

湖 北 大 学

本 科 毕 业 论 文 (设 计)

题 目: 基于大模型高质量情感对话虚拟人

姓 名: _____

学 号: _____ XXXX

专业年级: _____ XXXX 级

学 院: _____ XXXX 学院

指导老师: _____ XXXX

二 零 二 四 年 三 月 九 日

基于大模型的高质量情感人虚拟人系统

摘要

随着深度学习的发展，虚拟人的技术也迎来了巨大的突破。虚拟人，也常被称为虚拟角色或数字人物，是通过计算机技术生成的三维或二维的数字化人类形象。这些人物可以是完全虚构的，也可以是基于现实生活中的人物建模的。在技术上，一般采用的做法是，使用文本或者音频对一张图片或者一个视频进行驱动，生成逼真的人物视频。虚拟人在娱乐产业以及新闻产业等产业已经取得广泛应用，并且产生的了巨大的价值。在计算机视觉研究领域中，已是一个比较前沿的热点话题。但是目前虚拟人在商业领域取得了巨大的成功，仍然存在很多的问题，例如，感情表达的不真实，分辨率较低等，没有一个相对较好的交互系统流程等。本文在以下几个方面开展了研究：

1. 为了使交互流程更加清晰，本文提出了一个交互流程结构，通过 ASR(Automatic Speech Recognition，即语音转文本模块)、LLM (Large Language Model，即大语言模型模块)、TTS(Text to Speech，即文本转语音模块) 以及 Talking Head Generation(说话人脸生成模块)，实现用户与电脑对话，可以得到一个指定表情的说话人脸视频。
2. 为了使 EAT 模块能够更好的集成到系统中，我们对原有的 EAT 的推理方式进行了一些改变，通过将改进后的 EAT 推理模块拿进我们的系统主类中实现了对 EAT 推理模块的灵活运用。
3. 为了使人脸更加清晰，采用 GFPGAN 超分模块对生成的视频帧进行超分，GFPGAN 使用了 Tensorrt 模块进行加速，使得 GFPGAN 的速度提升到了一个“值”速度，使得系统的反应速度大大提升。
4. 使用了一个深度学习中常见的 Gradio 用户友好型的交互界面，用户通过上传音频和指定的图片，并可以选择不同的 pose 和表情标签，生成指定任人物的说话人视频。并且可以与 ChatGLM2-6B 大模型进行交互。

【关键词】 虚拟人 大模型 情感

High-Quality Emotional Conversational Avatars Based on Large Language Models

Abstract

Alongside the thriving development of the deep learning field, virtual humans have also seen tremendous breakthroughs. Virtual humans, also commonly referred to as virtual characters or digital avatars, are computer-generated 3D or 2D digital human images. These characters can be completely fictional, or they can be modeled after real-life people. Technologically, the typical approach is to use text or audio to drive a picture or video, generating a realistic human video. Virtual humans have found widespread application in the entertainment and news industries, generating significant value. In the field of computer vision research, it has become a relatively cutting-edge topic. However, while virtual humans have achieved great commercial success, there are still many problems, such as unrealistic emotional expression, low resolution, and a lack of a relatively good interactive system workflow. This paper conducts research in the following aspects:

1. To make the interaction process clearer, this paper proposes an interaction process structure, using ASR (Automatic Speech Recognition), LLM (Large Language Model), TTS (Text to Speech), and Talking Head Generation, to realize user-computer dialogue and obtain a talking head video with a specified expression.

2. In order to better integrate the EAT module into the system, we made some changes to the original reasoning approach of EAT. By incorporating the improved EAT inference module into our system's main class, we achieved flexible utilization of the EAT inference module.

3. To make the faces clearer, the GFPGAN super-resolution module is used to super-resolve the generated video frames. GFPGAN uses the Tensorrt module for acceleration, which significantly improves the response speed of the system.

4. A user-friendly Gradio interface, a common deep learning tool, is used. Users can upload audio and specify images, poses, and expression labels to generate talking head videos of the target person. Additionally, the system can interact with the ChatGLM2-6B large model.

【Key words】 Virtual Humans Large Language Model GFPGAN Emotion

目 录

图目录	vi
1 绪论	1
1.1 研究的背景和意义	1
1.2 国内外研究现状	1
1.2.1 基于规则引擎的虚拟人技术	1
1.2.2 基于三维建模的虚拟人技术	1
1.2.3 基于深度学习的虚拟人技术	2
1.3 研究内容及主要工作	2
1.3.1 研究内容	2
1.3.2 论文章节安排	2
1.4 本章小结	3
2 相关技术概述	3
2.1 生成对抗网络	3
2.1.1 数据分布与损失函数	4
2.1.2 优化目标	4
2.1.3 特性与问题	4
2.2 对话虚拟人系统	5
2.2.1 语音识别模块	5
2.2.2 对话系统模块	6
2.2.3 文本转语音模块	6
2.2.4 对话头像视频生成模块	6
2.3 本章小结	7
3 核心模块及改进方法	8
3.1 EAT 模型	8
3.1.1 EAT 模型提出背景	8
3.1.2 EAT 模块分析	8
3.1.3 EAT 第一阶段	9
3.1.4 EAT 第二阶段	10
3.1.5 EAT 模型改进	12
3.2 GPT-SoVits	13
3.2.1 优点及特性	13
3.2.2 训练和推理	13
3.3 ChatGLM2-6B	14
3.3.1 优点及特性	14
3.3.2 部署方式	14
3.4 FunASR	14
3.4.1 优点以及特性	15
3.4.2 模型调用	15
3.5 本章小结	15
4 高质量情感对话虚拟人系统设计与实现	15
4.1 系统设计	15
4.1.1 系统的整体架构	16
4.1.2 网络模型的部署	17

4.1.3 系统界面的展示	18
4.2 系统功能实现	18
4.2.1 前端界面设计	18
4.2.2 Gradio 组件介绍	18
4.3 功能模块的实现	18
4.3.1 驱动动作模块	18
4.3.2 表情模块	19
4.3.3 对话大模型模块	20
4.3.4 ASR 模块	21
4.3.5 TTS 模块	22
4.3.6 抠图模块	22
4.3.7 对话人视频生成	22
4.3.8 抠图模块	23
4.3.9 超分模块	24
4.4 本章小结	24
5 系统测试	25
5.1 功能测试	25
5.1.1 交互模块测试	25
5.1.2 驱动动作测试	27
5.1.3 表情测试	29
5.1.4 大模型测试	30
5.1.5 ASR 测试	31
5.1.6 TTS 测试	32
5.1.7 抠图测试	33
5.1.8 GFPGAN 超分辨率测试	34
5.2 本章小结	36
6 总结与展望	37
6.1 全文总结	37
6.2 致谢及展望	37
参考文献	38

图目录

图 2.1 GAN 基本结构	3
图 2.2 虚拟人系统结构图	5
图 3.1 EAT 结构图概览	8
图 3.2 EAT 详细结构图	9
图 3.3 EAM 模块流程图	11
图 3.4 ReposeNet 流程图	12
图 4.1 系统架构图	16
图 4.2 网络模型初始化	17
图 4.3 动作处理流程图	19
图 4.4 表情模块处理流程图	20
图 4.5 大模型模块流程图	21
图 4.6 ASR 模块	22
图 4.7 EAT 模块	23
图 4.8 EAT 数据处理流图	23
图 4.9 抠图流程	24
图 4.10 超分流程	24
图 5.1 音频模块界面和音频设备选择测试	25
图 5.2 音频裁剪与音频检阅测试	26
图 5.3 图片上传方式选择测试	27
图 5.4 驱动动作选择测试	28
图 5.5 三种不同的驱动动作结果	28
图 5.6 表情结果展示	29
图 5.7 主流模型结果比较	30
图 5.8 大模型对话测试	30
图 5.9 测试清除对话历史功能	31
图 5.10 双语 ASR 测试结果	32
图 5.11 双语 TTS 测试结果	33
图 5.12 抠图测试	34
图 5.13 视频帧超分结果展示	35

1 绪论

1.1 研究的背景和意义

现如今，随着人工智能的发展，越来越多成熟的人工智能的产品已经应用于各个领域。例如，在医疗领域，基于大数据，AI 可以帮助医生识别疾病模式，为患者提供个性化的治疗方案，提高诊断的准确性，典型的产品有 Google DeepMind 的眼科诊断工具以及 IBM 的 Watson Oncology 在癌症治疗领域提供的帮助。又例如，在直播领域，虚拟人也可以产生巨大的商业价值，近年爆火的虚拟主播“嘉然今天吃什么”已经拥有超过千万的粉丝，以及基于语音合成软件 VOCALOID 系列制作的女性虚拟歌手“洛天依”由于其巨大的影响力于 2024 年 1 月 26 日，入选“2023 年度十大虚拟数字人”；在如今短视频爆火的年代，虚拟人可以与短视频广告等结合发挥巨大的商业价值^[1]。

不仅如此，据文献^[2]，随着我国的老龄化加剧，到 2025 年，我国 60 岁及以上老年人口预计将达 3 亿人，占总人口比重将超过 20%，这一趋势将给我国养老保障体系带来巨大的压力，养老需求将持续增长。虚拟人技术在养老服务的应用可以从多个方面进行拓展，以应对老龄化加剧和空巢老人增多的挑战。一方面，虚拟人可以提供社交陪伴，通过高度仿真的交互，缓解老人的孤独感和社会隔离。另一方面，它们可以辅助日常生活，如提醒用药、健康监测和简单的家务助理，减轻传统养老服务的人力成本压力。此外，虚拟人还能提供定制化的娱乐和教育服务，如音乐、故事讲述和兴趣学习，丰富老年人的精神生活。综上所述，虚拟人技术的引入不仅能提升养老服务的质量和效率，还能在一定程度上降低整体服务成本，为解决老龄化社会问题提供新的解决方案。但是，现在主流的虚拟人主要是基于 Wav2lip^[3] 和 SadTalker^[4] 生成的，它们的主要问题是生成的人脸缺乏情感，交互体验差。但是，感情作为人与人交流的一个重要的纽带，在对话虚拟人系统中显然是不可或缺的组件。由此可见，情感虚拟人系统的巨大研究价值。

1.2 国内外研究现状

尽管，虚拟人的相关技术取得了很大的进步，但是目前仍然存在很多问题。如交互的时间比较长，生成的视频质量不高，生成的虚拟人表情缺失等问题，并且没有一个完整的系统可以做到正常的与人沟通。本课题将致力于解决情感缺失的问题，并搭建一个完整的虚拟人系统模块。

虚拟人系统的发展脉络可以追溯到 20 世纪中叶，随着计算机技术的进步和人工智能的发展，虚拟人系统逐渐成为研究的热点之一。

1.2.1 基于规则引擎的虚拟人技术

早期的虚拟人系统主要基于规则引擎，通过预先定义的规则和模式匹配来进行对话交互。1966 年，MIT 的 Eliza 程序^[5]引起了人们的关注，它是一种模拟心理医生的程序，它模拟了心理医生的对话过程，能够进行基于模式匹配的简单对话。Eliza 的成功吸引了广泛的注意，人们开始意识到计算机能够模拟人类语言交流的潜力。此后，随着计算机图形学和人机交互技术的发展，80 年代出现了一些简单的虚拟人系统，如 Julia 和 Racter。这一阶段标志着虚拟人系统的初步探索，为探索人机交互提供了可能性。

1.2.2 基于三维建模的虚拟人技术

在这一阶段，虚拟人系统得到了更大的关注和发展并且这个阶段的虚拟人系统开始向着虚拟现实领域发展，大多采用了三维建模技术来创建虚拟人的外观。1990 年代末期和 2000 年代初期，随着虚拟现实技术的兴起，出现了一些具有三维虚拟形象的虚拟人系统，如 Julie 和 Maxine。

1.2.3 基于深度学习的虚拟人技术

近现代的虚拟人系统具有更高的智能性和交互性。随着深度学习和自然语言处理技术的迅速发展，虚拟人系统不仅能够模拟情感，还能够理解语境并进行更自然的对话。现代虚拟人系统广泛应用于各种领域，包括虚拟助手、在线客服、虚拟偶像等。它们能够提供个性化的服务，为用户提供更好的体验的同时产生巨大的商业价值。

目前可以在一些社交媒体上看到一些基于 SadTalker^[4] 和 Wav2Lip^[3] 等方法的虚拟人形象，基于这些方法的虚拟人系统大多只和参考视频或参考图片的表情相关，结果呈现出单一和情感不可控的效果。同时由于分辨率的影响所以 SadTalker 分辨率 256x256) 和 Wav2Lip 分辨率 96x96) 的方法都需要对生成的视频结果帧做超分处理，但是前者中的超分辨率模块 face enhancer 处理起来非常的慢。本文提出的虚拟人系统是为了探索目前的虚拟人系统生成的视频缺乏情感并无法对生成人脸的表情进行控制的问题以及超分辨率模块加速的问题。

1.3 研究内容及主要工作

1.3.1 研究内容

本文旨在探索和开发一种高质量情感虚拟人系统。从实际交互场景出发，研究人机交互过程中实际需要的模块，将各个模块设计在一起，并对其中的一些重要模块进行改进，增强交互体验。

本文首先情感出发，选取一个合适的说话人脸视频生成模块 EAT^[6]，这个模块的作用是通过输入的音频驱动一个源图片生成一个情感对话头视频。然后从实际的场景出发，对 EAT 模块进行改进，加速推理过程和对结果进行优化，显著提高了交互的体验。本文主要的工作如下：

1. 探索并构建了一个完整的虚拟人交互系统，用户可以与一个由指定人物形象的大模型进行对话与交流，并且可以对虚拟人的形象进行指定，通过对感情标签的配置，实现对生成虚拟人表情的控制。大模型作为虚拟人的大脑可以回答用户的各种问题。

2. 为了提升视频结果的质量，本文采取了比较先进的超分辨率策略 GFPGAN 对结果视频帧进行处理，分辨率由原来的 256*256 到现在的 512*512，分辨率提升了 4 倍，并且鉴于 GFPGAN 的推理速度非常缓慢，本系统采取了使用 TensorRT 对 GFPGAN 的推理进行加速，使得整个交互过程的时间大大缩短。

