邱俊杰

智能行动中的信息建模

junjieqiu.com jay.junjie.qiu@outlook.com (+86) 191-2013-2361

01 个人简介

香港科技大学（广州）一年级博士生，关注**智能行动中的信息建模**，研究兴趣是智能系统如何学习、保留和利用信息，以理解不断变化的环境并指导行动。研究与工程经历包括从历史序列学习时序表征、维护叙事上下文，以及利用反馈指导视频生成；合作发表 NeurIPS 2025 Spotlight 论文。曾任 ByFusion · LuckyShort 架构师，直接向 CTO 汇报，曾在华为 2012 实验室和南科大科学与工程计算中心工作。

02 教育背景

香港科技大学（广州） 2026.09 - 至今

智能交通 · 博士研究生

南方科技大学 2021.09 - 2025.07

数据科学与大数据技术 · 理学学士

GPA：3.89/4.00　排名：4/53（前 10%）

荣誉：优秀毕业生、优秀本科毕业论文、树德书院十佳毕业生（2025.06）

03 工业界经历

ByFusion · LuckyShort 2026.03 - 2026.09

架构师 · AI 短剧生产

- 主导 AI 短剧视频智能体的架构设计，协调长篇叙事规划、视觉参考资产、可控生成和多模态评估，管理开发团队推进剧本生产智能体与视频生成工作台的核心功能开发，直接向 CTO 汇报。
- 设计长剧本生成的应用层记忆机制，保留角色设定与叙事状态，将新事件与历史上下文核对，并在修订后更新后续生成所需的上下文。
- 面向 AI 出海短剧，打通智能体驱动的剧本生产、视频生成、发布与投流链路，将核心产品连接为短剧生产与分发工作流；平台在 TikTok、YouTube 累计播放超 30 亿，多条视频单条播放超 1 亿。
- 将生成视频与其生成上下文、用户选择及拒绝原因对应记录，构建可回溯的多模态偏好数据；保留用户反馈与视频生成上下文的关联，用于视觉质量评估及长篇视频的跨镜头时序一致性检查。

华为 · 2012 实验室 2025.03 - 2025.12

研究实习生

- 将预训练时序模型适配至华为 Ascend（昇腾）设备，处理通信与电力业务的异构时序数据，准备模型训练和评估所需的数据输入，支持预训练模型在不同工业业务场景下的时序预测实验。
- 在华为内部通信与电力业务数据集上复现时序预测基线，并评估预训练时序模型的预测表现；以通信、电力业务数据集上已复现的基线结果作为对照，验证模型在已测试方法中取得最佳预测效果。
- 合作发表 FactoST（NeurIPS 2025 Spotlight）[1]；论文在公开基准上报告：使用 10% 训练数据、96 步上下文与预测长度，Electricity 相比 OpenCity 的 MAE 降低 12.5%，Weather 相比 Moirai 降低 13.7%。

南方科技大学 · 科学与工程计算中心 2024.03 - 2024.07

实习生 · 高性能计算

- 在南科大启明、太乙超算平台执行 HPL（LINPACK）及 HPCG 基准测试，评估计算吞吐与分布式执行效率，建立面向超算集群计算性能与运行效率的基准对照，作为平台性能评测的依据。
- 参与高性能计算集群新增节点的安装与调试，覆盖计算节点硬件组装、系统软件环境的安装与部署及集群接入调试，将新增计算能力接入既有超算平台，支持平台新增计算资源的部署与运行。
- 针对 2024 年 6 月新安装的四个计算节点，先通过 MPI 通信性能分析定位节点间的通信瓶颈，再优化节点间的交换机互连拓扑，依据通信耗时评估调整效果，使实测 MPI 通信时间降低约 20%。

04 研究经历

FactoST-v2 · 通用时序预训练 2025.10 – 2026.01

UTP 方法设计、预训练与实验 · 2026 年预印本 [2]

