AI技术赋能无障碍影视资源加工

Artificial Intelligence Powered Accessible Film Production

浙江大学

王 炜

2024年5月29日

什么是无障碍电影

无障碍电影是为了方便残障人士观看，经过加工的电影节目

- **面向听觉障碍人群的无障碍电影：增补字幕**
- **面向视觉障碍人群的无障碍电影：增补配音解说**

无障碍电影字幕

战 狼	
故事梗概：	今天放映的是电影《战狼》。该片主要讲述了小人物冷锋成长为拯救国家和民族命运的孤胆英雄的传奇故事。影片于2016年第33届大众电影百花奖斩获优秀影片及最佳新人奖。
主要演职人员：	影片导演吴京；冷锋由吴京饰演；龙小云由余男饰演；敏登由倪大红饰演；石青松由石兆琪饰演。
脚本：	
字幕	解说脚本
冷锋 男	
军衔 中士	
岗位 狙击手	
因在执行任务中擅自行动	
根据纪律条令	
予以行政看管	
————空行————	挺拔的身躯、整齐的正步、橄榄绿色的军装，冷风在两名士兵的随行下进入了问询室内。
2008年8月7日	
据线报	

无障碍电影旁白

无障碍直播解说

2021英雄联盟全球总决赛正式开赛。

S11作为最受关注的电竞赛事之一，B站在开赛期间正式推出了无障碍直播间，辅助听障人士更好地观赛。这也是电竞赛事中首个无障碍观赛直播间。

无障碍电影效果展示

浙江大学 ZHEJIANG UNIVERSITY

无障碍电影制作流程

浙江大学 ZHEJIANG UNIVERSITY

传统无障碍电影制作流程

手动制作

音频转录字幕

对于无字幕电影提取角色对话并转为文字字幕

查找插入区间

识别连续没有字幕的帧，作为可插入旁白的区间

理解制作旁白

理解区间内视觉内容，结合上下文生成旁白文字

旁白转录音频

将旁白内容文字人工朗读并录制生成旁白音频

音频插入合成

将旁白音频与电影原始音轨混合得到无障碍电影

制作成本高，制作周期长
难以推广

无障碍电影制作流程

浙江大学 ZHEJIANG UNIVERSITY

人工智能赋能无障碍电影制作流程

	音频转录字幕	查找插入区间	理解制作旁白	旁白转录音频	音频插入合成
手动	对于无字幕电影提取角色对话并转为文字字幕	识别连续没有字幕的帧，作为可插入旁白的区间	理解区间内视觉内容，结合上下文生成旁白文字	将旁白内容文字人工朗读并录制生成旁白音频	将旁白音频与电影原始音轨混合得到无障碍电影
半自动	使用语音识别技术自动识别转换角色对话	使用文字识别技术自动识别查找旁白区间	使用多模态理解技术自动生成旁白文字	使用语音合成技术自动生成旁白音频	使用模态对齐技术自动生成无障碍电影

EagleMovie无障碍电影制作工具

浙江大学 ZHEJIANG UNIVERSITY

根据中国盲文图书馆、上海光影之声等单位的需求：

1. 优化操作流程，增加步骤指引
2. 新增智能定位字幕位置，减少手工操作
3. 支持旁白区间后台检测功能
4. 旁白导入、导出功能新增支持格式：word，srt，ass
5. 新增字幕导出功能，可以导出word、srt、ass格式文件
6. 新增字幕导入，并把字幕文件合成进原视频功能

无障碍电影《野马分鬃》解说词+对白

说明：

1. 解说词均加粗黑体字凸显，前面标注时间码。
2. 建议以240-260字/分钟速度解说。
3. （）括号表示影片中的特定声响，或需要覆盖对白的提示。
4. 【】实心方括号表示解说速度和感情色彩的提示。

解说词+对白

00:10 影片《野马分鬃》由阿里巴巴影业出品。讲述一个大四学生和一辆二手车的故事。导演魏书钧，主演周游。片名《野马分鬃》指野马在奔驰时，鬃毛向两边分开的情景，也是太极拳中一个招式的名称。

