基于人工智能的文本和声音转教育视频系统

M. E. ElAlami[1]，*，S. M. Khater[2], **M. El. R. Rehan**[3]1埃及曼苏拉大学特殊教育学院计算机与信息系统教授2埃及曼苏拉大学特殊教育学院计算机教师培养系讲师3埃及曼苏拉大学特殊教育学院计算机教师培养系助教。电子邮箱：moh_elalmi@mans.edu.eg；shim_khater@mans.edu.eg；mohamedrehan@mans.edu.eg

摘要

技术发展催生了能够根据输入文本或声音生成教育视频的方法。近年来，深度学习技术在图像与视频生成中的应用受到广泛探索，尤其是在教育领域。然而，根据文本或语音等条件输入生成视频内容，仍然是一项颇具挑战的工作。本文提出一种适用于教育领域的新方法，即生成对抗网络（GAN），该方法构建逐帧框架，能够生成完整的教育视频。所提出的系统分为三个主要阶段：第一阶段通过语音识别将输入内容（文本或语音）转录为文字；第二阶段提取关键词，并使用 CLIP 和扩散模型等先进模型生成相关图像，以提升视觉质量和语义对齐效果；最后，将生成图像合成为视频格式，并配入预录或合成音频，从而生成完整的交互式教育视频。本文将所提出的系统与 TGAN、MoCoGAN 和 TGANS-C 等其他系统进行比较，取得了 28.75% 的弗雷歇特征距离（FID）得分，表明其视觉质量有所提升，且优于现有方法。

关键词 ： 智能系统；深度学习；生成对抗网络；文本到视频生成 n

1. 引言

视觉媒体，尤其是视频，在现代教育中发挥着核心作用，能够提升各学科领域学习者的参与度和理解力。研究证实，制作精良的教育视频可以简化复杂主题、帮助巩固记忆，并满足不同的学习风格 [1]。与此同时，人工智能（AI）正通过实现个性化学习、自动化教学任务并提供分析洞见，推动教育领域变革。人工智能的数据处理能力有助于打造更具适应性和成效的教育实践[2]。多模态学习的兴起，通过融合文本、音频、图像和视频，进一步丰富了教育体验。人工智能驱动的多模态学习分析（MMLA）有助于深入理解学习者的互动情况，为个性化反馈和学习设计提供支持。[3]。

生成式人工智能，尤其是多模态大语言模型（MLLMs），如今能够创作多种教育内容，包括动态内容。这些模型提供了可扩展、个性化且引人入胜的资源，显著推动了多模态教学方法的发展 [4]。尽管生成式人工智能具备诸多能力，要制作出高质量且能无缝融合文本与动态视觉效果的教育视频，在技术上仍颇具挑战，也十分耗时。传统制作方法需要丰富的专业知识和大量资源，限制了其在教育场景中的广泛应用 [5]。为应对人工制作带来的挑战并充分发挥生成式人工智能的能力，人们对开发能够自动完成教育视频内容创作流程的智能系统日益关注 [6]。

数字对象标识符：https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接受日期：2025年7月10日

本文旨在设计并开发一种创新型智能系统，能够自动将文本和听觉输入转换为一体化教育视频。本研究的主要贡献在于详细阐述系统架构和智能算法，以应对从多模态来源（文本和声音）生成视频所面临的挑战，并着重制作在教学上连贯、视觉上引人入胜的内容。本研究针对当前自动化教育视频制作工具的不足提出解决方案，助力教育工作者和开发者高效地制作高质量视觉学习材料，从而推动视频在教育中的应用不断提升并扩大规模。

2. 相关工作

Yan 等人（2024年）[7] 提出了一种视频合成时间表示的新方法，并开发了基于神经表示的连续视频生成器。他们利用位置嵌入构建运动表示，并探索了视频生成器的稀疏训练策略。他们重新设计了一种双结构视频判别器。该模型以 StyleGAN2 为基础，具有训练高效、图像质量出色以及潜在空间可调等优势。研究指出，这项工作为未来的视频生成发展奠定了坚实基础。

JayZhangjieWu 等人（2024年）[8] 提出了一种用于评估文本生成视频技术的独特评估指标。该指标由两部分组成：T2VScore-A 用于衡量视频与文本之间的一致性，T2VScore-Q 则根据技术和语义指标比较视觉质量。作者还创建了 TVGE 数据集，其中包含 2,500 多个人类评分的文本生成视频样本。实验结果表明，在与人类判断的相关性方面，T2VScore 显著优于现有指标，为评估人工智能生成的视频提供了一种更可靠、更全面的方法。

Mao 等人（2024年）[9] 提出了一种名为文本生成有声视频（TAVG）的新系统，可根据纯文本描述同步生成视频和音频。为此，研究人员利用 ChatGPT 工具及其增强功能，提供了一个全新的数据集，其中包含 170 多万个视频，并附有准确的自动音频和图像描述。他们还提出了一种用于衡量音视频兼容性的新指标。为提供标准模型，他们提出了一种生成模型，采用潜在扩散、对比学习和交叉注意力技术，确保输出内容同步。实验表明，该模型在内容质量和一致性方面优于传统方法，为多模态生成开辟了新途径。

