2024联想校企联合攻关课题清单

2024.09

课题发布清单

序号	课题名称	类别
人工智能		
1	智能体的任务拆解与规划技术研究	攻坚类
2	裸眼3D场景下2D视频转3D技术研究	应用类
3	裸眼3D场景下端到端意图理解的技术合作指南	攻坚类
4	低延时speech2Speech语音多模态AI模型	攻坚类
5	机器人视觉系统	攻坚类
核心部件		
6	面向未来的折叠屏材料技术研发	攻坚类
7	提升手机转轴开合折弯寿命的耐久性及稳定性	攻坚类
8	面向智能手机的高性能微泵风冷/液冷等主动散热方案	攻坚类
9	超薄轻质陶瓷材料开发	攻坚类
10	不锈镁——电泳涂装技术	攻坚类
11	不锈镁——涂料技术	攻坚类
12	铝合金表面抗指纹技术开发	攻坚类
智能服务与运维		
13	面向异构计算平台的GPU故障预测	应用类
14	面向大规模的大模型训推一体化异构集群的故障智能运维方法	攻坚类
智算平台		
15	面向RAG的融合知识图谱的向量语义检索	应用类
16	局域网内GPU AI算力分享	攻坚类
其他新兴技术		
17	基于用户认知交互的裸眼3D显示器3D空间界面设计研究	应用类

技术方向一：人工智能

课题1	智能体的任务拆解与规划技术研究	类型	攻坚类
研究背景	近年来，大模型的火热为人工智能注入新的浪潮，为计算机产品人机交互提供了新的范式。但是单纯的使用大模型进行任务的拆解与规划存在诸多问题，如计划的效用与准确性难以保证等。为了解决智能体在计算机交互中的核心问题，本课题提出了基于马尔科夫决策过程的研究方法，使用诸如强化学习或生成式人工智能技术构建复杂决策规程，并科学的构建衡量与评价标准，如奖励函数等，可以更好的辅助计算机对人复杂任务请求进行拆解与执行。 业务背景：AI agent作为研究院新财年重要的研究方向，对AIPC以及企业大模型等核心应用有重要的推动作用，AIPC已经在新pc机型上得到很好应用和反馈，企业大模型作为新的探索赛道也正在逐步设计与构建中。		
研究内容	针对当前研究方向与业务实际痛点，该项目主要包括以下三个子任务： 子任务一：构建基于马尔科夫决策过程的智能体任务拆解与规划的策略模型，能够完成单一任务（task）对多个技能（skills）的使用与规划； 子任务二：构建具有多任务强化学习能力的智能体任务拆解与规划的策略模型，能够完成多个任务（tasks）对多个技能（skills）的使用与规划； 子任务三：构建具有元强化学习能力的智能体任务拆解与规划的策略模型，能够完成没有见过的多个任务（tasks）对多个技能（skills）的使用与规划。		
研究目标	智能体的任务拆解与规划是大模型技术的核心技术差一点之一，可以为联想的AIPC与企业智能体注入更强技术竞争力，并且很好地补全了AI lab中大模型规划的核心模型，合作项目主要面对智能体的任务拆解与规划技术并实现以下目标： (1) 构建基于马尔科夫决策过程的智能体任务拆解与规划的策略模型 (2) 构建具有多任务强化学习能力的智能体任务拆解与规划策略模型 (3) 构建具有元强化学习能力的智能体任务拆解与规划策略模型		