3. 为了个性化的声音，本系统借鉴 GPT-SoVits 的方法，可以实现对指定人物的声音的克隆，用户可以指定虚拟人的声音，满足更加多样化的需求。

4. 本系统引入了情感分析模块，通过分析文本的情感，对用户传入的音频经过 ASR 转换后分析出用户当前的心情，从而使大模型做出更加人性化的反馈。

5. 为了提高易用性，本文基于 gradio 设计了一个美观的，易于交互的界面。并对相应功能进行了测试。

1.3.2 论文章节安排

本文主要使用五个章节对高质量情感人系统进行设计与研究，每个章节安排如下：

第一章，绪论，首先说明本文的研究背景与意义。随后介绍了对话虚拟人的相关技术在国内外的研究现状和技术发展脉络，最后对本文研究的内容和主要的工作做了总结说明。

第二章，相关技术概述，首先是说话人脸生成领域比较常用的 GAN 网络，它具有强大的生成能力，能够生成逼真的图像或者视频；其次是本系统的涉及到的相关技术的总体介绍，包括自动语音识别（ASR）、文本转语音（TTS）、大语言模型（LLM）、说话人脸视频生成（Talking Head Generation）

第三章，系统核心算法，主要是介绍本系统的核心算法 EAT，即情感人脸生成模块算法，详细介绍算法的核心思想和对其的改进方法。在其原本的推理基础上使用 Tensorrt 加速推理，极大的缩短了交互时间。并且，为了更加个性化，引入了克隆 TTS 的模块，可以对语音进行定制化的克隆操作，实现指定声音的高质量情感虚拟人的生成。

第四章，系统设计，首先对高质量情感虚拟人系统进行功能设计，主要研究使用的框架和具体的技术方案等，其次根据具体的交互需求，设计功能模块，以及相关的优化策略。

第五章，系统测试，对系统的功能模块进行测试。

第六章，总结和展望。主要是对本文的研究的高质量情感虚拟人系统做总结与对未来的展望。

1.4 本章小结

本章节为全文的绪论部分，首先，本文对目前虚拟人系统的相关技术和发展脉络要点进行了详细的介绍，其次阐明了本文的研究价值，最后对本文研究内容和论文的章节安排进行了简略说明。

2 相关技术概述

2.1 生成对抗网络

生成对抗网络（GAN，即 Generative Adversarial Network）是一种极具创新性的深度学习模型，自 2014 年由 Ian Goodfellow 等人首次提出以来，就在人工智能领域中引起了广泛的关注。GAN 网络通过内部的两个主要组件——生成器和判别器的对抗性协作，能够生成高质量、高度逼真的数据，广泛应用于图像生成、视频生成、语音合成等多个领域。如图 2.1 所示，GAN 网络包括两个基本的部分：生成器（Generator）和判别器（Discriminator）。我们将从数据分布的角度以及其优化目标和特性与问题，来解释生成器和判别器的相关概念：

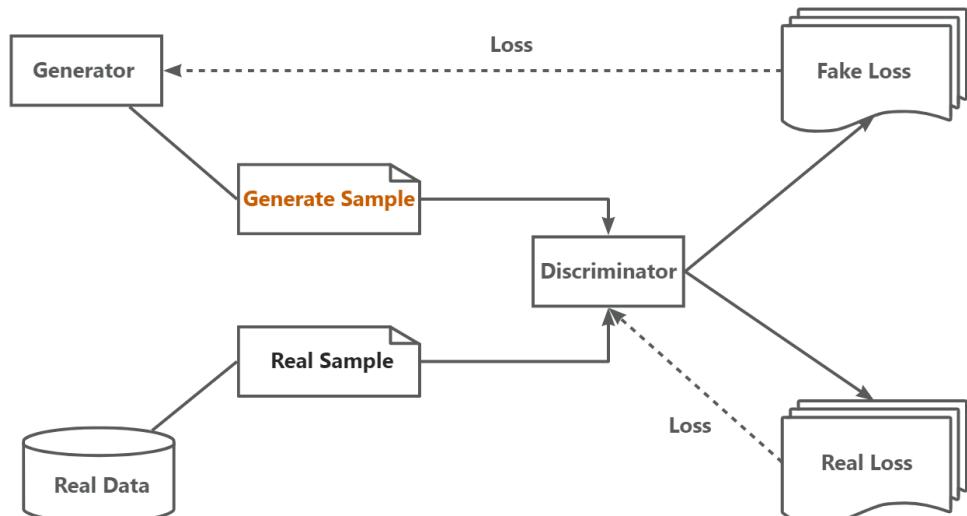


图 2.1: GAN 基本结构

2.1.1 数据分布与损失函数

假设有一个真实的数据集，我们可以用一个概率分布 p_{data} 来描述这个数据集中每一个数据点出现的概率。在理想情况下，这个分布完全代表了真实世界数据的统计特性。生成器 G 试图模仿这个真实的数据分布 p_{data} 。它通过接收一个先验分布 p_z （通常是高斯分布或均匀分布）上的随机噪声 z ，映射这个噪声到数据空间，从而创建假的数据点 $G(z)$ 。生成器的目标是调整其参数，使得 $G(z)$ 的分布 p_g 尽可能的接近 p_{data} 的分布。

GAN 的损失函数如下：

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

2.1.2 优化目标

GAN 的整个训练过程是交替进行的，第一个步骤训练判别器，第一个步骤训练生成器，这两个步骤在训练过程中交替进行，通常是固定一个组件（如固定 G 训练 D ，或固定 D 训练 G ）进行数次迭代更新，然后切换到另一个组件。这种交替策略有助于平衡生成器和判别器的学习速度，避免一方过早地占据优势导致训练失败。

对于判别器 D 而言，其优化目标是最大化其正确标记真实数据和生成数据的概率，所以它回去区分生成器生成的假数据和实际的数据样本，在训练过程中，我们优化 $\max_D V(D, G)$ ($\mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$)，对每轮的 D 的参数 θ_d 进行梯度相加。使损失函数 (G, D) 越大越好。

对于生成器 G 而言，生成器 G 的目标是生成尽可能逼真的数据以“欺骗”判别器，让判别器认为数据是真实的。在训练过程中，我们优化 $\min_G \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$ ，对每轮的 G 的参数 θ_g 进行梯度相减，使损失函数 $V(G, D)$ 的值越小越好。

在不断的训练过程中，当最后，在理想情况下，判别器 D 的最优策略会收敛到如下式子中：

$$D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}$$

其中 $p_{data}(x)$ 是样本 x 来自真实数据的概率密度， $p_g(x)$ 是样本 x 来自生成器的概率密度。此时，判别器已经无法区分真实数据分布和生成器生成的数据分布之间的差别，于是 $D(x) = \frac{1}{2}$ 。

2.1.3 特性与问题

GAN 是一种十分强大的无监督生成模型，生成器从随机噪声中开始生成样本，使得它可以极少的初始信息出发自动学习无法显式表示的原始数据集的复杂的样本分布。无论这个分布多么的复杂，只要模型训练的足够好，就可以生成机器逼真的样本集。并且相较于其他一些生成模型，GAN 生成的图和视频在细节上通常更加丰富。

但是，GAN 的训练过程非常不稳定，常常需要精心设计的网络结构和精确调整的训练参数。训练中，尝尝会出现梯度消失的问题和模式崩溃的问题。

梯度消失在某些情况下可能发生，当判别器可能会变得过于优秀，以至于生成器的梯度消失，使得生成器不能进一步学习和改进，即无论生成器生成的任何数据样本，判别器都会认为是“假”的。

模式崩溃也是 GAN 训练过程中的一个常见的问题，指的是生成器开始生成非常相似或完全相同的输出，而不是多样化的结果。这种结果产生的原因是，生成器找到了一种“欺骗”判别器的方法，通过输出某个特定的样本类型来最大化其得分。即使这些数据和真实数据看起来很相似，但是这些数据的多样性非常低。

目前，GAN 已经普遍应用于图像重建生成及图像编辑与增强领域、视频生成、视频编辑领域和风格迁移领域，具有非常巨大的影响力。

2.2 对话虚拟人系统

根据文献^[7]，多模态人机交互是虚拟人应用的一个重要的方向。该系统旨在使用深度学习模型生成具有自然特征的交互对象。如图所示，包括语音识别、对话系统、文本转语音和虚拟人视频合成等。下面将分别从这几个关键模块展开阐述。

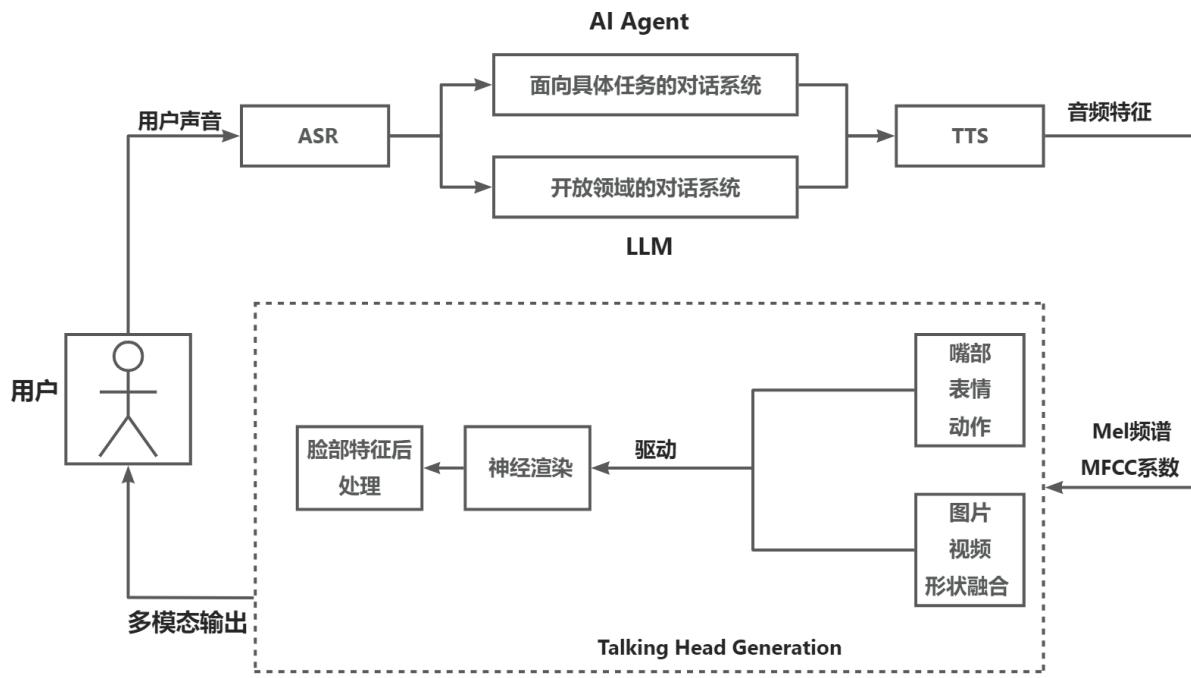


图 2.2: 虚拟人系统结构图

2.2.1 语音识别模块

语音识别 (Automatic Speech Recognition, 即 ASR)，指的是将语音信号转换为对应的文本的技术。它在过去几十年中经历了长足的发展，从最初的传统的机器学习方法，再到近年来的深度学习方法。

把统计模型的方法引入 ASR 的想法是由 Jelinek 等人最先提出的。Jelinek 等人提出了一个基于隐马尔可夫模型 (Hidden Markov Model, 即 HMM) 的语音识别系统^[8]，该系统使用 HMM 建模语音的时序性。他们将语音信号划分为较短的时间段，并使用 HMM 来建模每个时间段的声学特征。通过建立状态转移概率和发射概率矩阵，系统可以根据观察到的声学特征序列来推断出最可能的词序列。但是由于 HMM 的方法非线性建模能力有限，使用单个高斯分布往往无法捕捉到复杂的声学特征分布，很快便又有研究人员提出了高斯混合模型-隐马尔科夫模型 (Gaussian Mixture Model- Hidden Markov Model, 即 GMM-HMM) 的方法^[9]，使用 GMM 来建模语音的声学特征。GMM 可以更加灵活地建模语音信号的声学特征，不仅仅局限于单个高斯分布，通过将多个高斯分布组合成混合模型来建模复杂的声学特征分布，每个高斯分布代表声学空间中的一个聚类，从而可以更好地逼近实际的声学特征分布。

近年，研究人员通过将深度学习的方法引入 ASR 领域，使 ASR 的准确度又上了一个台阶，并且端到端的结构使得 ASR 的效率得到的很大的提升。目前端到端的 ASR，例如百度提出的 DeepSpeech^[10]，Openai 的 Whisper^[11]等都取得了比较好的效果，并且可以达到商用的水准。

2.2.2 对话系统模块

对话系统在虚拟人系统中是作为语义理解的核心模块，可以看作是虚拟人的“大脑”，它可以生成文本、翻译语言、进行内容摘要、分析情感、回答问题，并理解自然语言，功能非常的强大。例如，由 Google 的 Open AI 团队开发的 Chat GPT3.5 和 GPT-4^[12]；由 Facebook 的 Meta 团队开发的 Llama1^[13]和 Llama2^[14]；由清华大学团队开发中文大语言模型 Chatglm-6B^[15]具有非常强大的能力和实用性。同时，大模型微调的应用可以使得模型更具个性化，通过使用指定范围的数据对大模型进行训练，可以使得大模型成为某个领域的“专家”，解决更具有针对性的问题。

2.2.3 文本转语音模块

文本转语音（Text to Speech，即 TTS），TTS 是将自然语言文本转化为语音的过程。据文献^[16] TTS 根据研究的方法可以分为传统的机器学习的方法和深度学习的方法。传统的机器学习的方法又包括了连续语音合成（concatenative speech synthesis）和参数语音合成（parametric speech synthesis）等，受到模型的限制，这些方法的合成质量都较低。近年来，由于深度学习发展，更多高效的方法被引入到了 TTS 中。早期主要使用卷积神经网络（Convolutional Neural Network，即 CNN）加循环神经网络（Recurrent Neural Network，即 RNN）的方式，这种 TTS 模型存在长依赖问题、并行计算困难和编码器-解码器不一致等缺点^[17]，现在比较主流的方法是基于 Transformer^[18]的模型，这类模型可以在输入序列中捕捉全局的依赖关系，并且通过堆叠多个自注意力层（Multi-Head Attention），它能够有效地捕捉长序列的依赖关系，这对于处理长文本或较长的语音段落非常有益，提高了模型的语音合成质量和自然度。

