



Creating  
knowledge  
together

# Vision-Language-Action Models: Foundations and State of the Art

IKT526 Fall 2025

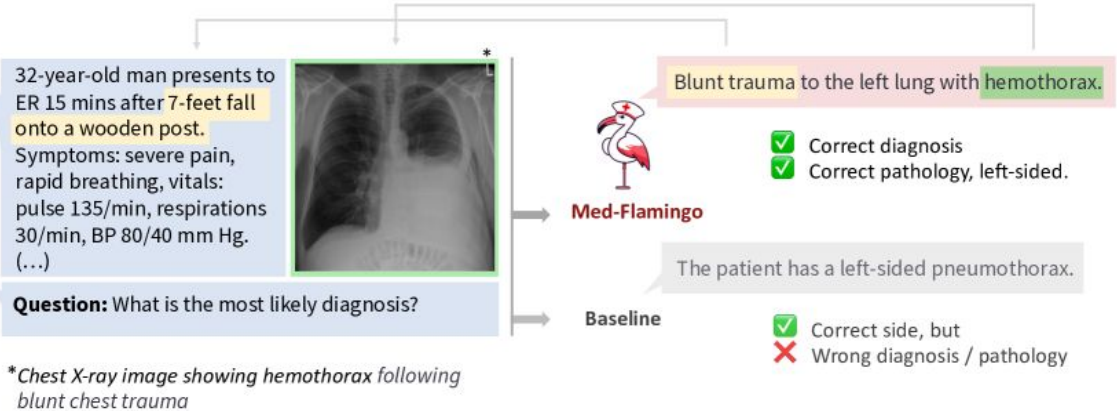
Philip André Haugen

# Introduction

# Background

- Early AI systems were unimodal, only designed for a single task.
- VLMs advanced multimodal learning by combining perception and language.
- CLIP, Flamingo, GPT-4V demonstrate effective integration of visual and linguistic information.
- These models can understand and describe, but cannot act in an environment.

# Background



Comparison of baseline vs. Med-Flamingo in clinical diagnosis from text and imaging [1].

## Sec. 4.1 Image Description on Diverse Domains

### Landmark Recognition and Description

**Prompt:**  
Describe the landmark in the image.



**GPT-4V:**  
The landmark in the image is the Space Needle, an observation tower located in Seattle, Washington. It stands 605 feet tall and offers panoramic views of the city and surrounding areas. The tower was built for the 1962 World's Fair and has since become a symbol of Seattle and a popular tourist attraction. The tower has a futuristic design with a large circular observation deck at the top and a restaurant that rotates 360 degrees. The base of the tower is surrounded by a park with trees and green spaces.

**Prompt:**  
Describe the landmark in the image.



**GPT-4V:**  
This image shows the Kinkaku-ji Temple, also known as the Golden Pavilion, located in Kyoto, Japan. The temple is a three-story Buddhist temple that is covered in gold leaf and sits on a pond, creating a stunning reflection on the water. The surrounding landscape is filled with trees and rocks, creating a peaceful and serene atmosphere. It is a popular tourist attraction and a designated National Special Historic Site.

**Prompt:**  
Describe the landmark in the image.



**GPT-4V:**  
The landmark in the image is the statue of George M. Cohan in Times Square, New York City. It is a bronze statue that stands on a granite pedestal in Duffy Square, which is the northern part of Times Square. The statue is a tribute to George M. Cohan, an American entertainer, playwright, composer, and producer, who is considered the father of American musical comedy. The statue was unveiled in 1959 and has since become a popular tourist attraction in Times Square.

**Prompt:**  
Describe the landmark in the image.



**GPT-4V:**  
The landmark in the image is the famous Lombard Street in San Francisco, California, known for its steep hairpin turns. The street is lined with colorful flowers and shrubbery, and there are houses and buildings on both sides. In the background, you can see the city skyline and the Coit Tower on Telegraph Hill. This street is a popular tourist attraction, and it is often considered one of the most crooked streets in the world.