00:28 驾校里，行车道两边绿树成荫，远处立着一座蓝色半圆形建筑。道路用黄色线条隔成左右两部分：其中一部分竖着两排标志杆，围出一条弧形通道。

十几辆白色比亚迪教练车排成一排，缓缓驶过，来到远处那座半圆形建筑下调了个头，往前行驶一小段后，排着队从标志杆围出的弧形通道里穿过。

01:06 其中一辆车缓缓绕了个半圆，穿过几个标志杆。驾驶座上的青年向右打方向盘，刚一加速，却撞上了标志杆停了下来。

01:20 坐在边上的教练和青年对视一眼，转过头，过了会儿又看向青年。

0:01:27 走啊

01:29 青年看了眼教练，又回头看看车外，打起方向盘，往后倒车，然后加速向前驶去。

01:39 干吗干停下

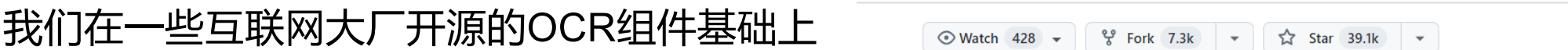
01:41 【快】车子压过标杆停下。

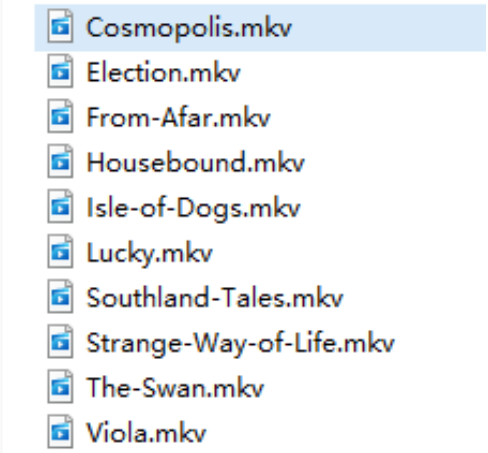
电影字幕识别能力优化

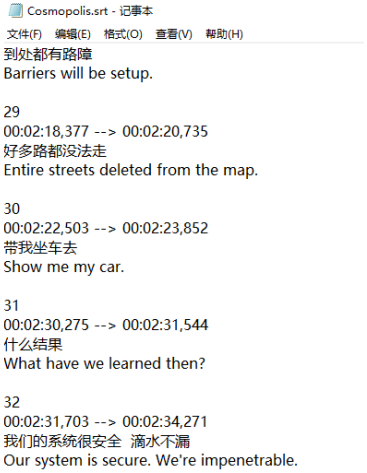
浙江大学 ZHEJIANG UNIVERSITY

目前，市面上缺少针对电影字幕识别的成熟开源组件，仅有通用文字识别(OCR)组件。

我们在一些互联网大厂开源的OCR组件基础上进行字幕识别（该组件在github网站有着39.1K的收藏量，是当前最主流的中文文字识别方法）







评价指标公式：$TF-IDF=\frac{\text{词汇在文本中出现的次数}}{\text{文本词汇的总个数}}*log(\frac{\text{语料库中文本的总个数}}{\text{包含该词汇的文本个数}+1})$

