Explaining Tokens — the Language and Currency of AI

Tokens are units of data processed by AI models during training and inference, enabling prediction, generation and reasoning.

March 17, 2025 by Dave Salvator

Share

Reading Time: 6 mins

Under the hood of every AI application are algorithms that churn through data in their own language, one based on a vocabulary of tokens.

Tokens are tiny units of data that come from breaking down bigger chunks of information. AI models process tokens to learn the relationships between them and unlock capabilities including prediction, generation and reasoning. The faster tokens can be processed, the faster models can learn and respond.

AI factories — a new class of data centers designed to accelerate AI workloads — efficiently crunch through tokens, converting them from the language of AI to the currency of AI, which is intelligence.

With AI factories, enterprises can take advantage of the latest full-stack computing solutions to process more tokens at lower computational cost, creating additional value for customers. In one case, integrating software optimizations and adopting the latest generation NVIDIA GPUs reduced cost per token by 20x compared to unoptimized processes on previous-generation GPUs — delivering 25x more revenue in just four weeks.

By efficiently processing tokens, AI factories are manufacturing intelligence — the most valuable asset in the new industrial revolution powered by AI.

What Is Tokenization?

Whether a transformer AI model is processing text, images, audio clips, videos or another modality, it will translate the data into tokens. This process is known as tokenization.

Efficient tokenization helps reduce the amount of computing power required for training and inference. There are numerous tokenization methods — and tokenizers tailored for specific data types and use cases can require a smaller vocabulary, meaning there are fewer tokens to process.

For large language models (LLMs), short words may be represented with a single token, while longer words may be split into two or more tokens.

The word darkness, for example, would be split into two tokens, “dark” and “ness,” with each token bearing a numerical representation, such as 217 and 655. The opposite word, brightness, would similarly be split into “bright” and “ness,” with corresponding numerical representations of 491 and 655.

In this example, the shared numerical value associated with "ness" can help the AI model understand that the words may have something in common. In other situations, a tokenizer may assign different numerical representations for the same word depending on its meaning in context.

For example, the word "lie" could refer to a resting position or to saying something untruthful. During training, the model would learn the distinction between these two meanings and assign them different token numbers.

For visual AI models that process images, video or sensor data, a tokenizer can help map visual inputs like pixels or voxels into a series of discrete tokens.

Models that process audio may turn short clips into spectrograms — visual depictions of sound waves over time that can then be processed as images. Other audio applications may instead focus on capturing the meaning of a sound clip containing speech, and use another kind of tokenizer that captures semantic tokens, which represent language or context data instead of simply acoustic information.

How Are Tokens Used During AI Training?

Training an AI model starts with the tokenization of the training dataset.

Based on the size of the training data, the number of tokens can number in the billions or trillions — and, per the pretraining scaling law, the more tokens used for training, the better the quality of the AI model.

As an AI model is pretrained, it's tested by being shown a sample set of tokens and asked to predict the next token. Based on whether or not its prediction is correct, the model updates itself to improve its next guess. This process is repeated until the model learns from its mistakes and reaches a target level of accuracy, known as model convergence.

After pretraining, models are further improved by post-training, where they continue to learn on a subset of tokens relevant to the use case where they'll be deployed. These could be tokens with domain-specific information for an application in law, medicine or business — or tokens that help tailor the model to a specific task, like reasoning, chat or translation. The goal is a model that generates the right tokens to deliver a correct response based on a user's query — a skill better known as inference.

How Are Tokens Used During AI Inference and Reasoning?

During inference, an AI receives a prompt — which, depending on the model, may be text, image, audio clip, video, sensor data or even gene sequence — that it translates into a series of tokens. The model processes these input tokens, generates its response as tokens and then translates it to the user's expected format.

Input and output languages can be different, such as in a model that translates English to Japanese, or one that converts text prompts into images.

To understand a complete prompt, AI models must be able to process multiple tokens at once. Many models have a specified limit, referred to as a context window — and different use cases require different context window sizes.

A model that can process a few thousand tokens at once might be able to process a single high-resolution image or a few pages of text. With a context length of tens of thousands of tokens, another model might be able to summarize a whole novel or an hourlong podcast episode. Some models even provide context lengths of a million or more tokens, allowing users to input massive data sources for the AI to analyze.

Reasoning AI models, the latest advancement in LLMs, can tackle more complex queries by treating tokens differently than before. Here, in addition to input and output tokens, the model generates a host of reasoning tokens over minutes or hours as it thinks about how to solve a given problem.

These reasoning tokens allow for better responses to complex questions, just like how a person can formulate a better answer given time to work through a problem. The corresponding increase in tokens per prompt can require over 100x more compute compared with a single inference pass on a traditional LLM — an example of test-time scaling, aka long thinking.

How Do Tokens Drive AI Economics?

During pretraining and post-training, tokens equate to investment into intelligence, and during inference, they drive cost and revenue. So as AI applications proliferate, new principles of AI economics are emerging.

AI factories are built to sustain high-volume inference, manufacturing intelligence for users by turning tokens into monetizable insights. That's why a growing number of AI services are measuring the value of their products based on the number of tokens consumed and generated, offering pricing plans based on a model's rates of token input and output.

Some token pricing plans offer users a set number of tokens shared between input and output. Based on these token limits, a customer could use a short text prompt that uses just a few tokens for the input to generate a lengthy, AI-generated response that took thousands of tokens as the output. Or a user could spend the majority of their tokens on input, providing an AI model with a set of documents to summarize into a few bullet points.

To serve a high volume of concurrent users, some AI services also set token limits, the maximum number of tokens per minute generated for an individual user.

Tokens also define the user experience for AI services. Time to first token, the latency between a user submitting a prompt and the AI model starting to respond, and inter-token or token-to-token latency, the rate at which subsequent output tokens are generated, determine how an end user experiences the output of an AI application.

There are tradeoffs involved for each metric, and the right balance is dictated by use case.

For LLM-based chatbots, shortening the time to first token can help improve user engagement by maintaining a conversational pace without unnatural pauses. Optimizing inter-token latency can enable text generation models to match the reading speed of an average person, or video generation models to achieve a desired frame rate. For AI models engaging in long thinking and research, more emphasis is placed on generating high-quality tokens, even if it adds latency.

Developers have to strike a balance between these metrics to deliver high-quality user experiences with optimal throughput, the number of tokens an AI factory can generate.

To address these challenges, the NVIDIA AI platform offers a vast collection of software, microservices and blueprints alongside powerful accelerated computing infrastructure — a flexible, full-stack solution that enables enterprises to evolve, optimize and scale AI factories to generate the next wave of intelligence across industries.

Understanding how to optimize token usage across different tasks can help developers, enterprises and even end users reap the most value from their AI applications.

Learn more in this ebook and get started at build.nvidia.com.

