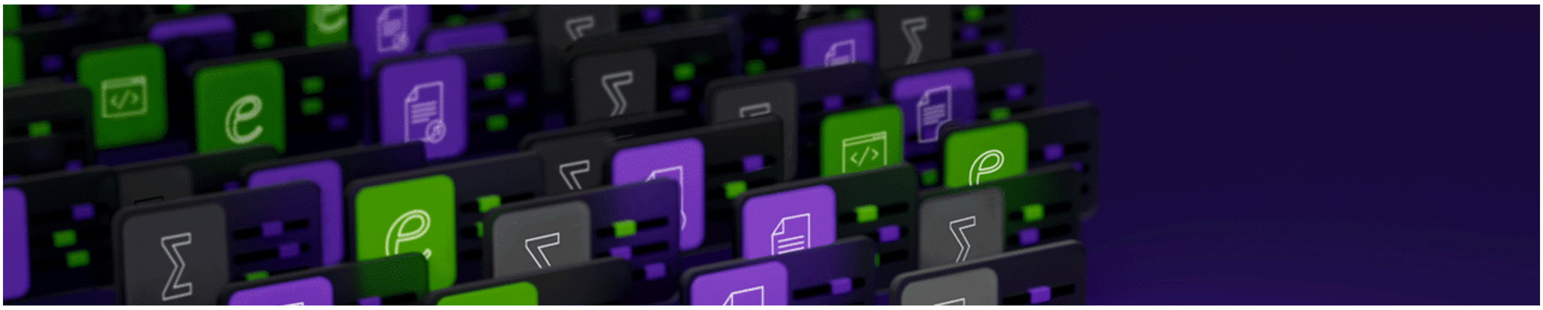
Mistral AI and NVIDIA Unveil Mistral NeMo 12B, a Cutting-Edge Enterprise AI Model

Mistral NeMo's ability to process and generate highly accurate content opens up new opportunities for companies.

July 18, 2024 by Kari Briski



Share

Feedback

Reading Time: 3 mins

Mistral AI and NVIDIA today released a new state-of-the-art language model, Mistral NeMo 12B, that developers can easily customize and deploy for enterprise applications supporting chatbots, multilingual tasks, coding and summarization.

By combining Mistral AI's expertise in training data with NVIDIA's optimized hardware and software ecosystem, the Mistral NeMo model offers high performance for diverse applications.

"We are fortunate to collaborate with the NVIDIA team, leveraging their top-tier hardware and software," said Guillaume Lample, cofounder and chief scientist of Mistral AI. "Together, we have developed a model with unprecedented accuracy, flexibility, high-efficiency and enterprise-grade support and security thanks to NVIDIA AI Enterprise deployment."

Mistral NeMo was trained on the NVIDIA DGX Cloud AI platform, which offers dedicated, scalable access to the latest NVIDIA architecture.

NVIDIA TensorRT-LLM for accelerated inference performance on large language models and the NVIDIA NeMo development platform for building custom generative AI models were also used to advance and optimize the process.

This collaboration underscores NVIDIA's commitment to supporting the model-builder ecosystem.

Delivering Unprecedented Accuracy, Flexibility and Efficiency

Excelling in multi-turn conversations, math, common sense reasoning, world knowledge and coding, this enterprise-grade AI model delivers precise, reliable performance across diverse tasks.

With a 128K context length, Mistral NeMo processes extensive and complex information more coherently and accurately, ensuring contextually relevant outputs.

Released under the Apache 2.0 license, which fosters innovation and supports the broader AI community, Mistral NeMo is a 12-billion-parameter model. Additionally, the model uses the FP8 data format for model inference, which reduces memory size and speeds deployment without any degradation to accuracy.

That means the model learns tasks better and handles diverse scenarios more effectively, making it ideal for enterprise use cases.

Mistral NeMo comes packaged as an NVIDIA NIM inference microservice, offering performance-optimized inference with NVIDIA TensorRT-LLM engines.

This containerized format allows for easy deployment anywhere, providing enhanced flexibility for various applications.

As a result, models can be deployed anywhere in minutes, rather than several days.

NIM features enterprise-grade software that's part of NVIDIA AI Enterprise, with dedicated feature branches, rigorous validation processes, and enterprise-grade security and support.

It includes comprehensive support, direct access to an NVIDIA AI expert and defined service-level agreements, delivering reliable and consistent performance.

The open model license allows enterprises to integrate Mistral NeMo into commercial applications seamlessly.

Designed to fit on the memory of a single NVIDIA L40S, NVIDIA GeForce RTX 4090 or NVIDIA RTX 4500 GPU, the Mistral NeMo NIM offers high efficiency, low compute cost, and enhanced security and privacy.

Advanced Model Development and Customization

The combined expertise of Mistral AI and NVIDIA engineers has optimized training and inference for Mistral NeMo.

Trained with Mistral AI's expertise, especially on multilinguality, code and multi-turn content, the model benefits from accelerated training on NVIDIA's full stack.

It's designed for optimal performance, utilizing efficient model parallelism techniques, scalability and mixed precision with Megatron-LM.

The model was trained using Megatron-LM, part of NVIDIA NeMo, with 3,072 H100 80GB Tensor Core GPUs on DGX Cloud, composed of NVIDIA AI architecture, including accelerated computing, network fabric and software to increase training efficiency.

Availability and Deployment

With the flexibility to run anywhere — cloud, data center or RTX workstation — Mistral NeMo is ready to revolutionize AI applications across various platforms.

