数字人体验质量评价

报告人：赵楷越 指导老师：闵雄阔

2023年10月23日

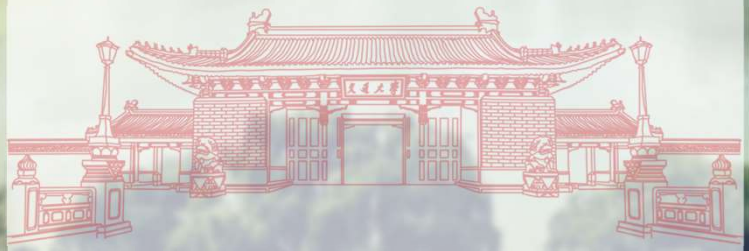
饮 水 思 源 • 爱 国 荣 校

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

目 录

01

课题概览与准备

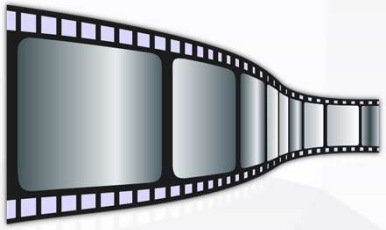


课题概览

Talking Head Generation

自动且有效的数字人说话视频质量评价，
能指引数字人生成算法进行改进。

OR

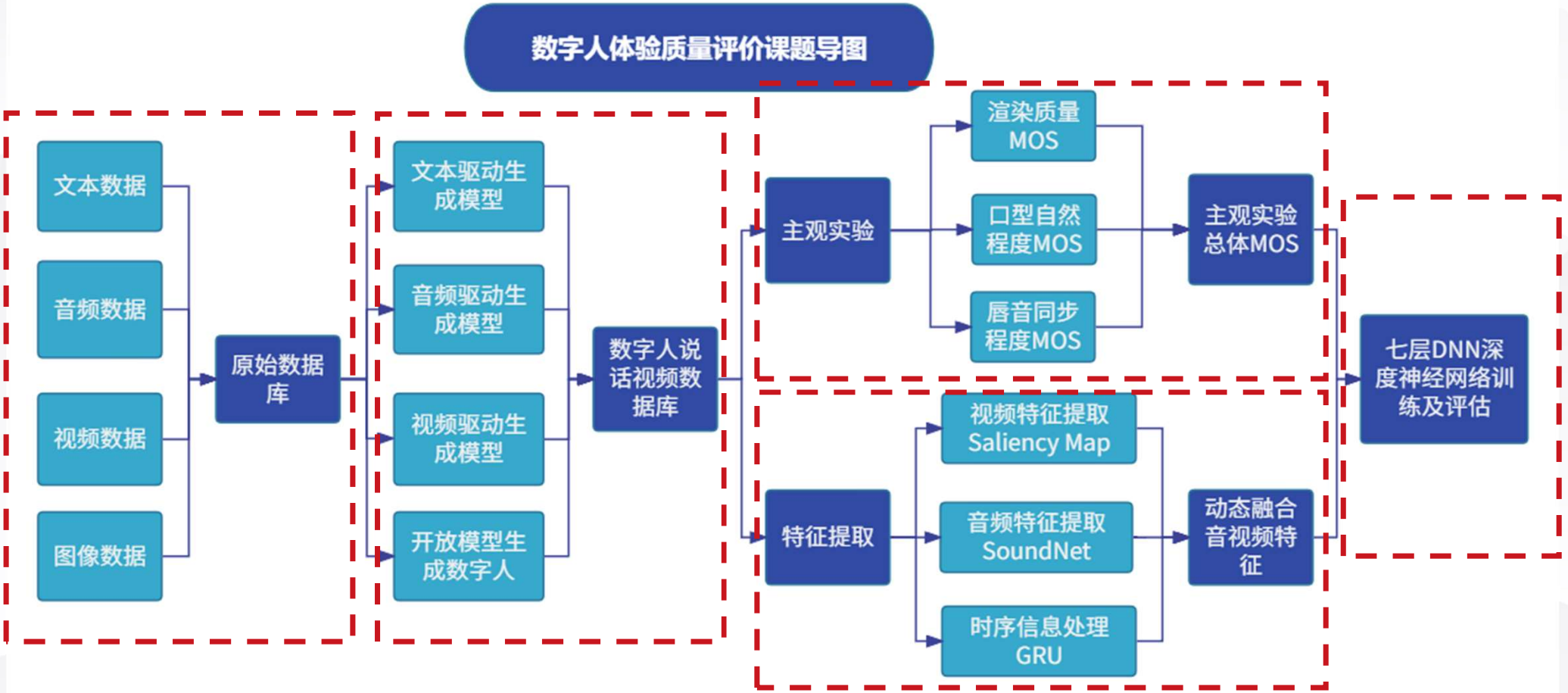


OR

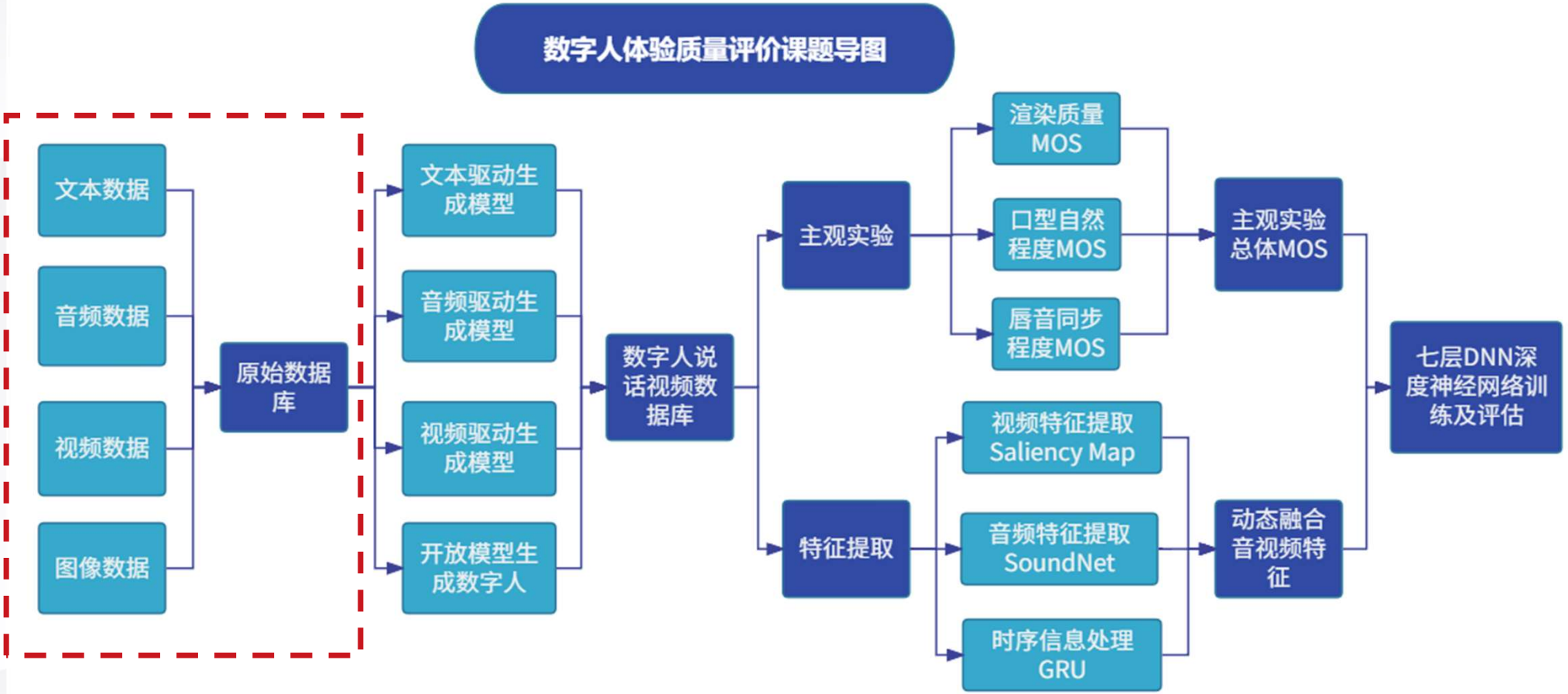
Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt. Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.

目前**缺少**对