技术方向一：人工智能

课题2	裸眼3D场景下2D视频转3D技术研究	类型	应用类
研究背景	裸眼3D技术作为一种不需要穿戴特殊设备即可观看立体图像的技术，该技术在教育、医疗、商业、文娱等领域具有重要的应用，近年来受到广泛关注。为了丰富裸眼3D的内容，将2D视频转化为3D视频成为关键挑战。传统的2D视频转换由于缺乏准确的深度信息和前景分割技术，存在立体感不足、时序稳定性差等问题。在此背景下，构建相应的大规模数据集，提升通用视频深度预测和前景分割模型的时序稳定性和精度，成为当前研究的重点。		
研究内容	本研究旨在构建高质量的数据集，研究通用的视频深度预测和目标分割模型，最终实现2D视频向裸眼3D视频的高效、高质量转换。具体目标包括： 1.提高视频深度预测模型的时序稳定性和精度，使物体帧间的深度变化规律与其Z轴运动轨迹一致。 2.提高视频前景分割模型的时序稳定性和精度，确保分割结果的连续性和一致性。		
研究目标	研究目标： 1.数据集构建：收集多种场景下的高质量3D视频数据；注视频中的深度信息和前景分割信息，形成用于训练和测试的数据集。 2.深度预测模型研究：开发和优化基于深度学习的深度预测模型，提升模型的精度和时序稳定性；设计算法确保物体帧间的深度变化规律与物体的Z轴运动轨迹一致，避免深度跳跃和闪烁现象；深度预测时序稳定性：运动意图与真实运动的平均端点误差（EPE） ≤ 5.0；深度预测精度:AbsRel<=0.07 3.前景分割模型研究：开发和优化基于深度学习的前景分割模型，提升模型的精度和时序稳定性；确保分割结果在连续帧间的稳定性，避免分割边界的不稳定和不连续现象；分割精度:交并比（Intersection over Union, IoU）>=0.88。 预计产出： 1.数据集：构建经过标注的高质量视频数据集，用于训练和测试深度预测和前景分割模型。 2.算法模型：提供经过优化的深度预测和前景分割算法模型。 3.学术论文与专利申请：1篇CCF-A类文章；2篇相关技术专利，保护研究成果。		

技术方向一：人工智能

课题3	裸眼3D场景下端到端意图理解的技术合作	类型	攻坚类
研究背景	随着裸眼3D技术的发展，用户可以在无需佩戴任何设备的情况下体验三维立体视觉效果。这种技术为各种应用场景提供了更加自然和直观的交互方式。然而，在裸眼3D场景下，传统的2D界面和键鼠输入方式已经无法满足3D场景的交互需求。用户通过自然的手势和视线操作与系统进行互动，这种方式更加直观和高效。但现有的交互技术在3D环境中存在较高的学习成本，用户需要花费大量时间来熟悉操作模式和界面。因此，亟需一种结合手眼信息的端到端意图理解技术，通过分析用户的视线和手势动作，自动识别其操作意图，降低用户的学习成本，提高交互效率和体验。		
研究内容	本研究将包括以下几个主要内容： 1.数据构建：构建裸眼3D场景下用户的视线和手势数据，包括用户在不同场景中的操作记录。 2. 模型开发：将视线和手势分析结果进行融合，建立综合意图理解模型。在单维度2cm目标选择任务中，与用户盯着注视行为相比，意图理解精度提高或无显著差异； 3.算法优化：针对端侧设备的计算资源限制，进行模型压缩和加速优化，确保实时性。 4. 系统集成：将意图理解模型集成到裸眼3D交互系统中，实现对用户意图的实时响应。 5. 技术验证：设计和开展多种使用场景下的实验，验证系统的性能和用户体验。		
研究目标	1. 开发一种基于裸眼3D场景的端到端意图理解技术。 2. 结合视线和手势分析等模型，精准识别用户的意图。 3. 降低用户的学习成本，提高用户体验。 4. 实现技术的实时端侧落地和验证。		

