|  |  |  |
| --- | --- | --- |
|  | **中国高校计算机大赛·AIGC创新赛** |  |

**ClipPersona——**

**AI赋能的个性化视频创作共生平台**

**项目计划书**

**创影共生**

**2025-07-06**

**目录**

[1. 项目概述 4](#_Toc202732258)

[1.1 项目背景与定位 4](#_Toc202732259)

[1.2 项目核心理念 5](#_Toc202732260)

[1.3 目标用户与应用场景 6](#_Toc202732261)

[1.4 市场需求与竞品分析 6](#_Toc202732262)

[1.5 作品核心价值主张与差异化优势 7](#_Toc202732263)

[2. 产品介绍 9](#_Toc202732264)

[2.1 交互设计 9](#_Toc202732265)

[2.1.1 交互设计理念 9](#_Toc202732266)

[2.1.2 交互设计特点 9](#_Toc202732267)

[2.1.3 交互设计优势 9](#_Toc202732268)

[2.2 产品原型图 9](#_Toc202732269)

[2.2.1 主页页面 9](#_Toc202732270)

[2.2.2 人格卡页面 10](#_Toc202732271)

[2.2.3 社区页面 10](#_Toc202732272)

[2.2.4 视频选择页面 10](#_Toc202732273)

[2.2.5 视频剪辑页面 10](#_Toc202732274)

[2.2.6 设置页面 10](#_Toc202732275)

[2.3 产品功能流程图 12](#_Toc202732276)

[2.4 产品功能详述 12](#_Toc202732277)

[2.4.1 ClipPersona Studio：个性建模系统 12](#_Toc202732278)

[2.4.2 对话式交互系统：自然语言控制剪辑意图 12](#_Toc202732279)

[2.4.3 自迭代剪辑副本机制：行为反向学习 12](#_Toc202732280)

[2.4.4 副本共享平台：构建风格开放生态 12](#_Toc202732281)

[2.5 数据流程图 1](#_Toc202732282)

[3. 技术方案 3](#_Toc202732283)

[3.1 关键技术 3](#_Toc202732284)

[3.1.1 集成开发环境：Android Studio 3](#_Toc202732285)

[3.1.2 开发框架：React Native 3](#_Toc202732286)

[3.1.3 语音识别：Whisper 3](#_Toc202732287)

[3.1.4 视频剪辑处理：MobileFFmpeg, GPUImage 3](#_Toc202732288)

[3.1.5 语言指令解析/背景生成：蓝心大模型 4](#_Toc202732289)

[3.1.6 视觉语言模型：MobileVLM-v2，Qwen-VL 4](#_Toc202732290)

[3.1.7 实例分割：YOLO11，SAM2 4](#_Toc202732291)

[3.1.8 目标消除：E2FGVI 4](#_Toc202732292)

[3.1.9 模型轻量化：ONNX/TensorRT 5](#_Toc202732293)

[3.2 其他相关技术 5](#_Toc202732294)

[3.2.1 数据库：SQLite/Firebase 5](#_Toc202732295)

[3.2.2 风格建模：VAE 5](#_Toc202732296)

[3.2.3 行为反向学习：PPO 5](#_Toc202732297)

[3.2.4 UI设计工具：Pisxo 5](#_Toc202732298)

[3.3 大模型应用能力分析 6](#_Toc202732299)

[3.3.1 表达理解能力 6](#_Toc202732300)

[3.3.2 风格建模能力 6](#_Toc202732301)

[3.3.3 交互优化能力 6](#_Toc202732302)

[3.3.4 生态构建能力 6](#_Toc202732303)

[4. 项目计划 7](#_Toc202732304)

[4.1 可行性分析 7](#_Toc202732305)

[4.1.1 技术可行性 7](#_Toc202732306)

[4.1.2 市场可行性 7](#_Toc202732307)

[4.1.3 经济可行性 7](#_Toc202732308)

[4.2 团队介绍 8](#_Toc202732309)

[4.2.1 导师介绍 8](#_Toc202732310)

[4.2.2 团队分工情况 8](#_Toc202732311)

[4.2.3 排期规划 8](#_Toc202732312)

[5. 总结与展望 9](#_Toc202732313)

[5.1 项目价值总结 9](#_Toc202732314)

[5.2 项目完成度总结 9](#_Toc202732315)

[5.3 展望 9](#_Toc202732316)

[5.3.1 功能拓展 9](#_Toc202732317)

[5.3.2 技术创新 10](#_Toc202732318)

[5.3.3 市场拓展 10](#_Toc202732319)

# 项目概述

## 项目背景与定位

随着短视频、自媒体与知识内容产业的高速发展，视频剪辑已成为内容创作的核心环节。然而，当前主流剪辑工具依然以时间线操作为核心，严重依赖手动流程，效率低、重复性高，难以适应高频创作的内容生产需求。同时，创作者的表达风格难以在工具中被准确捕捉、建模与复用，导致剪辑结果缺乏个性一致性，影响内容辨识度与品牌延续性。

AIGC 技术，特别是以大语言模型为代表的语义理解与行为建模能力，正在为剪辑流程的智能化和个性化提供现实基础。模型已具备对用户语气、节奏、情绪倾向和风格偏好的理解与转化能力，使“可学习的表达型 AI”从概念走向可用。

基于此，ClipPersona 项目提出了“剪辑人格副本”的新模式：通过训练可成长、可协作的数字副本，系统能够持续学习用户的表达逻辑与审美偏好，生成风格一致、语义贴合的个性化剪辑版本，并支持自然语言交互与行为反馈迭代，构建表达进化闭环。

项目致力于构建一个以大模型为核心、以人格化风格表达为导向的智能剪辑共生系统，使 AI 从被动工具转化为能够理解、协作、共创的“表达代理体”，帮助创作者在提升创作效率的同时，实现内容风格的沉淀与表达的一致性。