- 设计并预训练纯编码器分位数预测模型，采用归一化序列片段与随机历史前缀遮蔽，结合门控注意力和部分旋转位置编码构建支持变长上下文的时序主干，配置支持最长 2,048 步输入与 256 步预测。
- 在未参与预训练的基准上固定模型权重，评估 $12 \to 12$ 和 $96 \to 96$ 设置下的零样本预测；PEMS08 的 $96 \to 96$ 消融中，历史遮蔽使 MAE 从 107.22 降至 91.09，相比无遮蔽模型降低 15%（表 IX-X [2]）。

FactoST · 跨域时序预训练 2025.02 – 2025.05

跨域迁移评估 · NeurIPS 2025 Spotlight [1]

- 在多类时序预测数据集上开展零样本与少样本迁移评估，复用预训练获得的时序表征开展下游目标域预测任务的模型适配，并比较模型在跨域预测及有限目标域训练数据条件下的泛化和适配表现。

05 技术项目

视频生成智能体 · 跨镜头一致性 2026.07 – 2026.09

ByFusion · LuckyShort · 内部项目

- 针对跨镜头身份与场景漂移，构建参考资产条件下的关键帧生成流程，用 Gemini 3.8 Flash 检查生成内容并反馈至图像生成智能体，定向重生成异常帧或镜头，保留已接受帧及共享视觉参考。
- 集中管理长篇脚本规划与图像、视频生成各阶段的创作方向，结合成片反馈迭代调整；以模型先验对候选搜索进行热启动，围绕角色和叙事约束探索不同创作方案，协调叙事方向与视觉生成结果。

稀土定价 · 动作条件下的情景预测与重规划 2026.01 – 2026.03

中国移动研究院 · 中国稀土集团合作项目

- 参与长时域稀土价格建模的方法设计与实验，将地缘政治、产业链观测及历史价格序列作为决策上下文，以不同定价动作作为预测条件，推演未来市场走向，比较市场情景并支持后续重规划。

腾讯云 · 基于用户行为的算力资源分配 2025.05 – 2025.08

腾讯犀牛鸟基金合作项目

- 参与腾讯云资源分配的时序建模方法设计，结合用户行为画像与服务器集群资源分配的历史序列，研究面向不同用户的资源分配决策，将行为与历史分配信息作为算力资源分配模型的决策上下文。

ASC24 · 大模型推理基础设施与优化 2024.04

300 余支队伍中总排名第 10 · 一等奖及分组竞赛奖

- 在竞赛限定的四节点、每节点两张 GPU 集群上搭建分布式大模型推理服务，集成自建 RDMA 网络、vLLM 推理实例和 Ray 调度，使多个模型副本在固定硬件资源下并发执行大模型推理请求。
- 在 Ray 管理的 vLLM 副本间实现动态请求调度，通过可用工作实例队列将任务分配给当前空闲副本，在任务完成后将实例归还队列，供后续请求复用，平衡固定 GPU 资源下多个副本的推理负载。

06 技能与研究主题

物理智能	熟悉概念：Diffusion / Rectified Flow、Multimodal Conditioning、DreamerV3、JEPA、RSSM、Latent Dynamics、Latent Imagination、Model-based RL、Action Abstraction、Steerable VLA Policies
记忆与上下文	熟悉概念：Belief-State Modeling、Multimodal Memory、Event-centric Memory、Prospective Memory
训练、推理与仿真	项目实践：vLLM、Ray、RDMA、Huawei Ascend；熟悉：4D（DP/TP/PP/CP）、Continuous Batching、MuJoCo

07 参考文献

[1] S. Zhong, J. Qiu, Y. Wu, X. Zou, Z. Rao, B. Yang, C. Guo, H. Xu, and Y. Liang, "Learning to Factorize Spatio-Temporal Foundation Models," in Advances in Neural Information Processing Systems (NeurIPS), 2025. **Spotlight**.

[2] S. Zhong, J. Qiu, Y. Wu, Y. Liu, Y. He, Z. Rao, B. Yang, C. Guo, H. Xu, and Y. Liang, "Learning to Factorize and Adapt: A Versatile Approach Toward Universal Spatio-Temporal Foundation Models." arXiv preprint arXiv:2601.12083, 2026.