CIPS-LMG 华为云脑洞基金
2025 年申报课题介绍

本年度将围绕基于昇腾算力的大模型训推、Agent、多模态、具身智能等研究方向，优先面向以下课题及研究目标征集 Proposal，包括但不限于：

课题类别	课题
模型训推优化	面向高效推理的大模型注意力机制探索
	长序列能力增强技术探索
	Reasoning 模型高效思考机制探索
	极低 bit 量化压缩机制探索
多模态	多模态文生视频扩散模型加速方案探索
	云上多模态生成推理跨请求优化探索
昇腾云算子	融合算子的“化学反应”研究
AI 应用	AI 自动生成竞赛赛题
具身智能	具身智能 Real2Sim2Real 可用性探索
自由探索	包括不限于 AI 基础设施、大模型、媒体、具身智能等方向

课题 1：面向高效推理的大模型注意力机制探索

项目背景：

随着能力的不断增强，大模型已被广泛应用于搜索、对话、代码生成等场景，推理效率成为衡量大模型价值的关键指标。当前大模型主要基于 Softmax Attention 机制，但其计算复杂度随序列长度呈二次方增长，导致该架构长文本处理面临严重瓶颈；其次，由于自回归生成需要缓存大量的历史 Key Value 状态，显存压力巨大，对于昇腾等架构不友好。业界不断探索对大模型架构的创新以优化推理效率，如 GQA, MLA，线性 Attention，稀疏 Attention、混合架构等。

然而，结构探索的试错成本高昂，重新训练百亿级新架构模型，耗时可达数月，新的结构难以得到充分的验证。最近，研究者开始探索使用迁移学习的

方式，将现有架构的大模型权重迁移至新的架构的方式。该策略能大幅削减预训练开销，为新架构的快速迭代与实际落地提供了切实而有前景的方向。

技术挑战：

1、推理高效的架构设计：基于昇腾硬件，设计高效大模型架构，具有与 full softmax attention 接近的准确率、同时具有更低的计算复杂度，并且是硬件亲和的，如支持分块并行计算、序列并行等特性。

2、高效跨架构迁移学习：不同的模型架构无法直接复用权重，需要针对性设计迁移学习方案，在成本可控的前提下，减少精度损失。

3、算子硬件协同：新的模型架构存算特征与现有架构不同，当前算子效率低甚至无法使用，需要构建高效高性能的算子，最大化利用昇腾硬件的能力。

项目目标：

1、设计面向昇腾硬件高效推理的注意力机制及配套训练推理算子实现，实现 XX B 级别模型在同样集群大小、SLA 限制下，单卡吞吐提升 X 倍。

2、设计端到端迁移学习方案，使用原始预训练 2%以内的成本，实现 XX B 级开源 SOTA 模型迁移至新架构，在常见下游任务评测集上精度下降不超过 1 % 。

课题 2：长序列能力增强技术探索

项目背景：

传统大语言模型在处理长序列时面临效果瓶颈：随着输入长度增加，模型捕捉远距离依赖的能力显著下降。这主要源于注意力机制稀释、位置编码外推失效、信息累积丢失、甚至文本中存在较多噪声数据的问题。这直接导致在长文档问答、摘要、连贯性生成等任务中出现事实错误、前后矛盾、关键信息遗漏等问题。为提升长序列下的建模效果，研究者重点突破方向包括：改进位置编码，使模型能更好地理解长距离词序关系；引入层次化或稀疏注意力机制（如 Longformer），聚焦关键区域；设计递归或记忆机制（如 Transformer-XL 的分段记忆），显式保留历史重要信息；以及长文本针对性训练（如课程学习、构造长依赖任务）。尽管像 GPT-4、Claude 等模型已支持超长上下文（如 100K+ token），但在保持超长文本中事实一致性、减少“幻觉”以及复杂多跳推理等方面，效果仍有待提升。

技术挑战：

1、长距离依赖建模失效：注意力稀释，随着序列增长，标准注意力机制中关键信号被大量无关词稀释，导致模型难以捕捉远距离词间关系。位置编码外推崩溃，传统位置编码（如 Sinusoidal、RoPE）在训练长度外的位置失效，破坏词序理解。

