



重庆大学 / 计算机科学与技术 (卓越专业)

个人主页 [/ xiaowei.hua.cn](https://xiaowei.hua.cn)

本科在读 / 2023.09 - 2027.06

GitHub [/ Will-hxw](https://github.com/Will-hxw)

GPA 3.64/4.00 · 排名 35/309

邮箱 [/ xiaowei.huacqu@gmail.com](mailto:xiaowei.huacqu@gmail.com)

均分 88.69 · 综合测评年级前 3% · TOEFL 110

电话 / 微信 [/ 15730810078](https://t.me/15730810078)

实习经历

Tencent 腾讯 · TRS 推荐中台 / 推荐系统基础架构中心

向量数据库实习生 2025.06 - 2025.08

C++ / CUDA / Python / Golang / Faiss / CUDA VMM

overview

参与面向推荐召回与 RAG 场景的 GPU 向量检索基础设施建设, 围绕大 top-k selection、实时向量增删改、在线 batch 调度和 CUDA VMM 显存管理, 解决大规模向量召回中的吞吐、尾延迟、动态更新与多索引部署稳定性问题。

optimize

针对 Faiss GPU 在大 top-k 检索中 selection 阶段计算开销快速上升、原有实现难以适配分块距离结果的问题, 参考 RadiK 的 radix selection 思路, 重构距离分块与 top-k selection 的衔接流程; 在 top-k=2048 场景下将 GPU 检索 QPS 提升约 13x, P99 延迟由秒级降至百毫秒级。

realtime

参与 Faiss GPU 实时向量增删改能力建设, 基于 CUDA VMM 管理可重映射物理显存页, 并通过哈希索引维护文档 ID 与向量槽位之间的映射关系, 减少删除和更新触发的数据压缩及批量内存搬移, 将单条文档删除延迟由毫秒级优化至微秒级。

infra

参与 GPU 检索 batch 调度组件设计, 使用无锁队列承接高并发请求, 由专用 batch 线程聚合零散单-query 请求, 并通过线程池执行检索任务, 在控制排队延迟的同时扩大有效 batch, 提高 GPU 利用率与系统吞吐。

memory

参与 CUDA VMM 显存管理优化, 通过虚拟地址预留、物理页按需映射和分块距离计算降低索引扩容期间的瞬时显存峰值, 支持同卡多索引实例稳定运行, 减少扩容及数据更新阶段的 OOM 风险。

科研经历

跨视角几何一致性驱动的稀疏视角 3D Gaussian Splatting

科研经历 2025.08 - 至今

PyTorch / CUDA / COLMAP / 3D Gaussian Splatting 跨视角几何一致性 · 高斯置信度建模 · 生命周期优化 · 新视角合成

problem

研究稀疏输入条件下 3DGS 的几何欠约束问题。原始 3DGS 主要依据图像梯度、尺度和不透明度控制 densification 与 pruning, 难以区分表达不足的真实结构与缺少多视角支持的错误几何, 易产生漂浮高斯、边界破碎和纹理泄漏。

evidence

构建高斯级观测证据模型, 为每个 Gaussian 统计跨视角可见次数、相机基线覆盖、重投影残差、深度一致性及渲染贡献; 对各项证据归一化并加权融合, 得到几何支持度与不确定性评分, 衡量其跨视角可信程度。

optimization

提出置信度与重建残差联合驱动的生命周期优化策略: 对高支持度、高残差区域执行 split/clone, 补充欠表达的场结构; 对低支持度、高不一致性高斯进行 opacity decay 与 pruning; 结合图像与深度边缘执行各向异性分裂, 减少跨边界纹理扩散。

result

相较原始 3DGS, 在多组真实场景及不同输入视角下稳定提升 PSNR、SSIM 与边界区域重建质量, 明显减少漂浮伪影、边界断裂和前景纹理泄漏, 并改善高斯数量演化与新视角渲染的稳定性。

