预印后稿

引用格式：

Balestri, R., Cascarano, P., Degli Esposti, M., & Pescatore, G. (2025年). 重新构想预告片：一种创新的、由大语言模型驱动的表现性自动电影摘要框架（TRAILDREAMS）. 《在线传播与媒体技术期刊》, 15(3), e202524. https://doi.org/10.30935/ojcmt/16669

重新构想预告片：一种创新的、由大语言模型驱动的表现性自动电影摘要框架（TRAILDREAMS）

Roberto Balestri [1]
0009-0000-5008-2911

Pasquale Cascarano [1]
0000-0002-1475-2751

Mirko Degli Esposti [1]
0000-0003-0316-3449

Guglielmo Pescatore [1]
0000-0001-5206-6464

[1] 博洛尼亚大学，博洛尼亚，意大利* 通讯作者：
roberto.balestri2@unibo.it

摘要
本文介绍了 TRAILDREAMS，这是一个利用大语言模型（LLM）自动制作电影预告片的框架。大语言模型的作用是挑选关键视觉片段和具有感染力的对白，并协助 TRAILDREAMS 生成音乐和旁白等音频元素。其目标是高效制作引人入胜、视觉效果出色的预告片。在对比评估中，TRAILDREAMS 的观众评分超过了当前最先进的预告片生成方法。然而，与真实的人工制作预告片相比，它仍有差距。尽管 TRAILDREAMS 展现出巨大潜力，并推动了自动化创作流程的发展，但仍需进一步改进，才能缩小其与传统预告片之间的质量差距。接受日期：2025年3月8日

关键词： 大语言模型、GPT、生成式人工智能、电影营销、电影传播、电影预告片、多媒体

引言

电影预告片通过简短而引人入胜的叙事，塑造电影的市场反响和票房成功。
预告片不仅是宣传工具，还融合了艺术表现力与战略规划精心制作而成。这种创意与策略的结合涉及场景选择和叙事浓缩，从而确保预告片体现电影的核心主题。预告片的制作通常约占电影广告预算的5%，凸显了其在电影营销中的重要作用（Wasko，2003年）。

自2008年以来，预告片风格从传统旁白发展为配合音乐的蒙太奇形式，体现出制作技术的显著变化。这些转变反映出一种趋势，即预告片不再只是直白的广告，而是被制作成发挥微型叙事功能的作品，标志着电影宣传进入了一个新阶段（Richards，2018年）。

像生成式预训练变换器（GPT）（Gallifant等，2024年）系列这样的大语言模型（大语言模型），利用大量文本数据生成连贯且符合语境的文本，推动了自然语言生成的发展。从历史上看，循环神经网络（循环神经网络）在处理序列数据方面发挥了关键作用，尤其是在自然语言任务中（Marhon等，2013年；Piccolomini等，2019年）。然而，循环神经网络面临梯度消失以及处理长距离依赖关系等挑战（Tarwani和Edem，2017年）。这促使研究转向大语言模型；大语言模型利用变换器架构和自注意力机制，使其能够以卓越的准确性和流畅性理解并生成自然语言。

大语言模型融入各类创意产业目前仍处于试验阶段，但已经在改变内容创作的传统方式（Degli Esposti 与 Pescatore，2023年）。在音乐、文学和视觉艺术领域，人们正在测试这些模型生成创意内容、协助创作，以及辅助编辑和制作的能力。艺术家和创作者正在探索新的表达形式与创新方式。对大语言模型的这种试验性应用，是持续开展的研究的一部分，旨在确定如何最好地利用这些技术支持和提升艺术创造力（Epstein 等人，2023年）。

在作者此前开展的一项研究（Balestri 等，2024a）的基础上，本文提出了 TRAILDREAMS——一个将大语言模型的能力与其他先进人工智能技术相结合的新框架，用于自动完成电影预告片制作流程。这种整合能够自动筛选并剪辑关键电影场景和对白，确保预告片准确呈现电影的故事情节，同时吸引观众。

TRAILDREAMS 运用这些技术深入分析电影内容，帮助识别关键的叙事元素和类型特征。这种系统化的方法确保预告片不仅能吸引潜在观众，也忠实于电影的叙事精神。

为展示这项技术的能力，我们制作了混合式预告片，将传统旁白技巧与富有动感的电影对白相结合。这些混合式预告片展示了 TRAILDREAMS 如何通过将人工智能驱动的分析与传统电影制作技术相结合，使预告片制作焕然一新。

我们首先回顾自动预告片生成领域的现有文献，并将其归纳为三种主要方法。我们还探讨大语言模型驱动的方法在视频合成中的应用，并着重说明其日益凸显的重要性。随后，我们详细介绍 TRAILDREAMS 框架，阐述其组成部分及运作方式。接下来，我们分析 TRAILDREAMS 生成的结果，并将其与该领域已有的方法进行比较，例如 PPBVAM（Xu 等，2015）、MOVIE2TRAILER（Rehusevych 和 Firman，2020）以及传统官方预告片。之后，我们对生成的预告片进行批判性分析，讨论框架各个组件可能的改进之处，以提升输出质量。最后，我们总结研究并分享心得与反思。

相关研究

尽管视频内容摘要领域已得到广泛研究，自动制作电影预告片的任务却较少受到关注。现有的视频摘要研究大多旨在精简内容，提供管理不断增长的视频数据量的工具，而非像电影预告片那样打造引人入胜的叙事（Brachmann 等，2009；Zhou 等，2018）。

电影预告片自动生成是一项跨学科研究领域，涵盖自然语言处理、计算机视觉、叙事提取和多媒体内容创作。以往研究探索了各种预告片制作技术，但大多主要关注视觉与听觉方面，而非全面的叙事方法。

若按自动生成电影预告片所采用的方法对相关研究进行归类，可将其分为三种主要方法：

视觉与听觉特征分析

Smeaton 等人（2006年）提出了一种开创性的电影预告片制作方法，根据视觉活动、电影摄影和音频特征挑选精彩片段，并将其剪辑组合成连贯的预告片。

Zhou 等人（2010年）推进了预告片生成研究，重点关注用于场景分类的时间视觉特征，提升了场景检测和类型分类能力，但缺乏叙事整合。

Mahasseni 等人（2017年）提出了一种结合对抗式长短期记忆网络的点过程模型，用于预告片镜头选择，重点关注视觉吸引力和摘要，但未采用涵盖音频和文本分析的多模态方法。

Rehusevych 和 Firman（2020年）推出了“movie2trailer”，利用异常检测识别并汇集独特、引人注目的帧，制作成预告片。尽管其视听特征提取方法颇具创新性，但未能充分解决全面音频策略的整合问题，因此预告片的音频部分仅限于现有配乐片段，且未与上下文进行对齐（Rehusevych 和 Firman，2020年）。

情感与内容分析

Irie 等人（2010年）推出了 VID2TRAILER 系统，运用情感内容分析来选取具有情感冲击力的场景和音频提示，并考虑色彩、动作和对白语调等因素。该系统还融入象征性元素，增强主题表现和品牌识别，从而制作出能够有效传达电影情绪与精髓的预告片（Irie 等人，2010年）。

IBM研究团队开发了一种人工智能辅助预告片制作系统，结合了人类与人工智能的努力。该系统利用多模态分析识别引人入胜的片段，但编辑和叙事构建仍需人工参与，凸显了预告片制作中人工智能能力与人类创造力协作的方式（Smith等人，2017年）。

叙事与情境分析

Pavel等人（2014年）开发了一种视频分割和文本摘要生成方法，以提升内容浏览和导航体验。尽管这增强了观众与信息类视频内容的互动，但仍未满足电影预告片创作的特定需求，因为这需要更深入地整合叙事、音频和视觉元素（Pavel等人，2014年）。

Hesham 等人（2018年）提出的 S-Trailer 框架利用字幕分析来识别类型和关键场景，从而自动生成预告片，为预告片制作提供了一种以文本为中心的方法。然而，这种方法可能无法充分涵盖传统预告片制作中全面的视觉与听觉方面（Hesham 等人，2018年）。

