MoternAI 2024年总结

该报告旨在总结MoternAI于2024年度开展的算法项目

多模态llm领域

出于实现精彩秀智能剪辑中视频感知能力的目的，我们尝试训练一个可以感知台球场景视频的语言模型：

HiLight一期：视频语言模型（VLM）	**24.1-24.7**

我们首先实现一个更细粒度模态对齐的VideoEncoder：在CLIP-ViP基础上引入Patch与Token之间的对比学习损失。其次使用改进后的CLIP-ViP和Long-CLIP组成视频-图像双塔视觉提取模块，经过特征融合后作为Gemma-2B语言模型的视觉输入，实现基于视频理解的对话能力。

项目技术报告：https://arxiv.org/abs/2407.07325
项目地址：https://github.com/motern88/HiLight/tree/main

积累技术储备：

- 预训练对比学习Loss改进
- 模型特征可视化监控
- 适配大规模多模态数据的Dataloader
- torch DDP和DeepSpeed单机多卡分布式训练
- 多模态LLM多轮对话的推理实现
- LoRA微调

该项目缺陷复盘：

- 相关领域经验不充分，且缺少领域内资深专家的技术指导，我们直到项目末期才意识到只有Grounding级别的VLM才能解决该问题。导致我们最终的结果严格上没有对齐项目开始时的目标（目标经过三次变更，即：1. 对对比学习训练后，特征相似性；2. 基于语言理解的VLM ； 3. Grouding级别的VLM。到目前为止，完成到第二阶段，且效果不好）。
- VisionEncoder预训练经验的不充分，与LLM微调经验的缺乏，导致期间反复debug并重新训练。
- 没有正确评估团队产出，导致错误的项目进度规划，此后项目应当分阶段目标。

该项目收益总结：

- 从0到1积累LLM及多模态模型的设计、训练和推理实现经验。掌握拥抱LLM时代的门槛级技术的运用。
- 从0到1建立MoternAI团队协作机制。从算法进展表、Issue记录表的共享管理，实现明确任务总量，步骤拆分，以及状态追踪。

具身智能领域

为了能在将来面对具身智能产业大爆发时 “上桌吃饭”，MoternAI针对具有代表性的相关机器人算法开展研究，积累技术储备：

Robot Action项目：机器人动作生成算法	**24.7-24.11**

我们在不连续的时间内先后探索了Aloha机器臂上ACT算法相关超参数调试，探索和验证了在该机械臂上移植DiffusionPolicy算法。探索了基于瑞尔曼搭建自制机器人（可移动底座+双臂，预计加装头部可动摄像头），同时规划仿真平台端-实体机器人端-人类操作端的三端操作互通方案，以支持多样跨源数据的训练和更强的跨环境算法迁移能力。最后尝试在Aloha机械臂上部署RDT-170m。

积累技术储备：

- 了解具身智能的基本实现框架，增进对ACT/DiffusionPolicy/RDT等主流机器人算法及相关领域进展的了解。

该项目缺陷复盘：

- 硬件以及通信知识缺乏
- 缺乏对仿真环境搭建与操作的经验。
- 最终RDT-170m的部署止于训练和推理代码跑通，未真正执行成功过任务；DiffusionPolicy的迁移止于训练和推理代码跑通，未真正执行成功过任务。

该项目收益总结：

- 在Aloha机械臂搭建的投入下，我们先后完成了一系列机器人动作生成算法的部署和调试，熟悉了该领域主流的相关算法，积累了MoternAI在具身智能领域相关经验。

视频对象追踪VOS领域

SAM2的出现使得根本性解决单视角物体追踪成为了可能，我们利用已有的检测模型最大化发挥SAM2的能力，并在此基础上实现一系列工程优化：

Det-SAM2项目：自提示长视频分割框架	**24.8-24.11**

我们基于SAM2实现了一个实时推理视频流并由检测模型自动提供条件提示的长视频分割pipeline，是第一个功能完备的长视频自提示SAM2框架。其支持推理过程中在线添加新类别、将推理上一段视频的记忆用于新视频推理（预加载记忆库）等新功能。同时，我们实现了在完整Det-SAM2-pipeline上恒定的显存和内存开销，从而支持一次性推理无限长的视频。