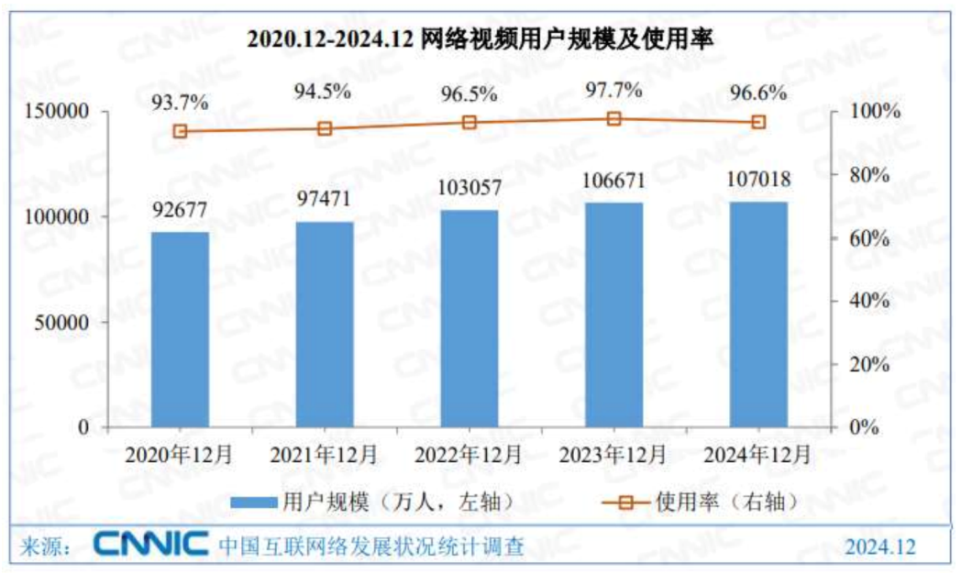


图 1‑1网络视频用户规模及使用率趋势图

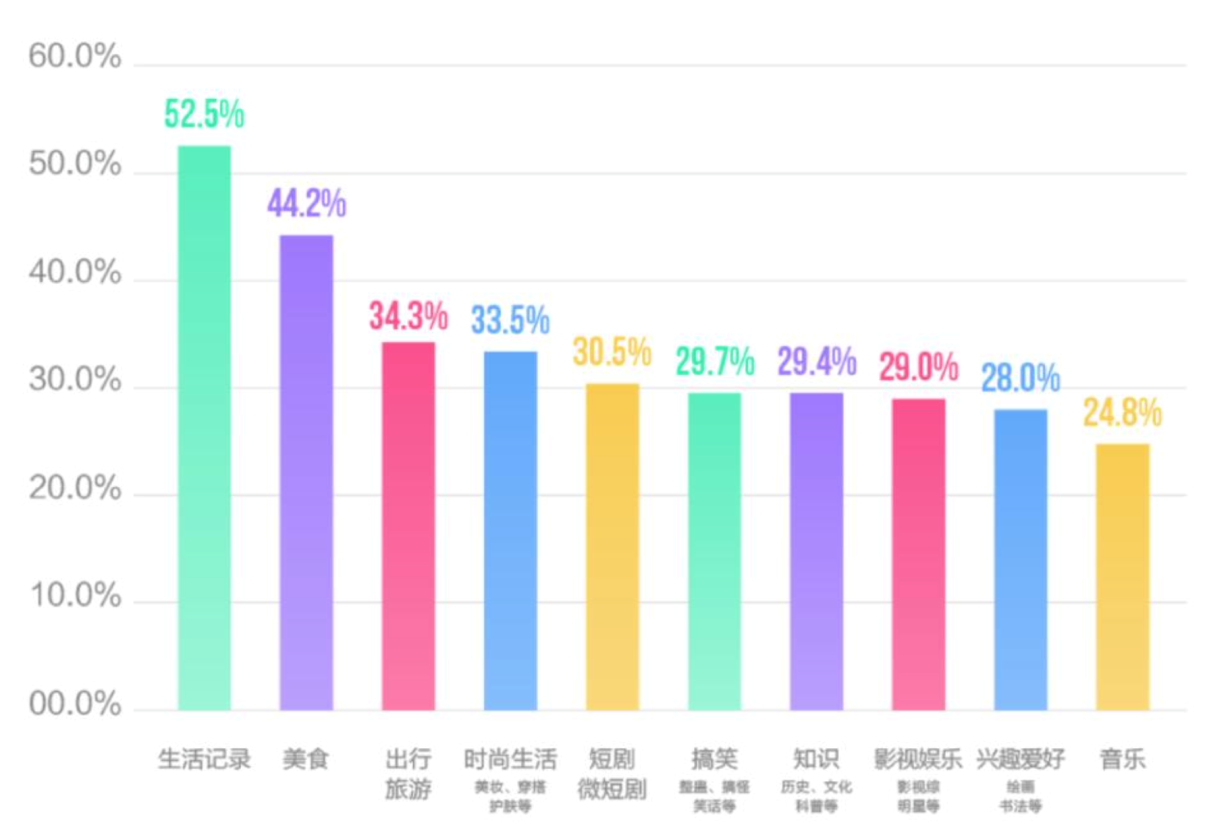


图 1‑2 网络视听内容创作者投稿类型分布图

## 项目核心理念

ClipPersona 致力于打造一种全新的剪辑体验：让剪辑不只是高效执行，而是风格表达的延伸。

它不是传统意义上的剪辑工具，而是一个具备理解力、成长性和协作能力的“表达型剪辑人格体”。创作者可以通过行为标注与对话交互，训练出属于自己的“剪辑副本”，系统会自动学习其节奏偏好、语言逻辑与视觉习惯，生成具有高度个性化风格的剪辑结果。

每一个“剪辑副本”都不是静态模板，而是可持续进化的数字人格。系统通过语言模型对语义意图的深度解析，结合反馈学习机制，不断优化“理解——执行——反思”的剪辑链路，实现真正意义上的个性化表达建模。

项目核心理念体现在三方面：

①表达理解力：大模型驱动的语义识别与情绪解析，精准还原创作者意图；

②行为可塑性：用户每一次调整，都会被系统吸收并内化为风格参数；

③风格资产化：以“剪辑人格卡”形式保存、调用、复用创作风格，实现风格资产的长期沉淀与迁移。

此外，ClipPersona 不仅支持个体表达，更致力于构建一个剪辑人格共享社区。创作者可以选择将人格卡私有存储，或公开分享至平台，供他人下载、复用或进行风格融合。这一机制将个人风格沉淀为可交易、可演化的创作资产，为 AIGC 内容创作构建出一个去中心化的“表达样式市场”。

ClipPersona 将剪辑从“技术操作”重构为“表达延续”，使 AI 从工具角色跃迁为创作者的风格共生体，构建一种前所未有的“创作人格协作机制”。

## 目标用户与应用场景

ClipPersona 面向的核心用户群体是以风格表达为核心竞争力的高频内容创作者，主要包括自媒体视频博主、知识类讲师、Vlog 与情感类创作者，以及小型内容运营团队与自由剪辑师。这类用户普遍对表达质量、剪辑效率和内容一致性有高度要求，但传统工具难以满足个性化风格建模与快速生成的双重需求。ClipPersona 提供可训练、可进化的剪辑副本机制，帮助创作者实现风格沉淀、剪辑一致性与效率统一，是当前内容创作生态中需求与契合度最高的目标群体。

在实际应用中，ClipPersona 具备高度的场景适应能力。对于知识讲解类内容，可构建“教学型剪辑人格”提升教学视频的统一性与专业感；在情绪表达类内容中，可结合共情剪辑引擎，实现节奏与情感的动态适配；而对于内容运营团队与自由剪辑师，通过风格卡系统可快速切换项目风格，实现批量化、多账号内容生产协同。系统支持移动端轻量拍摄与初剪，PC端深度训练与管理，能够覆盖自媒体创作、教育教学、品牌内容营销等多个内容密集型行业场景。