技术方向一：人工智能

课题4	低延时speech2Speech语音多模态AI模型	类型	攻坚类
研究背景	随着AIPC的发展，自然语音交互是一种必不可少的交互方式。传统的语音处理系统通常由多个独立的模块组成，以实时翻译为例：需要语音识别+机器翻译+语音合成3个模板组合而成，这些模块之间的衔接和优化往往较为复杂，整体系统延迟较高，无法满足实时性要求。 用一个大模型（可以基于MIT开源模型）实现语音入到语音出的交互方式，可以避免中间繁琐的模块处理，简化处理流程。搭配台式机第二代dNPU算力，可以实现多场景下低延迟的语音实时交互。		
研究内容	1. 可以基于开源的MIT license 模型进行模型调优、量化等方式实现最终目标，也可以自行搭建和训练模型已达到预期目标。 2. 模型实现的子功能： 2.1 语音克隆功能，实现男、女声，以及跨语种的语音克隆。 2.2 翻译至少实现中英互翻。 2.3 基于情绪识别的TTS，情绪识别可使用语音信息或者也可以引入图像信息增强情绪判断的准确性。 2.4 通话模式下的声音美化（降噪，去混响，过于刺耳和低沉声音修复是最基本需求）		
研究目标	1. 构建高性能的端到端语音交互模型，在语音识别准确率、语义理解准确率和语音生成自然度等方面达到行业领先水平。 1.1 类比于语音识别的词错率小于10%。 1.2 语音克隆相似度达到90%，FAD＜3。 1.3 在翻译场景中，翻译评分 BLUE＞0.8 1.4 基于情绪识别的TTS至少支持几种特有的带情绪的语音生成：停顿，笑声，叹气等。 1.5 实现语音的降噪、修复和美化。 2. 模型可实现的场景： 2.1 克隆后TTS场景 2.2 实时翻译场景(声音可以使用克隆后的音色) 2.3 声音美化场景 3. 模型实现ONNX转换和C++推理 4. 发表论文1篇，专利1-2个		

技术方向一：人工智能

课题5	机器人视觉系统	类型	攻坚类
研究背景	目前机器人领域通常使用激光slam技术完成室内地图构建，通过多源传感器信息处理支持机器人从点到点的自主避障和移动功能，通过设置工作点位或巡查点位进行移动，并通过加装摄像头，基于特定算法完成目标人物/物体的状态识别、行为动作的识别和告警，完成指定工作任务。 但雷达SLAM成本高昂，特别是高精度激光雷达传感器的价格限制了其普及；同时其较高的功耗也限制了机器人的运行里程，而且由于雷达slam技术自身的局限性，其本身也存在长时间运行和大规模场景中造成累积误差和漂移问题。在维护上，由于雷达SLAM和V-SLAM都需要不定期对雷达和camera进行标定校准，一定程度上拉低了机器人的易用性。本次申请研究的v-slam室内导航技术计划于2025年CES展会亮相，2025年2季度进行产品发布。		
研究内容	研究适用于C端机器人产品的V-SLAM导航整体解决方案。根据机器人使用需求研究如何使用v-slam方式完成室内地图建模和导航。输出包括但不限于算法模型/demo、摄像头模组推荐选型或技术规格要求、满足算法使用要求的摄像头安装位置、拍摄角度及数量，运行算法模型的硬件配置要求等，保证达到算法条件的硬件配置环境要求或推荐的配置等。具体算法模块包括： 1. V-SLAM算法：基于机器人视觉技术，完成在室内（主要是家用环境中）的路径规划及地图建模功能。可输出便于操作的二维/三维地图供操作点位使用。 2.目标检测算法：通过摄像头拍摄到人体（主要以小孩为主）局部图像完成目标检测，甄别是否为目标识别物。 3.目标跟踪算法：根据目标物持续移动速度及方向进行路径及指令优化，使机器人能够在目标物移动过程中启动跟随功能，跟随目标移动。 4.目标围绕算法：根据目标物动作判断目标物是否为相对停止状态，是否具备拍照/视讯条件，给机器人下发目标点位矢量信息和路径规划指令，最终完成角度拍摄/视讯。 5.上述算法完成所需要的软硬件规格及推荐的配置，包括摄像头参数要求及推荐型号，算法平台要求及推荐型号，算法平台要求以及对配套软硬件的规格要求和建议。 6.摄像头自校准方案：设计工装方案使机器人自带的雷达/摄像头能够在家用条件下完成自动标定，使机器人V-slam摄像头识别能力维持在较高水准。		
研究目标	1. 完成机器人可应用的v-slam算法模型，使机器人在0~10m识别范围内，v-slam导航精度可以达到±50mm以内，硬件成本不得高于同类型雷达slam机器人。 2.完成可部署的目标检测算法模型。使之基于摄像头FOV范围完成目标人体检测，在0~5m范围内，拍摄图像覆盖率不大于50%的情况下可以准确识别到目标人物，目标识别率不小于90%。 3. 完成目标跟踪算法。使目标在以0~1.5m/s内任意速度移动时，机器人具备识别和跟踪能力，响应时间小于2s，跟踪完成后不会发生碰撞，考虑刹车距离及磕碰干扰情况。 4.完成目标围绕算法模型，目标在相对停止时，机器人自带拍照/视讯摄像头能够对准目标物，目标物头部移动过程中，机器人具备动态调整功能。 5.输出整体解决方案boom清单及原材料规格清单，包含型号、规格、连接方式及主要控制板程序、算法模型及代码。 6.完成自校准工装设计及demo交付，使机器人自带摄像头能够完成自校准，校准完成时间在1min以内，校准成功率99%。 7.专利、论文：发表人工智能领域顶级期刊会议2篇，申请发明专利2项，提交标准提案1项		