数字人说话视频的质量评价方法。

课题概览


数字人体验质量评价课题导图
文本数据
音频数据
视频数据
图像数据
原始数据库
文本驱动生成模型
音频驱动生成模型
视频驱动生成模型
开放模型生成数字人
数字人说话视频数据库
主观实验
渲染质量MOS
口型自然程度MOS
唇音同步程度MOS
主观实验总体MOS
特征提取
视频特征提取Saliency Map
音频特征提取SoundNet
时序信息处理GRU
动态融合音视频特征
七层DNN深度神经网络训练及评估

原始数据库的建立

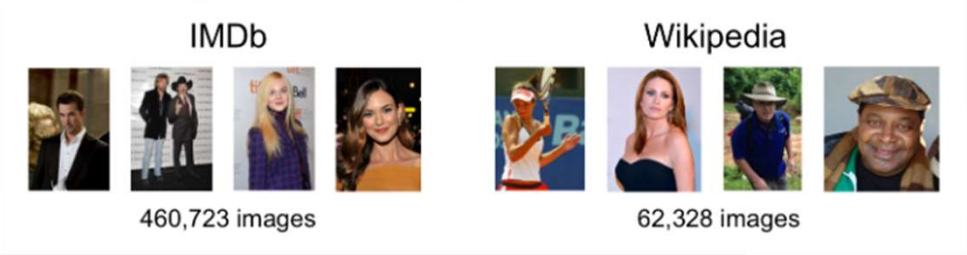

数字人体验质量评价课题导图
文本数据
音频数据
视频数据
图像数据
原始数据库
文本驱动生成模型
音频驱动生成模型
视频驱动生成模型
开放模型生成数字人
数字人说话视频数据库
主观实验
特征提取
渲染质量MOS
口型自然程度MOS
唇音同步程度MOS
主观实验总体MOS
视频特征提取Saliency Map
音频特征提取SoundNet
时序信息处理GRU
动态融合音视频特征
七层DNN深度神经网络训练及评估