此外，ClipPersona 也具备向潜力人群拓展的可扩展性，包括KOL孵化平台、品牌创意团队及社交平台内容机构等。这些用户可通过风格卡市场与多版本剪辑机制，完成内容A/B测试、人设风格构建与批量化分发等任务，进一步扩展项目的行业影响力与平台化潜力。

## 市场需求与竞品分析

当前视频内容市场处于高速扩张阶段，短视频、知识传播、自媒体教育等领域对高质量、风格统一的内容需求持续增长。然而，创作者在高频创作过程中普遍面临剪辑耗时、风格不可控、内容迭代效率低等问题，尤其在风格沉淀与个性表达方面，现有工具难以提供系统化解决方案。AIGC 技术的发展为“表达自动化”提供了技术路径，但市面产品仍集中于低门槛模板生成或半自动剪辑工具，缺乏对“表达风格人格化”的深入支持。

在现有竞品中，如 Runway、Descript 等产品聚焦于视频处理效率与语音编辑自动化，主流国产工具如剪映、万兴喵影等则强调素材库与模板拼接体验。这些工具虽提升了剪辑便捷度，但大多采用通用化处理逻辑，无法满足用户个性化风格表达与内容差异化运营需求。与此相比，ClipPersona 的核心竞争力在于其“剪辑人格副本”机制，通过大模型能力理解用户表达习惯，沉淀可复用的风格资产，实现内容风格的标准化、差异化与资产化并存。

随着创作者个体品牌意识增强与AI表达能力的发展，市场对“风格建模 + 内容协作 + 多版本表达”的剪辑系统将产生长期需求。ClipPersona 正是在这一趋势下，率先构建“表达协作型AI”产品范式，具有明确的市场切入价值与产品壁垒。

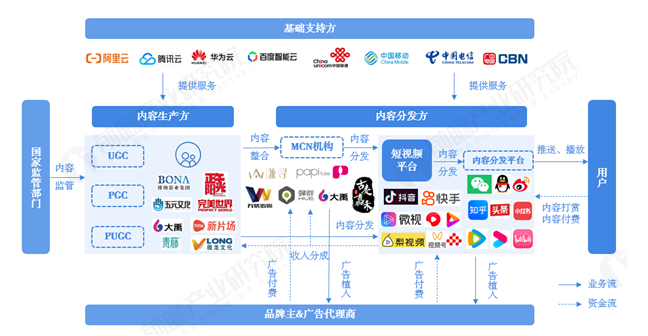


图 1‑3 中国短视频产业全景图



图 1‑4 全球短视频产品图

## 作品核心价值主张与差异化优势

ClipPersona 的核心价值在于以剪辑风格人格化为抓手，构建一个具备理解力、成长性与协作能力的 AI 表达系统，为创作者提供真正“像自己”的剪辑结果。这不仅是一种工具层的效率优化，更是一种创作方式的范式重构：系统以“剪辑人格副本”形式，承载用户语言节奏、视觉习惯与审美偏好，通过大模型驱动的语义理解与行为学习，实现内容风格的自动化生成与持续迭代。

相较于传统模板驱动或参数式剪辑工具，ClipPersona 有三大差异化优势：

①表达记忆体机制：每位创作者可训练独立剪辑人格，系统自动记录其风格逻辑并转化为可调用的个性化剪辑引擎，形成表达风格的 AI 记忆体；

②共情剪辑系统：基于语言高光、情绪节奏与观众行为预判，系统主动识别观众心理曲线，生成具有情感节奏感知能力的片段推荐与剪辑策略；

③风格资产平台机制：通过“剪辑人格卡”系统，用户可管理、共享、复用或出售其风格副本，构建风格人格市场，实现表达方式的资产化沉淀。

ClipPersona 不仅定义了一个智能剪辑工具，更提出了“AI表达代理人”的产品新形态，为创作者提供一种可持续成长、可个性建模、可协作表达的全新创作方式。

# 产品介绍

## 交互设计

### 交互设计理念

ClipPersona 的交互设计始终坚持以用户为中心，界面布局与操作流程充分贴合用户的使用习惯。整个产品围绕“剪辑目标”展开，确保每一步操作都服务于任务的顺利完成。同时，系统通过智能提示与实时反馈，让用户始终掌控操作进度，并通过语音与文字双通道输入，带来高效且灵活的多模态融合体验。

### 交互设计特点

基于上述理念，ClipPersona 在具体设计上展现出鲜明的特点。首页采用卡片式管理，历史视频以缩略图和时间戳直观呈现，方便用户快速浏览和定位。剪辑界面则注重沉浸式体验，隐藏冗余控件，突出内容本身和操作主线。无论是语音还是文字输入，用户都可以无缝切换，保证任务流程的连贯性。此外，AI 输出的视频支持一键预览和编辑回退，关键交互节点还配有动画动效，有效提升了操作引导性和产品的温度感。

### 交互设计优势

正因为有了上述设计理念和特点，ClipPersona 实现了极高的易用性和效率。用户无需专业基础即可零门槛上手，核心任务操作步数极少，每一步都伴有清晰的进度提示和交互反馈。关键功能如“继续编辑”、“导出”、“回首页”等始终固定在易于记忆的位置，页面风格与控件交互高度一致，极大降低了认知负担。由此，ClipNova 为用户带来高效、友好且专业的一站式视频剪辑体验。