Yue Ma 等人（2023年）[10] 提出了利用无姿态视频进行姿态引导的文本生成视频方法，以满足日益增长的需求：生成可通过文本编辑且姿态可控的角色视频。这是创建多样化数字人的关键环节。长期以来，这项任务受到两方面限制：缺乏包含成对视频姿态描述的全面数据集，以及缺少适用于视频合成的生成式先验模型。为克服这些限制，他们提出了一种新颖的两阶段训练方案，结合使用易于获取的数据集（如图像姿态对和无姿态视频）以及预训练的文本到图像（T2I）模型，从而有效生成姿态可控的角色视频。

Zhu 等人（2023年）[11] 提出了《文本图像到视频生成：基于静态图像和文本描述的可控视频合成》，旨在解决生成符合用户意图的可控视频这一挑战——这项任务在计算机视觉领域极具吸引力，但也十分复杂。他们提出了一种独特的任务，称为文本-图像到视频生成（TI2V），以便对外观和运动进行精细控制。该方法能够根据静态图像与文本描述的组合生成视频，灵活控制合成视频中的视觉元素和动态运动，从而满足用户的特定需求。

Singer 等人（2022年）[12] 提出了 Make-A-Video，这是一个将文本到图像（T2I）生成扩展至文本到视频（T2V）的系统。该系统从文本-图像对中学习视觉外观，并从无监督视频数据中学习运动。该模型加快了训练速度，无需配对的文本-视频数据，同时继承了多样化的图像生成能力。系统采用了包括分解式时间 U-Net 和注意力张量在内的时空模块，并通过时空流水线生成高分辨率、高帧率视频。该方法在文本到视频生成方面创下了新的最先进水平，并得到了定性和定量结果的共同验证。

在审阅以往相关研究后，发现现有系统存在若干局限，包括文本输入与音频输出缺乏整合、处理复杂教育内容的能力有限、生成的视频为静态或时长较短且语义对齐效果不佳，以及视频质量欠佳。为克服这些挑战，本文提出一种新型系统——文本到视频对抗视频生成网络。该系统扩展了传统的文本到视频对抗模型，使其能够根据文本输入生成动画教育视频。系统采用文本编码、使用 YAKE 算法进行关键词提取等先进技术，同时将视觉元素与预定义声音数据库中的适当音频片段关联起来。其目标是生成语义丰富、连贯且高质量的教育视频，使其与输入文本紧密匹配。

数字对象标识符： https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接受日期：2025年7月10日

3. 所提出的系统

本研究采用深度学习将文本或声音转换为教育视频。该系统由三个阶段组成。第一阶段，用户提供一个输入句子，可以是文本或音频形式。随后，系统将声音转换为文本。系统使用来自变换器的双向编码器表示（BERT）模型，从输入内容中提取重要术语和领域专用术语。对于每个提取出的术语，系统使用生成对抗网络技术从图像数据库中生成相应图像。第二阶段，系统根据用户设定的顺序，将每个术语对应的选定图像组合起来，生成一段由这些图像构成的无声视频。第三阶段，系统为每个术语选择适当的音频文件，然后将音频文件与视频合并。最后，视频会显示出来，用户也可以将其下载到自己的设备上。下一段将详细介绍所提出的方法。

用户通过文本框或麦克风输入基本词语或句子。语音借助 Google Assistant 转换为文本，随后以文本形式显示在文本框中。之后，系统会在图像数据集中搜索与该句子或词语相关的图像。每个关键词都对应一组已存储的图像。图像由向量量子生成对抗网络（VQGAN）生成。与传统生成对抗网络相比，该过程有助于生成结构更清晰、边缘更锐利的图像。系统使用 CLIP模型根据相似性和语义对齐来匹配图像与文本。此外，系统还使用扩散模型，通过去噪、细化细节和应用视觉滤镜来提升图像质量。处理后的图像会被编排并渲染为连贯的视频序列，随后将音轨整合到视频中，从而营造交互式且沉浸式的多媒体体验。所提出系统的框架如图（1）所示。

