NVIDIA

NTRsync2025

Emerging Trends in HPC and AI

Andy Grant, EMEA Director – HPC & AI, Higher Education and Research

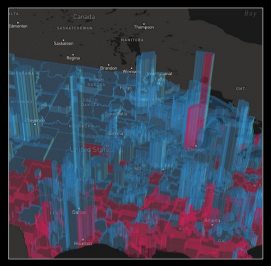
26th March 2025

Agenda

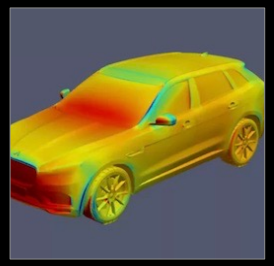
- Advanced Digital Twins
- Generative AI – Moving to Production
- Managing Complexity: Nvidia Inferencing Microservices (NIM)
- Agentic AI and NIM Blueprints
- Questions

NVIDIA AI Accelerated Computing Platform

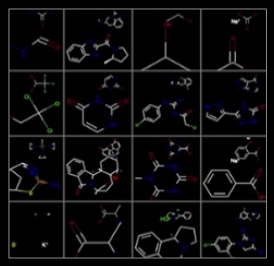
Hardware and Software Acceleration Across Every Workload and Vertical



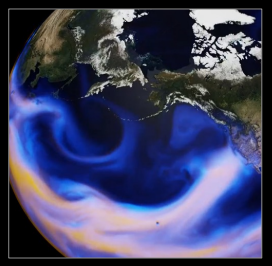
Data Processing



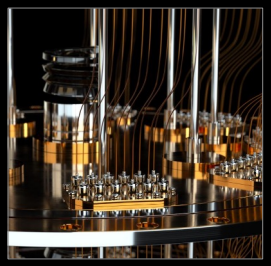
CAD, CAE, SDA



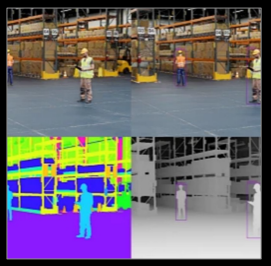
Computer-aided Drug Design



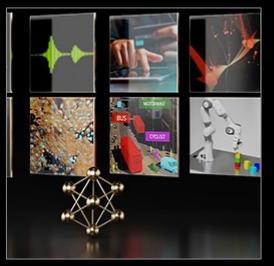
Climate Simulation



Quantum Simulation



Robotics & Industrial Digital Twins

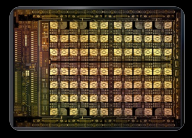


Enterprise AI

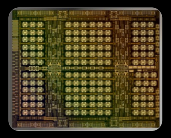
NVIDIA AI Enterprise

OMNIVERSE

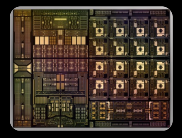
CUDA-X Libraries



CPU

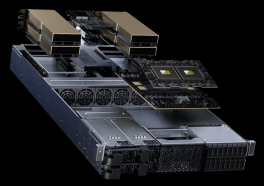


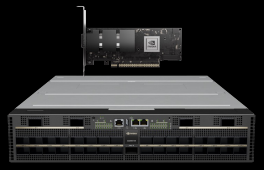
GPU



DPU

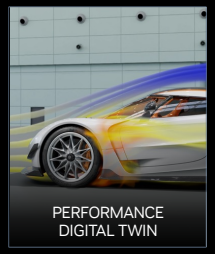
Accelerated Computing





The World's Largest Industries are Racing to Digitalize

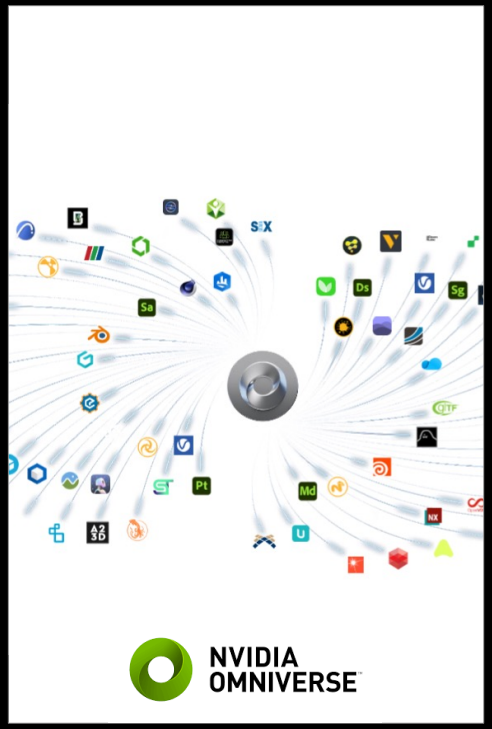
The era of AI-enabled digital twins has arrived



NVIDIA OMNIVERSE™

Key Technologies are Needed to Accelerate Digitalization

Enabling enterprises to unlock new possibilities and drive growth, productivity, and sustainability



OpenUSD

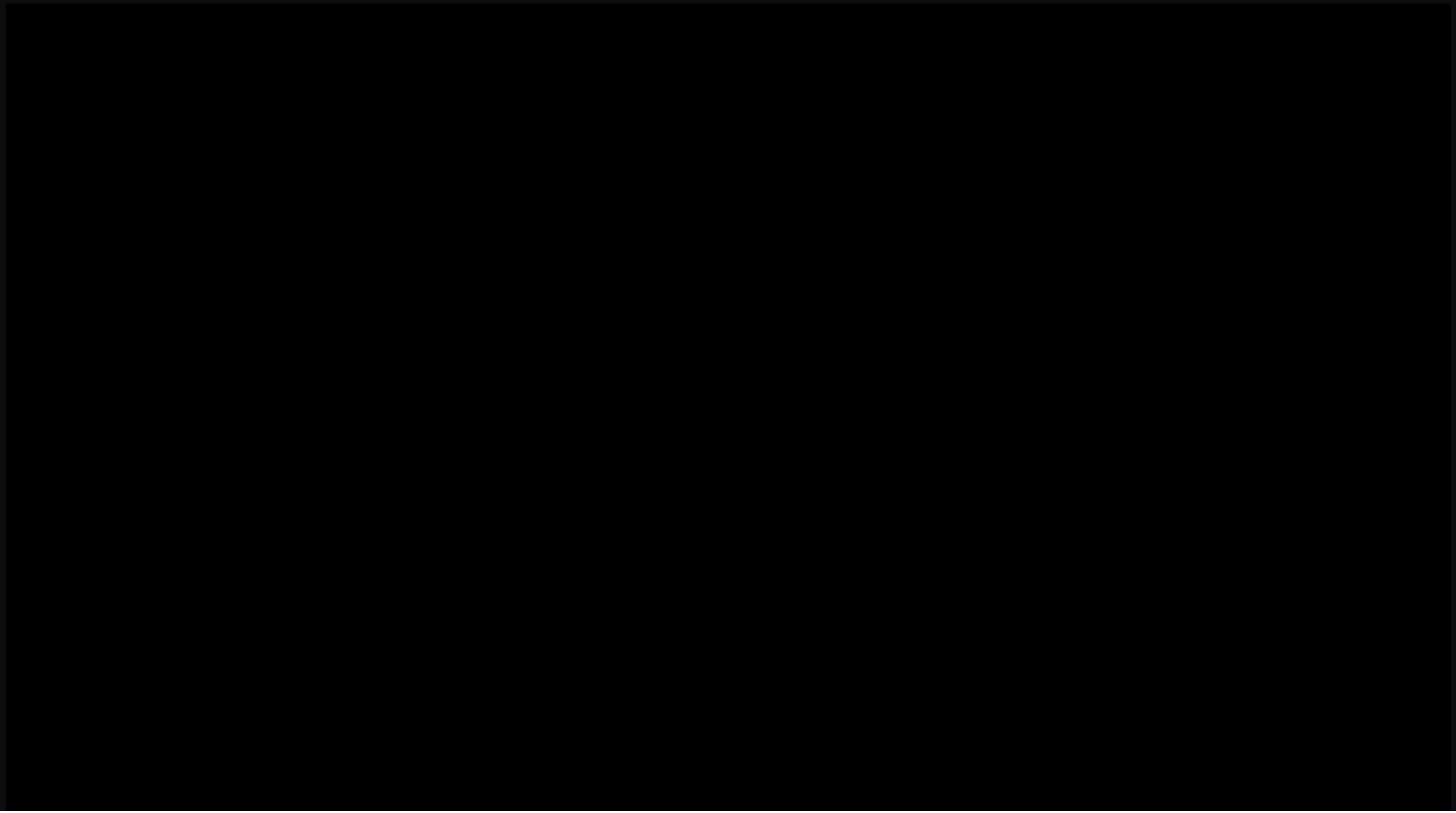
RTX

Accelerated Computing

Generative AI,
Computer Vision

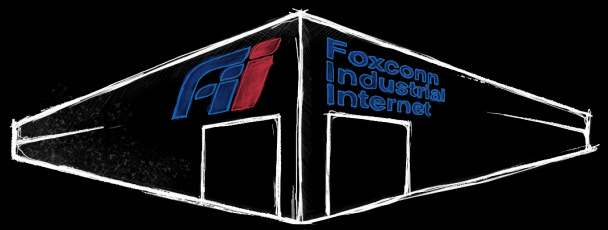
https://www.nvidia.com/en-us/omniverse/

Generative AI breeding a new generation of advanced digital twins



Emergence of Digital Twins

FACTORY ECOSYSTEM



INTEGRATOR FOXCONN

EDGE AI AND ROBOTS

ASM
EPSON
FARobot
JUSDA
kurtz ersa
TM ROBOT
TRI innovation
ViTrox

SOFTWARE

Ansys
ASPEED
AUTODESK
cadence
creo
FlexSim An AUTODESK Company
SIEMENS
SketchUp
VISUAL COMPONENTS