目前的 TTS 主要有两种架构，一种是两阶段的生成方式，另一种是端到端的生成方式。对于两阶段的生成方式，第一阶段是利用语言模型或者声学模型将输入的文本序列转换为语言特征或声学特征等中间特征，第二阶段利用声码器将中间特征转换为原始音频波形。比较常用的二阶段模型如 Tacotron2^[19]和 WaveNet^[20]等。另外一种架构就是完全端到端的方式，这种方式将整个语音合成过程作为一个单一模型进行建模和训练，从输入的文本直接生成对应的语音输出，而无需将任务分解为多个子任务。典型的例子就是 FastSpeech2^[17]。

2.2.4 对话头像视频生成模块

对话人头像视频生成亦称为唇动序列生成，其核心目标是根据一个驱动源——无论是音频片段还是文本内容——来合成相匹配的唇动序列。这一过程不仅涉及到唇部运动的精确模拟，还必须综合考虑对话头像的多种面部特征，包括表情变化和头部动作，以实现更自然、更生动的视频效果。（主要的任务）

在对话头像视频生成技术的早期探索中，研究人员依赖于跨模态检索和基于隐马尔可夫模型（Hidden Markov Model，即 HMM）的方法来构建驱动源与唇动数据之间的动态关系。这些方法虽然初步实现了唇动的合成，但它们对模型部署的环境条件、视觉音素的标注精度等方面有较高的要求，限制了其广泛应用。

为了克服这些限制，研究人员提出了一种新颖的基于图像的唇动合成技术^[21]。该技术通过从预先收集的样本库中检索和选取最合适的唇形，以生成接近真实的唇部图像。尽管这种方法通过文本到语素的转换进行唇形检索，但它并未充分考虑到语言内容的上下文信息，可能影响合成视频的自然度和准确性。另外，研究人员提出了一种基于卷积神经网络的架构^[22]，专门设计用于处理音频和视频之间的同步问题。这种架构能够学习音频信号和视频中口型变化之间的关联，使得嘴部同步取得较好的效果。进一步地，通过引入关键姿势的插值和平滑处理模块，改进了基于跨模态检索的姿势序列合成。此外，他们还利用生成对抗网络^[23]（即 GAN）模型，增强了视频合成的质量和真实感。

目前可以根据通过一段音频生成讲话头部视频的维度，将方法分为两类：采用 2D 以及采用 3D 的合成人脸方法。

2D 方法生成讲话头部视频主要使用了人脸关键点（即 landmarks）、语义图（即 semantic maps）或其他类似图像的表示方法来解决问题。在 IPLAP^[24]这篇文章中，研究人员提出了一个两阶段框架来生成与音频同步的保持身份信息的说话面部视频，首先通过一个基于 Transformer^[18]的模型从音频中预测嘴唇和下巴的标志点，随后使用一个视频渲染模型将生成的标志点转换为面部图像。此过程中，通过静态参考图像提取的外观信息，以及目标面部的部分遮挡图像来辅助生成真实且保持身份特征的视觉内容。

另外，唇部合成还可以使用图像到图像的转换来生成，这是 2D 方法的一个扩展。例如 Wav2Lip^[3]中提出了一种非常简单的方法（如图 1-2 所示），直接使用一个训练好的同步网络专家对输入的一半遮挡的嘴部的重建内容进行同步约束，取得了非常好的同步效果。还有一种比较常见的 2D 方法是一种变形-融合手（即 Deformation-Inpainting）的方法，例如，DINet^[25]通过对齐源图像和参考图像的特征并与音频特征一起送入变形模块得到变形的特征，然后融合部分通过一个特征编码器根据变形特征对源图进行修复得到说话人的帧。

尽管 2D 方法可以比较方便地表示说话人，但是仍然存在很多问题，例如 Wav2Lip^[3]，虽然实现了较高的同步效果，但是只关注于唇部的动作，且其分辨率大小为 96*96，生成的结果很模糊不真实。又如 DINet 虽然在高分辨率下取得了较好的效果，但是对于侧脸或者头部具有一定角度的图片，嘴部的边界框（Bounding Box）会非常的明显生成的图片会丢失掉一部分的特征。根据文献^[4]，2D 方法存在这些问题的原因是神经网络是从耦合的 2D 运行场（Motion Field）中进行学习。利用 3D 的方法可以更好的解耦各个特征分量信息，从而达到更好的效果。

3D 的方法可以更好的解耦出各个信息分量。目前主流的 3D 方法主要有参数化的方法和神经渲染的方法，参数化的方法主要是基于 3DMM 的方法^[26]，这种方法是一种基于统计学的方法，它基于主成分分析（PCA）的思想，通过分析一组大量详细的三维面部扫描，从中学习面部形状和纹理的变化模式。这些扫描覆盖了不同性别、年龄、种族的人群，以及一系列的表情变化。主要建立两个模型，分别为形状模型和纹理模型。3DMM 将面部形状和纹理表示为一个平均面部形状或纹理和一系列主成分的线性组合。通过对 3DMM 系数进行渲染可以合成说话人人脸。另外，FLAME^[27]是一个基于 3DMM 的面部建模方法，它支持面部表情、头部姿势和下巴运动的建模。FLAME 模型将人脸建模成由头部、下巴、双眼这四个可以绕轴旋转的关节和 5023 个面部的顶点组成的三维结构。通过线性蒙皮（Linear Blend Skinning，即 LBS）算法，模型可计算出面部顶点发生位移后皮肤的位置，进而渲染出整个人脸。神经渲染的方法都是基于 NERF^[28]衍生而来，它是一种用于 3D 场景重建的深度学习方法，它通过使用多层感知机（MLP）网络，将 3D 空间坐标和观察方向作为输入，输出该点的颜色和体积密度。MLP 通过学习这些点的密度分布和辐射亮度，使得从任意视角渲染的图像能够准确反映场景的几何和光照信息。这种方法能够捕捉到场景中的细微变化，包括阴影、反射和透明度等复杂光效，从而实现高度逼真的 3D 场景渲染。

2.3 本章小结

本章主要介绍对话虚拟人系统的相关技术，首先引入在图像和视频生成领域比较常用的 GAN 网络，它是一种常用的编码器-解码器结构，通过生成器和鉴别器这种对抗的方式来不断提高生成器的“造假能力”，最后其生成的图像非常的逼真。接着介绍了虚拟人对话系统的相关技术，包括 ASR、TTS、LLM、Talking Head Generation 等，系统通过 ASR 模块可以将语音转为文本，然后输入到大模型中进行处理，处理返回的结果输入到 TTS 模块得到响应音频，音频再驱动人脸图像或者视频动起来，最后得到目标说话人多模态信息反馈给用户。

3 核心模块及改进方法

3.1 EAT 模型

本系统采用 EAT 模型^[6]作为本系统的核心模块。EAT 模型的输入是一个参考视频、一张原图和一个表情提示（Emotional Label），通过提取参考视频的音频以及 3D 潜在编码表示对输入的图片进行驱动，合成 256*256 的情感说话人视频。

3.1.1 EAT 模型提出背景

传统的音频驱动的虚拟人头像生成模型在情感表达上存在限制，通常不会直接在训练过程中整合情感数据，导致生成的模型虽能发声，但往往无法自然地传达人类讲话者的情感细微差别。EAT 在不大规模重训模型的基础上，实现了对说话人的表情的控制，极大的节约了资源，提升了模型生成的效率。

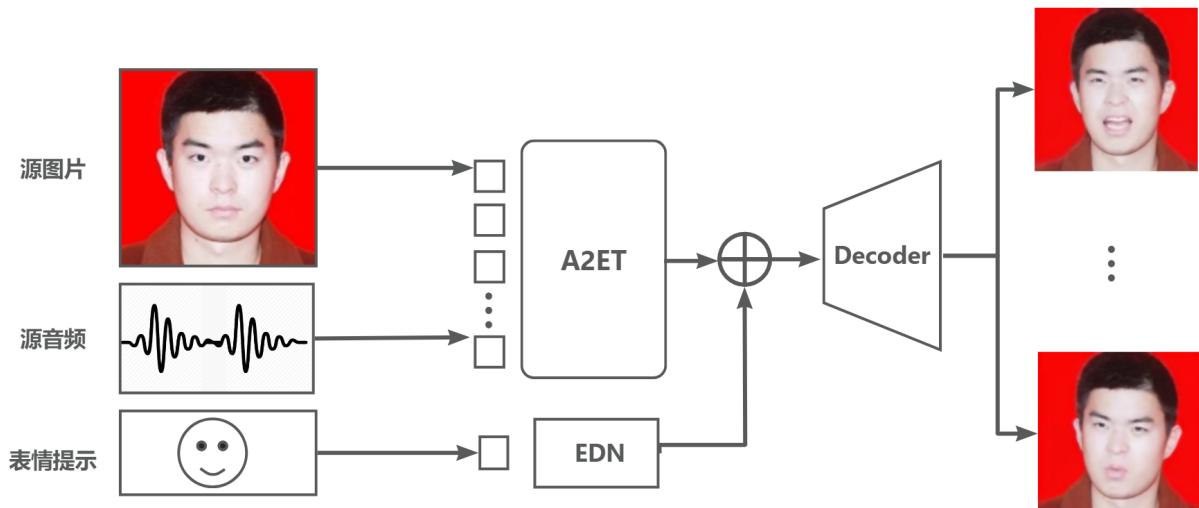


图 3.1: EAT 结构图概览

3.1.2 EAT 模块分析

EAT 模型主要是在预训练好的一个通用的说话头模型上加入了三个适应性的调整，如图 3.1 所示，EAT 包括深度情感提示网络（Deep Emotional Prompts）、情感变形网络（Emotional Deformation Network）以及情感适应模块（Emotional Adaptation Module）。模型包含了两个阶段，第一阶段主要训练一个音频到 3D 潜在关键点的预训练模型，第二个阶段对这个预训练好的模型的进行微调使其能表达情感。

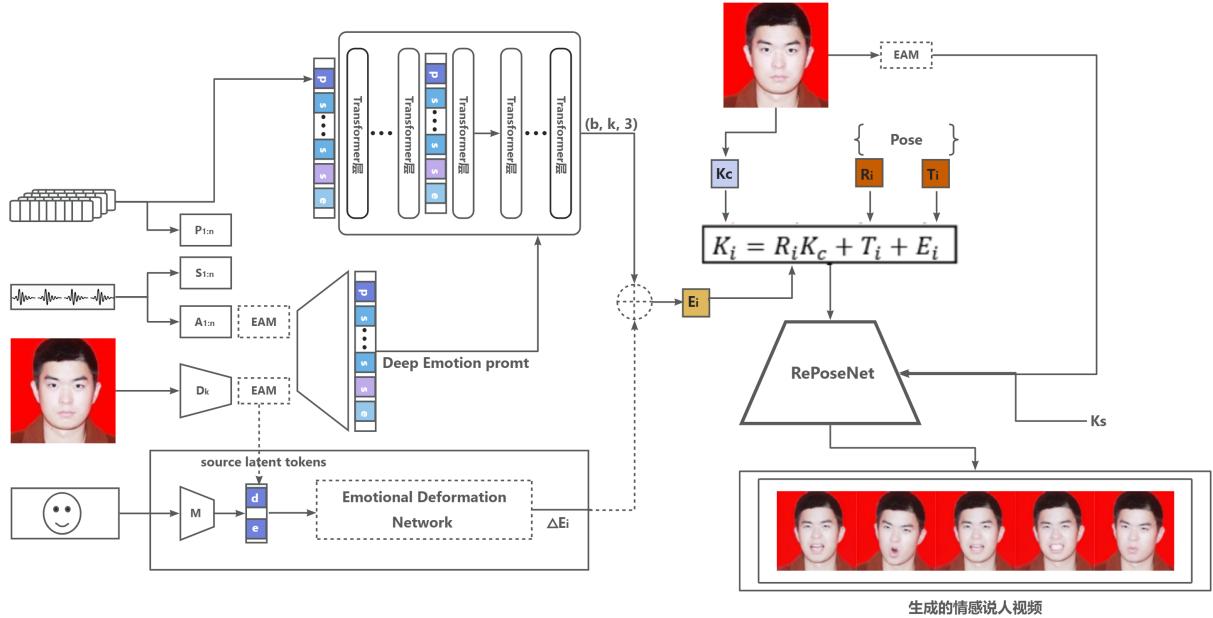


图 3.2: EAT 详细结构图

3.1.3 EAT 第一阶段

EAT 第一阶段主要完成两个目标，第一个目标是增强无监督的 3D 潜在关键点表示，使其能够更好的捕捉情感的表达；第二个目标是训练一个 A2ET (Audio to Expression Transformer) 模型，学会将输入的音频映射到关键点中控制面部表情的表情形变中。

3D 潜在关键点 它不仅包含二维平面上的位置信息，还包含深度信息，从而能够更加精确地表示面部的三维结构和运动，它可以将人物的身份特征和表情运动相关的信息解耦开来，从而实现更好的建模效果。3D 潜在表示 (3D Latent Representation) 一般由三个部分组成，分别由 3D 规范关键点 (3D Canonical Keypoints)、预估的头部姿态 P_e 和表情形变 E_{deform} 。头部姿态是由一个旋转矩阵 $R_s \in \mathbb{R}^{3 \times 3}$ 和一个平移向量 $T_s \in \mathbb{R}^3$ 表示。对于说话头视频中的每帧，可以无监督学习得到 3D 潜空间中的点 K_i ，可以表示成如下的式子：

$$K_i = R_i K_c + T_i + E_i \quad (3.1)$$

它由以下四个部分组成，表示个体独特面部特征的标准关键点 K_c 、相对于标准关键点进行旋转的旋转矩阵 R_x 、表示头部偏移的平移矩阵 T_i 以及表示表情变化引起关键点位置变化的表情变形因子 E_i 。

为了捕捉更多的表情细节，模型对人脸的潜在表示做了一系列改进，包括使用情感标签更多视频质量更高的 MEAD 数据集以及只计算脸部的 loss 从而减少背景的影响等。

音频到表情的转换器 A2ET (Audio to Expression Transformer) 这个模块是为了训练从音频 (Audio) 预测 3D 潜在关键点的表情形变 E_i 。音频信号与面部表情之间存在复杂的关联，并且往往涉及到从语音的细微变化中提取情感和语调的能力，Transformer 的自注意力机制能够分析整个音频序列，识别出对生成特定表情最关键的音频特征，如音调、声强等，从而可以更加精确地映射这些音频特征到相应的潜在关键点。具体的解释如下：

