

AI基础设施研究中心

中国电信人工智能研究院 (TeleAI)

面向大模型时代的下一代训练、推理与系统基础设施

范元凯 AI基础设施研究中心

AI Infra 是什么？

把模型真正跑起来的技术底座



AI Infra =

把模型真正跑起来，并跑得

更快、更稳、更省



为什么越来越重要？

因为模型能力再强，如果跑不起来、跑不稳、跑不便宜，最终也很难真正落地。



AI Infra 是 AI 时代的“水电煤”

(算力 + 网络 + 数据)

中心方向与总体布局

语义模型

语音模型

视觉模型

世界模型（多模态理解与生成统一大模型）

具身大小脑协同

软硬件一体化设计

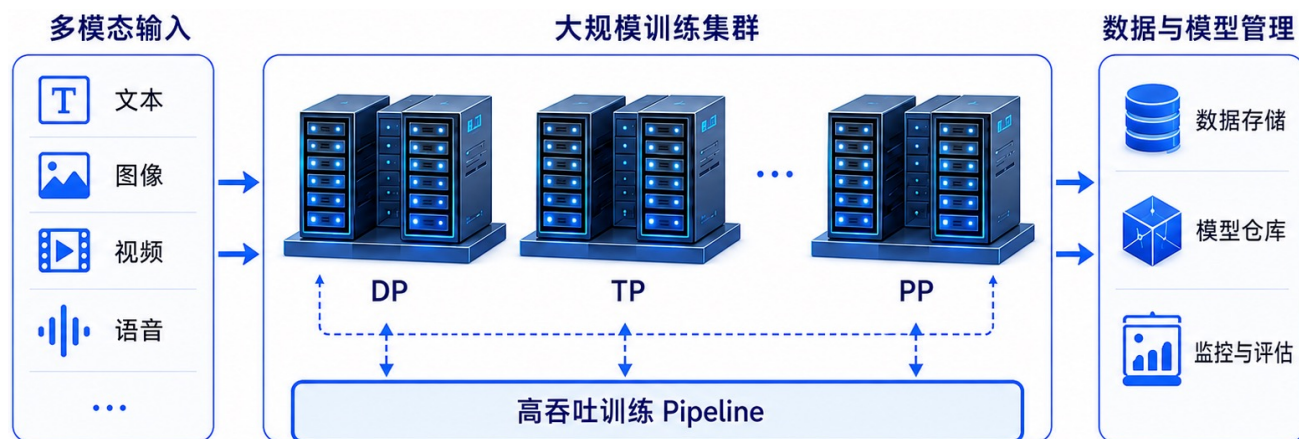


智传网 (AI Flow)

