2025-1

华中科技大学
计算机科学与技术学院
本科实验班科研信息手册

(2025 年)

计算机学院教学办公室

撰写:　教学办公室

编辑:　教学办公室

一、学院科研团队简介

学院现有专任教师136人，其中教授/研究员54人，副教授/副研究员56人，拥有1个基金委创新群体，3个教育部创新团队、1个科技部创新团队和1个湖北省创新团队。

学院设有并行分布式计算、数据存储、数字媒体、数据工程、人工智能与优化5个研究所和1个教学实验中心，建有大数据技术与系统国家地方联合工程研究中心等11个国家或省部级研究基地，孵化了达梦数据库等4个高新技术企业。学院科学研究始终站在学科前沿，在分布式计算、信息存储发展过程中产生了重要影响；在现代数据库理论与技术、数字媒体、网络系统、信息安全和NP难度问题求解等方面形成了自己的特色，积极承担包括国家科技重大专项在内的多项重要科研任务，取得包括国家自然科学二等奖、国家科技进步二等奖、国家技术发明二等奖在内的一系列高水平研究成果，具有突出的综合实力。

学院坚持“面向系统，软硬协同”的全栈式系统能力人才培养理念，建设多门线上、线下国家一流课程，获得国家级、省部级教学成果一等奖，获得挑战杯金奖、互联网+金奖、世界超算大赛总冠军、SAT国际算法竞赛第一名、图计算挑战赛全球总冠军、EDA工业布局布线设计全球冠军等竞赛成绩，培养了包括华为“天才少年”在内的一大批创新性人才。（数据统计截止2024年3月）

二、学院科研团队信息

团队序号	团队名称	联系人	联系邮箱
1	信息存储及应用团队	王芳老师	wangfang@hust.edu.cn
2	数据高效存储与计算团队	曹强老师	caoqiang@hust.edu.cn
3	多媒体流计算与存储团队	郭红星老师	guohx@hust.edu.cn
4	智能数据存储与管理团队	刘渝老师	liu_yu@hust.edu.cn
5	先进可扩展计算与系统团队	蒋文斌老师	wenbinjiang@hust.edu.cn
6	分布式系统团队	余辰老师	yuchen@hust.edu.cn
7	大模型与智能系统团队	张腾老师	tengzhang@hust.edu.cn
8	人机物系统与安全团队	王蔚老师	2014612548@hust.edu.cn
9	网络认知计算团队	莫益军老师	moyj@hust.edu.cn
10	现代数据工程与实时计算团队	袁凌老师	cherryyuanling@hust.edu.cn
11	智能与分布计算团队	李瑞轩老师	rxli@hust.edu.cn
12	现代数据库技术团队	朱虹老师	zhuhong@hust.edu.cn
13	嵌入与普适计算团队	胡龙老师	hulong@hust.edu.cn
14	嵌入式与人工智能团队	涂刚老师	tugang@hust.edu.cn
15	智能大数据管理与分析团队	郑渤龙老师	bolongzheng@hust.edu.cn
16	视觉计算与智能认知团队	李平老师	lpshome@hust.edu.cn
17	图形与视觉计算团队	李丹老师	lidanhust@hust.edu.cn
18	智能媒体计算与网络安全团队	杨卫老师	weiyangcs@hust.edu.cn
19	认知计算与智能信息处理团队	魏巍老师	weiw@hust.edu.cn
20	EDA 与工业优化团队	苏宙行老师	suzhouxing@hust.edu.cn
21	数据挖掘与机器学习团队	何琨老师	brooklet60@hust.edu.cn
22	智能与实时计算团队	李剑军老师	jianjunli@hust.edu.cn
23	智能信息与大数据团队	张瑞老师	ruizhang6@hust.edu.cn
24	智能计算与强化学习团队	金燕老师	jinyan@hust.edu.cn

三、学院科研团队及项目介绍

1．信息存储及应用团队

计算机学院“信息存储及应用团队”依托于计算机系统结构国家重点学科，其前身为“外存储系统”国家专业实验室。拥有信息存储系统教育部重点实验室、数据存储系统与技术教育部工程研究中心等科研基地平台。建有华中科技大学-华为技术有限公司新型存储技术创新中心、华科大-浪潮电子共建“新存储联合实验室”、华中科技大学-杭州海康威视数字技术股份有限公司“海量信息存储联合实验室”、荣耀-华中科技大学先进存储联合实验室、华中科技大学计算机科学与技术学院-新华三技术有限公司存储创新联合实验室等。

团队现有教授 7 人，副教授 5 人，讲师 1 人，博士后 3 人。其中长江特聘教授国家级人选 1 人，国家杰青 / 万人计划科技领军人才 2 人，973 首席科学家 1 人，中组部青年拔尖人才 1 人，重点研发计划青年科学家 2 人。2014 年获“信息存储系统与技术”教育部创新团队评优获滚动支持、2017 年获批“面向大数据的新一代存储技术研究”湖北省创新群体、2018 年获批“大数据存储系统与技术”基金委创新群体。

团队承担了国家重点研发计划项目、国家 973 计划项目（首席）、863 项目、国家自然科学基金项目（杰青、重点、优青、面上、青年）、国防预研项目、省重点研发等，并与华为、浪潮、海康威视、荣耀、OPPO、阿里云、字节跳动等企业开展长期深入合作。

团队主要学术方向及研究内容有：

非易失性存储技术：包括相变存储、阻变存储新机理，非易失性内存体系结构、持久化内存管理及文件系统，基于忆阻的存算一体化技术研制有存算一体 MRAM、RRAM 芯片等；

固态盘及盘阵列技术：研究存储控制器技术，各类通道接口技术、数据布局算法、数据容错技术、数据恢复重建算法、性能优化技术及节能技术等，自主研发有安全固态盘、通用固态盘阵列和便携磁盘阵列原型系统等；

海量存储系统及技术：研究分布式并行存储系统技术、对象存储系统、异构融合存储体系结构，近数据处理方法、数据去重技术、高可靠保障技术等，自主研发了海量分布式对象存储系统(CapFS)；

云存储及其服务保障技术：包括存储虚拟化、数据备份、灾难恢复、数据保护技术等；

存储应用与优化：研究复杂网络环境下的存储组织模式及存储软件，包括K-V存储、存储管理、存储服务软件等，面向大规模图数据存储与计算优化、面向AI大模型应用的系统优化技术等。

团队围绕信息存储，从存储器件/体系结构、存储设备、存储系统不同层面开展协同研究，支持以数据为中心的高效信息处理平台，研究工作涉及存内计算体系结构、非易失存储技术、大规模网络存储系统、云存储、AI技术、存储安全等诸多相关领域。研究成果已发表在FAST、ISCA、OSDI、SC、ATC、DAC、HPCA、MSST、ICDE、VLDB等知名国际会议和IEEE TC、IEEE TPDS、ACM TOS等权威期刊上发表系列论文；团队牵头制定国家标准11项，电子行业标准3项。获国家技术发明二等奖2项，国家科技进步二等奖1项，省部一等奖3项；获国际存储竞赛决赛奖,华为奥林帕斯先锋奖等。核心技术应用于我国骨干企业的存储产品，为推动我国存储产业由弱变强迈入全球领先行列做出了突出贡献。

团队培养的研究生工作在Intel、IBM、微软、百度、阿里、腾讯、华为、中兴等国内外著名IT企业和科研院所或自己开创存储公司,部分毕业生在美国、英国、加拿大、新加坡等国家学习或工作。培养博士获全国优博、电子学会优博、ACMChina优秀博士学位论文全国奖等奖励，培养博士已有3位入选华为“天才少年”计划。

团队成员：

<table>
<tr><td>冯　丹</td><td colspan="3">团队负责人
教授，主要研究领域为计算机系统结构、非易失存储技术、大规模网络存储系统、云存储、存内计算体系结构</td></tr>
<tr><td>王　芳</td><td>教授，主要研究领域为计算机系统结构、分布式存储、键值存储、非易失存储、池化内存系统、图计算等</td><td>胡燏翀</td><td>教授，主要研究领域为数据可靠性、纠删码、分布式存储、大模型存储、云存储、大数据存储等</td></tr>
<tr><td>华　宇</td><td>教授，主要研究领域为计算机系统结构、网络存储、数据管理、分布式计算等</td><td>谭支鹏</td><td>教授，主要研究领域为分布式存储,大数据存储与管理、智能存储、移动存储等</td></tr>
<tr><td>秦磊华</td><td>教授，主要研究领域为计算机系统结构、工程创新人才培养</td><td>谢雨来</td><td>教授，主要研究领域为大数据存储安全、入侵检测、云安全、舆情分析等</td></tr>
</table>

陈俭喜	副教授，主要研究领域为磁盘阵列体系结构、混合存储系统、分离式内存系统以及面向AI的存储技术等	童　薇	副教授，主要研究领域为计算机系统结构、计算机存储系统、非易失存储、存算融合等
施　展	副教授，主要研究领域为计算机系统结构、云存储、大数据存储系统等	刘　康	副教授，主要研究领域为芯片设计自动化，AI数据压缩，AI安全与隐私等
汪承宁	副教授，主要研究领域为DDR存算一体内存储器电路设计与EDA	李晓露	讲师,主要研究领域为存储系统可靠性，包括大规模分布式存储系统可靠性以及SSD可靠性
吴　兵	博士后,主要研究领域为非易失存储器及其存算一体系统	魏学亮	博士后,主要研究领域为非易失存储器及异构内存系统
方　鹏	博士后,主要研究领域为面向AI的高性能异构存算系统		

团队联系方式：

联系邮箱：wangfang@mail.hust.edu.cn，王芳老师

地址：华中科技大学光电信息大楼B区五楼

项目 1：以算力为中心的层次化存算融合（指导老师：方鹏）

一、项目背景

随着人工智能技术的迅猛发展，尤其是大模型、图学习、深度学习等新兴应用的兴起，对计算能力和存储能力的需求呈现指数级增长。GPU（图形处理单元）和 NPU（神经网络处理单元）等高性能异构计算设备在处理这些复杂任务时表现出色，但其内存容量和带宽的限制逐渐成为瓶颈，引发了所谓的“内存墙”问题。内存墙问题指的是计算设备的计算能力与内存带宽、容量之间的不匹配，导致计算设备无法充分利用其算力，进而限制了整体系统的性能。

在大规模 AI 应用中，数据量巨大且计算密集，传统的存储架构难以满足高效的数据访问需求。通常，GPU/NPU 等计算设备需要通过主机 DRAM（动态随机存取存储器）进行数据中转，才能访问 SSD（固态硬盘）等大容量存储设备。这种层次化的存储访问模式虽然在一定程度上缓解了内存容量不足的问题，但也带来了显著的数据传输延迟和带宽瓶颈，尤其是在处理大规模数据集时，数据在存储层次间的频繁迁移严重拖累了系统的整体性能。

为了应对这一挑战，提出了“以算力为中心的层次化存算融合”方案。该方案的核心思想是将 HBM（高带宽内存）、DRAM 和 SSD 等不同层次的存储资源整合为一个统一的内存空间，使得 GPU/NPU 等计算设备能够直接访问这些异构存储资源，减少对主机 DRAM 的依赖，从而提升数据传输效率。此外，随着 AI 应用的复杂性和多样性不断增加，单一的硬件架构已经无法满足多样化的计算需求。因此，如何将 CPU、GPU、NPU 等异构计算资源进行池化，并通过解耦 AI 应用算子的方式，在算力池中进行算子粒度的并行与调度，成为了提升系统整体性能的关键。

层次化存算融合不仅能够缓解内存墙问题，还能够为未来的 AI 应用提供更加灵活和高效的存储与计算架构。随着数据量的持续增长和计算任务的日益复杂，传统的存储和计算架构已经难以满足需求。通过将存储与计算资源深度融合，系统可以在有限的硬件资源下实现更高的性能，为大规模 AI 应用提供强有力的支持。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力和内存的服务器，配备高性能 GPU/NPU、HBM、DRAM 和 SSD 等存储设备。

2．软件平台

基于 Ubuntu、PyTorch 和 CUDA 等平台进行编程。

3．技术基础

本项目基于存储系统架构（NVMe、HBM 等）、异构计算/存储、AI 应用与算子优化相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．具备较强的自学能力，能够主动探索和解决问题，并有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch、CUDA 基础；

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．异构存储资源整合与优化：研究如何将 HBM、DRAM 和 SSD 等异构存储资源整合为统一的内存空间，优化存储资源的访问效率，减少数据中转的开销。

2．算力池化与算子调度：设计并实现 CPU-GPU/NPU 算力池化方案，解耦 AI 应用算子，在算力池中进行算子粒度的并行与调度，提升算力利用率。

3．基于 GPU 的 IO 堆栈设计：设计并实现一种基于 GPU 的 IO 堆栈，旨在最大化 GPU 直接访问磁盘（如 SSD）的吞吐量，同时减少 GPU 内核的占用，提升整体计算与存储的协同效率。

项目 2：面向大语言模型推理系统的优化方法研究（指导老师：方鹏）

一、项目背景

随着人工智能技术的不断突破，大语言模型在自然语言处理、机器翻译、文本生成等领域的应用价值日益凸显。以 OpenAI 的 ChatGPT 为代表的大模型凭借其强大的语言理解能力和生成能力，正在改变着我们的生活方式和工作模式。

大语言模型的应用价值体现在多个方面。在自然语言处理领域，它们能够高效地进行文本分类、情感分析、命名实体识别等任务，为企业提供智能化信息处理解决方案。在机器翻译领域，大语言模型实现了高质量、多语种的实时翻译，极大地促进了跨语言交流。此外，在文本生成方面，大语言模型的应用使得新闻报道、广告创意、文学作品等内容的创作效率大幅提升。而在智能客服领域，它们的应用更是为企业降低了运营成本，提升了客户服务体验。

在上述背景下，如何构建高效的大语言模型推理系统成为提升用户体验，降低企业成本的重要方式。大语言模型推理服务的广泛应用带来了对于推理系统吞吐量和延迟的挑战。面对大规模的用户请求，推理系统需要处理大量并发任务，这对系统的吞吐量提出了较高要求。同时，在实时交互场景下，如在线翻译、智能问答等，用户对响应速度的期望越来越高，降低推理延迟成为提升用户体验的关键。

为构建高效的大模型推理系统，对算力和显存的充分利用也是提升性能的重要因素。模型的参数规模庞大，训练和推理过程中的计算量巨大，高性能计算资源成为提升模型性能的瓶颈。同时，推理过程中对显存的大量占用，使得显存不足成为影响模型正常运行的一大问题。

多模态大模型的应用也日益广泛。它们在图像识别与生成、视频理解与生成、跨模态检索等领域展现出强大的能力，为自动驾驶、人脸识别、短视频创作、视频监控等信息处理任务提供了新的解决方案。发掘多模态大模型推理系统的特点和对算力和显存的特殊需求成为可能的研究方向。

本项目旨在研究面向大模型推理系统的优化方法。通过优化推理系统架构，设计更高效的请求调度算法，优化算子实现等方式，提高推理系统的吞吐量、降低延迟，同时减少模型对算力和显存的需求。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

基于 Ubuntu 操作系统，vLLM 等推理框架，进行学习和优化研究。

3．技术基础

本项目基于大模型推理计算，CUDA 并行程序设计，服务质量保障等理论和变成技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．对大模型有足够的兴趣，有较强的自学能力，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．多模态大模型数据集构建：针对特定任务领域，收集和整理高质量的问答数据集，特别是多模态训练和推理测试数据，确保数据集覆盖任务需求的多样性与复杂性，以适应不同的应用场景需求。

2．多模态大模型推理系统环境搭建：了解主流大模型推理系统的搭建方式并构建本地大模型推理服务。了解主流大模型推理服务框架的工作流程，测试分析不同阶段的时间占比情况。针对多模态大模型，分析不同模态推理任务的计算和访存特性。

3．高效的请求调度方法研究：基于大模型推理任务的特点，在了解现有的请求调度方式的基础上，研究现有调度方式的瓶颈，尝试提出可能的优化方案，并设计实现，进行初步的测试对比。

4．大模型底层算子优化：了解大模型推理系统的运行流程，梳理大模型推理计算过程中上层矩阵计算到底层算子的调用传递过程，在理解现有算子实现方式的基础上实现算子优化方案。

项目 3：面向大语言模型的存储系统优化（指导老师：胡燏翀）

一、项目背景

随着人工智能技术的飞速发展，大规模深度学习模型的研究和应用呈现爆发式增长。这些模型在语言处理、计算机视觉、生物信息学等多个领域展现了强大的能力，推动了科技的进步。然而，大模型的高效存储和管理成为了一个迫切需要解决的关键问题。以下从大模型存储的需求、挑战以及现有技术现状三方面阐述科研背景。

大模型存储的需求

大模型通常包含数亿到数千亿个参数，其训练数据集的规模也以 TB 甚至 PB 为单位计。这种规模的模型和数据给存储系统带来了巨大的压力，主要体现在以下几个方面：

1．高容量需求：如 GPT-4 等语言模型，其参数量已超过 1 万亿，加上训练数据、优化日志和中间结果，存储需求呈指数级增长。

2．高吞吐量与低延迟：训练和推理过程需要频繁访问模型参数，存储系统需支持高带宽和低延迟的读写操作，以避免成为性能瓶颈。

3．可靠性与容错性：大规模存储系统需保证数据的一致性与可靠性，尤其是在硬件故障或网络波动情况下，需具备快速恢复能力。

4．分布式与并发性：模型训练或推理往往在数千张 GPU 或 TPU 上并行地进行，这要求系统具备强大的分布式能力。

对于高容量需求，大语言模型的显存优化背景在于其庞大的参数量和复杂的计算需求，导致显存消耗巨大，限制了模型的规模和训练效率。例如，仅激活值就可能给每个设备带来超过 180GB 的开销。为应对显存开销大这一挑战，研究者们提出了多种显存优化方法。例如，使用张量并行和管道并行技术，在不同维度上切分模型，减少单个设备的显存负担；使用 Gradient Checkpointing 技术，通过在反向传播时重新计算激活值来减少显存使用；使用 FlashAttention 方法则专注于优化注意力计算过程中的显存消耗等。

对于高吞吐和低延迟需求，大语言模型追求高吞吐量与低延迟，以提升处理效率和用户体验。例如，在数据预处理 IO 加速方面，使用英伟达 DALI 技术可将预处理从 CPU 迁至 GPU，显著提升速度。在训练加速方面，使用梯度压缩加速训

练，通过动量校正、本地梯度裁剪等手段，实现高达 270-600 倍的梯度压缩，降低通信带宽需求，提高分布式训练的可扩展性。这些方法共同助力大语言模型在资源受限的情况下，实现更高效的训练和推理。

对于可靠性与容错性：模型训练通常涉及多个计算节点和长时间的运行，节点故障和计算失败是常见问题。为应对这些问题，容错机制如检查点（checkpoint）被广泛采用。检查点允许在节点故障时捕获异常并从正常节点同步最新参数，恢复训练，但是检查点可能会打断正常的模型训练过程，现有的方法如 Checkfreq，Gemini 通过将训练和检查点高效的流水线并行化来加强分布式训练中的容错能力。

对于分布式与并发性需求，分布式推理通过多节点并行计算提升吞吐量，如 DeepSpeed Inference 和 Xinference 等框架。并发性方面，vLLM 等框架通过连续批处理和动态批处理机制，大幅提高并发吞吐量和推理速度。调度优化如 Continuous Batching，实现 step 级别调度，提升效率和用户体验，已在多个推理框架中应用。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的多卡服务器

2．软件平台

基于 PyTorch 等深度学习及大模型库在 Ubuntu 系统上进行编程

3．技术基础

本项目基于存储系统、深度学习、大语言模型等技术开发。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术；了解分布式计算技术；了解存储系统相关知识理论。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1. 大模型训练中的显存优化
2. 大模型训练中检查点加速与容错能力增强
3. 大模型训练中的数据预取数据加载优化
4. 大模型训练中的分布式训练加速
5. 大模型推理中的显存优化
6. 大模型推理中的通信优化
7. 大模型推理中的调度优化
8. 大模型推理中的分布式推理加速

项目 4：使能失配补偿功能的高并行性原位计算 BLSA 及 Bank I/O 电路（指导老师：汪承宁）

一、项目背景

位线感测放大器（bitline sense amplifier, BLSA）即是 SRAM，其核心是一个正反馈晶体管电路，按照晶体管的类别可被划分为 pSA 和 nSA 两部分，是 RRAM/MRAM 存储器单元阵列读出电路的重要部件，具有三重功能：感测、驱动、锁存。传统为内存设计的感测放大器存储结点直接连接到位线，并与位线耦合，在应用于存算一体原位乘加计算时，面临着预充电一次只能读出一个有效比特的困境，没有并行性可言。本研究充分挖掘内存的命令级并行性（CLP）和数据级并行性（DLP），从位线与外围电路协同交互的角度，设计合适的位线与存储结点解耦电路，开发命令及数据比特并行性，实现原位乘累加计算延迟的缩减。由于制程变化性（process variation），RRAM/MRAM 外围位线感测放大器核心晶体管（core transistors）的阈值电压（Vth）会偏离预先设定的值，导致感测放大器左右两边支路不完全对称的问题。随着工艺节点的微缩，晶体管阈值电压变化率越来越大。如何通过额外的失配补偿（mismatch compensation）辅助电路对感测放大器的静态平衡（equilibrium）工作点进行补偿，成为了近两年来存储器厂商的聚焦问题。内存储器大厂 Samsung、SK Hynix 在最近的 DRAM 产品中均使用了带有失配补偿功能的偏移取消感测放大器（Offset-Cancellation SA, OCSA），Micron 厂商也提出了该电路设计，并强调了 OCSA 的重要性。如何通过电荷动态补偿，从电路上容忍工艺的适配，是先进内存工艺下存算一体原位计算芯片电路实用化的重要问题。研究高位和低位的偏移取消（OC）微操作的并行化机制，开发内存级并行性，实现时序重叠，降低总体延迟。

二、项目应用平台与基础

1．EDA 工业软件

本项目基于 HSPICE、Synopsys Design Compiler 等 EDA 工业软件开展。

2．技术基础

本项目基于 SPICE 面向对象硬件描述语言、Verilog-A 硬件描述语言等开展。大约 1 个星期即可上手。

三、项目要求

下列要求中 1 和 2 须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的计算机组成、Verilog 基础；

3．设计共源极负反馈晶体管电路，在感测放大器使能时，实现惠斯通电桥的负反馈，对感测放大器左右两边静态平衡电位补偿；

4．设计晶体管控制电路，通过左右两边互补存储结点的交叉，实现 nSA 左右两边栅极电压的对等；

5．撰写 HSPICE 集成电路模拟程序，对设计的感测放大器及其附属辅助电路进行延迟和功耗的仿真，分析所设计电路的版图（Layout）及面积。

四、项目开展

1．设计具备失配补偿功能的感测放大器，能够支持存算一体原位计算单元阵列“充一读多”的并行化功能；

2．分析电路微操作时序优化的比例；

3．分析所设计电路的 Layout 及面积；

4．实现面向存算一体原位计算，支持失配补偿功能的“充一读多”感测放大器电路拓扑及版图；

5．提出电路微操作时序优化机制。

项目 5：面向再生码分布式存储系统的数据修复性能优化（指导老师：李晓露）

一、项目背景

在国家数字经济大战略下，数据已经成为了重要的生产要素之一。随着海量数据的指数型增长，分布式存储系统作为提供海量数据存力的载体，其重要性不言而喻。根据中国信息通信研究院 2022 年 8 月发布的《中国存力白皮书》，分布式存储系统存力主要包括数据可靠性、存储开销、存储性能三大部分。

为了保障数据可靠性，分布式存储系统常采用多副本技术，即将一份数据复制多份存储在不同存储节点上以容忍节点故障，但这会导致巨大的存储开销。例如，三副本技术虽能容两错，但却需要消耗三倍于原数据量大小的存储空间。与之相比，纠删码在保障数据可靠性的同时，可大大降低存储开销，因而应用在了多个开源分布式存储系统中（如 HDFS，Ceph）。最常用的纠删码是 Reed Solomon（RS）码；（n，k）RS 码将 k 个原始数据块编码生成 n 个块并存储在 n 个节点上，使得其中任意 k 个块均能修复出 k 个原始数据块。与三副本技术相比，（4，2）RS 码也能容两错，但仅消耗两倍于原数据量大小的存储空间。

然而，RS 码的数据修复操作会在分布式存储系统内产生“修复带宽放大”的现象，即修复 1 个数据块需要读取 k 个块，使得修复开销极大，进而影响整个系统的存储性能。因此，本项目拟解决的科学问题是：如何在基于纠删码的分布式存储系统中进行高效的数据修复。

提升纠删码数据修复性能的方式主要有两种。第一种是设计新型纠删码以减少修复带宽，第二种是设计纠删码并行修复方案以减少修复时间：

再生码是一类可以减少修复带宽的新型纠删码。它将数据块切分为多个子块，通过重复利用下载的数据（本项目称为“数据复用”），使得数据修复节点仅从各节点下载少量数据即可修复 1 个数据块，以减少修复带宽。不少工业界开源分布式存储系统已部署再生码，如 Ceph 中已实现 Clay 码。

纠删码并行数据修复方案将修复操作进行拆分，并分发到不同的节点进行并行修复，从而减少修复时间。由于 RS 码具有编码结构简单、修复操作易被拆分等特点，学术界已有大量针对 RS 码并行修复方案设计的相关研究。

我们发现，可以通过结合再生码和并行修复，同时发挥两者的优势，进一步提升纠删码修复性能，以期更好的解决本项目的科学问题。然而，由于再生码编

码结构复杂，数据修复需要解一个复杂的线性系统，使得修复操作拆分困难，使得现有针对纠删码并行修复的研究工作依然存在大量挑战。例如，如何在异构环境下提升再生码数据修复的性能，以及如何在多故障场景下提升再生码数据修复的性能。

二、项目应用平台与基础

1．硬件平台

多节点的分布式集群。

2．软件平台

基于 ParaRC 进行本地测试和部署。

3．技术基础

图论、分布式存储系统、Redis 等。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定线性代数、离散数学基础知识；

3．具有较好的数据结构基础，以及较强的动手能力；

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．异构环境下再生码的数据修复性能优化：针对异构环境，构建数据修复有向图，并在有限时间内生成较好的解决方案。

2．多故障场景下再生码的数据修复性能优化：针对多故障场景构建数据修复有向图，并在有限时间内生成较好的解决方案。

项目 6：内存友好的高效 MoE（Mixture-of-Experts）架构（指导老师：施展）

一、项目背景

随着人工智能技术的飞速发展，大模型在各个领域展现出了强大的能力。然而，现有的 MoE（Mixture-of-Experts）架构在推理时需要将所有专家加载到内存中，导致内存占用巨大，难以在手机等内存受限的设备上运行。这限制了 MoE 模型在移动设备等资源受限环境中的应用。因此，设计一个内存友好的高效 MoE 架构，以降低推理时的内存占用，同时保持或接近原始 MoE 模型的效果，成为了亟需解决的重要问题。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器，以及内存受限的移动设备，以验证模型在不同硬件平台上的性能和内存占用情况.

2．软件平台

基于 Linux 操作系统、PyTorch 等深度学习框架进行编程和模型训练，同时可能需要使用 TensorRT 等模型推理加速工具，以提高推理效率.

3．技术基础

本项目基于深度学习、MoE 模型架构、内存优化技术、模型压缩与加速等相关理论与技术.

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 基础，熟悉深度学习框架如 PyTorch；

3．了解 MoE 模型架构和内存优化技术，具备一定的模型压缩与加速相关知识。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1. 内存优化策略研究

深入研究 MoE 模型在推理过程中的内存占用情况，分析不同组件和阶段的内存使用特点，提出针对性的内存优化策略。例如，探索参数共享、权重剪枝、量化等方法，减少模型在内存中的存储需求，同时保证模型的推理精度和效率.

2. 专家选择算法优化

优化 MoE 模型的专家选择算法，使其能够在推理时更加精准地选择所需的专家，减少不必要的专家加载。研究基于预测的专家选择方法，一次性预测出生成本次答案时所需要的专家，降低动态加载专家的频率和数量，从而减少内存占用和计算开销.

3. 模型推理加速技术

研究和实现 MoE 模型的推理加速技术，如模型剪枝、模型蒸馏、并行计算等，提高模型在推理时的效率，减少推理时间。同时，优化模型的计算图和内存访问模式，降低内存访问延迟，进一步提升推理性能.

4. 跨平台适配与测试

设计和实现 MoE 模型在不同硬件平台上的适配方案，确保模型能够在服务器和移动设备等不同环境下高效运行。进行跨平台的性能测试和内存占用测试，验证优化后的 MoE 模型在不同平台上的表现，为模型的实际应用提供可靠的数据支持。

通过以上研究和实践，旨在实现一个内存友好的高效 MoE 架构，推动 MoE 模型在更多场景和设备上的应用，为人工智能技术的发展贡献新的思路和方法。

项目 7：冗余数据消除技术（指导老师：王芳）

一、项目背景

随着信息技术的蓬勃发展和互联网的不断普及，企业与个人每天都会产生大量的数据。根据 IDC 的统计和预测，根据国际数据公司（IDC）的最新报告，全球数据量预计将持续快速增长，预计到 2028 年将超过 384.6 ZB（ZB，1 ZB = 10^{21} 字节），年复合增长率为 24.4%。存储容量增长速率远低于数据量的增长，大量数据被丢弃。而数据作为大数据应用以及 AI 应用的最重要驱动因素，有效保存管理更多数据具有重要意义。此外，在存储系统的各个层次上，在特定存储容量下存储更多数据，对于系统资源利用率、性能、能效、设备寿命等也均有提升作用。

以此为目标，人们提出了冗余数据消除技术。一般来说，无损的冗余数据消除技术可以分为三种：数据去重、差量压缩和传统压缩。传统压缩技术只能在有限的压缩窗口内查找冗余数据；差量压缩则可以通过相似性检索技术在整个存储系统内找相似的文件（或数据块），然后在相似文件（或数据块）之间查找冗余数据；数据去重则直接通过索引技术在整个存储系统内进行重复数据块的消除。

（一）数据去重

数据去重是一种无损压缩技术，通过维护重复数据的一个副本，消除重复的写数据以提高空间利用率，因此数据去重也被称作为单例存储。不同于以字符或字符串为单位的传统压缩技术，数据去重技术以文件或数据块为单位，实现了冗余数据的快速删除，很好地满足了企业对于冗余数据删除的吞吐量的需求。为了高效地消除存储设备中的冗余数据，降低存储开销，越来越多的企业已经将数据去重技术运用到存储系统中。传统的数据去重过程的四个阶段：数据分块、指纹计算、指纹索引和存储管理。其中数据分块和哈希计算占用大量的 CPU 资源，这成为数据去重系统中的潜在的瓶颈。

数据分块技术有定长分块和变长分块（即基于内容分块）两种，这里定长分块算法是最简单的实现方式，即根据文件内容的偏移位置决定分块；而基于内容分块算法是基于文件的内容来分块，这样有效地解决了文件修改导致的数据内容偏移的问题。相对而言基于内容分块的数据去重技术更能适应内容频繁修改的负载，所以能够发现和查找到更多的冗余数据。但是基于内容分块算法的计算开销

更大，因此常被用在延迟不敏感的场景，如备份和归档存储系统。

传统的判断数据块是否相同的方法是对数据块内容进行逐字节的比较。这种方法过于耗时，不适合应用于大规模存储系统中。为了解决这个问题，数据去重技术使用数据块的安全哈希摘要来代替数据块内容进行匹配，极大地简化了重复数据检测的过程。由于数据块的哈希摘要可以唯一地标识数据块，因此也被称为数据块指纹。数据去重技术使用了 MD5、SHA-1 等安全哈希摘要来唯一标识数据块，即给数据块装置了一个独一无二的指纹，通过这个指纹可以唯一标识数据块，这样简化了重复数据块的识别和匹配的过程：从全局的数据匹配缩小到数据块的哈希匹配。现有主流的数据去重技术采用了基于 Rabin 的分块算法、基于 SHA-1 的指纹算法，这使得数据去重技术在压缩存储空间的同时，不可避免地带来了计算开销和时延。

指纹索引指的是根据数据块的指纹查找匹配，如果有两个数据块的指纹匹配相等，那么它们所表示的数据块也相等。数据存储管理的时候，将非重复数据块直接写入存储系统；如果是重复数据块，则将重复的地址信息记录下来，以便于后续恢复操作。随着数据规模的持续增长，指纹的数据量会变得异常庞大，所以只能放在磁盘上存储和索引。如果将指纹索引存放在磁盘中会导致查询索引时需要频繁地访问磁盘，系统吞吐量过低。如果只索引部分指纹来减少指纹索引的空间占用又会损失去重率。

对于存储管理，在基于数据去重技术的备份系统中，数据块的数据内容只会保存一份并且被多个数据块所共享。对于非重复数据块，系统会将其数据内容保存在容器中。对于重复数据块，系统仅会记录之前存储过的数据内容所对应的地址信息。这种存储方式有效地提高了存储空间的利用率，但是同时也导致了碎片化问题的出现。碎片化问题指的是单个文件的数据块散落在不同的容器中，这会导致数据恢复 时需要更多的磁盘 I/O，降低了数据恢复的效率。

然而，根据 EMC 研究数据，经过数据去重后的存储系统中依旧存在大量的重复数据。这是因为数据去重只能实现粗粒度的冗余数据删除，无法消除数据块内部细粒度的冗余数据，这一类数据将被计算得出截然不同的 SHA-1 指纹，所以无法直接去重。为了解决这一问题，提出了差量压缩技术。

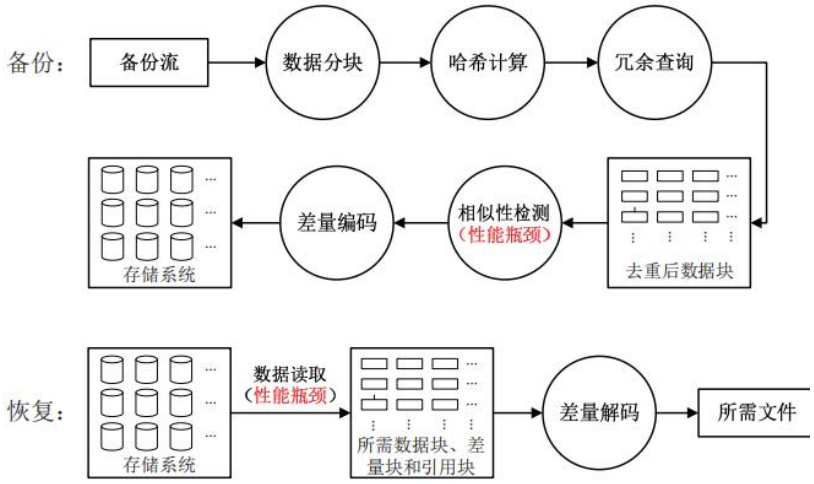
（二）差量压缩

差量压缩技术是一种细粒度的无损压缩技术，可以通过编码的方式有效地消除相似文件或相似数据块内部的数据冗余。由于两种技术具有互补性，差量压缩可以在数据去重的基础上进一步消除冗余数据，实现更好的冗余数据删除效果。整体分为四步，数据分块（与数据去重采用同一方法），相似性检测，差量编码，存储管理。具体来说，首先需要计算数据块的相似性特征值来检测潜在的差量压缩对象（相似性检测），接下来对于两个相似块，差量编码算法会利用滑动窗口技术寻找两个相似块之间的重复数据。对于重复的数据内容差量压缩会使用 COPY 命令进行编码，仅记录重复部分的位置和长度信息。 对于不重复的数据内容，差量压缩会使用 INSERT 命令进行编码，记录不重复部分的 数据内容和长度。以数据块 A 和 B 为例，假设数据块 A 是已经存储过的数据块，而 数据块 B 是数据块 A 的相似数据块（通过 sketch 比对），那么差量压缩可以根据数据块 A 的内容对数据 块 B 进行差量编码，从而只保存数据块 B 中与数据块 A 不重复的数据内容。在差量编码的过程中，被压缩的数据块被称为目标块，而另一个完全存储的数据块被称为引用块。

相似性检测是差量压缩的关键组成部分。主要使用文件或数据块的特征值来代替数据内容进行相似度判断。这一类相似性检测方法一般分为特征提取和特征匹配两个步骤。特征提取指的是从文件或数据块中提取出具有代表性的特征值的过程，而特征匹配指的是通过匹配特征值的方式查找相似文件或数据块的过程。相似性检测算法根据特征提取的粒度可以分为两大类，基于 shingle 和基于数据块。基于 shingle 的方法将数据对象划分成可重叠的定长子字符串，即 shingle。假设数据对象长度为 L，子字符串长度为 S，则有 L-S+1 个 shingle。基于 shingle 的相似性检测方法更准确，而基于数据块的相似性检测方法计算开销小。

最后的工作是对被判断为相似的两个数据块采用差量编码算法来缩减数据，消除数据块内部的重复信息。Xdelta 和 Zdelta 是目前存储系统中最常用的差量压缩算法。Xdelta 采用了一种基于滚动窗口的技术来产生字符串，这些字符串将被用做相似数据块之间冗余数据识别的最小单位。为了最大限度的寻找重复字符串，Xdelta 采用了逐字节滑动字符串窗口的方法。这样即使两个相似数据块之间的内容频繁地存在内容差异，这种滚动窗口技术还是能够足够精细地发现细粒度的重复字符串。其计算开销和索引开销较大。

由于差量压缩可以有效地消除相似对象之间的冗余数据，增强数据去重的冗余数据删除效果，因此二者常常联合起来进行冗余数据删除。对于联合使用数据去重和差量压缩技术的备份系统，其具体流程可参考如下：



联合使用数据去重和差量压缩技术的备份系统中，相似性检测和差量编解码是性能瓶颈。同时，在数据去重系统中，数据分块和指纹计算为潜在性能瓶颈。

因此，如何更好地设计实现数据分块、哈希计算、相似性检测和差量编解码模块，提升其吞吐量，是冗余数据删除这一领域亟需解决的核心问题。本研究旨在探索针对上述瓶颈的优化方法，通过改进数据分块算法、优化哈希计算过程，并设计更高效的相似性检测和差量编解码策略，从而提升冗余数据消除系统的整体性能。通过上述优化，期望能有效降低计算开销和时延，最终实现更加高效、快速的冗余数据删除，满足日益增长的数据存储需求。

二、项目应用平台与基础

1．硬件平台

本项目基于高性能计算服务器

2．软件平台

本项目基于 Ubuntu、Docker 平台，使用 C 语言、Rust 语言进行编程

3．技术基础

本项目基于文件存储备份系统、冗余数据删除等相关理论与技术。

三、项目要求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C 语言或者 Rust 语言基础、Ubuntu 系统基础；

3．具有一定的 Docker 使用基础，了解数据去重、差量压缩和数据压缩，文件存储及备份系统相关知识。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．分块算法优化研究：分块算法通过对文件的分割，让去重的粒度从文件级深入到块级，从而挖掘出更多的可以去重和压缩的数据内容。在部分仅需去重的数据缩减场景下，分块算法会因为其繁复的滚动哈希成为系统瓶颈，所以针对分块算法进行优化研究。可以硬件指令集加速分块算法的哈希滚动，也可以通过改变约束条件，跳跃分块等方式加速对数据的遍历。实践的重点是通过实验验证优化策略能够加速分块亦或是挖掘数据的重复内容。

2．缓存优化策略研究：数据缩减和恢复过程产生的元数据（如哈希值、特征值）和基块读取会占用一定的 IO 时间，在去重阶段，研究如何提升元数据的缓存效率，避免频繁从磁盘读取，可能涉及不同的缓存策略或对元数据访问模式的分析与优化。在差量压缩阶段，除了元数据，还需要探索如何优化基块的缓存方式，例如通过合理分配缓存空间或者利用基块访问的局部性规律提前加载所需数据。实践的重点是通过实验验证这些优化策略在减少读取时间和提升整体性能方面的效果。

3．相似性检测优化研究：相似性检测是数据缩减系统中的一个重要组成部分，决定这差量压缩的质量。同时，由于其吞吐量过低常常会成为整个数据缩减系统的瓶颈。在特征算法优化上，可以探索基于抽样的快速相似性检测算法，以降低线性变换的计算复杂度；在机器架构上，可以通过并行指令集加速特征提取，减少滚动哈希的时间；在空间局部性上，可以利用备份流的强局部性对相似块进行快速筛选。实践的重点是通过这些优化策略减少相似性检测的时间开销同时提升对相似块的检测能力。

