Stanford Tree Hacks 2026
Getting Started

Challenge Track: Edge AI

The Challenge: Build a multi-agent solution or system with NVIDIA open models.

The Tools

- *NVIDIA Nemotron open model & dataset*
- *NVIDIA Cosmos open model*
- *NVIDIA Brev credits*
- *40 ASUS Ascent GX10 w/ GB10*
- *20 NVIDIA Jetson Orin Nano Super*

Important Links

- DGX Spark
    a. https://build.nvidia.com/spark
- Jetson Orin Super Nano
    a. https://www.jetson-ai-lab.com/
- Cosmos
    a. https://github.com/nvidia-cosmos/cosmos-cookbook/tree/main
    b. Using Cosmos Reason on Cloud with NVIDIA Brev
        i. https://nvidia-cosmos.github.io/cosmos-cookbook/getting_started/brev/reason2/reason2_on_brev.html
- Nemotron
    a. https://github.com/NVIDIA-NeMo/Nemotron/tree/main

Nemotron

Robotics/Physical AI

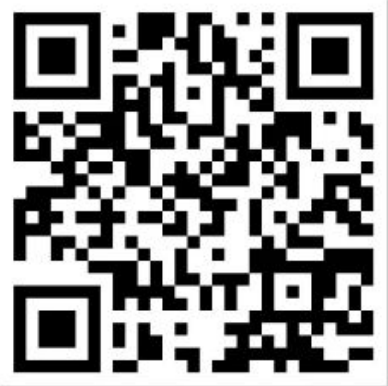
Speech

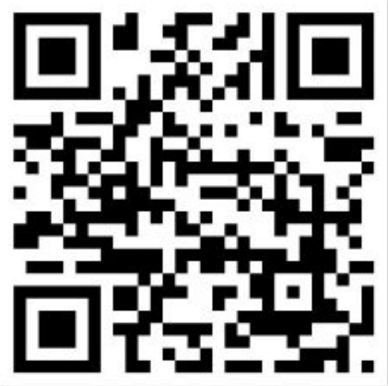
Get Started: Edge to Cloud

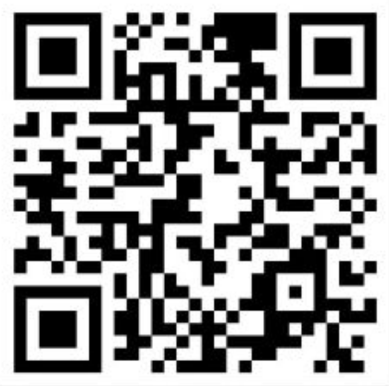
Spark

Jetson

NIM







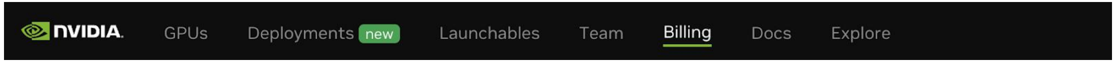
Discord for Support

Join # stanford-tree-hack-2026

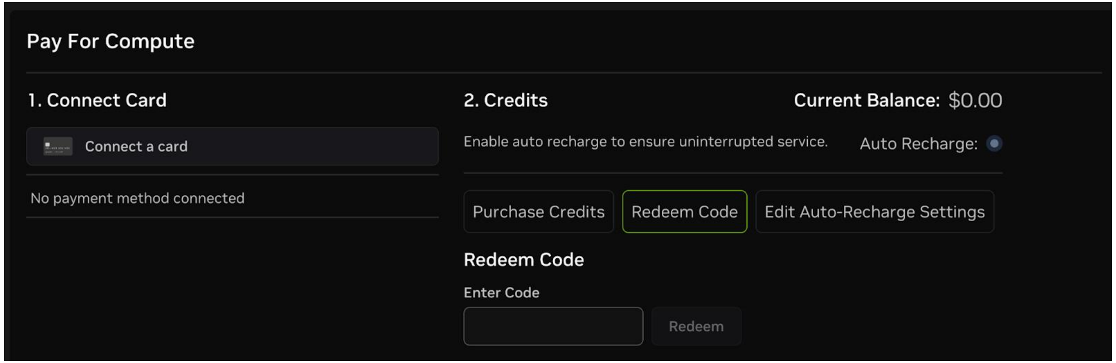
Steps for Redeeming your Brev Credit Coupon

Steps to redeem coupon:

Step 1: Please go to http://brev.nvidia.com and input your email to create an account.