多模态大规模训练系统

Multimodal Large-scale Training

支撑大规模参数模型**预训练**与**后训练**，提升大规模训练集群的稳定性与模型算力利用率



架构创新

异构训练

分离式训练

异步RL训练

系统优化

动态并行

通算掩盖

策略优化

混合精度

极致性能

算子融合

编译优化

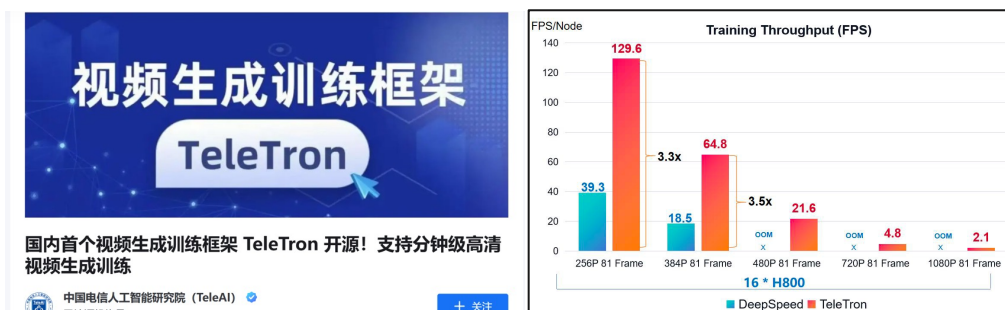
高性能算子库

开源训练框架、贡献核心技术
综合性能位于开源第一梯队

TeleTron 分钟级长高清视频生成大模型**预训练**框架

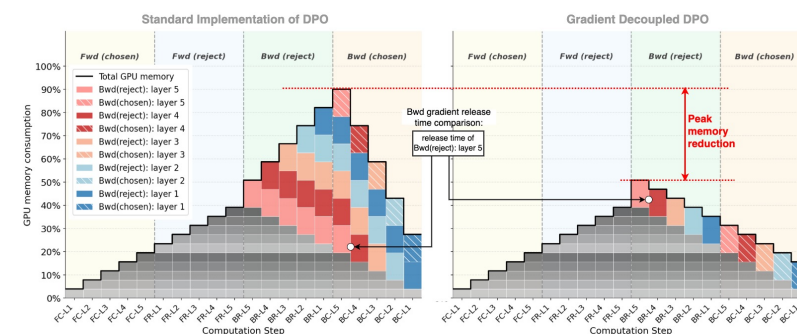
<https://github.com/Tele-AI/TeleTron>

<https://zhuanlan.zhihu.com/p/1907030055512671098>



TeleBoost 扩散模型**后训练**统一对齐框架

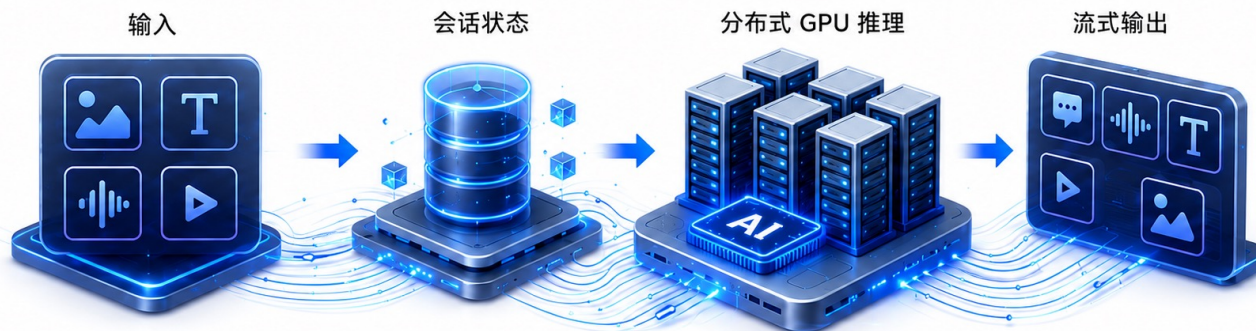
<https://github.com/Tele-AI/TeleBoost> (即将开源)



世界模型在线推理系统

World Model Online Inference

支撑生成式世界模型**在线服务**，加速推理过程，实现实时地流式音视频输出



◆ 新型范式

流式生成

分离式推理

协同推理

◆ 系统优化

请求调度

负载均衡

◆ 框架加速

稀疏

量化

异步编排

缓存复用

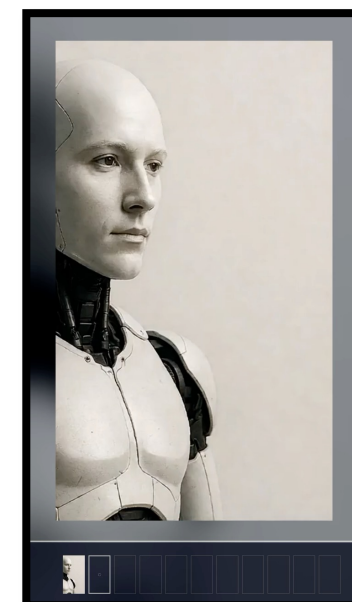
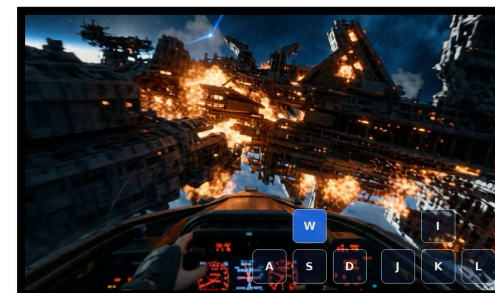
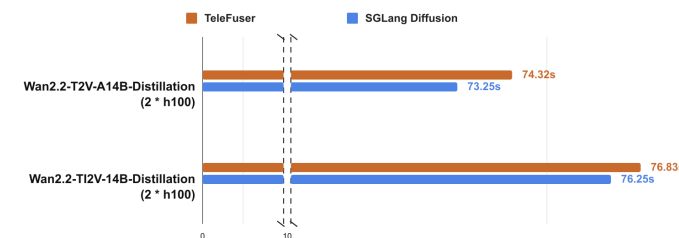
◆ 极致性能