10部电影

共8248行字幕

识别精度90.72%

电影字幕识别能力优化

浙江大学 ZHEJIANG UNIVERSITY

开源组件错误案例分析

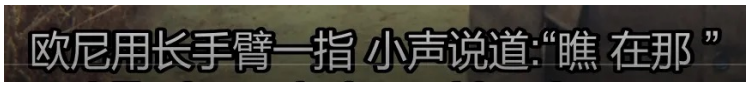
1、误识别画面中的文字

识别结果：NRVANA

2、漏识别文字

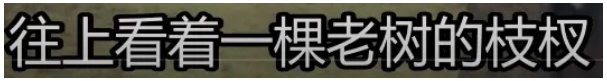
识别结果：切都蒸蒸日上（漏识别：“一”）

识别结果：这些整日在忙碌的城市中巡（漏识别：“逡”）

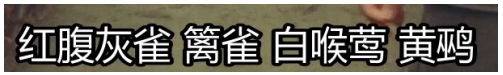


识别结果：欧尼用长手臂一指小声说道：在那（漏识别：”瞧“）

3、识别错误、无法识别难字



识别结果：往上看着一棵老树的枝权（将”枝杈“误识别为”枝权“）



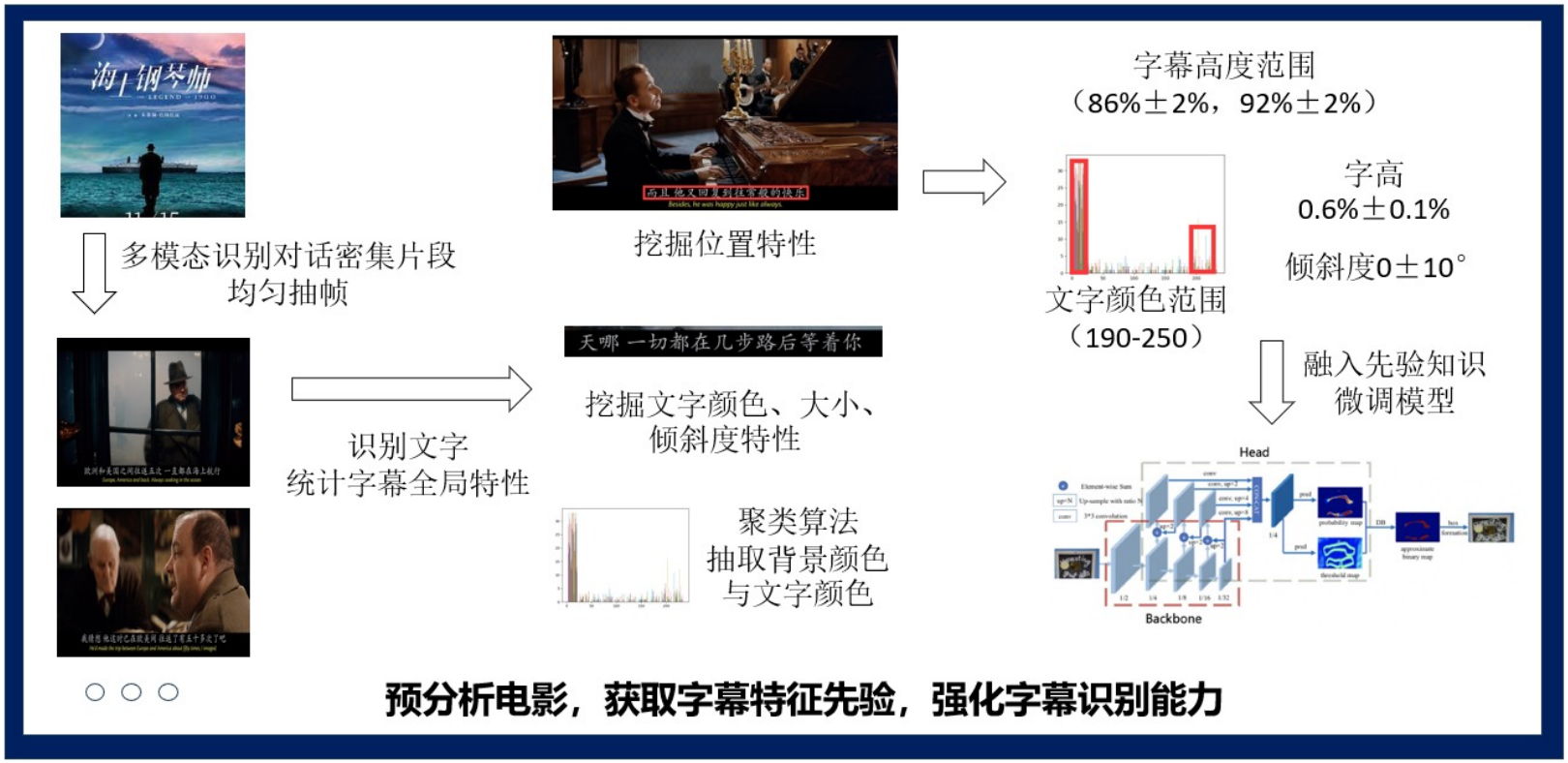
识别结果：红腹灰雀省白喉莺黄鸡（无法识别”篱“、”鸦“）

电影字幕识别能力优化

浙江大学 ZHEJIANG UNIVERSITY

优化方向：

- 针对误识别画面文字的问题，提出字幕先验信息提取算法



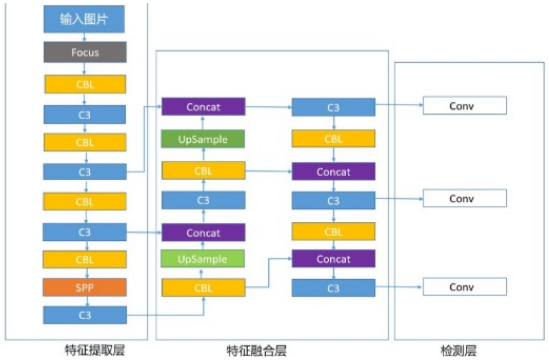
电影字幕识别能力优化

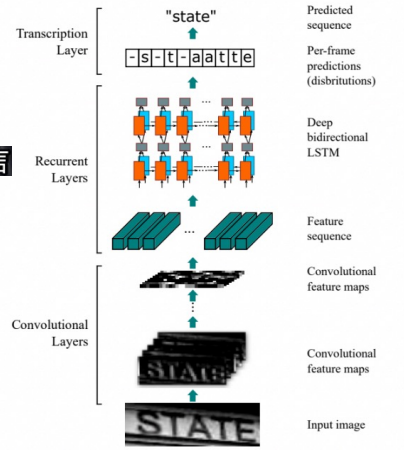
浙江大学 ZHEJIANG UNIVERSITY

优化方向二：

- 针对漏识别文字的问题，重新训练面向电影字幕识别的文本检测模型
- 针对无法识别难字、错误地识别字的问题，用海量电影文本数据微调文本识别模型

海量标注数据





识别结果： “因为外面一直有咱俩的传言”

