JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 12, pp. 73-85, December 2025
https://doi.org/10.9708/jksci.2025.30.12.073

Implementation of a Smart Lecture Video Generation System Using Open Source AI

SeongHun Kim*, Oh-Gyu Kwon**, Seong-Guk Nam***, Seung-Cheol Lee*, Tae-Young Yang*, Jae-Woo Ryu***, Chang-Hyeon Park****

*Researcher, Dept. of Computer Engineering, Yeungnam University, Gyeongsan, Korea
**CEO, MOBIPINTECHNOLOGY Co., Ltd., Daegu, Korea
***Researcher, R&D Center, NEARNETWORKS Co., Ltd., Daegu, Korea
****Professor, Dept. of Computer Engineering, Yeungnam University, Gyeongsan, Korea

[Abstract]

In this study, we propose an open-source AI-driven lecture video generation system that automatically produces instructional videos from university lecture materials. The proposed system integrates large language models (LLMs) and retrieval-augmented generation (RAG) to generate lecture scripts, synthesize speech through text-to-speech (TTS), and create lip-synced digital human videos in a unified pipeline. To construct a domain-specific knowledge base, lecture materials were collected from two Korean universities. Using the LangChain framework, the system connects HyperCLOVA X SEED 3B with a RAG module to generate lecture scripts tailored to an instructor's intent. The generated scripts are then processed by GPT-SoVITS for TTS synthesis and by a digital-human rendering module to produce realistic lecture videos. Experimental results show that integrating RAG improves performance over an LLM-only baseline by 5.6% in BERTScore, 3.3% in ROUGE-L, and 3.6% in BLEU. These improvements indicate that the proposed RAG structure enhances factual accuracy and logical consistency by referencing definitional and exemplar statements from external knowledge bases. This work demonstrates that even without fine-tuning, a RAG-based approach can be effectively applied to educational LLMs. It also provides foundational technology for future AI-driven smart-education platforms.

▸ **Key words**: LLM, RAG, Edutech, Open Source AI, Synthesis AI

• First Author: SeongHun Kim, Corresponding Author: Chang-Hyeon Park
*SeongHun Kim (hoonc-corgi@yu.ac.kr), Dept. of Computer Engineering, Yeungnam University
**Oh-Gyu Kwon (okkwon1021@gmail.com), MOBIPINTECHNOLOGY Co., Ltd.
***Seong-Guk Nam (nsg@aaf.co.kr), R&D Center, NEARNETWORKS Co., Ltd.
*Seung-Cheol Lee (fatalist316@gmail.com), Dept. of Computer Engineering, Yeungnam University
*Tae-Young Yang (xodud120016@gmail.com), Dept. of Computer Engineering, Yeungnam University
***Jae-Woo Ryu (wodn10110@aaf.co.kr), R&D Center, NEARNETWORKS Co., Ltd.
****Chang-Hyeon Park (park@yu.ac.kr), Dept. of Computer Engineering, Yeungnam University
• Received: 2025. 10. 29, Revised: 2025. 11. 11, Accepted: 2025. 11. 24.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 연구에서는 대학 강의자료를 기반으로 강의 영상을 생성하는 오픈소스 AI 기반 강의 영상 제작 시스템을 제안한다. 제안 시스템은 LLM과 RAG를 중심으로, 강의 대본 생성, TTS 기반 음성 합성, 디지털 휴먼 립싱크 영상 합성의 과정을 통합하여 End-to-End 방식으로 강의 영상을 생성한다. 이를 위해 국내 대학의 특정 교과목으로부터 강의자료를 수집하여 지식베이스를 구축한다. 구축된 지식베이스를 기반으로 LangChain 프레임워크를 이용해 HyperCLOVA X SEED 3B와 RAG를 연계하여, 교수자의 요구에 적합한 강의 대본을 생성한다. 생성된 대본은 GPT-SoVITS에 입력되며, TTS 음성 합성과 디지털 휴먼 립싱크 합성 과정을 통해 최종적으로, 현실감 있는 강의 영상이 제작된다. 제안 시스템은 RAG를 연계하여 LLM 단독 모델 대비 BERTScore 5.6%, ROUGE-L 3.3%, BLEU 3.6% 향상된 성능을 보였다. 이는 제안 시스템의 RAG 구조가 외부 지식베이스의 정의문과 예시문을 참조하여 모델의 사실적 정확성과 논리적 일관성을 향상시켰음을 의미한다. 본 연구는 별도의 파인튜닝 없이 RAG 기반 접근만으로 교육용 LLM의 실질적 활용 가능성을 입증하였으며, 향후 AI 기반 스마트 교육 플랫폼 구축의 핵심 기술적 기반을 제공한다.

▸ **주제어**: 대규모 언어모델, 검색증강생성, 에듀테크, 오픈소스 인공지능, 합성 인공지능

I. Introduction

최근 교육 환경은 디지털 기술의 발달과 사회적 요구의 변화에 따라 빠르게 변화하고 있다. 과거에는 대부분의 수업이 오프라인에서 대면으로 이루어졌지만, 이제는 온라인 플랫폼을 기반으로 한 비대면 수업이나 혼합형 수업인 블렌디드러닝 또는 플립러닝 등의 형태가 적극적으로 도입되고 있다[1]. 플립러닝은 학습자가 수업 전에 강의 영상을 시청하고, 수업 시간에는 토론이나 실습을 수행하는 참여 중심 학습 방식으로 주목받고 있다. 이러한 접근은 학습자의 이해도와 몰입도를 높일 수 있지만, 교수자가 수업마다 별도의 강의 영상을 직접 제작해야 한다는 한계가 존재한다. 특히 대학 강의의 경우 방대한 분량의 강의자료와 세부 개념을 영상으로 변환하는 과정이 많은 시간과 노력을 요구한다. 따라서 교육의 질을 유지하면서도 효율적으로 강의 콘텐츠를 제작할 수 있는 자동화 기술의 필요성이 대두되고 있다[2-4].

실제로 최근 많은 대학 및 교육 기관에서는 대면 수업을 복습용으로 다시 영상화하거나, 플립러닝 방식의 수업을 위해 별도의 강의 영상을 제작하는 사례가 증가하고 있다[5]. 이와 같은 강의 영상 제작은 단순한 녹화 작업을 넘어서, 대본 작성, 음성 녹음, 자료 시각화, 영상 편집 등 다양한 작업이 포함된다. 하지만 이러한 작업은 교수자의 상당한 시간과 노력을 요구하며, 강의 질의 일관성이나 학습자의 이해도에 직접적인 영향을 줄 수 있다. 특히 교수자가 영상 편집 기술에 익숙하지 않거나 전용 장비가 없는 경우, 양질의 강의 콘텐츠를 제작하는 데 큰 제약이 따른다. 또한 디지털 교육이 활성화되고 있음에도 불구하고, 교수자가 개별 강좌마다 영상을 반복적으로 제작하는 것은 교과 증설, 강의자료 개편 등을 고려할 때 비효율적인 교육 자원 사용으로 지속성이 낮다. 따라서 강의자료만으로 고품질 강의 영상을 제작할 수 있는 지능형 자동화 시스템의 필요성이 점차 커지고 있다[6-7].

본 연구는 기존 연구들이 개별 기술 요소에만 집중했던 것과 달리, 강의 자료 기반 멀티모달 분석, RAG 기반 대본 생성, TTS, 디지털 휴먼 합성을 하나의 시스템으로 완전 자동화한 최초의 강의 영상 자동 제작 구조를 제안한다는 점에서 차별적인 기술적 기여를 가진다. 이 통합 구조는 교수자의 제작 부담을 획기적으로 줄이고, 전문 기술 없이도 고품질의 강의 영상을 생성할 수 있도록 지원한다. 본 논문의 기여는 다음과 같다.

- 기존에 분리되어 존재하던 대본 생성, 음성 합성, 립싱크, 번역, 영상 제작을 하나의 End-to-End 시스템으로 통합한 최초의 자동 강의 영상 제작 도구를 제안하였다.
- 글로벌 학습자를 위해 전문용어 정규화 및 번역 모델 파인튜닝을 통해 고등교육 분야에서도 신뢰할 수 있는 번역 품질을 확보하였다.
- 교수자 이미지 기반 디지털 휴먼 합성과 자연스러운 TTS 음성을 자동 결합하여 실제 교수의 화법과 강의 스타일을 반영한 고품질 강의 영상을 제공한다.