Categories: Explainer | Generative AI

Tags: AI Factory | Artificial Intelligence | Inference

All NVIDIA News

Tarun Patil Engineers Success on NVIDIA's Circuit Silicon Correlation Team

Plug and Play: Build a G-Assist Plug-In Today

Hexagon Taps NVIDIA Robotics and AI Software to Build and Deploy AEON, a New Humanoid

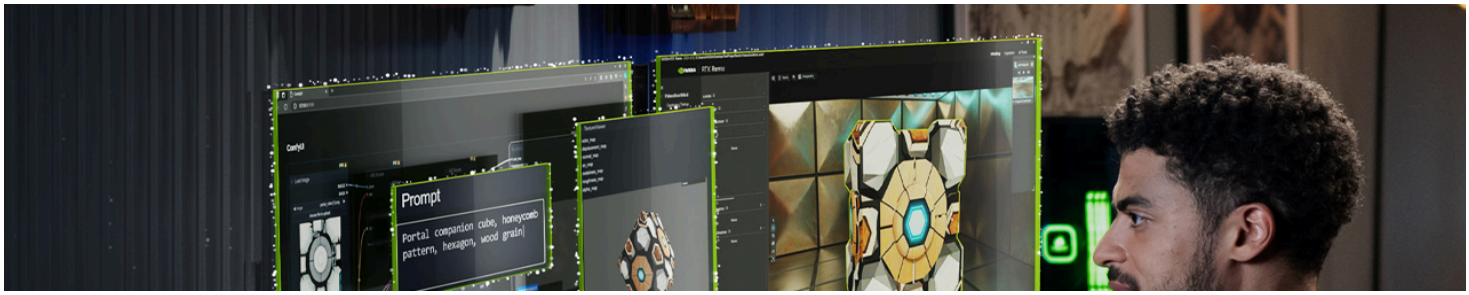
NVIDIA and Deutsche Telekom Partner to Advance Germany's Sovereign AI

Turn RTX ON With 40% Off Performance Day Passes

Drop It Like It's Mod: Breathing New Life Into Classic Games With AI in NVIDIA RTX Remix

Now generally available today, RTX Remix delivers NVIDIA DLSS 4, NVIDIA RTX neural rendering and AI-powered texture tools to modders worldwide.

March 13, 2025 by Nyle Usmani



Share

Reading Time: 3 mins

PC game modding is massive, with over 5 billion mods downloaded annually. Mods push graphics forward with each GPU generation, extend a game's lifespan with new content and attract new players.

NVIDIA RTX Remix is a modding platform for RTX AI PCs that lets modders capture game assets, automatically enhance materials with generative AI tools and create stunning RTX remasters with full ray tracing. Today, RTX Remix exited beta and fully launched with new NVIDIA GeForce RTX 50 Series neural rendering technology and many community-requested upgrades.

Since its initial beta release, RTX Remix has been experimented with by over 30,000 modders, bringing ray-traced mods of hundreds of classic titles to over 1 million gamers.

RTX Remix supports a host of AI tools, including NVIDIA DLSS 4, RTX Neural Radiance Cache and the community-published AI model PBRFusion 3.

Modders can build 4K physically based rendering (PBR) assets by hand or use generative AI to accelerate their workflows. And with a few additional clicks, RTX Remix mods support DLSS 4 with Multi Frame Generation. DLSS' new transformer model and the first neural shader, Neural Radiance Cache, provide enhanced neural rendering performance, meaning classic games look and play better than ever.

Generative AI Texture Tools

RTX Remix's built-in generative AI texture tools analyze low-resolution textures from classic games, generate physically accurate materials — including normal and roughness maps — and upscale the resolution by up to 4x. Many RTX Remix mods have been created incorporating generative AI.

Earlier this month, RTX Remix modder NightRaven published PBRFusion 3 — a new AI model that upscales textures and generates high-quality normal, roughness and height maps for physically-based materials.

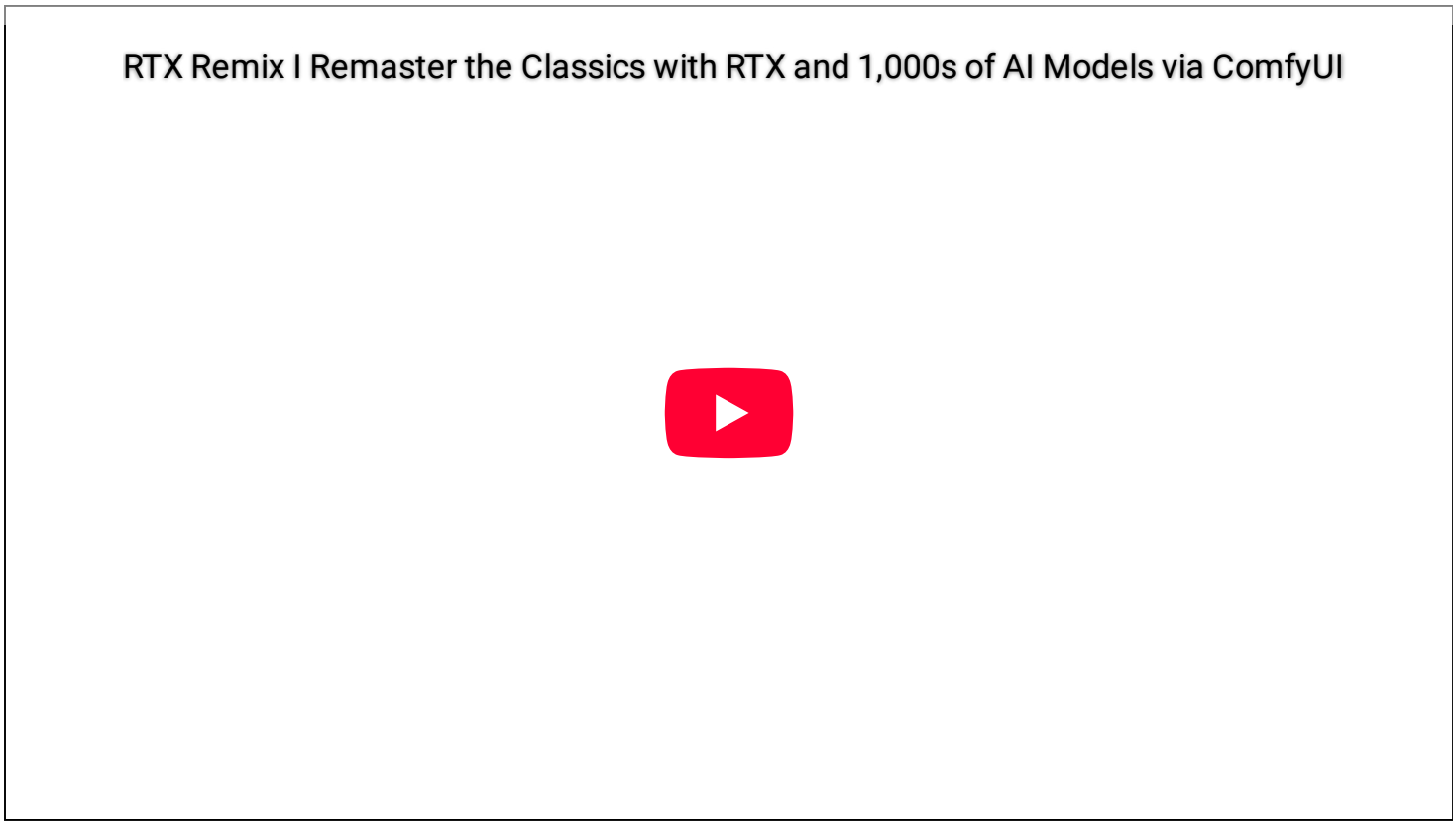
PBRFusion 3 consists of two custom-trained models: a PBR model and a diffusion-based upscaler. PBRFusion 3 can also use the RTX Remix application programming interface to connect with ComfyUI in an integrated flow. NightRaven has packaged all the relevant pieces to make it easy to get started.

The PBRFusion3 page features a plug-and-play package that includes the relevant ComfyUI graphs and nodes. Once installed, remastering is easy. Select a number of textures in RTX Remix's Viewport and hit process in ComfyUI. This integrated flow enables extensive remasters of popular games to be completed by small hobbyist mod teams.