ODMS

Innoconn
ingrasys
SIEMENS

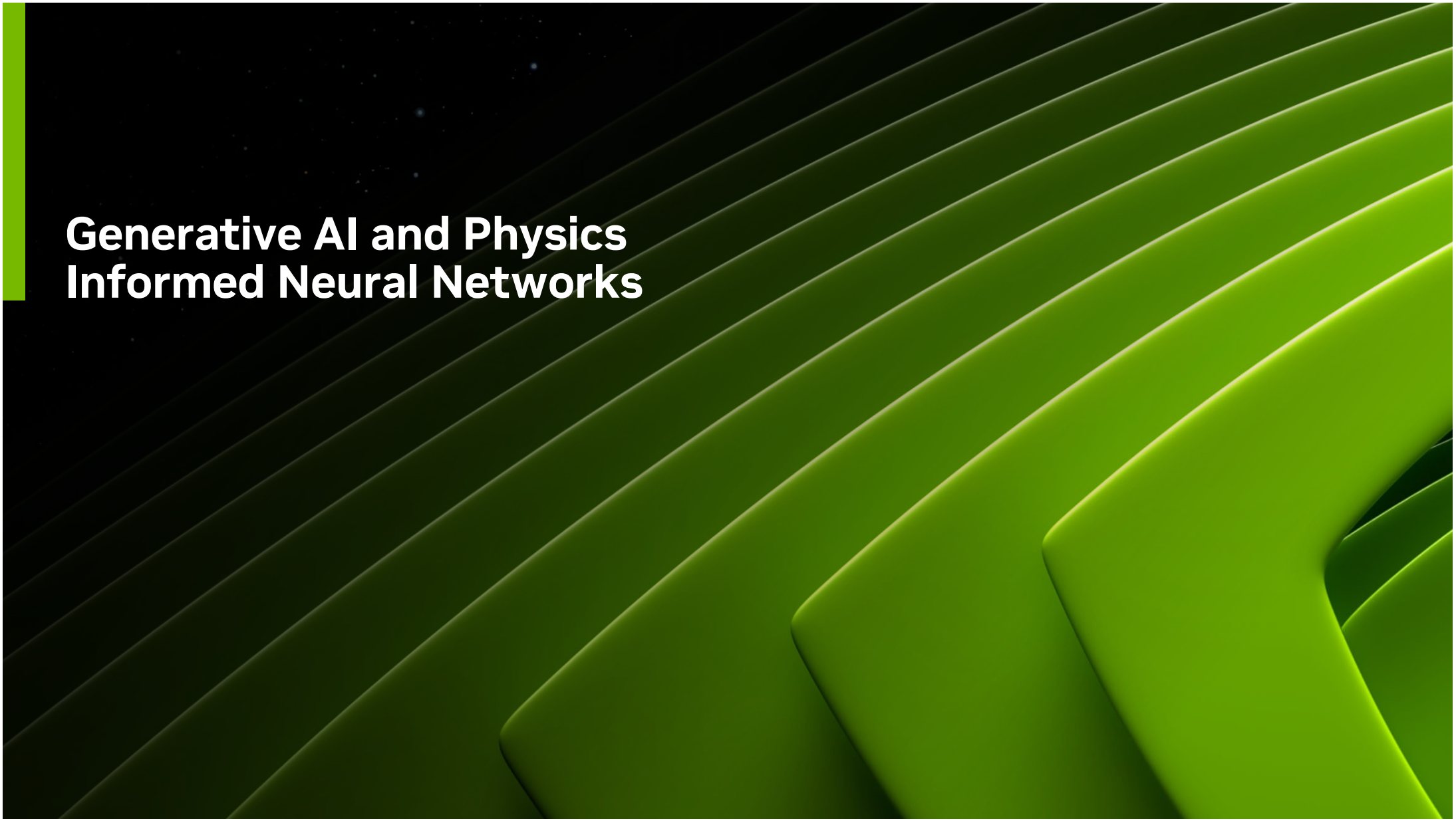
ISAAC DRIVE HOLOSCAN METROPOLIS

NVIDIA OMNIVERSE

NVIDIA


FANUC Robot CRX-20iA/L
FANUC Robot CRX-20iA/L
FANUC Robot CRX-20iA/L
FOXCONN

Generative AI and Physics Informed Neural Networks



From Predictive AI to Agentic and Physical AI

Predictive AI

Enables systems to observe and understand their surroundings but does not create or decide independently

Ex: Computer Vision such as facial recognition, object detection.

Generative AI

Produces novel outputs that are coherent and contextually relevant, often augmenting or assisting human creativity

Ex: ChatGPT, customer service chatbots

Agentic AI

Acts independently to achieve specified objectives, often adapting to new information dynamically.

Ex: Virtual Agents capable of managing schedules, making purchases, or solving problems

Physical AI

Integrates AI algorithms with physical systems, enabling machines to interact with and adapt to the real world.

Ex. Humanoid robots capable of undertaking a sequence of complex tasks, autonomous vehicles

.

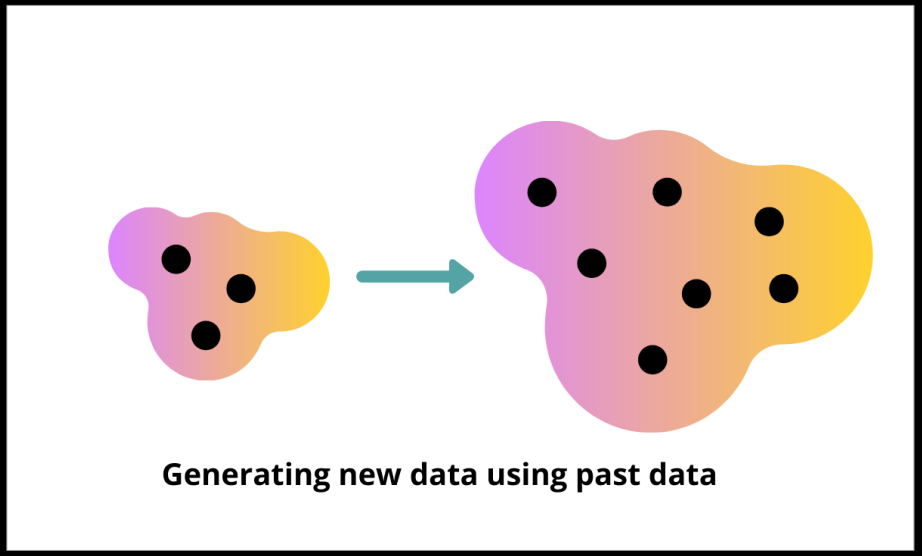
What is Generative AI?

Tips to Get Started

Generative AI refers to machine learning algorithms that enable computers to use existing or past content like text, audio and video files, images, and even code to generate new possible content.

The main idea is to generate completely original artifacts that would look like the real deal.

Often using Large Language Models (LLMs)



Generating new data using past data

An LLM is a Deep Neural Network

Map from “all previous words” to “next word”

A few thousand previous words for context

Predict the next word or group of words

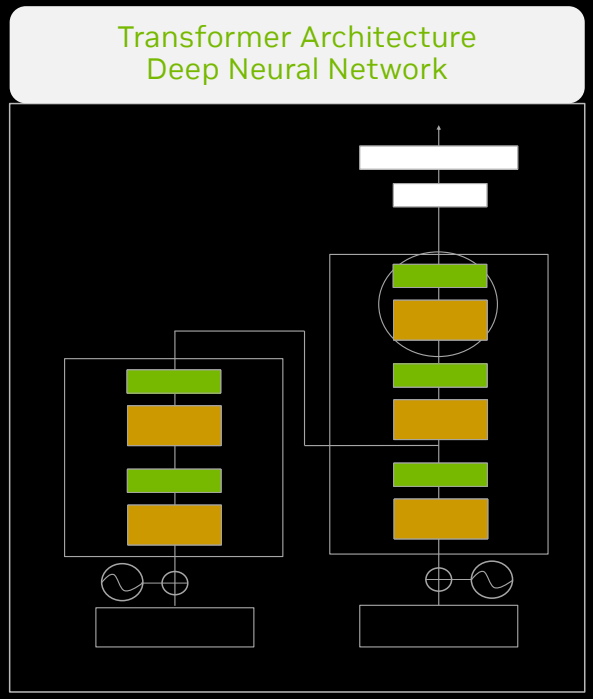
Through hard work, he supported himself and his •••

Because it crossed state lines, that criminal behavior attracted the attention of the •••

Joe Biden, who in 2011 was the •••

```
// loop over the string
int i;
for (i = 0; i < •••
```

This restaurant was fabulous! My star rating is •••



“family”

“*FBI*”