项目技术报告：https://arxiv.org/abs/2411.18977
项目地址：https://github.com/motern88/Det-SAM2/tree/main

积累技术储备：

- 对SAM2系列模型结构原理掌握。

  SAM2模型构成与推理传播方式，以及其在视频模态中通过记忆机制完成时间维度的跨帧关联的实现方式。

- SAM2的工程优化经验。

  通过不断释放旧帧缓存与记忆来控制恒定的显存和内存开销；通过控制存在检测模型输入的条件帧与其他非条件帧比例来调整SAM2对提示依赖程度；通过准确的帧索引计数器防止预加载记忆库与本次推理新生成的记忆库的混淆。

该项目缺陷复盘：

- 纯工程优化，没有本质解决SAM2的运行效率以及其在专有场景上精度问题。
- 对SAM2整体理解仅限于组件构成和推理传播方式，对于SAM2提示交互机制与掩码模糊匹配的核心（prompt encoder + memory bank）几乎不了解。
- 没有解决SAM2接受实例对象与DetectionModel输出类别对象的衔接问题。当前仍然需要保证检测模型中的每个类别仅出现一个样本。根本解决方式就是基于DetectionModel的追踪框架从头训练实现一个专属于此的MemoryBank组件，而非依靠SAM2的MemoryBank与DetectionModel强行拼接。

该项目收益总结：

- 通过对SAM2在工程优化期间的深入掌握和了解，奠基了我们对SAM2模型层面轻量化的后续优化思路。
- Det-SAM2项目切入准确开展及时，是首个实现持续性全流程自动化（包括初始提示与后续修正）的SAM2框架，其包含的一系列工程优化，也是同期SAM2后续工作中最靠近长视频VOS业务落地的开源框架。

小铁基础业务算法

小铁基础业务算法群多采用成熟、主流、稳定的算法框架，其完成了对小铁台球AI相关功能的基本支持，也实现了对部分新产品线的技术原型支持：

PS：时间周期超过一个月、技术难度跨度较大且属于突破类的项目将不在下列表中记录。

小铁基础业务算法	**24全年**
锁球器迭代优化	24.1-24.9
智能剪辑【一期】进球检测	24.3-24.5
AI教练/裁判【一期】（球桌、台球、袋口检测）	24.4-24.9
AI教练【二期】（球杆方向估计）	24.4-24.9
智能剪辑【一期】工程化移植	24.7-24.8
导游机：RAG+LLM	24.8-24.8

小铁基础业务算法	24全年
AI教练【二期】新增色卡修正方案	24.9-24.9
AI教练【二期】（手架、架杆检测）	24.10-24.10
AI裁判【二期】新增双摄像头追踪方案	24.12-25.1
AI教练【二期】摄像头/投影仪自动校准	24.12-25.1
打卡机：SD图像转绘工作流	24.12-25.1

项目细节：

- 锁球器迭代优化

  共更新模型权重3次，目前预测球数accuracy稳定在~96%。

- 智能剪辑【一期】进球检测

  将进球事件抽象为“球消失”，基于目标检测的结果增加后处理逻辑，通过每局球局中球数量的变化，给出正常球局中的进球时间点。部署方式为云端部署。

- AI裁判/I教练【一期】（球桌、台球、袋口检测）

  前期教练模型所有球为一个类别ball，后期由于业务需求，区分每一个球类别的裁判模型更通用，目前模型层已采用一个裁判模型合并两个需求。

  共在3个环境采集了1K+图片训练了19分类模型。

- AI教练【二期】（球杆方向估计）

  训练了一个球杆分割模型，可用于估计球杆出杆方向。

- 智能剪辑【一期】工程化移植

  将智能剪辑算法移植至Linux分析盒，实现视频解码、与终端通讯等相关需求。

- 导游机：RAG+LLM

  调研多种RAG框架，尝试llama.cpp与rwkv.cpp连接langchain的方案。

  详情见：https://github.com/motern88/AI_Conversation_in_Motern/blob/main/24-8-26/24-8-26.md RAG部分

