Stein

17280801850 25210980109@m.fudan.edu.cn 上海

正式党员



教育经历

复旦大学(推免) | **大数据学院** | **人工智能方向** | **知识工场实验室**-硕士 2025.09 - 2028.06

- 主修课程：nlp&llm、多模态大模型、AI前沿模型算法、最优化理论、深度学习等

杭州电子科技大学 | 卓越学院 | 计算机科学与技术 实验班-本科 2021.09 - 2025.06

- **成绩：** GPA 4.59/5（**专业前0.6%**） **国家奖学金 | 国家级大创**负责人| 省政府奖学金 | 校优秀毕业生
- 竞赛：数学建模 国赛 **国二** &省一 & 美赛 M奖 | 华为软件精英赛三等奖 等 10+ 项

专业技能

- 计算机与数学基础： 杭电 ACM 集训队成员，具备扎实编程与算法基础（C++ / Python / Matlab）；本科核心课程（数学分析、线性代数、数据结构、操作系统、计算机组成原理等）成绩均 90+。
- AI 框架与工程： 熟练使用 PyTorch、Transformers 进行模型训练与推理，具备 GPU 集群实战经验；熟悉 Accelerate、DeepSpeed 等分布式训练框架，掌握 Docker、Git、Linux 等工程化能力。
- AI 工具与工程能力： 熟练使用 Claude Code、OpenClaw 等 AI 编程工具，具备快速学习与工程落地能力，曾获部门月度代码提交第一并进行 AI Coding 经验分享。

科研项目经历

Locality-aware Diffusion Language Modeling（一作 | ICML under review） 2025.10 - 2026.03

系统研究 Masked Diffusion Language Models 可训练性问题，对比和 Autoregressive LLMs 在不同结构化生成任务中的能力差异，探究模型归纳偏置与任务依赖结构之间的关系。

1. 设计实现受控实验任务：**in-context linear regression、star-graph path-finding、Sudoku**，分别对应局部精确绑定、逆向规划和全局约束满足，以系统验证不同生成范式的适配边界。
2. 提出两种 locality-aware blockwise diffusion 架构：Scatter 通过“块内自回归 + 跨块同步更新”保留局部顺序性，Jigsaw 通过“块内自回归 + 跨块熵引导动态规划”兼顾局部建模与全局迭代优化。
3. 实验结果表明，模型性能与任务的 dependency structure 密切相关：在 linear regression 这类强局部依赖任务上，Jigsaw 显著提升了收敛性与训练稳定性，性能接近 AR；而在 path-finding 与 Sudoku 这类全局规划与约束满足任务上，Diffusion 范式整体优于 AR。
4. 本研究进一步说明，AR 更适合局部绑定与顺序生成任务，Diffusion 更适合全局规划与约束满足任务，为 dLLM 的结构设计、推理建模及后续 Agent 规划能力研究提供了实验依据。

基于时空距离与分帧的动态图欺诈检测（国家级大创 | CCF-B 一作 | 发明专利） 2023.04 - 2024.08

- 内容：针对传统GNN欺诈检测模型存在无法高效利用时空信息等问题，提出动态图欺诈检测模型STM：利用图浓缩、基于时空距离和差异聚合的帧内聚合和帧间聚合等多个模块，使模型充分利用长短期时间信息和空间信息，提升检测精度和效率，在多个动态图benchmark上取得**SOTA效果**。

Semantic Diffusion Language Modeling（二作 | ICML under review）

- 研究离散扩散语言模型中 noising kernel 设计对训练稳定性与生成质量的影响，提出基于偏差-方差权衡的统一理论分析框架。
- 提出 SemDLM，通过语义邻域扩散 + shared refresh branch + 词频先验校正降低训练偏差与方差，同时增强采样阶段误差修复能力。在LM1B 上取得 27.19 Test PPL，优于多类离散扩散基线模型。

