

NVIDIA Nemotron 3 Nano Omni: One Open Model for Vision, Audio, and Text at 9x the Throughput

Kabui, Charles

2026-05-11

[Read at ToKnow.ai](#)

The graphic features a dark blue background with a light blue grid pattern. In the top right corner, there is a faint line graph with three data points. The main title is in large white and light blue font. Below the title is a subtitle in smaller white font. Three key metrics are presented in separate boxes with vertical bars on the left: a red bar for '9.2x', a yellow bar for '3B', and a light blue bar for '6'. The date 'May 11, 2026' is at the bottom left, and the 'ToKnow.ai' logo is at the bottom right.

Nemotron 3 Nano Omni: Vision, Audio, and Text at 9x Throughput

Open 30B-A3B hybrid MoE unifying multimodal agent perception

- 9.2x**
Higher throughput vs open omni models (video)
- 3B**
Active params per token (of 30B total MoE)
- 6**
Leaderboards topped for docs, video, and audio

May 11, 2026

ToKnow.ai

NVIDIA released [Nemotron 3 Nano Omni](#) on April 28, a 30-billion-parameter open multi-modal model that handles text, images, audio, and video in one architecture. The “30B-A3B” label means 30 billion total parameters but only about 3 billion active per token, thanks to

a hybrid Mixture-of-Experts design that blends [Mamba layers](#) (for memory-efficient sequence processing) with transformer layers (for precise reasoning). NVIDIA reports [9.2x higher system throughput](#) for video tasks and 7.4x for multi-document tasks compared to other open omni models at the same per-user responsiveness threshold. It tops six leaderboards for document intelligence, video understanding, and audio comprehension, including [OCRBenchV2](#) and [VoiceBench](#). The model ships with open weights, training data, and full recipes, and runs on everything from [Jetson edge hardware](#) to cloud GPUs via [Ollama](#), vLLM, and 25+ partner platforms.

The payoff is collapsing three or four separate models (vision, speech, language, document parsing) into one. An AI agent that needs to watch a screen recording, read a chart, and listen to a voice note can now do it in a single inference loop instead of chaining separate models. [H Company's computer-use agent](#), built on Nano Omni, processes full 1920x1080 screen recordings in real time. Since only 3B parameters fire per token, the model fits on hardware that a dense 30B model never could.

Multimodal AI is moving from “bolt separate encoders onto a language model” to unified architectures where vision, audio, and text share the same reasoning loop. [LongCat-Next](#) showed a similar approach at 74B total (3B active), and now NVIDIA's entry brings full-stack deployment support from edge to cloud.

Sources:

- [Nemotron 3 Nano Omni Technical Report \(arXiv\)](#)
- [NVIDIA Technical Blog: Nemotron 3 Nano Omni](#)
- [NVIDIA Blog: Launch Announcement](#)
- [Nemotron 3 Nano Omni 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](#)