FRONTIER DEEPTECH ACCELERATOR QUALIFIED

ACCELERATED SPATIAL FOUNDATION MODELS

NeuroTensor Labs

Real-Time Spatial World Models for 100 Million Autonomous Robots

Developing sub-8ms Vision-Language-Action (VLA) foundation models accelerated natively on low-latency TensorRT runtimes, CUDA-X, and physics simulators for autonomous mobile robots and physical edge systems.

FOUNDING LEADERSHIP

Tanaji Vishnu Yadav

Founder & Chief Executive Officer

ex-Stanford AI Lab • 12+ Publications

TARGET COMPUTE ASK

$250K Inception Tier

Distributed 80GB SXM5 GPU Clusters

Accelerating 70B Foundation Model

The 50ms Latency Wall in Physical Robotics

Current multimodal models (LLaVA, RT-2, open-source VLMs) are designed for cloud chat and operate at 200ms–1500ms cycle times. In the physical world, an AMR moving at 2 m/s travels 1 meter before perceiving a moving human obstacle.

Unacceptable Latency

Standard PyTorch models take >50ms on edge chips, causing jitter, frequent emergency stops, and robotic navigation failure.

VRAM Memory Bloat

Standard 7B models consume 14GB+ VRAM, leaving zero compute headroom for path planning and motor control on embedded edge boards.

Sim-to-Real Gap

Models trained only on 2D internet photos fail catastrophically when confronted with changing industrial lighting and 3D depth variations.

NeuroVoxel: Native Spatial VLA Foundation Model

A breakthrough multimodal transformer that treats spatial geometry as a primary modality alongside RGB vision, executing in sub-8ms directly on Edge AGX and next-gen Thor edge hardware.

KEY INNOVATION 1

Unified Geometry-Action Tokens

Rather than decoupling perception, mapping, and motion planning into 3 separate pipelines, NeuroVoxel processes RGB-D sensor streams into unified 6-DoF trajectory primitives within a single forward pass.

`Latency: 5.8ms Glass-to-Action`

KEY INNOVATION 2

Hardware-Aware TensorRT Compilation

Our compiler directly generates TensorRT 10 execution engines using FP8 SmoothQuant and FlashAttention-3 kernels, compressing memory footprints down to under 2GB VRAM.

`VRAM Footprint: 1.92 GB (Fits Jetson Orin Nano)`

Built for High-Throughput Accelerated Computing

Every layer of the NeuroTensor pipeline leverages hardware-aware software libraries and GPU primitives for maximum TFLOPS utilization.

TensorRT 10.0

Automatic kernel fusion & FP8 execution graph optimization.

Isaac Sim & USD

Photorealistic synthetic digital twin training data generation.

Triton Server

Dynamic batching and multi-tenant GPU deployment across edge fleets.

CUDA 12.6 Kernels

Hand-tuned warp-level reduction for 3D point cloud voxelization.

Compatible with: Edge AGX Orin, Thor, Enterprise L40S, and 80GB SXM5 Clusters

100% CUDA Native

$48.5 Billion Total Addressable Market by 2030

The convergence of generative AI and physical machines is creating an exponential demand for low-latency foundation models that can be embedded into mass-market robots.

$48.5B

TOTAL ADDRESSABLE MARKET (TAM)

Global intelligent robotics software, spatial perception sensors, and autonomous machine intelligence by 2030 (34% CAGR).

$12.4B

SERVICEABLE MARKET (SAM)

Autonomous mobile robots (AMRs) in logistics, factory automation, drone delivery, and agricultural robotic fleets.

$1.8B

SERVICEABLE OBTAINABLE (SOM)

Edge foundation model licenses and per-device runtime royalties targeting our initial 50 enterprise OEM pipeline.

Modular Architecture for Robotics Developers

Robotics engineering teams can integrate our technology as pre-trained foundation weights, custom edge compilation engines, or synthetic training pipelines.

