

Lucas Chen

 [example-lucas-chen](#) |  [example-lucas-chen](#) |  lucas.chen@example.com
 [lucaschen.example.com](#) |  +41 44 555 0103 |  Zurich, Switzerland

SUMMARY

AI Research Engineer specializing in multimodal learning, efficient large language models, and production-grade inference systems. Experienced in building reproducible evaluation pipelines, optimizing neural networks for GPU deployment, and translating research prototypes into reliable tools. Combines experimental rigor with systems engineering across PyTorch, CUDA, TensorRT, and cloud-native machine learning infrastructure.

WORK EXPERIENCE

VectorMind Research

Apr 2023 - Present

- *AI Research Engineer, Zurich, Switzerland*
- Built a multimodal evaluation suite for retrieval, visual question answering, grounding, and document understanding.
- Improved inference throughput by 2.8x through quantization, operator fusion, and TensorRT deployment.
- Co-authored three research papers on parameter-efficient adaptation and compact vision-language models.
- Created standardized experiment tracking and reproducibility tools adopted by four research teams.

VisionWorks AI

Aug 2021 - Mar 2023

- *Machine Learning Engineer, Remote*
- Developed computer-vision models and serving infrastructure for medical imaging and industrial inspection.
- Improved segmentation accuracy by 8.4 percentage points over the previous production baseline.
- Reduced model size by 62% using pruning, knowledge distillation, and post-training quantization.
- Deployed GPU inference services using Triton Inference Server, Docker, and Kubernetes.

PROJECTS

TinyLLM Lab

[Repository](#)

An experimentation toolkit for benchmarking quantization, knowledge distillation, and parameter-efficient fine-tuning methods on compact language models. Built with PyTorch, Transformers, PEFT, ONNX, and TensorRT.

MedVision Bench

[Repository](#)

A reproducible benchmark for evaluating medical-image segmentation under domain shift and limited-label conditions. Includes versioned datasets, experiment tracking, and standardized metrics using PyTorch, MONAI, DVC, and Weights & Biases.

CrossModal Search

[Repository](#)

A semantic search prototype that retrieves images and documents from natural-language queries using compact multimodal embeddings, FAISS indexing, FastAPI, and a React interface.

SELECTED PUBLICATIONS

- 2025 *Compact Vision-Language Models through Layer-Wise Distillation*, International Workshop on Efficient Deep Learning.
- 2024 *Evaluating Parameter-Efficient Adaptation under Domain Shift*, Workshop on Foundation Models in Practice.

EDUCATION

- 2019 - 2021 Master of Science in Artificial Intelligence, ETH Zurich. Thesis focused on efficient transformer architectures for multimodal retrieval.
- 2015 - 2019 Bachelor of Computing in Computer Science, National University of Singapore. Specialization in machine learning and computer vision.

SKILLS

Research Areas	Multimodal learning, large language models, representation learning, knowledge distillation, quantization.
ML Frameworks	PyTorch, JAX, TensorFlow, Hugging Face Transformers, MONAI.
Inference Systems	CUDA, ONNX, TensorRT, Triton Inference Server, FAISS.
MLOps & Infrastructure	Docker, Kubernetes, Weights & Biases, MLflow, DVC, Linux.
Programming Languages	Python, C++, SQL.