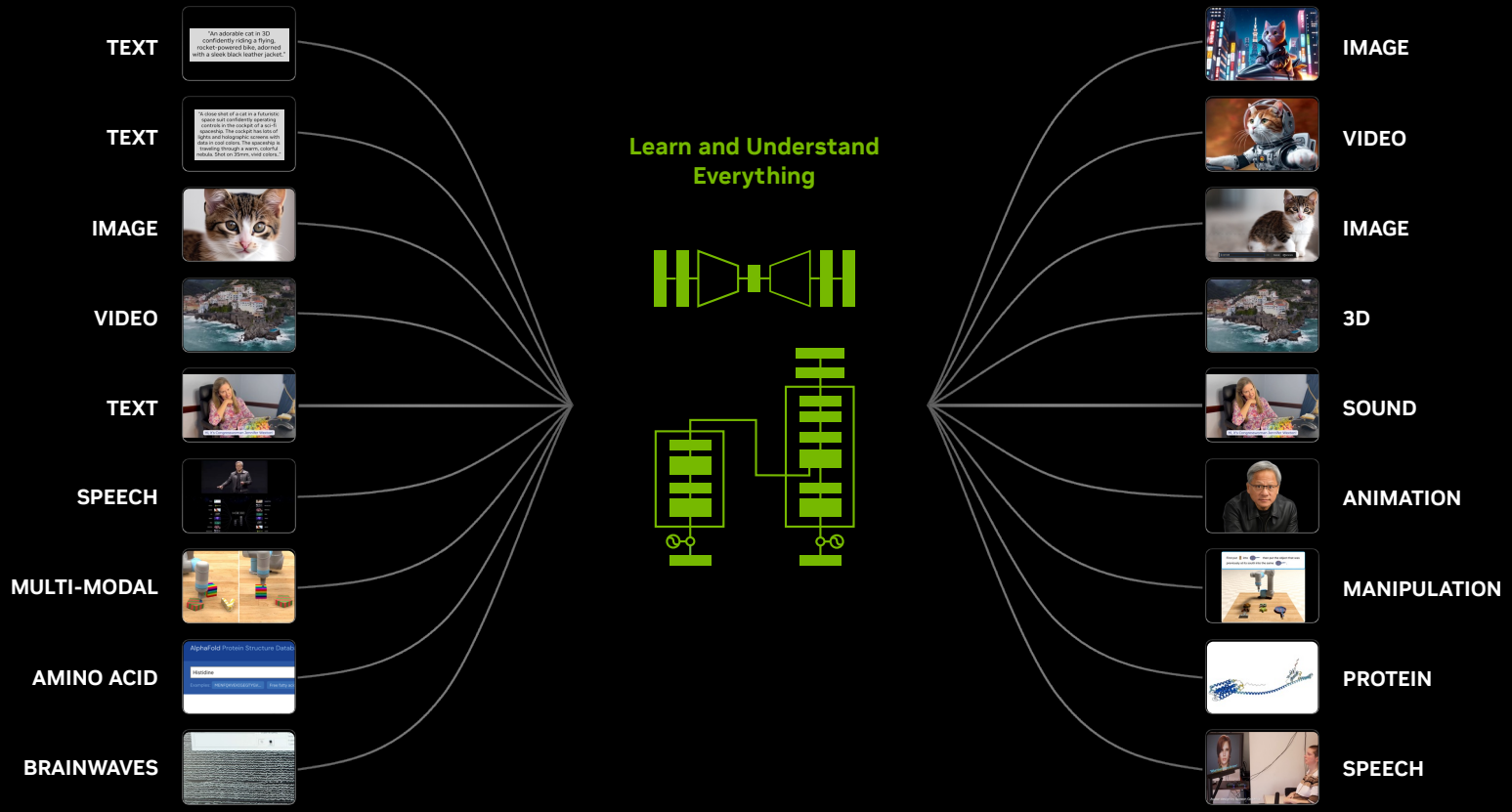
“Vice”

“strlen”

“five”

Generative AI

The most important technology of our generation



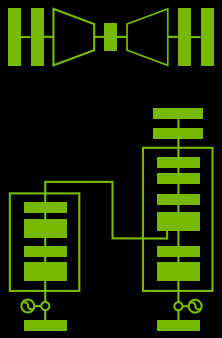
Generative AI

The most important technology of our generation

"A close shot of a cat in a futuristic space suit confidently operating controls in the cockpit of a sci-fi spaceship. The cockpit has lots of lights and holographic screens with data in cool colors. The spaceship is traveling through a warm, colorful nebula. Shot on 35mm, vivid colors.."

TEXT

Learn and Understand Everything



VIDEO

NVIDIA

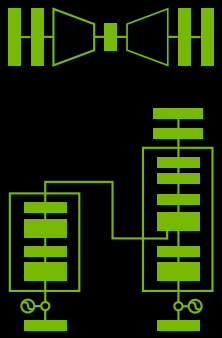
Generative AI

The most important technology of our generation

"A close shot of a cat in a futuristic space suit confidently operating controls in the cockpit of a sci-fi spaceship. The cockpit has lots of lights and holographic screens with data in cool colors. The spaceship is traveling through a warm, colorful nebula. Shot on 35mm, vivid colors.."

