Title: Vision-language-action model

URL: https://en.wikipedia.org/wiki/Vision-language-action_model

PageID: 79446353

Categories: Category:Deep learning, Category:Natural language processing, Category:Robotics

Source: Wikipedia (CC BY-SA 4.0).

In robot learning , a vision-language-action model (VLA) is a class of multimodal foundation models that integrates vision , language and actions. Given an input image (or video) of the robot's surroundings and a text instruction, a VLA directly outputs low-level robot actions that can be executed to accomplish the requested task. [1]

VLAs are generally constructed by fine-tuning a vision-language model (VLM, i.e. a large language model extended with vision capabilities) on a large-scale dataset that pairs visual observation and language instructions with robot trajectories. [2] These models combine a vision-language encoder (typically a VLM or a vision transformer), which translates an image observation and a natural language description into a distribution within a latent space , with an action decoder that transforms this representation into continuous output actions, directly executable on the robot. [3]

The concept was pioneered in July 2023 by Google DeepMind with RT-2, a VLM adapted for end-to-end manipulation tasks, capable of unifying perception , reasoning and control . [4]

Overview of architecture

VLAs share a common high-level architecture articulated in two stages:

In the first stage, a pre-trained VLM serves as the perception and reasoning core. It encodes one or more camera images together with a language instruction into a sequence of language tokens in a shared latent space. VLMs are specifically trained on large multimodal datasets and can perform a variety of tasks such as image understanding , visual-question answering and reasoning . In order to directly control robots, VLMs must be extended to output robot actions. [5]

In the second stage, an action decoder maps those tokens to discrete symbols that are then de-tokenised into continuous robot commands. These output actions are represented in the same way as language tokens, but specifically refer to the number of degrees of freedom (DoF) of the robot's end effector . Considering a 6-DoF end-effector, the action space usually includes end-effector displacements (positional and rotational) and gripper positions. For instance, in RT-2, each action vector covers 6-DoF in addition to the gripper state and a termination flag, all quantized into 256 bins. [2]

VLAs usually rely on off-the-shelf VLMs, giving the robot a prior understanding of images and text. During the training process, the model is then fine-tuned on data in the form of (text instruction, visual observation, action trajectory), and so it learns to map visual observations and text instructions to robot actions. The training dataset consists of robot demonstrations which may be gathered from real robots, human teleoperation, or even synthetically generated in a simulation environment. Due to end-to-end learning, VLAs inherently learn to associate high-level concepts (e.g. object categories and spatial relations) with low-level actions, eliminating the partitioning typical of traditional robotic systems. [2] [6]

Action representation

A crucial design choice for the architecture of a VLA is the format in which robot actions are encoded.

'Discrete Token Output' is the most common approach, used by VLAs such as RT-2 and OpenVLA, and it represents each motion primitive as a sequence of discrete tokens. In this way, the model encodes the robot actions as an action string, and the VLA model learns to generate these sequences just as a language model generates text. This token-based approach keeps the same output layer and makes training straightforward. However, converting continuous trajectories into vocabulary symbols can limit spatial accuracy or temporal resolution. RT-2 demonstrates that this can be mitigated using special tokens that, for instance, mark the end of an action segment. [2] [7]

'Continuous Output' (Diffusion/Flow) is an alternative approach used by VLAs such as π 0 that, in order to achieve accurate dexterity and high frequency control, forego discrete tokens and directly output continuous actions. This is achieved through the use of diffusion models or flow-matching networks that act as the action decoder. π 0 exploited this strategy to output continuous joint trajectories up to 50 Hz . Practically, continuous output tends to scale better to robots with many degrees of freedom, where discretization for every DoF would be impractical. [8]

Single-model versus dual-system design

VLAs can be organized either as a single end-to-end network or as a dual-system that employs two coupled models.

The single-model design, employed by RT-2, OpenVLA and π 0 , simultaneously understands the scene and the language instruction to produce robot actions in a single forward pass, keeping the architecture simple and reducing latency . [2] [7] [8]

The dual-system design, adopted by Helix and Groot N1, decouples the architecture into two components. The first component is usually slower and handles image observation and text instructions received as input. The second component runs at a faster rate and produces the robot's actions. The two components are trained end-to-end to communicate. This split improves dexterity and latency at the cost of increased computational complexity. [9] [10]

History

2023

Robotic Transformer 2 (RT-2)

Robotic Transformer 2 (RT-2) was developed by Google DeepMind in mid-2023 and established the vision-language-action model paradigm in robotics. It builds on two state-of-the-art VLMs, respectively PaLI-X [11] and PaLM-E, [12] by fine-tuning them on real robot demonstration data. RT-2 takes as input camera images paired with a text description and outputs discretized robot action encoded as discrete tokens. Compared to its predecessor RT-1, [13] which was trained only on robotic data, RT-2 exhibits stronger generalization for new tasks, being also able to perform multi-step reasoning using chain-of-thought . [4]