高性能算子库

开源**推理框架**、贡献核心技术
综合性能位于开源第一梯队

TeleFuser 面向下一代世界模型的实时高效**推理框架**

<https://github.com/Tele-AI/TeleFuser>



分布式通信与缓存系统

Distributed Communication & Caching

🎯 全场景支持海量数据处理、分布式多机训练和在线推理，提供统一分布式系统底座

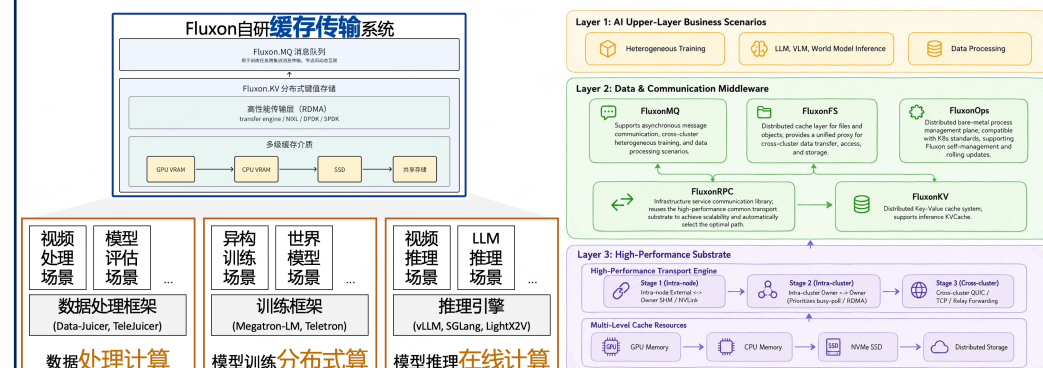
分布式通信与缓存系统



- ◆ **高性能**
 - 零拷贝
 - 共享内存
 - 跨界点P2P
 - 路由自适应
- ◆ **高可用**
 - 统一生命周期管理
 - 弹性解耦
 - 网络可用性
- ◆ **多场景**
 - 在线推理
 - 多机训练
 - 数据处理
 - 模型评测

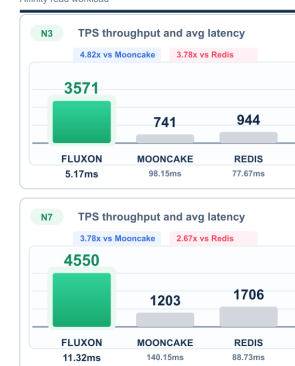
开源**统一通信缓存系统**，全链路**Rust**
综合性能超开源/主流系统

Fluxon FluxonKV 缓存 | FluxonMQ 消息队列 | FluxonFS 文件对象
<https://github.com/Tele-AI/Fluxon>
<https://zhuanlan.zhihu.com/p/2054332509052262216>