TEXT

Learn and Understand Everything

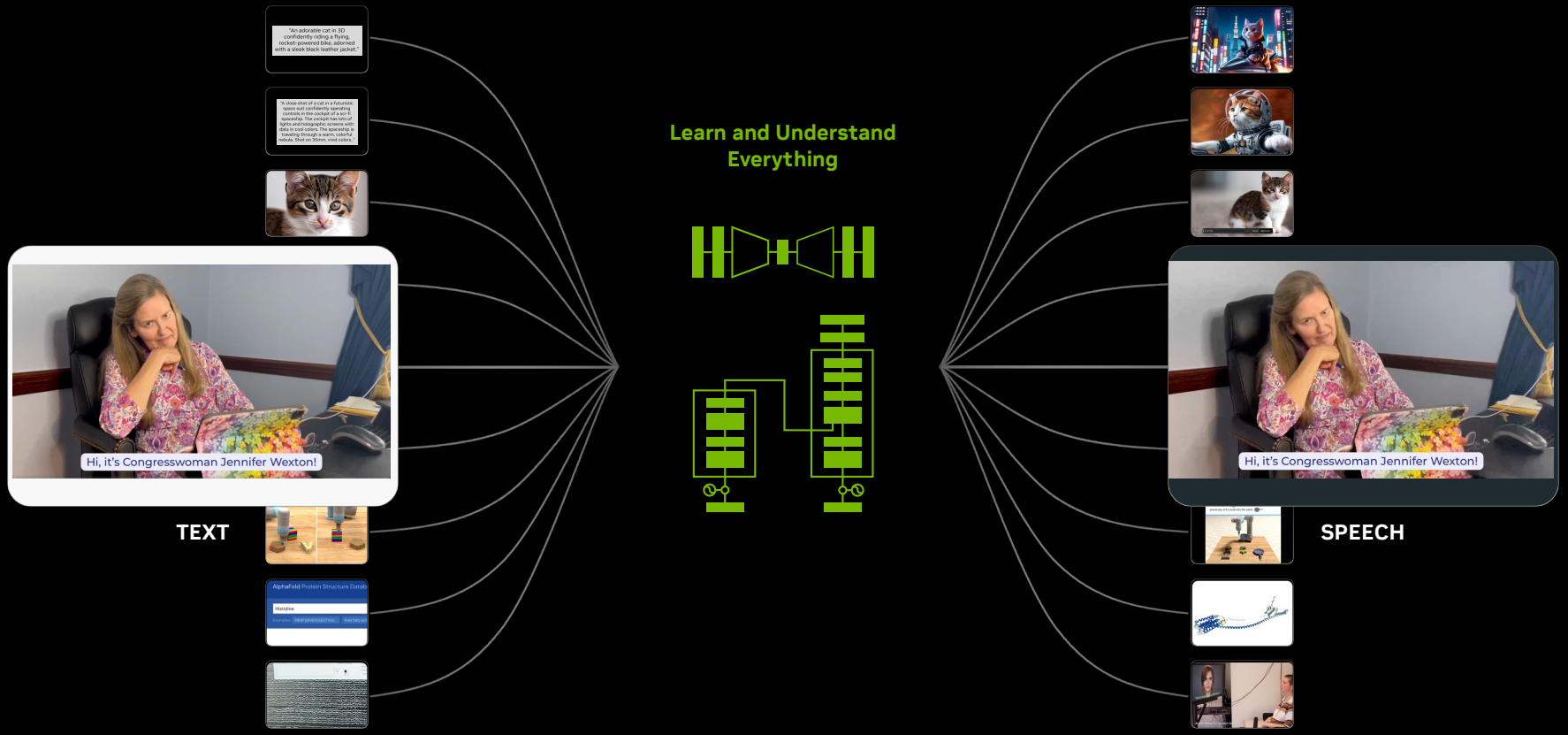


VIDEO

NVIDIA

Generative AI

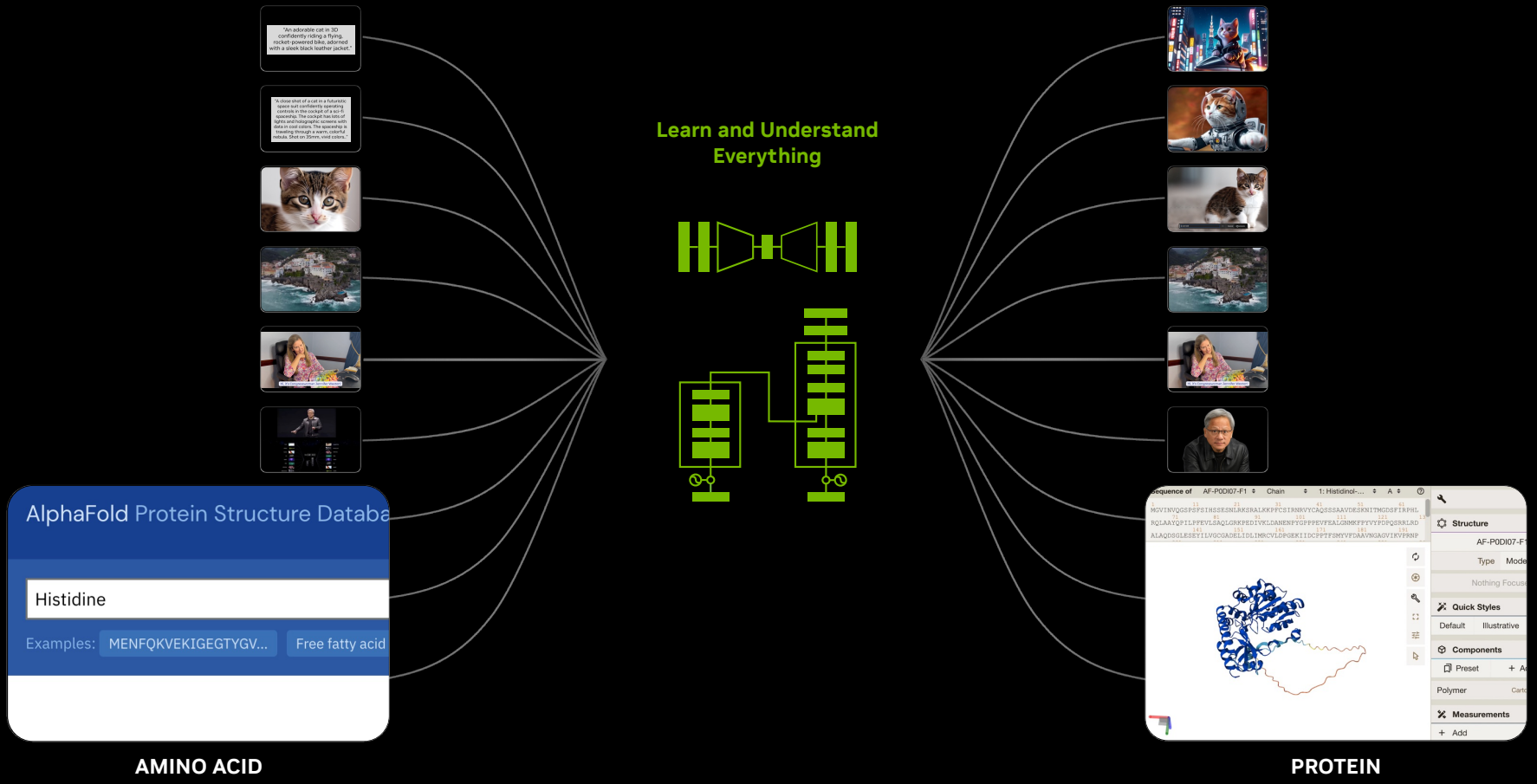
The most important technology of our generation



NVIDIA

Generative AI

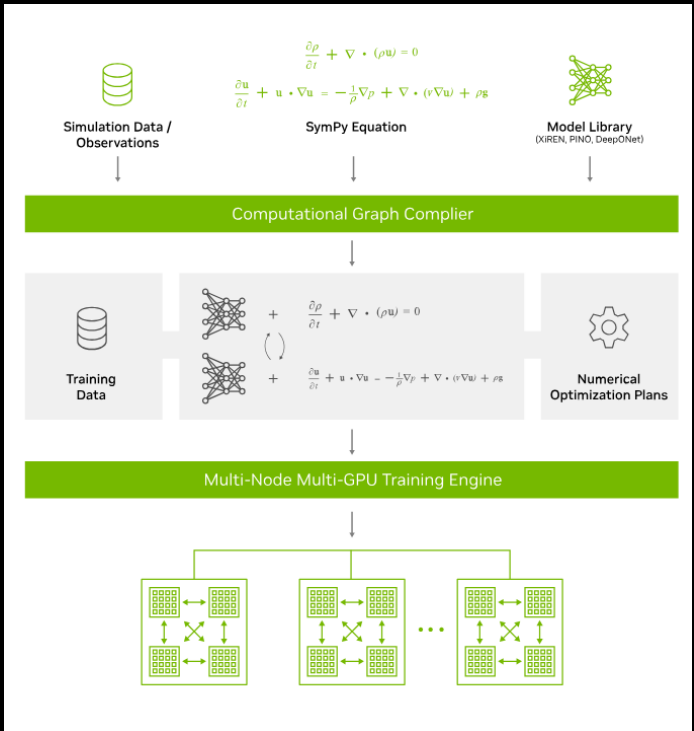
The most important technology of our generation



Physics-informed Neural Networks

NVIDIA Modulus: Framework for Developing Physics Machine Learning Neural Network Models

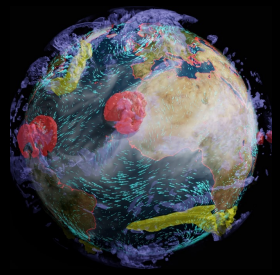
TRAINING NEURAL NETWORKS USING BOTH DATA AND THE GOVERNING EQUATIONS



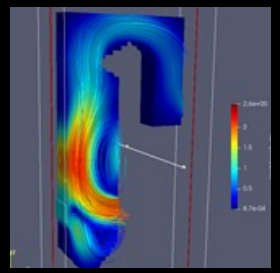
ADVANCING SCIENTIFIC DISCOVERY WITH MODULUS

RENEWABLE ENERGY
Siemens Gamesa: Up to 4000X Speedup of Wind Turbine Wake Optimization

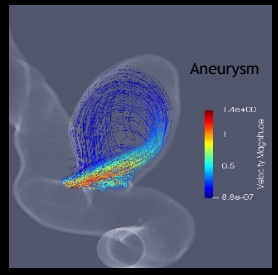
CLIMATE CHANGE
45,000X Speedup of Extreme weather Prediction with FourCastNet



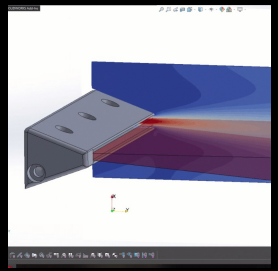
INDUSTRIAL HPC
NETL: 10,000X Faster Build Of high-fidelity surrogate models



HEALTHCARE
Achieve high-fidelity results faster for blood flow in inter-cranial aneurysm



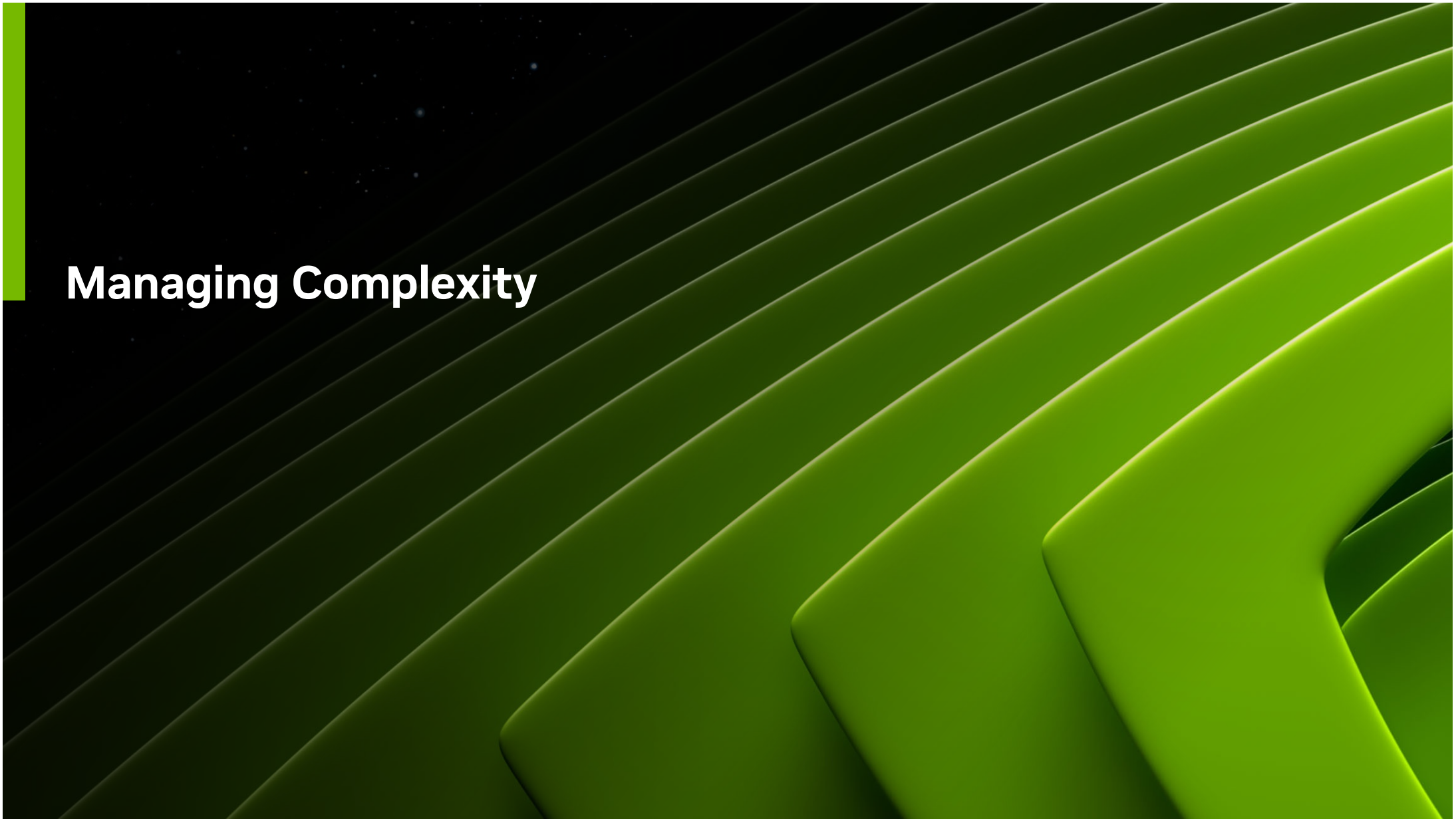
DIGITAL TWINS
Kinetic Vision: Design Optimization Using parameterized models