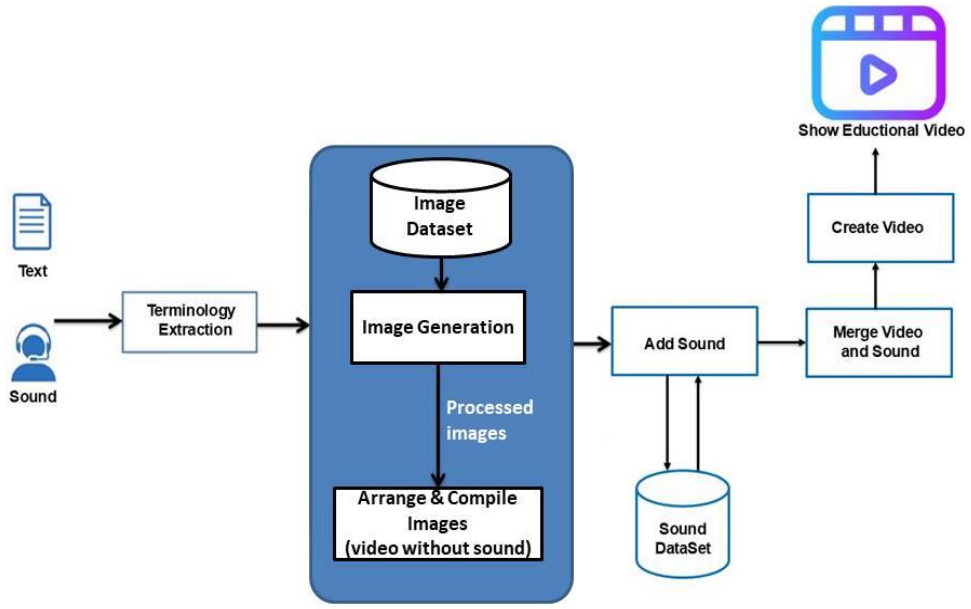


图 1. 所提出的系统框架

生成对抗网络生成清晰图像的能力已得到证实。从这个角度来看，如果我们从文本到图像阶段开始训练文本到视频网络，就有望更高效地生成视频。基于这些概念，我们将训练过程分为两个阶段：文本到图像生成和进化生成。下面介绍所提出架构的整体流程。我们首先学习如何从文本生成单张图像，然后逐步增加生成的图像数量。该训练过程会持续进行，直至达到所需的视频时长。这就是我们采用的主要范式。在详细介绍这两个阶段之后，我们将在各小节中说明用于稳定学习的方法。图（2）展示了所提出系统的流程图。

数字对象标识符：https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接收日期：2025年7月10日

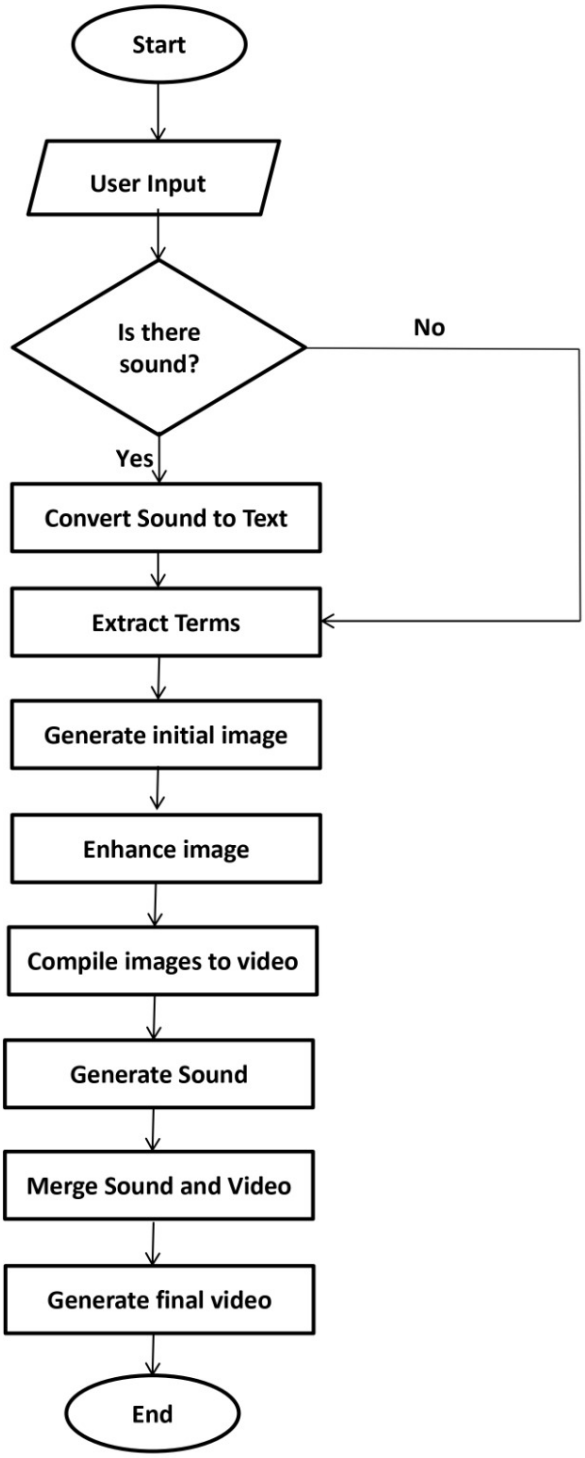


图2. 所提出的系统的流程图

以下算法概述了所提出的教育视频生成系统的系统化步骤。该系统集成了先进的人工智能模型，例如用于文本分析的 BERT、用于图像生成的 VQGAN 和扩散模型，以及用于语义对齐的 CLIP。其目标是自动将用户输入的文本或音频转换为引人入胜且语义准确的教育视频。表1展示了该系统的系统化工作流程。

数字对象标识符： https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接收日期：2025年7月10日

