基于开源大模型的短视频与直播 AI 实训系统的设计与实现

史永恒[1]
SHI Yongheng

摘　要　提出了一种基于开源大模型的短视频与直播 AI 实训系统设计方案，旨在通过集成 Llama-3、Stable Diffusion、Stable Video Diffusion 等先进技术，构建一个集教学与实践于一体的智能化平台，促进新媒体人才的培养。系统设计覆盖账号定位分析、剧本策划、分镜头生成至视频合成的全链路，利用 ComfyUI 提供用户友好的操作界面，NestJS 实现高效的服务器管理，以及 MySQL 和 MongoDB 支持数据存储。经过功能测试与用户体验评估，所设计的系统有效提升了学生的实践技能与内容创作水平，具有良好的应用前景。未来发展方向包括模型的持续优化、技术升级，以及加强与产业界的合作，以更好地适配行业需求，开拓新媒体教育的未来空间。

关键词　短视频直播；AI 实训系统；开源大模型；Llama-3；Stable Diffusion；Stable Video Diffusion；ComfyUI；Sora

doi：10.3969/j.issn.1672-9528.2024.08.045

0 引言

随着移动互联网的飞速发展，短视频与直播行业已成为当代数字文化的重要组成部分，极大地丰富了人们的娱乐方式与信息获取渠道。据统计，近年来短视频平台的用户量持续井喷，直播市场规模亦呈现出爆发式增长态势。这一现象不仅催生了新的职业形态，如网红、直播带货主播等，也对传统媒体传播方式提出了革新要求。教育领域对此积极响应，众多高校及职业院校纷纷开设短视频、直播等相关课程，以期培养符合市场需求的新型媒体人才。

近年来，随着 CHATGPT、lLAMA-3、DALL-E、stable Diffusion、stable Video Diffusion 等大模型的发布，AI 辅助视频内容创作的研究取得了显著进展。Sora 的宣传片一出，更是震惊业界，还没发布，就有了“世界模拟器”的外号。

ChatGPT 等语言模型凭借其在语言理解和生成上的卓越表现，为内容创作者提供了撰写直播脚本、剧本的新途径。用户只需提出大致想法或设定，CHATGPT 就能生成连贯、富有创意的文本内容，大大缩短了构思和初稿阶段的时间。

DALL-E 和 Stable Diffusion 等图像生成模型，则正在改变视频分镜头制作的工作流程。以往，分镜头脚本的创作高度依赖于导演或分镜师的个人经验和创意直觉，过程耗时且主观性强。而现在，通过训练 AI 模型学习海量优秀影片的分镜头脚本，AI 可以为创作者提供创意灵感和初步的分镜设计。导演、分镜师不再受限于自己的绘图技巧或素材库的局限，可以直接将脑海中抽象的概念转化为具体的图像，无论是复古未来主义的城市景观，还是梦幻般的生物设计，都能一键生成。例如，输入一个文字版分镜头，AI 可以自动生成视觉化的故事板草图，展示不同的镜头构图、角度切换和动作设计，这有助于导演在实际拍摄前快速迭代和优化视觉叙事方案。

视频内容的制作也迎来了革新。Sora、Stable Video Diffusion 等技术的突破，使得从简单的动态图像到复杂的剧情短片，都能通过 AI 生成。这些模型通过学习大量的视频数据，理解运动规律、场景过渡以及叙事逻辑，能够创造出具有一定连贯性和故事性的视频片段，虽然还不如图片生成技术成熟，但也让人们看到了 AI 在视频创作方面的潜力。

在国内，虽然也有部分院校开设了新媒体运营、短视频制作、直播营销等相关课程，但整体上仍处于起步阶段，教学资源分散，且缺乏与产业界的紧密合作，导致学生毕业后难以快速适应岗位需求。此外，针对 AI 技术在短视频与直播教育中的具体应用的研究尚不多见，尤其是针对短视频与直播实训系统集成的全面解决方案研究仍是一片蓝海。

1 研究目的与目标

本研究拟通过整合多种 AI 技术，构建一个全面支持短视频与直播实训的智能化平台，提供一个集教学、模拟实践于一体的 AI 实训系统，让学生亲历 AI 驱动的内容创作全链路，能够体验强化实战技能，培养符合行业需求的新媒体

1. 青岛港湾职业技术学院 山东青岛 266404
[基金项目] 山东省职业教育教学改革研究立项项目（2021424）

人才。

通过系统，学生能够体验短视频创作中从账号定位、剧本策划到分别生成文字版、图片版分镜头，直至最终由图片分镜头生成视频的全流程，还能够用 AI 直接生成有营销力的直播文案。