## 产品原型图

产品原型图设计采用低保真——中保真——高保真的设计模式。

### 主页页面

①视频卡片组件：用于展示每个历史视频的缩略图、标题和时间戳，方便用户快速浏览和选择。

②顶部/底部导航栏：为用户提供全局导航入口，便于在各主要功能间切换。

③快捷入口按钮：一键发起新建剪辑、导入视频等常用操作，提升效率。

### 人格卡页面

人格卡列表/卡片组件：以卡片形式展示用户已创建的人格卡，便于管理和选择。

编辑/新建人格卡按钮：支持用户新建或编辑人格卡，灵活调整个性化参数。

统计/使用频率展示：可视化展示各操作的使用频率，帮助用户了解自己的编辑习惯。

### 社区页面

①帖子/作品列表组件：以列表或卡片形式展示社区用户发布的内容，便于浏览和互动。

②评论区：支持用户对作品进行评论和交流，增强社区氛围。

③点赞/收藏按钮：允许用户对喜欢的内容进行点赞或收藏，提升参与感。

④社区导航标签：帮助用户按分类浏览不同类型的社区内容。

⑤用户头像/昵称展示：展示内容发布者的身份信息，增强社交属性。

### 视频选择页面

①媒体文件列表/缩略图网格：以缩略图形式展示本地可用媒体文件，便于选择。

②导入/确认按钮：一键导入所选素材，快速进入编辑流程。

③文件信息展示：显示文件的大小、时长等详细信息，辅助用户决策。

### 视频剪辑页面

①视频预览播放器：实时播放当前编辑的视频，便于用户查看剪辑效果。

②剪辑操作：集合裁剪、分割、字幕、音乐等常用剪辑功能，操作直观。

③输入区（语音/文字）：支持用户通过语音或文字输入剪辑指令，提升交互灵活性。

④操作历史/撤销重做按钮：记录并允许用户撤销或重做编辑操作，保障编辑安全。

⑤动效/动画反馈区：在关键操作时提供动画提示，增强操作的可感知性。

### 设置页面

设置项列表：集中展示所有可配置项，便于用户查找和调整。

切换控件（开关/单选/多选）：以直观的控件形式支持用户快速切换设置。

账号信息展示与编辑：显示并允许用户修改个人信息。

关于/帮助/反馈入口：为用户提供产品信息、使用帮助和反馈渠道。



图 2‑1 主页页面原型图



图 2‑2 社区页面原型图（1）



图 2‑3 人格卡页面原型图



图 2‑4 视频选择页面原型图

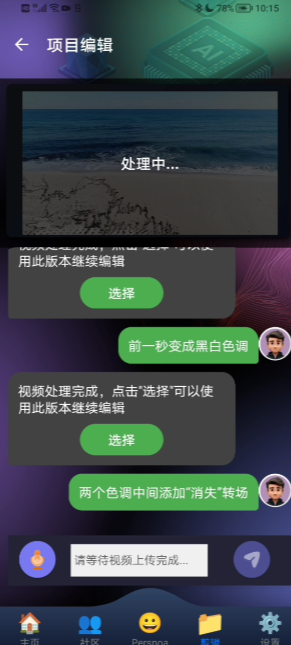


图 2‑5 视频剪辑页面原型图



图 2‑6 设置页面原型图

## 产品功能流程图

产品主要分为用户交互、业务逻辑、模型训练、视频处理和输出管理五个核心模块。

用户交互模块负责提供直观的界面和多样的交互功能。通过移动端界面，用户可以上传视频、输入自然语言或语音指令、选择风格卡并查看观众心理曲线。通过实时反馈增强用户体验，为后续处理奠定基础。

业务逻辑模块作为系统的中枢，协调各模块之间的数据流和功能实现。该模块处理用户指令的语义解析、风格卡的管理以及剪辑决策的优化，确保指令与Persona风格的精准匹配。它还支持对话式反馈交互，解释剪辑决策并提供优化建议，增强系统的智能化和用户友好性。

模型训练模块负责剪辑人格副本（Persona）的生成和优化。通过用户上传的视频和标注偏好（如“喜欢/不喜欢”片段），该模块利用视觉特征提取和强化学习技术，构建个性化的风格模板，不断提升剪辑精准度。

视频处理模块实现具体的剪辑操作，如目标消除、背景替换、风格化剪辑和共情剪辑。目标消除通过目标检测和背景修复移除指定对象；背景替换支持自定义或AI生成背景，并确保光照与时空一致性；风格化剪辑根据Persona参数优化节奏、转场和特效；共情剪辑通过语音情绪分析和语义解析，预测观众心理，优化观众留存率。

输出管理模块负责最终视频的生成与分发。系统支持本地和云端渲染，生成符合用户需求的视频成品，并通过社区平台实现风格卡的共享与下载。用户可以将剪辑成果转化为视频，分享至社交媒体，或在社区中与其他用户互动，形成创作生态。

在五个模块之外，还有一些基础的系统功能，如注册登录账号、新手指南、数据统计等。这些功能配合五大模块共同构成了一个高效智能的视频剪辑系统。

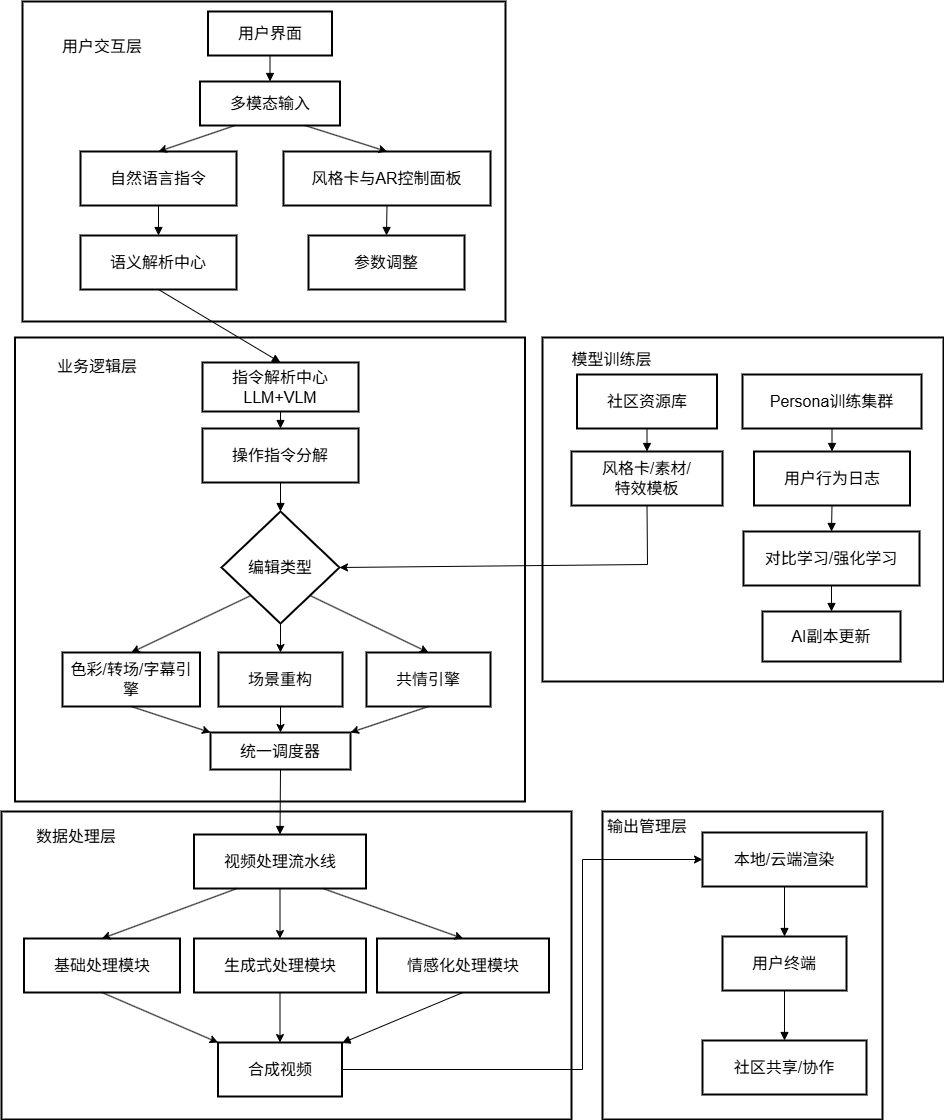


图 2‑7 产品功能流程图

## 产品功能详述

ClipPersona 的系统架构围绕“表达风格学习——内容生成执行——行为反馈——资产共享”四大核心能力，构建了一套剪辑智能化与表达个性化高度融合的功能体系。各模块既相互独立，形成行为——风格——资产的表达闭环。

