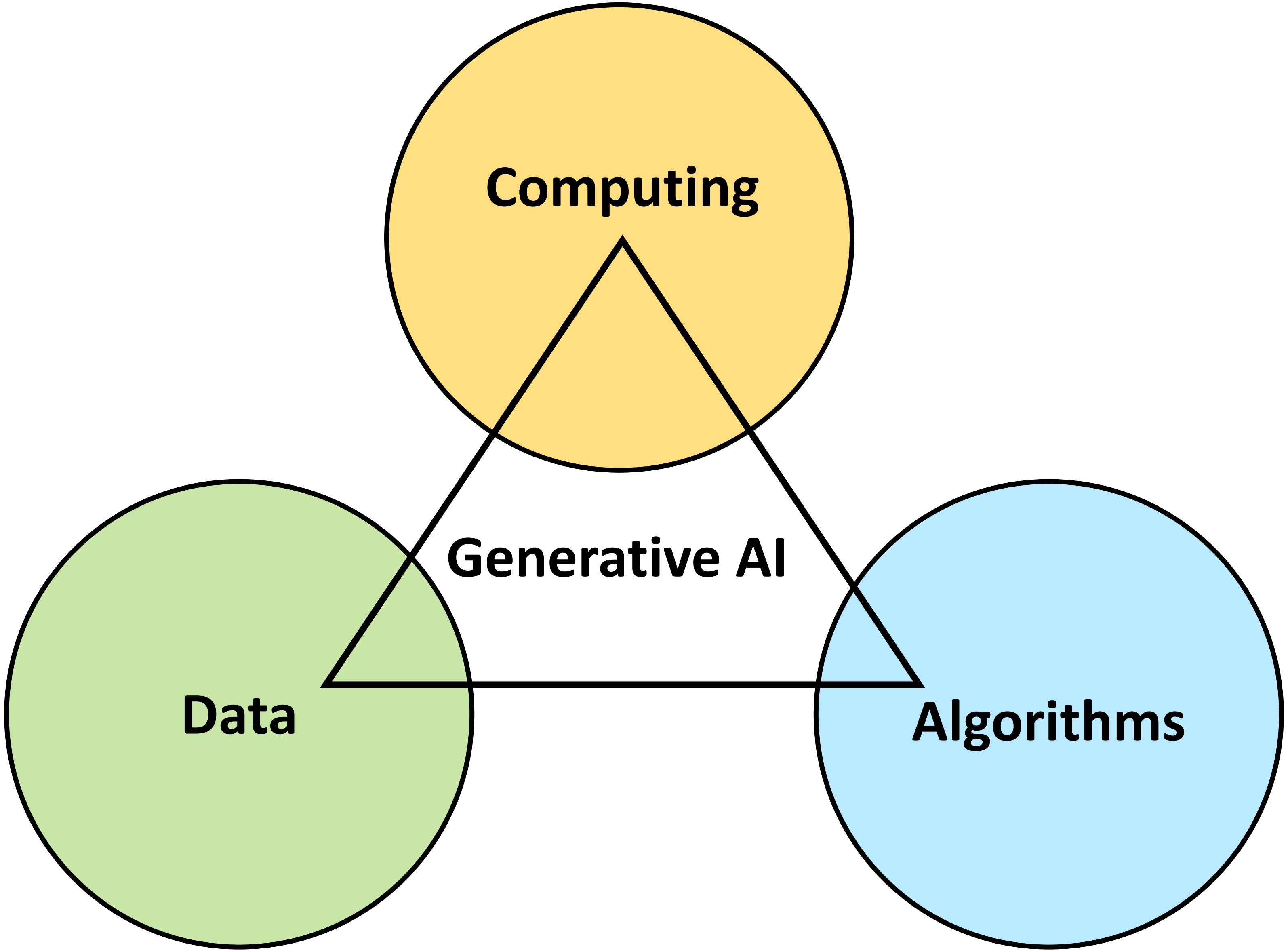
Industrial Technology Research Institute (ITRI),
Fintech and Green Finance Center (FGFC, NTPU),
NTPU-114A513E01, 2025/03/01~2025/12/31

AI, ML, DL, Generative AI

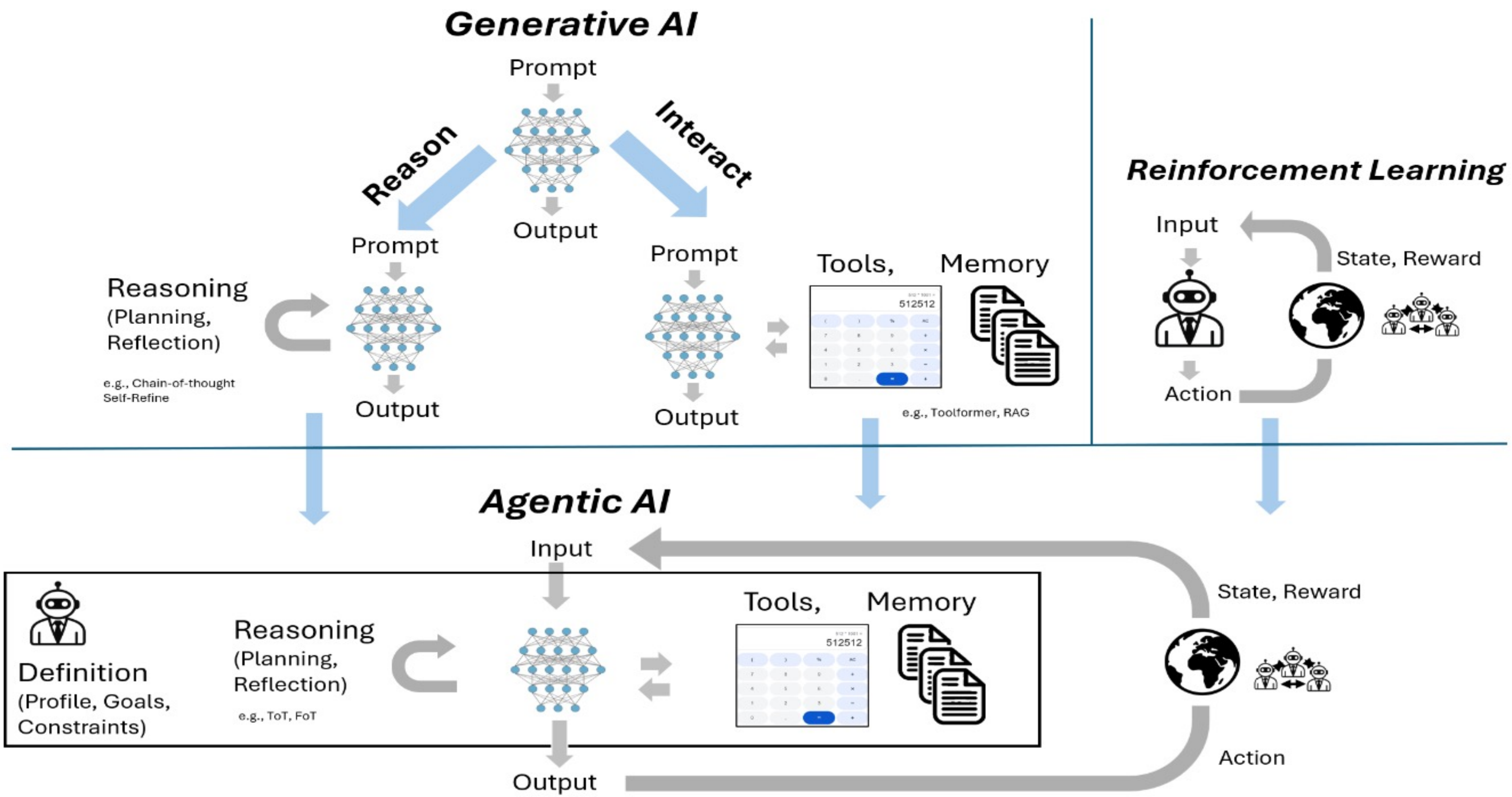


Source: Jeong, Cheonsu. "A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture." arXiv preprint arXiv:2309.01105 (2023).

Generative AI



From Generative AI to Agentic AI



Source: Schneider, Johannes. "Generative to Agentic AI: Survey, Conceptualization, and Challenges." arXiv preprint arXiv:2504.18875 (2025).

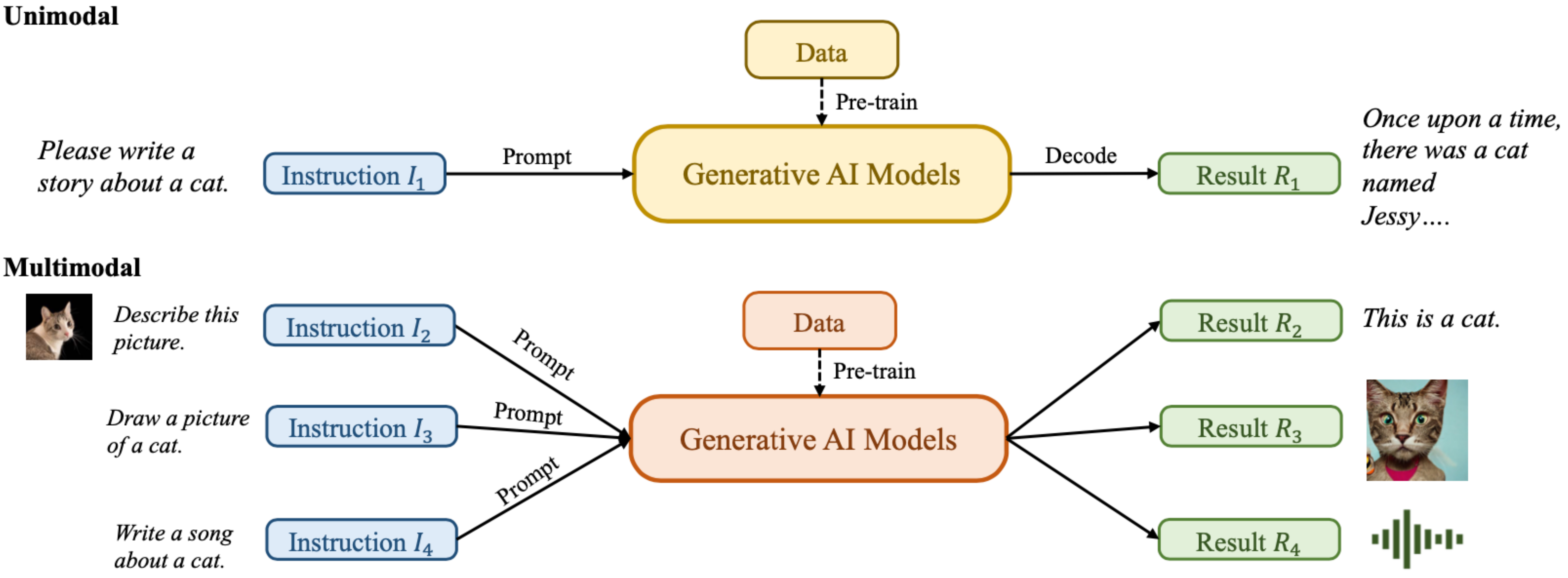
Generative AI

Text, Image, Video, Audio

Applications

Generative AI (Gen AI)

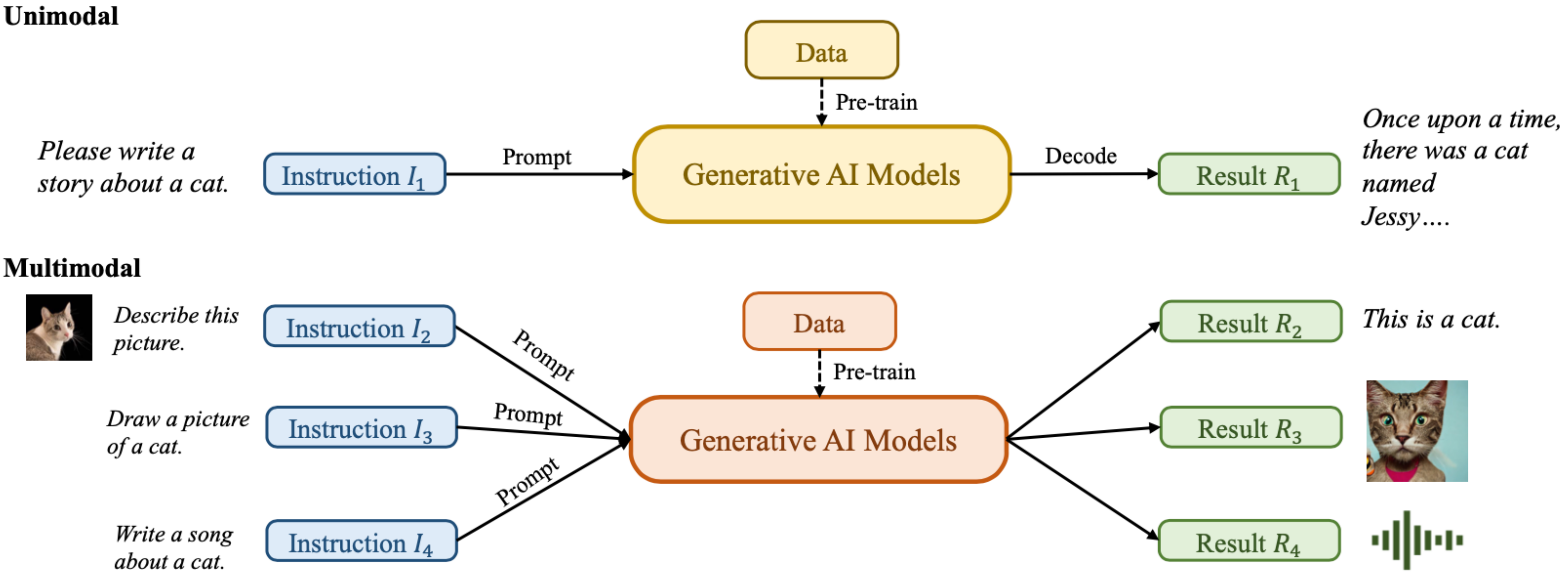
AI Generated Content (AIGC)



Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

Generative AI (Gen AI)

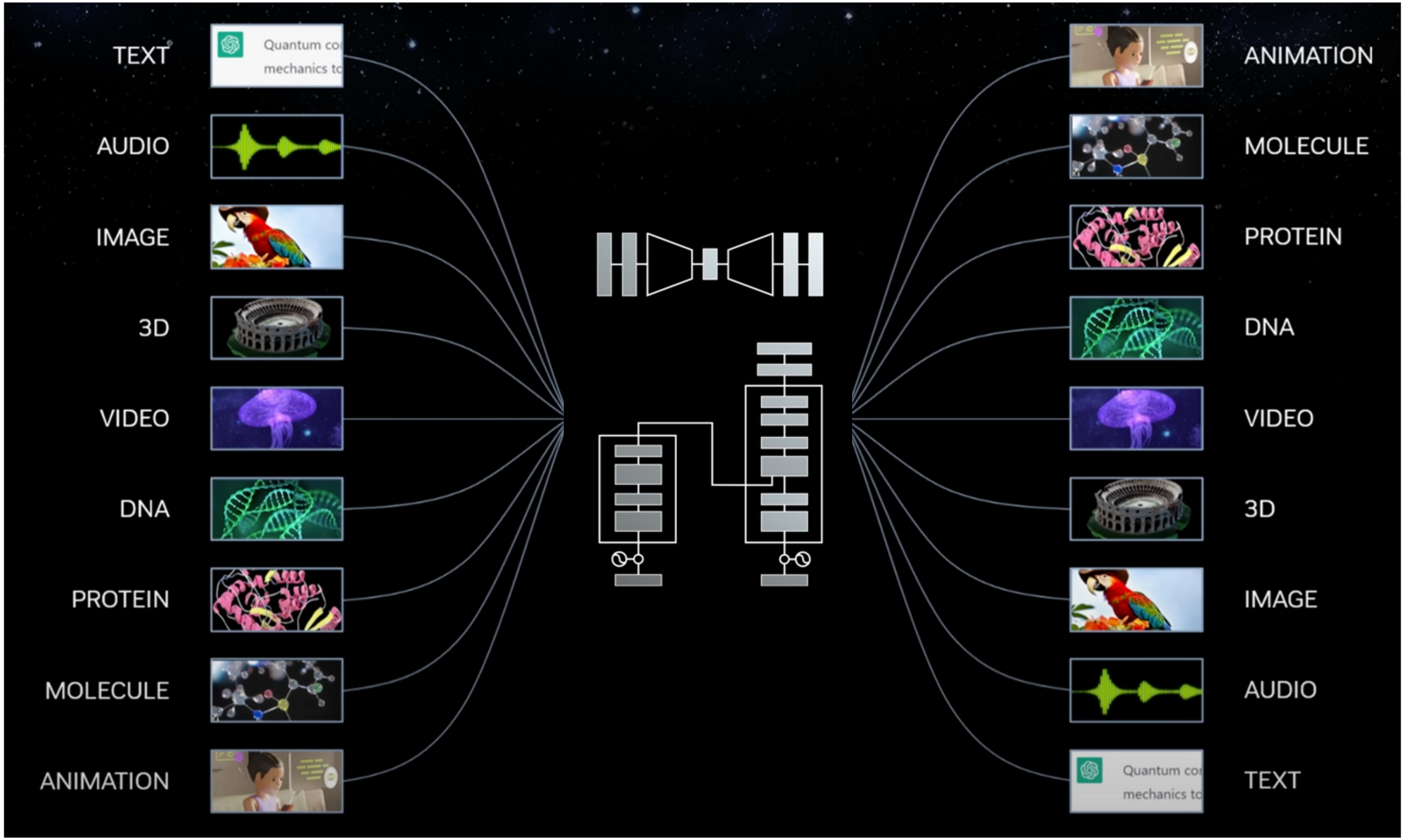
AI Generated Content (AIGC)



Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

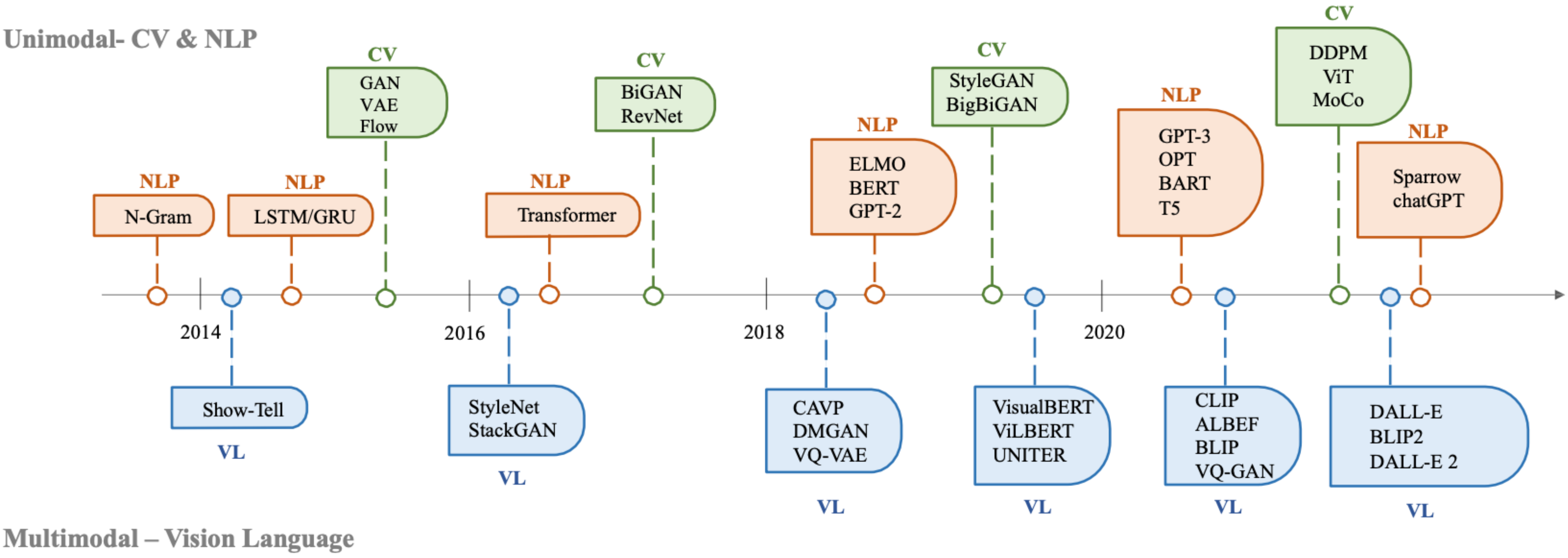
Modular Modalities

Where Can The Transformer Fit?



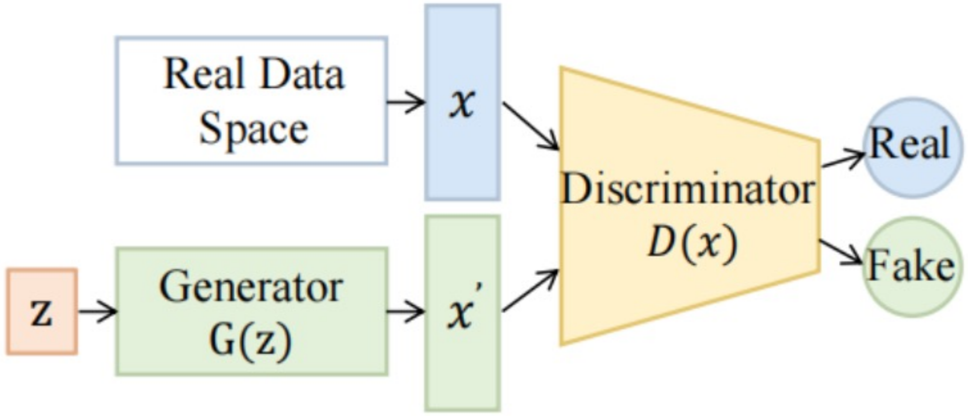
Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

The history of Generative AI in CV, NLP and VL

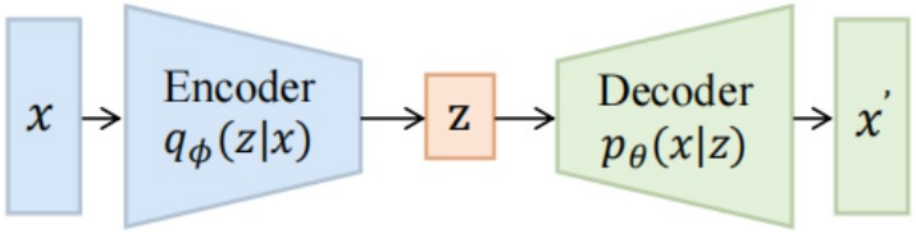


Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

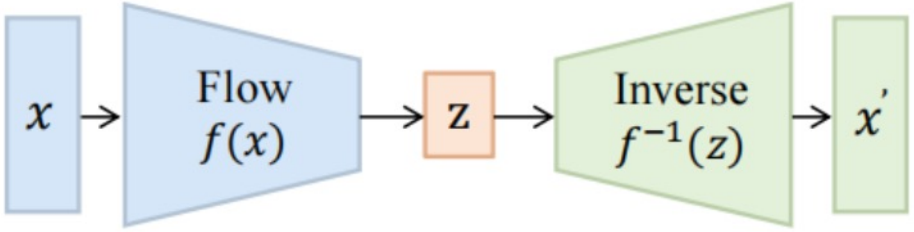
Categories of Vision Generative Models



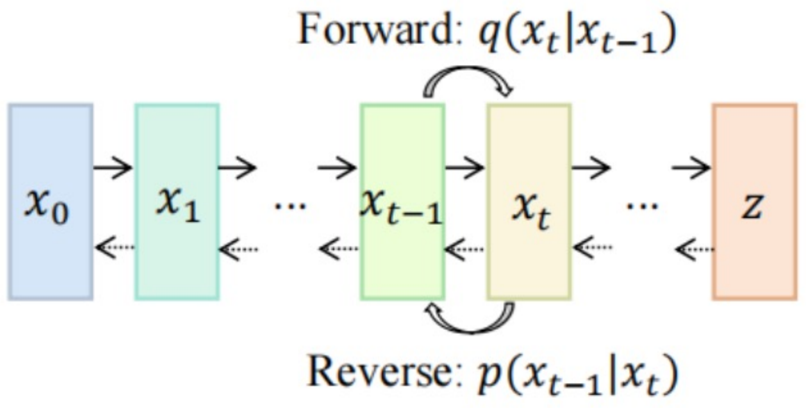
(1) Generative adversarial networks



(2) Variational autoencoders



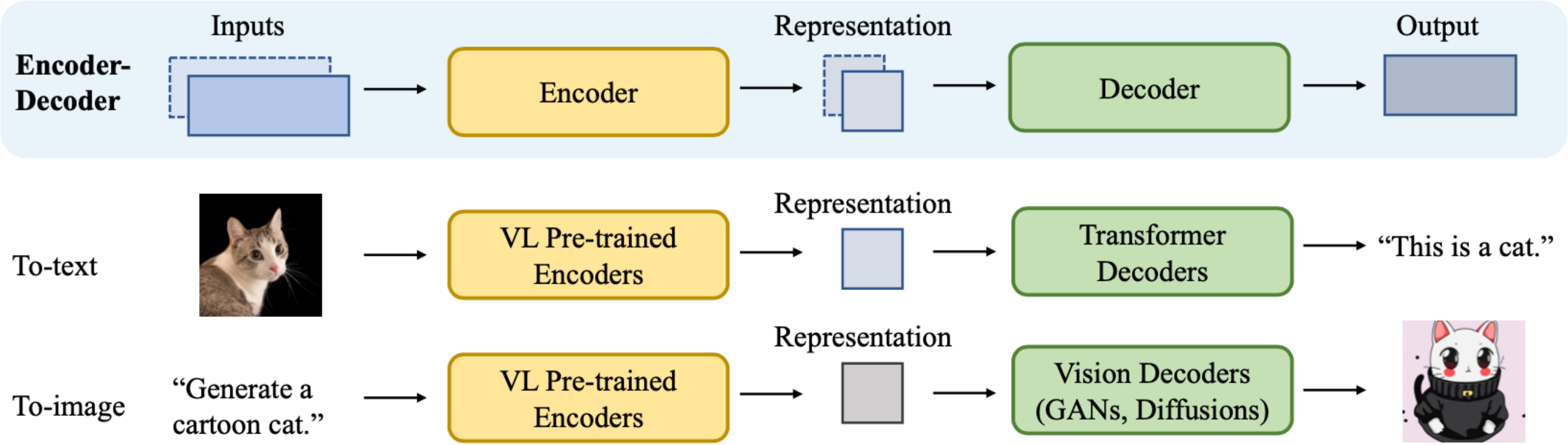
(3) Normalizing flows



(4) Diffusion models

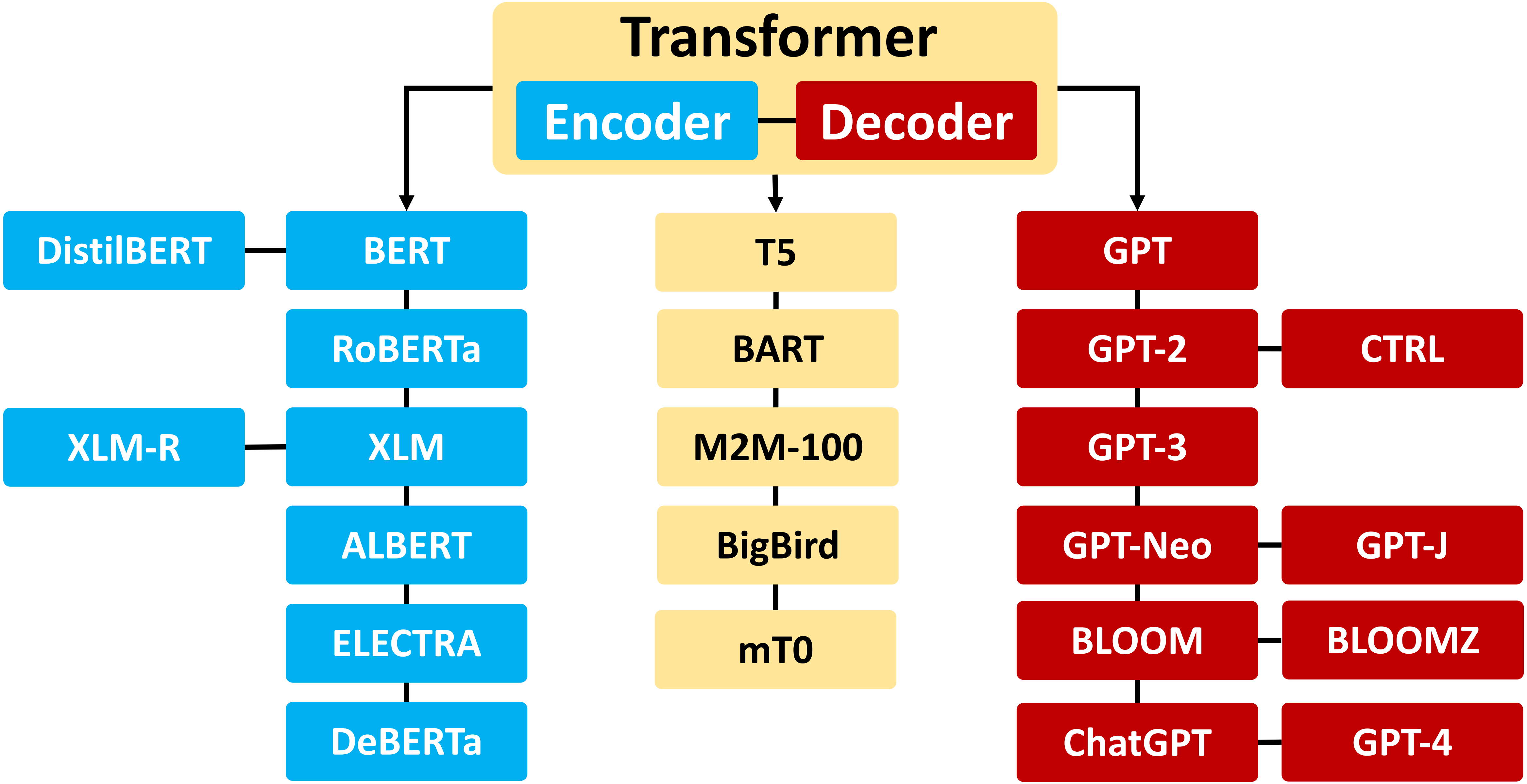
Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

The General Structure of Generative Vision Language



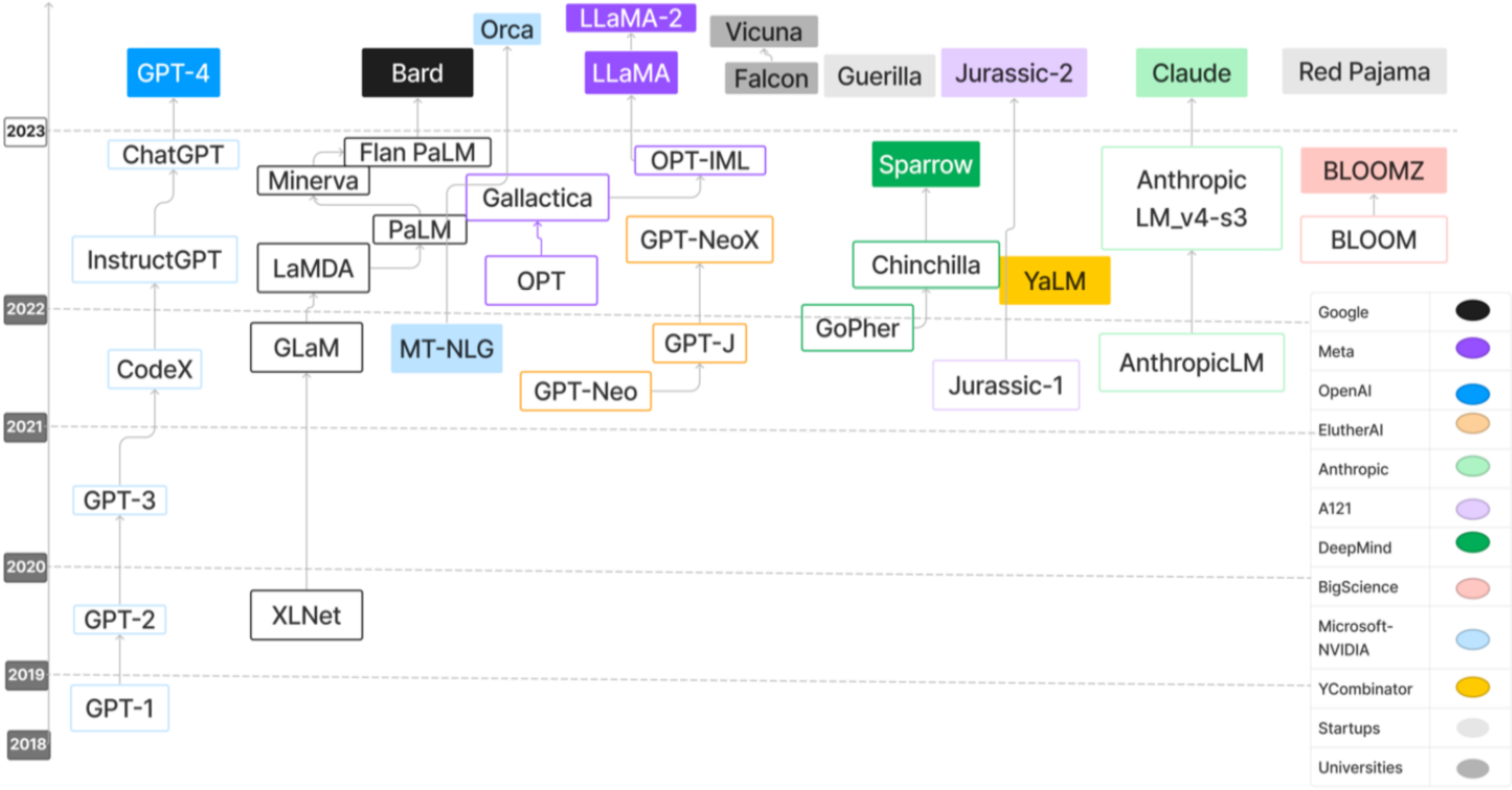
Source: Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." arXiv preprint arXiv:2303.04226.

Transformer Models



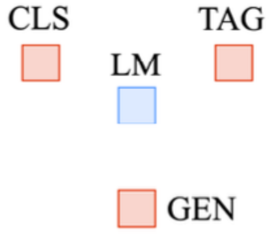
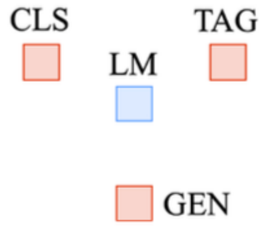
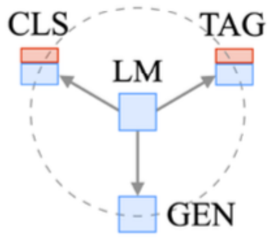
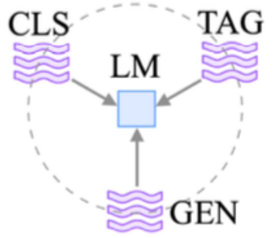
Source: Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.

Large Language Models (LLMs)



Source: Hadi, Muhammad Usman, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. "Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects." Authorea Preprints (2023).

Four Paradigms in NLP (LM)

Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Feature (e.g. word identity, part-of-speech, sentence length)	
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	
Transfer Learning: Pre-training, Fine-Tuning (FT) c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	
GAI: Pre-train, Prompt, and Predict (Prompting) d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	

Source: Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. (2023) "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." ACM Computing Surveys 55, no. 9 (2023): 1-35.