Papalampidi 等人（2021年）提出了一种用于预告片生成的基于图的电影表示方法，根据源自剧本的场景信息对电影进行分段。这种创新方法侧重于叙事进展和角色动态，但可能受限于剧本数据的可用性（Papalampidi 等人，2021年）。

Hu 等人（2022年）利用图卷积网络，依据叙事理论将电影划分为有意义的部分，并借鉴 Snyder（2005年）的理论，旨在突出关键叙事时刻和情感时刻。尽管这种方法提升了场景代表性，但由于采用人工分割，也包含半自动化环节（Hu 等人，2022年）。

Xie 等人（2023年）提出了一种多模态美学引导方法，用于预告片创作。该方法整合了多种类型的信息和美学评估，以确保叙事与视觉连贯性。此方法旨在实现预告片制作的全面自动化，提升故事讲述效果与美学吸引力（Xie 等人，2023年）。

特别介绍：大语言模型驱动的视频生成

在利用大语言模型进行视频生成的领域，一些值得关注的创新虽然并非直接针对自动预告片生成，却彰显了大语言模型在更广泛的视频内容创作领域中的潜力。

例如，Zhu 等人（2023年）的“MovieFactory”是一个能够将文本描述转化为完整电影的框架。该系统借助 ChatGPT 创作详尽的剧本，随后运用先进的视频与音频生成技术将其呈现出来（Zhu 等人，2023年）。

类似地，Long 等人（2024年）的“VideoDrafter”框架利用大语言模型将文本提示转化为详尽的多场景剧本。

虽然这些框架并非专为制作电影预告片而设计，但其方法凸显了大语言模型在创作引人入胜的叙述方面的灵活性与强大能力。

总而言之，现有文学研究已从纯视觉技术发展到针对电影预告片生成任务的有限多模态分析。我们提出的 TRAILDREAMS 框架推动了该领域的

通过集成大型多模态模型来驱动多模态分析，实现预告片生成，并统筹电影中的音频、视频、旁白和对白。这款软件能够管理并合成这些多样化输入，具有独特优势。TRAILDREAMS 确保从场景选择到音频创作的每个环节，都能为自动化预告片制作带来前所未有的叙事整合水平。

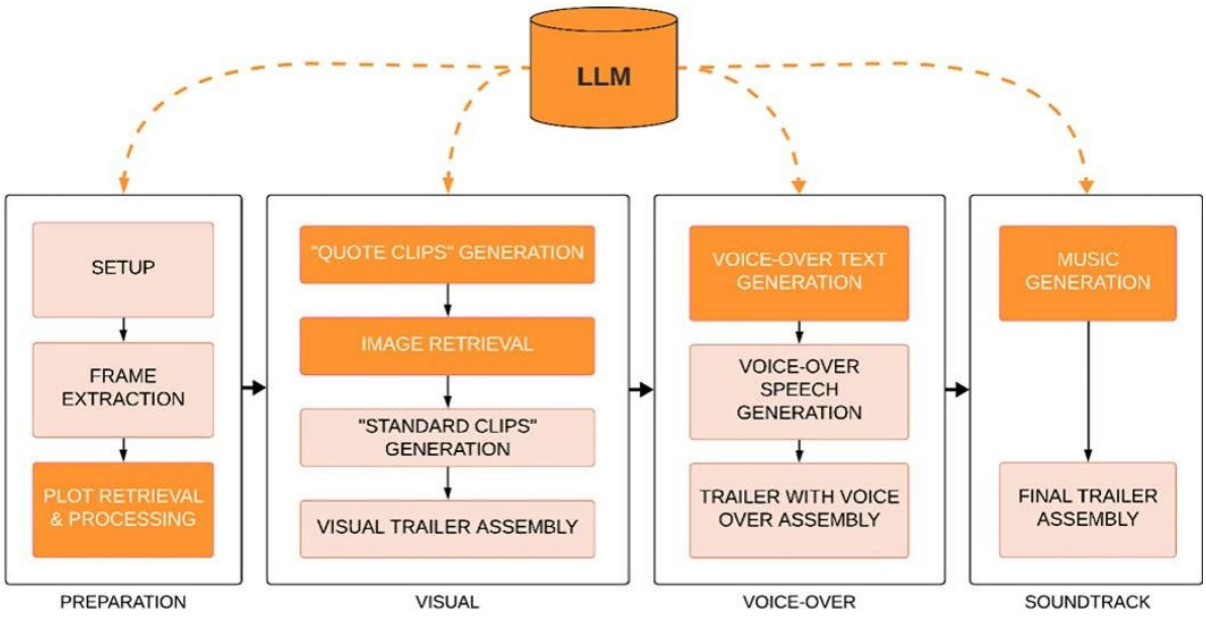

图 **1**。 *TRAILDREAMS* 框架流程 *[*经 *Balestri* 等人（*2024a*）许可转载。© *IEEE* *]*

TRAILDREAMS 框架

本节　本文介绍了所提出的 TRAILDREAMS：一个用于情境感知自动化电影预告片生成的框架　.

图 1 展示　了 TRAILDREAMS 架构的概念图，该架构在方法论上组织为四个核心阶段　es:

1. **准备阶段：**包括初始设置、从电影中提取帧，以及使用大语言模型将电影梗概划分为场景。2. **视觉阶段：**包括创建"引言片段（QC）"和"标准片段（SC）"。前者是由富有感染力的对白组成的片段（由大语言模型选取），用于捕捉电影叙事的精髓；后者则是故事的视觉骨架，在没有对白的情况下引导观众了解情节的发展。随后，这些元素被组装成连贯的视觉预告片。3. **旁白阶段：**大语言模型生成与视觉内容相辅相成的旁白脚本。该脚本被转换为音频，并与视觉预告片同步。4. **配乐阶段：**音乐生成模型在大语言模型的音乐指导下，创作出与电影主题相契合的独特配乐。最终，通过将该配乐与旁白及视觉内容整合，组装成完整的预告片，确保呈现效果协调且富有感染力。

"TRAILDREAMS"框架使用的大语言模型是 OpenAI（2023a）的 GPT-4，通过 API 调用接入。该框架的设计也支持适配开源模型，这表明该方法既具创新性，也易于使用，研究人员可将其应用于更广泛的场景。

准备阶段

阶段 ***0***：设置

在设置阶段，用户可以配置关键设置，以引导电影预告片生成流程。其中包括指定互联网电影数据库（IMDb）的唯一标识符、电影文件的存放位置，以及用于整理项目的名称。

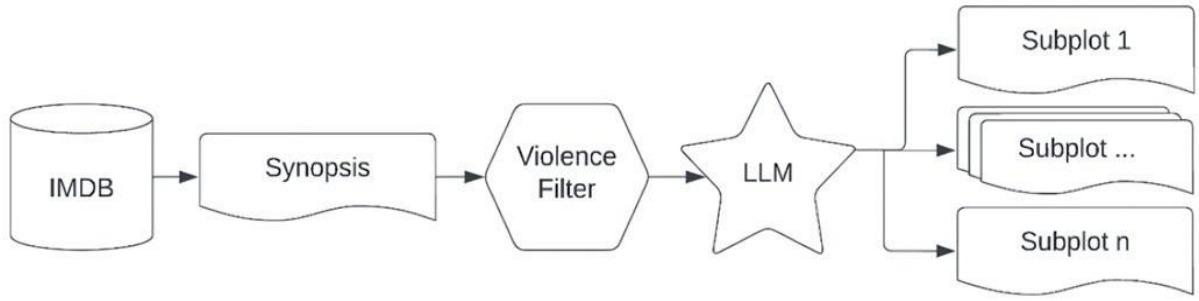

图***2***。剧情简介经过筛选，并拆分为若干子情节（来源：作者自行整理）

IMDb 编码是必需的，因为它能让系统自动抓取并存档全面的电影信息，为后续处理步骤做准备。系统利用 CINEMAGOER Python 库（Alberani，2006年）获取电影的剧情简介、最相关的引文、上映日期、导演、彩色技术、片名和类型等详情。此外，系统还通过专用脚本访问电影的 IMDb 页面，提取相关引文。整合 IMDb 信息后，TRAILDREAMS 便能获得生成预告片所需的全部电影数据。

