第十七届全国大学生软件创新大赛

文档编号：SWC2024-哇哈哈

TA-交互式陪伴与抚慰平台

TA- Interactive companion and comfort platform

技术研究报告

Version: 1.0.0

哇哈哈

2024.3.18

目录

文档修订历史

序号	修订原因	版本号	作者	修订日期	备注
1	创建文档	V0.1.0	队员 1	2024.3.15	
2	撰写文档	V0.2.0	队员 1	2024.3.16	
3	撰写文档	V0.3.0	队员 2	2024.3.16	
4	撰写文档	V0.4.0	队员 3	2024.3.17	
5	修改部分内容	V0.5.0	队员 2	2024.3.18	
6	定稿	V1.0.0	队员 1	2024.3.18	

1 问题聚焦

1.1 问题描述

至亲的离世对任何人来说都是巨大的打击，它带来的往往是无数个日夜的思念，无数抑制不住的眼泪。突然的变故往往会给身处这一状态的人带来极大的痛苦，甚至产生心理创伤，过渡期往往是痛苦的，难受的。另外，由于种种原因，还有人没能在至亲生命的最后时刻陪在他们身边，想说的话，想表达的爱都还未说出口，至亲便匆匆离开，留下了一生都无法弥补的遗憾。

1.2 问题抽象

我们想要通过 AIGC 的方式在一定程度上给以温暖和陪伴，并给出一个跨越时空对话的窗口，让有遗憾的人弥补遗憾，想表达爱的人表达出爱。

我们希望通过 AIGC 技术，在突发变故时给以目标群体安慰。这一点对应的技术问题为需要训练针对该类群体的**特定大模型**，针对目标群体特点生成适合的回复。我们的要求是具有**共情能力**，且需要能够**鼓励**其回归现实生活。我们还希望能给他们提供 AI **心理咨询**接口，使用户能够随时随地发现潜在的心理问题，及时主动干预。

我们还希望给这类群体提供吐露心声，弥补遗憾的窗口。这一问题对应需要我们解决模拟具有一定性格特征的人物回复问题，另外还需要解决**声音克隆**及根据图片的**面部动作生成**问题，且以上问题还具有一定的**实时性**要求。

整个生态需要一个 app 来呈现给用户使用，且针对目标用户特点，设计应尽量**温暖**，使用方式应尽可能**简洁**。这些均为软件**设计**问题。

1.3 问题定位

- 业务领域：心理关怀，生活服务
- 技术领域：AIGC，深度学习，共情大模型，面部动作生成，声音克隆

1.4 问题评估

- **技术性**：该问题需要对通用语言生成模型进行特定应用场景下的调优，同时结合计算机视觉、自然语言处理等多方面 AIGC 新兴技术，具有很强的技术性。

- **普适性**：该问题对应的目标群体庞大，几乎是人人都需要面对的问题。该问题的想法来源于真实需求，也来源于我们自身经历与网络评论的真实反馈，具有很大的普适性。
- **价值点**：解决该问题能在一定程度上使很多人打开心结，拥抱现实生活；同时也能在陷入痛苦时获取温暖与安慰，相信经过我们不断的完善与版本迭代，这款产品可以产生很大的价值。

1.5 问题分解

1.5.1 适应特定场景的大模型训练与提示词工程

我们项目的目标群体明确，且具有鲜明的特征，在不同功能下所需大模型的能力有不同的侧重，具有清晰的要求。因此，我们需要针对再见一面、AI 小伴及 AI 心理咨询三个需要大模型生成的功能分别针对性的微调模型，使其适应各个模块的特定需求。其中再见一面模块注重回复内容的正向鼓励，且回复需要尽量拟合用户对个人的描述；AI 小伴模块注重共情能力，同时也需要具有正向鼓励的倾向性；AI 心理咨询注重专业性，能从专业视角给出判断。