4．元数据索引优化研究：元数据索引优化研究的目标是解决随着总数据量

增大导致的元数据开销问题，包括元数据存储占用的增长和索引构建开销的提升。可以利用批量操作或异步构建策略，将索引构建的高峰负载分摊到系统运行的低负载时间。同时，研究元数据压缩与预取策略，将高频访问的元数据索引预先加载到缓存中，减少从磁盘读取的频次。实践的重点在于设计高效的索引策略以降低系统资源消耗。

5．数据编码的优化研究：数据编码是挖掘数据重复性最细粒度的手段，可以更进一步提升数据的压缩比。去重和差量压缩所能做到的压缩比会被块级限制，所以数据编码能有效利用字符串，字节的重复性进行冗余数据的删除。然而复杂的熵编码以及随窗口增大造成的索引开销使得数据编码拖慢了系统整体运行效率。所以实践的重点在于如何通过现有的硬件条件（指令，缓存）配合高效的编码算法完成快速数据编码。

项目 8：异构内存系统的设计和管理（指导老师：魏学亮）

一、项目背景

数据中心应用对内存的需求正呈现爆发式增长，而大模型的兴起更是加剧了这一趋势。随着人工智能和机器学习等领域的快速发展，模型的规模和复杂性不断提升，所需处理的数据量也急剧增加。以深度学习为例，模型参数数量可能达到数十亿甚至数千亿级别，这不仅需要海量内存来存储参数和中间计算结果，还要求高效的内存访问和数据传输能力，以支持模型的训练和推理过程。

与此同时，数据中心还需同时运行多种应用和任务，包括数据分析、实时处理、虚拟化等。这些应用对内存的需求也在持续攀升。例如，在数据分析领域，海量数据的存储和处理是数据挖掘和预测分析的基础；在实时处理领域，快速响应用户请求和处理实时数据流，则要求内存具备足够的容量和速度来支持高效的数据访问。

面对内存资源的巨大需求，传统的单一内存架构已难以满足日益增长的性能和容量需求。这推动了异构内存技术的出现与发展。异构内存技术通过将不同类型的内存（如 DRAM 内存、持久内存、CXL 内存等）有机结合，构建分层内存架构。在异构内存架构中，不同内存类型的性能各有特点：DRAM 通常有着最好的访问性能，用于存放最频繁访问的热数据；持久内存有着较大容量，但访问性能相对较差，通常用于存储访问频率相对较低的设计；CXL 内存有着更大容量，但访问性能也更差。这种分层设计为内存系统的优化带来了新的机遇，同时也对内存管理提出了更高的要求。

异构内存调度。异构内存系统中的内存调度需要根据应用程序的访问模式和不同内存介质的特性来合理分配和管理内存资源。为了最大化系统性能和资源利用率，内存调度策略必须能够动态识别应用程序中的热数据和冷数据。热数据是指频繁访问的数据，对系统响应时间和性能有直接影响，因此需要将其放置在快速的 DRAM 内存，以减少数据访问延迟并提高应用执行效率。相对地，冷数据是不常访问的数据，如历史记录或不常用文件，将其迁移到大容量持久内存中，不仅能释放 DRAM 空间给更需要的热数据使用，还能利用持久内存的持久性在断电后存储数据，同时避免因 DRAM 容量限制导致的频繁数据交换和页面错误。

为实现高效内存调度，系统可采用智能页面迁移策略，如硬件或软件管理的

页面迁移机制。硬件管理的页面迁移开销低、迁移粒度灵活，而软件管理方案可提供复杂的迁移决策逻辑。无论哪种方案，都需要准确识别数据热度，并及时将其迁移到合适的内存层级。此外，内存调度还需考虑应用程序的动态变化，因为其访问模式可能随时间改变，如在不同工作负载下，某些数据可能从热数据变为冷数据，反之亦然。因此，内存调度策略应具备自适应能力，根据实时访问模式变化调整数据分布，以维持系统高性能和资源高效利用。

异构内存去重。异构内存去重技术是一种优化内存使用效率的方法，旨在减少不同内存设备（如 DRAM 内存和持久内存）中重复数据的存储。其核心原理是通过识别和消除重复数据，将相同内容的数据合并存储，从而节省内存空间。该技术通常包括以下步骤：首先，对内存中的数据进行扫描和哈希计算，识别重复数据；其次，通过指针或引用将重复数据指向同一存储位置；最后，释放冗余数据占用的内存。异构内存去重技术能够有效降低内存占用，提升系统性能，尤其在内存资源有限或数据冗余较高的场景中表现显著。

异构内存去重技术通过消除重复数据，显著节省内存空间，提高内存利用率，尤其在数据冗余较高的场景中效果突出。该技术能够降低内存访问压力，提升缓存命中率，从而加速数据处理和应用程序运行效率，优化系统性能。同时，在异构内存中，去重技术减少了对持久内存的写入次数，延长了硬件寿命。此外，通过减少内存占用和数据传输量，该技术还能降低内存子系统的能耗，特别适用于能效敏感的场景，如数据中心。在大数据、虚拟化或云计算环境中，去重技术有效管理海量数据，优化资源分配，降低成本。总之，异构内存去重技术在性能、能效和成本等方面带来了显著收益。

二、项目应用平台与基础

1．硬件平台

本项目基于高性能计算服务器和持久性内存服务器。

2．软件平台

本项目基于 Ubuntu 平台进行以及 Gem5，Qemu，Mess 等体系结构模拟器进行编程。

3．技术基础

本项目基于异构内存中的调度与去重相关理论与技术，在项目研发过程中需要使用 C++，python 等编程语言。

三、项目要求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python、Ubuntu 系统基础；

3．具有一定的 Gem5 等体系结构模拟器使用基础，了解计算机系统结构相关知识。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．负载热度特征分析：对不同负载的内存访问模式进行分析，识别数据的热点区域（频繁访问的数据）和冷点区域（较少访问的数据），以及各区域的数据冷热变化规律。

2．异构内存模拟器对比研究：对现有的异构内存模拟器（包括 Gem5、Qemu、Mess 等）进行对比研究，分析它们在模拟异构内存架构时的性能、精度和适用场景。

3．NUMA 架构下的去重算法实现：在 NUMA（非统一内存访问）架构下，实现去重算法，以减少跨节点访问的开销，提升整体性能。

4．页面热度预测算法实现：实现页面热度预测算法，用于预测内存页面的访问热度，以便在异构内存系统中优化数据迁移策略。

5．内存页面调度算法实现：在异构内存架构下，实现页面调度算法，用于在异构内存系统中优化数据的迁移和放置策略。

6．内存去重算法实现：在异构内存架构下，实现内存去重算法，用于消除异构内存中的冗余数据，从而节省内存空间并提高利用率。

7．持久性内存服务器性能评测：基于实际的持久性内存硬件平台，对分层异构内存调度机制进行性能测试，评估其在实际环境中的表现。

完成以上项目后，可选择进一步开展以下研究：

1．NUMA 架构下的去重算法优化：在实现 NUMA 架构下的去重算法基础上，进一步减少跨节点访问开销，提升内存访问效率。

2．页面热度预测算法优化：在实现页面热度预测算法的基础上，进一步优化预测精度和计算效率，增强数据迁移策略的准确性。

3．内存页面调度算法优化：在实现内存页面调度算法的基础上，优化数据迁移与放置策略，提高异构内存系统的整体性能。

4．内存去重算法优化：在实现内存去重算法的基础上，提升去重效率并降低计算开销，进一步提高内存利用率。

2．数据高效存储与计算团队

计算机学院“数据高效存储与计算”团队长期从事海量数据存储与处理、新型存储介质、设备及系统、新兴应用驱动的系统优化研究。依托于计算机系统结构国家重点学科和湖北省重点学科，是武汉光电国家研究中心存储研究部、“信息存储系统”教育部重点实验室、数据存储系统与技术教育部工程研究中心的重要组成部分。建有华中科技大学-华为光存储创新中心、变革性存储创新中心和云存储创新中心，已经和国内知名企业共建实验室。

团队历史悠久，由我国著名计算机存储技术先驱裴先登教授于上世纪八十年代创立，培养了众多优秀人才，其中不少人成为我国信息科技领域的顶梁之才，包括一批国际国内知名的教授和学者（包括 IEEE Fellow），国内著名头部企业的总裁和上市公司创始人等。信息存储教育部重点实验室首任主任谢长生教授是第二任团队负责人，带领团队取得了大量国际先进水平的成果。本团队为我国信息存储产业的起步和发展做出了重要的贡献，培养了大批优秀学生，不少毕业生成为国内外知名大学的教授和工业界的领军骨干。

团队现有教授 9 人，副教授/高级工程师 7 人，博士后 2 人。承担了国家重点研发计划项目（首席）、国家 973 计划项目、863 项目、国家自然科学基金项目（重点、面上、青年），并与华为、浪潮、阿里巴巴、腾讯等一大批知名企业开展了广泛合作。

目前研究领域主要包括下列四个方面：

面向人工智能的新型存算传架构及系统：面向人工智能不断增涨的高算存需求，研究新型异构存算传系统结构和关键器件、算法和加速器，服务人工智能国家战略需求。

分布式存储系统：针对云计算和大数据发展，研究分布式存储系统，服务大数据处理国家需求。

高性能存储设备及系统：针对对于存储性能无限需求，研究高性能存储设备及存储系统，服务国家自主可控的信息技术战略。

温冷数据长效新型存储：面向为了数据要素驱动的新型社会运行模式，研究新型存储介质及海量温冷数据长期存储及应用机制，实现原创长期低成本数据存储技术突破。

团队研究领域实现了存储介质、设备及系统核心技术的全面覆盖，并广泛新型体系结构、操作系统、大模型加速、区块链等诸多相关领域。相关的研究成果已发表在 IEEE/ACM Trans.等期刊和 ISCA、FAST、ASPLOS、VLDB、ATC、DAC、Eurosys，SOCC，ICS、ICDCS、DATE、ICCD 等重要国际学术会议上。团队连续在国际超算大会 ISC22、ISC23，超算存储 IO500 十节点排行榜连续两届世界第一名，团队连续三年获得华为云优秀合作伙伴奖；培养了多名硕士和博士获得“华为天才少年”，第 12 届亚太区大学生 RDMA 编程竞赛冠军等。

团队秉承自由科学探索和高水平工程实践相结合的风格，瞄准国际学术前沿进行创新性研究，同时也面向国家重大需求解决关键工程技术难题。作为高校科研团队，践行立德树人教育方针，实验室具有良好的科研氛围，系统性地培养学生创新能力、工程能力、合作能力和表达能力，努力提升每个学生的专业技术水平、科研能力和综合素质。

团队主要成员			
姓名	**职称**	**研究方向**	**联系方式**
曹　强	教授/博导	新型计算机系统结构，高性能存储，温冷长效存储	caoqiang@hust.edu.cn
谢长生	教授/博导	负责团队战略规划，技术指导，目前不招收新的研究生	cs_xie@ hust.edu.cn
吴　非	教授/博导	智能存储，存算传融合系统，非易失存储系统	wufei@ hust.edu.cn
万继光	教授/博导	计算机系统结构，分布式存储及云存储，键值存储系统，大模型推理与训练系统	jgwan@hust.edu.cn
谭志虎	教授/博导	计算机系统结构，嵌入式系统设计	stan@mail.hust.edu.cn
万胜刚	教授/博导	计算机系统结构，区块链存储，分布式存储系统	sgwan@hust.edu.cn
胡迪青	教授/博导	计算机系统结构，嵌入式系统设计，数据编解码技术	hudq024@hust.edu.cn
张静宇	研究员/博导	机器学习与大数据处理，超大容量新型存储技术	jy_z@hust.edu.cn
甘棕松	教授/博导	工业软件设计，超分辨存储系统	ganzongsong@hust.edu.cn
周　健	副教授/博导	存算传融合系统，新型计算机体系结构，操作系统	jianzhou@hust.edu.cn
姚　杰	副教授	计算机体系结构，嵌入式系统，磁光电融合存储系统	jackyao@hust.edu.cn
王海卫	副教授	计算机系统结构，新存储原理与设备	hiway@hust.edu.cn

周　游	副教授	固态硬盘，软硬件协同存储系统	zhouyou2@hust.edu.cn
张　猛	副教授	存储可靠性和纠错编码	zgmeng@hust.edu.cn
李国宽	副教授	计算机系统结构，机器学习与大数据处理，边缘智能存储与分析	liguokuan@hust.edu.cn
肖　亮	高级工程师	计算机系统结构，数字媒体技术	xiaoliang@mail.hust.edu.cn
鲁　凯	博士后	大模型推理与训练系统，分布式存储系统，键值存储系统	Kailu@hust.edu.cn

团队联系方式：

联系邮箱：caoqiang@hust.edu.cn，曹强老师

地址：华中科技大学光电信息大楼 B520 房间。

项目 1：TinyKV 分布式键值存储系统

一、项目背景

在信息技术飞速发展的今天，数据已成为企业最重要的资产之一。随着互联网、物联网和大数据技术的不断进步，企业面临着海量数据的存储、处理和分析挑战。传统的关系型数据库虽然在数据一致性和复杂查询方面表现出色，但在处理大规模数据时却常常显得力不从心，尤其是在高并发访问和动态扩展需求日益增长的情况下，性能瓶颈和扩展性不足的问题愈发突出。

分布式键值存储系统作为一种新兴的数据存储解决方案，正是为了解决这些问题而设计的。它通过将数据分散存储在多个节点上，实现了高可用性和高容错性，能够在节点故障时自动进行数据恢复，确保系统的持续运行。

但是，分布式键值存储系统在实现过程中同样要面临诸多挑战。分布式系统节点间的真实网络环境可能复杂多变，一个节点发出的数据包可能由于网络波动无法顺利到达目的地。更特殊的情况下，存储系统中可能还会出现节点故障、网络分区等状况。要在多个节点间快速地进行数据同步，应对复杂多变、高并发的网络环境，兼顾容错性和数据一致性，对存储系统的设计提出了要求。

TinyKV 来自于开源数据库企业 PingCAP 推出的 Talent Plan 开源数据库系列课程，旨在为学习和研究分布式键值存储系统提供一个简单而实用的框架和流程，使学习者无需搭建真实的分布式环境，无需详尽地了解成熟分布式键值存储系统中的每个组件，无需自主寻找可能使用到的开发依赖包，即可体验到分布式键值存储系统核心代码的开发。TinyKV 被广泛用于计算机科学课程和研究项目，帮助学生和开发者理解分布式系统的基本原理和实现技术。其代码结构清晰，易于阅读和修改，适合用于学习和实验。

Diego Ongaro 和 John Ousterhout 于 2013 年提出的 Raft 算法是一种用于实现分布式系统中一致性的共识算法，旨在解决在分布式环境中如何确保多个节点之间的数据一致性问题，以确保在节点故障或网络分区的情况下，系统仍然能够保持一致性和可用性。Raft 算法广泛应用于分布式数据库、分布式文件系统和其他需要高可用性和一致性的分布式系统中。TinyKV 使用基于 Raft 算法的分布式一致性协议，通过选举和日志复制机制，实现不可靠网络环境中多节点间的数据一致性。

本项目基于 TinyKV 提供的代码框架、实验流程和测试内容，旨在实现一个分布式键值存储系统中的共识算法和键值服务等核心模块，使其具备较好的容错、容灾能力和可扩展性。

二、项目应用平台与基础

1．硬件平台

本项目基于具有足够内存和 CPU 性能的服务器进行开发。

2．软件平台

本项目基于 Linux 操作系统和 TinyKV 分布式键值存储系统框架进行编程。

3．技术基础

本项目基于键值存储、数据库、共识算法相关理论及技术，项目开发使用 Go 编程语言。

三、项目要求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有面向对象程序开发基础，具有一定的阅读代码、项目开发能力，具有一定的自主解决开发过程中遇到的较有挑战性困难的能力；

3．了解 Go 语言基础语法，了解 Go 语言项目管理方式和测试方式；

4．具有一定的分布式系统开发基础。

四、项目开展

本项目包括 Standalone KV、Raft KV、Multi-raft KV 和 Transaction 四个部分。开展项目时，可依据各部分难度及各部分间的依赖关系，选取合适的开展顺序完成至少一项的项目内容。

Standalone KV 部分需要实现一个单节点的基本键值操作服务。这一部分的实现较容易，在这一部分中可以熟悉 Go 语言编程、键值操作和 Column Family 等基础概念。

Raft KV 部分需要实现一个具有容灾能力的键值服务，大致可以分为以下三个步骤：首先，实现基本的 Raft 一致性共识算法及对应的接口；然后，调用 Raft

算法接口实现在多服务器之间复制日志及回复的流程；最后，实现 Raft 的日志压缩和快照功能。

Multi-raft KV 部分需要在 Raft KV 部分的基础上，实现成员变动及领导权变动机制。为达成这一目的，需要对基础的 Raft 算法进行必要的功能扩展，然后基于此使集群支持领导权转移、成员变动、存储分区分裂等管理命令，使存储系统能够应对更加复杂的网络环境。这一部分还需要为平衡集群负载的调度器实现信息更新等功能。

Transaction 部分需要在数据库中构建一个事务系统，保证多客户端同时对相同键读写时不产生一致性冲突，并提供相应事务调用接口。

项目 2：基于 GPU-CPU 异构的大模型推理加速

一、项目背景

大模型是指具有大规模参数和复杂计算结构的机器学习模型。这些模型通常由深度神经网络构建而成，拥有数十亿甚至数千亿个参数。大模型的设计目的是为了提高模型的表达能力和预测性能，能够处理更加复杂的任务和数据。大模型在各种领域都有广泛的应用，包括自然语言处理、计算机视觉、语音识别和推荐系统等。大模型通过训练海量数据来学习复杂的模式和特征，具有更强大的泛化能力，可以对未见过的数据做出准确的预测。在过去的几年里，语言模型的参数规模从数亿增长到数千亿，甚至达到万亿级别。例如 OpenAI 的 GPT-3 拥有 1750 亿个参数，而 GPT-4 据称已经突破了万亿参数的大关。这种增长速度使得大模型在处理复杂任务、理解语言和逻辑推理方面表现出了显著的优势。

端侧大模型定义为运行在设备端的大规模人工 智能模型，这些模型通常部署在本地设备上， 如智能手机、IoT、PC、机器人等设备。与传统的云端大模型相比，端侧大模型的参数量更小，因此可以在设备端直接使用算力进行运行， 无需依赖云端算力。 端侧大模型在成本、能耗、可靠性、隐私和个性化方面相比云端推理具有显著优势，并能够以低能耗提供高效且安全的 AI 处理，减少延迟并保护用户隐私，适合个性化的 AI 应用。取决于行业对数据安全、隐私保护的需求、行业本身智能设备的普及程度以及 AI 大模型技术的成熟度，这些因素的相互作用和共同推动，端侧大模型将推动各行业智能化发展的步伐。

端侧大模型指在终端设备（如智能手机、平板、PC、智能穿戴设备、自动驾驶及具身智能等）上运行的大型预训练模型。相较于云端大模型，端侧大模型需要在资源有限的设备上高效运行，这对模型压缩、推理加速及能耗优化提出了更高的要求，其核心技术特点在于轻量化，通过量化、模型稀疏性等技术手段减少运行时需要的模型参数量以及计算量，但由于损失了参数信息，会使得模型的准确率显著下降。另一种方向是异构推理，利用 CPU 本地的内存甚至是更慢速的存储来放置模型，通过实时的参数加载来支持 GPU 进行模型推理，能显著减少显存容量对可选模型的限制，但由于需要频繁的传输参数，会使得 IO 成为严重的瓶颈。

因此，如何利用有限的硬件条件运行超出显存容量限制的大模型，以及如何

针对具有远低于云端算力的本地 GPU 算力的端侧设备进行推理加速，满足个人应用级别的延迟表现，成为了端侧大模型推理领域亟需解决的核心问题之一。本研究旨在探索面向端侧的大语言模型推理系统优化，在保证准确度要求的前提下，通过轻量化手段减少模型参数，以及利用 GPU-CPU 进行协同推理，充分利用端侧的异构算力完成推理过程。

二、项目应用平台与基础

1．硬件平台

基于具有足够内存和算力的服务器

2．软件平台

基于 Centos7、PyTorch 等深度学习及大模型库进行编程

3．技术基础

本项目基于深度学习、大语言模型、大语言模型推理、KV 缓存等相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．GPU-CPU 异构推理框架：CPU 内存相对于 GPU 显存更大且更廉价，利用类似于页面换入换出的思路，设计并搭建高效的异构推理系统，使 GPU 能够运行超过显存容量限制的大模型。

2．GPU-CPU 并行计算调度：异构推理系统具有 GPU 和 CPU 两种算力，基于模型参数切分和矩阵分块计算原理，设计高效的 GPU-CPU 间任务调度策略，同时

发挥两种的算力进行推理，提高推理性能。

3．模型参数量化算法设计：模型量化能够有效减少参数的空间占用。提出并实现新的模型量化算法，在一定参数精度保证下提高量化速度以及压缩比，从而加速推理过程。

4．稀疏化 KV 缓存算法：KV 缓存会随着上下文长度的增加线性增加，由于注意力机制加权平均的特点，少部分的 KV 缓存会决定注意力层的大部分的输出结果。设计一套稀疏化 KV 缓存算法，根据历史注意力分数识别 KV 缓存中的关键词，利用关键词来进行注意力机制的计算，提高推理速度。

5．参数稀疏性：大模型网络结构中存在激活函数来控制神经元的激活或响应，实际推理时大部分的神经元处于未激活状态，提前识别并规避不激活神经元，能有效加速全连接层的计算过程。设计并实现一种神经元激活预测算法，能够根据当前上下文识别全连接层中的激活神经元，并仅利用激活神经元完成推理过程。

6．融合算子设计和优化：大模型推理过程中不同的算子完成不同的计算任务，通过将算子的功能进行整合，减少中间值的来回访存和写回，以及充分利用局部性和高速的寄存器缓存，使得单个融合算子通过更少的 IO 开销完成更复杂的计算逻辑，从而提高推理性能。

3．多媒体流计算与存储团队

本团队隶属计算机系统结构国家重点学科和信息存储系统教育部重点实验室，团队现有教师5人，研究生20余人。本团队已承担和完成国家自然科学基金项目7项，湖北省自然科学基金5项及国防预研重点项目等其他各类项目30余项，在多媒体数据的处理、存储和传输及网络存储系统方面进行了大量的研究与开发工作，共计发表学术论文300余篇，其中被SCI、EI和ISTP三大索引收录100余篇次，申请专利30多项，已授权20余项。本团队还是AVS工作组的发起单位之一，积极参与了AVS视频编码标准的制定工作。该标准目前已经被国家标准化管理委员会正式批准为国家标准。本团队与美国、加拿大等国多个从事多媒体和智能无线感知的研究小组建立了良好的科研合作关系，近年来先后有30多人次出国参加国际学术会议和项目研发，因而能够及时了解国际前沿问题，在多媒体计算、网络存储和计算机视觉和人工智能等领域开展创新性的研究。

团队目前的主要研究方向包括网络流媒体系统与应用、计算机视觉与行业大数据分析等两个方面。具体研究内容与进展如下：

网络流媒体系统与应用

研究内容包括新一代视频编码标准VVC的优化与应用，3D视频编码，3D点云数据压缩，流媒体QoE的建模与评估，流式存储、流化处理与传输机制，流媒体调度与缓冲策略及三维实时建模与沉浸式虚拟会议室等。

本方向承担包括2项国家自然科学基金、3项湖北省自然科学基金，1项中国博士后科学基金，1项华为高校科技基金在内的多项研究课题，是中国音视频编解码标准AVS第二部分：视频的主要起草人之一。申请专利10余项，已授权8项。负责开发出实用化的嵌入式视频服务器，已在实际的视频监控系统中得到推广应用。在ACM Multimedia、IEEE ICME、IEEE ICCCN、Springer Multimedia Systems、《计算机学报》、《电子学报》、《计算机辅助设计与图形学学报》等国内外权威期刊和重要国际会议上发表论文50 余篇，其中近20余篇被SCI和EI 收录。在IEEE多媒体领域最重要会议ICME 2011上发表的论文获最佳提名奖。

计算机视觉与行业大数据分析

研究内容包括自然场景下对象表观表示方法，目标跟踪算法中的数据驱动机制，高分辨率图像中特定小目标识别技术，基于深度学习的排样算法，复杂场景

下的微人脸检测与识别，基于 RGB-D 相机的密集人群计数与定位，文字识别 OCR 解决方案。支持各种复杂票据的识别，支持结构化数据提取，以及识别后的智能处理。

本方向承担包括 1 项国家自然科学基金、1 项湖北省自然科学基金在内的多项研究课题，同时致力于计算机应用技术的研究与工程化。Springer Multimedia Tools and Applications、Neucomputing 等国内外权威期刊和重要国际会议上发表论文 10 余篇，其中近多篇被 SCI 和 EI 收录。在工程化方面，得到了华中科技大学产业孵化器的支持，专注于保险行业信息化与数据服务，应用人工智能及大数据分析技术，为保险行业提供信息化产品开发、控费增效整体方案解决，致力于保险+数字化赋能。完成了多项科研成果，包括：保险理赔大数据分析算法、移动视频技术在保险理赔中的应用系统、医疗票据的识别系统、理赔风控系统、理赔控费系统等。拥有与保险行业信息化与数据服务相关的 5 个软件产品证书、43 个软件著作权证书、5 个专利。

团队成员：

郭红星	团队负责人 副教授，主要研究领域为网络流媒体系统与应用		
孙伟平	副教授，主要研究领域为计算机视觉	范晔斌	博士，主要研究领域大数据分析
李　榕	博士，主要研究领域大数据分析	夏　涛	博士，主要研究领域大数据分析

团队联系方式：

联系邮箱：guohx@hust.edu.cn

地址：华中科技大学南一楼中 407 房间。

4. 智能数据存储与管理团队

计算机学院“智能数据存储与管理团队”依托于计算机学院存储所，武汉光电国家研究中心，是信息存储系统教育部重点实验室、数据存储系统与技术教育部工程研究中心、光谷实验室的重要组成部分。建有华中科技大学--腾讯公司智能云存储技术联合研究中心。是科技部重点领域创新团队、教育部“长江学者和创新团队发展计划”创新团队的重要组成部分。

团队现有教授 2 人，副教授 1 人，副研究员 1 人，讲师 1 人，博士后 2 人，其中长江特聘教授国家级人选 1 人。获评国家自然科学基金委创新研究群体。承担了国家重点研发计划项目（首席）、国家 973 计划项目、863 项目、国家自然科学基金项目（重点、面上、青年），并与华为、腾讯、浪潮、达梦、中移动等一大批企业开展了项目上的深入合作。

目前研究领域以 AI for Storage 展开，主要包括下列四个方面：

认知存储系统：基于语义存储研究基础，夯实面向智能应用的存储系统的系统研发，服务智能应用数据基座建设的国家需求。

智能云存储系统：基于智能缓存与云盘装箱研究基础，升级开展面向智能应用的智能存储系统研发，服务智能应用数据基座建设的国家需求

动态互联分布式系统：基于新硬件研究基础，夯实面向高性能处理的分布式系统的系统研发，服务大数据处理国家需求。

自治数据库系统：基于智能数据库运维与索引推荐研究基础，升级数据库面向大模型推理时的精确性与稳定性，服务国家信创产业升级。

团队是全球最早开展 AI for Storage 研究的团队，并在数据库调参领域具有不可撼动的领先地位，并广泛涉及智能算法、暗数据、内存调度、学习型索引、大模型加速、系统运维等诸多相关领域。相关的研究成果已发表在 IEEE/ACM Trans.等期刊和 SIGMOD、VLDB、ICDE、ASPLOS、DAC、MM、IJCAI、EuroSys、ATC、PACT、CIKM、ICME、DATE、ICPP、ICCD 等重要国际学术会议上，实现了数据库和系统领域顶级会议的全覆盖。相关研究成果在通讯、航天和社交分析等诸多复杂且重要现实场景中进行了应用，解决了腾讯、浪潮、华为、中移动、达梦等服务器提供商的卡脖子技术难题，构建了极致性价比的下一代云存储系统，获湖北省科学发明奖一等奖、中国电子协会科学发明二等奖。团队三次获得中国国际大

学生创新大赛全国金奖（原中国国际“互联网”+大学生创新创业大赛），两次获得“创青春” 全国大学生创业大赛全国金奖，获得第十三届“挑战杯”中国大学生软件设计大赛全国一等奖，获得第一届全国大学生信息存储技术竞赛特等奖，获得2024年川渝大学生人工智能大赛暨腾讯开悟人工智能全球公开赛AI芯片算子开发赛道全国一等奖。

团队毕业生质量获得企业一致认可，在业界具有良好口碑。每年毕业生被华为、腾讯、阿里、美团、快手、抖音等一流IT企业“抢人”，张霁-华为天才少年，张煜-腾讯技术大咖，张光钰、张嘉伟-华为星，年薪不是问题，前进永不停息。

团队成员：

周　可	团队负责人 教授，主要研究领域为大数据处理，云存储		
王　桦	教授,主要研究领域为云存储、智能存储、智能资源调度	李春花	副教授,主要研究领域为智能云存储、索引优化、存储安全
刘　渝	副研究员，主要研究领域为相似性哈希、认知存储、智能运维、暗数据	张东映	讲师,主要研究领域为卫星图像数据存储与管理、智能运维
洪志明	博士后，主要研究领域为向量数据库、学习型数据压缩	王　鹏	博士后，主要研究领域为智能缓存、内存资源管理与调度

团队联系方式：

联系邮箱：liu_yu@hust.edu.cn，刘渝老师

网址：http://idsm.wnlo.hust.edu.cn/index.htm

地址：华中科技大学新光电大楼C536。

项目 1：基于 FPGA 加速的学习型高倍率 SSD 透明压缩研究

一、项目背景

随着信息技术的迅猛发展，数据的产生速度和规模不断增加，尤其是在大数据、云计算和物联网等领域，数据量已经达到了前所未有的水平。作为重要的存储设备，SSD 在性能、效率、节能、可靠性等方面已全面超越 HDD。考虑到我国的 HDD 产业接近空白，处于被卡脖子的地位，国内专家学者呼吁启动以固态硬盘 SSD 取代机械硬盘 HDD 的存储革命，打破垄断。

然而，SSD 也存在价格较高、写入寿命有限、写放大等问题。SSD 透明压缩通过集成计算单元（如 FPGA 等）在存储路径上完成数据的压缩和解压缩操作，有助于提升 SSD 存储效率，减少对系统资源的占用。然而，经典透明压缩算法对数据冗余度较大的视频、图像、音频等数据的压缩能力有限，而提高该类型数据的压缩率，有助于更大程度上扩充 SSD 存储效率，延长其使用寿命，降低写放大。然而支持高压缩率的学习型算法推理速度难以匹配 SSD 的高带宽，造成带宽资源浪费。为解决经典透明压缩算法压缩率较低，而支持高压缩率高感知质量的学习型压缩算法编解码速度难以匹配 SSD 的高带宽，造成 SSD 带宽资源的浪费传统的数据存储方式面临着存储容量不足、访问速度慢和成本高等问题。本研究基于 FPGA 计算加速器特性，调整学习型数据压缩模型结构，提高模型的数据处理并行度，替换推理短板模块，形成适配于 FPGA 的学习型（AI）数据压缩方案，在 AI 模型--计算单元--存储单元之间进行综合优化。探索基于 FPGA 加速的学习型数据压缩方案，实现高倍率高质量高速度的学习型 SSD 透明压缩算法。

二、项目应用平台与基础

1．硬件平台

GPU、FPGA 以及 SSD 硬盘。

2．软件平台

基于 PyTorch 或 Tensorflow 深度学习框架以及 Vivado、Vitis HLS 等 FPGA 开发平台进行编程

3．技术基础

深度学习、学习型数据压缩、FPGA 硬件加速方法等相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足，满足 4 者优先：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习等相关理论与技术。

4．具有一定的 FPGA 开发经验或知识储备。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．学习型图像/视频/音频压缩研究：选择一种类型数据，如图像、视频、音频，开展基于深度学习的数据压缩研究。研究方向为高压缩率、高压缩质量、变码率、渐进式压缩、重要性感知。

2．面向高压缩解压缩速度的模型优化：通过多种手段，如引入小波变换、傅里叶变换、模块融合、Mamba、分块压缩等方法提升模型压缩/解压缩速度。

3．学习型数据压缩模型轻量化研究：在不显著影响压缩质量的前提下，对学习型数据压缩模型开展轻量化研究，包括模型蒸馏、减枝、量化，减少模型体积、提升模型推理速度、探索定点数据训练方法。

4．面向 FPAG 的压缩模型推理加速研究：在 FPGA 上实现学习型压缩模型的推理加速，研究影响推理速度的硬件资源优化方法以及模型优化方法。

5．基于学习型数据压缩技术的 SSD 透明压缩研究：研究将 FPGA 赋能的学习型压缩算法整合为 SSD 透明压缩模块，提升 SSD 存储空间，延长其寿命。

项目 2：面向新型计算模式的 AI 使能分布式存储架构及关键技术

一、项目背景

随着新型计算模式的兴起，新型存储和网络硬件技术被广泛应用。新型存储介质正在模糊传统存储层级架构中分层的界线，新型高速网络正在削除本地和远程访存的差异，通过 RDMA 访问远程非易失性内存甚至快于访问本地慢速存储设备，这二者为分布式存储系统进行层次化和平行化访问提供了多种 I/O 路径选择，然而传统架构固化的 I/O 路径无法发挥新硬件的技术优势。同时，新型计算模式丰富多变的负载特性，驱动着分布式存储系统内各个部分的智能化演进，传统架构中孤岛式的 AI 使能模块缺乏全局的统筹，存在训练资源瓶颈和优化目标不一致等问题，限制了 AI 使能技术的真正落地。如何设计一种新型分布式存储架构，能够充分发挥多 I/O 路径和多任务 AI 使能优势，是本课题拟解决的关键技术问题。

为解决上述关键科学问题和关键技术问题，本课题的研制目标包括：建立面向新型计算模式的分布式存储架构，支持多 I/O 路径的动态互连，实现多种存储介质与 DRAM 的层次化和平行化灵活组合； 研究“存储 AI 大脑”，实现分布式控制面和数据面多任务 AI 使能的统筹调度，以满足新型计算模式数据访问的多样化和动态性需求，为系统研制及其它课题提供架构支撑与智能决策服务。

课题的研究内容着重从以下几个方面展开。首先，在 AI 使能的动态互联分布式存储架构方面，重点探索“存储 AI 大脑”的组成结构与工作机制，包括全局统一进行多源数据采集、多模型训练、推理，以及根据应用负载变化与当前资源可用情况，实现智能化决策和多任务统筹调度的方法。在动态互联架构中，结合层次化和平行化设计原则，研究存储结构中多种存储介质之间的相对位置、连接和包含关系、以及多 I/O 路径选择算法和可重定向映射方法等。同时，通过引入任务驱动的存储系统软件架构，将存储硬件和软件资源进行抽象，建立任务驱动的统筹调度模块，对用户 I/O 任务和系统非 I/O 任务进行统一并发调度，以实现全局任务统筹。

其次，在 AI 使能的分布式存储控制面技术方面，本课题通过分析用户负载动态变化情况并进行负载预测，在控制面实现系统自优化。包括：用户负载分析与预测技术，通过分析用户负载特征，构建用户负载、多维度资源和系统性能的

关系模型，完成对未来一段时间内多尺度、多维度的用户负载预测；系统性能自动调优技术，针对系统参数进行重要性排序，筛选影响力高的参数，然后构建参数模型并进行安全性评估，减少调优过程中系统故障和性能异常发生的次数。

最后，在 AI 使能的分布式存储数据面技术方面，通过分析不同应用数据的访问特点并采用智能缓存策略提高访问效率。包括：冷热数据感知技术，根据数据的访问频率、最近访问时间和重要性，感知各存储节点上的数据温度，将数据划分为不同的冷热等级，实现冷热数据的高效管理和优化访问；智能缓存技术，包括基于工作负载特征和访问模式的智能预取算法、多级缓存体系结构中不同级别之间的数据迁移和准入策略、以及多节点层次化和平行化动态变化环境下的智能缓存替换策略。

总体而言，本课题通过从架构设计、控制面优化到数据访问处理的一系列研究，提出了一种创新的面向新型计算模式的 AI 使能分布式存储架构，旨在克服传统系统在新型存储硬件和复杂计算负载条件下所面临的瓶颈问题。其核心技术成果有望提高系统的存储效率、数据访问灵活性和智能化水平，并为新型计算模式的实际应用提供坚实的技术支撑。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的高性能存储服务器集群，配有 100Gbps IB 互联网络，以及相关的存储硬件设备。

2．软件平台

基于 Ubuntu 操作系统、Ceph 分布数存储系统，以及深度学习及大模型库等进行编程

3．技术基础

本项目基于深度学习、强化学习、大语言模型、参数高效微调、向量检索、RDMA、新型存储硬件等相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

4．对其他项目相关的技术有兴趣并且有相关理论和技术基础。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．基于大模型的分布式存储系统 Ceph 的异常检测功能研制：本研究结合大模型语言理解与深度学习，构建 Ceph 异常检测模块。包括：分析 Ceph 架构及常见异常，搭建数据结构化框架处理日志与数据，运用深度学习与机器学习实现异常检测，并通过大模型进行根因分析与解决方案推荐。研究目标是构建智能异常检测模块，实现 Ceph 系统的异常识别、预测与根因分析，提升系统性能与故障恢复能力。

2．基于近数据存储的大模型推理加速研究：探索使用 FPGA 模拟带计算资源的 SSD 进行大模型推理的架构，优化内存与 SSD 的数据传输和管理策略，降低 CPU 占用率,提升推理速度。研究目标是实现大模型推理速度提升至 20 tokens/s，内存占用降低至 5GB 以下，以提高端侧和集群侧的推理效率和降低成本。

3．多属性过滤条件下的向量检索研究：现有工作难以高效处理多过滤条件下的向量检索本，研究将标量和向量合并成高维的向量，以实现向量与其属性之间的统一，避免分阶段搜索。通过依赖于属性值的距离计算函数实现图的构建，查询感知的方式引导图的遍历路径，实现多过滤条件下的高效近似检索。研究目标是在多过滤条件的向量检索场景下，性能优于现有向量数据库 milvus 等；在单范围过滤条件下，性能不弱于现有最好的工作 SeRF。

4．基于进化算法的文件系统调参：研究内容是解决强化学习在文件系统调参任务中由于稀疏奖励与环境不稳定导致的收敛时间波动较大的现象。研究目标是优化文件系统的调参方法实现收敛时间稳定在较低水平且调参效果不变。

5．基于新型 LLM Agent 架构的存储系统智能化控制：设计新型的 LLM Agent 架构实现存储系统中各模块的调度与调优。研究目标是通过 LLM 的统筹决策，实

现运行存储系统中各模块的合理调度以实现整体性能最优。

6．基于 ZNS SSD 和 FDP 的高效数据管理关键技术研究：针对分区命名空间固态硬盘的存储特性，研究数据管理策略和性能优化方法，旨在提升存储系统的读写性能和设备寿命，为数据中心应用提供更好的存储解决方案。本研究主要包含但不限于以下内容：1）设计新型数据放置和调度策略；2）探索 ZNS SSD 作为缓存介质的可行性；3）优化垃圾回收算法，元数据管理方法等。预期目标：1）利用 ZNS 的特性实现更好的数据放置，并有效降低数据迁移时的开销；2）实现高效处理大模型计算中的海量小文件的多介质缓存组件。

7．新一代互联技术探索：随着摩尔定律的减缓，登纳德缩放定律的失效，我们对多设备间的互联有了更迫切的需求。单个设备在如今海量数据的时代已很难有所发挥，但当多个设备互联成一个系统之后，一切就发生了改变。当前，以 CXL 为代表的新一代互联技术正处于蓬勃发展阶段。从以 Intel、Nvidia 为首的芯片侧，到以 Meta、微软为首的应用侧，再到以三星、镁光为首的存储侧，工业界前沿纷纷积极投入 CXL 协议的制定和研发，技术前景不言而喻。预期目标：当前 CXL 技术离真正落地还有一段距离，互联的难点还有很多。本研究将探索以下几个研究点：1）高效的缓存一致性协议；2）针对新协议设备间的通信模式优化；3）更高效的索引机制。

8．智能缓存策略及系统关键技术研究：随着接入存储系统的应用数目和系统规模的不断增大，存储系统数据访问模式越来越复杂，访问数据流呈现时空局部性交错等现象，传统的启发式缓存策略无法很好适应动态变化的应用负载，相比之下，基于机器学习的智能缓存策略能够有效识别复杂的访问模式，适用于多样性混合负载，其缺陷在于算力开销大、推理时延高。预期目标：1）提出基于机器学习的智能缓存策略，并设计缓存系统；2）相关技术在华为公司实际应用，提升存储系统整体性能；3）进一步探索基于数据内容的缓存策略研究