配置还会确定 SC 和质量控制（QC）的预期数量，不过这些目标可以调整。程序能够灵活修改最终的片段选择，并可根据预告片的整体质量排除部分片段。

用户可以为片段设置最短和最长时长。通过调整音频设置，可为旁白、SC 音频和背景音乐设定合适的音量，从而确保声音设计均衡。

阶段***1***：帧提取

在帧提取阶段，系统会从电影中生成一组帧，以供后续的预告片生成环节使用。该提取过程由 FFMPEG（Bellard，2000年）自动执行。FFMPEG 是一款功能强大且用途广泛的开源软件，专为处理视频和音频而设计。具体而言，在阶段1中，FFMPEG 用于按预定间隔准确捕获帧。

这一阶段必须在帧采样上达到最佳平衡。采样过少的帧可能会限制可用于匹配电影各种支线情节的选项，从而可能遗漏关键的视觉时刻。相反，过度采样可能导致冗余，并增加计算与存储成本。

为解决这一问题，该算法采用一种策略性方法：将电影时长除以九，以确定待提取帧的总数（即每九秒提取一帧）。该比例基于经验得出，旨在最大化帧的表征多样性，同时确保对电影叙事弧的全面覆盖。

这一流程首先调整电影的时间线，剔除影片开头和结尾的片段，以避免片头片尾字幕以及可能泄露结局的内容。随后，算法在调整后的时间线上均匀分布帧提取时间戳。

第二阶段：情节检索与处理

在第二阶段，我们使用大语言模型将电影的剧情简介改写为结构化的情节大纲，为预告片的叙事框架奠定基础。这一关键流程包括剖析预告片中需要着重呈现的核心叙事元素，确保有效传达电影的精髓。

首先，对剧情简介进行过滤，删除可能触发大语言模型严格内容过滤器、导致处理流程中断的内容，例如暴力、色情或种族歧视词语。比如，我们会过滤掉“killed”“cadaver”“corpse”等词，并将其替换为“REDACTED”。

随后，大语言模型将过滤后的剧情简介划分为一系列“子情节”。这一步超越了简单的摘要，因为大语言模型会创造性地识别并提取关键叙事元素，以增强预告片的叙事弧（**图2**）。

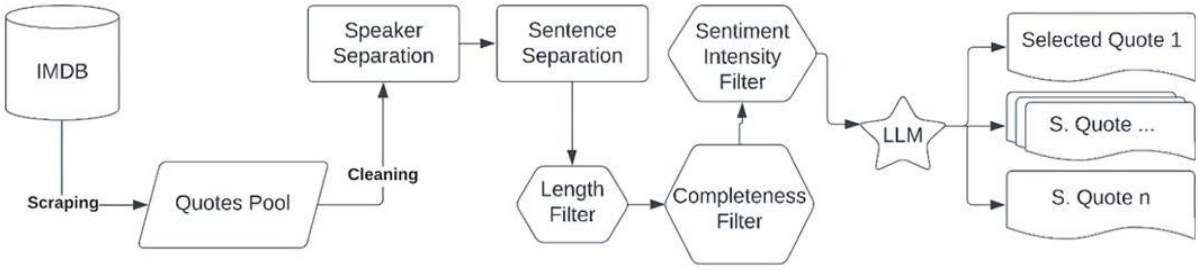

图*3.* 引文选择（来源：作者自行整理）

每个衍生的子情节都被构想为预告片中的一个视觉场景或序列，从而确保汇集出重要、连贯且引人入胜的叙事片段；这些片段组合在一起时，能够有效传达电影的故事情节。

所得子情节被细致地编目到各自的文件夹中，为后续视觉匹配与预告片剪辑阶段的结构化处理奠定基础。

例如，以下是为电影《碟中谍》（De Palma，1996）生成的前三个支线情节：

- 化装后的特工伊森·亨特正准备执行秘密行动。
- 间谍小队在昏暗的房间里整装待发，空气中弥漫着紧迫感。
- 布拉格的背景下，一场神秘的公文包交换出了岔子。

视觉阶段

第三阶段：引语片段生成

在“QC生成”阶段，系统会对电影中的关键对白进行筛选和优化，以增强预告片的叙事深度。收集到一批初始引文后，系统会进行彻底的清理和筛选，在交由语言模型分析之前先对这些引文进行优化。

流程从说话人分离开始，算法会识别并分离不同说话人的引文。具体来说，说话人分离是指整理从 IMDb 提取的引文；这些引文最初会像这样被归在一起，随后拆分成两条独立的引文，分别归属于弗罗多和甘道夫，如下所示：

> 弗罗多：“你迟到了。”
>
> 甘道夫：“巫师从不迟到，弗罗多·巴金斯。他也不会早到。他总会在自己打算到的时候准时到达。”

系统会先清理文本，去除多余字符和空格，以确保格式统一；随后检查每条引文的长度，确保符合预设标准（不少于 12 个字符且不超过 80 个字符）。

系统使用 ‘en_core_web_sm’ spaCy（Explosion AI，2016年）NLP模型验证每个句子的结构完整性，确认句子中同时包含主语和谓语。随后，系统使用 TextBlob Python 库进行情感强度分析，确保每个句子的情感权重都超过最低阈值。这两个步骤对于筛除很可能无关的引文至关重要。处理后的引文按长度排序，

优先保留简洁且具有情感冲击力的引文，并限制为仅保留最短的前 200 条，供后续大语言模型评估。

筛选出所选引文中的暴力用词后，将其交给大语言模型，请模型选出与电影核心主题更加契合的对白（**图3**）。

后续步骤包括提取音频并进行转录，使用 StableWhisper（jianfch，2023年）转录音频。StableWhisper 是 OpenAI（2023b）的 Whisper 语音识别模型的衍生版本。鉴于语音识别可能存在误差，系统采用一种受格式塔模式匹配（Ratcliff & Metzener，1988年）启发的序列匹配算法，将所选引文与对应的音频片段准确对齐。

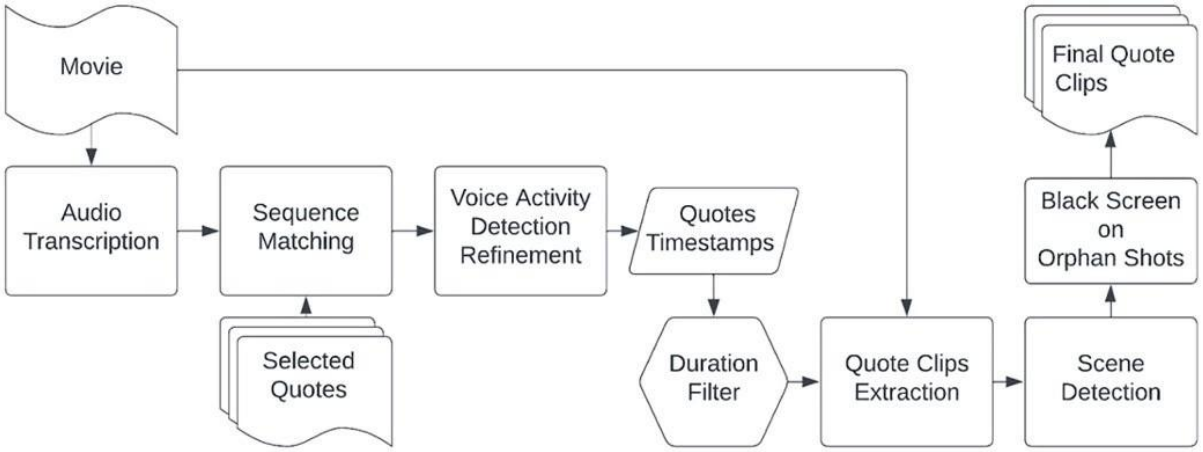

图*4*。 引语片段生成（来源：作者自行整理）

为优化音频片段的时间戳，提高其准确性并降低语句被截断的可能性，系统使用语音活动检测模型 Pyannote（Bredin 等，2020年）。为避免生成过长的质量控制（QC）片段，系统会舍弃时长超过 12 秒的区间（**图4**）。