- 제안 시스템은 다양한 전공, 교육 수준, 교수자 환경에 유연하게 적용 가능하며, 향후 LMS 및 온라인 교육 플랫폼과의 연계를 통해 대규모 교육 운영에도 활용될 수 있다.
- RAG 구조를 통해 외부 지식베이스를 실시간으로 참조함으로써 강의자료의 사실성, 예시의 정확성, 개념 설명의 논리성을 강화한 강의 대본을 작성한다.

이러한 기술적 기여는 다음과 같은 교육적 효과로 이어진다. 우선, 촬영, 녹음, 편집과 같은 복잡한 제작 과정이 완전히 자동화됨에 따라 교수자는 강의자료 준비와 수업 운영에 집중할 수 있고, 학습자는 다양한 언어와 형태의 고품질 학습 콘텐츠에 접근할 수 있다. 또한 디지털 휴먼 기반 강의 영상은 높은 몰입도와 지속성을 제공하며, 다국어 자막 지원은 외국인 학습자의 학습 장벽을 크게 완화한다. 결과적으로 본 연구는 교육의 질, 효율성, 접근성을 동시에 향상시키는 최초의 End-to-End AI 기반 강의 영상 제작 시스템으로서, 스마트 교육 환경 전반에 확장 가능한 핵심 기술적 기반을 마련한다.

이하, 본 논문의 구성은 2장에서 외부 지식 통합, 데이터 전처리, 음성 및 버추얼 휴먼 합성, 다국어 자막 생성, 강의 영상 제작과 관련하여 수행된 연구를 검토한다. 3장에서는 제안하는 강의 영상 제작 시스템을 상세히 기술하며, 4장에서는 실험 구성과 정량적 평가 결과를 제시한다. 마지막으로, 5장에서는 결론과 기대 효과를 논의한다.

II. Related Work

2.1 External Knowledge Integration for Lecture Script Generation using LLM

최근 LLM은 자연어 처리 분야에서 뛰어난 성과를 거두었다. 특히 GPT, LLaMa, PaLM, HyperCLOVA 등의 다양한 모델들은 방대한 범용 텍스트 데이터를 기반으로 사전 학습되어, 문장 생성, 질의응답, 요약, 번역 등 다양한 작업에서 뛰어난 성능을 달성했다[10-13]. 특히 자기 회귀(Autoregressive) 기반 트랜스포머 구조를 채택한 모델들은 문맥 이해와 텍스트 생성 측면에서 기존 접근법보다 우수한 성능을 보인다[13, 14]. 그러나 이와 같은 오픈소스 기반 범용 LLM은 일반적인 주제에는 강점을 보이나, 특정 도메인의 지식이 요구되는 작업에서는 성능 저하를 보인다[15]. 특히 고등교육 분야의 강의자료는 전문용어, 논리적 설명 구조, 시각 자료 해설 등 복잡한 구성을 포함하고 있어, 범용 LLM으로는 학습자의 이해도 향상에 효과적인 강의 대본을 생성하기 어렵다. 또한, 교육 자료는 정확하고 검증된 지식 전달을 전제하므로, LLM이 생성한 텍스트에 허위정보(Hallucination)가 포함되거나 최신 정보가 반영되지 않을 시 학습자의 오개념 형성을 초래할 수 있다. 이를 보완하기 위한 접근법으로 RAG가 주목받고 있다[16].

RAG는 외부 지식베이스에서 관련 정보를 검색하고 이를 프롬프트에 통합함으로써, 신뢰성 있는 문장 생성을 가능하게 한다. 결론적으로, 범용 LLM의 능력만으로는 고등교육 콘텐츠와 같은 도메인 특화 텍스트 생성 및 정확성에 한계가 있으며, 이를 해결하기 위해 외부 지식 통합 기술이 병행되어야 한다.

2.2 Pre-processing of Documents for Knowledge Base Construction

대부분의 강의 문서는 PDF, PowerPoint, Word 등 다양한 형식으로 제공되며, 페이지 또는 슬라이드 단위의 시각적 레이아웃과 텍스트 및 이미지가 혼합된 멀티모달 문서 구조를 가진다. 따라서 이러한 복합 구조에 적합한 분석 기법과 처리 전략이 요구된다. 즉, 강의자료로부터 신뢰성 있는 지식베이스를 구축하기 위해서는 텍스트와 시각적 요소를 정확히 분리, 추출하여 구조화하는 전처리 과정이 선행되어야 한다[17].

텍스트 처리 단계에서는 문서의 논리적 구조를 보존하면서 정보를 정확히 추출하는 것이 중요하다. 이를 위해 강의자료를 PDF 형식으로 변환한 뒤 PDFMiner.six와 같은 도구를 활용하여 텍스트 내용뿐 아니라 위치, 폰트 크기, 글꼴 속성 등 레이아웃 정보를 함께 추출함으로써 구조 인식 기반 분석을 수행할 수 있다[18]. 또한 인코딩 오류를 방지하기 위해 chardet 등의 도구를 이용한 텍스트 정제 및 표준화 과정이 필요하다[19].

이미지 처리의 경우, 슬라이드 내 그래프나 도식은 강의 내용의 의미적 이해에 중요한 단서를 제공하므로 정밀한 인식이 요구된다. Tesseract와 같은 OCR(Optical Character Recognition) 도구를 사용하여 이미지 내 삽입된 텍스트를 추출할 수 있으며, 인식 정확도 향상을 위해 해상도 보정, 노이즈 제거, 이진화 등의 전처리 기법을 병행한다[20].

이렇게 독립적으로 추출된 텍스트와 이미지 정보는 페이지 또는 슬라이드 단위에서 통합되어 강의 내용의 의미 구조를 반영하는 통합 표현으로 구성된다. 이러한 표현은 이후 지식베이스의 구성 단계에서 핵심 입력으로 활용되어, 대본 생성, 요약, 질의응답 등 하위 모듈이 사실적 근거를 기반으로 작동할 수 있도록 지원한다.

2.3 Synthesize Speech and Virtual Human

텍스트 기반으로 생성된 강의 대본은 문서 형태보다는 시청각 콘텐츠로 제공될 때 학습자의 몰입도와 이해도를 높일 수 있다. 특히 비대면 환경에서는 실제 교수자의 발화와 유사한 음성과 시각적 피드백이 학습 집중력과 지속성에 긍정적인 영향을 미친다. 이에 따라 최근에는 TTS 및 디지털 휴먼 기술이 교육 콘텐츠 제작에 도입되고 있으며, 이는 강의 자동화 시스템의 주요 구성 요소로 작용한다. 기존 TTS 모델은 특정 음성을 학습하기 위해 대량의 음성 데이터를 요구하거나 표현력의 한계가 있었으나, 최근에는 Few-shot 기반 기술의 개발으로 소량의 화자 데이터로도 특정 음성을 학습할 수 있게 되었다[21].

GPT-SoVITS는 대표적인 멀티스피커 음성 모델로, 새로운 화자의 음성을 별도로 학습하지 않고도 음성 샘플로부터 화자의 특징을 추출함으로써 해당 인물의 음색과 말투를 재현할 수 있다. 특히 수백 ms 단위의 phoneme-level alignment를 활용하여 고품질·고정밀 립싱크 음성 합성을 구현할 수 있는 경량화된 구조를 가진다. 이러한 기술적 특성은 짧은 음성 샘플만으로도 개인화된 음성을 효과적으로 생성할 수 있어, 실제 교수자의 목소리를 반영한 콘텐츠 제작에 활용될 수 있다.

Wav2Lip은 시청각 동기화를 위한 립싱크 기술으로, 영상에서 추출한 얼굴과 생성된 음성을 결합해 실제 발화와 유사한 입 모양을 구현할 수 있다. Zero-shot 방식으로도 동작하므로 학습되지 않은 이미지에도 자연스럽게 적용할 수 있다. 이러한 음성 및 영상 기술은 단순한 시청각 효과를 넘어, 학습자에게 친숙하고 일관된 강의 경험을 제공하고, 대면 수업과 유사한 실재감을 부여하며, 동일 콘텐츠의 다국어 확장이나 접근성 향상의 기반이 된다.

2.4 Multilingual Lecture Script Generation

디지털 교육 콘텐츠는 언어적 배경과 지역에 관계없이 학습 기회를 제공함으로써 글로벌 학습 환경을 지원하는 핵심 매체로 작용한다. 이에 따라 동일한 교육 콘텐츠에 대하여 다양한 언어권의 학습자들이 동일한 학습 경험을 가질 수 있도록 하기 위한 다국어 자막의 수요가 증가하고 있다. 자막은 청각 보조뿐 아니라 학습자의 이해를 돕는 수단으로서, 교육 접근성과 포용성을 높이는 필수 요소로 고려된다. 기존 자막 생성은 수작업 또는 단순 기계 번역 기반으로 이루어졌으나, 고등교육 콘텐츠의 전문성과 용어 복잡성으로 인해 일반 번역 모델로는 품질 보장이 어렵다. 이로 인한 전문용어의 부정확한 번역은 학습자의 강의 이해를 저해하므로, 번역 모델 정제와 용어 정규화 체계를 활용하여 고품질 자막을 생성하여야 한다.

