基于消费类视听终端操作系统的AI代理终端实现技术研究

应用设计报告

目录

[一、项目背景与目标 2](#_Toc188629359)

[1.1 项目背景 2](#_Toc188629360)

[1.2 项目目标 2](#_Toc188629361)

[二、应用设计内容与方案 2](#_Toc188629362)

[3.1 需求分析 2](#_Toc188629363)

[3.1.1 特定场景 3](#_Toc188629364)

[3.2 功能设计 3](#_Toc188629365)

[3.3 交互设计 3](#_Toc188629366)

[3.4 技术选型 4](#_Toc188629367)

[3.4.1 技术框架 4](#_Toc188629368)

[3.4.2 编程语言 5](#_Toc188629369)

[3.4.3 其他技术选择 5](#_Toc188629370)

[三、测试与评估 5](#_Toc188629371)

[四、实施计划 5](#_Toc188629372)

# 一、项目背景与目标

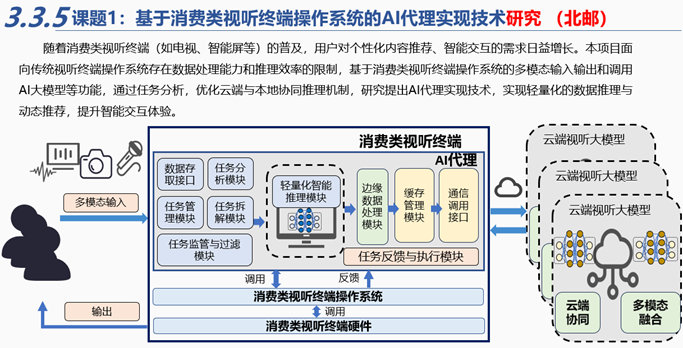
## 1.1 项目背景

消费类视听终端算力难以满足人工智能、大模型在消费类视听终端应用需求问题。操作系统要探索云端协同方式，解决消费类视听终端算力不能满足人工智能和大模型技术终端侧应用需求的问题。

## 1.2 项目目标

面向传统视听终端操作系统数据处理能力限制，基于消费类视听终端操作系统多模态输入输出和调用AI大模型等功能，研究AI代理实现技术，开发面向操作系统的应用层AI代理应用，通过任务监管与过滤模块、任务分析模块、轻量化智能推理模块、任务拆解模块、云端视听大模型交互模块、反馈和执行模块，结合云端与本地协同推理提升智能交互体验。

# 二、应用设计内容与方案



## 3.1 需求分析

* 追求智能化体验：用户期望AI代理终端能够具备智能识别、理解及响应能力，能够根据用户习惯和需求提供定制化服务。
* 注重交互便捷性：用户希望与AI代理终端的交互过程简单、直观，能够通过语音、控制器输入等多种方式实现交互。
* 偏好个性化内容：用户倾向于接收符合个人兴趣的内容推荐，包括影视、音乐、新闻等。
* 关注隐私保护：用户在使用AI代理终端时，对个人信息及隐私的保护意识较强，要求系统具备严格的隐私保护措施。

### 3.1.1 特定场景

**场景一：影视内容智能检索**

* 用户需求：多模态输入，精准检索，实时反馈，个性化推荐
* 功能需求：多模态输入处理，任务分析与拆解，云端协同推理，任务监管与过滤

**场景二：影视内容个性推荐**

* 用户需求：个性化需求，多层次需求解析，动态更新
* 功能需求：需求解析与提取，任务分解与分配，多源数据整合，动态更新与反馈

**场景三：影视内容实时交互**

* 用户需求：实时响应，精准感知，多模态交互。
* 功能需求：多模态感知与捕获，任务分析与管理，轻量化本地推理，云端协同优化

## 3.2 功能设计

1. **任务监管与过滤模块**：负责监管用户发起的任务请求，通过预设的规则或算法对请求进行过滤，确保只有合法、合理的请求被进一步处理。

* 接收：接收用户通过多模态输入方式发起的任务请求。
* 规则匹配：根据预设规则对请求进行匹配，判断其合法性和合理性。
* 过滤决策：基于规则匹配的结果，决定是否将请求传递给下一个处理模块。

1. **任务分析模块**：对经过监管与过滤的任务请求进行深入分析，理解用户的真实意图和需求。利用自然语言处理（NLP）技术解析用户请求的语义内容。
2. **轻量化智能推理模块**：在本地执行轻量化的智能推理任务，以提高响应速度和降低对云端的依赖。
3. **任务拆解模块**：将复杂的任务请求拆解成多个可执行的小任务，以便逐一处理。
4. **云端视听大模型交互模块**：与云端的大模型进行交互，获取更高级、更复杂的处理结果。

* 构建：根据本地推理和任务拆解的结果，构建向云端发送的请求。
* 解析：解析云端返回的处理结果，并将其转化为用户可理解的形式。

1. **反馈和执行模块**：将处理结果反馈给用户，并执行相应的操作以满足用户需求。

目前没有详细设计，牵涉到与中间件硬件交互。

决定轻量化智能推理还是云端视听大模型交互，需要判断规则：用**规则引擎**设计，如文本推理由本地完成，涉及图像音视频则交给云端视听大模型。

## 3.3 交互设计

说明用户与应用之间的交互方式、操作流程及反馈机制。

**场景一：影视内容智能检索**

用户通过智能电视语音输入“我想看刘德华主演的动作片”，系统首先通过任务监管模块去除噪音和冗余，将输入解析为“刘德华 动作片”。任务分析模块评估任务复杂度后将任务拆解为两个部分：“刘德华主演”的筛选任务由本地数据库完成，“动作片”分类的筛选任务由云端协同处理，系统结合用户历史偏好和评分进行排序，整合云端结果，生成推荐列表如《无间道》《拆弹专家》等，并附加预览图和简介，并支持自动加载指定影片并进入播放界面，实现精准、实时和高效的智能交互体验。

