Nebius acquires Inferize to strengthen Nebius Token Factory's production inference stack

- Brings technology that cuts time to launch and scale large AI models, improving token economics and capacity utilization
- Adds deeply experienced systems engineering talent to Nebius's inference team
- Continues production inference build-out on Nebius with optimizations across the stack

Amsterdam, October 1, 2026 — Nebius (Nasdaq: NBIS), the AI cloud company, today announced it has acquired Inferize, an inference optimization company whose technology shortens the time needed to launch and scale large AI models, making inference workloads elastic.

Inferize's technology and team have joined Nebius Token Factory, Nebius's managed inference platform for production AI.

Cold starts – the time models need to load before serving a single request – impose additional time and costs on teams running AI models in production at scale. This leaves assigned GPUs idle at launch time, when demand spikes and new instances spin up, and when weights are updated mid-run for example during reinforcement learning – and forces platforms to hold spare capacity just to hit their service-level targets.

Inferize's technology cuts this “idle GPU tax”, enabling capacity to scale much more closely with actual usage and driving higher capacity utilization and better token economics.

Danila Shtan, Chief Technology Officer of Nebius, said:

> “Running inference well takes more than fast GPUs and optimized models. The whole system needs to respond when demand changes, including how quickly additional capacity is ready to serve customers. Inferize brings technology that accelerates that process and a team with deep expertise in GPU systems. We’re bringing both into Token Factory to make it more responsive to customer demand and get more useful work out of our infrastructure. The team’s contribution will extend well beyond this first integration.”

Inferize adds another layer to how Nebius runs production inference. Eigen AI brought optimization at the model, kernel and system levels into Nebius Token Factory, while Clarifai's core team and licensed technology added system-level inference and compute orchestration.

Guy Bortnikov, co-founder and CEO of Inferize, said:

> "Keeping spare GPUs running is the price of being ready for demand. Removing that cost is what we built Inferize to do, and Nebius is where it can go straight into the platform. Our team will work across the stack with one objective: serving more customer demand from every GPU."

Founded in January 2026, the Inferize team had a working prototype within three months. Inferize's engineers will work across Nebius Token Factory, starting with the integration of their technology.

The terms of the transaction were not disclosed.

About Nebius

Nebius, the AI cloud company, is building the full-stack platform for developers and companies to take charge of their AI future — from data and model training to production deployment. Founded on deep in-house technological expertise and operating at scale with a rapidly expanding global footprint, Nebius serves startups and enterprises building AI products, agents, and services worldwide.

Nebius is listed on Nasdaq (Nasdaq: NBIS) and headquartered in Amsterdam.

For more information please visit nebius.com.
Media kit nebius.com/media-kit.

Contacts

Media relations
media@nebius.com

Investor relations
askIR@nebius.com

Disclaimer

Forward-looking statements

This press release contains forward-looking statements within the meaning of the Private Securities Litigation Reform Act of 1995, which involve risks and uncertainties. All statements contained in this press release other than statements of historical fact, including, without limitation, statements regarding our integration of the Inferize team and technology; our future financial and business performance, strategy, expected growth, planned investments and capital expenditures, capacity expansion plans, anticipated future financing transactions and expected financial results, are forward-looking statements. The words "anticipate," "believe," "continue," "estimate," "expect," "guide," "intend," "likely," "may," "will" and similar expressions and their negatives are intended to identify forward-looking statements.

These forward-looking statements are subject to risks, uncertainties and assumptions, some of which are beyond our control. Actual results may differ materially from the results predicted or implied by such statements, and our reported results should not be considered as an indication of future performance. The potential risks and uncertainties that could cause actual results to differ from the results predicted or implied by such statements include, among others: risks inherent in the integration of acquired businesses; market, macroeconomic and geopolitical conditions; our ability to build, operate and manage our businesses to the desired scale; competitive pressures; technological developments; our ability to secure and retain clients; our ability to secure additional capital to enable the growth of the business; unpredictable sales cycles; and potential pricing pressures; as well as those risks and uncertainties related to our continuing businesses included under the captions "Risk Factors" and "Operating and Financial Review and Prospects" in our Annual Report on Form 20-F for the year ended December 31, 2025, filed with the SEC on April 30, 2026, which is available on our investor relations website at https://nebius.com/investor-hub and on the SEC website at www.sec.gov.

All information in this press release is as of the date hereof (unless stated otherwise). Except as required by law, we undertake no obligation to update or revise publicly any forward-looking statements, whether as a result of new information, future events or otherwise, after the date on which the statements are made or to reflect the occurrence of unanticipated events.

In addition, statements that "we believe" and similar statements reflect our beliefs and opinions on the relevant subject. These statements are based upon information available to us as of the date hereof and, while we believe such information forms a reasonable basis for such statements, such information may be limited or incomplete, and our statements should not be read to indicate that we have conducted an exhaustive inquiry into, or review of, all potentially available relevant information. These statements are inherently uncertain, and investors are cautioned not to unduly rely upon these statements.