RTX Remix and REST API

RTX Remix Toolkit capabilities are accessible via REST API, allowing modders to livelink RTX Remix to digital content creation tools such as Blender, modding tools such as Hammer and generative AI apps such as ComfyUI.

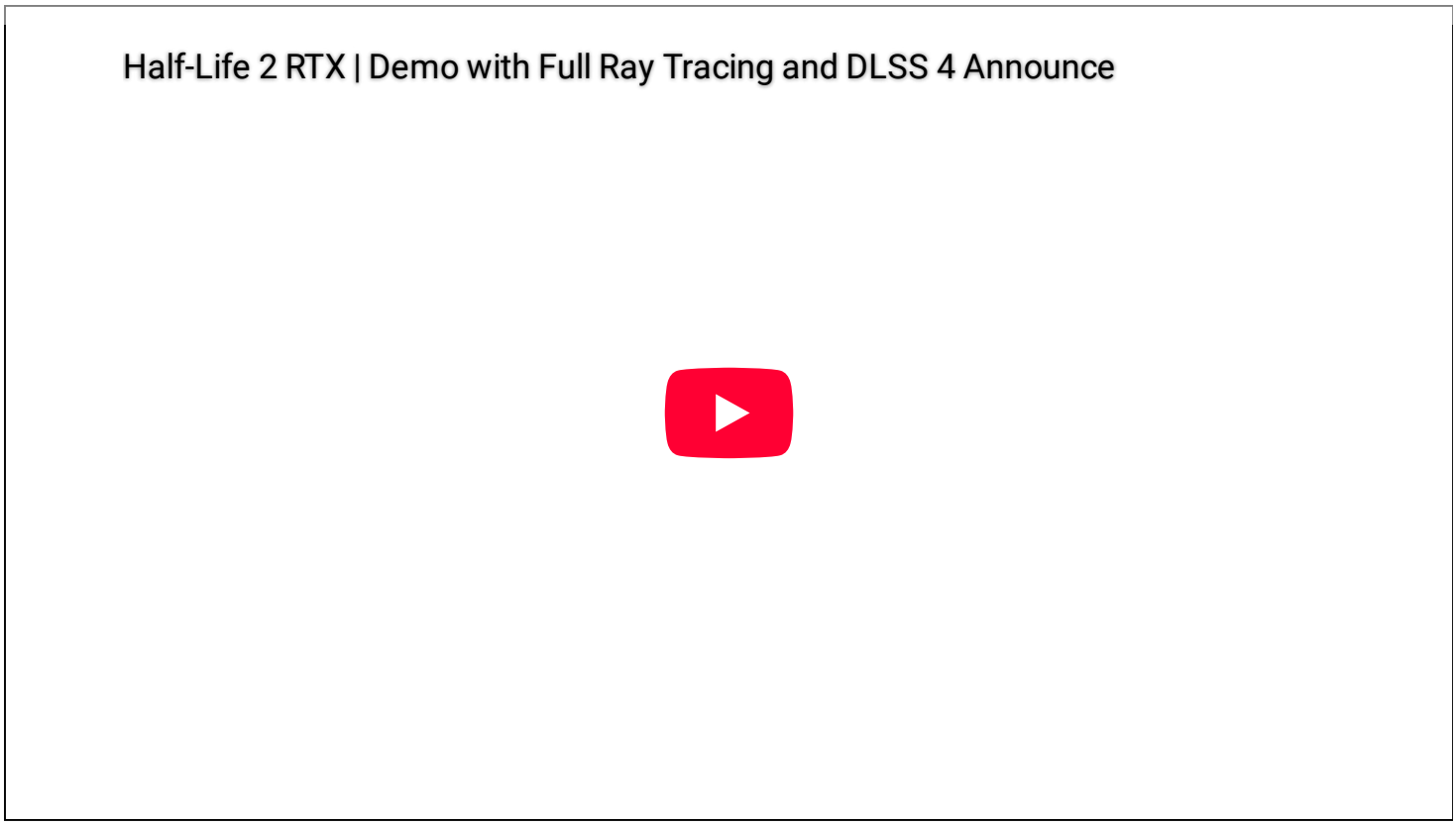
For example, through REST API integration, modders can seamlessly export all game textures captured in RTX Remix to ComfyUI and enhance them in one big batch before automatically bringing them back into the game. ComfyUI is RTX-accelerated and includes thousands of generative AI models to try, helping reduce the time to remaster a game scene and providing many ways to process textures.



Modders have many super resolution and PBR models to choose from, including ones that feature metallic and height maps — unlocking 8x or more resolution increases. Additionally, ComfyUI enables modders to use text prompts to generate new details in textures, or make grand stylistic departures by changing an entire scene's look with a single text prompt.

'Half-Life 2 RTX' Demo

Half-Life 2 owners can download a free *Half-Life 2 RTX* demo from Steam, built with RTX Remix, starting March 18. The demo showcases Orbifold Studios' work in Ravenholm and Nova Prospekt ahead of the full game's release at a later date.



Half-Life 2 RTX showcases the expansive capabilities of RTX Remix and NVIDIA's neural rendering technologies. DLSS 4 with Multi Frame Generation multiplies frame rates by up to 10x at 4K. Neural Radiance Cache further accelerates ray-traced lighting. RTX Skin enhances Father Grigori, headcrabs and zombies with one of the first implementations of subsurface scattering in ray-traced gaming. RTX Volumetrics add realistic smoke effects and fog. And everything interplays and interacts with the fully ray-traced lighting.

What's Next in AI Starts Here

From the keynote by NVIDIA founder and CEO Jensen Huang on Tuesday, March 18, to over 1,000 inspiring sessions, 300+ exhibits, technical hands-on training and tons of unique networking events — NVIDIA's own GTC is set to put a spotlight on AI and all its benefits.

Experts from across the AI ecosystem will share insights on deploying AI locally, optimizing models and harnessing cutting-edge hardware and software to enhance AI workloads — highlighting key advancements in RTX AI PCs and workstations. RTX AI Garage will be there to share highlights of the latest advancements coming to the RTX AI platform.