5. 先进可扩展计算与系统团队

计算机学院“先进可扩展计算与系统团队”依托于计算机系统结构国家重点学科和计算机软件与理论湖北省重点学科，是大数据技术与系统国家地方联合工程研究中心、服务计算技术与系统教育部重点实验室、集群与网格计算湖北省重点实验室的重要组成部分。建有华中科技大学-华为数据中心架构创新中心、华中科技大学医疗健康大数据中心、华中科技大学数据流-大数据中美联合研究中心等。是科技部重点领域创新团队、教育部“长江学者和创新团队发展计划”创新团队的重要组成部分。

团队现有教授 8 人，副教授 3 人，讲师 2 人，博士后 3 人。其中长江特聘教授国家级人选 2 人，国家杰青 / 万人计划科技领军人才 2 人，973 首席科学家 2 人，中组部青年拔尖人才 2 人，基金委优青 2 人，教育部青年长江 1 人，省杰青 3 人。获评科技部重点领域创新团队、教育部创新团队、湖北省自然科学基金创新群体。承担了国家重点研发计划项目（首席）、国家 973 计划项目（首席）、863 项目、国家自然科学基金项目（杰青、重点、优青、国际合作、面上、青年），并与华为、浪潮、曙光、中移动、Intel 等一大批企业开展了广泛合作。

目前研究领域主要包括下列四个方面：

分布式异构内存服务器：基于异构内存研究基础，夯实面向大数据处理的分布式系统的系统研发，服务大数据处理国家需求。

高效动态图专用计算机：基于静态图加速器研究基础，升级开展面向动态图专用计算机的研发，服务大数据处理国家需求

高效能大模型推理加速器：立足大模型服务成本高重大需求，研究高能效大模型推理机及关键技术，服务国家人工智能产业升级

多层数据流架构及系统：面向新型数据流模式，研究多层数据流融合架构及软件系统，瞄准新需求，建立新的增长点。

团队研究领域实现了图计算机核心技术的全面覆盖，并广泛涉及内存计算、大模型加速、数据流等诸多相关领域。相关的研究成果已发表在 IEEE/ACM Trans. 等期刊和 ISCA、MICRO、HPCA、FAST、ASPLOS、ICS、SC、ATC、CGO、PACT、HPDC、ICDCS、VEE、MASCOTS、DATE、ICCD 等重要国际学术会议上，实现了系统结构领域顶级会议的全覆盖。相关工作入选 2023 年度“CCF 优秀博士学位论文激励计

划”、“CCF 体系结构优秀博士学位论文激励计划”，2024 年度 CCF 高性能计算“博士学位论文激励计划”。研究成果在电力、金融和社交分析等诸多复杂且重要现实场景中进行了应用，解决了浪潮、华为、曙光等服务器提供商的卡脖子技术难题，构建了极致性价比的下一代内存型存储系统，获湖北省科学技术进步奖一等奖、世界存储领域顶级奖项“奥林帕斯先锋奖”。团队连续在 Graph Challenge 竞赛（图计算领域最具影响力的国际赛事之一）获得全球总冠军（国内首次，2021-2022），研发的图计算机部分指标连续四年登顶 Graph 500 及 GreenGraph 500 全球最权威图计算榜单（2021-2024），连续四年获得中国国际大学生创新大赛全国金奖（原中国国际“互联网”+大学生创新创业大赛）、获得第十三届“挑战杯”中国大学生创业计划竞赛全国金奖。

团队毕业生质量获得企业一致认可，在业界具有良好口碑。每年毕业生被华为、腾讯、阿里巴巴、百度、微软、亚马逊、IBM 等一流 IT 企业“抢人”，年薪屡创新高，系统级人才供不应求。

团队成员：

金　海	团队负责人 教授，主要研究领域为大数据处理，并行分布式计算、云计算		
廖小飞	团队负责人 教授，主要研究领域为并行分布式计算，大数据处理，图计算		
蒋文斌	教授,主要研究领域为深度学习系统、图计算	刘海坤	教授,主要研究领域为内存计算、近数据处理
邵志远	教授，主要研究领域为图计算机、分布式计算	郑　龙	教授,主要研究领域为图计算机、运行时优化
张　宇	教授，主要研究领域为图计算、体系结构	张书豪	教授，主要研究领域为大模型系统
郑　然	副教授,主要研究领域为深度学习、分布式优化	叶晨成	副教授,主要研究领域为持久内存、近存计算
赵　进	副教授,主要研究领域为图计算	毛伏兵	讲师,主要研究领域为系统软件和体系结构
姚鹏程	讲师,主要研究领域为体系结构、AI/图计算芯片设计	黄　禹	博士后,主要研究领域为图计算、存算一体

王庆刚	博士后,主要研究领域为图计算	段卓辉	博士后,主要研究领域为内存计算

团队联系方式：

联系邮箱：wenbinjiang@hust.edu.cn，蒋文斌老师

地址：华中科技大学南一楼西 122 房间。

项目 1：面向医疗领域的大模型智能体设计及部署（指导老师：李钦宾）

一、项目背景

近年来，随着深度学习技术的突破性进展，大模型（如 GPT、Qwen 等）凭借其大规模参数和预训练能力，在自然语言处理等领域表现出卓越的性能。大模型不仅能够处理复杂的多模态数据，还具备强大的上下文理解和知识推理能力，为人工智能系统带来了颠覆性的变革。在各行业中，大模型逐渐成为推动智能化发展的核心引擎。

大模型智能体（Agent）是一种以大模型作为大脑的智能系统，能够理解用户意图并进行复杂的推理和决策。大模型智能体通常包含记忆、规划及工具使用模块，通过结合外部知识库和各类工具，大模型智能体可以实现从数据获取到智能推理的一体化服务，在实际应用中展现出强大的灵活性和适应性。大模型智能体已经在代码生成领域得到了大量应用。

在医疗领域，大模型智能体正逐步展现其在诊疗辅助、健康管理和医学研究等方面的应用潜力。例如，智能体可以通过处理患者病历文本，为医生提供诊断建议；智能体还能为患者提供医疗建议，帮助健康管理。这些能力有助于提升医疗效率、优化资源配置，并改善患者的医疗体验。

大模型智能体在医疗领域的应用前景广阔，也面临以下诸多挑战：1）如何设计多智能体协作框架以完成不同种类的医疗任务（例如挂号，医疗问答等）；2）如何选择合适的数据源和工具赋能智能体，使其能更好的完成推理任务；3）如果提高智能体的效率，降低计算资源开销。

本项目旨在设计并部署一个面向医疗领域的大模型智能体，充分发挥其在自然语言处理和知识推理方面的能力。通过构建医疗知识库、设计多智能体协作流程等方法，本项目期望解决大模型智能体在医疗领域的适应性和效率问题。最终目标是实现一个高效、可扩展的医疗智能体系统，推动人工智能技术在医疗领域的实际应用落地。

二、项目应用平台与基础

1．硬件平台

具有充分算力的 GPU 服务器集群。

2．软件平台

基于 Python 进行开发。

3．技术基础

本项目基于深度学习、大语言模型、检索增强生成等相关理论与技术。

三、项目需求

以下要求为必须满足的条件：

1．熟练掌握 Python 编程语言；

2．踏实肯干，有充分时间投入；

3．对大语言模型有基础的了解。

四、项目开展

本项目针对医疗智能体的核心功能，提供以下研究方向：

1．多智能体协作框架设计：通过设计多个不同角色的智能体，开发多智能体的高效协作框架，能实现医疗相关的任务。

2．医疗知识库的构建与优化：系统化整理和编码医学领域的知识，构建高效的检索式知识库和工具库，以支持大模型的推理任务。

3．多智能体的分布式部署：在分布式集群上部署及优化多智能体，在任务量大的场景能及时响应并完成任务需求。

项目 2：稀疏矩阵乘法加速策略研究（指导老师：蒋文斌）

一、项目背景

稀疏-密集矩阵乘法（Sparse-Dense Matrix Multiplication, SpMM）是科学计算、机器学习、图像处理等领域中的一项核心且关键的计算操作。由于稀疏矩阵中大多数元素为零，传统的矩阵乘法算法在计算过程中往往会浪费大量资源。因此，如何高效加速 SpMM，成为了当前研究的一个重要方向。

近年来，GPU 因其强大的并行计算能力，成为了加速稀疏矩阵乘法的重要平台。特别是张量核心（Tensor Cores, TCs）被广泛用于矩阵乘法加速，具有显著的性能优势。然而，张量核心通常仅支持分块矩阵乘法，并且对于稀疏矩阵的计算，依赖于对矩阵进行分块处理和压缩。随着 GPU 硬件架构的不断发展，安培架构 GPU 中的张量核心引入了全新的稀疏张量核心指令（Sparse Tensor Cores, SpTCs），这些指令能够跳过对零值元素的计算，从而有效地提高了张量核心的计算吞吐量。然而，当前的稀疏张量核心指令仅支持横向 2:4 结构化稀疏矩阵的运算，而在实际的科学计算中，稀疏矩阵通常是非结构化的，二者之间存在显著差异。因此，在实际应用中，如何将非结构化的稀疏矩阵转换为符合 2:4 结构化稀疏格式的矩阵，成为了一个亟待解决的关键问题。

与此同时，斯特拉森矩阵乘法通过分治法优化了通用矩阵乘法，显著减少了乘法操作的次数。与传统的矩阵乘法方法（时间复杂度为 O(n^3)）相比，斯特拉森算法将计算复杂度降低到约 O(n^2.81)，这是通过将每个矩阵分割成更小的子矩阵并递归地计算七个子问题来实现的。但由于稀疏矩阵的特性，斯特拉森算法直接应用至 SpMM 中可能存在内存开销大和冗余计算的问题。如何将斯特拉森算法与稀疏矩阵的稀疏性相结合，从而提高计算效率，降低内存需求，成为了一个重要研究方向。

因此，如何在 GPU 上高效地加速稀疏矩阵乘法，充分利用稀疏矩阵的特性，减少计算资源浪费，是本项目的研究目标。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器 (RTX 4090)

2．软件平台

Visual Studio Code，C/C++、Python、CUDA 开发环境

3．技术基础

本项目基于 GPU 和稀疏矩阵的相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C/C++、Python 基础，并对 GPU 编程充满浓厚兴趣；

3．对矩阵运算操作有基本了解。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．面向稀疏矩阵的最大匹配策略研究：本项目结合二分图最大匹配问题，设计并计划实现一种能够将非结构化稀疏矩阵转化为 2:4 结构化稀疏矩阵的算法。具体方法是将稀疏矩阵按行划分为若干个窗口，并在每个窗口中压缩零列，仅保留非零列。接着，将两个非零列组合成一个单元，并通过最大匹配算法寻找符合 2:4 结构化稀疏格式的分块。这一策略能够有效地适配 GPU 的稀疏张量核心指令（SpTCs），从而显著提升稀疏矩阵的计算性能。目前，该项目已取得显著进展。我们验证了 SpTCs 能够实现比传统张量核心（TCs）高出两倍的吞吐量，并且已成功实现简化版本的匹配策略。尽管如此，仍需要进一步优化算法，以提高其计算效率和适用性。

2．稀疏斯特拉森矩阵乘法研究：结合斯特拉森矩阵乘法，设计并实现一种高效的 SpMM 计算内核。分析不同算法之间的计算复杂度，并比较新算法与现有的稀疏矩阵乘法算法（如 cuSPARSE）在不同稀疏度和矩阵规模下的执行效率。

6. 分布式系统团队

计算机学院“分布式系统团队”依托于计算机系统结构国家重点学科和计算机软件与理论湖北省重点学科。

团队现有教授 4 人，副教授 4 人，讲师 2 人，博士后 5 人。其中万人计划科技领军人才 1 人，海外优青 1 人。获评科技部重点领域创新团队、教育部创新团队、湖北省自然科学基金创新群体。承担了国家重点研发计划项目（首席）、国家 973 计划项目（首席）、863 项目、国家自然科学基金项目（杰青、重点、面上、青年），并与阿里、华为、腾讯、浪潮、曙光等一大批企业开展了广泛合作。

目前研究领域主要包括下列四个方面：

云计算的核心系统软件：包括强隔离可定制的云操作系统内核、异构资源虚拟化、高可扩展服务器无感知计算架构等关键领域和前沿方向。

智能算网融合：构建分布式大模型训练和推理系统，设计单/跨数据中心分布式算网协同框架，提高面向算力网的大模型训练和推理效率。

边缘智能：尝试构建针对异构边缘设备的统一机器学习框架，探索基于异构计算资源的动态调度,研究智算任务的云原生化。

可扩展区块链：开展弹性图式区块链的基础理论与关键技术自主创新研究，探索零知识证明和多方安全计算等技术在提升区块链系统性能和保护数据隐私方面的应用。

多层数据流架构及系统：面向新型数据流模式，研究多层数据流融合架构及软件系统，瞄准新需求，建立新的增长点。

团队定位于云计算、数据中心、边缘计算、云端融合、区块链等分布式系统领域，研究内容包括云操作系统与容器虚拟化、数据中心资源管理与绿色计算、软件定义网络与网络虚拟化、边缘计算与新型云端融合架构、区块链存储与网络等，同时在工程计算、视频处理、制造服务、智慧交通与智能家电等应用领域开展了一系列应用。在云计算核心支撑技术——虚拟化方面的研究成果获得 2020 年国家自然科学二等奖和 2018 年教育部自然科学一等奖。近年来，在轻量级虚拟化，尤其是容器技术方面研发了一系列核心技术，成果应用于百万级容器实例的阿里容器云服务、国产深度操作系统、航天二院仿真云平台等；针对跨域分布式多云数据中心资源管理的可扩展性瓶颈，与华为合作研发了多云级联技术开源

模块Tricircle,进入了全球最大的云计算开源生态标准OpenStack官方发行版；在区块链存储与网络优化方面，提出轻量化弹性存储机制Jidar及高效BlockP2P协议，显著减少数据访问延迟与存储开销，提升区块链可扩展性，并系统地阐述了跨链交互的概念和关键挑战，探索并提出了同构/异构链间交互架构；在智慧交通和智能家电等应用领域取得的研究开发成果已应用于海尔智能家庭中央空调和西门子智能交通决策支持系统上。上述研究成果发表于IEEE/ACM Trans.、JSAC、Proceedings of the IEEE、ISCA、ATC、HPDC、WWW、INFOCOM、ICDE、SoCC、ICDCS、IPDPS、MSST、ICNP等重要学术期刊和会议上。

团队成员：

金　海	团队负责人 教授，主要研究领域为大数据处理，并行分布式计算、云计算		
吴　松	团队负责人 教授，主要研究领域为并行分布式计算，云计算与虚拟化		
余　辰	教授,主要研究领域为边缘智能、工业互联网	何　强	教授，主要研究领域为边缘智能
肖　江	教授,主要研究领域为区块链、分布式系统	王多强	教授,主要研究领域为并行分布式计算、高性能计算
姚德中	教授，主要研究领域为机器学习系统优化、模型压缩与计算加速	顾　琳	副教授，主要研究领域为边缘智能和算力网性能优化
王　雄	副教授,主要研究领域为机器学习系统、分布式学习、数据中心网络	张晓今	讲师,主要研究领域为可信机器学习、大模型和理论计算机
杜冰倩	讲师,主要研究领域为分布式深度学习系统与算法优化	戴小海	博士后,主要研究领域为区块链、分布式系统
樊　浩	博士后,主要研究领域为云原生核心系统软件	黄　卓	博士后,主要研究领域为轻量级虚拟化技术
罗瑞坤	博士后,主要研究领域为边缘计算，边缘智能	余庚花	博士后,主要研究领域为边缘智能、算力网络

团队联系方式：

联系邮箱：yuchen@hust.edu.cn，余辰老师

地址：华中科技大学东五楼222房间。

项目 1：面向大语言模型的专业领域知识注入与推理

一、项目背景

快速辅助决策在许多特定领域的应急处置中得到了广泛应用，如火灾、地震等。然而，传统的决策方法主要依赖规则知识图谱推理，这虽然能够在一定程度上保证决策的逻辑性和可解释性，但也面临着一些显著问题，包括误差传播、推理复杂度过高以及知识更新不便等。特别是在快节奏的应急处置过程中，这些问题显得尤为突出，严重影响了决策的效率和准确性。

随着人工智能技术的发展，尤其是大模型的出现，决策和处置领域迎来了全新的发展契机。特别是在民用领域，如医学和法律等，大模型通过知识注入的方式，能够从大量实例数据（如病例、判决书等）中学习，并展现出处理复杂推理任务的潜力。与传统方法不同，这些大模型不再依赖显式的规则图谱，而是通过理解数据中隐含的规则，为决策和处置提供更加合理的方案。另外，在某些特定场景下，尽管规则类知识体系较为完备，但实例数据的稀缺性仍然是一个关键问题。往往难以获得大量的数据来覆盖所有可能的情况，这导致了在模型训练时的困难，并限制了模型在处理未知或新颖场景时的性能，难以有效泛化。我们希望可以研究出新方法让大模型在这两种情况下都能快速适应新领域任务，该构想分为两个方面：①在有规则集而缺少实例信息时，使用检索增强生成方法并优化；②在实例信息充足时，使用参数高效微调方法并优化。

在有规则集而缺少实例信息时，使用检索增强生成方法并优化。当缺少实例数据时，可以使用检索增强生成方法，将规则知识与现有数据集结合，从而生成合适的输出。该方法的关键步骤包括：首先，通过构建一个规则知识库，系统地存储领域规则，并使用检索模型（如基于向量空间模型的检索器）从中提取相关知识。接着，将这些检索到的规则知识与输入数据结合，作为条件输入传递给生成模型（如 GPT 系列）进行推理。通过优化生成模型的检索机制，使其能够更有效地结合外部规则数据，并生成合理的输出。最后，使用强化学习等方法对生成过程进行优化，以提高规则与生成内容的契合度和生成结果的质量。

在实例信息充足时，使用参数高效微调方法并优化。在实例数据充足的情况下，参数高效微调方法的核心思想是只针对预训练大模型的部分参数进行微调，从而避免完全训练整个模型并节省计算资源。关键步骤包括：首先，选择模型中

与任务最相关的部分（如某些特定层或注意力头），对其进行微调，以适应新的实例数据。在微调过程中，采用低秩适应（LoRA）等方法，仅更新少量的模型参数，同时保持原模型大部分参数不变。其次，使用梯度累积或分层学习率调整等技术，以提高训练效率并确保模型在有限数据下的泛化能力。最后，通过实验验证微调后的模型性能，确保模型能够高效地学习到新任务的特征，同时最大化计算资源的利用效率。

因此，当实例数据充足时，应该如何处理大模型才能使其快速适应新领域任务；当实例数据不足时，如何弥补规则知识与实例数据之间的鸿沟，成为了这一领域亟需解决的核心问题。本研究旨在探索基于大语言模型的知识注入技术，通过将规则知识与实例数据相结合，提升决策的准确性和效率。通过实例推理结果与规则的映射，进一步增强推理结果的可解释性，使得决策过程更加透明和可靠，最终实现更加高效、精确的应急决策。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

基于 Centos7、PyTorch 等深度学习及大模型库进行编程

3．技术基础

本项目基于深度学习、强化学习、大语言模型、参数高效微调、检索增强生成相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．大模型数据集构建：针对特定任务领域，收集和整理高质量的标注数据集，确保数据集覆盖任务需求的多样性与复杂性，为模型微调和优化提供坚实的数据基础。

2．大模型环境搭建及网络结构更新：设计并搭建适用于大规模计算的软件环境，优化计算资源配置，更新深度学习网络结构以适应具体任务，确保训练和推理的高效性。

3．构建高效的规则知识库：通过对特定领域规则知识的系统化整理和编码，设计一个高效的检索知识库，能够动态地与生成模型结合，并在缺少大量实例数据时提供决策支持。

4．检索与生成的优化算法：提出和实现新的检索增强生成算法，结合深度学习中的生成模型（如 GPT）和信息检索技术，改进检索机制，提高生成结果的质量与效率。

5．规则知识与生成模型的融合框架：设计一个框架或方法论，允许不同的规则集在生成模型的推理过程中被有效利用，确保生成的内容符合领域规则和约束。

6．低秩适应（LoRA）或其他微调技术优化：优化低秩适应（LoRA）等技术，使得在有限的实例数据下，能够高效地微调大模型并保持较高的性能，减小计算和存储开销。

7．参数共享与增量微调框架： 设计基于增量学习的微调框架，使得模型能够在不断增加新的任务数据时高效更新，避免重新训练整个模型，提升训练效率。

项目 2：迈向通用具身智能

一、项目背景

人工智能的持续发展，使得各路学者开始思考人工智能如何才能像人一样感知、认知、学习和决策。传统的人工智能乃至各种大模型，仍旧通过。赋人工智能以身体，以人工智能技术为基础，让它与物理世界产生真实的交互，从而理解环境并自主决策，这被认为是人工智能走向通用人工智能的必经之路——具身智能。

具身智能关注与通过物理世界学习，让机器具有自主认知和自主行为的能力，本研究小组将目光集中在具身智能的落地实现上，我们选用机器狗和轮式机器人作为物理基础，以将具身智能落地实践为目标，计划构建具身智能的自主感知集和自主任务集。我们希望让机器狗/轮式 AI 机器人像人一样拥有自主行为决策。

该构想分为三个阶段：

程序规定的基础行为。完成轮式机器人/机器狗的基础定位和导航算法。利用激光雷达和摄像头完成即时定位与导航。此过程需依赖于先验地图/自建地图，因此衍生出基于雷达和相机的建图任务。可以考虑流行的 ROS 系统建图算法进行地图构建、传统的 AMCL 等算法完成定位、A*等其他路径规划算法完成导航规划。在定位与导航基础上，我们可以在理想环境中，为机器狗/机器人分配指定任务。

半自主任务。上述基础行为能力依赖于较理想的实际环境，但在现实物理世界中，环境极为复杂，因此需要考虑各种算法的动态适应能力，在传统的 AMCL 算法中融入各种传感器的数据（如相机）对定位过程进行优化、定位结果进行修正，或利用视觉感知进行定位；在定位导航算法中，需要额外考虑不同的路况、障碍物情况、社会规则等等因素，确保移动过程的通过性（定位导航过程可以类比当前自动驾驶）；半自主任务比如取快递、送快递等等，需要再基础行为能力的基础上，丰富任务逻辑，利用各种识别算法完成后继行为，可能涉及的技术有图像识别、目标检测、基于强化学习和视觉定位机械关节控制等等。据此，我们可以让机器狗在实际物理世界中，完成类似取送物品、搜救等任务。

基于多源感知的全自主决策与交互。为了让机器狗/机器人的感知和行为更像人类，我们考虑附加更强的感知、交互和决策能力，利用具身智能所搭载的各种传感数据（音频、视频、雷达点云、压力值等）完成对环境的自主探知、交互、

行为决策，从而与人类对话、理解三维环境、训练出更灵活和完备的决策方式。该过程中，我们需要使用自然语言处理（NLP）、人机交互、视觉导航、综合各种识别结果的强化学习决策、地图维护修正等技术/方法。到此，我们构建出一个全生命周期内可以基本完成自己决策和行为的具身智能体。

二、项目应用平台与基础

1．硬件平台

本项目基于较为成熟的硬件平台，以宇树机器狗 go2 EDU、蔚蓝 alpha 机器狗、大疆 AI 机器人为主；另外会使用到激光雷达、三维雷达、工业相机、深度相机、各种算力计算平台。

2．软件平台

本项目基于 Ubuntu、 ROS、 PyTorch 等平台进行编程。

3．技术基础

本项目基于深度学习、强化学习相关理论与技术，在项目研发过程中需要使用 Python、C++等编程语言。

三、项目要求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python、ROS 系统基础，了解机器学习相关内容；

3．具有一定的 PyTorch、深度学习、强化学习基础

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．机器狗激光雷达定位实现

2．机器狗传统导航算法实现

3．机器狗三维雷达和视觉实时建图

4．定位、导航、建图等算法优化

5．机器狗 gazebo 模拟环境建立

6．机器狗基于视觉和雷达数据的强化学习导航算法实现

7．2V2 轮式 AI 机器人对抗系统——策略部分实现

8．机器狗与 AI 机器人协同“行人伴随”

9．机器狗/AI 机器人在密集移动障碍环境中的动态路轨规划

10．机器狗/AI 机器人基于视觉的复杂道路情况自适应导航（探究细粒度行为规划）

11．机器狗/AI 机器人上语音识别与交互模块实现

12．机器狗根据语言交互的任务创建和规划

13．多任务强化学习模型

7. 大模型与智能系统团队

计算机学院“大模型与智能系统团队”依托于计算机系统结构国家重点学科和计算机软件与理论湖北省重点学科，是大数据技术与系统国家地方联合工程研究中心、服务计算技术与系统教育部重点实验室、集群与网格计算湖北省重点实验室的重要组成部分。建有华中科技大学-青岛泰屹投资发展有限公司“异构计算技术联合研究中心”。是科技部重点领域创新团队、教育部“长江学者和创新团队发展计划”创新团队的重要组成部分。

团队现有教授/研究员 2 人，副教授 2 人。承担了一项湖北省重大科技攻关项目、两项国家重点研发计划课题和多项国家自然科学基金项目，并与华为、浪潮、曙光、中移动等一大批企业开展了广泛合作。

目前研究领域主要包括下列四个方面：

大模型计算框架与平台支撑：针对大模型在计算平台的“卡脖子”问题，开展面向国产 AI 芯片的大模型计算框架与平台支撑系统研发，服务构建自主可控的大模型基础设施。

多模态大模型及人机交互研究：面向多源异构大数据，研究以语言为中心的多模态大模型，实现自然语言与代码、表格、视觉等多模态数据的智能化交互研究。

大模型高效演进技术：针对现有大模型部署后是静态的，知识难以更新，无法及时适应动态变化环境的问题，进行基于增量数据精化、参数高效微调、检索增强生成等一系列大模型演进技术的研究。

机密计算软硬协同加速：面向全同态加密、零知识证明和安全多方计算等新兴密码学系统，研究高效算法优化与硬件加速技术在数据治理中的应用，旨在提升国家数据安全与隐私保护能力，服务国家数据治理体系和治理能力现代化战略。

团队研究领域实现了大模型和智能系统从底层平台支撑、中间层模型构建、上层落地应用全栈领域核心技术的全面覆盖，并广泛涉及大模型加速、异构计算等诸多相关领域。相关的研究成果已发表在 IEEE/ACM Trans.等期刊和 ICML、NeurIPS、ICLR、AAAI、IJCAI、KDD、SIGMOD、ACL、ICSE、FSE、ASE、ESORICS、SOSP、HPCA、ASPLOS 等重要国际学术会议上，实现了从人工智能、算法、软件工程、体系结构领域顶级会议的全覆盖。相关工作获 CCF 自然科学一等奖（排名

第一)、湖北省自然科学优秀学术论文一等奖、中国人工智能学会吴文俊人工智能科学技术奖优秀博士学位论文提名、IEEE BigComp 最佳论文等。

团队研究成果应用于字节头条、神威太湖之光高性能计算机等大型互联网企业与国家大型服务平台。研发的内存计算系统 Mammoth 被 IEEE TKDE 大数据调研报告作为全球 5 个代表性内存计算系统之一推荐。成果被 IEEE Computer 亮点推荐、入选 IEEE Spectrum 的 IEEE Journal Watch。团队本科生还多次在 ICML、NeurIPS、ICLR 等人工智能顶级会议上发表论文。

团队毕业生质量获得企业一致认可，在业界具有良好口碑。毕业生多入职华为、腾讯、阿里巴巴、字节跳动、百度、微软、亚马逊、IBM 等一流 IT 企业。

团队成员：

金　海	团队负责人 教授，主要研究领域为大数据处理，并行分布式计算、云计算		
石宣化	团队负责人 教授，主要研究领域为智能计算，异构计算，大数据处理		
华强胜	研究员,主要研究领域为并行分布式计算理论、算法及应用	张　腾	副教授,主要研究领域为机器学习与人工智能
万　瑶	副教授，主要研究领域为多模态大模型、代码智能、表格智能		

团队联系方式：

联系邮箱：tengzhang@hust.edu.cn，张腾老师

地址：华中科技大学东五楼 220 房间。

项目 1：面向国产平台的大模型支撑系统

一、项目背景

AI 大模型的国产化需要面向国产体系结构的大模型支撑系统。现有的大模型支撑系统存在内存资源利用率低、模型规模难以扩展、并行效率低等问题，亟待面向国产 AI 芯片研发高效的大模型计算框架与平台。

针对缓解大模型对内存和算力的需求，一些研究团队已经进行了初步的探索。例如微软发布的针对大模型训练的优化库 DeepSpeed，旨在优化数据并行内存占用，提升训练效率。英伟达发布的 Megatron 框架，利用张量级别的模型并行进行分布式训练提升系统扩展能力。华为的 MindSpore 智能计算框架通过对后端计算图的编译优化实现高效计算。然而，由于国产 AI 芯片并行层次相对复杂，现有框架在大规模分布式训练方面存在自动并行切分耗时，并行效率低的问题，同时自动并行也缺乏与计算图编译优化等关键性能优化手段的协同，最终导致大模型训练效率受限。同时，对于通用的多级异构内存管理框架，例如 BaM，GMT 等工作主要面向大数据应用，不适用于大模型训练存算系统。DeepSpeed 框架中的 ZeRO-infinity，利用 SSD 的存储容量来实现大规模模型训练。但缺乏对多级异构内存的统一管理，内存卸载以模型层为粒度，数据迁移调度不够灵活，产生了较大 I/O 开销。

针对上述问题，需要构建一套专门面向国产平台的大模型支撑系统。该系统通过面向国产平台的多级模型划分方法，自适应不同模型、不同硬件网络层级结构。解决大模型自动并行切分方法耗时长、并行效率低、缺乏计算图编译协同优化的问题。同时，支持基于国产 AI 加速卡的远程存储直通技术，融合了高带宽内存、DRAM、非易失性存储、固态硬盘等多级存储，构建统一的多级存储层次结构。实现单卡支持训练模型参数最高达到千亿，多卡并行训练规模最高超过万亿，降低数据 I/O 时间，提高训练效率。

具体来说，该研究包含以下子目标：1）设计面向国产平台上大模型训练的设备内存高效池化机制。2）实现面向复杂大模型训练的自适应并行切分方法。3）实现计算图编译和自动并行的协同优化。4）结合内存层次结构和训练数据特征，设计面向大模型训练的数据布局技术，实现训练数据在不同内存层次的合理放置，提高数据访问效率和内存资源利用率。5）针对具有不同生命周期的张量数据，

设计训练过程张量数据在各级内存间的高效迁移策略，实现各级内存资源的充分利用，并使用计算掩盖数据传输开销。6）基于大模型训练数据访问特性，设计训练数据在高访问带宽内存中的缓存机制和数据访问重定向机制，降低数据访问开销，提高内存资源利用率与模型训练吞吐量。

二、项目应用平台与基础

1．硬件平台

本项目主要基于国产 AI 服务器构建硬件基础，例如昇腾 910B 等。

2．软件平台

本项目基于 CentOS、PyTorch 等平台进行编程。

3．技术基础

本项目基于大语言模型、大模型并行内存优化、数据建模技术、任务调度与并行处理技术、芯片适配与算子优化技术以及智能内存管理技术相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术、并行内存优化等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．面向国产平台的并行训练性能模型构建：深入探究国产平台的硬件架构细节与大模型训练的具体要求，测试并收集各类模型在该平台上训练时的多方面数据，涵盖计算资源的占用情况、内存的使用动态、数据传输的效率表现等，进而构建起能精确反映模型训练性能与国产平台软硬件环境内在联系的性能模型。

2．面向国产平台的大模型训练关键算子实现和优化：围绕国产 AI 芯片的独特指令集架构和计算资源特点，对大模型训练中的核心算子，如卷积、全连接、注意力机制等进行重新设计与实现。充分利用芯片所具备的特殊计算单元和强大

的并行处理能力，开发出高效的算子算法。在这个过程中，从多个角度对算子进行优化，包括优化计算逻辑以减少不必要的计算步骤，改进数据访问模式以提高数据读取的速度和效率，完善内存管理策略以降低内存使用的开销。

3．面向国产平台存储结构的内存调度机制研发：深入地研究国产平台复杂多样的存储层次结构，详细了解不同类型内存的容量大小、带宽限制、访问延迟等关键参数。在此基础上开发一套高效的内存调度机制，该机制能够根据大模型训练过程中数据的访问频率、时效性以及计算任务对数据的实时需求，动态地对内存资源进行合理的分配和管理。

4．基于张量生命周期的数据迁移机制：利用计算图上的依赖分析和运行时技术，构建轻量级的张量生命周期监测工具。设计数据迁移调度器，依据张量生命周期信息，利用启发式调度策略在运行时动态选择需要卸载和预取的张量，并确定数据迁移的目的地和时机。研究与数据迁移机制协同的重计算策略，提出数据迁移和重计算的性能评估模型，权衡张量不同决策对全局性能的影响。选择部分张量使用重计算策略来充分利用计算和存储资源，避免潜在的数据访问瓶颈，如存储器争用等。

5．基于访问重定向的数据缓存机制：基于内存重定向的设计思想，设计带有低速存储器-内存映射的内存访问机制。设计基于批量读取的内存填充机制，将低速存储器中的原始训练数据按照数据块的形式重新组织，并利用内存映射机制与内存建立起联系。针对分布式训练场景设计内存填充的分布式交互机制。并通过网络请求的方式实现各节点内存中数据的相互共享，以满足各节点的数据需求。

项目 2：基于高效参数微调的大语言模型混合专家系统构建

一、项目背景

近年来，随着人工智能领域的快速发展，大语言模型（Large Language Model，LLM）凭借其强大的推理和生成能力，逐渐在各种应用场景中得到广泛应用。例如在智能客服场景中，它能精准理解客户咨询，迅速给出专业解答，实现不间断服务，极大提升客户满意度，降低企业人力成本；在内容创作方面，辅助创作者构思文章框架、生成初稿，为写小说、新闻报道、学术论文等提供灵感源泉，让创作过程更加高效流畅。然而，随着应用场景的复杂程度不断提高，对于许多更加困难，需要更多领域知识共同支撑的情形，传统基于 Transformer 架构堆叠的模型难以给出良好的推理结果，且模型在演进更新和推理生成时效率低下。导致这些问题的原因一方面在于堆叠结构的大模型在训练过程中，虽然通过大量数据学习了不同类型和领域的知识，但是不同类型的知识会互相干扰。另一方面，堆叠结构模型在演进更新和推理生成时，需要激活几乎所有的参数，必然导致其效率低下。为了解决以上问题，我们想研究出新方法，使得大模型可以一方面实现高效的演进更新，另一方面可以使得大模型学习不同类型的知识，减少模型学习过程中知识互相干扰的影响。该构想分为两个方面：1）设计更高效的参数微调方法，实现 LLM 的高效演进更新；2）设计混合专家网络（Mixture-of-Experts，MoE），在推理过程中，只激活少量模型参数，基于不同的专家网络整合知识进行高效推理。

由于 LLM 的复杂结构和巨大参数规模，预训练所需的人力和资金成本巨大，因此，在算力不足的情况下，为了使 LLM 适应下游任务，需要采用参数高效微调技术。其关键步骤包括：首先，对模型不同的网络结构进行分析，对重要的网络结构层进行微调，以使其适应新数据。其次，在微调中，常用的低秩适应（LoRA）等方法，结构过于简单，导致微调模型的学习能力有限，因此需要对该结构进行优化，通过引入非线性映射或者加入其他网络结构的方式，提高微调模型的学习能力，以确保微调后的模型能准确学习新数据。最后，通过实验测试模型在微调后的性能，以及模型微调所需的参数规模，确保其有效性和高效性。

为了使得模型在更新演进和推理生成方面得到进一步提高，LLM 不再追求纵向堆叠更多的 Transformer 层，转而横向构建不同的小规模模型，通过路由网络

实现模型训练和任务分发，极大减少了推理过程中所需激活的参数规模。要构建更高效的 MoE 系统，其关键步骤包括：首先，对 MoE 的结构进行分析，不同网络层级的信息需要设计不同的融合方式。其次，对于 MoE 的路由网络，在微调阶段容易导致部分专家模型训练不足，因此需要设计相应的负载均衡损失函数，使每个专家网络得到充分训练。最后，通过设计辅助损失函数，能更准确地激活任务相关的 LoRA 微调专家，并整合专家输出进行推理生成。

本研究旨在探索基于高效参数微调的大语言模型混合专家系统构建技术，通过设计更高效的 LLM 高效参数微调方法，并结合混合专家系统，提升 LLM 的演进更新效率和处理复杂任务的能力。通过高效更新小规模的专家模型，并整合少数专家模型的知识，提高模型的更新效率和推理能力。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

基于 Centos7、PyTorch 等深度学习及大模型库进行编程

3．技术基础

本项目基于深度学习、大语言模型、参数高效微调、混合专家系统相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．大模型数据集构建：针对特定任务领域，收集和整理高质量的标注数据集，确保数据集覆盖任务需求的多样性与复杂性，为模型微调和优化提供坚实的数据基础。

2．大模型环境搭建及网络结构更新：设计并搭建适用于大规模计算的软件环境，优化计算资源配置，更新深度学习网络结构以适应具体任务，确保训练和推理的高效性。

3．大模型高效参数微调方法：通过对大模型各网络层级的功能分析，设计不同的高效参数微调结构，并引入不同类型的非线性映射，以进一步提升微调模型的学习能力。

4．MoE 负载均衡损失设计：提出和实现新的专家网络负载均衡损失函数，通过对网络中专家模型接受的 token 类型和数量进行分析，使得各专家网络可以获得充足的训练。

5．MoE 融合框架：设计一个框架或方法论，使得专家模型可以对不同类型任务进行准确响应，将任务相关的专家输出进行合理整合用于推理生成。

8．人机物系统与安全团队

计算机学院“人机物系统与安全团队”依托于华中科技大学计算机系统结构国家重点学科和计算机软件与理论湖北省重点学科，由加拿大国家工程院院士，加拿大工程研究院院士，欧洲科学院外籍院士，联合国科学院院士，华中科技大学国家千人计划特聘教授杨天若教授2010年牵头成立。团队目前主要的研究领域包括：**嵌入式系统、物联网、大数据安全、普适计算与移动计算、存储系统、高性能计算、云计算、区块链**等方向。团队现有教授1人，副教授2人，海外“客座教授”8人，博士、硕士研究生40余人。团队近年来承担了国家海外高层次人才引进计划、重点研发项目等多项研究课题，在CPSS理论与应用等领域积累了丰富的实践项目研发经验。

团队的研究成果及获奖情况包括：

杨天若教授2017年成功当选加拿大国家工程院院士，2018年成功当选加拿大工程研究院院士，荣获2017年IEEE TCSC 可扩展计算杰出成就奖，2018年IEEE TCPS信息-物理系统杰出领袖奖，2018年IEEE SCSTC智慧计算终身成就奖，2019 AMiner 物联网最有影响力学者奖，2021年当选欧洲科学院外籍院士、联合国科学院院士。

2021年和2023年连续两届”挑战杯”获三项大奖：《抗量子攻击的密钥分发同步系统的设计与实现》获第十七届揭榜挂帅赛道一等奖，《后量子安全的电子邮件系统》获第十七届揭榜挂帅赛道特等奖，《基于isQ语言的大整数因数分解》获第十八届揭榜挂帅赛道三等奖。

近5年有5篇论文获得最佳论文奖。

7名博士分别获得2017，2018，2019,2020,2022,2023,2024年度IEEE TCSC最佳博士论文奖(每年全球只有5个名额)。。

团队毕业生质量获得企业一致认可，在业界具有良好口碑。每年为华为、腾讯、阿里巴巴、百度、微软等一流IT企业输送大量人才，也有相当多的优秀毕业生在国内大学院校从事科研及教学工作。

团队成员：

杨天若	团队负责人 教授，主要研究领域为人工智能、大数据、区块链、高性能计算、嵌入和普适计算、人机物系统设计。

王　蔚	副教授，主要研究领域为密码学，大数据安全，隐私保护。
崔金华	副教授，主要研究领域为嵌入式系统，存储系统，大数据。

团队联系方式：

联系邮箱：20146125548@hust.edu.cn，王蔚老师

地址：华中科技大学南一楼东南 404 房间。

项目 1：针对 GAN 推理攻击的联邦学习防御

一、项目背景

随着互联网技术的迅猛发展，以深度学习为代表的全新一代技术正改变我们处理信息的方式。深度学习技术是基于海量的大数据之上，对数据进行处理、加工，并产生更加有价值的计算结果。然而深度学习的快速发展也伴随着一些问题。深度学习要求大量的训练数据，且所需的计算开销较高。同时，在分布式场景下运行深度学习，用户需要将自己的数据提交给中央服务器，这可能会导致用户数据遭到盗用、滥用，进而泄露用户隐私。