技术方向二：核心部件

课题6	面向未来的折叠屏材料技术研发	类型	攻坚类
研究背景	折叠屏手机作为创新焦点，正迅速成为市场新宠。然而，屏幕折痕、耐用性与防护力不足阻碍了用户体验与市场普及。当前技术虽有所进展，但仍未实现无痕、自修复与超强防护的理想状态。针对这个问题，我们希望通过与高校建立合作，攻关材料科学与工程创新，开发下一代折叠屏技术，实现屏幕无折痕、自我修复及耐冲击、抗划伤特性。目标是革新折叠屏手机性能，引领行业技术升级与市场变革。		
研究内容	1. 探索无痕折叠材料： 研究新型柔性材料配方，旨在实现屏幕在反复折叠后仍能保持平整无痕，解决现有折叠屏技术中最突出的折痕问题。如高回弹性能的光学胶材，以及基板（如基于高分子材料PET/PI、超薄玻璃等）材料的优化。 2. 增强耐冲击与抗划伤能力，包括并不限于： - 新型基材本体的开发（如PI/PET/TPU等） - 通过纳米技术与涂层工艺，提高屏幕的耐冲击和抗划伤性能（如hard coating） - 降低划痕可见性抗眩光涂层 - 自修复机制开发：开发屏幕表面的自我修复功能，确保屏幕在受到轻微损伤后能够自动恢复——在一定的温度或光照条件下重建化学键。例如TPU，基于氢键的聚合物，含有微恢复胶囊的复合材料等等。		
研究目标	1. 开发无痕折叠技术：实现一种新型屏幕材料，确保在数十万次折叠后，屏幕表面折痕轻微或不可见，显著改善折叠屏手机的外观。 2. 创新自修复机制：研制具备自我修复功能的屏幕涂层，使屏幕在遭受轻微划伤或撞击后，能在一定条件下（用户可复制）自动恢复原状，无需外部特殊干预。 3. 增强保护层性能：设计并验证一种或多种保护层配方，使屏幕的耐冲击性和抗划伤性提升，表面硬度提升至7H或以上。		

技术方向二：核心部件

课题7	提升手机转轴开合折弯寿命的耐久性及稳定性	类型	攻坚类
研究背景	目前在折叠手机开发过程中遭遇一些难题，主要是凹凸轮工作面在翻盖开合寿命测试中容易受到磨损，10万次后扭力衰减损失在35%- 40%，导致翻盖寿命测试20万次后，手机会出现闭合张嘴，打开力小的问题，长期来看，影响用户体验及用户对品牌的认可度及忠诚度。		
研究内容	转轴高强度高耐磨性新材料开发，以实现转轴尺寸小型化，提升折叠手机开合寿命耐久性		
研究目标	1. 基于BASF 420W材料属性进行改良或开发可替的金属合金材料 2. 热处理后力学性能需求：屈服强度1500- 1800Mpa，延伸率>6%，硬度>HV600，杨氏模量>200Gpa 3. 成型工艺：MIM （Metal Powder Injection Molding Technology）		