- 1) **数据集选择。** 根据 EAT 论文，由于 Transformer 模型具有大量的参数同时为了使模型更好地学习数据中的模式和特征，所以训练 Transformer 模型通常需要较大的数据。作者选择在大型数据集 Voxceleb2 上训练 A2ET 模型。Voxceleb2 是一个包含

大量来自各种语音背景的说话者的语音数据集，可以提供丰富的数据样本，有助于模型更好地学习和泛化。

- 2) **模型训练输入。**作者为了保证更好的利用音频信息，同时使用了音频的语义特征和声学特征。音频语义特征提取的事语音信号中的高级语义信息，例如说话者的语调、情感、语义内容等；声学特征是从原始语音信号中提取的低级特征，描述了语音信号的声学属性，如频谱、声音强度、语速等。对于第 i 帧，作者从连续的 $2w+1$ 个音频帧（左边右边各扩展 w 帧）中提取语义上下文。首先，作者将语音特征 $S_{i-w:i+w}$ 和头部姿势特征 $P_{i-w:i+w}$ 转换成相应的语音 token。接着，第 i 帧的 6DoF（自由度）被编码为姿势 token p 。这些 tokens 作为输入传递给 A2ET 的编码器。为了捕捉嘴部微动，我们使用音频编码器和关键点检测器 D_k 来编码声学特征 $A_{i-w:i+w}$ 和潜在源图像特征。这些特征被融合以获得声学 tokens，A2ET 解码器利用这些 token 来输出 $2w+1$ 个 tokens。
- 3) **优化策略。**表情形变通过每帧图片预测得到，如果直接优化这个表情形变会导致优化困难。经过探索，作者发现自监督学习得到的三维关键点之间存在内在的相互依赖性，并且只有少数关键点对面部表情有重要影响。于是，为了降维和除去冗余信息，作者对得到的表情形变 E_i 进行了主成分分析（PCA 检测），使得 E_i 的维度从 45 为降到了 32 维。这样，模型就方便优化了。根据论文补充材料，具体的公式如下：

$$E_i = PE_i * U^T + M \quad (3.2)$$

U 是训练集的主特征值矩阵， M 是均值向量， PE_i 是预测得到的 PCA 模型， E_i 是表情形变。将原始数据减去平均向量 M 并乘以主成分特征向量矩阵 U 的转置，可以将数据投影到主成分空间中，并实现降维和重建的目的。最终得到的 E_i 的维度是 (15,3)

3.1.4 EAT 第二阶段

第二阶段的主要任务是对一个已经训练好的 A2ET 模型进行情感微调，为实现此目的，作者提出了高效情感自适应模块，分别包括三个网络，深度情感网络（Deep Emotional Prompts）、情感变形网络（Emotional Deformation Network，EDN）、情感自适应网络（Emotional Adaptation Module，EAM）。下面分别对几个网络进行阐述：

深度情感网络 深度情感网络主要是为了将我们输入情感标签映射为提示字符（Prompt Tokens）送入到 Transformer 中从而实现对感情的注入。作者将注入网络第一层的字符称为浅层字符（Shallow Prompt Tokens），注入 Transformer 第一层以后的每层的提示字符称为深度提示字符（Deep Prompt Tokens），经过发现，注入深层字符可以实现更好情感表达，但是同样会带来不同步的副作用。

情感变形网络 前面我们介绍过，对于 3D 潜在关键点，我们可以通过 A2ET 模型预测音频和图片得到一个人脸关键信息表情形变 E_i ，现在我们需要控制这个表情形变来表现我们所需要的情感表达。并且，经过实验探究，表情形变是具有线性可加性的。于是作者提出使用为表情形变增加一个残差项用于表示目标情感的表情形变。计算公式见公式 3.3

$$E'_i = E_i + \Delta E_i \quad (3.3)$$

通过设计一个情感形变网络，我们可以从输入的表情提示和预测 3D 潜在表示中预测出表情形变 E'_i

情感自适应网络 这个模块主要是为增强视觉质量而设计。模块主要对输入的原图特征和情感提示的嵌入表示进行映射，得到一个表情条件下的特征。具体做法是，首先对输入的情感提示的嵌入表示进行处理，得到原图特征通道数的一组权重系数 γ 和对应偏移值 β 。随后对输入的原图特征和调整后的系数在通道维度上计算得到图片在表情条件下的特征。具体的计算公式见公式 3.4 和公式 3.5

$$\gamma, \beta = \tanh(\text{FC}(\text{ReLU}(\text{FC}(e)))) \quad (3.4)$$

$$EAM(x) = F_s(1 + \gamma, x) + \beta \quad (3.5)$$

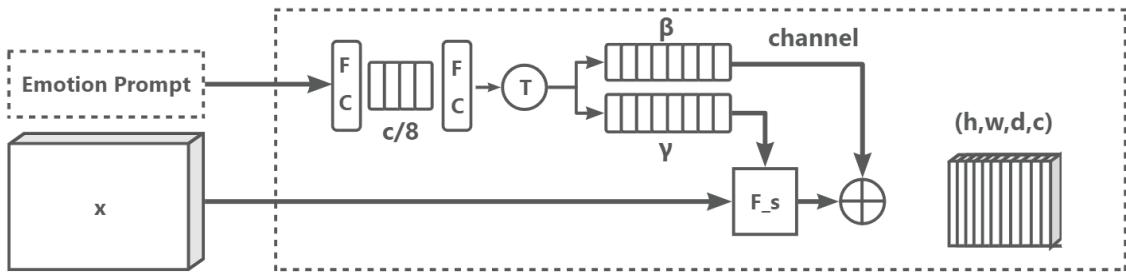


图 3.3: EAM 模块流程图

此模块具有很高的灵活性，可以在 EAT 模块的很多地方使用。此网络的主要的流程图如图 3.3 所示

重定位网络 重定位网络（即 RePosition Network）主要为了实现对输入的原图和输入的原始图像的潜在关键点和驱动关键点进行变换得到最终的输出图像，通俗来说就是可以将一个人的面部表情迁移到另一个人的面部。主要分为了两个步骤，第一个步骤是对输入的原始图像编码成我们上一步提到的图片在表情条件下的特征，第二个步骤是对输入的 3D 潜在关键点和驱动关键点进行计算，得到一个 3D 变形矩阵，通过这个矩阵可以对图片在表情条件下的特征进行变换得到输出的图片。主要的流程如图 3.4 所示。

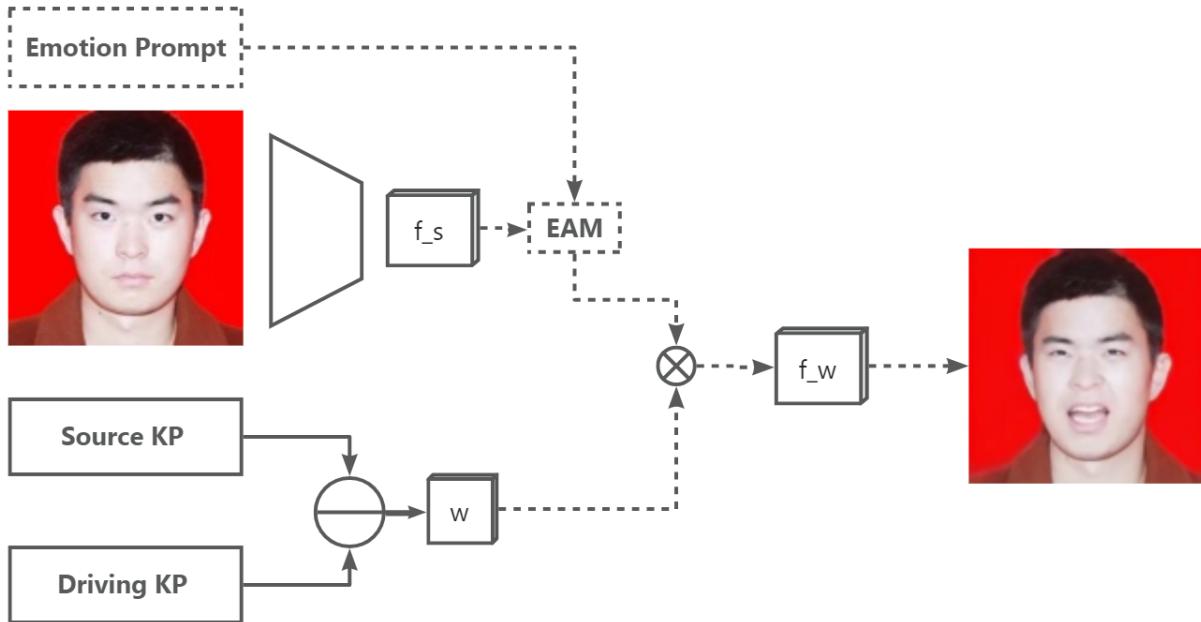


图 3.4: ReposeNet 流程图

3.1.5 EAT 模型改进

系统采取了原版 EAT 模型的架构，使用原版 EAT 模型的权重信息。但是对 EAT 的推理方式进行了修改。官方仓库的 EAT 模型推理阶段使用的是已经处理好的推理数据，提供了相应的处理脚本。具体而言，推理时，模型需要输入一张图片，一段驱动视频，驱动的头部动作、驱动音频、音频特征以及对人脸裁剪进行仿射变换时的一个人脸模版。本系统的需求是对指定的任意图片和任意的音频进行驱动，显然原版本的推理方式不符合系统的需求，于是我对推理代码进行了重构。主要的个性化的改进是以下几点。

自定输入驱动潜在表示和自定义头部动作 原版的 EAT 推理，需要输入原始图像的 3D 潜在表示和驱动视频的 3D 潜在表示作为驱动的模版。在本系统中，由于客户端并没有设置获取指定说话人的视频信息和 pose 信息。前面我们提到，3D 潜在表示主要是由多个部分组成的，我们在推理时，无需驱动视频的 3D 潜在关键点信息，因为 3D 潜在关键点信息由我们的原图像提供，只需要驱动视频的预估的头部姿态和表情形变。于是我选取了一个表情中性的驱动视频模版（减少驱动视频的表情干扰），通过预处理脚本得到这个驱动视频的潜在表示，然后再推理时将其头部姿态和表情形变送入重定位网络中推理得到目标的情感说话人视频。

为了增加个性化，我们将视频的头部动作参数拿出来作为一个可以送进模型推理的参数信息。通过对不同头部动作的驱动视频的提取，得到不同的头部动作参数，将这个头部动作参数作为客户端用户可以指定的参数，实现了 pose 的个性化定制效果。

自定义输入语音 在模型推理的过程中，使用的音频特征是用过 deepspeech 特征来表示的。但是在原版仓库的推理中，输入的音频和对应的特征是从已经处理好的驱动视频数据中直接获取的。如果系统想要实现对输入音频的自定义，需要手动将特征提取模块集成到系统。于是，我对原始提供的 deepspeech 特征提取模块进行了修改，将其作为一个函数引入。调用时通过对指定音频文件路径传入得到一个 deepspeech 的特征。同时原版的特征提取模块非常慢，我对其进行了相关的加速处理。

但是这样直接除了会带来两个问题。第一个 deepspeech 是基于 tensorflow 的，tensorflow 的 gpu 启动方式默认会占满显存，如果不及时释放会导致服务器的显存爆炸，系统

处理异常，于是我加上了一个针对 tensorflow 的显存动态申请模块，可以实现对显存的及时释放。当 TensorFlow 程序开始运行时，GPU 内存不会一次性全部分配，而是根据需要动态分配。这样可以在程序运行过程中更灵活地管理 GPU 内存，避免因为一开始分配过多内存而导致内存浪费或者与其他进程竞争资源。第二个问题是对于 GPT-SoVits 这个 TTS 模块返回的语音，其采样率是 32KHZ。但是为了实现音频与视频的同步，输入音频的采样率必须在 16KHZ。针对此问题我增加了音频重采样的处理，使得送入特征提取函数的音频符合预期输入。

预处理模块改进 为了使得推理网络的效率更高，我将构建模型操作和权重载入操作拿到了系统主类的初始化模块，同时将 EAT 模块集成成为一个功能类，这样这个类将可以如 API 一样在系统主类中进行调用。具体的改进有三点。第一点，针对原版 EAT 对图片进行预处理的操作和提取原图片的潜在表示合并在一起处理，减少中间文件的产生。第二点，对数据准备阶段的函数进行修改，使其能对长度不足的头部动作序列进行扩充以满足模型的需求。同时，由于原版 EAT 生成的图的背景会扭曲（作者说这是因为潜在关键点在全图分布造成的），所以我加入了一个 MODNet 的一个推理模块，这个模块可以将图片背景去掉，于是将背景转为纯色的背景解决了背景扭曲的问题。

3.2 GPT-SoVits

本系统的 TTS 部分选择了一个非常强大的的语音项目 GPT-SoVits 作为我们的基本 TTS 模块，这个 TTS 模型可以完成音色转换的任务，即可以给定一个目标说话人的短时间说话音频，可以训练出一个微调模型。然后指定一个文本输入和一段参考音频，模型能够完美克隆出目标说话人的音色。

3.2.1 优点及特性

GPT-SoVits 是一个神经编解码器语言模型，它使用变分推理框架进行训练，使用编码器解码器模型将文本编码成潜在表示，其声码器基于 GAN，合成的语音具有很高的质量。其主要的优点可以概括为以下几点：

实现零样本 TTS 模型可以使用很少量的音频数据，就可以合成任何的语音。这对于训练带来了极大的便利，可以通过少量的数据，实现模型的快速微调训练。并且对于基模型的微调不需要大量的计算资源即可实现比较好的效果。

音色克隆 基于少量数据进行模型微调后，推理时只需要输一条不超过 10s 的参考音频和对应的文本内容就可以完美克隆目标音色。音色与目标音色极其的相似，几乎实现以假乱真的地步。

情感丰富 以往的语音合成技术往往只基于 Vits 或者 FastSpeech，它们的模型大多是在录音室录制的，往往不自然，这样就导致这种模型推理的结果具有一个明显的特点，就是机械音十分的严重，可以明显听出是合成的。但是 GPT-SoVits 可以实现非常自然的语音合成效果，推理的速度也十分快速，可以应用于实时的对话系统。

3.2.2 训练和推理

为了个性化定制音色，我对 GPT-SoVits 进行了少样本的训练。我使用自己录制的两条中文音频和一条英文音频分别训练得到了两个指定任人物音色的 TTS 模型。同时，由于原模型的推理是集成在 webui 中，为了使这个模块可以更方便的引入到本系统中，我将推理模块的代码进行修改，使其更方便引入我们的系统中。

