

# NVIDIA Cosmos 3: One Open Model for Text, Images, Video, Audio, and Robot Actions

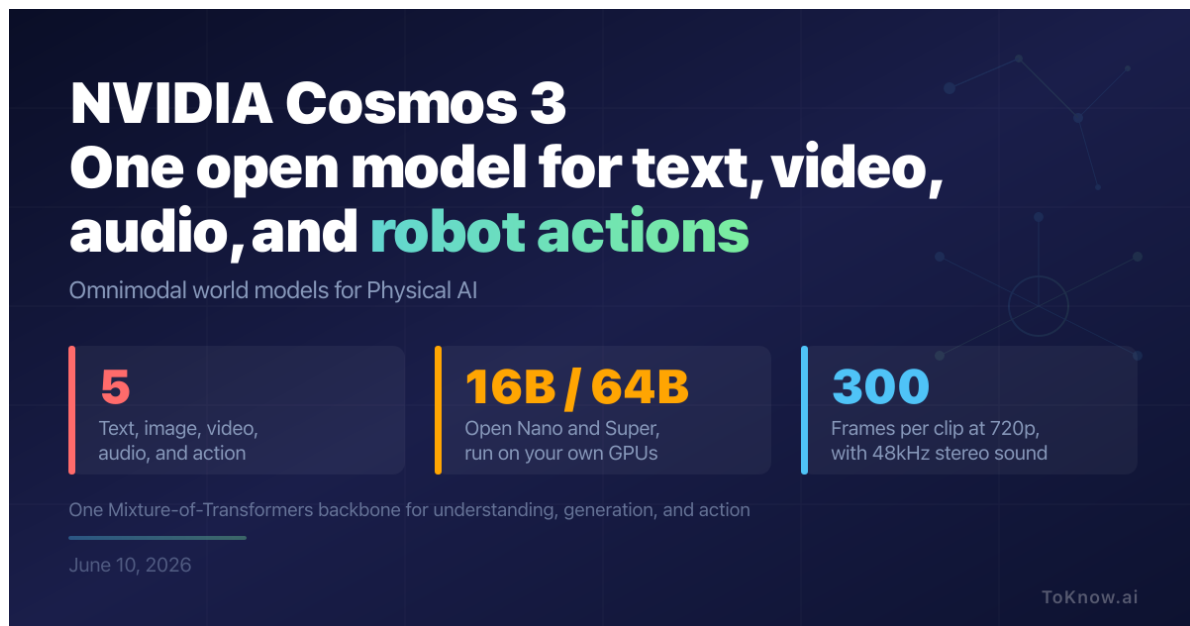
Kabui, Charles

2026-06-10

---

[Read at ToKnow.ai](#)

---



**NVIDIA Cosmos 3**  
**One open model for text, video, audio, and robot actions**

Omnimodal world models for Physical AI

- 5**  
Text, image, video, audio, and action
- 16B / 64B**  
Open Nano and Super, run on your own GPUs
- 300**  
Frames per clip at 720p, with 48kHz stereo sound

One Mixture-of-Transformers backbone for understanding, generation, and action

June 10, 2026

ToKnow.ai

NVIDIA released [Cosmos 3](#), a family of “omnimodal” world models that read and generate five things at once: text, images, video, audio, and robot actions. A single Mixture-of-Transformers design pairs an autoregressive transformer for understanding with a diffusion transformer for generation, so the same weights can describe a scene, imagine how it unfolds, and plan a robot’s moves. It comes in a 16B “Nano” and a 64B “Super” size, and can generate clips up to

300 frames at 720p with synchronized 48kHz stereo sound, plus action trajectories for robots and vehicles. NVIDIA reports the models [ranked best among open systems](#) for image and video generation, and topped a robot-policy benchmark.

Building a robot or self-driving stack usually means stitching together separate models: one to read the camera, one to predict what comes next, one to pick actions. Cosmos 3 folds all of it into one set of weights you can train together. Because NVIDIA released the code, checkpoints, synthetic training scenes, and benchmark under an open license, a small team can run the 16B model on its own GPUs instead of a closed API, and use it to generate the training data robots usually lack.

This moves world models from research demos toward a shared, open backbone for Physical AI, echoing efforts like [OpenWorldLib](#) to unify the many kinds of world model. The open question is no longer whether a model can add a modality, but whether one model can do all of them well enough to trust on a robot.

Sources:

- [NVIDIA Cosmos 3 technical report \(arXiv\)](#)
- [NVIDIA/cosmos on GitHub](#)
- [Cosmos 3 project website \(NVIDIA Research\)](#)
- [Cosmos 3 model collection on Hugging Face](#)

---

*Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. Read more: [/terms-of-service](#)*