Comparison of Generative AI and Traditional AI

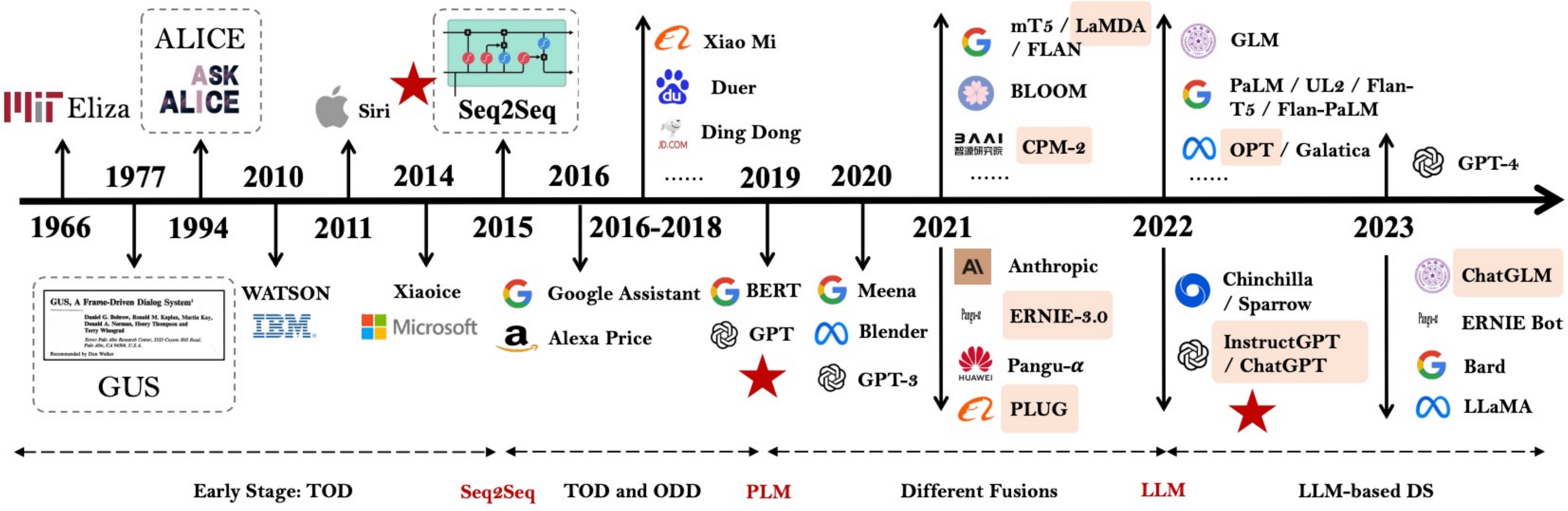
Feature	Generative AI	Traditional AI
Output type	New content	Classification/Prediction
Creativity	High	Low
Interactivity	Usually more natural	Limited

Generative AI

Text, Image, Video, Audio Applications

The Development of LM-based Dialogue Systems

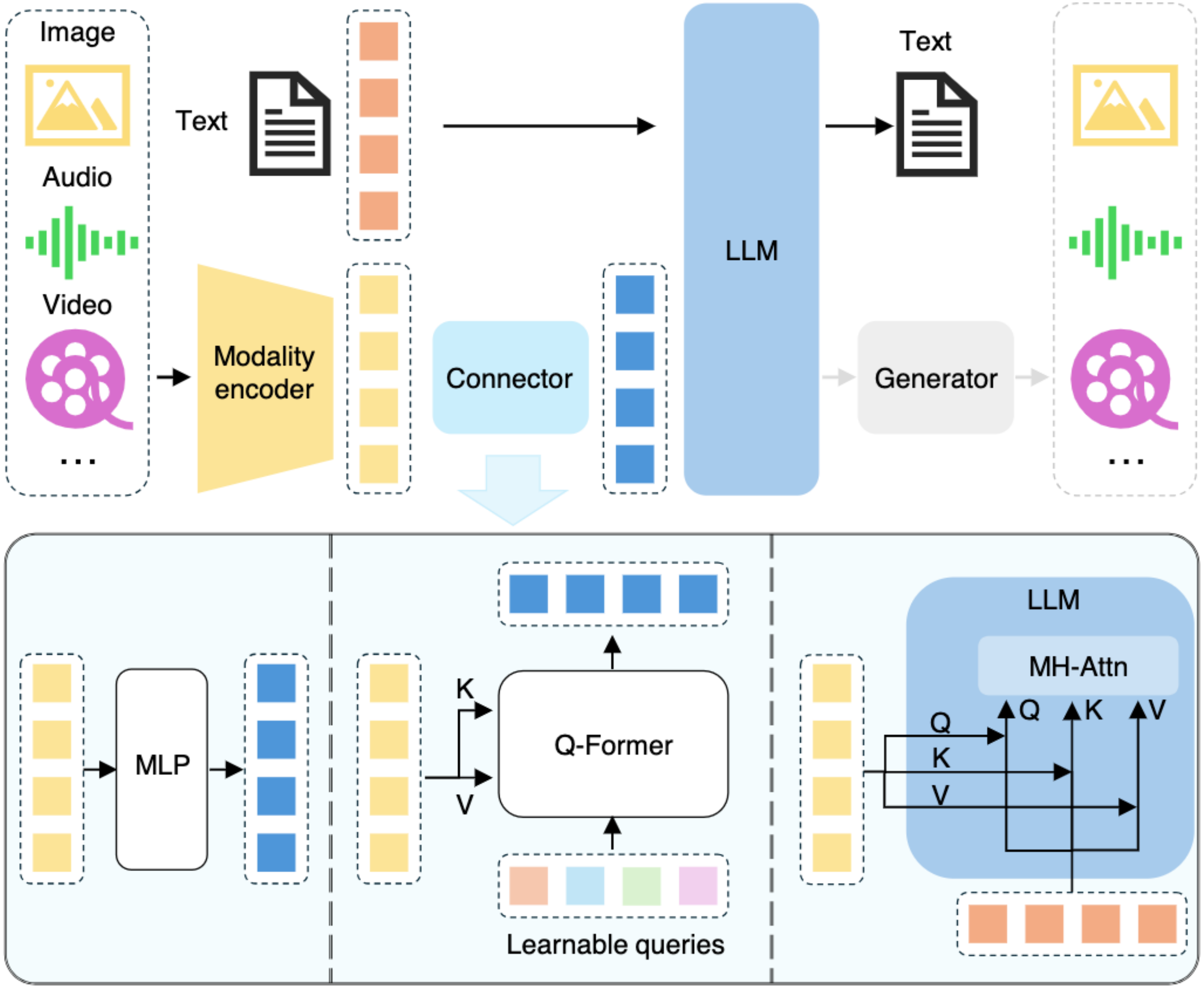
1) Early Stage (1966 - 2015)
2) The Independent Development of TOD and ODD (2015 - 2019)
3) Fusions of Dialogue Systems (2019 - 2022)
4) LLM-based DS (2022 - Now)



Task-oriented DS (TOD), Open-domain DS (ODD)

Source: Wang, Hongru, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. "A Survey of the Evolution of Language Model-Based Dialogue Systems." arXiv preprint arXiv:2311.16789 (2023).

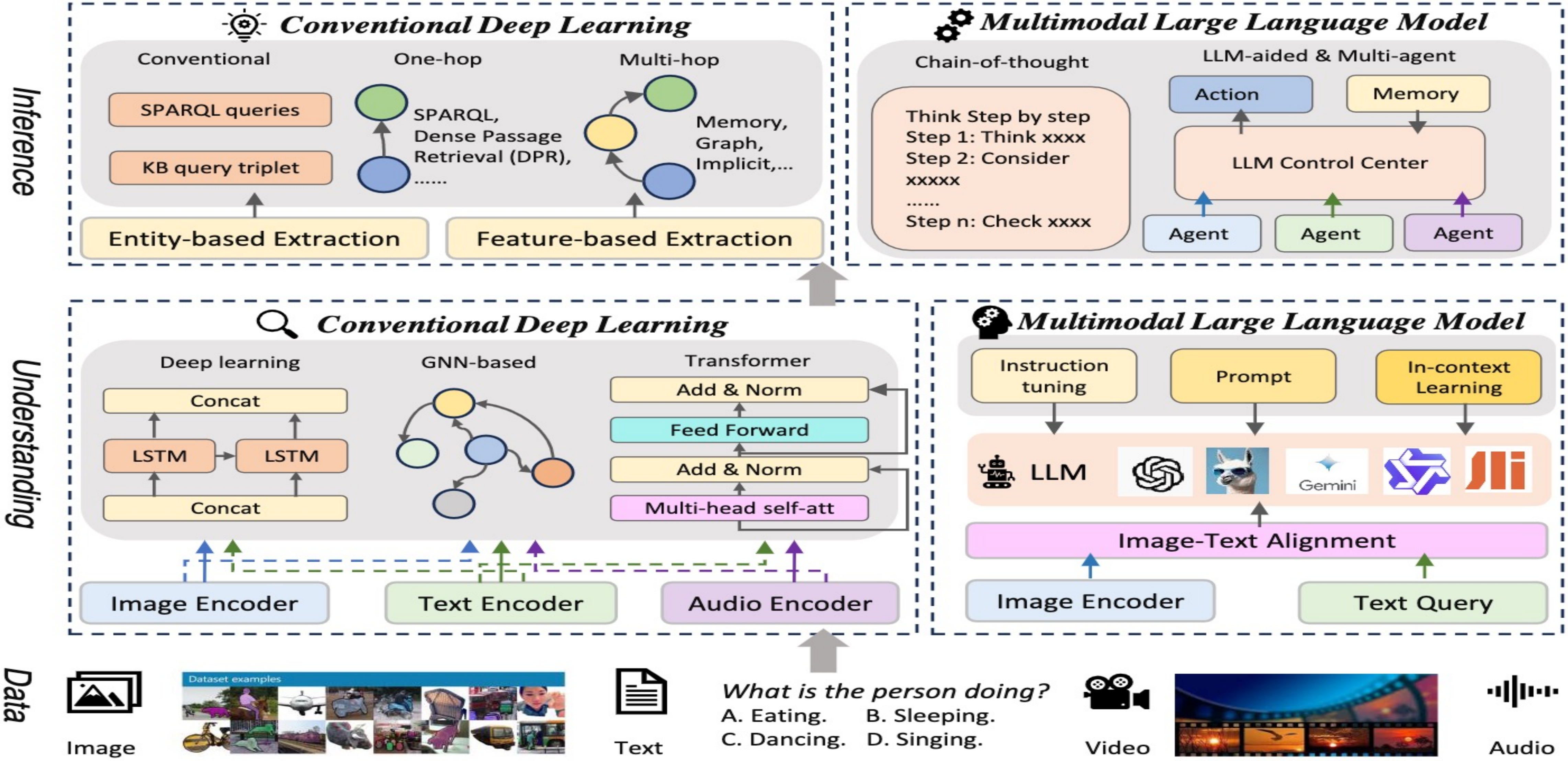
Multimodal Large Language Models (MLLM)



Multimodall LLM
Three types of connectors:
1. projection-based
2. query-based
3. fusion-based connectors

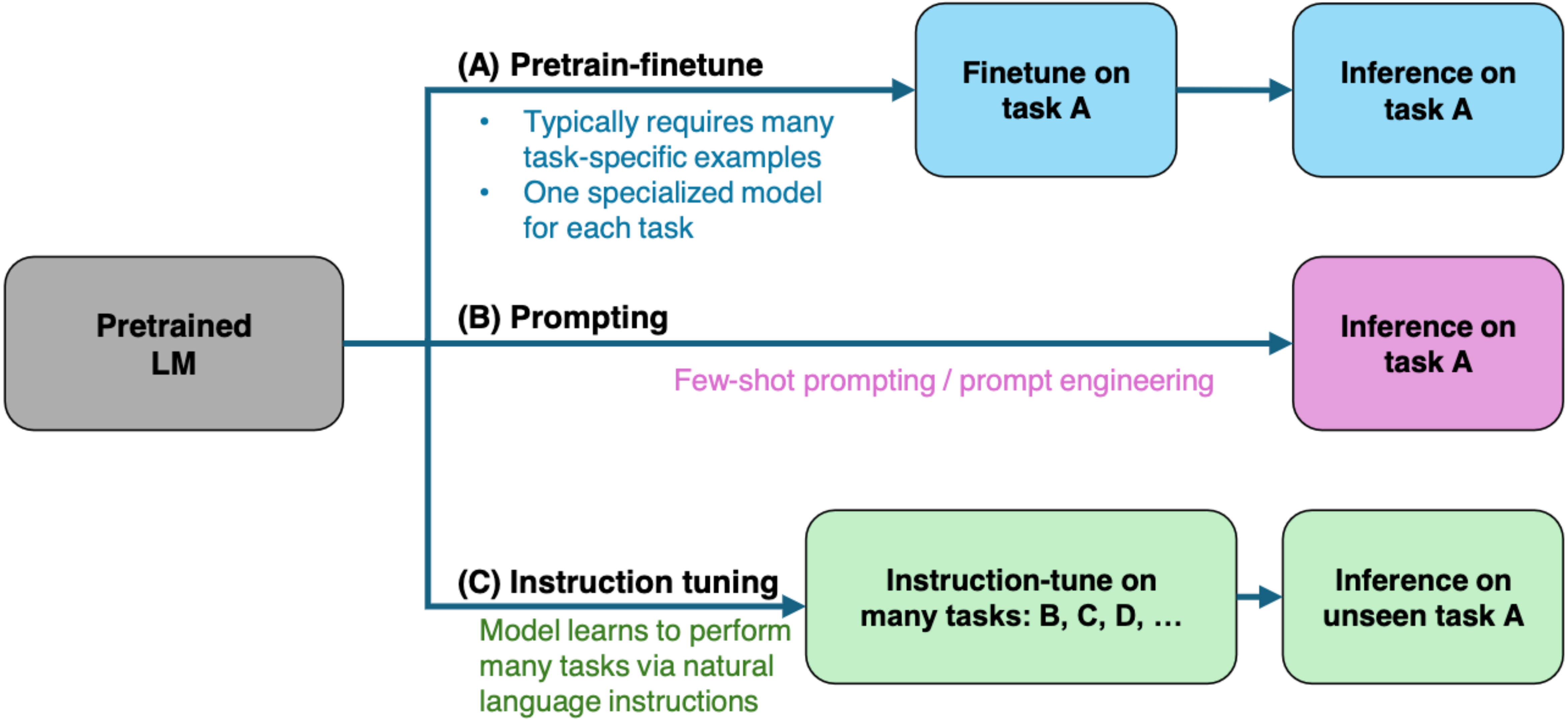
Source: Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. (2024) "A survey on multimodal large language models." National Science Review (2024): nwae403.

Multimodal Large Language Model (MLLM) for Vision Question Answering



Source: Jiayi Kuang, Jingyou Xie, Haohao Luo, Ronghao Li, Zhe Xu, Xianfeng Cheng, Yinghui Li, Xika Lin, and Ying Shen. (2024) "Natural Language Understanding and Inference with MLLM in Visual Question Answering: A Survey." arXiv preprint arXiv:2411.17558.

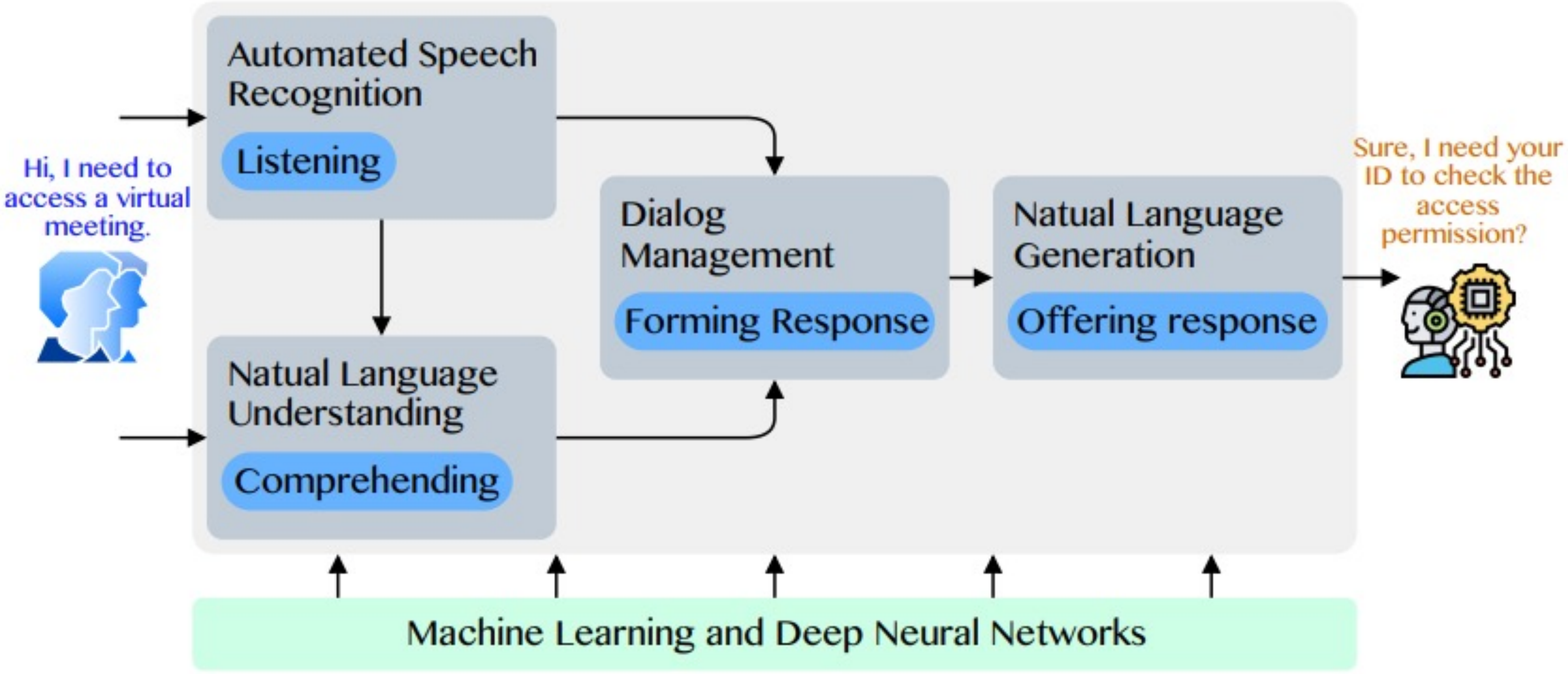
Large Language Models (LLM)
Three typical learning paradigms



Source: Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. (2024) "A survey on multimodal large language models." National Science Review (2024): nwae403.

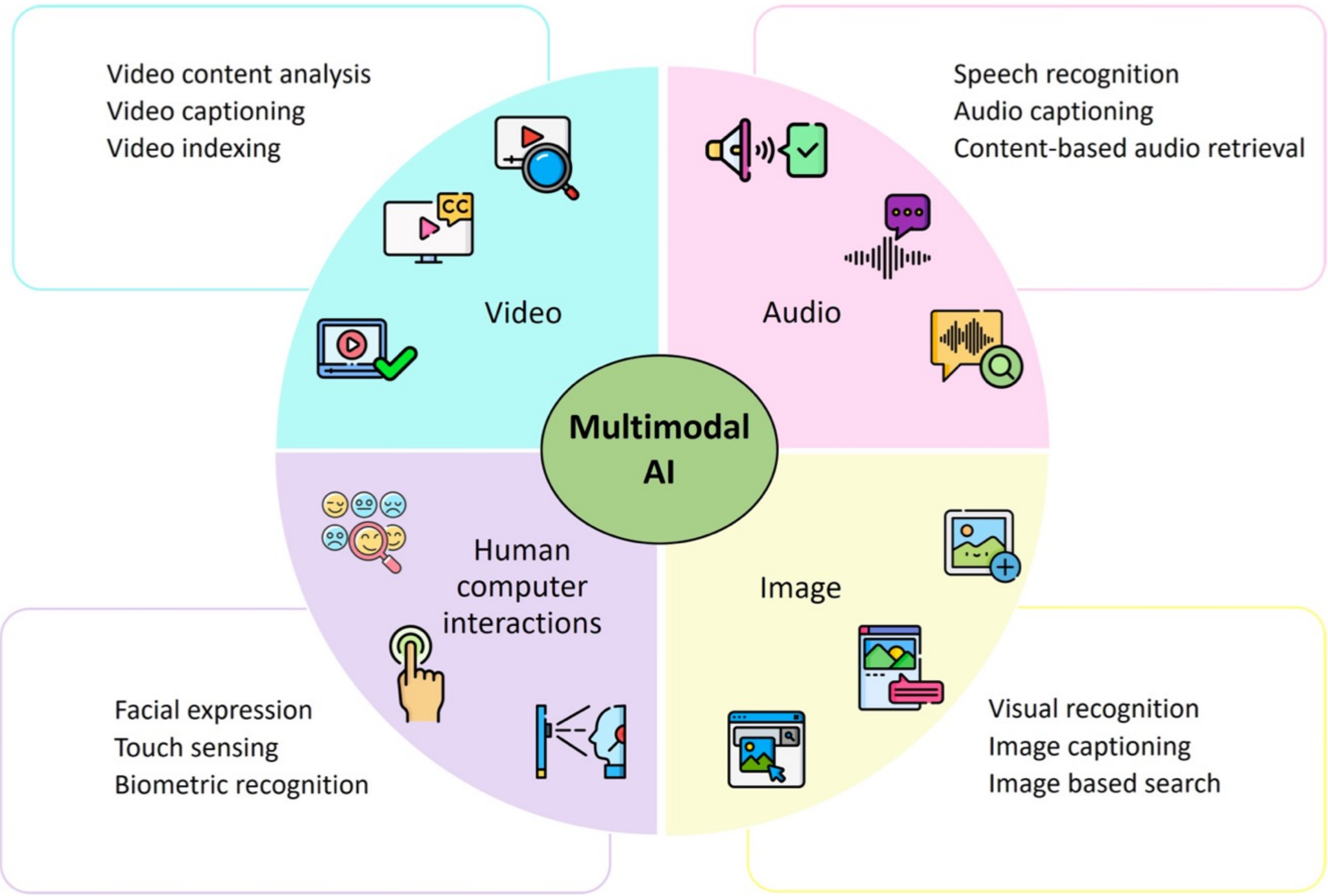
Conversational AI

to deliver contextual and personal experience to users



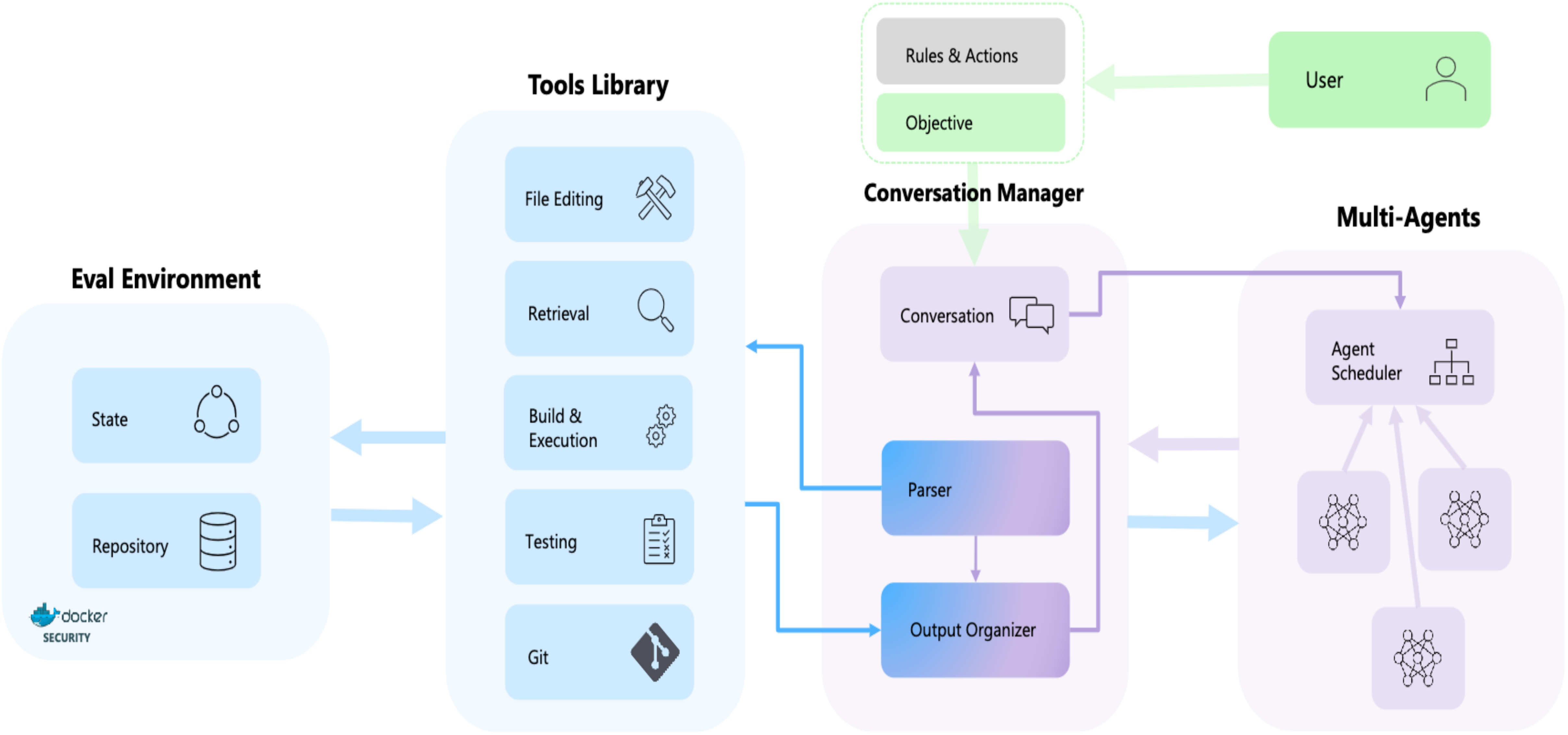
Source: Huynh-The, Thien, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022). "Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.

Technological Integration for Multimodal AI



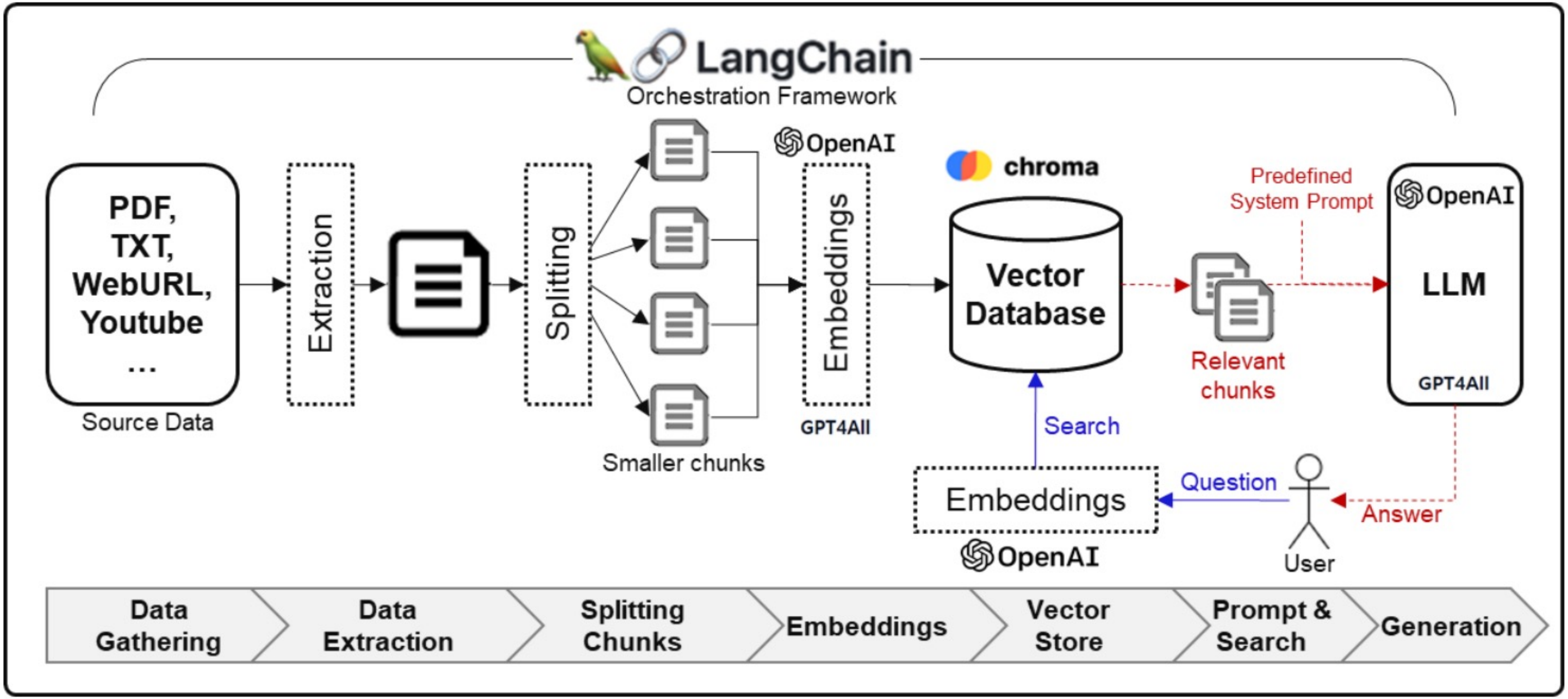
Source: Sohail, Shahab Saquib, Faiza Farhat, Yassine Himeur, Mohammad Nadeem, Dag Øivind Madsen, Yashbir Singh, Shadi Atalla, and Wathiq Mansoor. "The Future of GPT: A Taxonomy of Existing ChatGPT Research, Current Challenges, and Possible Future Directions." Current Challenges, and Possible Future Directions (April 8, 2023) (2023).

AutoDev: Automated AI-Driven Development



Source: Tufano, Michele, Anisha Agarwal, Jinu Jang, Roshanak Zilouchian Moghaddam, and Neel Sundaresan. (2024) "AutoDev: Automated AI-Driven Development." arXiv preprint arXiv:2403.08299 (2024).

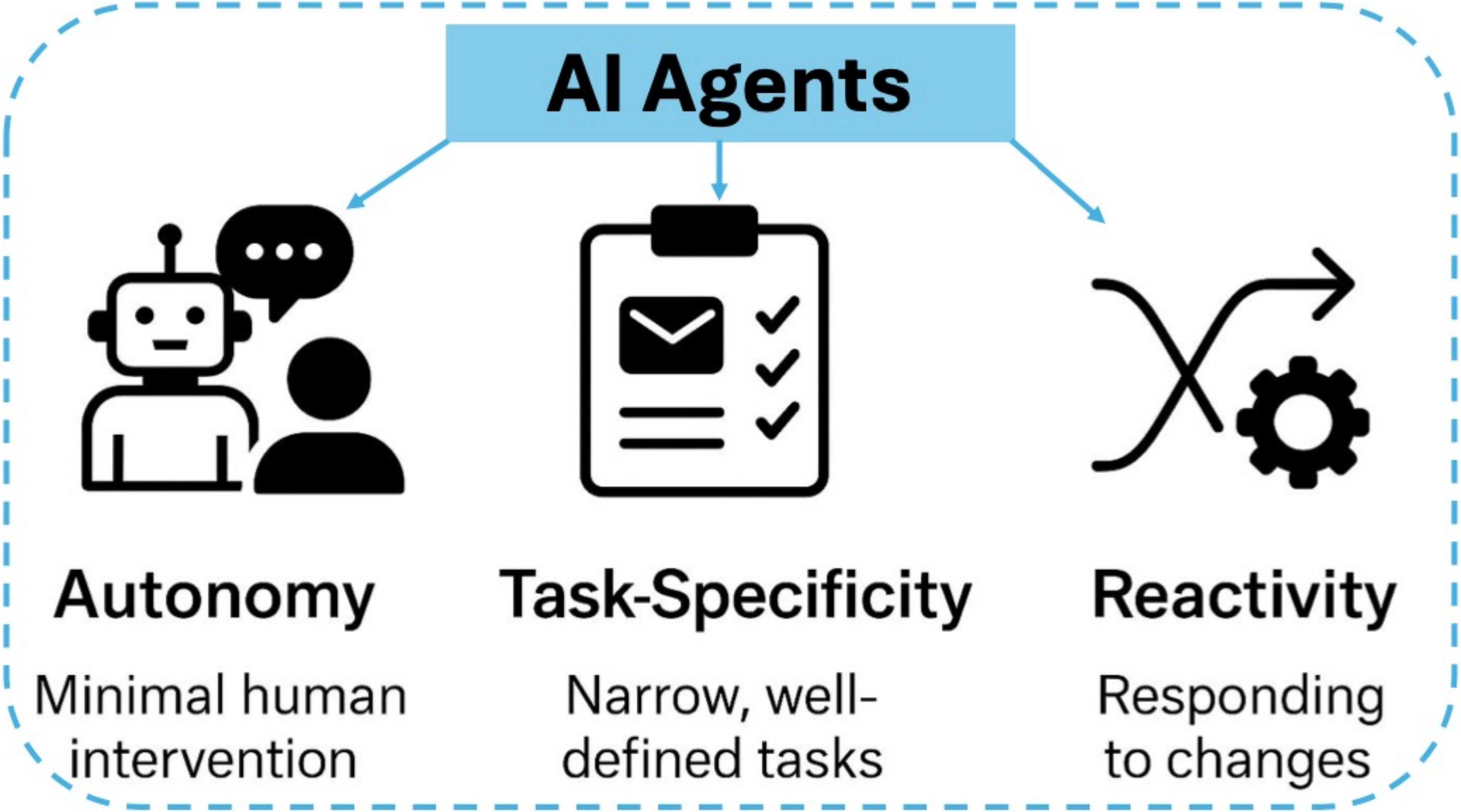
Framework for Implementing Generative AI Services using RAG Model



Source: Jeong, Cheonsu. "A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture." arXiv preprint arXiv:2309.01105 (2023).

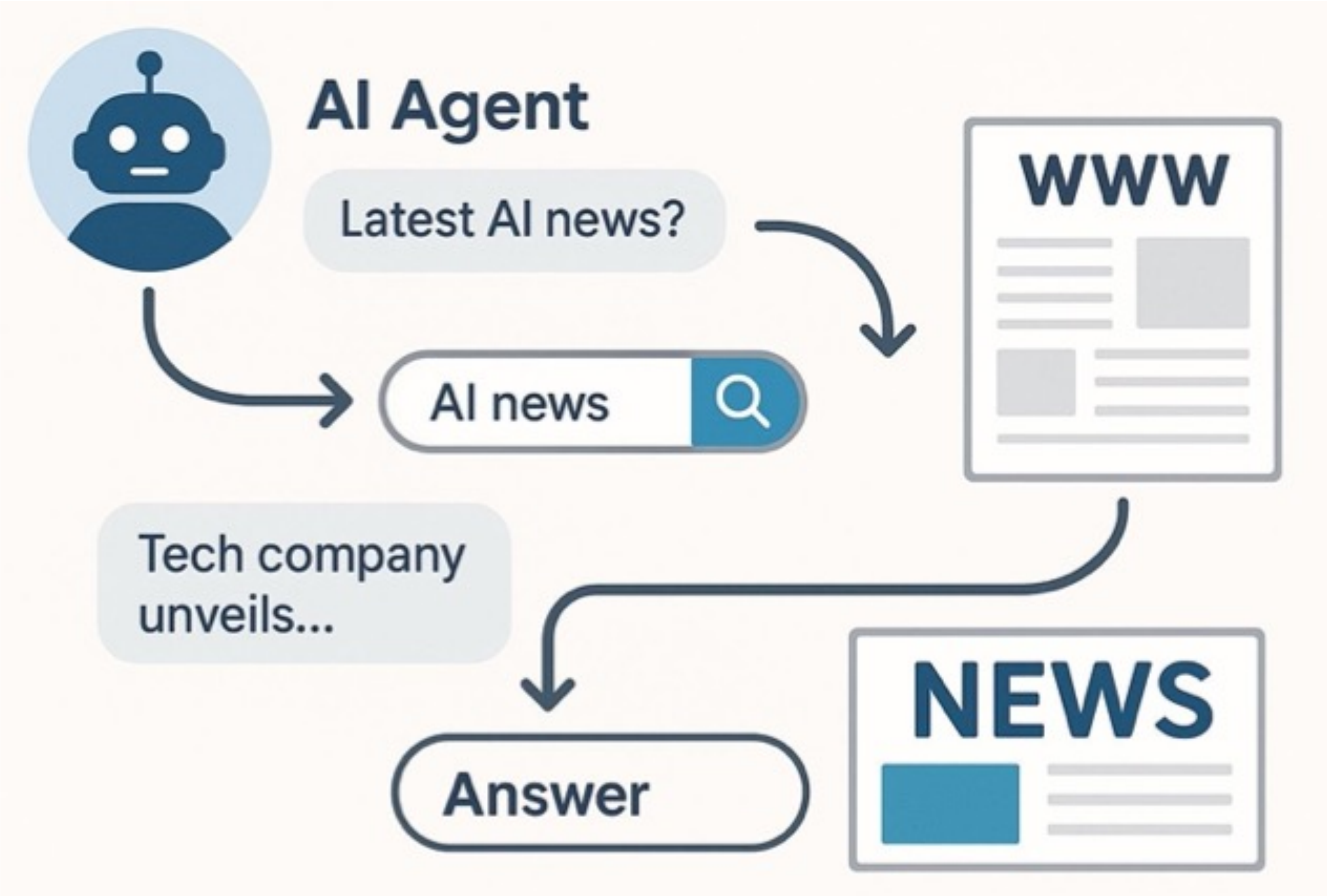
Agentic AI

AI Agents



Source: Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. (2025) "Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges." arXiv preprint arXiv:2505.10468 (2025).