2 系统架构与技术选型

本短视频与直播 AI 实训系统分为 AI 大模型、服务器端、数据库和前端四个模块，整体架构设计如图 1 所示。

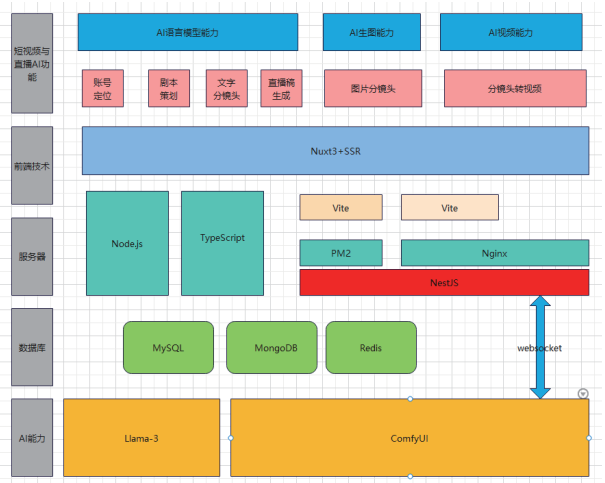

图 1 短视频与直播 AI 实训系统架构图

2.1 AI 能力

大模型选用了目前表现比较好的开源大模型 Llama-3。Llama-3 作为一款强大的语言模型，在文本生成上的表现已得到广泛认可。为了更好地适应短视频与直播内容创作的特定需求，本文对 Llama-3 进行了微调。

AI 绘画大模型选择了 Stable Diffusion。作为一种稳定的图像生成技术，Stable Diffusion 在系统中被用于将文字描述转化为具体的视觉图像。通过对其 API 的灵活调用，可以快速将剧本中的场景描述转换为概念草图，为后续的视频制作提供基础素材。

AI 视频大模型选择了 Stable Video Diffusion。Sora 目前还没有发布，而且发布了的 Sora 也会因收费等问题而不适合应用于教学实训。Stable Video Diffusion 作为除 Sora 外为数不多的视频大模型之一，代表了目前 AI 视频生成领域的较新成果。它能够进一步将静态图像序列转变为动态视频片段，生成连贯且具有艺术感的视频内容，极大地提升实训作品的专业度。

用户界面选用了 ComfyUI：作为用户友好的图形界面，ComfyUI 使得非技术人员也能轻松操作复杂的 AI 工作流。在本系统中，利用 ComfyUI 设计了多个预设工作流程，实训中学生只需选择对应的任务类型，即可一键触发从文本到图像再到视频的生成过程，极大简化了创作流程。

2.2 服务器

服务器部分的功能主要包括用户管理与 websocket 转发，技术上选择了 NestJS 。它为 Node.js 生态带来了类似 Spring Boot 的开发体验，即通过高度结构化的、可测试的、易于扩展的方式来进行服务器端开发。通过它提供的框架，可以快速实现数据库操作和接口搭建。

2.3 数据库

数据库使用了 mysql 和 mongooDB。因为 ConmfUI 的接口需要转发，而 Conmfui 的接口不支持跨域，任务队列管理也有所欠缺，所以采用了服务器中转，并且增加了队列管理功能，队列缓存使用 Redis。

2.4 前端

前端负责用户界面交互，负责用最简单的方式向用户提供定位工具、剧本策划工具、一键分镜头、文字分镜头生成图片分镜头、图片生成视频等功能。采用响应式设计，技术上采用了 nuxt3，UI 库使用 VUETIFY，确保了跨设备的良好体验。

借助微调后的 Llama-3，学生可以通过简单的输入引导，自动生成短视频账号的定位分析报告，明确内容创作方向。随后，系统提供剧本策划工具，帮助学生构思故事情节和角色设定。一旦剧本确定，系统就会自动生成文字版分镜头脚本，接着利用 Stable Diffusion 生成对应图片分镜。最后，通过 Stable Video Diffusion，这些分镜将被合成为动态视频，完成视频制作的闭环。

直播模块专注于直播稿的智能化生成，以 Llama-3 为核心，结合直播行业语言习惯和风格，设计直播词生成算法。该算法综合考虑直播主题、目标受众、品牌调性等因素，自动生成既符合语境又具有吸引力的直播台词。

3 系统功能模块和技术实现

3.1 账号定位与剧本策划：Llama-3 驱动的个性化内容策划策略

在高度竞争的短视频赛道中，Llama-3 的差异化定位建议是通过深度分析现有市场趋势与竞争环境来实现的。

差异化实施：Llama-3 会对同领域内的热门账号进行内容分析，挖掘未被充分探索的子领域或视角。然后，模型会生成具有前瞻性和差异化的定位建议，如专注于某个垂直细分市场，或采用独特的叙事风格。