联邦学习的诞生正是为了解决深度学习技术发展中产生的这些问题。联邦学习系统一般由一个中央服务器和多个参与者客户端组成。服务端和参与者协商一个深度学习任务目标，在联邦训练中需要迭代训练并交换参数，来共同训练完成一个全局模型。参与者在训练过程中，训练数据全程留在本地，并未向服务器上传个人敏感数据，仅仅将训练所产生的梯度信息上传服务器。这样的分布式深度学习训练，相比传统的中心化机器学习系统能够更好地保护用户隐私。除此以外，这还能进一步提高机器学习的精度，因为可以将边缘设备的数据利用起来，解决了机器学习训练数据碎片化的问题。这样，联邦学习系统能将处于“数据孤岛”的设备利用起来，在保障他们的隐私属性的前提下，将这些设备所具有的本地数据的价值发挥，来完成更多的机器学习任务以及人工智能程序。

然而，联邦学习也很容易受到各种攻击。这里我们关注一种基于生成对抗网络（Generative Adversarial Network，GAN）打造的攻击方式，在训练过程中攻击者选定一个特定的分类标签发起进攻，利用模型参数更新不断训练 GAN 模型，并生成该标签对应的训练数据，同时将生成的训练数据错误的标记成其他标签，进一步诱导受害用户释放更多的参数信息来加强其攻击效果。这种攻击方式由于其进攻难度低，窃取效果好、隐蔽性高，对联邦学习造成了极大的安全风险及隐私泄露危害。我们希望研究一种抵御这种基于 GAN 的推理攻击的联邦学习防御方法，使得攻击者在伪装成善意训练用户时无法获得其他用户的隐私数据。具体地，我们构想：①采用假标签注入的方式混淆 GAN 网络中判别网络的输出层，增大敌手从输出层发起进攻窃取用户数据隐私的难度；②在①的基础上，利用 Cycle-gan 所生成的混淆数据去训练假标签，使敌手在进攻到假标签时仅生成具

有语义性但不泄露用户隐私的数据。

对于假标签注入，服务器首先收集所有用户的标签信息，并生成大量随机的假标签加入其中，并依照注入后的标签集设计神经网络模型结构。此时所有参与者获得的模型都是注入后的，而对于参与者而言，仅了解自身持有的标签，对于其他标签，并不清楚标签与分类的对应关系。之后，服务器与用户可以根据原始联邦学习方式进行模型训练。而敌手由于无法确定标签与分类的对应关系，无法针对某一分类进行数据推理。

对于注入的假标签，敌手仍然能够一定程度进行区分，因此，服务器基于 Cycle-gan 生成假标签，而不是随机生成。服务器基于 Cycle-gan 生成一个混淆数据集，该数据集不包含用户的隐私数据。服务器基于该混淆数据集生成假标签并在之后将混淆集用于假标签的训练。这样可以欺骗敌手还原出来混淆集所对应的图片，还能让敌手误以为自己进攻成功。

针对联邦学习的推理攻击使得联邦学习下用户的隐私数据不再安全，因此，本研究旨在探索能抵御基于 GAN 的推理攻击的联邦学习方法。通过服务器生成并注入假标签，可以迷惑攻击者，使其难以攻击到真实的标签所对应的数据，同时该注入行为不影响正常的学习模型训练。在此基础上，服务器通过 Cycle-gan 生成混淆数据集，并依此生成假标签训练，进一步迷惑攻击者。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

基于 PyTorch 等进行编程

3．技术基础

本项目基于深度学习、联邦学习、生成对抗网络相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 基础；

3．具有一定的 PyTorch 基础，了解联邦学习、生成对抗网络等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．联邦学习环境搭建以及模型训练：设计并搭建联邦学习环境，选取各类数据集进行联邦学习模型训练，以便后续攻击与防御实验开展。

2．基于 GAN 的推理攻击：设计并搭建针对联邦学习的推理攻击，模拟攻击者对系统的攻击，以便后续研究与实验开展。

3．假标签注入的联邦学习环境构建：设计并搭建联邦学习模型，其中标签集中包含大量假标签，在此模型下进行联邦学习训练。最终输出不包含假标签的完成训练的模型。

4．基于 Cycle-gan 的混淆数据集构建：设计基于 Cycle-gan 的混淆数据集生成方案，并基与此生成假标签。

5．基于混淆数据集的假标签生成与训练：设计并搭建联邦学习模型，其中服务器生成大量假标签并利用混淆数据集进行训练。最终输出不包含假标签的完成训练的模型。

项目 2：基于同态加密的 K 均值聚类算法实现

一、项目背景

在当今机器学习领域，数据的重要性不言而喻，作为训练模型、做出预测和决策的基石，数据的使用和共享也日益频繁。然而，这种数据交流的增加同时也带来了隐私泄露的风险，尤其是对于包含敏感信息的数据。当敏感信息被泄露时，个人隐私可能受到侵犯，数据可能被滥用，这种情况不仅对用户和组织的信任构成威胁，也可能带来严重的法律后果。

在这一背景下，密态计算技术成为解决隐私泄露问题的重要前沿。密态计算结合了先进的加密技术和隐私保护协议，旨在在计算过程中维护数据的保密性。这种技术的广泛应用领域涵盖医疗保健、金融以及人工智能等领域。在机器学习领域，密态计算技术的运用可以有效保护数据隐私，确保在数据处理和分析过程中不会泄露敏感信息，从而增强数据安全性和隐私保护。

在实现机器学习算法保护的过程中，同态加密技术发挥着至关重要的作用。同态加密允许在加密数据上进行计算，得到的结果仍然是加密的，无需进行解密操作。这一特性使得在保护数据隐私的同时进行各种数据分析操作成为可能，例如聚类、简单统计分析等。同态加密技术的引入为 K 均值聚类等机器学习任务的隐私保护提供了强有力的支持，使数据所有者能够安全地利用数据进行模型训练和推理，避免数据泄露和隐私侵犯的风险。这种技术结合为机器学习领域提供了一种创新的隐私保护解决方案，有助于构建更安全和可靠的数据分析环境。

K 均值聚类（K-means Clustering）是一种无监督学习算法，主要用于数据的聚类分析。其核心思想是将数据集划分为 K 个簇，使得簇内的数据点尽可能相似，而簇间的数据点尽可能不同。相似性通常通过计算数据点之间的距离来衡量，如欧氏距离。K 均值聚类的方法原理包含：①初始化：随机选择 K 个数据点作为初始聚类中心，这些可以是数据集中实际存在的点，也可以是随机生成的点；②分配：将每个数据点分配给最近的聚类中心，形成 K 个簇；③更新：重新计算每个簇的中心，通常取簇内所有点的平均值；④迭代：重复步骤②和③，直到聚类中心不再发生显著变化或达到预设的迭代次数。

K 均值聚类在多个领域有广泛应用，包括：①图像处理：在图像分割中，K 均值聚类可以将图像划分为多个具有相似特征的区域，通过将图像的像素作为数

据点，基于颜色、纹理或空间位置等特征进行聚类，从而实现图像的分割；②市场细分：在商业中，K 均值聚类可以将客户根据其购买行为、年龄、收入等特征划分为不同的组，从而制定更加个性化的营销策略；③降维和矢量量化：K 均值聚类可以将高维特征压缩到一列当中，常用于图像、声音和视频等非结构化数据的处理。

因此，本项目旨在利用同态加密技术实现 K 均值聚类算法，即使用同态加密将明文空间的 K 均值聚类算法转换成密文空间的 K 均值聚类算法，以确保在保护上述应用的数据隐私的同时进行聚类分析。由于同态加密技术只支持加法、乘法、标量乘积、矩阵乘法以及近似除法，还需要额外实现一个同态排序功能支持分支（例如，if 语句）。另外，可以使用同态加密技术做 K 均值聚类算法的应用，例如，K 均值聚类算法训练好之后，当用户新来了一个数据，服务器可以用同态加密进行相似度计算判断这个数据属于哪个簇。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

基于 python 库（如 Pyfhel，sklearn）进行编程

3．技术基础

本项目基于同态加密、k 均值聚类算法相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 基础；

3．了解 k 均值聚类算法相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．研究同态加密：深入研究同态加密技术，了解如何在加密数据上执行聚类算法，并选择适合的同态加密库或工具。

2．数据集准备：准备用于聚类分析及应用的数据集，并将其转换为适合同态加密处理的格式。

3．算法开发：实现 K 均值聚类算法的同态加密版本，确保在加密数据上正确执行聚类操作。

4．相似度计算算法开发：实现相似度计算算法的同态加密版本，确保在加密数据上正确执行相似度计算操作。

5．训练模型应用：将训练好的同态加密版 K 均值聚类算法结合同态加密版相似度计算算法用于新的数据集分类应用。

9. 网络认知计算团队

计算机学院“网络认知计算团队”（NCCL）建有湖北省智能边缘计算专业型研究所和电网信息通信安全与智能技术联合实验室。团队现有教授 1 人，讲师 1 人，博硕士研究生 30 余人。承担了国家重点研发计划项目/课题、国家重大专项项目/任务、自科基金和国际合作项目 10 余项。与腾讯、华为、西门子、大疆、移动、电信、国网等龙头企业开展了广泛合作。

目前研究领域主要包括两个方面：

算网融合：面向大模型训练、远程控制网络（含工业互联网、触觉互联网和远程驾驶网络等）、空天地一体网络和命名标识网络等，研究基于人工智能和大模型的服务质量保障、网络安全攻击防御和智能运营维护管理决策等关键技术，服务未来多样化应用场景和异构融合承载网络。

认知计算：面向异构多源多模态数据，研究图建模与图计算理论，研究垂直行业大模型的参与调优策略和剪枝优化策略，研究思维链决策推演策略和方法，开发多智能体协同系统和行业应用。

团队相关研究成果形成各类标准 10 余项，申请并授权发明专利 40 余项，在 IEEE/ACM Trans.等期刊和 ACL、EMNLP、MM、DySPAN 和 ICC 等重要国际学术会议上发表高水平论文 20 余篇。学生依托相关成果获得国家级竞赛一等奖和二等奖多项。研究成果在能源、教育和建造等行业得到应用。

团队依据学生特点进行差异化培养，鼓励学生多样化就业，得到用人单位好评。学生去向不仅包括腾讯、字节、阿里和百度等头部互联网企业，还包括金融量化机构，以及国央企、高校研究所和政府选调生。

团队成员：

莫益军	团队负责人 教授，主要研究领域为边缘计算、图计算、认知计算、算网融合
刘辉宇	讲师，主要研究领域为自然语言处理、大数据、网络安全

团队联系方式：

联系邮箱：moyj@hust.edu.cn，莫益军老师

地址：华中科技大学南一楼中 601 房间。

项目 1：基于大模型智能体的思维链构造与推理增强

一、项目背景

随着人工智能技术的不断进步，大模型在自然语言处理、计算机视觉等领域取得了显著的成就。然而，传统的模型推理能力主要依赖于数据驱动和模式匹配，缺乏深层次的逻辑推理和思维能力。思维链（Chain of Thought）作为一种新兴的推理方法，通过模拟人类的思维过程，使大模型能够进行更加复杂和抽象的推理任务。思维链的核心在于将问题分解为一系列逻辑步骤，逐步构建推理链条，从而实现从已知信息到未知结论的推导。

在实际应用中，基于思维链的大模型推理能力具有广泛的应用前景。例如，在智能问答系统中，大模型可以通过思维链对用户的问题进行深入分析，理解问题背后的逻辑关系和隐含条件，从而提供更加准确和全面的答案。在知识图谱构建中，大模型可以利用思维链对知识进行推理和扩展，发现新的知识关联和潜在规律。此外，在科学研究、金融分析、法律推理等领域，基于思维链的大模型推理能力也展现出巨大的潜力，能够帮助人类解决更加复杂和棘手的问题。

然而，基于思维链的大模型推理能力在发展中也面临诸多挑战。首先，在大模型微调层面，构建高质量的思维链数据集是一个难题，需要大量的标注数据和专业知识。其次，在模型设计层面，如何有效地将思维链融入大模型的架构中，使其具备良好的推理能力和可解释性，仍需进一步探索。此外，在推理效率方面，思维链推理过程可能较为复杂和耗时，如何提高推理速度和效率，满足实际应用的需求，也是一个亟待解决的问题。

二、项目应用平台与基础

1．硬件平台

GPU 算力集群和多个大模型平台 Token

2．软件平台

基于 Langchain、OpenPrompt、DeepSeek-V3、DeepSpeed 等框架研究思维链构造与推理

3．技术基础

本项目基于思维链、多步推理、强化微调和基准等相关理论与技术。

三、项目需求

满足以下要求：

1．自驱力强、思维活跃、有好奇心和探究精神；

2．熟悉 C++和 python 开发语言；

3．了解深度学习、大模型智能体技术等相关理论与技术，了解各种大模型微调策略，熟悉图处理。

四、项目开展

从以下项目中选取一个：

1．强化微调与思维链生成：研究如何通过强化学习对大模型进行微调，以生成更加准确和连贯的思维链，提升模型在复杂推理任务中的表现。

2．智能体思维链：研究如何为智能体设计和实现思维链推理机制，使其能够在多任务环境中自主思考和决策，提高智能体的适应性和灵活性。

3．信息检索与思维链重构：探索利用信息检索技术从大规模数据中提取相关信息，并结合思维链重构方法，构建完整的推理链条，以支持更加全面和深入的推理过程。

4．思维链搜索优化与迭代收敛：针对通用推理思维链构造搜索空间过大，及思维链发散和由此产生的幻觉问题，研究搜索优化和迭代收敛加速策略。

5．数据科学思维链：构建面向数据科学的思维链推理，研究如何从数据处理中发现潜在的逻辑关系和规律，构建数据科学领域的思维链，以支持数据自动化分析和决策制定。

项目 2：人机协同智能系统

一、项目背景

人机协同智能系统，指人类与人工智能系统通过紧密合作，共同完成任务并提升整体效率与效果的智能系统。该系统结合人类的创造力、情感理解与复杂决策能力，以及 AI 的计算速度、数据处理和模式识别优势，广泛应用于医疗、制造、金融、教育等领域。随着增强学习、情感计算和多模态交互技术的突破，人机协同将更加智能化、个性化和自然化，

人机协同智能系统面临的主要挑战在于如何设计高效的自然交互界面，实现人类与 AI 的无缝协作，同时克服 AI 在复杂场景中的技术局限性。研究增强学习、边缘计算和多模态交互关键技术，旨在提升无人系统（设备）的适应性、情感理解能力和实时响应效率，建立人类对 AI 决策的信任、确保系统的透明性与可解释性。

二、项目应用平台与基础

1．硬件平台

博创尚和 Aviator 410 人工智能无人机、阿木实验室 P450 人工智能无人机、嘉创飞航 UE-01F 人工智能无人机、六爻飞梦 Magpie 360 人工智能无人机、小米 CyberDog 2 机器狗和启创远景 QC-8KT 组合式机器人等。

2．软件平台

基于 RViz、Gazebo 和各无人机设备厂家提供的基础库进行开发控制

3．技术基础

本项目基于接口控制、ROS、嵌入式操作系统、深度学习、决策优化、大模型和多智能体等相关理论与技术。

三、项目需求

满足以下要求：

1．自驱力强、思维活跃、有好奇心和探究精神；

2．熟悉 C++和 python 开发语言；

3．对机械控制等硬件平台有浓厚兴趣。

四、项目开展

从以下项目中选取一个：

1．无人设备的自主导航与识别控制：针对无人机在复杂环境中和受电磁干扰情况下自主寻找并命中目标的问题，研究自主定位、SLAM 建图、路径规划、避障及穿越、目标识别等技术。

2．无人设备姿态与行进鲁棒控制：针对无人设备受外部环境和外力干扰情况下的平衡和行进问题，研究其鲁棒性控制技术，保证系统（如无人机/狗的翻滚恢复、抗跌倒）的稳定性和性能。

3．无人设备的集群编队与协同控制：研究无人机集群编队与协同控制的编队策略、路径规划、密集通信、任务分配和协同决策等。实现多个无人设备在复杂环境中高效、稳定地协同工作。

4．人机协同交互控制系统：研究包括文字、声音、手势和脑电等各种模式人机协同交互控制技术，建立复杂任务的人机协同分配调度机制。

项目 3：智算网络的感知与认知

一、项目背景

智算网络的感知与认知能力得益于人工智能（如深度学习和强化学习）、5G/6G 通信技术、边缘计算与云计算的快速发展，以及物联网的普及，这些技术为其提供了强大的数据处理和智能决策支持。人工智能技术赋予网络自主学习和推理的能力，使其能够从海量数据中提取有价值的信息；5G/6G 通信技术提供了高速、低延迟的网络环境，为实时感知和响应奠定了基础；边缘计算与云计算的结合则实现了计算资源的分布式协同，提升了数据处理效率和灵活性；而物联网的普及则带来了海量设备和数据的接入，进一步推动了智算网络对感知与认知能力的需求。在应用层面，智算网络在智能交通、智慧城市、工业互联网和智能医疗等领域展现出巨大潜力。

在智能交通中，智算网络能够实时感知交通流量并优化信号控制，缓解拥堵问题；在智慧城市中，通过感知环境数据，智算网络可以优化资源分配和城市管理，提升居民生活质量；在工业互联网中，智算网络能够实现设备状态的实时监控和预测性维护，提高生产效率；在智能医疗中，通过感知患者数据，智算网络可以为个性化医疗方案的制定提供支持。同时，网络架构正从传统的静态配置和集中式管理向动态优化和分布式协同演进。传统网络依赖固定的规则和配置，难以应对复杂多变的环境，而智算网络通过感知与认知能力，能够动态调整网络配置、资源分配和路由策略，实现网络的自我优化和适应。

智算网络在发展中面临多方面的挑战。在数据层面，海量、异构数据的处理对存储和计算能力提出了高要求，同时数据质量与可靠性问题（如噪声、缺失或错误）可能影响感知与认知的准确性。在算法与模型层面，智算网络需要满足实时性要求，在极短时间内完成分析与决策，同时模型需具备良好的泛化能力以适应复杂多变的网络环境，并确保决策过程的可解释性以支持网络管理与故障排查。在网络架构层面，网络高效协同对通信与资源管理提出了挑战，资源受限问题也限制了智能感知与认知的实现，而网络拓扑、流量模式和用户需求的动态变化则要求智算网络具备快速适应能力。

网络感知是智算网络的基础，主要通过网络测量技术高效采集预测网络状态、环境数据及用户行为，并实时监控网络流量、设备状态和资源利用率，同时整合

多源数据以形成全局视图。网络认知则是智算网络的核心，利用机器学习和深度学习技术对数据进行分析与建模，构建网络行为模型，并基于感知数据动态调整网络配置、资源分配和路由策略，通过强化学习等技术实现网络的自我优化与自适应，以应对复杂多变的环境需求。

二、项目应用平台与基础

1．硬件平台

GPU 算力集群和可编程交换机

2．软件平台

基于 K8s 和 PyTorch 编程框架、DeepSpeed 等分布式训练框架和 P4 语言等

3．技术基础

本项目基于 K8s 容器调度、网络测量测绘、VR 或 AR 平台、知识图谱与图处理、大语言分布式训练、多模态大模型等相关理论与技术。

三、项目需求

满足以下要求：

1．自驱力强、思维活跃、有好奇心和探究精神；

2．具有较强的系统抽象和建模能力；

3．熟悉 C++和 python 开发语言；

4．对操作系统和网络架构有浓厚兴趣。

四、项目开展

从以下项目中选取一个：

1．高效网络测量测绘：学习大规模网络的高效测量测绘技术，分析其瓶颈和问题挑战，构建网络表征与可视化分析工具。

2．可编程网络转发优化：学习 P4 编程语言和 RDMA 协议，面向分布式大模型训练和推理，优化智算网络传输协议、流量和队列机制。

3．面向智算网络应用的多模态网络大模型：针对 AR+VR、全息通信、分布式训练和智算网络集群调度，学习构建多模态网络大模型及其运维管理知识体系。

4．网络虚拟化及编排调度：学习网元虚拟化和容器编排技术，构建弹性可扩展和可迁移的智算网络部署及任务分配调度体系。

10．现代数据工程与实时计算团队

现代数据工程与实时计算团队依托于华中科技大学计算机学院计算机软件与理论湖北省重点学科建设，拥有开放的学术氛围和国际前沿的研究方向。目前研究领域主要包括：现代数据工程、人工智能、实时计算、软件工程等。

团队师资力量雄厚，现有教授 3 名，副教授 4 名，讲师 3 名，博士后 1 名，具有博士学位者 11 人，其中新世纪优秀人才 1 人，湖北省杰出青年基金获得者 1 人，湖北省优秀博士论文获得者 1 人，具有海外留学背景 5 人。目前在读博硕士研究生近 100 人，拥有近 500 平方米实验基地，主要实验设备资产总值 300 余万元。团队发展与建设为高层次人才的培养提供了良好的基础设施与外部条件。

团队承担了 80 余项科研项目，包括国家重点研发计划、国家 863 项目、国家自然科学基金重点/面上/青年项目、国防预研重点项目、国防预研基金、企业重大横向应用项目等。发表国内外学术期刊及国际学术会议论文近 200 篇，获得国家发明专利 15 项，国家软件著作版权 6 项；获得湖北省科技进步一等奖 1 项。团队坚持开放与联合，与美国、德国、英国等国家和香港、台湾地区的大学，以及华为、烽火科技集团、中船重工、芯动科技、武汉精伦电子、武汉蓝星科技股份有限公司、北京捷报金峰数据技术有限公司等知名企业保持着密切合作。此外，团队成功举办了包括第 37 届 CCF 中国数据库学术会议(NDBC 2020)在内的多个学术会议。团队秉承“明德、厚学、求是、创新”的华科大精神，倡导“专心致志做事，自由自在做人”的原则，不断开拓进取，勇攀科学高峰，致力成为国内一流、国际知名的研发团队和人才培养基地。

团队包括四个研究方向：

现代数据工程

现代数据工程方向主要研究跨模态数据组织和检索、图数据处理、时空数据管理、知识图谱和大数据分析处理等。首次研究了路网空间移动对象的连续反向 k 近邻查询和广播环境下路网空间中的连续近邻查询问题，提出了高效的查询索引；为应对大数据背景下传统字符串相似度搜索算法时间复杂度高空间消耗大的问题，结合先进的学习索引技术和时空数据系统，提出了基于步长和草图构建的字符串搜索算法以及基于字符串搜索的活动轨迹查找算法；针对现有知识图谱管理系统分布式架构下的相似性检索、子图匹配等问题，设计并实现了面向图谱管

理分布式机群的顶层管理服务，构建了大规模图谱管理机群，基于图分割实现了在总体规模千万级节点、上亿条边的图谱数据集上关键词检索的秒级响应、子图检索的分钟级响应。本方向承担了国家自然科学面上、青年基金，国防预研重点基金等一批项目，并和一批企业开展了广泛合作。牵头的《数据结构》课程获评国家精品在线开放课程，并入选首批“国家线上线下混合式一流本科课程。相关研究成果已发表在 IEEE TKDE, TDS，ACM TWEB 等期刊和 ICDE、CIKM、DASFAA 等重要国际学术会议上，获授权国家发明专利 4 项，国家软件著作版权 1 项。

人工智能

人工智能方向的研究主要包括自然语言处理、推荐系统、Fintech、小样本学习、强化学习等。在自然语言处理方面，主要在针对多模态任务型对话系统面临的语义对齐、知识推理和意图检测等问题进行研究，分别提出了一种基于跨模态联想学习的视觉-语言预训练框架，一种基于领域知识细粒度推理的端到端任务型对话模型和一种基于统一 Transformer 嵌入的多模态对话生成框架。在推荐系统方面，主要研究基于知识增强的推荐和跨域推荐，针对推荐数据稀疏问题并考虑文本信息对推荐数据稀疏性的影响，提出了一种基于文本信息的深度强化学习交互式推荐框架；针对跨域推荐问题，研究了跨域数据特征交互对单域数据的影响，提出了一种基于图卷积的跨域迁移模型。本方向的研究工作得到了总装预研重点基金、航天科学基金、以及企业重大横向项目等课题资助，研究成果发表在 ACL，EMNLP，MM，ECAI，CIKM，DASFAA，IEEE/ACM Trans.等重要学术期刊和会议上，获授权国家发明专利 3 项。

实时计算

实时计算方向的研究内容包括嵌入式实时系统的任务调度、资源管理、节能降载、调试技术与环境等。针对多处理器全局实时任务调度难题，提出了目前具有最高精度的静态优先级响应时间分析方法，研究成果获嵌入式系统领域顶级会议 EMSOFT 2021 最佳论文提名奖。提出了首个多处理器环境下的保证数据时序一致性的任务调度方法，成果获 IEEE TC 期刊亮点论文推荐。针对嵌入式软件开发调试面临的高效率、高健壮性、低能耗、低负载等多种挑战，突破了多项关键技术难题，研制了绿色高效健壮的嵌入式软件开发平台，显著提升了开发效率及环境友好性，推动了嵌入式软件行业的发展。项目成果应用于中船重工第七〇九研

究所、武汉征原电气、武汉天喻信息、武汉精伦电子、上海富友支付等企业重要产品研发，有效降低了软件开发调试成本，产生了显著的社会和经济效益，并于2017年获湖北省科技进步一等奖。本方向承担了国家自然科学重点、面上、青年基金,国防预研重点基金,企业重大横向项目等一批课题，研究成果发表在IEEE TC，TPDS，TMC，TCAD，ACM TODAES，TECS，RTSS，RTAS，EMSOFT等重要学术期刊和会议上，获授权国家发明专利6项，国家软件著作版权2项。

软件工程

软件工程方向主要研究软件安全性分析与质量评估方法。与中国船舶重工集团公司第七〇九研究所深度合作，在软件安全性分析方面，ZK系统代码规范性审查工具提出了支持多种语言的漏洞静态检测框架，对程序源码进行预处理，能有效地减少漏洞检测中的误报率；同时采用基于安全规则语言进行漏洞检测，用户通过该语言自定义安全规则来描述待检测的漏洞模式,并使用该规则来检测源码中用户所关注的特定安全漏洞，该方法已应用于基于自定义规则库的ZK系统软件专用测试工具。在软件质量评估方面,提出了基于关联漏洞的安全评估方法，应用于ZK系统代码健壮审查工具，解决了大部分评估方法不能准确评估多个漏洞联合利用对系统造成的潜在影响，使得评估系统的结果更加客观、全面。本研究方向得到了总装预研基金、企业横向项目等课题的资助，获国家发明专利2项，国家软件著作版权3项。

团队成员：

李国徽	团队负责人 教授，主要研究领域为现代数据工程、实时计算、人工智能		
袁　凌	教授,主要研究领域为大数据分析、人工智能	潘　鹏	副教授,主要研究领域为大数据分析、深度学习
瞿彬彬	副教授，主要研究领域为图计算机、分布式计算	杨茂林	副教授,主要研究领域为软件工程、人工智能
赵小松	讲师，主要研究领域为实时计算、人工智能	杨　中	博士后，主要研究领域为数据工程
阳富民	教授,主要研究领域为嵌入式系统	胡贯荣	副教授,主要研究领域为嵌入式系统

张　杰	讲师,主要研究领域为嵌入式系统	周正勇	讲师,主要研究领域为嵌入式系统

团队联系方式：

联系邮箱：袁凌老师，cherryyuanling@hust.edu.cn

团队主页：http://ade.cs.hust.edu.cn

地址：华中科技大学南一楼中 413，西南 501、502

11．智能与分布计算团队

计算机学院“智能与分布计算团队”依托于计算机应用国家重点学科和计算机数据科学与智能科学学科，是汉江国家重点实验室、湖北省大数据应用工程企校联合创新中心、湖北省大数据安全工程技术研究中心、分布式系统安全湖北省重点实验室、大数据与国家传播战略教育部哲学社会科学实验室、华中科技大学-财富趋势大数据智能与安全联合研究中心的重要组成部分。

团队现有成员 10 人，教授 4 人，副教授 2 人，讲师 2 人，博士后 2 人。其中国家级人才 1 人。主持承担了国家重点研发计划项目、国家自然科学重点基金、国家自然科学基金面上、青年等多项项目，并与华为、中移动、阿里、蚂蚁、国家重点研究所等一大批企业开展了广泛合作。

目前研究领域主要包括下列四个方面：

分布式机器学习：研究联邦学习、云边大小模型协同智能、边缘设备智能等方向，服务大数据处理国家需求。

数据挖掘：研究推荐、图数据分析、多模态数据分析等方向，服务企业现实落地需求

小样本学习：针对样本不足难学习的问题，研究少量样本下模型构建技术，服务自动驾驶等重点前沿领域。

可信人工智能：研究可解释性的人工智能、密态环境下的模型推理与加速，解决人工智能的痛点问题。

团队研究领域集中在智能数据科学与分布式智能计算，包括数据挖掘、分布式学习、小样本学习等诸多相关领域。项目相关技术获授权发明专利上百项，参与起草国家标准 5 项，发表 SCI/EI 论文数百篇，论文总引用数万次，国际同行院士及 IEEE Fellow 给予高度评价，相关的研究成果已发表在 IEEE/ACM Trans. 等期刊和 NeurIPS、ICML、ICLR、KDD、WWW、CVPR、MM、AAAI 等重要国际学术会议上，实现了人工智能领域顶级会议的全覆盖。相关工作入选 2023 与 2024 年度“ACM 中国武汉优秀博士学位论文”与“ACM 中国北京优秀博士学位论文”，1 人获得华为天才少年。研究成果在金融、社交网络、推荐等诸多复杂且重要现实场景中进行了应用，解决了蚂蚁、阿里、百度、华为等互联网厂商的卡脖子技术难题，构建了一系列新型大数据处理技术，获湖北省科学技术进步奖一等奖等多项

奖项。团队连续在国家信息安全竞赛、机器人人工智能大赛、挑战杯等多个赛事获得一等奖。

团队毕业生质量获得企业一致认可，在业界具有良好口碑。每年毕业生被华为、腾讯、阿里巴巴、百度、微软、亚马逊、IBM 等一流 IT 企业“抢人”，年薪屡创新高，人工智能人才供不应求。

团队成员：

李瑞轩	团队负责人 教授，主要研究领域为隐私计算，人工智能，大数据处理，数据挖掘		
李玉华	教授,主要研究领域为数据挖掘	周　潘	教授,主要研究领域为机器学习，可信人工智能
文坤梅	副教授，主要研究领域为数据挖掘，情感计算	辜希武	副研究员,主要研究领域为大数据处理，分布式计算
邹逸雄	讲师，主要研究领域为计算机视觉、小样本学习	王号召	讲师，主要研究领域为分布式机器学习、推荐系统
齐伊宁	博士后,主要研究领域为隐私计算	刘　伟	博士后,主要研究领域为可解释人工智能

团队联系方式：

联系邮箱：hz_wang@hust.edu.cn，王号召老师

地址：华中科技大学南一楼西 442 房间。

项目 1：大模型的跨域垂直领域小样本学习

一、项目背景

近年来，以深度学习与大模型为代表的人工智能技术在通用任务上取得了长足的发展，例如监控视频分析、自动驾驶等，因此对于我国信息化建设尤为重要。然而，在现实场景中，人工智能技术在垂直领域应用落地时仍面临巨大挑战，难以可靠便利地应用于各行各业的发展。在《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》、《新一代人工智能发展规划》、《“十四五”国家信息化规划》中，人工智能技术落地被列为“十四五”的重要内容。因此，研究面向现实场景垂直领域应用的人工智能技术具有重大意义。

视觉任务作为一类重要的应用，因其广泛性与廉价性，在许多落地应用中被寄予厚望。在视觉语言大模型风靡的当下，泛化并微调通用的视觉与语言大模型已逐渐成为落地视觉任务的通用方案。然而，现有应用方案通常针对通用数据领域，而许多垂直应用场景与通用数据域存在巨大的差异，直接泛化大模型到这些领域仍然面临巨大挑战。例如，大模型通常在自然图片、通用文本下进行预训练（上游），随后泛化到同为自然图片、通用文本驱动的任务下（下游，例如自动驾驶）。当泛化模型到跨度更大的任务（例如医学影像、卫星遥感等）上进行微调时，效果通常不尽如人意，主要体现在如下三方面：

1．上下游数据的风格差异：现有大模型深度学习方法依赖于海量训练数据，由于通用数据（例如自然图片）易于获取，因此大模型通常在通用数据上进行上游的预训练。但是，下游垂直领域通常呈现出极高的专业性（例如医疗影像），在图片风格、像素分布上可能与上游通用数据呈现出巨大差异。例如，自然图片像素呈现出彩色、低频的分布，而 X 光图片像素呈现出黑白、高频的分布等，这极大影响了上游模型在下游数据上的泛化迁移。

2．上下游数据的语义差异：类似风格差异问题，现有大模型深度学习方法通常在通用语义文本（例如社交媒体）上进行上游预训练，而下游垂直领域存在大量专业知识（例如“冠状肺炎”），难以涵盖在上游文本知识中，导致大模型对于下游知识难以有良好的表征，限制了上游模型到下游的垂直泛化迁移。

3．下游数据通常更难收集：不同于上游通用数据，下游数据通常更难收集，例如医疗数据可能需要昂贵的设备、专业的人员操作、珍稀的病例等，卫星遥感

数据可能涉及隐私机密，因此难以像上游一样轻易获得大量数据，因此需要考虑在下游样本不足的情况下高效微调上游迁移来的模型。

针对上述问题，为了促进现有大模型深度学习技术在现实的跨域垂直领域落地，本项目旨在研究泛化上游大模型到下游垂直领域上、并利用少量样本高效微调的方法，让大模型技术高效、可靠地垂直落地。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

基于 Centos7、PyTorch 等深度学习及大模型库进行编程

3．技术基础

本项目基于深度学习、强化学习、大语言模型、参数高效微调、检索增强生成相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．视觉跨域泛化：通过训练模型在不同视觉域（如自然场景到艺术画作）间有效迁移知识，实现对新域的稳健识别与理解，提升模型泛化能力。

2．语义跨域泛化：通过构造中层语义表达、知识增强等技术，使模型能够从通用领域（如日常沟通）学习到的信息，成功应用于另专业领域（如医疗问答），跨越不同领域间的语义鸿沟。

3．跨域小样本学习：在目标域只有极少数标注样本的情况下，结合源域迁移的信息与高效微调技术，实现下游任务的少样本学习。

4．零样本学习：在完全不依赖目标类别训练样本的情况下，利用类别间的语义关系（如属性描述、文本标签）进行学习与预测，实现未见类别的识别。

5．小样本增量学习：面对新类别数据持续到来的场景，模型能够在仅提供少量新样本的情况下，不断学习并扩展其识别能力，同时保持对旧知识的记忆，减少遗忘。

项目 2：大模型微调数据选择方法研究

一、项目背景

近年来，以深度学习与大模型为代表的人工智能技术在通用任务上取得了长足的发展，例如监控视频分析、自动驾驶等，因此对于我国信息化建设尤为重要。然而，在现实场景中，人工智能技术在垂直领域应用落地时仍面临巨大挑战，难以可靠便利地应用于各行各业的发展。在《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》、《新一代人工智能发展规划》、《“十四五”国家信息化规划》中，人工智能技术落地被列为“十四五”的重要内容。因此，研究面向现实场景垂直领域应用的人工智能技术具有重大意义

对于一般的科研机构而言，训练大模型的成本难以接受。微调（Fine-tuning）通过在目标任务的数据集上对模型进行进一步训练，能够显著提升模型的性能。因而科研机构的研究方式一般是使用在通用数据上大规模预训练的大模型在专用数据上进行微调来构建垂域大模型。然而，由于微调数据的质量和内容对模型的表现有直接影响，选择合适的微调数据具有重要意义。

具体而言，数据选择具有如下方面的意义：（1）**提升模型性能**。针对目标任务的高质量数据能够使模型更好地捕捉任务相关的特征，从而提高准确性和效率。优化的微调数据可以减少模型的过拟合风险，提升泛化能力。（2）**增强模型对领域知识的掌握**。通过加入目标领域的专用数据，模型能够更好地理解领域相关术语、语境以及问题结构。例如，法律文本、医学报告或科技论文等具有独特的语言风格和语义特点，微调数据能够帮助模型适应这些差异。（3）**提高数据利用效率**。数据选择可以聚焦于任务最相关的数据，避免冗余信息，提升训练效率。通过挑选具有代表性的数据，减少标注数据需求，降低成本。（4）**提升鲁棒性与公平性**。数据选择有助于去除偏见和噪声，改善模型对少数群体或特殊情况的处理。确保数据分布的多样性，可以避免模型对特定语言模式或群体的过度偏向。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器。

2．软件平台

基于 ubuntu 系统，Pytorch 等深度学习以及大模型库进行编程。

3．技术基础

本项目基于机器学习、深度学习、大语言模型等相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．有足够的自驱力，对未来发展有明确目标和计划，有在顶级会议上发表论文的目标，有充足的自由时间投入；

2．熟悉深度学习，熟练掌握 PyTorch 编程；

3．参加过深度学习相关比赛，有大模型微调经验。

四、项目开展

针对项目描述中的场景需求，逐条或并行开展：

1．基础知识学习与领域文献调研：熟悉大模型及微调相关的基础知识，理解数据选择的重要性。了解大模型（GPT、BERT 等）的基本原理、预训练和微调机制。学习自然语言处理（NLP）任务（如分类、生成、问答等）的主要数据类型和特点。学习机器学习和深度学习中的数据处理与评价指标。阅读指导老师给定的论文，并自行扩展调研相关论文。

2.动手能力训练：掌握 NLP 工具和库的使用，如 Hugging Face Transformers、TensorFlow、PyTorch。使用 Hugging Face 加载预训练模型（如 BERT 或 GPT-2），尝试对简单任务（如文本分类）进行微调。熟悉数据预处理技术，如分词、清洗、格式转换。

3．深入理解数据选择：实现如下 4 个方法的数据选择策略，观察各自优缺点。基于任务相关性：TF-IDF、语义相似度计算。基于表示学习：利用预训练模型提取特征，选择高质量数据。基于不确定性：通过模型的不确定性得分选择训练数据。基于多样性：通过聚类或覆盖方法构建具有代表性的数据集。探索特定任务的数据集（如 SQuAD、GLUE、COCO）。

4．前沿方向学习：自适应数据选择：根据模型的学习阶段动态调整数据。数据质量评价：开发自动化数据质量评估指标。数据合成与增强：利用生成模型

（如 GPT-3）生成补充数据。数据去偏：设计方法减少数据集的偏见对模型的影响。

5．实际问题研究：选择一个特定的任务（如医疗文本分类），探索数据选择的优化。比较不同领域数据（如多语言、多模态）的选择方法。关注学术会议（ACL、EMNLP、NeurIPS）中的数据选择相关论文。总结经验、提炼方法，撰写高质量的论文或技术报告。整理实验代码，发布到 GitHub。

项目 3：基于大模型的人工智能应用异常识别技术

一、项目背景

人工智能技术的高速发展，给各个产业带来了革命性的新面目。以数据为生产要素的生产力变革正成为推动社会经济发展的重要动力。金融、零售、医疗等传统产业与人工智能应用的结合，产生了商业与社会治理的新范式，为社会的持续进步注入了新活力。但另一方面，人工智能应用高度智能自动化的特点，结合不同于人类思维的智能模式，给行业监管与社会治理带来了新的挑战与风险。以金融行业为例，近年来，以人工智能应用为载体的异常乃至违法行为层出不穷。在人工智能高频自动化操作条件下，高价超募、抱团压价、信息披露违规、操纵市场、重大舆情等异常行为发生更迅速，监管时间窗口大大缩短，且违规手段更加隐蔽多样，有必要针对性地研究新型异常识别技术。

近年来，基于大语言模型（LLM，如 Llama、GLM、DeepSeek 等）的异常识别技术引起了学术与产业界的高度关注。利用大模型卓越的上下文理解能力，准确把握纷繁复杂的行为数据，理解行为主体意图，构建不同智能体进行自动化协同判定，显著提升异常识别的效率和精准性。然而，LLM 在处理复杂行为识别时仍面临一系列挑战，特别是垂直领域知识的匮乏和幻觉风险。此外，在监管领域，由于训练数据的稀缺性和人工智能决策的黑盒性，LLM 任何误判（假阳性）都有可能导致严重的后果，导致其在社会治理中的应用受到限制。