优化效果：精度从90.72%提升至98.10%

无障碍电影制作流程

浙江大学 ZHEJIANG UNIVERSITY

人工智能赋能无障碍电影制作流程

半自动

音频转录字幕

使用语音识别技术自动识别转换角色对话

减少无字幕电影的字幕生成和判断时间

查找插入区间

使用文字识别技术自动识别查找旁白区间

减少查找可插入旁白区间耗时，预估旁白字数

理解制作旁白

使用多模态理解技术自动生成旁白文字

减少旁白制作耗时和成本，提升制作效果

旁白转录音频

使用语音合成技术自动生成旁白音频

减少人工录制旁白音频的耗时和成本

音频插入合成

使用模态对齐技术自动生成无障碍电影

减少人工合成音轨和视频的时间和成本

基于多模态理解的电影旁白生成

浙江大学 ZHEJIANG UNIVERSITY

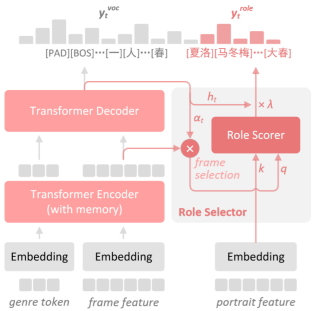
旁白

多模态理解与生成模型

电影主要内容

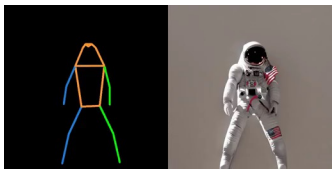
前后帧内容

场景视觉信息



场景音频信息