AI Agents



Source: Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. (2025) "Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges." arXiv preprint arXiv:2505.10468 (2025).

Comparison of Generative AI and Traditional AI

Feature	Generative AI	Traditional AI
Output type	New content	Classification/Prediction
Creativity	High	Low
Interactivity	Usually more natural	Limited

AI Agent / Agentic AI, Generative AI, Traditional AI

Feature	AI Agent / Agentic AI	Generative AI	Traditional AI
Core Concept	To autonomously perceive its environment, make decisions, and take actions to achieve specific goals.	To create new, original content (text, images, code, etc.) that resembles its training data.	To execute specific tasks based on pre-programmed rules or statistical patterns.
Primary Function	Action & Goal Achievement. Executes a series of tasks to complete an objective (e.g., "Book me a flight to Taipei next Tuesday.").	Creation & Synthesis. Creates novel outputs in response to a prompt (e.g., "Write a poem about rain.").	Classification & Prediction. Answers questions with a known range of outcomes (e.g., "Is this spam?").
Decision Making	Based on a continuous loop: Perceive -> Plan -> Act. It reasons about its goal, breaks it down, and executes steps.	Based on probabilistic patterns learned from massive, unstructured datasets. It predicts the next most likely word, pixel, or note.	Based on explicitly programmed logic (if-then rules) or learned patterns from structured data.
Key Characteristic	Autonomous & Goal-Oriented. Proactively takes steps and can adapt its plan based on new information.	Creative & Probabilistic. Can produce a wide variety of unique outputs from the same prompt.	Deterministic & Logic-Based. Given the same input, it will almost always produce the same output.

AI Agent / Agentic AI, Generative AI, Traditional AI

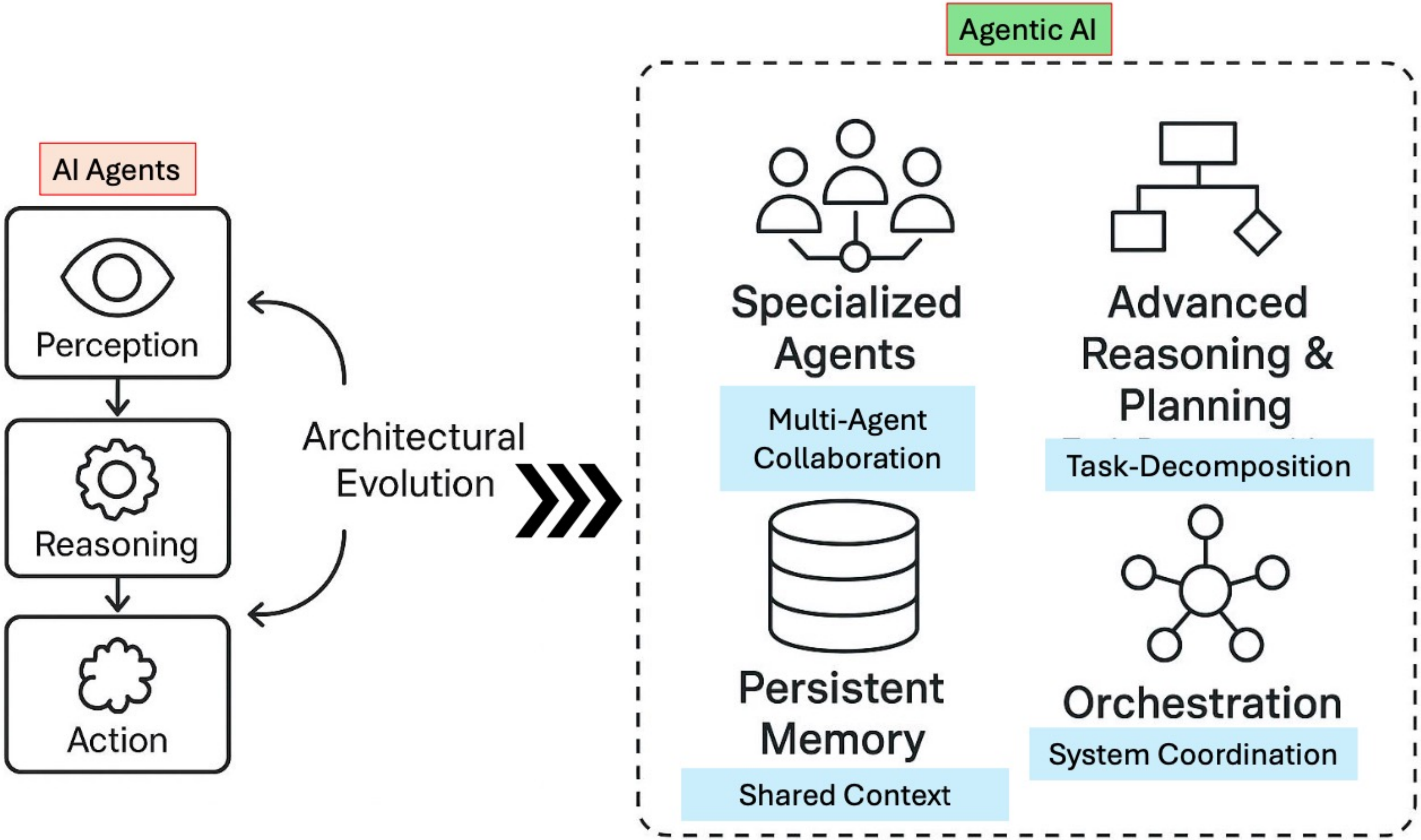
Feature	AI Agent / Agentic AI	Generative AI	Traditional AI
Interaction Model	Proactive & Interactive. Actively observes its environment (digital or physical) and takes actions to change it.	Responsive. Engages in a dialogue or responds to a user's prompt to generate content.	Reactive. Responds to a direct input or query. It doesn't act on its own.
Example Technologies	Architectural frameworks like ReAct (Reason + Act), and systems that combine LLMs with tools and memory.	Large Language Models (LLMs) like GPT-4, Diffusion Models (for images), Generative Adversarial Networks (GANs).	Expert systems, decision trees, linear regression, traditional machine learning (ML) models.
Common Use Cases	Self-driving cars, autonomous trading bots, smart assistants that manage calendars, customer service agents that process refunds.	ChatGPT, Google Gemini, Midjourney (image generation), Copilot (code generation), music composition.	Spam filters, chess engines, recommendation systems (e.g., Netflix), credit scoring, medical diagnosis from scans.
Relationship to Others	An architecture or system that often uses Generative AI to reason and Traditional AI for specific sub-tasks to accomplish a goal.	Can serve as the "brain" or reasoning engine for an AI Agent, enabling it to understand, plan, and generate actions.	The foundation for modern AI. Its techniques can be components within larger AI systems.

AI Agents vs Agentic AI

Feature	AI Agents	Agentic AI
Definition	Autonomous software programs that perform specific tasks.	Systems of multiple AI agents collaborating to achieve complex goals.
Autonomy Level	High autonomy within specific tasks.	Broad level of autonomy with the ability to manage multi-step, complex tasks and systems.
Task Complexity	Typically handle single, specific tasks.	Handle complex, multi-step tasks requiring coordination.
Collaboration	Operate independently.	Involve multi-agent information sharing, collaboration and cooperation.
Learning and Adaptation	Learn and adapt within their specific domain.	Learn and adapt across a wider range of tasks and environments.
Applications	Customer service chatbots, virtual assistants, automated workflows.	Supply chain management, business process optimization, virtual project managers.

Source: Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. (2025) "Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges." arXiv preprint arXiv:2505.10468 (2025).

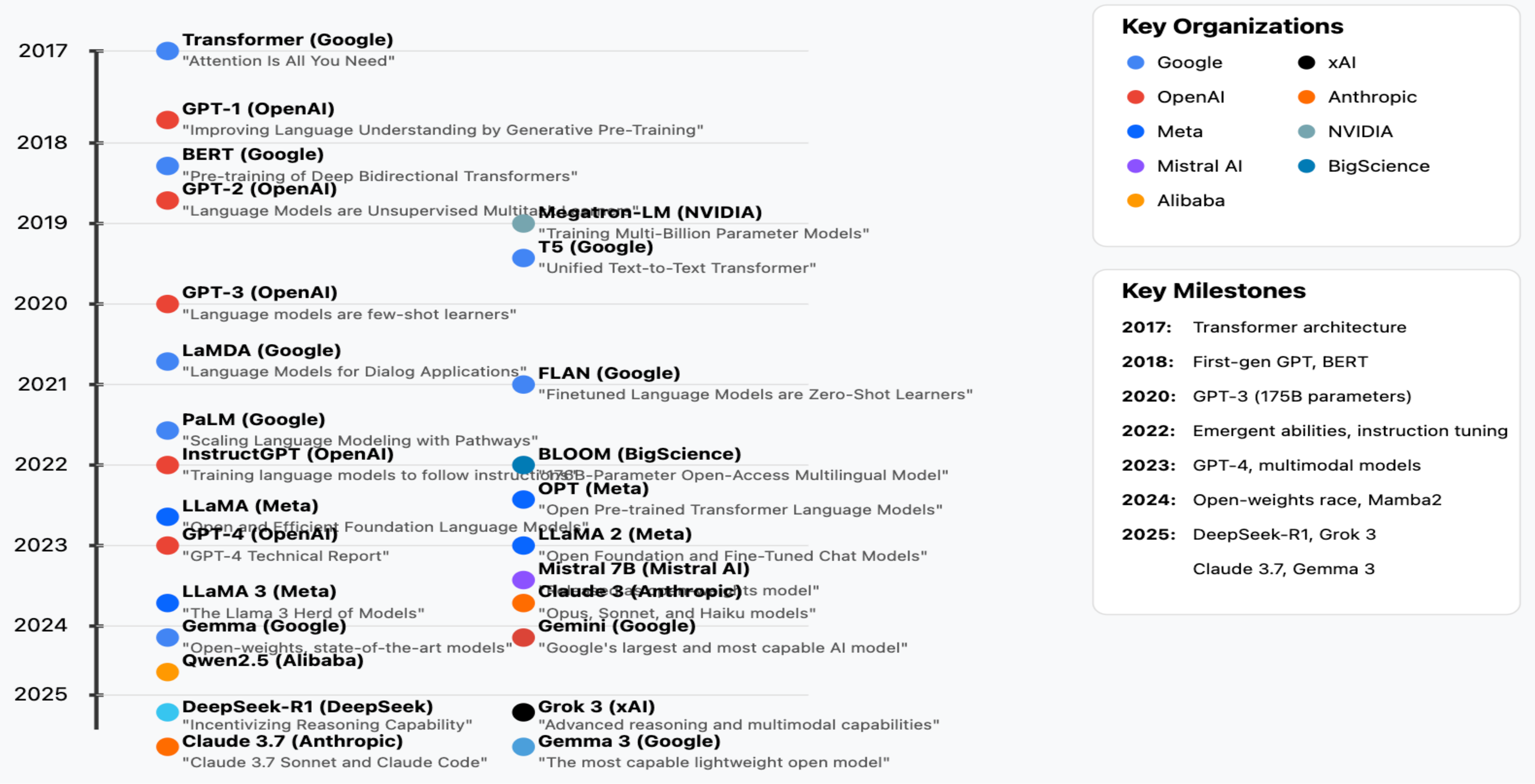
AI Agents vs Agentic AI



Source: Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. (2025) "Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges." arXiv preprint arXiv:2505.10468 (2025).

AI Agents and Large Multimodal Agents (LMAs)

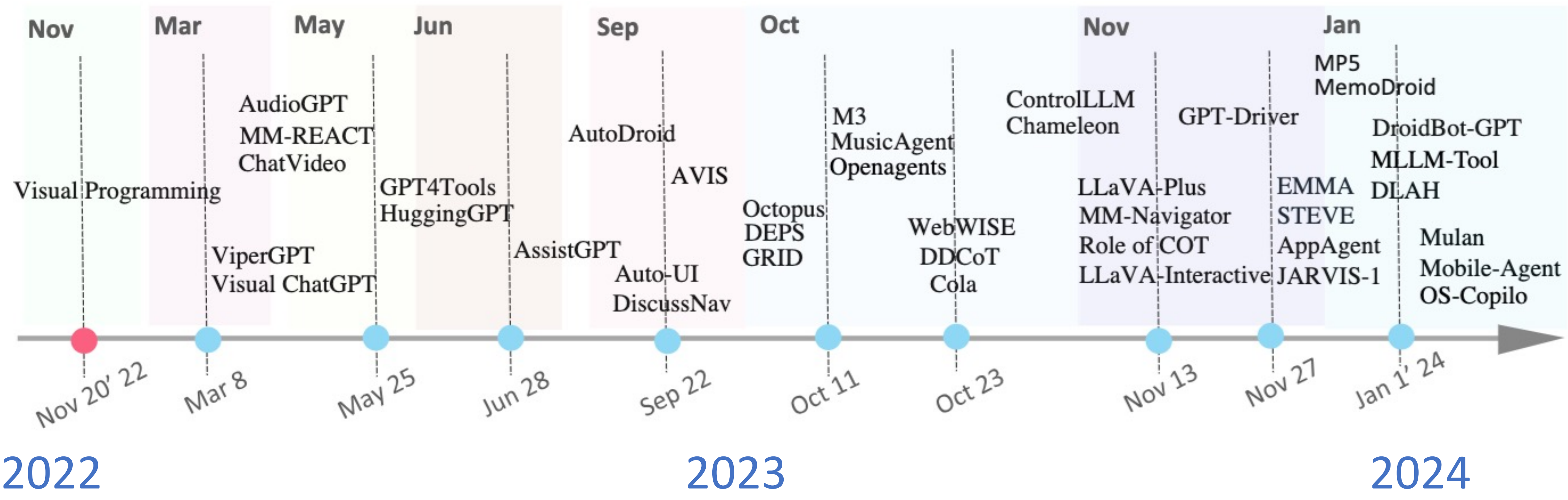
Generative AI LLMs (2017-2025)



Source: https://github.com/Hannibal046/Awesome-LLM

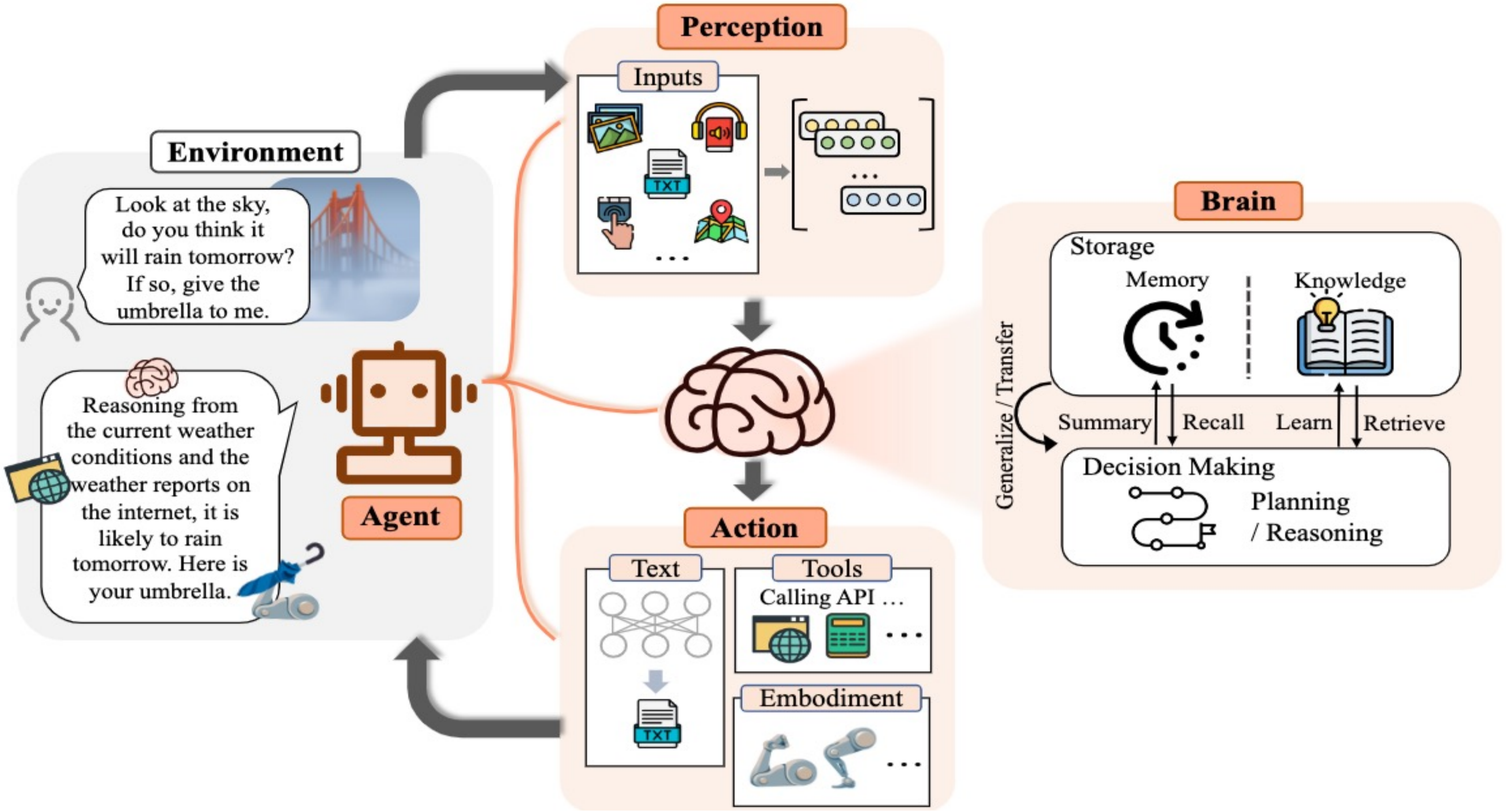
LLM-powered Multimodal Agents

Large Multimodal Agents (LMAs)



Source: Xie, Junlin, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. "Large Multimodal Agents: A Survey." arXiv preprint arXiv:2402.15116 (2024).

Large Language Model (LLM) based Agents

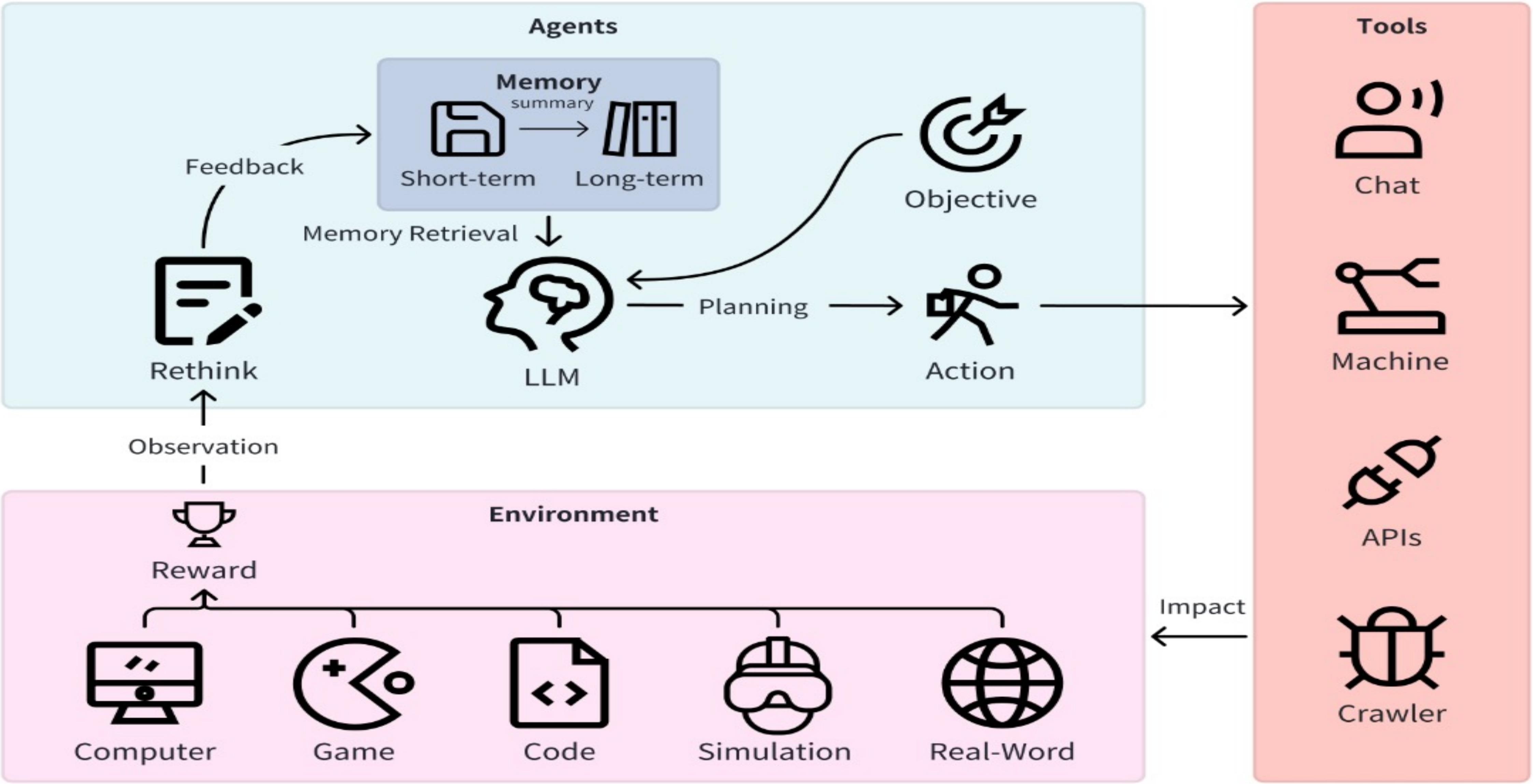


Source: Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2023). The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864.

LLM-based Agents

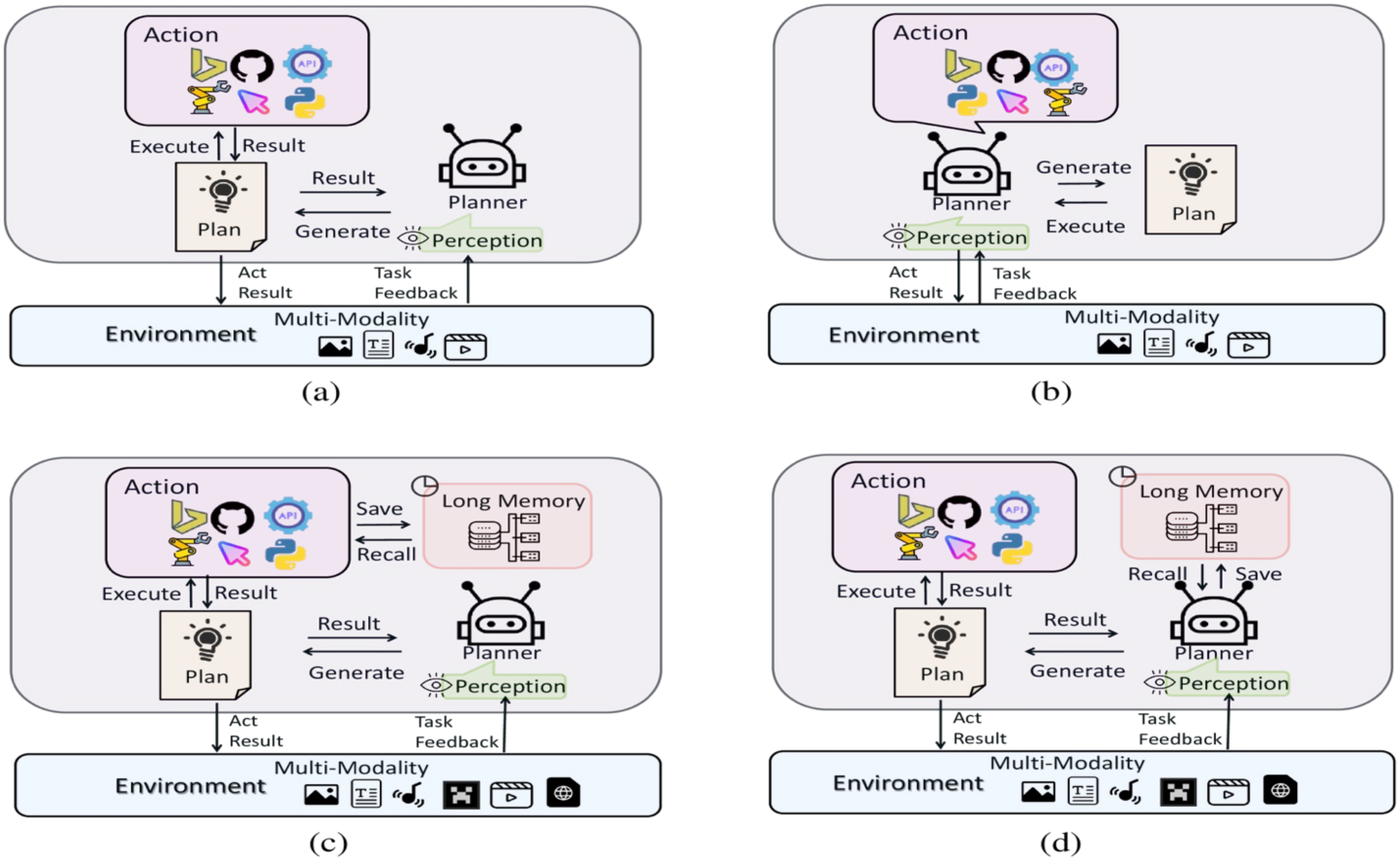
- Definition: AI agents that use Large Language Models as their core decision-making mechanism
- Key Features:
  - Natural language interface
  - Vast knowledge base
  - Ability to understand context and nuance
  - Generalize to new tasks with minimal additional training

LLM-based Agents



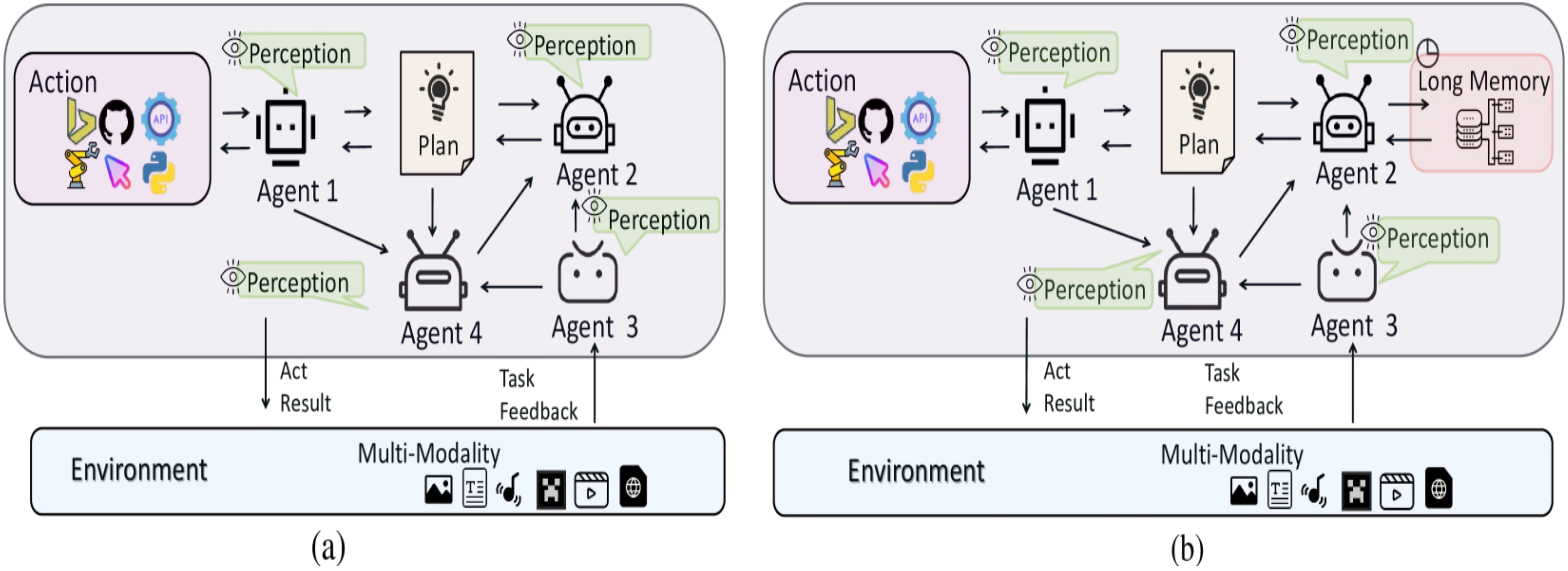
Source: Cheng, Yuheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang et al. "Exploring large language model based intelligent agents: Definitions, methods, and prospects." arXiv preprint arXiv:2401.03428 (2024).

Large Multimodal Agents (LMA)



Source: Xie, J., Chen, Z., Zhang, R., Wan, X., & Li, G. (2024). Large Multimodal Agents: A Survey. ArXiv, abs/2402.15116.

Large Multimodal Agents (LMA)



Source: Xie, J., Chen, Z., Zhang, R., Wan, X., & Li, G. (2024). Large Multimodal Agents: A Survey. ArXiv, abs/2402.15116.

Agentic AI Cloud Architecture

Microservices and Serverless Architecture

Containers (Docker, Kubernetes)

Serverless platforms (AWS Lambda, Google Cloud Functions)

APIs and Tooling Integration via MCP

Agents access tools (e.g., databases, APIs, CRMs, payment gateways) using Model Context Protocol (MCP)

Enhances tool-using behavior of LLM agents

Tools and Frameworks

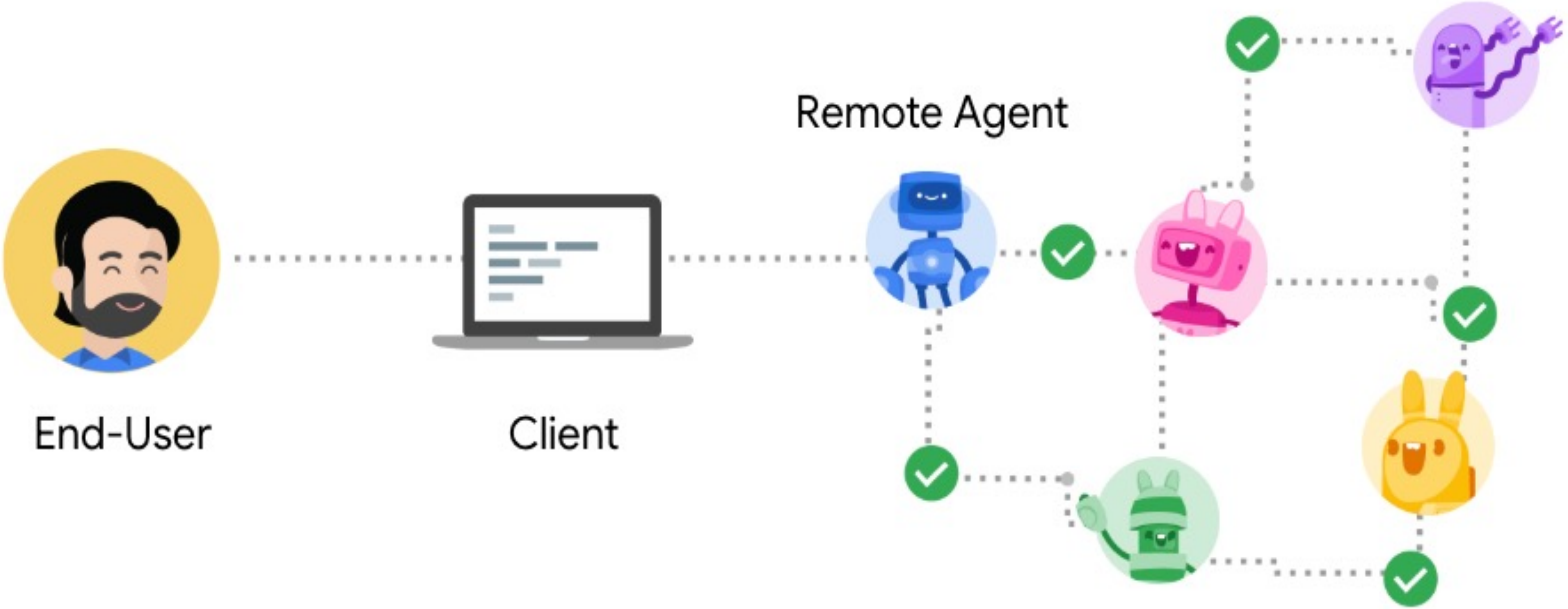
LangChain, AutoGen, CrewAI: for orchestrating LLM agents

Anthropic's MCP, Google's A2A: communication protocols

Vector DBs (Pinecone, Weaviate): for agent memory

Agent2Agent Protocol (A2A)

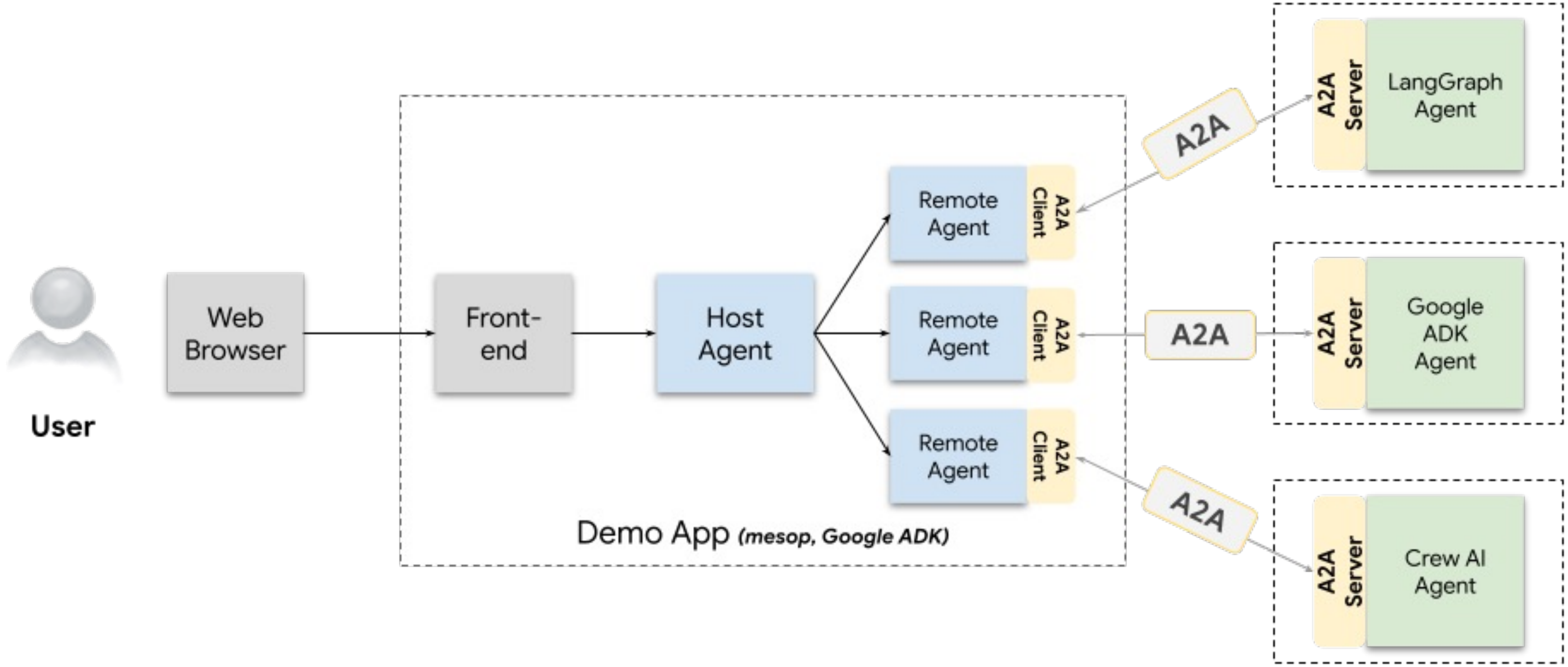
An open protocol enabling Agent-to-Agent interoperability, bridging the gap between opaque agentic systems