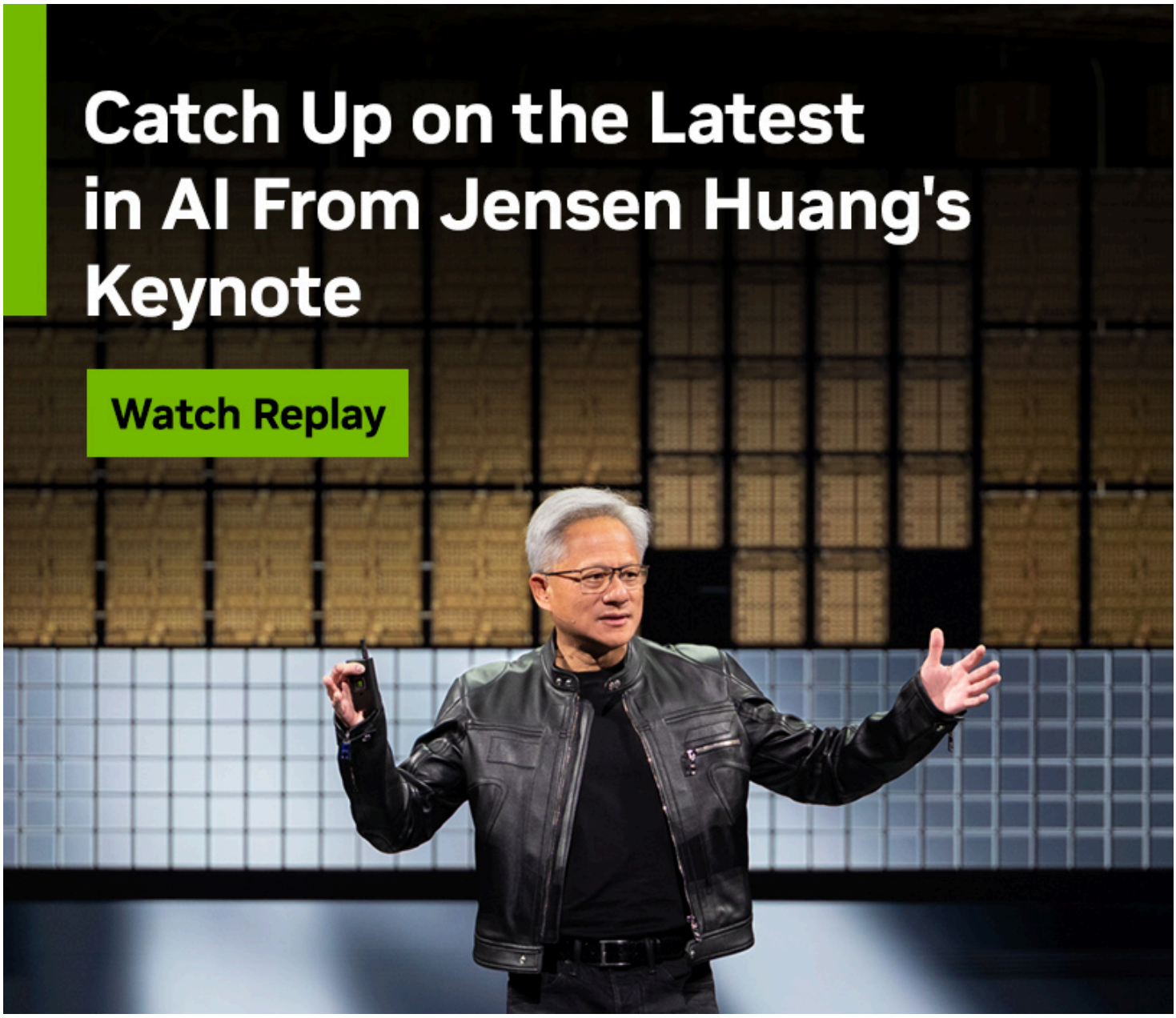
Follow NVIDIA AI PC on Facebook, Instagram, TikTok and X — and stay informed by subscribing to the RTX AI PC newsletter

---

Categories: Corporate | Generative AI

Tags: Artificial Intelligence | GeForce | NVIDIA RTX | RTX AI Garage | RTX Mobile Workstations

Load Comments



All NVIDIA News

NVIDIA Research Showcases the Future of Robotics at RSS

Step Inside the Vault: The 'Borderland' Series Arrives on GeForce NOW

Tarun Patil Engineers Success on NVIDIA's Circuit Silicon Correlation Team

Plug and Play: Build a G-Assist Plug-In Today

Hexagon Taps NVIDIA Robotics and AI Software to Build and Deploy AEON, a New Humanoid

Gaming Goodness: NVIDIA Reveals Latest Neural Rendering and AI Advancements Supercharging Game Development at GDC 2025

New neural rendering tools, rapid NVIDIA DLSS 4 adoption, ‘Half-Life 2 RTX’ demo and digital human technology enhancements are among NVIDIA’s announcements at the premier conference for game developers.

March 13, 2025 by Angie Lee

Share

Reading Time: 4 mins

AI is leveling up the world’s most beloved games, as the latest advancements in neural rendering, NVIDIA RTX and digital human technologies equip game developers to take innovative leaps in their work.

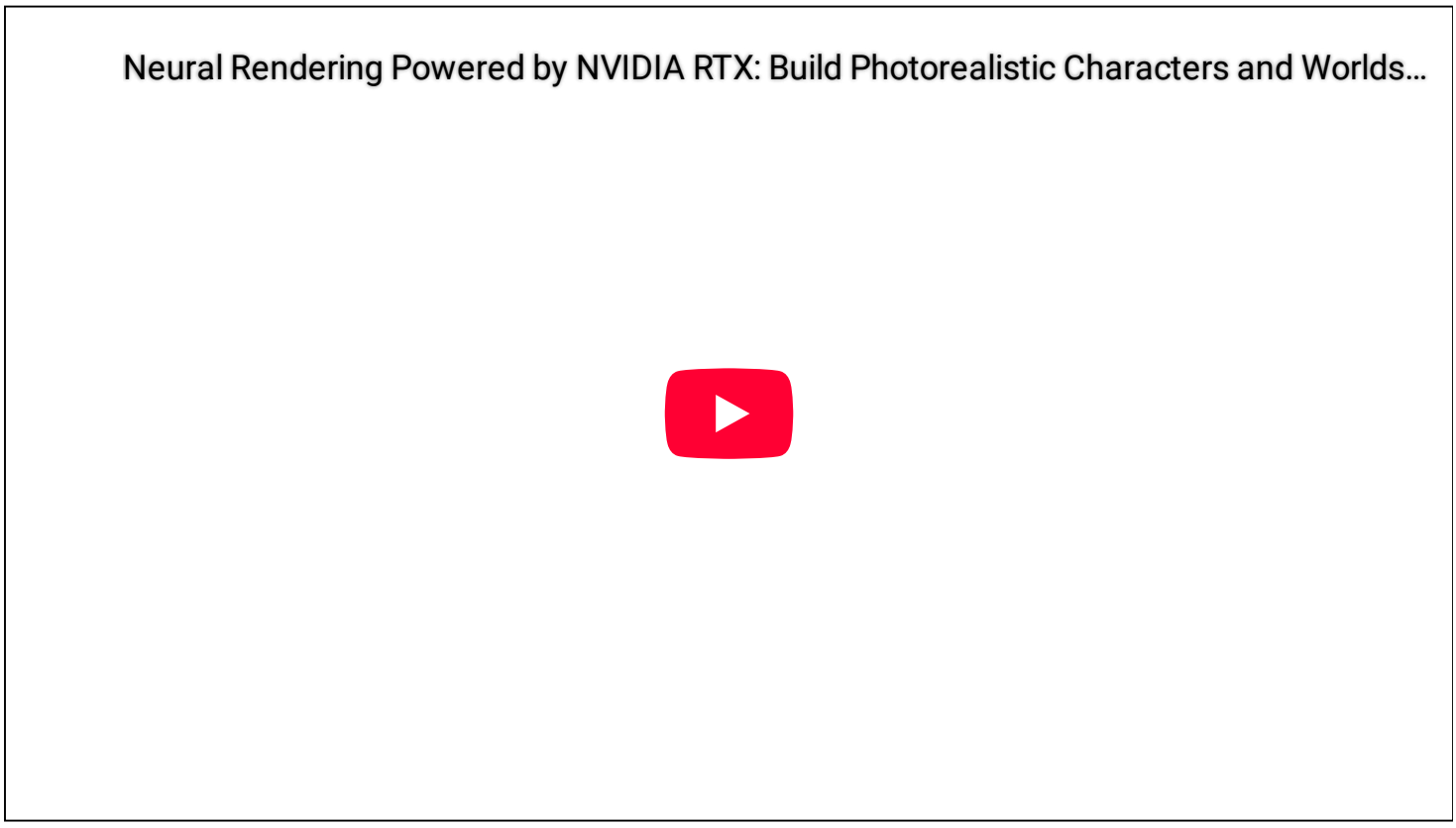
At this year's GDC conference, running March 17-21 in San Francisco, NVIDIA is revealing new AI tools and technologies to supercharge the next era of graphics in games.