최근 위스퍼와 같은 음성 인식 기술과 함께 기계 번역 기술은 최근 트랜스포머 기반의 신경망 모델 등장 이후 비약적인 성능 향상을 이루었으며 품질 높은 자막을 제공함에 큰 기여를 하고 있다[22]. 특히 DeltaLM(Decoder Enhanced Lossless Transformer Language Model)은 고도화된 인코더-디코더 구조를 바탕으로, 다국어 문장 간의 의미적 정렬을 효과적으로 학습하며 한국어, 영어, 중국어, 베트남어, 불어 등 다양한 언어 쌍에 대해 뛰어난 번역 성능을 보인다[23]. 또한 이 모델이 형성한 다국어 의미 공간은 교육 콘텐츠에서 흔히 나타나는 길고 복잡한 문장 구조에서도 문맥 보존과 의미 전달의 정확도를 높이는 데 강점을 가진다. 그러나 전문용어 중 “Compiler”, “Stack” 등과 같은 표현은 번역에 따라 컴퓨터공학 분야의 “컴파일러”와 “번역기”, 자료구조의 일종인 “스택”과 “더미” 등으로 번역될 수 있다. 이처럼 선택된 번역 표현의 의미와 도메인의 일치 여부에 따라 번역 품질이 크게 달라질 수 있다. 이에 따라 본 시스템은 사전 정의된 전문용어 데이터셋을 기반으로, 자주 사용되는 전문용어들의 다양한 표현과 그에 대응하는 표준 번역어를 정규화하는 작업을 수행한다. 이 과정은 번역 결과의 정확도뿐만 아니라, 교육 콘텐츠 전반의 용어 일관성과 신뢰성을 확보하는 데 필수적이다.

2.5 Lecture Video Generation System

최근 인공지능 기술의 발달과 교육 환경의 변화에 따라 강의 영상 제작을 자동화하려는 다양한 시도가 이루어지고 있다. 공리 외의 연구에서는 생성형 AI를 기반으로 디지털 휴먼이 강의를 진행하는 시스템을 제안하였으며, 실제 교수자의 외형과 유사한 영상 표현이 학습자의 몰입도와 만족도에 긍정적인 영향을 미칠 수 있음을 실증하였다[24]. 특히 텍스트 입력만으로 강의에 활용 가능한 시청각 콘텐츠를 구현함으로써, 기존 강의 영상 제작에 필요한 물리적 자원의 대체 가능성을 제시하였다. 한 편, 박동연 외의 연구에서는 시각장애 학습자를 위한 온라인 강의 해설 시스템을 제안하였으며, 슬라이드 내 이미지, 표, 텍스트 등을 자동 분석하여 TTS 기반 음성 해설을 생성하고 기존 강의 영상에 삽입함으로써 정보 접근성 향상의 기술적 기반을 마련하였다[25]. 오령의 연구에서는 AI 기반 영상 제작 도구의 사용자 수용성에 대한 연구를 통해, 효율성과 혁신성이 사용자 인식과 활용 의도에 긍정적인 영향을 미친다는 점을 규명하였다. 이 연구는 실제 교육자와 1인 창작자 집단을 대상으로 하여, AI 영상 제작 시스템이 교육

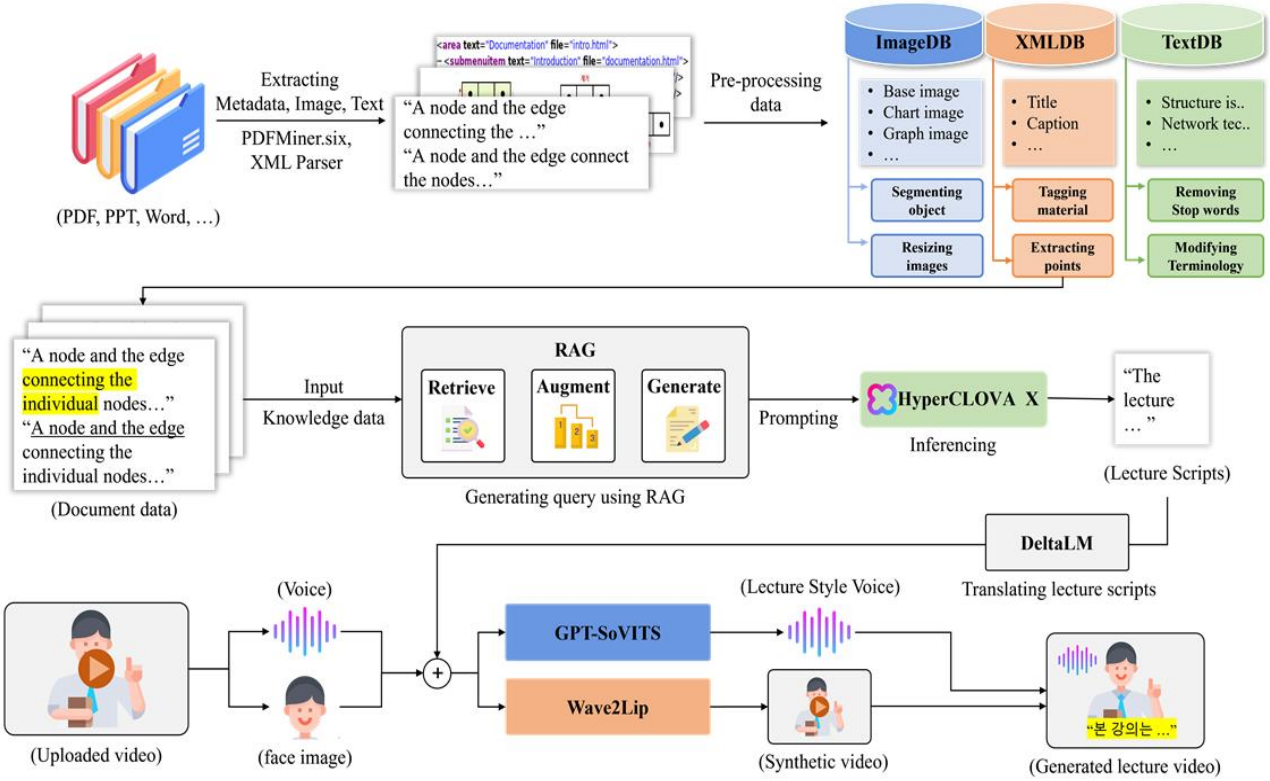


Fig. 1. End-to-End workflow of the proposed intelligent lecture video generation system integrating RAG, LLM, and synthetic video generation.

현장과 콘텐츠 산업 전반에 확산될 수 있는 가능성을 보여주었다[26]. 이미영의 연구는 초등교육에서 생성형 AI 영상 도구를 적용한 수업 사례를 개발하고, 학생들의 영상 제작 활동을 통해 창의성과 디지털 표현 능력의 향상 가능성을 확인하였다[27]. 또한 디지털콘텐츠학회 등에서는 최근 상용화된 다양한 AI 영상 제작 도구들을 기능별로 분석하고, 이들이 실제 교육 콘텐츠 제작에 활용될 수 있는 구조를 제시하였다[28]. 이처럼 기존 연구들은 각각의 맥락에서 AI 기반 강의 영상 제작의 가능성과 효과를 입증하고, 실제 온/오프라인 교육 현장에서의 실용성을 제시하고 있다. 그러나 이러한 연구들은 대부분 영상 제작의 특정 요소에 국한되거나, 완성된 영상에 대한 후처리 또는 도구 활용의 효과 분석에 초점을 두고 있어, 강의자료로부터 대본 생성, 음성 합성, 시각 자료 해설, 디지털 휴먼 연동, 자막 번역에 이르는 End-to-End 자동화 과정 전체를 구현한 사례는 부족한 상황이다. 또한 고등교육 콘텐츠의 전문성과 논리적 구조를 반영할 수 있는 LLM 기반의 정확한 대본 생성, 교수자의 발화 성격을 반영한 음성 합성, 시각 자료 중심의 강의 구성 등은 충분히 고려하지 않거나 기술적 제약이 존재한다. 이에 본 연구는 이러한 한계점을 보완하기 위해 다양한 기술 요소를 포함하는 인공지능 기반 강의 영상 제작 시스템을 제안한다. 기존 연구들은 강의 콘텐츠 제작을 위한 다양한 과정 중 개별 단계에 초점을 맞춘 반면, 본 연구는 대본 생성, 도메인 특화 지식 반영, 음성 합성 및 영상 제작의 전 과정을 자동화하는 통합 시스템을 구현한다. 이러한 통합적 접근은 강의자료 기반 자동화의 효율성과 생성 결과의 신뢰성을 동시에 향상시키며, AI 기반 교육 콘텐츠 제작의 방향성을 제시한다는 점에서 의의가 있다.