技术方向二：核心部件

课题8	面向智能手机的高性能微泵风冷/液冷等主动散热方案	类型	攻坚类
研究背景	Generative AI(GAI)的出现，大大推动了智能手机On Device AI技术的发展。GAI需要大量算力从而会产生大量热量，但手机体积非常小，仅靠被动散热技术很难解决散热问题，主动散热(风冷/液冷)技术如果能应用于智能手机，就会很好的解决这个问题。目前市面上的风冷/液冷解决方案体积大，功耗高，不适用于智能手机，急需设计出创新型的适用于手机的主动散热方案。		
研究内容	1. 研究适合智能手机的微型微泵风冷/液冷方案 2. 研究适合智能手机的其他低功耗主动散热方案 可以重点讨论微型风冷系统研究，比如微型收缩扩张型无阀微泵风道设计或其他高可靠性的有阀/无阀微泵风冷系统设计		
研究目标	1.新型主动散热方案需要小尺寸(厚度最好小于1mm)，低功耗(运行最大功耗<400mW) 2.主动散热方案需要降低智能手机表面温升2 ° C以上(与被动散热方案相比)		

技术方向二：核心部件

课题9	超薄轻质陶瓷材料开发	类型	攻坚类
研究背景	陶瓷材料具有高硬度，高耐磨，耐高温，耐腐蚀等特点。随着生活品质的提升，陶瓷用于笔记本外观件在提升外观品质的同时提升耐磨性并避免屏蔽信号等性能。同时陶瓷材料也是中国文化的象征，在笔电产品使用陶瓷材料提升品牌认可度。优化陶瓷材料现有配方及工艺，或开发陶瓷类外观效果品质的材料，使之从重量，强度，价格上适合笔记本电脑产品的应用。		
研究内容	基于陶瓷外壳满足笔记本电脑外壳使用，从以下四个方面优化开发新材料： （1）研究材料配方,满足笔记本电脑产品的使用 （2）研究材料的晶体结构，平衡晶相和玻璃相配比 （3）材料加工工艺研究，新材料具备量产性 （4）新材料后制程工艺研究测试并满足可靠性测试要求		
研究目标	关键技术指标如下： （1）外观效果达到传统陶瓷的质感 （2）产品尺寸：长度大于360mm，宽度大于250mm，厚度小于0.7mm，平面度小于0.5mm （3）材料密度小于2.0 g/cm3，弹性模量大于90GPa，表面硬度大于9H （4）材料能够进行CNC,能热弯成形四边曲造型,材料软化点小于800度 （5）单片（360×250×0.7mm）材料成本不高于50元（RMB） （6）2025年12月前具备量产性 （7）申请相关发明专利2-3项，专利归联想所有		

技术方向二：核心部件

课题10	**不锈镁（挤压态、压铸态）高品质、高颜值、高亮铝合金阳极外观表面处理技术研究、开发和产品化——电泳涂装技术**	**类型**	攻坚类
研究背景	随着消费笔记本轻量化、高品质外观日益强烈的追求，新一代高亮不锈镁材料得到市场的认可与追捧，其不仅具有高品质的金属外观，还大大降低了整机的重量，但随着用户对外观需求的不断升级，提升公司产品客户的满意度，急需在该材料基础上攻关新一代的表面处理技术从而在产品成本、外观品质、产品质量、用户满意度等方面实现产品竞争力。		
研究内容	研究新一代可适用于不锈镁合金的阴/阳极电泳配方及相关前处理工艺，达到可实现不锈镁高亮化处理的效果，实现用户对3C产品高品质需求的电泳涂装工艺，并实现镁合金材料领域中的创新，通过该课题开发可实现不锈镁高品质颜值的进一步优化，并进一步提升产品品质，降低产品成本，实现绿色涂装、节能环保的ESG要求。		
研究目标	目标需求: • 提供新一代阴/阳极水性电泳涂料及前处理配方及相关加工参数 • 涂装精致外观要求——铝合金喷砂阳极外观效果或其他丝绸质感效果 • 适合量产的低成本要求——电泳原漆（固含量60%）成本＜RMB120/Kg • 涂膜性能满足联想产品的信赖性测试要求 【研究成果】 1. 实现课题2025.12具备量产性，导入消费笔记本产品一个机型的应用落地。 2. 发表SCI期刊会议1-2篇，申请发明专利2项		