Experience Mistral NeMo as an NVIDIA NIM today via ai.nvidia.com, with a downloadable NIM coming soon.

See notice regarding software product information.

Categories: Generative AI

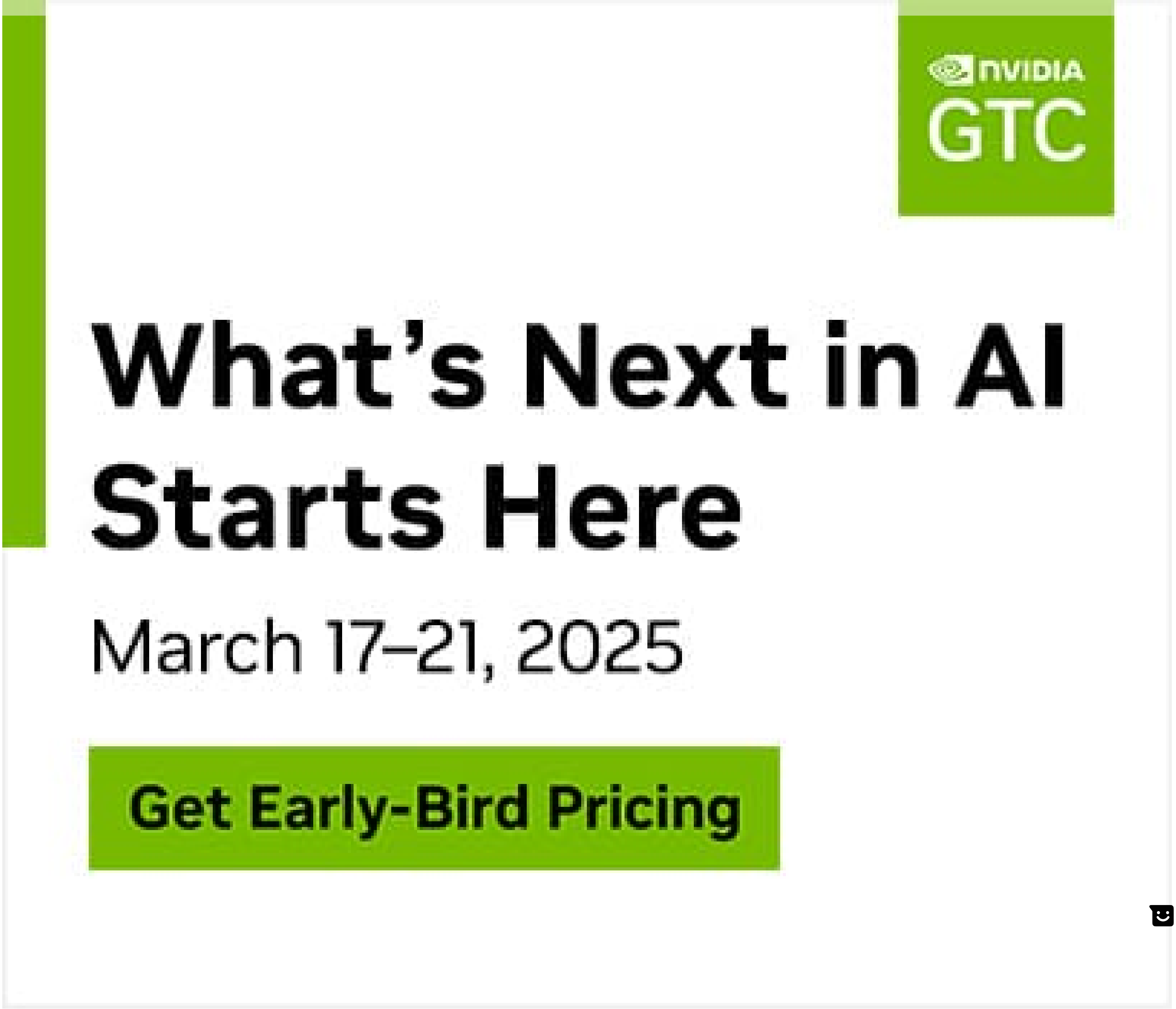
Tags: NVIDIA NeMo

SHARE THIS

FACEBOOK

LINKEDIN

EMAIL



Recommended for You

- NVIDIA RTX Video Super Resolution Update Enhances Video Quality, Detail Preservation and Expands to GeForce RTX 20 Series GPUs
- For the World to See: Nonprofit Deploys GPU-Powered Simulators to Train Providers in Sight-Saving Surgery
- Into the Omniverse: Marmoset Brings Breakthroughs in Rendering, Extends OpenUSD Support to Enhance 3D Art Production
- Foxconn and NVIDIA Amp Up Electric Vehicle Innovation
- Making Machines Mindful: NYU Professor Talks Responsible AI

Stay up to date on the latest enterprise news.

Hot Deal, Cool Prices: GeForce NOW Summer Sale Offers Priority and Ultimate Memberships Half Off

Members can look forward to nine games joining the cloud this week, including Capcom's latest release, 'Kunitsu-Gami: Path of the Goddess.'

