

학술 논문 자동요약을 위한 멀티모달 조합별 성능 분석

주민선*, 온병원**

Ablation Study of Multimodal Summarization on Scientific Articles

Min Sun Ju*, Byung-Won On**

본 논문은 2023학년도 국립군산대학교 연구년 연구비 사업에 의하여 연구되었음

요약

HeterGNN, PEGASUS-X 등 기존 학술 논문 요약 모델들이 텍스트 위주의 접근을 취하는 가운데, 최근 GPT-4o, Gemini-2.5 등 멀티모달 언어 모델의 등장으로 도표와 표 등 시각적 정보를 활용한 요약 성능 향상에 대한 관심이 증가하고 있다. 그러나 기존 연구들은 멀티모달 정보 추가가 성능 향상을 보장한다는 가설을 체계적으로 검증하지 않았다. 본 연구는 GPT-4o, Grok-2, Gemini-2.5를 활용하여 논문의 텍스트, 도표, 표를 다양한 조합으로 구성한 멀티모달 요약의 효과를 체계적으로 분석하였다. 실험 결과, GPT-4o와 Gemini-2.5에서는 각각 6.8%와 5.9%의 성능 향상을 보인 반면 Grok-2에서는 5.2% 감소하여 모델별 상이한 결과를 확인하였다. 이를 통해 멀티모달 정보 추가가 항상 성능 향상을 보장하지 않으며, 기존 평가 지표의 한계를 발견하고 새로운 평가 프레임워크의 필요성을 제시하였다.

Abstract

Existing academic paper summarization models such as HeterGNN and PEGASUS-X primarily adopt text-based approaches, while recent emergence of multimodal language models like GPT-4o and Gemini-2.5 has increased interest in leveraging visual information such as figures and tables to improve summarization performance. However, previous studies have not systematically verified the hypothesis that adding multimodal information guarantees performance improvement. This study systematically analyzes the effectiveness of multimodal summarization by extracting text, figures, and tables from papers and feeding various combinations to GPT-4o, Grok-2, and Gemini-2.5. Experimental results show performance improvements of 6.8% and 5.9% for GPT-4o and Gemini-2.5 respectively, while Grok-2 exhibits a 5.2% decrease, confirming varying results across different models. These findings demonstrate that adding multimodal information does not always guarantee performance improvement, and reveal limitations of existing evaluation metrics, presenting the necessity for a new evaluation framework.

Keywords

multimodal summarization, multimodal language models, ablation study, scientific article summarization

* 국립군산대학교 소프트웨어학과 학사과정
- ORCID: <https://orcid.org/0009-0002-4080-7629>
** 국립군산대학교 소프트웨어학과 교수(교신저자)
- ORCID: <https://orcid.org/0000-0001-6929-1388>

· Received: Oct. 15, 2025, Revised: Nov. 26, 2025, Accepted: Nov. 29, 2025
· Corresponding Author: Byung-Won On
Dept. of Software Science and Engineering, Kunsan National University,
558, Dachak-ro, Gusan, Jeollabuk-do, Korea
Tel.: +82-63-469-8913, Email: bwon@kunsan.ac.kr

1. 서론

최신 기술을 습득하기 위한 최적의 자료는 학술 논문이다. 하지만 여러 학회에서 수많은 논문들이 발간되기 때문에, 빠른 시간 내에 사람이 직접 논문을 읽으며 핵심을 파악하기에는 어려움이 있다. 이에 따라 학술 논문의 핵심 내용을 빠르게 전달할 수 있는 자동 요약 기술의 중요성이 점점 더 부각되고 있다. 하지만 학술 논문 요약에는 여러 어려움이 존재한다. 먼저 학술 논문은 전문 용어와 복잡한 개념이 많다[1]. 기본 지식이나 배경이 부족할 경우 학술 논문을 요약하는 데 어려움이 있다. 두 번째로는 분량이 매우 많다는 점이다[2]. 짧게는 10페이지에서 길게는 수십 페이지의 분량으로 구성되어 있는 논문이 많기 때문에 방대한 분량 또한 학술 논문을 파악하기 어렵게 만든다. 또한 수식, 도표, 표 등 다양한 멀티모달 정보가 존재하고 중복된 서술과 핵심 정보가 분산되어 있는 경우가 많아 좋은 성능의 요약문을 자동으로 생성하기에는 어려움이 존재한다[2]-[4].

기존 주요 요약 모델인 BART, T5와 같은 Transformer 기반 모델은 텍스트 위주의 요약 모델이다. 이러한 텍스트 위주의 요약 모델은 짧은 텍스트 요약에서는 좋은 성능을 보이지만, 멀티모달 논문 요약에는 부적합하다. 이는 시각적 모달리티를 충분히 반영하지 않기 때문인데, 논문과 같은 환경에서는 멀티모달 정보(도표와 표)가 핵심 정보를 담고 있어 이를 반영하지 않으면 요약이 부정확해진다. 따라서 학술 논문 요약에는 멀티모달 정보가 반드시 필요하며, 더 중요한 것은 어떤 텍스트 섹션과 시각적 요소의 조합이 최적의 요약 성능을 달성하는지에 대한 체계적 분석이 필요하다.

최근 뉴스, 리뷰, 논문 등 다양한 도메인에서 문서 자동 요약 시 멀티모달 정보를 활용하는 연구가 점점 많아지고 있다. 이러한 연구들은 멀티모달 정보를 활용하면 기존 텍스트 중심 요약 방식보다 성능이 향상될 것이라는 가설을 바탕으로 요약 모델을 제시하고 있다. 하지만 대부분의 연구에서는 이러한 가설이 체계적이고 정확한 검증 과정을 거치지 않은 채 당연한 것으로 받아들여지고 있는 실정

이다. 멀티모달 정보의 융합이 실제로 어떤 조건에서, 어느 정도의 성능 향상을 가져오는지에 대한 명확한 검증이 부족한 상황이다.

따라서 본 연구에서는 이러한 가설을 정확히 검증하고 최적의 조합을 찾기 위해 다음과 같은 체계적인 Ablation Study 접근 방법을 제안한다. 첫째, 자연어처리 분야의 학회 논문들을 표준 데이터셋으로 활용하여 실험의 재현성과 신뢰성을 확보한다. 둘째, Introduction, Methodology, Proposal Experimental Setup & Results, Conclusion의 4가지 섹션과 도표, 표 등 시각적 요소의 가능한 조합을 포괄적으로 실험한다. 셋째, GPT-4o, Grok-2, Gemini-2.5 등 최신 멀티모달 거대 언어 모델들을 활용하여 객관적인 성능 비교를 수행한다. 넷째, 다양한 입력 조합으로 인해 과도하게 증가하는 실험 데이터의 문제를 해결하기 위해 클러스터링과 필터링 기법을 적용하여 효율적인 실험 환경을 구축한다. 다섯째, 정량적 평가와 더불어 실제 멀티모달 요약의 품질을 다각도로 분석하고, 평가 지표의 한계점을 명확히 규명한다.

이러한 체계적인 Ablation Study를 통해 멀티모달 학술 논문 요약에서 성능을 최대화하는 최적 조합 전략을 도출하고, 주요 멀티모달 요약 모델에 대한 실증적 근거를 제시하고자 한다.

II. 관련 연구

최근 문서의 자동 요약 분야에서 멀티모달을 활용한 연구가 활발히 진행되고 있다. 특히 2022년 이후 텍스트와 이미지 정보를 통합적으로 활용하는 연구들이 급격히 증가하고 있는 추세를 보인다.