### ClipPersona Studio：个性建模系统

ClipPersona Studio是一个生成人格副本的系统，创作者进入ClipPersona Studio后，可导入过往视频，系统引导其进行偏好标注。系统将语言节奏、镜头节选逻辑、内容结构偏好等行为转化为风格向量，生成初始风格参数组，为后续剪辑任务提供风格基础，通过类似“自我训练空间”的过程，生成用户专属的“剪辑人格副本”。

### 对话式交互系统：自然语言控制剪辑意图

用户可以通过文本或语音输入自然语言指令，与剪辑副本进行互动，直接发出剪辑需求，系统利用语义解析技术理解用户指令，将模糊语言指令映射为结构化的剪辑行为，将其转化为具体的剪辑操作，生成剪辑方案。这一功能通过直观的语言交互方式，降低用户操作门槛，提升剪辑效率，实现灵活的剪辑控制和个性化优化。

### 自迭代剪辑副本机制：行为反向学习

副本生成剪辑方案并呈现给用户后，用户可以选择接受、调整或拒绝该方案，或者对方案进行进一步调整。用户无需手动设置参数，系统会自动记录用户的选择，并在后台重新拟合风格模型，系统通过分析用户反馈数据，动态调整副本的剪辑风格模型，使其逐步更贴近用户的个性化需求。

### 副本共享平台：构建风格开放生态

用户可以将自己构建的剪辑副本分享到社区平台，设置为公开或私有状态，其他用户可以浏览、下载并使用这些副本。平台支持用户对副本进行评分和收藏，用户还可以将他人副本与自己的风格结合，生成新的剪辑风格。这一模块通过社区共享机制，构建用户之间的创作生态，促进风格交流与灵感碰撞。

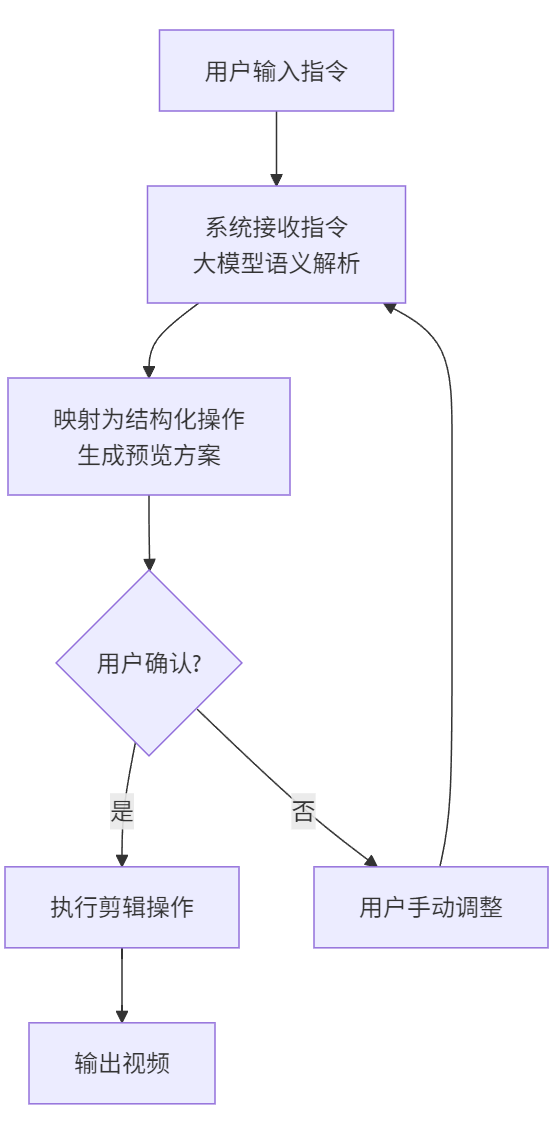


图 2‑8自然语言控制剪辑流程图

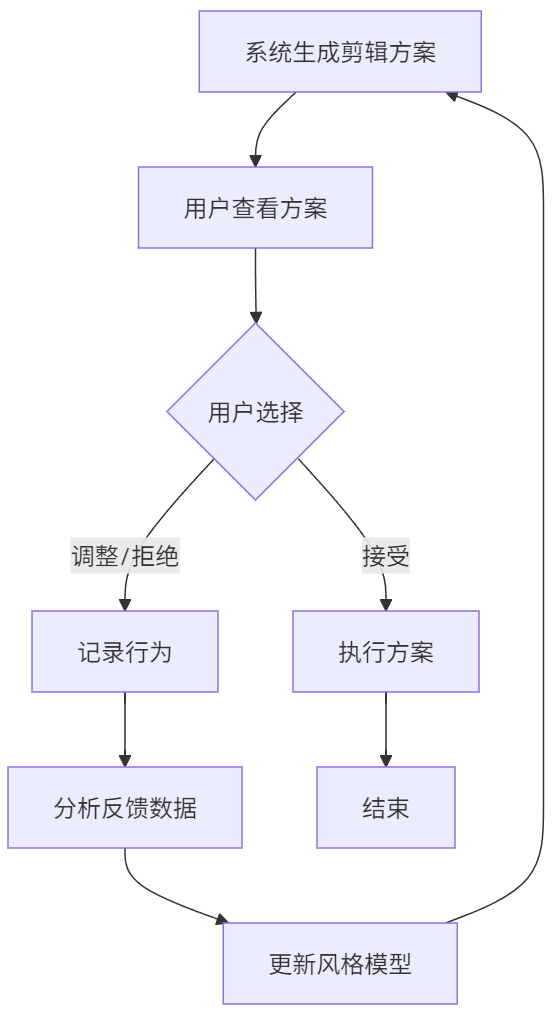


图 2‑9 系统行为学习流程图

## 数据流程图

数据流程实现了从用户输入、AI解析、智能编辑、多模块协同处理，到视频合成、结果展示、用户行为数据回流和模型自我进化的闭环，确保了产品的智能性、个性化和可持续优化能力。