Examples of GPT-4V producing multimodal landmark recognition and descriptive outputs [2].

# Vision-Language-Action Models

- Extend multimodal models with an action space in addition to vision and language.
- Inputs: visual observations and textual instructions
- Outputs: action representations
- Training: pretraining on large-scale vision-language data, followed by fine-tuning on action trajectories.

# Motivation

- Many real-world tasks follow a see-think-act loop: perception, reasoning, and action.
- Natural language is an intuitive way for humans to specify tasks.
- Robots must ground commands in high-dimensional sensory input and map them to low-level motor control.
- Reinforcement learning can optimize actions, but is challenged by sparse rewards, sample inefficiency, and safety risks.
- Need for models that can generalize across tasks, environments, and modalities with limited real-world training.

# Key Ideas

- Unify perception, language understanding, and control policies within a single model.
- Enable generalization: from web-scale data to unseen robotic tasks.
- Represent a step toward embodied intelligence.
- Move beyond multimodal understanding to multimodal decision-making and execution.

# Vision-Language-Action

# Core Architecture

Encoders:

- Vision encoder, language encoder, optional modalities.
- Outputs modality-specific embeddings in latent space.

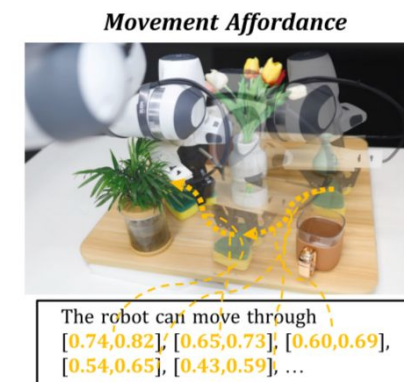
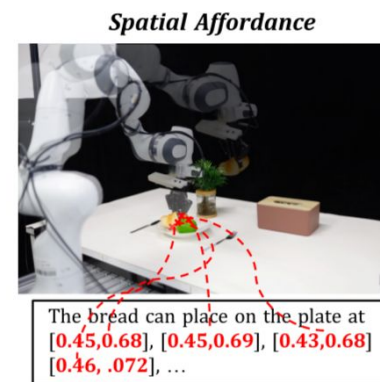
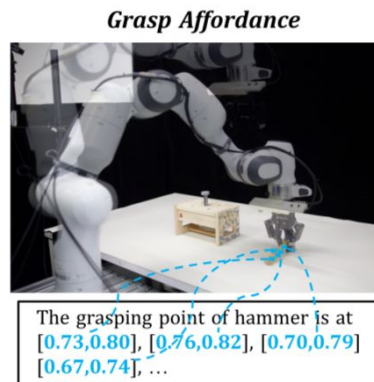
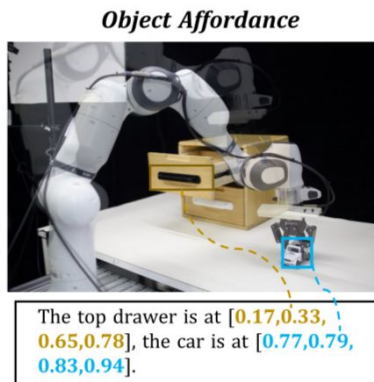
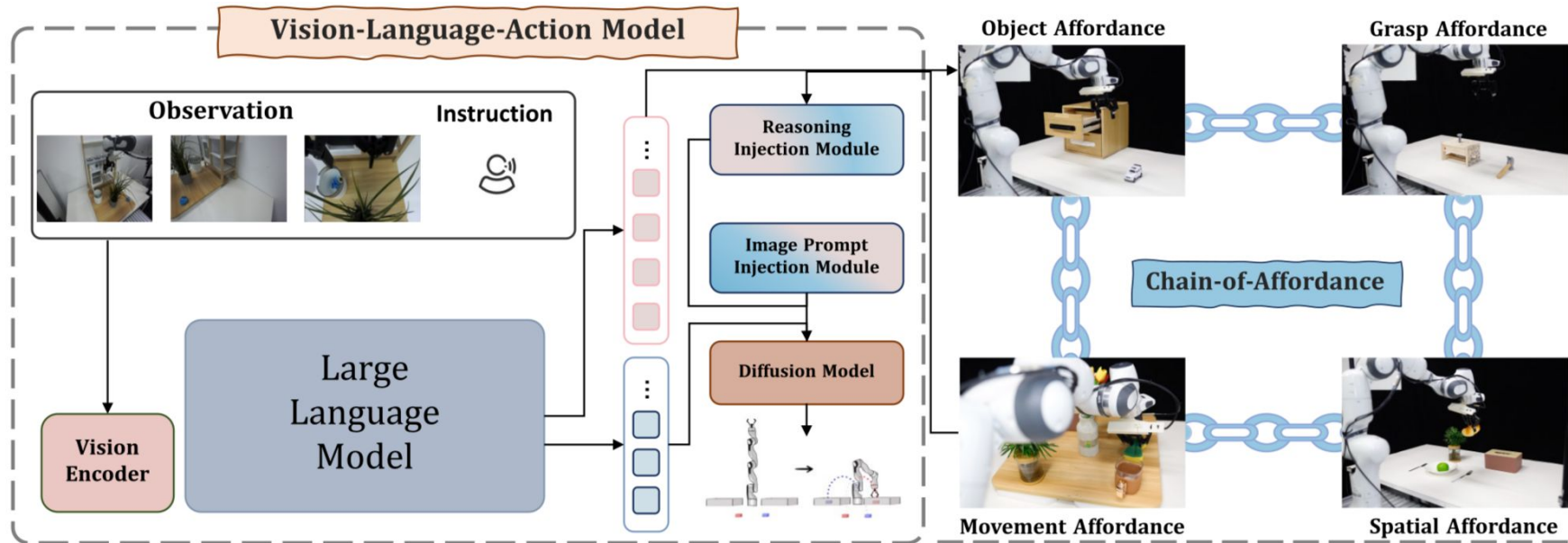
Fusion:

- Attention mechanism, concatenation & projection, token-level alignment.

Action:

- Fused representation into action space.
- Architectures such as discrete action tokens, continuous control outputs, hierarchical planners.
- The link between symbolic task specification and robot control signals.

# Core Architecture



[3]

# Training and Inference

## Training

Multimodal Data sets (Image, Text, Action)

Encoders + Fusion Module + Policy Head

Loss Functions (contrastive, action prediction)

Backpropagation & Optimization

## Inference

Live Inputs (Camera Feed + Task Prompt)

Encoders + Fusion Module + Policy Head

Forward Pass Only (no weight updates)

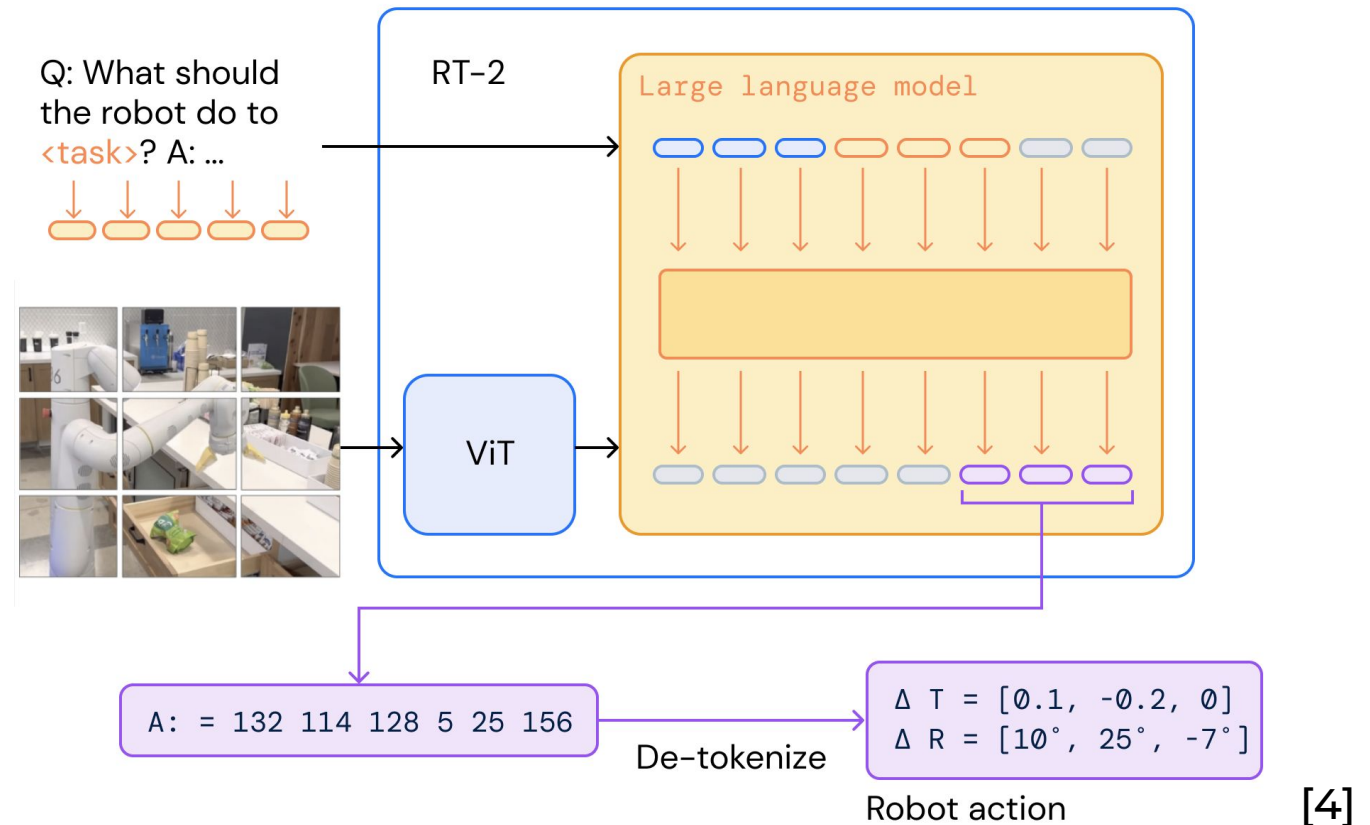
Predicted Action Sequence (Robot Control)

# Examples and SOTA

# Examples

RT-2 by DeepMind [4]:

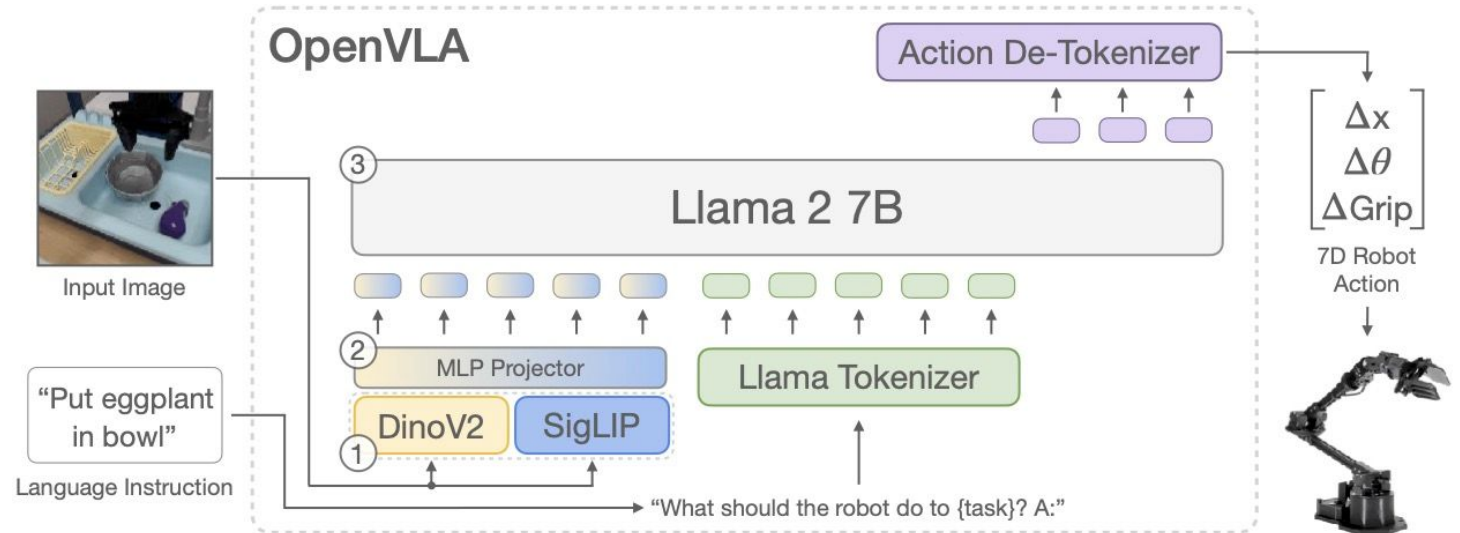
- Major breakthrough.
- Actions are tokenized like words.
- Generates code-like instructions.



# Examples

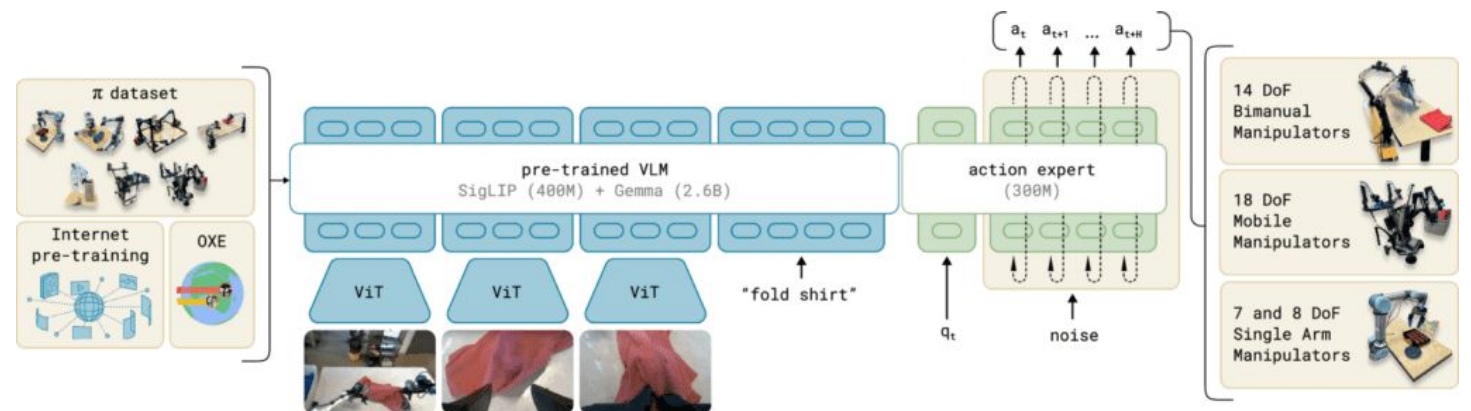
OpenVLA [5]:

- Open source community reimplementation of the RT-2 architecture.
- Democratized research.



$\pi$ 0 Physical Intelligence [6]:

- Generalist robot policy.
- Flow matching.
- Continuous motor commands.