技术方向二：核心部件

课题11	**不锈镁（挤压态、压铸态）高品质、高颜值、高亮铝合金阳极外观表面处理技术研究、开发和产品化——涂料技术**	**类型**	攻坚类
研究背景	随着消费笔记本轻量化、高品质外观日益强烈的追求，新一代高亮不锈镁材料得到市场的认可与追捧，其不仅具有高品质的金属外观，还大大降低了整机的重量，但随着用户对外观需求的不断升级，提升公司产品客户的满意度，急需在该材料基础上攻关新一代的表面处理技术从而在产品成本、外观品质、产品质量、用户满意度等方面实现产品竞争力。		
研究内容	研究新一代可实现不锈镁基材表面无需经过前处理（化学转化、微弧氧化）直接进行美观涂装的新型涂料，通过该课题开发可实现优化镁合金工件的加工工序，提升工厂制程产能与良率，进一步优化产品成本，满足公司ESG在环保、节能方面的要求。		
研究目标	目标需求: • 提供无需前处理工序实现直接在不锈镁表面涂装的新一代涂料配方与加工参数（实现简化加工工序，提升制程产能与良率） • （高品质、高外观）涂装精致外观要求——铝合金喷砂阳极外观效果 • 涂装量产低成本化要求——14寸笔记本壳体＜RMB10 • 涂膜性能满足联想产品的信赖性测试要求 研究成果 1. 实现课题2025.12具备量产性，导入消费笔记本产品一个机型的应用落地。 2. 发表SCI期刊会议1-2篇，申请发明专利2项		

技术方向二：核心部件

课题12	铝合金表面抗指纹技术开发	类型	攻坚类
研究背景	随着人们对数码产品精致度的要求越来越高，笔记本外壳的颜色体系越来越多，针对某些颜色较深的阳极外观存在着行业内普遍的指纹印问题，如何通过低成本的技术方案有效的解决指纹印问题，降低指纹印的残留，提升指纹印的易擦拭能力，具有非常重要的意义。		
研究内容	1. 研究基于铝合金阳极表面处理技术的抗指纹工艺方案； 2. 研究铝合金原材合金成分、晶粒状况等因素对表面抗指纹是否存在影响； 3. 研究表面微观结构对抗指纹的影响； 4. 研究指纹印的系统性检测方案； 5. 研究抗指纹性能的有效检验标准。		
研究目标	1. 开发低成本表面处理抗指纹技术方案，基于360*250mm的壳体在一次阳极基础上单道工序目标价格＜3 RMB/pcs 2. 表面处理后，水滴角＞130°，钢丝绒耐磨3000次后，水滴角＞120° 3. 表面处理后，油滴角＞80°，钢丝绒耐磨3000次后，油滴角＞70° 4. 标准皮脂测试下，皮脂区和非皮脂区△E＜0.5，△G＜1 5. 申请发明专利≥1项，提交标准提案≥1项 6. 2025年12月技术具备可量产性		

技术方向三：智能服务与运维

课题13	面向异构计算平台的GPU故障预测	类型	应用类
研究背景	GPU作为其主要的计算加速器，在大规模集群中扮演着至关重要的角色。大模型训练时间比较长，2022年Meta 使用千卡A100训练175B OPT模型耗费了2个月，共发生了105次重启，最长健康持续训练时间为2.8天，零一万物千卡GPU集群故障频率为15~20次/月，GPU相关的故障作为集群故障中最频繁发生的故障，影响巨大。在异构计算平台中，针对异构的GPU的可能发生的故障，提前做出预测，调度训练任务实现智能化的断点续训。		
研究内容	研究GPU相关故障的类型和成因； 研究GPU相关故障的预测方法和预测模型； 研究GPU相关故障的恢复方法和断点续训策略。		
研究目标	1. GPU相关故障覆盖率100%，包括相关的软硬件故障；覆盖英伟达和至少一款国产GPU。 2. GPU相关故障实现提前5分钟预测，预测的准确率> 80%，召回率>60% 3. 发表人工智能领域顶级期刊会议的论文1-2篇（CCF-B以上），申请发明专利2项		

