NEBIUS

The ultimate AI cloud for media innovators

One partner for your entire AI lifecycle, from secure model training to low-latency inference, backed by deep media expertise.

Learn more

Full speed, zero setbacks

Train at the speed the hardware allows. Automated fault detection and checkpoint recovery make a failure cost minutes, not days.

Economics that scale with you

Scale up in minutes for a launch and back down after. Serving costs track what you use, not what you provisioned to be safe.

A co-engineering partner

Our engineers work on your side, down to the InfiniBand config and the NCCL patch. We ship no competing media models.

Behind the defining models

The model makers defining what comes next already build here, from Higgsfield and Recraft to Black Forest Labs.

Sovereign infrastructure, your rules

Your data stays in-region, with the compliance and content-security posture the biggest studio deals gate on.

Purpose-built for media AI

A full-stack AI cloud tuned for high-volume media training and inference, not a general-purpose box.

Production-ready inference

Scale open models from prototype to millions of generations a day on private, high-availability endpoints.

NEBIUS TOKEN FACTORY

Engineered with NVIDIA

Reference-platform partner on the AI factories powering large-scale inference and agentic media workflows.

NEBIUS | NVIDIA

Higgsfield

Training at hypergrowth scale

Scaled from zero to 15M+ users in nine months, while training a multi-billion-parameter diffusion model on NVIDIA HGX B200, with no service disruption.

$200M
run rate in nine months

4.5M+
generations per day

<1 hr
to first training run

"Nebius felt like a co-engineer, not just a cloud provider."

Alex Mashrabov
CEO
Higgsfield AI

RECRAFT

A 20B-parameter model for designers

When training ran 8x slower, Nebius configured GPU Direct RDMA and a patched NCCL build, reaching 4x network speedup and roughly 6x faster training.

20B
parameters (8 × SDXL)

~6x
faster training

>50%
PartiPrompts win

Higgsfield · RECRAFT · World Labs · Black Forest Labs · Photoroom · REACTOR · Microsoft · Meta · Dubformer · krisp · Pika · wubble