Step 2: Once you have signed up, please go to the 'Billing' tab located at the top of the Brev console (as shown in the screenshot below).



Pay For Compute

1. Connect Card

No payment method connected

2. Credits

Current Balance: $0.00

Enable auto recharge to ensure uninterrupted service.

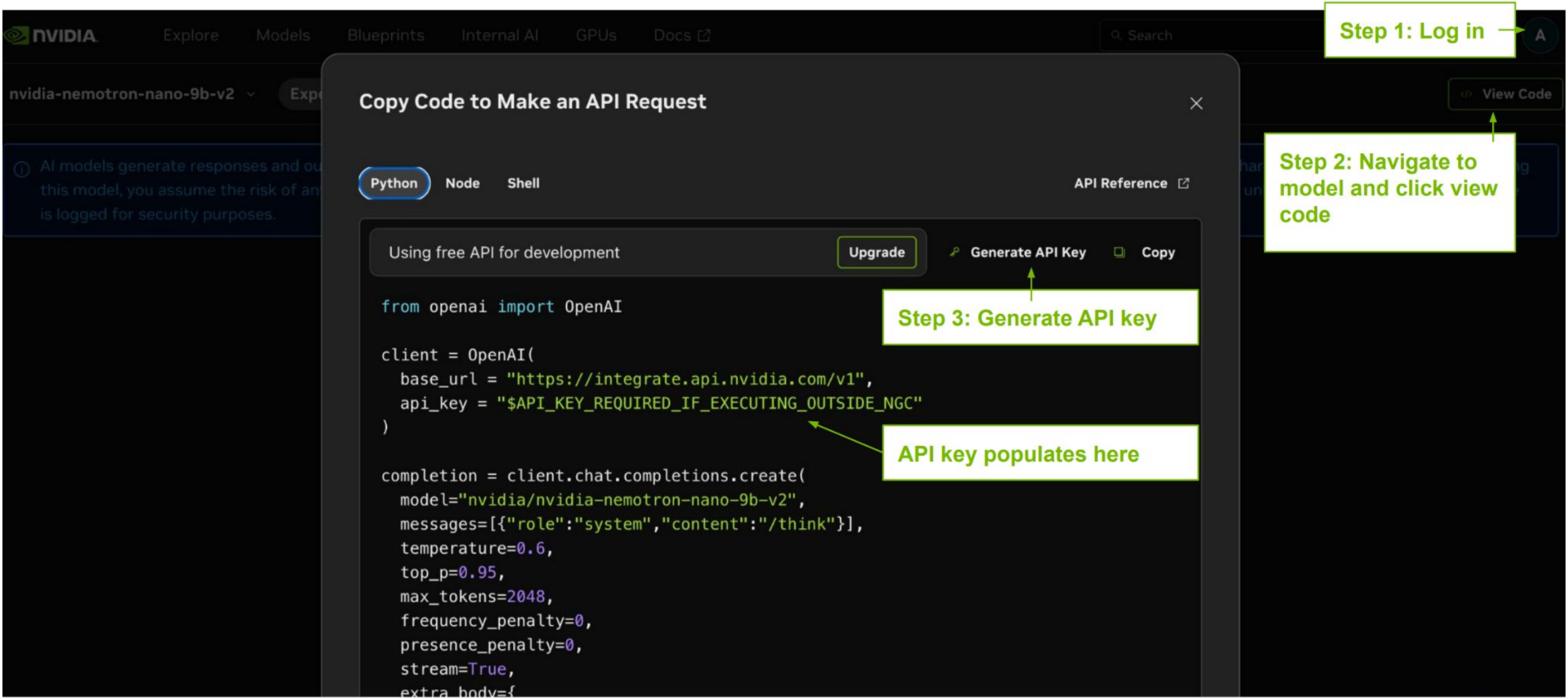
Auto Recharge: ●

Redeem Code

Enter Code

Step 3: Once you have navigated to the 'Billing' tab, scroll down and click 'Redeem Code'. Then, enter the code you were provided in the 'Enter Code' field and click 'Redeem.'