1.5.2 语音克隆

语音克隆模型用以支撑再见一面功能，需要根据用户输入的样本音频尽可能模拟音频中的声音，并从大模型得到的文本输出模仿音频。

1.5.3 基于图片的面部动作生成

面部动作生成用以支撑再见一面功能，需要根据用户输入的图像及语音克隆模块输出的音频输出一段图像中人物说话的视频。

1.5.4 内容检测与过滤

该功能用以解决漂流瓶模块中潜在的内容安全威胁，防止恶意中伤等问题出现。

1.5.5 软件本体开发

软件本体开发用以集合所有服务，构建整体生态。其中的重点包括前后端开发，界面设计，简介的交互逻辑及隐私协议、使用声明等内容。

其中，大模型、语音克隆和面部动作生成为串接关系，后者的输入依赖于前者的输出。所有内容均以软件本体为载体。

2 相关工作

2.1.1 大模型微调

当前大模型发展迅猛，2023 年以来各大巨头均针对大模型发力，如百度文心、智谱 GLM、OpenAI GPT 等均提供了可自输入数据集进行微调的大模型接口。我们将使用他们提供的微调接口进行调整。另外的一条途径是通过提示词工程来直接给予大模型指令，进而使其扮演特定角色。

2.1.2 语音克隆技术

语音克隆技术最近也有较大发展，出现了如MockingBird、Bark-Voice-Cloning及最新的 GPT-SoVITS 等。其中效果较好的是 GPT-SoVITS，其他如 MockingBird 均会产生一定程度的杂音，影响效果。

2.1.3 面部动作生成技术

目前有一些数字人生成技术能够对某个人进行模拟，但有些需要的数据量过于复杂，并不适合我们面向的场景。经过调研，目前适合我们场景的面部动作生成技术包括 wav2lips、SadTalker 等。

2.1.4 内容检测与过滤技术

目前的内容检测与过滤技术大多仍集中于关键词屏蔽，有部分团队正在基于大模型做语义层面的内容过滤。

另外，针对问题本身，目前市场上的陪伴类软件几乎都不含有 AIGC 元素或使用相关技术，大都是基于人与人直接交互的方式提供倾诉渠道，来缓解用户孤单情绪或是倾诉欲望。

3 技术方案

3.1 技术方向

我们所研究的技术方向主要为以下四点

- **自然语言处理**，即大模型在特定应用场景下的微调
- **计算机视觉**，即面部动作生成相关内容
- **语音信息处理**，即语音克隆相关内容
- **软件本体**，即保证软件本身运行安全可靠，满足性能要求

3.2 技术选择

- 在大模型微调上，我们根据首轮测试效果，选择百度文心大模型进行微调，使其适应我们的场景要求。
- 在语音克隆上，我们选择 GPT-SoVITS 模型，但由于目前其使用方式为 webui 或命令行图形界面，且需要用户自行调整参数进行训练，因此我们需要对其主要内容进行抽取与组合，并测试一个较好的参数集合。
- 在面部动作生成上，我们选择 SadTalker 作为生成模型，因为其效果较好，至于实时性的要求，我们正在进行模型的 TensorRT 转换工作，将会极大提升其生成速度。
- 内容检测与过滤方面，我们将首先部署关键词检测，并在词库中添加我们目标场景下的不合适的关键词。在此基础上尝试基于语义的不安全内容识别与过滤。

3.3 结果期望

在再见一面功能模块中，对于用户输入的性格特征、面部信息、声音信息，能够输出合适的文本、生成的视频和语音效果逼真，做到可接受范围内一定的实时性，后期进行调优。

在 AI 小伴及 AI 心理咨询模块中，对于用户输入，能够生成符合基本要求的回复。AI 小伴能给用户以鼓励和陪伴，AI 心理咨询师能帮助用户解决某些实际的问题，并能鼓励其及时就医。

