Yanda Cheng

+1-859-338-3379 | nickcp39@gmail.com | https://yanda-cheng.com | LinkedIn | Google Scholar

Education

PhD in Biomedical Engineering	*University at Buffalo (SUNY), Buffalo, NY*	GPA: 4.0/4.0	2021–2026 (Expected)
MEng in Biomedical Engineering	*Cornell University (Ivy League), Ithaca, NY*		2020–2021
BS in Electrical & Computer Engineering	*University of Kentucky, Lexington, KY*		2015–2020

Work Experience

SGLang (LLM/VLM Serving Framework) Mountain View, CA (remote)
Open-Source Contributor *Aug 2025 – Feb 2026*

- Added **Sequence Parallelism (SP)** support for **GLM-Image** in SGLang diffusion pipeline (**#18077**): enabled multi-GPU sharded execution for DiT-style diffusion, established reproducible latency/VRAM baselines, and validated behavior across SP/TP settings to unblock high-resolution generation
- Fixed high-concurrency file-descriptor leaks in HTTP utils (**#12047**) by ensuring urllib responses are always closed on success and error paths, preventing OSError (too many open files); stress-tested with 5,000 requests (100 threads) achieving 0 percent failures and p99 under 30 ms with stable FD count
- Unblocked out-of-the-box diffusion launches by fixing missing dependencies in `sglang[diffusion]` (**#17900**; related **#17671**): added **accelerate** and **ftfy** to extras, enabling `device_map="cuda/auto"` model loading for Wan2.1 and FLUX-class models in clean-room installs
- Built and shared **Qwen** performance benchmarks focusing on **p99 latency** across multiple **SP/TP** configurations, providing actionable baselines for parallelism trade-offs and regression detection

HydroSense Tech Beijing / Singapore
Co-founder & Core Engineer *2015 – Present (part-time)*

- Co-founded an IoT precision-sensing startup and led ML for field-deployed devices; scaled the product line from pilot deployments to **multi-million-dollar annual revenue** across municipal and industrial customers in 5+ countries
- Built an end-to-end deployment stack: C++ firmware on STM32-based controllers plus Python services on edge gateway, with secure OTA updates over RS-485 and Bluetooth/Wi-Fi, enabling **fleet-wide remote monitoring and configuration** for 5000+ devices
- Designed and shipped an **on-device 1D-CNN** for temperature-drift compensation using PyTorch training and INT8 quantization, reducing weight-estimation MAE by **15 percent in production** (20k+ installed sensors)
- Built an AI-driven calibration and prediction framework (Python/MLflow) to model nonlinearities and temperature dependence; reduced field drift by **30 percent** while tracking long-term device health across deployments
- **Patents:** Co-inventor on 3 granted patents covering rainfall sensing hardware and ML-based drift compensation / EMI mitigation algorithms (2017–2023)

Seno Medical New York, NY
Applied ML Engineer Intern *May 2024 – Aug 2024*

- Collaborated with clinicians to curate 2,000+ medical imaging QA datasets with annotation guidelines; fine-tuned LLaMA (LoRA) and built a RAG system, cutting hallucination by 50 percent and raising Recall at 5 to 90 percent
- Designed an end-to-end evaluation harness with label schema (factuality, coverage, style), automatic metrics, and Python error-analysis scripts; reduced hallucination rate from 30 percent to 15 percent and increased Recall at 5 to 90 percent
- Implemented a modular RAG backend: chunking and ingestion jobs, FAISS-based vector store, retriever/re-ranker layer, and LLM orchestration, served via FastAPI and containerized with Docker for PHI-compliant on-prem deployment
- Set up Git-based workflows and observability (structured logs, latency/error-rate dashboards); shipped reproducible Docker images and deployment runbooks for downstream clinical teams

Roswell Park Comprehensive Cancer Center Buffalo, NY
Research Scientist *May 2023 – Aug 2023*

- Designed a multi-modal data pipeline aligning imaging with structured text reports (de-identification, schema normalization, automated QC) to produce clean image–text pairs for generative modeling
- Built dataset generation jobs (filtering, sampling, augmentation, metadata tagging) to create balanced image–report datasets for conditional generation and downstream training
- Implemented a domain-specific latent diffusion model (Stable Diffusion–style, DDIM/DPMS sampling), trained on AWS, for conditional image-to-text report generation and synthetic data augmentation
- Automated GPU training and experiment tracking with standardized configs, logging, and comparison scripts, enabling fast iteration over model variants and dataset settings on the AWS stack

Selected Research Projects

Predictive modeling of chronic foot ulcer outcomes using longitudinal photoacoustic imaging *npj Imaging, 2026*