Getting Started - build.nvidia.com



Step 1: Log in

Step 2: Navigate to model and click view code

Step 3: Generate API key

API key populates here

AI models generate responses and ou... this model, you assume the risk of an... is logged for security purposes.

Copy Code to Make an API Request

Using free API for development

```
from openai import OpenAI

client = OpenAI(
  base_url = "https://integrate.api.nvidia.com/v1",
  api_key = "$API_KEY_REQUIRED_IF_EXECUTING_OUTSIDE_NGC"
)

completion = client.chat.completions.create(
  model="nvidia/nvidia-nemotron-nano-9b-v2",
  messages=[{"role":"system","content":"/think"}],
  temperature=0.6,
  top_p=0.95,
  max_tokens=2048,
  frequency_penalty=0,
  presence_penalty=0,
  stream=True,
  extra_body={
```

Ideal Projects for this Track

Winning projects will showcase true agentic behavior:

Multi-Agent Systems: Build teams of specialized AI agents (like Report Generator: Research Agent → Outline Agent → Writer Agent → Editor)

Agentic RAG: Systems that intelligently decide WHEN to retrieve information, not just HOW (perfect for domain-specific assistants)

ReAct Pattern Workflows: Agents that Reason → Act → Observe in loops to solve problems iteratively (like automated debugging or technical support)

Tool-Calling Applications: Leverage Nemotron's exceptional ability to use external APIs and tools (finance analysis, DevOps automation, content creation)

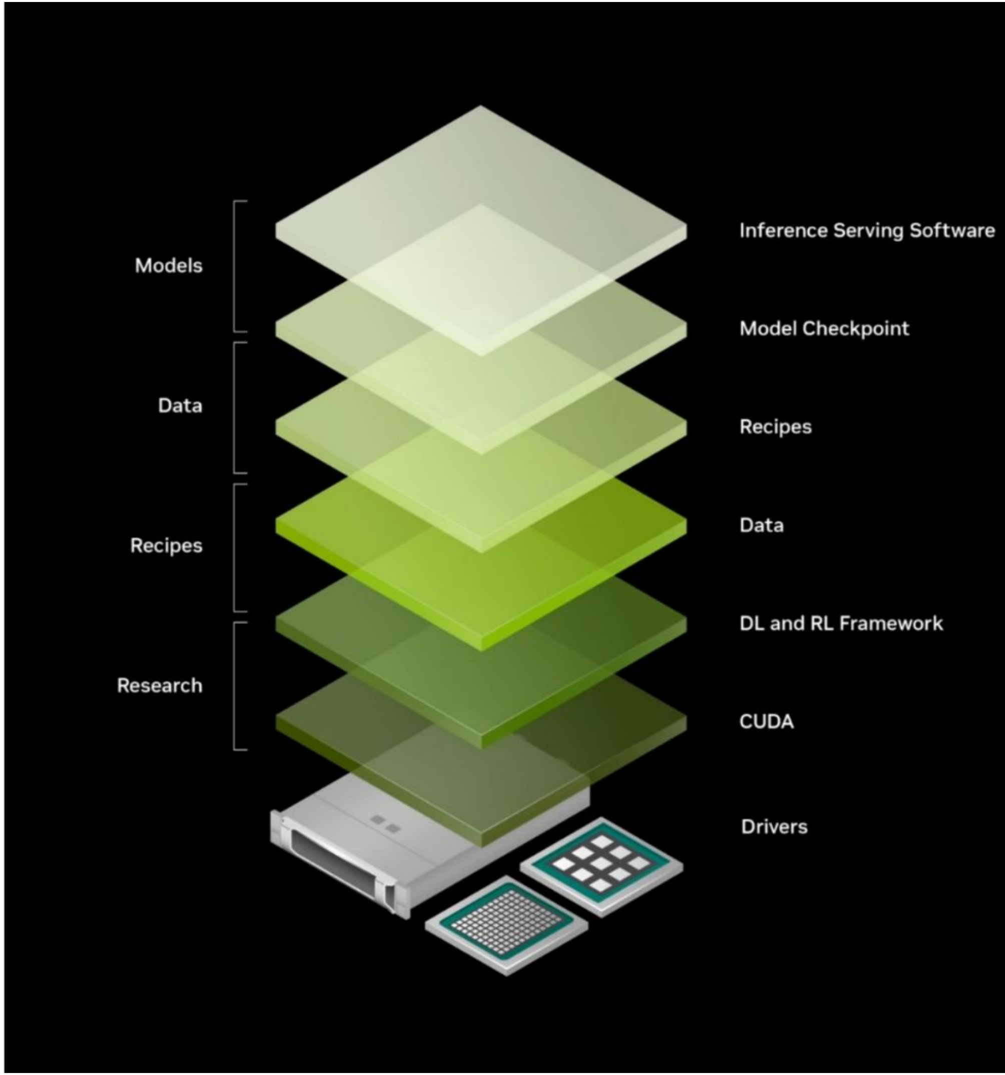
Multi-Modal Agents: Combine Nemotron reasoning with VLMs (visual analysis + logical decision-making.

