

EDUCATION

University of Toronto

Bachelor of Applied Science in Computer Engineering
Dean's Honours List (All Semesters)
Top 30 ECE Student in 2023–2024

Expected May 2028

GPA: 3.91/4.0

PUBLICATIONS (*equal contribution)

- **Zhengkao Ni**, Weimin Qiu, Meng Tang. “Learning via Self-Consistency for Diffusion-based Video Reasoning.” *arXiv:2609.36826*, 2026.
- Shidu Ren*, Qilin Gu*, **Zhengkao Ni***, Junhan Sun, Jiaqi Wang, Damien Scieur, Yunze Liu. “FlexiWorld: Learning and Planning via Flexible Action Chunks Across Multiple Time Scales.” *arXiv:2609.35138*, 2026.
- Mai Ali*, **Zhengkao Ni***, Deepa Kundur. “Interpretable-MTLNet: A Kolmogorov–Arnold Network for Multitask Mental Health Prediction.” *NeurIPS 2025 Workshop BrainBodyFM*.

WORK & RESEARCH EXPERIENCE

Machine Learning Engineer Intern, Tenstorrent

May 2026 – Present

Toronto, Ontario, Canada (Hybrid)

- Brought up **DiffusionGemma**, **Wan2.2**, and **MiniMax-H3** on Tenstorrent hardware in [tt-metal](#) (1.6K-star open-source repository), taking each model from initial porting to functional inference.
- Optimized compute kernels for LLM workloads, improving throughput and latency on Tenstorrent accelerators.
- Integrated and tuned **vLLM** for Tenstorrent hardware to enable efficient, high-throughput LLM serving.

Undergraduate Researcher, University of California, Merced

Feb 2026 – Present

@ [GSI Lab](#), Prof. Meng Tang

- First-authored a paper on learning via self-consistency for diffusion-based video reasoning.
- Designed training-free test-time scaling that aggregates paths, points, and masks across MiniMax-H3 (33B) rollouts, raising visual search accuracy from 48.4% to 99.0%; early readout cuts denoising steps by over half.
- Proposed consensus-based Rejection Fine-Tuning with LoRA to distill multi-rollout consensus into a single generation, improving segmentation gIoU from 0.353 to 0.513 and 4×4 maze success from 72.0% to 84.0%.

Undergraduate Researcher, University of Toronto

Mar 2026 – Jul 2026

@ [Prof. Baochun Li's iQua Research Group](#)

- Contributed to a paper under review at **IEEE INFOCOM 2027** on reliable and efficient infrastructure for distributed multi-agent systems.
- Designed and evaluated workflow orchestration mechanisms to improve fault tolerance, recovery, and resource efficiency.

Undergraduate Researcher, University of Toronto

May 2025 – Sep 2025

@ [Kundur Research Group](#)

- Developed a KAN-based multitask model in PyTorch to detect depression and anxiety from wearable data, achieving a macro-AUROC of 0.731 and outperforming CNN-LSTM, Transformer, and GBDT baselines.
- Built a symbolic surrogate retaining 98.2% accuracy with interpretable rules (**NeurIPS 2025 Workshop**).

Research Assistant Intern, National University of Singapore

May 2024 – Aug 2024

- Fine-tuned Qwen2 and Llama 3 with LoRA as AI teaching assistants for programming contests.
- Deployed models on Alibaba Cloud ECS with vLLM behind a FastAPI service with async batch processing.