梁广

+86 189-7573-2667
LIANGG@LAMDA.NJU.EDU.CN
个人主页 谷歌学术
LINKEDIN GITHUB

研究兴趣

- 强泛化与高效能模型设计
  涵盖 FP8/FP4 训练，量化感知训练、更高的数据效率的优化器等前沿技术的设计与实现，旨在在相同训练计算资源下，大幅提高模型智能能力和性能。
- 大模型的高效训练与推理
  涵盖 LLM、VLM 等基础模型的 Pretrain、Midtrain、SFT、RL。以更快的训练速度（FP8 模型训练，千卡集群预训练）和更强大的模型性能（提升模型泛化性能）更低廉的部署成本（模型量化和蒸馏）。

教育背景

北京中关村学院 (ZGCA), 北京 2024 年 9 月 – 至今

联合培养博士生

- 学院/项目导师: **刘铁岩**教授 (北京中关村学院党委书记、院长)

南京大学, 江苏南京 2024 年 9 月 – 至今

博士研究生，计算机科学与技术 **LAMDA** 实验室 (由**周志华**院士领导)

- 导师: **吴建鑫**教授 (CVPR 2024 程序委员会主席)

西安交通大学, 陕西西安 2020 年 9 月 – 2024 年 6 月

工学学士 人工智能学院 (人工智能拔尖人才试验班)，钱学森书院

实习经历

上海人工智能实验室 (Shanghai AI Laboratory), 中国上海

***OpenDATALab* 实习生** *(负责人: 何聪辉)* 指导老师: 王斌

主要负责更高效的 OCR 模型和算法的设计与研发。

- DocLayout-YOLO (2k ★): 共同第一作者，负责模型架构创新。
- MinerU (55k ★): 核心贡献者 (负责公式识别模块 & 参与高效的 VLM 训练)。
- UniMERNet (450 ★): 模型架构和实验 & 共同第一作者 (核心架构创新，提出 R-S 注意力机制，主要实验运行)。

项目经历

北京中关村学院科研项目, 中国北京 2025 年 11 月 – 至今

项目学生负责人 指导老师: 刘铁岩、郑书新、何纪言、吴建鑫

主要致力于构建面向下一代算力的通用 FP8/FP4 大模型高效训练与推理框架。

- **核心动机与创新**：针对大模型低精度训练中“激活值异常 (Outliers)”导致的崩溃问题，跳出传统分块量化的工程路径，旨在从数学层面“物理消除”异常值，大幅提高大模型吞吐量，并初步在数百卡的集群上验证。
- **软硬协同设计**: 基于前期 A100 上的基础模型验证经验，目前正针对下一代 GPU 集群 Hopper (H100)、Blackwell (RTX 5090/B200) 架构进行算法与底层硬件的协同设计，探索小值域算力下的智能涌现极限。

发表论文

(= 表示共同第一作者)

- **TWEO: Transformers Without Extreme Outliers Enables FP8 Training And Quantization For Dummies**
  系统性解决了 Transformer 激活值中的离群值问题，显著增强了在通用模型预训练中的稳定性和训练速度并降低了量化损失。并展现出了强大的领域通用性，在视觉基础模型和语言基础模型上都验证了吞吐量的大幅提高和推理性能的大幅提升。
  Guang Liang, Jie Shao, Ningyuan Tang, Xinyao Liu, Jianxin Wu
  CVPR 2026 | arXiv:2511.23225

- **GPLQ: A General, Practical, and Lightning QAT Method for Vision Transformers**
  提出“激活先行”的 ViT 量化框架，速度较传统方法提升 100 倍，并展现出 SOTA 级别的泛化性能。
  Guang Liang, Xinyao Liu, Jianxin Wu
  NeurIPS 2025 | Camera Ready | arXiv:2506.11784 | 代码
- **Unimernet: A universal network for real-world mathematical expression recognition**
  提出一种通用的真实世界数学公式识别网络，在多种复杂场景下实现了高精度的公式解析与识别。
  作为共一作者设计了模型并运行了大部分实验
  Zhuangcheng Gu=, **Guang Liang**=, Bin Wang, Chao Xu, Bo Zhang, Botian Shi, Conghui He
  CVPR 2026 | arXiv:2404.15254 | 代码

- **MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing**
  1.2B 参数的文档解析模型。在复杂公式、表格和密集文本上实现了 SOTA 级别的识别精度。
  Co-author in MinerU Team
  Tech Report | arxiv: 2509.22186 | HuggingFace

在投论文

(= 表示共同第一作者)

- **DocLayout-YOLO: Towards Robust and Efficient Document Layout Analysis through Diverse Document Synthesis and Perception Enhancement**
  通过多样化的文档合成和感知增强技术，提出了一种鲁棒且高效的文档版面分析模型，极大提升了对复杂版面元素的泛化解析能力。
  Zhiyuan Zhao=, Hengrui Kang=, **Guang Liang**=, Linke Ouyang, Weijia Li, Bin Wang, Conghui He
  在投 | 代码
- **MontageAug: Enhancing Long-tail Robustness And Semantic Consistency of VLMs**
  提出一种针对 VLM 语义一致性的训练中图像增强方法，有效提升了 VLM 在**长尾数据**分布下的性能。并展现出通用有效性，在医学眼底彩照领域/OCR(公式识别) 等多种 VLM 任务上取得显著提升，尤其是难例少样本。
  Xinyao Liu=, **Guang Liang**=, Yanlin Qu, Huixun Jia, Xiaodong Sun, Diping Song
  在投

获奖情况

冠军 ***(最佳解决方案奖，奖金 20 万日元)***, MVA 2025 挑战赛 2025 年

- Small Multi-Object Tracking for Spotting Birds (SMOT4SB) 挑战赛 冠军 (排名 1/78 支队伍，击败其余队伍 308 次提交)。[获奖证书] [论文] [代码]
- 共同第一作者、通讯作者。作者: Xiang Yu=, **Guang Liang**=, Xinyao Liu=。

冠军 ***(一等奖，奖金 1000 美元)***, TRAUMATHOMPSON@MICCAI 挑战赛 2023 年

- 在两个竞赛赛道均获得第一名 (超过 20+ 支队伍参加):
  - 动作识别 (Action Recognition) [排行榜] & 动作预测 (Action Anticipation) [排行榜]
- 与 Xinyao Liu 为共同第一作者。

胡保生奖学金 2023 年
邱昌荣奖学金 2022 年
优秀学生 2022 年
省级一等奖, 全国大学生数学建模竞赛 2022 年
校级奖学金 2021 年

其他经历

审稿人: CVPR 2026 ECCV 2026 2025 年 – 至今
助教: 模式识别与计算机视觉 (课程主页) 2025 年春季
助教: 从 *0* 到 *1* 构建大模型 (课程老师主页) 2026 年春季