为克服这些局限，检索增强生成（RAG）与智能体技术逐渐成为关注的焦点。RAG 通过为大语言模型提供外部知识库，为决策过程提供实时、相关的上下文支持，显著提升了决策的准确性和相关性。但传统 RAG 方法在处理复杂领域需求时仍存在不足：在捕捉人类专家知识中的实体关系，理解特定垂直领域语义方面，现有 RAG 技术存在明显的短板；检索在捕捉语义依赖性和模糊查询方面，导致检索结果无法完全精准识别特定垂直领域语义。

智能体技术以 AI 为核心，构建了一个立体感知、全域协同、精确判断、持续进化和开放的智能系统，具体而言，智能体技术主要由以下三个部分组成：

1．规划（Planning）：智能体会把大型任务分解为子任务，并规划执行任务的流程；智能体会对任务执行的过程进行思考和反思，从而决定是继续执行任务，或判断任务完结并终止运行。

2．记忆（Memory）：短期记忆，是指在执行任务的过程中的上下文，会在子任务的执行过程产生和暂存，在任务完结后被清空。长期记忆是长时间保留的信息，一般是指外部知识库，通常用 RAG 来存储和检索。

3．工具使用（Tool use）：为智能体配备工具 API，比如：计算器、搜索工具、代码执行器、数据库查询工具等。有了这些工具 API，智能体可以与外部世界交互，解决实际的问题。

智能体技术能够大大提高大模型异常识别的决策可控性，有助于构建人类可理解的判别系统，从而更好地融合人类专家知识，使之符合特定垂直领域需求。

通过将 RAG 与智能体技术相结合，本项目致力于构建精准高效的人工智能异常识别系统，提升金融监管领域的社会治理能力和监管效率。这种融合方法既发挥了 RAG 在信息检索方面的优势，又利用智能体的高效专家系统，有效提升了系统对复杂需求的理解和处理能力。在实践中，RAG 为智能体提供精准的领域知识支持，而智能体则通过高效的决策能力来实现精准异常识别。这种创新性的技术融合不仅提升了异常识别的效率和质量，更为社会治理的智能化发展提供了新的技术路径，同时也为其他复杂专业领域的异常识别系统建设提供了可借鉴的技术方案。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器。

2．软件平台

基于 ubuntu 系统，pytorch 等深度学习以及大模型库进行编程。

3．技术基础

本项目基于机器学习、深度学习、大语言模型、检索增强生成、智能体等相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 编程基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．领域数据收集：通过 Python 爬虫等技术手段，结合项目所提供的数据，针对资本市场监管领域，整合多种数据来源并进行清洗，包括领域专业文献、行业案例分析、市场技术规范、用户反馈、相关政策法规等。

2．智能体构建：基于领域数据和人类专家知识，设计智能体工作流程，并开发相应的工具小模型，增强大模型异常识别系统对特定领域应用的适应性。

3．检索向量数据库构建：设计并实现一个高效的检索向量数据库，通过对领域数据进行分块和向量化处理，实现语义级别的检索。

4．检索与生成算法实现：实现检索增强生成（RAG）流程，实现向量检索和知识图谱检索，并将其深度融合，提升上下文表征能力与生成结果的相关性。

5．大模型环境搭建与优化：搭建大模型运行环境，包括基础算力支持和软件环境配置。优化计算资源分配，确保大规模模型推理的高效执行。

6．全阶段框架实现：开发一个完整的资本市场人工智能应用异常识别系统，涵盖数据收集、智能体构建、检索与生成等多个阶段，确保系统各模块无缝衔接，并支持多轮对话中的动态需求解析和更新。

项目 4：基于多模态大模型的中文金融文档信息抽取研究

一、项目背景

文档是信息抽取通常包括文档识别与解析、文档内容理解两个子任务。文档解析的目标是把一篇文档图像转为机器可读的 JSON 或 markdown 格式。首先通过版面分析得到一页文档图片中表格、文字、插图等内容的位置，再把不同内容交给对应的模块去解析，比如传统技术使用 OCR 工具提取文字、使用表格解析模型解析表格结构和内容等。最后把这些内容拼接并排版形成机器可读的文档格式，交给下游信息提取、文档问答、文档分类等任务使用。传统的文档解析范式依赖多个外部模块，难以保证整体流程的效率和准确率。最近几年出现了基于多模态大模型的端到端的解决方法，这些多模态模型可以直接接收文档图像和文字输入，展现出了很强的竞争力。

文档理解过去通常作为文档解析的下游任务，近几年有大量研究工作尝试端到端地解决问题。随着大语言模型的发展，基于 Transformer 解码器的架构逐渐流行。文档图像可以直接输入到编码器中，随后通过 Transformer 解码器来解析信息，并最终以自回归的方式输出结果。这一领域的开创性工作包括 Naver 公司 2022 年发布的 Donut 模型[9]，以及百度公司在 ICLR-2023 会议上介绍的 StrucTexTv2 模型[10]。

自从 GPT4V 发布之后，视觉理解领域很多研究工作都是基于大语言模型 LLM 以利用其强大的世界知识和推理能力。这样可以利用已有的大量单模态训练数据训练得到的单模态模型，减少对于高质量图文对数据的依赖，并通过特征对齐、指令微调等方式打通两个模态的表征。例如，微软推出的 LLaVA 模型首次尝试将指令调优扩展到语言图像多模态空间[16]，使用 GPT-4 根据生活场景图片生成了大量的文本-图片预训练数据，并使用简单的线性层作为视觉空间到文字空间的映射。

文档理解与常见的视觉理解领域有所不同，文档图像中的文字和图表通常都是丰富且密集的，需要模型具有细粒度的视觉感知能力以及推理能力。由此出现了面向文档的多模态大模型，通常是在原有模型的基础上使用文档数据微调或者在结构上做出改进，使得他们能更好地处理文档数据。比如，mPLUG-DocOwl1.5 系列模型是在 mPLUG-Owl 的基础上使用文档指令微调得来[17]。Monkey 是在

Qwen-VL 的基础上增强了模型处理高分辨率文档图像的能力[18]，高分辨率图像使得模型输出更有把握的答案，但带来了视觉 token 过多的问题。

在金融领域中，模型对于金融文档的理解和信息提取能力是一项常见且需求广泛的基础能力，近期针对金融领域大模型研究逐渐增多。例如，复旦大学 DISC 团队推出的中文金融大语言模型 FinLLM[19]，旨在为用户提供专业、智能和全面的金融咨询服务。同时，也有一些研究专注于评估大语言模型在金融领域的能力和知识，开发了如 Fin-Eva 和 FinEval 等中文语言数据评测集。然而，目前的研究大多集中于大语言模型，对多模态大模型（Multimodel Large Language Model）以及金融文档领域的应用探讨仍显不足。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器。

2．软件平台

基于 pytorch 等深度学习以及大模型库进行编程。

3．技术基础

本项目基于机器学习、深度学习、多模态大语言模型等相关理论与技术。

三、项目需求

下列要求中 1 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 编程基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．领域数据收集：结合项目所提供的数据，针对金融领域，整合多种数据来源并进行清洗，包括上市公告、金融财报、公司年报等。

2．多模态大模型架构设计：基于大语言模型与 ViT 等视觉模型主干网络，设计多模态大模型

3．模型微调：基于 LoRA 等微调方法，设计金融垂域大模型的微调算法。

4．提示工程设计：设计并实现金融文档抽取的提示词，针对不同金融信息类型自适应抽取。

5．金融大模型加速：金融信息抽取要求高效性，需要采用参数量化、令牌剪枝等技术加速大模型对金融信息的处理。

参考文献

[1] Kim G, Hong T, Yim M, et al. Ocr-free document understanding transformer[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 498-517.

[2] Tang Z, Yang Z, Wang G, et al. Unifying vision, text, and layout for universal document processing[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 19254-19264.

[3] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2024, 36.

[4] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding, 2024.7, 8

[5] Li Z, Yang B, Liu Q, et al. Monkey: Image resolution and text label are important things for large multi-modal models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 26763-26773.

[6] Chen W, Wang Q, Long Z, et al. DISC-FinLLM: A Chinese financial large language model based on multiple experts fine-tuning[J]. arXiv preprint arXiv:2310.15205, 2023.

项目 5：基于知识图谱与大模型检索增强的智能问答系统实现

一、项目背景

随着全球数字经济的蓬勃发展，数据已成为推动经济增长与产业升级的关键生产要素，承载着巨大的社会与商业价值。数据要素交易市场通过统一、开放的数据共享机制，成为实现数据资源高效配置和价值流通的核心平台，为企业和机构搭建供需对接的桥梁，促进产业协同创新。然而，这一平台的高效运作面临诸多技术挑战，尤其在智能问答系统方面：如何精准处理复杂的领域知识、应对数据类型的异构性，以及在多轮动态交互中实现高效、准确的响应，已成为亟待解决的难题。传统方法过度依赖表面特征进行语义理解，难以深入挖掘深层语义关系和上下文信息，导致用户意图理解欠佳且系统可扩展性有限。

近年来，基于大语言模型（LLM，如 GPT、Claude、T5 等）的方法，通过大规模数据预训练和自监督学习，展现了卓越的上下文理解能力。LLM 能够准确把握复杂的语言现象，理解用户意图，并生成连贯流畅的回答，显著提升了智能问答任务的交互效率和用户体验。然而，LLM 在处理复杂领域知识时仍面临显著局限，特别是在专业领域知识缺乏和生成内容可能出现幻觉方面。由于训练数据的多样性以及缺乏针对特定领域的深入学习，LLM 生成的回答有时可能不准确或偏离实际，导致其在工业生产中的应用受到限制。

为克服这些局限，检索增强生成（RAG）技术逐渐成为突破口。RAG 通过融合外部知识库与语言模型，为生成过程提供实时、相关的上下文支持，显著提升了回答的准确性和相关性。但传统 RAG 方法在处理复杂领域需求时仍存在不足：检索阶段难以有效处理语义依赖性和模糊查询，导致检索结果无法全面覆盖复杂领域语义；生成阶段因缺乏结构化约束而逻辑性不足。此外，现有 RAG 方法在多轮交互中，特别是在捕捉实体关系与理解专业领域语义方面，仍有明显短板。

知识图谱（KG）的引入为 RAG 技术开辟了新的发展空间。知识图谱通过结构化建模领域实体及其关系，为问答任务提供精准的专业知识支持，并通过推理能力增强回答的条理性与可解释性。在数据要素交易场景中，知识图谱弥补了传统 RAG 方法在语义覆盖和逻辑生成上的不足，尤其在处理复杂语义表达和推理任务时展现独特优势。其动态适配能力在多轮交互中提供稳定的知识支撑，使回答更贴合用户需求和上下文语境。

通过将 RAG 技术与知识图谱相结合，我们致力于构建高效的智能问答系统，提升数据要素交易平台的智能问答能力和用户体验。这种融合方法既发挥了 RAG 在信息检索与生成方面的优势，又利用知识图谱的结构化表征和推理能力，有效提升了系统对复杂需求的理解和处理能力。在实践中，知识图谱为 RAG 提供精准的领域知识支持，而 RAG 则通过灵活的生成能力将这些知识转化为流畅、准确的回答。这种创新性的技术融合不仅提升了问答交互效率和回答质量，更为数据要素交易市场的智能化发展提供了新的技术路径，同时也为其他复杂专业领域的智能问答系统建设提供了可借鉴的技术方案。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器。

2．软件平台

基于 ubuntu 系统，pytorch 等深度学习以及大模型库进行编程。

3．技术基础

本项目基于机器学习、深度学习、大语言模型、检索增强生成、知识图谱等相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 编程基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．领域数据收集：通过 Python 爬虫等技术手段，针对数据要素交易领域，整合多种数据来源，包括领域专业文献、行业案例分析、数据市场技术规范、用

户反馈、相关政策法规等。

2．知识图谱构建：基于领域数据，开展实体和关系的自动抽取，构建结构化的知识图谱框架，并探索动态更新机制，增强知识图谱对新场景的适应性。

3．检索向量数据库构建：设计并实现一个高效的检索向量数据库，通过对领域数据进行分块和向量化处理，实现语义级别的检索。

4．检索与生成算法实现：实现检索增强生成（RAG）流程，实现向量检索和知识图谱检索，并将其深度融合，提升上下文表征能力与生成结果的相关性。

5．大模型环境搭建与优化：搭建大模型运行环境，包括基础算力支持和软件环境配置。优化计算资源分配，确保大规模模型推理的高效执行。

6．全阶段框架实现：开发一个完整的数据要素领域智能问答系统，涵盖数据收集、知识图谱构建、检索与生成等多个阶段，确保系统各模块无缝衔接，并支持多轮对话中的动态需求解析和更新。

12．现代数据库技术团队

华中科技大学计算机学院数据库与多媒体技术研究所(亦称现代数据库技术团队）成立于1992年，是国内首批从事数据库管理系统研发的单位，最早推出了自主知识产权的国产数据库管理系统——达梦数据库，创办、孵化了专业研发数据库管理系统的武汉达梦数据库股份有限公司，并在2005年联合申报成立了湖北省数据库工程研究中心，形成了学研产一体化发展模式，有力推动了数据库产品研发、服务等领域的发展，培养了大批研究人才。近年来，随着互联网的发展，研究所在云数据管理、大数据、网络空间安全等方面也展开了较为全面深入的研究。

通过承担国家重大专项分课题“大型通用数据库管理系统与套件研发及产业化—安全数据库管理系统与前沿技术研究”研发完成了虚拟化数据库管理平台、云数据库管理平台的原型系统，原型系统通过了由工信部指定的测试单位《中国软件评测中心》的测试，并通过了工信部组织的课题验收。

在面向大数据的分布式数据管理方面，取得的主要研究成果包括：1）针对传统的关系数据库可扩展性存在的问题，在深入研究数据库内核的基础上研究了存储和计算分离的分布式数据库架构及其查询优化技术；2）针对大数据下存在大量不确定数据的管理问题，研究了概率数据库上基于约束的更新算法，在大数据量环境下的计算性能得到了较大提高。

通过承担一系列数据安全方面的课题，包括：国家自然基金项目“移动社交网络中关联社交关系的位置隐私保护研究”、武汉市科技局项目“移动社交网络中的社交关系隐私保护研究”、湖北省科技厅项目“国产密码解决方案的推广应用”;科技部863计划项目“高安全等级数据库管理系统及其测评关键技术研究”、863计划项目“多级安全数据库管理系统技术研究”、国家密码管理局密码基金项目“云计算中加密数据的密钥管理方法研究”、以及上述国家重大专项分课题“安全数据库管理系统与前沿技术研究”中的云数据安全等，本团队已经在数据库系统安全、数据安全、隐私保护、云端密文检索等方面取得了大量研究成果。其中，依据数据库安全等级国标五级的要求研究的安全数据库管理系统原型、云环境下多租户的安全保障机制、在大数据量和分布式环境下的隐私保护方法以及私有信息快速检索方法等成果具有代表性。

秉承科技服务于社会的原则，研究所在国内最早推出了国产数据库产品——达梦数据库，创办、孵化了专业研发数据库管理系统的武汉达梦数据库有限公司，并与达梦公司联合申报成立了湖北省数据库工程研究中心，形成了学研产一体化的良性发展模式，大大推动了数据库产品研发、服务等领域的发展，也为数据库产业培养输送了大量人才。

团队现有教授 1 人，副教授 2 人，讲师 6 人。

主要研究方向：

1.分布式云数据管理系统架构研究

主要包括两个方面：云数据库架构以及基于人工智能的数据库性能优化研究。

云数据库架构研究，主要研究基于关系型 DBMS 的新型云数据库架构及核心技术，包括：计算与存储分离技术、日志即数据库技术、分布式共识协议、新型硬件在 DBMS 上的有效应用等，以支持多计算节点和多存储节点可灵活配置的、低同步延迟、高可用、高可扩展的分布式数据库集群。

基于人工智能的数据库性能优化，主要研究将人工智能技术应用到分布式数据库系统自治和性能调优上，包括：对分布式数据库集群的参数自调优、分布式数据库查询优化、缓冲区自适应管理、系统负载均衡、集群内并发调度等方面。通过人工智能与数据库内核技术的结合，以达到既能充分利用分布式数据库的资源优势，又能提供高性能数据库服务的目标。

2.大数据环境下数据分析

主要研究将人工智能技术应用于大数据分析，以便进行预测、预警、数据关联性分析等，包括：将来自不同数据源的数据进行整理、清洗，根据不同的应用需求，评估数据在进行各类数据分析之前数据是否满足完整性、准确性、一致性等特性，以及是否存在重复数据的方法。数据分析和挖掘相关算法、策略研究；数据分析结果的可视化展示等。

3.数据库安全与隐私保护技术

安全数据库的研究主要研究高安全等级的数据库管理系统实现的关键技术，如形式化安全分析技术、隐蔽通道分析技术、安全审计分析技术以及当数据库管理系统面临入侵时的入侵发现、数据受损评估和受损数据恢复等技术。

隐私保护与数据发布技术，主要针对来自多个数据源的数据进行数据分析和

数据挖掘引起的隐私泄露问题，在将关系数据、位置数据以及社交网络数据等发布提交出去之前，先进行隐私保护处理，在保障安全性的同时又能保持处理后的数据可用性。

4.面向大数据平台的测试技术

主要包括两个方面的研究，一是国产基础软件（操作系统和数据库）安全性测试工具的研发；二是面向测试基准（如 TPC 系列），研发数据库以及大数据平台的测试工具。如：数据库稳定性测试工具、Oracle 数据库符合性测试工具、大数据平台测试工具(TPC-x.bb)、数据库性能测试工具(TPC-E，TPC-DS)等。

团队成员：

朱　虹	团队负责人 教授，主要研究领域为新一代云数据库管理系统、大数据环境下的隐私保护、大数据分析与数据挖掘以及高安全等级系统的测试等		
曹忠升	副教授,主要研究领域为数据库管理系统、多媒体处理技术、云计算与大数据技术等	周英飚	副教授，主要研究领域为数据库管理系统、多媒体处理技术、云计算与大数据技术、密码技术等
左　琼	讲师,主要研究领域为现代数据库技术、云计算和大数据管理与技术、数据库安全技术等	谢美意	教授,主要研究领域为数据库安全、现代数据库理论与技术
班鹏新	讲师，主要研究领域为现代数据库理论与技术，数据库安全，大数据，机器学习	李　专	讲师，主要研究领域为现代数据库技术
张　勇	讲师,主要研究领域为现代数据库理论与技术、互联网应用与移动互联网应用的设计、硬盘故障分析与修复、硬盘固件分析、数据及文件系统恢复	李海波	讲师,主要研究领域为现代数据库技术

团队联系方式：

联系邮箱：zhuhong@hust.edu.cn，朱虹老师

实验室地址：华中科技大学南一楼一楼中厅数据库研究所

项目 1：数据库集群上的智能多参调优（指导老师：左琼）

一、项目背景

数据库管理系统中存在大量可调参数，传统数据库管理员针对负载特点根据既有经验和规则对这些参数进行调参以提升系统整体性能。随着数据库规模的增大，人工调优在整体调优效果和时效上存在不足明显，基于机器学习的数据库参数自动化调优方法开始流行。

在数据库参数调优中，从调整参数到调优效果可见是需要有一个时间间隔的。随着参数维度的增加，自动化数据库调参是一个很耗时的过程。而且调优结果具有时效性，随着负载变化，已有的调参结果将不一定仍是最优解，需要实时动态调整。现有的基于机器学习的数据库多参调优方法主要可分为 3 大类：启发式方法、基于贝叶斯的调优方法和基于强化学习的调优方法。在模型训练过程中，通常需要依赖大量的先验样本数据来进行模型训练或重要参数筛选，以提升调优的准确型和性能。但面向不同负载，重要参数的选择和调优程度是不同的，实例数据的稀缺性仍然是一个关键问题。往往难以获得大量的数据来覆盖所有可能的情况，这导致了在模型训练时的困难，并限制了模型在处理未知或新场景时的性能，难以有效泛化。我们希望可以研究出优化方法来提升数据库多参调优的效率和对漂移负载调优的有效性，该构想分为两个方面：①在有调优规则集而缺少样本数据时，使用规则增强生成方法并优化；②在前期调优基础上，利用已有调优历史信息，针对负载变化或硬件部署的差异，使用迁移学习方法来进行参数调优。

在有规则集而缺少实例信息时，使用规则增强生成方法并优化。当缺少样本数据时，可以将规则知识与强化学习方法相结合，从而快速生成合适的输出。该方法的关键步骤包括：首先，通过构建一个规则知识库，系统地存储各种负载特点和数据库状态下的数据库调参规则。接着，在强化学习模型中根据当前数据库状态选择合适的数据库规则给出调参推荐。与强化学习方法相结合，生成合理的参数推荐输出。最后，使用强化学习等方法对调优过程进行优化，以缩减参数空间、提升有效规则的选择率和生成调优结果的质量。

在前期调优基础上，利用已有调优历史信息，针对负载变化或硬件部署的差异，使用迁移学习方法来进行参数调优。关键步骤包括：首先，选择模型中与调优任务最相关的部分（如某些特定层或注意力头），对其进行微调，以适应新的

负载数据。在微调过程中，仅更新少量的模型参数，同时保持原模型大部分参数不变。其次，使用梯度累积或分层学习率调整等技术，以提高训练效率并确保模型在有限数据下的泛化能力。最后，通过实验验证微调后的模型性能，确保模型能够高效地学习到漂移负载的特征，同时最大化调优效果。

因此，在数据库集群多参调优场景中，针对负载样本不充足和动态变化的场景，如何有效动态调参才能使其快速适应负载的变化，成为了这一领域亟需解决的核心问题。本研究旨在探索经验规则与强化学习相结合的技术，通过将规则知识与强化学习有机结合，提升参数推荐的准确性和效率。通过调参推荐结果与规则的映射，进一步增强调参结果的可解释性，最终实现更加高效、精确的数据库集群调参推荐。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

数据库参数接口编程（C++）、基于 LLamaTune、DDPG、TD3、SAC 等数据库参数调优模型进行编程（Python）。

3．技术基础

本项目基于贝叶斯模型、强化学习、参数高效微调、启发式规则等相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．了解深度强化学习、贝叶斯模型等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．数据库调优规则集构建：针对特定负载类型和数据库类型（Mysql 或 PostgreSQL），收集和整理包含负载特点、数据库状态和控制参数的调参规则及规则间关系，覆盖调参任务需求的多样性与复杂性，为模型调优提供有效的经验规则集。

2.数据库集群调参环境的搭建及调参模型更新：搭建基 Mysql 或 PostgreSQL 的数据库集群环境（共享存储架构），构建基于贝叶斯模型或深度强化的数据库集群调参模型，确保训练的高效性。

3．规则知识与调参模型的融合方法：设计一个框架或方法论，允许不同的规则集在调参模型的推理过程中被有效利用，确保对于特定负载场景的有效规则选取和利用。

4．基于迁移学习的负载偏移动态调参方法：包括：低秩适应（LoRA）或其它增量微调技术优化，使得在有限的负载样本下，能够高效地微调调参模型并保持较高的性能，避免重新训练整个模型，提升训练效率。

13．嵌入与普适计算团队

华中科技大学“嵌入与普适计算团队”于2012年建立，现有IEEE Fellow一名，全球高被引学者2名，荣誉教授1名，副教授3名，硕博研究生42名。团队硬件配备有最新显卡，自建服务器，性能良好的电脑、显示器等，提供良好的办公环境。

团队拥有高水平的研究团队和自由开放的学术氛围,致力于最前沿的研究方向，主要从事**情感计算、无人系统智能博弈、三维视觉、织物计算、智慧医疗、6G**等方向的研究。

EPIC团队主页：https://mmlab.snu.ac.kr/~mchen/epic2022/。

团队成员：

胡　龙	团队负责人 副教授，主要研究领域为情感计算、无人系统智能博弈，人工智能，6G等方向		
郝义学	副教授,主要研究领域为边缘计算、多智能体强化学习、智慧医疗等	李贤芝	副教授,主要研究领域为三维视觉、智能感知等方向

团队联系方式：

联系邮箱：hulong@hust.edu.cn 胡龙老师

地址:华中科技大学南一楼西216

14．嵌入式与人工智能团队

计算机学院“嵌入式与人工智能团队”依托于计算机软件与理论湖北省重点学科，是数据工程研究所的重要组成部分。团队现有副教授 1 人。主持过国家高技术研究发展计划（863 计划）、国家自然科学基金、国家发改委信息安全专项、湖北省信息产业发展专项、文化部科技创新项目、湖北省光电子专项等重点及广电总局重大项目；获得湖北省科技进步奖（三等奖）、国家广电总局科技创新奖（二等奖）、武汉市科技进步奖（三等奖）等；是武汉市青年科技晨光计划资助对象；已发表国内及国际高水平学术会议及期刊论文近二十余篇。团队在“互联网”+大学生创新创业大赛、第二十二届中国计算语言学大会（CCL2023）汉语高考阅读理解对抗鲁棒测评、第二十四届中国机器人及人工智能大赛、第二十三届中国计算语言学大会评测获得一等奖；

目前研究领域主要包括下列三个方面：

检索增强系统：基于向量数据库，实现大语言模型的记忆增强，服务国家人工智能产业升级。

代码生成：利用大语言模型，从自然语言描述或代码示例中自动生成高质量代码，服务国家人工智能产业升级。

大语言模型可解释性研究：致力于揭示大语言模型模型的内部工作机制，服务国家人工智能产业升级。

团队毕业生质量获得企业一致认可，在业界具有良好口碑。每年毕业生被华为、腾讯、阿里巴巴、百度等一流 IT 企业“抢人”，年薪屡创新高，系统级人才供不应求。

团队成员：

涂刚	团队负责人 副教授，主要研究领域为大语言模型、深度学习、机器学习

团队联系方式：

联系邮箱：tugang@hust.edu.cn，涂刚老师

地址：华中科技大学南一楼中 404 房间。

15. 智能大数据管理与分析团队

智能大数据管理与分析团队（IDEAL, Intelligent Big Data Management and Analysis）拥有自由开放的学术氛围和国际前沿的研究方向，致力于探索大数据存储、管理、挖掘与分析相关前沿技术，积极推进学术成果在产业界的落地。

目前主要研究领域包括：时空、时序和高维大数据管理与分析等。

时空数据管理与分析：研究智慧城市交通的数据管理分析算法与系统，包括智能导航、车辆调度与路径规划等；

时序数据管理与分析：研究面向工业物联网的智能运维算法与系统，包括实时监控、异常检测和根因诊断等；

高维数据管理与分析：研究高效率高精度检索基础算法与理论，包括近似最近邻和最大内积问题等。

团队已在国际知名学术期刊和会议上发表 50 余篇学术成果，包括 SIGMOD，VLDB，ICDE，SIGKDD，WWW，TKDE，VLDBJ，TOIS 等。成果获得数据库领域 CCF A 类顶级会议 VLDB 2024 年最佳论文提名、VLDB 2020 年和 ICDE 2019 年优秀论文（One of the Best Papers）。获得 2021 年度 ACM 武汉新星奖，2020 年度 ACM SIGSPATIAL 中国新星奖。团队同学积极参加各类国际/国内大赛，获得第十七届挑战杯揭榜挂帅专项赛全国特等奖，ACM SIGSPATIAL CUP 比赛 2020 年全球总冠军和 2019 年荣誉奖（Honorable Mention）。

团队承担了 10 余项科研项目，包括国家自然科学基金、国家重点研发计划子课题、湖北省自然科学基金以及行业头部企业横向项目等。实验室坚持开放与联合，与美国、澳大利亚、丹麦、新加坡等国家和香港地区的大学，以及微软、华为、腾讯、阿里巴巴、国家电网、滴滴等国内外知名企业保持着密切合作。

团队成员：

郑渤龙	团队负责人 教授，国家级青年人才，主要研究领域为时空、时序和高维大数据管理与分析等

团队联系方式：

联系邮箱：bolongzheng@hust.edu.cn，郑渤龙老师

团队主页：http://ideal.cs.hust.edu.cn

地址：华中科技大学南一楼东 407

16．视觉计算与智能认知团队

计算机学院“视觉计算与智能认知团队”依托于计算机软件与理论湖北省重点学科、计算机应用和智能科学与技术等特色专业，是图像信息处理与智能控制教育部实验室的重要组成部分。

团队现有教授 1 人，副教授 3 人，工程师 2 名，博士后 1 人，博士生 10 余人，硕士生 40 余人。

团队拥有自由开放的学术氛围和国际前沿研究方向。主要从事计算机视觉与人工智能方面的研究，具体包括多模态智能、人工智能安全、大数据智能及智能交通等方面的研究工作。

多模态智能方向：主要聚焦在多模态大模型、机器人具身智能、跨模态指代识别、个体识别与追踪、多模态视觉行为理解、大规模视觉搜索等；

人工智能安全方向：主要是针对人工智能中的安全问题开展研究，具体包括基于深度学习的信息隐藏及数字水印、大模型水印、对抗样本攻击与防御、深度伪造与检测等；

大数据智能方向：主要是基于历史大数据进行自动化建模，对当前生产大数据进行智能预警和预测，包括基于时序大数据的异常检测、基于行业大数据的预测预警等。

智能交通方向：主要从事与智能交通与车路协同有关的计算机视觉与人工智能研究，包括：点云目标检测、目标语义分割、目标轨迹跟踪、驾驶员专注检测等。

团队具备实力雄厚的师资力量、充满活力的科研团队以及良好的硬件设施环境。近 10 年来团队承担了国家自然科学基金重点项目、国家重点研发计划课题、国家 863 计划、国家自然科学基金面上项目、国家电子发展基金、湖北省重大科技创新计划、湖北省重点研发计划等国家级/省部级项目 30 余项。与华为、腾讯、斗鱼、中石油、工行、顺丰、Intel 亚太中心、天喻信息等多个行业龙头企业开展技术合作。与英国、美国、澳大利亚、瑞士、新加坡等多个大学和研究机构开展合作。团队在 AAAI、IJCAI、CVPR、ACM MM、TIP、TMM、TIFS 等知名会议和期刊上发表论文 300 余篇，申请专利 40 多项。获得省部级科技奖励 2 项，指导研究生获得国际比赛前 3 名 10 余项。团队追求卓越的研究成果，加大科技

成果实际落地和转化，研究成果在公安、通信、新能源、石油、银行等行业诸多复杂且重要现实场景中进行了应用。

团队秉承“明德、厚学、求是、创新”的华科大精神，倡导“天行健，君子以自强不息”的原则，不断开拓进取，勇攀科学高峰，致力成为国际一流研发团队和人才培养基地。近年来培养了多名博士，在中科大苏州研究院、国防科大、武汉理工、华为、腾讯等工作，众多硕士毕业生任职于腾讯、阿里、华为、字节跳动、国泰君安、花旗银行、政府机构等知名企事业单位。

团队成员：

<table>
<tr><td>凌贺飞</td><td colspan="3">团队负责人
教授，主要研究领域为多模态智能、人工智能安全、大数据智能</td></tr>
<tr><td>陈加忠</td><td>副教授，主要研究领域为计算机视觉、智能交通</td><td>李 平</td><td>副研究员，主要研究领域为人工智能安全、大数据智能</td></tr>
<tr><td>刘 辉</td><td>副教授，主要研究领域为机器人具身智能、人工智能安全</td><td>史宇轩</td><td>博士后，主要研究领域为多模态智能</td></tr>
</table>

团队联系方式：

联系邮箱：lpshome@hust.edu.cn，李平老师

地址：华中科技大学南一楼 423/432/411/410 房间。

项目 1：多模态行人检索（指导老师：凌贺飞）

一、项目背景

在智能监控系统中，行人检索技术对于提高公共安全至关重要。传统的单模态行人检索主要依赖于可见光摄像头拍摄的 RGB 图像，然而，在复杂环境条件下如光照变化、遮挡或恶劣天气下，单模态数据可能不足以提供准确的行人匹配。随着传感器技术和深度学习的发展，多模态行人检索成为了解决这些问题的关键方法，通过结合多种类型的传感信息如 RGB-D 图像、红外图像、音频、步态等可以显著提升行人检索的鲁棒性和准确性。行人检索技术广泛应用于公共安全、智慧城市管理、零售分析等多个领域，例如在公共场所的安全监控中能够快速定位特定个体有助于预防犯罪和应急响应；在商业环境中可以通过识别常客来优化客户服务体验。此外，该技术还对寻找失踪人员以及灾难救援有着重要的应用价值。尽管基于 RGB 图像的传统行人检索已经取得了显著进展，但在实际部署时仍面临诸多挑战，不同时间段或天气状况下的光照差异会导致图像质量不稳定影响识别效果，部分身体部位被障碍物遮挡会减少可用于比对的有效特征点，雨雪雾等不良气候条件会影响相机成像质量降低检索精度，行人在不同场景中更换服装会使基于外观特征的匹配变得困难。

为克服上述局限性，研究人员提出了多模态行人检索方案，利用来自多个传感器的数据融合以增强系统的适应性和可靠性。具体优势包括增加冗余度，即使某一模态失效其他模态仍能提供补充信息确保整体性能稳定；互补特性，不同传感器捕捉的信息具有不同的侧重点，例如红外相机不受光照限制而深度传感器可以获取三维结构，这些信息相互补充提高了识别准确性；情境感知，结合非视觉信息如音频或动作模式，使得系统能够在更多维度上理解行人行为从而做出更准确的判断；跨域适应能力，通过引入多源数据模型更容易适应新环境的变化减少了对特定场景的依赖。然而，多模态行人检索不仅需要处理来自不同模态的数据融合问题，还需要解决一系列的技术挑战与研究热点，这些挑战直接关联到具体的项目开展方向。

首先，多模态数据集构建是基础，保证数据集的多样性和代表性为后续的模型训练提供可靠的数据支持。为了应对行人穿着改变或者面部被遮挡的情况，将步态信息辅助识行人重识别作为一个重要方向，这涉及到将步态特征与其他视觉

特征相结合，开发能够忽略衣服变化而专注于不变人体特征的算法。文本信息融入也是一个关键点，特别是在没有视频或图片的情况下进行更精确的搜索，这要求利用自然语言处理(NLP)技术，如行人描述（身高、体型、服装颜色等）作为额外输入信息，增强检索结果的准确性。换衣行人重识别则针对行人在不同场景中更换衣物的问题，研究如何让模型学会专注于不变的人体特征实现准确的检索。多模态数据集生成旨在扩充训练数据量，并测试模型对新情况的泛化能力，使用现有的大语言模型或图像生成模型（如 DALL-E、Stable Diffusion 等）创建合成但真实的行人图像和相应的多模态数据。跨域适应确保模型在未见过的环境中仍然表现良好，开发能够处理来自不同分布的数据的方法涉及领域自适应(domain adaptation)和领域泛化(domain generalization)的研究。最后，轻量化模型设计是为了使多模态行人检索能够在边缘设备上实时运行，减少计算资源消耗，这需要研究高效且紧凑的神经网络架构，确保模型在资源受限环境下依然具备高性能。

综上所述，多模态行人检索不仅是学术界的研究热点，也是工业界亟待突破的技术瓶颈。它不仅有望大幅提升现有系统的性能，而且将推动智能监控系统向更加智能化、人性化方向发展，通过有效解决技术挑战并探索新的研究方向，可以进一步增强多模态行人检索系统的实用性和可靠性

二、项目应用平台与基础

1．硬件平台

本项目将采用配备有 RGB 摄像头、红外摄像头、深度传感器等多种感知设备的智能监控系统，并可能涉及便携式穿戴设备或其他可获取人体特征信息的装置。高性能 GPU 服务器，以支持大规模深度学习训练。

2．软件平台

本项目基于 Ubuntu 操作系统，使用 Python 编程语言，结合深度学习框架如 PyTorch 或 TensorFlow 进行算法开发。此外，还将用到 OpenCV 等计算机视觉库来处理图像数据。

3．技术基础

本项目基于深度学习、卷积神经网络（CNN）、循环神经网络（RNN）、注意力

机制、跨模态学习等相关理论和技术，在项目研发过程中需要使用 Python 编程语言。

三、项目要求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入，具备良好的团队合作精神，愿意接受挑战并致力于科研创新；

2．具有一定的 Python 编程语言基础，了解机器学习相关内容；

3．具有一定的 PyTorch、深度学习，对计算机视觉领域有一定了解，特别是行人检测、跟踪及重识别相关知识。

四、项目开展

针对多模态行人检索的应用场景，可以选择以下方向之一或者多个方向展开研究：

1．多模态数据集构建：收集并整理高质量的多模态行人数据集，包括但不限于 RGB 图像、深度图、步态剪影图、3D 图像等，保证数据集的多样性和代表性，为后续的模型训练提供可靠的数据支持。

2．步态信息辅助识行人重识别：步态分析是一种基于行走模式的身份验证方法。将步态特征与其他视觉特征相结合，可以在行人穿着改变或者面部被遮挡的情况下提高识别精度。

3．文本信息融入：将自然语言处理(NLP)技术应用于行人检索，比如利用行人描述（如身高、体型、服装颜色等）作为额外的输入信息。这可以帮助在没有视频或图片的情况下进行更精确的搜索。

4．换衣行人重识别：当行人在不同场景中更换衣物时，传统的行人重识别算法可能会失效。研究如何让模型学会忽略衣服的变化而专注于不变的人体特征，实现准确的检索。

5．多模态数据集生成：使用现有的大语言模型或图像生成模型（如 DALL-E、Stable Diffusion 等）来创建合成但真实的行人图像和相应的多模态数据。这种方法可以扩充训练数据量，并有助于测试模型对新情况的泛化能力。

6．跨域适应：开发能够处理来自不同分布的数据的方法，确保模型在未见过的环境中仍然表现良好。这涉及到领域自适应(domain adaptation)和领域泛化(domain generalization)的研究。

7．轻量化模型设计：为了使多模态行人检索能够在边缘设备上实时运行，需要研究高效且紧凑的神经网络架构，以减少计算资源消耗。

项目 2：高精度 AI 数字人（指导老师：凌贺飞）

一、项目背景

随着人工智能（AI）技术的飞速发展，虚拟形象的应用已经从科幻电影走进了现实世界。高精度 AI 数字人作为这一领域的前沿成果，结合了计算机视觉、自然语言处理、语音合成等先进技术，为用户提供了一种全新的交互体验。在当前数字化转型的大背景下，各行各业对高效、智能的服务需求日益增长，而高精度 AI 数字人凭借其高度逼真的外观、自然流畅的对话能力和多模态交互特性，正逐渐成为提升用户体验和服务质量的重要工具。本项目旨在开发一款能够广泛应用于客户服务、教育辅导、娱乐互动等多个场景的高精度 AI 数字人，以推动各行业的智能化升级，并为用户带来更加个性化、人性化的服务。

高精度 AI 数字人的核心技术之一是计算机图形学与深度学习相结合的人脸建模和表情生成。通过采集大量真实人物的面部图像数据，利用深度神经网络进行特征提取和重建，可以创建出具有极高相似度的 3D 人脸模型。更重要的是，借助于先进的运动捕捉技术和情感识别算法，这些数字人不仅能够在外观上做到栩栩如生，还能根据对话内容实时调整表情，展现出丰富的情感变化。例如，在客服场景中，当客户表达不满时，数字人可以及时做出同情或安慰的表情；而在教育辅导过程中，则能表现出鼓励和支持的态度，增强师生之间的情感连接。这种高度拟真的交互方式，使得数字人在提供信息的同时，也传递出了温暖的人文关怀，从而拉近了人机之间的距离。

除了视觉表现外，高精度 AI 数字人在语言交流方面同样具备卓越的能力。借助于自然语言处理（NLP）技术，数字人可以理解并回应用户的提问，实现双向沟通。更进一步地，通过引入知识图谱和语义推理机制，它们还能够基于上下文进行逻辑思考，给出更具针对性的回答。比如，在旅游咨询场景下，如果游客询问某个景点的历史背景，数字人不仅可以准确地描述相关信息，还能推荐其他相关的名胜古迹，甚至规划出一条完整的游览路线。此外，为了确保对话的质量和连贯性，项目团队特别注重训练数据的选择和标注工作，确保数字人能够应对各种复杂情况下的交流需求，为用户提供流畅且自然的语言交互体验。

安全性与隐私保护是高精度 AI 数字人在实际应用中必须重视的问题。由于涉及到大量的个人信息收集和处理，如何保证数据的安全性和用户隐私成为了关

键挑战之一。为此，我们在设计之初就遵循严格的数据安全标准，采用了加密传输、访问控制等一系列技术手段来防范潜在风险。同时，在法律法规允许范围内，尽可能减少不必要的敏感信息收集，并给予用户充分的选择权和知情权。另外，针对可能存在的伦理问题，如数字人是否会被误认为真人等问题，我们也进行了深入研究，制定了相应的解决方案，例如在适当场合明确标识数字人身份，避免误导公众。通过这些措施，我们致力于构建一个既安全又可靠的高精度 AI 数字人平台，让用户放心使用这项创新技术。