READ_AFFINITY

Affinity read workload



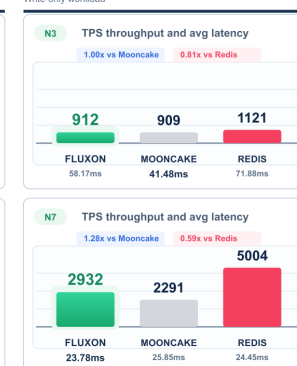
READ_ZIPF

Zipf skewed read workload



PUT_ONLY

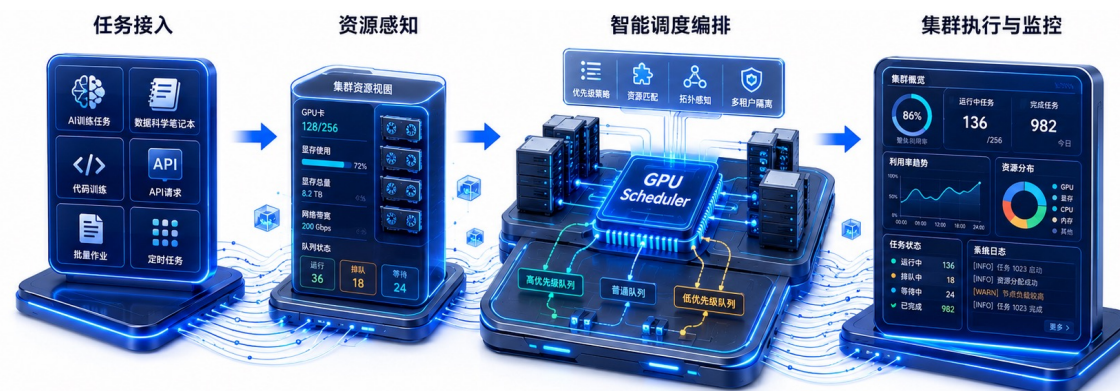
Write-only workload



大规模集群算力管理平台

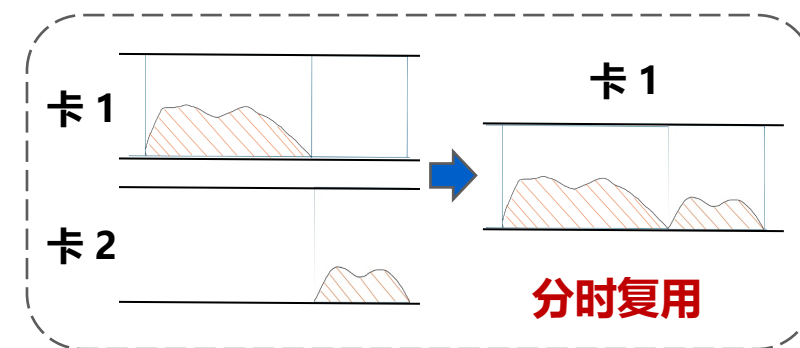
Large-Scale GPU Cluster Management

 统一纳管万卡算力集群，实现动态算力资源调度，算力超分、虚拟化



- ◆ **架构创新**
 - 细粒度资源切分
 - 资源池化
 - 组调度
- ◆ **系统优化**
 - 显存优化
 - 算力流动
 - 远程调度
 - 任务优先级

基于开源技术，自研**算力流动**系统
综合性能超主流解决方案



通过算力流动和任务迁移，实现资源高效利用与灵活调度



中心科研进展与成果

聚焦前沿技术，沉淀具有影响力的科研成果（目前团队已发表顶会论文共**十余篇**，投稿**二十余篇**）



十余篇

国际顶级学术会议期刊论文

发表于国际顶级学术会议与高水平期刊



代表性论文成果

- 高水平论文持续产出
- 聚焦系统与前沿技术方向
- 论文发表于国际/国内重要学术平台



科研项目与平台

- 承担多项重点科研任务
- 联合建设科研平台
- 推动产学研协同创新



知识产权与成果转化

- 形成专利与软著储备
- 推动技术原型落地
- 支撑业务创新与应用



学术影响力与团队建设

- 积极参与学术交流
- 培养青年科研骨干
- 形成持续创新的研究梯队

开放合作与交流

期待与大家共同探索下一代AI Infra

应用

多模态模型



大语言模型



扩散模型



框架/模型

训练框架



推理框架



稀疏



蒸馏

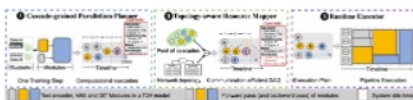


系统

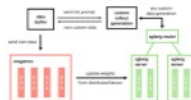
大模型推理系统



大规模训练系统

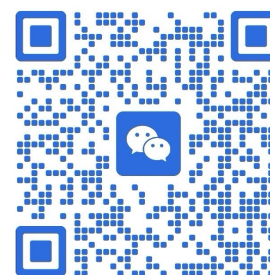


智能体强化学习系统



以系统创新， 支撑下一代智能涌现。

—— 欢迎交流合作 ——



个人联系方式



个人主页

实力对标工业界头部大厂，争当**科研界第一**的AI Infra团队！期待各位同学加入！