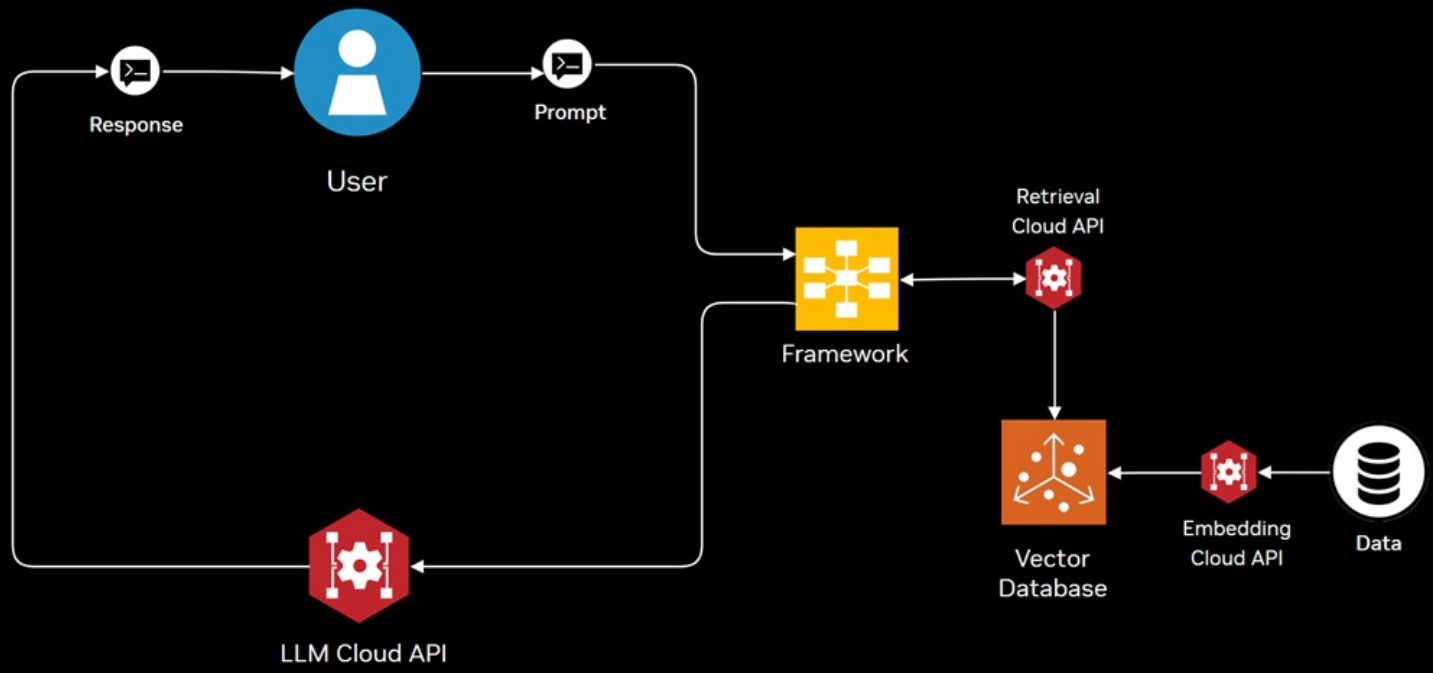
Maximizing Wind Energy Production Using Wake Optimization

NVIDIA Modulus and Omniverse

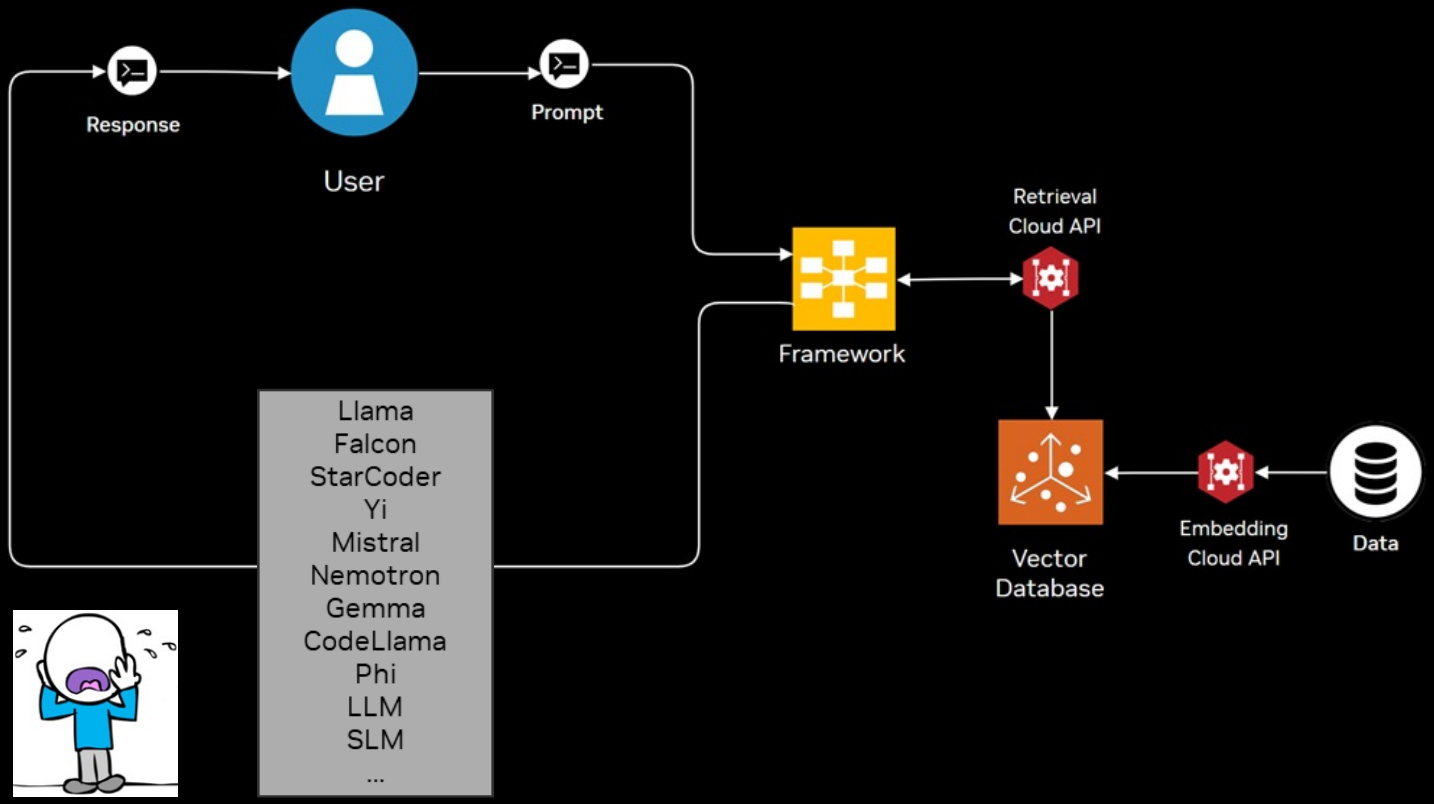
Managing Complexity



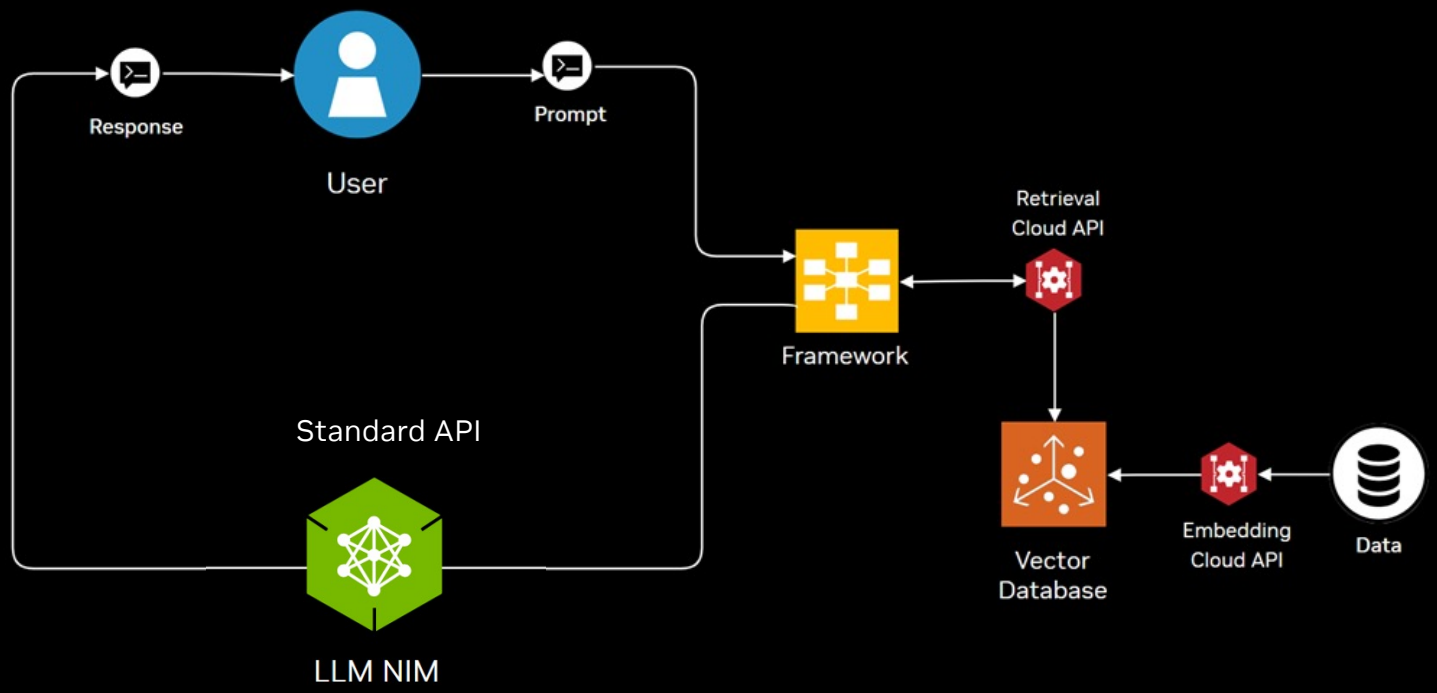
GenAI: Managing the complexity



The challenge: ”How do I keep up with the pace of innovation?"