Key announcements include new neural rendering advancements with Unreal Engine 5 and Microsoft DirectX; NVIDIA DLSS 4 now available in over 100 games and apps, making it the most rapidly adopted NVIDIA game technology of all time; and a *Half-Life 2 RTX* demo coming Tuesday, March 18.

Plus, the open-source NVIDIA RTX Remix modding platform has now been released, and NVIDIA ACE technology enhancements are bringing to life next-generation digital humans and AI agents for games.

Neural Shaders Enable Photorealistic, Living Worlds With AI

The next era of computer graphics will be based on NVIDIA RTX Neural Shaders, which allow the training and deployment of tiny neural networks from within shaders to generate textures, materials, lighting, volumes and more. This results in dramatic improvements in game performance, image quality and interactivity, delivering new levels of immersion for players.



At the CES trade show earlier this year, NVIDIA introduced RTX Kit, a comprehensive suite of neural rendering technologies for building AI-enhanced, ray-traced games with massive geometric complexity and photorealistic characters.

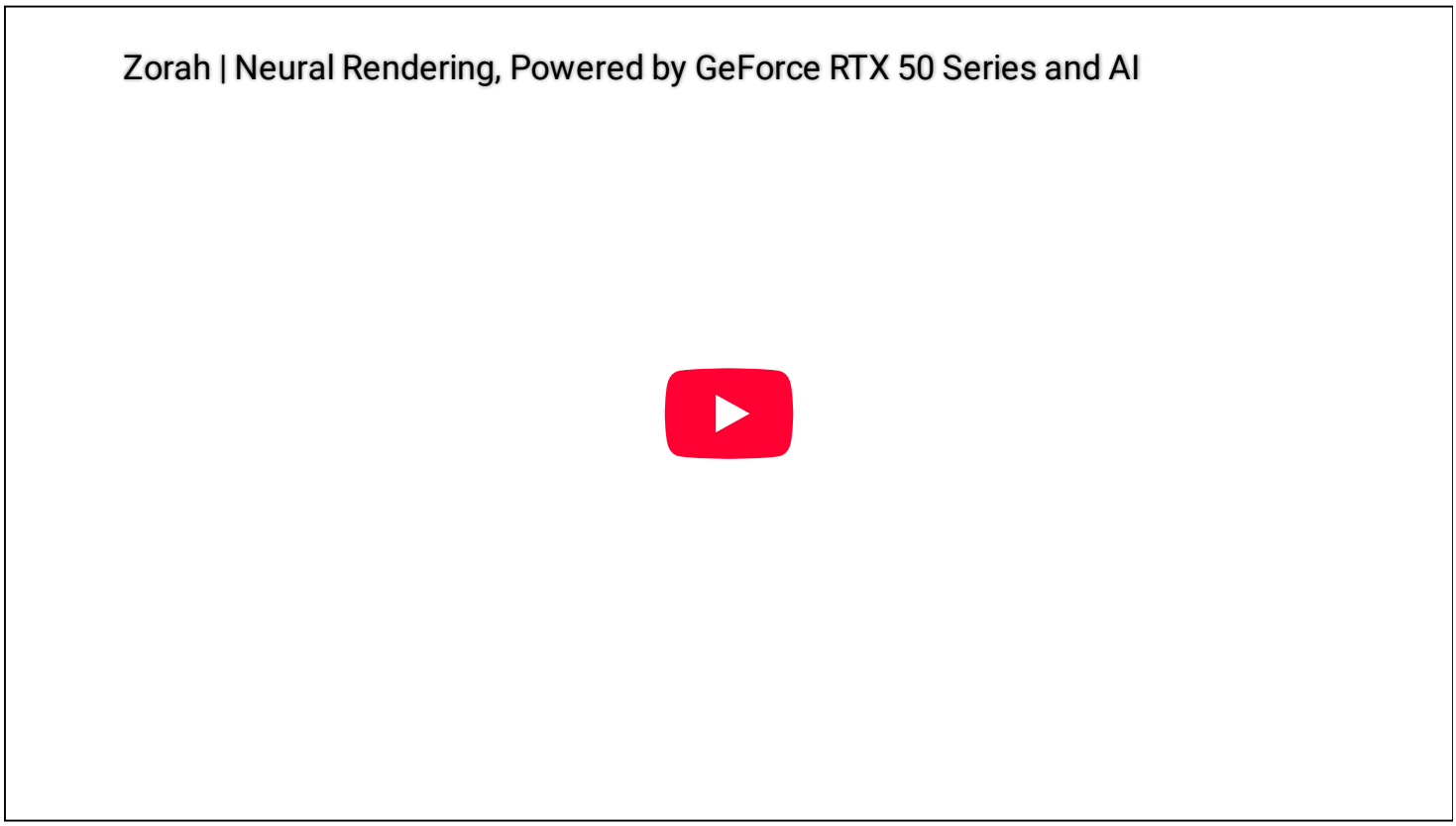
Now, at GDC, NVIDIA is expanding its powerful lineup of neural rendering technologies, including with Microsoft DirectX support and plug-ins for Unreal Engine 5.

NVIDIA is partnering with Microsoft to bring neural shading support to the DirectX 12 Agility software development kit preview in April, providing game developers with access to RTX Tensor Cores to accelerate the performance of applications powered by RTX Neural Shaders.

Plus, Unreal Engine developers will be able to get started with RTX Kit features such as RTX Mega Geometry through the experimental NVIDIA RTX Branch of Unreal Engine 5. These enable the rendering of assets with dramatic detail and fidelity, bringing cinematic-quality visuals to real-time experiences.

Now available, NVIDIA's "Zorah" technology demo has been updated with new incredibly detailed scenes filled with millions of triangles and cinematic lighting in real time — all by tapping into the latest technologies powering neural rendering, including:

- ReSTIR Path Tracing
- ReSTIR Direct Illumination
- RTX Mega Geometry



And the first neural shader, Neural Radiance Cache, is now available in RTX Remix.