III. Lecture Video Generation System

본 장에서는 대학교에서 활용되는 강의자료를 바탕으로 교수자가 학습자에게 제공 가능한 오픈소스 AI 기반 강의 영상 제작 시스템을 제안하고, 그 개발 방법을 단계별로 설명한다. 본 연구에서 활용되는 오픈소스 AI의 하이퍼 파라미터는 기본 설정값을 활용하였다. 제안 시스템은 Figure 1과 같이 데이터 수집, 강의자료 전처리, 강의 대본 생성, 강의 영상 제작의 네 단계로 구성한다. 먼저 실제 대학 강의에서 활용되는 다양한 교과목의 강의자료와 영상을 수집하여, 지식베이스 구축을 위한 데이터셋을 구성한다. 수집된 데이터셋으로부터 텍스트와 이미지를 추출하고, 불필요한 정보를 제거한 뒤 메타데이터를 포함한 문서

단위로 전처리한다. 전처리한 데이터는 벡터 인덱스로 변환하여 검색 가능한 형태의 지식베이스로 저장한다. 이후, 저장한 데이터는 RAG와 연동하여, 질의에 따라 관련 내용을 검색함으로써 사실 기반의 강의 대본을 자동으로 생성한다. 마지막으로, 생성한 대본을 바탕으로 음성 합성 기술과 디지털 휴먼 영상 생성 기술을 통해 자연스러운 강의 영상을 제작한다.

이러한 과정을 진행하여 제안 시스템은 기존 강의자료로부터 자동으로 지식 기반 강의 영상을 생성하는 완전한 파이프라인을 구현한다. 단, 최종적으로 생성된 강의 영상의 모든 저작권 및 소유권은 해당 영상을 제작한 교수자와 그 소속 기관에 귀속되며, 생성된 디지털 휴먼, 음성 합성 결과물, 대본 등의 2차적 저작물 또한 동일한 권리 범위 내에서 보호됨을 반드시 고지한다.

3.1 Data Collection

HyperCLOVA X SEED 3B는 네이버에서 상업용으로 공개한 오픈 엑세스 모델로, 약 30억 파라미터 규모의 경량 멀티모달(텍스트+이미지) 모델이다. 또한 이 모델은 본 연구에서 필수적인 생성·요약·분석에 최적화된 인스트럭션 튜닝이 적용되어 있다. 이러한 모델 구조로, 텍스트와 이미지 처리에 특화되어 있으며, 질의와 검색, 생성을 통합한 RAG 구조와 함께 사용하여 외부 지식베이스의 정보를 동적으로 반영 가능하다. 따라서 교수자가 제시하지 않은 내용도 관련 자료를 검색 및 참조하여 대본을 생성할 수 있다. 이에 따라, 지식베이스 구축을 위해 국내 2개 대학과 협력하여 컴퓨터공학과의 핵심 교과목인 자료구조, 알고리즘, 컴퓨터구조, 운영체제, 데이터베이스 등 5개 과목의 강의자료 및 강의 영상을 수집하였다. 과목별 1학기에 해당하는 분량의 자료를 확보하여 총 560건의 강의자료 및 영상 데이터를 구축하였으며, 주요 교과목별 데이터 수량은 Table 1과 같다. 수집된 자료는 PDF, PPT, Word 형식으로 정규화되어, 이후 RAG를 위한 지식베이스 구축 단계에서 활용된다.

Table 1. Summary of Collected Lecture Materials for RAG-based Script Generation

Subject Name	Number of data
Data Structure	Lecture pdf/ppt : 14 * 5 Lecture video : 14 * 5
Operating Systems	Lecture pdf/ppt : 14 * 3 Lecture video : 14 * 3
Database	Lecture pdf/ppt : 14 * 5 Lecture video : 14 * 5
Computer Architecture	Lecture pdf/ppt : 14 * 2 Lecture video : 14 * 2
Algorithm	Lecture pdf/ppt : 14 * 5 Lecture video : 14 * 5

3.2 Pre-processing Lecture Data

본 절에서는 RAG 구조에서 효율적인 검색과 문맥 통합이 가능하도록 데이터를 전처리하고 지식베이스를 구축하는 과정을 설명한다. 전체 프로세스는 콘텐츠 추출, 텍스트 정제, 메타데이터 구조화, 지식베이스 인덱싱 순으로 진행된다.

콘텐츠 추출 단계에서는 PDFMiner.six와 PyMuPDF 라이브러리를 활용하여 각 강의자료로부터 텍스트와 이미지를 분리 추출한다. PowerPoint 자료의 경우 XML Parser 기반으로 내부 텍스트와 이미지 객체를 직접 식별하며, Word 자료는 동일 포맷으로 변환 후 처리한다[29].

이후 텍스트 정제 단계에서 텍스트의 인코딩을 감지하여 UTF-8로 통일하고, 특수문자, 불필요한 공백, 페이지 번호, 저자 정보 등을 제거한다. 강의 내용과 직접 관련이 없는 문장은 필터링하고, 문단 단위로 구조화하여 검색 가능한 형태의 문서 단위로 저장한다.

3.3 Lecture Script Generation

본 절에서는 전처리 된 강의자료 데이터를 활용하여 강의 대본을 생성하기 위한 LLM과 RAG 시스템의 연계 과정을 설명한다. 본 연구에서는 상업용 오픈소스 모델인 Hyper CLOVA X SEED 3B를 LLM 모델을 사용하였으며, 추가적인 파인튜닝을 수행하지 않고, RAG를 연계하여 모델이 사실 기반의 강의 대본을 생성할 수 있다. 기존 LLM은 사전 학습 시점에 한정된 정보를 포함하므로, 최신 학문적 내용이나 교육 자료의 변화에 대응하기 어렵다. 이를 해결하기 위해 본 연구에서는 LangChain 프레임워크를 활용하여 RAG 를 구축한다. LangChain은 외부 지식베이스로부터 검색된 문서를 LLM 입력 프롬프트에 통합함으로써 모델이 사실 기반의 정보에 근거한 응답을 생성할 수 있도록 지원한다.

제안하는 시스템의 전체 구조와 실행 흐름은 Figure 2와 같이 지식베이스 구축, 검색 메커니즘, 생성 통합의 세 단계로 구성된다. 지식베이스 구축 단계에서는 수집 및 전처리된 강의자료를 기반으로 벡터 형태의 검색 가능한 데이터베이스를 생성한다. 이를 위해 강의자료의 텍스트, 이미지 요약 정보, 그리고 강의 영상의 음성 인식 결과를 통합하여 문단 단위로 구조화한다. 그리고 각 문단은 문장 트랜스포머 기반 한국어 특화 임베딩 모델인 KoSimCSE[30]을 통해 다차원 벡터로 변환되고, FAISS [31]에 인덱싱 한다. 이 데이터베이