表 **1**：教育视频生成系统的拟议算法

步骤编号		步骤功能
		初始化
步骤 1	I.	使用输入文本框和麦克风按钮初始化图形用户界面。
	II.	加载预训练模型：BERT、VQGAN、CLIP、扩散模型。
	III.	准备图像和视频数据的存储路径。
		输入获取
步骤 2	I.	如果用户输入音频，请使用 Google Assistant 或 Whisper 将其转换为文本。
	II.	如果用户输入文本，则直接将其传递给系统。
	III.	在输入框中显示生成的文本。
		关键词提取
步骤 3	I.	使用 BERT 从输入文本中提取关键领域专用术语。
	II.	将所有关键词存入列表，以便进一步处理。
		图像生成•
		对于每个关键词：
第4步	I.	使用 VQGAN模型生成图像。
	II.	使用 CLIP模型从语义上匹配生成图像与文本。
	III.	应用扩散模型提升图像质量。
	IV.	按顺序存储每个术语对应的最终图像。
		视频合成
步骤 5	I.	根据关键词顺序或用户输入，依次排列所选图像。
	II.	将图像编译成无声视频流。
		音频集成
步骤 6	I.	检索或生成与提取的术语相对应的音轨。
	II.	将音频与视频合并，制作最终的教育视频。
		视频输出
步骤 7	I.	在图形用户界面上显示最终视频。
	II.	提供一个按钮，供用户下载生成的视频。
	III.	流程结束。

数据集来源：

所提出的系统使用公开可用的数据集。对于图像生成，我们使用来自 Kaggle 的数据集，例如 Common Objects in Context（COCO），该数据集提供经过分类的图像用于训练。对于语音处理，我们采用 LibriSpeech 数据集（可在 Kaggle 上获取），其中包含 1,000 小时的英语音频。

算法 1：训练流程

(I)数据：

• 输入文本：t

• 对应的真实视频帧：V_R = (X1, X2, ..., X_{2n}) 网络组件：

数字对象标识符：https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接收日期：2025年7月10日

• 生成器：G

• 门控循环单元（GRU）：R

• 判别器：D_I、D_{S1}、D_{S2}、...、D_{Sn}

（II）文本到图像生成阶段

目标：根据文本输入生成单张图像，并训练网络 G、R 和图像判别器 D_I。

当尚未收敛时，循环执行

获取 z0、z1 ∈ (0, 1) 生成 I1 = G(R(z0, (φ_t, z1))) 从 V_R 中随机选择一张真实图像 X 根据公式 3 更新 G、R，以最小化损失 使用独立样本配对，根据公式 3 更新 D_I，以最大化损失 结束

（III）进化生成阶段

目标：通过进化过程迭代生成图像序列，并改善空间和时间一致性。

对于 m ⇒ 1 到 n，执行 当尚未收敛时，循环执行 获取 z1、...、z_{2^m} ∈ $\mathcal{N}$ (0, 1) 通过重复执行 R 和 G 生成 I1、...、I_{2^m} 从 V_R 中随机选择连续的 2^m 个真实帧 X1、...、X_{2^m} 根据以下损失函数更新 G、R，以最小化损失：- 图像级判别损失函数 - 时间序列级判别损失函数 使用独立样本配对，根据公式 3 更新 D_I，以最大化损失 根据公式 4 更新 D_{S_m}，以最大化损失 结束 结束

其中：

- t：输入文本
- V_R：真实视频帧序列
- X_i：真实视频中的第 i 帧
- G：生成器网络
- R：用于处理编码文本的 GRU 网络
- D_I：用于评估单幅图像质量的判别器
- D_{S_m}：用于评估 m 级时间序列一致性的判别器
- z：从正态分布 $\mathcal{N}$(0,1) 中采样的噪声向量
- φ_t：文本 t 的编码表示
- I_i：生成图像

数字对象标识符： https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接受日期：2025年7月10日

训练流程

首先，使用图像判别器 D1 初始化新的步判别器 DS_1。所有步判别器都必须具有相同的架构，唯一的区别在于其第一层的输入通道数量。具体而言，DS_m 以 2^m 个图像作为输入，而 DS_{m-1} 则接收 2^{m-1} 个图像。因此，这些步判别器的初始化过程可描述如下：

在我们的实践中，生成器和判别器均采用了极为特殊的错误训练架构。然而，与普通的生成对抗网络不同，判别器的两个分支之间的顺序进行了微小调整。在此，判别器经过多个转换模块，以实现高水平的函数映射。其中一个分支通过在 Sankranti 上进行 Sankranti，计算文本图像的纹理壁纸匹配；另一个分支则在没有文本求解器的情况下进行判别。这两个独立分支的运作提升了图像质量，并使图像与课程精准匹配。课程匹配模型采用三种损失进行训练，与 Reed 等人的方法类似。真实对（x, φt）、伪造对（I, φt）和错误对（x, t），其中 φt 是不同于 φt 的特定文本向量。

由 G、R、D 和 D_S 生成的惩罚项是对以下内容的总体描述：

L_obj(G, R, D_I, D_S) = L_I(G, R, D_I) + L_S(G, R, D_S) **(1)**

该公式结合了两个组成部分：

- L_I(G, R, D_I)：用于单个图像级判别的对抗损失。

- L_S(G, R, D_S)：用于序列级或视频片段判别的对抗损失。

图像级损失：L_I(G, R, D_I)

L_I(G, R, D_I) = Σ_i [log(D] _ _ I(X^i)) + log(DI(X^i, φ_i)) + log(1 - D_I(I^i)) + log(1 - D_I(I^i, φ_i)) + log(1 - D_I(X^i, φ̂_i)) (2)