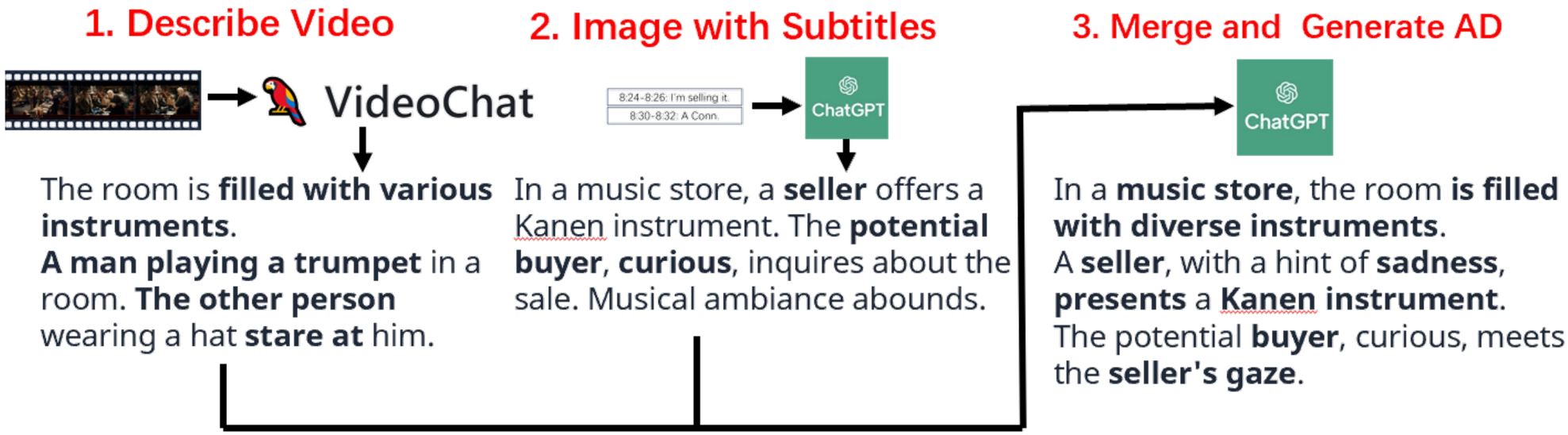
角色动作信息



目前我们的成果

基于多模态理解的电影旁白生成

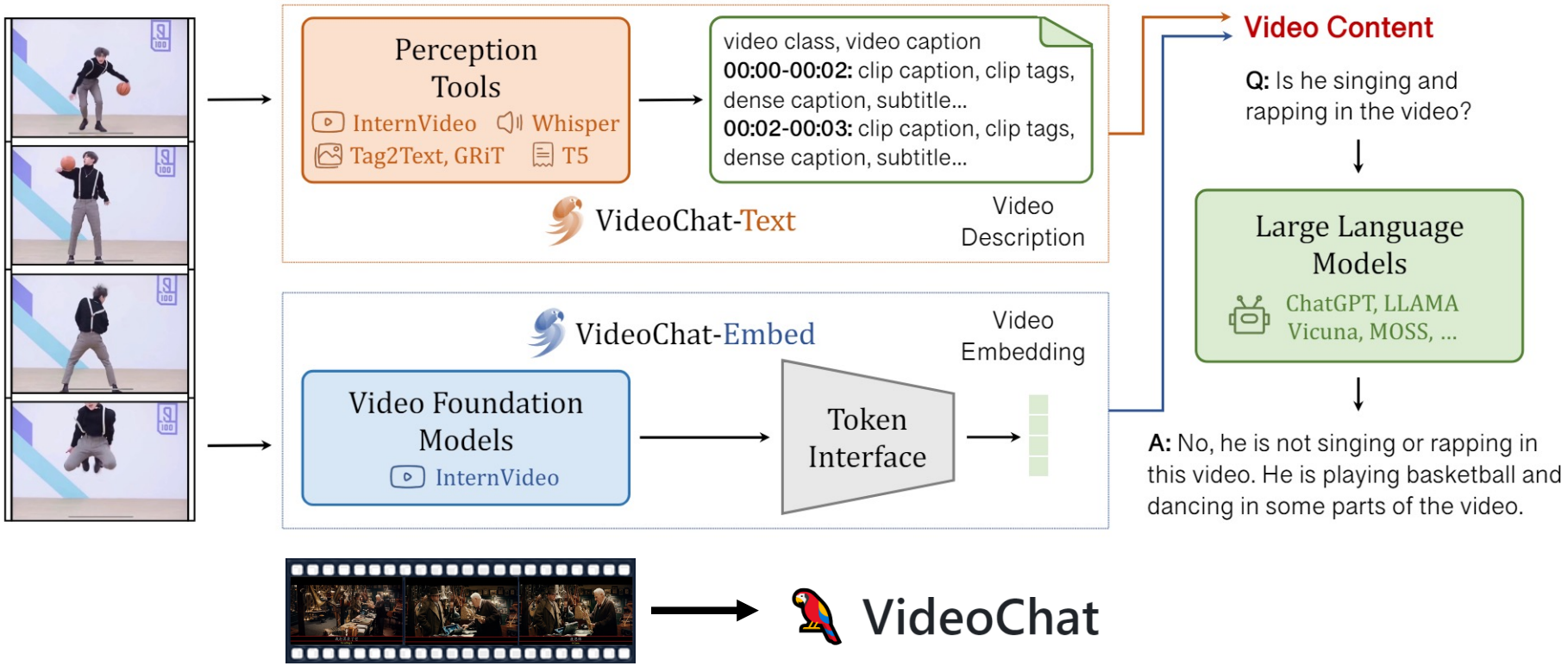
针对现有多模态大模型生成的视频描述缺乏视频中的深层语义信息的问题，我们提出了融合多模态大模型VideoChat和大语言模型ChatGPT，根据旁白片段前后的字幕推断深层语义信息，将两者融合以生成更高质量的旁白。



目前我们的成果

浙江大学 ZHEJIANG UNIVERSITY

- 集成多模态大模型VideoChat来生成旁白文本



视频片段通常非常短，VideoChat 很难正确理解深层语义信息

目前我们的成果

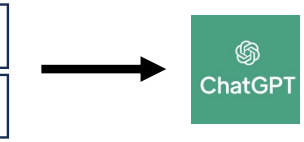
VideoChat 的输出：房间里摆满了各种乐器。一个男人在房间里吹着小号。另一个戴着帽子的人盯着他看。

目前我们的成果

为了解决理解深层语义信息的困难，我们利用 ChatGPT 根据前后字幕合理推断语音间隙的情节，并生成描述。

8:24-8:26: I'm selling it.