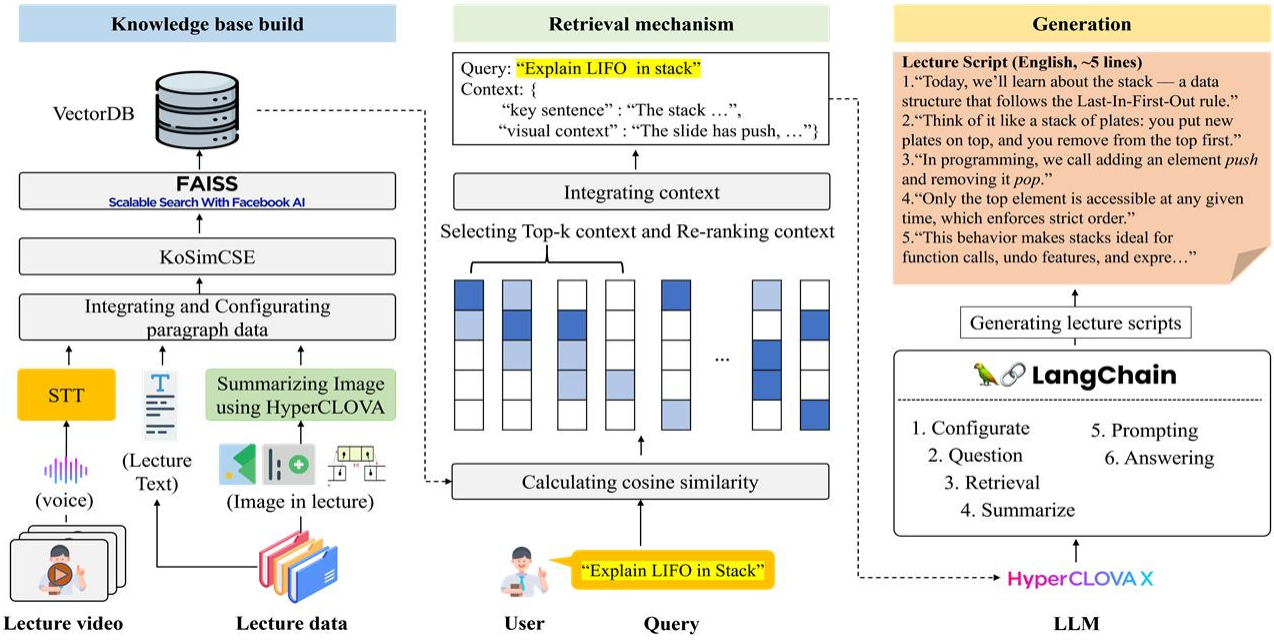


Fig. 2. Flow of Proposed RAG System

스는 RAG 질의 시 교육 도메인별 사실 기반 문맥을 검색하는 데 활용된다. 또한, 강의 슬라이드 내 그래프나 도표와 같은 시각적 요소는 Hyper CLOVA SEED Vision Encoder를 통해 텍스트 설명으로 변환하고, 해당 정보를 메타데이터로 함께 저장한다. 이러한 구조를 통해 강의자료 내 시각 정보와 텍스트 정보가 통합된 형태의 검색이 가능하다.

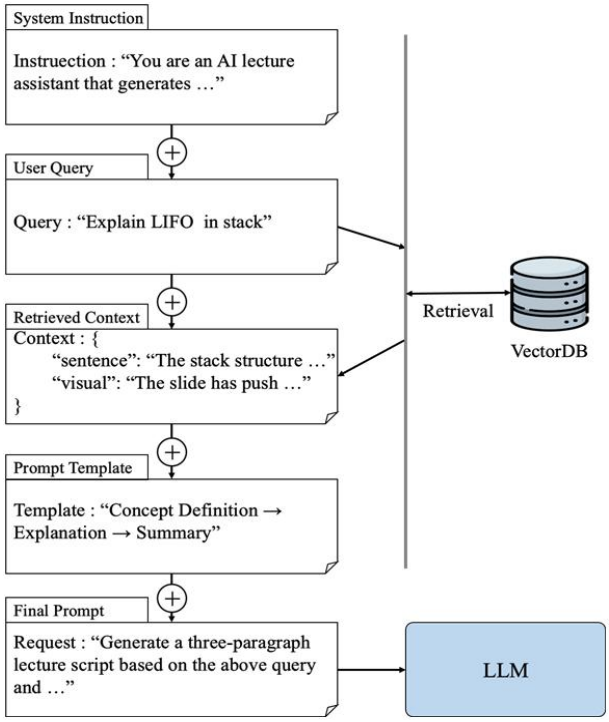
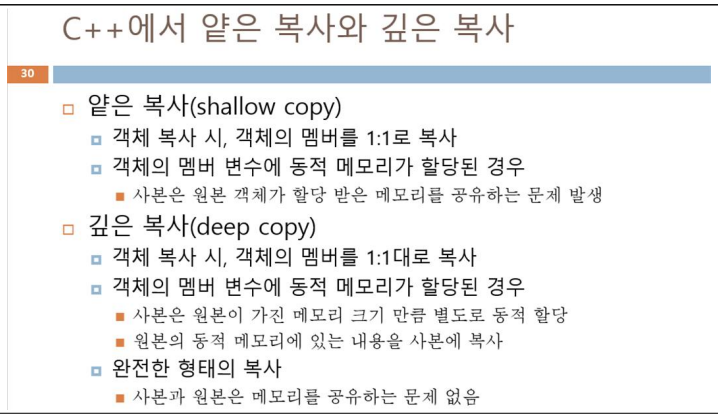


Fig. 3. Process of prompt construction for lecture generation using LLM

검색 단계에서는 "스택의 후입선출 구조를 설명하라"와 같은 사용자의 질의를 입력으로 받아, LangChain의 검색 모듈이 벡터 DB에서 관련 문서를 검색한다. 유사도 계산에는 코사인 유사도를 사용하며, 검색된 결과는 Top-k 후보로 선별된 뒤, 재정렬(Re-ranking) 과정을 거쳐 가장 관련성이 높은 문서를 선택한다. 이후 선별된 문서의 핵심 문장, 시각적 요약, 관련 개념을 하나의 Context로 통합하고, 이를 HyperCLOVA X SEED 3B 모델의 입력 프롬프트에 포함시킨다. 프롬프트 생성 과정은 Figure 3과 같이 단계적으로 구성한다. 먼저 모델의 역할을 명시하는 시스템 지시문을 통해 일관된 응답 톤과 생성 목표를 설정한다.

이후 사용자 질의가 입력되면, RAG 검색 모듈에서 반환된 관련 문맥과 시각적 정보가 함께 통합되며, 모델이 구성할 강의 대본의 논리적 전개 구조를 정의하는 프롬프트 템플릿이 추가적으로 결합 된다. 마지막으로 모든 요소들을 결합한 최종 프롬프트를 모델에 입력하여, 검색 결과에서 반환된 지식을 활용하여 사실적 근거와 논리적 흐름을 갖춘 강의 대본을 생성한다. Figure 4는 제안모델에 입력된 슬라이드로부터 생성한 대본을 나타내고 있으며 실제 대본과 비교하여 문맥상 유사함을 나타낸다.

생성된 대본은 문단 단위로 구성되며, 각 슬라이드 단위의 강의 스크립트로 저장된다. 이 과정에서 LangChain의 프롬프트 템플릿 기능을 이용하여, 슬라이드별 설명, 핵심 개념, 예시 설명, 요약의 구조를 자동으로 형성하도록 프롬프트를 설계하였다. 이를 통해, HyperCLOVA X SEED 3B 모델은 파인튜닝 없이도 외부 데이터베이스의 지식을



[실제 대본]
이번에는 얕은 복사와 깊은 복사에 대해 살펴보겠습니다.
먼저 얕은 복사는 우리가 일반적으로 사용하는 복사 연산자를 통해, 객체가 가진 각 멤버 변수의 값을 그대로 복사하는 방식입니다. 문제는 객체 내부에 포인터 멤버가 포함되어 있을 때 발생합니다. 이 경우 단순히 주소 값만 복사되기 때문에, 원본 객체와 복사된 객체가 동일한 메모리 자원을 공유하게 됩니다. 이렇게 메모리를 공유하는 상태에서는 한쪽에서 값을 수정하거나 메모리를 해제했을 때, 다른 객체에서도 예상치 못한 오류나 충돌이 발생할 수 있습니다.
깊은 복사는 이러한 문제를 해결하기 위해 사용자가 직접 복사 과정을 정의하는 방식입니다. 포인터 멤버가 있다면, 먼저 필요한 크기만큼 새로운 메모리를 따로 할당하고, 그 안에 원본이 가지고 있던 실제 데이터를 일일이 복사해야 합니다.
결론적으로 깊은 복사는 얕은 복사보다 구현 과정이 복잡하고 시간이 더 들 수 있지만, 각 객체가 독립적인 메모리 공간을 가지므로 훨씬 안전한 복사 방식입니다.