S. Syed et al.[5]는 텍스트 요약 분야의 다면적 문헌 탐색을 통해 요약 기술의 발전 과정을 체계적으로 분석하였다. J. Zhu et al.[6]은 멀티모달 출력을 가진 멀티모달 요약 시스템을 제안하여 과학 문서의 텍스트와 시각 정보를 통합적으로 생성하는 프레임워크를 제시하였다. H. Shakil et al.[7]은 추상적 텍스트 요약의 현황과 도전 과제, 개선 방향을 종합적으로 검토하여 요약 기술 발전의 이론적 기반을 마련하였다.

한국어 문서 요약 분야에서도 다양한 연구가 진행되었다. G. H. Lee et al.[8]은 뉴스 기사와 소셜 미디어를 활용하여 약 425,000개의 한국어 문서 요약 데이터를 구축하였으며, 제목, 부제를 요약으로 활용하는 새로운 방법론을 제안하였다[8]. Y. Lee et al.[9]은 한국어 특화 문서 요약을 위한 TextRank 기반 기법을 설계하여 KoNLPy의 okt 형태소 분석기와 SentenceTransformer를 활용한 한국어 최적화 요약 시스템을 제안하였다.

과학 논문 요약 분야에서는 구조적 특성을 고려한 연구들이 진행되었다. A. Cohan and N. Goharian [10]은 인용 문맥과 논문의 담화 구조를 활용한 과학 논문 요약 기법을 제안하였으며, G. Sharma et al.[11]은 추출적 접근과 추상적 접근을 결합하여 과학 요약을 생성하는 방법론을 제시하였다. 국내에서는 완전성과 간결성을 고려한 텍스트 요약 품질의 자동 평가 기법 연구가 요약 품질 평가의 새로운 관점을 제시하였다[12].

텍스트 요약의 성능 향상을 위한 다양한 접근법도 연구되었다. R. C. Belwal et al.[13]은 주제 기반 벡터 공간 모델과 의미적 측정을 활용한 텍스트 요약 기법을 제안하였고, T. Wang et al.[14]는 지역 주제와 계층적 정보를 통합하여 긴 문서의 추출적 요약 성능을 향상시키는 방법을 연구하였다.

멀티모달 요약 연구에서는 시각적 요소의 활용이 주목받고 있다. T. Gigant et al.[15]는 비전-언어 모델을 활용한 멀티모달 프레젠테이션 요약에서 모달리티와 구조의 영향을 분석하였으며, P. Janjani et al.[16]은 다중 소스, 다중 모달, 다중 언어 융합을 통한 정보 추출 및 요약 기술을 제안하여 멀티모달 정보 통합의 새로운 방향을 제시하였다. A. Saha et al.[17]은 다양한 관점을 통합하여 단일 요약을 생성하는 접근법을 제안하였다.

하지만 기존 연구들은 몇 가지 중요한 한계점을 보인다. 첫째, 대부분의 멀티모달 요약 연구가 시각적 정보 추가의 효과를 당연한 것으로 가정하고 있어, 실제로 멀티모달 정보가 요약 성능에 미치는 영향에 대한 체계적이고 실증적인 검증이 부족하다. 둘째, 논문의 어떤 섹션 조합이 최적의 요약을 생성하는지에 대한 포괄적인 분석이 이루어지지 않았으

며, 섹션별 기여도와 상호작용 효과가 명확히 규명되지 않았다. 셋째, 다양한 멀티모달 언어모델(LVLM) 간의 멀티모달 처리 능력에 대한 객관적이고 체계적인 성능 비교가 부족하다. 넷째, 기존 ROUGE와 BERTScore 등의 평가 지표가 멀티모달 요약의 품질을 제대로 측정하지 못하는 근본적인 한계에 대한 분석이 미흡하다.

이러한 한계점들을 극복하기 위해 본 연구에서는 다음과 같은 차별화된 접근 방법을 제안한다. 첫째, 멀티모달 정보 활용의 효과를 실증적으로 검증하기 위해 Text Only와 Text+Visuals 방식을 체계적으로 비교 분석한다. 둘째, 15가지 서로 다른 섹션 조합(Introduction, Methodology, Proposal Experimental Setup & Results, Conclusion)에 대해 포괄적인 성능 분석을 수행하여 최적 조합을 도출한다. 셋째, GPT-4o, Gemini-2.5, Grok-2 등 3개 주요 LVLM의 멀티모달 처리 능력을 동일한 조건에서 객관적으로 비교 평가한다.

III. 멀티모달 요약 분석 프레임워크

3.1 데이터 전처리

본 연구는 자연어처리 분야에서 권위 있는 국제 학회인 EMNLP 2024에서 발표된 1,267편의 논문을 데이터셋으로 활용하였다. EMNLP 2024 논문을 선택한 이유는 해당 논문들에 포함된 이미지에 대한 라벨링 작업을 사전에 완료한 데이터를 보유하고 있어, 멀티모달 요약 연구를 위한 시각적 요소 분석에 필수적인 구조화된 데이터를 즉시 활용할 수 있었기 때문이다. 1,267편의 전체 논문을 대상으로 포괄적인 분석을 수행하는 것은 현실적 제약으로 인해 어려움이 있어, 효율적이고 체계적인 실험을 위해 논문을 카테고리별로 분류한 후 두 가지 선별 기준을 적용하여 각 카테고리에서 4편씩 선정하였다.

먼저 1,267편의 논문을 주제별로 분류하기 위해 키워드 기반 규칙 분류(Rule-based classification)를 수행하였다. 각 논문의 PDF 첫 페이지에서 제목을 자동 추출한 후, 사전에 정의한 13개 토픽별 키워드 목록과 매칭하여 분류하였다. 제목 추출 시에는

"Proceedings", "Author", "Copyright" 등 메타데이터 정보를 제외하고, 8~180자 길이의 대문자로 시작하는 자연스러운 문장 형태의 텍스트를 제목 후보로 선정하였다. 분류 토픽은 Language Modeling, Information Extraction, Question Answering, Machine Translation, Summarization, Text Generation, Sentiment & Opinion, Information Retrieval, Syntax & Parsing, Speech & Multimodal, Ethics & Fairness, Resource & Benchmark, Application의 13개 주제와 기타를 포함한 Others로 구성하였다. 각 토픽은 해당 분야를 대표하는 핵심 키워드로 정의되었으며, 예를 들어 Language Modeling 토픽은 "language model", "pre-training", "transformer", "bert", "gpt-4o", "llm" 등의 키워드를, Summarization 토픽은 "summarization", "summarize", "summary", "abstractive", "extractive" 등의 키워드를 포함하였다. 논문 제목에 해당 키워드가 포함되면 해당 토픽으로 분류되며, 어떤 토픽에도 속하지 않는 경우 Others로 분류되었다. 이러한 키워드 기반 분류 방식은 각 논문의 주제를 명확하고 일관되게 구분할 수 있는 장점이 있다.

분류 결과, 1,267편의 논문은 총 14개 토픽(13개 토픽+Others)으로 분류되었다. 이 중 Others 토픽은 명확한 주제 분류가 어려운 다양한 논문들이 혼재되어 있어 일관된 분석 기준을 적용하기 어렵다고 판단하여 실험 대상에서 제외하였다. 따라서 13개의 명확한 토픽을 대상으로 실험을 진행하였으며, 실험 논문 수를 체계적으로 줄이기 위해 2가지 선별 기준을 설정하였다.

첫 번째 기준은 논문의 시각적 요소를 정량화하는 것이다. PDF 스캔을 통해 각 논문의 Figure 개수와 Table 개수를 카운트하여 Total(Figure+Table) 수를 계산한 후, 토픽별로 Total이 가장 많은 상위 4편의 논문을 선정하였다. 이를 통해 총 52편의 논문을 확보하였다. 도표와 표가 많이 포함된 논문일수록 텍스트 단독 요약과 멀티모달 요약 간의 성능 차이를 명확히 관찰할 수 있을 것으로 예상하였다.