2024

OpenVLA

OpenVLA is a 7b-parameter open-source VLA model introduced in June 2024 by researchers at Stanford . It was trained on the Open X-Embodiment dataset, a collaboration between 21 institutions that collected over one million episodes on 22 different embodiments. The model fuses image features using DINOv2 [14] and CLIP , with a Llama-2 language backbone, and outputs discrete actions tokens. Despite its smaller size with respect to Google DeepMind's RT-2, OpenVLA outperforms RT-2 on a suite of manipulation tasks. It also supports parameter-efficient fine-tuning methods and quantization for resource-constrained deployment. [7] [15] [16]

Octo (Open Generalist Policy)

Octo is a lightweight open-source generalist robot policy from UC Berkeley . Originally trained on Open X-Embodiment, it was released in smaller configurations (27M and 93M parameters). Octo encodes text instructions and image observations respectively with a language model and a lightweight convolutional neural network . Additionally, instead of an autoregressive decoder, Octo uses a diffusion policy that outputs continuous joint trajectories, enabling smoother motion and fast task adaptation. During fine-tuning, the block-wise attention structure of the architecture employed by Octo, allows to add new observations without modifying the parameters. [17]

TinyVLA

TinyVLA is a compact VLA designed for fast inference and efficient training. TinyVLA addresses the computational requirements and the heavy reliance on large datasets of its predecessors by initializing the policy with a smaller multimodal backbone and then fine-tuning on robotics data. This work demonstrated potential for more efficient VLAs, focusing on architecture and data curation without the computational cost of very large models. [18]

π 0 (pi-zero)

π 0 (pi-zero) is a large-scale generalist VLA, announced in late 2024 by the startup Physical Intelligence. [8] [better source needed] π 0 incorporates Paligemma [19] as a pre-trained VLM backbone, built from SigLIP [20] and Gemma [21] encoders, with an action expert trained on robot trajectories from Open X-Embodiment. Trained on robot trajectories from 8 different embodiments, it is able to generalize cross-embodiment, control different robotic arms (single-arm, dual-arm) and tackle a wide variety of tasks. π 0 also introduced flow-matching model to generate high-frequency continuous actions, up to 50 Hz, while the action head takes advantage of a diffusion policy. [22] [23] π 0 -FAST, an extension of π 0 , takes advantage of Frequency-space Action Sequence Tokenization (FAST), [24] a novel time-series compression approach that transform continuous tokens from time domain to frequency domain using discrete cosine transform .

2025

Helix

Helix, unveiled in February 2025 by Figure AI , it is a generalist VLA specifically tailored for humanoid robots. It is the first VLA able to control at a high frequency the entire upper body of a humanoid (i.e. arms, hands, torso, head, fingers). It uses a dual-system architecture, with two complementary systems trained to communicate in an end-to-end manner. System 2 (S2) is an internet-scale VLM specialized in scene understanding and language comprehension, while System 1 (S1) is a visuomotor policy that translates the latent representations produced by S2 into continuous robot actions. This decoupled architecture allows to achieve both broad generalization and fast low-level control. Helix is trained on ~500 hours of robot teleoperation paired with automatically generated text descriptions. The Helix model underscored the ability of VLAs to scale to complex embodiments such as humanoids. [9]

GR00T N1

GR00T N1, released by NVIDIA in March 2025, is a VLA for humanoid robots that adopts the same dual-system architecture employed by Helix. It is composed of a System 2, a VLM responsible for the perception of the environment, and a System 1, which generates motor action. Different from other VLAs, it includes a heterogeneous mixture of data comprising robots' trajectories, human videos and synthetic datasets. [10]

Gemini Robotics

Gemini Robotics , introduced in 2025 by Google DeepMind , is a VLA that builds on top of the capabilities of Gemini 2.0. While Gemini is inherently able to process multimodal data such as text, images, videos and audio, Gemini Robotics extends these capabilities to the physical world, allowing robots to take actions. The reasoning capabilities of the Gemini 2.0 VLM backbone, paired with learned low-level robot actions, allow the robot to perform highly dexterous tasks such as folding origami, as well as playing with cards. The model exhibits a high degree of generalization and is able to adapt to entirely new platforms. In June 2025, the authors released Gemini Robotics On-Device, a lightweight version of the previous model, optimized to run locally on a real robot with low-latency and high reliability while preserving dexterity. [6] [25]

SmolVLA

SmolVLA is an open-source compact VLA with 450 million parameters released by Hugging Face . It represents an effort to democratize research on VLAs. It was trained entirely on LeRobot, an open-source dataset collected and curated by the community. Despite its compact size, SmolVLA achieved comparable performances with much larger VLAs such as Octo, OpenVLA and π 0 . The architecture of SmolVLA employs flow-matching for continuous control, and asynchronous inference to decouple the VLM backbone from the action execution. SmolVLA can be fine-tuned and used on a single consumer GPU. [26] [27] [28]