- AI教练【二期】新增色卡修正方案

  使用摄像头中色卡采样值来均衡摄像头画面色差，尝试修正画面色彩检测模型精度问题。

- AI教练【二期】（手架、架杆检测）

  为支持教练考级、裁判球权轮转等需求，模型新增了“手架”、“架杆”等类别，使得模型为21分类。

  待改进：目前“手架”类别 由于各人架杆姿势不同，出现在桌台边缘等原因，存在一定概率漏检（~10%）。

- AI裁判【二期】新增双摄像头追踪方案

  为解决球高速移动下产生形变拖影，无法通过单一的RGB摄像头实现目标检测的问题，新增了一个红外（黑白）高帧率摄像头，在于充分利用两个摄像头的优势，互补不足，建立一个稳健的多传感器跟踪系统。

  技术方案详见：https://github.com/motern88/AI_Conversation_in_Motern/blob/main/%E5%8F%8C%E6%91%84%E5%83%8F%E5%A4%B4%E6%96%B9%E6%A1%88_%E8%BF%BD%E8%B8%AA%E7%AE%97%E6%B3%95%E8%B0%83%E7%A0%94/12.04.md

- AI教练【二期】摄像头/投影仪自动校准

  获取摄像头视角下球与投影成像的偏差数据，计算摄像头视角偏差与投影仪设备固定偏差，求解单映性矩阵。自动校准求解步骤见：AI_Conversation_in_Motern/摄像头-投影仪自动化校准求解/自动化求解计算步骤.md at main · motern88/AI_Conversation_in_Motern

- AI打卡机：SD图像转绘工作流

  本地部署ComfyUI，并基于自带的API集成进AI打卡机应用

  积累技术储备：熟悉了工作流的使用及调参，增加自定义节点、工作流的导入导出

团队共识

- MoternAI在24年上半年的主要发展瓶颈是相关领域经验不足，因此曾强烈希望为团队引入技术专家以弥补；然而下半年随着MoternAI的团队成长，我们的协作能力和技术储备经验上升，当前团队发展的主要瓶颈转变为产能不足。

  具体而言产能不足是指，要克服当前公司产品推进的算法瓶颈，就要求我们以前沿技术原型为基础完成一系列调优、改进、轻量化部署等工作。在大部分情况下，由于产能不足我们只能实现阶段性成果，无法一次性在一个周期内完成全流程的研发。在这种状况我们会面临是否继续开展后续工作的尴尬抉择，这会对产品迭代和公司理想中的投入形成负反馈循环。

  我们希望增加用于推进产品算法瓶颈项目的资源，一方面减少业务上扯皮的消耗的时间，一方面有新的人员...

- 一些新产品线的开展在早期验证阶段严重缺乏AI功能开发的评估，往往在AI功能开发遇到困难时MoternAI团队才首次接入。这种时候已经没有时间走正常的开发流程搭建算法demo了，只能临时东拼西凑，欠下许多后续的技术债。但同时我们也不希望过于琐碎的AI功能都由MoternAI参与，因此新立项产品线早期AI功能的合理评估很重要（及时地预见问题，如果必须要介入开发就提早介入开发）。

- 一个误区是将我们今年在多模态LLM/视频追踪VOS领域的项目定义为与业务无关的探索。实际上，如果没有智能剪辑（精彩秀）的需求，如果没有AI裁判算法准确度的瓶颈，我们不会开展HiLight与Det-SAM2项目。两个项目的出发点均是与业务需求的强关联，而其阶段性成果暂时不能落地在业务中也是事实。根本原因我们在团队共识的第一点有所阐述——产能不足致使我们没法在一个时间周期内完成其所需的所有研发。

- 团队内部的沟通可以增强团队的协作与技术储备的共享，而团队与公司的沟通依然缺乏：如何将团队技术储备转化为公司的技术储备；如何在具体任务上对齐团队与公司的目标?

- 最后一点的团队共识是，尽管在一些任务上，当前学术界最前沿领先的模型也无法满足需求，在我们场景下我们也可以突破这个上限。我们实施的算法项目不会受到领域研究瓶颈的限制。