Top 10 most powerful Physical AI robot-controlling models of 2026

Discover the top 10 physical AI models that help robots operate in real-world settings such as factories, warehouses, and research.

Over the past two years, the gap between the capabilities of language models and the deployment of robots in the real world has narrowed significantly. A new class of platform models is emerging—no longer focused on text generation, but on physical action.

These systems have now been deployed on real-world hardware in factories, warehouses, and research labs. They include robot control policies, experimental vision-language-action (VLA) models, open-source models, and even a world model used to extend training data.

Below are the 10 most important models in the field of 'Physical AI' as of 2026.

NVIDIA Isaac GR00T N-Series (N1.5 / N1.6 / N1.7)

NVIDIA launched the GR00T N1 at GTC in March 2025 as the first open-source foundation model for humanoid robots with reasoning capabilities and general skills.

The N-series then developed very rapidly. GR00T N1.5 (COMPUTEX 5/2025) added a 'freeze' VLM, improved grounding with Eagle 2.5, introduced the FLARE training objective allowing learning from human-view video (ego video), and introduced the GR00T-Dreams blueprint, reducing the time to create synthesized data from

several months to approximately 36 hours.

GR00T N1.6 (December 15, 2025) is upgraded with an NVIDIA Cosmos-2B VLM backbone supporting flexible resolution, doubling the DiT scale (32 layers compared to 16 in N1.5), adding state-relative action chunks for smoother movement, and thousands of hours of teleoperation data from various robotic systems such as YAM bimanual, AGIBot Genie-1, and Unitree G1. This version has been validated on real-world bimanual and locomanipulation tasks.

The latest version , GR00T N1.7 Early Access (April 17, 2026), is a 3B parameter VLA, open-licensed, built on the Cosmos-Reason2-2B backbone with a dual-system Action Cascade architecture. The breakthrough is EgoScale—training on 20,854 hours of human-view video across more than 20 task groups, far exceeding previous teleoperation data. NVIDIA states this is the first time a 'scaling law' has been established for robot dexterity: increasing data from 1,000 to 20,000 hours more than doubles the task completion rate. N1.7 is available on HuggingFace and GitHub with an Apache 2.0 license and is being tested by partners such as AeiRobot, Foxlink, NEURA Robotics, and Lightwheel.

Google DeepMind Gemini Robotics 1.5

Google DeepMind developed Gemini Robotics as a VLA model based on Gemini 2.0, adding physical actions as a new form of output to directly control the robot.

Ra m?t tháng 3/2025 cùng Gemini Robotics-ER (Embodied Reasoning), phiên b?n Gemini Robotics 1.5 (9/2025) b? sung kh? n?ng agentic: chuy?n ??i thông tin hình ?nh và ch? d?n thành l?nh ?i?u khi?n ??ng c?, ??ng th?i làm rõ quá trình suy lu?n ?? robot có th? x? lý các tác v? nhi?u b??c m?t cách minh b?ch h?n.

Mô hình hi?n ch? dành cho ??i tác nh? Agile Robots, Agility Robotics, Boston Dynamics và Enchanted Tools. Nhánh Gemini Robotics-ER ti?p t?c phát tri?n v?i b?n 1.6 (14/4/2026), c?i thi?n suy lu?n không gian và hi?u ?a góc nhìn, ??ng th?i b? sung kh? n?ng ??c thi?t b? ?o (gauge, sight glass) h?p tác cùng Boston Dynamics. Phiên b ?n này có th? truy c?p qua Gemini API và Google AI Studio.

Physical Intelligence ?0 / ?0.5 / ?0.7

?0 ?? xu?t ki?n trúc flow matching trên n?n mô hình vision-language, k? th?a tri th?c ng? ngh?a quy mô internet. Nó ???c hu?n luy?n trên nhi?u lo?i robot có ?? khéo léo cao nh? robot m?t tay, hai tay và mobile manipulator, và ?ã ???c open-source t? tháng 2/2025.

?0.5 (4/2025) không t?p trung t?ng ?? khéo léo mà h??ng t?i generalization trong môi tr??ng m?. Mô hình s? d? ng co-training trên nhi?u nhi?m v?, nhi?u robot, k?t h?p d? ?oán ng? ngh?a c?p cao và d? li?u web ?? x? lý các môi tr??ng ch?a t?ng th?y nh? b?p ho?c phòng ng? m?i. Phiên b?n ti?p theo áp d?ng ph??ng pháp RECAP (RL v ?i Experience & Corrections), h?c t? demonstration, c?i thi?n qua ch?nh s?a và tr?i nghi?m t? ??ng, giúp t?ng g? p ?ôi throughput ? các tác v? nh? l?p filter máy pha cà phê, g?p ?? hay l?p h?p carton.