3.2 剧本生成：微调模型以专业生成剧本与分镜头

为了使语言模型成为专业级的剧本与分镜头生成器，对模型进行了微调，涉及训练模型理解叙事结构、角色发展、

戏剧冲突等剧本创作的核心元素。

微调策略：通过收集大量高质量的剧本样本作为训练数据，特别是那些在目标领域内获得高评价的作品。微调时，强调剧本的结构完整性，如引入、发展、高潮和结局的布局，同时教授模型如何构建引人入胜的对话和场景描述。微调过程中，持续评估并优化模型对特定类型或风格剧本的生成能力，剧本生成模块如图 2 所示。

图 2 剧本生成模块前端界面设计

3.3 文字分镜头脚本到图片分镜：前端调用 ComfyUI 工作流，实现高效文生图

为了让用户更加专注于创作，在前端设计了符合视频创作流程的用户界面，剧本生成后，点击一键分镜，就能进入分镜创作界面。在这里，文字版分镜头已经由语言大模型按照场景自动生成，点击提示词生成按键，可以一键生成文生图提示词。

用户也可以根据需要，自由改动已经生成的提示词。

图片分镜的生成同样不需要复杂的操作，在前端通过调用预置的 ComfyUI 工作流 API（图 3），就可以生成图片分镜。

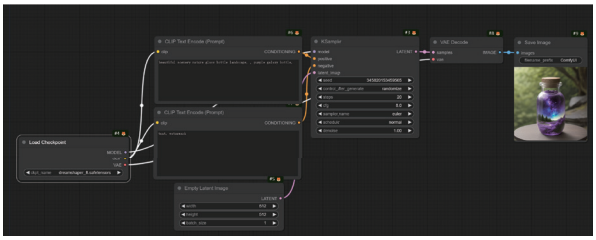

图 3 ComfyUI 文生图工作流

如果对分镜不满意，还可以再次修改提示词，重新生成图片。

3.4 图片分镜到动态视频合成：Stable Video Diffusion 与 ComfyUI 图生视频工作流

Stable Video Diffusion 模型专为连续图像序列到视频的平滑转换而设计，与 ComfyUI 的结合可以简化这一过程。

视频合成工作流：通过 ComfyUI，构建一个工作流，该工作流不仅整合了 Stable Video Diffusion 模型的调用，还包括视频帧率设定、过渡效果添加、时间序列平滑处理等功能。如图 4 所示，用户上传图片分镜后，系统自动处理，应用动态模糊、光影变化等特效，确保视频流畅自然。此外，可集成音频合成选项，让用户在生成视频的同时匹配背景音乐和音效，最终输出一个完成度高的短视频作品。

图 4 ComfyUI 图生视频工作流

通过这一系列智能化流程，从内容策划到视频合成的每一步都得到了 AI 技术的强力支持，极大地提高了短视频创作的效率与创意水平，同时也让内容制作者能够更专注于创意与故事本身。

4 系统测试与评估

4.1 系统功能测试

针对前述各个功能模块，设计了一系列的测试，全面验证了系统的功能性。

账号定位模块：通过不同类型的用户输入，如目标受众特征、内容风格等，检验 Llama-3 生成的定位报告是否符合预期，分析报告的准确性和针对性。

剧本策划模块：输入不同复杂程度的剧本元素，评估 Llama-3 生成的初始剧本初稿在情节设计、人物塑造、台词表达等方面的质量。

分镜头生成模块：验证 Stable Diffusion 和 Stable Video Diffusion 在将文字分镜头转换为图像和视频时的效果，查看生成内容是否贴合原始剧本的视觉表达。

通过大量的测试样例验证，确保了系统各个功能模块的稳定可靠性。

4.2 用户体验评估

除了功能测试之外，还组织了一系列的用户体验测试，邀请了不同背景的学生群体参与。

交互体验：评估前端界面的友好性、操作便捷性，了解用户在使用各项功能时的满意度。

内容生成效果：获取学生、教师对 Llama-3、Stable Diffusion 等 AI 模型生成内容的主观评价，包括创意性、贴合度、专业性等。

整体反馈：通过问卷调查和交流讨论，全面收集学生对于该 AI 实训系统的总体体验感受。

根据测试结果，发现大部分学生对系统功能的易用性和生成内容的质量表示满意，但也有部分学生反映在操作流程和视觉呈现方面仍有进一步优化的需求。

4.3 实践能力培养效果

为进一步评估该 AI 实训系统在提升学生实践技能方面的成效，针对参与实训的学生群体进行了对比分析。对比实训前后学生自主创作的短视频作品，发现学生在镜头设计、剧情构建、视觉表达等方面均有较大进步。