展望未来，高精度 AI 数字人的应用场景将不断扩展，为社会带来更多便利和发展机遇。一方面，随着 5G、物联网等新兴技术的发展，数字人有望实现在更多终端设备上的无缝接入，打破时空限制，随时随地为用户提供服务。另一方面，通过与其他 AI 技术的深度融合，如增强现实（AR）、虚拟现实（VR），数字人将变得更加生动有趣，为用户创造出沉浸式的互动体验。此外，考虑到不同行业对于数字人的特殊要求，我们将继续探索定制化解决方案，满足多样化的需求。总之，高精度 AI 数字人不仅仅是一个技术创新项目，更是连接人类与数字世界的桥梁，它代表着未来智能化生活的无限可能性。通过持续的技术研发和社会合作，相信我们可以共同开启一个充满活力的新时代，让每个人都享受到科技进步带来的美好生活。

二、项目应用平台与基础

1．硬件平台

提供可供充足使用的 gpu 算力

2．软件平台

基于 Huggingface、PyTorch 等深度学习及深度视觉库进行编程

3．技术基础

本项目基于深度学习、扩散模型、参数高效微调、视频生成相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的C++或者Python基础；

3．具有一定的PyTorch基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．数据收集与预处理：协助从公开资源或合作机构获取真实人物的面部图像数据，确保数据来源合法合规。参与对收集到的数据进行标注，如标记面部特征点、表情分类等，为后续的人脸建模提供高质量的训练集。去除噪声数据，整理并标准化数据格式，保证数据集的质量和一致性。

2．模型开发与优化：基于现有的深度学习框架（如 PyTorch），参与初步的3D人脸重建模型训练，理解并应用卷积神经网络（CNN）等算法。探索如何通过语音和文本分析用户情绪，辅助开发能够根据对话内容实时调整表情的数字人。

3．应用场景实现：设计并实现简单的客服场景，让数字人能够在虚拟环境中练习处理常见的客户问题，评估其表现。研究现有数据加密技术和访问控制机制，提出适合本项目的实施方案。调查高精度AI数字人在不同行业中的应用潜力，探讨其对就业市场、日常生活等方面的影响，撰写研究报告。

项目 3：基于字体水印生成的文档保护（指导老师：凌贺飞）

一、项目背景

在当今数字化信息爆炸的时代，各类文档在网络上广泛传播，涵盖了商业机密文件、学术研究成果、创意作品等重要资料。文档的安全保护与版权追溯成为了维护信息秩序、保障创作者和所有者权益的关键需求。

当前，常见的文档水印方式主要分为可见水印和隐式水印两类。可见水印通常以图案、文字等形式直接呈现在文档表面，如在文档角落添加公司标志或版权声明。这种方式虽然能直观地展示版权信息，但严重影响文档的美观度和整洁性，尤其对于需要高质量展示的文档，如设计作品、商务演示文稿等，可见水印会降低文档的专业性和视觉效果。同时，可见水印极易被恶意篡改或裁剪，抗攻击能力较弱，无法为文档提供可靠的安全防护。

隐式水印则将信息嵌入到文档的像素、格式等元素中，在正常阅读时不被察觉。然而，传统的隐式水印存在诸多问题。例如，基于图像像素的隐式水印在文档经历格式转换、压缩、打印扫描等操作后，水印信息容易丢失或损坏，导致难以准确提取。部分基于格式的水印技术应用场景有限，不适用于大多数纯文本或图文结合的文档。此外上述方法对文档的修改较大，仍然易于察觉。

字体水印作为一种相对新颖的隐式水印方式，具有独特的优势。它通过对字体的内部结构进行微小修改来嵌入水印信息，在不影响文档正常阅读体验的前提下，实现版权标识和信息追踪。字体水印不会像可见水印那样破坏文档的视觉效果，且由于其与文档的文字内容紧密结合，相比基于像素的水印更难被发现和去除。然而，现有的字体水印技术也面临挑战。一方面，在保证水印隐蔽性的同时，难以兼顾水印的鲁棒性。当文档遭遇打印、扫描、格式转换等处理后，水印信息往往难以完整准确地提取出来。另一方面，目前的字体水印生成过程大多较为复杂，缺乏高效的自动化手段，需要人工进行大量的参数设置和调试，无法满足大规模文档处理的需求。

随着人工智能技术的迅猛发展，深度学习、机器学习等技术在图像识别、数据处理等领域取得了显著成果，为字体水印领域带来了新的机遇。本项目旨在借助人工智能技术，开发一种能自动生成且具备高隐蔽性、强鲁棒性的字体水印系统。本项目计划从以下三个阶段推进：

基础水印生成算法构建：利用 CNN，Transformer 等方式对大量字体样本进行分析，提取字体的结构特征，如笔画走向、曲率变化等。在此基础上，开发基于特征映射的水印嵌入算法，将水印信息转化为与字体特征相匹配的形式，并嵌入到字体的关键部位，如笔画的控制点、曲线参数等。同时，设计相应的水印提取算法，能够从嵌入水印的字体中准确还原水印信息。在这一阶段，重点关注算法的准确性和基本性能，确保水印的嵌入和提取过程稳定可靠。

水印性能优化与自适应调整：针对文档可能面临的打印、扫描、压缩等操作，引入模拟噪声层，对抗网络等手段对水印进行优化。通过让生成器生成更具抗干扰能力的水印字体，判别器区分原始字体和水印字体，不断迭代训练，提升水印的鲁棒性。此外，考虑到不同类型文档的特点和需求，开发自适应水印调整机制。根据文档的格式（如 PDF、Word、PPT 等）、用途（商业文档、学术论文、个人资料等）以及用户设定的安全级别，自动调整水印的嵌入强度、位置和内容，以实现最佳的水印效果。

智能化水印管理与交互系统集成：构建智能化的水印管理平台，实现对水印字体的集中管理、分发和更新。用户可以通过该平台方便地生成、添加和检测水印，同时平台具备用户权限管理、水印日志记录等功能，方便对文档的版权进行跟踪和管理。打造一个集水印生成、嵌入、检测、管理和交互于一体的智能化系统。

二、项目应用平台与基础

1．硬件平台

本项目依托高性能的计算硬件，有充足的算力资源，为深度学习模型的训练和复杂算法的运行提供强大的算力支持。同时配备高性能的服务器，用于存储海量的字体样本数据、训练模型以及部署水印生成与检测系统。

2．软件平台

项目基于 Python 语言进行开发，利用其丰富的机器学习和深度学习库，如 TensorFlow、PyTorch 等，实现 AI 算法的搭建和训练。操作系统选用 Linux，其稳定的性能和良好的开源生态环境，有利于项目的开发和部署。

3．技术基础

项目核心技术为深度学习和图像处理技术以及部分图形学知识。 深度学习技术用于字体特征提取、水印生成与优化模型的构建。图像处理与图形学知识用于对不同类型的字体进行建模与渲染。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 基础；

3．有一定的 PyTorch 基础，了解深度学习、数字水印、字体渲染等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．大规模字体特征建模与字体生成

2．矢量字体的水印嵌入策略实现

3．针对不同文档格式的自适应水印调整机制

4．字体水印模型训练过程中的渲染优化

5．水印嵌入后的字体视觉质量评估

6．不同场景下字体水印的鲁棒性调优

7．水印管理平台的水印分发与溯源管理机制设计

8．多语言环境下的水印系统适配与优化

9．字符定位与水印检测算法优化

项目 4：抗屏摄隐形水印技术（指导老师：凌贺飞）

一、项目背景

随着数字内容的快速传播与多媒体技术的不断进步，版权保护问题成为数字经济和创意产业的重要议题。在图像、视频等数字资源的应用中，从高分辨率图片到在线教育课程，从影视作品到敏感商业资料，如何有效防止未经授权的录屏、截图以及后续传播，已成为内容安全领域亟待解决的问题。为了应对这一挑战，抗屏摄隐形水印技术（Anti-Screen Capture Invisible Watermarking Technology）应运而生。该技术不仅能够嵌入难以察觉的隐形水印，还具备抗屏摄录制的能力，有效防止通过屏幕录制方式获取并去除水印，从而保障数字内容的版权安全。抗屏摄隐形水印技术通过对内容的微小变动或冗余数据的巧妙利用，使水印在人眼不可见的同时，具备防屏摄录制的能力。具体而言，当用户通过相机设备拍摄屏幕，录像或截图工具录制带有隐形水印的内容时，水印会保持不变，且在屏幕录制的图像中可被识别或提取。

传统水印技术的目的是在不显著影响内容质量的前提下，将版权信息嵌入到多媒体内容中。显性水印（如文字、标志）虽然易于识别，但在内容传播时通常被人为裁剪或覆盖；隐形水印则通过对内容细节的微小调整，将信息嵌入到内容中，人眼不可见但可通过算法提取。这种隐形水印在版权追踪、防伪认证等领域取得了广泛应用。然而，随着屏幕拍摄、录制技术和截图工具的普及，现有水印技术逐渐显露出不足之处：

1．鲁棒性问题：水印嵌入后经拍摄、录制或截图处理之后，往往因分辨率变化、压缩失真，摩尔纹，透视变换和光照影响等原因导致提取水印信号失败，丧失溯源和版权认证能力。

2．动态内容难嵌入：现有的抗屏摄方案要分为固定模板和内容自适应两种形式。前者通过生成固定水印样式模板叠加到保护内容中，实现了快速高效的水印嵌入，但是由于水印模板的固定形式在动态内容上容易出视觉伪影。后者通过生成与图像内容相关的水印可以避免伪影的产生，但是由于屏幕显示的动态内容多样，还需要进一步提升现有方法的嵌入速度。

3．多设备兼容性不足：拍摄后的图像的质量受到，显示设备、相机成像优化算法的复合影响导致图像之间存在不同程度的差异。导致了现有算法难以实现

多相机-显示器之间的水印信号提取。

抗屏摄隐形水印技术应运而生，其核心目标是在各种条件下嵌入的水印能够：

1．保持隐蔽性：对用户而言，水印完全不可见，不影响内容的视觉质量。

2．抵抗屏摄、录制：即便经过相机拍摄和录制处理，水印依然可通过算法恢复或提取。

3．跨设备鲁棒性：水印算法能够适应不同屏幕分辨率、颜色表现以及显示设备的特性，在各种设备上可以有效实现水印信号的提取。

4．动态环境适配：对于视频内容，水印需随内容变化而动态调整，确保其在不同帧间的稳定性。

抗屏摄隐形水印技术的实现涉及多领域交叉，包括图像处理、信号处理、机器学习和硬件设计。其核心挑战包括：

1．如何在内容嵌入水印的同时，确保对人眼的隐蔽性且不降低多媒体质量。

2．如何设计高鲁棒性的水印，使其在经过屏幕拍摄、压缩、模糊、噪声等干扰之后任然可以提取出嵌入的水印信号。

3．如何利用深度学习优化水印的生成与提取过程，进一步提升水印嵌入效率和不同设备下的适应能力。

抗屏摄隐形水印技术在多个行业中具有广泛的应用前景，尤其在内容版权保护、金融信息保密、广告防盗版、社交平台内容保护等领域具有重要作用。随着技术的不断发展，结合深度学习、图像处理、加密技术等多种技术手段，抗屏摄隐形水印将能够更好地应对日益复杂的数字内容盗版问题，提升内容的安全性和版权追溯能力。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

基于 Ubuntu、PyTorch 等深度学习及图像处理库进行编程

3．技术基础

本项目基于深度学习、数字图像处理、数字水印相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、数字图像处理和数字水印等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．抗屏摄水印嵌入算法：研究通过使用深度学习方法来优化传统空域与频域嵌入算法，实现并提升水印在屏摄、压缩等失真场景下的的鲁棒性。

2．高隐蔽性抗屏摄模板水印样式研究：研究通过深度学习方法来生成抗屏摄模板，减少水印模板对于内容的视觉影响，提升水印方法的视觉质量。

3．水印弱信号同步算法：设计水印信号的样式和形态，提出和实现水印弱信号在屏摄场景中的同步问题，确定水印信号在屏摄图像中的位置和方向。

4．高效动态水印扰动生成：研究针对视频多媒体，基于视觉和时间序列的扰动优化策略，以提升水印在动态环境中的适配能力。

5．嵌入算法与设备协同优化：研究针对不同显示屏和相机之间的显示和成像差异，优化水印嵌入策略，提高跨设备适应性。

6．屏摄水印实时嵌入与提取：开发可实时处理的水印嵌入与检测系统，提升实时性和精度。

7．水印强度与视觉感知平衡优化：研究水印嵌入对视觉感知的影响，提升隐蔽性和鲁棒性。

8．基于多模态数据的水印增强算法：结合视觉、音频等多模态信息，提升水印防护能力。

9．复杂环境中的屏摄水印鲁棒性测试：构建和模拟标准化的评估测试环

境，用于评估水印在屏摄、压缩、几何失真等恶劣条件下的鲁棒性。

项目 5：面向垂直领域的 AI 多模态大模型实践与应用（指导老师：凌贺飞）

一、项目背景

随着人工智能（AI）技术的迅速进步，其在各行各业的应用正从广度转向深度，特别是在特定垂直领域中展现出巨大的潜力。传统的 AI 模型往往侧重于处理单一类型的数据，如文本、图像或音频，但在现实世界中，信息通常是多模态的，即包含了多种形式的数据。这促使了多模态 AI 模型的发展，这些模型能够整合和分析来自不同来源的数据，提供更全面、精准的服务。本项目专注于面向垂直领域的 AI 多模态大模型的开发与应用。

AI 多模态大模型可以在多个垂直领域应用。在旅游推荐场景下，多模态 AI 模型的引入将为用户带来前所未有的便利性和准确性。例如，当一位旅行者计划访问一个新的目的地时，该系统可以通过分析社交媒体上的照片标签、旅游博客的文章、在线评论中的情感分析以及天气预报等多源数据，来识别出最受欢迎且适合当前季节的景点。此外，通过结合用户的过往旅行记录和个人兴趣点，系统可以提供定制化的行程安排建议，包括最佳交通方式、住宿选择以及活动预订等。这种高度个性化的服务不仅能节省用户的时间和精力，还可能激发他们探索未曾考虑的目的地的兴趣。为了确保推荐结果的相关性和质量，我们将特别关注数据的质量控制和算法的优化，以期建立一个既高效又可靠的智能旅游助手，帮助每一位旅行者发现世界的美好。

政务办公作为另一个重要的垂直领域，同样可以从多模态 AI 的应用中受益匪浅。政府部门每天需要处理大量的文件、报告、会议记录以及其他形式的信息。多模态 AI 可以帮助简化这些流程，提高工作效率和服务质量。例如，在政策制定过程中，AI 系统可以综合分析立法文献、公众意见调查、专家访谈录音等多种资料，辅助决策者形成更加科学合理的法规。同时，在日常办公中，智能文档管理系统可以通过自然语言处理技术和图像识别技术自动分类、归档文件，并支持快速检索，减少人工错误。对于公共服务而言，基于多模态 AI 的聊天机器人可以 24 小时不间断地回答市民咨询，处理简单事务，增强政府与民众之间的互动交流。总之，通过引入多模态 AI，我们可以构建一个更加透明、高效和人性化的政务服务环境，进一步推动智慧城市建设。

然而，要实现上述目标并非易事，多模态 AI 在实际应用中面临着诸多挑战。

首先是数据收集和标注的问题。为了训练出高性能的多模态模型，必须获取高质量、多样化的数据集，并对其进行精确标注，这需要投入大量的人力物力资源。其次是跨学科知识的融合难题，因为多模态 AI 涉及到计算机视觉、自然语言处理、机器学习等多个领域的专业知识。再者，如何保证模型的安全性和隐私保护也是一个不容忽视的问题，尤其是在处理敏感信息时。最后，随着技术的发展和社会对公平性的要求日益增高，我们必须不断改进算法，确保它们不会产生偏见或歧视。面对这些挑战，本项目将采取一系列措施，如加强数据管理、培养复合型人才、严格遵守法律法规等，力求打造一个安全可靠、开放包容的 AI 生态系统，使多模态 AI 真正造福社会各个层面。

面向垂直领域的 AI 多模态大模型实践与应用是一个充满机遇和挑战的研究课题。它不仅有助于推动旅游业和政务办公等行业的数字化和智能化进程，也为 AI 技术本身的发展提供了新的方向。未来，我们将继续深化研究，不断完善模型架构和技术框架，力求在更多实际场景中验证其价值。与此同时，我们也意识到，成功的背后离不开社会各界的支持与合作。因此，本项目还将积极寻求与学术界、产业界以及政府机构的合作机会，共同探讨多模态 AI 的最佳实践路径，努力构建一个更加美好的未来。通过各方共同努力，相信我们一定能够克服困难，开创出一片新的天地，让多模态 AI 成为连接人与世界的桥梁，助力各行各业实现更大的飞跃。

二、项目应用平台与基础

1．硬件平台

提供可供 7B、13B 参数大模型全量微调的 GPU 计算平台

2．软件平台

基于 Huggingface、PyTorch 等深度学习及大模型库进行编程

3．技术基础

本项目基于深度学习、强化学习、大语言模型、参数高效微调、检索增强生成相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++或者 Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．构建面向特定任务的大模型数据集：为了支持特定领域的高级应用，专注于创建和整理高质量、经过精心标注的数据集。这些数据集不仅涵盖了该领域内任务需求的广泛多样性和复杂性，还为后续的模型微调和优化工作奠定了坚实的基础。通过确保数据的全面性和准确性，可以更好地训练模型以理解和应对真实世界中的挑战，从而提升其在实际应用中的表现。

2．优化低秩适应（LoRA）及其他微调技术：针对现有大模型在有限实例数据条件下的高效微调需求，特别强调对低秩适应（LoRA）技术的改进。这些优化旨在使模型能够在少量样本的支持下快速学习新任务，并维持高水平的表现力。同时，我们致力于减少计算资源消耗和存储需求，使得即使是在资源受限环境中也能顺利实施微调过程，提高模型部署的灵活性和成本效益。

3．参数共享与增量学习微调框架：考虑到持续引入新任务数据时保持高效更新的需求，开发一个基于参数共享和增量学习的微调框架。此框架支持模型在面对新增任务时仅需调整部分参数而非重新训练整个网络结构，极大地提高了训练效率并节省了时间和资源。此外，它还促进了模型的知识积累，随着更多任务数据的加入，模型将变得更加智能且适应性强，为实现长期稳定的性能提供了坚实的基础。

4．模型开发与优化：构建能够处理多种类型输入（如文本、图像、音频）的多模态模型，实现信息的有效融合。对模型进行调优，提高预测精度和服务响应速度，基于用户行为和偏好构建智能推荐系统，提供定制化的服务建议。

5．检索与生成的优化算法：提出和实现新的检索增强生成算法，结合深度学习中的生成模型（如 GPT）和信息检索技术，改进检索机制，提高生成结果的

质量与效率。

项目 6：面向具身智能机器人动作生成（指导老师：凌贺飞）

一、项目背景

随着人工智能技术的发展，赋予机器人具身智能（Embodied Intelligence），即让其能够像人类一样理解并互动于物理世界，已成为一个重要的研究方向。为了推进这一目标，我们计划研究面向具身智能机器人的动作生成，即机器人能够在当前环境中，依据语言指令和实时观测图像来动态地生成相应的动作。

根据不同方法在训练过程中对数据的使用形式，现有具身智能动作生成方法主要可以分为三种范式：

1．强化学习（Reinforcement Learning，RL）：强化学习是实现机器人自主动作生成的技术之一。通过与环境的交互，机器人可以学习到如何通过试错不断优化自己的动作策略，达到预定目标。RL 能够处理高维数据，学习复杂的行为模式，适合于决策和控制。HDPG 提出了混合动态策略梯度，用于两足运动，使得控制策略可以同时由多个准则动态优化。DeepGait 是一种用于地形感知运动的神经网络策略，它结合了运动规划和强化学习方法。

2．模仿学习（Imitation Learning）：RL 的缺点是需要大量实验数据。为了解决这个问题，引入了模仿学习，其目的是通过收集高质量的演示来最小化数据使用。通过学习专家演示的动作来让机器人生成类似的动作。从人类演示中模仿学习最具有代表性的方法是 ALOHA 和 Mobile ALOHA。Mobile ALOHA 对于每个操作任务，使用平台采集的 50 条专家示范数据进行模仿学习，经过协同训练后任务成功率可达到 70%。

3．多模态动作生成学习(Vision-Language-Action Learning，VLA)：模仿学习局限于小规模数据和简单任务，在复杂任务中表现出狭窄的泛化和失败。因此，具身操作不能仅仅依赖视觉这单一感知通道和小规模数据模型，越来越多的研究集中在将预训练的视觉语言模型迁移到具身操作领域，实现多模态感知与动作生成的结合，即如何在大规模数据集上，结合来自视觉、触觉、文本等多方面的信息，生成更加精确和鲁棒的动作。

从早期使用自主数据收集进行缩放策略训练的工作，到最近探索将基于 transformer 的现代策略与大型演示数据集相结合的工作，许多工作使用从机器人收集的大型轨迹数据集来训练具身智能动作生成策略。现有工作并没有提供将

VLA 部署和适应新的机器人、环境和任务的最佳实践，特别是在商品硬件(例如,消费级 GPU)上。因此，实现具身智能动作生成的一个有希望的方向是通过利用大规模机器人数据集上预训练的基础模型，针对具体机器人操作任务进行模型定制化设计。然而这面临两个关键挑战，如何将定制化的机器人任务数据集适应到预训练基础模型的动作空间，如何处理不同的机器人感知、传感器设置、任务规范等；如何充分利用数据集信息进行有效且轻量化的动作预测。

本研究团队专注于具身智能机器人动作生成技术的前沿探索，旨在开发轻量化、高性能的视觉语言动作模型，以适应边缘计算设备的限制，并确保在真实世界应用中的高效部署。我们的工作主要围绕视觉语言动作模型建立、轻量化微调以及落地部署三个核心方面展开：

1．视觉语言动作模型建立。首先，我们致力于设计或选择一个适用于具身智能动作生成任务的基础模型，该模型将具备良好的表达能力和较低的计算复杂度。例如，RDT、openVLA 等，这些模型在大规模机器人动作数据集上进行训练，对于理解特定任务的动作至关重要。此阶段的工作将集中于优化模型架构，使其能够在保持高精度的同时减少参数数量和运算量。

2．轻量化微调设计。接下来，在基础模型之上，我们将引入一系列先进的轻量化技术进行微调。这包括但不限于：1）模型剪枝，通过去除冗余连接来简化网络结构；2）知识蒸馏，利用大型预训练模型指导小型模型的学习过程，从而传递关键特征表示；3）量化感知训练，允许模型在训练过程中适应低精度数值格式，如 8 位整数运算，以加速推理并节省存储空间。最终目标是获得一个紧凑而高效的视觉语言动作生成器，它可以在资源受限环境下运行。

3．落地部署。最后，为了证明所开发技术的实际价值，我们会将其部署到多种实际应用场景中测试性能，比如服务机器人的人机交互、智能家居的自动化控制等。在此过程中，我们将根据具体需求对模型做进一步定制化调整，确保其能应对不同环境下的挑战。

二、项目应用平台与基础

1．硬件平台

本项目基于较为成熟的硬件平台，以 mobile aloha 机器人、koch 系列机械

臂为主；在 Isaac Sim 和 Mujoco 等虚拟平台以及现实世界进行实验；另外会使用到激光雷达、三维雷达、工业相机、深度相机和各种算力计算平台。

2．软件平台

本项目基于 Ubuntu、 ROS、 PyTorch、TensorFlow 等平台进行编程。

3．技术基础

本项目基于深度学习、强化学习相关理论与技术，在项目研发过程中需要使用 Python、C++等编程语言。

三、项目要求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python、ROS 系统基础，了解机器学习相关内容；

3．具有一定的 PyTorch、深度学习、强化学习基础

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．基于模仿学习的机械臂动作生成算法优化

2．基于视觉语言动作模型的机械臂动作生成算法实现以及轻量化研究

3．机械臂操作任务仿真模拟环境搭建技术

4．现实世界机械臂操作任务专家演示收集

5．从虚拟环境到真实世界的迁移策略研究

项目 7：轻量级目标检测与识别技术（指导老师：凌贺飞）

一、项目背景

随着人工智能技术的快速发展，尤其是计算机视觉领域，目标检测和识别技术在众多应用场景中扮演着至关重要的角色。然而，在资源受限的环境中（如移动设备、嵌入式系统），传统的目标检测模型由于其计算复杂度高、参数量大等问题，难以直接部署。因此，研究轻量级目标检测与识别技术成为当前的一个重要课题。轻量级模型不仅需要保持较高的准确率，还需满足实时性要求，以适应各种边缘计算场景。

本研究小组致力于探索和发展适用于边缘设备的轻量级目标检测与识别技术，旨在降低模型的计算成本和存储需求，同时不牺牲检测精度。我们将从网络架构优化、量化方法、剪枝策略等多个角度出发，结合最新的深度学习研究成果，构建高效且紧凑的模型，实现快速、准确的目标检测和识别。

该构想分为三个阶段：

1．基础模型设计。首先，我们选择或设计一个具有代表性的轻量级网络结构作为基础模型，例如 MobileNet、ShuffleNet 等。这些模型通过特殊的卷积操作（如深度可分离卷积）来减少计算量，并保持较好的性能表现。此阶段将重点关注于如何平衡模型的精度与速度。

2．模型压缩与加速。在此基础上，进一步采用模型压缩技术，包括但不限于权重剪枝、低秩近似、知识蒸馏等手段，对基础模型进行瘦身。此外，利用量化技术将浮点运算转换为整数运算，以显著提升推理速度并降低功耗。这一阶段的目标是创建一个既小又快的检测器，适合部署在资源有限的硬件平台上。

3．实际应用部署。最后，为了验证提出的轻量级模型的有效性和实用性，将在多个真实世界的应用场景下测试模型的表现，如智能监控、无人驾驶、无人机视觉导航等。期间，还会考虑针对特定任务定制化调整模型，确保其能够满足不同领域的特殊需求。同时，探索如何将模型集成到现有的软件框架中，比如 TensorFlow Lite、ONNX Runtime 等，以便更方便地应用于各种终端设备。

二、项目应用平台与基础

1．硬件平台

基于 ARM 架构的处理器、GPU 加速卡、Raspberry Pi 等嵌入式系统。

2．软件平台

Linux 操作系统、PyTorch、OpenCV 等库用于模型训练与评估。

3．技术基础

本项目依赖于深度学习理论、神经网络架构设计、模型压缩与优化技术，在项目研发过程中需要使用 Python 编程语言，项目部署过程中需要使用 c++编程语言。

三、项目要求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python、基础；

3．具有一定的 PyTorch 基础，了解深度学习等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．基于 yolo 系统目标检测算法轻量化研究

2．基于 detr 系统目标检测算法轻量化研究

3．轻量化网络结构模型优化技术

4．基于 onnxruntime 工具的模型部署优化技术

5．基于 ncnn 工具的模型部署优化技术

项目 8：基于文本引导扩散模型的人脸对抗样本攻击（指导老师：凌贺飞）

一、项目背景

人脸识别技术在当今社会广泛应用，从安防监控到移动支付等领域都扮演着关键角色。然而，其安全性备受关注，对抗样本攻击成为威胁人脸识别系统稳定性与可靠性的重要因素。传统的对抗样本生成方法在攻击效果、隐蔽性及可迁移性上存在不足。

近年来，扩散模型在图像生成领域取得显著进展，能够生成高质量、多样化的图像。结合文本引导机制，扩散模型可根据文本描述生成相应图像。将文本引导扩散模型应用于人脸对抗样本攻击，有望打破传统方法的局限。通过精准的文本描述指导扩散模型生成针对人脸识别系统的对抗样本，不仅能提高攻击效果，还可增强攻击的隐蔽性与迁移性，为深入理解人脸识别系统的脆弱性及推动防御技术发展提供新途径。

本研究致力于探索文本引导扩散模型生成人脸对抗样本的方法，研究文本描述的设计以及扩散模型的优化，以实现对人脸识别系统更具有效性的攻击。

二、项目应用平台与基础

1．硬件平台

选用搭载 NVIDIA GeForce RTX 3090 的服务器，以满足扩散模型训练与对抗样本生成的高算力需求，以及大容量存储设备保存数据集与模型参数。

2．软件平台

基于 Ubuntu 操作系统，运用 PyTorch 深度学习框架进行模型开发与训练。结合 Diffusers 库调用和定制扩散模型，使用 OpenCV 库进行图像的预处理和后处理操作，利用自然语言处理工具包（如 NLTK、Hugging Face Transformers）处理文本信息。

3．技术基础

项目基于深度学习、计算机视觉、自然语言处理以及扩散模型相关理论与技术。深度学习用于构建和训练人脸识别模型及评估对抗样本攻击效果；计算机视觉技术负责图像采集、预处理与特征提取；自然语言处理技术实现文本描述的理解与生成；扩散模型则是本课题中生成对抗样本的核心技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．掌握扎实的 C++、Python 编程技能，能够独立完成复杂算法的代码实现，熟悉面向对象编程思想，保证项目软件部分的高效开发。

2．拥有一定的 PyTorch 基础，了解深度学习模型的构建、训练和优化流程；熟悉扩散模型原理与应用，掌握自然语言处理和计算机视觉相关技术，为项目研究提供技术支撑。

3．具备吃苦耐劳、踏实肯干的品质，能投入充足时间开展项目研究，面对研究中的困难与压力保持积极态度，确保项目按时推进。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．文本引导扩散模型的训练与优化：收集整理大量与攻击人脸识别系统相关的文本描述和对应图像数据，对基础扩散模型进行训练与微调。优化模型结构和超参数，提高模型根据文本生成对抗样本的准确性与效率。

2．文本描述设计与生成策略制定：研究如何设计有效的文本描述以引导扩散模型生成高质量对抗样本。制定生成策略，如结合人脸识别系统的漏洞信息、模型结构特点等生成针对性文本，提高对抗样本的攻击效果。

3．对抗样本的可迁移性研究与提升：探究生成的对抗样本在不同人脸识别系统和场景下的可迁移性。分析影响迁移性的因素，如模型结构差异、数据分布不同等，提出改进措施，提升对抗样本的通用性。

4．生成对抗网络与扩散模型的协同攻击：结合生成对抗网络与文本引导扩散模型，构建一种协同攻击架构。利用生成对抗网络生成具有潜在攻击能力的人脸图像作为初始样本，再通过文本引导扩散模型对这些样本进行进一步优化，以此提升攻击样本的视觉质量和可迁移性。

5．多模态融合的文本引导生成：将图像、语音等多模态信息融入文本引导过程。例如，结合人脸图像的特征向量和语音指令，生成更加精准、丰富的文本描述，从而指导扩散模型生成对抗样本。通过多模态融合，模型能够获取更全面的信息，生成的对抗样本可以针对不同模态感知的人脸识别系统进行攻击，增加

攻击的多样性和有效性

6．强化学习驱动的攻击策略探索：运用强化学习算法，让智能体在与不同人脸识别系统的交互过程中，探索最优的文本引导和对抗样本生成策略。智能体的动作空间包括文本描述的生成和对抗样本的调整操作，奖励函数基于攻击成功率、隐蔽性等指标设计。通过不断的试验和学习，智能体能够发现新颖且高效的攻击策略，进而提高生成的对抗样本的可迁移性。

项目 9：基于雷视融合的 3D 点云目标检测（指导老师：陈加忠）

一、项目背景

基于 LiDAR 的 3D 目标检测旨在使用从 LiDAR 捕获的点云来预测 3D 目标边界框。由于 LiDAR 提供了精确的在深度和几何信息方面，已有学者提出了大量的 3D 检测方法，并在各种基准测试中取得很好的性能。现有方法要么直接在点云上进行预测，要么将点云转换为体素。PointNet 是第一个以端到端方式处理点云的框架，它将无序点云集作为直接输入，并保留点云的空间结构。VoxelNet 将点云离散化为体素，并使用密集卷积以获得鸟瞰图（BEV）特征。然而，由于 LiDAR 的固有局限性，点云通常很稀疏，无法提供足够的上下文来识别远离传感器的区域，从而导致性能欠佳。

最近，激光雷达（LiDAR）和摄像头是两种广泛使用的传感器。基于 LiDAR 和相机融合 3D 目标检测取得了一些进展。早期实现激光雷达相机的方法通过将 LiDAR 点云或区域提议投影到 2D 图像上进行融合。但这些方法忽视两种模式之间的信息差距。最近的研究采用不同的查询生成策略或创建统一的 BEV 中间特征来融合多模态特征。例如，TransFusion 应用两阶段管道来融合相机和 LiDAR 特征，但其性能依赖于查询初始化策略。BEVFusion 探索了统一表示对于 BEV 特征，通过视图变换，既保留了稀疏 LiDAR 点云的空间信息，又将 2D 图像提升为 3D 特征，有效地保持了两种模态之间的一致性。然而，视觉模态仍然与 LiDAR 获取的几何感知信息之间的合作并不融洽，这限制了激光雷达和相机之间的互补性。

目前先进的多模态方法主要是进行全局融合，其中图像特征和点云特征在整个场景中进行融合。这种做法缺乏细粒度的区域级信息，产生次优的融合性能。为了处理来自不同模态的异构数据，现有方法通常预先定义与两种模态兼容的统一空间（即自我-车辆坐标系中的鸟瞰图 BEV），然后在该共享空间上执行特征对齐和融合。BEV 表示将复杂的 3D 空间简化为 2D 平面，从而更容易理解场景。然而，从整个 BEV 场景级别执行融合忽略了前景实例和背景区域之间的固有几何差异，这可能会破坏性能。例如，与在自然图像中观察到的目标实例相比，BEV 中表示的目标实例通常显示较小的大小。此外，前景实例占用的 BEV 网格单元数量显著低于背景实例占用的网格单元数量，导致前景和背景样本之间的严重不平衡。因此，已有方法很难捕获目标实例周围的局部上下文，或者在解码阶段主要依赖

于额外的网络来迭代地细化检测。虽然一些方法旨在执行目标级编码，但它们忽略了场景和实例特征之间的潜在协作。例如，通过与共享相似语义信息的实例的交互来增强其特征，可以潜在地纠正场景中的假阴性目标。因此，如何同时制定实例级和场景级上下文，以及通过利用多模态融合来精确地集成它们，仍然是一个开放的问题。

状态空间模型（SSM）以其良好的性能、线性复杂度和在语言和图像领域的长序列建模能力而备受关注。然而，由于 SSM 的因果关系要求以及点云的无序性和不规则性，将 SSM 扩展到点云场是不平凡的。已有研究表明基于 SSM 的点云处理主干网络，具有因果感知的排序机制，能够构造因果依赖关系，利用序列化使得无序点可以有序地表达，并保留它们的空间邻近性。在 ModelNet40 分类数据集和 ScanNet 语义分割数据集上，分别具有较好的分类和语义分割性能。然而，仍需更深入地研究是否存在更有效的点排序方法，及更大尺寸和参数的点 Mamba 拥有的表达能力。特别是，基于点 Mamba 架构的更具应用价值的大规模 3D 任务尚未得到研究，比如交通场景点云 3D 目标检测与语义分割、3D 点云场景的自然语言理解。

为了解决 3D 点云目标检测面临的挑战，提出了一种新的 3D 多模态目标检测方法。具体地，研制一种 LiDAR 引导模块，该模块由稀疏深度引导和 LiDAR 占用引导组成。前者将 LiDAR 点云生成的稀疏深度与相机特征相结合，生成深度感知特征，增强相机特征对深度信息的敏感性。受其他占用任务的启发，后者使用占用特征引导视图转换生成的 3D 特征体积，并聚焦 3D 特征体积中的目标，从而为融合提供更有价值的信息。然后，构造了一个多尺度双路径 Mamba 模块来改善物体周围的相互作用，并扩展三维特征体的感受野。通过以上设计，相机模态具有足够的语义特征和更准确的深度分布。为了在 LiDAR 模态中获得丰富的特征，我们对 LiDAR 点云执行 FPS 下采样，并使用稀疏深度压缩来聚合不同尺度的特征。该操作以较少的计算和内存消耗提供较大的感受野。此外，提出了一种 LiDAR 引导的自适应融合 Mamba 模块，用于有效融合由激光雷达点云和图像生成的 BEV 特征。在该模块中，LiDAR 特征自适应地引导相机 BEV 特征，以从全局范围加强跨模态交互。

为了为不同距离的目标提供更精细的区域级信息，并将原始位置信息保留在

更精细的粒度内，我们提出了具有位置信息编码能力的局部融合模块，以在每个方案的均匀划分网格中编码原始点云的位置信息，并在图像平面上投影网格中心以采样图像特征。然后，通过交叉注意模块融合采样图像特征和编码的局部网格特征。为了在每个方案的全局融合特征和局部融合 ROI 网格特征之间实现更多的信息交互，通过自关注提出了特征动态模块，以生成更多信息的多模态特征用于第二阶段细化。

二、项目应用平台与基础

1．硬件平台

基于具有 RTX 4060Ti 及以上 GPU 的机器，实验室提供部分算力。

2．软件平台

基于 Linux 或 Windows 的 TensorFlow、PyTorch 等深度学习及大模型库进行编程

3．技术基础

本项目基于深度学习、计算机视觉、Point Transformer、Point Mamba、雷视融合相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、点云分析与 3D 目标检测等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．面向 3D 目标检测的点云分析状态空间模型（SSM）：研究一种具有因果排序机制的基于 SSM 的点云处理骨干网。实现一种基于树结构的不规则点排序策略，对 z 阶序列中的点进行全局排序，保留它们的空间邻近性，提供全局建模能力。

构造因果依赖关系，同时显著降低计算成本。

2．基于多视角图像 BEV 的 3D 目标检测：研究增强组件来解决现有基于密集 BEV 的 3D 目标检测器的缺点，包括增强目标级一致性的 CRF 调制深度估计模块、具有扩展感受野的长期时间聚集模块以及结合感知技术和 CRF 调制深度嵌入的两级目标解码器。

3．基于多引导 LiDAR 和视觉特征融合的 3D 目标检测：研究基于 LIDAR 引导的全局交互和自适应融合的多模态三维目标检测方法。引入稀疏深度引导和 LiDAR 占用引导来生成具有足够深度信息的 3D 特征。开发 LIDAR 引导的自适应融合状态空间模型，以自适应序列化增强不同模态 BEV 特征的交互。

4．基于实例-场景写作融合的多模态 3D 目标检测：研究一种新的多模态融合框架，可以联合捕获实例和场景级别的上下文信息。不同于仅关注 BEV 场景级融合的现有方法，它通过显式合并实例级多模态信息，从而促进以实例为中心的 3D 对象检测任务，通过挖掘实例候选，探索它们的关系，并为每个实例聚合本地多模态上下文。

5．基于聚类 Mamba 的 3D 目标检测：研究采用 Cluster-Mamba 将每个对象视为一个 3D 空间聚类，聚类主要由属于同一目标的非空体素组成，并利用聚类进行 Mamba 序列化建模，直接从稀疏体素特征中生成提议框。基于聚类的 Mamba 结构通过利用聚类中包含的目标先验信息，有效地提高 3D 目标检测器的性能和收敛速度。

6．Graph 匹配多模态 3D 目标检测：研究一种更精确的特征对齐策略，用于通过图匹配检测 3D 目标，融合图像分支中语义分割编码器的图像特征和 LiDAR 分支中 3D 稀疏 CNN 的点云特征。通过计算划分为点云特征子空间内的曲面距离来构造最近邻关系，以节省计算开销。

项目 10：基于状态空间与多模态的 3D 点云补全（指导老师：陈加忠）

一、项目背景

点云补全作为 3D 视觉领域的一项基础性研究问题，其核心目标是从部分观测的点云数据重建完整的 3D 形状。在实际应用场景中，由于传感器固有的局限性，如视角限制、自遮挡、物体表面材料的反射率差异以及传感器分辨率等因素，采集到的原始点云数据往往存在大量缺失和不完整的问题。这种数据质量缺陷严重制约了点云在自动驾驶感知、机器人操作、增强现实(AR/VR)和 3D 场景重建等下游应用中的使用效果。因此，设计能够准确理解 3D 形状并有效恢复缺失几何信息的点云补全技术变得尤为重要。