[생성 대본]
이번 슬라이드에서는 매우 중요한 개념인 얕은 복사와 깊은 복사의 차이를 설명드리겠습니다.
먼저 얕은 복사입니다. 얕은 복사는 객체를 복사할 때 멤버 변수의 값을 그대로 1:1로 복사하는 방식입니다. 그런데 이 멤버 변수 안에 동적 메모리 포인터가 들어 있는 경우, 문제가 생길 수 있습니다. 복사된 객체와 원본 객체가 동일한 메모리 공간을 공유하게 되기 때문입니다. 이 상황에서는 한 객체에서 메모리를 해제하면 다른 객체에서도 문제가 발생하는, 이른바 이중 해제 같은 오류가 생길 수 있습니다.
반면, 깊은 복사는 이러한 문제를 해결하는 방식입니다. 깊은 복사는 마찬가지로 객체의 멤버를 1:1로 복사하지만, 포인터 멤버가 있다면 원본이 가지고 있던 동적 메모리의 크기만큼 새롭게 메모리를 할당합니다. 그리고 원본 메모리의 내용까지 그대로 복사해서, 두 객체가 완전히 독립적인 메모리 공간을 갖도록 만들어 줍니다. 그래서 원본과 사본이 서로에게 영향을 주지 않으며 안전한 복사 방식이라고 할 수 있습니다.
정리하면, 얕은 복사는 빠르지만 위험 요소가 있고, 깊은 복사는 비용은 더 들지만 안전합니다.

Fig. 4. Comparison of Actual Script and Proposed-Model-Generated Script

실시간으로 참조하여, 사실에 기반한 최신 정보가 포함된 강의 대본을 동적으로 생성할 수 있다.

3.4 Digital Human-based Lecture Video

본 절에서는 3.3절에서 생성된 강의 대본을 활용하여, 교수자의 실제 음성을 모사한 TTS 음성과 디지털 휴먼의 얼굴 영상을 결합한 현실감 있는 강의 영상을 제작하는 과정을 설명한다.

연구적 활용목적으로 사전에 허락을 받은 약 30분 분량의 교수자 음성 데이터를 이용하여 발화 특성과 억양 패턴을 반영하며, 생성된 강의 대본을 입력으로 하여 실제 교수자의 목소리와 유사한 억양과 리듬을 갖는 음성을 합성한다. 특히 GPT-SoVITS는 발음 단위의 세밀한 타이밍 정보를 조절할 수 있어, 문장 길이나 쉼의 위치에 따른 발화 리듬까지 자연스럽게 구현할 수 있다. 결과적으로 생성된 음성은 교수자의 실제 화법(교수식 톤, 설명의 완급 조절, 문장 말미 강세 등)을 충실히 재현하여 학습자의 몰입감을 높이는 데 기여한다.

이후 디지털 휴먼 영상 합성 단계에서는 생성된 음성과 연구적 목적으로 사전에 허락을 받은 교수자의 얼굴 영상을 입력으로 받아, 자연스러운 강의 영상으로 합성한다. 본 연구에서 사용한 Wav2Lip 모델은 입력된 음성과 얼굴 영상의 입술 영역을 정밀하게 합성하여, 음성의 발화 타이밍에 맞게 입술 모양을 생성한다. 또한 화자의 음향 스펙트로그램을 분석하여 프레임 단위로 입술 움직임을 예측하고, CNN 기반 얼굴 인코더와 오디오 디코더를 통해 시각적 자연스러움을 유지할 수 있다. 특히, 학습되지 않은 새로운 인물의 얼굴에도 적용 가능한 Zero-shot 합성 기능을 지원하므로, 다양한 교수자 얼굴 영상에 손쉽게 립싱크를 적용할 수 있다.

강의 영상 합성은 다음과 같은 순서로 진행된다. 먼저, 3.3절에서 생성된 강의 대본을 기반으로 GPT-SoVITS에서 생성된 음성 파일을 입력한다. 다음으로, 약 3초 길이로 촬영된 교수자 얼굴 영상을 Wav2Lip의 입력으로 제공하면, 음성의 시간축에 맞춰 자연스럽게 입술이 움직이는 립싱크 영상이 생성된다. 마지막으로, 생성된 립싱크 영상

Table 2. Experimental Configuration Parameters

Category	Configuration	Detail
RAG Configuration	Framework	LangChain
	Retrieval Mechanism	Integrated pipeline of Retriever and VectorStore
	Retrieval Method	Top-k search based on cosine similarity
	Top-k	5
	Prompt Integration Method	LangChain PromptTemplate + Context Augmentation
Embedding Model	Model Name	KoSimCSE
	Vector Dimension	768
	Training Status	Pretrained
Vector DB	Idnexing Library	FAISS(Facebook AI Similarity Search)
	Indexing Algorithm	HNSW(Hierarchical Navigable Small World)
	Retrieval Speed	Average 50ms per query
Metrics	Target	Three lecture scripts written by actual lecturers
	Function	BERTScore(Precision/Recall/F1), ROUGE-L, BLEU

과 강의자료의 슬라이드, 그래프, 도식 등을 시간 동기화하여 하나의 완성된 강의 영상으로 렌더링한다.

이렇게 완성된 디지털 휴먼 기반 강의 영상은 실제 교수자의 음성과 얼굴 움직임을 사실적으로 재현함으로써, 기존의 텍스트 또는 슬라이드 중심 강의보다 학습자의 주의집중도와 감정적 몰입도를 효과적으로 향상시킨다. 또한, 동일한 강의 대본을 다양한 언어와 화자 스타일로 자동 변환할 수 있어, 국제화된 교육 환경에서 활용 가능성을 높인다. 따라서, 본 시스템은 인적 자원의 한계를 보완하면서도 이질감 없는 교수자의 형상을 유지하는 형태의 지능형 강의 콘텐츠 생성 기술로 활용될 수 있다.

IV. Experiments

본 장에서는 제안하는 RAG 구조를 LLM과 연계함으로써, 강의 대본 생성 시 사실적 정확성과 의미 일관성이 향상되는지를 분석하기 위해 제안 시스템의 성능 평가를 수행한다. 이를 위해 실제로 교수자가 작성한 대본과 비교하여 LLM이 생성한 대본과 RAG를 연계한 LLM이 생성한 대본을 평가 및 비교 분석한다.

4.1 Experimental Setup

본 절에서는 제안 시스템의 구성 방식, 실험 데이터, 비교 모델 및 평가 지표를 포함한 실험 환경을 상세히 설명한다. 제안하는 RAG 구조는 Table 2와 같이 LangChain 프레임워크를 기반으로 구축되었으며, 강의자료로부터 생성된 지식베이스를 HyperCLOVA X SEED 3B 모델과 연계하여 문맥 검색 및 생성 과정을 수행한다. 또한, LLM을 단독으로 사용한 모델을 비교군으로 설정하여, 동일한 질의 입력에 대해 두 모델의 출력 품질을 정량적으로 비교한다.

지식베이스는 수집된 컴퓨터공학 핵심 교과인 자료구조, 알고리즘, 운영체제, 컴퓨터구조, 데이터베이스의 강의자료를 기반으로 구축하였다. 각 강의자료는 슬라이드 단위로 분리된 텍스트, 이미지 요약, 강의 음성 STT 결과로 구성되어 있으며, 이를 문단 단위로 구조화하여 벡터화하였다. 본 실험에서는 이와 같이 다섯 개의 과목을 평가에 사용한다.

텍스트 임베딩에는 한국어 문장 의미 표현에 최적화된 KoSimCSE 모델을 사용하였으며, 생성된 벡터는 FAISS 라이브러리를 이용하여 인덱싱하였다. 검색 단계에서는 코사인 유사도를 기준으로 top-k를 5로 설정하여 문서를 탐색하고, 이를 LangChain의 검색 모듈을 통해 LLM의 입력 프롬프트에 통합하였다.

평가 지표는 BERTScore[32], ROUGE-L[33], BLEU를 사용한다. BERTScore는 사전학습된 언어 모델의 문장 임베딩을 활용하여 문장 간의 의미적 유사도를 계산하는 지표로, 본 연구에서 제안한 시스템이 강의 대본 생성 시 의미 일관성과 개념 정확도를 얼마나 유지하는지를 평가할 수 있다. ROUGE-L은 두 문장 간의 LCS(Longest Common Subsequence)를 기반으로 문장의 구조적 일치도를 측정하므로, 제안 시스템이 실제 교수자의 설명 구조나 서술 순서를 충분히 반영했는지를 검증하는 데 적합하다. BLEU는 생성된 문장의 표현 정확도와 사실적 일치성을 n-gram 단위의 어휘 일치를 바탕으로 평가하는 대표적인 지표이며, 모델이 참조 대본의 기술적 용어나 표현 패턴을 얼마나 정확하게 재현하는지를 분석할 수 있다. 이 세 지표를 통해, 제안하는 RAG 시스템의 단순한 문장 생성 품질을 넘어 의미적 일관성, 문장의 논리적 구조 유지 및 사실 기반의 정확성을 종합적으로 검증하고자 한다.