**场景二：影视内容个性推荐**

用户通过消费类视听终端向AI Agent输入需求：“我想看电影，放假过年的时候看点年轻人爱看的热门片子”，任务分解模块解析用户需求，提取以下关键信息：时间范围：2025年春节假期；电影类型偏好：小年轻爱看的热门片子；电影数量推算：预测假期内用户可能观影的电影数量，将任务拆解为独立子任务，包括调用电影记录平台（如豆瓣）API，筛选评分高、观影人数多的热门电影；分析平台中1825岁用户的观影偏好，生成符合“年轻人爱看”特征的电影画像；利用网页信息提取模型获取2025年中国春节假期的具体时间范围；结合平台历史数据和用户个体行为数据，预测用户春节期间的在线时长及观影时间。系统将上述模块的处理结果整合为结构化信息，包括每部电影的名称、评分、热度等；用户假期预计的观影时间及可观看电影数量；假期日期信息：假期的具体时间范围。

AI Agent通过与用户交互，呈现推荐结果，同时支持用户调整需求。如用户希望增加推荐电影数量或更改类型偏好，系统会利用上下文信息与推理缓存，快速重新分配任务并生成新的推荐结果。

**场景三：影视内容实时交互**

用户正在电视上观看中国 vs 日本的足球比赛，画面中中国队球员武磊正持球推进。用户提出问题：“武磊的进攻路线是什么？” 系统通过多模态输入模块捕获当前画面，结合武磊的动态位置和运动轨迹，由任务分析模块判断为实时画面解析问题，利用轻量化智能推理模块本地推断进攻路线为“右路推进，目标禁区”。AI Agent在不暂停直播的情况下，在屏幕右下角动态显示：“武磊当前进攻路线：右路推进，目标禁区。” 信息框短暂显示后自动消失。随后，用户提出问题：“中国队在本场比赛中有出现的可能性吗？” AI Agent通过消费类视听终端操作系统感知比分、时间和赛事背景等外部条件，任务分析模块判断为复杂问题，并通过通信调用接口将数据传至云端。云端大模型分析积分情况和同组比赛结果后返回完整分析。AI Agent在直播画面角落显示：“晋级形式分析：1:0胜利晋级，平局需看其他比赛。” 同时，画面底部以滚动字幕显示：“若中国队获胜：直接晋级；若平局：取决于沙特 vs 韩国比赛结果。

## 3.4 技术选型

### 3.4.1 技术框架

**AI Agent：**基于Plan-and-Execute或ReAct等推理模式，实现任务的规划与执行分离或思考-行动循环机制。开源AI搜索引擎（如Search4all）可作为任务监管模块的一部分，用于整合多个大模型的结果，并提供搜索和推荐功能。简单的任务可以通过workflow实现。

**TVOS：**采用C/S架构，支持组件的裁剪与组合，方便跨应用平台重用。选择Python和Java作为开发语言（可根据组件特性调整），并将应用APK的开发分为多个模块分别实现，实现功能组件模块的分离，提高模块可复用度。

**云端与本地协同推理：**

* **云端模型**：任务处理阶段的Flamingo（多模态输入->文本输出），BERT（文本输入输出），执行阶段的sora（text->video），DALL.E（text->picture），以及国内的云试听相关模型阿里云通义千问，百度文心一言，字节跳动豆包
* **本地推理**：通过轻量化智能推理模块，在本地设备上执行部分推理任务，减轻云端负担。目前设计为文本推理，选择中文化的DistilBERT作为NLP本地推理模型（DistilBERT 的参数量为 **66M**）

### 3.4.2 编程语言

**Python**：负责模型相关

**Java**：负责架构和数据相关

### 3.4.3 其他技术选择

**多模态交互**：利用音频处理（语音识别、合成、音频分析）和视觉处理（视频理解、图像处理、内容生成）技术，实现用户与应用的视听结合交互。需要与TVOS的各类中间件协同进行开发。

**开源模型与接口**：阿里云通义千问、百度文心一言等国内云试听相关模型，可作为云端视听大模型交互模块的技术支持。同时，ChinaDRM技术为应用层提供公共服务支撑能力，Agent系统开发时需要大量调用该服务提供的接口。

# 三、测试与评估

目前只给出测试评估标准，后续开发到一定阶段时编写测试方案。

* **测试方法**：采用单元测试、集成测试、系统测试等多种测试方法，确保系统功能的正确性和稳定性。
* **测试用例**：针对每个功能模块设计详细的测试用例，覆盖所有可能的使用场景。
* **测试环境**：搭建与生产环境相似的测试环境，确保测试结果的准确性。

# 四、实施计划

**阶段一：技术选型与环境搭建（第1-2个月）**

明确项目目标，细化功能需求，调研并选择合适的技术框架、编程语言及本地文本推理模型。完成调研报告，确定技术路线，搭建基础开发环境。

**阶段二：系统设计与开发（第3-7个月）**

设计AI Agent架构，包括任务监管与过滤、任务分析、轻量化智能推理、任务拆解、云端视听大模型交互、反馈和执行等模块。开发各模块功能，集成到TVOS系统中。完成系统核心功能的开发，初步实现AI代理在TVOS系统上的应用。

**阶段三：功能优化与集成测试（第8-11个月）**

对系统进行优化，提升性能，修复bug。进行集成测试，确保各模块协同工作正常。在TVOS硬件上进行大量功能测试，调整模块和代码。最终形成测试报告。

**阶段四：编写文档（第12个月）**