



NVIDIA Brings AI Assistants to Life With GeForce RTX AI PCs

Project G-Assist, NVIDIA ACE NIMs for Digital Humans, and Generative AI Tools Deliver Advanced AI Experiences on RTX Laptops; Plus, RTX-Accelerated APIs for Small Language Models Coming to Windows Copilot Runtime

COMPUTEX—NVIDIA today announced new [NVIDIA RTX™](#) technology to power AI assistants and digital humans running on new [GeForce RTX™](#) AI laptops.

NVIDIA unveiled Project G-Assist — an RTX-powered AI assistant technology demo that provides context-aware help for PC games and apps. The Project G-Assist tech demo debuted with *ARK: Survival Ascended* from Studio Wildcard. NVIDIA also introduced the first PC-based NVIDIA NIM™ inference microservices for the [NVIDIA ACE](#) digital human platform.

These technologies are enabled by the [NVIDIA RTX AI Toolkit](#), a new suite of tools and software development kits that aid developers in optimizing and deploying large generative AI models on Windows PCs. They join NVIDIA's full-stack RTX AI innovations accelerating over 500 PC applications and games and 200 laptop designs from manufacturers.

In addition, newly announced RTX AI PC laptops from ASUS and MSI feature up to GeForce RTX 4070 GPUs and power-efficient systems-on-a-chip with Windows 11 AI PC capabilities. These Windows 11 AI PCs will receive a free update to Copilot+ PC experiences when available.

"NVIDIA launched the era of AI PCs in 2018 with the release of RTX Tensor Core GPUs and NVIDIA DLSS," said Jason Paul, vice president of consumer AI at NVIDIA. "Now, with Project G-Assist and NVIDIA ACE, we're unlocking the next generation of AI-powered experiences for over 100 million RTX AI PC users."

Project G-Assist, a GeForce AI Assistant

AI assistants are set to transform gaming and in-app experiences — from offering gaming strategies and analyzing multiplayer replays to assisting with complex creative workflows. Project G-Assist is a glimpse into this future.

PC games offer vast universes to explore and intricate mechanics to master, which are challenging and time-consuming feats even for the most dedicated gamers. Project G-Assist aims to put game knowledge at players' fingertips using generative AI.

Project G-Assist takes voice or text inputs from the player, along with contextual information from the game screen, and runs the data through AI vision models. These models enhance the contextual awareness and app-specific understanding of a large language model (LLM) linked to a game knowledge database, and then generate a tailored response delivered as text or speech.

NVIDIA partnered with Studio Wildcard to demo the technology with *ARK: Survival Ascended*. Project G-Assist can help answer questions about creatures, items, lore, objectives, difficult bosses and more. Because Project G-Assist is context-aware, it personalizes its responses to the player's game session.

In addition, Project G-Assist can configure the player's gaming system for optimal performance and efficiency. It can provide insights into performance metrics, optimize graphics settings depending on the user's hardware, apply a safe overclock and even intelligently reduce power consumption while maintaining a performance target.

First ACE PC NIM Debuts

[NVIDIA ACE technology](#) for powering digital humans is now coming to RTX AI PCs and workstations with NVIDIA NIM — inference microservices that enable developers to reduce deployment times from weeks to minutes. ACE NIM microservices deliver high-quality inference running locally on devices for natural language understanding, speech synthesis, facial animation and more.

At COMPUTEX, the gaming debut of NVIDIA ACE NIM on the PC will be featured in the [Covert Protocol tech demo](#), developed in collaboration with Inworld AI. It now showcases [NVIDIA Audio2Face™](#) and [NVIDIA Riva](#) automatic speech recognition running locally on devices.

Windows Copilot Runtime to Add GPU Acceleration for Local PC SLMs

Microsoft and NVIDIA are collaborating to help developers bring new generative AI capabilities to their Windows native and web apps. This collaboration will provide application developers with easy application programming interface (API) access to GPU-accelerated small language models (SLMs) that enable [retrieval-augmented generation](#) (RAG) capabilities that run on-device as part of Windows Copilot Runtime.

SLMs provide tremendous possibilities for Windows developers, including content summarization, content generation and task automation. RAG capabilities augment SLMs by giving the AI models access to domain-specific information not well represented in base models. RAG APIs enable developers to harness application-specific data sources and tune SLM behavior and capabilities to application needs.

These AI capabilities will be accelerated by NVIDIA RTX GPUs, as well as AI accelerators from other hardware vendors, providing end users with fast, responsive AI experiences across the breadth of the Windows ecosystem.

The API will be released in developer preview later this year.

4x Faster, 3x Smaller Models With the RTX AI Toolkit

The AI ecosystem has built hundreds of thousands of open-source models for app developers to leverage, but most models are pretrained for general purposes and built to run in a data center.

To help developers build application-specific AI models that run on PCs, NVIDIA is introducing [RTX AI Toolkit](#) — a suite of tools and SDKs for model customization, optimization and deployment on RTX AI PCs. RTX AI Toolkit will be available later this month for broader developer access.