July 18, 2024 by GeForce NOW Community

Share

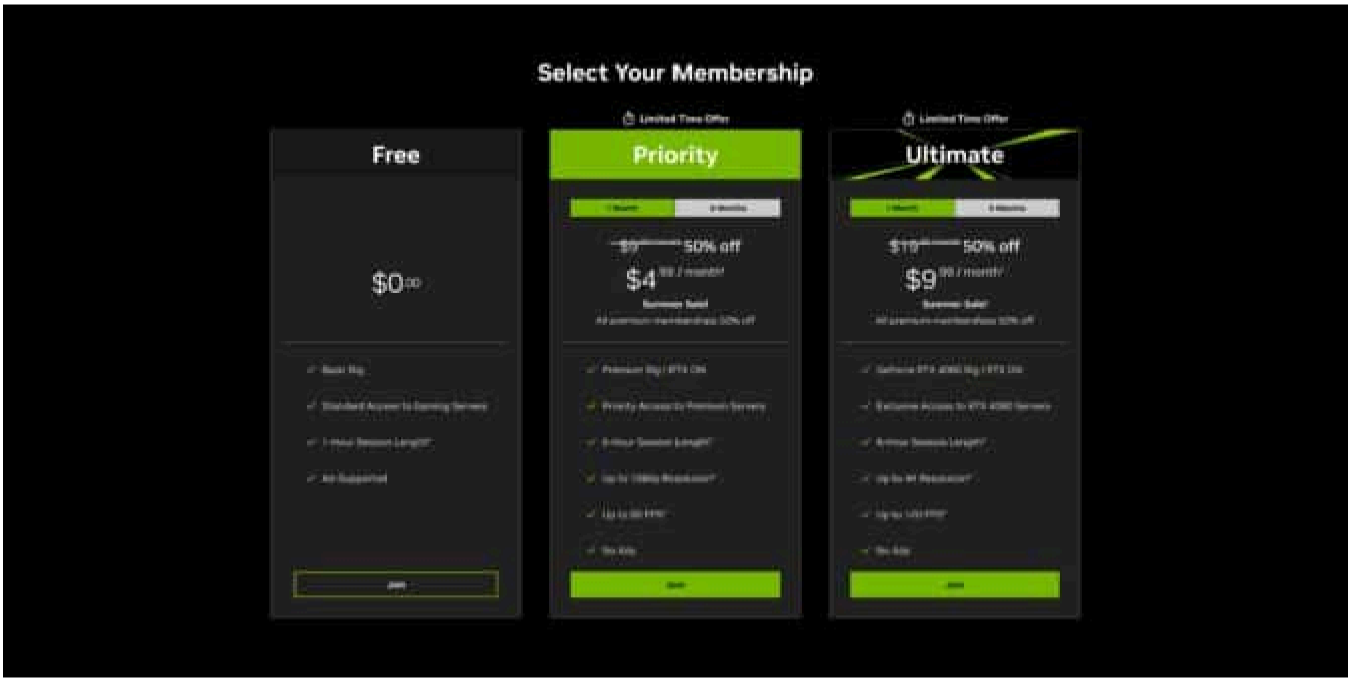
Reading Time: 2 mins

It's time for a sweet treat — the GeForce NOW Summer Sale offers high-performance cloud gaming at half off for a limited time.

And starting today, gamers can directly access supported PC games on GeForce NOW via Xbox.com game pages, enabling them to get into their favorite Xbox PC games even faster.

It all comes with nine new games joining the cloud this week.

We Halve a Deal



Unlock the power of cloud gaming with GeForce NOW's sizzling summer sale.

Take advantage of a special new discount — one-month and six-month GeForce NOW Priority or Ultimate memberships are now 50% off until Aug. 18. It's perfect for members wanting to level up their gaming experience or those looking to try GeForce NOW for the first time to access and stream an ever-growing library of over 1,900 games with top-notch performance.

Priority members enjoy more benefits over free users, including faster access to gaming servers and gaming sessions of up to six hours. They can also stream beautifully ray-traced graphics across multiple devices with RTX ON for the most immersive experience in supported games.

For those looking for top-notch performance, the Ultimate tier provides members with exclusive access to servers and the ability to stream at up to 4K resolution and 120 frames per second, or up to 240 fps — even without upgraded hardware. Ultimate members get all the same benefits as GeForce RTX 40 series GPU owners, including NVIDIA DLSS 3 for the smoothest frame rates and NVIDIA Reflex for the lowest-latency streaming from the cloud.

Strike while it's hot — this scorching summer sale ends soon.

Path of the Goddess

Rinse and repeat.

Capcom's latest release, *Kunitsu-Gami: Path of the Goddess* is a unique Japanese-inspired, single-player Kagura Action Strategy game.

The game takes place on a mountain covered in defilement. During the day, purify the villages and prepare for sundown. During the night, protect the Maiden against the hordes of the Seethe. Repeat the day-and-night cycle until the mountain has been cleansed of defilement and peace has returned to the land.

Walk the path of the goddess in the cloud with extended gaming sessions for Ultimate and Priority members. Ultimate members can also enjoy seeing supernatural and human worlds collide in ultrawide resolutions for an even more immersive experience.

Slay New Games

Having a holiday in Hinterberg.

In *Dungeons of Hinterberg* from Microbird Games, play as Luisa, a burnt-out law trainee taking a break from her fast-paced corporate life. Explore the beautiful alpine village of Hinterberg armed with just a sword and a tourist guide, and uncover the magic hidden within its dungeons. Master magic, solve puzzles and slay monsters — all from the cloud.

Check out the list of new games this week:

- *The Crust* (New release on Steam, July 15)
- *Gestalt: Steam & Cinder* (New release on Steam, July 16)
- *Nobody Wants to Die* (New release on Steam, July 17)
- *Dungeons of Hinterberg* (New release on Steam and Xbox, available on PC Game Pass, July 18)
- *Flintlock: The Siege of Dawn* (New release on Steam and Xbox, available on PC Game Pass, July 18)
- *Norland* (New release on Steam, July 18)
- *Kunitsu-Gami: Path of the Goddess* (New release on Steam, July 19)
- *Content Warning* (Steam)
- *Crime Boss: Rockay City* (Steam)

What are you planning to play this weekend? Let us know on X or in the comments below.