- Designed longitudinal follow-up protocol for over 400 diabetic foot ulcer patients; coordinated **IRB approval**, multi-center patient enrollment, and data collection with endocrinology, vascular surgery, and wound care teams
- Combined radiomics features with LASSO/Random Forest models achieving AUC over 0.85 for quantitative ulcer healing prediction, directly supporting clinical decision-making
- *Applicable to:* longitudinal outcome prediction across medical imaging modalities, risk stratification in clinical cohorts, and time-series modeling in CV/ML pipelines

OneTouch automated photoacoustic and ultrasound imaging of breast in standing pose*IEEE Trans. Medical Imaging, 2025, cite*

- Co-developed a multi-modal clinical imaging platform integrating AI-assisted diagnosis for early breast cancer screening; coordinated with radiologists and surgeons on clinical protocol design and image interpretation validation
- *Applicable to:* multi-modal medical imaging, computer vision for cancer screening, and deployment of AI tools in radiology workflows

Dual-scan photoacoustic tomography for the imaging of vascular structure on foot*IEEE Trans. UFFC, 2023, cited 21 times*

- Developed quantitative 3D vascular imaging protocol for peripheral arterial disease (PAD) assessment; established clinical imaging standards in collaboration with a vascular surgery team
- *Applicable to:* quantitative vascular imaging, angiography analysis, and 3D vessel segmentation in medical imaging and CV applications

Selected Publications (First / Co-Author, Peer-Reviewed)

- **Cheng Y.**, Huang C., et al. Predictive modeling of chronic foot ulcer outcomes using longitudinal photoacoustic imaging. *npj Imaging* 4(1), 12, **2026**.
- **Cheng Y.**, Huang C., Bing R.W., et al. Dysphagia assessment based on photoacoustic imaging: a pilot ex vivo and in vivo study in infant swine. *Med-X* 3(1), 1–9, **2025**.
- **Cheng Y.**, Zheng W., et al. Unsupervised denoising of photoacoustic images based on the Noise2Noise network. *Biomed. Optics Express* 15(8), 4390–4405, **2024**. (14 citations)
- **Cheng Y.**, Huang C., et al. Dual-scan photoacoustic tomography for the imaging of vascular structure on foot. *IEEE Trans. UFFC* 70, **2023**. (21 citations)
- Huang C., **Cheng Y.**, et al. Enhanced clinical photoacoustic vascular imaging through skin localization and adaptive weighting. *Photoacoustics* 42, 100690, **2025**. [4 citations]
- Zhang H., ..., **Cheng Y.**, et al. OneTouch automated PA/US breast imaging in standing pose. *IEEE Trans. Medical Imaging*, **2025**. (8 citations)
- Liu X., **Cheng Y.**, et al. Simultaneous tissue blood flow and oxygenation with a wearable fiber-free optical sensor. *J. Biomed. Optics* 26(1), 012705, **2021**. **[38 citations]**

Total: about 100 citations (Google Scholar); 11 SCI papers; Venues: npj Imaging, IEEE TMI, IEEE TUFFC, Photoacoustics, BOE, Med-X, JBO

Technical Skills

Medical Imaging: Image reconstruction, denoising (DL/classical), registration, segmentation, 3D visualization; CT, PA, US, multiphoton microscopy
AI / Machine Learning / LLMs: CNN, UNet, Transformer, LSTM, Radiomics, LASSO, SVM, Random Forest, deep learning denoising; large language models, RAG systems, prompt engineering, multimodal models
Clinical Data Analysis: Longitudinal analysis, Kaplan-Meier survival analysis, ROC/AUC, regression, statistical testing (Python / R / SPSS)
Research Workflow: IRB/ethics protocol submission, multi-site data management, SCI manuscript preparation & submission, academic presentation
Programming: Python, C/C++, MATLAB, R, SQL; PyTorch, scikit-learn, OpenCV, Pandas, Docker
Languages: Mandarin Chinese (native), English (fluent – academic writing, cross-national team collaboration)

Awards & Honors

- **$3,000 Startup Seed Grant**, University at Buffalo Entrepreneurship Award – PA-based Breast Cancer Detection (2023)
- Key Contributor, NIH-Funded Research Project (PI: Prof. Jun Xia, University at Buffalo; ongoing)
- Outstanding Teaching Assistant – BME 503 Image Processing & BME 302 Medical Devices, University at Buffalo

Yanda Cheng

+1-859-338-3379 | nickcp39@gmail.com | https://yanda-cheng.com | LinkedIn | Google Scholar