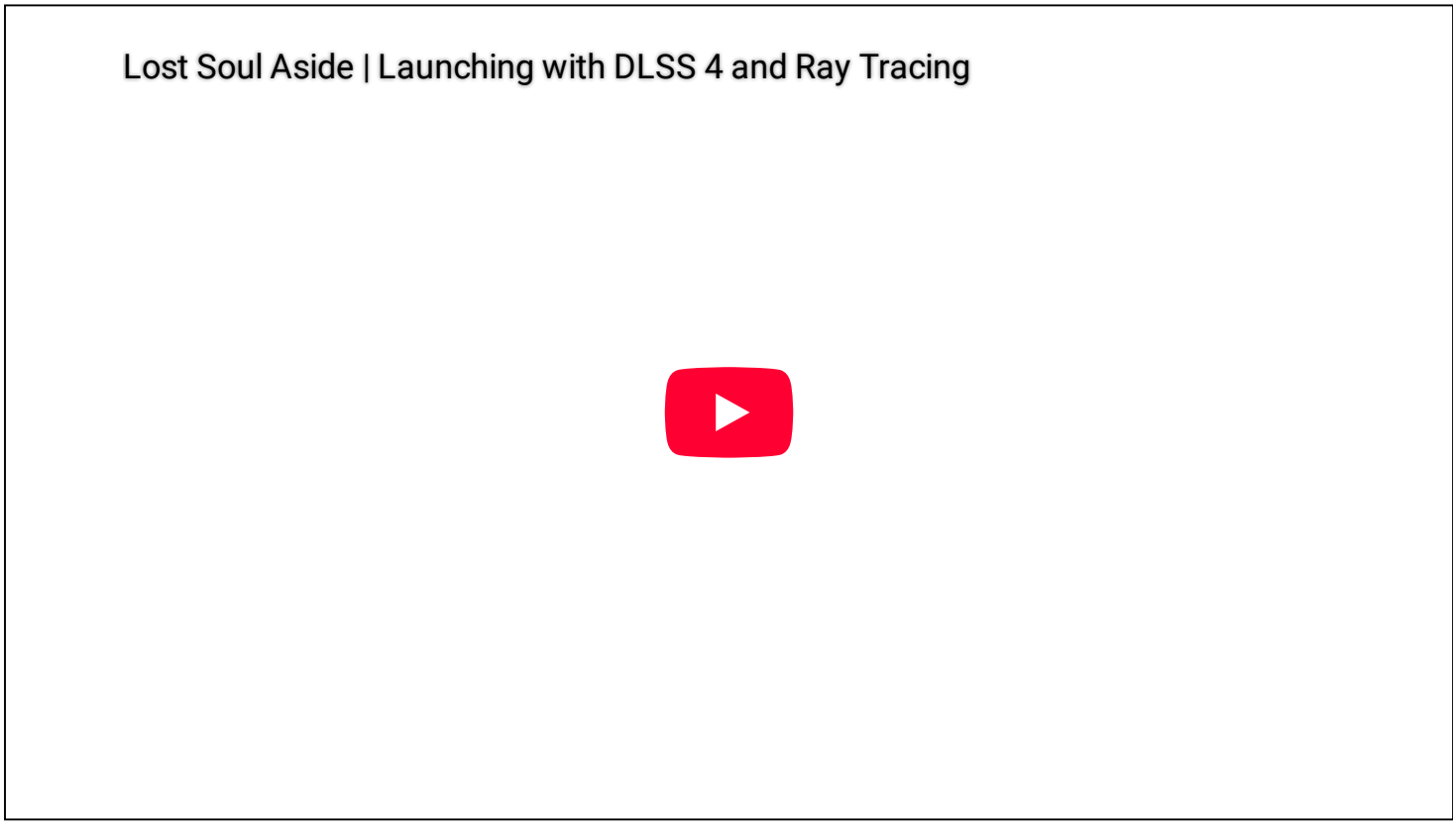
Over 100 DLSS 4 Games and Apps Out Now

DLSS 4 debuted with the release of GeForce RTX 50 Series GPUs. Over 100 games and apps now feature support for DLSS 4. This milestone has been reached two years quicker than with DLSS 3, making DLSS 4 the most rapidly adopted NVIDIA game technology of all time.

DLSS 4 introduced Multi Frame Generation, which uses AI to generate up to three additional frames per traditionally rendered frame, working with the complete suite of DLSS technologies to multiply frame rates by up to 8x over traditional brute-force rendering.

This massive performance improvement on GeForce RTX 50 Series graphics cards and laptops enables gamers to max out visuals at the highest resolutions and play at incredible frame rates.

In addition, *Lost Soul Aside*, *Mecha BREAK*, *Phantom Blade Zero*, *Stellar Blade*, *Tides of Annihilation* and *Wild Assault* will launch with DLSS 4, giving GeForce RTX gamers the definitive PC experience in each title. Learn more.

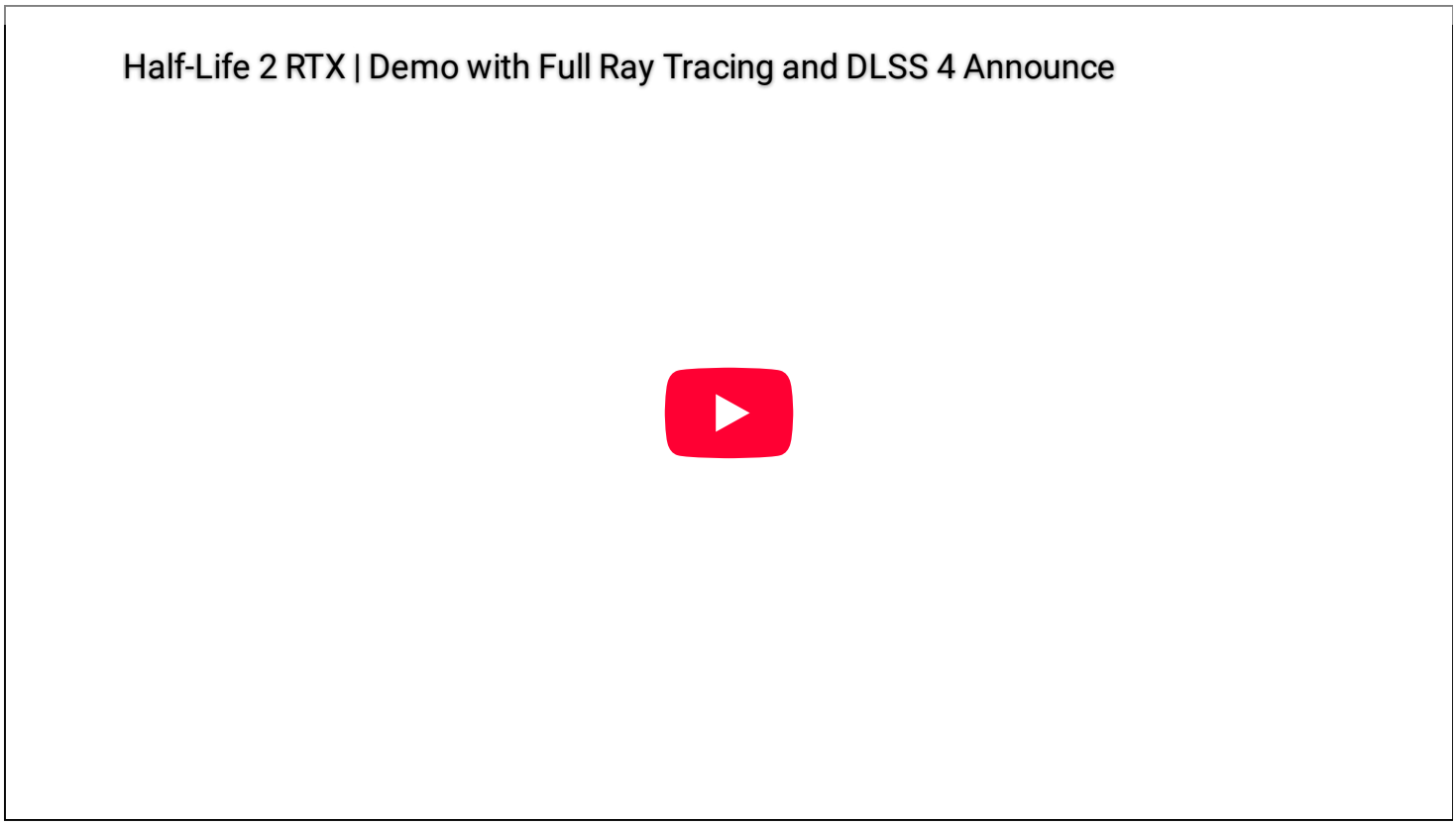


Developers can get started with DLSS 4 through the DLSS 4 Unreal Engine plug-in.