项目经历

Synthos-QQ · IM 消息语义检索与整合 Agent

GitHub [项目唯一负责人](#) 2026.02 - 至今

Agent / RAG / bge-m3 / TypeScript / Node.js / MongoDB / React

problem

针对 IM 消息碎片化、话题交织及跨时段信息关联困难的问题, 构建 Agentic Retrieval 系统, 使用户能够围绕兴趣获取有效信息整合、恢复事件上下文并追踪话题演化。

retrieval

设计查询规划—混合召回—相关性重排—上下文构建链路; 由 Agent 识别问题意图并生成时间、群组与话题约束, 基于 bge-m3 融合稠密与稀疏检索, 经 reranker 筛选和聚合跨时段证据, 提升复杂查询与多跳信息的召回质量。

agent

将语义检索、结构化筛选、话题追踪和摘要生成封装为可调工具, 引入证据充分性判定与 corrective retrieval; 在证据缺失或冲突时自动改写查询并迭代检索, 生成带原始消息定位的回答与结论。

system

设计消息—事件—主题三级知识组织与增量更新机制, 配合任务编排完成数据同步、索引更新和周期性知识合成, 实现跨时段问答、话题演化追踪及可追溯摘要生成。

🏆 学科竞赛与奖学金

算法 · 数学 · 综合

- 2025 年 ACM-ICPC 国际大学生程序设计竞赛 银牌
- 2026 年 团体程序设计天梯赛 个人三等奖
- 2025 年 百度之星程序设计大赛 银奖
- 2021 年 高中数学联赛 江西省一等奖
- 2023 年 全国大学生数学竞赛 重庆赛区一等奖
- 2025 年 国家励志奖学金
- 校级甲等奖学金 3 次 (综合评测年级前 3%)

🔧 技术能力与研究兴趣

🔧 技术能力

数学与 Coding · 专业基础 · 科研能力 · 工程能力

- **数学与 Coding**: 具备手搓算法、复杂代码调试、论文公式推导与实验验证能力; 获高中数学联赛省一等奖、全国大学生数学竞赛省一等奖及 ICPC 银牌。
- **专业基础**: 熟悉 C++ / Python / PyTorch / CUDA / TypeScript, 掌握数据结构、OS、计组、计网与数据库; 了解优化算法、Transformer、Pre-training、SFT、RLHF / DPO、RAG、Agent 与 AI Infra。
- **科研能力**: 能够系统检索和阅读论文, 从已有工作中提炼问题、积累 idea 与 insight, 并通过实验迭代验证和改进方案; 具备英文论文阅读与技术写作能力。
- **工程能力**: 具有复杂系统开发、架构设计与性能调优经验; 参与 GPU 向量检索、动态索引、batch 调度和显存管理优化, 独立完成 RAG / Agent 全链路系统。

🎯 AI Infra · Agent Systems · 具身智能

研究兴趣 · 未来方向

- **研究兴趣**: 聚焦 AI Infra、Agent Systems 与具身智能。AI Infra 方面, 关注大模型训练与推理效率、GPU 资源管理、在线调度及高性能向量检索; Agent Systems 方面, 关注长时程自进化 Agent 的 Agentic Retrieval、长期记忆、任务规划统一高效协同; 具身智能方面, 关注三维场景重建、空间表示与环境建模。希望研究如何为复杂智能体构建高效、可靠且可扩展的系统基础设施。
- **未来方向**: 研究生阶段计划以 AI / Agent Infra 为主要研究主线, 延续 GPU 向量检索、显存管理、batch 调度与 Agent 系统开发经验, 重点探索检索、记忆、推理与工具执行之间的系统协同, 以及长时程 Agent 任务中的性能与可靠性问题。同时继续深化三维视觉基础, 研究三维场景表示在智能体空间记忆、环境理解和持续交互中的作用。希望通过问题分析、系统设计与实验形成稳定的科研能力, 并以攻读博士为长期规划。