训练过程 训练过程比较简单，在查阅相关的资料之后，大概掌握了训练的流程。训练总共分为 3 个流程：收集数据、数据处理、开始训练。收集数据，对于中文数据集的准备，我使用了我自己手机录制 3 分 30 秒的录音；对于英文数据的准备，我使用了 1 分 30 秒的自己录制的英文音频。数据集处理阶段，首先使用降噪工具对训练数据进行人声与环境音分离，这样可以减少环境音的干扰；随后将训练数据集切分为小段的音频，并利用降噪工具对人声中可能的设备噪音进行降噪；接着使用音频标注工具以及 ASR 工具，对音频和对应的识别文本进行标注，对于音频中的停顿的地方以及不对应的音频段进行删除；然后使用音频格式化工具将音频数据集进行格式化，得到最终模型可以接受的输入方式。具体流程如下图：

3.3 ChatGLM2-6B

本系统的大模型部分采取的是清华大学提出的一个基于预训练的 GLM 模型的开源项目 ChatGLM2-6B。它在 MMLU、CEval、GSM8K、BBH 等数据集上的性能取得了大幅度的提升，在同尺寸 6B（60 亿参数量）开源模型中非常亮眼。

3.3.1 优点及特性

ChatGLM2-6B 是在第一代双语对话模型 ChatGLM-6B 的基础上开发的。在保留初代模型对话流畅、部署门槛较低等众多优秀特性的基础之上，ChatGLM2-6B 引入很多新的特性。

性能提升 ChatGLM2-6B 采用了 GLM 的混合目标函数，在 1.4T 中英标识符的预训练与人类偏好对齐训练下，性能得到大幅提升。相较于初代模型，MMLU 提升 23%，CEval 提升 33%，GSM8K 提升 571%，BBH 提升 60%，在同尺寸开源模型中具备更强竞争力，并且利用 FlashAttention 技术，模型的上下文长度从 2K 扩展到 32K，对话阶段使用 8K 上下文长度进行训练，支持更多轮次的对话，增强了模型对长篇对话或复杂对话场景的理解和表达能力。

推理效率提升 基于 Multi-Query Attention 技术，ChatGLM2-6B 具有更高效的推理速度和更低的显存占用。在官方的实现中，推理速度提升 42%，用户可以更快的获取对话的结果，拥有更加流畅的对话体验。同时在 INT4 模型量化下，6G 显存支持的对话长度从 1K 提升到了 8K，这使得量化的模型可以支持处理更长的对话历史记录，在生成对话时获得更加全面和深入的上下文信息，为硬件受限的长对话和复杂对话下的对话场景提供了一个合适解决方案，在保证精度的同时，极大缓解了模型部署的硬件的压力。

3.3.2 部署方式

目前，官方仓库提供了很多种部署方式，本系统主要采用的是本地服务器部署的方式。系统使用到了 uvicorn 服务器，它是基于 uvloop 和 http tools 构建的非常快速的 ASGI 服务器。uvloop 用于替换标准库 asyncio 中的事件循环，使用 Cython 实现，它非常快，可以使 asyncio 的速度提高 2-4 倍。我们可以采用 http 请求的方法，通过将模型推理转为 FastAPI 的方式，可以向服务器指定端口发送请求，得到模型推理的结果。

3.4 FunASR

FunASR 是由阿里巴巴通义实验室语音团队开源的一款语音识别基础框架，集成了语音端点检测、语音识别、标点断句等领域的工业级别模型。在这个工具包中，整合了很多十分高效易用的模型，我们采用了的是达摩院提出的 Paraformer 模型^[?]。

3.4.1 优点以及特性

Paraformer 模型是一个优秀的多语言 ASR 模型，这个模型采用了一种高效的非自回归端到端语音识别框架，在很大的程度上提升了非自回归（NAR）模式在 ASR 中的性能。下面介绍一下 FunASR 的有点以及特性：

强大的长序列处理 Paraformer 模型通过引入全局注意力机制和自注意力机制，能够更好地处理长序列数据，如长篇文本、长时间语音等。传统的 Transformer 模型在处理长序列时存在着性能下降和计算复杂度增加的问题，而 Paraformer 通过改进的机制有效地解决了这些问题。全局注意力机制允许模型在整个序列范围内进行全局扫描，从而更好地捕捉序列中远距离的依赖关系和语义信息。此项特性对于需要处理长距离依赖关系的任务非常重要，例如对话系统、长篇文本处理等。

跨领域适应性 Paraformer 模型是一种通用的语言建模模型，适用于多种语言和领域的应用场景。它可以很好地适应不同领域的数据特点和语言特点，具有一定的通用性和灵活性。这种跨领域适应性使得 Paraformer 模型可以广泛应用于自然语言处理领域的各种任务，包括语音识别、文本生成、机器翻译等，为不同领域的应用提供了一种统一的模型框架。

易于部署 并且集成后的模型可以非常方便的在本地部署，并且有中文版本和英文版本，根据测试都具有较高的准确度。并且这个模型可以比较方便的在本地进行相关的部署，在使用时，直接调用 api 接口即可实现快速语音识别功能。并且模型具有非常高效的推理能力，可以用于实时对话领域。

3.4.2 模型调用

直接使用集成好的模型库直接调用可以大大降低系统的复杂度。鉴于目前 ASR 技术已经非常的成熟，并且有非常易用的模型，所以系统的 ASR 模块直接采用的是达摩院在 modelscope 上发布的中文语言识别“Paraformer 语音识别-中文-通用-8k-离线”模型和英文语音识别“Paraformer 语音识别-英文-通用-16k-离线-长音频版”模型。

推理时，模型允许两种推理方式，可以采用基于 ModelScope 进行推理或者给予 FunASR 的推理。在进行测试时，系统采用基于 ModelScope 的方式推理时会报错，并且未找到合适解决方案，所以系统采用的是基于 FunASR 的推理方式。基于 FunASR 的推理有很多种功能，可以实现语音识别、语音端点检测、标点恢复以及时间戳的预测。

本系统在集成 ASR 模块时，只利用了非实时的语音识别功能，并将推理模块集成为一个主算法类的方法。方法的输入是音频路径，输出是语音识别的文本字符串结果。音频路径是客户端传递给服务器的音频数据的存储路径。

3.5 本章小结

本章主要介绍本系统用到的核心算法特性和相应的部署方式及改进策略。

4 高质量情感对话虚拟人系统设计与实现

4.1 系统设计

本系统主体使用深度学习中的 Gradio 框架，本系统可划分为一种简单的基于 B/S 架构的系统。系统主要实现根据用户传入的音频和图片，用户指定头部姿态和表情标签，

最终可以生成不同风格的情感虚拟人。主要利用 python 中的一系列的库对系统进行实现。

4.1.1 系统的整体架构

系统采用依托于 Gradio 的简单的 B/S 架构，即浏览器/服务器架构，是一种比较常见的网络架构模式，用于构建应用程序和系统。Gradio 大致可以分为界面层、服务层和集成层；界面层提供了丰富的交互式组件接口，使得用户可以方便的通过这写组件与模型进行交互；服务层主要依托 Python Web 或者 Flask 等后端框架，处理用户的输入，执行模型推断，并将结果返回到前端界面；集成层主要集成各种深度学习库（PyTorch、Tensorflow 等），允许用户直接在这些库的基础上构建交互式界面。在这种架构中，系统可以被分为了两个部分：浏览器和服务器。浏览器主要使用目前深度学习部署中比较流行的 Gradio 框架的界面，服务端集成相关算法模块的推理阶段，用于处理用户请求。系统架构图如 4.1 所示。

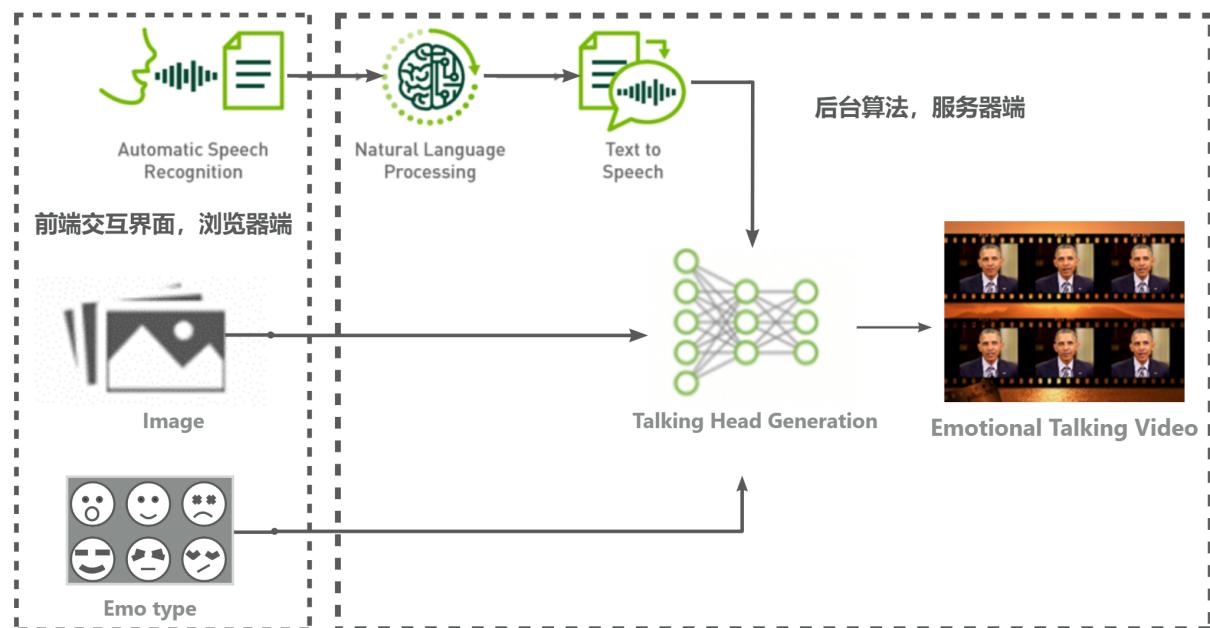


图 4.1: 系统架构图

浏览器 浏览器主要是与用户进行交互的界面，Gradio 框架提供了一个简单的 Python API，使我们能够在本地创建交互式应用程序。我们可以使用 Gradio API 来定义模型的输入和输出，并配置界面的布局和样式。可以通过将数据或者事件与相应的控件进行绑定，从而达到与服务器进行数据传输和与用户进行交互的目的。

服务端 服务端主要用于接收浏览器的请求，并对请求进行处理。使用的也是基于 Gradio 的框架，Gradio 的服务器最后将结果返回给浏览器。后台服务主要的接口是对用户的图片数据、音频数据、表情标签以及头部姿态进行接收并载入到内存中，服务端集成的推理算法类初始化后，可以使用接收的数据进行推理，最终生成高质量的情感对话虚拟人视频，后台会将情感对话人的视频传递到前台界面进行显示。

具体数据处理流程 描述一下整个系统的数据交互过程，系统没有设置特殊的缓存中间件，对中间的结果，直接保存在服务器的指定文件夹中，方便进行调试，以下是数据交换过程：

1. 用户通过浏览器访问本地服务器上的 Gradio 应用。
2. 用户在浏览器中输入数据、调整参数等操作，然后通过绑定的控件触发事件，最后将这些数据发送到本地服务器。
3. 本地服务器接收来自浏览器的请求，并进行处理。
4. 服务器执行与远程服务器相同的操作，包括数据预处理、模型推断等。
5. 处理完成后，本地服务器将结果返回给客户端，用户在浏览器中可以看到结果。

4.1.2 网络模型的部署

在基于大模型的高质量情感对话虚拟人系统中，需要对四个模型和一个接口进行部署和请求，这个四个模型分别是 GPT-SOVITS 模型、ChatGLM2-6B 模型、GPFGAN 模型和 EAT 模型。4.2展示了模型初始化的过程。

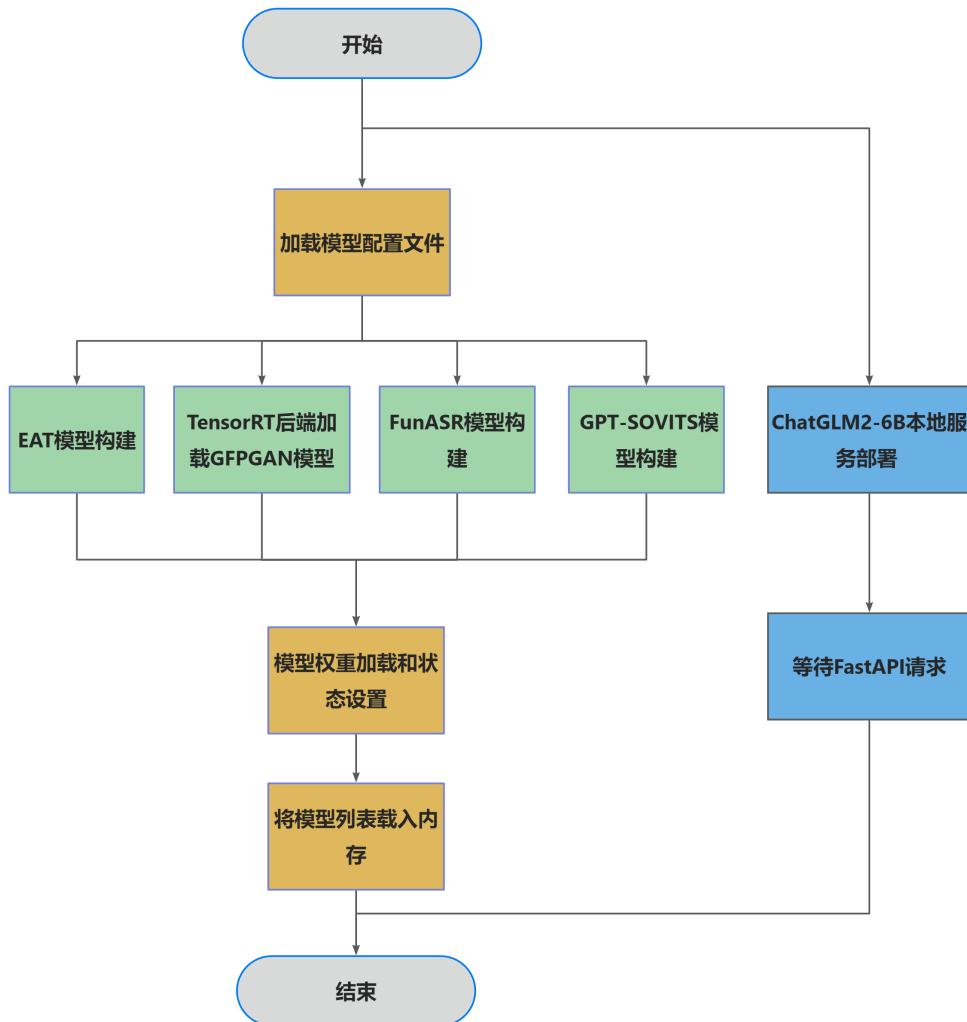


图 4.2: 网络模型初始化

初始化时首先初始化四个模型并将 ChatGLM2-6B 的后端部署到服务器上，指定 8000 端口，后续需要用到大模型时，使用 FastAPI 请求的方式对端口发起请求可以获取目标输出。