Developers can customize a pretrained model with open-source QLoRa tools. Then, they can use the [NVIDIA TensorRT™](#) model optimizer to quantize models to consume up to 3x less RAM. NVIDIA TensorRT Cloud then optimizes the model for peak performance across the RTX GPU lineups. The result is up to 4x faster performance compared with the pretrained model.

The new [NVIDIA AI Inference Manager](#) SDK, now available in early access, simplifies the deployment of ACE to PCs. It preconfigures the PC with the necessary AI models, engines and dependencies while orchestrating AI inference seamlessly across PCs and the cloud.

Software partners such as Adobe, Blackmagic Design and Topaz are integrating components of the RTX AI Toolkit within their popular creative apps to accelerate AI performance on RTX PCs.

“Adobe and NVIDIA continue to collaborate to deliver breakthrough customer experiences across all creative workflows, from video to imaging, design, 3D and beyond,” said Deepa Subramaniam, vice president of product marketing, Creative Cloud at Adobe. “TensorRT 10.0 on RTX PCs delivers unprecedented performance and AI-powered capabilities for creators, designers and developers, unlocking new creative possibilities for content creation in industry-leading creative tools like Photoshop.”

Components of the RTX AI Toolkit, such as TensorRT-LLM, are integrated in popular developer frameworks and applications for generative AI, including Automatic1111, ComfyUI, Jan.AI, LangChain, LlamaIndex, Oobabooga and Sanctum.AI.

AI for Content Creation

NVIDIA is also integrating RTX AI acceleration into apps for creators, modders and video enthusiasts.

Last year, NVIDIA introduced RTX acceleration using TensorRT for one of the most popular Stable Diffusion user interfaces, Automatic1111. Starting this week, RTX will also accelerate the highly popular ComfyUI, delivering up to a 60% improvement in performance over the currently shipping version, and 7x faster performance compared with the MacBook Pro M3 Max.

[NVIDIA RTX Remix](#) is a modding platform for remastering classic DirectX 8 and DirectX 9 games with full ray tracing, NVIDIA DLSS 3.5 and physically accurate materials. RTX Remix includes a runtime renderer and the RTX Remix Toolkit app, which facilitates the modding of game assets and materials.

Last year, NVIDIA made RTX Remix Runtime open source, allowing modders to [expand game compatibility](#) and advance rendering capabilities.

Since RTX Remix Toolkit launched earlier this year, 20,000 modders have used it to mod [classic games](#), resulting in over 100 RTX remasters in development on the [RTX Remix Showcase Discord](#).

This month, NVIDIA will make the RTX Remix Toolkit open source, allowing modders to streamline how assets are replaced and scenes are relit, increase supported file formats for RTX Remix’s asset ingestor and bolster RTX Remix’s AI Texture Tools with new models.

In addition, NVIDIA is making the capabilities of RTX Remix Toolkit accessible via a 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. NVIDIA is also providing an SDK for RTX Remix Runtime to allow modders to deploy RTX Remix’s renderer into other applications and games beyond DirectX 8 and 9 classics.

With more of the RTX Remix platform being made open source, modders across the globe can build even more stunning RTX remasters.

[NVIDIA RTX Video](#), the popular AI-powered super-resolution feature supported in the Google Chrome, Microsoft Edge and Mozilla Firefox browsers, is now available as an SDK to all developers, helping them natively integrate AI for upscaling, sharpening, compression artifact reduction and high-dynamic range (HDR) conversion.

Coming soon to video editing software Blackmagic Design's DaVinci Resolve and Wondershare Filmora, RTX Video will enable video editors to upscale lower-quality video files to 4K resolution, as well as convert standard dynamic range source files into HDR. In addition, the free media player VLC media will soon add RTX Video HDR to its existing super-resolution capability.

Learn more about RTX AI PCs and technology by joining [NVIDIA at COMPUTEX](#).

About NVIDIA

[NVIDIA](#) (NASDAQ: NVDA) is the world leader in accelerated computing.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, and availability of our products, services, and technologies, including NVIDIA RTX technology, GeForce RTX AI laptops, Project G-Assist, NVIDIA NIM inference microservices, NVIDIA ACE digital human platform, NVIDIA RTX AI Toolkit, GeForce RTX 4070 GPUs, RTX Tensor Core GPUs, DLSS, NVIDIA Audio2Face, NVIDIA Riva, NVIDIA TensorRT, NVIDIA AI Inference Manager, NVIDIA RTX Remix, NVIDIA DLSS 3.5, RTX Remix Runtime, and NVIDIA RTX Video; the benefits and impact of NVIDIA's collaboration with third parties, and the features and availability of their services and offerings; third parties using or adopting NVIDIA's products or technologies and the benefits thereof; RAG APIs enabling developers to harness application-specific data sources and tune SLM behavior and capabilities to application needs; and NVIDIA unlocking the next generation of AI-powered experiences for over 100 million RTX AI PC users, are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features, or functions set forth herein.

© 2024 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, Audio2Face, GeForce RTX, NVIDIA NIM, NVIDIA RTX and TensorRT are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

Jordan Dodge
SHIELD, GeForce NOW
NVIDIA Corp.
+1-408-506-6849
jdodge@nvidia.com