此函数用于衡量判别器 D_I 区分真实图像及其正确描述与生成图像或不匹配图像的能力。

其中：

- X^i：一张真实图像。

- I^i：一张生成图像。-

φ_i：匹配文本或条件特征。
- φ̂ _i：不匹配或无关特征。- D_I：
图像判别器。

序列级损失：L_S(G, R, D_S)

L_S(G, R, D_S) = Σ_i [log(D_S(X^i,S)) + log(D_S(X^i,S, φ_i)) + log(1 - D_S(I^i,S)) + log(1 - D_S(I^i,S, φ_i)) + log(1 - D_S(X^i,S, φ̂_i))] (3)

其中：

- X^i,S：真实视频/图像序列。

- I^i,S：生成视频/图像序列。

- D_S：序列（或视频）判别器。

该损失函数不仅确保单帧具有真实感，还能保留序列间的时间一致性和语义对齐。

4. 应用与结果

所提出的系统使用 HTML、CSS、PHP、MySQL 和 Node.js 设计。该系统采用生成对抗网络算法和 TiVGAN方法，根据文本或声音输入创建教程视频。系统还设有独立的图像数据库和声音文件数据库，通过提供相关的视觉内容和声音内容，以及系统内部的处理流程，来改进生成过程。系统根据提取出的术语从数据库中提取图像。每个术语都对应一组图像，然后从该术语显示的图像中进行选择。

视频由这些选定的图像生成。随后，视频与从声音数据库中选取的声音文件进行整合。用户选择文件中的每个术语，最后生成完整的视频。运行

数字对象标识符： https://doi.org/10.54216/FPA.210115收稿：2025年2月25日 修订：
2025年5月31日 接收：2025年7月10日

在计算机上运行所提出的系统，需要安装 XAMPP 软件以将其转换为本地服务器，并需要一些 Node.js 库。

所提出的系统应用于配置了 16 GB RAM、主频为 2.70 GHz 的英特尔酷睿 i7-6820HQ 处理器，以及 4 GB 显存独立显卡的计算设备。系统运行于 64 位操作系统架构。这些规格足以高效执行所提出架构中涉及的视频生成、图像处理和多媒体同步等计算功能。

图3展示了用户输入文本或音频的界面。点击“从输入句子中提取术语”按钮后，术语将显示在表格中。第一列包含术语，第二列显示相应的音频文件，第三列则显示每个术语对应的图像。随后，用户使用 TiVGAN 技术从显示的图像中选择所需图像。点击“创建视频”按钮后，视频创建过程便会开始。