原始数据库的建立

原始图像数据库

IMDB-WIKI Dataset

属性：随机背景、年龄、身份；含有自然表情



原始音频数据库

VoxCeleb2 dataset

属性：视频有明显的头部移动，在自然状态下录制

原始视频数据库

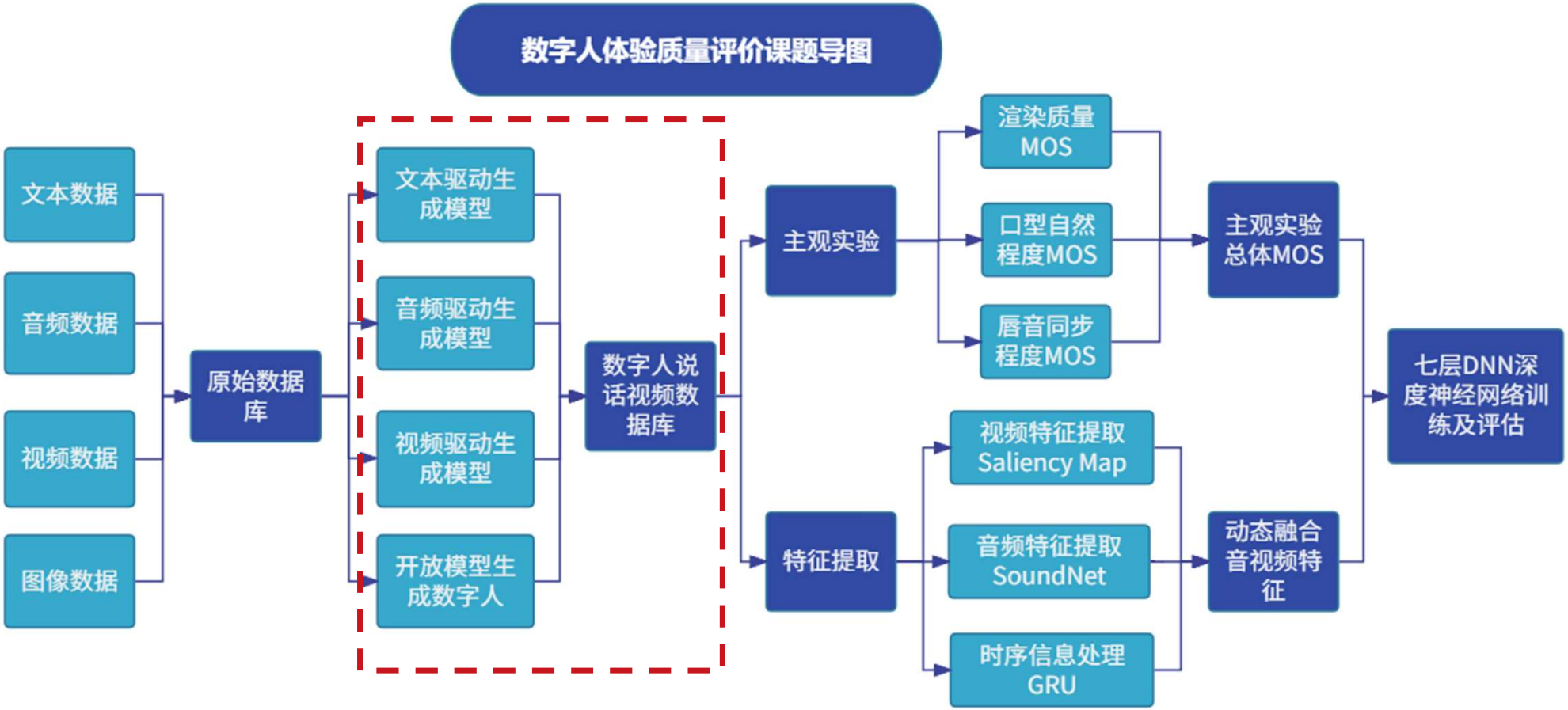
Lip Reading Sentences

数据属性：包含千段的TED演讲的视频片段

原始文本数据库

属性：由具有**不同音素**的英文短句组成。

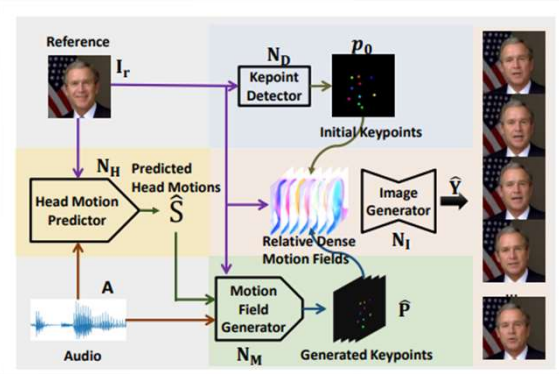
数字人说话视频数据库的建立



数字人说话视频数据库的建立

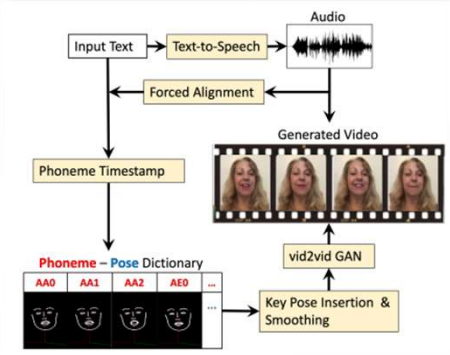
音频驱动生成

采用模型：Sad-Talker，Audio To Obama



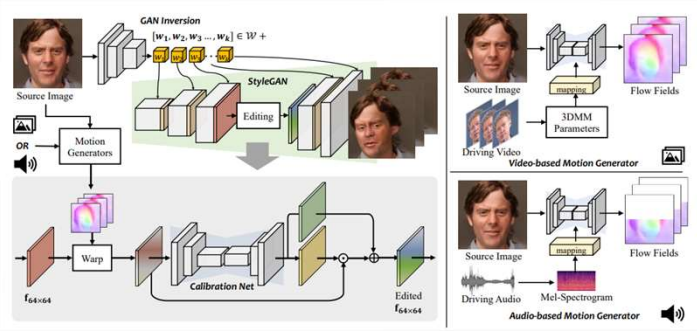
文本驱动生成

采用模型：One-Shot，Multilingual TTS



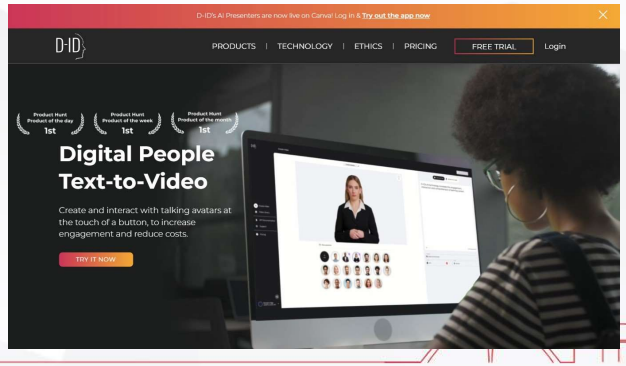