Agent Simulation & Evaluation: Use Nemotron to generate realistic test scenarios and evaluation pipelines.

Video Analytics AI Agents: Analyze live or recorded video streams with Cosmos Reason VLM

Nemotron

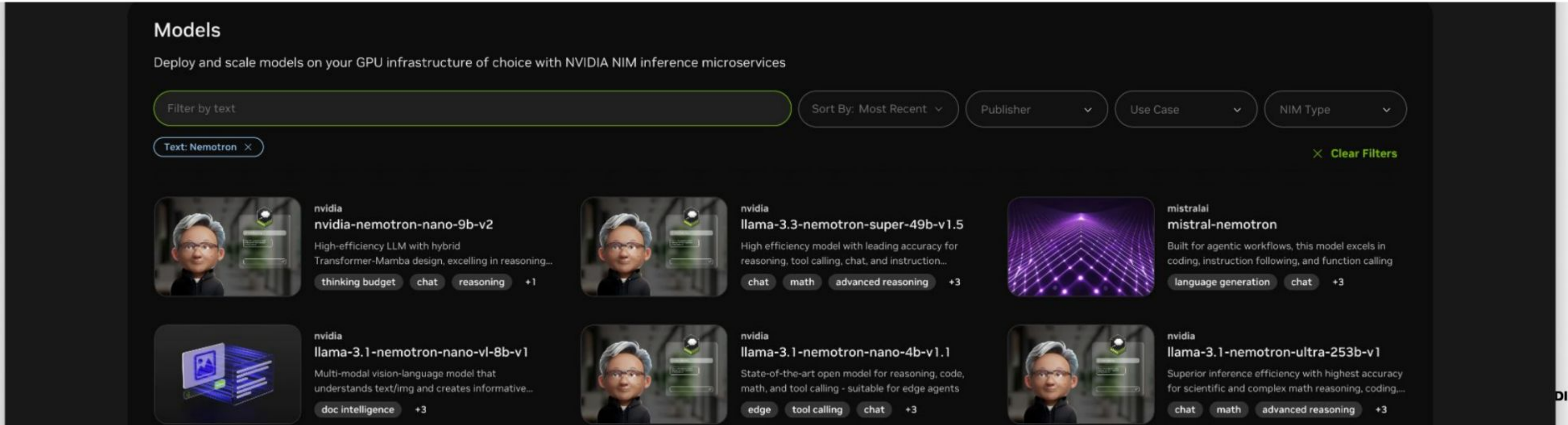
Is more than models. It’s a focus on optimizing for our AI ecosystem to provide the best quality of results across our technology stack.



What is Nemotron Best For?

