YIANNIS HADJIYIANNI

RESEARCH ENGINEER | FOUNDATION MODELS & DISTRIBUTED ML

me@yiannisha.dev • +1 (415) 605-5658 • GitHub • LinkedIn

Research focus. Scalable foundation-model training and numerical reliability. Experience in distributed PyTorch training, RL systems, and real-time sensory perception/control.

SELECTED RESEARCH

Harbin Institute of Technology (HIT) | *Research · Prof. Dimitrios Stamoulis · 2026–Present*

- Numerical reliability of robotics foundation models (RSS 2027 incoming). Identified batch-dependent inference drift across 10+ robotics models and demonstrated changes in task outcomes. Code

DreamZero: high-throughput WAM rollout engine | *in progress*

- Designing a distributed, asynchronous rollout system for world-action-model RL; focusing on high-throughput sampling, actor/learner execution, and numerical parity between rollouts and training.

Low-/mixed-precision adaptation of VLAs and WAMs | *in progress*

- Investigating FP8/FP4 full-model and LoRA fine-tuning, measuring training efficiency, optimization stability, and downstream policy behavior.

PROFESSIONAL EXPERIENCE

Cybernetic Physics Sep 2026 – Present

ML Systems Research Engineer · Robotics foundation models

- Optimized distributed VLA training on an 8-GPU node, achieving 30% higher training throughput.
- Developed and profiled BF16/FP8 CUDA, Triton, and HIP implementations for π0/GR00T-class robot models across NVIDIA and AMD GPUs.

Huawei Summer 2026

Research · Continual learning / ML training systems

- Studied SDPO-based continual learning and online distillation, with attention to training throughput and reliable post-training execution.

Tools for Humanity Mar – May 2026

Machine Learning & Rust Software Engineer

- Built and instrumented real-time perception, eye-tracking, and feedback-control systems; used hardware-in-the-loop evaluation to test tracking stability on physical devices.

Progressive Robotics Oct 2025 – Mar 2026

Machine Learning & Robotics Engineer

- Reduced RL training time by 2×; developed distributed, GPU-parallel RL environments, synthetic-data generation, and transformer/RL pretraining workflows.
- Implemented CUDA/C++ geometry and physics heuristics for learned combinatorial optimization.

Delian Alliance Industries Feb – Oct 2025

Robotics Software Engineer

- Built military-grade, low-latency real-time video streaming and teleoperation infrastructure; optimized on-device perception for NVIDIA Jetson.

EDUCATION

National Technical University of Athens (NTUA) | Electrical & Computer Engineering MEng

Focus: Machine learning and robotics systems

SELECTED PUBLICATIONS, WRITING & OPEN SOURCE

MLSys 2026 (Young Professionals Symposium) — Small ML Runtime Stack for Raspberry Pi 5 QPUs · paper
NeurIPS 2025 Workshop — CORE: Full-Path Evaluation of LLM Agents Beyond Final State · paper
NeurIPS 2026 (PALM Workshop, accepted) — We Should Evaluate Agentic Memory as Inference, Not as a Tape Recorder · paper
Research essay (2026) — Reconstructive Memory for Learning Systems · article

Open source: FlashRT (AMD BF16/FP8 kernels) · Vx compiler

TECHNICAL Python · PyTorch · Distributed training / FSDP · RL · CUDA · Triton · HIP/ROCm · C++ · Rust · Nsight