在漂流瓶模块中，除基本的发送、接收漂流瓶外，应能够对可能触及心理健康的词进行识别和屏蔽，避免不安全因素。

4 技术实践

4.1 使用的开发框架及依赖的库

4.1.1 软件本体开发部分

所在部分	主要关键技术/框架/工具	技术解释
后端	Spring Boot	开源的 Java 框架，它简化了基于 Java 的应用程序的构建、部署和管理过程。帮助开发人员更快地创建独立的、生产级别的应用程序，而无需过多的配置和样板代码。
	Spring MVC	Spring MVC 是 Spring 框架提供的一款基于 MVC 模式的轻量级 Web 开发框架，将 Web 应用进行职责解构，简化了 Web 应用的开发。
	RabbitMQ	是一个开源的消息中间件，用于在分布式应用程序之间传递消息。它实现了高级消息队列协议（AMQP）标准，提供了可靠的消息传递机制，广泛用于构建分布式和异步通信系统。
	MyBatis Plus	基于 MyBatis 的增强工具，简化了基于 Java 的持久层开发。它提供了一组强大的功能和工具，可以帮助开发人员更快地进行数据库操作和 CRUD（创建、读取、更新、删除）操作，减少了繁琐的重复代码，提高了开发效率。
	Redission	Redission 是一个在 Redis 的基础上实现的 Java 驻内存数据网格（In-Memory Data Grid）。提供了一系列的分布式的 Java 常用对象，以及许多分布式服务。
	Knife4j	Knife4j 是基于 Spring Boot 构建的一个文档生成工具，它可以让开发者为我们的应用生成 API 文档。可以更加方便的基于 API 文档进行测试。
	Hutool	Hutool 是一个开源的 Java 工具库，旨在简化 Java 开发中的常见任务，提供了丰富的工具类和方法。

数据存储	MySQL	MySQL 是基于 SQL 查询、强大、稳定且易于使用的关系型数据库管理系统（DBMS）。
	Redis	Redis 是一个开源的内存数据库，也被称为数据结构服务器，它提供了高性能、持久化的数据存储和检索。
	COS 对象云存储	存储海量文件的分布式存储服务，可通过网络随时存储和查看数据。具备高扩展性、低成本、可靠和安全的特性。
测试	Spring JUnit	基于 Spring Boot 的单元/集成测试工具，提高项目可靠性
前端	Cordova	cordova 是一个混合式应用开发框架，其中有丰富的插件库支持。开发者编写一套代码，可发布到 iOS、Android、Web（响应式）等多个平台。
	Vue3	Vue 是一个流行的开源 JavaScript 框架，用于构建用户界面和单页面应用程序。
项目依赖管理工具	Maven	Maven 是一个开源的项目管理和构建工具，用于管理 Java 项目的构建、依赖管理和发布。提供一种标准化的项目结构和构建流程。
项目版本管理工具	Git	开源的分布式版本控制系统，可以有效、高速地处理从很小到非常大的项目版本管理。
项目部署工具	Docker	Docker 是一个开源的应用容器引擎，它允许开发者将应用程序及其依赖项打包到一个轻量级、可移植的容器中。容器可以在任何支持 Docker 的操作系统上运行

4.1.2 文本生成部分

文本生成最终依赖百度文心大模型和智谱 AI GLM4 模型。

4.1.3 语音克隆部分

语音克隆使用参数调整后的 GPT-SoVITS 克隆模型。

4.1.4 视频生成部分

使用 SadTalker 对 GPT-SoVITS 生成的语音进行对应的面部动作生成。

4.2 技术实践过程

在文本生成上，我们首先对大量的大语言模型进行测试分析，发现国内模型对中文文本的理解和生成能力要更强，根据对比，我们选择了百度文心大模型。然后我们开始进行 prompt 提示词的构造。构造出一个基本的能满足条件的 prompt 后，我们进入模型微调阶段，通过组内同学的讨论和不断思考，我们自己构造了一个小型数据集，但初次微调结果不佳，模型并未有明显改变，后来我们通过加大数据集规模等方式，看到了一定效果。在此基础上我们进一步优化 prompt 并进行带有攻击性、破解性的测试，最终确定 prompt。但心理咨询师总存在回复过长，不会引导的问题，因此我们又尝试使用智谱最新模型 glm4，获得了较好的效果。