# State of the Art

<b>DeeR-VLA (2024)</b> [9]	Open-source 7B VLA with Llama 2 + dual vision encoders (DINOv2, SigLIP), modular vision-language fusion.	General manipulation across 29 tasks, +16.5% success vs. 55B RT-2-X model, easy LoRA fine-tuning and quantization.
<b>OTTER (2025)</b> [10]	Text-guided visual feature selection; uses a frozen VLM and extracts only instruction-relevant image tokens.	Instruction following with strong zero-shot generalization, outperforms fine-tuned models on novel objects & scenes.
<b>CoA-VLA (2025)</b> [11]	Chain-of-affordance reasoning: prompts model with sequential affordances (object, grasp, place, move) in visual and textual form.	Multi-step manipulation tasks, surpasses OpenVLA and Octo on success rates, better handling of unseen object poses and obstacles.
<b>RoboMamba (2024)</b> [12]	Integrates a linear-time state-space language model (Mamba) with a vision encoder, aligns visual tokens via co-training.	Broad reasoning and manipulation, 3x faster inference than transformer VLAMs, needs only 0.1% parameters fine-tuned to learn new SE(3) tasks.
<b>ShowUI (2025)</b> [13]	VLA for GUI automation; graph-based UI token pruning to reduce redundant screen pixels, plus interleaved vision-language-action sequence training.	Digital (web/mobile) interface tasks, 75% zero-shot accuracy on UI element grounding, 1.4x faster training via token pruning.

# Applications

# Applications



[7]

# Applications



General



Autonomous 1x

[8]

# Challenges

# Challenges

Generalization vs specialization:

- Broad generalization without sacrificing performance.

Long-horizon reasoning and memory:

- Memory architectures such as recurrence and external memory.

Data and simulation:

- Data distribution is a limitation.

Safety and ethical considerations:

- Avoid harmful actions. MDP formalisms proposed.

# Future Directions

# Future Directions

Generalization without forgetting:

- Balance fine-tuning for tasks with retention of broad multimodal knowledge.

Structured reasoning for action:

- Hybrid design with end-to-end learning.

Efficient real-time inference:

- Optimize for embedded hardware in edge devices.

Safety, alignment, and evaluation:

- Develop standardized VLAM benchmarks.

# Thank you! Questions?

# References

- [1] M. Moor et al., "Med-Flamingo: a Multimodal Medical Few-shot Learner," arXiv:2307.15189, Jul. 2023.
- [2] Z. Yang et al., "The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)," arXiv:2309.17421, Oct. 2023.
- [3] J. Li et al., "CoA-VLA: Improving Vision-Language-Action Models via Visual-Textual Chain-of-Affordance," arXiv:2412.20451, Jul. 2025.
- [4] A. Brohan et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," arXiv:2307.15818, Jul. 2023.
- [5] M. J. Kim et al., "OpenVLA: An Open-Source Vision-Language-Action Model," arXiv:2406.09246, Jun. 2024.
- [6] K. Black et al., " $\pi_0$ : A Vision-Language-Action Flow Model for General Robot Control," arXiv:2410.24164, Nov. 2024.
- [7] A. James, "DeepRoute.ai debuts Vision-Language-Action model in Korea," ADAS & Autonomous Vehicle International, Apr. 2025.
- [8] Google DeepMind, "Gemini Robotics: Generality in action," YouTube, Mar. 2025.
- [9] Y. Yue et al., "DeeR-VLA: Dynamic Inference of Multimodal Large Language Models for Efficient Robot Execution," arXiv:2411.02359, Nov. 2024.
- [10] H. Huang et al., "OTTER: A Vision-Language-Action Model with Text-Aware Visual Feature Extraction," ICML, Jul. 2025.
- [11] J. Li et al., "CoA-VLA: Improving Vision-Language-Action Models via Visual-Textual Chain-of-Affordance," arXiv:2412.20451, Jul. 2025.
- [12] J. Liu et al., "RoboMamba: Efficient Vision-Language-Action Model for Robotic Reasoning and Manipulation," arXiv:2406.04339, Dec. 2024.
- [13] K. Q. Lin et al., "ShowUI: One Vision-Language-Action Model for GUI Visual Agent," arXiv, Nov. 2024.