NVIDIA Nemotron models are purpose-built for agentic AI - the next generation of intelligent systems that don't just answer questions, but reason, plan, and take action. Unlike general-purpose LLMs, Nemotron excels at:

- Advanced reasoning and multi-step problem-solving
- Function calling to interact with external tools and APIs
- Autonomous decision-making within agent workflows
- Retrieval-augmented generation (RAG) for knowledge-based tasks
- Multi-agent orchestration where specialized agents work together together



Models

Deploy and scale models on your GPU infrastructure of choice with NVIDIA NIM inference microservices

Text: Nemotron

nvidia
nvidia-nemotron-nano-9b-v2
High-efficiency LLM with hybrid Transformer-Mamba design, excelling in reasoning...
thinking budget · chat · reasoning · +1

nvidia
llama-3.3-nemotron-super-49b-v1.5
High efficiency model with leading accuracy for reasoning, tool calling, chat, and instruction...
chat · math · advanced reasoning · +3

mistralai
mistral-nemotron
Built for agentic workflows, this model excels in coding, instruction following, and function calling
language generation · chat · +3

nvidia
llama-3.1-nemotron-nano-vl-8b-v1
Multi-modal vision-language model that understands text/img and creates informative...
doc intelligence · +3

nvidia
llama-3.1-nemotron-nano-4b-v1.1
State-of-the-art open model for reasoning, code, math, and tool calling - suitable for edge agents
edge · tool calling · chat · +3

nvidia
llama-3.1-nemotron-ultra-253b-v1
Superior inference efficiency with highest accuracy for scientific and complex math reasoning, coding,...
chat · math · advanced reasoning · +3

How to get DGX Spark and Jetson

Hardware will be available on the first-come, first-serve basis.

There are 40 DGX Spark units, and 20 Jetson Orin Nano Super units.

Request a unit by talking to either an NVIDIA employee at the NVIDIA booth, or a Stanford TreeHacks staff member.

How to get Started with Nemotron

- Nemotron GitHub (models, examples, configs)
- Nemotron models on build.nvidia.com (model cards, playground, API)
- “What is Nemotron Best For?” model overview (reasoning, tool-calling, RAG, agents)

How to do Visual Search and Summarization

- Combine vision and language models to search visual content and generate natural language summaries
  - Extract key information from images, videos, or documents
  - Generate contextual summaries of visual data at scale
  - Enable semantic search across image/video libraries
- Extract frames/features with vision model → Generate embeddings → Store in vector DB → Query with natural language → Summarize results with LLM
- Important links:
  - Production-ready AI Blueprint: https://github.com/NVIDIA-AI-Blueprints/video-search-and-summarization
  - Community example for video dataset indexing: https://github.com/NVIDIA/GenerativeAIExamples/tree/main/community/video-dataset-search

How to do Smart Query Routing

- Automatically route incoming queries to the most appropriate model based on query complexity, domain, or cost requirements. For instance:
  - **Simple questions** → smaller, faster models
  - **Complex reasoning** → larger, more capable models
- Optimize for latency, cost, and quality simultaneously
- Use semantic embeddings or lightweight classifier to analyze query intent, then route to appropriate model tier
- Try It Yourself
  - Demo on DGX Spark with vLLM Semantic Router - https://github.com/vllm-project/semantic-router/tree/main/deploy/examples/multi-model-routing

Connect Two DGX Sparks for Inference

Spark Your Ideas with DGX Spark

- https://build.nvidia.com/spark/connect-two-sparks/overview

What is Cosmos Reason 2

State-of-the-art reasoning vision language model (VLM) for Physical AI

- Open, customizable, commercial-ready reasoning VLM
- Excels at navigating diverse real-world scenarios
- Enhanced spatial-temporal understanding and visual perception
- Flexible deployment with 2B and 8B model sizes
- Improved long-context understanding with 256K input tokens

#1

Open model
Physical Reasoning,
Physical AI Bench

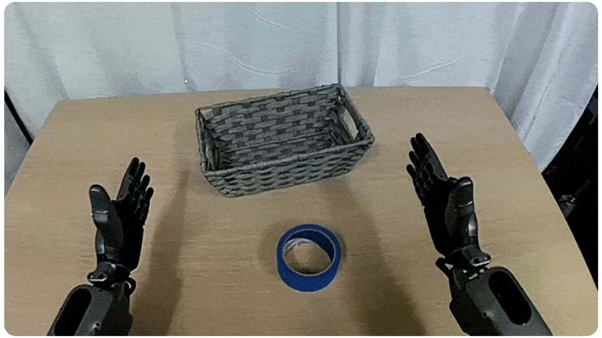
2M+

Downloads
Hugging Face

What is Cosmos Reason Best Used For?