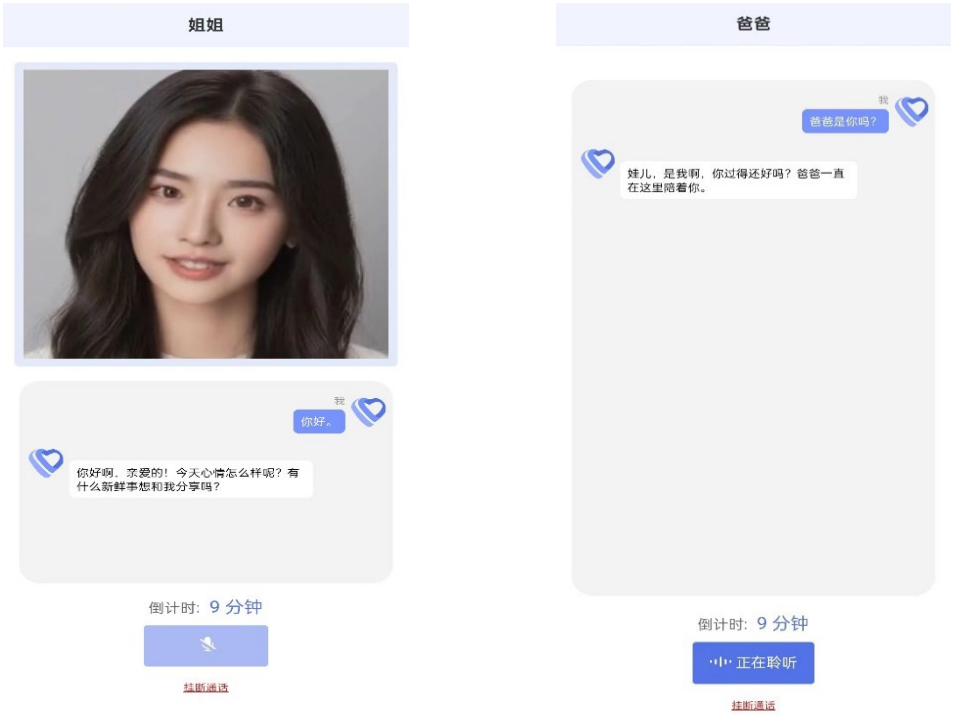
在音频生成上，我们首先使用 MockingBird，但经过测试发现其电流杂音问题严重，甚至影响视频生成效果。因此我们决定替换 GPT-SoVITS。但 GPT-SoVITS 的使用形式为 webui 或命令行界面，需要用户自行调整参数进行训练，这并不符合我们的需求。为此，我们从其源码中提取相关函数并组成我们需要的序列，实现模型的自动化训练以及根据用户 id 的自动化推理。在测试中我们又发现了自动化生成的模型存在一定概率效果不佳的问题，我们又通过不断调优参数使这种情况发生的概率降低，达到要求。

在视频生成上，我们原想通过给出参考动作的方式预生成一组可播放视频，在语音生成后随语音一起播放，伴有嘴部的微动，这种实现方式可以保证绝对的实时性。但经过测试发现，对不上嘴型是非常影响整体效果的，因此我们果断放弃这种方式，转而使用 SadTalker 作为我们的视频生成模型，其需要更高的算力才能支撑实时性，但其效果也是不可比拟的。且我们通过预生成过渡视频的方式使交流更加流畅和沉浸，从用户的角度看，视频是不会停止的，极大增强了真实感，同时一定程度上弥补了实时性稍弱的缺陷。

最终，我们将几个技术进行串接，对外开放视频和音频交流方式下的训练、推理接口，实现了整套流程。

5 结果验证

上述整套流程涉及的技术均已实现，语音和视频交互的最终界面如下，由于我们拥有的算力不足，目前语音的响应时间在 10s 左右，视频的响应时间在 30s 左右。针对该问题我们还在使用硬件升级和模型的 TensorRT 加速，预计加速倍数将达到 5-7 倍，能够很大程度压缩响应时间。



AI 小伴及 AI 心理咨询模块的最终呈现如下，模型已经可以输出符合需求的

文本，给用户以情感安慰和正向鼓励。