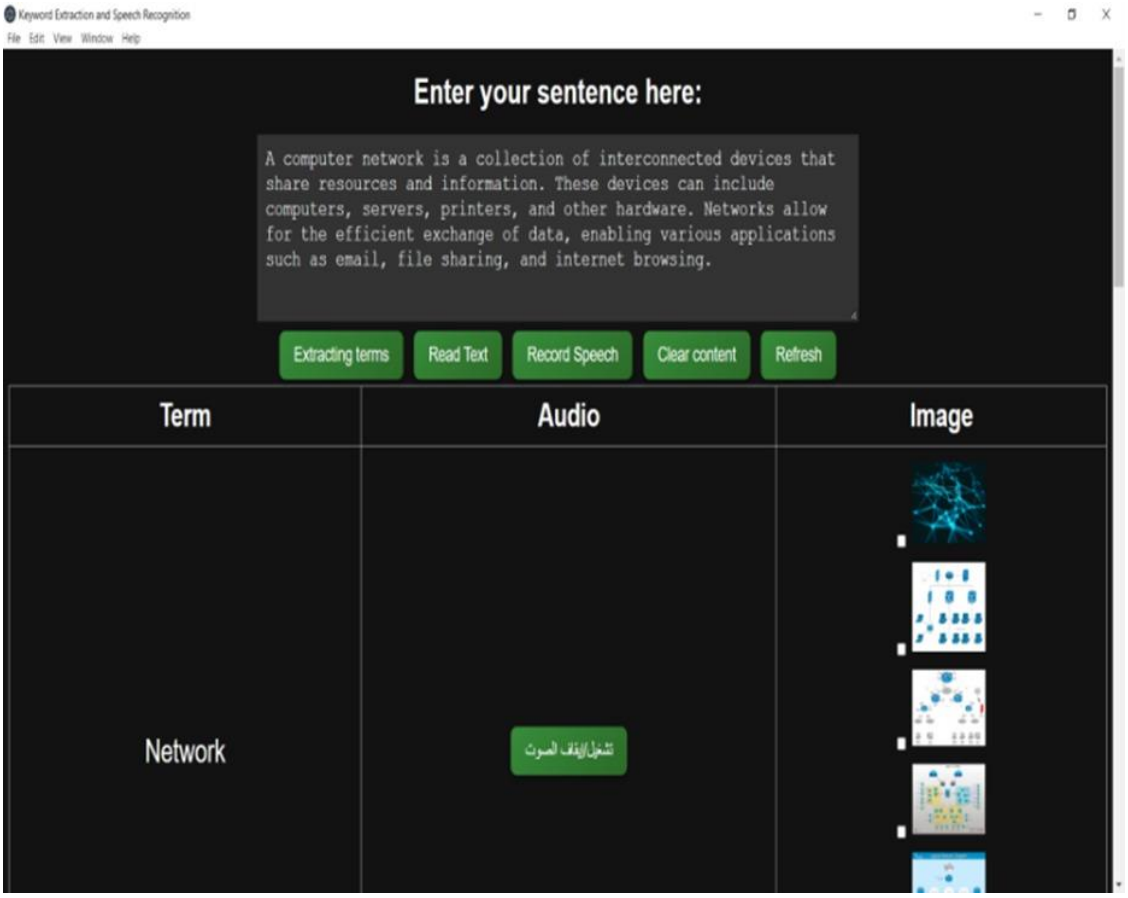


图 3. 从句子中提取术语的图形用户界面

图 4 展示了执行图像整合后的结果：从图像数据库中为每个术语选择相应的图像后，系统会按照所示格式生成视频，且视频不含声音。在下一阶段，按下音频合并按钮后会打开一个窗口，其中显示每个术语对应的音频文件。合并这些音频文件后，再将无音频的视频与合并后的音频文件进行合并。

为基于 FFmpeg 的混音技术提取的每个术语，从私有音频数据库中选取音频文件混音后，使用由 Node.js 代码控制的外部工具将视频和音频文件合并，而不对视频进行重新编码

数字对象标识符：https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接收日期：2025年7月10日