随后，系统根据优化后的时间戳从电影中提取视频片段。系统对每个片段应用镜头边界检测算法（PySceneDetect 库中包含该算法）（Castellano，2014年），以识别并修正“孤立镜头”——即可能会使最终预告片显得有误的短暂、孤立镜头。系统将检测到的孤立镜头替换为黑屏，以保持预告片的视觉连续性。

质量控制与第二阶段生成的支线情节并无关联；它们之所以至关重要，是因为带来了情感冲击，而非承担叙事功能。

例如，从电影《星际穿越》（Nolan，2014年）中检索到的三段引文如下：

- “爱是我们能够感知到的唯一能够超越时间与空间的事物。”
- “我们一直以来都是凭借克服不可能之事的能力来定义自己。”
- “别让我离开，*Murph*。”

阶段***4***：图像检索

在此阶段，受 Oliveira（2024年）先前工作的启发，系统启动一个过程，将电影叙事与相应的视觉表征对齐。大语言模型从情节线（我们在第二阶段生成的子情节）中提取关键词。这些关键词作为后续步骤的锚点。

本阶段的核心采用多模态句子转换器模型 Clip-ViT-L-14（Radford 等人，2021年；Radford 等人，2022年；Sentence Transformers，2022年），旨在将文本与视觉内容嵌入共享语义空间。该模型衍生自 Sentence-BERT（Reimers 与 Gurevych，2019年），使系统能够理解并将文本描述符（关键词）与视觉内容（电影帧）相匹配，从而促进两种模态之间的语义一致性。

在嵌入过程之后，系统遍历现已转化为结构化嵌入数组的大量电影帧集合。目标是识别那些与关键词具有高度语义相似度的帧。这种相似度通过余弦相似度度量进行量化，以确保所选帧在语义上与关键词提供的叙事线索对齐。为确保电影各部分都得到呈现，我们检查不同所选帧之间的时间间隔（以秒计）至少为电影总时长的 1.5%（以避免在最终预告片中出现来自同一电影场景的镜头）。

为了增强我们生成的预告片的连贯性，并尽可能减少图像检索过程中的错误，我们采用了一种基于帧时间分布的筛选方法。虽然我们承认，线性推进并非人工制作的预告片固有的特征

但经验证据表明，将生成的支线情节中前 40% 的内容与电影最初 40% 的帧相匹配，并将其余场景与随后的 60% 相匹配，可以提高生成的预告片的质量。

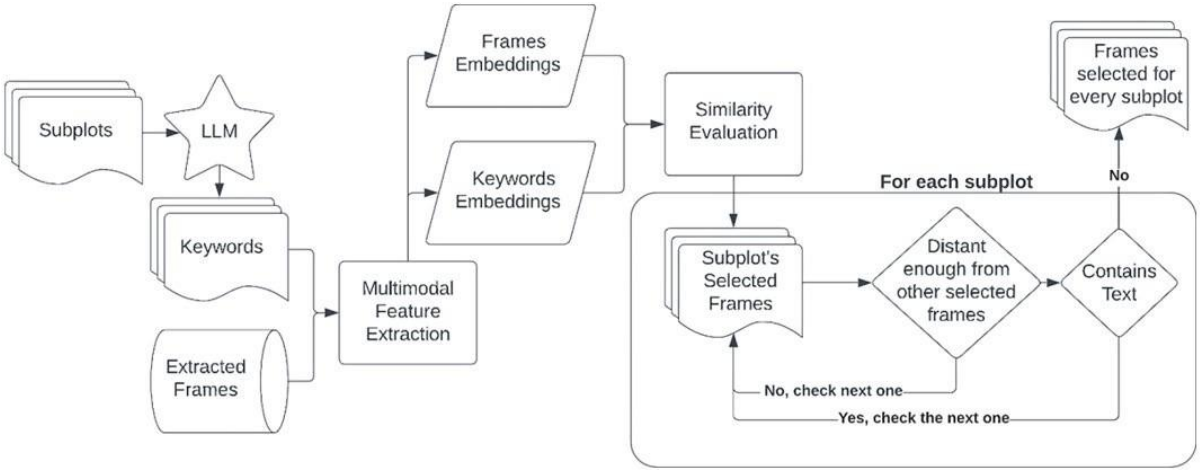

图*5.* 图像检索过程（来源：作者自行整理）

我们还使用基于 CRAFT 算法（Baek等人，2019年）和 CRNN 识别模型（Shi等人，2017年）的 Python库 EASYOCR（JaidedAI，2020年），确保所选帧不含叠加文字。如果检测到文字，系统便会选择下一帧最相似的帧。

现在，每个场景选出的帧都不含无关文字，并且在语义上与电影叙事相契合（**图5**）。

阶段***5***：生成标准片段

在此阶段，系统从上一阶段选出的帧开始生成视频片段。

系统会围绕每个选定帧建立一个“缓冲区”，即向该帧前后延伸的时间窗口。该缓冲区可确保生成的片段涵盖关键时刻前后的完整情境与精髓。

在此缓冲区内，系统采用镜头边界检测算法来确定片段的精确起点。此步骤确保每个片段都是连贯且完整的叙事单元，在视频内容的自然断点处开始和结束，并尽量使片段时长保持在设置阶段指定的最小与最大长度之间。

为进一步优化片段，系统通过更激进的镜头边界检测来评估每个片段内镜头的一致性。它识别并修正任何“孤立镜头”，从而增强片段的叙事流和视觉连续性。在这种情况下，系统不会像质量控制那样用黑屏覆盖孤立场景，而是缩短或延长片段以使其在视觉上更具吸引力，同时尽量保持先前指定的最小和最大长度。

阶段***6***：视觉预告片组装

在此阶段，算法将标准内容与质量控制相结合，以构建连贯的叙事流。首先，系统对标准场景片段进行排序，使其按反映原始叙事顺序的序列排列。这种组织方式有助于防止图像检索阶段可能产生的不一致。

在所提出的方法中，首先使用 Rouard等人（2023年）所述的混合变换器模型对 QC 进行处理，以实现音源分离。该模型会从背景中分离出人声元素，只保留必要的人声部分。随后，通过系统化的间隔策略，将这些 QC 整合到标准场景片段序列中。具体而言，将场景片段总数除以 QC 数量加一，借此确定 QC 在序列中的最佳插入位置（见**图6**）。例如，如果一支预告片总共有 7 个片段，其中 3 个是 QC，4 个是 SC，那么序列将如下所示：SC、QC、SC、QC、SC、QC 和 SC。