4.1.3 系统界面的展示

在服务器后台运行编写好的 python 脚本，脚本调用 `launch()` 函数来启动 Gradio 的交互式界面。`launch()` 函数会在本地启动一个 Web 服务器，它会监听一个可用的端口如 7890。一旦 Web 服务器启动，Gradio 会自动打开默认的 Web 浏览器，并将应用程序的界面呈现在浏览器中。

4.2 系统功能实现

本章节基于前一章提到的核心算法的应用进行了交互式的开发，基于 Gradio 框架实现了基于大模型的高质量情感对话虚拟人系统开发。本系统主要实现从用户输入音频、图片，选择情感配置和头部动作，输出一个高质量的情感说话头视频。

4.2.1 前端界面设计

Gradio 是一个 Python 库，旨在帮助开发者快速构建和部署机器学习模型的图形用户界面（GUI）。它提供了一个简单的接口，允许用户定义输入和输出的界面元素，然后自动生成一个交互式的界面，以便用户可以通过浏览器或本地运行的方式与模型进行交互。

4.2.2 Gradio 组件介绍

这个地方需要一个表格，`latex` 转 `word` 之后再画。

4.3 功能模块的实现

主要是有八个功能模块，交互模块、驱动动作模块、表情选择模块、ASR 模块、TTS 模块、抠图模块、对话大模型模块、超分模块。本节将专门介绍这个几个模块的实现方式，从整个模块的设计和源码实现方面进行阐述。

4.3.1 驱动动作模块

驱动动作模块的逻辑比较简单，主要是前端的 gradio 的单选组件进行控制。对用户选择 `pose` 进行解析。在编写的 gradio 脚本中，可以将单选组件的值作为参数传递给主类的成员方法，在接收到变量后，我们对变量进行解析，从而选择不同的驱动动作载入到我们的模型中，实现对驱动动作的控制。如图 4.3 即为我们的驱动动作模块流程图。

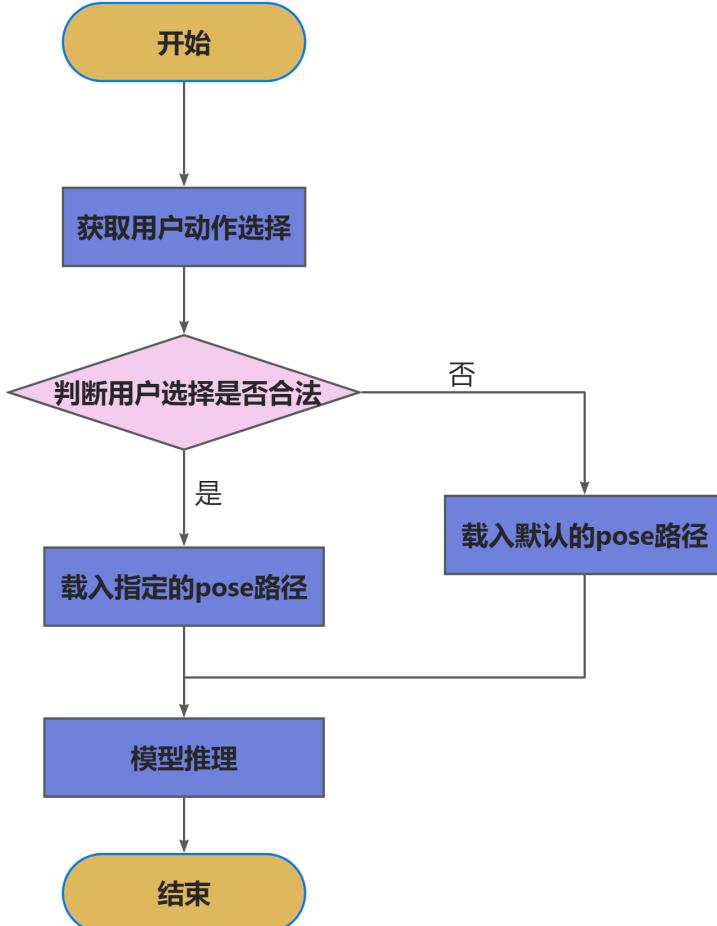


图 4.3: 动作处理流程图

4.3.2 表情模块

表情模块与动作模块类似，都是比较简单的逻辑，通过单选按钮控制参数的输入，然后后台有专门的函数对这个输入进行转义，转换侧好难过模型可以接受的输入方式。如图 4.4 我们的表情功能模系统流程图。

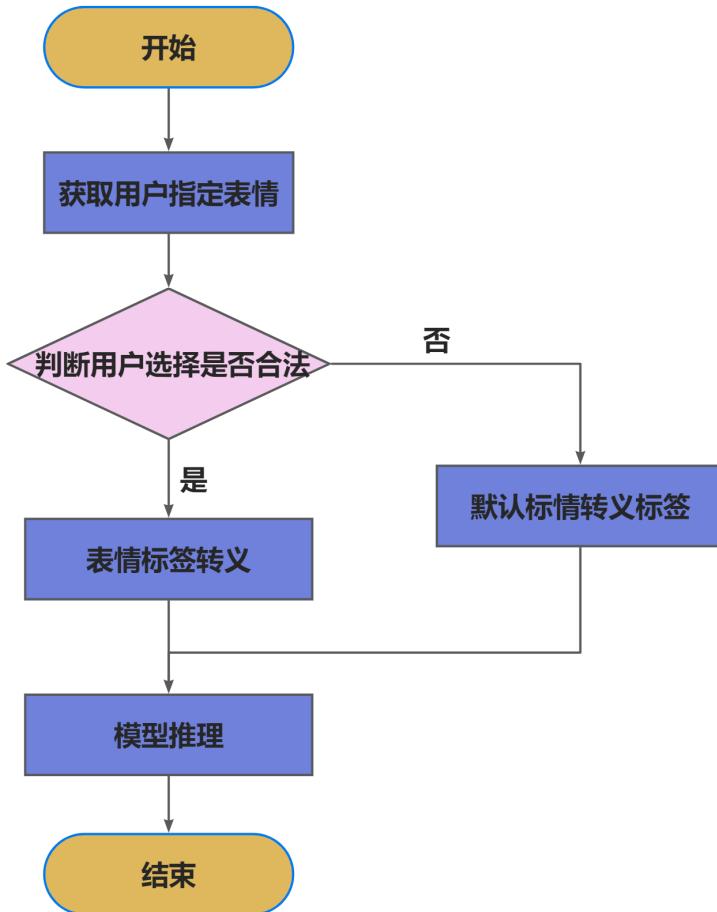


图 4.4: 表情模块处理流程图

4.3.3 对话大模型模块

对话大模型主要是 uvicorn 本地部署，uvicorn 是一个基于 asyncio 开发的一个轻量级高效的 web 服务器框架，为 Python 异步框架提供一个快速、轻量级的服务器，其还支持 HTTP/1.1、HTTP/2 和 WebSockets 等协议。本模块定义了一个异步的处理函数，这个允许模型在获取请求数据或者推理时不会阻塞主线程，提高了程序的响应速度和并发能力，具体流程图如图 4.5

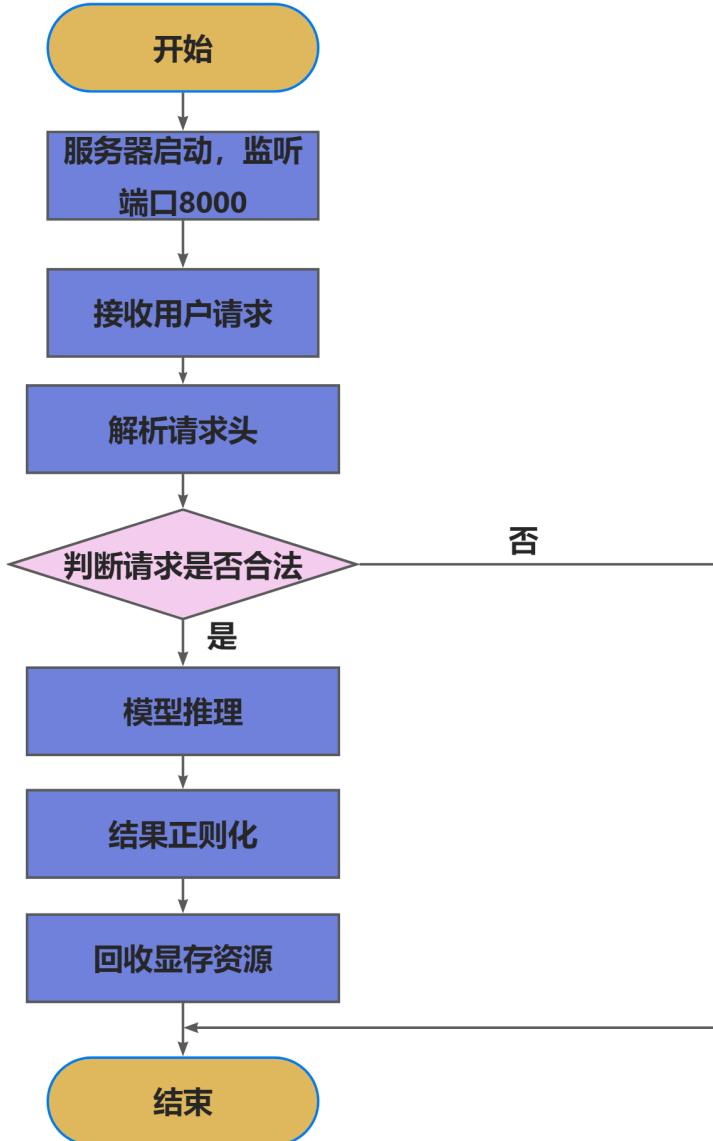


图 4.5: 大模型模块流程图

4.3.4 ASR 模块

ASR 模块直接将对应的推理模块拿过来了，然后使用函数进行包装，作为主类的一个成员方法，可以接收来自 gradio 的音频数据。音频数据客户端传过来的事字节类型的数据，我们使用一个函数模块对其进行解码保存，然后将保存的路径返回给推理模块进行推理。主要流程图如图 4.6。

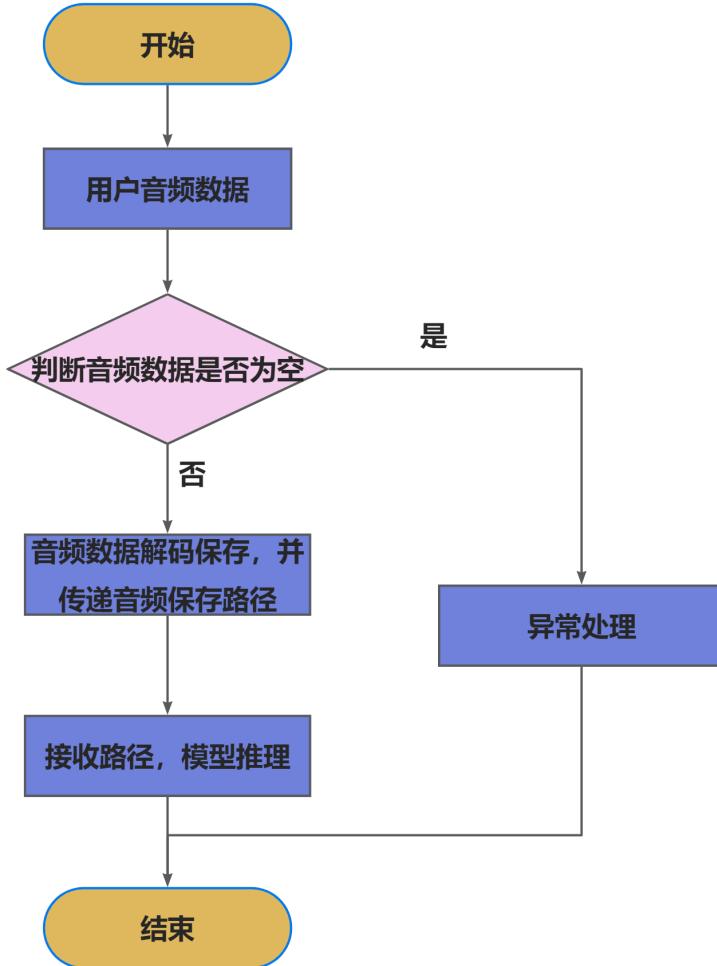


图 4.6: ASR 模块

4.3.5 TTS 模块

TTS 模块是依托于 GPTSoVits 的推理模块，将其包装为一个类进行使用。

4.3.6 抠图模块

4.3.7 对话人视频生成

依托于 EAT 模型的推理模块，推理接受的输入主要是来自 gradio 以及其前面模块处理的结果。从 gradio 直接拿到图片数据，这个图片数据，gradio 可以设置默认传递方式为 numpy 类型，这样这个数据就可以直接传递给推理模块进行处理，同时推理模块还需要表情模块传递的转移情感标签信息、驱动动作模块传递的驱动动作以及 TTS 返回的音频路径。具体的流程图如图 4.7 所示。