①输入阶段：用户首先输入剪辑指令或参数，作为整个流程的起点。

②指令解析与分解：输入的指令会被送入指令解析中心，经过剪辑指令数据的处理后，由操作指令分解器进一步细化为具体的操作。

③编辑类型标注与分流：分解后的操作会被打上编辑类型标签，根据不同的标签，流向不同的处理分支。

④统一调度：各分支处理后的结果会被调度器整合，进入下一个处理阶段。

⑤多模块协同处理：调度后的数据会进入不同模块，进行更深层次的视频内容生成与优化。

⑥视频处理流水线：各模块输出的数据在视频处理流水线中汇聚，最终合成为完整的视频内容。

⑦结果输出与数据回流：合成后的视频可以在本地或云端进行展示。同时，用户的行为数据会被记录，生成用户行为日志；风格卡的生成结果也会被存入社区资源库。

⑧模型持续优化：用户行为日志和社区资源库的数据会被用于模型训练和更新，实现产品的持续自我优化和智能进化。

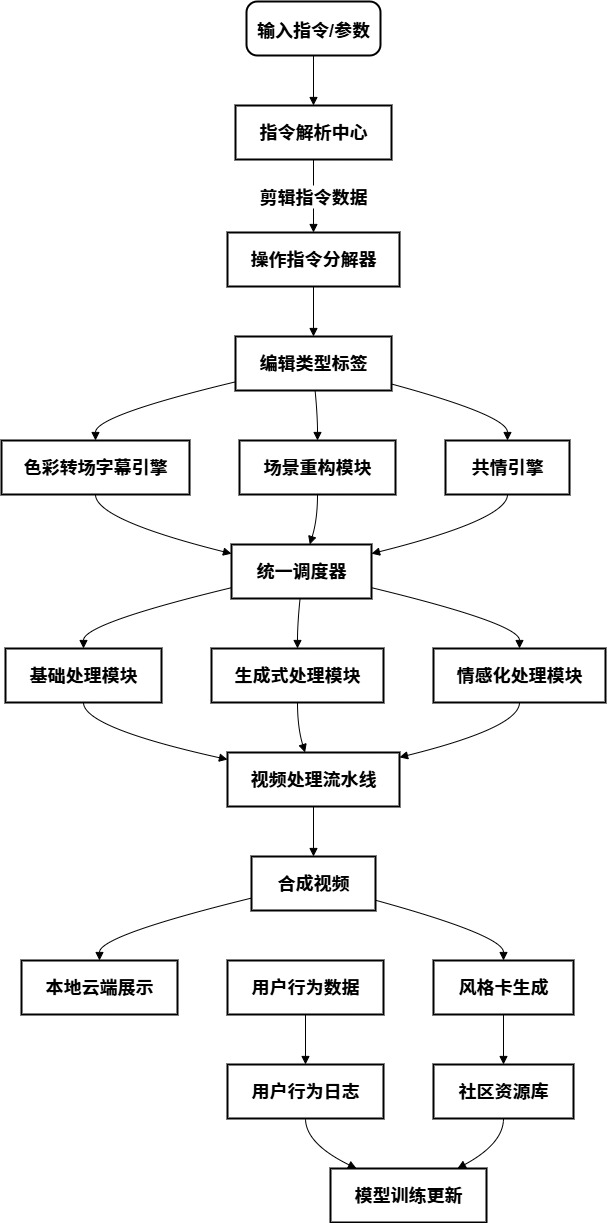


图 2‑10 数据流程图

# 技术方案

## 关键技术

### 集成开发环境：Android Studio

Android Studio 是由谷歌公司开发的集成开发环境（IDE），专为 Android 应用开发设计，集成了开发者所需的工具，包括 Android SDK、代码编辑器、调试器和性能分析工具。Android Studio 提供强大的代码编辑器，支持 JavaScript 和 TypeScript 的开发，结合 React Native 框架，实现跨平台应用的快速构建。其布局编辑器支持通过拖拽方式创建用户界面，大幅降低 UI 开发成本。此外，Android Studio 提供 Android Virtual Device（AVD）管理器，允许开发者在计算机上模拟 Android 设备，便于测试和调试。Gradle 构建工具则优化了编译、依赖管理和应用打包流程，确保开发效率。

### 开发框架：React Native

React Native 是支持跨平台移动应用开发的框架，支持 iOS 和 Android 平台，提供丰富的 UI 组件和高效的渲染性能。其基于 JavaScript 的开发模式结合原生组件，支持快速迭代和热更新，缩短开发周期。React Native 的 Flexbox 布局系统确保界面适配不同屏幕尺寸，搭配 React Navigation 实现平滑的界面切换（如 Studio 界面、社区平台）。通过桥接机制调用原生模块（如 GPUImage），优化视频预览和交互体验。

### 语音识别：Whisper

Whisper 是一种由 OpenAI 开发的多语言语音转文字模型，具备强大的鲁棒性，能够准确处理多种口音、方言和语言输入。其基于 Transformer 架构的设计支持高效的语音特征提取和序列建模，特别适合处理复杂用户环境下的自然语言指令输入。Whisper 还支持多种预训练配置，可根据具体需求进行微调，以优化剪辑相关任务中的语音识别性能。

### 视频剪辑处理：MobileFFmpeg, GPUImage

MobileFFmpeg 是一款功能强大的跨平台视频处理工具，支持视频裁剪、拼接、转码和封装，兼容主流编码格式，为剪辑任务提供底层支持。GPUImage 基于 OpenGL ES 技术，实现了实时滤镜、转场效果和颜色调整功能，充分利用 GPU 加速，提升移动端视频渲染的流畅性和视觉质量。

### 语言指令解析/背景生成：蓝心大模型

蓝心大模型由vivo开发，是一款高性能的多模态 AI 模型，集成了强大的视觉和语言理解能力，能够深度解析自然语言指令并结合视觉上下文进行语义分析。其高效的推理能力和多语言支持使其在对话系统、意图识别和内容生成领域表现出色，特别适合复杂剪辑任务中的语义映射和背景生成需求。蓝心大模型通过云端部署和本地优化相结合的方式，兼顾了性能与实用性，为个性化的视频剪辑提供了技术支撑。

### 视觉语言模型：MobileVLM-v2，Qwen-VL

MobileVLM-v2 是一种专为移动端优化的视觉语言模型，结合 MobileNetV3 和 Transformer 结构，实现了高效的自然语言指令解析和多模态内容理解。其轻量化设计使其能在资源受限的环境中运行，特别适用于将文本指令与视频内容结合进行智能交互。MobileVLM-v2 通过预训练和微调，能够处理复杂的视觉场景和语言指令，适配对话式剪辑控制和多模态语义分析需求，显著提升移动设备上的智能化体验。

Qwen-VL是由阿里巴巴云开发的高性能视觉语言模型，集成了多模态输入输出能力，支持图像、文本和边界框的处理，具备多语言对话、图像中文本识别和多图像交错理解功能。Qwen-VL 采用视觉接收器与大语言模型相结合的架构，通过三阶段训练，在视觉问答、图像描述和视觉定位等任务中表现出色。其高效的推理能力和多语言支持使其特别适合复杂剪辑任务中的语义解析和内容生成需求。Qwen-VL 通过云端部署，高质量实现视觉相关任务。

### 实例分割：YOLO11，SAM2

YOLO11 是一种先进的实时目标检测模型，基于改进的 backbone 结构和增强的 neck 设计，在保持高精度检测能力的同时，优化了计算效率，特别适合移动端环境。SAM2（Segment Anything Model 2）进一步扩展了实例分割功能，通过生成像素级分割掩码，显著提升了目标识别的精细度。其轻量化设计结合最新优化技术，使得在资源受限的设备上也能实现高效处理。