Table 3. Performance Comparison between Baseline LLM and RAG-integrated Model

Models	BERT Score	ROUGE-L	BLEU
Model 1 (Llama-only)	0.752	0.426	0.263
Model 2 (Llama + RAG)	0.792	0.472	0.293
Model 3 CIOVAX only	0.827	0.493	0.318
Proposed (CLOVAX+RAG)	0.873	0.526	0.354

Table 4. Script and Video Generation Time Comparison

Models	Script Generation Time	Video Generation Time (include Script)
Model 1 (Llama-only)	182 sec	561 sec
Model 2 (Llama + RAG)	195 sec	574 sec
Model 3 CIOVAX only	106 sec	484 sec
Proposed (CLOVAX+RAG)	113 sec	492 sec

4.2 Experimental Result

4.2.1 Analyzing Quality of Generated Script

Table 3은 Llama 기반 모델(Llama 3 Bllossom 3B)과 Clova X 기반 모델 간의 성능 및 RAG 적용유무에 대한 성능을 나타낸 것이다. 네 가지 모델을 BERTScore, ROUGE-L, BLEU 지표로 비교한 결과로 전반적으로 제안 시스템(CLOVA X + RAG)이 모든 지표에서 가장 높은 값을 기록하였으며, 이는 강의 대본 생성 품질의 의미적, 구조적, 표현적 정확도가 기존 모델 대비 전반적으로 향상되었음을 보여준다.

먼저 BERTScore 기준으로 Llama-only 대비 0.752에서 0.873로 크게 향상되었으며, 동일 모델 내에서 RAG 적용 시에도 각각 Llama, Clova X에서 일관된 개선이 나타났다. 특히 전문용어, 알고리즘 절차, 정의문 등 개념적, 기술적 설명이 포함된 항목에서 향상 폭이 더 컸는데, 이는 제안 시스템이 강의자료에서 추출된 정의문, 예제, 시각요약 정보를 프롬프트 단계에서 통합적으로 주입함으로써 도메인 지식을 반영한 의미적 정합성을 강화했기 때문으로 판단된다.

ROUGE-L 지표에서도 Llama-only 대비 0.426에서 0.526으로 약 10.0% 향상이 확인되었으며, 이는 생성 문장이 기존 대본의 문장 구조와 핵심 표현을 더 높은 비율로 재현함을 보였다. 마찬가지로 BLEU 지표에서도 0.263에서 0.354로 상승하며, 이는 모델이 기술적 용어와 특정 개념을 기존 대본과 유사한 표현으로 생성하는 능력이 높아졌음을 의미한다. 특히 RAG 결합 효과는 Llama에서 0.03 향상, Clova X에서 0.036 향상으로, 대형 모델뿐 아니라 소형 모델에서도 지식 기반 증강이 유효하게 작동함을 확인하였다.

4.2.2 Evaluation of Script and Lecture Video Generation Time

본 절에서는 제안 시스템의 처리 효율을 분석하기 위해, 10장으로 구성된 강의 슬라이드(슬라이드당 약 1분 분량 대본 생성 프롬프트 입력)를 입력하고 모델별 대본 생성(Script Generation) 및 최종 영상 제작(Video Generation) 시간을 측정, 비교한다. Table 4는 각 모델별 강의대본 생성시간 및 최종 동영상 생성시간을 나타낸 결과로, 제안모델은 대본생성 단계에서 113초, Video Generation 단계에서 492초를 기록하였다. 특히 RAG 적용에 따른 지연은 약 6% 수준에 불과해, 지식 기반 확장 구조를 사용하면서도 실시간 제작에 가까운 처리 속도를 유지할 수 있음을 확인할 수 있다. 반면 LLaMA 기반 모델(Model 1, Model 2)은 각각 182초, 195초로 이는 제안 모델 대비 약 1.6배 더 긴 시간을 기록하였다. 이러한 결과는 LLaMA3 Blossom 3B 모델이 한국어 처리 시 더 많은 토큰을 생성함으로써 나타나는 차이로 해석된다.

영상 생성 시간에 대해서 Proposed 모델은 492초가 측정되었는데, 이는 기존의 강의 제작 방식과 비교하여 매우 높은 시간 효율성을 가진다고 할 수 있다. 일반적으로 교수자가 직접 강의를 촬영하는 경우, 강의 내용 준비와 촬영만으로도 최소 수십 분에서 한 시간 이상의 시간이 소요될 수 있다는 점에서 높은 시간적 효율성을 보인다고 할 수 있다.

V. Conclusions

본 연구는 기존 대학 강의자료를 활용하여 자동으로 강의 영상을 생성할 수 있는 시스템을 제안하고, 이를 실현하기 위한 기술적 구조와 실험적 검증 결과를 제시하였다. 제안 시스템은 강의 대본 생성, TTS 기반 음성 합성, 디지털 휴먼 립싱크 합성, 그리고 다국어 자막 생성으로 구성하였으며, 특히 HyperCLOVA X SEED 3B 모델과 RAG를 활용하여, 사실 기반의 고품질 강의 대본을 자동 생성하는 것을 목표로 하였다.

제안한 시스템은 실제 강의자료로부터 텍스트, 이미지를 추출, STT 변환을 거친 음성과 통합하여 구조화된 지

식베이스를 구축하고, LangChain 기반의 RAG 프레임워크를 통해 이를 HyperCLOVA X SEED 3B 모델에 연계하였다. 이를 통해 모델은 외부 지식을 실시간으로 참조하며, 교수자의 설명 구조를 반영한 사실 중심의 강의 대본을 생성할 수 있었다. 결론적으로, RAG를 활용하여 지속적인 파인튜닝 없이도 LLM의 교육적 실용성을 입증하였으며, 향후 AI 기반 강의 자동화와 스마트 교육 플랫폼 구현의 핵심 기술 기반이 될 것으로 기대한다. 본 연구의 의의는 강의자료 기반 지식 확장과 대본, 음성, 영상 생성 과정을 단일 파이프라인으로 통합함으로써, 교육 콘텐츠 제작의 효율성과 정확성을 동시에 향상시킬 수 있는 실증적 가능성을 제시한 데 있다.

다만, 본 연구는 강의자료 기반 자동 대본 생성, 음성 합성, 디지털 휴먼 립싱크, 다국어 자막 생성으로 이어지는 통합 파이프라인의 기술적 타당성을 중심으로 평가하였기 때문에, 생성된 콘텐츠가 실제 학습자에게 제공하는 만족도나 교육적 효과를 직접 측정하지 못한 한계가 있다. 또한 강의자료 전처리 과정에서 슬라이드의 텍스트를 추출하였으나, 이 단계에서 발생할 수 있는 문자 인식 오류 또는 의미 손실에 대한 정량적 분석을 수행하지 못했다는 점도 한계점이다. 텍스트 추출 품질은 RAG 검색 신뢰도와 최종 생성 품질에 직접적인 영향을 미치므로, 향후 연구에서는 추출정보 정확도 평가, 추출 오류의 RAG 전이 효과에 대한 검증이 반드시 필요하다.

이에 따라 향후 연구에서는 다양한 학문 분야의 강의자료를 적용하여 시스템을 고도화하여 학습자 맞춤형 강의 콘텐츠 제작이 가능하게 하며, MOS(Mean Opinion Score)를 활용한 음성, 영상 품질 평가, 사용자 만족도 설문, 학습 효과 분석 등 사용자 중심 평가를 수행할 계획이다. 더 나아가, 전처리 단계에서의 추출된 텍스트의 품질을 세밀하게 분석하고, 오류 보정 기법을 적용하여 전체 파이프라인의 신뢰성과 재현성을 향상시키고자 한다. 이를 향후 연구로 남긴다.

ACKNOWLEDGEMENT

This work was supported by the Digital Innovation Hub project supervised by the Daegu Digital Innovation Promotion Agency(DIP) grant funded by the Korea government(MSIT and Daegu Metropolitan City) in 2025(No.225C000366, 2025년 R&BD협업 프로젝트(상용화 지원)).