技术方向三：智能服务与运维

课题14	面向大规模的大模型训推一体化异构集群的故障智能运维方法	类型	攻坚类
研究背景	随着人工智能技术的快速发展，特别是深度学习模型的规模不断扩大，对计算资源的需求急剧增加。大模型（如GPT-3、BERT等）的训练通常需要数千甚至数万张GPU卡，这对集群的运维管理提出了前所未有的挑战。 现代计算集群通常包含多种类型的计算资源，如CPU、GPU、TPU等，以及不同架构和性能的硬件。这种异构性增加了运维的复杂性，需要系统能够有效地管理和调度这些不同类型的资源。因此传统的运维方式在面对大规模、高复杂度的集群时显得力不从心。人工运维不仅效率低下，而且容易出错。因此，需要开发智能化的运维系统，以提高运维效率，减少人为错误，并确保集群的稳定运行。此外，大模型训练和推理的成本高昂，包括硬件投资、能源消耗和维护成本等。高效的智能运维系统可以帮助优化资源使用，减少不必要的能源消耗和硬件损耗，从而降低总体成本。最后，在大规模集群中，硬件和软件故障的发生几乎是不可避免的。智能运维系统需要能够快速定位故障原因，并自动或半自动地进行故障恢复，以减少故障对训练和推理任务的影响。因此，通过引入智能化、自动化的运维方法，结合最新的云原生技术和AI技术，来构建一个高效、可靠、安全的大规模AI集群运维系统，成为当前AI基础设施领域的一个重要研究方向。		
研究内容	研究大规模异构AI集群的故障指标与日志序列建模算法； 研究基于机器学习的多维度日志分析与故障预测算法： 通过开发先进的故障指标日志序列建模算法和多维度日志分析与故障预测算法，我们能够显著提升大规模异构AI集群的智能运维能力。这些方法不仅能够高效处理和理解海量复杂的日志数据，还能准确预测潜在故障并提供深入的根因分析。这将大幅提高系统的可靠性、可用性和资源利用率，同时降低运维成本和人力需求。		
研究目标	1. 研究适用于千卡乃至万卡规模异构AI集群的故障指标日志序列建模算法，实现每秒处理至少百万条日志的能力，日志语义理解准确率较传统方法提升20%以上。 2. 研发基于机器学习的多维度日志分析与故障预测系统，故障预测准确率达到90%以上，预测提前时间不少于30分钟，相比传统规则基础方法，降低误报率50%以上。 3. 发表人工智能/分布式系统领域顶级期刊会议论文1-2篇，申请发明专利2-3项。		

技术方向三：智算平台

课题15	面向RAG的融合知识图谱的向量语义检索	类型	应用类
研究背景	RAG（Retrieval Augmented Generation）通过检索阶段获取相关信息，然后在生成阶段生成准确、连贯的文本，帮助大模型更好地理解和生成高质量的内容。传统RAG通过向量检索基于语义相似度对大规模文本进行高效检索，但无法完全捕捉复杂的语义关系。GraphRAG利用知识图谱获得更准确和更全面的信息，但面临成本和计算效率的极大挑战，难以应用落地。		
研究内容	研究高效的知识图谱与向量语义检索的融合技术，实现在检索效率和检索质量方面的动态平衡，实现向量数据库自适应索引技术		
研究目标	1. 研究文本向量与已有知识图谱之间的关联和映射 2. 研究向量语义检索过程中如何融合考虑语义相似性及在知识图谱中的概念相关性 3. 开发原型验证系统，支持百万级向量及百万实体/关系的知识图谱的融合搜索，实现低于20ms的单个向量搜索时延，及对比基于单纯向量语义检索的RAG实现10%-20%生成质量提升 4. 发表数据库/人工智能领域内CCF-A期刊/会议论文1篇		