NVIDIA GeForce NOW
@NVIDIAGFN · **Follow**

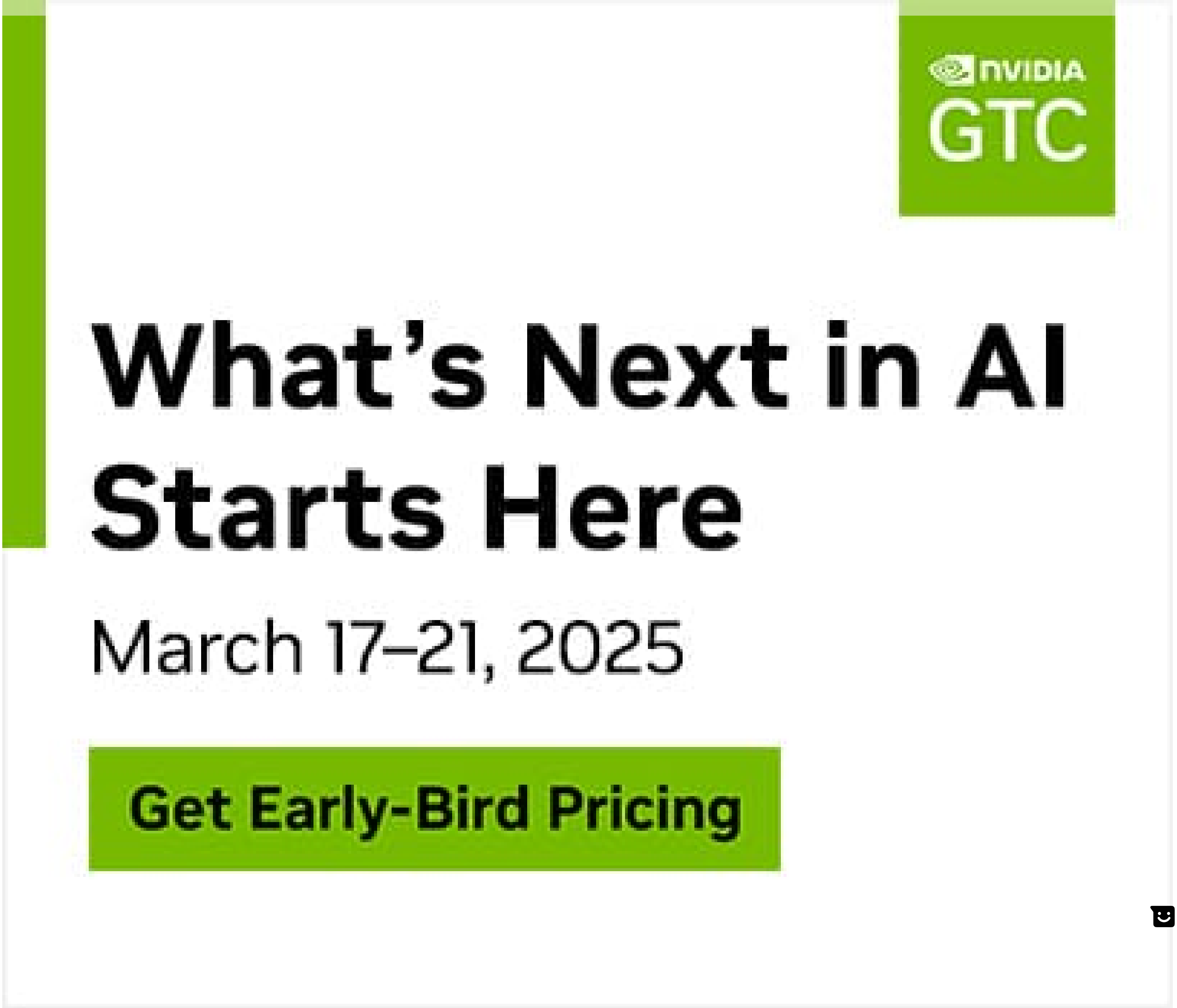
Come sale away this summer ⛵

10:00 AM · Jul 17, 2024

Categories: Gaming

Tags: Cloud Gaming | GeForce NOW

Load Comments



Recommended for You

NVIDIA RTX Video Super Resolution Update Enhances Video Quality, Detail Preservation and Expands to GeForce RTX 20 Series GPUs

For the World to See: Nonprofit Deploys GPU-Powered Simulators to Train Providers in Sight-Saving Surgery

Into the Omniverse: Marmoset Brings Breakthroughs in Rendering, Extends OpenUSD Support to Enhance 3D Art Production

Foxconn and NVIDIA Amp Up Electric Vehicle Innovation

Making Machines Mindful: NYU Professor Talks Responsible AI

Decoding How AI-Powered Upscaling on NVIDIA RTX Improves Video Quality

July 17, 2024 by Brian Choi

Share

Reading Time: 4 mins

Editor's note: This post is part of the AI Decoded series, which demystifies AI by making the technology more accessible, and showcases new hardware, software, tools and accelerations for RTX PC and workstation users.

Video is everywhere — nearly 80% of internet bandwidth today is used to stream video from content providers and social networks. While screens have become bigger and support higher resolutions, nearly all video is only 1080p quality or lower.