2、信息累积与一致性衰减：关键信息丢失，模型在超长文本中难以持续记忆和整合分散信息，导致生成内容前后矛盾。认知偏差放大：长上下文加剧"幻觉"问题，尤其在多跳推理任务中错误累积。

项目目标：

1、面向大模型长序列场景，提出一种去噪方案，显著提升大模型在超长文本中的关键信息捕获能力，在下游任务（例如 LongBench-E）上精度提升 10%+。

课题 3：Reasoning 模型高效思考机制探索

项目背景：

当前，推理大模型（Large Reasoning Models, LRMs）在数学、编程等复杂任务中展现出卓越的多步推理能力，但同时也普遍存在思考效率低下的问题。这类模型在生成推理链（Chain-of-Thought, CoT）时，往往产生大量冗余内容，例如反复复述问题定义，这些内容对最终答案帮助不大，却增加了计算成本。这些模型还会对简单问题进行不必要的过度思考，难以根据任务复杂度有效分配思考预算，即使是简单问题比如简单的加法运算，也可能生成多轮冗余的验证步骤。同时，LRMs 还存在欠思考问题，模型会频繁地切换思考方向，推理过程变得浅显、碎片化 ，导致推理过程冗长且低效。

技术挑战：

1、推理效率量化：如何平衡精度和推理效率是一个巨大的挑战，当前推理大模型依赖较长的思维链去反复思考，思维链每一步的必要性和对最终结果的贡献很难评估，这使得如何精确地判断哪些部分可以压缩成为了一个难题。同样地，无法量化推理效率也会引发推理长度难以控制的问题，如何让模型思考得恰到好处，既不太浅以致遗漏逻辑，也不太深以致浪费计算，是推理大模型的重要挑战。

2、超越 Transformer 架构瓶颈：现有 LRMs 大多基于 Transformer，其二次复杂度在处理数千甚至更多 token 的长推理链时成为严重瓶颈。开发能够处理长序列的新架构或高效近似方法至关重要。

3、跨任务泛化：不同任务需要不同的推理深度。单一的推理策略或长度策略难以适应所有任务。如何在保证跨领域鲁棒性的同时实现效率，是一个复杂挑战。

项目目标：

1、提出一种新的 Reasoning 模型高效思考机制，量化描述 LRMs 思维链推理效率，在数学、编程等复杂推理任务上 JCT (Job Completion Time) 降低 50%。

2、提升 LRMs 推理效率的同时，保持精度下降在 1%以内。

课题 4：极低 bit 量化压缩机制探索

项目背景：

DeepseekV3/R1 中 FP8&BF16 混合精度、MLA 等的应用展示了低精度数值计算在实际业务中的显著价值，可以帮助大模型在实施过程中降低成本、提升性能、扩展应用场景。低 bit 量化是一种关键技术手段，可以带来：

1、 成本降低：极低 bit 量化压缩可以带来计算资源节省、存储需求减少。

2、 效率提升：通过极低 bit 压缩机制，满足模型推理效果的同时，通过低精度计算加速推理，有效提升推理性能。

3、 更大规模模型部署：降低了模型的内存大小和计算资源依赖，使得有限资源下大规模模型的部署和运行成为可能。

技术挑战：

1、 精度保持：压缩率越高信息丢失越多。在极低 bit 下如何设计压缩机制和推理计算策略使得模型信息更多的保留、保持更高的精度是难点问题。

2、 算法优化：复杂的算法虽然精度更高，但一定程度上以损失性能为代价，需要寻找算法复杂度与和推理性能的平衡，优化量化算法和推理计算，确保推理过程高效运行。

3、 多样性支持：不同模型或同一模型不同层/结构对极致压缩的敏感程度不同，如何快速找到最佳压缩机制，满足精度和性能的极致要求。

4、 硬件亲和：量化后的模型高效推理往往需要硬件配套低精度计算的支持，极低 bit 量化压缩机制如何做到硬件加速亲和，如下一代昇腾芯片的支持。

5、 评价体系：探索模型有损的各种优化措施对模型性能的定量影响，构建完善评价体系。

项目目标：

1、 设计极低 bit（FP8/FP4/INT8/INT4/INT2 等单一或混合）量化压缩机制，支持 DeepSeek、LLaMA 等多种结构热点模型， 支持可选量化 bit 输入(单个或多个），较 BF16 精度模型，挑战 3-6 倍压缩。

