

Xinyang Wu

Zürich, Switzerland (Work Permit B) | xinyangwuethz@gmail.com | xinyangwuethz.github.io | 076 527 85 77
 linkedin.com/in/xinyangwu | github.com/XinyangWuEthz

Profile

- **Software Engineer focused on AI Infrastructure and Backend Systems**, with an MSc from **ETH Zürich**, building production **Python / FastAPI** platforms for grounded LLM assistants, high-throughput data, and scientific workflows.
- Production AI reliability across **routed retrieval**, **source grounding**, **human-in-the-loop evaluation**, **NDJSON streaming**, **session management**, **OpenTelemetry / Datadog observability**, and CI/CD release controls.
- Infrastructure foundation in **Docker**, **distributed multi-GPU Linux pipelines**, **dataset versioning**, and **reproducible environments**, currently building a production-style **Go / Kubernetes** service platform.

Experience

Software Engineer, Miltenyi Biotec – Zürich, Switzerland Sept 2024 – present

- Built a grounded **Documentation and Evaluation Assistant platform** across **FastAPI** and **Next.js**: **deterministic evidence routing** across 8 intents and 7 claim types under a fixed 6-document budget, streaming **NDJSON** responses with source metadata.
- Architected a production **agentic LLM workflow service** with typed API boundaries, asynchronous streaming, output validation, and robust session management, improving automated spec-generation throughput by **35%**.
- Built **human-in-the-loop evaluation infrastructure**: a versioned, content-hash-pinned golden set of **100 reviewed cases** and **100 labeled evidence chunks**, enforced by **25 regression assertions** and **32 deterministic retrieval gates** in CI.
- Designed a high-throughput ingestion and indexed-retrieval platform for biosensor time-series using **Python / FastAPI**, where storage-layer indexing and query optimization cut query latency by **10%** while serving **30** concurrent internal users.
- Established **GitLab CI / Docker** release pipelines with automated tests, linting, data validation, and guardrail verification; cut deployment cycle time by **10%** and saved **3 hours/week** of manual setup.
- Instrumented FastAPI endpoints with **OpenTelemetry / Datadog** RED dashboards for p95 / p99 latency, cutting regression triage from **2 min to 30 s**, and diagnosed production failures via stage-tagged traces, converting **46** unhandled HTTP 500 paths into typed errors across **120** call sites.

Machine Learning Engineer Intern, BOSCH – Stuttgart, Germany Mar 2024 – Aug 2024

- Engineered scalable dataset versioning and automated data-pipeline tooling on a **distributed multi-GPU Linux cluster**, cutting model-training data-preparation time by **50%** and ensuring **100% reproducibility** for downstream computer-vision stacks.
- Developed a distributed **Python / PyTorch** compute pipeline for 3D video reconstruction with Gaussian Splatting, improving GPU memory utilization by **30%** during large-scale spatial-temporal upsampling.
- Automated Docker-based environment provisioning and container validation for cross-team workloads; work productized into **first-author patent US 19/402,148** (EU / CN counterparts).

Research Assistant (Generative AI), ETH Zürich - Computer Vision Lab – Zürich, Switzerland Mar 2023 – Sept 2023

- Built a controllable **Stable Diffusion 1.5 / ControlNet** image-generation pipeline with **Hugging Face Diffusers / PyTorch**, fine-tuning trainable adapters on a **24 GB RTX 3090** and improving mean color-accuracy AUC by **3.97 points** over prior work with reproducible training and evaluation runs.

Selected Projects and Patent

- **Current Build - Go / Kubernetes Service Platform**: production-style Go REST service packaged in a multi-stage non-root Docker image and deployed with **Helm**, dev / prod values, probes, resource controls, ConfigMap / Secret separation, **HPA**, and GitHub Actions validation.
- **Trust & Safety ML Prototypes - Review Router / Abuse Signals**: content-policy routing under a fixed human-review budget over **63,978** Jigsaw test rows, and account-level abuse detection over **20K** synthetic accounts; both enforced by CI test gates.
- **Project - AI Reading Partner**: shipped a multimodal Chrome extension using the **Gemini API**, DOM extraction, backend orchestration, citation-grounded context, bounded tool calling, and live TypeScript UI updates.
- **First-Author Patent (US 19/402,148)**: *Gaussian Flow for Video Perception and Temporal Dataset Upsampling* - distributed spatial-temporal upsampling for large-scale video data; EU / CN counterparts filed.

Education

ETH Zürich, MSc in Information Technology and Electrical Engineering – Zürich, Switzerland Sept 2022 – June 2025

- Coursework: **Big Data for Engineers**, **Probabilistic AI**, Dynamic Programming and Optimal Control, Image Analysis & Computer Vision.

Sun Yat-sen University, BEng in Biomedical Engineering – Shenzhen, China Sept 2018 – June 2022

Skills

AI Infrastructure & LLM Systems: Grounded LLM assistants, RAG, question routing, retrieval policy, source attribution / coverage, evaluation harnesses, golden sets, regression gates, structured outputs, guardrail verification, Gemini and OpenAI-compatible APIs.

Backend & Distributed Systems: Python, FastAPI, REST APIs, asynchronous / NDJSON streaming, session management, SQL, indexing, query optimization, high-throughput ingestion, distributed multi-GPU Linux pipelines, dataset versioning.

Reliability & Observability: OpenTelemetry, OTEL Collector, Datadog, RED metrics, p95 / p99 latency, structured logging, runbooks, performance debugging, data validation, reproducible environments.

Cloud-Native Delivery: Docker, GitLab CI, GitHub Actions, multi-stage builds, Go, Kubernetes, Helm, HPA, health probes, ConfigMaps, Secrets, container validation, CI/CD release gates.

Languages & Core CS: Python, Go, TypeScript / JavaScript, SQL, Java, C, Bash, MATLAB; data structures, algorithms, concurrency, API design, system design, Linux debugging.

Languages

- English (C1), German (B1), Mandarin Chinese (Native), Japanese N1 (C1).