视频驱动生成

采用模型：Video2Head， PC-AVS



开放模型生成

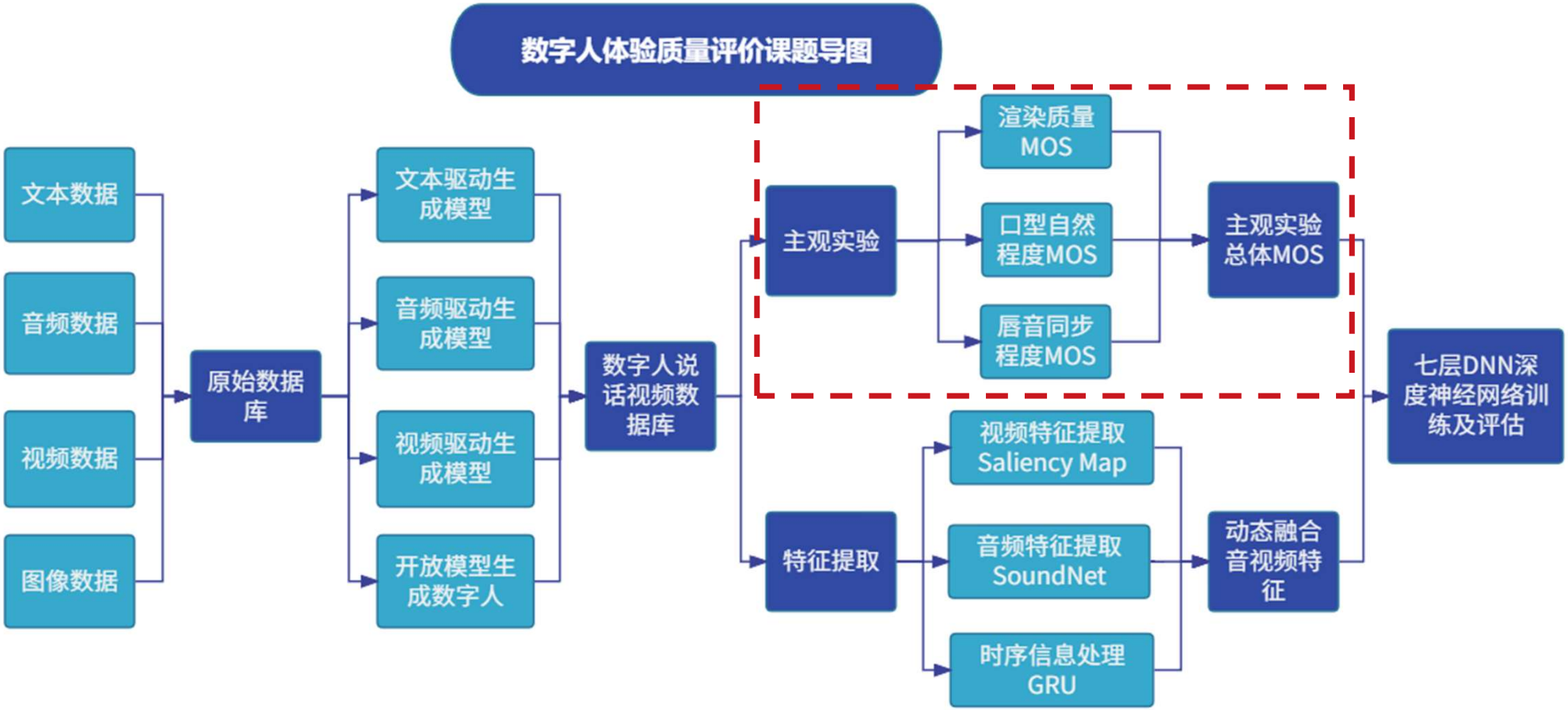
采用模型：D-ID，Synthesia AI



02

主观实验

主观实验


数字人体验质量评价课题导图
文本数据
音频数据
视频数据
图像数据
原始数据库
文本驱动生成模型
音频驱动生成模型
视频驱动生成模型
开放模型生成数字人
数字人说话视频数据库
主观实验
特征提取
渲染质量MOS
口型自然程度MOS
唇音同步程度MOS
主观实验总体MOS
视频特征提取Saliency Map
音频特征提取SoundNet
时序信息处理GRU
动态融合音视频特征
七层DNN深度神经网络训练及评估

主观实验准备

测试方法：开始测试之前会先让测试人员观看真人视频，并告知**根据生成视频与真人视频的差异程度**进行打分。MOS值范围为[0, 100]之间的整数。

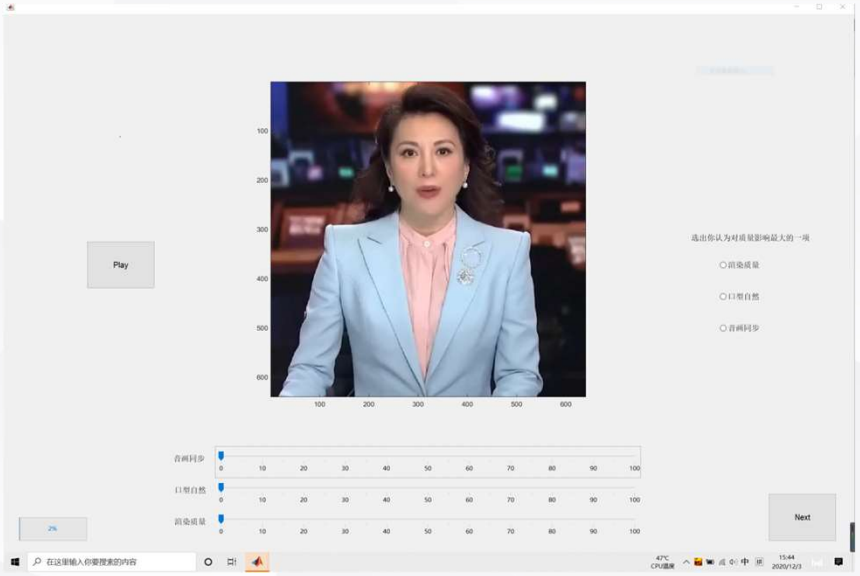
主观测试人员：**10名交大在读学生**（本科生6人，硕士生2人，博士生2人），均不具有数字人质量评价相关知识背景。