Source: https://google.github.io/A2A/

A2A Demo Web App

Agents talking to other agents over A2A



Source: https://github.com/google/A2A/blob/main/demo/README.md

A2A

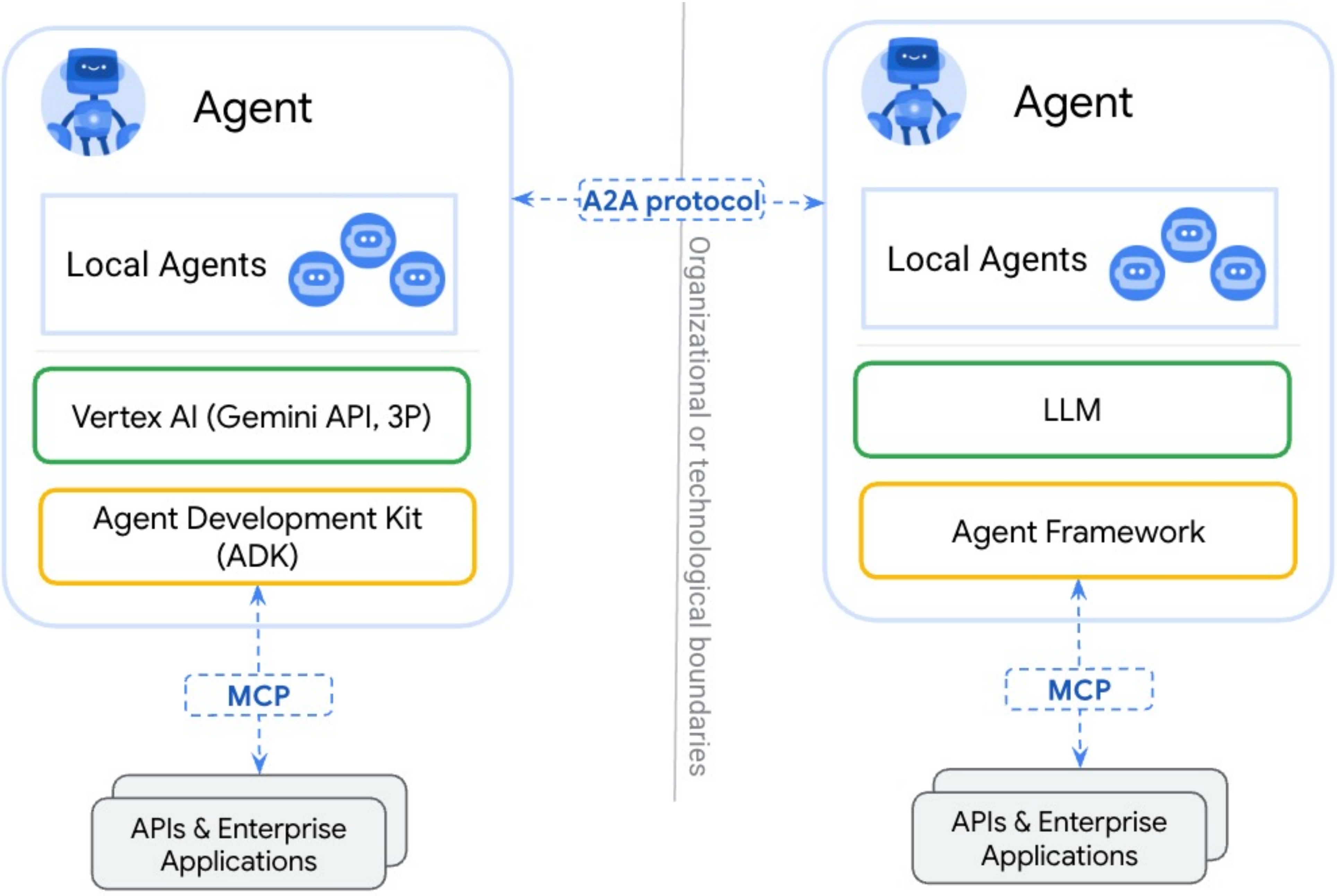
(Agent2Agent Protocol)

for agent-agent collaboration

MCP

(Model Context Protocol)

for tools and resources



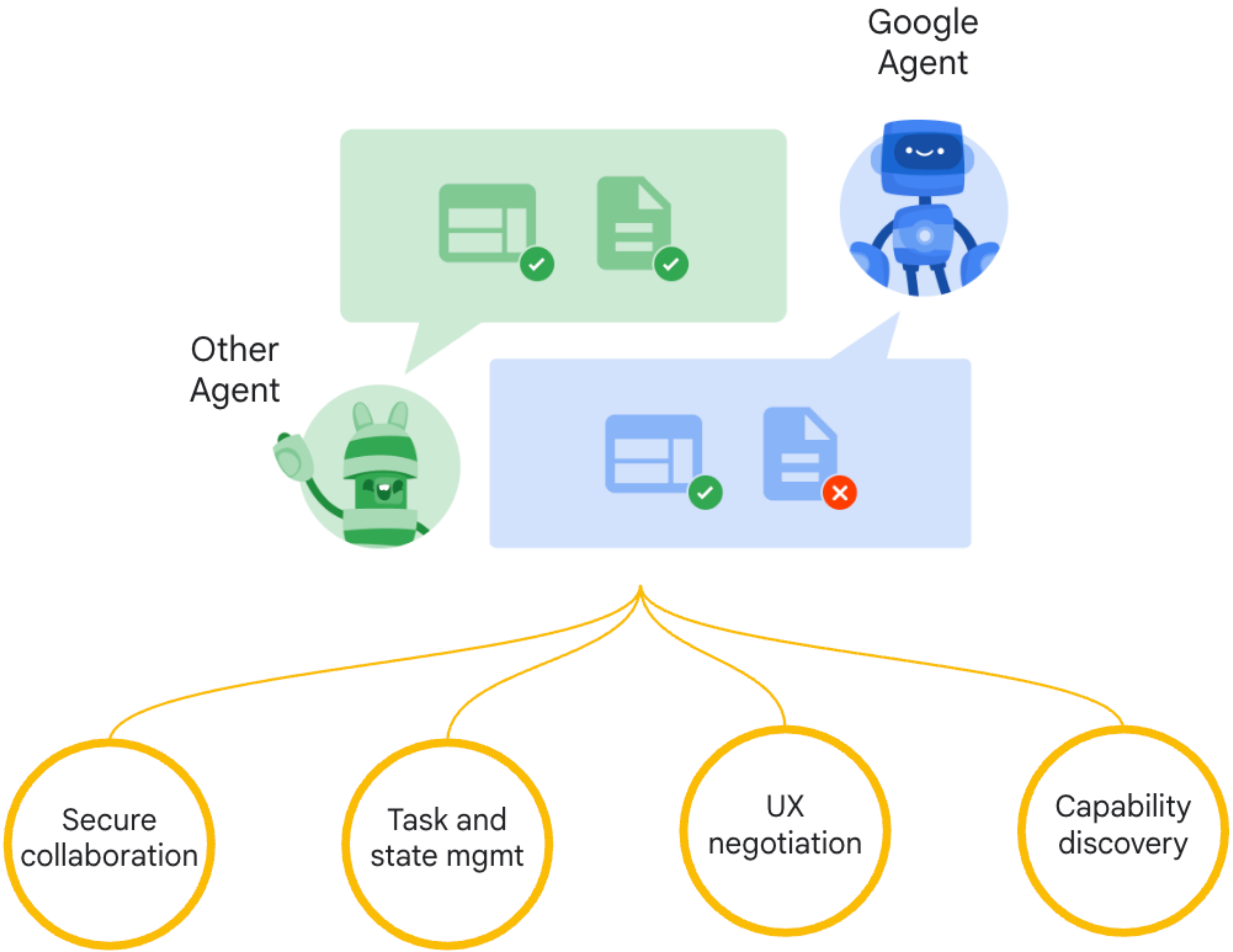
Source: https://google.github.io/A2A/

Google A2A (Agent2Agent Protocol)

Seamless Agent Collaboration

Simplifies Enterprise Agent Integration

Supports Key Enterprise Requirements

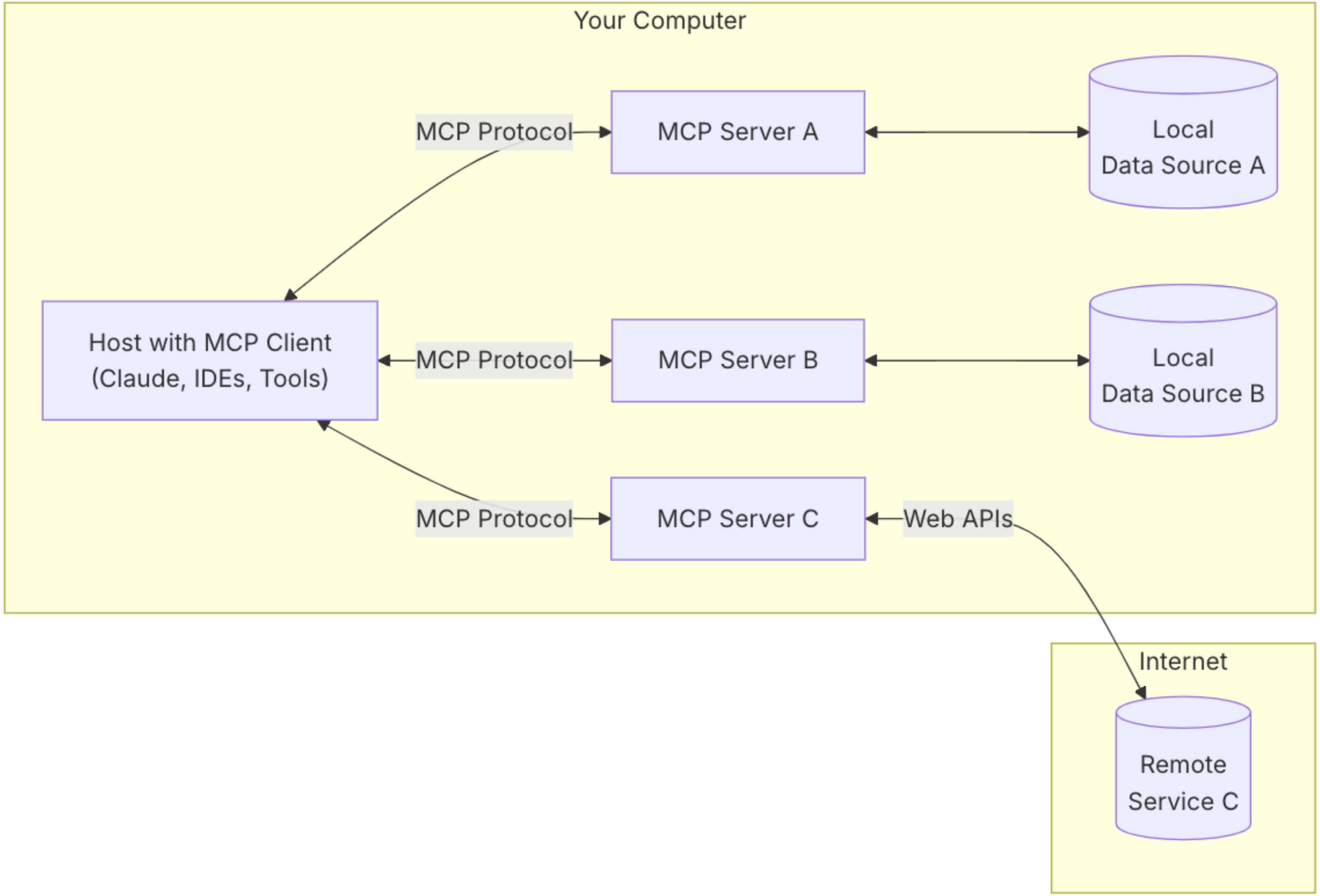


Source: https://google.github.io/A2A/

MCP (Model Context Protocol)

MCP is a open protocol that standardizes how applications provide context to LLMs.

MCP: USB-C port for AI applications.



Source: https://modelcontextprotocol.io/introduction

MCP and A2A

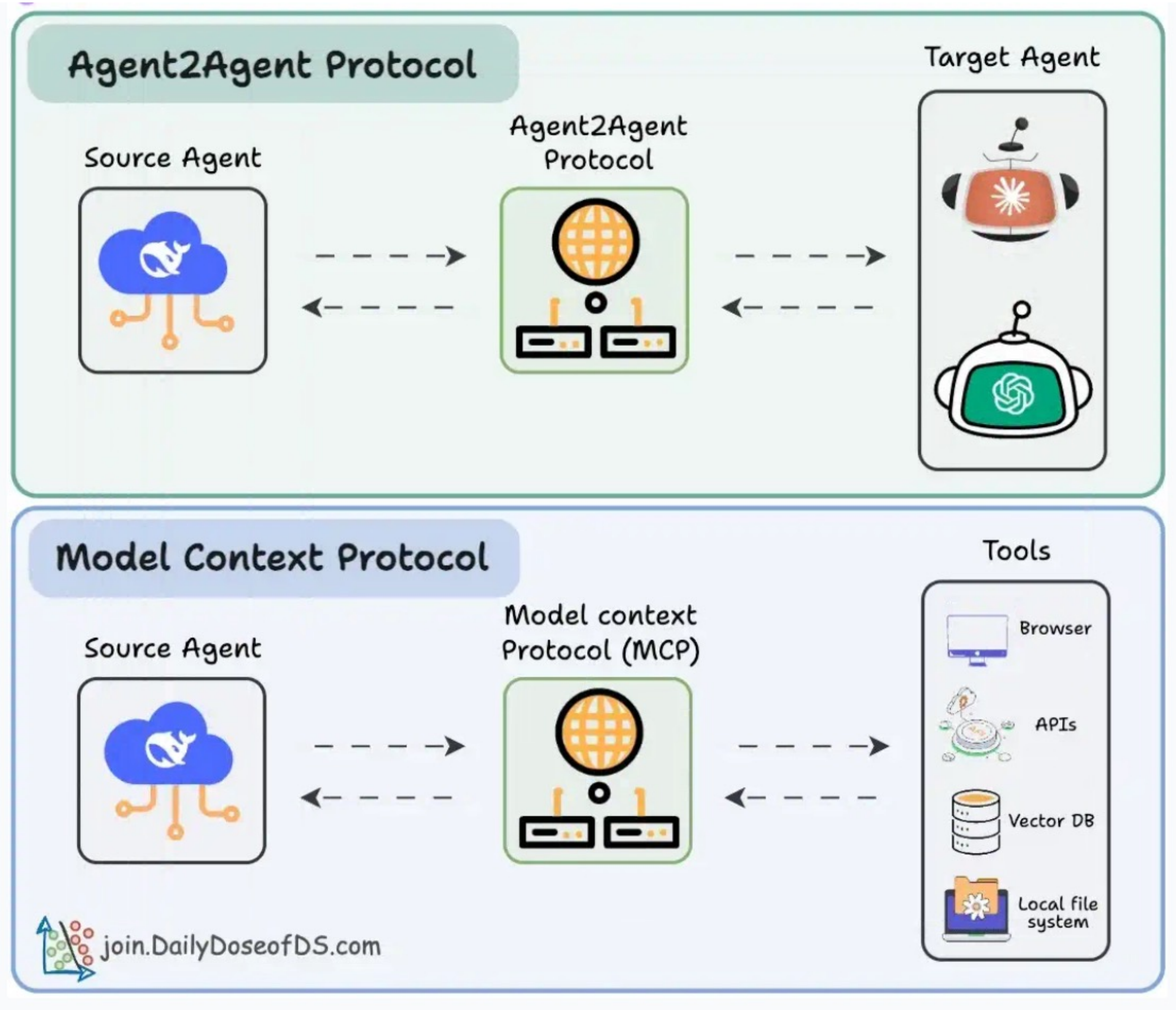
- MCP (Model Context Protocol) for tools and resources
  - Connect agents to tools, APIs, and resources with structured inputs/outputs.
  - Google ADK supports MCP tools. Enabling wide range of MCP servers to be used with agents.
- A2A (Agent2Agent Protocol) for agent-agent collaboration
  - Dynamic, multimodal communication between different agents without sharing memory, resources, and tools
  - Open standard driven by community.
  - Samples available using Google ADK, LangGraph, Crew.AI

Source: https://google.github.io/A2A/

Agentic applications require both A2A and MCP

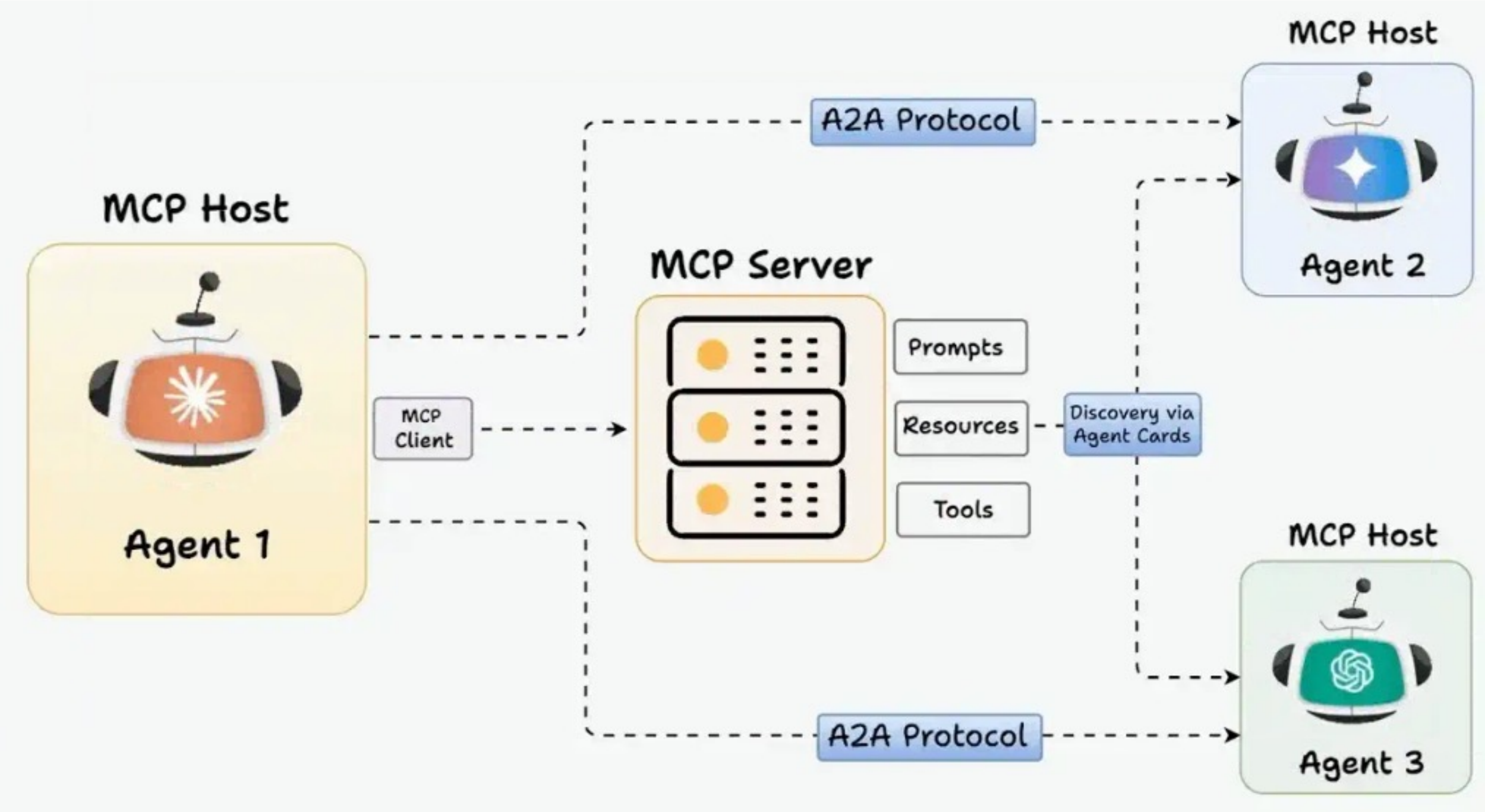
A2A allows agents to connect with other agents and collaborate in teams.

MCP provides agents with access to tools



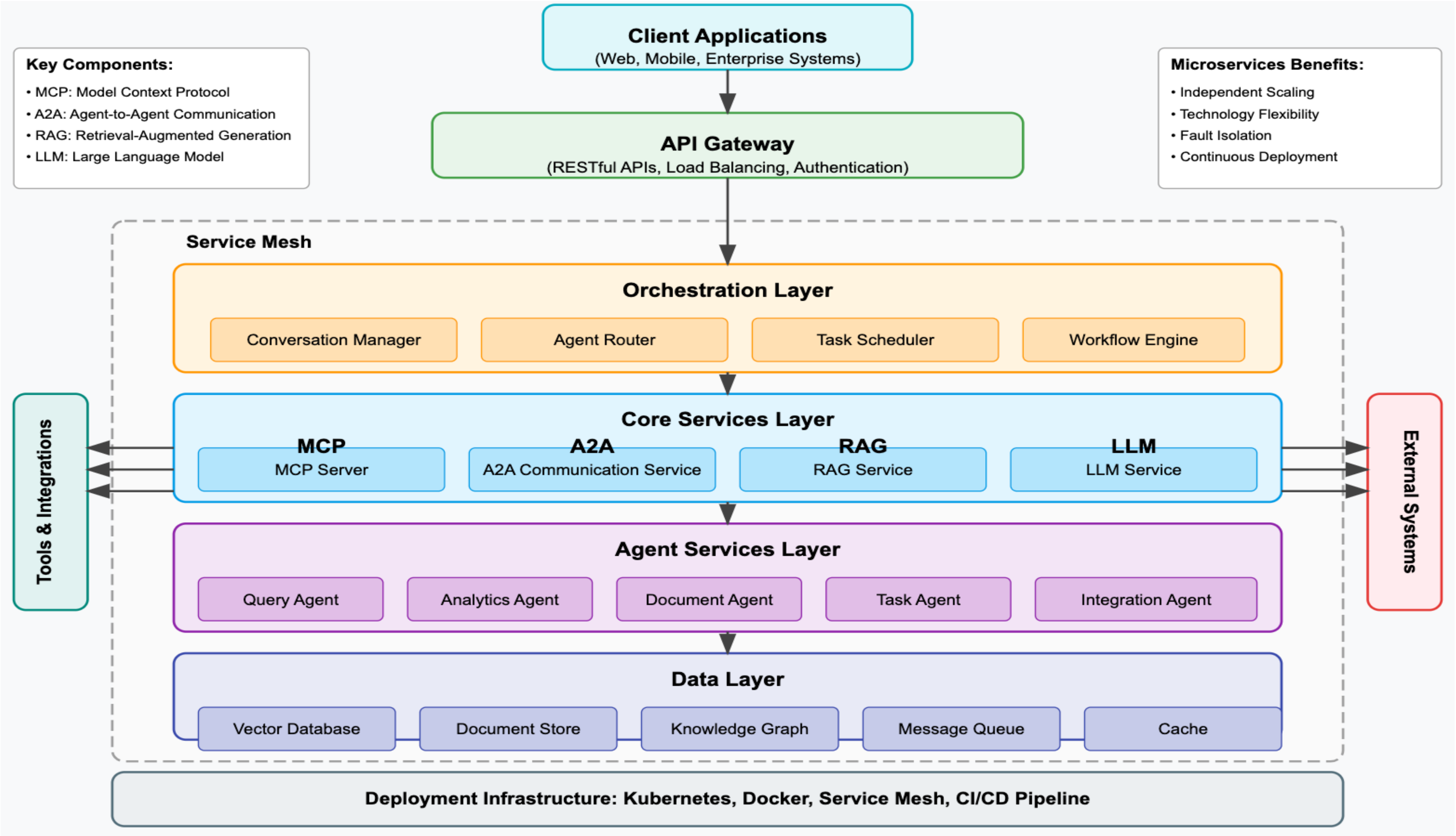
Source: https://blog.dailydoseofds.com/p/a-visual-guide-to-agent2agent-a2a

MCP and A2A Protocol for AI Agents

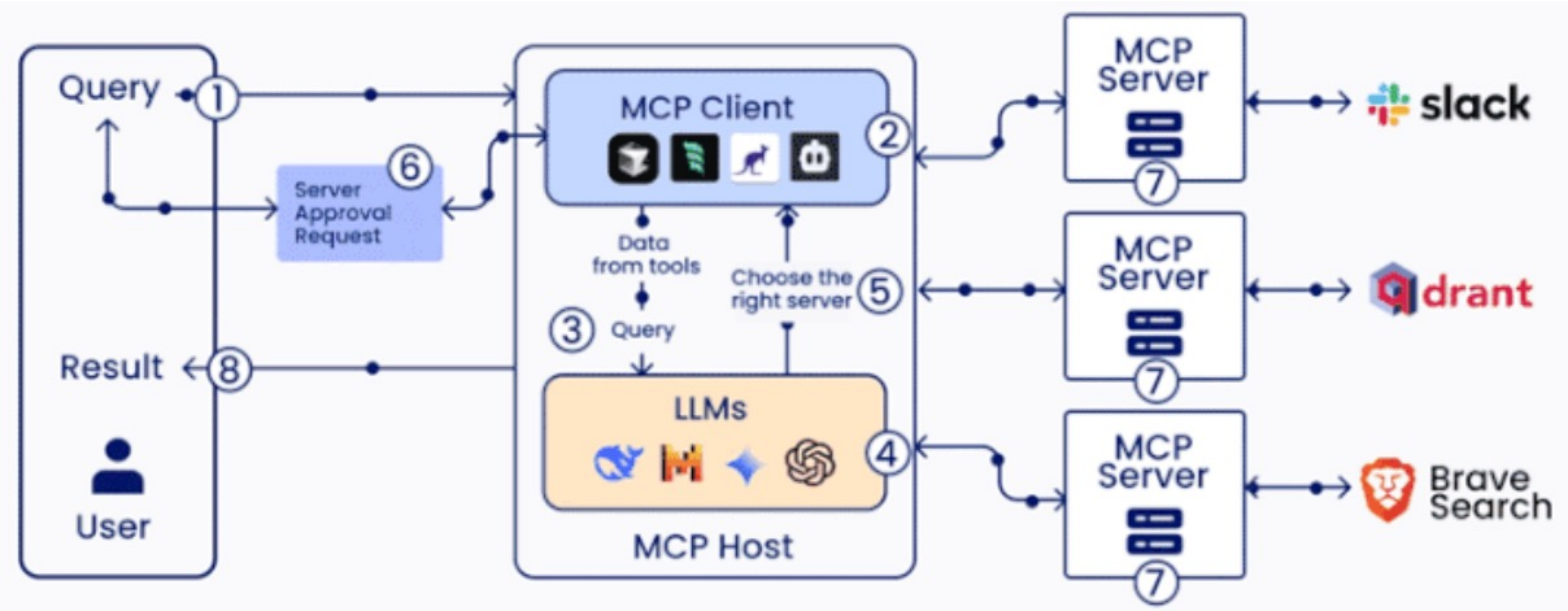


Source: https://blog.dailydoseofds.com/p/a-visual-guide-to-agent2agent-a2a

Agentic AI System with Microservices Architecture

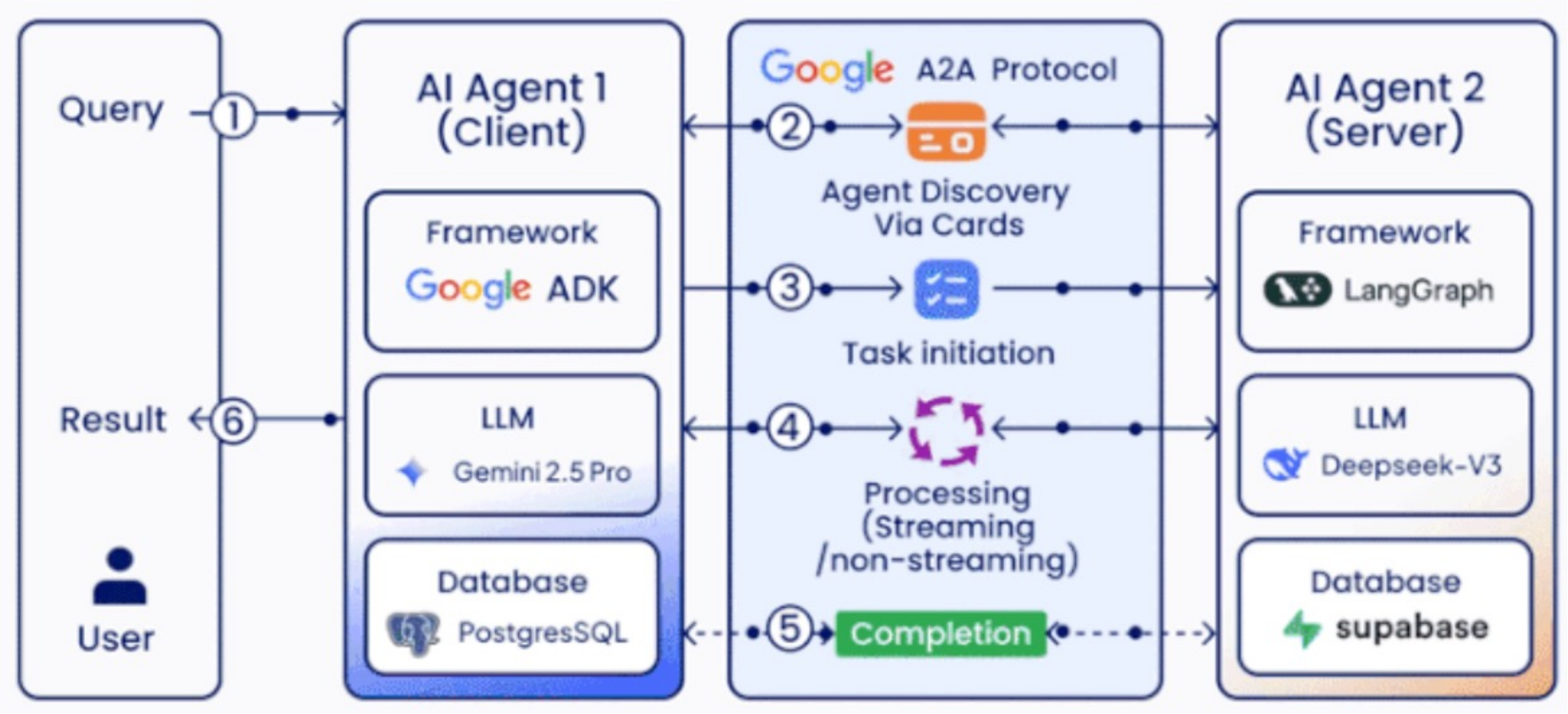


MCP (Model Context Protocol)



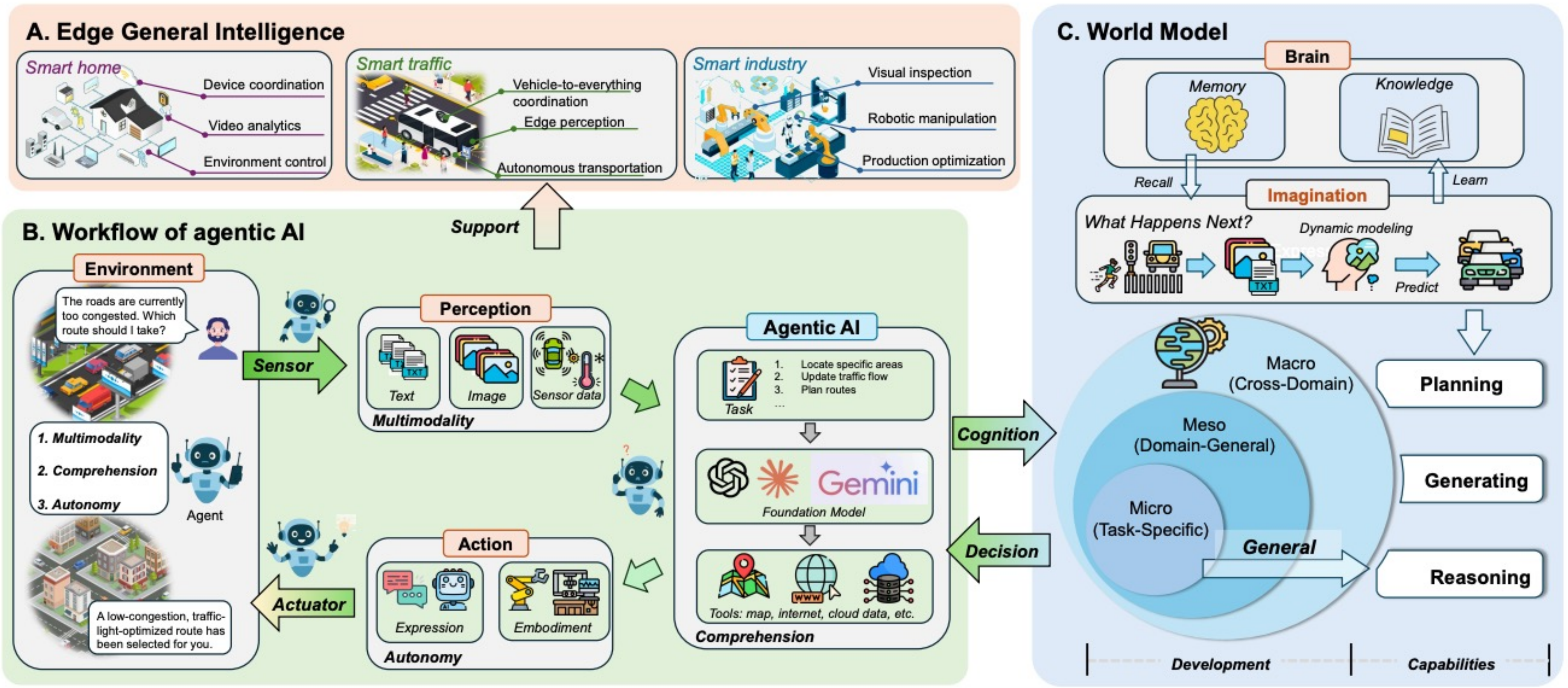
Source: Houssem Eddine Lassoued (2025), How MCP and A2A Protocols Are Revolutionizing AI Agent Development, https://medium.com/@houssemeddinelassoued/

A2A (Agent2Agent Protocol)



Source: Houssem Eddine Lassoued (2025), How MCP and A2A Protocols Are Revolutionizing AI Agent Development, https://medium.com/@houssemeddinelassoued/

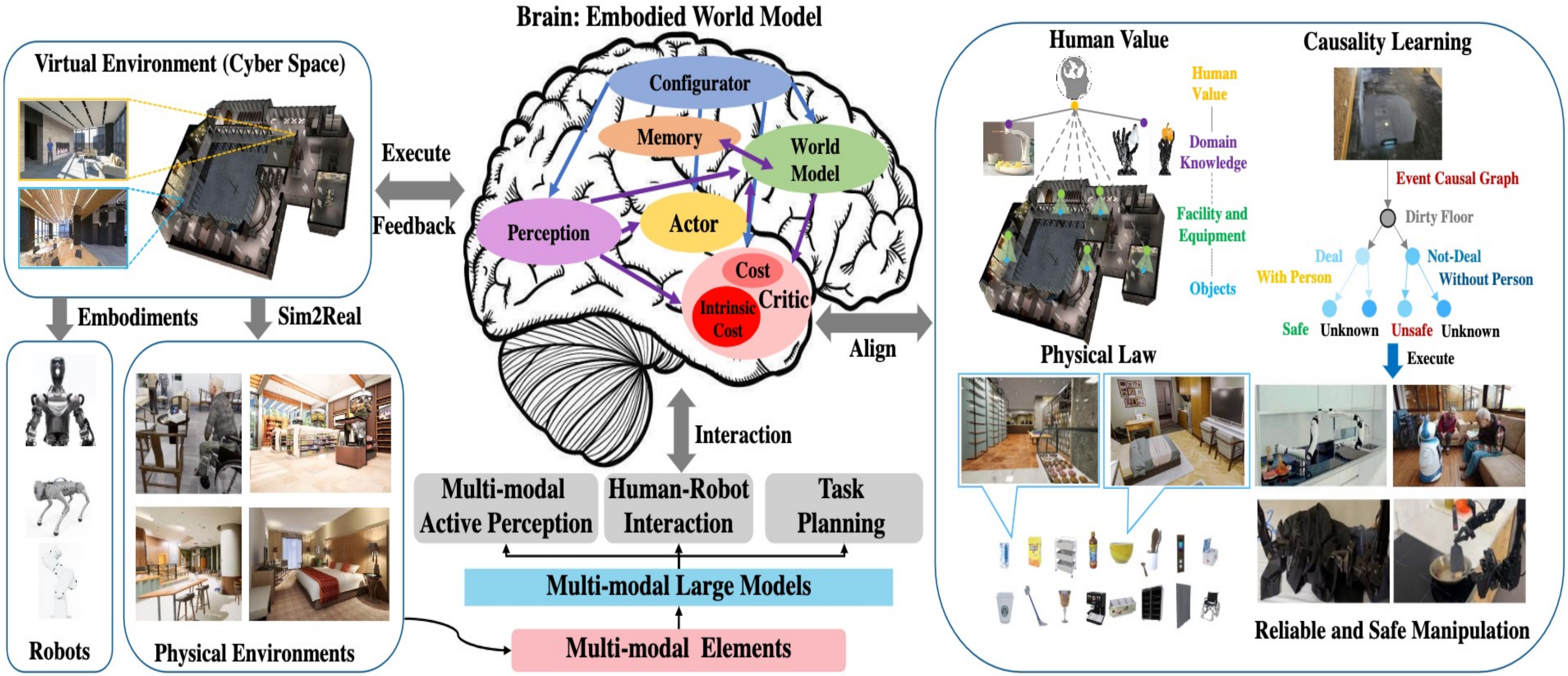
Agentic AI and World Model for Edge General Intelligence



Source: Changyuan Zhao, Guangyuan Liu, Ruichen Zhang, Yinqiu Liu, Jiacheng Wang, Jiawen Kang, Dusit Niyato et al (2025), "Edge general intelligence through world models and agentic AI: Fundamentals, solutions, and challenges." arXiv preprint arXiv:2508.09561.