为确保听觉过渡流畅，每段质量控制内容都会加入快速的淡入和淡出效果，使听觉变化更加平滑，并尽可能减少不同音频片段之间令人感到突兀的切换。

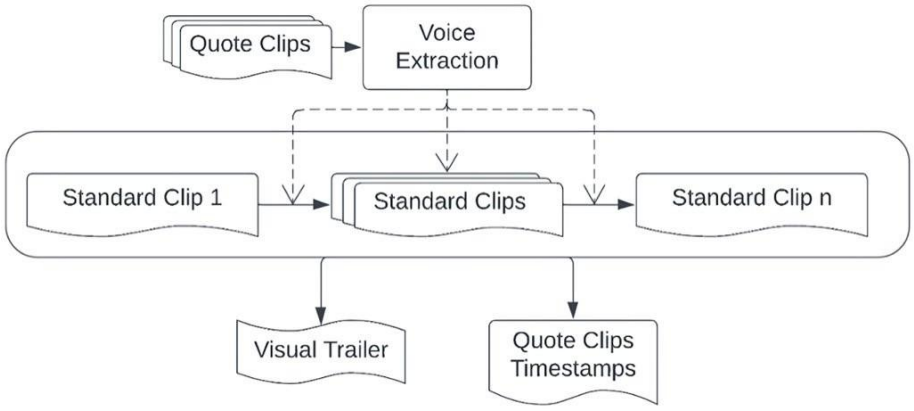

图*6* .视觉预告片组装（来源：作者自行整理 ）

最后，系统会将所有片段拼接成一个统一的视频文件，并生成时间戳日志。该日志记录生成的预告片中每段质量控制内容的起止时间，为后续音频编辑提供信息。 **旁白阶段**

阶段*7*：旁白文本生成

在此阶段，大语言模型会利用电影情节概要、导演信息和上映日期，生成一系列适合用于预告片旁白的引人入胜、发人深省的短句。我们会确保这些短句富有感染力且富有诗意，同时避免剧透，从而保持电影的悬念与魅力。生成短句的数量会根据前几个步骤制作的预告片时长进行调整，以确保叙事与视觉元素和谐融合。例如，为电影《霍比特人：意外之旅》（Jackson，2012）生成的四句短句如下：

- “每个传奇都有一个开端，每场冒险都有一位意想不到的英雄。”
- “在这场伟大征途的边缘，中土世界的命运掌握在最渺小的双手中。”
- “阴影中低语着，邪恶正在蠢蠢欲动，等待着那被黄金触染的烈焰。”
- “与彼得·杰克逊一同，迎接十二月命运的觉醒。”

阶段 *8*：旁白语音生成

在此阶段，算法利用 Coqui xtts-v2 文本转语音（文本转语音）开源模型（Coqui AI，2021年），将脚本文本转换为音频，并采用让声音选择与电影类型相匹配的方法。算法从 LibriTTS 数据集（Zen 等，2019年）提供的五种预设声音类型中，选择与电影类型相符的一种声音。

对于涵盖多种类型的电影（电影在 IMDb 上被标注为多种类型的情况非常常见），算法会采用类型到声音的映射策略，为每种类型分配一种特定的声音类型。当一部电影被标注为多个类型时，算法会统计与这些类型对应的各声音类型出现的次数，并选出与该电影类型关联最频繁的声音类型。随后，文本转语音模型会为每段旁白脚本生成音频文件。从现在起，我们将这些文件称为“语音片段”。

阶段 *9*：带旁白合成的预告片

在此阶段，算法将语音片段整合到预告片中，重点是使其与视觉内容对齐，并调整质量控制的音量，以保持音量平衡。算法首先检索之前保存的时间戳，这些时间戳标示了质量控制在预告片中的位置。利用这些时间戳，算法将语音片段放置在预告片的时间线上，确保它们之间不会发生重叠。

其中一个关键步骤是计算语音片段的平均音量。该数值用于调整质量控制的音量，确保其能够良好地融入预告片的音频中。调整音量后，算法将语音片段与预告片当前的音轨合并，并将音频重新附加到视频上。

配乐阶段

阶段 *10*：音乐生成

算法执行一套详细流程，为预告片合成背景音乐，确保音频能够契合故事氛围与主题。大语言模型如同一位创新型作曲家，根据电影情节创作详细的音乐描述。该描述概括了叙事的精髓，为音乐创作过程奠定基础。例如，大语言模型为《碟中谍》（De Palma，1996）生成的音乐描述如下：

- *In*乐器：交响乐团，突出运用铜管乐器和打击乐器，并以电子元素丰富音色 .

- 调性：多变，运用了调式交替。
- 速度：不同段落的速度范围从行板到快板。
- 力度：力度变化范围广，从隐秘场景中的极弱到动作序列中的极强。
- 织体：复杂，各段落之间交织着不同的动机。
- 节奏元素：运用切分音和混合拍子营造紧迫感与张力。
- 和声元素：运用不协和和弦与半音进行营造神秘感和悬念。
- 旋律元素：为角色设计主导动机，并运用小调音阶和八度跳进增强力度。

随后，提示 MusicGen（Copet 等人，2024年）文本转音频模型生成与大语言模型生成的文本相对应的音乐。

阶段 ***11***：最终预告片剪辑

该系统将创作的音乐与预告片画面及现有音频整合，并采用音频闪避技术来保持音频平衡。实际上，当出现质量控制音频或语音片段时，算法会降低音乐音量，以确保叙事清晰。完成闪避处理后，系统会将调整后的音频与视频合并。此步骤还包括为音频和视频添加淡入和淡出效果，以确保带来流畅的观看体验。

系统规格与效率

在开发和测试 TRAILDREAMS 框架时，我们使用的设备包括一台 HP Omen 16 笔记本电脑，配备 Intel i7 13700HX 处理器、NVIDIA RTX 4080 显卡和 16GB 内存，运行 Windows 11 操作系统。我们使用 Python 3.10 满足编程需求。值得注意的是，尽管我们的框架功能先进，但要高效运行，TRAILDREAMS 并不需要如此高端的平台。我们的案例研究充分体现了这一效率：研究对象《金刚狼》是一部片长为 2 小时 18 分钟的电影。在上述设备上使用我们的框架，仅用 24 分钟便完成了这部电影的预告片生成过程。这表明该框架易于使用，并且即使在性能一般的系统上也能提供良好的表现。

评估与结果

本节将进行比较分析，重点关注 TRAILDREAMS 框架及自动预告片生成领域的领先竞争者。我们的评估基于这些方法生成的预告片，这些预告片可供直接比较。虽然我们希望纳入较新论文生成的预告片，以进行更全面的评估，但由于无法获取这些预告片，因此比较仅限于现有数据。最初纳入比较的竞争者包括：

- VID2TRAILER（Irie 等人，2010年），
- MUVEE——用于视频摘要的商业软件，
- PPBVAM（基于点过程的视觉吸引力模型）（Xu 等，2015年），• MOVIE2TRAILER（Rehusevych 和 Firman，2020年），以及
- RT——原始官方真实预告片。

然而，根据 Rehusevych 和 Firman（2020）报告的结果，VID2TRAILER 和 MUVEE 的性能均明显落后，因此我们将这两种方法排除在当前分析之外，仅将 PPBVAM、MOVIE2TRAILER 和官方真实预告片作为比较对象。

为确保评估不偏不倚，我们采用了严格的方法。我们招募了 16 名电影偏好各异的志愿者（媒体研究专业的本科生和硕士生）来评审预告片。为确保评审公正，所有参与者此前都未看过这些预告片，也不知道各方法的应用顺序。为保持一致，所有预告片均统一为 480x360 分辨率。与采用原始电影原声带的 PPBVAM 和 MOVIE2TRAILER 等其他方法不同，我们的 TRAILDREAMS 框架通过融入自动生成的旁白和原声带脱颖而出。这种方法彰显了我们致力于提供完全自动化、沉浸式视听体验的决心。此外，虽然可以多次运行 TRAILDREAMS 并挑选最佳输出，但我们向志愿者展示的预告片仅由软件运行一次生成。这可能与竞争方法有所不同，因为竞争者可能从其框架生成的结果中挑选了经人工评估后表现最佳的版本。与 Irie 等人（2010）、Xu 等人（2015）以及 Rehusevych 和 Firman（2020）的做法类似，志愿者从恰当性、吸引力和兴趣三个方面评估每部预告片，以确定观看预告片后他们观看原电影的可能性：

- 恰当性：“这段预告片看起来与真正的预告片有多相似？”

- 吸引力：“这段预告片有多吸引人？”
- 兴趣：“看过这段预告片后，你有多大可能会去看原电影？”

他们采用 Likert（1932）量表进行评分，范围从 1（最低）到 7（最高）。分析聚焦于《金刚狼》（2013年）、《霍比特人：史矛革之战》（2013年）和《300勇士：帝国崛起》（2014年）的预告片。选择这些电影，是因为其他研究提供了这些电影的预告片，且复现竞争对手算法的某些组件受到限制。