评价维度：

1）**渲染质量**：视频整体渲染的感知质量

2）**口型自然**：关注数字人口型抖动情况

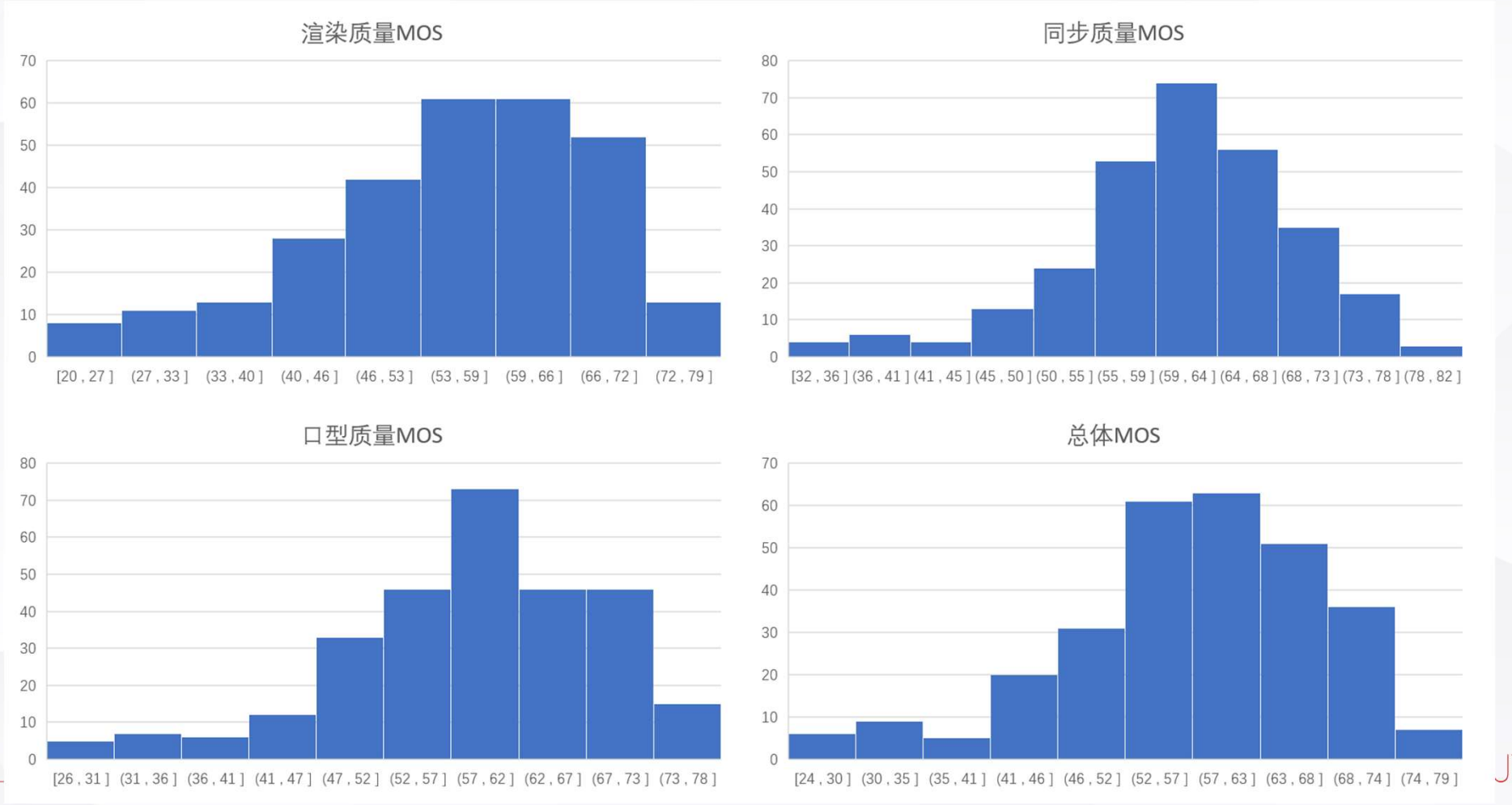
3）**唇音同步**：数字人与音频的同步情况



主观实验UI界面

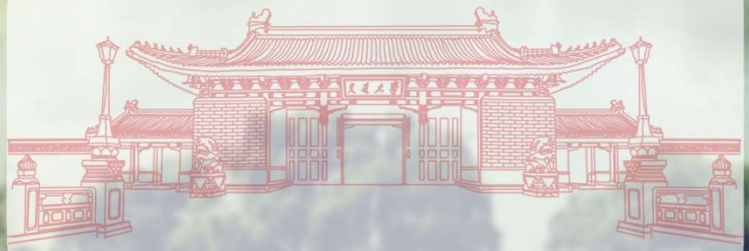
主观实验结果

整体MOS评分 = **0.5***渲染MOS + **0.25***同步MOS + **0.25***口型MOS

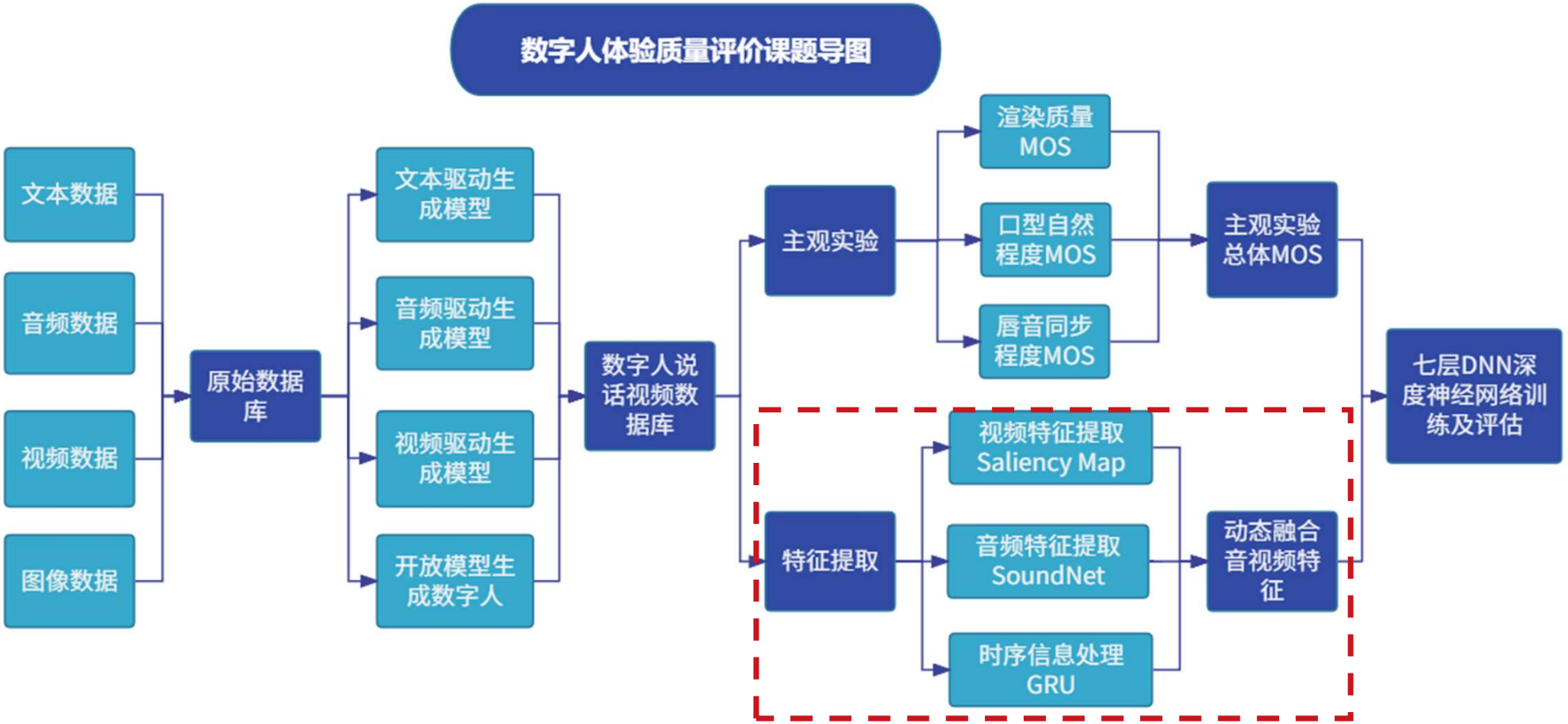