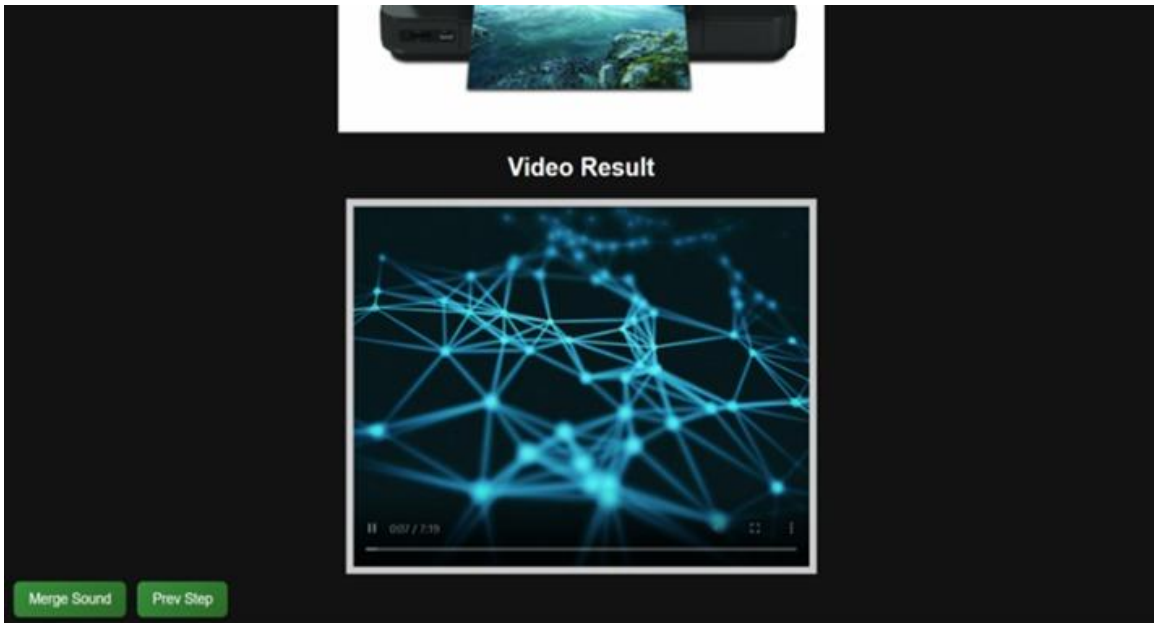


图 **4**。图形用户界面通过 TiVGAN 技术展示生成的视频。

视频的最终形式体现了通过 FFMPEG 库将无声视频与音频文件进行整合后的效果。通过此界面，可以在系统中查看视频，并将其下载到用户的电脑上，如图（5）所示。

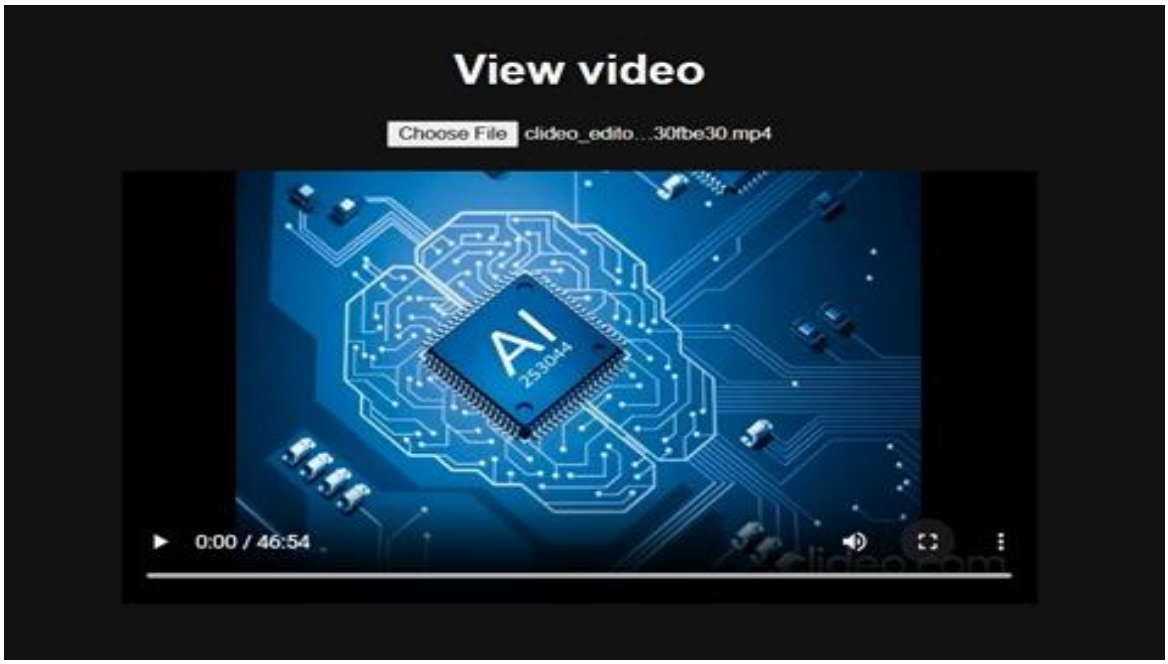


图 **5**。图形用户界面展示了使用 TiVGAN 将无声画面与音频结合后生成的最终视频。

定量结果：

在定量评估中，使用弗雷歇起始距离（FID）衡量生成图像的质量，如公式 4 所示。FID 用于衡量两个图像集之间的相似度。每种方法各收集六帧视频帧，然后计算每个图像集中相同数量的视频帧与训练数据集之间的 FID。结果如表 2 所示，其中展示了最佳性能。

$$FID = \|\mu r - \mu g\|2 + Tr(\Sigma r + \Sigma g - 2(\Sigma r \Sigma g)^{1/2}) \quad (4)$$

数字对象标识符： https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接受日期：2025年7月10日

其中：

• μr：从真实图像中提取的特征的均值向量。• μg：从生成图像中提取的特征的均值向量。• Σr：真实图像特征的协方差矩阵。• Σg：生成图像特征的协方差矩阵。• Tr：迹算子，用于对矩阵的对角元素求和。• (ΣrΣg)1′2：Σr 与 Σg 的乘积的矩阵平方根。

为计算弗雷歇起始距离（FID），首先使用预训练的 Inception V3 网络从图像或视频帧中提取特征。然后，分别计算真实图像和生成图像的图像分布。最后，使用标准 FID 公式计算这两种分布之间的弗雷歇距离，从而根据分布相似度定量衡量真实图像与生成图像之间的相似性。

表**2** ：不同模型的 FID（弗雷歇特征距离）评估

	无文本或声音 有文本	有文本和声音	
TGAN **[13]**	70.10	59.59	55.30
MoCoGAN **[14]**	83.07	81.85	78.45
TGANS-C **[15]**	75.50	69.92	65.80
所提出的系统	40.25	35.10	28.75

所提出的系统显然表现最佳，在所有情况下都取得了最低的 FID，尤其是在使用文本和声音时 d. FID 越低，生成视频的质量越好，也越接近现实。

加入声音对大多数模型都有帮助，但对于所提出的系统而言，这成为一项优势，使其表现尤为出色。MoCoGAN 对声音的利用较少，因为该模型无法从附加信息中获得多少收益。所提出的系统在同时使用文本和声音时表现出色，这意味着它能恰当地利用附加信息来提升视频质量。

为评估所提出系统的有效性，我们将其与四种最先进的视频生成模型进行了比较：TGAN、MoCoGAN、TGANS-C 和所提出的系统。评估基于定量性能指标（例如 FID 分数或用户满意度），并在三种不同的输入配置下进行：不使用文本或声音、仅使用文本，以及同时使用文本和声音。

结果表明，在全部三种输入条件下，所提出的系统都显著优于其他模型。分数越低，质量越好，这说明所提出的系统生成的教育视频更加准确、真实，且在语义上更加一致。

结果表明，在全部三种输入条件下，所提出的系统都显著优于其他模型。分数越低，质量越好，这说明所提出的系统生成的教育视频更加准确、真实，且在语义上更加一致。

TGAN 和 MoCoGAN 在通用视频生成方面表现出色，尤其是在没有特定输入指导（文本或声音）的情况下。然而，由于缺乏深层语境理解，它们在需要与语义内容保持一致时，性能会下降。TGANS-C 引入了语义条件控制，与文本的对齐效果有所提升，但仍缺乏多模态集成，因此在丰富教育内容中的适用性受限。所提出的系统（尽管目前的评估尚不完整）旨在支持双模态输入（文本 + 图像），但未能充分利用声音线索，因此不太适合旁白教育视频生成。