2、 极致压缩精度误差保持在 1%以内（较比 BF16），推理性能提升 1.4+倍（较比 W8A8）。

3、 量化速度快，600B 及以上模型，实现<1 天级别极速压缩效率。

4、 量化压缩技术对昇腾芯片亲和，支持昇腾芯片高效运行。

课题 5：多模态文生视频扩散模型加速方案探索

项目背景：

随着生成式人工智能的快速发展，多模态文生视频（Text-to-Video，T2V）扩散模型在影视制作、虚拟数字人、互动娱乐等领域展现出巨大潜力。然而，此类模型因需处理高维时空数据，存在计算复杂度高、推理速度慢、显存占用大等瓶颈。当前主流模型生成数秒视频往往需分钟级时间，严重限制了实时交互与应用落地。针对这一挑战，探索高效的模型加速方案成为关键。本项目旨在系统研究面向文生视频扩散模型的优化技术，涵盖步数蒸馏、分段推理、cache 寻优、稀疏 Attention 等方向，突破计算效率与生成质量的平衡难题，推动 AIGC 技术从实验室走向实际应用场景，为产业提供低成本、高性能的视频生成解决方案。

技术挑战：

1、动态质量评估失准：传统图像指标无法捕捉运动合理性与时空连贯性，加速后模型因细节简化导致评估分数虚高，掩盖物理规则违反（如违背重力定律的运动）和语义失真（如动作对象错位）。

2、时空连续性崩塌：加速操作（步数蒸馏/Cache 复用）破坏帧间运动微分约束，引发跨帧抖动链式反应，空间抖动：物体边缘闪烁（像素偏移≥15px/帧）；时间失真：运动轨迹异常（如自由落体出现悬停）。

3、语义对齐问题：由于自然语言具有复杂性和多义性，步数蒸馏/稀疏 Attention 等方案会加剧语义信息与视频内容的对齐问题。

4、计算与显存平衡：cache 复用需要保存多步推理的结果，会显著增加推理显存，存在显存溢出风险。

项目目标：

1、针对 Diffusion 模型架构，提出一种即插即用的 Diffusion 模型加速方案，在 wan2.1 14B 模型上实现推理性能提升 5 倍，并且在照片生成视频场景下合成视频质量下降控制在 3%以内。

课题 6：云上多模态生成推理跨请求优化探索

项目背景：

随着多模态生成模型能力的增长，多模态推理从 demo 逐步转为大规模生产。本课题主要针对现在相对成熟的 DiT 模型。DiT 模型通过把在 latent space 下进行多轮去噪，迭代产生完整的图片或者视频，为严重的计算密集型任务（类比 LLM 推理的 Prefill 而不是 Decode 阶段），而且隐空间的序列长度随着图片/视频分辨率以及视频的长度增长。

现在存在大量针对推理过程中进行优化的算法，包括对模型推理过程中，跨层[1] 或者跨去噪周期[2] 复用的工作。但是这些工作通常忽略了挖掘请求之间的相关性。就像 LLM 推理中的多轮对话、前缀复用、以及 batching 策略等工作一样。在云上场景，不同于学术工作针对的少量显卡，是大的资源池，处理大量、多组任务。并行、请求间关系的挖掘至关重要。

技术挑战：

1、针对用户多模态生成请求，通过挖掘请求之间关系，实现中间状态复用等优化。比如两个租户生成类似的视频（比如基于同样模板生成视频），可以复用前几步骤的中间结果。但是该优化的一个重点挑战在于，语义级别的复用，不同于 LLM 推理的前缀匹配，如何决定什么样的请求，复用多少中间状态，保证结果的质量。