두 번째 기준은 논문의 텍스트 분량을 고려한 것이었다. PDF에서 텍스트를 추출하여 워드 토큰을 카운트한 후, 토픽별로 워드 토큰이 가장 많은 상위 4편의 논문을 선정하였다. 텍스트 분량이 많은 논문

을 선택한 이유는 충분한 텍스트 정보를 바탕으로 한 요약 성능 평가가 가능하며, 동시에 멀티모달 정보가 추가되었을 때의 성능 향상을 보다 명확히 측정할 수 있기 때문이다. 또한 텍스트 분량이 많은 논문일수록 요약문의 품질 차이가 통계적으로 유의미하게 나타날 가능성이 높다고 판단하였다.

이 2가지 기준을 통해 각각 52편의 논문을 선정하여 최종적으로 총 104편의 논문을 확보하였다. 이는 Others를 제외한 13개 토픽에서 각각 8편씩 균등하게 분포되도록 구성되었으며, 시각적 요소가 풍부한 논문과 텍스트 분량이 많은 논문으로 구분되어 멀티모달 요약의 효과를 다각도로 분석할 수 있는 균형 잡힌 실험 환경을 제공한다.

선별된 104편의 논문에 대해서는 PDF 텍스트 추출 및 섹션 분리 과정을 수행하였다. 먼저 각 논문에서 References를 제외한 본문 텍스트를 추출한 후, 4개의 주요 섹션으로 분류하였다. Introduction과 Conclusion 섹션은 논문에 명시된 해당 섹션의 내용을 그대로 활용하였고, Proposal Experimental Setup & Results 섹션은 논문 내 실험 설계 및 실험 결과와 관련된 내용을 포함하였다. Methodology 섹션은 위 3개 섹션에 해당하지 않는 나머지 내용으로, 주로 방법론, 이론적 배경, 관련 연구 등의 내용이 포함되었다. 분리된 각 섹션은 후속 실험에서 개별적으로 사용되거나 다양한 조합으로 구성되어 멀티모달 요약 성능 분석을 위한 기초 데이터로 활용되었다.

이러한 체계적인 데이터 전처리 과정을 통해 분야별로 요구되는 시각적 정보와 텍스트 정보의 비중이 상이함을 확인할 수 있으며, 이는 멀티모달 요약 성능 평가에서 분야별 특성을 고려한 실험 설계의 필요성을 시사한다.

3.2 제안 프레임워크

멀티모달 정보의 효과를 정량적으로 분석하기 위해 2가지 주요 실험 조건을 설정하였다. 그림 1은 본 연구의 전체 실험 설계 과정을 보여준다.

첫 번째는 논문 선정 기준에 따른 비교로, 3.1절에서 설명한 바와 같이 EMNLP 2024 논문 1,267편

을 14개 토픽으로 클러스터링한 후 Others를 제외한 13개 토픽에서 논문을 선정하였다. 그림 1의 상단에서 볼 수 있듯이, 도표와 표 개수가 많은 논문 상위 4편(이미지가 많은 논문)과 단어 수가 많은 논문 상위 4편(토큰이 많은 논문)을 토픽별로 선정하여 총 104편의 논문을 확보하였다. 이를 통해 이미지가 많은 논문 52편과 토큰이 많은 논문 52편으로 구분하였다.

두 번째는 요약 생성 방식에 따른 비교로, 텍스트 정보만 활용한 Text Only 방식(Text section)과 텍스트와 이미지 정보를 함께 활용한 Text+Visuals 방식(도표와 표 추출)으로 구분하였다. 이미지 데이터는 사전에 라벨링이 완료된 Figure와 Table 이미지 데이터셋을 활용하였으며, 각 논문의 이미지는 카테고리별 폴더 구조로 체계적으로 정리되어 있어 파일명에 포함된 키워드를 통해 유형을 자동으로 식별할 수 있었다.

각 논문에서 추출한 4개 섹션(Introduction, Methodology, Proposal Experimental Setup & Results, Conclusion)을 총 15가지 섹션 조합을 생성하였다. 그림 1의 우측에 표시된 바와 같이, 1개 섹션, 2개 섹션 조합, 3개 섹션 조합, 전체 섹션 조합의 4가지 유형으로 구성하였다. 이러한 조합들은 논문 104편에 대해 프롬프트 입력으로 구성되어, 그림 1 하단의 LVLM(GPT-4o, Grok-2, Gemini-2.5) 3개 모델에 각각 입력되었다. 실험에 사용된 모델들은 현재 시

중에서 가장 대중적이고 널리 활용되는 멀티모달 대규모 언어 모델들로 선정하였으며, 구체적으로 GPT-4o, Grok-2, Gemini-2.5 버전을 선택한 것은 API 기반 대규모 실험의 비용 효율성을 고려하여 각 회사에서 제공하는 가장 경제적인 모델을 활용하기 위함이었다.

멀티모달 프롬프트 구성을 위해서는 API 전송에 적합한 이미지 정규화 처리를 수행하였다. 모든 이미지는 최대 1024×1024 픽셀 크기로 조정되고 RGB 색상 공간으로 변환된 후, JPEG 형식으로 압축하여 base64로 인코딩하였다. 각 모델의 API 특성에 맞춰 GPT-4o와 Grok-2는 OpenAI 호환 API를 통해 이미지를 전송하였으며, Gemini-2.5는 Google의 GenerativeAI API를 활용하였다. 프롬프트는 "Based on the following original abstract and additional images, write an improved abstract" 형식으로 표준화하였고, Figure와 Table을 개별적으로 처리하여 각각의 효과를 분석할 수 있도록 구성하였다.

생성된 요약문은 Abstract를 정답 요약문(Reference summary)로 하여 ROUGE-1/2/L와 BERTScore 지표로 성능을 측정하였다. 평가 지표 선정의 근거로, ROUGE는 텍스트 요약 분야에서 가장 널리 사용되는 평가 지표로, ROUGE-1은 단어 수준의 일치도를, ROUGE-2는 연속된 두 단어 조합의 일치도를, ROUGE-L은 가장 긴 공통부분 수열을 기반으로 문장 구조의 유사성을 측정한다.

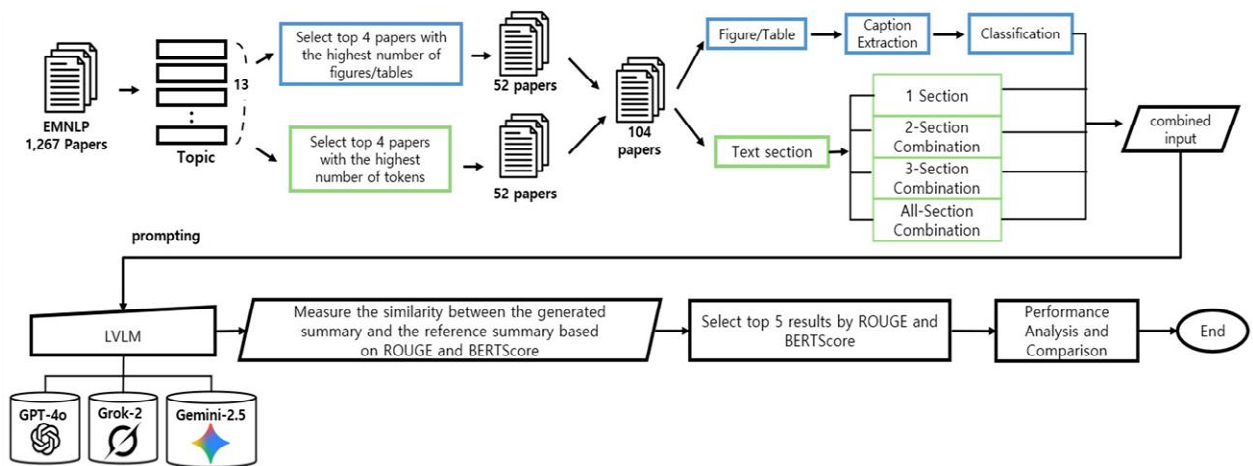


그림 1. 멀티모달 학술 논문 요약 실험 설계 및 평가 프로세스

(입력 구성 예시: PDF에서 추출된 텍스트 섹션 조합과 해당 논문의 도표 및 표를 결합하여 LLM에 요약문 생성 요청)