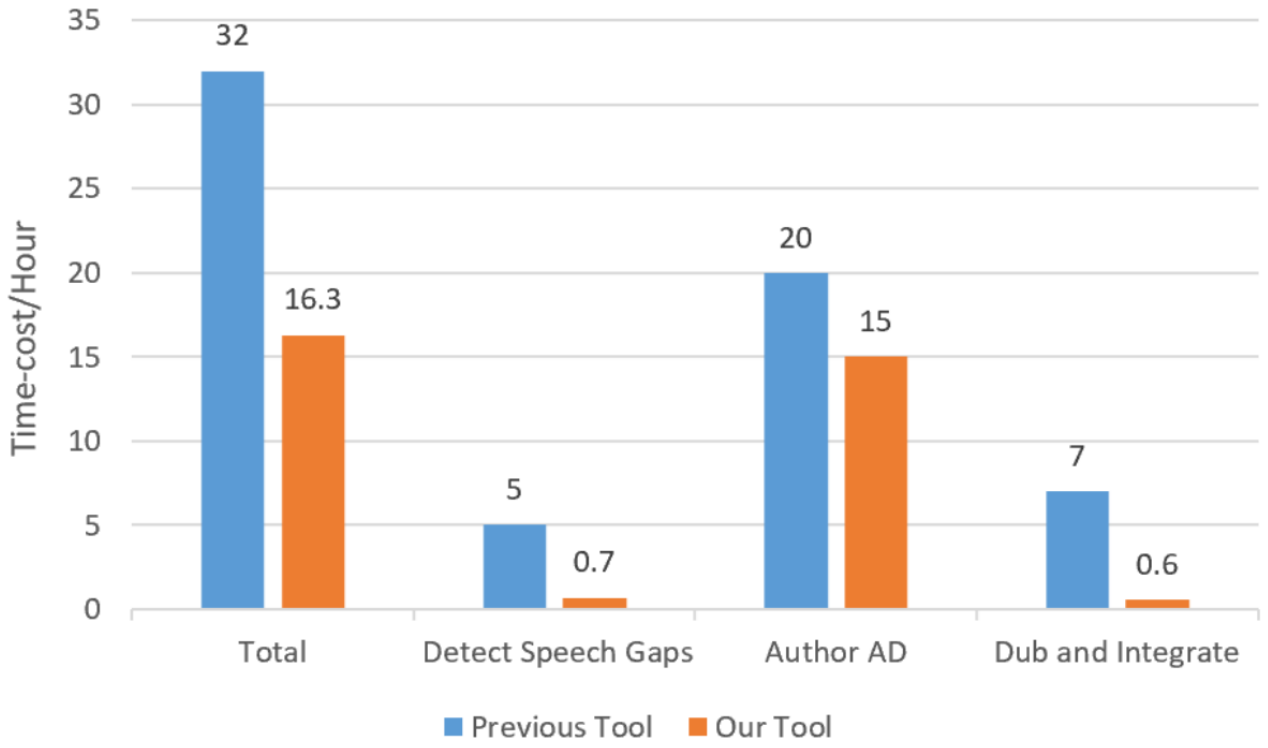
8:30-8:32: A Conn.



最后，我们使用 ChatGPT 合并这两个描述，并提示 ChatGPT 保留更合理的部分。

在一家音乐商店里，房间里摆满了各种不同的乐器。一位销售员带着一丝悲伤，展示着一种卡嫩乐器。一个好奇的潜在买家与销售员的目光相遇。

目前我们的成果



我们邀请了6位志愿者测试并使用我们的系统和之前的系统
对于制作同一部电影，新系统耗时减少了50%

目前我们的成果

浙江大学 ZHEJIANG UNIVERSITY

赞赏👍

- 他们不再需要花费大量精力在上述过程上。
- 生成的旁白描述令人印象深刻，描述了重要的视觉内容，并为他们提供了重要参考。
- 他们充分肯定了该工具对于电影制作的可访问性所做出的贡献。

建议

- 他们建议推理速度应当进一步加快，生成的旁白应更加关注视觉细节，如面部表情。
- 他们应提供各种语言版本，以造福各种文化背景的视障人士。

目前我们的成果

浙江大学 ZHEJIANG UNIVERSITY

Making Accessible Movies Easily: An Intelligent Tool for Authoring and Integrating Audio Descriptions to Movies

Ming Shen[1], Gang Huang[2], Yuxuan Wu[1], Shuyi Song[3,2], Sheng Zhou[1], Liangcheng Li[2,*], Zhi Yu[1], Wei Wang[2], Jiajun Bu[2]
[1]School of Software Technology, Zhejiang University
[2]College of Computer Science and Technology, Zhejiang University
[3]DBAPPSecurity Ltd.
{shenming2023,david.huang,wux521,brendasoung,zhousheng_zju,liangcheng_li,yuzhirenzhe,wangwei_eagle,bjj}@zju.edu.cn

ABSTRACT

Blind and visually impaired (BVI) individuals encounter significant challenges in perceiving the visual content of movies. Audio descriptions (AD) are inserted into speech gaps to describe visual content and storyline for BVI individuals. However, the processes of authoring and integrating AD are laborious, involving tasks such as identifying speech gaps, authoring AD scripts, dubbing, and integrating them into the movie. To streamline these processes, we introduce EasyAD, an intelligent tool to automate these processes. EasyAD utilizes character recognition technology to identify speech gaps and utilizes speech synthesis technology for AD dubbing. EasyAD addresses the misidentification of the background music of existing methods, and for the first time applies a multimodal large language model in the tool to generate AD. EasyAD is currently operational at the China Braille Library and we invite 6 AD authors for a user study. The results demonstrate that with the use of EasyAD, the processing time for a medium-difficulty movie is reduced by nearly 50%, reducing the workload of AD authors and accelerating accessible movie production in China. EasyAD leverages the advantages of AI technologies, especially multimodal large language models, for accessible movie production and benefits BVI individuals.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility systems and tools**.