随着深度学习技术的快速发展，点云补全研究已经从传统的基于几何先验和对称性假设的方法，逐渐发展到了基于深度神经网络的数据驱动方法。早期的 PCN(Point Completion Network)率先提出了一种端到端的编码器-解码器架构，通过将部分点云编码为全局特征向量，然后解码生成完整的点云。然而，由于编码阶段的最大池化操作不可避免地导致精细几何信息的丢失，解码器难以从压缩的全局特征中恢复丰富的细节。为了克服这一限制，TopNet 等工作提出了更加复杂的解码器结构，通过遵循分层根树结构来约束点云的生成过程。这些开创性工作虽然实现了端到端的点云补全，但在保持几何细节真实性和结构完整性方面仍有较大提升空间。

近年来，随着 Transformer 在计算机视觉领域的成功应用，研究人员开始探索将注意力机制引入点云补全任务。PoinTr 的提出标志着点云补全研究进入了新阶段。该工作首次将点云补全重新定义为集合到集合的转换问题，创新性地采用 Transformer 编码器-解码器架构。通过将点云表示为具有位置嵌入的无序点组集合，PoinTr 将输入转换为点代理序列，并设计了几何感知模块来显式建模局部几何关系。这种设计使得网络能够更好地学习结构知识并保留细节信息。在此基础上，AdaPoinTr 进一步改进了查询生成机制，并在补全过程中引入去噪任务，显著提升了补全效果并减少了训练时间。PointAttN 则完全依赖于交叉注意力和自注意力机制，摒弃了传统的 KNN 等显式局部区域划分方法，实现了对点云结构更灵活的建模。这类基于注意力的方法能够更好地捕捉点云中的长程上下文信息，但同时也带来了较大的计算开销。

为了更好地解决点云补全中的细节生成问题，研究者们提出了一系列创新的网络架构。SnowflakeNet 将完整点云的生成建模为 3D 空间中点的雪花状生长过程，通过引入跳跃式 Transformer 来学习最适合局部区域的点分裂模式，实现了从粗到精的渐进式生成。SeedFormer 创新性地引入了 Patch Seeds 形状表示，不仅能够从部分输入中捕捉一般结构，还能保留局部模式的区域信息。这些方法通过精心设计的层次化生成策略，在提升补全质量方面取得了显著进展。

多模态融合是近期点云补全研究的另一个重要方向。传统方法主要依赖单一的点云数据进行补全，而忽视了其他模态数据可能提供的有价值信息。BEVFusion 探索了统一的 BEV 特征表示，通过视图转换将 2D 图像信息提升为 3D 特征，有效地保持了不同模态之间的一致性。SVDFormer 则创新性地设计了自视图融合网络，利用多视角深度图像信息来观察不完整的自身形状并生成紧凑的全局形状表示。这些工作很好地展示了如何通过多模态信息的互补性来提升补全效果。

最新的研究进展引入了更多创新的技术范式。CP3 首次将预训练-提示-预测范式引入点云补全，通过引入不完整中的不完整(IOI)预训练任务来提高生成的鲁棒性，并设计了语义条件细化网络来实现更精确的细节恢复。ODGNet 通过引入正交字典学习来获取形状先验，有效补偿了缺失的几何信息。AGFA-Net 则通过设计空间和通道注意力块来替代传统的 KNN 操作，自适应地聚合全局特征。VAPCNet 提出了无监督的视角表示学习方案，避免了显式视角估计的复杂性。这些工作从不同的技术角度推动了点云补全技术的发展。

尽管已取得了显著进展，点云补全仍面临诸多挑战。首要问题是如何在保证全局形状合理性的同时重建精细的局部几何细节，这需要网络具备对整体形状的理解能力并能准确把握局部结构特征。传统方法往往将这两个层面割裂开来，要么通过编码器-解码器架构先获取全局特征再生成细节，要么通过局部特征聚合来重建完整形状。这种割裂的处理方式忽视了形状特征的层次性和连续性。针对此问题，提出设计"多尺度协同注意力机制"，该机制允许特征在不同尺度间自由流动。具体而言，可以构建一个层次化的特征金字塔，每个层级都维护全局和局部两个视角的特征表示。通过设计双向特征调制模块，使全局特征能够为局部细节重建提供语义指导，同时让局部特征的聚合反过来验证和细化全局假设。这种协同机制可以通过引入能量最小化框架来实现，将特征一致性作为优化目标之一。

在处理多模态信息方面，现有方法往往采用简单的特征拼接或注意力融合，这难以处理模态间的不一致性。为了进一步利用多模态信息，可以构建"跨模态一致性学习框架"。该框架不是直接融合不同模态的特征，而是先在一个共享的语义空间中建立模态间的对应关系。可以设计一个基于对比学习的预训练任务，让模型学会将不同模态的信息映射到一个语义一致的特征空间。这样，即使在某些模态缺失的情况下，模型也能够利用学习到的语义对应关系进行合理的补全。其次，在技术实现层面，计算效率与补全质量之间的权衡也是一个关键问题。最近的一些工作如 OctFormer 开始探索将无序点云转换为有序序列的新范式。不同于传统直接在无序点集上进行特征提取的方法，序列化方法首先通过特定的排序策略将点云转换为有序序列，在简化处理流程的同时尽可能保持空间结构信息。这类方法通过选择记录序列化映射关系，显著降低了计算开销。进一步探索 Mamba 此类线性复杂度的点云分析模型，通过不同的排序策略以及精心设计的局部特征提取组件，以进一步权衡计算效率与补全质量。

二、项目应用平台与基础

1．硬件平台

基于具有 RTX 4060Ti 及以上 GPU 的机器，实验室提供部分算力。

2．软件平台

基于 Linux 或 Windows 的 TensorFlow、PyTorch 等深度学习及大模型库进行编程

3．技术基础

本项目基于深度学习、计算机视觉、Point Transformer、多模态信息融合、状态空间模型相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1 吃苦耐劳，踏实肯干，有充分时间投入；

2 具有一定的 C++、Python 基础；

3 具有一定的 PyTorch 基础，了解深度学习、点云分析与点云补全等相关理

论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1. 基于状态空间模型的时空一致性点云补全：研究利用状态空间模型(SSM)来建模点云补全的时序和空间依赖关系。设计自适应时空序列化机制，将无序点云转换为具有因果依赖关系的序列表示；构建层次化 SSM 结构，在不同尺度捕获局部到全局的几何特征；引入空间注意力机制来增强特征之间的长程依赖。与现有工作相比，该方向更强调建模点云特征的时序演化过程。

2. 自监督动态字典学习的点云补全：研究一种自监督学习构建动态形状字典来指导补全过程。通过设计动态字典更新机制，自适应学习形状先验；构建基于注意力的字典查询模块，根据局部几何相似性检索相关形状模式；引入字典正交约束，确保字典元素的多样性。该方法能更好地适应不同类型的形状缺失。

3. 基于对比学习的语义感知点云补全：研究关注如何将语义信息引入点云补全过程。通过设计对比学习框架，学习形状的语义感知表示；构建语义引导的特征增强模块，根据语义相似性调制特征；引入语义一致性约束，确保补全结果与已知语义信息的一致性。这种方法能更好地保持物体的语义特征。

4. 自适应注意力引导的跨多模态点云补全：研究一种新的多模态信息点云补全网络，设计自适应特征融合模块，根据不同模态信息的可靠性动态调整融合权重；引入跨模态注意力机制，建立不同模态特征之间的对应关系；构建层次化特征对齐策略，实现从粗到细的渐进式补全。有效提升对具有噪声和遮挡的真实场景的处理能力。

5. 基于扩散模型与状态空间融合的多模态 3D 点云补全：研究将扩散概率模型（Diffusion Probabilistic Models）与状态空间模型（State Space Models, SSM）相结合，将扩散模型的生成能力与状态空间模型的时空依赖建模能力结合，逐步生成完整点云的同时保持时空一致性，采用层次化的扩散步骤，首先生成全局结构，再逐步细化局部几何细节，确保补全结果的整体性和精细度。

6. 基于多尺度 Transformer 预测与语义分割引导的点云补全网络：研究利用将 Transformer 用于缺失点云的推理预测，设计一个多尺度的 Transformer

编码器-解码器架构，分别在不同尺度上处理点云的全局和局部特征。通过层次化的特征提取，捕捉点云的多层次几何信息，确保补全过程中的细节和整体结构的平衡。进一步引入语义分割信息作为补全过程的引导。设计一个语义引导模块，将预先分割的语义标签与点云特征进行融合，确保补全结果在语义上与输入一致。例如，在补全建筑物点云时，能够准确区分并补全墙体、窗户等不同语义区域。

17. 图形与视觉计算团队

华中科技大学计算机学院图形与视觉计算团队，是图像处理和智能控制教育部重点实验室成员，在数字媒体方向有较为坚实的基础。

团队主要研究领域为**三维重建与建模、虚拟现实与增强现实、计算机图形学**。主持国家自然科学基金、湖北省自然科学基金、中央高校科研业务专项研究等纵向项目，并作为子课题负责人参与完成了多项国家重点研发计划、国家支撑计划以及省部级课题中的虚拟现实研究课题。同时完成了多项国防军工联合项目，并在近年来与腾讯游戏进行深度的图形图像领域的科研及人才培养等合作。在国内外权威刊物或会议上发表学术论文30余篇，论文已被SCI和EI收录20余篇。

团队教师曾多次荣获湖北省优秀学士学位论文指导教师奖、大学生科技创新活动优秀指导教师奖，华中科技大学教学质量奖，华中科技大学“优秀教师班主任”等荣誉。

团队所指导研究生学生曾获国家奖学金，三好学生等奖项；就业大多奔赴国内头部游戏公司，如腾讯游戏、网易游戏，以及拼多多、美团、小米、华为等国内外知名互联网公司。团队日常活动安排包括但不限于每周研究生学术研讨，羽毛球、乒乓球等文娱活动，聚餐等。

李　丹	团队负责人 副教授，主要研究领域为虚拟现实、计算机图形学等方面

团队联系方式：

联系邮箱：lidanhust@hust.edu.cn，李丹老师

地址：华中科技大学南一楼。

项目 1：面向虚拟试衣的服装模型重建方法研究

一、项目背景

虚拟试衣技术能够为用户提供便捷、直观的试衣体验，帮助消费者在购买服装前更好地了解服装的合身度和外观效果。传统的虚拟试衣方法主要依赖于二维图像处理技术，通过将服装图片叠加在人体模型上进行简单的变形和渲染，但由于缺乏对服装三维形态的准确建模，容易出现服装与人体不贴合、试衣效果不自然等问题，影响试衣的真实感和可信度。

随着三维重建和计算机视觉技术的发展，基于三维模型的虚拟试衣方法逐渐受到关注。服装模型重建是实现高质量虚拟试衣的关键环节之一，其目标是从二维服装图像或三维扫描数据中重建出准确、详细的服装三维模型。然而，服装模型重建面临着诸多挑战，其几何形态复杂多变，易受人体姿态、动作及材质影响，需要考虑服装与人体的复杂贴合关系，兼顾人体动态变化和服装物理特性，以实现自然贴合效果。因此，研究面向虚拟试衣的服装模型重建方法，为虚拟试衣技术提供更加准确和真实的三维服装模型，具有重要的应用价值和市场前景。

二、项目应用平台与基础

1．硬件平台

基于具有 GPU 足够算力的服务器

2．软件平台

基于 Python 编程语言和 PyTorch 深度学习框架进行编程

3．技术基础

本项目基于计算机视觉、深度学习和计算机图形学相关理论与技术。

三、项目需求

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础。

四、项目开展

设计或优化现有的服装模型重建算法，提升算法速率或细节表现，重建的模

型应能正确反映服装真实的颜色、样式、尺寸等信息，能够自然地穿着于人体模型表面，实现虚拟试衣功能。

项目 2：面向三维打印件的偏差检测系统的设计与实现

一、项目背景

3D 建筑打印作为一种新的智能建造技术，具有废物产生量低、可持续性、施工工期短、施工成本低和工人安全等优点，目前正逐渐引起广泛关注。3D 建筑打印技术在打印时使用包含多种外加剂的专门混合物，以确保每一层硬化得更快，从而保持足够的粘性，并且不会凝固得太快而无法充分粘附到下一层。该技术在打印材料或墙壁时能够减少几何方面的限制，从而增强了建筑设计的自由度。

尽管 3D 建筑打印有很多优势，但仍存在一些挑战。3D 建筑打印的质量会受到环境重力、材料变化和打印系统稳定性等诸多因素的影响，传统干预方法是需要人工操作人员在打印对象出现偏差或失败时对打印系统进行干预，效率低下，精度不高，难以满足 3D 建筑打印闭环控制的需求。因此，具备自动化、智能化优点的计算机视觉技术应用到 3D 建筑打印几何质量检测具有重要的应用价值。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器，深度相机

2．软件平台

基于 C++，OpenCV 等图像工具包进行编程

3．技术基础

本项目基于计算机视觉、图像处理相关理论与技术。

三、项目需求

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

四、项目开展

构建深度相机视觉平台，根据采集的深度图数据和彩色图数据，进行数据处理实现三维打印件的形貌测量算法，设计偏差检测算法来检测打印模型和标准模型的偏差。

项目 3：基于 3D 高斯的三维人体重建方法研究

一、项目背景

三维人体重建技术作为一种先进的数字化建模技术，在虚拟试衣、远程医疗、影视制作等领域具有广泛的应用前景。传统的三维人体重建方法主要依赖多视角图像或深度传感器数据，这些方法虽然能够获得较为精确的三维模型，但通常需要复杂的设备和较长的重建时间，限制了其在实时应用中的可行性。近年随着深度学习技术的迅猛发展，基于单目图像的三维人体重建方法逐渐受到关注。然而，单目图像重建面临着深度信息缺失问题，导致重建结果往往存在较大的几何误差和细节缺失，难以满足高质量重建的需求。

3D 高斯（3D Gaussian Splatting）作为一种新兴的三维表示方法，通过将场景建模为一组离散的三维高斯分布，能够有效地捕捉场景的几何结构和细节特征。其在处理动态场景和实时渲染方面展现出独特的优势，为三维人体重建提供了新的思路。基于 3D 高斯的三维人体重建方法，旨在利用单目视频数据，快速、准确地重建出具有丰富几何细节的三维人体模型。该方法在处理人体姿态变化、服装细节以及复杂光照条件下的重建问题时，能够更好地适应人体的动态特性和视觉表现，为实现高质量的三维人体重建提供了可能，具有重要的应用价值。

二、项目应用平台与基础

1．硬件平台

基于具有足够 GPU 算力的服务器

2．软件平台

基于 Python 编程语言和 PyTorch 深度学习框架进行编程

3．技术基础

本项目基于计算机视觉、深度学习和计算机图形学相关理论与技术。

三、项目需求

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础。

四、项目开展

设计或优化现有的三维人体重建算法，提升重建精度，优化模型在细节、体型和姿态等方面的表现，实现通过少量图像序列或低成本传感器快速获取用户体型数据，并生成精确的三维人体模型。

18．智能媒体计算与网络安全团队

华中科技大学“智能媒体计算与网络安全团队”依托计算机科学与技术学院计算机应用湖北省重点学科和网络空间安全一级学科、下一代互联网接入国家工程实验室、中国教育科研网华中地区网络中心和华中科技大学网络与计算中心，拥有开放自由的学术氛围和国际前沿的研究方向。

团队现有教授 2 人，副教授 2 人，在读全日制博士、硕士研究生 50 余人，团队承担了国家重点研发计划 5 项、国家自然科学基金 7 项、国防预研基金 5 项、国家 863 重点项目子课题 1 项、国家科技支撑计划 1 项、国防预研计划 1 项、教育部博士点基金 1 项、湖北省杰出青年基金项目 1 项以及各类企业合作项目 50 余项；申请发明专利 39 项，其中授权 29 项；获得软件著作权 28 项；获省科技进步二等奖 2 项、武汉市科技进步奖 1 项、日内瓦国家发明奖 1 项。

目前团队研究领域主要围绕下列五个方面：

运动视频分析：针对体育运动视频尤其是足球运动的智能分析，构建了海量的细粒度足球视频标注数据集，囊括了足球视频中的绝大部分视觉任务；利用球员视觉跟踪、比赛精彩程度分析、运动事件检测、球场三维重建等深度学习运动分析方法，对体育比赛进行智能分析，为行业提供了专业可视化的动态比赛数据；

三维视觉：旨在对三维场景的理解、重建以及编辑开展研究，研究内容包括：基于多视角立体匹配（MVS）、神经辐射场（NeRF）和 3D 高斯（3DGS）的三维场景重建方法，主要包含重建速度和重建质量改进；实现物体材质以及环境光照估计；动态三维场景的重建；场景的编辑和风格变换；三维表征方法与传统渲染管线的结合等；

数字孪生与建模仿真：旨在为军事/军工/泛军工领域研训人员研发易用、高效、可信的虚实孪生联合战场底座，研究内容包括：多维战场环境数值模拟、大空间多物理辐射场高效解算、作战装备孪生体生成、大规模多智能体泛化智能决策以及多场景并行的分层式容器化集群引擎研究；

生成模型：旨在对生成模型（如扩散模型和自回归模型）的原理以及应用进行研究。研究内容包括：扩散模型的数学原理及改进；文本引导的图片生成；人体动作生成以及跨模态增强；三维几何模型的生成，包括以 Field 为表征和 Mesh 表征的几何生成方法；

医学图像处理：旨在对基于医学影像的多项关键技术开展研究，包括：图像分割，研究使用生成式模型对脑部 MRI 图像实现结构分割；图像配准，研究基于深度学习的 MRI 或 CT 图像配准技术；图像关键点检测，研究针对脑部 MRI 图像的关键点检测算法；大语言模型，研究设计面向多类型医学图像的大语言模型；

网络空间安全：旨在对网络虚拟化技术以及多模态下的网络舆情分析技术开展研究。网络虚拟化技术包括真实源地址验证、网络拓扑保护、P4 可编程交换机；多模态下的网络舆情分析技术包括网络流媒体视频异常事件检测、社交网络图片篡改分析、虚假新闻视频舆情传播分析、舆情分析模型鲁棒性检测。

围绕以上研究方向，团队发表论文 250 余篇，其中发表在 IEEE/ACM trans 等重要刊物和 CCF-A 类顶级会议上的论文 50 余篇。更多相关信息可在团队主页 http://media.hust.edu.cn 上获取。

团队培养的毕业生获得外界的一致认可，毕业生去向包括国家机构和国有企业，如江西省人民政府、国家电网等；IT 龙头企业，如华为、腾讯、字节、阿里、百度等；以及各大高校，如浙江大学、华中科技大学、四川大学等。

团队成员：

<table>
<tr><td>于俊清</td><td colspan="3">团队负责人
教授，博导，主要研究领域为计算机网络、视频智能分析与搜索、多核计算与流编译、教育信息化等</td></tr>
<tr><td>管　涛</td><td>教授，博导，主要研究领域为数字孪生与建模仿真、三维渲染引擎、基于图像的三维建模以及增强与虚拟现实</td><td>杨　卫</td><td>副教授，博导，主要研究领域为计算摄影学，三维视觉，人体动作捕捉，神经表征和人工智能</td></tr>
<tr><td>何云峰</td><td>副教授，主要研究领域为基于内容的视频信息处理与检索以及高维数据索引研究</td><td>王跃嵩</td><td>未定级教师，主要研究领域为计算机视觉和三维重建</td></tr>
<tr><td>宋子恺</td><td>博士后，主要研究领域为计算机视觉，运动估计和社会网络分析</td><td></td><td></td></tr>
</table>

团队联系方式：

联系邮箱：weiyangcs@hust.edu.cn，杨卫老师

地址：华中科技大学南六楼 306 房间。

项目 1：基于 3D 高斯泼溅技术的室内场景重建以及编辑

一、场景描述

在增强现实（AR）和虚拟现实（VR）等领域，高质量的室内场景能够显著增强观众的沉浸感和代入感。传统的三维建模过程通常需要复杂的处理流程，需要设计师花费大量时间进行绘制草图、搭建结构、创建纹理等步骤，成本高昂且效率低下。相比之下，基于图像的三维场景重建技术以其成本低、自动化程度高、模型逼真等优势，近年来受到广泛关注。然而，尽管该领域研究已经取得了一些重要突破，现有算法在应对复杂多变的室内场景时仍然表现不佳，例如对光照变化、遮挡和纹理信息不足等常见室内场景问题的适应性不足。

针对上述问题，本项目旨在利用近年来备受关注的高斯泼溅（3D Gaussian Splatting，简称 3DGS）技术，探索实现高效且高质量的三维场景重建方法。我们期望通过改进该技术，提升室内场景的重建效果，为复杂室内场景的三维重建提供可靠的技术支持。

另一方面，三维场景的语义理解和编辑正逐渐成为研究热点。语义理解旨在基于自然语言或其他提示信息，精确定位目标对象在三维空间中的区域。而三维场景编辑包括风格化编辑和物理操作编辑两大方向。风格化编辑主要根据指定的文本或图像提示，改变场景的视觉外观；物理操作编辑则着重于场景对象的空间调整和几何变换。这些编辑能力可以帮助人类用户通过直观的自然语言交互，轻松实现对三维场景的定制化改造。

随着 CLIP、Diffusion 等基础模型的不断发展与改进，针对二维图像的语义理解与生成技术已经取得了显著进展。然而，由于三维数据训练需求的高成本与高算力限制，专门针对三维数据的大规模基础模型尚未成熟。因此，如何有效利用现有二维基础模型的能力，实现三维场景的语义理解与编辑，已成为当前研究的核心问题之一，并展现出重大的实际应用价值。

综上所述，本项目研究可分为四个方面：

1.解决稀疏重建分块问题：针对室内场景下，Structure-from-Motion (SfM) 过程中常见的分块问题，拟采用 MASt3R-SfM 技术对分块的小场景独立重建，首先求解块内相机位姿，然后利用 3D 高斯泼溅技术针对该分块进行高效的三维场景重建，并生成虚拟视角的渲染结果。接着，通过虚拟视角渲染的特征匹配，将

分块的小场景注册到整体的大场景中，从而有效提升全局场景的相机注册率和稀疏点云的一致性，为复杂室内场景的完整重建奠定基础。

2．提升室内场景重建质量：为改善室内场景重建质量低下的问题，拟引入几何先验对重建过程进行约束，包括利用单目深度估计大模型或 MASt3R 的预测结果对重建进行约束和指导。通过几何先验的引导，解决室内场景中的光照变化、遮挡以及纹理不足等问题，增强复杂室内场景的重建效果。

3．实现高质量的三维场景理解：在三维场景的语义理解方面，拟基于 SAM 等基础模型的分割结果，通过空间聚类等算法解决多视图分割结果不一致的问题，提升三维空间实例分割的精度与质量。同时，结合 CLIP 等语义大模型，从二维图像中提取语义信息并映射到三维空间中，实现高质量的三维场景语义理解，以提供语义化交互与场景编辑的支持。

4．实现高质量的三维场景编辑：为满足三维场景编辑的多样化需求，拟研究基于二维生成模型的三维场景编辑方法。结合 Diffusion 等生成模型，从二维图像生成目标样式或外观，并将其映射到三维场景中，通过引入多视图一致性约束，确保编辑过程中三维场景的几何和纹理一致性，以完成高质量的风格化编辑与物理操作编辑，赋予用户对三维场景直观、高效的编辑能力。

本项目将围绕上述问题展开深入研究，为三维场景的高效重建、智能化语义理解与直观交互编辑提供创新性解决方案。

二、项目应用平台与基础

1．硬件平台

具有足够算力的服务器

2．软件平台

Ubuntu 系统、PyTorch 等深度学习及大模型库

3．技术基础

多视图立体几何、深度学习、基础模型、参数高效微调。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入

2．熟练使用 C++、Python

3．会用 PyTorch、CUDA，了解三维视觉、深度学习、大模型技术

四、项目开展

1.室内场景数据集构建：收集和整理高质量的室内场景数据集，确保数据集覆盖任务需求的多样性与复杂性，为实验提供坚实的数据基础。

2．实验环境搭建：设搭建适用于上述任务的软件环境。

3．提升相机注册率和解算精度：基于上述研究路线，提升室内场景下，SfM 的相机注册率以及相机解算精度。

4．提升室内场景的重建效果：基于上述研究路线，利用几何约束，提升室内场景的三维重建效果。

5．实现室内场景下的语义理解和编辑：基于上述研究路线，设计算法，实现室内场景下的三维场景语义理解以及编辑效果。

19．认知计算与智能信息处理团队

团队依托于计算机学院，拥有自由开放的学术氛围，探索国际前沿的研究方向。团队主要从事认知科学与智能信息处理的交叉与结合研究，包括基于人工智能技术的自然语言处理与图像处理的理论与方法，算法设计与复杂性分析、形式化方法与具体应用研究等；主要研究方向包括信息检索、文本数据挖掘、自然语言处理、图像处理、机器学习、人工智能、推荐系统及社交媒体分析等。

团队现有教授 2 名，副教授 2 名，讲师 1 名，70 余名硕博士研究生。团队构成具有计算机科学、语言学、认知科学等多学科交叉、多层次系统、产学研结合典型特色，具备实力雄厚的师资力量、充满活力的科研团队以及良好的硬件设施环境。团队参与并承担多项国家科研攻关项目，包括国家自然科学重点/专项基金、国家 863 计划、科技部火炬计划/重大专项、国家工信部电子发展基金、国家信息产业部电子发展基金、国家“十二五”/“十三五”预研项目、国防纵向项目、湖北省科技重大专项、千万级校企联合实验室等多个项目，建立了华科-平安产险人工智能等多家校企联合实验室。团队目前已出版学术专著和教材 15 本，发表国内外高水平学术期刊及国际学术会议论文 250 余篇，获得国家发明专利 30 余项，获得国家软件著作版权 40 余项；获得国家技术发明二等奖 1 项、国家教委科技进步二等奖 1 项、国家教委科技进步三等奖 1 项、湖北省科技进步一等奖 1 项/二等奖 2 项，武汉市科技进步一等奖 1 项；湖北省教学成果一等奖 1 项、湖北省多媒体课件二等奖 1 项、校教学成果一等奖 1 项、校教学质量一等奖多项等。

目前研究领域主要包括下列三个方面：

自然语言处理与智能计算：聚焦于研究使计算机能够高效处理和理解人类自然语言的基础理论及分析方法，重点关注大模型的对齐机制、微调方法及推理过程中的优化策略等关键问题，旨在解决模型输出与用户预期不一致，模型参数量大，复杂任务中的长文本理解和生成，跨领域推理能力不足等难题。具体研究方向包括：自然语言处理基础理论（如分词、句法分析、语义理解、篇章分析、指代消歧等）、机器翻译、信息抽取与过滤、信息检索、知识表示与推理、多源异质文本数据分析与挖掘、大模型对齐、大模型推理、大模型高效微调技术、知识增强大模型等。

多模态融合与智能感知：聚焦于研究多模态信息的智能感知与认知理论方法，重点关注多模态大模型在跨模信息对齐、人类偏好对齐及跨模态数据一致性建模等方面的关键问题，旨在实现对复杂对齐目标的理解（如指令、偏好和伦理规范等）与高质量的响应生成。针对数据模态异质性，对齐目标复杂性，数据偏见及有害信息的存在，数据分布不均衡性等特点，具体研究方向包括：多模信息表征与理解、跨媒体关联理解、多模/跨模感知融合、多模态大模型对齐、文生图对齐、多模态大模型去幻等。

大数据处理与智能推荐：聚焦于研究将现实社会应用场景与虚拟互联网进行深度融合，以实现认知层面的人机交互的方法，重点关注大数据技术在实际应用中的创新、智能系统对复杂用户需求的精准理解与响应、多视角跨模态的个性化推荐等关键问题，旨在开发高效的分布式计算模型，解决大规模数据处理中的性能瓶颈，特别针对图数据、跨模态数据和时间序列数据等复杂数据类型。具体研究方向包括：大数据分布式计算模型和框架、大数据知识表示与推理、智能问答、智能推荐等。

团队的研究领域全面覆盖自然语言处理与智能计算的核心技术，涵盖从基础理论到应用技术的全方位前沿性探索。团队已在信息检索、自然语言处理、文本数据挖掘、社交媒体分析与计算、人工智能等领域的权威期刊（IEEE TKDE、ACM TOIS、IEEE T-Cyber 等）和人工智能领域与数据挖掘顶级会议（ACL、AAAI、IJCAI、SIGIR、SIGKDD、WWW、ACM MM、CVPR、ICDE、EMNLP、CIKM）等国内外高水平会议和期刊上发表论文 250 余篇，其中 CCEA/CCFB 及 IEEE/ACM Trans 论文 100+篇。团队相关研究成果应用于平安产险、蚂蚁科技、国家电网、南方电网等多家企业。团队学生曾荣获 2023 年 NLPCC 大赛“对话方面级情感分析”赛道全球冠军，ACL-SIGHAN 2024 中文方面级情感计算全球亚军，2023 年第十八届“挑战杯”揭榜挂帅“智能数字人（数字分身）生成技术研究”赛道全国一等奖等。团队毕业生凭借高素质、扎实的专业技能和创新思维赢得了企业的一致认可。每年华为、腾讯、阿里巴巴、百度、字节跳动等顶尖 IT 企业为团队毕业生提供了丰厚的薪资待遇，充分证明了团队教育成果的高价值以及对行业需求的精准契合。

团队成员：

魏　巍	团队负责人 教授，主要研究领域为人工智能，自然语言处理，信息检索与推荐，多模计算，数据挖掘(文本挖掘/社交媒体分析与挖掘)，机器学习		
向　文	教授，主要研究领域为工业互联网及安全，嵌入式系统及应用，自然语言处理（知识图谱、语义理解、自动问答）	卢　萍	副教授，主要研究领域为信息存储理论与技术，大数据分析与图像处理
李　开	副教授，主要研究领域为嵌入式系统，硬件加速，边缘计算	江　胜	讲师，主要研究领域为软件工程、大数据处理、知识图谱

团队联系方式：

联系邮箱：weiw@hust.edu.cn，魏巍老师

地址：华中科技大学东一楼专利中心 343 房间。

项目 1：迈向通用具身智能系统

一、项目背景

随着人工智能技术的飞速发展，学术界与工业界正在探索如何让人工智能具备更高水平的认知、学习、感知和决策能力。传统的人工智能依然局限于静态的数据处理或单一任务的自动化执行，缺乏对真实物理世界的全面理解和自适应能力。为了实现具备通用智能的人工智能系统，必须让 AI 不仅拥有“思考”的能力，还能通过身体与环境进行实时互动与自我学习，从而使其具备更为灵活和全面的感知、决策与执行能力。

本项目的目标是构建一个具备通用具身智能的机械臂系统，使其能够自主感知环境、理解语音指令，并完成抓取、搬运等复杂任务。系统将结合深度学习、计算机视觉、语音识别和机械臂控制技术，让机械臂具备更高的自主性和灵活性，能够在各种实际应用中进行任务执行。

该构想分为以下三个阶段：

第一阶段：基础环境感功能实

在这一阶段，我们的目标是让机械臂能够理解人类的语音指令，并拥有基本的环境感知能力。通过机械臂配备的视觉系统（如 RGB 摄像头、深度摄像头等）和音频传感器，将环境信息输入给机械臂，通过基础的深度学习模型完成语音识别，目标检测等基础感知功能。

第二阶段：简单环境下的任务实现

在完成了第一阶段的基础任务后，学会利用基础的方法，搭建一个完整的控制系统来完成简单任务。例如可以在仿真平台上，通过集成现有的强化学习算法和传统的控制方法，实现简单的机械臂抓取任务。与此同时，，通过将语音识别模块与机械臂控制系统结合，使得机械臂能够根据人类的语音指令执行特定任务。例如，当用户说出“抓取红色杯子”时，系统能够通过语音识别提取目标信息，并通过计算机视觉识别物体，最终完成抓取动作。此外，还可以结合触觉反馈，确保机械臂在抓取物品时不造成损坏或过度的用力。

通过这一阶段的工作，机械臂将具备初步的任务执行能力，能够根据环境变化和人类指令，完成简单的抓取和搬运任务。

第三阶段：全自主决策具身智能实现

在这一阶段，机械臂将具备全面的自主决策和学习能力，能够根据任务需求、环境变化以及与人类的协作需求，独立做出决策并执行复杂任务。本阶段的核心目标是通过整合感知、任务规划和决策过程，实现机械臂的高度自主性。为了实现这一目标，我们将学习如何利用大模型，将感知信息、任务规划与决策过程有机地统筹起来，确保机械臂在多变的环境中能够高效且智能地完成任务。到此，我们构建出一个全生命周期内可以基本完成自己决策和行为的具身智能体。

二、项目应用平台与基础

1．硬件平台

本项目将提供较为成熟的算力平台作为支撑，包括多卡 4090 等算力服务器。

2．软件平台

本项目基于 PyTorch 等平台进行编程。

3．技术基础

本项目基于深度学习、强化学习相关理论与技术，在项目研发过程中需要使用 Python、C/C++等编程语言。

三、项目要求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的深度学习基础，对 python 等编程语言较为熟悉；

3．具有一定的强化学习基础

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．基于音频传感器和深度学习模型的语音识别模块构建：设计并训练融合音频-自然语言双模态的深度学习模型，并依托此模型构建一个能够精确接收和识别音频的语音识别模块。

2．基于视觉系统和多模态大模型的环境感知模块构建：利用多模态大模型技术，选择合适的多模态大模型，并在机械臂模拟环境下完成微调，构建具备对

模拟环境中物体的识别与定位能力的环境感知模块。

3．基于仿真平台的仿真机械臂基础控制系统实现：在仿真平台上，集成强化学习算法和传统控制理论，构建基础的机械臂控制系统，完成物体抓取任务。

4．基于已有模块的智能仿真机械臂系统实现：实现语音识别模块和环境感知模块至仿真平台的迁移，并与仿真机械臂控制系统进行集成，打造一个具备高级环境感知能力和指令理解智能的机械臂系统。

5．基于大语言模型的具身智能体构建：利用大语言模型和现有模块，构建一个具备任务规划、环境分析及行为反思等复杂思维能力的高阶具身智能体。

20．EDA 与工业优化团队

计算机学院“EDA 与工业优化团队”是国内最早从事 NP 难问题算法研究的团队之一，在成立至今 40 余年一直专注于 NP 难问题的求解算法研究及其工业落地应用。团队聚焦高端工业设计与工业制造领域的关键工业软件核心优化算法研发，与华为、蚂蚁、海思、中兴、中船、嘉立创等头部企业在 EDA、先进制造、云计算、通信等领域开展广泛合作，建有华中科技大学计算机学院—海思半导体有限公司 EDA 联合研究中心、华中科技大学—弥费科技先进制造人工智能算法研究中心等。团队参加人工智能、EDA 等领域的国际算法竞赛获得 10 余项全球冠军，全球前三名 30 余项，部分竞赛成果获得央视新闻、新华社、新闻联播、央视元旦晚会、人民日报、文汇报等主流媒体的关注和报道。团队现有教授 2 人，副教授 4 人，讲师 1 人，博士后 1 人，同时与美国、加拿大、英国、法国、德国等国家的著名学者（如禁忌算法提出者、冯•诺依曼理论奖的获得者、美国工程院院士 Fred Glover 教授）开展了深入合作。

目前主要研究领域主要包括：

经典组合优化问题求解和黑盒优化：研究逻辑、图论、计算几何中的经典 NP 难组合优化问题，提出高效智能优化算法对其进行求解，刷新当前国际最好结果；研究基于黑盒优化的通用模型与求解算法及其在模拟电路调参等场景中的应用，设计高效的启发式算法、贝叶斯优化算法等对其进行求解。

芯片设计自动化关键工业软件：研究芯片设计流程中的逻辑综合、网表划分、版图规划、全局布局、详细布局、线网布线、时钟树综合、原理图画图等多个环节的 NP 难组合优化问题求解，将网表（拓扑图）映射到三维空间（几何图）中，最小化芯片的面积、时延、功耗（PPA）等关键性能指标。覆盖数字芯片、FPGA、模拟芯片、量子芯片、先进封装、PCB 等多种场景。

芯片仿真验证关键工业软件：研究芯片设计与制造环节中的缺陷检测自动化，包括设计阶段的等价验证、硬件加速仿真，以及制造阶段的自动化测试向量生成（ATPG）等。其中，等价验证与 ATPG 可归约为经典的 NP 完全问题—布尔表达式可满足性问题（SAT）；硬件加速仿真旨在将复杂的逻辑函数映射到超大规模处理器阵列或 FPGA 上，通过合理的任务分配与调度最大化仿真系统的时钟频率。

先进制造领域关键工业软件：研究先进制造中的生产排程、任务调度、路径

规划、装箱、切割、下料等多个环节中的NP难组合优化问题求解，为企业降本增效；为智能工厂与智慧仓储等场景提供多智能体路径规划（MAPF）、AGV/叉车/机器人/天车的任务分配、路径规划与交通拥堵控制等核心算法内核，提供多智能体调度系统的仿真与优化一体化解决方案。

团队以智能优化算法求解复杂工业优化问题为研究特色，致力于培养掌握复杂系统建模和核心算法设计能力的复合型人才。注重算法设计与实践能力培养，为学生提供系统的算法能力培养流程：从重现经典问题的经典方法，到使用经典方法求解其他问题，再到改进经典方法，最后针对实际应用问题设计新算法；从编程训练到学术竞赛再到各类工业应用项目，帮助学生循序渐进地完成由易到难、从理论到实践的进阶。

团队近年来的代表性算法竞赛获奖包括2017年SAT竞赛全球冠军、ICCAD 2021布局布线全局优化算法竞赛全球冠军、GECCO 2020-2021相机布局与集合覆盖算法竞赛连续两届蝉联全球冠军、GECCO 2022与2024堆场调度算法竞赛两届全球冠军、CG:SHOP 2023凸多边形覆盖竞赛青年组全球冠军、DIMACS 2022车辆路由算法竞赛两项全球冠军、2023集成电路EDA设计精英挑战赛FPGA Die级系统布线算法设计赛道全国一等奖、GECCO 2024多模态优化的小生境方法竞赛全球冠军、2024中国研究生创“芯”大赛·EDA精英挑战赛FPGA面向布线优化的详细布局赛道全国一等奖等。相关高水平成果经整理形成学术论文，在人工智能会议和期刊（如AAAI、IJCAI、AI）上发表论文200余篇。

团队毕业生质量获得企业一致认可，在业界具有良好口碑。毕业生主要加入华为、字节跳动、拼多多、美团、小红书、腾讯、阿里巴巴、百度、中兴、中船等一流IT企业。

团队成员：

吕志鹏	团队负责人 教授，主要研究EDA等关键工业软件算法内核、NP难问题		
丁俊文	副教授,主要研究任务调度、路径规划、智能优化算法	苏宙行	副教授,主要研究复杂系统建模、EDA软件关键算法
石　柯	教授，主要研究物联网、工业大数据、人工智能	许贵平	副教授，主要研究数据库与大数据分析

黄　志	讲师,主要研究 NP 难问题高效算法、运筹学	张庆雲	博士后,主要研究 NP 难问题求解、启发式优化算法

团队联系方式：

联系邮箱：suzhouxing@hust.edu.cn，苏宙行老师

地址：华中科技大学南一楼 608 室。

项目 1：多智能体路径规划

一、项目背景

多智能体路径规划（Multi-Agent Path Finding，MAPF）是指在一个共享的封闭环境中，为多个智能体制定行驶路径，以使其从各自的起点顺利到达目标位置，同时避免时空碰撞。图 1 给出了一个代表性的小规模多智能体路径规划的示例，3 个智能体沿着各自的路径每个时间步移动一格到达目标点，并在任意时刻都没有产生碰撞。智能体是具有自主行为的实体，通常是机器人、无人机、自动驾驶车辆、视频游戏中的虚拟角色等。它们在共享环境中执行任务，并根据自己的感知、目标和策略来做出决策。智能体的数量通常超过一个，它们需要在物理空间中协同工作或并行执行任务。MAPF 不仅仅是单一智能体路径规划问题的扩展，更是在多个智能体之间进行有效协调的复杂优化问题。随着智能体数量的增加，问题的规模和复杂性迅速上升，因此，MAPF 在理论研究和实际应用中都面临着许多挑战。

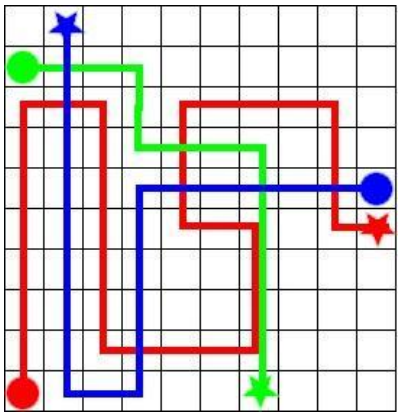


图 1 多智能体路径规划示例

多智能体路径规划问题在许多实际领域都有广泛应用，以下是几个典型的应用场景：

1．仓储物流与机器人调度：在仓储自动化系统中，多个机器人需要协同工作来搬运货物。MAPF 可以帮助设计每个机器人从起点到目标的路径，避免机器人之间发生碰撞，提高系统的整体效率。例如，亚马逊的 Kiva 系统就利用多智能体协作进行自动化货物搬运。