NeuroVoxel-1 (7B / 32B VLA Foundation Model)

Pre-trained weights with open fine-tuning adapters for custom sensor rigs.

GA v1.2

EdgeSync TensorRT 10 Compiler

Automated quantization and engine builder producing deterministic sub-8ms binaries.

PROPRIETARY TOOL

OmniSynthetic Studio (Physics Digital Twin & USD Sim)

Procedural USD world generator for rapid domain randomization & zero-shot transfer.

ISAAC SIM POWERED

Proven Superiority Across Hardware & Latency

Rigorous testing on Edge AGX Orin 64GB, Enterprise 80GB SXM5, and Enterprise L40S demonstrates clear performance leadership over baseline open models.

THROUGHPUT

128.4 FPS

vs. 31 FPS baseline

INFERENCE LATENCY

5.8 ms

vs. 48.2 ms baseline

VRAM FOOTPRINT

1.92 GB

vs. 14.8 GB PyTorch

OBSTACLE ACCURACY

99.4%

Zero-shot 3D mAP

Validated across 3 commercial pilot trials in logistics fulfillment centers (Q1 2026).

High-Margin Software & Royalty Model

Monetizing through per-device edge runtime licensing, annual enterprise SDK access, and cloud training compute subscriptions.

TIER 1 • OEM EMBEDDED

Per-Device Royalty

$18 - $45 / device / yr

Embedded runtime license per deployed AMR, drone, or industrial robot camera node. Recurring annual activation.

Target: 250,000 devices by 2028

TIER 2 • ENTERPRISE SDK

Annual Platform License

$120,000 / enterprise

Full access to NeuroVoxel-1 weights, EdgeSync compiler toolchain, and custom hardware calibration assistance.

Target: 35 OEM customers

TIER 3 • CLOUD DIGITAL TWINS

Synthetic USD Simulation

$0.05 / sim scenario

On-demand GPU-accelerated simulation hours via cloud physics simulation and digital twin orchestration.

High-volume usage based

Executive Leadership & Direction

NeuroTensor Labs is founded and led by Tanaji Vishnu Yadav, directing core spatial AI architecture, low-latency GPU acceleration, and enterprise commercial deployment.



FOUNDER & CEO

Tanaji Vishnu Yadav

Lead AI Systems Architect • ex-Stanford AI Lab

Pioneer in low-latency spatial geometry models and embodied robotics vision. 12+ peer-reviewed papers across CVPR, ICCV, and NeurIPS with 2,400+ citations.

tanaji.yadav@neurotensor.ai

Foundation Model Architecture & VLA

Directs the design of NeuroVoxel spatial transformer weights, multimodal tokenization, and end-to-end 6-DoF trajectory planning for autonomous mobile robots.

GPU Acceleration & TensorRT Compilation

Oversees FP8 SmoothQuant compiler pipelines and CUDA 12.6 kernel execution on Edge AGX Orin, Thor, and 80GB SXM5 clusters.

Capitalization & Compute Expansion Roadmap

Accelerating development of our 70B parameter spatial world model with dedicated ultra-scale compute resources and co-marketing opportunities.

Compute & Hardware Request

- ✓ **$100K–$250K Cloud Credits:** Allocation for Distributed 80GB SXM5 GPU cluster hours.
- ✓ **Hardware Dev Kits:** 4x Edge AGX Orin 64GB & Next-Gen early-access dev units for edge testing.
- ✓ **TensorRT Technical Liaison:** Direct sync with low-latency TensorRT & physics simulation compiler teams.

12-Month Key Milestones

- **Q2 2026:** Pre-train NeuroVoxel-70B on 1 Trillion physical & Isaac Sim USD tokens.
- **Q3 2026:** Deploy live pilots across 500+ commercial warehouse AMRs with sub-6ms latency.
- **Q4 2026:** Present real-time spatial foundation benchmarks at premier AI & robotics summits.

Direct Founders Contact: founders@neurotensor.ai

Return to Website

CONNECT WITH CEO TANAJI VISHNU YADAV