KEYWORDS

accessible movie, accessibility, blind and Visually impaired, audio descriptions, multimodal large language model

ACM Reference Format:
Ming Shen[1], Gang Huang[2], Yuxuan Wu[1], Shuyi Song[3,2], Sheng Zhou[1], Liangcheng Li[2,*], Zhi Yu[1], Wei Wang[2], Jiajun Bu[2]. 2024. Making Accessible Movies Easily: An Intelligent Tool for Authoring and Integrating Audio Descriptions to Movies. In *Proceedings of The 21st International Web for All Conference (W4A'24).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author: liangcheng_li@zju.edu.cn.

W4A'24, May 13–14,2024, Singapore

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

"That of all the arts, the most important for us is the cinema."– Vladimir Lenin

Movies serve as a significant form of art and play a crucial role in our lives. However, blind and visually impaired (BVI) individuals encounter significant challenges in perceiving the visual content of movies. To address this issue, audio descriptions (AD) [18], which describe visual content and storyline, are integrated into movies at speech gaps. The Web Content Accessibility Guidelines [9] recommends that AD should be provided for all online videos.

Nevertheless, the process of authoring and integrating AD is laborious. Authors have to invest considerable effort in identifying speech gaps, authoring AD scripts, dubbing and editing them into the movie [25]. Besides, condensing the crucial content for comprehension within the limited speech gap also requires advanced skills from authors.

To alleviate the workload for authors, various methods and tools have been proposed by researchers. For instance, tools like CinAD [5] automate the identification of speech gaps and the generation of AD by analyzing movie scripts and subtitle files. However, the challenge arises in obtaining movie scripts and subtitle files on many occasions. Alternatively, tools like 3PlayMedia [1] and descript [7] employ speech-to-text recognition, enabling authors to edit the text for AD creation, which are automatically dubbed and integrated. Nevertheless, there is a risk of misinterpreting background music as text. Other tools like CrossA11y [14], which identifies inaccessible segments through multimodal model alignment. However, the drawback is that these segments might not align perfectly with speech gaps. Methods such as AutoAD [10, 11] initially utilize multimodal large language models to generate AD, achieving commendable results. However, these methods have not yet been practically implemented in tools.

China has a substantial population of nearly 20 million BVI individuals [16]. However, the existing tools have technical shortcomings, and their support for the Chinese language is particularly limited. In the absence of suitable AD auxiliary tools, the primary accessible movie producer in China, China Braille Library [13], relies on general video editing software such as PremierePro [3] and CapCut [6] to author and integrate AD into movies. These manual processes consume significant effort, resulting in a limited

Web4All 2024
21st International Web for All Conference

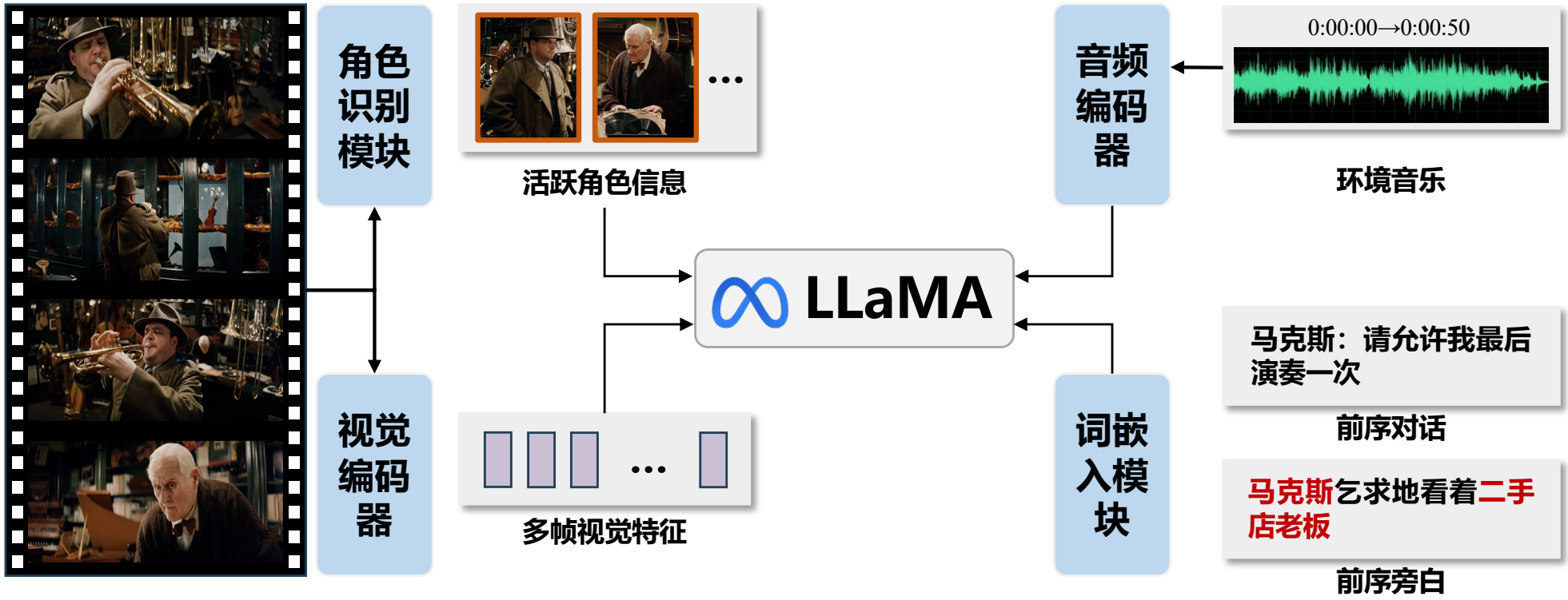
基于多模态理解的电影旁白生成的工作被信息无障碍领域知名国际会议Web4All2024接收
获评最佳通讯论文奖候选