Upscalers can help sharpen streamed video and, powered by AI on the NVIDIA RTX platform, significantly enhance image quality and detail.

What Is an Upscaler?

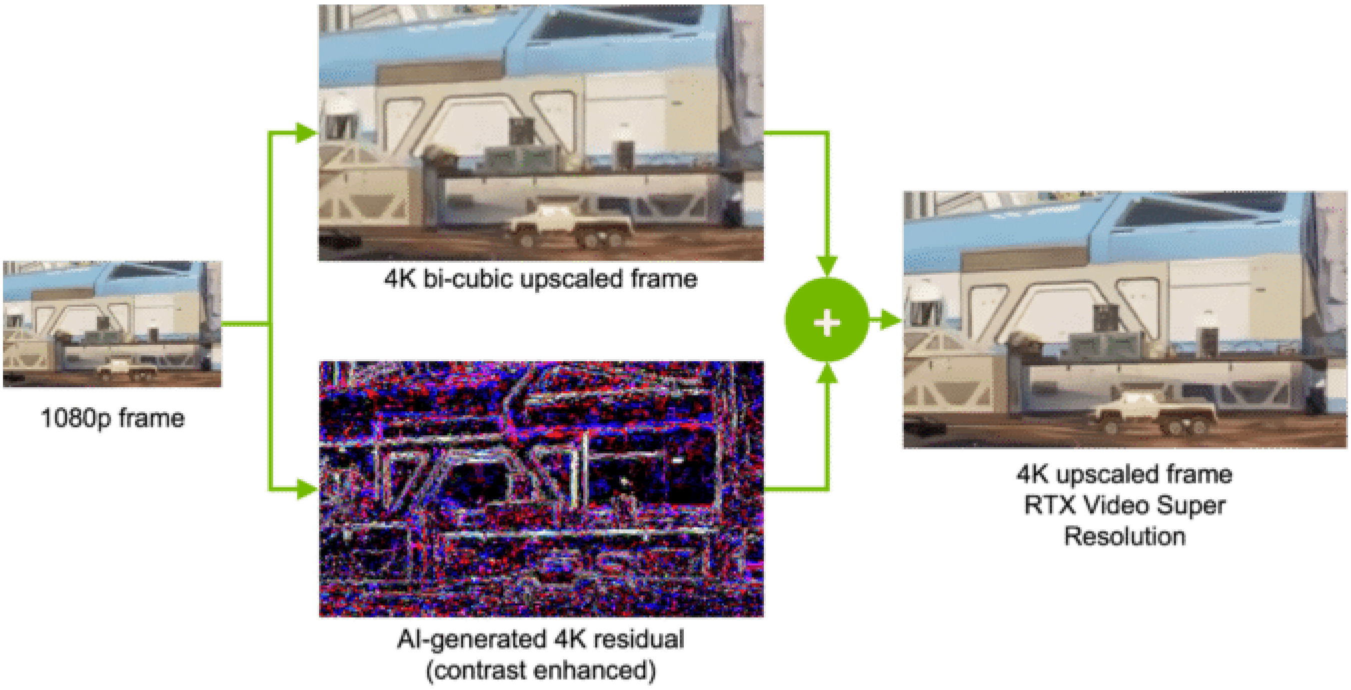
The larger file size of videos makes it harder to compress and transmit compared to images or text. Platforms like Netflix, Vimeo and YouTube work around this limitation by encoding video — the process of compressing the raw source of a video into a smaller container format.

The encoder first analyzes the video to decide what information it can remove to make it fit a target resolution and frame rate. If the target bitrate is insufficient, the video quality decreases, resulting in a loss of detail and sharpness and the presence of encoding artifacts. The smaller the file, the easier it is to share on the internet — but the worse it looks.

Typically, software on the viewer's device will upscale the video file to fit the display's native resolution. However, these upscalers are fairly simplistic, merely multiplying pixels to meet the desired resolution. They can help sharpen the outlines of objects and scenes, but the final video typically carries encoding artifacts and sometimes looks over-sharpened and unnatural.

AI Know a Better Way

The NVIDIA RTX platform uses AI to easily de-artifact and upscale videos.



Easily de-artifact and upscale videos with RTX.

The process of AI upscaling involves analyzing images and motion vectors to generate new details not present in the original video. Instead of merely multiplying pixels, it recognizes the patterns of the image and enhances them to provide greater detail and video quality.

Images must be first de-artifacted before any processing begins. Artifacts — or unwanted distortions and anomalies that appear in video and image files — occur due to overcompression or data loss during transmission and storage.

NVIDIA AI networks can de-artifact images, helping remove blocky areas sometimes seen in streamed video. Without this first step, AI upscalers might end up enhancing the artifacted image itself instead of the desired content.

Super-Sized Video

Just like putting on a pair of prescription glasses can instantly snap the world into focus, RTX Video Super Resolution, one of NVIDIA's latest innovations in AI-enhanced video technology, gives users a clearer picture into the world of streamed video.

Click the image to see the differences between bicubic upscaling (left) and RTX Video Super Resolution (right).