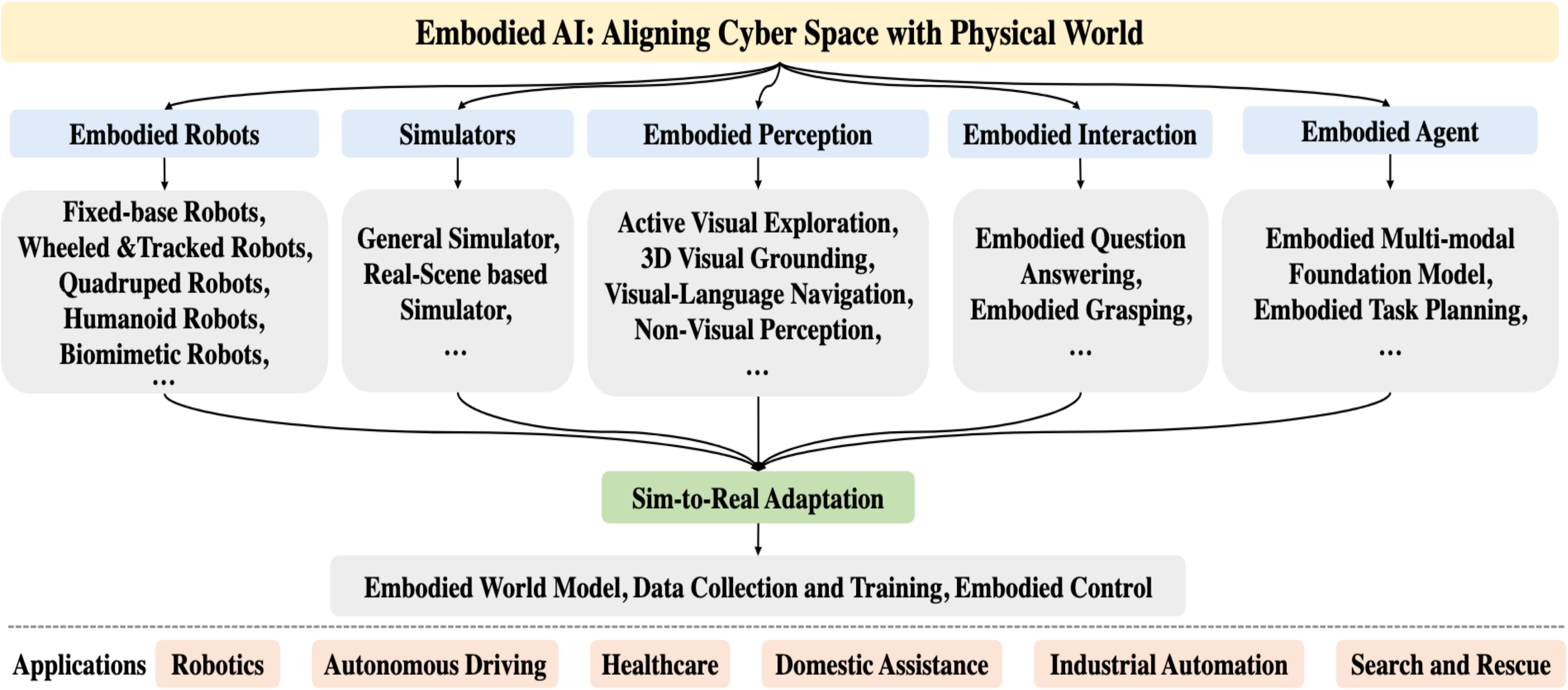
Physical AI (Robotics)

Framework of the Embodied Agent based on MLMs and WMs



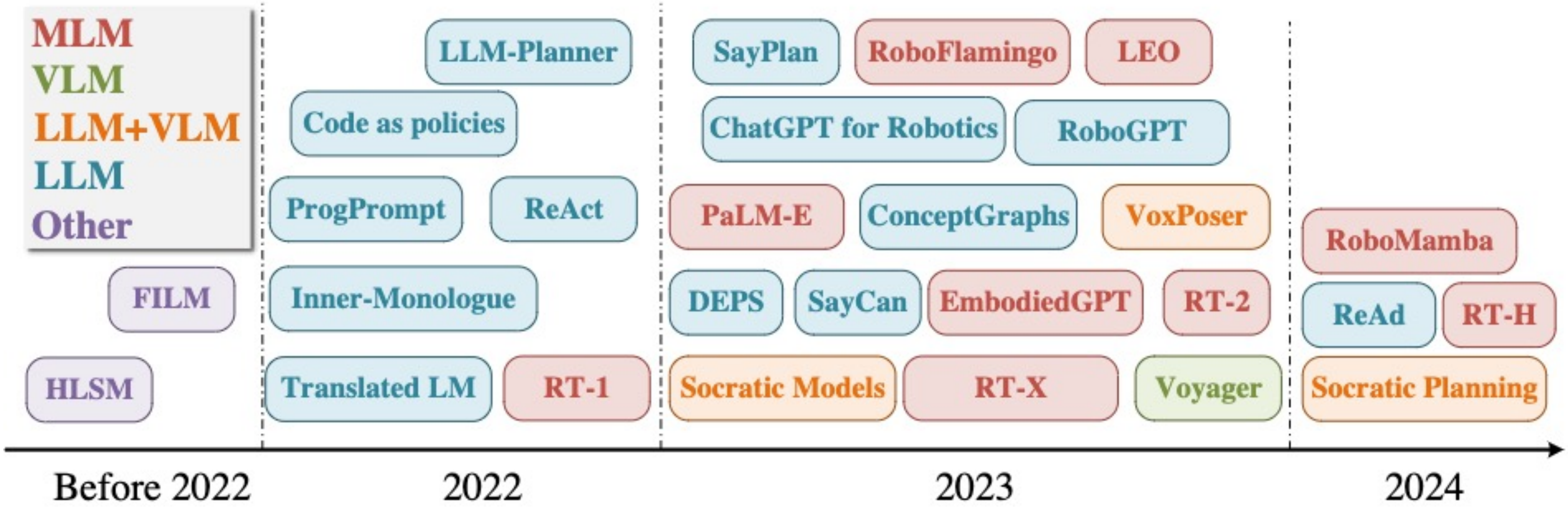
Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Embodied AI



Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Embodied Agents



MLM: Multimodal Language Model, which directly perceive the world and control the embodiment
VLM: Visual-Language Model with the outer policy models
LLM + VLM: LLM-based agent that perceives the world utilizing the VLM, and LLM
means the Large-Language Model with visual context and outer policy models.

Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Boston Dynamics: Spot

Automate sensing and inspection, capture limitless data, and explore without boundaries.

Source: https://www.bostondynamics.com/spot

Boston Dynamics: Atlas

The world's most dynamic humanoid robot

Atlas is a research platform designed to push the limits of whole-body mobility

Source: https://www.bostondynamics.com/atlas

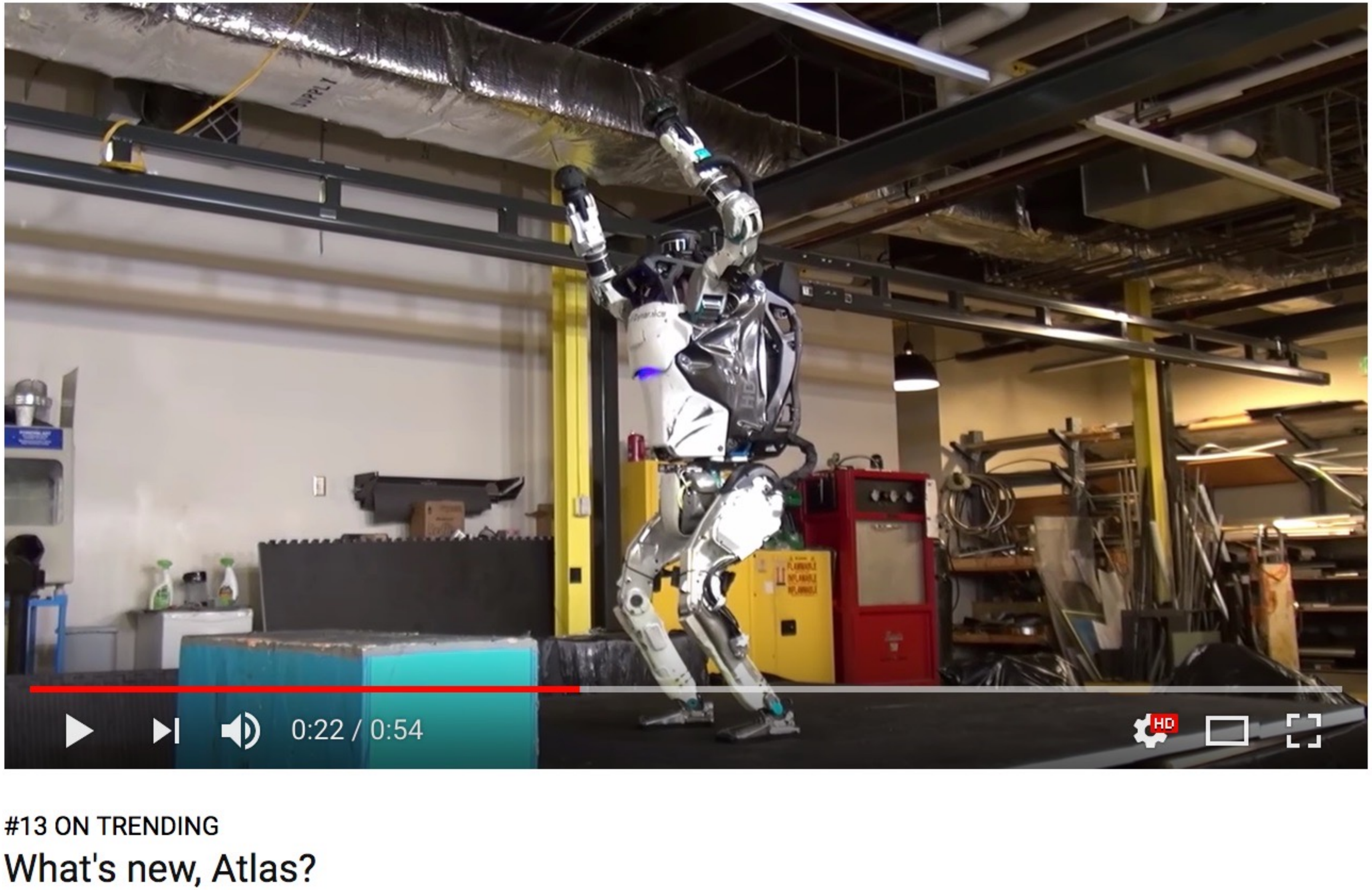
Boston Dynamics: Atlas Goes Hands On

Atlas uses a machine learning (ML) vision model to detect and localize the environment fixtures and individual bins.

The robot uses a specialized grasping policy and continuously estimates the state of manipulated objects to achieve the task.

Source: https://bostondynamics.com/video/atlas-goes-hands-on/

Boston Dynamics: Atlas



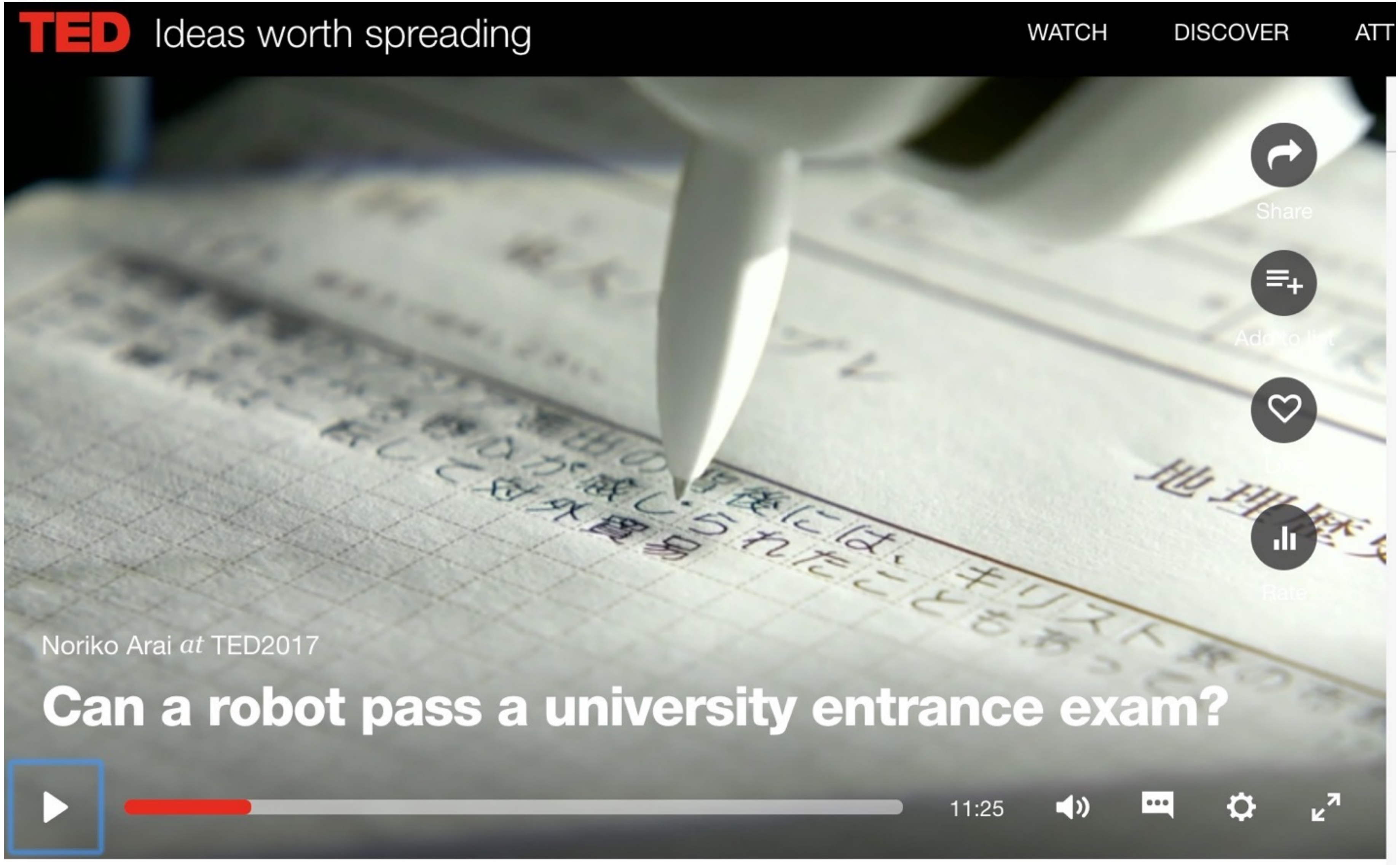
https://www.youtube.com/watch?v=fRj34o4hN4I

Humanoid Robot: Sophia

https://www.youtube.com/watch?v=S5t6K9iwcdw

Can a robot pass a university entrance exam?

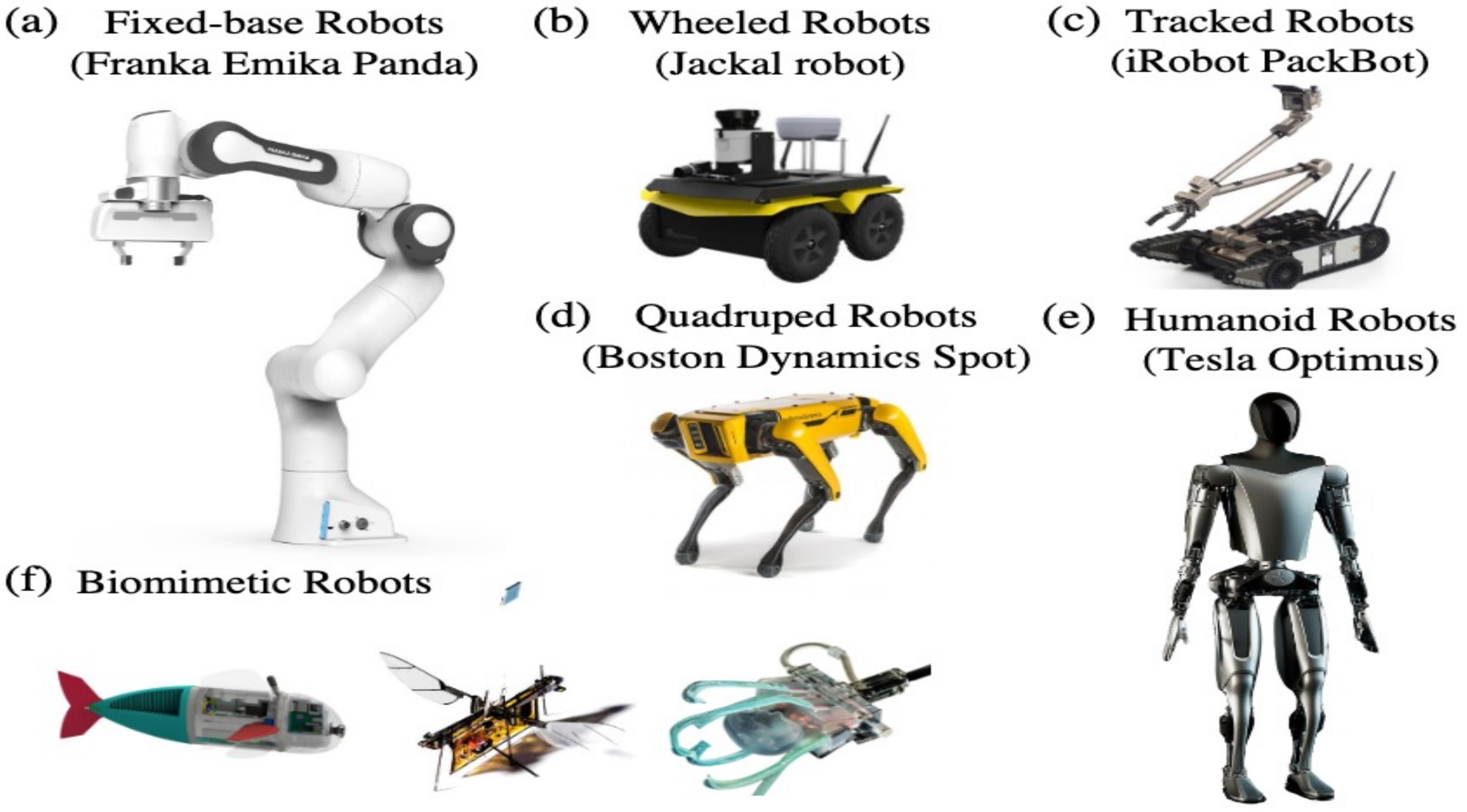
Noriko Arai at TED2017



https://www.ted.com/talks/noriko_arai_can_a_robot_pass_a_university_entrance_exam

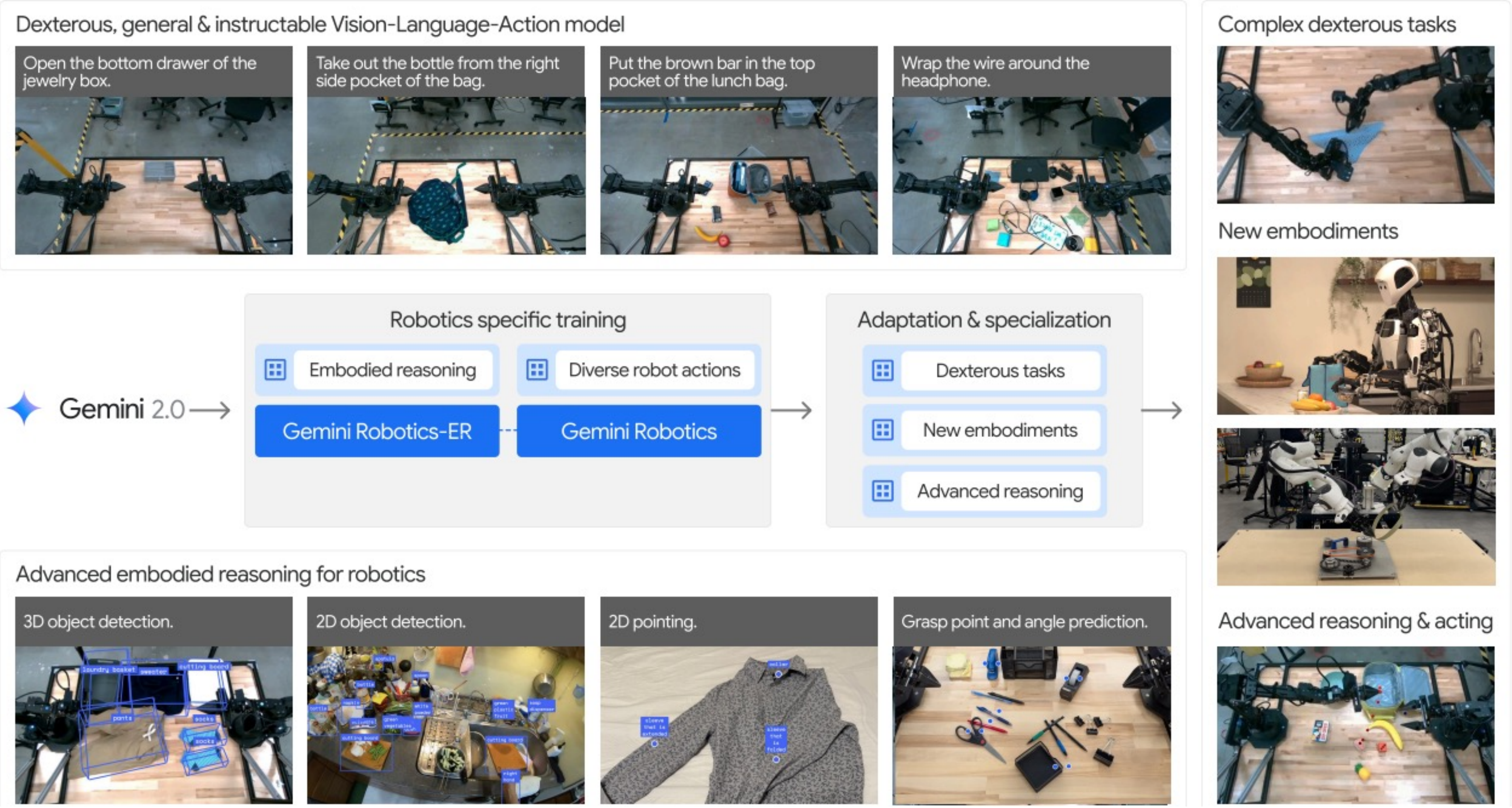
https://www.youtube.com/watch?v=XQZjkPyJ8KU

Embodied Robots



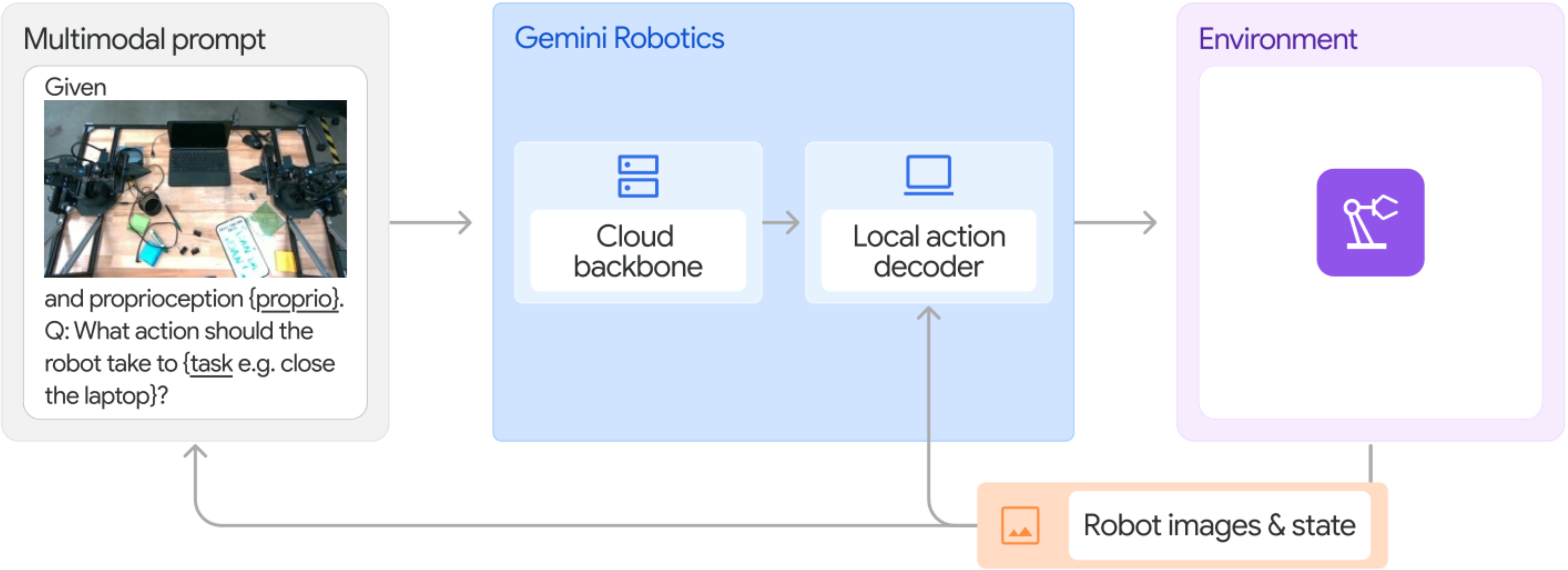
Source: Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.

Gemini Robotics:
Bringing AI into the Physical World



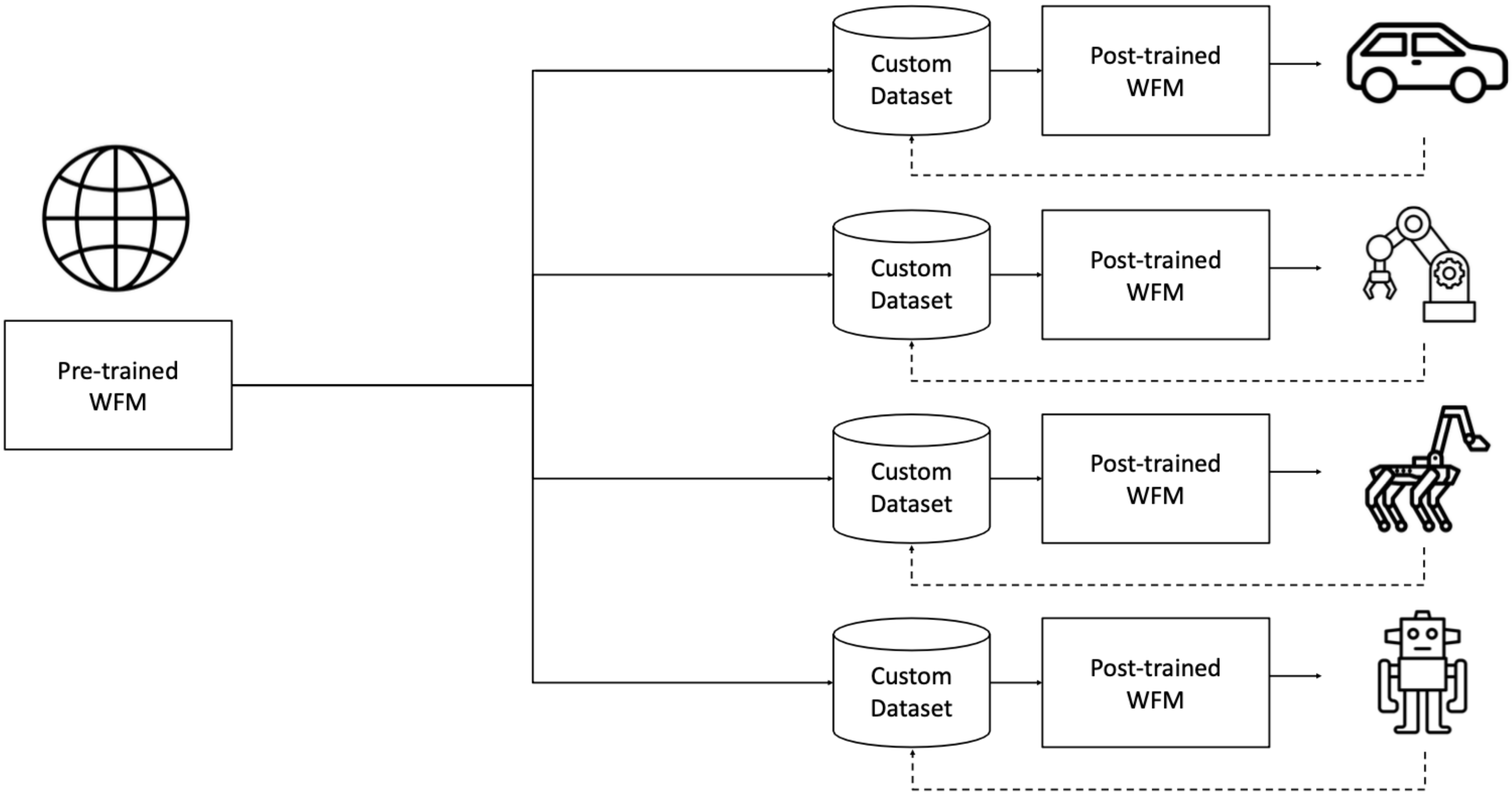
Source: Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna et al.(2025) "Gemini robotics: Bringing ai into the physical world." arXiv preprint arXiv:2503.20020 (2025).

Gemini Robotics Models: Architecture, Input and Output



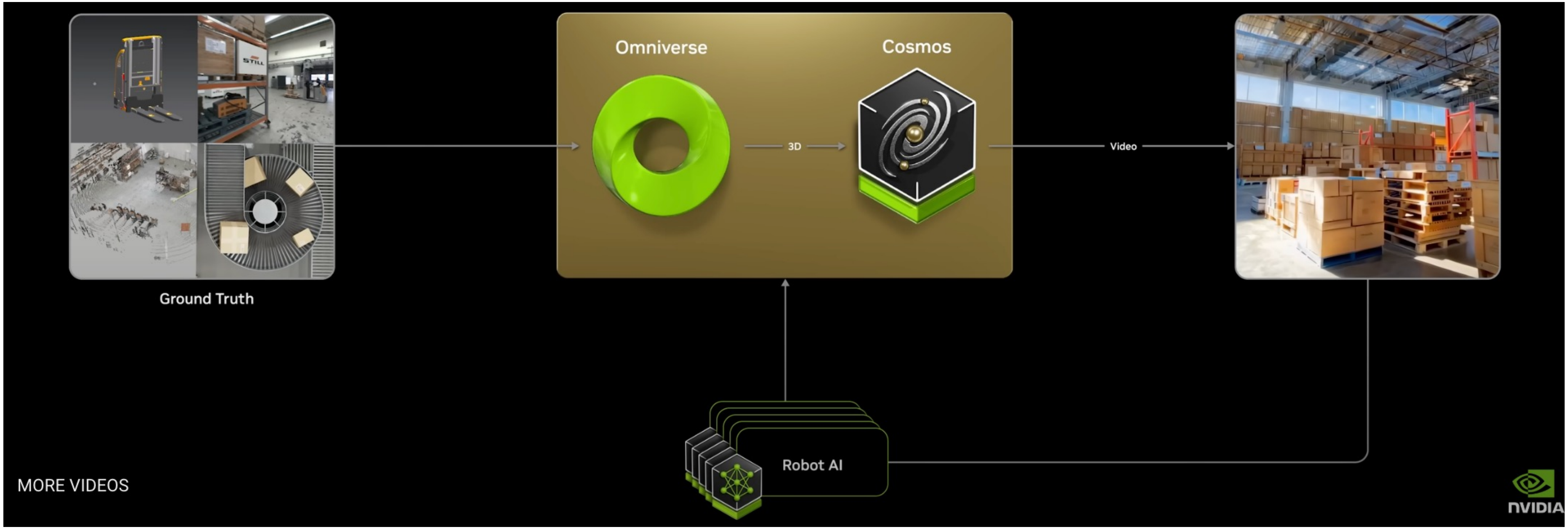
Source: Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna et al.(2025) "Gemini robotics: Bringing ai into the physical world." arXiv preprint arXiv:2503.20020 (2025).