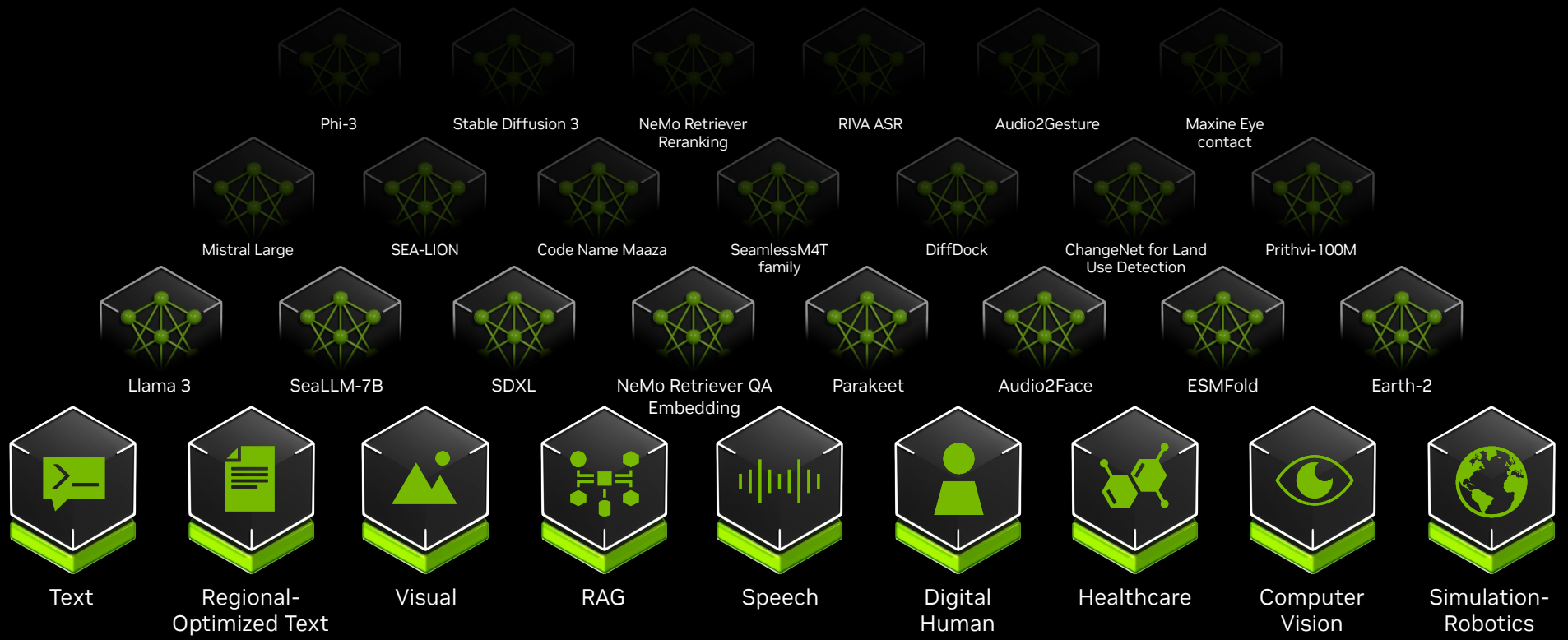
NVIDIA NIMS: All about easing the path to production



NVIDIA INFERENCE MICROSERVICE

Pre-Trained AI Models
Packaged and Optimized to Run Across CUDA Installed Base

NVIDIA NIM For Every Domain

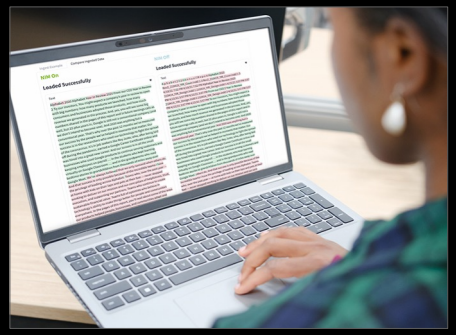


NVIDIA AI Blueprints

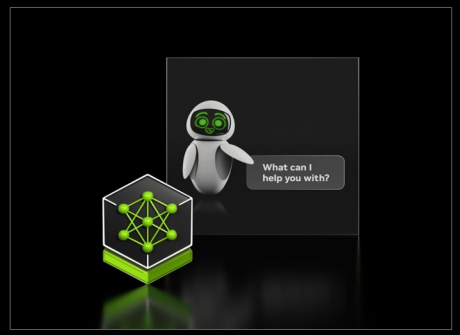
NIM Workflows for application deployment