03

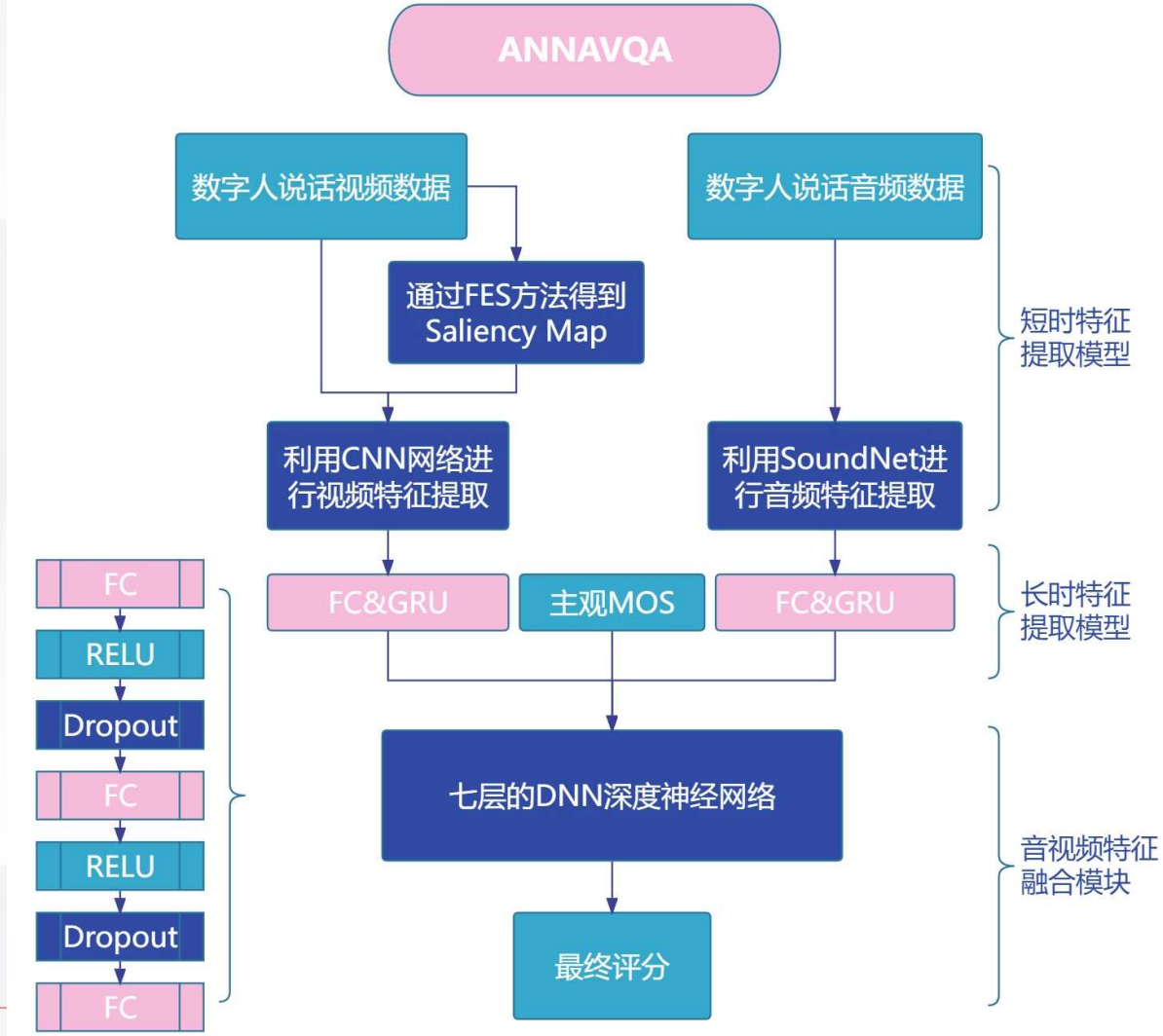
特征提取



特征提取


数字人体验质量评价课题导图
文本数据
音频数据
视频数据
图像数据
原始数据库
文本驱动生成模型
音频驱动生成模型
视频驱动生成模型
开放模型生成数字人
数字人说话视频数据库
主观实验
渲染质量MOS
口型自然程度MOS
唇音同步程度MOS
主观实验总体MOS
特征提取
视频特征提取Saliency Map
音频特征提取SoundNet
时序信息处理GRU
动态融合音视频特征
七层DNN深度神经网络训练及评估