‘Half-Life 2 RTX’ Demo Launch, RTX Remix Official Release

Half-Life 2 RTX is a community-made remaster of the iconic first-person shooter *Half-Life 2.*

A playable *Half-Life 2 RTX* demo will be available on Tuesday, March 18, for free download from Steam for *Half-Life 2* owners. The demo showcases Orbifold Studios’ work in the eerily sensational maps of Ravenholm and Nova Prospekt, with significantly improved assets and textures, full ray tracing, DLSS 4 with Multi Frame Generation and RTX neural rendering technologies.



Half-Life 2 RTX was made possible by NVIDIA RTX Remix, an open-source platform officially released today for modders to create stunning RTX remasters of classic games.

Use the platform now to join the 30,000+ modders who've experimented with enhancing hundreds of classic titles since its beta release last year, enabling over 1 million gamers to experience astonishing ray-traced mods.

NVIDIA ACE Technologies Enhance Game Characters With AI

The NVIDIA ACE suite of RTX-accelerated digital human technologies brings game characters to life with generative AI.

NVIDIA ACE autonomous game characters add autonomous teammates, nonplayer characters (NPCs) and self-learning enemies to games, creating new narrative possibilities and enhancing player immersion.

ACE autonomous game characters are debuting in two titles this month:

In *inZOI*, "Smart Zoi" NPCs will respond more realistically and intelligently to their environment based on their personalities. The game launches with NVIDIA ACE-based characters on Friday, March 28.

And in *NARAKA: BLADEPOINT MOBILE PC VERSION*, on-device NVIDIA ACE-powered teammates will help players battle enemies, hunt for loot and fight for victory starting Thursday, March 27.

Developers can start building with ACE today.

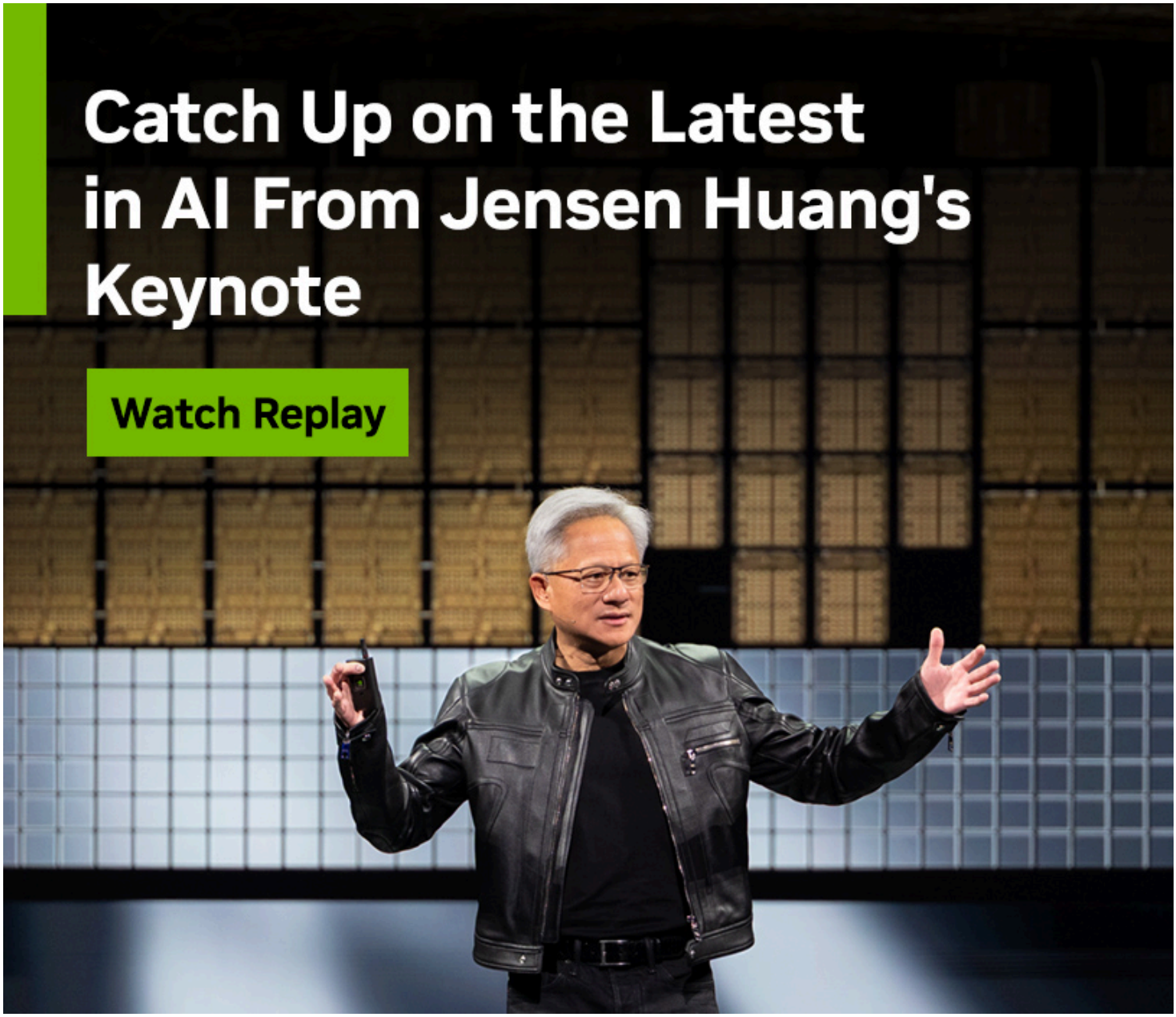
Join NVIDIA at GDC.

See notice regarding software product information.

Categories: Gaming | Generative AI

Tags: Artificial Intelligence | Creators | Events | Game Development | Gaming | GeForce | GPU | NVIDIA RTX | Open Source | Ray Tracing | Rendering | Visual Computing

Load Comments



All NVIDIA News

NVIDIA Research Showcases the Future of Robotics at RSS

Step Inside the Vault: The ‘Borderland’ Series Arrives on GeForce NOW

Tarun Patil Engineers Success on NVIDIA’s Circuit Silicon Correlation Team

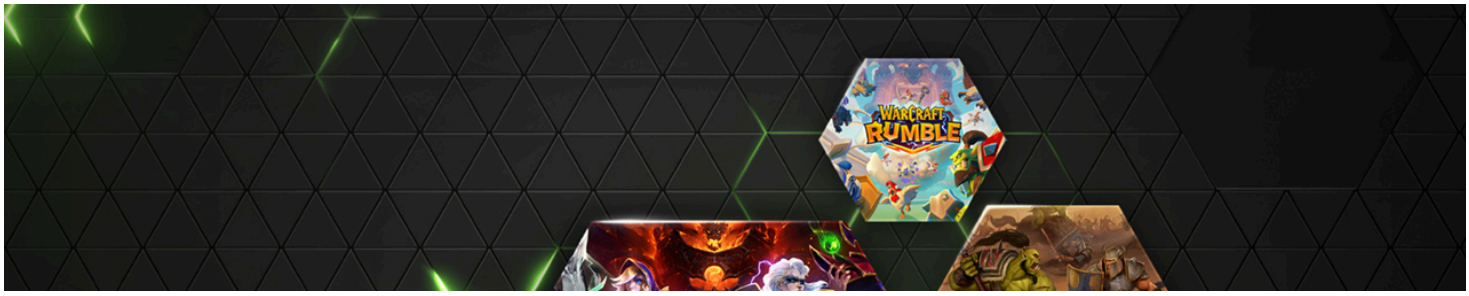
Plug and Play: Build a G-Assist Plug-In Today

Hexagon Taps NVIDIA Robotics and AI Software to Build and Deploy AEON, a New Humanoid

Relive the Magic as GeForce NOW Brings More Blizzard Gaming to the Cloud

Stream the 11 games joining the cloud, along with the latest update of ‘Zenless Zone Zero.’

