**2.3 项目任务内容**

**任务10：研发AI代理应用**

**研究目标**

随着消费类视听终端（如电视、智能屏等）的普及，用户对个性化内容推荐、智能交互的需求日益增长。本项目面向传统视听终端操作系统存在数据处理能力和推理效率的限制，基于消费类视听终端操作系统的多模态输入输出和调用AI大模型等功能，通过任务分析，优化云端与本地协同推理机制，研究提出AI代理实现技术，实现轻量化的数据推理与动态推荐，提升智能交互体验。

**主要研究内容**

开展基于消费类视听终端操作系统的 AI 代理实现技术研究。基于操作系统的多模态输入输出能力，调用云端视听大模型等功能，深入研究 AI 代理实现技术，开发面向该操作系统的 AI 代理应用。通过构建任务监管与过滤、任务分析、任务拆解、轻量化智能推理、任务执行与反馈等多个功能模块，结合本地与云端协同推理架构，使各模块紧密协作，实现云端与本地协同执行任务，提升用户的智能交互体验。

技术方案

1. 研究AI代理应用框架

该研究的关键在于如何基于消费类视听终端操作系统的应用场景构建AI代理框架，达成系统智能交互、高效处理任务和稳定运行的效果。针对消费类视听终端的多模态输入输出场景，AI代理应用框架如图1所示：

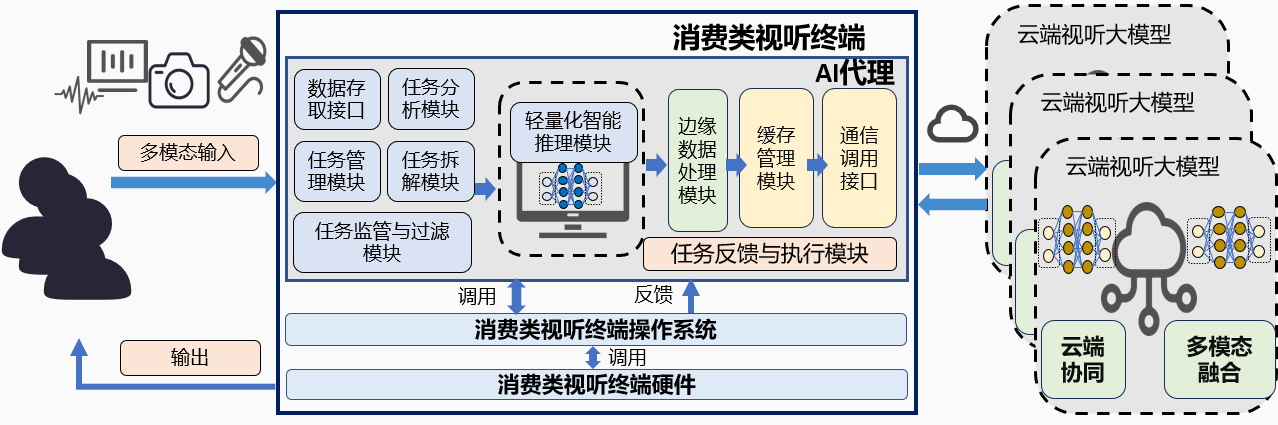
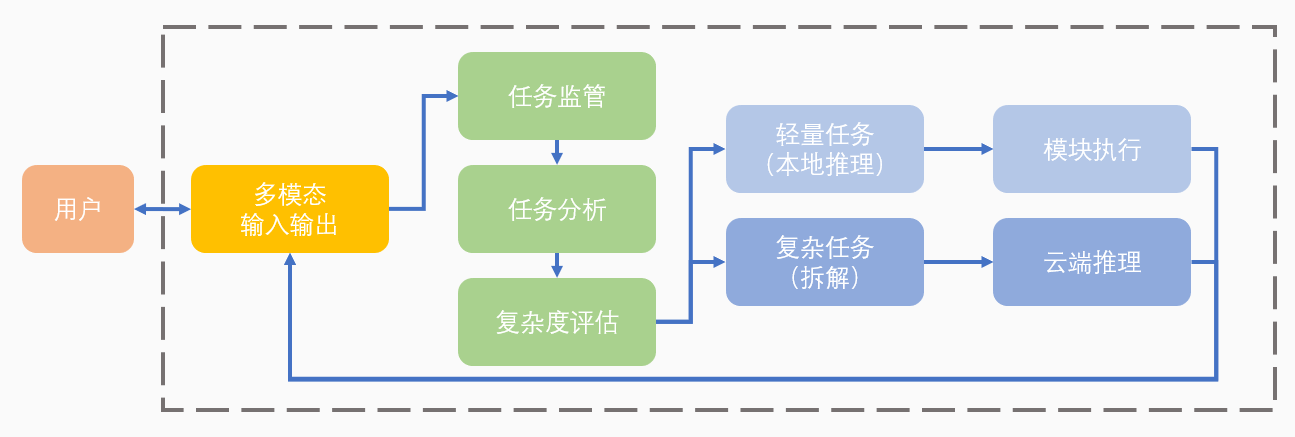