实习经历1

AI4S/多模态大模型-阿里巴巴 2025.10 - 至今

研究方向：蛋白质/小分子**多模态统一大模型**、蛋白质生成/理解评测框架、Protein Agent 后训练。

一：ProtText-Eval —— 文本指导蛋白质设计统一评测框架

- **角色：负责独立完成搭建完整统一评测基础设施**
- 面向蛋白质生成/理解任务，构建统一评测框架，解决实验指标不一致、结果不可比等问题。
- 构建四层 26 项指标评测体系，覆盖生物合理性过滤、指令对齐与任务专项能力等全链路评估；设计模块化中间协议 ProteinSample → Adapter → ProteinPrediction，支持模型、指标、数据集**解耦接入与扩展**；开发基于 MD5 去重与持久化缓存的 PDBPool，显著提升结构预测指标计算效率。
- 已集成 5+个基线模型的推理与评测链路，统一封装模型输入输出与结果解析流程，已完成 8+benchmark 基线评测，框架支持生成模式与编辑模式双评测范式，为后续多模态蛋白模型训练提供标准化评估基础设施。

二：蛋白质功能理解 Agent —— ReAct 推理与强化学习

- **角色：核心设计与开发，负责 Agent 架构设计、工具系统搭建与 RL 训练链路设计**
- 面向蛋白质理解任务，从零搭建 Agent 后训练工程框架，使模型能自主决定工具调用顺序并基于中间观测进行推理调整。
- 设计多轮 Thought→Action→Observation rollout 引擎，实现 Teacher Backend 抽象层，支持本地启发式与外部 LLM 教师模型动态决策。开发 18 个生物信息学工具注册与执行系统，完成 6+个核心工具（mmseqs2/hmmscan/tmbed/signalp等）的 native backend 移植，设计分层依赖机制消除工具间隐式耦合。
- 基于阿里 ROLL & ROCK 框架打通 rollout → 轨迹导出 → SFT/RL 配置生成 → 双评测（ROCK 调用指标 + Protext 生物合理性）的完整数据闭环
- 设计多维度奖励函数（答案质量、工具选择、推理质量、执行效率），适配 GRPO 强化学习训练。

三：蛋白质多模态统一大模型（理解 + 生成）

- **角色：参与蛋白模态训练策略设计与 SFT / 推理链路开发**
- 模型基于 MoT（Mixture-of-Transformer）架构，构建蛋白序列理解与生成一体化多模态大模型，支持 protein ↔ text 双向建模。
- 设计三阶段训练体系：
  Stage1（UniRef50，~5900万序列）蛋白预训练；Stage2（SwissProt）protein-text 双向对齐；
  Stage3 SFT 支持 QA / caption / design。
- 搭建统一预训练框架：实现 protein/text 混合 packed sequence、统一 next-token loss（text CE + protein CE）、token-level 路由 MoT。
- 构建蛋白理解与推理能力：支持 protein+question → answer 推理、序列 likelihood scoring 与 ProteinGym benchmark 评测
- 实现两阶段解耦训练 pipeline（Stage1→Stage2 checkpoint 初始化），支持大规模分布式训练。

技术栈：Python、Linux、PyTorch、Transformers；分布式训练（Accelerate、DeepSpeed）；推理与加速（FlashAttention-2、vLLM）；工程工具（Docker、SLURM、WandB）；RLHF（verl / ROLL，GRPO）。

实习经历2

智能驾驶算法工程师-天翼交通科技 VLA部门　　2025.04 - 2025.07

- 参与设计并落地面向自动驾驶 4D 数据的自动预标注 pipeline，联合多种开源模型完成地图/道路预测、3D 检测、语义生成与 Occupancy 预测等任务。
- 负责多类开源算法的调试、适配与集成，解决不同模型在输入输出格式、推理流程、结果对齐上的兼容问题，构建统一化自动标注链路。
- 通过工具链开发与自动化流程建设，提升预标注效率与多任务协同能力，为大规模自动驾驶数据生产和模型训练提供高质量数据基础。

技术栈：PyTorch、CUDA、Linux、Docker、OpenPCDet、MMDetection3D、OpenCV。