World Foundation Model Platform for Physical AI



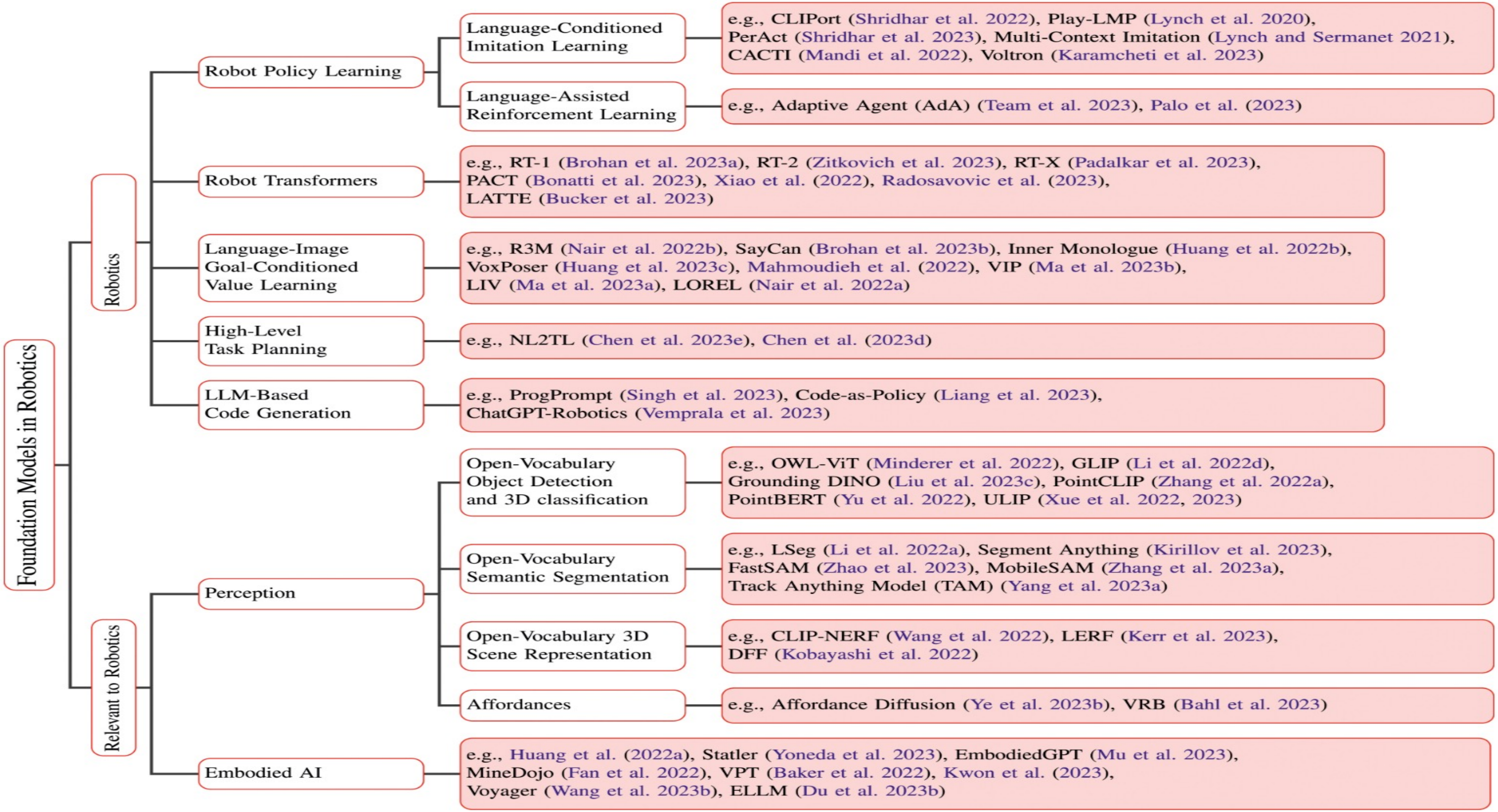
Source: Agarwal, Niket, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay et al. "Cosmos world foundation model platform for physical ai." arXiv preprint arXiv:2501.03575 (2025).

NVIDIA Cosmos
World Foundation Model Platform for Physical AI



Source: Agarwal, Niket, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay et al. "Cosmos world foundation model platform for physical ai." arXiv preprint arXiv:2501.03575 (2025)., https://research.nvidia.com/labs/dir/cosmos-predict1/

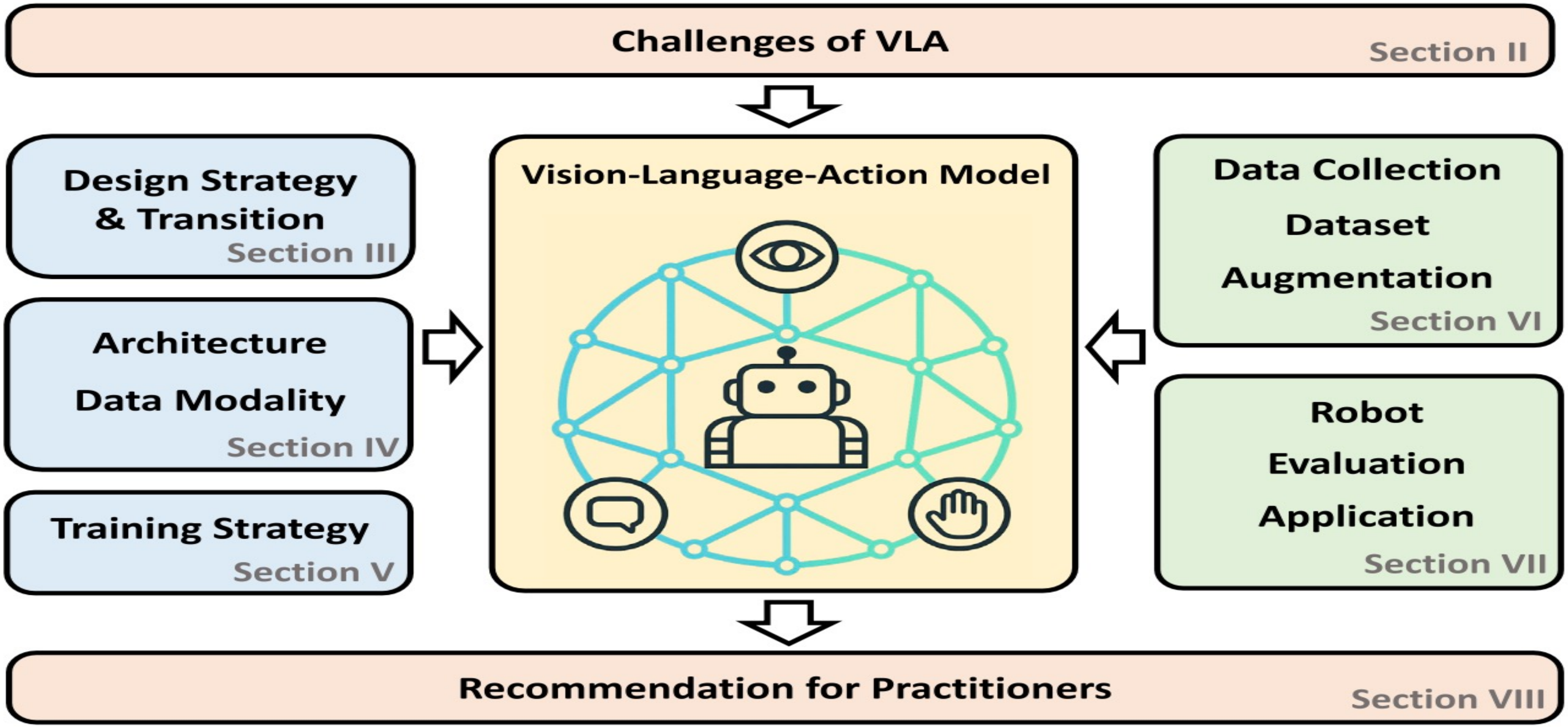
Foundation Models in Robotics



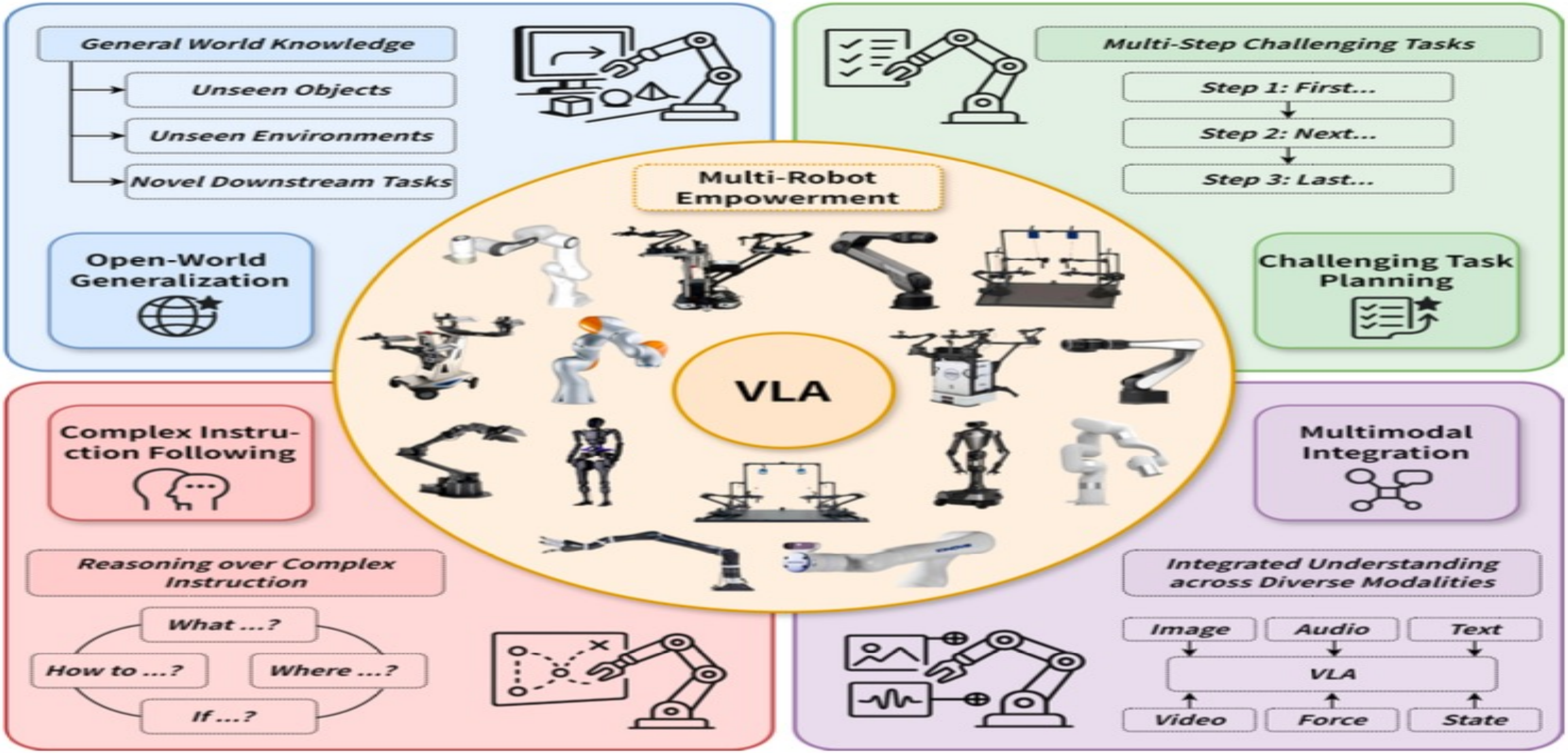
Source: Firoozi, Roya, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu et al. "Foundation models in robotics: Applications, challenges, and the future. " The International Journal of Robotics Research 44, no. 5 (2025): 701-739.

Vision Language Action (VLA) Models for Robotics

Vision-Language-Action (VLA) Models for Robotics

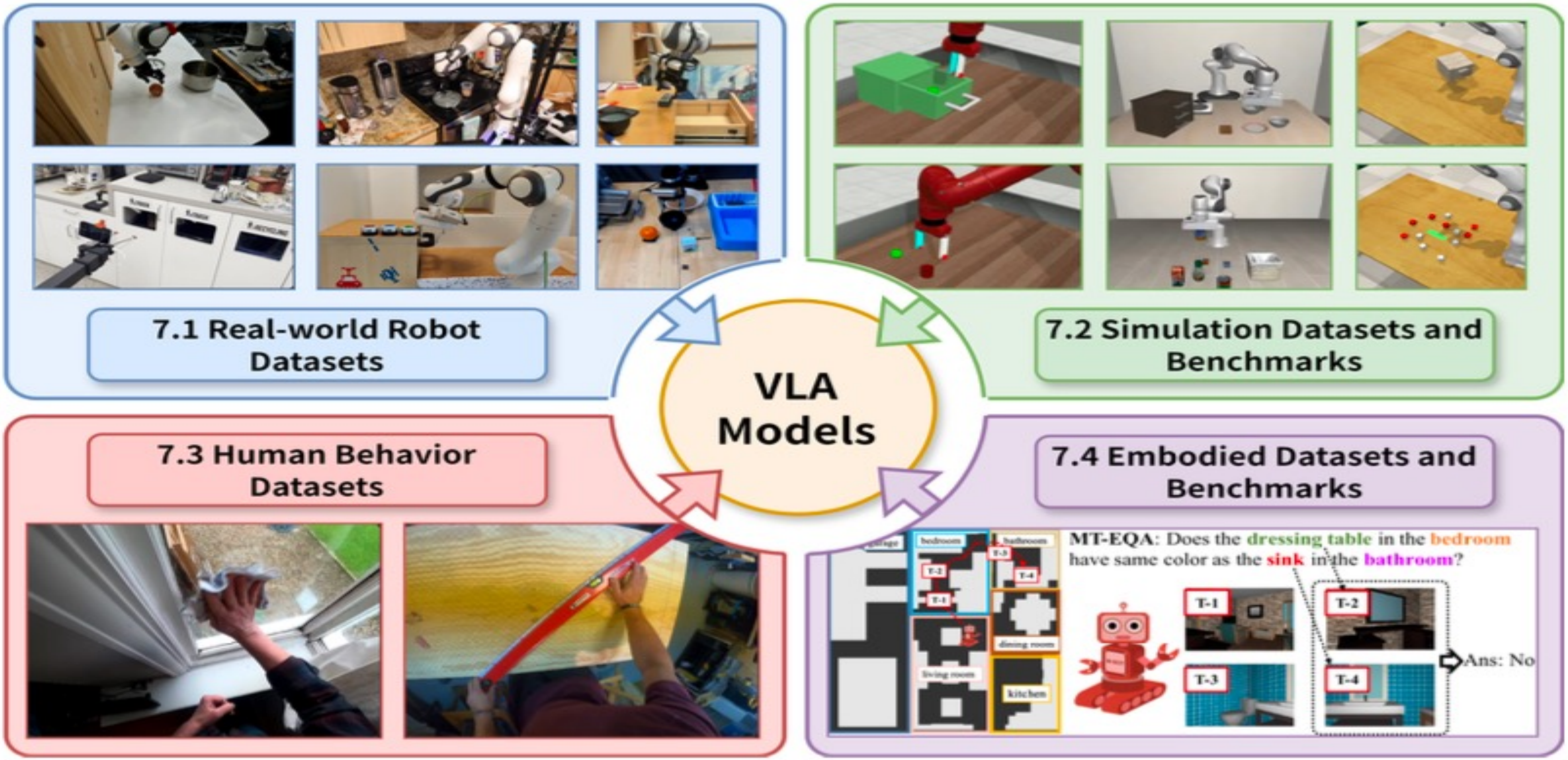


Large VLM-based Vision-Language-Action Models for Robotic Manipulation



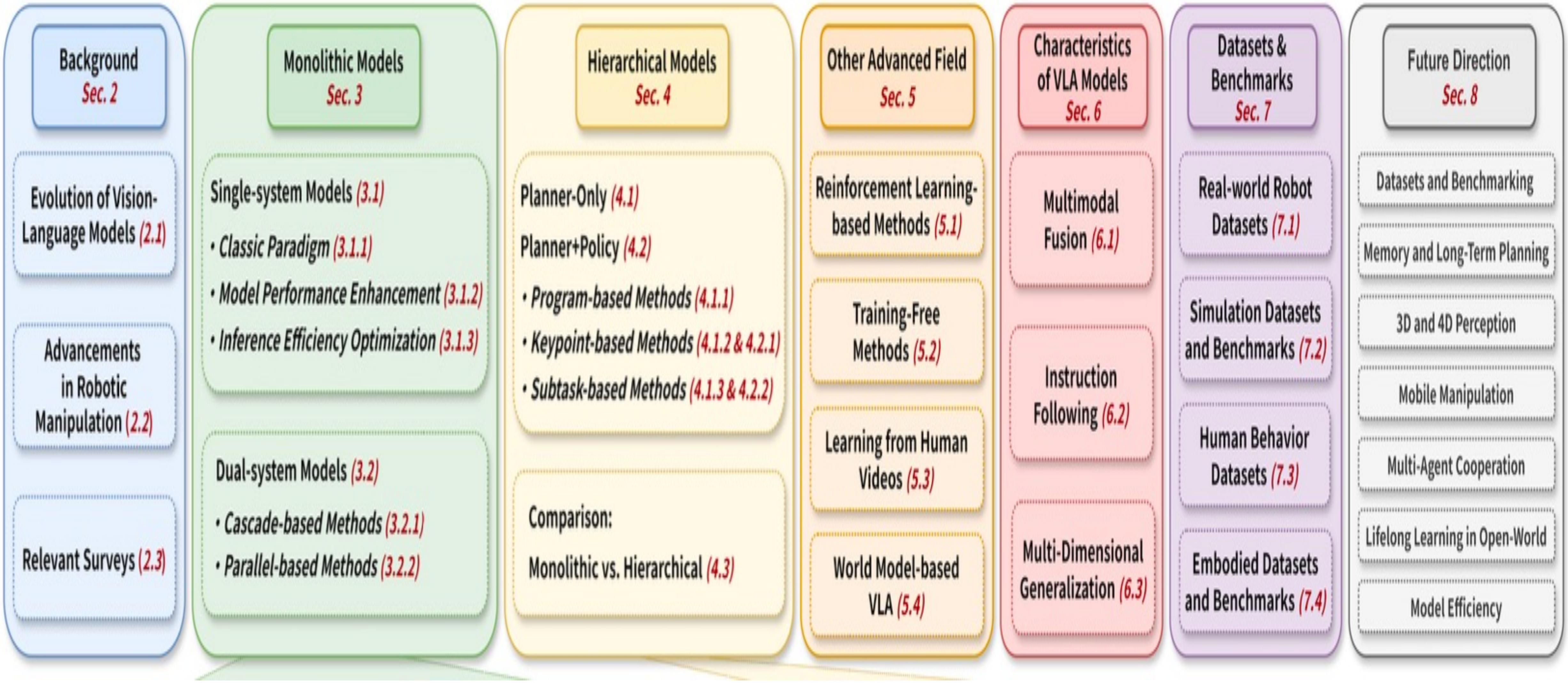
Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

Large VLM-based Vision-Language-Action Models for Robotic Manipulation



Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

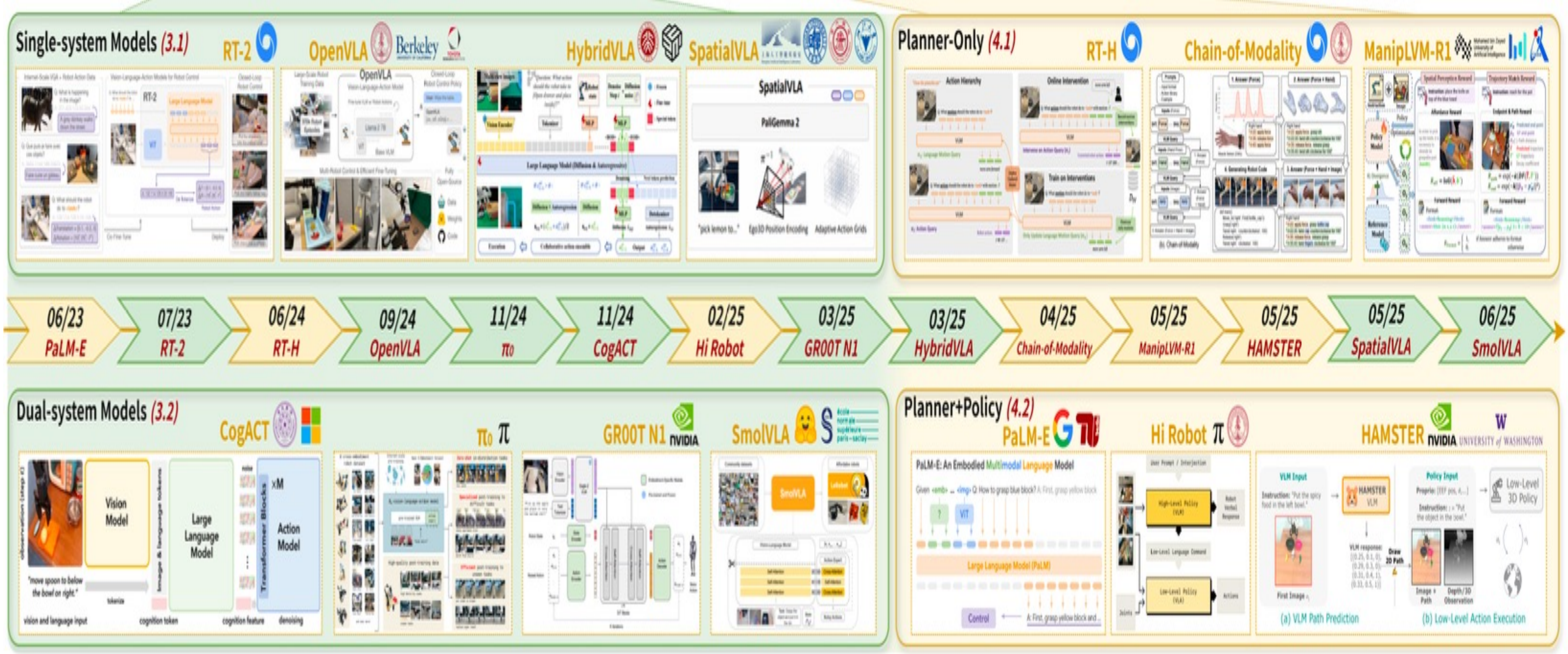
Large VLM-based Vision-Language-Action Models for Robotic Manipulation



Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

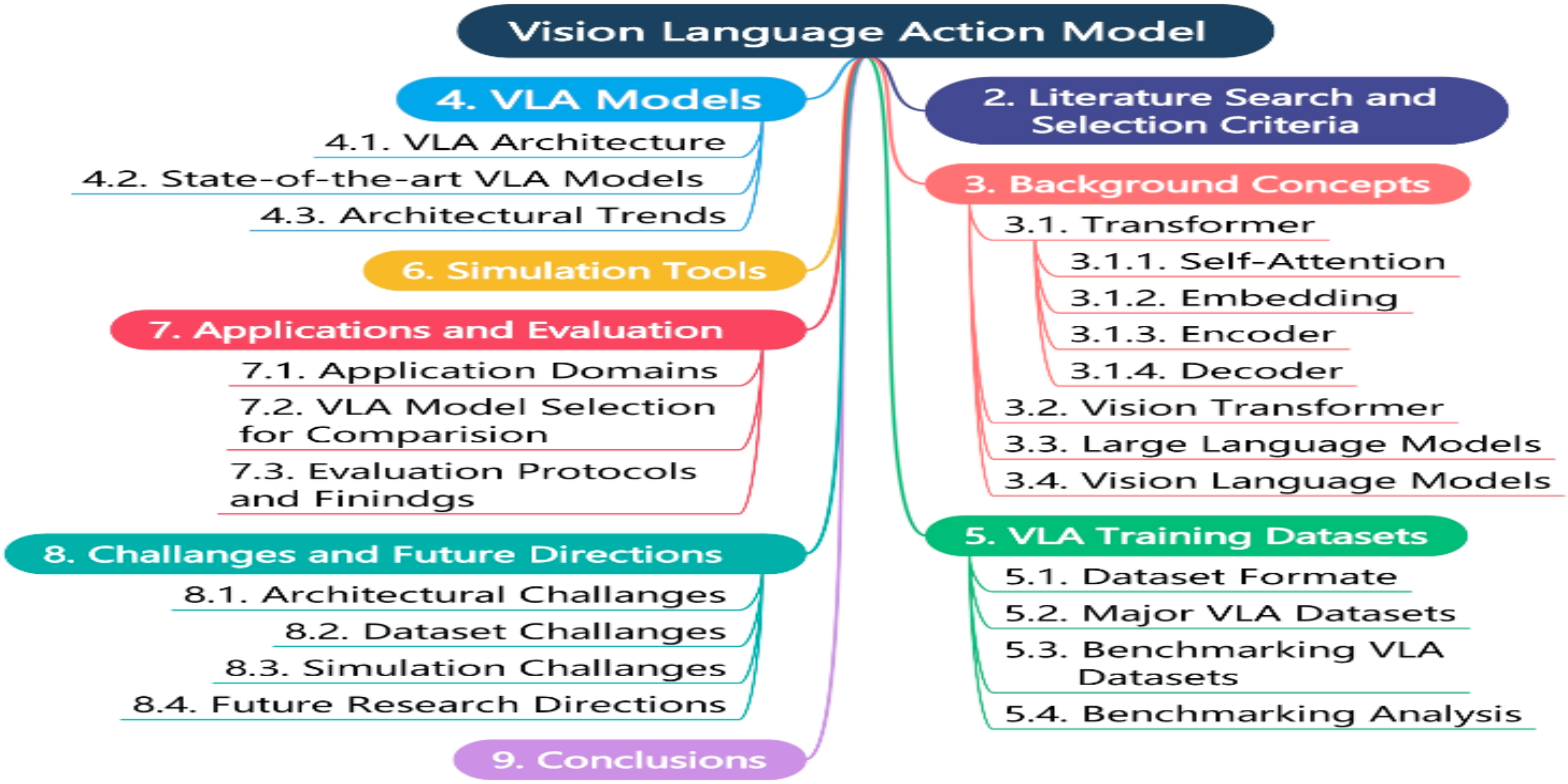
Large VLM-based Vision-Language-Action Models for Robotic Manipulation (Timeline)

Monolithic models and Hierarchical Models



Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

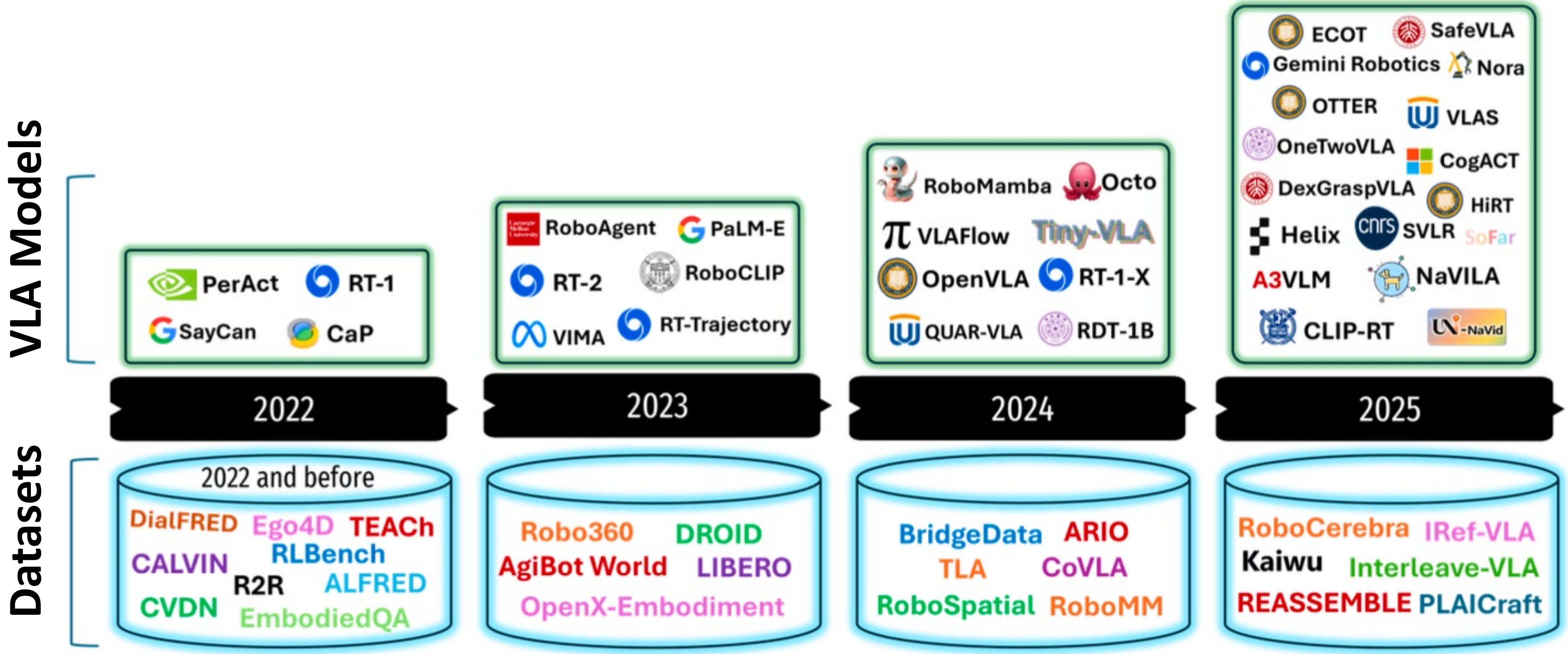
Vision Language Action Models in Robotic Manipulation



Source: Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025).

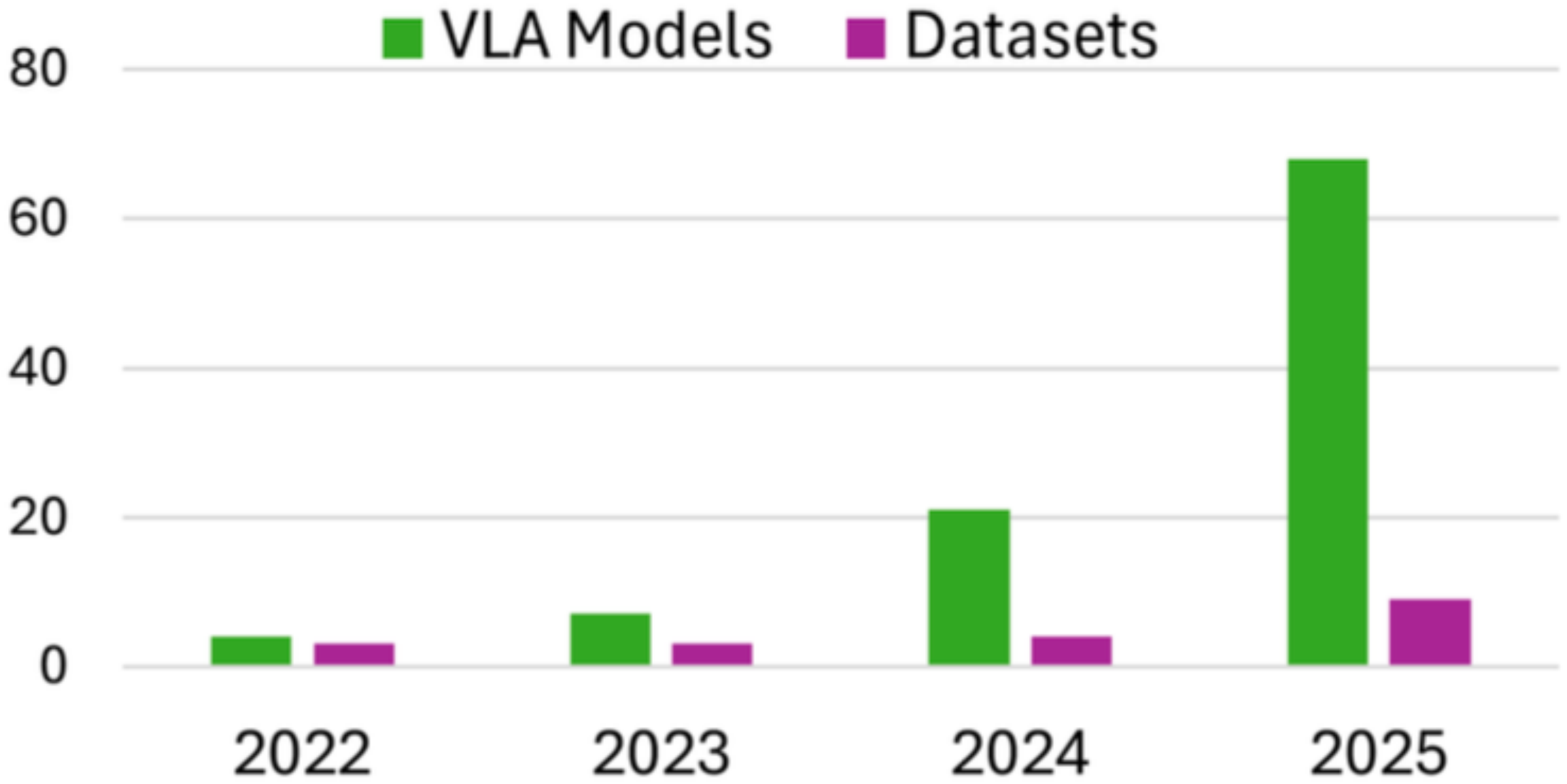
Vision Language Action (VLA) Models, Datasets

Contributing institutions: Academic (e.g., CMU, CNRS, UC, Peking Uni)
Industrial Labs (e.g., Google, NVIDIA, Microsoft)



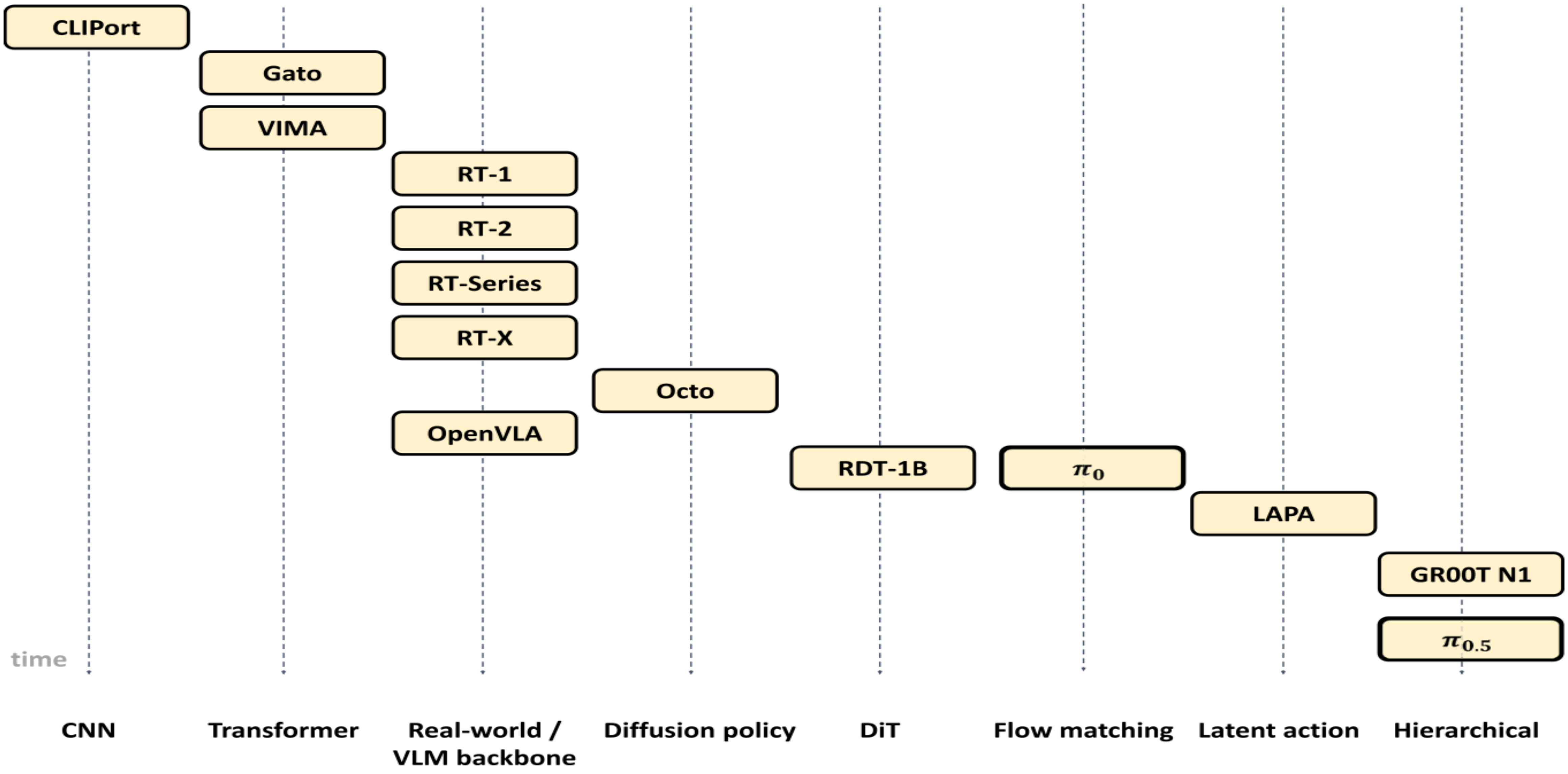
Source: Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025).