- "Evaluating the Effectiveness of STEM Images Captioning" Maurizio Leotta, Marina Ribaudo
- **Best Communication Paper Candidate: "**Making Accessible Movies Easily: An Intelligent Tool for Authoring and Integrating Audio Descriptions to Movies" Ming Shen, Gang Huang, Yuxuan Wu, Shuyi Song ,Sheng Zhou, Liangcheng Li, Zhi Yu, Wei Wang, Jiajun Bu
- "Does ChatGPT Generate Accessible Code? Investigating Accessibility Challenges in LLM-Generated Source Code" Wajdi Aljedaani, Abdulrahman Habib, Ahmed Aljohani, Marcelo Eler, Yunhe Feng

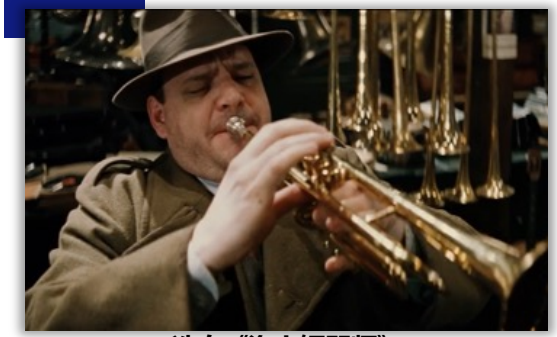
基于多模态大模型的无障碍电影

浙江大学 ZHEJIANG UNIVERSITY

MMAD: Multi-modal Movie Audio Description （COLING 2024）



部分图像识别结果



选自《海上钢琴师》

信息描述准确度不足

CLIP-Caption-Reward	一个戴帽子的男人在刷牙，背景是一把金属勺子。
BLIP	一个人一个小号一个房间。
LLaVA	一个戴帽子的人在吹长号。
标准答案	**带着帽子的马克斯在吹小号。**

选自《海上钢琴师》

缺少时序信息和音频模态

CLIP-Caption-Reward	一群年轻人在看海报，背后是照片。
BLIP	许多人围坐在一起，其中一人拿着手机。
LLaVA	一群妇女坐在一个房间里。
标准答案	**大家在欣赏1900的钢琴演奏，包括缝补衣服的阿姨**

部分图像识别结果

GPT-ClipCap（我们的方法）

一名男子戴着帽子，穿着风衣，在一家陈列着大量号角的商店前吹奏金色小号

整体描述：一名男子戴着帽子在陈列着喇叭的商店前吹小号。

OpenAI ChatGPT4.0

prompt

描述1：带帽子，穿着棕色皮夹克的男人

描述2：吹金色的小号

描述 3：一个有大量号角的店

图中无文本

目前我们的成果

The talented pianist, 1900, mesmerized the audience with his virtuosic performance of "Christmas Eve" while wearing a pristine white tuxedo and bow tie.

Chris Gardner, a man with a box in his hand, runs frantically through the city, dodging people and cars while being chased by a taxi driver who is honking.

- 基于多模态语义信息的互补性，我们提出了一个新颖的框架，该框架善于利用多种模态输入来增强AD 生成，为视障人士提供更丰富的信息。
- 我们设计了叙述者时间间隔检测模块，用于精确定位适当的时间间隔，以便将语音和文本识别同时纳入AD插入。
- 针对多模态输入，我们使用单模态训练方法设计了音频感知环境特征增强模块、演员跟踪感知故事链接模块和电影片段上下文对齐模块，并在框架的输入层设计了多模态转换器以实现多模态融合。

- 无障碍电影数据集构建
- 个性化提示工程
- 无障碍电影大模型

无障碍环境建设法/马拉喀什条约 + 人工智能技术赋能无障碍影视资源（电影、电视、赛事直播）


谢谢！
浙江大学
ZHEJIANG UNIVERSITY