REFERENCES

[1] W. Yang, X. Zhang, X. Chen, J. Lu, and F. Tian, “Based case based learning and flipped classroom as a means to improve international students’ active learning and critical thinking ability,” BMC Medical Education, Vol. 24, No. 759, July 2024. DOI: 10.1186/s12909-024-05758-8

[2] M. I. Baig and E. Yadegaridehkordi, “Flipped classroom in higher education: a systematic literature review and research challenges,” International Journal of Educational Technology in Higher Education, Vol. 20, No. 61, November 2023. DOI: 10.1186/s41239-023-00430-5

[3] A. A. Omoniyi, L. C. Jita, and T. Jita, “Teachers’ Experiences with Flipped Classrooms in Senior Secondary Mathematics Instruction,” Computers, Vol. 14, No. 5, p. 180, May 2025. DOI: 10.3390/computers14050180

[4] Y. Wang, G.-J. Wang, L.-J. Yan, J. Gao, C. Fu, and H.-M. Ren, “Qualitative research on the flipped classroom cognition of undergraduate nursing students,” BMC Medical Education, Vol. 24, No. 1460, December 18 2024. DOI: 10.1186/s12909-024-06426-7

[5] R. Deng, S. Feng, and S. Shen, “Improving the effectiveness of video-based flipped classrooms with question-embedding,” Education and Information Technologies, Vol. 29, pp. 12677-12702, 2024. DOI: 10.1007/s10639-023-12303-5

[6] D. Romanow, M. K. Cline, and N. P. Napier, “A Response to COVID: From Traditional to Remote Learning Using a Flipped Classroom Pedagogy and Its Impact on BI Skills Attainment,” Journal of Information Systems Education, Vol. 35, No. 1, pp. 99-111, 2024. DOI: 10.62273/NOBF5942

[7] A. Holmberg, “Generating narrated lecture videos from slides with synchronized highlights,” arXiv preprint arXiv:2505.02966, May 2025. DOI: 10.48550/arXiv.2505.02966

[8] GPT-SoVITS, 2024. [Online]. Available: https://github.com/RVC-Boss/GPT-SoVITS

[9] K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, “A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild,” Proceedings of the 28th ACM International Conference on Multimedia, pp. 484-492, Seattle, USA, October 2020. DOI: 10.1145/3394171.3413532

[10] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” OpenAI Technical Report, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[11] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey et al., “The Llama 3 Herd of Models,” arXiv preprint arXiv:2407.21783, November 2024.

[12] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra et

al., “PaLM: Scaling Language Modeling with Pathways,” Journal of Machine Learning Research, Vol. 24, pp. 240:1-240:113, 2023.

[13] NAVER Cloud HyperCLOVA X Team, “HyperCLOVA X THINK Technical Report,” arXiv preprint arXiv:2506.22403, June 2025. DOI: 10.48550/arXiv.2506.22403

[14] MLP-Lab, LLaMA-3 Bllossom, 2024. [Online]. Available: https://github.com/MLP-Lab/Bllossom

[15] Z. Song, B. Yan, Y. Liu, M. Fang, M. Li, R. Yan, and X. Chen, “Injecting Domain-Specific Knowledge into Large Language Models: A Comprehensive Survey,” arXiv preprint arXiv:2502.10708, 2025. DOI: 10.48550/arXiv.2502.10708

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems, Vol. 33, pp. 9459-9474, 2020.

[17] D. W. Lee, C. Ahuja, P. P. Liang, S. Natu, and L.-P. Morency, “Multimodal Lecture Presentations Dataset: Understanding Multimodality in Educational Slides,” Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision, pp. 1-12, Seoul, Korea, October 2023. DOI: 10.1109/ICCV51070.2023.01838

[18] Pdfminer, 2024. [Online]. Available: https://github.com/pdfminer/pdfminer.six

[19] Chardet, 2024. [Online]. Available: https://github.com/chardet/chardet

[20] TesseractOCR, 2024. [Online]. Available: https://github.com/tesseract-ocr/tesseract

[21] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems, Vol. 33, 2020. DOI: 10.48550/arXiv.2005.14165

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” arXiv preprint arXiv:2212.04356, December 2022. DOI: 10.48550/arXiv.2212.04356

[23] S. Ma, L. Dong, S. Huang, D. Zhang, A. Muzio, S. Singhal, H. H. Awadalla, X. Song, and F. Wei, “ΔLM: Encoder-Decoder Pre-training for Language Generation and Translation by Augmenting Pretrained Multilingual Encoders,” arXiv preprint arXiv:2106.13736, June 2021. DOI: 10.48550/arXiv.2106.13736

[24] K. Li Gong, S. Bae, and O. Kwon, “The Effect of Virtual Human Lecturer’s Human Likeness on Educational Content Satisfaction: Focused on the Theory of Experiential Economy,” The Journal of the Korea Contents Association, Vol. 22, No. 7, pp. 524-539, 2022. DOI: 10.5392/JKCA.2022.22.07.524

[25] Dong-Yeon Park, So-Jeong Kang, Yu-Jin Kim, and Soon-Bum Lim, “Automatic Commentary System of Online Video Lectures for Visually Impaired Students,” *Journal of the HCI Society of Korea*, Vol. 17, No. 2, pp. 31-39, June 2022. DOI: 10.17210/jhsk.2022.06.17.2.31

[26] Ryeong Oh, “A Study of Factors Influencing Intention to Use AI Video Production Technology: Focusing on One-Person Media Producers,” *Korean Journal of Broadcasting and Telecommunication Studies*, Vol. 38, No. 3, pp. 133-173, May 2024. DOI: 10.22876/kab.2024.38.3.004

[27] Mi-young Lee and In-kee Ahn, “Development Study of Elementary Art Classes Using Video Generation AI,” *Art and Education*, Vol. 26, No. 1, pp. 51-77, January 2025.

[28] Se-Hui Yi and Jin Lee, “Types and Trends in Artificial Intelligence-Based Video Content Authoring Tools,” Korean Journal of Digital Content Society, Vol. 25, No. 6, pp. 1589-1600, June 2024. DOI: 10.9728/dcs.2024.25.6.1589

[29] XML Parser, 2024. [Online]. Available: https://github.com/leoz0214/XML-Parser

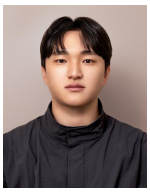
[30] BM-K, “Sentence-Embedding-Is-All-You-Need[Korean Sentence Embedding Repository),” GitHub repository, Mar. 2023. [Online]. Available: https://github.com/BM-K/Sentence-Embedding-is-all-you-need

[31] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, “The FAISS library,” IEEE Transactions on Big Data, 2025.

[32] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” arXiv preprint arXiv:1904.09675, 2019. [Online]. Available: https://arxiv.org/abs/1904.09675

[33] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in Text Summarization Branches Out, Barcelona, Spain, 2004, pp. 74-81. [Online]. Available: https://aclanthology.org/W04-1013.pdf

Authors



SeongHun Kim received B.S. degrees in the Department of Computer Engineering from Yeungnam University, Korea, in 2025. He is currently a M.S. student in the Department Computer Engineering at Yeungnam University. His current research interests include IEEE 802.11 MAC protocol, multi-media streaming over wireless networks.

Oh-Gyu Kwon received his B.S. degree in Computer Engineering from Yeungnam University, Gyeongsan, Korea, in 2002, and his M.S. degree in 2009. He is currently pursuing a Ph.D. in Computer Engineering while serving as the CEO of Mobifin. His recent research interests focus on AI-based application software.

Seong-Guk Nam received the B.S. and M.S. degrees in Computer Engineering from Yeungnam University, Gyeongsan, Korea, in 2021 and 2023, respectively. Since 2023, he has been with the R&D Center at Nearnetworks, where he is currently a research engineer. He is interested in deep learning and Large Language Model.

Seung-Cheol Lee received the M.S. degrees in computer engineering from Yeungnam University, South Korea, in 2021, where he is currently pursuing the Ph.D. degree with the Department of Computer Engineering. His research interests include deep learning, machine learning, big data analysis, natural language processing, and software engineering

Tae-Young Yang is currently pursuing the B.S. degree in the Department of Computer Engineering at Yeungnam University, Korea, from 2020 to 2025, and he is expected to graduate in 2026. He is an undergraduate research student in the Artificial Intelligence and Intelligent Information Systems Laboratory. His current research interests include computer vision and artificial intelligence.

Jae-Woo Ryu received the B.S. degree in Computer Engineering from Yeungnam University College, Daegu, Republic of Korea, in 2014. He has been pursuing an M.S. degree in Computer Engineering at Yeungnam University, Gyeongsan, Korea, since 2025. He is currently a Senior Researcher at the Corporate R&D Center of Nearnetworks. His research interests include AI-based application software, intelligent systems, and data-driven artificial intelligence technologies.

Chang-Hyeon Park received the B.S. degree in Electronics Engineering from Kyungpook University, Korea, in 1986 and M.S. and Ph.D degrees in Computer Science from Seoul University, Korea, in 1988 and 1992, respectively. Dr. Park joined the faculty of the Department of Computer Engineering at Yeungnam University, Gyeongsan, Korea, in 1993. He is currently a Professor in the Department of Computer Engineering at Yeungnam University. He is interested in artificial intelligence, data mining, and embedded system.