?0.7 (16/4/2026) t?p trung vào compositional generalization — k?t h?p k? n?ng t? nhi?u ng? c?nh ?? gi?i quy?t nhi?m v? ch?a t?ng hu?n luy?n. ?ây là mô hình có kh? n?ng '?i?u h??ng' (steerable) v?i các n?ng l?c emergent,

???c xem là b??c ti?n h??ng t?i robot ?a n?ng, dù v?n ?ang ? giai ?o?n nghiên c?u.

Figure AI Helix

Helix (20/2/2025) là VLA ??u tiên có th? xu?t ?i?u khi?n liên t?c v?i t?n s? cao cho toàn b? ph?n thân trên robot humanoid, bao g?m c? tay, thân, ??u và t?ng ngón tay.

H? th?ng g?m hai ph?n: System 2 là VLM 7B tham s? ch?y ? 7–9 Hz ?? hi?u ng? c?nh, System 1 là transformer 80M tham s? ch?y ? 200 Hz ?? chuy?n ??i bi?u di?n thành hành ??ng chính xác. Mô hình ???c hu?n luy?n trên kho?ng 500 gi? d? li?u teleoperation ?a robot, ?a ng??i v?n hành.

Helix ch?y hoàn toàn trên GPU nhúng tiêu th? ?i?n th?p, phù h?p tri?n khai th?c t?. Nó s? d?ng m?t b? tr?ng s? duy nh?t cho t?t c? hành vi, không c?n fine-tune theo t?ng task, và ?ã ???c th? nghi?m trong thao tác gia ?ình và phân lo?i hàng hóa logistics. Ngoài ra, nó có th? ?i?u ph?i ??ng th?i hai robot thông qua ki?n trúc supervisory.

OpenVLA

OpenVLA là mô hình VLA mã ngu?n m? 7B tham s?, hu?n luy?n trên 970.000 demonstration robot th?c.

It combines Llama 2 with image encoders using DINOv2 and SigLIP. Despite being seven times smaller, OpenVLA still outperformed RT-2-X (55B) by 16.5 percentage points in success rate across 29 tasks.

The OFT (Optimized Fine-Tuning) method accelerates inference speed by 25–50 times and achieves 97.1% on the LIBERO benchmark. The OFT+ version adds FiLM conditioning to improve grounding and support high-frequency bimanual control. OpenVLA supports LoRA, quantization, and ROS 2 integration.

Octo

Octo is an open-source policy robot from UC Berkeley with two versions: 27M and 93M parameters.

The model uses transformers with diffusion decoding and was trained on over 800,000 episodes from the Open X-Embodiment dataset. It supports diverse inputs (language, images) and adapts to various sensor and action types without requiring architectural changes.

Octo is designed for fast fine-tuning. With around 100 demonstrations, it outperformed initial training by an average of 52% across multiple benchmarks and achieved performance comparable to the RT-2-X in zero-shot, albeit on a much smaller scale.

AGIBOT BFM and GCFM

AGIBOT has unveiled two foundation models in its 'One Robotic Body, Three Intelligences' architecture.

BFM focuses on learning behavior from demonstrations, while GCFM creates actions based on multimodal input (text, audio, video). The company also built the AGIBOT WORLD 2026 dataset from real-world environments and deployed 10,000 robots by March 2026.

Gemini Robotics On-Device

This version is optimized for running directly on robots with low latency and no network required.

It inherits capabilities from Gemini Robotics , primarily training the ALOHA robot and adaptable to the FR3 robot or the Apollo humanoid. The new task learning model uses only 50–100 demonstrations and is currently in a limited testing phase.

NVIDIA Cosmos World Models

Cosmos is not a policy that controls the robot, but rather a world model that creates simulation data.

It can generate trajectory from images and language descriptions, helping robots learn in new environments without needing actual teleoperation data. Cosmos Predict 2 is used in GR00T-Dreams and has been released on HuggingFace.

SmolVLA (HuggingFace LeRobot)

SmolVLA is a compact 450M parameter VLA model from Hugging Face, trained entirely from open-source data.

It uses the SmolVLM-2 backbone combined with transformer flow-matching and was trained on 10 million frames from 487 datasets. SmolVLA runs on mainstream GPUs and MacBooks, with a fine-tuning time of about 4 hours on the A100.

In real-world testing, SmolVLA achieved approximately 78.3% after fine-tuning and performed comparably to or better than larger models in the LIBERO and Meta-World benchmarks. This is the most accessible starting point for teams with limited resources.

The emergence of Physical AI models represents a major shift: AI is no longer just processing information, but is beginning to interact directly with the physical world.

These systems are ushering in a new era where robots can learn, adapt, and perform complex tasks in real-world environments. While challenges remain, the overall trend is clear: AI is moving from 'language' to 'action'.