Fig. 1. Framework of ablation study for multimodal academic paper summarization

(Example input: PDF text sections combined with figures and tables from the paper for LLM summarization)

그러나 ROUGE는 표면적인 어휘 일치에만 의존하여 의미적 유사성을 충분히 반영하지 못하는 한계가 있기 때문에 BERTScore를 추가로 활용하였다. BERTScore는 사전 훈련된 BERT 모델의 임베딩을 활용하여 의미적 유사성을 측정하는 지표로, 학술 논문의 경우 동일한 개념을 다양한 전문 용어로 표현하는 특성을 고려할 때 의미 기반 평가가 필수적이다. 이러한 이중 평가 체계를 통해 표면적 일치도와 의미적 일치도를 모두 검증하여 단일 지표의 편향성을 최소화하고 요약 품질에 대한 균형 잡힌 평가를 수행하였다.

IV. 실험 결과 및 분석

4.1 Text Only vs Text+Visuals

멀티모달 정보 추가가 요약 성능에 미치는 영향을 분석하기 위해 논문 특성별로 Text Only 방식과 Text+Visuals(Text+Multimodal) 방식의 최고 성능을 비교하였다. 본 연구에서는 EMNLP 2024에서 발표된 1,267편의 논문 중 이미지 개수와 텍스트 분량을 기준으로 104편을 선정하여 실험을 수행하였다. 선정된 논문들은 "UNIGEN: Universal Domain Generalization"(Information Extraction 분야), "What the Harm? Quantifying the Tangible Impact of Gender Bias"(Ethics & Fairness 분야), "Is C4 Dataset Optimal for Pruning?"(Resource & Benchmark 분야) 등 13개 학술 카테고리에 걸쳐 분포되어 있으며, 각 논문은 제목, 카테고리, 파일명 정보와 함께 체계적으로 관리되었다. 표 1은 이미지가 많은 논문과 텍스트가

많은 논문에서 각각 최고 성능을 달성한 조합의 평가 지표와 멀티모달 정보 추가에 따른 변화율을 보여준다.

표 1에서 볼 수 있듯이, 이미지가 많은 논문에서는 멀티모달 정보 추가 시 모든 평가 지표에서 일관된 성능 향상을 보였다. ROUGE-L은 0.1855에서 0.1924로 3.72% 향상되었으며, ROUGE-1은 5.73%, ROUGE-2는 1.95%, BERTScore는 0.07% 향상되었다. 이러한 수치들의 실질적 의미를 해석하면, ROUGE-L의 3.72% 향상은 생성된 요약문이 정답 요약문과의 최장 공통부분 수열(Longest Common Subsequence) 일치도에서 약 17개 단어 수준의 추가 매칭이 발생했음을 의미한다. 이는 평균 300단어로 구성된 Abstract에서 실질적으로 체감 가능한 개선 수준이다. ROUGE-2의 1.95% 향상은 bi-gram 수준에서 약 11개의 추가 단어 쌍이 정확히 일치함을 의미하며, 이는 핵심 개념과 전문 용어의 정확한 재현이 개선되었음을 나타낸다. 이러한 결과는 이미지가 많은 논문(예: "SEG2ACT: Global Context-aware Action Generation" 등 Information Extraction 분야 논문들)에서 시각적 정보가 요약 품질 향상에 실질적으로 기여할 수 있음을 의미한다.

텍스트가 많은 논문에서는 표 1의 하단 행을 보면, ROUGE-L이 0.1820에서 0.2411로 무려 32.47%라는 극적인 향상을 보였다. 이 수치의 실질적 의미는 매우 중요하다. ROUGE-L 점수가 0.1820에서 0.2411로 상승했다는 것은 bi-gram 수준의 정확한 표현 일치도가 절대값으로 약 18개의 단어 쌍이 추가로 매칭되었음을 의미한다.

표 1. 논문 특성별 멀티모달 정보 추가 결과

Table 1. Performance comparison of LLMs using text-only and text+visuals

Paper Characteristics	Method	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Papers with many Images	Text Only	0.2899	0.5015	0.1855	0.8537
	Text+Visuals	0.3065	0.5113	0.1924	0.8543
	Improvement	+5.73%	+1.95%	+3.72%	+0.07%
Papers with many Tokens	Text Only	0.2814	0.5189	0.1820	0.8678
	Text+Visuals	0.2930	0.5559	0.2411	0.8767
	Improvement	+4.12%	+7.13%	+32.47%	+1.03%

일반적으로 학술 논문 요약에서 ROUGE-L의 0.05 이상 향상은 요약 품질의 실질적 개선으로 간주되는데, 본 실험에서는 그 기준의 약 12배에 달하는 개선을 보였다. 이는 멀티모달 정보가 "proposed method", "baseline model", "significant improvement"와 같은 학술적 표현의 정확한 재현을 크게 개선시킴을 의미한다. ROUGE-2도 7.13% 향상되어 이미지가 많은 논문(1.95%)보다 더 큰 향상폭을 보였는데, 이는 개별 키워드 수준에서 약 25개의 추가 단어가 정확히 매칭되었음을 나타낸다. 그러나 ROUGE-1 향상폭은 4.12%, BERTScore 향상폭은 1.03%로 상대적으로 미미하여, 평가 지표 간 불일치가 관찰되었다. 특히 주목할 점은 BERTScore의 1.03% 향상이 절대값으로는 약 0.009포인트에 불과하다는 것인데, BERTScore의 일반적 변동 범위가 ± 0.005 포인트임을 고려하면 이는 실용적 측면에서는 미미한 수준이다. 이러한 ROUGE-L의 극적인 향상과 다른 지표들의 미미한 향상 간의 격차는 현재 평가 방법론의 한계를 시사하는 중요한 발견이다. ROUGE-2는 bi-gram 수준의 표면적 일치만을 측정하므로 멀티모달 정보로 인한 표현의 다양성이나 의미적 정확성은 충분히 반영하지 못한다.

4.2 시각 정보 유형별 성능 비교

멀티모달 정보가 요약 성능에 미치는 영향을 정량적으로 분석하기 위해, Figure와 Table을 개별적으로 사용했을 때와 함께 사용했을 때의 성능을 비교하였다. 표 1에서 선정된 상위 5개 조합에 대해 동

일한 논문과 섹션 조합에서 시각 정보 유형만 달리 하여 실험을 수행하였다. 분석 대상 논문에는 "UNIGEN: Universal Domain Generalization" (Information Extraction 분야), "Is C4 Dataset Optimal for Pruning?" (Resource & Benchmark 분야) 등이 포함되었다. 각 조합당 GPT-4o, Grok-2, Gemini-2.5 3개 모델에 대해 실험하여 평균값을 산출하였다.

