Cenwei Zhang

cwzhang2001@gmail.com | blog.cenweizhang.github.io | github.com/cenweizhang

Educations

The Chinese University of Hong Kong | Information Engineering *Postgraduate* — Sep. 2026 (Expected)

Shanghai Jiao Tong University | BME (EE Track) *Undergraduate* — Sep. 2022 — Jun. 2026

Skills

- **Languages&Workflow**: Python, C++; proficient in the PyTorch deep learning framework. Linux, Shell, Git, GitHub, Docker.
- **Technical Foundation**: Familiar with the theory of **Generative Modeling** (Diffusion & Flows) as well as manifold optimal transport mechanisms; familiar with visual image processing model components and their variants, such as ViT, CNN, and DiT; familiar with basic LLM frameworks, including compression, quantization, and inference optimization methods; familiar with fundamental VLM frameworks.

Work Experience

Ubiquant-IQuest Research |*LLM algorithm intern* — Apr. 2026 — Present

- Conducted research on trustworthy reasoning for multimodal LLMs, exploring a **Best-of-Evidence** (BoE) framework that decomposes VLM candidates into verifiable evidence factors, selectively verifies high-value evidence under limited budgets, and updates candidate credibility. The work connects BoN, PRM, and self-consistency methods, with applications to multimodal QA, decision systems, and Flow/Diffusion image generation, especially in clinical medical scenarios.

DolphinAI Co., Ltd. |*Algorithm Research Intern | Medical Image Generation* — Apr. 2025 — Dec. 2025

- Tackled high-frequency noise and semantic decoupling by integrating **structure-aware** strategies into the generative backbone, effectively filtering artifacts while preserving the **Manifold Consistency** of fine-grained pathological features.

Selected Publications

MedCore: Boundary-Preserving Medical Core Pruning for MedSAM |*Preprint*

Cenwei Zhang, *Suncheng Xiang*†, *Lei You*†

- Proposed MedCore, a structured pruning framework for MedSAM that mitigates boundary degradation under compression by preserving both SAM-to-MedSAM adaptation-critical structures and high-boundary-leverage components. MedCore combines dual-intervention scoring (zeroing vs. resetting groups to SAM weights) with boundary-aware Fisher estimation, and introduces a boundary leverage principle explaining compression-induced boundary shifts through logit perturbations and boundary gradients. On polyp segmentation benchmarks, MedCore achieves about **60%** parameter reduction and **58.4%** FLOPs reduction while maintaining Dice **0.9549**, BF1 **0.6388**, and HD95 **5.14**; an extreme setting further reaches **86.6%** parameter reduction with strong boundary quality.

From Baselines to Transport Geodesics: Axiomatic Attribution via Optimal Generative Flows |*Preprint*

Cenwei Zhang*, *Lin Zhu**, *Manxi Lin*, *Lei You*†

- Co-developed the spatiotemporal gait analysis model GraphGaitNet, fusing anatomy-based Graph Convolutional Networks (GCN) and Transformer temporal modeling, achieving a high diagnostic accuracy of **96.8%** for Parkinson's disease.
- Proposed a manifold Shapley attribution framework based on optimal generative flows, decomposing attribution into fixed-path credit allocation and data-transport path selection. We prove the uniqueness of the Aumann–Shapley line integral and approximate OT paths with RF/Reflow, deriving stability bounds that link vector-field error to attribution error. Experiments show that low-action, transport-consistent paths produce more stable and structured explanations while maintaining competitive deletion faithfulness.

Anatomically-Informed GNN on Ground Reaction Force for Parkinson's Disease Diagnosis |***Accepted to IEEE JBHI*** *(CCF-C)*

Wenhao Li, ***Cenwei Zhang***, *Shi Chang*, *Jingtong Zhao*, *Guanning Lin*†

- Co-developed the spatiotemporal gait analysis model GraphGaitNet, fusing anatomy-based Graph Convolutional Networks (GCN) and Transformer temporal modeling, achieving a high diagnostic accuracy of **96.8%** for Parkinson's disease.

Projects

LLM Pre-training and Fine-tuning System based on nanoGPT |*Independent Project & Implementation*

Jan. 2026 — Mar. 2026

- Independently implemented the GPT-2 foundation model from scratch using PyTorch, integrating modern architectures like RoPE and the **FlashAttention** acceleration operator. Executed end-to-end pre-training on a single H100 GPU utilizing `bfloat16` mixed precision, achieving a 31% Zero-shot accuracy on the downstream HellaSwag commonsense reasoning task, outperforming the original GPT-2 baseline by approximately 10%.

Personal Summary

- Professional working proficiency in English (IELTS 7.0).
- **Personal Life**: Passionate about city walks, hiking, and amateur photography. I enjoy honing logical deduction and task planning skills through Go (inspired by AlphaGo), and often draw inspiration for constructing digital ”world models” from the open-world environment of Genshin Impact.