教育背景

博士生物医学工程 纽约州立大学布法罗分校，美国纽约州布法罗 平均绩点 4.0 / 4.0 2021–2026（预计毕业）

工程硕士生物医学工程 康奈尔大学，美国纽约州伊萨卡 2020–2021

学士电气与计算机工程 肯塔基大学，美国肯塔基州列克星敦 2015–2020

工作经历

SGLang（大语言模型与视觉语言模型推理框架） 美国加州 Mountain View（远程）
开源贡献者 *2025* 年 *8* 月 – *2026* 年 *2* 月

- 为 SGLang 扩散推理流程中的 **GLM-Image** 实现 **Sequence Parallelism（SP）**多 GPU 并行支持（**#18077**），使 DiT 风格扩散模型能够进行多 GPU 分片执行，并建立可复现的推理延迟与显存占用基线，验证不同 SP 与 TP 配置下的行为表现，从而支持高分辨率图像生成
- 修复 HTTP 工具中的高并发文件描述符泄漏问题（**#12047**），通过确保 urllib 响应在成功与异常路径中均被正确关闭，避免出现 OSError（too many open files）；在 5000 次请求和 100 线程压力测试下实现 0 失败，且 p99 延迟低于 30 ms，文件描述符数量保持稳定
- 通过修复 `sglang[diffusion]` 的缺失依赖问题（**#17900**，关联 **#17671**），补充 **accelerate** 与 **ftfy** 依赖，使 Wan2.1 与 FLUX 类模型能够在全新环境中开箱即用，并支持 `device_map="cuda/auto"` 的模型加载方式
- 构建并分享 **Qwen** 模型在多种 **SP/TP** 并行配置下的性能基准测试，重点关注 **p99 延迟**，为并行策略取舍与性能回归检测提供可操作的基线

HydroSense Tech 北京 / 新加坡
联合创始人兼核心工程师 *2015* 年 – 至今（兼职）

- 联合创立一家物联网高精度传感创业公司，负责现场部署设备的机器学习系统；推动产品线从试点部署扩展至多个国家的市政与工业客户，形成数百万美元级年收入规模
- 构建端到端部署体系，包括基于 STM32 控制器的 C++ 固件与边缘网关上的 Python 服务，并通过 RS-485 与 Bluetooth/Wi-Fi 实现安全 OTA 更新，支持 5000 台以上设备的远程监控与统一配置管理
- 设计并落地用于温度漂移补偿的 **端侧 1D-CNN** 模型，使用 PyTorch 训练并进行 INT8 量化部署，使重量估计的平均绝对误差在生产环境中降低 **15 percent**（覆盖 2 万台以上已部署传感器）
- 构建基于 Python 与 MLflow 的 AI 驱动校准与预测框架，用于建模非线性误差与温度依赖关系；在实际部署中将设备现场漂移降低 **30 percent**，同时实现长期设备健康状态跟踪
- **专利：**作为共同发明人拥有 3 项已授权专利，覆盖雨量传感硬件设计、基于机器学习的漂移补偿以及 EMI 抑制算法（2017–2023）

Seno Medical 美国纽约州纽约市
应用机器学习工程师实习生 *2024* 年 *5* 月 – *2024* 年 *8* 月

- 与临床医生合作整理 2000 条以上医学影像问答数据，并制定标注规范；完成 LLaMA 模型的 LoRA 微调与 RAG 系统构建，使 hallucination 降低 50 percent，并将 Recall at 5 提升至 90 percent
- 设计端到端评估框架，包括标签体系（事实性、覆盖度、风格）、自动化指标与 Python 误差分析脚本；将 hallucination 率从 30 percent 降低至 15 percent，并将 Recall at 5 提升至 90 percent
- 实现模块化 RAG 后端，包括文本切分、数据摄取任务、基于 FAISS 的向量存储、检索与重排序层以及 LLM 编排系统，并通过 FastAPI 提供服务、利用 Docker 完成容器化部署，以满足 PHI 合规的本地部署需求
- 建立基于 Git 的协作流程与系统可观测性方案（结构化日志、延迟与错误率监控面板），并为后续临床团队交付可复现的 Docker 镜像与部署文档

Roswell Park **综合癌症中心** 美国纽约州布法罗
研究科学家 *2023* 年 *5* 月 – *2023* 年 *8* 月

- 设计多模态数据流程，将医学影像与结构化文本报告进行对齐（包括去标识化、数据模式规范化与自动质量控制），以生成用于生成式建模的高质量图像文本配对数据
- 构建数据集生成任务，包括过滤、采样、增强与元数据标注，从而形成用于条件生成与下游训练的平衡图像报告数据集
- 实现面向特定医学领域的潜空间扩散模型（Stable Diffusion 风格，采用 DDIM 与 DPMS 采样），并在 AWS 上完成训练，用于条件式图像到文本报告生成以及合成数据增强
- 通过标准化配置、日志记录与对比脚本，实现 GPU 训练与实验跟踪自动化，从而支持不同模型变体与数据集设置的快速迭代