图1 基于消费类视听终端操作系统的AI代理应用框架

该框架采用采用模块化设计理念：多模态输入输出模块支持文本、语音、图像、视频等多种输入方式；任务监管与过滤模块负责针对用户发起的任务请求进行监管和筛选；任务分析模块针对经过监管与过滤的任务请求进行深度语义解析；轻量化智能推理模块在本地执行轻量化的智能推理任务；任务拆解模块针对复杂任务请求进行拆解；云端视听大模型交互模块针对高复杂度任务与云端的大模型进行交互；任务执行与反馈模块整合本地与云端的计算结果，执行任务操作并反馈给用户。各模块分工明确、协同工作，框架的技术路线如图2所示：



**图** **2 AI代理应用框架技术路线**

\*\*1. 多模态输入输出模块\*\*

针对消费类视听终端多模态交互需求，本模块采用分层处理架构。

多模态输入方面，文本输入处理基于蒸馏版BERT模型（DistilBERT）实现语义理解，通过领域词库扩充和上下文注意力机制提升专业术语识别准确率并提炼核心信息；语音输入利用Moonshine 开源语音识别将语音音频准确转换为文本，最终实现多模态输入文本理解。

多模态输出方面，利用语音合成（TTS）技术，如 F5-TTS 或 OuteTTS，将处理结果转换为自然流畅的语音输出；文本输出则结合本地语言模型实现结构化信息自然语言转换，最终实现系统反馈多模态输出。

\*\*2. 任务监管与过滤模块\*\*

本模块采用双引擎架构确保请求合规性。规则引擎基于Easy-Rules框架实现，包含语法校验、敏感词过滤等12类基础规则集，支持动态加载更新。智能过滤引擎采用用户行为分析模型，通过时序卷积网络（TCN）分析历史交互数据，可识别恶意请求模式。特别针对影视推荐场景构建了分级过滤机制，第一级硬件加速规则匹配耗时<5ms，第二级智能分析平均处理时间控制在50ms内。系统维护动态黑白名单，对连续异常请求自动触发安全熔断机制。

\*\*3. 任务分析与拆解模块\*\*

语义分析层采用领域适配的DistilBERT模型，在通用中文模型基础上注入影视领域实体识别能力。任务复杂度评估引入多维度量化体系，综合考虑模态类型、数据规模、计算复杂度等7项指标。拆解引擎实现混合策略，常规任务采用基于模板的快速拆解，创新性任务启用Octo-Planner进行多步推理。拆解过程生成带权重依赖关系的任务DAG图，支持优先级动态调整。针对影视推荐场景特别优化了跨平台API调用拆解策略，减少30%的冗余请求。

\*\*4. 轻量化智能推理模块\*\*

本地推理运行时采用模块化设计，核心包含模型管理、预处理加速、推理引擎三部分。模型管理支持热加载多种压缩模型格式，包括TensorFlow Lite、ONNX Runtime等。预处理环节利用TVOS专用DSP实现图像音频特征快速提取。文本推理使用定制版DistilBERT，通过层间共享和8bit量化技术，将模型尺寸压缩至45MB。系统实现智能计算资源分配，根据任务优先级动态调整CPU/GPU负载，确保高优先级任务响应延迟稳定在150ms内。

\*\*5. 云端视听大模型交互模块\*\*

云端接入层实现多模型路由机制，根据任务特征自动选择最优服务提供商。通信协议采用gRPC流式传输，支持大尺寸视频数据分块上传。结果缓存设计三级存储策略：内存缓存热数据（TTL 5分钟）、本地存储近期结果（TTL 24小时）、持久化存储用户偏好数据。特别针对实时交互场景开发了渐进式结果返回机制，云端处理过程中可返回部分计算结果。模块集成智能重试和降级策略，在网络波动时自动切换备用接入点。

\*\*7. 反馈与执行模块\*\*

交互界面实现动态布局引擎，支持语音、遥控器、手机APP等多通道统一控制。反馈内容生成采用模板+生成的混合模式，结构化数据自动转换为自然语言描述。执行引擎封装了TVOS系统200+个原子API，提供事务型操作保障。实时通信层基于WebSocket实现双向交互，支持操作指令的即时撤销重做。针对残障用户开发了多模态互补反馈机制，视觉信息自动转换为语音提示。

\*\*8. 测试验证体系\*\*

构建三级自动化测试框架：单元测试覆盖所有核心算法；集成测试验证模块间接口协议；系统测试模拟真实用户场景。性能测试采用渐进式负载策略，从单用户逐步增加到万级并发。健壮性测试包含网络抖动、服务降级等20种异常场景。特别开发了影视领域测试数据集，包含10万+标注样本覆盖各类边缘案例。质量门禁设置68项关键指标，包括核心路径响应时间、资源占用率、异常恢复时间等。