See also

Robot learning

Large language model

Foundation model

Natural language processing

References

Further reading

Brohan, Anthony; Brown, Noah; Carbajal, Justice; Chebotar, Yevgen; Chen, Xi; Choromanski, Krzysztof; Ding, Tianli; Driess, Danny; Dubey, Avinava (July 28, 2023), RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control , arXiv : 2307.15818

Black, Kevin; Brown, Noah; Driess, Danny; Esmail, Adnan; Equi, Michael; Finn, Chelsea; Fusai, Niccolo; Groom, Lachy; Hausman, Karol (2024), $π_0$: A Vision-Language-Action Flow Model for General Robot Control , arXiv : 2410.24164

Ma, Yueen; Song, Zixing; Zhuang, Yuzheng; Hao, Jianye; King, Irwin (March 4, 2025), A Survey on Vision-Language-Action Models for Embodied AI , arXiv : 2405.14093

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model

Model Context Protocol

Neural network

Prompt engineering

Reinforcement learning from human feedback

Retrieval-augmented generation

Self-supervised learning

Stochastic parrot

Synthetic data

Top-p sampling

Transformer

Variational autoencoder

Vibe coding

Vision transformer

Waluigi effect

Word embedding

Character.ai

ChatGPT

DeepSeek

Ernie

Gemini

Grok

Copilot

Claude

Gemini

Gemma

GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5

1

2

3

J

4

4o

4.5

4.1

OSS

5

Llama

o1

o3

o4-mini

Qwen

Base44

Claude Code

Cursor

Devstral

GitHub Copilot

Kimi-Dev

Qwen3-Coder

Replit

Xcode

Aurora

Firefly

Flux

GPT Image 1

Ideogram

Imagen

Midjourney

Qwen-Image

Recraft

Seedream

Stable Diffusion

Dream Machine

Hailuo AI

Kling

Midjourney Video

Runway Gen

Seedance

Sora

Veo

Wan

15.ai

Eleven

MiniMax Speech 2.5

WaveNet

Eleven Music

Endel

Lyria

Riffusion

Suno AI

Udio

Agentforce

AutoGLM

AutoGPT

ChatGPT Agent

Devin AI

Manus

OpenAI Codex

Operator

Replit Agent

01.AI

Aleph Alpha

Anthropic

Baichuan

Canva

Cognition AI

Cohere

Contextual AI

DeepSeek

ElevenLabs

Google DeepMind

HeyGen

Hugging Face

Inflection AI

Krikey AI

Kuaishou

Luma Labs

Meta AI

MiniMax

Mistral AI

Moonshot AI

OpenAI

Perplexity AI

Runway

Safe Superintelligence

Salesforce

Scale AI

SoundHound

Stability AI

Synthesia

Thinking Machines Lab

Upstage

xAI

Z.ai

Category

v

t

e

Outline

Glossary

Index

History

Geography

Hall of Fame

Ethics

Laws

Competitions

AI competitions

Aerobot

Anthropomorphic Humanoid Android Cyborg Gynoid

Humanoid

Android

Cyborg

Gynoid

Claytronics

Companion

Automaton Animatronic Audio-Animatronics

Animatronic Audio-Animatronics

Audio-Animatronics

Industrial

Articulated arm

arm

Domestic

Educational

Entertainment

Juggling

Military

Medical

Service

Disability

Agricultural

Food service

Retail

BEAM robotics

Soft robotics

Biorobotics

Cloud robotics

Continuum robot

Unmanned vehicle aerial ground

aerial

ground

Mobile robot

Microbotics

Nanorobotics

Necrobotics

Robotic spacecraft Space probe

Space probe

Swarm

Telerobotics

Underwater remotely-operated Robotic fish

remotely-operated

Robotic fish

Tracks

Walking Hexapod

Hexapod

Climbing

Electric unicycle

Robotic fins

Motion planning

Simultaneous localization and mapping

Visual odometry

Vision-guided robot systems

Evolutionary

Kits

Simulator

Suite

Open-source

Software

Adaptable

Developmental

Human–robot interaction

Paradigms

Perceptual

Situated

Ubiquitous

ABB

Amazon Robotics

Anybots

Barrett Technology

Boston Dynamics

Doosan Robotics

Energid Technologies

FarmWise

FANUC

Figure AI

Foster-Miller

Harvest Automation

HD Hyundai Robotics

Honeybee Robotics

Intuitive Surgical

IRobot

KUKA

Rainbow Robotics

Starship Technologies

Symbotic

Universal Robotics

Wolf Robotics

Yaskawa

Critique of work

Powered exoskeleton

Workplace robotics safety Robotic tech vest

Robotic tech vest

Technological unemployment

Terrainability

Fictional robots

Category

Outline