张岑蔚

cwzhang2001@gmail.com | blog.cenweizhang.github.io | github.com/cenweizhang

教育经历

香港中文大学 | 计算机科学与技术 (信息工程类)| 硕士研究生 2026.09(预计入学)

上海交通大学 | 生物医学工程 (电子信息类)| 工学学士 2022.09—2026.06

技术能力

- **语言 & 工作流**: 常用 Python,C++；熟悉深度学习框架 Pytorch。熟悉 Linux, Shell, Git, GitHub，docker。
- **工程底座**: 熟悉 **Generative Modeling** (Diffusion & Flows) 理论及流形最优传输机制；熟悉 **ViT，CNN，DiT** 等视觉图像处理的模型构件及其变体；熟悉基础 LLM 框架及其压缩与量化和推理优化方法；熟悉基础的 vlm 框架。

工作经历

九坤投资至知创新研究院 | 大模型算法实习生 2026.04 至今

- 围绕多模态大模型可信推理开展研究，探索 Best-of-Evidence（BoE）框架，将 VLM 多候选输出拆解为可验证证据因子，在有限预算下选择高价值证据校验并更新候选可信度，分析其在多模态 QA、决策系统及 Flow/Diffusion 图像生成, 尤其是医疗临床场景中的应用。

DolphinAI **海豚智声有限公司** | 算法研究实习生 / 图像生成 2025.04—2025.12

- 深入剖析医学影像的数据分布特性，针对高频噪声扰动与细粒度语义解耦难题，在生成式底座中探索 Structure-aware 策略。确保模型在滤除高频伪影的同时，严格保持病理级细粒度特征的流形一致性（Manifold Consistency）。

科研论文

MedCore: Boundary-Preserving Medical Core Pruning for MedSAM |*Preprint*

Cenwei Zhang, *Suncheng Xiang*†, *Lei You*†

- 提出一种面向 MedSAM 的结构化剪枝框架，针对医学分割中 “Dice 保持较高但边界质量退化” 的压缩风险，联合保留 SAM→MedSAM 适应过程中关键结构与高边界杠杆结构。方法通过双重干预评分比较结构归零与重置为原始 SAM 权重的影响，并结合边界感知 Fisher 估计识别边界敏感组件；进一步提出边界杠杆原理，从 logit 扰动与边界梯度角度解释压缩导致的边界位移。在息肉分割基准上，MedCore 在恢复微调后实现约 **60%** 参数压缩与 **58.4%** FLOPs 降低，同时达到 Dice **0.9549**、BF1 **0.6388**、HD95 **5.14**；极限设置下可实现 **86.6%** 参数缩减并保持较强边界质量。分析表明 MedSAM 处于头部脆弱边界区，注意力头剪枝的边界杠杆显著高于 MLP，解释了边界指标退化机制。

From Baselines to Transport Geodesics: Axiomatic Attribution via Optimal Generative Flows |*Preprint*

Cenwei Zhang*, *Lin Zhu**, *Manxi Lin*, *Lei You*†

- 提出基于最优生成流的流形 Shapley 归因框架，将归因分解为固定路径贡献分配与数据传输路径选择；证明 Aumann–Shapley 线积分唯一性，并以 RF/Reflow 近似 OT 路径，推导了将向量场误差与归因误差联系起来的稳定性界。实验表明，低作用量且传输一致的路径能够产生更加稳定且结构化的解释，同时保持有竞争力的删除忠实性。

Anatomically-Informed GCN on Ground Reaction Force for Parkinson’s Disease Diagnosis |***CCFC-JBHI*** *Accept*

Wenhao Li, ***Cenwei Zhang***, *Shi Chang*, *Jingtong Zhao*, *Guanning Lin*†

- 参与研发时空步态分析模型 GraphGaitNet，融合基于解剖学的图卷积网络与 Transformer 时序建模，在帕金森病诊断任务中实现 **96.8** 的高准确率。

项目经历

基于 nanoGPT 的 LLM 模型预训练微调 | 独立项目与实现 2026.01—2026.03

- 独立基于 PyTorch 手写复现 GPT-2 底座，融入 RoPE 及 **FlashAttention** 加速算子，借助 `bfloat16` 混合精度于单张 H100 跑通端到端预训练；在下游 HellaSwag 常识推理任务中实现 31% 的 Zero-shot 准确率，较原版模型提升约 10%。

个人总结

- 可以使用英语进行无障碍工作交流（IELTS 7.0）。
- 热爱 Citywalk & hiking；amateur photographer；喜欢在围棋中的逻辑推演与任务规划 (from AlphaGO)；Genshin Impact 开放世界也是构建灵感的过程。~