March 13, 2025 by GeForce NOW Community



Share

Reading Time: 2 mins

Bundle up — GeForce NOW is bringing a flurry of Blizzard titles to its ever-expanding library.

Prepare to weather epic gameplay in the cloud, tackling the genres of real-time strategy (RTS), multiplayer online battle arena (MOBA) and more. Classic Blizzard titles join GeForce NOW, including *Heroes of the Storm, Warcraft Rumble* and three titles from the *Warcraft: Remastered* series.

They're all part of 11 games joining the cloud this week, atop the latest update for hit game *Zenless Zone Zero* from miHoYo.

Blizzard Heats Things Up

Heroes (and save data) never die in the cloud.

Heroes of the Storm, Blizzard's unique take on the MOBA genre, offers fast-paced team battles across diverse battlegrounds. The game features a roster of iconic Blizzard franchise characters, each with customizable talents and abilities. *Heroes of the Storm* emphasizes team-based gameplay with shared experiences and objectives, making it more accessible to newcomers while providing depth for experienced players.

The cloud is rumbling.

In *Warcraft Rumble*, a mobile action-strategy game set in the *Warcraft* universe, players collect and deploy miniature versions of the series' beloved characters. The game offers a blend of tower defense and RTS elements as players battle across various modes, including a single-player campaign, player vs. player matches and cooperative dungeons.

Old-school cool, new-school graphics.

The Warcraft Remastered collection gives the classic RTS titles a modern twist with updated visuals and quality-of-life improvements. *Warcraft: Remastered* and *Warcraft II: Remastered* offer enhanced graphics while maintaining the original gameplay, allowing players to toggle between classic and updated visuals. *Warcraft III: Reforged* includes new graphics options and multiplayer features. Both these remasters provide nostalgia for long-time fans and an ideal opportunity for new players to experience the iconic strategy games that shaped the genre.

New Games, No Wait

New agents, new adventures.

The popular *Zenless Zone Zero* gets its 1.6 update, "Among the Forgotten Ruins," now available for members to stream without waiting around for updates or downloads. This latest update brings three new playable agents: Soldier 0-Anby, Pulchra and Trigger. Players can explore two new areas, Port Elpis and Reverb Arena, as well as try out the "Hollow Zero-Lost Void" mode. The update also introduces a revamped Decibel system for more strategic gameplay.

Look for the following games available to stream in the cloud this week:

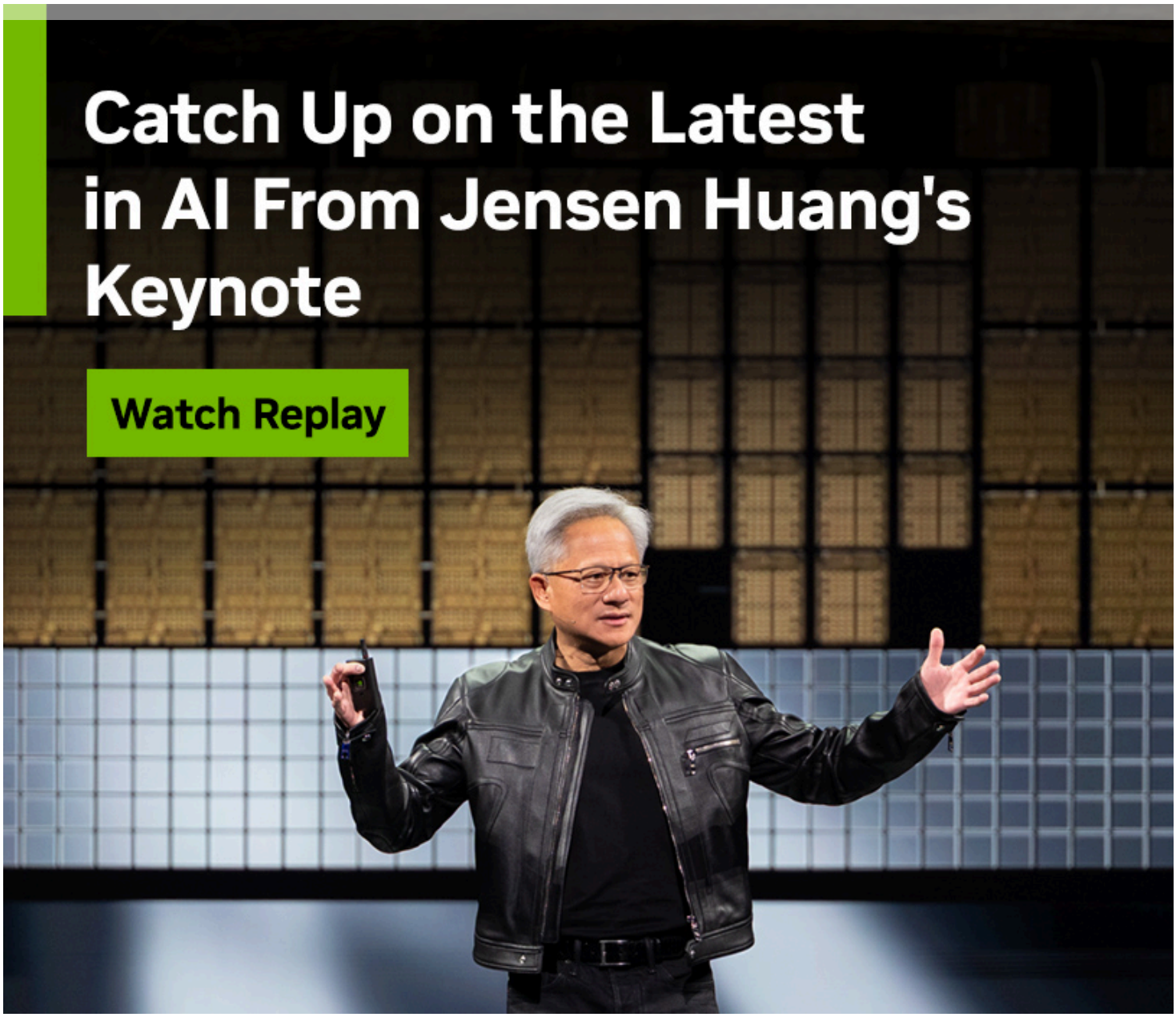
- *Citizen Sleeper 2: Starward Vector* (Xbox, available on PC Game Pass)
- *City Transport Simulator: Tram* (Steam)
- *Dave the Diver* (Steam)
- *Heroes of the Storm (Battle.net)*
- *Microtopia* (Steam)
- *Orcs Must Die Deathtrap* (Xbox, available on PC Game Pass)
- *Potion Craft: Alchemist Simulator* (Steam)
- *Warcraft I Remastered* (Battle.net)
- *Warcraft II Remastered* (Battle.net)
- *Warcraft III: Reforged* (Battle.net)
- *Warcraft Rumble* (Battle.net)

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

Categories: Gaming

Tags: Cloud Gaming | GeForce NOW

Load Comments



All NVIDIA News

NVIDIA Research Showcases the Future of Robotics at RSS

Step Inside the Vault: The ‘Borderland’ Series Arrives on GeForce NOW

Tarun Patil Engineers Success on NVIDIA’s Circuit Silicon Correlation Team

Plug and Play: Build a G-Assist Plug-In Today

Hexagon Taps NVIDIA Robotics and AI Software to Build and Deploy AEON, a New Humanoid