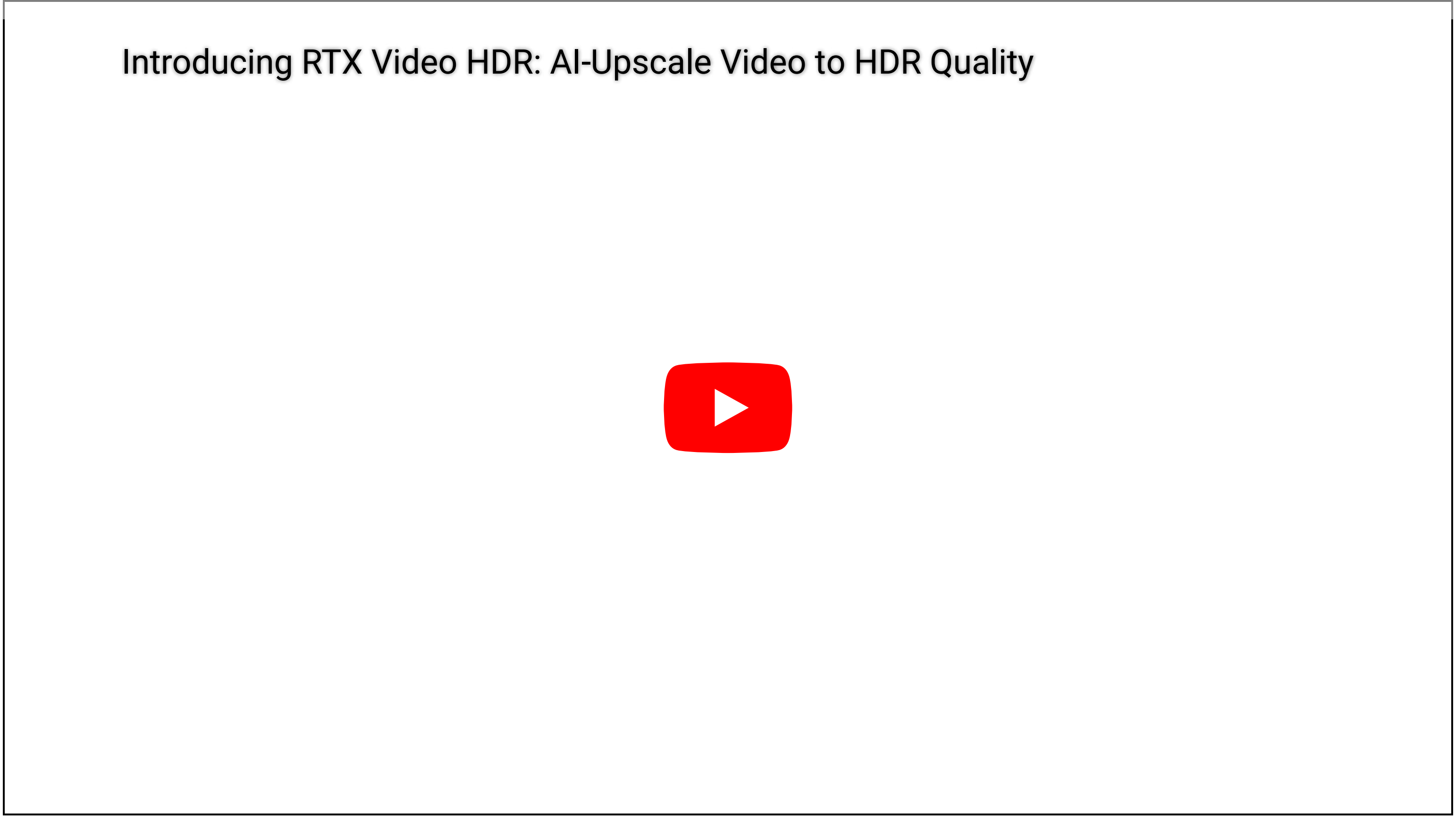
Available on GeForce RTX 40 and 30 Series GPUs and RTX professional GPUs, it uses AI running on dedicated Tensor Cores to remove block compression artifacts and upscale lower-resolution content up to 4K, matching the user's native display resolution.

RTX Video Super Resolution can be used to enhance all video watched on browsers. By combining de-artifacting with AI upscaling techniques, it can make even low-bitrate Twitch streams look stunningly clear. RTX Video Super Resolution is also supported in popular video apps like VLC so users can apply the same upscaling process to their offline videos.

Creators can soon use RTX Video Super Resolution in editing apps like Black Magic's Davinci Resolve, making it easier than ever to upscale lower-quality video files to 4K resolution, as well as convert standard-dynamic range source files into high-dynamic range (HDR).

Say Hi to High-Dynamic Range

RTX Video now also supports AI HDR. HDR video supports a wider range of colors, lending greater detail especially to the darker and lighter areas of images. The problem is that there isn't that much HDR content online yet.



Enter RTX Video HDR — by simply turning on the feature, the AI network will turn any standard or low-dynamic-range content into HDR, performing the correct tone mapping so the image still looks natural and retains its original colors.