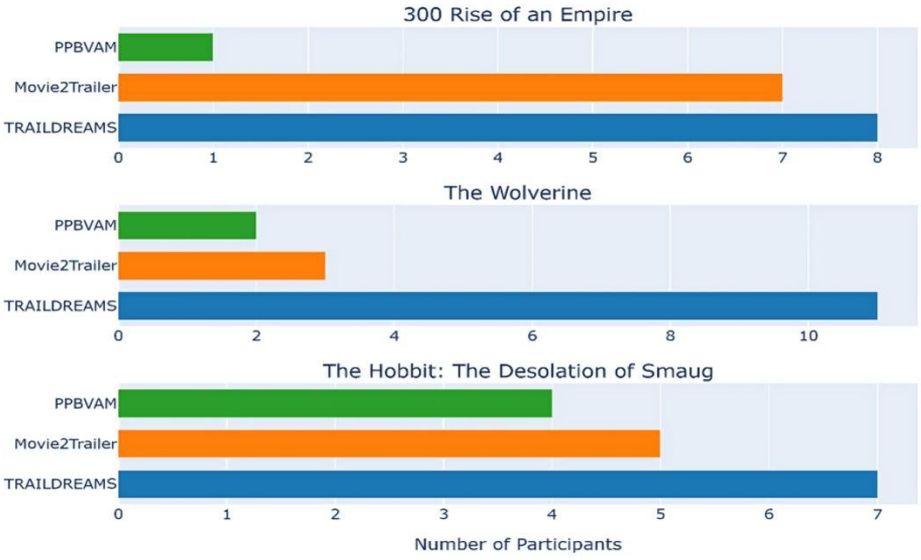

图 ***7***。参与者评选出的最佳电影预告片 *[经 Balestri 等人（2024a）许可转载。© IEEE]*

我们使用总分指标评估了不同预告片生成方法的效果，该指标为三项指标评分之和。对于每位参与者，总分最高的方法被视为效果最佳的方法。**图 7**展示了条形图，显示在所考察的电影中，认为各方法最佳的参与者人数，从而直观比较各方法在不同电影情境下满足观众标准的表现。TRAILDREAMS 始终优于竞争对手，表明其说服观众的能力更强。值得注意的是，TRAILDREAMS 为《金刚狼》生成的预告片获得了较高评分，这或许是因为与另外两部电影相比，为该片生成的配乐和旁白质量更好。

此外，我们计算了各方法的平均得分和中位数得分，结果见**表1**和**表2**。TRAILDREAMS 在所有类别中的表现均优于 PPBVAM，并且在恰当性和兴趣方面优于 MOVIE2TRAILER。吸引力得分略有差异，这主要可能是因为 TRAILDREAMS 使用的人工智能生成的配乐不如竞争者使用的原创人工创作配乐那样优质。尽管如此，TRAILDREAMS 的表现仍值得称赞，尤其是在保持吸引力方面的竞争力，同时在恰当性方面也取得了优异成绩。后者凸显了 TRAILDREAMS 制作符合传统预期的预告片的能力，即电影预告片应具备怎样的视觉效果和观感。

表 *1.* 各方法在三部电影中各项质量指标的平均得分。

平均分	**TRAILDREAMS**	**Movie2Trailer**	**PPBVAM**
恰当性	**3.56**	3.15	2.81
吸引力	2.96	**3.08**	2.88
兴趣	**2.90**	2.85	2.62

表 *2.* 各方法在三部电影中各项质量指标的中位数得分。

中位数得分	**TRAILDREAMS**	**MOVIE2TRAILER**	**PPBVAM**
恰当性	**4.0**	**3.0**	2.0
吸引力	3.0	**3.0**	**3.0**
兴趣	3.0	2.0	2.5

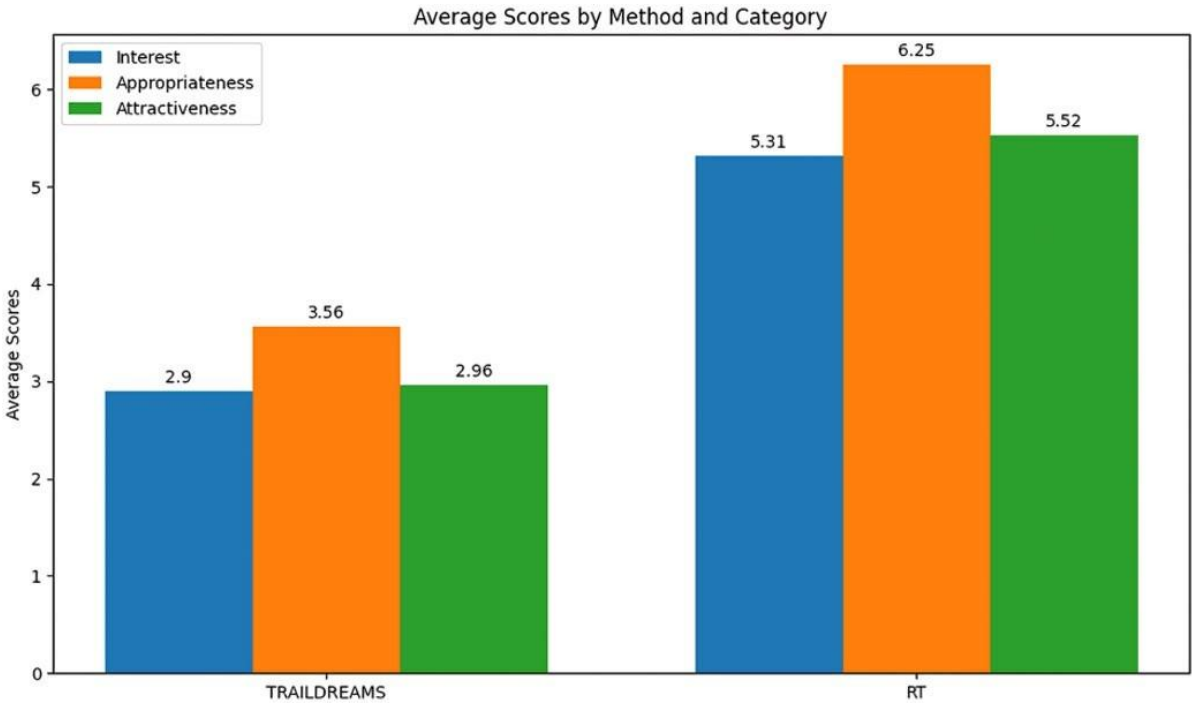

图*8*。 *TRAILDREAMS* 框架生成的预告片与真实预告片的评分（来源：作者自行整理）

分析与未来工作

TRAILDREAMS 框架标志着自动预告片生成领域的一项重大进展，展现出巨大的潜力。然而，要增强其功能，并使其输出更接近人工制作的预告片，还需要进行多项改进。**图8**展示了三部电影中 TRAILDREAMS 生成的预告片与真实预告片的汇总得分差异，凸显了自动化与传统预告片制作方法之间存在的显著差距。

我们进行了深入的定性分析，以辨别我们的预告片与真实预告片之间的主要差异。这项探索帮助我们确定了若干有待改进的关键方面。针对每个已发现的问题，我们都提出了可行的解决方案，旨在提升预告片的整体质量。**标准片段优化**

问题

目前，SC 可能会包含角色正在说话、却没有对应音频的场景，导致观众看到的内容与听到的声音不匹配。

未来工作

集成动作识别深度学习模型有助于识别出人物在特写镜头中明显正在说话的片段。该模型可使系统避免选择这些片段，或者确保选取说话场景时包含相应的对白或旁白。通过更准确地匹配视听元素，这项改进将带来更加连贯、更具吸引力的观看体验。

语音生成改进

问题

旁白　　当前文本转语音模型生成的旁白有时缺乏人类语音的表现力和自然度　　h.