Digital Avatar

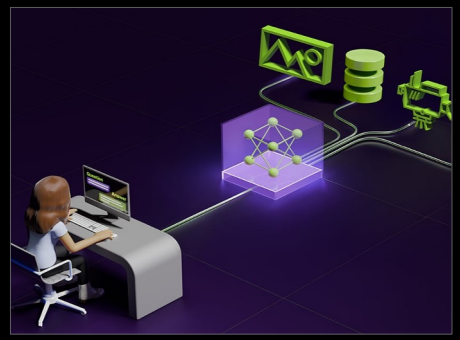
PDF to Podcast



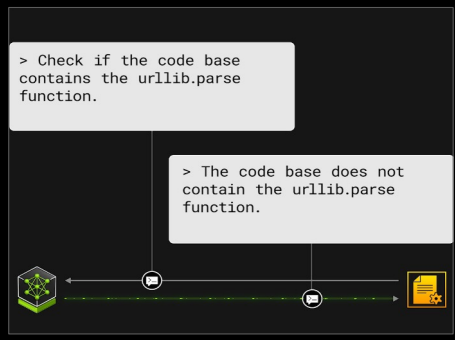
Multimodal PDF Data Extraction



Customer Service AI Assistant



Video Search & Summarization



Security Vulnerability Analysis

NVIDIA AI Blueprints

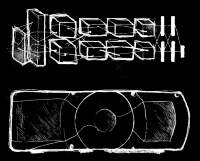
NVIDIA

Agentic AI with Advanced Reasoning

Agentic AI

AI Agents that can act independently to achieve a specified task

2012 ALEXNET



PERCEPTION AI

SPEECH RECOGNITION
DEEP RECSYS
MEDICAL IMAGING

GENERATIVE AI

DIGITAL MARKETING
CONTENT CREATION

AGENTIC AI

CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE

PHYSICAL AI

SELF-DRIVING CARS
GENERAL ROBOTICS

NVIDIA

NVIDIA Models and Libraries to build Agentic AI

NVIDIA AI Blueprints

Software Security
AI Agent

NVIDIA NeMo

- Curator
- Customizer
- Evaluator
- Guardrails
- RAG Retriever

NVIDIA NIM

Understanding & Reasoning

Information Retrieval

AI Safety

Digital Human

Visual Content Generation

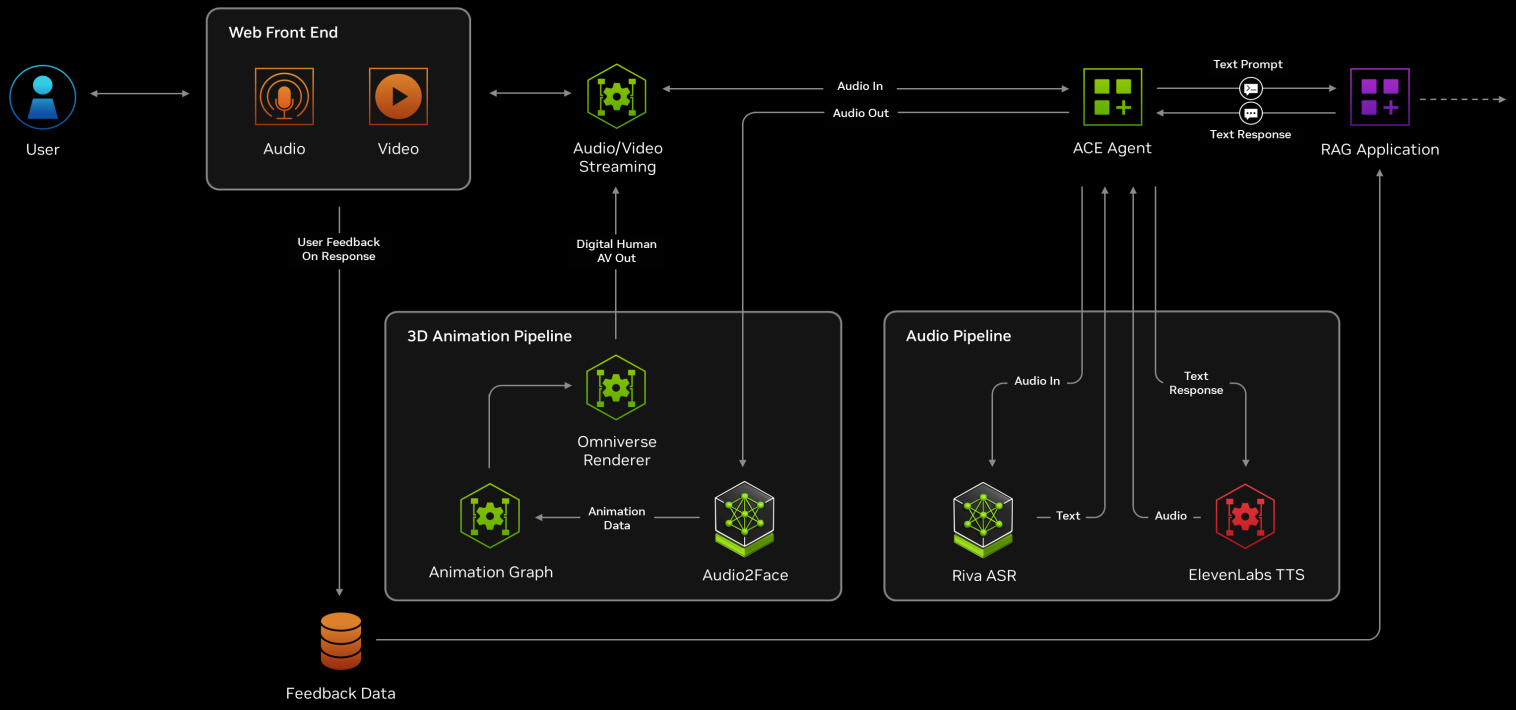
Digital Biology

Physical AI

NIM Agent Blueprints – Digital Human Customer Services Agent

https://build.nvidia.com/nvidia/digital-humans-for-customer-service/blueprintcard

- Tasks
- Allows a human like interface to generative AI applications using various ASR, TTS and Audio2Face technologies

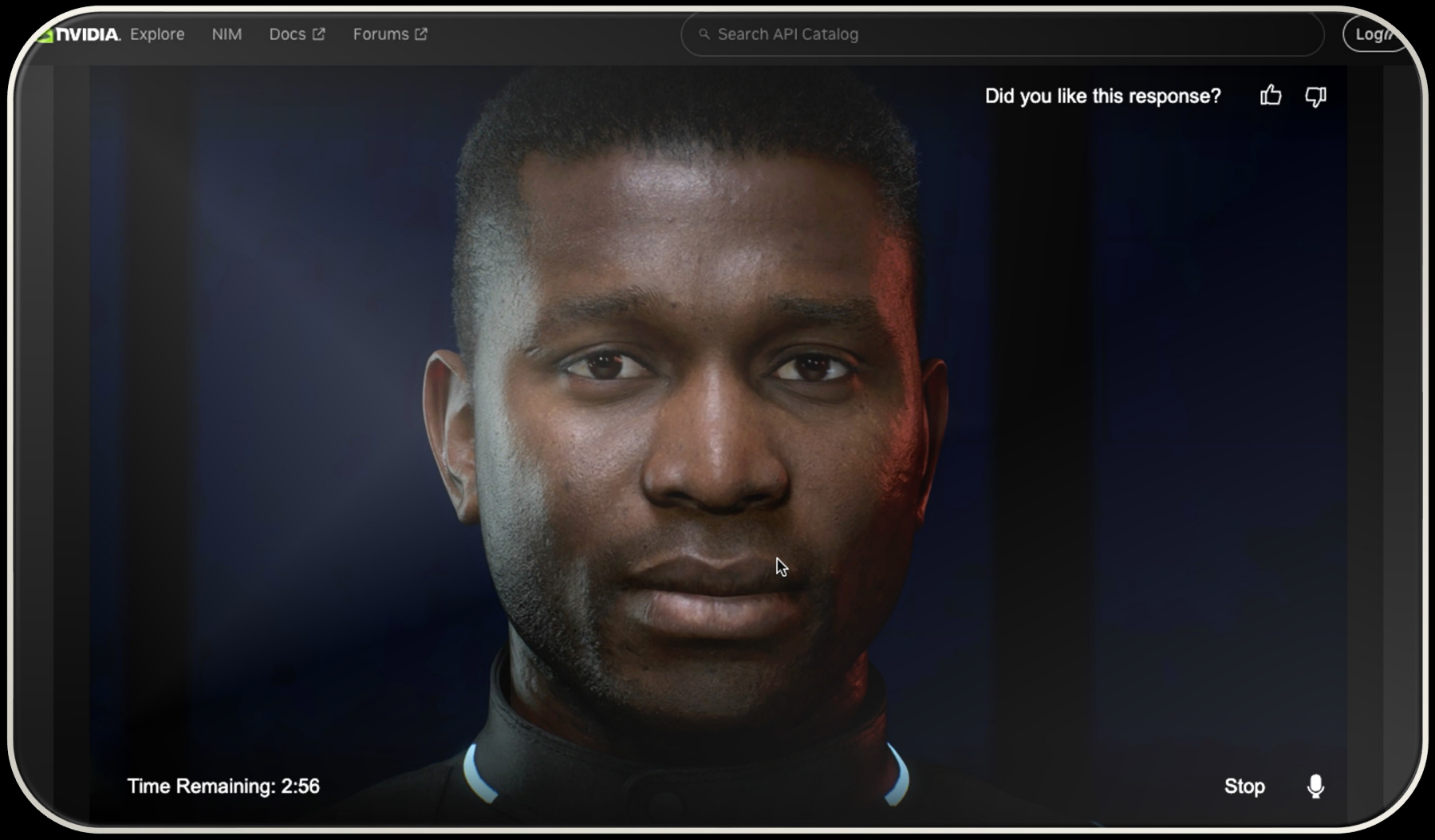


- The following NIMS are used:

- nv-embedqa-e5-v5
- nv-rerankqa-mistral4b-v3
- Llama3-8b-instruct
- Parakeet-ctc-1.1b-asr
- Audio2face
- Other ACE Microservices
-

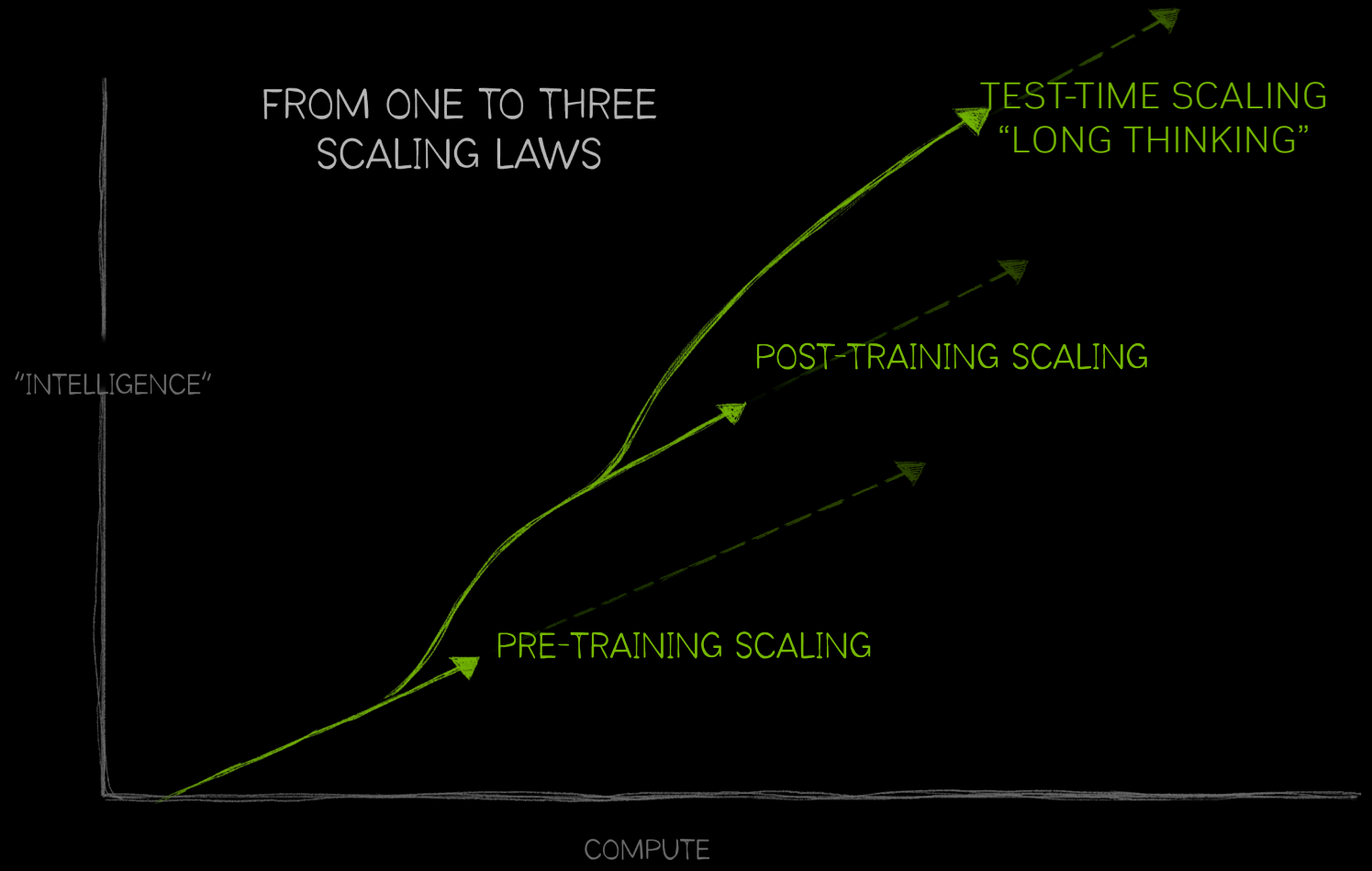
NIM Agent Blueprints

Digital Human – Customer Services Example



AI Scaling Laws Drive Exponential Demand for Compute

New “long thinking” supercharges inference scaling



Pre-Training: Teaching AI Models Knowledge

What is NVIDIA?