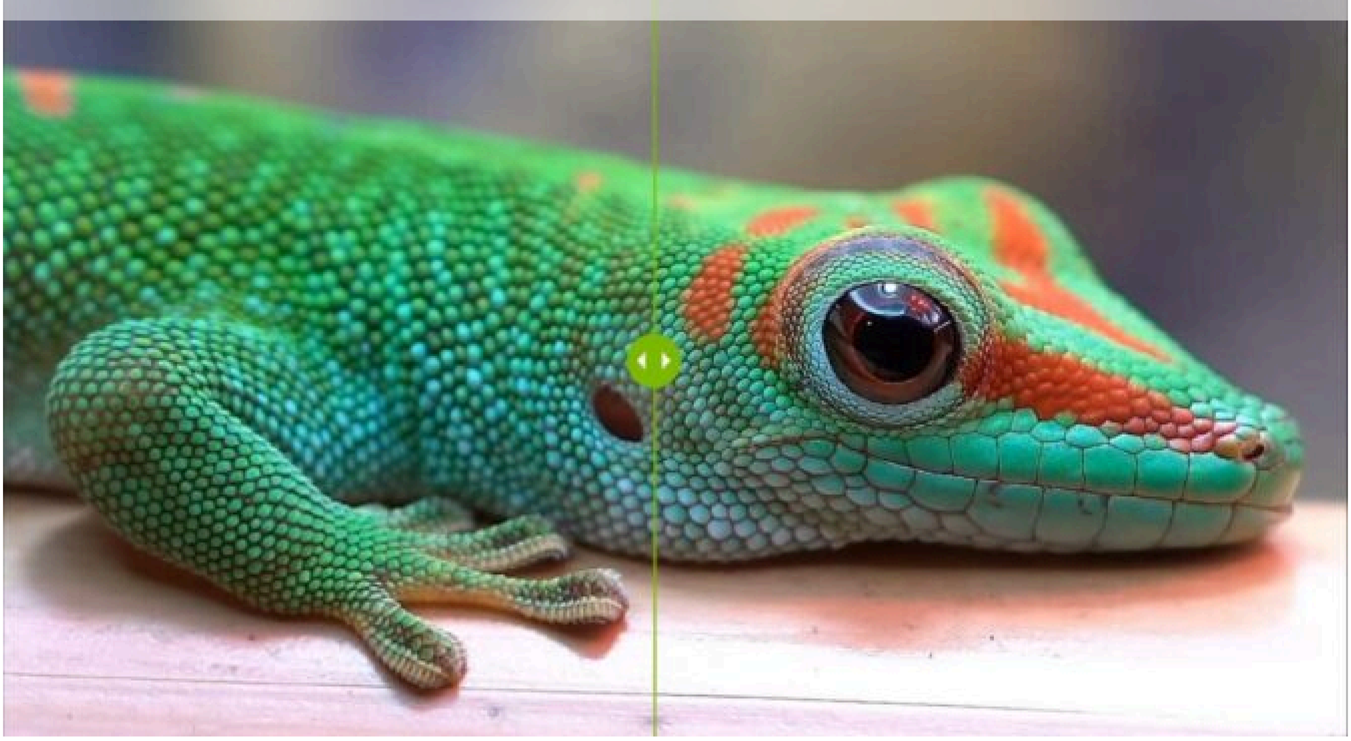
AI Across the Board

RTX Video is just the latest implementation of AI upscaling powered by NVIDIA RTX.

Members of the GeForce NOW cloud streaming service can play their favorite PC games on nearly any device. GeForce RTX servers located all over the world first render the game video content, encode it and then stream it to the player's local device — just like streaming video from other content providers.

Members on older NVIDIA GPU-powered devices can still use AI-enhanced upscaling to improve gameplay quality. This means they can enjoy the best of both worlds — gameplay rendered on servers powered by RTX 4080-class GPUs in the cloud and AI-enhanced streaming quality. Get more information on enabling AI-enhanced upscaling on GeForce NOW.

The NVIDIA SHIELD TV takes this one step further, processing AI neural networks directly on its NVIDIA Tegra system-on-a-chip to upscale 1080p-quality or lower content from nearly any streaming platform to a display's native resolution. That means users can improve the video quality of content streamed from Netflix, Prime Video, Max, Disney+ and more at the push of a remote button.



SHIELD TV is currently available for up to $30 off in North America and £30 or 35€ off in Europe as part of Amazon's Prime Day event running July 16-17. For Prime members in Europe, eligible SHIELD TV purchases also include one month of the GeForce NOW Ultimate membership for free, enabling GeForce RTX 4080-class PC gameplay streamed directly to the living room.

Nvidia Shield TV: Why it's still the BEST Android TV box!

AI has enabled unprecedented improvements in video quality, helping set a new standard in streaming experiences.

Generative AI is transforming gaming, videoconferencing and interactive experiences of all kinds. Make sense of what's new and what's next by subscribing to the AI Decoded newsletter.

Categories: Generative AI

Tags: AI Decoded | Artificial Intelligence | GeForce | GeForce NOW | NVIDIA RTX | SHIELD

Load Comments

Recommended for You

NVIDIA RTX Video Super Resolution Update Enhances Video Quality, Detail Preservation and Expands to GeForce RTX 20 Series GPUs

For the World to See: Nonprofit Deploys GPU-Powered Simulators to Train Providers in Sight-Saving Surgery

Into the Omniverse: Marmoset Brings Breakthroughs in Rendering, Extends OpenUSD Support to Enhance 3D Art Production

Foxconn and NVIDIA Amp Up Electric Vehicle Innovation

Making Machines Mindful: NYU Professor Talks Responsible AI

Meet the Designer Creating Spaces Where NVIDIANs Do Their Life’s Work

July 16, 2024 by Haley Hirai

Share

Reading Time: 2 mins

Early in her career, Jennifer Marko moved from Tokyo to San Francisco with just three suitcases and the ambition to make a difference in the interior design field.

Driven by a passion for creating innovative workplace designs that enable people to do their best work, Marko has spent the past three decades leading interior design and strategy teams at top architecture firms and large global companies.