未来工作

随着文本转语音技术的发展，采用更先进的模型将至关重要。未来版本的 TRAILDREAMS 应探索整合新一代文本转语音引擎，以提供更加逼真且富有表现力的语音输出。这项升级将显著提升预告片的听觉吸引力，使其更具感染力，更贴近人类表达。

音乐生成增强

问题

生成音乐有时无法与预告片的情感基调或叙事流完美契合，听感也未必总是令人愉悦。

未来工作

开发更先进的音乐生成模型，将能够创作出更加和谐的配乐，从而提升整体观看体验。

图像检索流程优化

问题

在副情节生成阶段，图像检索过程有时会选取与大语言模型所规划的目标场景不对应的帧。

未来工作

为了提升预告片的视觉连贯性，增强图像检索过程的鲁棒性至关重要。一种可行的解决方案是采用多模态大语言模型，或称“VLLM”（视频大语言模型），进行“二次检查”，以确认所选帧准确呈现了目标叙事片段。

音效的引入

问题

与人工制作的预告片不同，TRAILDREAMS 目前尚未加入音效，而音效对于营造动感且身临其境的预告片体验至关重要。

未来工作

为弥合这一差距，未来版本的 TRAILDREAMS 应加入一个模块，用于在预告片中融入相关音效。这可能需要开发一种智能系统，能够识别预告片中适合通过音效增强叙事或情感冲击的片段。实现这一功能将为预告片增添新的层次感和真实感，使其更接近传统的人工制作的预告片。

结论

本研究介绍了 TRAILDREAMS，这是一个利用大语言模型推动自动预告片生成领域发展的创新框架。TRAILDREAMS 的独特之处在于采用多模态方法，统筹音频、视频、旁白和对白，从而制作出能引发观众深切共鸣的预告片，其叙事丰富度超越了当前自动预告片制作的标准。比较分析进一步印证了我们的研究发现：与竞品相比，TRAILDREAMS 在制作更具吸引力、更恰当、更引人入胜的预告片方面表现更优。该框架的高效性体现在能够快速生成一部完整的《金刚狼》预告片，这也展示了其显著简化预告片制作流程的潜力。

尽管取得了诸多进展，TRAILDREAMS 与所有现有自动化系统一样，在高预算制作方面仍逊于人类创造力；在这一领域，艺术性与技术精湛度的结合定义了行业标准。然而，对于独立电影人和低预算制作而言，TRAILDREAMS 尤其具有价值。在资源匮乏的情况下，TRAILDREAMS 提供了一种可行的选择。

展望未来的改进方向——提升语音生成、优化片段选择、改进图像检索并加入音效——TRAILDREAMS 将进一步缩小自动生成预告片与人工制作预告片之间的差距。

总之，TRAILDREAMS 为自动预告片生成领域树立了新的标杆，展现了人工智能增强、并在某些情况下变革电影产业创作流程的潜力。

数据可用性：本文的源代码以及各种精心设计的提示词，均可在“TRAILDREAMS-framework”（Balestri 等人，2024b）代码库中获取，访问地址为 https://github.com/ robertobalestri/TRAILDREAMS-Framework。由于版权限制，无法直接提供许多电影的生成预告片。不过，可以在 https://youtu.be/O9fS8s2LRqM 观看乔治·A·罗梅罗执导的公有领域电影《活死人之夜》（1968）的预告片。经合理请求，可向通讯作者索取其他生成的预告片。

参考文献

Alberani, D.（2006年）。Cinemagoer [计算机软件]。GitHub。https://cinemagoer.github.io/ Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H.（2019年）。用于文本检测的字符区域感知。载于《2019年IEEE/CVF计算机视觉与模式识别会议论文集》（第9357–9366页）。IEEE。https://doi.org/10.1109/CVPR.2019.00959 Balestri, R., Cascarano, P., Degli Esposti, M., & Pescatore, G.（2024a）。一种通过大语言模型生成预告片的自动深度学习方法。载于《2024年第9届信号处理前沿国际会议》（ICFSP），法国巴黎（第93–100页）。https://doi.org/10.1109/ICFSP62546.2024.10785516 Balestri, R., Cascarano, P., Degli Esposti, M., & Pescatore, G.（2024b）。TRAILDREAMS-framework [计算机软件]。GitHub。https://github.com/robertobalestri/TRAILDREAMS-Framework Bellard, F.（2000年）。FFmpeg [计算机软件]。https://ffmpeg.org/ Brachmann, C., Chunpir, H. I., Gennies, S., Haller, B., Kehl, P., Mochtarram, A. P., Möhlmann, D., Schrumpf, C., Schultz, C., Stolper, B., Walther-Franks, B., Jacobs, A., Hermes, T., & Herzog, O.（2009）。基于语义视频模式的自动电影预告片生成。载于 I. Maglogiannis、V. Plagianakos 和 I. Vlahavas（编），《人工智能：理论与应用。SETN 2012》。计算机科学讲义（第7297卷，第345–352页）。Springer。https://doi.org/10.1007/978-3-642-30448-4_44Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, Di., Titeux, H., Bouaziz, W., & Gill, M. P.（2020年）。Pyannote.Audio：用于说话人分离的神经网络构件。载于《ICASSP 2020–2020年IEEE声学、语音与信号处理国际会议论文集》（第7124–7128页）。IEEE。https://doi.org/10.1109/ICASSP40776.2020.9052974 Castellano, B.（2014年）。PySceneDetect [计算机软件]。https://www.scenedetect.com/ Copet, J., Kreuk, F., Gat. Itai, Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A.（2024年）。简单且可控的音乐生成。arXiv。https://doi.org/10.48550/arXiv.2306.05284 Coqui AI。（2021年）。XTTS [计算机软件]。https://docs.coqui.ai/en/latest/models/xtts.html De Palma, B.（导演）。（1996年）。《碟中谍》[电影]。派拉蒙影业。Degli Esposti, M., & Pescatore, G.（2023年）。通过数据驱动方法探索电视连续剧特性与电视研究。载于《媒体变革国际会议论文集》。https://doi.org/10.21428/93b7ef64.ec022085 Epstein, Z., Hertzmann, A., Akten, M., Farid, H., Fjeld, J., Frank, M. R., Groh, M., Herman, L., Leach, N., Mahari, R., Pentland, A., Russakovsky, O., Schroeder, H., & Smith, A.（2023年）。生成式人工智能的艺术与科学。《科学》，380(6650)，1110–1111。https://doi.org/10.1126/science.adh4451 Explosion AI。（2016年）。spaCy英语模型 [计算机软件]。https://spacy.io/models/en Gallifant, J., Fiske, A., Levites Strekalova, Y. A., Osorio-Valencia, J. S., Parke, R., Mwavu, R., Martinez, N., Gichoya, J. W., Ghassemi, M., Demner-Fushman, D., McCoy, L. G., Celi, L. A., & Pierce, R.（2024年）。对 GPT–4 技术报告和系统卡的同行评审。《PLOS数字健康》，3(1)，文章编号 e0000417。https://doi.org/10.1371/ journal.pdig.0000417 Hesham, M., Hani, B., Fouad, N., & Amer, E.（2018年）。智能预告片：仅使用字幕自动生成电影预告片。载于《深度与表征学习国际研讨会论文集》（第26–30页）。https://doi.org/10.1109/IWDRL.2018.8358211 Hu, Y., Jin, L., & Jiang, X.（2022年）。用于生成预告片的基于GCN的框架。载于《计算与人工智能国际会议论文集》（第610–617页）。https://doi.org/10.1145/ 3532213.3532306 Irie, G., Satou, T., Kojima, A., Yamasaki, T., & Aizawa, K.（2010年）。自动预告片生成。载于《ACM国际多媒体会议论文集》（第839–842页）。ACM。https://doi.org/10.1145/ 1873951.1874092 Jackson, P.（导演）。（2012年）。《霍比特人：意外之旅》[电影]。华纳兄弟。JaidedAI。（2023年）。EASYOCR [计算机软件]。GitHub。https://github.com/JaidedAI/EASYOCR jianfch。（2023年）。Stable Whisper [计算机软件]。GitHub。https://github.com/jianfch/stable-ts Likert, R.（1932年）。一种态度测量技术。《心理学档案》，22(140)，文章编号55。Long, F., Qiu, Z., Yao, T., & Mei, T.（2024年）。VideoDrafter：基于大语言模型的内容一致多场景视频生成。arXiv。https://doi.org/10.48550/arXiv.2401.01256 Mahasseni, B., Lam, M., & Todorovic, S.（2017年）。使用对抗式长短期记忆网络进行无监督视频摘要。载于《2017年IEEE计算机视觉与模式识别会议论文集》（第2982–2991页）。IEEE。https://doi.org/10.1109/CVPR.2017.318 Marhon, S. A., Cameron, C. J. F., & Kremer, S. C.（2013年）。循环神经网络。载于 M. Bianchini、M. Maggini 和 L. Jain（编），《神经信息处理手册》。智能系统参考文库（第49卷，第29–65页）。Springer。https://doi.org/10.1007/978-3-642-36657-4_2 Nolan, C.（导演）。（2014年）。《星际穿越》[电影]。派拉蒙影业。