표 2는 상위 5개 조합 각각에 대해 동일한 논문과 섹션 조합에서 시각 정보 유형만 변경하여 측정된 ROUGE-L 성능을 비교한 결과이다. Figure+Table을 함께 사용한 경우 평균이 0.4137로 가장 높았으며, Figure만 사용한 경우(0.3250) 대비 27.29%, Table만 사용한 경우(0.3382) 대비 22.33%의 성능 향상을 보였다. 이 수치의 실질적 의미를 구체적으로 해석하면, ROUGE-L 0.4137은 생성된 요약문이 Abstract와 평균 124단어(300단어 기준 41.37%)의 최장 공통 부분 수열을 공유함을 의미한다. 반면 Figure만 사용시 0.3250은 약 98단어(32.50%)의 일치율을 나타내므로, Figure+Table 조합은 약 26개 단어의 추가적인 정확한 순서 매칭을 제공한다. 학술 논문 요약에서 26개 단어는 일반적으로 23개의 핵심 문장에 해당하며, 이는 연구 방법론이나 주요 결과를 추가로 정확히 표현할 수 있는 수준이다. 선행 연구들이 보고한 ROUGE-L 개선폭이 일반적으로 5~10% 범위임을 고려할 때, 본 연구의 22~27% 향상은 실용적으로 매우 유의미한 수준이다. 이는 시각 정보를 언어로 변환하는 과정에서 발생하는 Abstract 표현 방식과의 불일치를 Figure와 Table의 상호보완적 정보 제공으로 극복할 수 있음을 시사한다.

표 2. 도표와 표 결합 사용 시 ROUGE-L 성능 향상 결과

Table 2. ROUGE-L performance improvement when using combined figures and tables

Combination	Figure+Table	Figure Only (Avg)	Improvement over Figure Only	Table Only (Avg)	Improvement over Table Only
introduction + conclusion	0.4221	0.2938	+43.66%	0.3365	+25.43%
introduction	0.4208	0.3352	+25.53%	0.3247	+29.60%
introduction+methodology+experimental setup & results	0.4249	0.3153	+34.76%	0.3431	+23.84%
conclusion	0.4211	0.3364	+25.18%	0.3516	+19.77%
introduction+methodology+conclusion	0.3798	0.3445	+10.25%	0.3355	+13.20%
Average	0.4137	0.3250	+27.29%	0.3382	+22.33%

동일 조합 기준으로 비교한 결과, 모든 조합에서 Figure+Table이 단독 사용 대비 우수한 성능을 보였으며, 향상률은 조합에 따라 10.25%에서 43.66%까지 크게 차이가 났다. Figure+Table은 Figure만 사용한 경우 대비 평균 27.29%, Table만 사용한 경우 대비 평균 22.33%의 성능 향상을 나타냈다. Table 단독 사용이 Figure 단독 사용보다 평균 4.06%포인트 높은 성능을 보였으나, 조합에 따라 우열이 바뀌어 일관된 경향을 보이지 않았다.

Introduction+Conclusion 조합에서는 Figure+Table이 Figure만 사용 시 대비 43.66%의 향상으로 가장 큰 개선 효과를 보인 반면, Introduction+Methodology+Conclusion 조합에서는 10.25% 향상으로 상대적으로 작은 향상폭을 나타냈다. 이러한 향상폭의 차이는 실질적으로 중요한 의미를 가진다. Introduction+Conclusion 조합에서 43.66% 향상은 ROUGE-L 점수가 0.2938에서 0.4221로 약 0.1283포인트 증가함을 의미하며, 이는 약 38개 단어의 추가 매칭에 해당한다. 반면 Introduction+Methodology+Conclusion 조합에서 10.25% 향상은 0.3445에서 0.3798로 약 0.0353포인트 증가로 약 11개 단어의 추가 매칭을 의미한다. 전자의 경우 요약문에 약 3~4개의 핵심 문장이 추가로 정확히 표현되는 반면, 후자는 1개 문장 수준의 개선이다. 이는 섹션 조합의 특성과 시각 정보 유형 간의 상호작용이 성능에 영향을 미침을 시사한다.

Table만 사용한 경우가 Figure만 사용한 경우보다

대부분의 조합에서 우수한 성능을 보였으나, Figure+Table 대비 개선 효과의 차이는 평균 4.96%(27.29% 대 22.33%)포인트로 제한적이었다. 절대값으로 환산하면 Figure+Table과 Table 단독 사용의 성능 차이는 평균 0.0755포인트(0.4137 대 0.3382)이며, 이는 약 23개 단어의 차이에 해당한다. 이는 2~3개 문장 수준의 표현 차이로, 실용적 관점에서 유의미한 개선이다. 이는 Table의 구조화된 수치 데이터가 텍스트 표현으로 변환하기 용이하여 Abstract의 서술 방식과 더 유사한 반면, Figure의 시각적 패턴은 언어화 과정에서 상대적으로 더 많은 정보 손실이 발생하기 때문으로 해석된다.

주목할 만한 사례로, Introduction+Methodology+Proposal Experimental setup & results 조합의 일부 논문에서 Table만 사용하고 Gemini-2.5 모델을 적용한 경우 Figure+Table을 함께 사용한 경우보다 높은 성능을 기록하였다[18]. 이는 해당 논문의 핵심 정보가 Table에 집중되어 있어 Figure 정보가 오히려 노이즈로 작용한 것으로 분석되며, 시각 정보의 품질과 관련성이 단순한 정보량보다 중요함을 보여준다.

표 3은 동일한 조합에 대해 BERTScore로 측정된 의미적 유사도 성능을 비교한 결과이다. Figure+Table을 함께 사용한 경우 평균이 0.8633으로 가장 높았으며, Figure만 사용한 경우(0.8546) 대비 1.02%, Table만 사용한 경우(0.8566) 대비 0.78%의 향상을 보였다. 이는 절대값으로 각각 0.0087포인트, 0.0067포인트의 증가를 의미한다.

표 3. 도표와 표 결합 사용 시 BERTScore 성능 향상 결과
Table 3. BERTScore performance improvement when using combined figures and tables

Combination	Figure+Table	Figure Only (Avg)	Improvement over Figure Only	Table Only (Avg)	Improvement over Table Only
introduction + conclusion	0.8728	0.8562	+1.94%	0.8570	+1.84%
introduction	0.8692	0.8590	+1.19%	0.8552	+1.64%
introduction+methodology+experimental setup & results	0.8621	0.8493	+1.51%	0.8556	+0.76%
conclusion	0.8544	0.8493	+0.60%	0.8541	+0.04%
introduction+methodology+conclusion	0.8582	0.8592	-0.12%	0.8611	-0.34%
Average	0.8633	0.8546	+1.02%	0.8566	+0.78%

BERTScore는 0~1 범위의 점수이므로, 1% 내외의 향상은 실질적으로 미미한 수준이다. BERTScore의 전형적인 표준편차가 ± 0.005 임을 고려할 때, 이러한 향상은 통계적 유의성이 제한적일 수 있다. 의미적 임베딩 공간에서 0.01 미만의 차이는 일반적으로 동일한 의미 영역 내의 미세한 변동으로 해석된다.

특히 Introduction+Methodology+Conclusion 조합에서는 오히려 Figure+Table이 Figure만 사용 시 대비 0.12%, Table만 사용 시 대비 0.34% 낮은 성능을 나타냈다. 이는 의미적 유사도 측면에서는 멀티모달 정보의 추가가 항상 긍정적인 효과를 보이는 것은 아님을 의미한다. 그러나 ROUGE-L의 큰 향상폭(평균 22~27%)과 대조적으로 BERTScore의 작은 향상폭은, Figure+Table 조합이 의미적 유사성보다는 Abstract의 정확한 표현 재현에서 더 큰 우위를 보임을 입증한다. 이는 멀티모달 정보가 의미의 정확성보다는 표현의 정확성 개선에 더 효과적임을 시사하며, 실용적으로는 생성된 요약문이 Abstract와 유사한 문체와 용어를 사용하게 됨을 의미한다.