2、另外在调优算法的过程中，也缺乏自动化的方法来评判复用的质量。

3、在此基础上，如何设计并行策略，让相关请求并行或者在相同资源上进行运行，实现极致性价比。

项目目标：

1、提出一种系统化识别请求之间相似性、并进行复用推理优化的系统化方法。（鉴于提升性能与请求负载高度强相关，在没有一个基线负载情况下，不做量化指标，没有意义）

2、提出一种质量评估机制，能够自适应调优复用策略。

3、提出一种基于复用的并行化策略，性价比进一步提升 50%。

课题 7：融合算子的“化学反应”研究

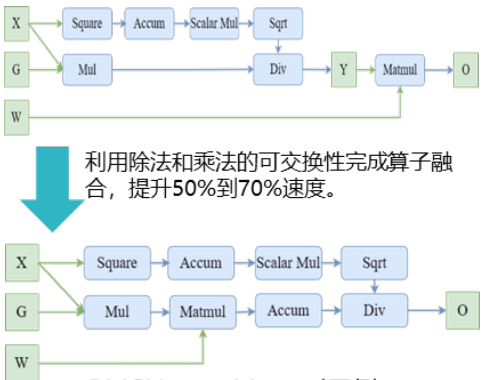
项目背景：

在 NPU 算子优化过程中，**人工进行单算子优化通常只能带来边际性提升。**然而，在算子融合的广泛空间中，某些算子组合对性能的提升具有显著影响。通过对算子进行联合优化，可以突破单独优化带来的瓶颈，进而实现“**1+1>2**”的性能增益。当前，由于缺乏有效的方法来**发现融合算子空间中的关键组合并构建联合优化技术方案**，算子性能优化受到限制，进而影响整体模型的性能表现。

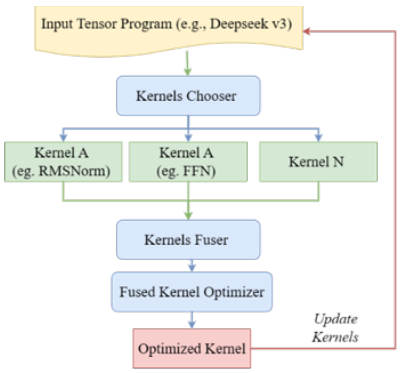
技术挑战：

1、搜索空间庞大（Chooser and Optimizer ）：整个网络通常由大量算子组成，因此优化的搜索空间包含了两个主要部分：1）对已有算子的组合空间，2）融合后算子的优化方案空间。如何高效的设计两个部分的搜索是找到化学反应的关键。

2、通用性挑战（Generalization）：在完成一个融合算子的“化学反应”研究后，如何进一步将其推广应用于其他算子组合，进而提升整个优化过程的通用性和可扩展性，仍是一个重要挑战。



RMSNorm+Matmul示例



从网络中自动识别算子组合进行“化学反应”优化

项目目标：

1、构建 AscendC 编码大模型，根据自然语言/python 表达需求，生成的基于 AscendC 的 NPU 内核代码编译通过且功能正确率> 80%.

2、构建 AscendC 优化智能体，通过多智能体协作、演化学习等方式，优化当前算子代码，算子性能平均提升> 30%+。

课题 8：AI 自动生成竞赛赛题

项目背景：

大语言模型（LLM）在教育、招聘等场景应用深化，但传统赛题开发依赖人工、成本高且原创性不足。华为等企业招聘需大量原创编程题，而高校课程、竞赛等场景也存在庞大题库需求，市场规模约130亿。当前LLM生成题目存在经典题重复、难度控制不稳、测试用例不可靠等问题。亟需自动化生成高原创性、难度可控且评测完整的赛题方案，以提升效率满足多元需求。

技术挑战：

LLM 的逻辑推理能力迅速提升：公开测试显示，部分旗舰模型在算法解题任务上已达到国际信息学奥赛（IOI）金牌水平。这使得 LLM 的应用从 AIGC、代码补全逐步扩展到“问题生成（Problem Generation）”领域。尽管当前主流大模型可在专家提示下生成题目，但仍存在以下局限：