### 目标消除：E2FGVI

E2FGVI（End-to-End Flow-Guided Video Inpainting）是一种基于流引导的视频补全模型，采用光流估计技术（如 RAFT）捕获帧间运动信息，确保修复区域与前后景的时序一致性。该模型利用 GAN 结构和时空注意力机制，能够生成高质量的背景修复效果，特别适用于复杂场景下的目标消除任务，具有广泛的应用潜力。



图 3‑1 目标消除效果预览图

### 模型轻量化：ONNX/TensorRT

ONNX（Open Neural Network Exchange）是一种开放的神经网络交换格式，支持跨平台模型部署，兼容 iOS 和 Android 等框架，降低了模型迁移的复杂性。TensorRT 则通过模型量化、层融合和动态内存管理等技术，显著提升了推理速度和效率，确保实时性与低功耗。

## 其他相关技术

### 数据库：SQLite/Firebase

SQLite 提供轻量级本地存储，适配离线场景；Firebase 提供云端实时数据库和文件存储，确保数据同步和安全性。

### 风格建模：VAE

变分自编码器（VAE）通过潜在空间建模生成风格向量，捕捉用户行为中的个性化特征；CLIP-ViT 结合视觉和音频特征提取，构建个性化的剪辑风格表达，为表达建模提供技术基础。

### 行为反向学习：PPO

PPO（Proximal Policy Optimization）是一种高效的强化学习算法，通过用户行为反馈优化模型，适合动态调整剪辑意图和风格适应。

### UI设计工具：Pisxo

Pixso 是一款云端协作 UI/UX 设计工具，支持实时编辑、原型设计和多平台适配，适合快速构建交互界面。在项目中用于设计风格卡管理面板和对话窗口等UI界面，优化用户体验并支持团队协作。

## 大模型应用能力分析

ClipPersona 充分利用大语言模型的语义理解、行为建模和多模态处理能力，构建以个性化表达为核心的智能剪辑系统。以下从自然语言理解、视频行为理解和人格卡形成能力三个维度，分析大模型在项目中的具体应用能力，确保系统高效、精准地支持个性化视频创作。

### 自然语言理解能力

大模型通过深度语义解析，精准捕捉用户在自然语言指令中的创作意图和情感倾向，为 ClipPersona 提供强大的交互基础。系统利用大模型的上下文分析功能，将用户模糊的语言描述转化为结构化的剪辑指令，结合视频内容和语义上下文，生成具体的剪辑方案。同时，大模型支持实时对话交互，分析用户反馈并提供优化建议，例如当用户对剪辑结果不满意时，模型能快速解析调整意图，生成更贴合需求的方案。这种自然语言理解能力显著降低操作门槛，使系统从被动工具转变为主动协作的表达代理，增强用户创作的流畅性和灵活性。

### 视频行为理解能力

大模型通过分析用户上传的视频和行为数据，深入理解创作者的表达逻辑和风格偏好，为个性化剪辑提供核心支持。在 ClipPersona 中，模型从用户过往视频中提取语言节奏、镜头选择逻辑和情绪表达特征，构建个性化的表达逻辑模型，确保剪辑结果与用户风格高度契合。此外，大模型结合用户对剪辑方案的反馈（如接受、调整或拒绝），通过强化学习算法动态优化行为模型，持续提升剪辑精准度。这种视频行为理解能力使系统能够捕捉用户独特的创作习惯，为后续生成动态、个性化的剪辑副本奠定基础。

### 人格卡形成能力

大模型通过持续学习和多模态生成能力，支持 ClipPersona 构建并优化“剪辑人格副本”及风格卡共享生态。基于用户上传视频和偏好标注，模型提取语言节奏、内容结构等特征，形成可动态调整的风格向量，作为剪辑任务的核心参数。这些风格向量不仅支持跨场景复用，还能通过用户反馈迭代优化，从静态模板进化到可持续进化的数字人格。此外，大模型的多模态能力助力风格卡共享平台，支持风格融合、推荐和预览功能，用户可快速体验、调整或交易他人分享的风格卡。这种人格卡形成能力推动了去中心化的创作生态构建，促进创作者间的风格交流，奠定了“表达样式市场”的技术基础。

### 交互优化能力

大模型驱动的对话式交互系统极大降低了用户操作门槛。用户可以通过自然语言或语音直接与系统沟通，表达剪辑需求，模型通过语义解析，将指令转化为具体的剪辑操作。同时，大模型支持实时反馈交互，能解释剪辑决策并提供优化建议。例如，当用户对某段剪辑不满意时，模型能分析用户调整行为，生成更贴合需求的方案。这种基于大模型的交互机制，使系统从被动执行工具转变为主动协作的表达代理。

# 项目计划

## 可行性分析

### 技术可行性

ClipPersona项目基于vivo提供的蓝心大模型矩阵API，构建了一套完整的技术实现方案。在核心技术层面，项目采用蓝心大模型的多模态理解能力实现语义解析与行为建模，通过MobileVLM-v2视觉语言模型和Whisper语音识别技术构建用户风格画像。

在系统架构方面，项目采用React Native跨平台框架与模型压缩技术，实现端侧轻量化部署方案。蓝心大模型提供的语义解析能力与MobileVLM-v2的结合，确保了自然语言指令与视频内容的精准映射。同时，项目集成了YOLOv11实例分割和E2FGVI目标消除技术，为视频处理提供了可靠的技术保障。

针对关键的技术挑战，制定了明确的技术路线：通过ONNX/TensorRT实现模型量化与加速，采用PPO强化学习算法优化反馈学习机制，持续提升人格卡系统的风格适配能力。

从开发周期来看，项目技术方案与赛事时间节点高度契合。核心模块开发阶段将完成人格训练、对话交互等关键功能，优化迭代阶段重点解决性能调优问题，最终阶段进行系统集成与压力测试。这种分阶段实施策略能够有效控制技术风险，确保项目按时交付。

### 市场可行性

目标用户群体（视频创作者、教育博主、情感类Vlogger等）对个性化剪辑工具的需求明确且持续增长。当前市场主流工具仍聚焦于基础剪辑功能，缺乏对用户表达风格的深度学习和个性化支持。ClipPersona提出的“人格副本”机制直击创作者风格沉淀与效率提升的双重需求，差异化优势显著。赛事合作方vivo的渠道资源与品牌支持，为产品后续发展提供了有力保障。项目设计考虑了从个人创作者到内容机构的扩展可能，具备商业化潜力。

### 经济可行性

项目开发以人力投入为主，充分利用赛事提供的技术资源，有效控制开发成本。收益模式设计兼顾短期赛事目标和长期发展路径，通过分阶段商业化策略实现可持续运营。项目技术成果具备可复用性，可延伸至其他AIGC应用场景，创造衍生价值。整体财务模型显示项目具备良好的投入产出比和可持续发展潜力。