VLA Models and Foundational VLA Datasets



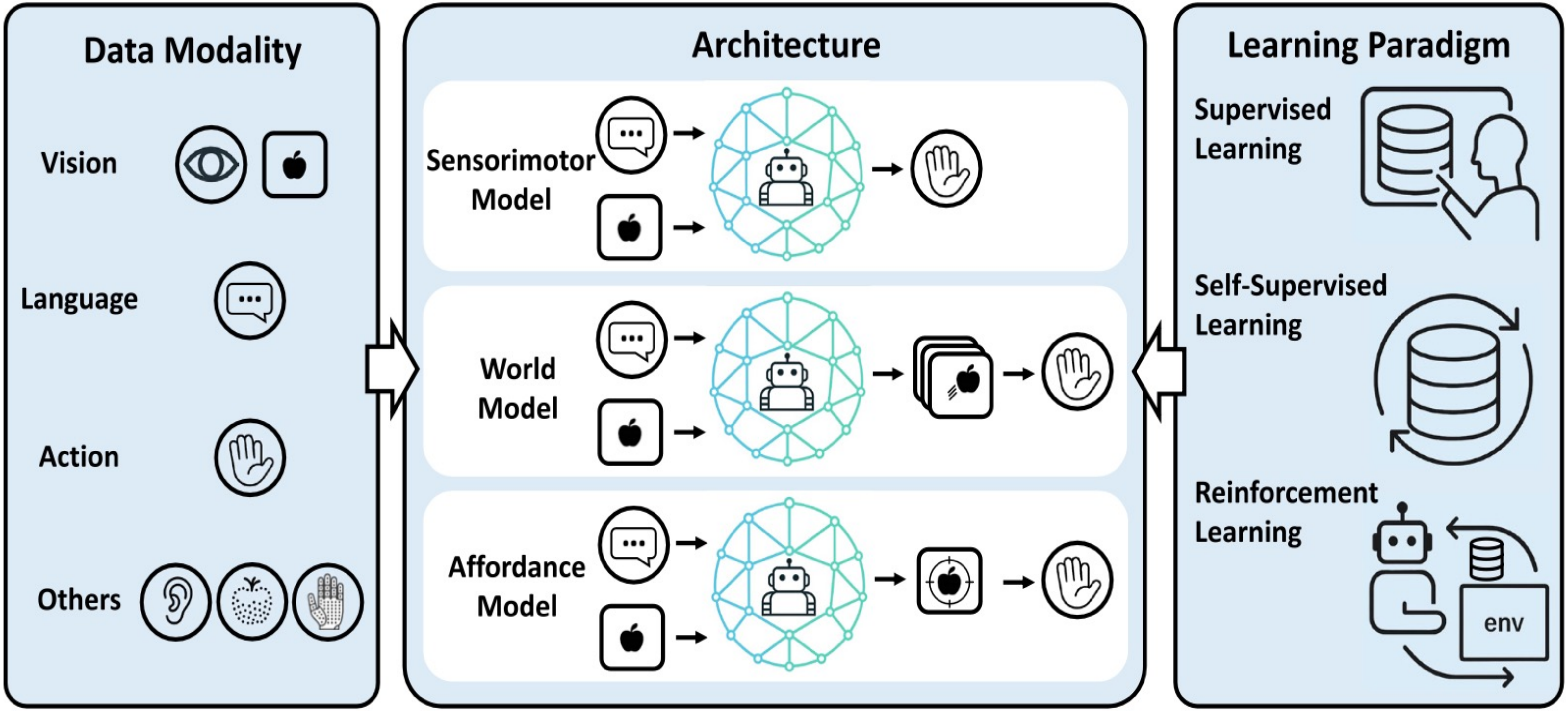
Source: Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025).

Timeline of Vision-Language-Action (VLA) Models



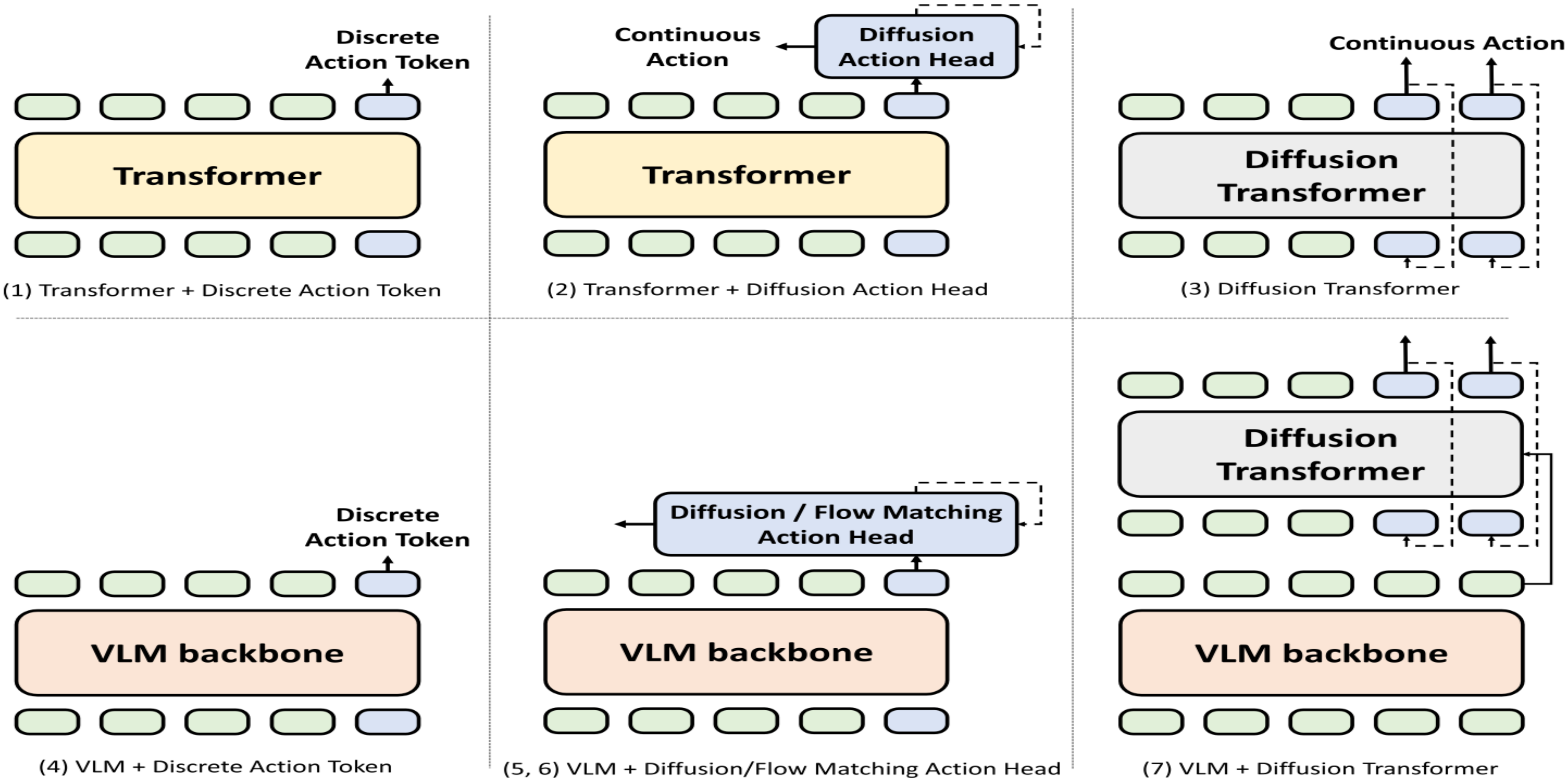
Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

VLA Model Components and Training Paradigms



Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

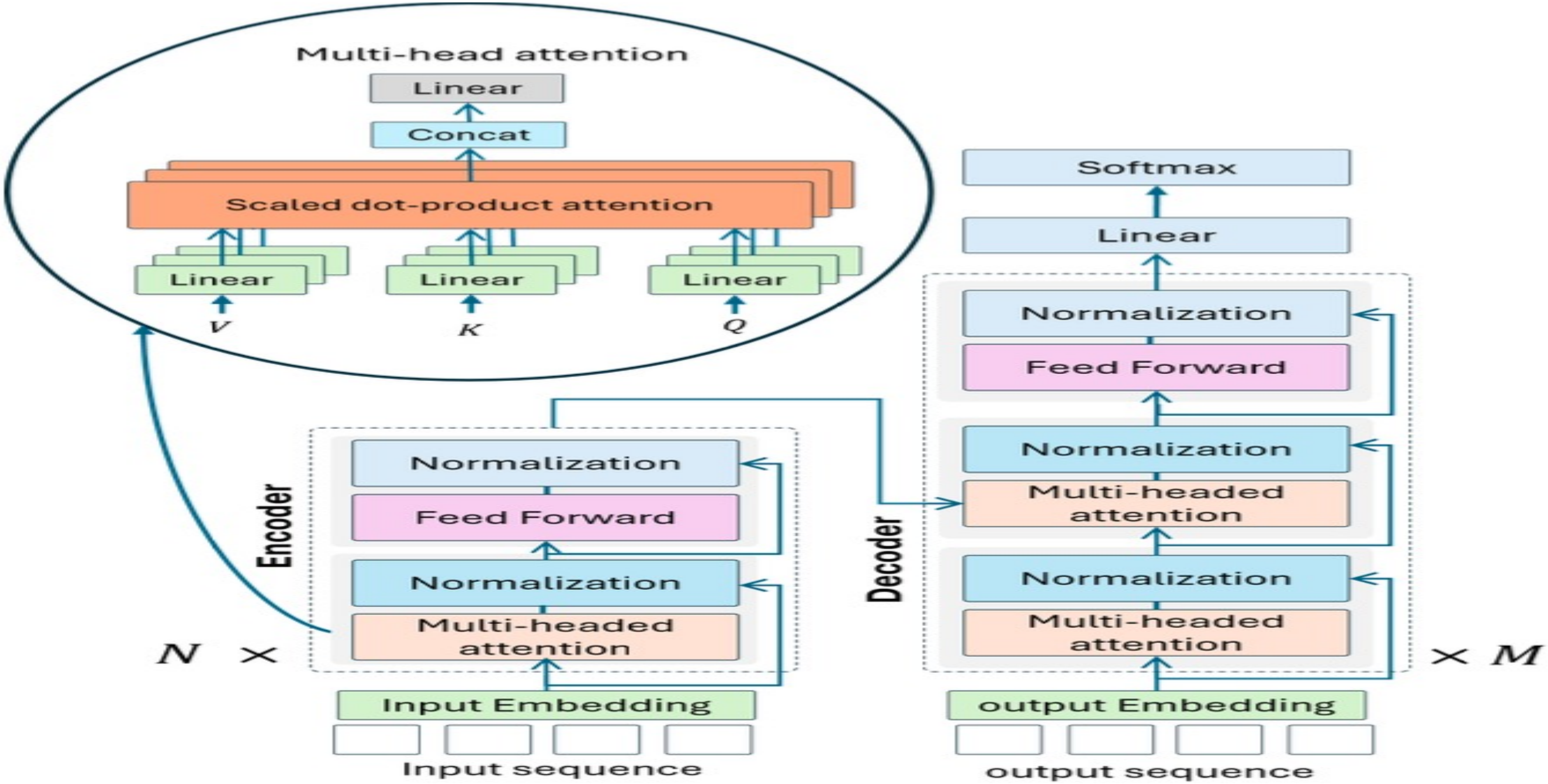
Architecture of Sensorimotor Models for VLA



Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

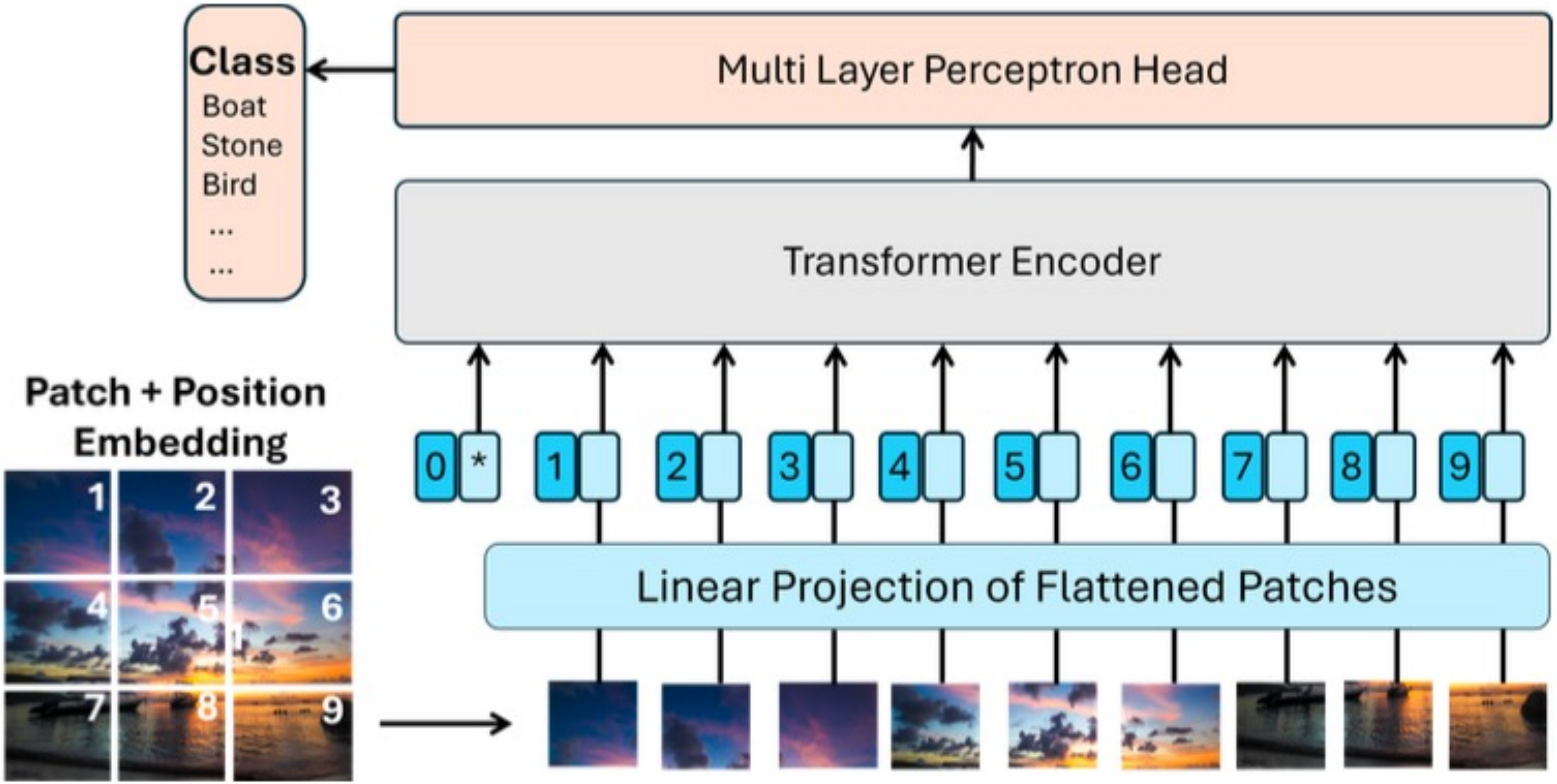
Transformer Architecture:

Encoder-decoder structure and the internal mechanism of multi-head attention



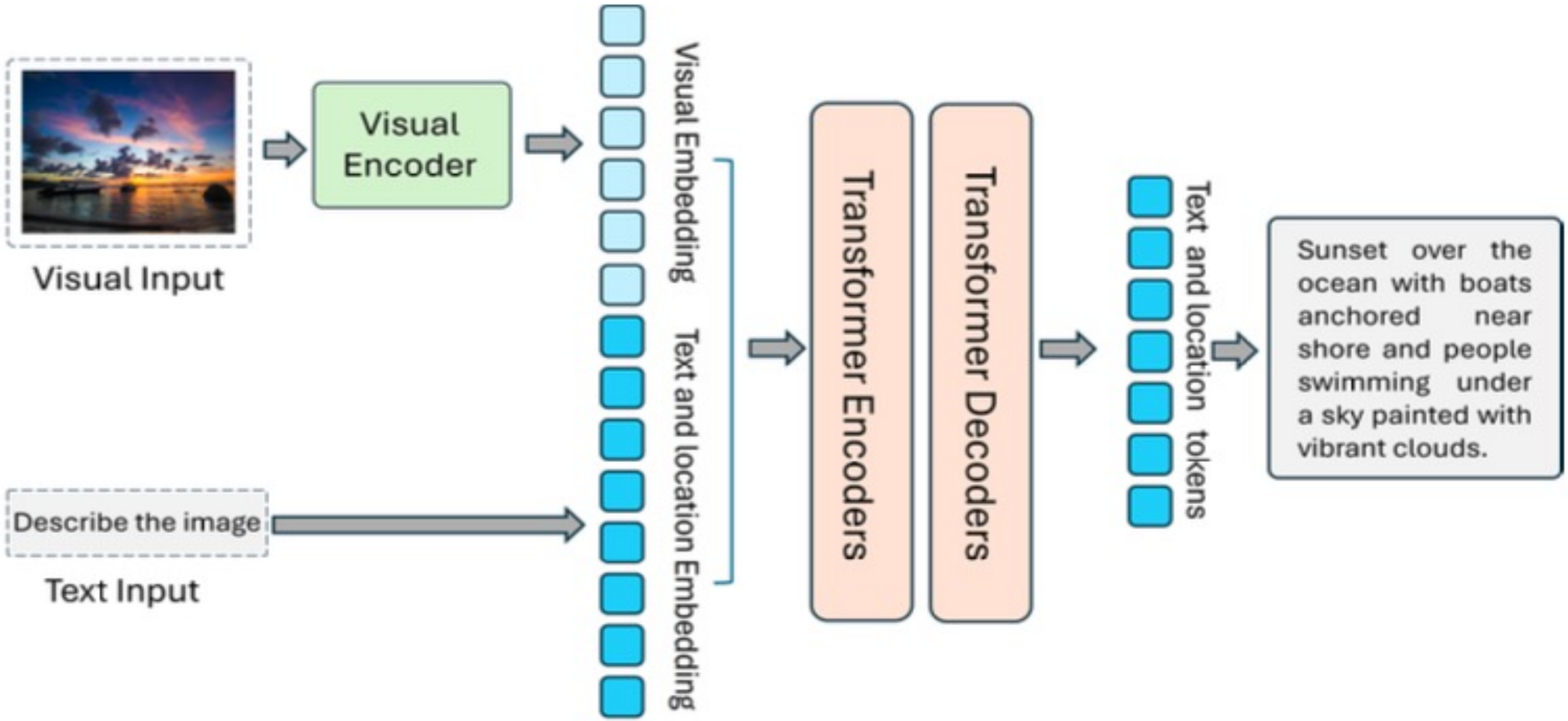
Source: Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025).

Architecture of the ViT



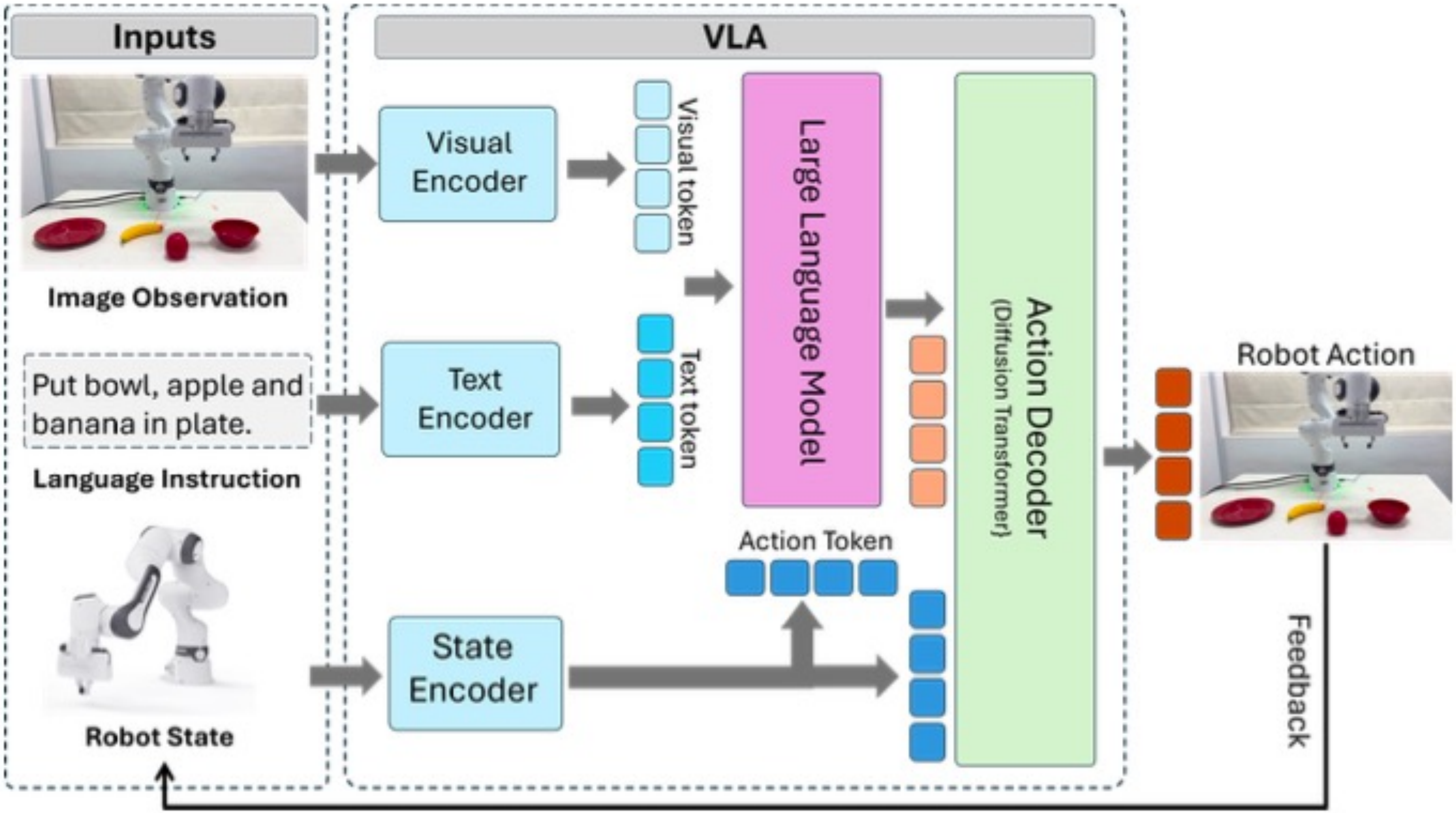
Source: Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025).

Architecture of VLM
for Image Captioning and Semantic Understanding

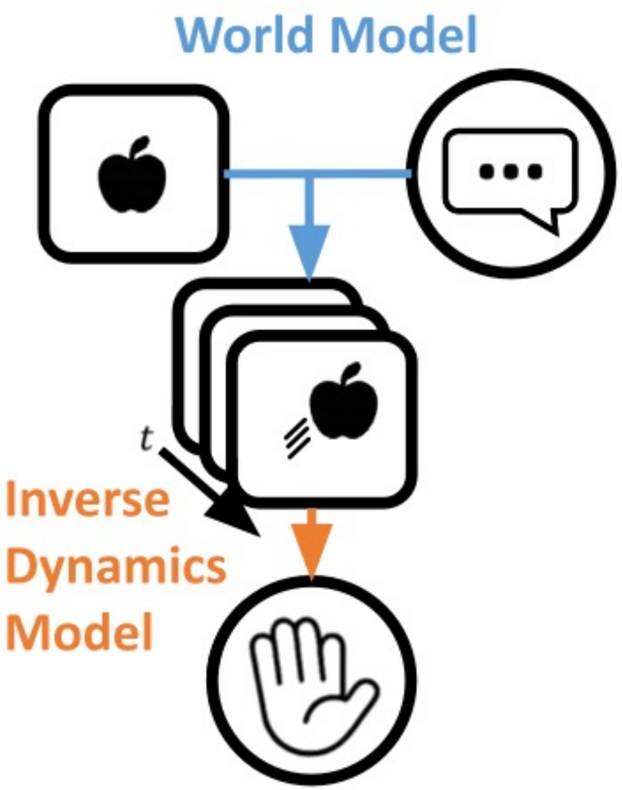


Source: Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025).

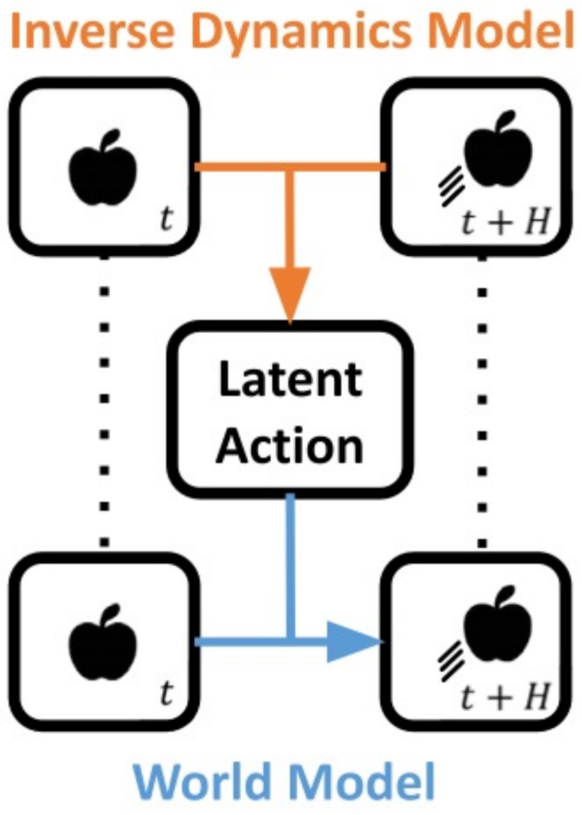
Architecture of a VLA System for Robotic Manipulation



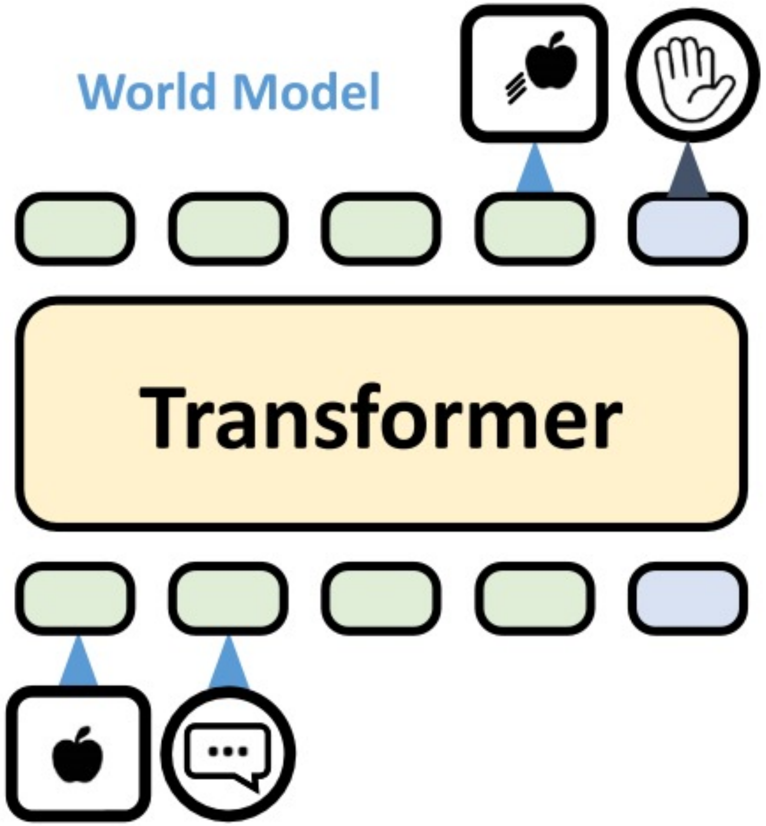
Design Patterns for Incorporating World Models in VLA



(1) Action generation in world models



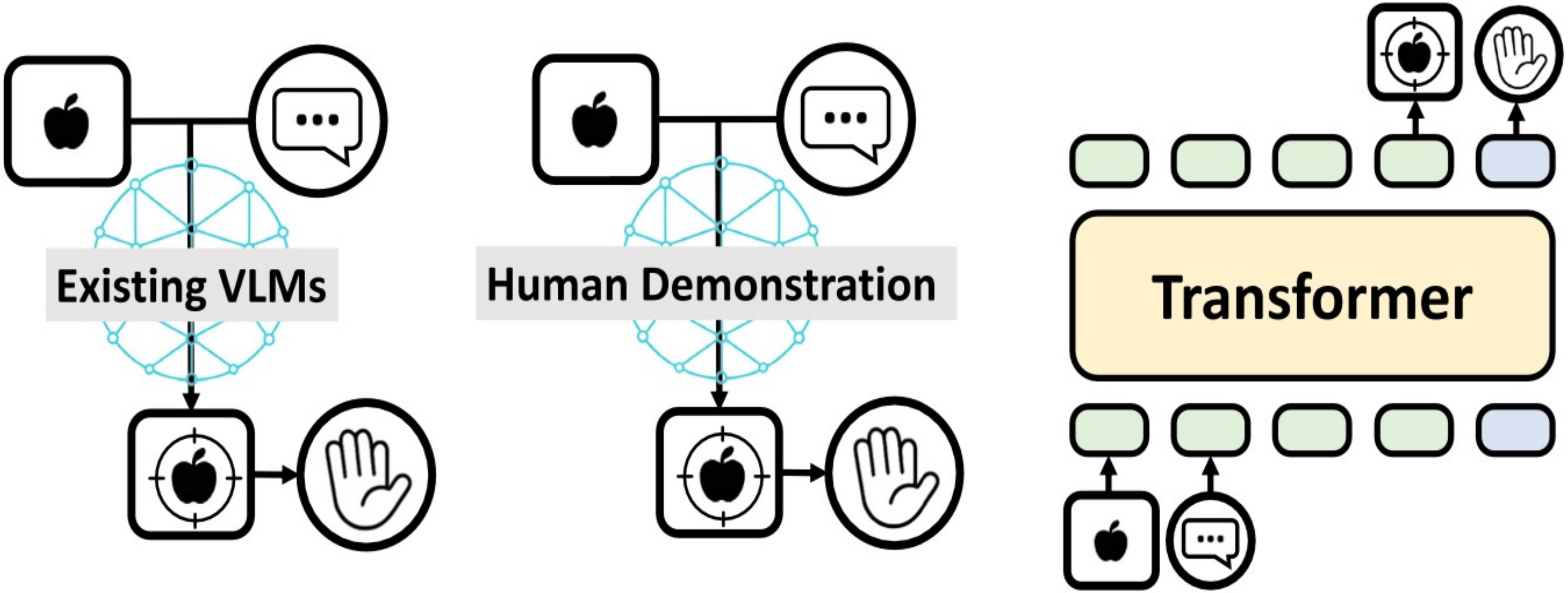
(2) Latent action generation via world models



(3) Sensorimotor models with implicit world models

Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

Design Patterns for Incorporating Affordance-based Models in VLA



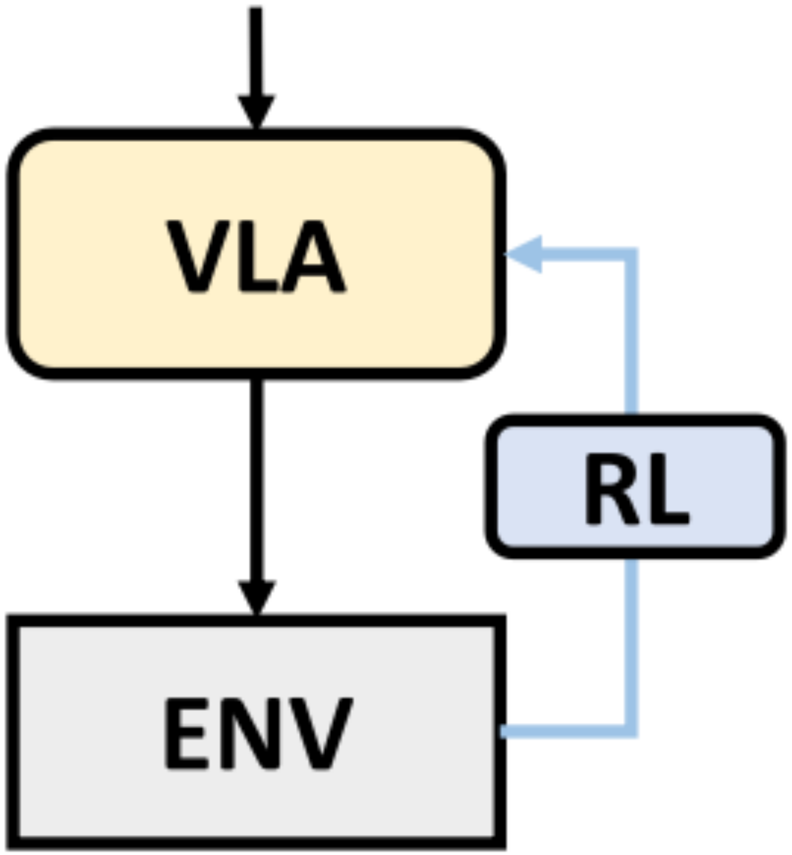
(1) Affordance prediction and action generation using VLMs

(2) Affordance extraction from human datasets

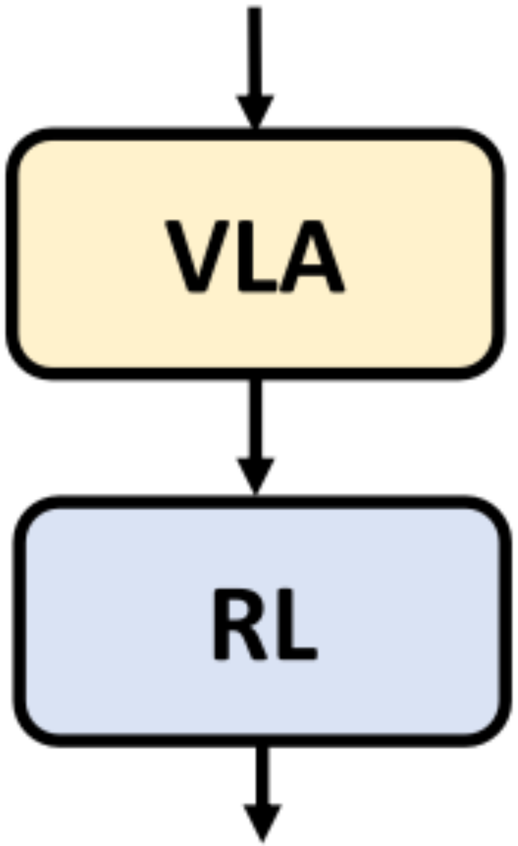
(3) Integration of sensorimotor models and affordance-based models

Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

Integrating RL with VLA Models



(1) Improving VLA using RL



(2) Using VLAs as high-level policies and RL for low-level control

Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

Robots Used in VLA Research

Robot

Manipulator

Hand/Gripper

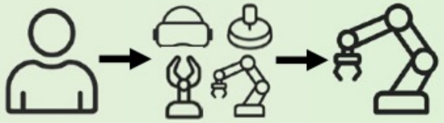
Mobile Robot

Quadruped Robot

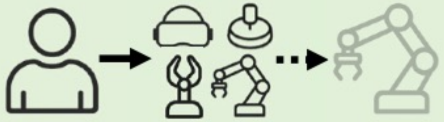
Humanoid Robot

Data Collection

Teleoperation



Proxy Devices



Human Data Collection

Dataset

Human Video Datasets

Ego4D *Ego-Exo4D*

HOI4D *ARCTIC*

Simulation Datasets

RoboTurk *MimicGen*

Real Robot Datasets

QT-Opt *RT-X*

BC-Z *DROID*

...