特征提取



基于**ANNAVQA模型**

两个显著优势：

1、考量**视觉注意力**（Visual Attention）

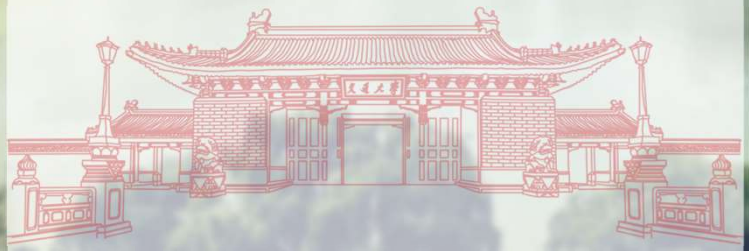
人的注意力会更多放在视频中**运动的实体**，而对视频中的**静态背景关注较少**。

2、考量**时序关注度**（Temporal Memory）

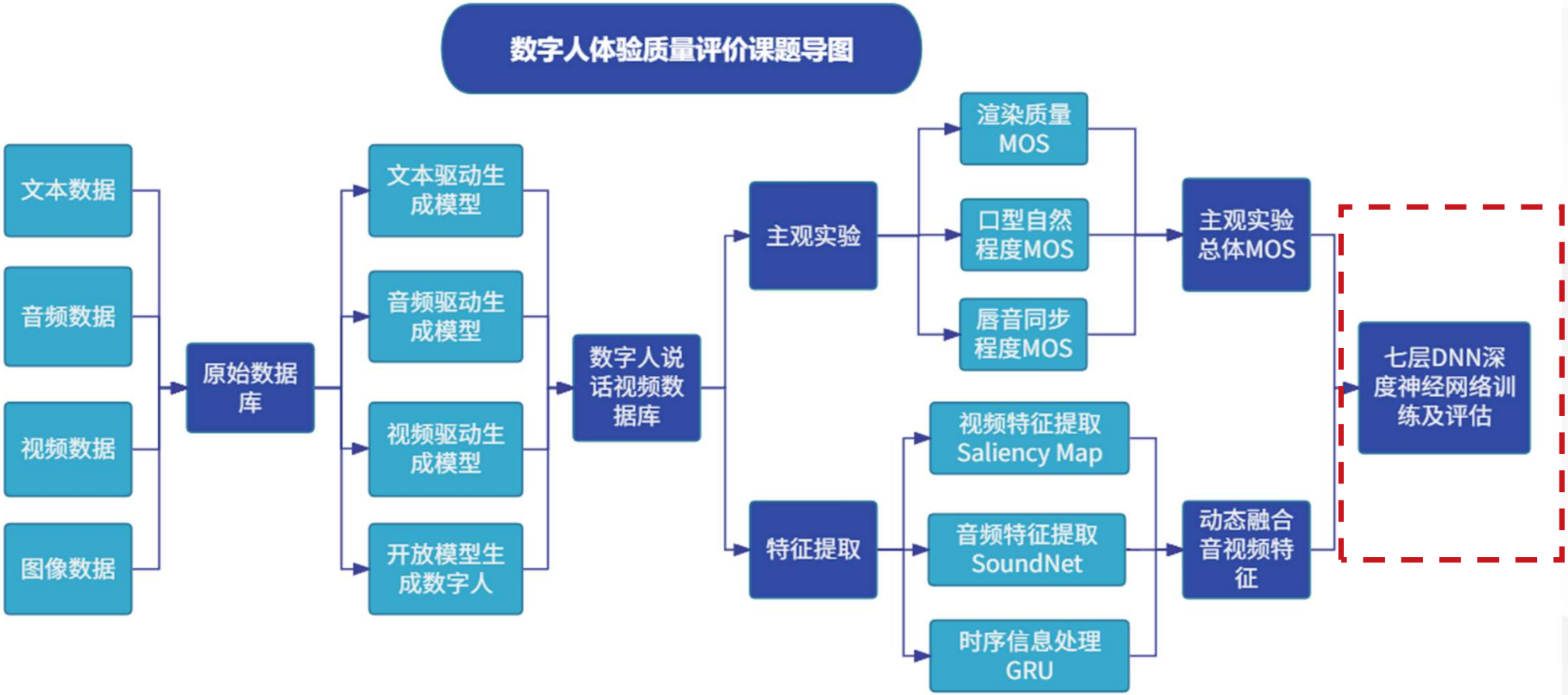
用户评价视频质量时，即使视频质量在大部分时间里不错，但用户更容易受到其中**短暂低质量画面的影响**而给视频打上低分。

04

结果分析



结果分析


数字人体验质量评价课题导图
文本数据
音频数据
视频数据
图像数据
原始数据库
文本驱动生成模型
音频驱动生成模型
视频驱动生成模型
开放模型生成数字人
数字人说话视频数据库
主观实验
特征提取
渲染质量MOS
口型自然程度MOS
唇音同步程度MOS
主观实验总体MOS
视频特征提取Saliency Map
音频特征提取SoundNet
时序信息处理GRU
动态融合音视频特征
七层DNN深度神经网络训练及评估

结果分析（一）

取模型十次随机训练结果的平均值，计算以下四种IQA模型性能指标如下：

因此，在基于大量数字人说话视频数据及对应视频的主观评分MOS值结合训练后，模型能够自动对批量的数字人视频进行客观且准确的打分，达到了本课题的目的。

PLCC=0.7435（准确性-皮尔逊线性相关系数）

$$PLCC = \frac{\sum_{i=1}^{N}(s_i - \bar{s})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2 \sum_{i=1}^{N}(p_i - \bar{p})^2}}$$