行业匹配度：通过走访当地相关企业，收集他们对学生毕业作品的反馈，发现学生作品的专业水准和商业价值较实训前显著提升。

总的来说，该 AI 实训系统切实帮助学生全面掌握了短视频与直播内容的创作技能，为他们未来的职业发展奠定了坚实的基础。

4.4 系统优势与不足

经过系统测试和实践效果评估，总结出该 AI 实训系统的主要优势和不足之处。

充分利用前沿 AI 技术，为学生提供了全流程的内容创作体验，大幅提升了实践动手能力。系统设计灵活，可根据教学需求快速调整，为多样化的课程实践提供了有力支撑。用户友好的交互界面降低了学生的使用门槛，有利于顺利开展实训活动。

目前还存在的问题是：部分 AI 模型在生成内容的连贯性和专业性方面仍有提升空间，需要进一步优化训练数据和微调策略。

4.5 未来改进方向

针对上述不足，计划在未来进一步优化该 AI 实训系统。持续关注 AI 技术发展，及时引入新的生成模型，不断提升内容创作的专业水准。加强与行业企业的合作，结合实际需求完善系统功能，以确保学生培养的技能与市场需求高度匹配。

5 应用前景与展望

5.1 研究总结

本研究致力于探索 AI 技术在短视频与直播内容创作教育中的应用，设计了一套集脚本生成、视觉内容创作、视频编辑于一体的智能辅助系统。该系统的核心在于 Llama-3 模型的微调、ComfyUI 的集成以及多模态内容生产的自动化流程，这些关键点共同构成了一个高效、创新的教学辅助平台。

系统在教育实践中的价值主要体现在以下几个方面。首先，极大地提升了教学效率与学生创作的积极性，使学生能够在短时间内掌握视频内容创作的全流程，同时激发了他们的创意潜能；其次，通过提供个性化学习资源和即时反馈，增强了教学的针对性与互动性，促进了学生个性化学习路径的发展；再次，系统引入的行业前沿技术和实际操作经验，有效缩短了理论与实践的距离，为学生未来进入职场奠定了坚实的基础。综上所述，该系统不仅优化了教学方法，还为新媒体教育领域带来了一场技术驱动的变革。

5.2 展望未来

随着 AI 技术的持续演进，其在内容创作教育领域的应用前景广阔，未来将更加注重人机协作的深度与广度，以及学习体验的个性化与智能化。AI 将不再局限于辅助工具的角色，而是成为学生创作过程中的智慧伙伴。

5.3 对教育改革的启示

本研究的系统及其应用，为新媒体教育模式的革新提供了重要启示。它表明，AI 技术的融入不仅能够提升教学质量和效率，还能促进教育的公平与包容性，尤其是在教育资源分配不均的背景下，AI 作为强大的辅助工具，可以跨越地域限制，为偏远地区的学生提供高质量的教育内容和创作机会。此外，系统强调的实践导向和跨学科融合，对于培养未来媒体人才至关重要。它要求学生不仅要掌握传统媒体素养，还需具备数字技能、创新能力及跨文化沟通能力，这正是未来媒体行业所需的核心竞争力。因此，该系统不仅是一次技术上的尝试，更是对未来教育理念和模式的一种探索和示范。

参考文献：

[1] 胡泳 .AI 视频的兴起 :Sora 类生成式平台的可能性与风险 [J]. 传媒观察 ,2024(4):5-19.

[2] 张秋月 . 多屏传播语境下短视频创作课程教学的创新路径 [J]. 传媒 ,2024(6):82-84.

[3] 李昕婕 . 人工智能时代的影视行业变革与影视教师能力转向 [J]. 现代视听 ,2023(10):33-36.

[4] 何文睿 , 高丹阳 , 周羿旭 , 等 . 基于扩散模型的多模态引导图像合成系统 [J]. 北京信息科技大学学报 (自然科学版), 2023, 38(6):80-87.

[5] 鄢凯杰 , 孙略 . 人工智能在电影制作领域的应用探究——以分镜头脚本生成工具为例 [J]. 现代电影技术 ,2024(2):35-42.

[6] 曹磊 , 俞剑红 .AIGC 技术在电影数字化创作与制作平台的创新应用 [J]. 北京电影学院学报 ,2023(11):80-91.

【作者简介】

史永恒（1982—），男，山东兖州人，本科，副教授，研究方向：短视频策划、直播营销、人工智能应用。

（收稿日期：2024-05-12）