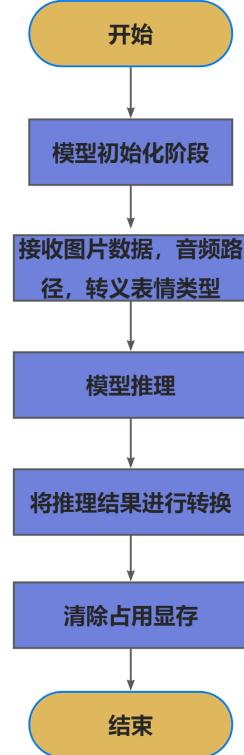


图 4.7: EAT 模块

为了更加详细的了解这个数据处理过程，我绘制了一张数据流处理图。如图 4.8 所示，分别对输入的音频，图片以及图片进行处理得到我们想要的中间表示。

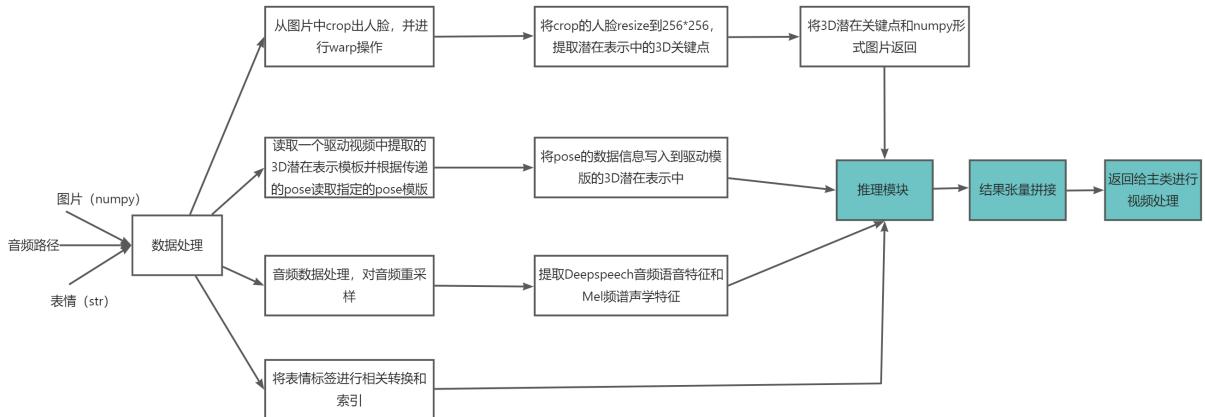


图 4.8: EAT 数据处理流图

4.3.8 抠图模块

抠图模块的引入主要是为了解决 EAT 模型输出结果背景抖动的问题，利用抠图将人像抠出，然后替换上纯色的背景，就可以实现我们的模块的基本功能。抠图模块主要是参考 MODNet^[29]这篇文章提出的方法，使用的是 onnx 加速版本的模型权重。基于原推理模块进行了相关的改进，原推理模块是得到 matte 图片，即一张黑白的图片，人像部分的颜色为白色，十六进制值为 (255,255,255)；背景部分为黑色，十六进制值为 (0,0,0)。

$$output = foreground * mask + background * (1 - mask) \quad (4.1)$$

根据公式 4.1，我们通过前景图，也就是我们的需要进行融合的原图去乘拓展了透明度通道的 matte 图片再加上我们的不同颜色的背景图乘对应的不透明度值就可以得到指定背景的图片。主要的流程图如图 4.9

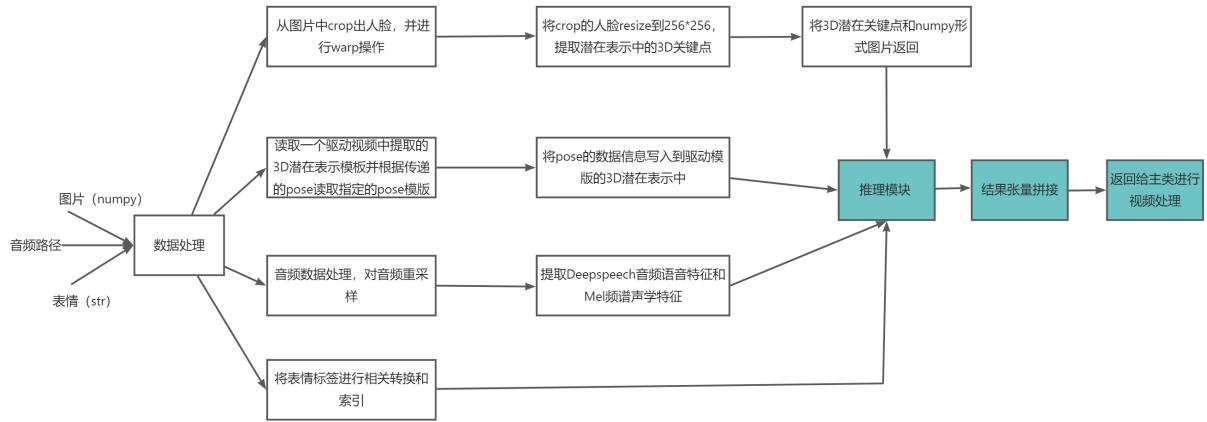


图 4.9: 抠图流程

4.3.9 超分模块

超分模块主要是流程与上面的几个模块相差不大，也是模型初始化之后进行相关的数据处理操作，主要是要将数据转为 GFGAN 模型的指定输入格式，需要对图片进行缩放到 512x512 的大小。具体流程参考 4.10。



图 4.10: 超分流程

4.4 本章小结

本章主要介绍系统的基本结构和相应功能模块的实现。主要包括交互模块、驱动动作模块、表情控制模块、对话大模型模块、ASR 模块、TTS 模块、抠图模块以及对话人

视频生成模块和超分模块。每个模块都对其流程作了详细的阐述。

5 系统测试

本系统主要完成根据用户的输入音频和相关自定义配置完成对高质量情感对话虚拟人视频的生成。由于本系统涵盖了很多的功能模块，包括交互模块、驱动动作模块、表情选择模块、ASR 模块、TTS 模块、抠图模块、对话大模型模块、超分模块。本章主要是对系统的功能模块进行测试。

5.1 功能测试

系统的主要功能是用户可以选择不同的驱动 pose 和表情配置标签来通过指定图片和音频生成带有情感的说话人视频。测试方法是选择不同的对话虚拟人的参数配置，根据结果观察系统是否正常完成了功能。

5.1.1 交互模块测试

交互模块主要是用户输入数据的采集模块。例如允许用户录制音频或者上传音频，允许用户使用摄像头捕捉图片或者自定义上传图片等操作。下面对这些功能进行测试。

音频上传和录制模块 首先进入浏览器界面，如图 5.1a 所示，点击向上的箭头图标可以上传用户本地音频文件，用户还可以选择使用本地麦克风设备进行音频录制。如图 5.1b 点击麦克风图标可以选择设备开启录制。点击“record”按钮开启录制，点击“stop”按钮停止录制，如果用户需要对音频进行裁剪可以点击“剪刀”图标对其进行裁剪，裁剪后按下“trim”确认裁剪，如果想取消操作，点击“cancel”即可取消裁剪操作。如 5.2b 为使用户能知道自己上传的音频时候正确，用户可以点击按钮调用服务器的 ASR 接口进行音频转文字操作，从而实现对结果的检阅。



图 5.1: 音频模块界面和音频设备选择测试



图 5.2: 音频裁剪与音频检阅测试

图片上传和摄像头捕捉 同样进入浏览器界面，如图 5.3a 所示点击上传箭头可以对本地图片进行选择；如图 5.3b 点击摄像头按钮可以使用摄像头对图片进行捕捉。具体而言，上传图片时，点击图片框内或者框下方的向上箭头上传图片；使用摄像头捕捉画面时，点击框下面的圆圈唤醒摄像头，然后再点击“相机”按钮进行捕捉即可得到照片。



图 5.3: 图片上传方式选择测试

5.1.2 驱动动作测试

进入界面操作如图 5.4, 上传音频和图片后, 选择相同的表情和不同的驱动动作进行测试。测试时, 对同一段音频, 同一个人物的不同的 pose 图片切帧组合成图像进行观察。测试基于控制变量的原理。由于 TTS 的生成每次具有不确定性的时长和情感, 为了保证嘴型一致比较头部的动作, 我们直接在服务端送入一段音频进 EAT 模型进行测试, 确保只有驱动动作是变化的。

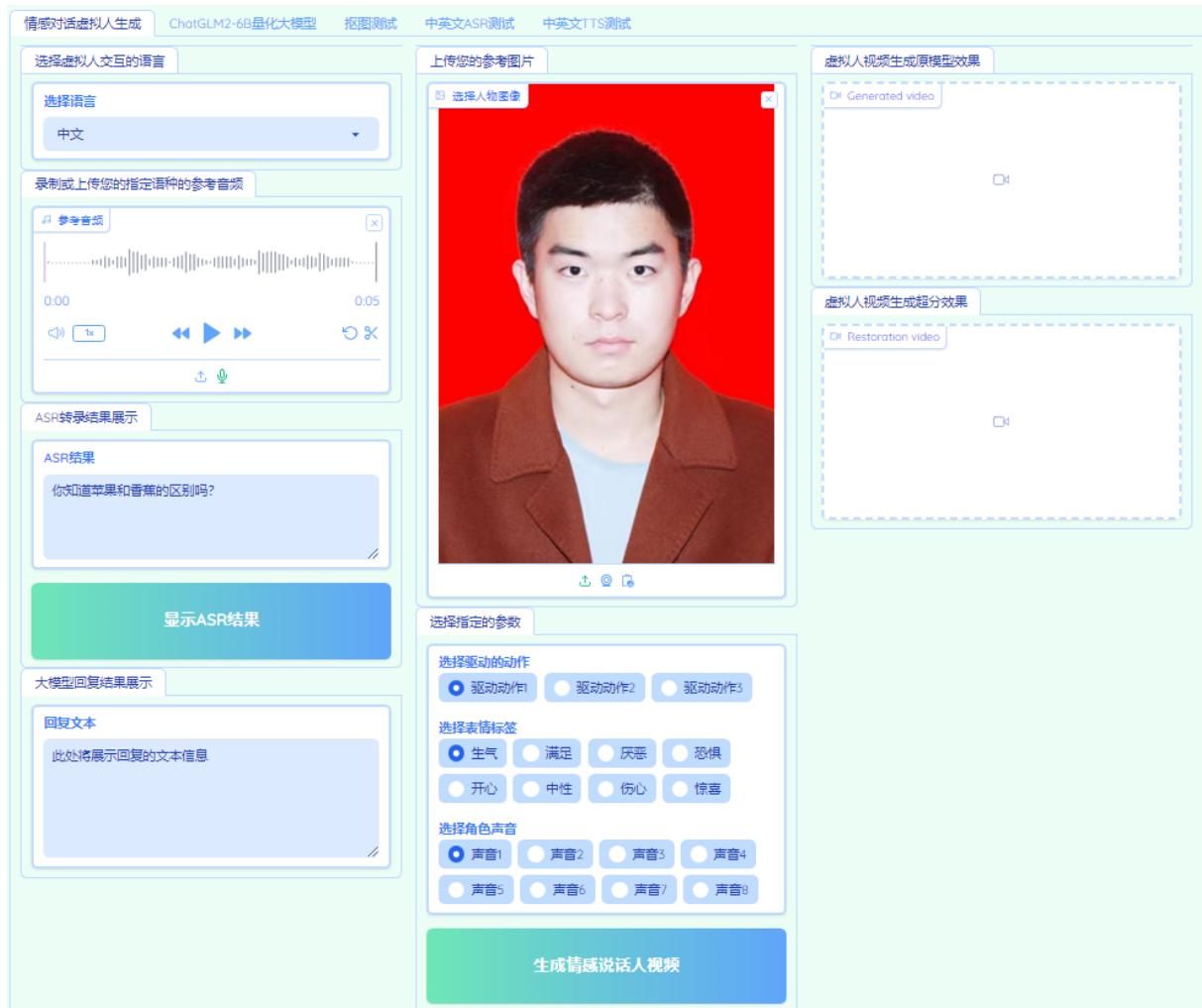


图 5.4: 驱动动作选择测试



图 5.5: 三种不同的驱动动作结果

具体而言，测试时，对比方式是准备一段 1 分钟的音频直接送入 EAT 得到每个驱动动作的结果视频，然后对视频裁剪成帧，每 100 帧 (4s) 取一帧，一共得到 8 帧图片进行对比。每一排为不同的驱动动作表示。这个 pose 的控制经过实验对头部的控制效果似乎控制效果并不是很理想。

5.1.3 表情测试

基于不同的表情，本系统也参照着驱动动作的方法进行控制变量对比。但是测试其表情控制的能力，我们选取不同的特征进行测试。具体而言，我使用了固定的音频对EAT模型进行驱动，得到不同表情的视频，将视频帧，然后按照每100帧取1帧的方式，一共取10帧作为每一行的比较对象。得到的测试结果如下图所示：



图 5.6: 表情结果展示

第一行至第八行的表情标签分别为生气 (Angry), 满足 (Content), 厌恶 (Disgusted), 恐惧 (Fear), 开心 (Happy), 中性 (Neutral), 伤心 (Sad), 惊讶 (Surprised)。为了体现出 EAT 在表情上的优越性，我使用同样的音频和视频对 Wav2Lip 和 SadTalker 的结果进行了对比。



图 5.7: 主流模型结果比较

如图 5.7 所示，第一行为本系统的选取第一个驱动动作和生气表情的结果；第二行为 SadTalker 对同驱动图同音频，默认第一个驱动动作的推理结果；第三行为 Wav2Lip 对同驱动图同音频驱动，选择增强质量模型 wav2lip-gan 的结果。可以明显看出，本系统的结果具有明显的表情，并且具有头部动作，SadTalker 虽然可以生成出质量很高的说话头，但是没有表情，Wav2Lip 同步效果非常好，但是图片的分辨率非常低。

5.1.4 大模型测试

测试对话功能 大模型使用主要是对本地部署的 ChatGLM2-6B 服务进行 FastAPI 的请求。为了方便测试，我同样将 ChatGLM2-6B 集成到了前端进行测试。如图所示，为我们的 ChatGLM2-6B 前端测试界面，我们可以输入进行测试。如图 5.8，我们输入信息后点击确认即可对大模型进行测试。



图 5.8: 大模型对话测试

在测试时，测试信息输入为“你好”和“苹果有什么功效”，模型回复为“你好！我

是人工智能助手 ChatGLM2-6B，很高兴见到你，欢迎问我任何问题。”和“苹果富含多种维生素、矿物质和抗氧化剂，因此具有许多健康功效。以下是一些常见的苹果功效：...”等信息。表明我们正确的向本地服务器部署的大大语言模型发送了请求。