조합별로 살펴보면, Introduction+Conclusion 조합에서 Figure+Table은 Figure만 사용 시 대비 1.94%, Table만 사용 시 대비 1.84%의 향상을 보여 가장 큰 개선 효과를 나타냈다. 반면 Conclusion 단독 조합에서는 Figure만 사용 시 대비 0.60%, Table만 사용 시 대비 0.04%의 미미한 향상을 보였다. 이는 섹션 조합에 따라 멀티모달 정보의 의미적 기여도가 다르게 나타남을 보여준다. Table이 Figure보다 Figure+Table 대비 약간 작은 성능 격차를 보이는 이유는 다음과 같이 분석된다. 첫째, Table은 구조화된 수치 데이터로 구성되어 "accuracy 92.3%"와 같이 텍스트로 직접 표현이 가능하다. 둘째, Abstract도 주로 수치 기반 서술을 사용하므로 Table의 정보가 Abstract의 표현 방식과 더 유사하다. 셋째, LVLM이 표를 텍스트로 변환하는 과정이 명확하고 일관적이어서 정보 손실이 적다.

반면 Figure는 시각적 패턴과 트렌드를 언어로 표현하는 과정에서 상대적으로 더 많은 정보 손실이 발생한다. 그래프의 추세나 다이어그램의 구조적 관계는 추상적 개념이므로 Abstract의 간결한 서술과 불일치하기 쉽다. 또한 "Figure 3에서 보듯이"와 같은 표현은 Abstract에 포함되지 않아 ROUGE 점수

향상의 제약 요인이 된다. Figure와 Table을 함께 사용할 때 큰 성능 향상을 보이는 이유는 상호보완성에 있다. Figure는 시각적 트렌드와 패턴을 제공하고, Table은 정량적 수치와 비교 데이터를 제공하여 질적 맥락과 양적 근거가 결합된 완전한 정보를 구성한다. Abstract는 "제안 방법은 baseline 대비 15.3% 향상을 보였으며, 안정적인 성능을 유지했다"와 같이 수치(Table)와 트렌드(Figure)를 모두 포함하므로, 단독 사용으로는 Abstract의 일부만 커버하게 된다.

그러나 멀티모달 정보의 통합이 항상 최대 성능 향상을 보장하지는 않는다. 본 실험에서 멀티모달 방식의 개선 효과가 제한적인 경우가 있는 근본적인 원인은 이미지 임베딩과 텍스트 임베딩 간의 정렬(Alignment) 문제로 분석된다. 현재 대부분의 LVLM은 텍스트 임베딩 공간에서 학습되었으며, 이미지를 포함한 멀티모달 입력을 처리할 때 이미지 임베딩과 텍스트 임베딩이 동일한 의미 공간에서 병렬적으로 정렬되어야 한다. 그러나 Figure와 Table에서 추출된 시각 정보가 LVLM의 입력으로 변환되는 과정에서 의미적 일관성이 손실되고, 이미지 임베딩과 텍스트 임베딩 간의 매핑이 불완전하여 모델이 두 모달리티의 정보를 효과적으로 통합하지 못한다. 특히 표 2에서 확인된 바와 같이 Figure+Table 대비 ROUGE-L 기준 22~27%의 상대적 성능 격차는 단순히 정보량의 문제가 아니라, 관련성 있는 텍스트와 이미지가 병렬적으로 정렬되지 않아 발생하는 멀티모달 임베딩 공간의 불일치를 반영한다. Figure의 시각적 패턴과 Table의 수치 데이터가 텍스트 설명과 의미적으로 연결되어야 하지만, 현재의 멀티모달 처리 방식에서는 이러한 연결이 명시적으로 학습되지 않아 두 모달리티 간의 정보가 독립적으로 처리되는 한계가 있다. 따라서 향후 연구에서는 이미지와 텍스트 임베딩을 명시적으로 정렬하는 학습 방법이나, 멀티모달 정보를 통합된 표현 공간으로 사상하는 기법의 개발이 필요하다.

본 실험 결과는 멀티모달 학술 논문 요약에서 Figure와 Table을 함께 사용하는 것이 ROUGE-L 기준으로 유의미한 성능 향상을 가져옴을 실증적으로 입증하였다. ROUGE-L 기준 Figure+Table이 단독 사용 대비 평균 22~27%의 상대적 향상을 보이며, 조합에 따라 최대 43.66%까지 개선 효과가 나타나 실용

적 가치가 높다. 이는 절댓값으로 약 23~26개 단어가 추가적으로 정확히 매칭된 것을 의미하며, 학술 논문 요약 시스템의 실용화 관점에서 의미 있는 개선이다. 반면 BERTScore 기준으로는 평균 0.78~1.02%의 미미한 향상에 그쳐, 멀티모달 정보가 의미적 유사정보보다는 표현의 정확성 개선에 더 효과적임을 확인하였다. 따라서 멀티모달 요약 시스템 개발 시 Figure와 Table을 모두 처리할 수 있는 아키텍처 구현이 권장되며, 동시에 멀티모달 임베딩 간의 정렬을 개선하는 방법론 연구가 병행되어야 한다.

4.3 평가 지표별 최고 성능 조합

ROUGE-L과 BERTScore를 기준으로 최고 성능을 달성한 상위 3개 조합을 분석하여 각각 표 4와 표 5에 제시하였다. 두 평가 지표에서 선정된 최적 조합을 비교 분석함으로써 멀티모달 요약의 성능 특성을 다각도로 이해할 수 있다. 분석 대상에는 EMNLP 2024의 다양한 카테고리(Information Extraction, Ethics & Fairness, Resource & Benchmark 등)에 속한 논문들이 포함되었으며, 각 논문의 특성에 따라 최적 조합이 달라지는 양상을 확인할 수 있었다.

표 4에서 볼 수 있듯이, ROUGE-L 기준 최고 성능은 이미지가 많은 논문에서 Conclusion 섹션만을 Gemini 모델로 멀티모달 방식 처리한 경우로 0.3065를 기록하였다. 이 수치의 실질적 의미는 생성 요약문이 Abstract의 약 92개 단어(300단어 기준 30.65%)와 정확히 순서대로 일치함을 나타낸다. 일반적으로 학술 논문 요약에서 ROUGE-L 0.30 이상은 "good quality" 수준으로 평가되므로, 이는 실용적으로 충분히 활용 가능한 성능이다. 2순위는 텍스트가 많은 논문에서 Introduction+Proposal Experimental setup & results+Conclusion의 섹션 조합을 Grok 모델로 멀티모달 방식 처리한 경우로 0.2930을 달성하였다. 1순위와의 차이는 0.0135포인트로, 이는 약 4개 단어의 차이에 해당하며 실용적으로는 거의 유사한 수준이다. 3순위는 이미지가 많은 논문에서 Introduction+Conclusion 조합을 Gemini 모델로 Text only 방식 처리한 경우로 0.2899를 기록하였다. 1순위와의 격차는 0.0166포인트로 약 5개 단어 차이이며, 2순위와도 0.0031포인트(약 1개 단어) 차이로 세 조합 모두 매우 근접한 성능을 보였다.

결과를 정리해보았을 때 상위 2개 조합이 모두 멀티모달 방식이었으며, 1순위와 3순위가 동일하게 이미지가 많은 논문에서 나왔다는 점이 주목할 만하다. 특히 1순위와 3순위 모두 Gemini 모델을 사용했으나, 멀티모달 정보 추가로 0.0166포인트(약 5.72%)의 성능 향상을 달성한 것은 시각적 정보 활용의 효과를 입증한다. 또한 1순위 조합은 Conclusion 섹션 단독으로 최고 성능을 달성했는데, 이는 학술 논문의 핵심 내용이 결론 부분에 집약적으로 표현되며, 여기에 멀티모달 정보를 결합할 경우 Abstract와 높은 유사도를 보이는 요약을 생성할 수 있음을 시사한다.

