基于LMM多模态大模型的针对BLV人群的具身智能EI交互系统设计

- **背景调研**
  - AI类产品
    - 目目前尚缺乏有效的AI产品以显著改善视觉障碍（BLV）人群的日常生活质量
    - AI驱动的穿戴设备已经初具规模，且其未来发展前景被广泛看好
      - 参考产品
        - 以AI技术为基础的便携式设备（如AI Pin）
        - 各类智能穿戴设备（如Rabbit R1）为用户提供辅助功能
        - 集成AI技术的智能眼镜（如Ray-Ban Meta）为用户提供实时信息
  - BLV人群
    - 视觉障碍人群的规模相当庞大，亟需相关技术的关注与支持
  - LMM多模态大模型
    - 动态学习能力日益增强的LMM多模态大模型正在迅速发展并日趋成熟
    - 目前，多模态大模型在特定领域（例如，针对BLV人群的医疗器械）中的识别能力仍显不足，且准确率亟需提升
  - 具身智能EI
    - 具身智能（Embodied Intelligence, EI）已成为全球研究的前沿热点
    - 该技术已被广泛应用于人形机器人、自动驾驶汽车等多个领域
- **概念提出**
  - 市场需求
    - 视觉障碍人群对基于AI的辅助装置的需求愈发迫切
  - 基本形式
    - 该装置应具备良好的便携性，设计为一种新型可穿戴智能设备
  - 使用方式
    - 设备能够主动推测用户需求，并提供相应建议
    - 用户可通过语音指令进行控制

- 目标效果
  - 旨在为视觉障碍人群提供便利，促进其更好地融入社会生活与工作环境，提高其社会接纳度
  - 助力提升其自信心与自尊感

感知环境

- 语音
  - 设备通过语音唤醒功能激活
    - 利用关键词检测技术启动录音功能
  - 通过语音活性检测（VAD）技术判断录音内容的有效性
    - 判断录音内容是否包含人类语音信号
  - 进行语音识别，将录音内容转化为文本信息
    - 录音内容将被转换为可读的文字形式
    - 针对特定场景进行热词检测以增强语音理解能力
    - 应用降噪录音算法，优化人声识别效果
- 图像识别
  - 同时进行环境定位与三维模型构建
    - 在未知环境中创建带有定位信息的三维地图
    - 综合使用深度相机、激光雷达、红外相机及鱼眼相机等传感器进行环境感知
  - 视觉类LMM多模态大模型
    - 利用视觉领域的LMM多模态大模型进行图像信息的分析与理解
- 其他识别
  - 采用其他感知方式增强设备对环境的理解

决策与执行

- 具身智能Agent框架
  - 感知器（感知模块负责接收并处理环境信息）
    - 功能
      - 提供决策所需的数据支持
    - 硬件
      - 深度相机
      - 激光雷达

- 红外相机
- 鱼眼相机
- 光线传感器
- 陀螺仪传感器
- 磁力传感器
- 加速度传感器
- 温度传感器
- 气压传感器
- 心率传感器
- 麦克风

- 记忆系统
  - 存储底层子任务相关的短期记忆
  - 存储高层任务规划所需的长期记忆
- 决策引擎（云端的大型多模态模型）
  - 功能
    - 负责自然语言处理、计算机视觉和跨模态任务
  - 应用方式
    - 自动提供信息
      - 上传可供推断的传感器数据至云端
      - 根据周围环境推断应该进行何种主动行为
      - 根据历史行为推断应该进行何种主动行为
      - 根据其他传感器获得的环境及人体数据进行综合推断应该进行何种主动行为
      - 将决策的结果传输至终端
    - 按需提供信息
      - 利用麦克风获取使用者需求
      - 上传可供推断的传感器数据至云端
      - 根据周围环境推断语义
      - 根据历史行为推断语义
      - 根据其他传感器获得的环境及人体数据进行综合推断语义

- 过滤无用信息，找到使用者最为关心的信息
- 将决策的结果传输至终端

- LMM多模态大模型
  - CLIP
    - 识别图像，关联文本
  - OWL-ViT
    - 视觉对象检测
  - AudioLM
    - 理解输出音频内容
  - GPT
    - 文字理解
- 执行器
  - 功能
    - 为BLV人群提供方向指引、物品搜索及日常生活支持
  - 硬件
    - 震动马达
      - 指引方向
      - 危急警告
    - 扬声器及听筒耳机
      - 语音告知信息

- 具身智能数据集
  - 离散控制类数据
    - 记录人在不同场景下的离散动作序列
    - 模型学习得到更精准的执行方案
  - 图片数据
    - 多样化视觉输入
    - 复杂光照条件下的图像
    - 质量低、模糊、无意义的图像
    - 学习物体、环境、人类行为，形成更精准和个性化的视觉理解能力

- 语言数据
  - 自然语言指令
  - 对话历史
  - 情境描述
  - 收集自然语言指令和对话历史以优化人机交互
- 传感器数据
  - 收集各种传感器的数据以增强环境感知能力
- 端对端训练与微调
  - 进行模型的端对端训练与持续优化
- 具身实现
  - 上层任务规划
    - 从具体任务转化为可执行技能
  - 下层理解执行
    - 从技能转换为具体的响应行为

模拟场景

- 场景一：日常出行
  - 通过语音唤醒设备
  - 实时提供导航指引
  - 识别周围环境并构建三维地图
- 场景二：社交活动
  - 识别社交环境中的人物
  - 提供关于对话的背景信息
  - 协助用户进行主动交流
- 场景三：工作场所
  - 提供任务相关信息与指引
  - 辅助完成工作任务与沟通
- 场景四：紧急情况
  - 识别紧急情况并发出警报
  - 提供快速逃生路径指引

未来展望

- 技术进步
  - 致力于提升多模态大模型的识别精度
  - 优化硬件性能以增强设备的感知与响应能力
- 社会影响
  - 通过技术进步显著改善视觉障碍人群的生活品质
  - 通过普及相关技术，推动社会对视觉障碍人群的理解与接纳
- 可能的应用扩展
  - 探索该技术在教育和娱乐领域的潜在应用
  - 研究该技术与智能家居系统的整合可能性

结论

- 对本研究的主要成果及其社会意义进行总结
- 突出研究对视觉障碍人群的积极影响
- 未来研究方向和建议进行展望