相比之下，所提出的系统将自然语言处理、文本转语音和生成对抗网络整合到统一流水线中，使其能够提取文本的语义含义、生成同步的声音旁白，并以具有教学意义的方式呈现概念。这种协同作用使内容与视觉效果更加契合，尤其是在同时提供文本和声音时。

数字对象标识符： https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接受日期：2025年7月10日

在教育视频合成领域，所提出的系统表现最佳，尤其是在准确性、相关性和语义深度至关重要的情况下。它能够适应文本和音频输入，因此优于 TGAN 和 MoCoGAN 等模型；这些模型并未针对结构化的教育应用场景进行优化。

向专家展示所提出的系统后，大多数专家表示高度认同，其中45%表示“非常同意”，30%表示“同意”。较小一部分（15%）持中立态度，而表示反对的仅占少数，其中5%选择“不同意”，另有5%选择“非常不同意”。总体而言，专家对该程序的评价普遍积极，但如图7所示，尚未达成完全共识。

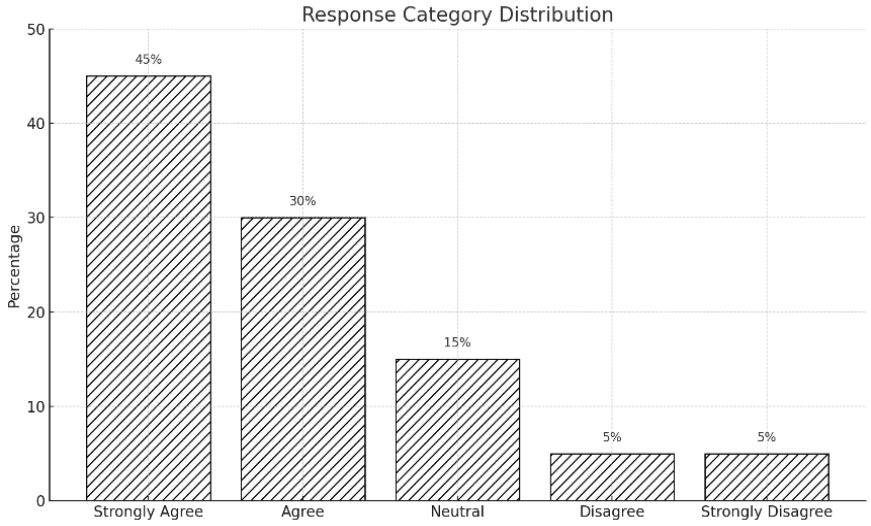


图7. 展示专家对该程序问卷的回答分布情况的图表 e

本研究邀请了15名计算机科学和教育技术领域的教师专家参与。旨在评估用于生成教学视频的计算机系统。反馈分为五个等级：非常同意、同意、中立、不同意和非常不同意。表（3）展示了针对所提出的系统、依据14项目标标准制定的专家评估问卷条目。

表3：所提出系统的专家评估问卷条目

无评	价项目	非常同意	同意	中立	不同意	非常不同意
1	帮助计算机科学学员生成教学视频。					
2	支持用户控制内容排序和导航。					
3	以逻辑清晰、有条理的方式呈现视频内容。					
4	提供用户友好界面和直观的互动。					
5	实现预期的教育成果。					

数字对象标识符: https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接收日期：2025年7月10日

6	能够妥善处理用户输入错误。					
7	运行可靠，不会出现技术故障。					
8	兼容各种计算系统。					
9	避免视频中出现不必要的视觉杂乱。					
10	充分利用屏幕空间。					
11	提供清晰的音频质量。					
12	提供适当的系统响应时间。					
13	使视频时长与学习内容相匹配。					
14	向用户清晰传达系统目标。					

75%的同意率凸显了该设备稳健的教育设计和技术可靠性。专家指出了诸多优点。该设备有助于提升学习者自主性（条目1、2）。它确保指导清晰并明确需求（条目3、5、14）。此外，它还体现出良好的技术稳定性（条目7、8、12）。仅有10%的反馈指出了需要改进的方面。这些问题较为轻微，并未影响整体效果。研究结果支持在教师培训项目中使用该设备，以促进技术支持的学习。

5. 结论

本文介绍了一种从文本或音频输入生成教学视频的新方法。系统首先从输入中提取关键词，然后生成与各关键词对应的图像，制作成无声视频。最后，系统从数据库中添加相关音频，生成高质量的教育视频。所提出的系统融合了生成对抗网络、自动语音识别、自然语言处理及其他算法，并将音频集成作为一项关键特色。专家对该系统给予了积极评价。与其他模型相比，该系统的错误率最低（28.75%），展现了其在结合文本与音频方面的卓越性能。其他系统的改进效果较差，未能达到可比的结果。未来的工作可考虑将该系统拓展至其他领域。

参考文献

[1] M. S. Noorderwier 和 M. van der Schoot，《教育视频的有效性：文献的元分析综述》，《教

数字对象标识符：https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接受日期：2025年7月10日

[8] J. Zhangjie Wu、Y. Li 和 B. Zhou, 《T2VScore：文本到视频生成评估的可靠指标》，载于

数字对象标识符： https://doi.org/10.54216/FPA.210115收稿日期：2025年2月25日 修订日期：2025年5月31日 接收日期：2025年7月10日