표 5를 보면, BERTScore 기준으로는 완전히 다른 패턴이 나타났다. 1순위는 텍스트가 많은 논문에서 Introduction+Proposal Experimental setup & results+Conclusion 조합을 Grok 모델로 멀티모달 방식 처리한 경우로 0.8767을 달성하였다. BERTScore 0.8767은 의미적 임베딩 공간에서 매우 높은 유사도를 나타내며, 일반적으로 0.85 이상은 의미적으로 거의 동일한 수준으로 해석된다. 2순위는 같은 텍스트가 많은 논문에서 Introduction+Conclusion 조합을 Gemini 모델로 Text only 방식 처리한 경우로 0.8678을 기록하였다. 1순위와의 차이는 0.0089포인트인데, BERTScore에서 0.01 미만의 차이는 의미적으로 거의 구분되지 않는 수준이다. 3순위 역시 텍스트가 많은 논문에서 Introduction 섹션만을 Gemini 모델로 Text only 방식 처리한 경우로 0.8662를 달성하였다. 1순위와의 격차는 0.0105포인트로 역시 의미적으로는 미미한 차이이며, 2순위와 3순위의 차이는 0.0016포인트에 불과하여 실질적으로 동등한 수준이다.

BERTScore 기준에서는 상위 3개 조합이 모두 텍스트가 많은 논문에서 나왔으며, 1순위만 멀티모달 방식이고 2, 3순위는 Text only 방식이었다. 이는 ROUGE-L 결과와 극명한 대조를 이루는 부분이다. 특히 주목할 점은 1순위 조합이 표 4의 2순위 조합과 동일하다는 것인데, 이는 해당 조합이 ROUGE-L과 BERTScore 모두에서 우수한 균형 잡힌 성능을 보임을 의미한다. 구체적으로 이 조합은 ROUGE-L 0.2930(2위)과 BERTScore 0.8767(1위)을 달성하여, 표면적 표현의 정확성과 의미적 유사도를 동시에 확보한 유일한 조합이다. 두 평가 지표 간의 이러한 차이는 몇 가지 중요한 시사점을 제공한다.

표 4. ROUGE-L 기준 상위3개 조합

Table 4. Results based on ROUGE-L scores

Rank	Paper Type	Section Combination	LVLN	Method	ROUGE-L
1	Paper with many Images	conclusion	Gemini	Multimodal Added	0.3065
2	Paper with many Tokens	introduction+experimental setup & results+conclusion	Grok	Multimodal Added	0.2930
3	Paper with many Images	introduction+conclusion	Gemini	Text only	0.2899

표 5. BERTScore 기준 상위3개 조합

Table 5. Results based on BERTScore scores

Rank	Paper Type	Section Combination	LVLN	Method	BERTScore
1	Paper with many Tokens	introduction+experimental setup & results+conclusion	Grok	Multimodal Added	0.8767
2	Paper with many Tokens	introduction+conclusion	Gemini	Text only	0.8678
3	Paper with many Tokens	introduction	Gemini	Text only	0.8662

첫째, ROUGE-L과 BERTScore는 요약 품질의 서로 다른 측면을 평가한다는 것이다. ROUGE-L은 순차적 문장 순서와 n-gram 일치도를 측정하는 반면, BERTScore는 의미적 임베딩 공간에서의 유사도를 평가한다. 표 4의 결과를 보면 이미지가 많은 논문에서 멀티모달 방식이 최고 성능(0.3065)을 달성했으며, 이는 시각 정보가 Abstract의 표현 방식과 순차적 구조를 재현하는 데 효과적임을 보여준다. 특히 Conclusion 섹션 단독으로 이러한 성능을 달성한 것은, 결론 부분의 간결한 서술과 시각 정보의 결합이 Abstract의 핵심 문장 구조와 높은 n-gram 일치도를 만들어냄을 의미한다. 반면 표 5에서는 텍스트가 많은 논문이 상위권을 독점했는데, 이는 풍부한 텍스트 정보가 의미적 정확성 확보에 더 유리함을 시사한다. 실용적으로 해석하면, 멀티모달 정보는 요약문의 표면적 표현을 Abstract와 유사하게 만드는 데 탁월하지만, 의미적 깊이 향상은 텍스트 정보의 양과 질에 더 크게 의존한다는 것이다.

둘째, 논문 특성에 따른 최적 조합이 평가 지표에 따라 달라진다는 점이다. ROUGE-L에서는 이미지가 많은 논문이 최고 성능과 3위를 차지했으나(1위, 3위), BERTScore에서는 텍스트가 많은 논문이 상위권을 독점하였다(1위, 2위, 3위). 이는 텍스트가 많은 논문에서 생성된 요약문이 의미적 정확성 측

면에서 더 안정적임을 의미한다. 구체적으로, 텍스트가 많은 논문은 더 풍부한 문맥 정보를 제공하여 LVLN이 의미적으로 일관된 요약을 생성하기 용이한 반면, 이미지가 많은 논문은 시각 정보의 언어화 과정에서 표면적 표현은 크게 개선되지만 의미적 임베딩 공간에서의 유사도는 상대적으로 제한적일 수 있다.

셋째, 섹션 조합의 복잡도와 성능 간의 관계도 주목할 만하다. ROUGE-L에서는 단일 섹션(Conclusion)이 최고 성능을 보인 반면, BERTScore에서는 3개 섹션 조합(Introduction+Proposal Experimental setup & results+Conclusion)이 최고 성능을 달성했다. 이는 순차적 표현 일치치를 위해서는 간결하고 집약적인 정보가 효과적이지만, 의미적 정확성을 위해서는 포괄적인 맥락 정보가 필요함을 시사한다. ROUGE-L 1위 조합이 단일 섹션으로 0.3065를 달성한 반면, BERTScore 1위 조합은 3개 섹션으로 0.8767을 달성하여, 평가 목표에 따라 섹션 선택 전략을 차별화해야 함을 보여준다.

넷째, 모델별 특성도 중요한 요인으로 작용한다. Gemini 모델은 ROUGE-L 기준 상위 3개 중 2개(1위, 3위)에서 선정되어 순차적 표현 재현에 강점을 보였으며, Grok 모델은 두 평가 지표 모두에서 2순위 이상에 선정되어 균형 잡힌 성능을 나타냈다. 특히

Grok 모델의 Introduction+Proposal Experimental setup & results+Conclusion 조합은 ROUGE-L 0.2930(2위), BERTScore 0.8767(1위)을 달성하여 두 지표 모두에서 우수한 성능을 보였다. 이는 포괄적인 섹션 조합과 멀티모달 정보가 결합될 때 표현의 정확성과 의미의 정확성을 동시에 확보할 수 있음을 시사하며, 실용적 관점에서 균형 잡힌 요약 품질이 필요한 경우 해당 조합의 활용을 권장할 수 있다.

4.4 섹션 조합별 성능 패턴 분석

논문 요약 시 어떤 섹션을 선택하고 조합하는가에 따라 성능이 어떻게 달라지는지 분석하기 위해, 섹션 개수별 평균 성능을 계산하였다. 표 6은 1개부터 4개까지 각 섹션 개수별 ROUGE-L과 BERTScore의 평균 성능을 보여준다. 본 분석에는 EMNLP 2024의 104편 논문 전체가 포함되었으며, 각 섹션 개수 그룹별로 최소 15개 이상의 조합이 평가되었다.

표 6을 보면, 섹션 개수와 성능 간에 비선형 관계가 나타났다. 가장 주목할 만한 발견은 2개 섹션 조합이 ROUGE-L 평균 0.2856으로 가장 높은 성능을 달성했다는 점이다. 이는 1개 섹션(0.2748), 3개 섹션(0.2794), 4개 섹션(0.2701)보다 우수한 결과이다. 이 수치의 실질적 의미를 해석하면, 2개 섹션 조합은 평균 86개 단어가 Abstract와 정확히 순서대로 일치하는 반면, 1개 섹션은 82개 단어, 3개 섹션은 84개 단어, 4개 섹션은 81개 단어가 일치한다. 2개 섹션과 4개 섹션 간의 차이는 5개 단어로, 이는 하나의 핵심 개념이나 수치 표현의 차이에 해당한다.