测试清除历史记录功能 通过点击清楚历史记录按钮用户可以，实现对历史对话的清除。如图 5.9所示

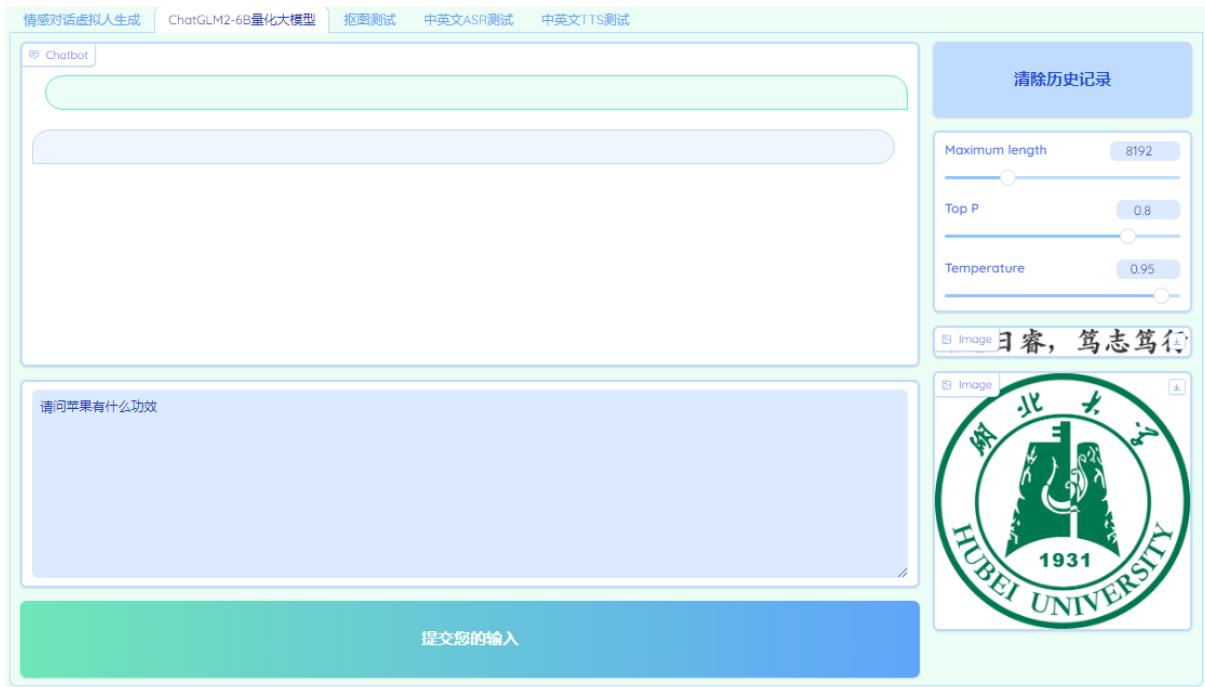


图 5.9: 测试清除对话历史功能

5.1.5 ASR 测试

ASR 自动语音识别模块，此模块主要是将输入的语音转为文字。为了方便测试，将 ASR 模块可视化到前端界面，方便进行测试和查看结果。进入如图界面，我们分别中文和英文语言进行测试。



(a) 中文 ASR 测试

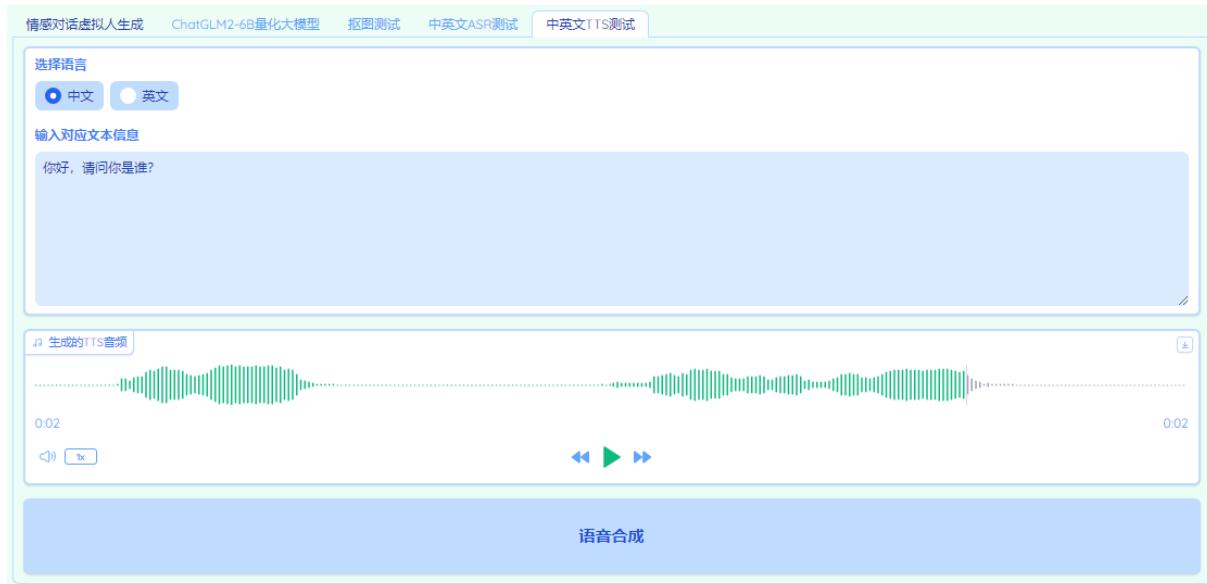


(b) 英文 ASR 测试

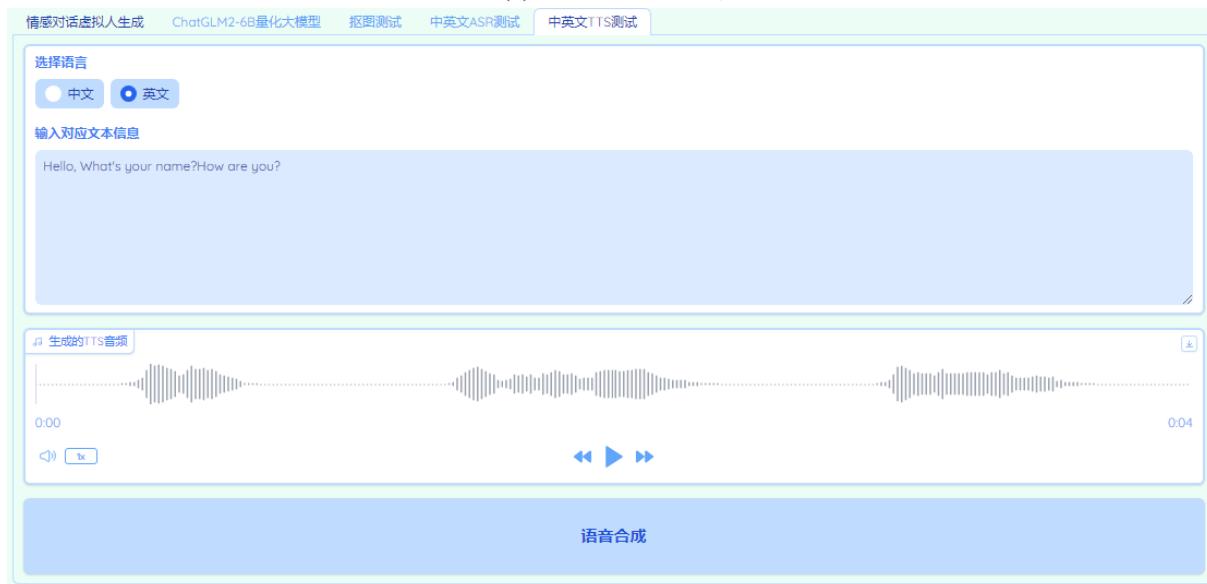
图 5.10: 双语 ASR 测试结果

5.1.6 TTS 测试

本系统设计了双语 TTS，支持中文和英文现对 TTS 进行测试。分别选择中文和英文，输入对应文本信息，然后推理得到对应的音频结果。



(a) 中文 TTS 测试



(b) 英文 TTS 测试

图 5.11: 双语 TTS 测试结果

5.1.7 抠图测试

基于上第三章提到的 EAT 模块的改进，增加的抠图模块可以方便用户对不是纯色背景的人物图片进行抠取，并替换上纯色背景，然后用户可以选择下载图片，并上传作为目标推理图片。如图 5.12a 为抠图测界面展示。



(b) 对测试图片进行抠图结果

图 5.12: 抠图测试

5.1.8 GFPGAN 超分辨率测试

基于 GFPGAN，我实现了在 TensorRT 上的部署，使其速度大大提升，在不牺牲大量精度的情况下达到了极快的推理速度，接近 100 帧每秒。为了更加的直观的进行观察，我们还是选取同一个视频的效果进行比较。如图 5.13 所示，展示 TensorRT 框架部署的 GFPGAN 模型推理结果。



(a) 视频帧超分结果 1



(b) 视频帧超分结果 2



(c) 视频帧超分结果 3

图 5.13: 视频帧超分结果展示

测试时，采用了官方仓库原版 EAT 的推理模型的视频结果进行对比，第一排为原 EAT 结果，第二排为 tensorRT 部署的 GFPGAN 超分结果。可以明显看到，超分后的结果更加的清晰，因为分辨率由原来 224x224 到超分后的 512x512。

5.2 本章小结

本章主要是对系统中相关的功能模块进行测试，通过将功能模块集成到前端方便测试。

6 总结与展望

6.1 全文总结

虚拟人在如今的社会生活中起到越来越重要的作用，虚拟人不仅可以带来巨大的商业价值，并且其作为一种交流的实体可以为人们提供良好的交互体验，由此说明开发一个情感对话虚拟人的必要性。首先本文介绍了虚拟人研究的国内外现状，在介绍完虚拟人的发展脉络之后，本文选择基于深度学习的方式对相关的技术进行研究。随后对虚拟人的相关技术进行调研，以及实验，最后开发出了一套基于大模型的情感对话虚拟人系统，可以实现对话和情感说话人视频生成的功能。

6.2 致谢及展望

在完成本篇毕业论文之际，我要向所有在我大学生涯中给予支持和帮助的人们表示衷心的感谢。

首先，我要感谢我的导师 XXX 教授的指导和鼓励，帮助我修改论文和进行纠错。同时，我还要感谢我的家人朋友，特别是我的父母。感谢你们对我无私的支持和理解。你们的鼓励和信任是我不断前行的动力。我要感谢我的朋友们，你们在我最困难的时刻给予了我力量和鼓舞。最后，感谢 XXXX 的指导和提供实验的设备。在实验室我成长了很多，对于很多前沿的研究有很多的了解，对于看问题的方式也有了更多的改变，这也激励着对技术的不懈追求和探索。大学生涯即将结束，我希望自己在今后的生活中更加坚定自我，踏踏实实，不断追求上进，牢记湖大校训“日思日睿，笃志笃行”，成为一个对社会有用的人。

参考文献

- [1] 崔雪婷, 姚叶子, 陈婧. 浅析元宇宙视域下虚拟数字人在短视频中的应用——以抖音“天好 TianYu”账号为例[J/OL]. 数字技术与应用, 2023, 41(05): 41-44. DOI: 10.19695/j.cnki.cn12-1369. 2023.05.13.
- [2] 本刊编辑部. 锚定未来, 破局立新——2024 养老行业观察[J]. 城市开发, 2024(01): 38-41.
- [3] PRAJWAL K, MUKHOPADHYAY R, NAMBOODIRI V P, et al. A lip sync expert is all you need for speech to lip generation in the wild[C]//Proceedings of the 28th ACM international conference on multimedia. 2020: 484-492.
- [4] ZHANG W, CUN X, WANG X, et al. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 8652-8661.
- [5] WEIZENBAUM J. Eliza—a computer program for the study of natural language communication between man and machine[J/OL]. Commun. ACM, 1966, 9(1): 36–45. <https://doi.org/10.1145/365153.365168>.
- [6] GAN Y, YANG Z, YUE X, et al. Efficient emotional adaptation for audio-driven talking-head generation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023: 22634-22645.
- [7] ZHEN R, SONG W, HE Q, et al. Human-computer interaction system: A survey of talking-head generation[J]. Electronics, 2023, 12(1): 218.
- [8] JELINEK F. Continuous speech recognition by statistical methods[J]. Proceedings of the IEEE, 1976, 64(4): 532-556.
- [9] RABINER L R, JUANG B H. Speech recognition: Statistical methods[J]. Encyclopedia of language & linguistics, 2006: 1-18.
- [10] HANNUN A, CASE C, CASPER J, et al. Deep speech: Scaling up end-to-end speech recognition[A]. 2014.
- [11] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]// International Conference on Machine Learning. PMLR, 2023: 28492-28518.
- [12] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[A]. 2023.
- [13] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models [A]. 2023.
- [14] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models [A]. 2023.
- [15] DU Z, QIAN Y, LIU X, et al. Glm: General language model pretraining with autoregressive blank infilling[A]. 2021.
- [16] KHANAM F, MUNMUN F A, RITU N A, et al. Text to speech synthesis: A systematic review, deep learning based architecture and future research direction[J]. Journal of Advances in Information Technology Vol, 2022, 13(5).
- [17] REN Y, HU C, TAN X, et al. Fastspeech 2: Fast and high-quality end-to-end text to speech[A]. 2020.
- [18] VASWANI A, SHAZER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [19] SHEN J, PANG R, WEISS R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 4779-4783.

- [20] OORD A V D, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[A]. 2016.
- [21] THIES J, ZOLLHOFER M, STAMMINGER M, et al. Face2face: Real-time face capture and reenactment of rgb videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2387-2395.
- [22] CHUNG J S, ZISSERMAN A. Out of time: automated lip sync in the wild[C]//Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. Springer, 2017: 251-263.
- [23] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
- [24] ZHONG W, FANG C, CAI Y, et al. Identity-preserving talking face generation with landmark and appearance priors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 9729-9738.
- [25] ZHANG Z, HU Z, DENG W, et al. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 37. 2023: 3543-3551.
- [26] BLANZ V, VETTER T. A morphable model for the synthesis of 3d faces[M]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023: 157-164.
- [27] LI T, BOLKART T, BLACK M J, et al. Learning a model of facial shape and expression from 4d scans. [J]. ACM Trans. Graph., 2017, 36(6): 194-1.
- [28] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.
- [29] KE Z, SUN J, LI K, et al. Modnet: Real-time trimap-free portrait matting via objective decomposition [C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 36. 2022: 1140-1147.