- 倾向输出经典或已公开的“老题”，原创性不足；
- 难以精准控制题目难度与知识点覆盖；
- 生成的测试用例和参考答案可靠性不稳定，需要人工大量复核。

综上所述，业界急需一种能够自动化、批量化生成高原创性、难度可控且评测完整的赛题技术方案，以降低人工成本、缩短出题周期，并满足教育、竞赛与招聘等多元场景的快速迭代需求。

项目目标：

1、建立一个评估出题的评估体系和方法论，建立对应Benchmark：确定关于评估出题质量，比如难度，原创性，标程正确性，测试用例覆盖度等评价指标。

2、实现一个AI自动出题的workflow流水线，可自动生成题目、题解、用例和标准程序。借助大语言模型、Agent、强化学习等技术，自动生成原创性和质量都较高的题目、题解。题目应有不同难度，研究比较国内外不同的大语言模型，

对不同题目难度和不同算法知识点方面的优势，进行自动的推荐，或是给出多个题目供出题专家选择。

- 出题难度基准目标：可对标校招笔试题和ICPC校赛难度或leetcode上的hard难度。
- 出题难度挑战目标：ICPC区域赛难度。
- 自动生成测试用例目标：应覆盖足够的数据范围和边界条件，以适应不同算法的时间复杂度和空间复杂度要求。测试用例的文件大小需要控制在一定范围，以满足不同OJ的需要。
- 自动生成标准程序目标：可选择C++、java、Python语言。

3、牵引挑战目标：在AI自动出题领域，可以产出有影响力的论文或专利成果，只作为目标牵引，不作为交付要求。

课题 9：具身智能 RealSim2Real 可用性探索

项目背景：

针对具身智能领域 VLA 模型训练面临的高成本数据采集问题，云端仿真平台可通过 7×24 小时持续运行，以低成本生成多样化的场景数据。然而，传统仿真环境与真实物理世界间存在动力学参数、传感器噪声、环境扰动等多维度的偏差。通过将真实环境数据引入仿真器中，能够构建一种通用的仿真器校正范式，缩小 Real2Sim2Real 的鸿沟，为后续的真实环境部署奠定坚实基础。

技术挑战：

1、构建虚拟仿真与物理世界的通用校正范式：业界现有校正方法多针对特定场景、特定任务或特定参数设计，且存在大量人工介入，难以应对多模态数据差异与复杂环境的泛化挑。

2、如何使用真实数据减小物理仿真 gap：仿真器的物理参数与真实场景下的差异较大，业界现有缩小 sim2real gap 的系统辨识法和残差网络法可扩展性较差。

3、如何使用真实数据减小渲染仿真 gap：业界主要是通过给机器人观测添加域随机化来降低 sim2real gap，无法还原现实中的光照、材质、传感器噪声等信息。

项目目标：

1、通用仿真器物理动态校准方法：提出一种不针对具体任务的通用校准方法，通过采集真实机器人在操作过程中的传感器数据（如轨迹和视频等）或者是人类遥控操作的数据，通过数据驱动的神经网络校正物理引擎参数，并补偿仿真与现实之间的差异，使得轨迹误差降低 25%，提升仿真器的物理精度。

2、通用仿真器渲染动态校准方法：采用生成式模型、三维高斯泼溅等方法学习真实图像传感器的数据分布，提升机器人在仿真器中观测的真实性，仿真和真实的图像观测相似度指标(如 FVD)提升 30%。

3、基于校准后的仿真器生成的数据经过模仿学习/强化学习方法训练出来的策略迁移到真实场景的成功率提升 50%。

自由探索

本基金特设立自由探索类 topic，围绕云计算领域，鼓励提出改变“游戏规则”、打破常规，甚至颠覆云计算领域现有架构、技术、商业模式的 IDEA。创新方向包括但不限于 AI 基础设施、大模型、媒体、具身智能等方向，需考虑创新性、可行性、商业价值等多个维度。