Augmentation

Vision

Language

Action

Evaluation

CALVIN, *AI2-THOR*, *Meta-World*

Habitat, *Habitat 2.0*, *Habitat 3.0*

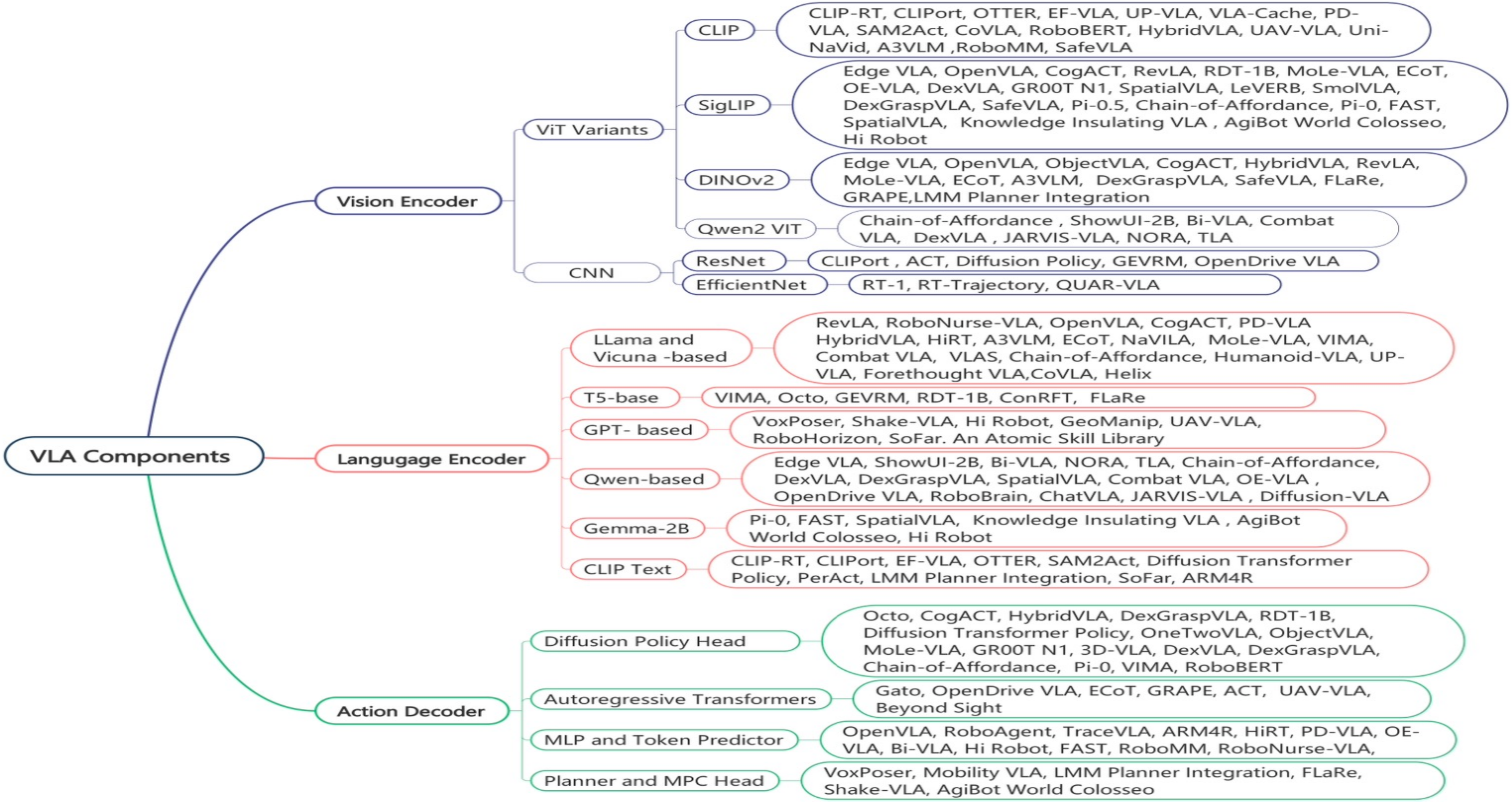
robosuite, *robomimic*, *RoboCasa*, *LIBERO*

ManiSkill, *ManiSkill 2*, *ManiSkill 3*, *ManiSkill-HAB*

RLBench, *COLOSSEUM*, *SIMPLER*, *RoboArena*

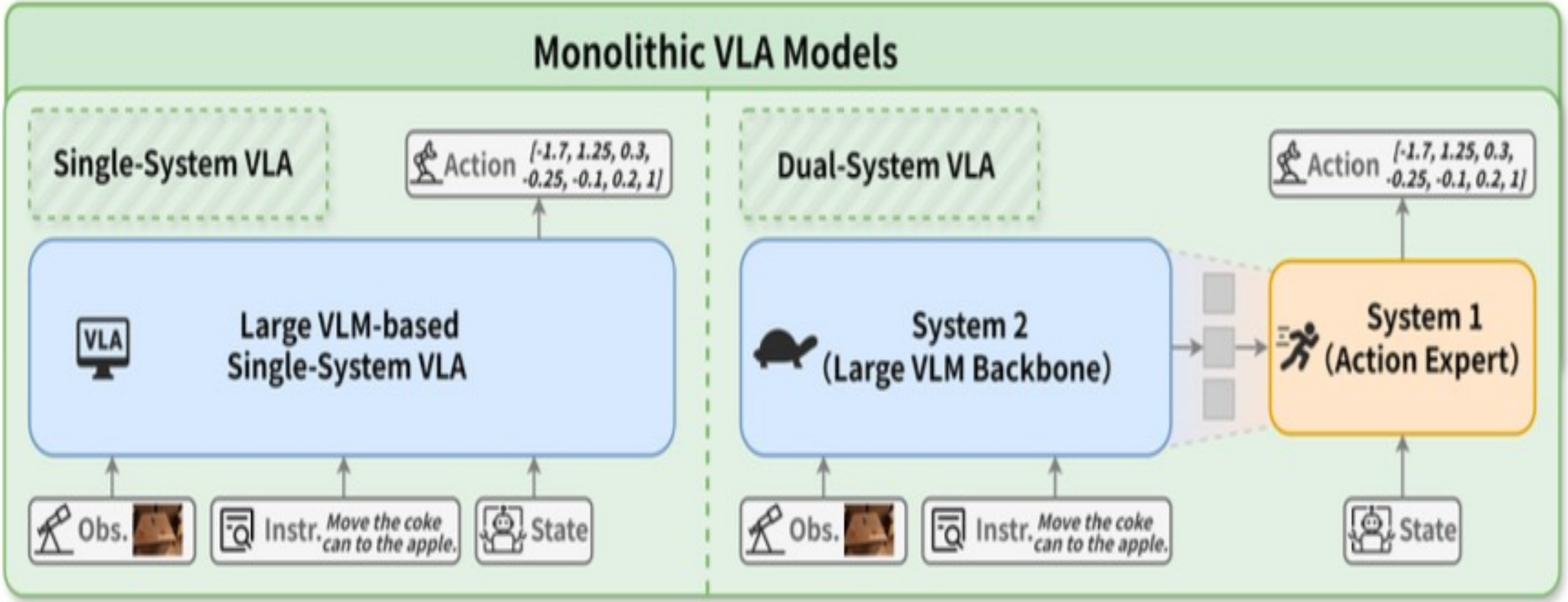
Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

Vision Language Action (VLA) Components



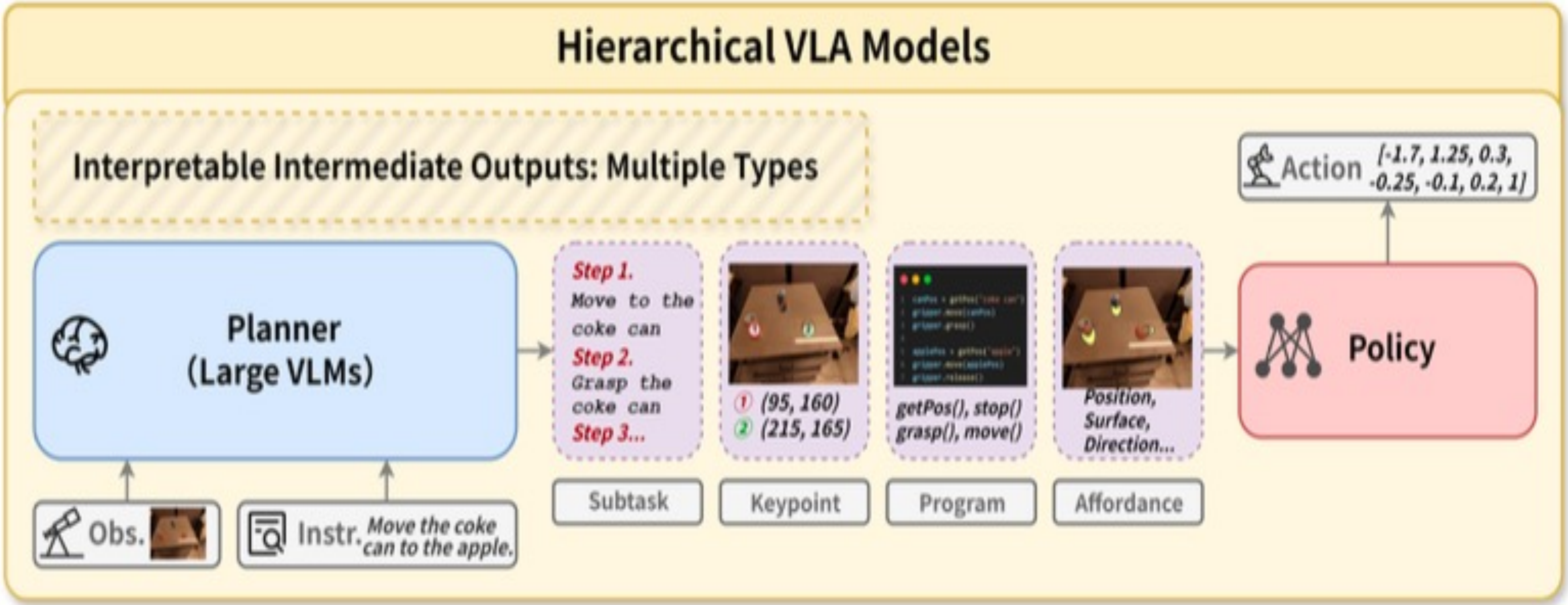
Source: Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025).

Large VLM-based Vision-Language-Action Models
Monolithic VLA Models



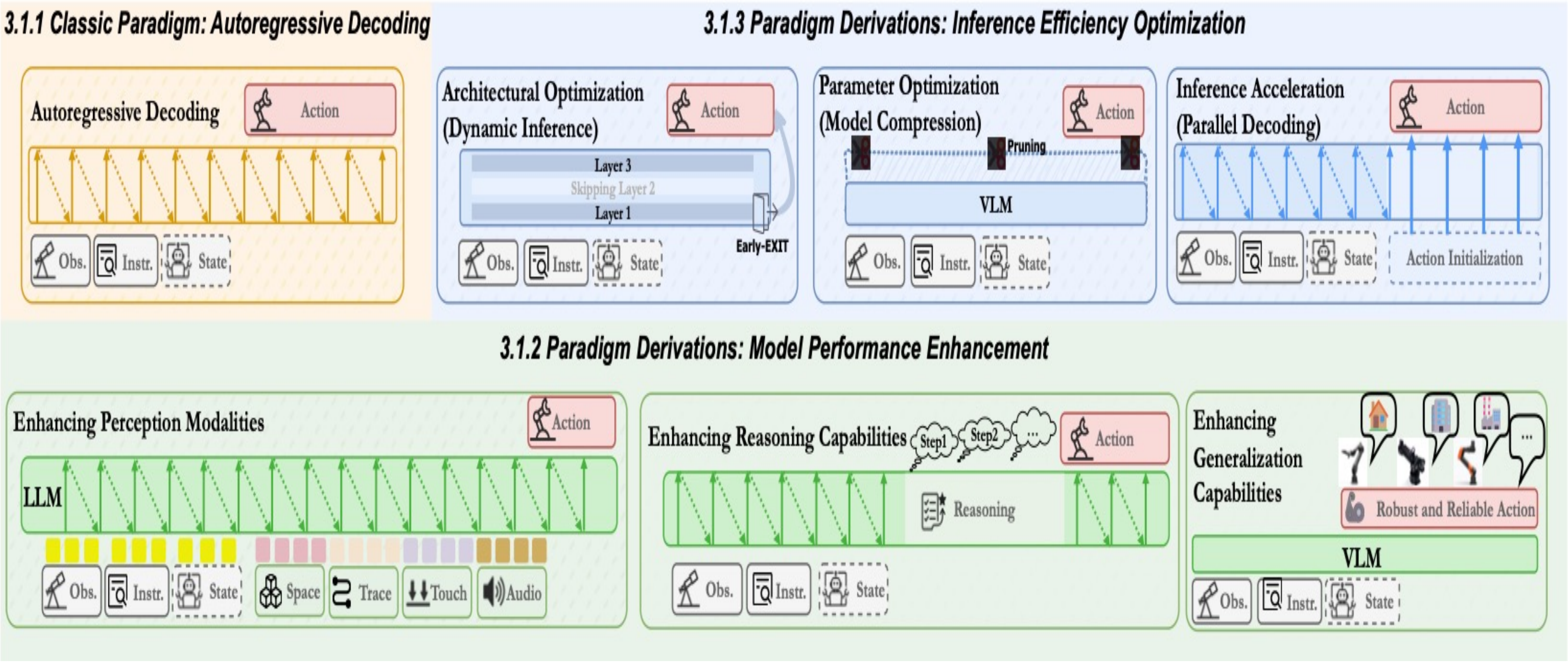
Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

Large VLM-based Vision-Language-Action Models
Hierarchical VLA Models



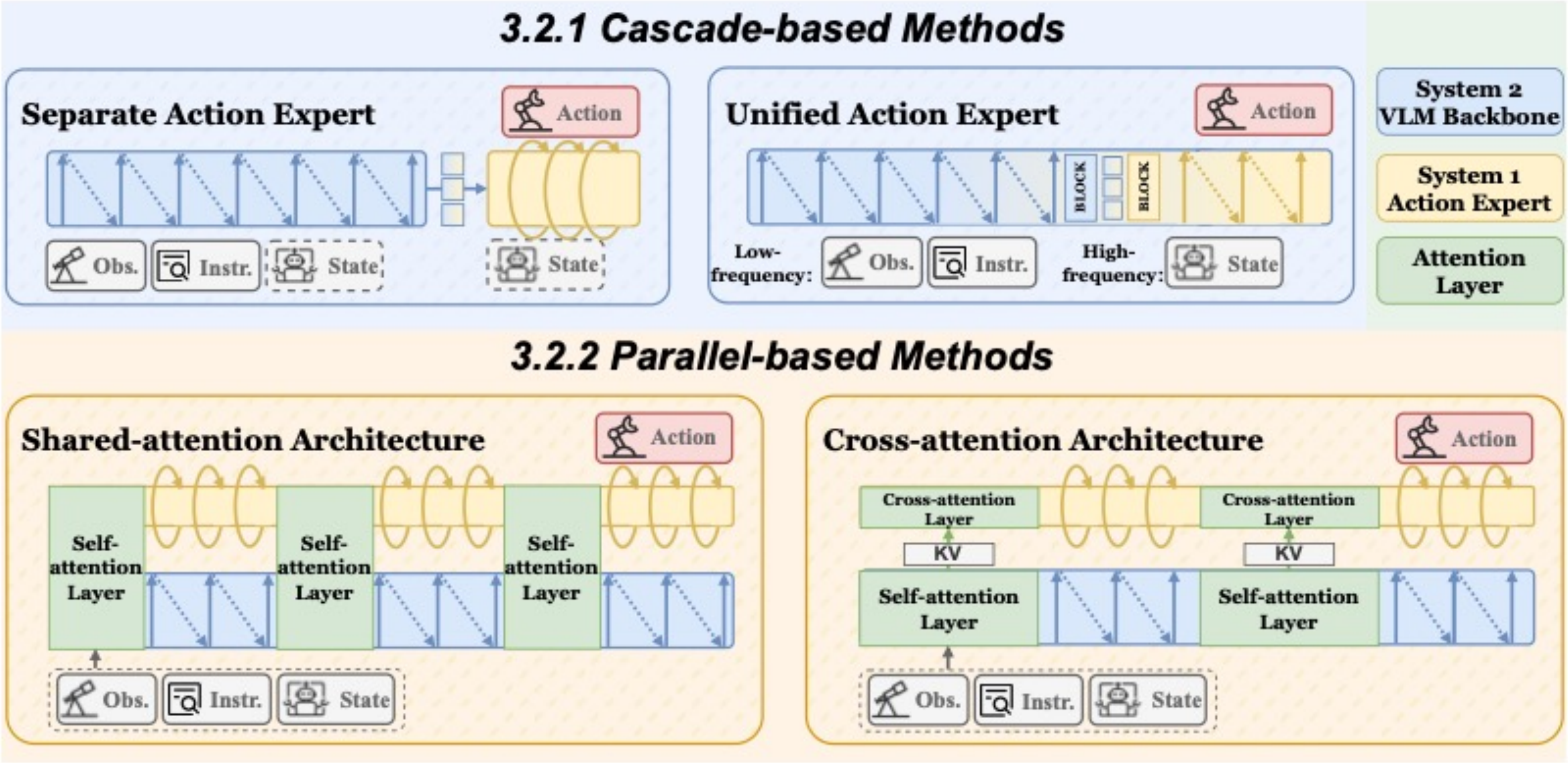
Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

Large VLM-based Vision-Language-Action Models
Paradigms in Monolithic Single-system Models



Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

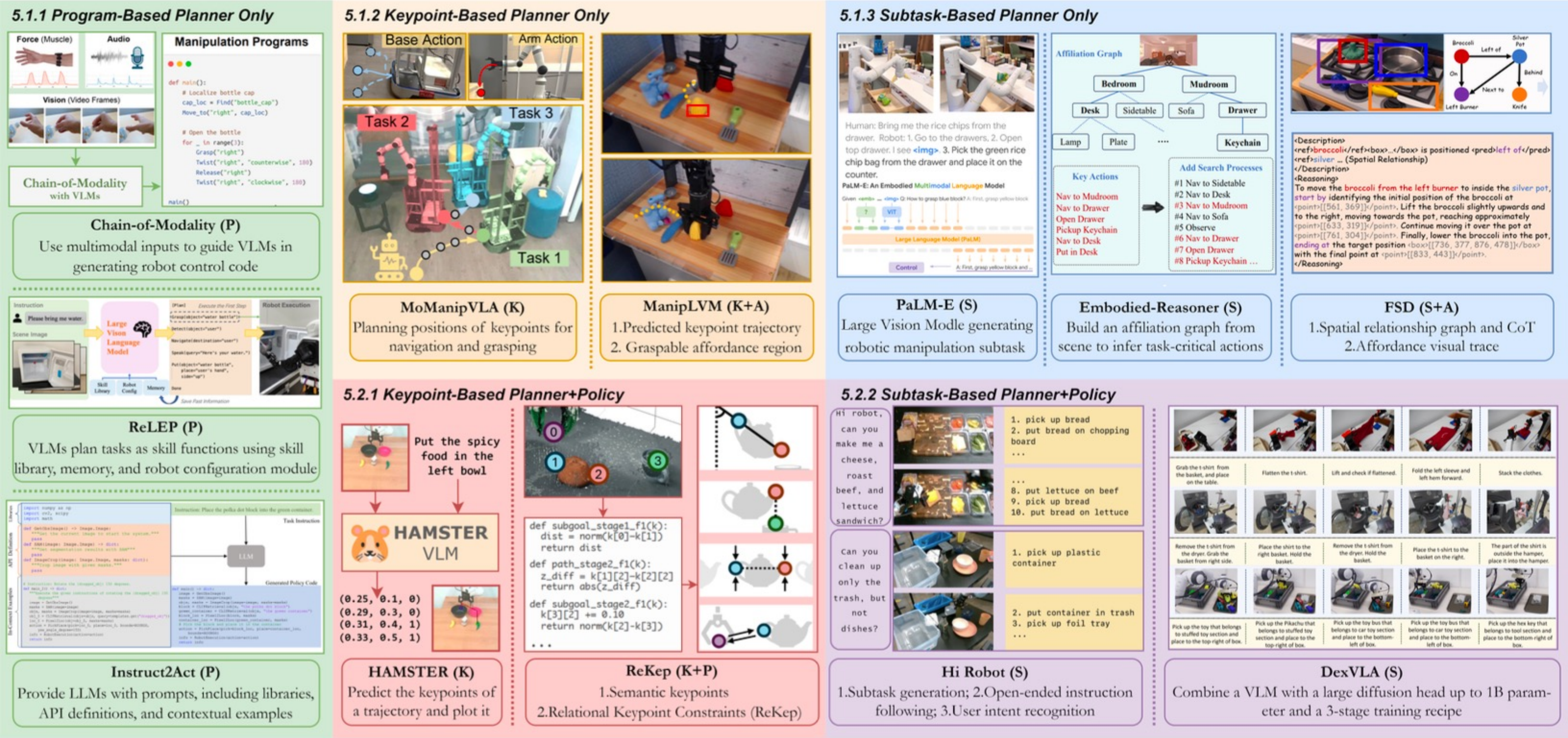
Large VLM-based Vision-Language-Action Models Paradigms in Monolithic Dual-system Models



Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

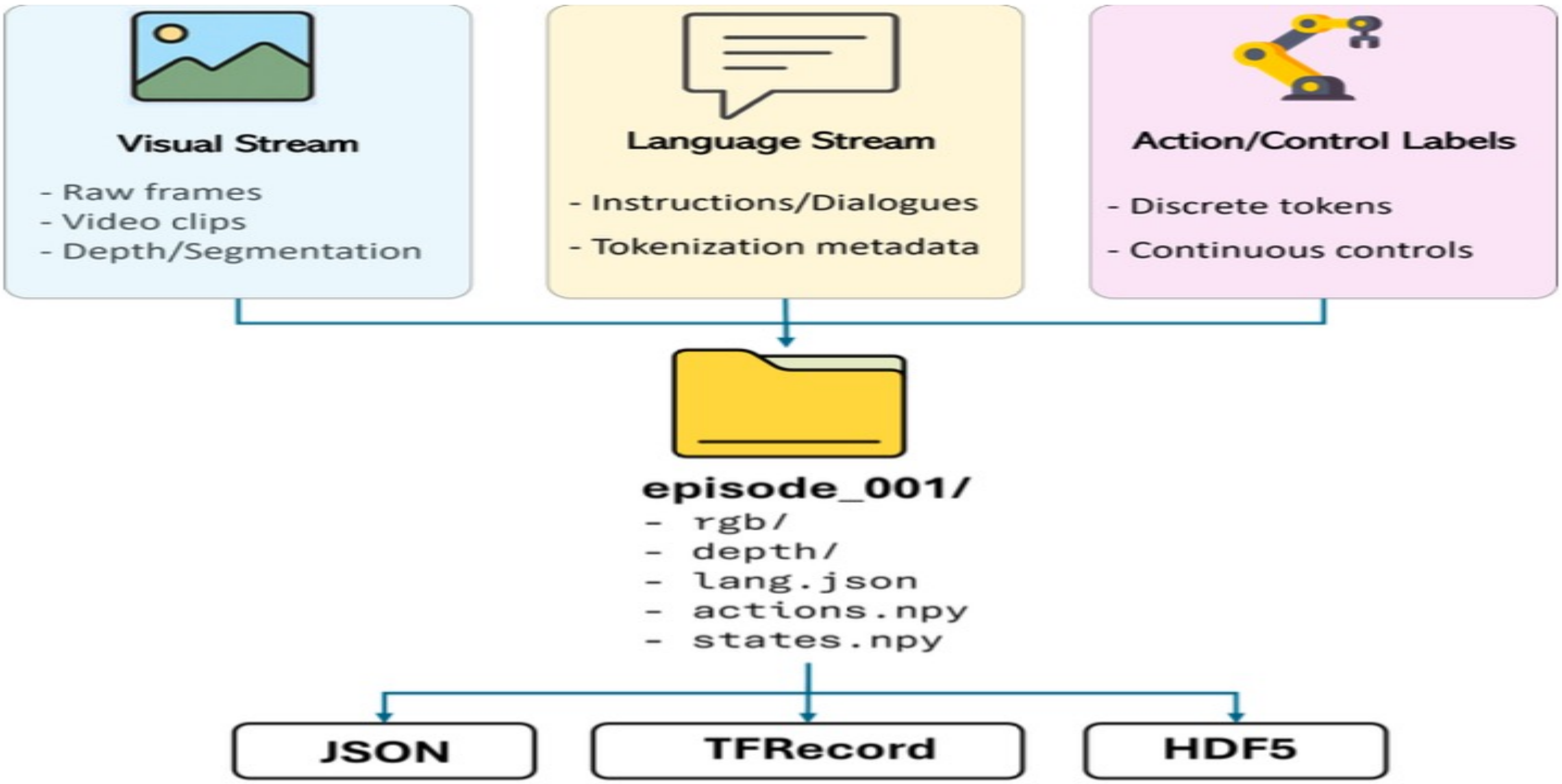
Large VLM-based Vision-Language-Action Models

Hierarchical Models



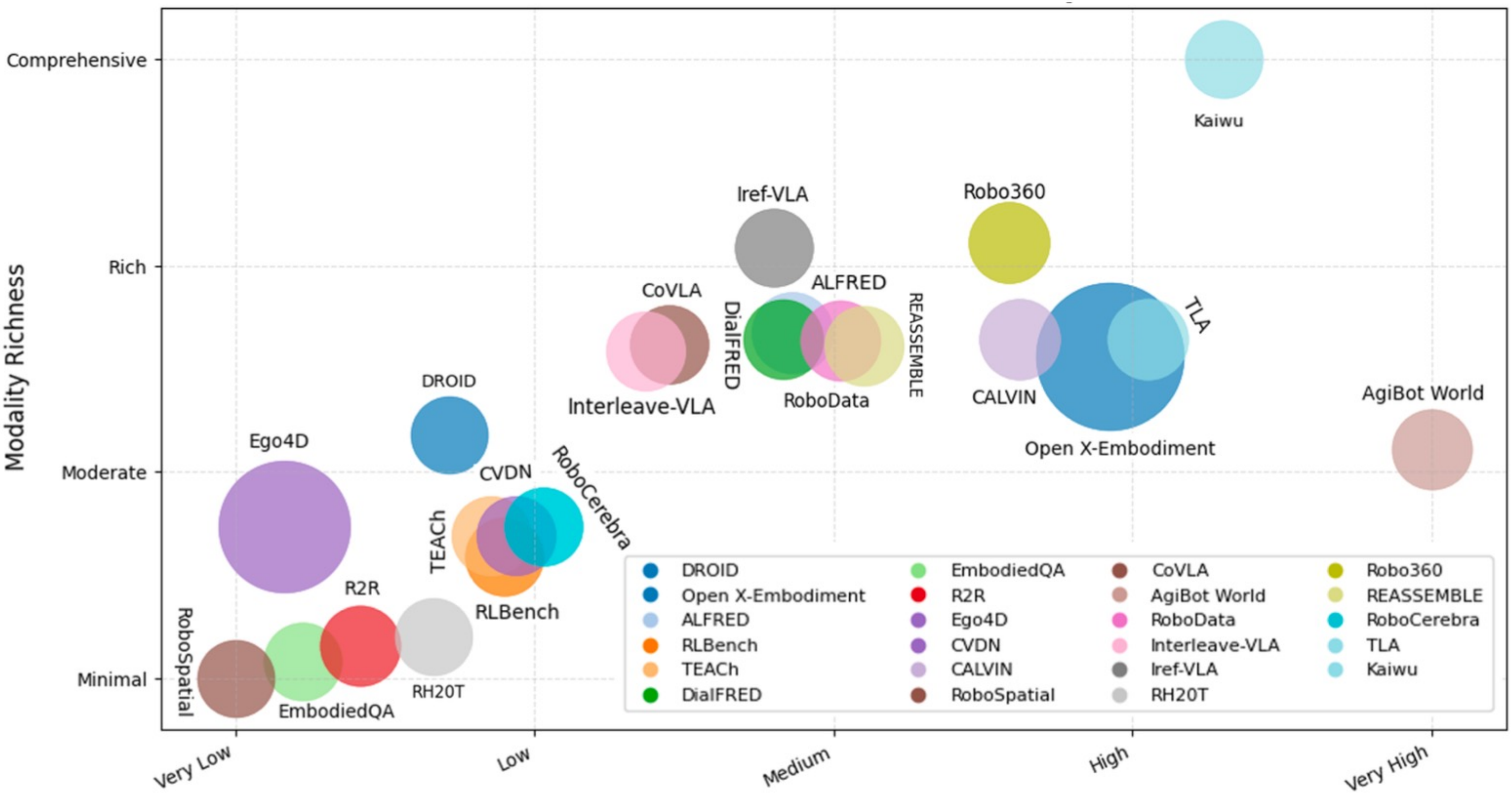
Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

Schematic of the Unified VLA Training Data Format



Source: Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025).

Benchmarking VLA Datasets
by Task Complexity and Modality Richness



Source: Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025).

Real-World Robot Datasets for VLA Research

Name	Episodes	Skill	Task	Modality	Embodiment	Collection
QT-Opt	580K	1 (Pick)	NA	RGB	KUKA LBR iiwa	Learned
MT-Opt	800K	2	12	RGB, L	7 robots	Scripted, Learned
RoboNet	162K	NA	NA	RGB	7 robots	Scripted
BridgeData	7.2K	4	71	RGB, L	WidowX 250	Teleop
BridgeData V2	60.1K	13	NA	RGB-D, L	WidowX 250	Teleop
BC-Z	26.0K	3	100	RGB, L	Google EDR	Teleop
Language Table	413K	1 (Push)	NA	RGB, L	xArm	Teleop
RH20T	110K	42	147	RGB-D, L, F, A	4 robots	Teleop
RT-1	130K	12	700+	RGB, L	Google EDR	Teleop
OXE	1.4M	527	160,266	RGB-D, L	22 robots	Mixed
DROID	76K	86	NA	RGB-D, L	Franka	Teleop
FuSe	27K	2	3	RGB, L, T, A	WidowX 250	Teleop
RoboMIND	107K	38	479	RGB-D, L	4 robots	Teleop
AgiBot World	94K	87	217	RGB-D, L	AgiBot G1	Teleop

Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

Benchmarks for Vision-Language-Action Evaluation

Simulation Environments: Navigation (Nav), Manipulation (Manip), and Whole-Body Control (WBC)

Name	Task	Scenes / Objects	Observation	Physics	Built Upon	Description
robosuite	Manip	NA / 10	RGB-D, S	MuJoCo	NA	Modular framework, 11 tasks
robomimic	Manip	NA / NA	RGB	MuJoCo	robosuite	Offline learning, 8 tasks
RoboCasa	Manip	120 / 2.5K	RGB	MuJoCo	robosuite	100 kitchen tasks, photorealistic
LIBERO	Manip	NA / NA	RGB	MuJoCo	robosuite	130 tasks in 4 task suites
Meta-World	Manip	1 / 80	Pose	MuJoCo	NA	50 Manip tasks for Meta-RL
LeVERB-Bench	Nav, WBC	4 / NA	RGB	PhysX	Isaac Sim	Humanoid control

Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

Benchmarks for Vision-Language-Action Evaluation

Simulation Environments: Navigation (Nav), Manipulation (Manip), and Whole-Body Control (WBC)

Name	Task	Scenes / Objects	Observation	Physics	Built Upon	Description
ManiSkill	Manip	NA / 162	RGB-D, PC, S	PhysX	SAPIEN	4 tasks, 36K demos
ManiSkill 2	Manip	NA / 2.1K	RGB-D, PC	PhysX	ManiSkill	Extended task diversity
ManiSkill 3	Nav, Manip, WBC	NA / NA	RGB-D, PC, S	PhysX	ManiSkill 2	GPU-parallelized simulation
ManiSkill-HAB	Manip	105 / 92	RGB-D	PhysX	ManiSkill 3, Habitat 2.0	HAB tasks from Habitat 2.0
RoboTwin	Manip	NA / 731	RGB-D	PhysX	SAPIEN	Dual-arm tasks
Ravens	Manip	NA / NA	RGB-D	PyBullet	NA	10 tabletop tasks
VIMA-BENCH	Manip	NA / 29	RGB, S	PyBullet	Ravens	17 multimodal prompt tasks
LoHoRavens	Manip	1 / 3	RGB-D	PyBullet	Ravens	Long-horizon planning
CALVIN	Manip	4 / 7	RGB-D	PyBullet	NA	Long-horizon lang-cond tasks

Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

Benchmarks for Vision-Language-Action Evaluation

Simulation Environments: Navigation (Nav), Manipulation (Manip), and Whole-Body Control (WBC)

Name	Task	Scenes / Objects	Observation	Physics	Built Upon	Description
Habitat	Nav	185 / NA	RGB-D, S	Bullet	NA	Fast, Nav only
Habitat 2.0	Nav, Manip	105 / 92	RGB-D	Bullet	Habitat	Mobile manipulation (HAB)
Habitat 3.0	Nav, Manip	211 / 18K	RGB-D	Bullet	Habitat 2.0	Human avatars support
RLBench	Manip	1 / 28	RGB-D, S	PyBullet	V-REP	Tiered task difficulty
THE COLOSSEUM	Manip	1 / 107	RGB-D	PyBullet	RLBench	20 tasks, 14 env variations
AI2-THOR	Nav, Manip	NA / 118	RGB-D, S	Unity	NA	Object states, task planning
CHORES	Nav	191K / 40K	RGB	Unity	AI2-THOR	Shortest-path planning
SIMPLER	Manip	4 / 17	RGB	PhysX	SAPIEN, Isaac Sim	Real-to-sim evaluation
RoboArena	Manip	NA / NA	RGB	Real	NA	Distributed real-world evaluation

Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

Summary

- **Generative AI**
- **Agentic AI**
- **Physical AI (Robotics)**

References

- Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson.
- Denis Rothman (2024), Transformers for Natural Language Processing and Computer Vision - Third Edition: Explore Generative AI and Large Language Models with Hugging Face, ChatGPT, GPT-4V, and DALL-E 3, 3rd ed. Edition, Packt Publishing
- Aurélien Géron (2022), Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 3rd Edition, O'Reilly Media.
- Steven D'Ascoli (2022), Artificial Intelligence and Deep Learning with Python: Every Line of Code Explained For Readers New to AI and New to Python, Independently published.
- Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. (2022) "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." arXiv preprint arXiv:2207.02696.
- Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. (2025) "Yolov9: Learning what you want to learn using programmable gradient information." In European Conference on Computer Vision, pp. 1-21. Springer, Cham.
- Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, and Abdeltawab Hendawi. (2024) "Evaluating the Evolution of YOLO (You Only Look Once) Models: A Comprehensive Benchmark Study of YOLO11 and Its Predecessors." arXiv preprint arXiv:2411.00201.
- Ranjan Sapkota, and Manoj Karkee. (2024) "Yolo11 and vision transformers based 3d pose estimation of immature green fruits in commercial apple orchards for robotic thinning." arXiv preprint arXiv:2410.19846.
- Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. (2021) "Learning transferable visual models from natural language supervision." In International Conference on Machine Learning, pp. 8748-8763. PMLR.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. (2021) "Vilt: Vision-and-language transformer without convolution or region supervision." In International Conference on Machine Learning, pp. 5583-5594. PMLR.
- Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. (2022) "Attention mechanisms in computer vision: A survey." Computational Visual Media ,:1-38.
- Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann.(2020) "Blazepose: On-device real-time body pose tracking." arXiv preprint arXiv:2006.10204.
- Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna et al.(2025) "Gemini robotics: Bringing ai into the physical world." arXiv preprint arXiv:2503.20020 (2025).
- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay et al. (2025) "Cosmos world foundation model platform for physical ai." arXiv preprint arXiv:2501.03575 (2025).
- Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu et al. (2025) "Foundation models in robotics: Applications, challenges, and the future." The International Journal of Robotics Research 44, no. 5 (2025): 701-739.
- Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).
- Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025)
- Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025)..
- Min-Yuh Day (2025), Python 101, https://tinyurl.com/aintpupython101