代表性研究项目

基于纵向光声成像的慢性足部溃疡结局预测建模 *npj Imaging*，*2026*

- 为 400 名以上糖尿病足溃疡患者设计纵向随访方案；协调 **IRB 审批**、多中心患者招募以及与内分泌科、血管外科和创面护理团队的数据采集工作
- 将 radiomics 特征与 LASSO 及 Random Forest 模型结合，实现 AUC 超过 0.85 的定量溃疡愈合预测，直接支持临床决策
- 适用方向：适用于跨医学影像模态的纵向结局预测、临床队列风险分层以及 CV 和 ML 流程中的时间序列建模

站立位乳腺自动化光声与超声成像系统 *IEEE Trans. Medical Imaging*，*2025*，被引用 *8* 次

- 共同开发多模态临床成像平台，并集成 AI 辅助诊断能力，用于乳腺癌早期筛查；与放射科医生和外科医生协同完成临床方案设计与图像解读验证
- 适用方向：适用于多模态医学影像、癌症筛查中的计算机视觉方法，以及 AI 工具在放射学工作流中的部署

双扫描光声断层成像用于足部血管结构成像 *IEEE Trans. UFFC*，*2023*，被引用 *21* 次

- 开发用于外周动脉疾病（PAD）评估的定量三维血管成像方案；与血管外科团队合作建立临床成像分析标准
- 适用方向：适用于定量血管成像、血管造影分析以及医学影像与计算机视觉中的三维血管分割

代表性论文（第一作者或共同作者，同行评审）

- **Cheng Y.**, Huang C., et al. Predictive modeling of chronic foot ulcer outcomes using longitudinal photoacoustic imaging. *npj Imaging* 4(1), 12, **2026**.
- **Cheng Y.**, Huang C., Bing R.W., et al. Dysphagia assessment based on photoacoustic imaging: a pilot ex vivo and in vivo study in infant swine. *Med-X* 3(1), 1–9, **2025**.
- **Cheng Y.**, Zheng W., et al. Unsupervised denoising of photoacoustic images based on the Noise2Noise network. *Biomed. Optics Express* 15(8), 4390–4405, **2024**. （14 次引用）
- **Cheng Y.**, Huang C., et al. Dual-scan photoacoustic tomography for the imaging of vascular structure on foot. *IEEE Trans. UFFC* 70, **2023**. （21 次引用）
- Huang C., **Cheng Y.**, et al. Enhanced clinical photoacoustic vascular imaging through skin localization and adaptive weighting. *Photoacoustics* 42, 100690, **2025**. （4 次引用）
- Zhang H., ..., **Cheng Y.**, et al. OneTouch automated PA/US breast imaging in standing pose. *IEEE Trans. Medical Imaging*, **2025**. （8 次引用）
- Liu X., **Cheng Y.**, et al. Simultaneous tissue blood flow and oxygenation with a wearable fiber-free optical sensor. *J. Biomed. Optics* 26(1), 012705, **2021**. （38 次引用）

总计：Google Scholar 约 100 次引用；SCI 论文 11 篇；期刊包括 npj Imaging、IEEE TMI、IEEE TUFFC、Photoacoustics、BOE、Med-X、JBO

技术技能

医学影像：图像重建、降噪（深度学习与传统方法）、配准、分割、三维可视化；CT、PA、US、多光子显微成像
人工智能 / 机器学习 / 大语言模型： CNN、UNet、Transformer、LSTM、Radiomics、LASSO、SVM、Random Forest、深度学习降噪；大语言模型、RAG 系统、提示词工程、多模态模型
临床数据分析：纵向分析、Kaplan-Meier 生存分析、ROC/AUC、回归分析、统计检验（Python、R、SPSS）
科研流程： IRB 与伦理审批流程、多中心数据管理、SCI 论文撰写与投稿、学术汇报
编程： Python、C/C++、MATLAB、R、SQL；PyTorch、scikit-learn、OpenCV、Pandas、Docker
语言：中文（母语）、英文（流利，可进行学术写作与跨国团队协作）

奖励与荣誉

- **3000 美元创业种子基金**，University at Buffalo 创业奖，用于光声乳腺癌检测项目（2023）
- NIH 资助科研项目核心成员（项目负责人：Jun Xia 教授，University at Buffalo，项目持续进行中）
- 优秀助教，University at Buffalo，BME 503 图像处理课程与 BME 302 医疗设备课程