技术方向三：智算平台

课题16	局域网内GPU AI算力分享	类型	攻坚类
研究背景	随着边缘计算技术的快速发展，DT PC在各类应用场景中发挥着越来越重要的作用。AIGC等需要高性能计算和实时响应的场景也对边缘端算力提出了更高的要求。在AIGC等场景中，GPU因其强大的并行计算能力，成为实现高效渲染和AI模型推理等任务的核心硬件。然而，传统的集中式计算模式在面对实时处理需求时，往往会遇到算力不足等问题。因此，基于DT PC的边缘端算力分享方案成为一种理想的解决途径。 本课题旨在研究和开发一套局域网内GPU AI算力分享的方案，以提升边缘端设备的计算效率和性能。通过这种方案，局域网内的DT PC可以共享GPU资源，从而在渲染、AI模型推理等高计算需求的任务中实现低延迟、高带宽、快速响应的目标。		
研究内容	本课题主要聚焦于以下2个关键部分，以提升局域网内GPU资源的利用效率和任务执行性能，技术方向包括但不局限于以下描述： 1. GPU算力资源管理与多资源调度策略研究：研究针对单GPU多任务或多GPU多任务场景的GPU资源监控和调度策略，结合多资源调度（resource scheduling）理论和智能调度算法，优化多任务环境中的GPU算力利用率。 2. 计算任务数据传输优化方法研究：研究不同设备之间AI模型参数的传输方法优化，结合压缩算法、传输协议优化和并行传输技术，分析传输过程中的瓶颈问题，提出减少传输时间的技术方案，提高模型参数同步效率。结合数学模型和实验分析，验证传输优化方法在大规模数据处理中的性能提升，解决现有传输方法的效率低下问题。		
研究目标	1. 研究一种通用且高效的GPU远程调用机制，使其在多样化的硬件环境中表现稳定，同平台GPU调用速度提升到现有方案（CUDA库重定向）的1.5~2倍。 2. 研究任务切分、智能调度和负载均衡算法，优化传输协议、采用参数压缩算法等使多机AIGC任务（单GPU多任务或多GPU多任务）的资源利用率与单机GPU训练推理效率损失不超过30%。 3. 发表高性能计算/分布式计算/人工智能领域顶级期刊会议1-2篇，申请发明专利1项，提交标准提案1项		

新兴技术

课题17	基于用户认知交互的裸眼3D显示器3D空间界面设计研究	类型	应用类
研究背景	裸眼3D显示技术无需佩戴任何辅助设备即可呈现具有深度感的立体图像，显著提升了用户的视觉体验和交互感知。然而，当前在裸眼3D的3D空间界面设计上仍面临诸多挑战： 1、界面和体验的舒适性： 如何通过界面设计降低用户的认知负荷，提升操作的自然性和直观性，并确保长时间使用的舒适性？鉴于裸眼3D显示器的视场角（FOV）相对有限（通常在33度左右），需要研究如何在有限的FOV内设计出高效且用户友好的3D空间界面。 2、空间设计的沉浸感和新奇体验： 传统的2D界面设计方法无法充分利用裸眼3D显示技术的优势。单纯的2D界面或2D与3D结合的界面在视觉效果和交互方式上存在诸多限制。需要探讨如何通过界面设计营造出令人惊叹的空间感，从而提升用户的体验感受；并且利用先进的交互技术，如手势识别、眼动追踪和语音控制等，打造更加智能和便捷的交互方式。		
研究内容	1、用户空间认知模型与适应性设计： 研究用户在3D空间中的认知模型，包括深度感知和注意力焦点等方面； 2、用户体验和新颖的设计分析： 强调界面的互动性和沉浸感，提出创新的交互方式和用户体验；		
研究目标	1. 用户空间认知模型与适应性设计：研究用户在3D空间中的认知模型，包括深度感知和注意力焦点等方面；评估深度感知的准确性和用户的主观满意度。探索如何降低用户在3D空间中的认知负荷，确保界面操作的自然和直观；通过主观问卷或生理指标测量用户对空间界面的认知负荷，目标是NASA-TLX评分低于50，或其他同等指标。 2. 用户体验和新颖的设计分析：强调界面的互动性和沉浸感，提出创新的交互方式和用户体验；分析用户在3D空间中的行为模式，为界面设计提供数据支持。研究3D界面元素的新颖设计，如光标等的呈现方式；确保在不同视角范围内，界面的文字和图形保持清晰度。评估界面的新颖和惊奇体验，通过皮肤电导率增加0.05-0.5微西门子（μS）或心率增加10-20次/分钟等指标证明惊奇体验的存在。 交付： 1. 论文：发表联想认可的CCF-A类会议、期刊论文1-2篇 2. 报告：提交“裸眼3D显示器上界面设计的最佳实践”、“裸眼3D空间界面设计指导原则等报告1-2篇 3. 专利：提交专利2项		

赋能现在 布局未来