Analyze video data during training and runtime

Data Curation and Annotation



Robot Planning and Reasoning

Video Analytics AI Agent

Get Started with Cosmos Reason with a WebUI for live video streams

Edge deployment with DGX Spark

1. **Go to https://github.com/NVIDIA-AI-IOT to see full instructions**
2. Install using pip/uv
3. First run the vLLM running Cosmos Reason 2
4. For the easiest setup - run live-vlm-webui command on the same machine
5. vLLM will auto detect the model running
6. NOTE: Inference on USB cameras is supported only on the same host running the live vlm webui. Open the client in the browser of the same host or use RTSP streaming of the camera.

Getting Started with Metropolis VSS Blueprint

Use Cosmos Reason to build Video Analytics AI Agents for Cloud and Edge Deployments

Cloud Deployment

Brev Launchable

1. **Go to the Launchable page.**
2. Click on ‘Deploy Launchable’ on top right.
3. Click on ‘Go to Instance Page’
4. Click on ‘Open Notebook’ button after it is enabled (This could take a couple of minutes).
5. **Navigate and open `video-search-and-summarization/deploy/1_Deploy_VSS_docker_Crusoe.ipynb` notebook.** This notebook is designed to run as a launchable on 8XL40S GPU CRUSOE Cloud Provider with Ephemeral storage.
6. Add your NGC_API_KEY in the first code cell.
7. Restart Kernel and Run all cells.
8. Follow the instructions in the notebook to access video search and summarization blueprint UI.

https://docs.nvidia.com/vss/latest/content/cloud_brev.html#cloud-brev

Edge Deployment

DGX Spark

1. **Clone the VSS GitHub repository on your device**
2. **git clone https://github.com/NVIDIA-AI-Blueprints/video-search-and-summarization.git**
3. Explore the deployment scenarios in the documentation for local, hybrid and event reviewer deployments
4. Local deployment will deploy VSS with all features including video summarization, Q&A and live stream alerts
5. Event Reviewer will deploy a lightweight version of VSS suitable for low latency alerts by combining a CV and VLM pipeline.
6. Follow the deployment instructions for your desired profile
7. Once deployed access the VSS Web UIs to upload videos, connect streams and test VSS

https://docs.nvidia.com/vss/latest/content/vss_dep_docker_compose_arm.html

1st Place

- ASUS Ascent GX10 w/ NVIDIA GB10
- ASUS ZenScreen
- $450 NVIDIA Brev Credit Vouchers (2)
- NVIDIA Jetson Orin Nano Super (2)
- NVIDIA Hat (4)

2nd Place

- ASUS Ascent GX10 w/ NVIDIA GB10
- ASUS ZenScreen
- $300 NVIDIA Brev Credit Vouchers (2)
- NVIDIA Jetson Orin Nano Super
- NVIDIA Hat (4)

3rd Place

- ASUS ZenScreen
- $200 NVIDIA Brev Credit Vouchers (4)
- NVIDIA Jetson Orin Nano Super
- NVIDIA Hat (3)

Join Us for a 30 Minute Workshop

Saturday 2:00 PM

Jay Rodge

Sr. Developer Advocate

Chitoku Yato

Sr. Technical Product Marketing Manager

In this fast, 30-minute workshop, NVIDIA experts will show you how to get models running locally with the performance, latency, and privacy modern apps demand. You'll walk away with practical techniques you can immediately apply to your hack, whether you're building agents, vision apps, or real-time experiences.

Pause your coding — this session might be the upgrade your project needs to stand out.