Oliveira, D.（2024年1月7日）。用人工智能创作电影预告片。Towards AI。
https://towardsai.net/p/machinelearning/creating- movie-trailers-with-ai OpenAI.（2023a）。OpenAI–GPT-4。OpenAI。https://openai.com/gpt-4 OpenAI.（2023b）。Whisper [计算机软件]。GitHub。
https://github.com/openai/whisper Papalampidi, P.、Keller, F.与Lapata, M.（2021年）。通过任务分解进行电影预告片生成。arXiv。https://doi.org/10.48550/arXiv.2111.08774 Pavel, A.、Reed, C.、Hartmann, B.与Agrawala, M.（2014年）。视频摘要：面向信息性讲座视频的可浏览、可略读格式。载于*27th* 年度ACM用户界面软件与技术研讨会论文集（第573–582页）。ACM。https://doi.org/10.1145/2642918.2647400 Piccolomini, E. L.、Gandolfi, S.、Poluzzi, L.、Tavasci, L.、Cascarano, P.与Pascucci, A.（2019年）。将循环神经网络应用于GNSS时间序列以进行去噪与预测。载于*26th* 时间表示与推理国际研讨会论文集。
https://doi.org/10.4230/LIPIcs.TIME.2019.10 Radford, A.、Kim, J. W.、Hallacy, C.、Ramesh, A.、Goh, G.、Agarwal, S.、Sastry, G.、Askell, A.、Mishkin, P.、Clark, J.、Krueger, G.与Sutskever, I.（2021年）。从自然语言监督学习可迁移的视觉模型 [预印本]。arXiv。https://arxiv.org/abs/2103.00020 Radford, A.、Kim, J. W.、Hallacy, C.、Ramesh, A.、Goh, G.、Agarwal, S.、Sastry, G.、Askell, A.、Mishkin, P.、Clark, J.、Krueger, G.与Sutskever, I.（2022年）。clip-ViT-L-*14* [计算机软件]。
https://huggingface.co/sentencetransformers/clip-ViT-L-14 Ratcliff, J. W.与Metzener, D.（1988年）。模式匹配：格式塔方法。《Dr. Dobb's Journal》，13，第46篇。Rehusevych, O.与Firman, T.（2020年）。
movie2trailer：使用异常检测的无监督预告片生成。载于D. Tabernik、A. Lukezic与K. Grm（编），*25th* 计算机视觉冬季研讨会论文集。Reimers, N.与Gurevych, I.（2019年）。Sentence-BERT：使用孪生BERT网络的句子嵌入。载于EMNLP-IJCNLP 2019年–2019年自然语言处理实证方法会议暨*9th* 自然语言处理国际联合会议论文集。
https://doi.org/ 10.18653/v1/d19-1410 Richards, G.（2018年3月14日）。深入探讨：过去十年电影预告片发生了怎样的变化？Exit6电影节博客。
h
t
t
p
s://
www.exit6filmfestival.com/post/2018/03/14/going-in-deep-how-have-movietrailers-changed-in-the-last-decade Rouard, S.、Massa, F.与Défossez, A.（2023年）。用于音乐源分离的混合Transformer。载于ICASSP 2023年 - 2023年IEEE声学、语音与信号处理国际会议（ICASSP），希腊罗得岛（第1–5页）。
https://doi.org/10.1109/ICASSP49357.2023.10096956 Sentence Transformers.（2022年）。clip-ViT-L-14。
https://huggingface.co/sentence-transformers/clip-ViT-L-14 Shi, B.、Bai, X.与Yao, C.（2017年）。一种用于基于图像的序列识别的端到端可训练神经网络及其在场景文本识别中的应用。IEEE模式分析与机器智能汇刊，39（11），2298–2304。https://doi.org/10.1109/TPAMI.2016.2646371 Smeaton, A. F.、Lehane, B.、O'Connor, N. E.、Brady, C.与Craig, G.（2006年）。为动作电影预告片自动选择镜头。载于 *8th* ACM多媒体信息检索国际研讨会论文集（第231–238页）。ACM。https://doi.org/10.1145/1178677.1178709 Smith, J. R.、Joshi, D.、Huet, B.、Hsu, W.与Cota, J.（2017年）。利用人工智能增强创造力：应用于电影预告片创作。载于 *25th* ACM国际多媒体会议论文集（第1799–1808页）。ACM。https://doi.org/10.1145/3123266.3127906 Snyder, B.（2005年）。《救猫咪！：你唯一需要的电影剧本创作之书》。Michael Wiese Productions。Tarwani, K. M.与Edem, S.（2017年）。自然语言处理中循环神经网络研究综述。《国际工程趋势与技术期刊》，48（6），301–304。
https://doi.org/10.14445/ 22315381/IJETT-V48P253 Wasko, J.（2003年）。《好莱坞如何运作》。SAGE。
https://doi.org/10.4135/9781446220214 Xie, J.、Chen, X.、Zhang, T.、Zhang, Y.、Lu, S.-P.、Cesar, P.与Yang, Y.（2023年）。基于多模态且以美学为导向的叙事视频摘要。IEEE多媒体汇刊，25，4894–4908。
https://doi.org/ 10.1109/TMM.2022.3183394 Xu, H.、Zhen, Y.与Zha, H.（2015年）。通过基于点过程的视觉吸引力模型进行预告片生成。载于第24届人工智能国际会议（IJCAI'15）论文集（第2198–2204页）。AAAI Press。
https://dl.acm.org/doi/10.5555/2832415.2832554 Zen, H.、Dang, V.、Clark, R.、Zhang, Y.、Weiss, R. J.、Jia, Y.、Chen, Z.与Wu, Y.（2019年）。LibriTTS：一个源自LibriSpeech、用于文本转语音的语料库。载于国际语音通信协会年会论文集。https://doi.org/10.21437/Interspeech.2019-2441

Zhou, H.、Hermans, T.、Karandikar, A. V.与Rehg, J. M.（2010年）。通过场景分类进行电影类型分类。载于论文集 *18th ACM*国际多媒体会议（第747–750页）。ACM。https://doi.org/10.1145/1873951.1874068 Zhou, K.、Qiao, Y.与Xiang, T.（2018年）。面向无监督视频摘要、带有多样性—代表性奖励的深度强化学习。载于论文集 *32nd AAAI*人工智能会议。AAAI。https://doi.org/10.1609/aaai.v32i1.12255 Zhu, J.、Yang, H.、He, H.、Wang, W.、Tuo, Z.、Cheng, W.-H.、Gao, L.、Song, J.与Fu, J.（2023年）。MovieFactory：使用面向语言与图像的大型生成模型，从文本进行自动电影创作。载于论文集 *31st ACM*国际多媒体会议（第9313–9319页）。ACM。https://doi.org/10.1145/3581783.3612707