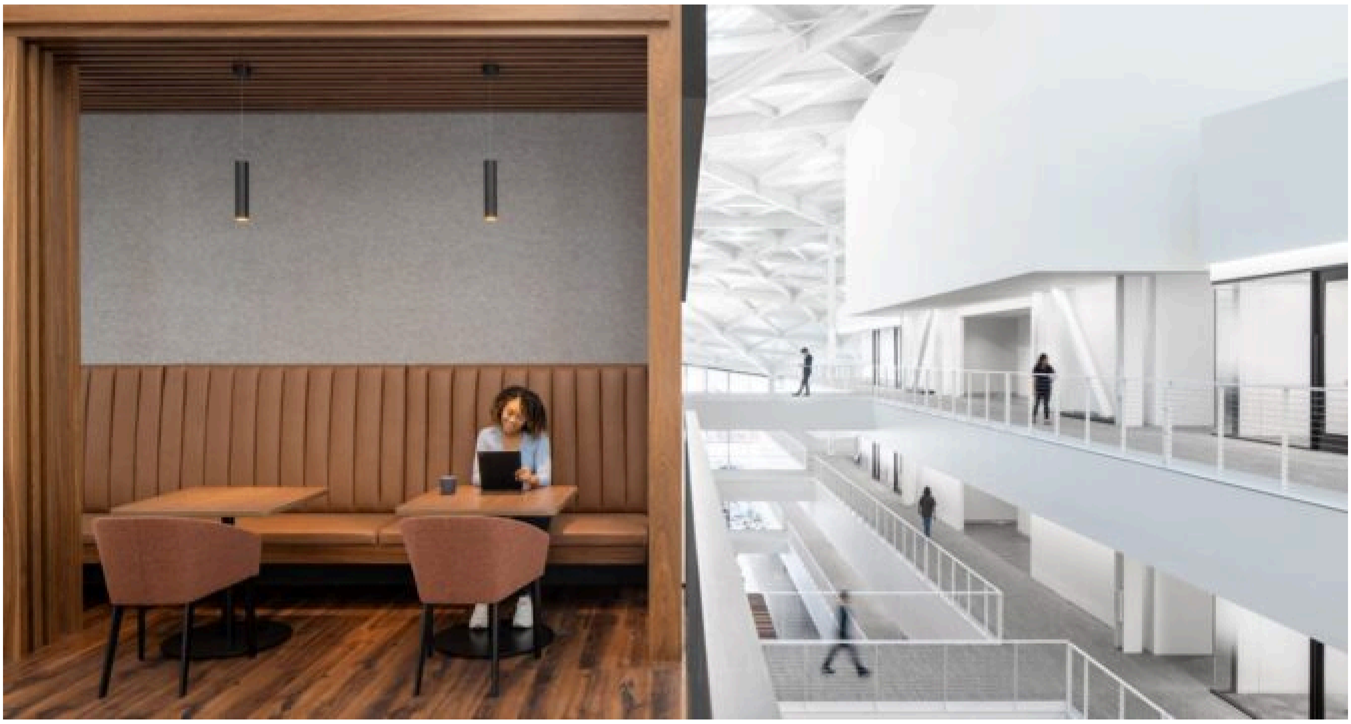
As NVIDIA founder and CEO Jensen Huang began ramping up plans to build the new Voyager building in Santa Clara, Marko joined the company in 2018 as leader of global real estate planning and workplace design with the unique opportunity to write her own job description and create a team from the ground up. Her mission was to help turn Huang’s vision into reality.

The central "mountain" in NVIDIA's Voyager building, part of the Santa Clara campus.

Since then, Marko has built NVIDIA's design, planning and space management teams, and together they've set out to create a cohesive, timeless and functional workplace design across the company's global offices, with custom nuances based on each location.

Data about how employees are using a space drives design decisions. With the rise of flexible work, Marko and team are focused on making sure NVIDIA's spaces reflect new ways of working and facilitate collaboration. Their design guidelines are agile, constantly evolving to meet changing needs — and paralleling the way NVIDIANs approach their work.

"NVIDIA treats its employees with so much respect, and we carry out that guiding principle to provide great workplaces," she said. "Our goal is to create the canvas for NVIDIANs to do the best work of their lives."



“Valleys” on both sides of the central “mountain” separate communal space from conference rooms and workspace in Voyager.

With an architect father who designed airports and Canadian embassies all over the world, Marko lived in Nepal, Malaysia, Indonesia, Japan and her home country of Canada before moving to the U.S. She brings a global perspective to her work and is passionate about unifying NVIDIA’s brand experience across its offices worldwide.

Marko is especially proud of the custom-built, 750,000-square-foot Voyager building that supports more than 3,000 employees at the company’s Santa Clara campus. Offices like this are the result of Huang’s vision and many cross-functional teams’ collaboration to bring the building and workspace to life.

“I love what I do and feel that if you always try to do what’s best for the greater good, success will come,” she said. “My proudest moments are the achievements that we’ve accomplished as a team.”

Learn more about NVIDIA life, culture and careers.

NVIDIANs meet in the company’s new Discovery campus in Bangalore, India.

Categories: NVIDIA Life

Recommended for You

NVIDIA RTX Video Super Resolution Update Enhances Video Quality, Detail Preservation and Expands to GeForce RTX 20 Series GPUs

For the World to See: Nonprofit Deploys GPU-Powered Simulators to Train Providers in Sight-Saving Surgery

Into the Omniverse: Marmoset Brings Breakthroughs in Rendering, Extends OpenUSD Support to Enhance 3D Art Production

Foxconn and NVIDIA Amp Up Electric Vehicle Innovation

Making Machines Mindful: NYU Professor Talks Responsible AI