NVIDIA is a multinational technology company that specializes in designing and manufacturing graphics processing units (GPUs), high-performance computing hardware, and artificial intelligence (AI) technologies. The company was founded in 1993 by Jensen Huang, Chris Malachowsky, and Curtis Priem.

NVIDIA is headquartered in Santa Clara, California, and has become one of the leading companies in the field of computer graphics, gaming, and AI.

100 Trillion Tokens

Knowledge is not the same as Thinking

I need to figure out where to seat my family of 8 for dinner at a round table.
My in-laws don't get along with my parents.
My sister needs to sit next my little brother to help him eat.
My wife really doesn't want to sit next to my mom.
Where should each person be seated?

A complex family dynamic!

Don’t worry, I am here to help you navigate this challenging situation. Continuing the relationships and seating preferences, here is a suggestion.

1.Separate tables: If possible, consider setting up a separate table for your in laws and parents. This way, they won’t have to interact with each other, and you can still have a family gathering.

Actual Llama 8B Response

Post Training: Teaching Models to Think

6000+ Thinking Tokens | 300 Answer Tokens

I need to figure out where to seat my family of 8 for dinner at a round table.
My in-laws don't get along with my parents.
My sister needs to sit next my little brother to help him eat.
My wife really doesn't want to sit next to my mom.
Where should each person be seated?

To seat your family of 8 harmoniously at a round table
while respecting all constraints, here's the optimal arrangement:

You Wife Dad Mom Sister Little Brother Mother-in-law Father-in-law

Key Constraints Met:

- In-laws vs. Parents: Separated by your sister, brother, and you.
- Sister & Brother: Adjacent for assistance.
- Wife & Mom: Not next to each other (separated by you and dad).

Physical AI

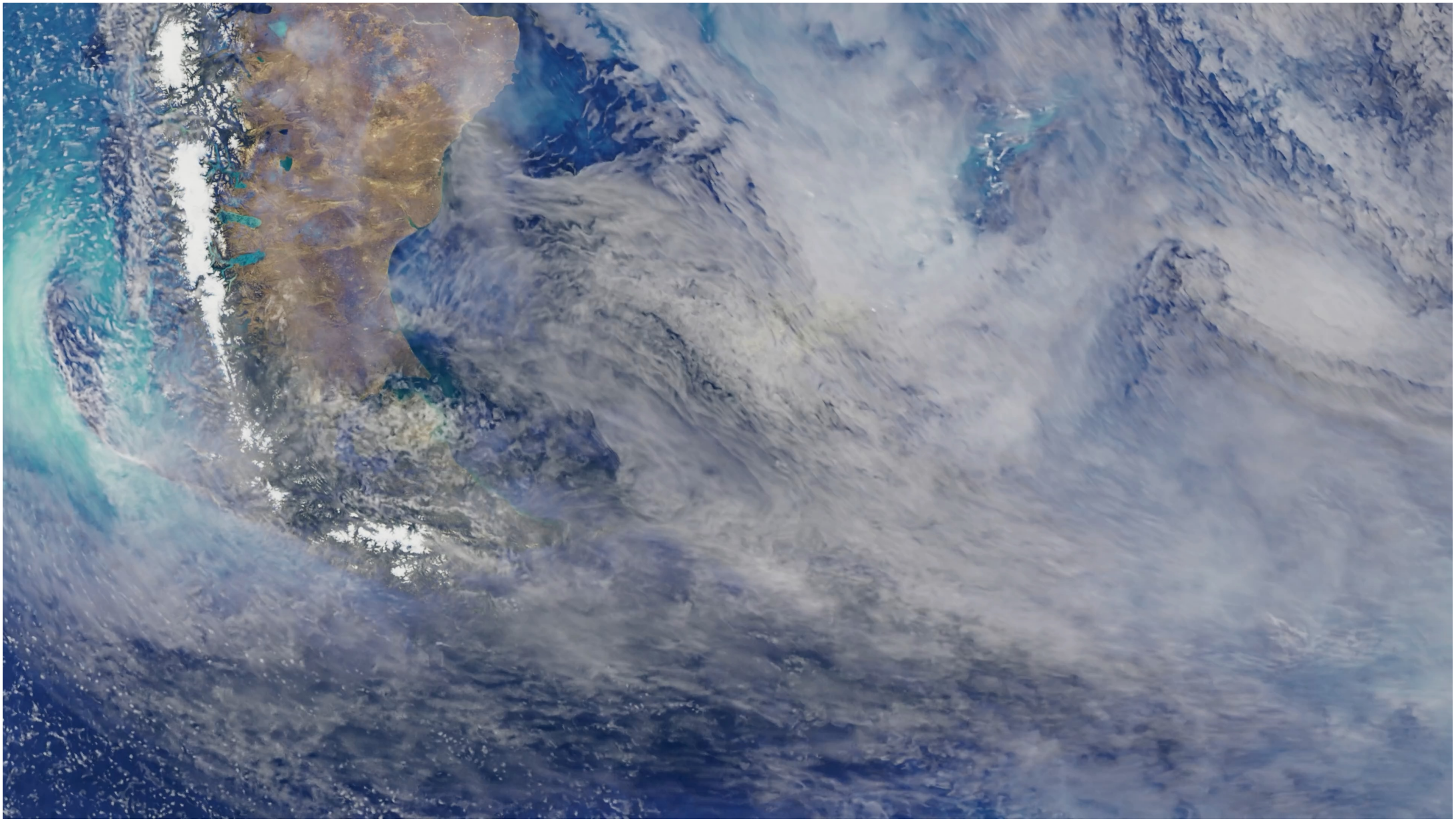
"Pick up the piece of toast from the toaster and place it on the white plate to the right of the toaster."

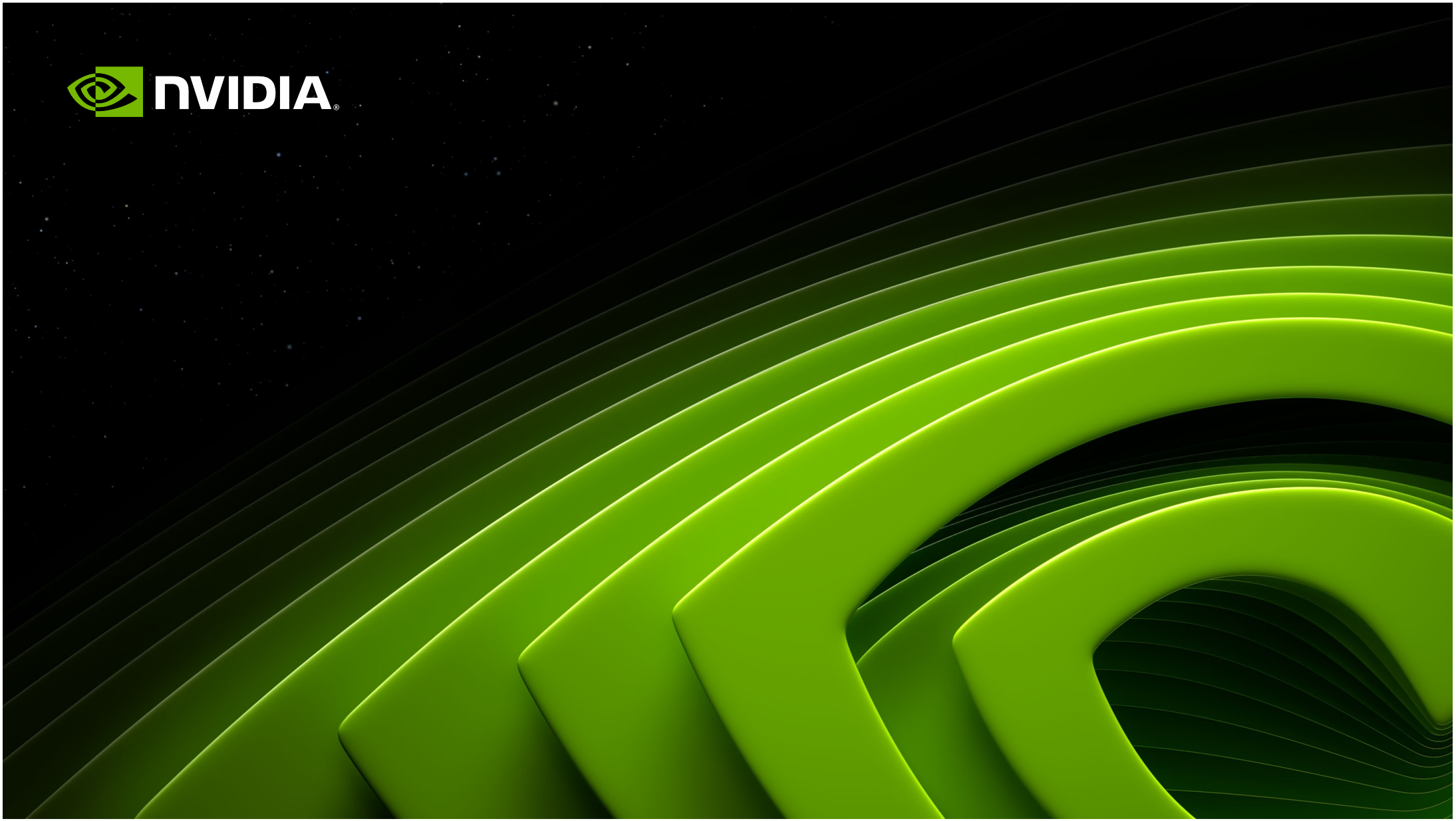
Image Tokens

Text Tokens

World Foundation Model

Action Tokens




NVIDIA