SROCC=0.6741（单调性-斯皮尔曼等级相关系数）

$$SROCC = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}$$

KROCC=0.4886（单调性-肯德尔等级相关系数）

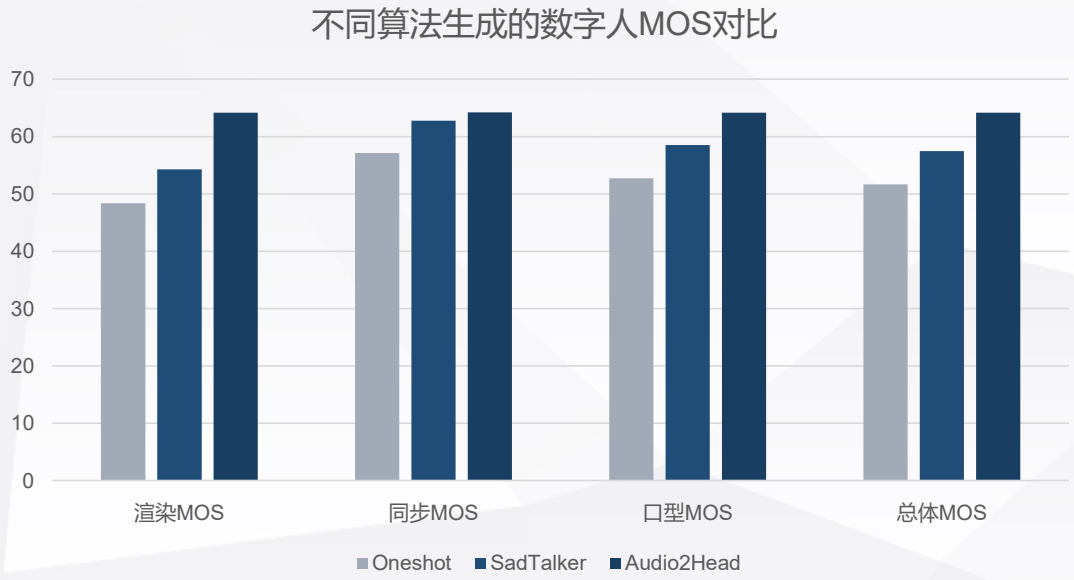
$$KROCC = \frac{2(N_c - N_d)}{N(N-1)}$$

RMSE=0.0792（一致性-均方根误差）

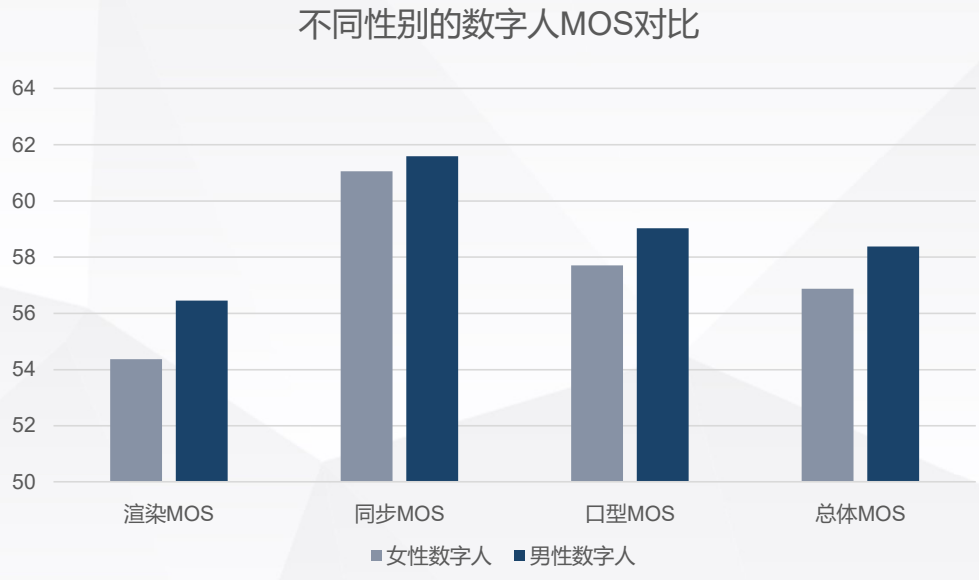
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(s_i - p_i)^2}$$

结果分析（二）

同时，也基于数字人主观MOS评分对不同算法提出了一定**改进意见**



文本 < 视频 < 音频

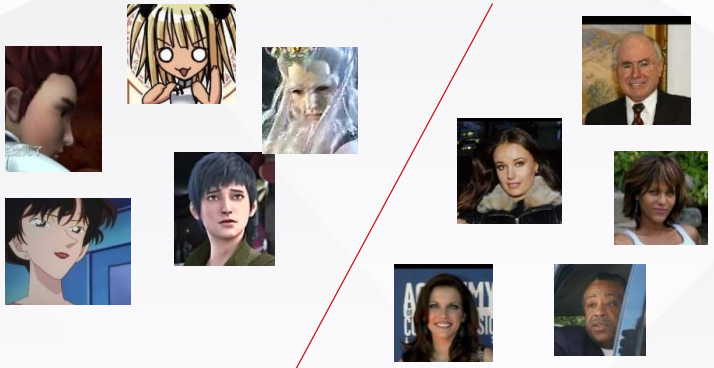


女性数组人MOS < 男性数字人MOS

未来研究方向

由于课题组人力资源紧缺，本课题的受制于数字人说话视频的**数据瓶颈**，难以在在课题研究周期内为数字人体验评价数据库提供**额外的大量充足有效的数字人说话视频原数据**作为模型的进一步训练测试。

虚拟形象/真实形象



未来研究方向：

1）**扩大数据库规模**并且**提升数字人体验评价模型的性能**，深化改良模型算法和方案。

2）进行对**虚拟数字人形象**的人脸逼真度评价与质量评价，增强与虚拟数字人的关系。

3）将本模型与目前其他的数字人质量评价模型**进行对比**，综合分析本模型的优劣势。

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

感谢老师！

饮 水 思 源　爱 国 荣 校