## 团队介绍

### 导师介绍

指导老师沈莹长期从事人工智能和机器学习领域的研究，尤其在语音信号处理和自然语言处理方面拥有深厚造诣，发表过SCI/EI论文50余篇，被引用2000余次。研究领域与ClipPersona项目中 Whisper 语音识别、大模型自然语言交互及语义解析等核心技术紧密契合，为项目的技术创新和商业化落地提供了重要的技术支持与战略指导。

### 团队分工情况

尹诚成：负责技术架构的实现与优化工作，包括主导蓝心大模型API的集成与调优，开发人格训练模块的核心算法，并确保端侧轻量化部署方案的落地。同时负责技术文档的编写和代码质量把控。

戴湘宁：负责将“剪辑共生体”理念转化为可执行的产品逻辑。主要工作包括设计人格卡系统的交互范式（如风格切换、融合机制）、制定用户行为数据分析方案，并确保技术实现与产品愿景的一致性。

胡宝怡：负责构建项目的视觉语言体系，完成人格系统UI设计，开发系统交互界面，并制作比赛相关物料。

吴瑞翔：承担对话式交互系统前端开发、基础功能联调测试等任务，同时负责用户数据采集模块实现与压力测试执行。协助组长完成性能调优，参与核心算法攻关。

### 排期规划

①核心模块开发阶段：此阶段聚焦于完成ClipPersonaStudio个性建模系统、对话式交互系统以及自迭代剪辑副本机制等关键功能的开发。目标是实现人格训练、自然语言控制剪辑意图和行 为反向学习的基础闭环。

②优化迭代阶段：在核心功能初步实现后，我们将重点进行系统性能调优、用户体验优化 以及Bug修复。此阶段还将完善风格卡管理系统和副本共享平台，提升系统的稳定性和可用性。

③系统集成与测试阶段：项目后期将进行全面的系统集成测试和压力测试，确保各模块协同工作顺畅，并能应对高并发场景。同时，也将着手准备最终的作品展示和路演材料。

# 总结与展望

## 项目价值总结

ClipPersona通过"剪辑人格副本"的创新机制，实现了视频创作领域的三重突破。在技术层面，项目构建了基于蓝心大模型的表达风格学习系统，将多模态理解技术应用于视频剪辑领域，开创了“AI表达协作体”的新范式。在创作赋能方面，项目解决了创作者在内容量产与风格一致性之间的核心矛盾，通过人格卡系统和共情剪辑引擎，帮助用户实现“效率-个性-共情”的创作闭环。在行业生态方面，项目开创了“数字表达资产”的新形态，使创作风格成为可沉淀、共享和演化的数字资产。

## 项目完成度总结

截至复赛阶段，ClipPersona项目已完成核心框架搭建与关键技术验证工作。

在项目规划方面，项目已完成竞品分析，确立了“人格副本”的差异化定位，并制定了完整的开发路线图。同时，项目团队完成了人格设计方案的论证和部分模型的视频消除功能验证，为后续开发奠定了坚实基础。

在技术层面，基于React Native的跨平台UI框架已开发完成，核心界面如人格训练空间、风格卡管理、视频剪辑、社区互动等均已实现并具备较高的交互一致性。页面和组件体系已搭建完善，支持卡片式管理、沉浸式剪辑、多模态输入等关键功能；同时，蓝心大模型 API 的集成工作进展顺利，后端已完成与大模型的对接，实现了自然语言指令的解析与结构化，具备基础的语义理解和剪辑意图识别能力。

整体来看，产品已实现从前端交互到后端智能解析的完整闭环，为后续功能扩展和模型能力提升奠定了坚实基础。

## 展望

在未来的发展中，我们将探索更多的可能性，以提升产品的价值和影响力。

### 功能拓展

未来将重点发展跨模态风格迁移能力，支持将音频播客、图文内容自动转化为视频人格化表达。计划引入实时协作剪辑功能，允许多用户共同训练“团队人格副本”，适配MCN机构协作场景需求。同时探索AR创作扩展方向，结合vivo AR眼镜开发空间化剪辑交互体验，实现虚实融合的表达训练环境。在人格卡系统方面，将增强风格融合的精细度控制，支持更复杂的风格权重分配与动态调整。

### 技术创新

技术演进将聚焦于提升系统的自适应能力，通过优化强化学习算法，使剪辑副本能够更快地适应用户的创作变化。我们计划开发实时风格推荐系统，基于当前剪辑内容自动推荐最匹配的风格调整方案。同时将研究轻量化技术，在保证功能完整性的前提下，进一步降低系统资源占用，提升移动端使用体验。这些技术创新将大幅提升系统的实用性和易用性。

### 市场拓展

项目将采取分阶段的市场推广策略，初期聚焦于核心创作者社群，通过精准运营建立口碑。中期计划与内容创作平台合作，将剪辑人格系统整合到创作生态中。长期目标是构建开放的风格共享平台，形成创作者之间的风格交流社区。我们将持续收集用户反馈，不断优化产品功能，确保项目能够真正解决创作者的实际需求，在竞争激烈的剪辑工具市场中建立差异化优势。

**参考文献**

[1] Chang, Y.-L., Liu, Z., & Hsu, W. (2019). VORNet: Spatio-temporally consistent video inpainting for object removal. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1785–1794). IEEE. <https://doi.org/10.1109/CVPRW.2019.00229>

[2] Karale, S., & Saini, J. R. (2024). A novel steady and redundant frames removal performance analysis of video background subtraction algorithms with YOLOv models. In 2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC) (pp. 1–6). IEEE. <https://doi.org/10.1109/ICMNWC63764.2024.10872195>

[3] Yoon, J., Yu, S., & Bansal, M. (2024). Raccoon: Remove, add, and change video content with auto-generated narratives. arXiv. <https://arxiv.org/abs/2405.18406>

[4] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (pp. 28492–28518). PMLR. <https://arxiv.org/abs/2212.04356>

[5] Khanam, R., & Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. arXiv. <https://arxiv.org/abs/2410.17725>

[6] Li, Z., Zhang, C., Jia, Q., Lei, Z., & Zhou, Y. (2022). Towards an end-to-end framework for flow-guided video inpainting. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 17541–17550). IEEE. <https://doi.org/10.1109/CVPR52688.2022.01703>

[7] Chu, X., Yang, L., Wang, H., Zhang, X., Chen, J., & Zhang, Y. (2023). MobileVLM: A fast, strong and open vision language assistant for mobile devices. arXiv. <https://arxiv.org/abs/2312.16886>

[8] Chu, X., Yang, L., Wang, H., Zhang, X., Chen, J., & Zhang, Y. (2024). MobileVLM V2: Faster and stronger baseline for vision language model. arXiv. <https://arxiv.org/abs/2402.03766>

[9] Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In 2nd International Conference on Learning Representations, ICLR 2014. <https://arxiv.org/abs/1312.6114>

[10] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv. <https://arxiv.org/abs/1707.06347>