2．交通与运输系统：无人驾驶汽车、无人机编队等需要在动态环境中协同移动。MAPF 可以用于自动规划车辆或飞行器的路径，避免交通事故或碰撞，提高交通效率和安全性。例如，无人机群体在执行搜索任务时，MAPF 算法可以确

保每个无人机都能够安全、高效地执行任务。

3．机器人竞技与模拟：在模拟环境或机器人竞技中，多个机器人需要在相互竞争的情况下规划路径，避免碰撞并争夺目标位置。这类应用中，MAPF 算法可以用来规划机器人在复杂环境中的行动轨迹。

4．搜索与救援任务：在灾难救援等高危环境下，多个机器人可能需要协作进行搜索与救援任务。MAPF 在这种场景下可以帮助机器人合理安排路径，确保任务完成的同时避免障碍物和其他机器人的冲突。

5．视频游戏与仿真：在大型多人在线游戏（MMO）中，虚拟角色的路径规划同样是游戏引擎中的关键问题。MAPF 可以用于规划游戏中的敌人或友军的运动路径，确保他们能够合理避开障碍和其他角色，提升游戏体验。

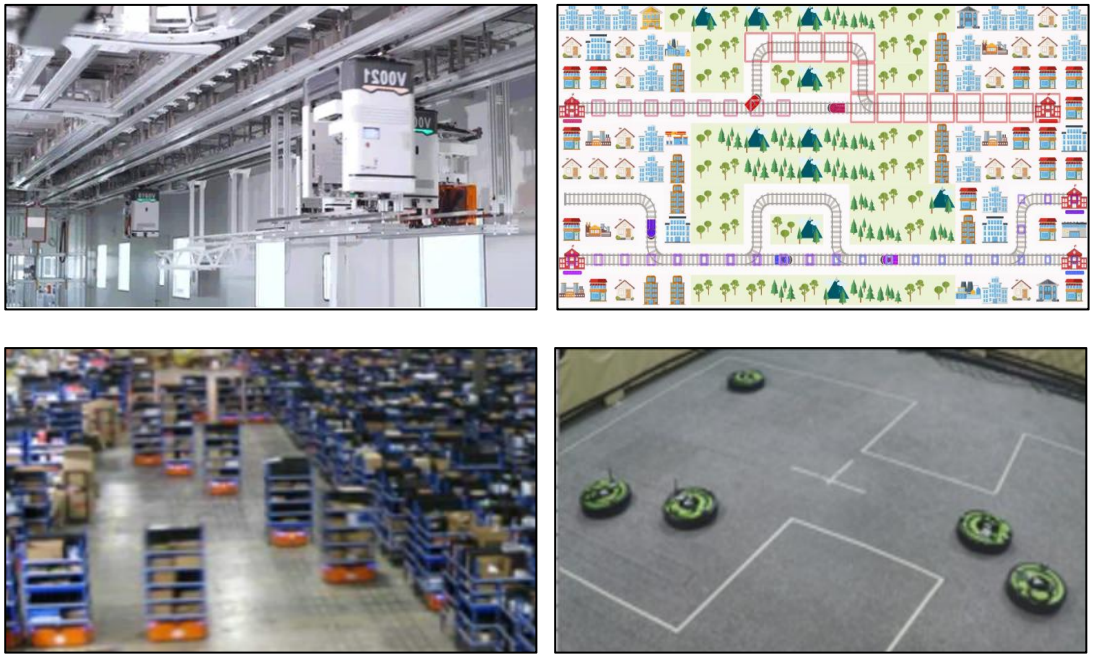


图 2 多智能体路径规划应用场景示例

二、项目应用平台与基础

1．硬件平台

本项目为算法研发，对硬件无特殊要求，实验室可提供一定的算力平台。

2．软件平台

本项目对操作系统和开发环境无特殊要求，推荐使用 C++语言进行编程。

3．技术基础

本项目掌握基本的 C/C++编程、数据结构与算法知识即可，在项目研发过程

中需要使用 C/C++等编程语言。

三、项目要求

下列要求中 1 和 2 必须满足：

1．积极主动，有充分时间投入；

2．敢于挑战难题，愿意持续钻研改进算法；

3．有一定的编程基础。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．经典多智能体路径规划的冲突驱动搜索算法实现

2．经典多智能体路径规划的大邻域搜索算法实现

3．经典多智能体路径规划的加权局部搜索算法实现

4．异构多智能体路径规划算法实现

5．变速多智能体路径规划算法实现

6．多智能体实时路径规划算法实现

7．带动态故障的多智能体实时路径规划算法实现

8．单向轨道网络上的多智能体路径规划算法实现

9．多智能体任务分配与路径规划算法实现

10．多智能体实时任务分配与路径规划算法实现

11．多智能体多周期任务分配与路径规划算法实现

12．多智能体多周期实时动态任务分配与路径规划算法实现

13．自由二维空间内多智能体路径规划算法实现

14．自由三维空间内多智能体路径规划算法实现

15．电池容量约束下多智能体路径规划算法实现

21．数据挖掘与机器学习团队

数据挖掘与机器学习团队依托于华中科技大学计算机学院，主要研究领域包括人工智能安全、图机器学习、智能优化与决策、深度学习、AI4Science 等。团队负责人何琨，华中卓越学者计划特聘岗教授，智能科学与技术专业带头人，华中科技大学霍普克罗夫特计算科学研究中心副主任、执行主任。团队在 NeurIPS、ICML、ICLR、CVPR、ICCV、ACL、AAAI、IJCAI 等国际顶级会议以及 AIJ、IJCV 等权威期刊上发表学术论文 200 余篇，何琨教授 2015 年以来多次入选 AI 2000 全球最具影响力学者榜单，并入选 2024 年全球前 2%顶尖科学家榜单。曾获约束与规划国际学术会议 CP 2021 最佳论文奖、SAT 可满足性问题 2022 国际算法竞赛主赛道冠军、MAXSAT 最大可满足性问题 2024 国际算法竞赛非完备组四个赛道冠军等奖项。所指导的研究生中培养了 2 名华为天才少年等优秀人才。团队长期与国内外顶尖高校如美国康奈尔大学、法国亚眠大学和知名企业如微软亚洲研究院、阿里安全、华为等保持密切交流与合作。团队拥有较为充分的高性能 GPU 服务器和高性能个人计算机，可进行大数据量的建模仿真及深度学习实验计算，以宽敞舒适的办公场所和充分的软硬件实力为团队提供良好的科研环境。

联系方式：何琨教授，brooklet60@hust.edu.cn

项目 1：AI 辅助新材料研究发现

一、项目背景

近年来，人工智能的高速发展推动了多领域科研的突破，尤其是在新材料的发现过程中。传统的材料研究方法通常需要经过漫长的实验步骤，而 AI 技术的引入大大加速了这些步骤，提升了效率和准确性。新材料的研究与开发是推动科技进步和产业升级的重要驱动力，然而，传统的研究方法往往依赖于实验的反复验证，周期长、成本高，且创新性有限。随着人工智能技术的不断成熟，AI 在新材料发现中的应用逐渐成为研究热点，尤其是在文献筛选、分子性质预测、分子设计、化学合成路径设计以及实验合成与检测等环节中，AI 展现出了巨大的潜力。

一种新材料的发现通常包括以下五个步骤：

文献筛选与提取：研究人员需要从海量的文献中筛选出与目标材料相关的信息，提取关键数据。传统方法依赖于人工筛选，效率低下且容易遗漏重要信息。AI 可以通过大模型结合检索增强生成（RAG）技术，快速分析和提取文献中的关键信息，大幅提升筛选效率。

分子性质预测：在材料设计过程中，分子的物理化学性质是决定其应用潜力的关键因素。传统方法通常依赖于实验数据的积累，耗时且成本高。AI 可以通过分子大模型，结合迁移学习或微调训练，快速预测分子的能级、导电性、热稳定性等性质，为新材料的设计提供理论支持。

分子重新设计：基于预测的分子性质，研究人员需要对分子进行重新设计，以满足特定的性能需求。AI 可以通过生成对抗网络（GAN），扩散模型（Diffusion）等技术，自动生成具有特定性能的新分子结构，极大地提高了设计效率。

设计化学合成路径：新材料的合成路径设计是实验阶段的关键步骤。传统方法依赖于经验丰富的化学家进行路径设计，周期长且容易出现错误。AI 可以通过分析已有的合成路径数据，结合化学反应规则，自动生成高效、可行的合成路径，减少实验失败的概率。

实验合成与检测：最后，研究人员需要通过实验验证设计的材料是否符合预期。AI 可以通过数据分析和反馈，优化实验参数，提升实验成功率。

我们希望通过 AI 加速并优化这些过程，提升研究效率，减少人工干预的同

时，提升创新性和发现新材料的概率。通过引入 AI 技术，研究人员可以在更短的时间内完成从文献筛选到实验验证的全过程，缩短研发周期，降低成本，并提高新材料的发现效率。

二、项目应用平台与基础

1．硬件平台

本项目将和公司合作提供 A100*2 GPU 服务器 用于大模型的分子建模和 AI 训练。

2．软件平台

本项目将与公司合作，提供历史实验数据，用于 AI 模型的训练和验证。

项目将结合 AI 辅助编程 Cursor 开发模式进行高效的代码生成和优化。

项目使用 Git 控制版本迭代。

3．技术基础

本项目基于深度学习和相关技术开发，主要使用 Python 编程语言和 PyTorch 框架。

由于项目涉及前端界面的开发，团队成员需要具备一定的 Web 框架基础。

三、项目需求

1．有充分时间投入，每周至少两次同步进度。

2．具备一定的 Python 基础，Web 基础，爬虫技术。

3．具备一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

目标 1：设计具有 RAG 增强搜索能力的 AI 知识库

1.精准与模糊搜索

利用分子结构（如 SMILE）、关键字、化学反应式等信息，设计一个高效的搜索引擎，能够在文献库中进行精准与模糊的快速检索，并将搜索结果实时整合至 AI 知识库中，方便后续研究使用。

2.文献自动总结

利用大型 AI 模型的文献阅读和总结能力，根据指定结构自动提取并归纳文献中的关键内容，包括“总结、反应、技术、重要结论”等要点，帮助科研人员快速获取关键信息。

3.高级功能

提供分子结构的高级嵌入技术，识别具有不同取代基的相同分子骨架，同时能够检索和识别这些分子是否已申请专利，进一步提升搜索效率与实用性。

目标 2：预测分子能级

1.迁移学习与微调

利用分子大模型进行迁移学习或微调，预测分子的能级，提升能级预测的准确性，并有效处理不同类型的分子数据。

2.AI 加速传统方法

研究并探索如何结合 AI 技术优化传统的分子能级预测方法，提升计算效率和结果精度，减少传统方法的局限性。

3.高级功能

探索分子间相互作用对能级的影响，进一步提高预测精度，考虑分子内部和分子间的非理想交互效应。

4.局部基团替换优化

基于模型的预测结果，提出局部基团替换的优化建议，帮助研究人员在分子设计中做出更为合理的结构调整。

目标 3：根据分子片段生成新分子

1.分子骨架生成

根据现有分子片段设计新的分子结构，提升分子设计效率，尤其在专利撰写过程中提供有力支持。此功能有助于加速新分子的构思和生成，减少传统设计中的人工干预。

2.反应合成路径搜索

针对目标分子、底物和合成条件，进行全面的反应合成路径分解搜索。此任

务面临较高的复杂性，需在多条件、多反应路径下进行优化，以达到高效且准确的预测。

3.基于专利与文献生成新分子

基于专利和文献中的分子结构、合成路线等信息，生成新的分子设计。该功能要求使用高效的分子大模型，能够根据已有文献内容提出创新性分子结构和合成路线，为新材料的发现提供强有力的支持。

项目 2：基于 Transformer 的图表示学习方法研究

一、项目背景

在过去的数十年间，图表示学习领域取得了长足的进步。早期的传统方法如基于矩阵分解、随机游走等技术尝试对图结构数据进行特征提取与降维，以获取节点或图的低维表示，从而便于后续的任务处理，例如节点分类、链路预测等。这些方法虽然在一定程度上提供了可行的解决方案，但往往受限于手工设计特征的局限性，难以自动捕捉图中复杂且多变的结构信息与语义关系，无法满足日益增长的复杂应用需求。

近年来，深度学习技术的蓬勃发展为图表示学习注入了新的活力，图神经网络（GNN）应运而生并成为该领域的研究热点。GNN 通过在图上定义卷积、循环等操作，能够有效地聚合邻居节点信息，层层传递并更新节点特征，从而实现对图结构特征的自动学习。然而，随着实际应用场景的愈发复杂，GNN 逐渐暴露出一些固有缺陷。一方面，GNN 在长距离依赖捕捉上存在瓶颈。尽管通过多层堆叠理论上可以扩展感受野，使节点能够获取更远距离节点的信息，但在实践过程中，由于梯度消失或梯度爆炸问题，随着网络层数的增加，长距离信息传递的有效性大打折扣；另一方面，GNN 对节点特征的利用方式相对单一。多数 GNN 模型在聚合邻居信息时，未能充分考虑节点自身属性特征的多样性以及不同属性之间的复杂交互关系。

与此同时，研究人员开始尝试将 Transformer 引入图表示学习领域，期望借助其强大的特征捕捉能力解决传统图学习方法面临的困境。基于 Transformer 的图表示学习方法初步展现出一些独特优势。其自注意力机制能够突破图结构中邻居节点的限制，直接计算任意节点之间的关联强度，为捕捉长距离依赖提供了新的途径。此外，Transformer 对于输入特征的灵活处理方式，有利于将节点的复杂属性信息与拓扑结构信息进行有机整合。

尽管如此，目前基于 Transformer 的图表示学习方法仍处于发展阶段，面临诸多挑战亟待解决。例如，如何设计适用于图结构的位置编码方式，使 Transformer 更好地建模复杂图结构特征；怎样优化自注意力机制的计算复杂度，以应对大规模图数据带来的高计算量挑战；以及如何在模型训练过程中避免过拟合，充分利用有限的图数据资源等。这些问题限制了基于 Transformer 的图表

示学习方法的进一步发展。

本项目聚焦于基于 Transformer 的图表示学习方法研究，通过深入探索和攻克上述关键技术难题，开发出更为高效、强大的图表示学习模型，为社交、生物等众多领域提供创新性的解决方案，推动相关行业的智能化升级与科学研究的深入发展。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

基于 PyTorch、DGL、PyG 等深度学习框架进行编程

3．技术基础

本项目基于深度学习、图表示学习、数据挖掘相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 基础；

3．具有一定的 PyTorch 基础，了解图深度学习和 Transformer 等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．位置编码的设计：针对图结构的数据特征，设计相对应的位置编码，为图 Transformer 的高效建模提供坚实的基础。

2．注意力计算的优化：在现有的线性 Transformer 的基础上，结合图结构的数据特征，设计一种新型的图线性注意力机制，以降低图 Transformer 模型的训练开销。

3．数据增强方法的设计：针对图 Transformer 的建模特性，设计相适配的

数据增强方法，以有效提升图 Transformer 在稀疏数据下的性能。

4．基于特定任务的图 Transformer 设计：结合具体的下游任务场景，如图分类、节点分类、链路预测和图聚类等，开展针对性的模型研究，以提升图 Transformer 在特定任务场景下的性能。

5．基于特定图结构的图 Transformer 设计：结合具体图结构数据，如符号图、同构图、异构图和有向图等，开展针对性的模型研究，以提升图 Transformer 对不同类型图结构数据的建模能力。

22．智能与实时计算团队

计算机学院“智能与实时计算团队”依托于华中科技大学计算机学院计算机软件与理论湖北省重点学科建设，拥有开放的学术氛围和国际前沿的研究方向。团队现有教授 1 名，副教授 1 名，武汉市“青年科技晨光计划”入选 1 人，湖北省优秀博士论文获得者 1 人。目前在读博硕士研究生约 30 余人，拥有近 200 平方米实验基地。团队的发展与建设为优秀人才的培养提供了良好的基础设施与外部条件。团队承担了 30 余项科研项目，包括国家重点研发计划课题、国家自然科学基金重点/面上/青年项目、国防预研重点项目、国防预研基金、企业横向应用项目等。发表国内外学术期刊及国际学术会议论文 80 余篇，获得国家发明专利 15 项，获得湖北省科技进步一等奖 1 项。团队坚持开放与联合，与美国、德国等国家和香港、台湾地区的大学，以及华为、中船重工等知名企业保持着密切合作。团队秉承“明德、厚学、求是、创新”的华科大精神，倡导“专心致志做事，自由自在做人”的原则，不断开拓进取，勇攀科学高峰，致力成为国内知名研发团队和人才培养基地。

目前研究领域主要包括下列两方面：

多模态人工智能：人工智能方向研究主要包括自然语言处理、信息检索与推荐、Fintech、小样本学习、强化学习等，聚焦在自然语言处理和信息检索与推荐两个领域。其中，自然语言处理方面主要研究基于预训练大语言模型的领域专有化模型微调以及训练方法、多模态数据的语义对齐与特征融合、多模态意图感知、多模态任务型对话系统、事实验证等；信息检索与推荐方面主要研究大规模跨模态信息的快速检索、多模态推荐系统和可解释性分析、以及大规模知识图谱构建与查询等。本方向承担了国家重点研发计划课题、国家自然科学基金、企业横向项目等一批课题，相关研究成果发表在 ACL，EMNLP，COLING，AAAI，SIGIR，ACM MM，CIKM，ECAI，DASFAA 等重要国际学术会议和 IEEE TKDE，TDS，ACM TWEB 等知名期刊上。

信息物理系统与实时系统：本方向主要研究如何通过传感器、PLC、软件系统采集外部事物信息，根据数据采集情况，进行实时处理和分析。已完成案例包括智慧工地系统、智慧矿山系统、智慧试验室管理系统、以及智慧能管系统。具体实现功能主要包括数据采集、数据分析、图像/视频识别（包含资源受限和非

受限环境，例如，嵌入式系统上的图像识别需要考虑计算资源、能耗与实时性限制)、设备健康度评估、设备损坏影响程度评估、设备运维推荐、节能方案推荐等。已实现系统支持千余传感器同步监控；数据采集/处理实时性保障（采集/处理周期用户自定义)；图像/视频识别支持服务器和嵌入式系统两类环境，其中，嵌入式系统环境中支持飞腾 2000+、英伟达、比特大陆等开发环境；设备健康度评估准确性 95%以上；节能方案推荐实现年均节能 60%以上。与普通操作系统相比，实时系统能够确保任务（或指令）在指定时间内完成，而非尽快完成（这种方式不具备时间上的确定性)。嵌入式系统是指计算、存储资源有限，便有穿戴与安装的操作系统。实时嵌入式系统通常是小型设备（例如智慧盒子、无人机、电子表、智能眼镜等）上的系统运行环境。任务调度是实时嵌入式系统中的关键技术，涉及系统实时性保障、系统能耗控制、系统可靠性与容错性等方面。在实时嵌入式系统方向，对于任务调度、任务实时性保障、任务实时性控制、任务响应时间优化具有较深入的研究。本方向承担了国家自然科学基金、湖北省自然科学基金、企业横向项目等一批课题，研究成果发表在 IEEE TC, TPDS, TMC, TCAD，ACM TODAES，TECS，RTSS, RTAS, EMSOFT 等重要学术期刊和会议上。

团队毕业生质量获得企业一致认可，在业界具有良好口碑。每年毕业生均到华为、字节、腾讯、阿里等一流 IT 企业就业，年薪屡创新高。

团队成员：

李剑军	团队负责人 教授，主要研究领域为自然语言处理，推荐系统，信息物理系统等
周　全	副教授，主要研究领域为信息物理系统，实时系统等

团队联系方式：

联系邮箱：jianjunli@hust.edu.cn，李剑军老师

地址：华中科技大学南一楼中 510 房间

项目 1：大语言模型赋能的推荐系统

一、项目背景

随着互联网技术的迅猛发展以及智能设备的普及，用户生成的数据量呈指数级增长，推荐系统成为应对信息过载的重要手段之一。在电商、视频平台、社交媒体等领域，推荐系统通过个性化服务满足了用户需求，并显著提升了用户体验。然而，传统的推荐技术在应对海量数据和多样化用户需求时，暴露出了一些局限性，比如在语义信息理解和不同模态的数据融合（ID 与文本）方面。

传统图推荐方法主要依赖于用户与物品之间的交互关系，例如点击、浏览或评分等行为。这些方法通过构建用户-物品图或用户之间的协同关系来预测用户的偏好，它们通常仅考虑用户与物品之间的直接交互关系，忽视了物品本身的语义关联。这导致推荐结果可能缺乏多样性和准确性，难以应对复杂的场景需求。同时，许多推荐对象（如书籍、电影、商品等）通常伴随有丰富的文本描述。传统方法难以充分理解这些文本数据的语义特征，从而限制了推荐性能。

近年来，大语言模型为推荐系统的发展带来了新的机遇。大规模预训练语言模型具有强大的语义理解和生成能力，能够从文本数据中提取深层次的语义特征，弥补传统推荐系统在语义信息处理上的不足，提升推荐系统的性能。一方面，大语言模型既能为推荐物品的文本属性提供编码；另一方面，大语言模型的生成式建模经过微调后也可用于推荐系统的物品生成上。

我们的目标是利用大语言模型的先验知识和文本的理解、生成能力，设计出全新的基于大模型增强的推荐系统。

该构想分为两种实现路径。

LLM as Recommender Systems，大模型作为基座的推荐系统。这一类推荐系统直接利用大模型的生成能力，借助提示学习等方式进行生成式推荐。通过适当微调或者设计对齐单元的方法，将大模型生成的文本替换为物品，作为序列推荐等场景下的推荐器；其次，对于大模型生成的表征，考虑构建更新颖的对齐方式或者寻找合适的相似度计算的方式与候选物品进行匹配，从而获取推荐结果。

LLM-enhanced Recommender Systems，大模型增强的推荐系统。大模型增强通过将 LLM 融合进传统高效高精度的协同过滤之中，使得传统的协同过滤推荐也可以得到大模型的增幅。这种方法的基本做法是将大模型生成的嵌入整合到图结构

中，以增强图的聚合表示能力。这种形式的增强一般包含下述三个步骤：

1．建立一个用大模型进行语义增强的图神经网络建模知识图谱推荐方法。通过大模型完善物品文本、大模型编码语义信息并结合知识图谱、思维链等方式，对于传统的图模型的表征进行增强

2．采用传统的或者改进过的图推荐模型，进行邻域信息的聚合

3．对于图中输出的结果，进一步结合对比学习等技术将 ID 和文本表征进行对齐，从而得到最终的推荐概率。

二、项目应用平台与基础

1．硬件平台

本项目的实践主要基于 Linux 服务器，配备 3090/4090/A 系列/H 系列显卡进行高效推理与计算。

2．软件平台

本项目可基于任何一项代码编辑器进行模型实现，推荐使用能够搭载 Python 语言的 vscode 或 Pycharm 进行

3．技术基础

本项目基于深度学习、推荐系统、、自然语言处理、大语言模型等相关知识，在项目研发过程中需要使用 Python 语言与深度学习框架 Pytorch。

三、项目要求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 Python 基础与工程实现能力，了解深度学习的基本模型与训练方式，并能熟练使用大语言模型（chatgpt、claude 等）辅助进行学习与科研工作。

3．熟悉项目协作开发工具 git 与远程代码仓库（github 等）的使用。

四、项目开展

可针对项目描述中的场景需求，可选取至少一项或结合几项开展：

1．LLM as Recommender Systems 场景下，设计特定于推荐系统的微调方式

训练大模型作为推荐器，在传统数据集上取得不错的推荐效果。

2．LLM as Recommender Systems 场景下，探究更高阶的 prompt 的方式以更好地微调大模型以适应推荐任务

3．LLM-enhanced Recommender Systems 场景下，大模型增强通过将 LLM 融合进传统高效高精度的协同过滤之中。结合知识图谱、思维链、GraphRAG 等技术，更充分地利用大语言模型的推理能力增强推荐系统。

4．LLM-enhanced Recommender Systems 场景下，探究使用数据集增强的方式来引入更多的信息辅助推荐。

5．LLM-enhanced Recommender Systems 场景下，探究使用数据集增强的方式来引入更多的信息辅助推荐。

6．LLM-enhanced Recommender Systems 场景下，对 base 图模型进行进一步优化，设计更高阶、更契合 LLM 增强表征的图传播算法。

7．在所有大模型推荐场景下，在现有研究基础上，进一步设计大模型生成结果的对齐、匹配算法，并尽可能让算法能在各种推荐场景下具有普适性。

项目 2：基于大语言模型的工程知识库构建及应用

一、项目背景

随着科技的飞速发展和项目复杂性的增加，各个工程领域都需要处理海量、多源的技术文档、设计方案和行业规范。然而，传统知识管理工具往往难以满足实时性和高效性的要求。与此同时，人工智能特别是大语言模型的崛起，为知识的智能化管理提供了新的可能。大语言模型具备强大的自然语言理解和生成能力，能够高效整合、组织和检索跨领域的工程知识，为构建基于大语言模型的知识库提供了基础。尽管在语言理解及文本生成等领域展现了优于传统模型的优势，现有大语言模型在理解和应用专业性较强的领域知识时仍然存在一些不足，这也影响了专业领域中基于大语言模型的工程知识库的构建。因此，如何构建领域专用的基于大语言模型的工程知识库仍然是一个关键问题。针对该问题，我们希望能够为专业领域的工程构建足以投入实际应用的基于大语言模型的工程知识库。该构想主要分为两个方面：①从收集到的某领域内的相关文本文件中抽取出用于构建工程领域知识库的知识并构建该领域的工程知识库；②通过检索增强生成及参数高效微调等技术，开发出一款可以根据构建的工程知识库中包含的专业领域知识以及通用知识进行问答的自然语言智能问答工具。

为了构建基于大语言模型的专项领域知识库，首先需要收集充足的包含相关领域知识的文本文件用于该领域相关的知识抽取。从专业领域文本中抽取知识是构建专业领域知识库的核心环节，旨在将复杂、分散的专业信息结构化为可用的知识资产。通过自然语言处理技术，尤其是基于大语言模型的方法，可以自动解析海量领域文本，提取文本中的关键实体、关系以及规则。这些知识经过语义理解和信息整合后，被转化为知识图谱或其他结构化形式，并存储在知识库中。这样的知识库能够高效地支持查询、推理和决策。例如，在工程领域，知识库可用于设计优化、故障诊断和技术推荐。通过这种方法，构建的知识库能够显著提升领域知识的可访问性和利用效率，为专家和非专业用户提供精准、实用的智能化支持，助力行业创新和高效发展。

在构建好工程领域的知识库后，为了用户可以与构建的基于大语言模型的工程知识库进行交互，还需要开发用于与构建好的知识库进行交互的自然语言智能问答工具。为了让现有的大语言模型能够理解工程相关的专业领域知识，可以对

基座大模型进行专业领域知识的参数高效微调。此外，为了让模型的回答更加可靠，还可以根据用户的问题对知识库进行检索并增强模型生成的回答。结合这两种技术，开发的工具既能动态访问最新的知识库内容，又保留语言模型的强大生成能力和自然交互特性。这种自然语言智能问答工具可以广泛应用于工程、医疗、法律等专业领域，为用户提供准确、高效、便捷的智能问答服务，助力专业领域的数字化和智能化发展。

综上所述，为专业性较强的工程领域构建领域专用的知识库是十分必要的。而为工程相关领域的用户构建能够根据领域专用知识库进行自然语言智能问答的工具可以让用户对知识库中的专业领域知识进行更好的应用。为了实验对专业领域知识更高效的应用，本研究通过探索知识抽取、推理及融合等技术来为工程相关的专业领域构建领域专用知识库，并利用参数高效微调以及检索增强生成等基础为工程知识库的应用提供用于和知识库进行交互的自然语言智能问答工具。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器；

2．软件平台

基于 Dify 等大语言模型应用开发框架及 Ollama、LLaMA Factory 等大模型相关工具；

3．技术基础

本项目基于大语言模型、参数高效微调、检索增强生成、数据库及知识图谱等相关理论与技术。

三、项目需求

1．踏实肯干，有充分时间投入；

2．具有一定的 Python 编程基础；

3．了解一定的大语言模型相关理论及技术；

4．具备一定的数据库及知识图谱等相关理论与技术。

四、项目开展

可针对项目描述中的场景需求，选取至少一项开展：

1．知识库数据的采集和加工：针对特定工程领域，收集和清洗大量高质量的文本数据，确保数据能够覆盖任务需求且具备一定的多样性，为工程领域知识库构建提供坚实的数据基础。

2．结构化及非结构化知识抽取：从结构化及非结构化文件中抽取出工程相关的实体、关系及规则等知识并将知识存入知识图谱或数据库中，为后续对这些知识的精确查询提供基础。

3．构建文本知识库：通过文本切片算法对文本进行分块，并使用 Embedding 模型对分块后的文本进行向量化处理，并存储到向量数据库中，以方便在知识问答场景中通过检索增强生成等技术从向量库中召回知识进行回答。

4．基于 text2sql 的数据库自然语言问答：通过大模型将用户的自然语言问题生成与问题等价的 sql 语句并根据生成的 sql 语句查询数据库，以此实现用户与数据库之间的自然语言问答及相关问题的精确查询。

5．训练及测试数据集生成：由于某些工程领域并不具备用于训练及微调大模型的大量数据，因此需要借助大模型根据收集到的工程领域相关文本数据，自动为后续的模型训练及微调生成相应格式的训练集和测试集。

6．大语言模型参数高效微调：通过优化低秩适应等参数微调技术对模型进行参数微调，使得在领域相关的训练数据有限的前提下，能够高效地微调大模型，使其能够理解领域相关的专业知识并保持模型本身的性能。

7．基于大语言模型的文档解析与生成：利用大语言模型的基础能力对输入文档进行解析，并能够对文档进行内容、风格及合规性等方面的审查。优化大模型的生成能力，使大模型可以根据提供的模板生成与模板风格一致的工程领域相关的大篇幅文本。

23．智能信息与大数据团队

智能信息与大数据团队由最近加入华中科技大学计算机学院的张瑞教授主导。张瑞教授是大数据和人工智能方向国际知名学者、华中卓越学者首席教授、之前是墨尔本大学（QS 2024 年世界排名 14）的终身教授、澳大利亚国家级人才项目获得者，同时也是清华大学客座教授。本科毕业于清华大学、博士毕业于新加坡国立大学，并在微软研究院、谷歌、AT&T 研究院、德国 Max Planck Institute 等世界顶尖研究院和学府担任访问学者。在人工智能（大模型、信息检索增强生成 RAG、多模态理解与生成、推荐系统、知识图谱、对话系统等）和大数据（数据库、索引查询、数据挖掘、图数据挖掘与管理、时间空间数据挖掘与管理等）方向取得了一系列具有国际影响力的创新成果，被国际头部 IT 公司例如微软、谷歌、亚马逊、华为、美国电信电报公司等广泛采用。获得多次国际重大奖项，包括谷歌学者奖、澳大利亚 Future Fellowship 人才计划、大洋洲计算机科学研究杰出贡献奖等。在多个国际顶级会议获得过最佳论文奖，包括信息检索领域顶级会议 WSDM 2024 最佳论文提名奖、数据挖掘领域旗舰会议 ACM SIGKDD 2016 年最佳论文奖。最近在大模型个性化多模态生成方向的成果 PMG 技术受到科技领域头部媒体“量子位”的报道，并在华为公司推进应用。在人工智能和大数据领域发表国际顶尖会议和期刊如 SIGIR、SIGKDD、 SIGMOD、 NeurIPS、 ICML、 ACL、AAAI、VLDB、ICDE、Web Conference、ACM Multimedia、 TPAMI、 TKDE、 TOIS、TODS、 VLDB Journal 等论文 200 多篇，其中 CCF A 类或者 JCR 一区论文 120 多篇。主持过多项科研探索项目以及工业合作项目，有丰富的项目主持以及落地经验。最新论文参考(https://www.ruizhang.info/publications/pubindex.htm)

团队主要研究方向如下：

人工智能在信息领域、自然语言处理方面的应用，具体包括：大模型及应用、大模型复杂推理、信息检索、检索增强生成（RAG）、多模态理解、生成推荐系统、知识图谱、知识抽取、关系抽取、实体对齐

大数据、数据挖掘，具体包括：数据库向量检索、查询，例如自然语言通过大模型进行数据库/表格操作，图挖掘、图数据管理，时间、空间数据管理、查询、挖掘

团队毕业生就职于国内外多所顶尖高校（美国马里兰大学、澳大利亚墨尔本

大学、新加坡国立大学、香港科技大学等）和国内外科技巨头公司（微软、谷歌、亚马逊、Twitter、腾讯、百度等），毕业生年薪屡创新高；所培养的多位学生在世界头部高校和科技巨头公司实习。

团队成员：

张　瑞	团队负责人 教授，主要研究领域为人工智能（大模型、信息检索增强生成 RAG、多模态理解与生成、推荐系统、知识图谱、对话系统等）和大数据（数据库、索引查询、数据挖掘、图数据挖掘与管理、时间空间数据挖掘与管理等）

团队联系方式：

联系邮箱：ruizhang6@hust.edu.cn，张瑞老师

地址：华中科技大学南一楼

项目 1：基于大模型的个性化图像生成中的人脸身份编辑与矫正

一、项目背景

随着人工智能技术的发展，大型语言模型在文本理解和生成方面取得了显著进展。这些模型不仅能够处理纯文本数据，还逐渐扩展到了图像、音频、视频等多模态领域。

得益于多模态大语言模型的蓬勃发展，特别是生成模型的兴起，生成式推荐系统应运而生。传统的推荐系统主要依赖于协同过滤和矩阵分解等方法，通过分析用户的历史行为和偏好来推荐内容。虽然这些方法在一定程度上提高了推荐的准确性，但通常只能提供静态的推荐列表，缺乏动态和个性化的交互体验。

为了填补这一研究领域的空白，个性化多模态生成（Personalized Multimodal Generation, PMG）方法被提出。PMG 通过对用户的行为历史（如点击记录、对话内容等）进行分析，构建离散和连续的用户偏好表征，生成更具个性化的多模态内容。特别在电影网站的海报展示、社交媒体的表情符号推荐以及在线广告的产品展示等多种场景下，PMG 都能够根据用户个人偏好生成极具个性化的图像。

然而，尽管 PMG 通过个性化图像生成已经实现了个性化、多模态的生成式推荐，但仍然有一些不足。现有的个性化图像生成模型在进行人物图像的生成时，对于人脸身份的指向总是随意、不够准确的。为了生成符合需求的个性化图像，我们需要对图像生成后的人脸身份信息进行编辑、矫正。

二、项目应用平台与基础

1. 硬件平台

基于具有足够算力的服务器

2. 软件平台

基于 Centos7、PyTorch 等深度学习及大模型库进行编程

3．技术基础

本项目基于深度学习、强化学习、大语言模型、参数高效微调、检索增强生成相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

研究现状分析：对当前个性化多模态生成方法的研究现状进行综述，特别关注现有个性化图像生成任务中，人脸身份生成方面的挑战。

数据集准备：收集和标注数据集，以支持模型训练。

算法设计：通过人脸检测算法识别原图中的人物特征和信息，构造 mask 图像。从数据库检索目标人脸图像作为身份信息控制条件，生成矫正图像。

模型训练：使用深度学习框架训练生成模型，优化人脸身份生成的准确性。

结果评估：通过实验评估模型在人脸身份编辑和矫正方面的性能和效果，并根据实验结果进行模型优化。

五、预期目标：

实现个性化图像生成任务中人脸身份编辑、矫正算法，能够根据人脸身份信息，生成准确的个性化图像。

撰写并完成关于个性化图像生成中人脸身份编辑、矫正的研究论文，详细描述研究背景、方法设计、实验结果及及总结展望，并计划在相关学术会议或期刊上投稿发表。

参考资料：

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, Xi Xiao, PMG : Personalized Multimodal Generation with Large Language Models, The Web Conference 2024.

Ye H, Zhang J, Liu S, et al. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models[J]. arXiv preprint arXiv:2308.06721, 2023.

Wang Q, Bai X, Wang H, et al. Instantid: Zero-shot identity-preserving generation in seconds[J]. arXiv preprint arXiv:2401.07519, 2024.

项目 2：基于轻量化微调的个性化大模型研究

一、项目背景

项目目标为探索轻量化微调方法如何提升大模型的个性化能力。开发和实现使用 LoRA 等方法的微调策略，以解决个性化信息和通用信息的解耦与整合问题。同时研究在不同场景下，动态决定是否引入个性化参数进行响应的机制。

二、项目应用平台与基础

1．硬件平台

基于具有足够算力的服务器

2．软件平台

基于 Centos7、PyTorch 等深度学习及大模型库进行编程

3．技术基础

本项目基于深度学习、强化学习、大语言模型、参数高效微调、检索增强生成相关理论与技术。

三、项目需求

下列要求中 1 和 2 必须满足：

1．吃苦耐劳，踏实肯干，有充分时间投入；

2．具有一定的 C++、Python 基础；

3．具有一定的 PyTorch 基础，了解深度学习、大模型技术等相关理论与技术。

四、项目开展

1．熟悉并掌握主流大模型和轻量化微调方法。

2．个性化数据的收集与处理

3．多层 LoRA 微调与模型开发。1）实现多层 LoRA 微调策略： 第一层存储个性化信息，第二层增强通用能力。2）通过训练验证个性化和通用信息的解耦效果，确保 LoRA 参数层之间的独立性和协同性。

4．个性化参数的自动化选择机制。设计基于场景的动态决策机制： 1）针

对通用任务，仅加载通用参数。2）针对个性化任务，结合个性化参数和通用参数生成响应。

5．输入/输出分析与模型比较。1）在各种个性化场景中比较模型表现，包括个性化答题和通用答题任务。2）分析微调成本和效率，确定最优方案。3）验证模型在平衡个性化和通用化方面的有效性。

五、预期目标

1．项目目标

（1）基于轻量化微调方法完成一个基础个性化模型，并验证多层 LoRA 策略的有效性。

（2）确定多层 LoRA 策略在提升个性化和通用化能力中的最佳参数配置，实现信息的高效解耦与整合。

（3）实现基于场景的自动化决策机制，动态加载参数以应对不同任务。

（4）分析并报告模型表现，明确个性化与通用化能力的平衡点，并优化方案。

2．学生能力目标：

（1）熟练掌握主流大模型框架和微调工具（如 Hugging Face）。

（2）调研并预处理个性化数据集，确保其在评估中的有效性和相关性。

（3）设计并验证多层 LoRA 微调策略，包括用于增强个性化能力的参数层以及用于进一步优化通用能力的参数层。

（4）实现解耦学习，将个性化信息和通用信息分离并更新到对应的参数中。

（5）开发基于场景的动态决策机制，在响应过程中动态加载个性化参数。

（6）分析模型在通用任务和个性化任务中的表现。

24. 智能计算与强化学习团队

计算机学院“智能计算与强化学习团队”旨在构建和谐奋进的团队文化，倡导团结协作、进取创新的理念。主要从事组合优化、运筹优化、工业优化、深度学习、强化学习、多智能体博弈对抗、大模型等领域的算法、关键技术与实际应用等方面的研究。在路径规划、深度强化学习、启发式算法等方面取得了一系列创新性成果。承担了国家自然科学基金项目（面上、青年）和多项横向项目。与国内外相关领域的专家有较为广泛的学术交流与合作。

团队目前研究领域主要包括**EDA领域的芯片布局布线、逻辑综合等算法研究，物流和供应链及低空经济的路径规划算法研究，多智能决策优化算法研究，大模型推理研究，模仿学习研究**等五个方面。

团队在AAAI、IJCAI等人工智能会议以及IEEE Transactions等计算机权威期刊发表学术论文30余篇，指导学生获得了国际竞赛GECCO赛道冠军和华为挑战赛、腾讯开悟、 Go-Bigger多智能体决策智能挑战赛、EDA设计精英挑战赛等多项奖项。

团队成员：

金　燕	团队负责人 教授，主要研究领域为组合优化、运筹优化、工业优化、深度学习、强化学习、多智能体博弈对抗、大模型等领域的算法、关键技术与实际应用等方面

团队负责人及联系方式：

负责人：金燕老师

联系邮箱：jinyan@hust.edu.cn

地址：华中科技大学南一楼东210房间