표 6. 섹션 개수별 평균 성능

Table 6. Performance comparison of LVLMs according to different number of sections in text

Number of Sections	Average ROUGE-L	Average BERTScore
1 Section	0.2748	0.8598
2 Section	0.2856	0.8612
3 Section	0.2794	0.8589
4 Section	0.2701	0.8567

BERTScore에서도 유사한 패턴이 관찰되었는데, 2개 섹션 조합이 0.8612로 가장 높은 점수를 기록하였으며, 이는 의미적 임베딩 공간에서 Abstract와 매우 높은 유사도를 보인다. 4개 섹션의 0.8567과 비교하면 0.0045포인트 차이인데, BERTScore에서 이 정도 차이는 미미하지만 일관된 경향성을 나타낸다.

특히 섹션 개수가 증가할수록 성능이 향상될 것이라는 일반적인 예상과 달리, 3개 섹션부터는 오히려 성능이 감소하는 경향을 보였다. 4개 섹션 전체를 포함한 조합의 경우 ROUGE-L이 0.2701로 가장 낮았으며, BERTScore도 0.8567로 최하위를 기록하였다. 2개 섹션 조합 대비 ROUGE-L은 5.4%포인트, BERTScore는 0.5%포인트 낮은 수치이다. 이 차이의 실질적 의미는 명확하다. 4개 섹션을 모두 포함하면 약 5개 단어 수준의 정확도가 감소하며, 이는 주요 연구 결과나 방법론 설명 중 일부가 부정확하거나 불필요하게 표현됨을 의미한다.

이러한 결과는 정보의 양보다 질적 선택이 더 중요함을 명확히 보여준다. 모든 섹션을 포함하면 더 완전한 정보를 제공할 것으로 기대되지만, 실제로는 불필요한 세부 사항이나 중복된 내용이 포함되어 오히려 요약의 핵심성을 떨어뜨리는 것으로 해석된다. 특히 4개 섹션 조합의 경우 Methodology 섹션에 포함된 방법론, 이론적 배경, 관련 연구 등의 상세한 내용이 요약문의 간결성을 저해하여 Abstract와의 유사도를 낮추는 요인으로 작용한 것으로 보인다. 이는 학술 논문 요약의 본질이 "완전성"보다는 "간결성과 핵심성"에 있음을 실증적으로 입증한다.

반면 1개 섹션의 경우 ROUGE-L 0.2748로 중간 수준의 성능을 보였다. 이는 단일 섹션만으로도 어느 정도의 요약 품질을 확보할 수 있음을 의미하나, 2개 섹션 조합에 비해서는 3.8%포인트 낮은 성능이다. 3.8%포인트는 약 11개 단어의 차이로, 이는 1~2개의 중요 문장이 누락되거나 부정확하게 표현됨을 의미한다. 단일 섹션의 경우 논문의 특정 측면만을 반영하게 되어 전체적인 맥락 파악에 한계가 있는 것으로 판단된다.

2개 섹션 조합이 최고 성능을 달성한 이유를 구체적으로 분석하면, 주로 Introduction+Conclusion 조합이 높은 성능에 기여한 것으로 나타났다. 이 조합

은 논문의 연구 배경과 동기를 제시하는 Introduction과 연구 결과 및 의의를 재진술하는 Conclusion을 결합하여, 논문의 핵심 메시지를 간결하면서도 완전하게 전달한다. 또한 두 섹션 모두 Abstract와 유사한 추상화 수준과 서술 방식을 가지고 있어 높은 ROUGE 점수로 이어진 것으로 분석된다. 실용적으로 해석하면, Introduction+Conclusion 조합은 "무엇을 연구했는가"와 "무엇을 발견했는가"라는 학술 논문의 핵심 질문에 대한 답을 제공하므로, Abstract 생성에 최적화된 정보 조합이다.

3개 섹션 조합의 경우 ROUGE-L 0.2794로 2개 섹션 조합보다 2.2%포인트 낮은 성능을 보였다. 2.2%포인트는 약 7개 단어의 차이로, 이는 실용적으로는 미미한 수준이다. 대표적인 3개 섹션 조합인 Introduction+Methodology+Conclusion이나 Introduction+Proposal Experimental Setup & Results+Conclusion의 경우, 추가된 섹션(Methodology 또는 Proposal Experimental Setup & Results)이 지나치게 구체적인 정보를 포함하여 요약문의 추상성을 저해하는 것으로 판단된다. 예를 들어, Methodology 섹션의 상세한 알고리즘 설명이나 Experimental Setup의 구체적인 파라미터 설정은 Abstract에서는 일반적으로 간략히만 언급되므로, 이러한 정보가 과다하게 포함되면 오히려 요약 품질이 저하된다.

이러한 분석 결과는 학술 논문 자동 요약 시스템 설계에 중요한 시사점을 제공한다. 단순히 많은 섹션을 포함하는 것보다 논문의 핵심 내용을 담고 있는 섹션을 전략적으로 선별하는 것이 더 효과적임을 확인하였다. 특히 2개 섹션 조합, 그중에서도 Introduction과 Conclusion의 조합이 가장 균형 잡힌 성능을 제공한다는 점은 요약 시스템 개발 시 고려해야 할 중요한 요소로 작용한다. 실용적으로, 이는

제한된 계산 자원으로 최대 성능을 얻기 위한 효율적인 전략을 제시한다.

4.5 모델별 성능 특성

GPT-4o, Gemini-2.5, Grok-2 등 3개 모델의 멀티모달 정보 활용 능력을 비교 분석하기 위해, 모델별로 Text Only 방식과 Text+Visuals 방식의 평균 성능을 계산하였다. 표 7을 보면, 모델별로 멀티모달 정보 처리 능력이 상이하게 나타났다.

Gemini-2.5는 Text Only 방식에서 ROUGE-L 0.2782를 기록하였고, Text+Visuals 방식에서는 0.2947로 향상되어 5.93%의 성능 증가를 보였다. BERTScore에서도 0.8655에서 0.8678로 0.27% 개선되었다. 이는 Gemini-2.5가 시각적 정보를 텍스트 정보와 효과적으로 통합하여 처리할 수 있음을 의미한다. 특히 ROUGE-L에서 약 6%에 가까운 향상은 멀티모달 정보가 Abstract의 표현 방식과 순차적 구조 재현에 기여함을 보여준다.

GPT-4o의 경우 더 큰 향상폭을 보였다. Text Only 방식의 ROUGE-L 0.2443에서 Text+Visuals 방식의 0.2608로 6.75% 증가하였으며, 이는 세 모델 중 가장 높은 향상률이다. BERTScore 역시 0.8584에서 0.8620으로 0.42% 개선되었다. 절대 향상폭으로 보면 ROUGE-L에서 0.0165포인트(약 5개 단어), BERTScore에서 0.0036포인트의 개선이 이루어졌다. 이러한 결과는 GPT-4o가 멀티모달 정보를 활용할 때 가장 큰 성능 이득을 얻을 수 있음을 시사한다. 특히 두 평가 지표 모두에서 가장 높은 상대적 향상률을 보인 점은, GPT-4o의 멀티모달 정보 통합 아키텍처가 효과적으로 설계되어 있음을 의미한다.

표 7. 모델별 평균 성능 비교

Table 7. Performance comparison of GPT, Gemini and Grok according to combined multimodal data

LVM	Text Only ROUGE-L	Text+Visuals ROUGE-L	Text Only BERTScore	Text+Visuals BERTScore
GPT	0.2443	0.2608	0.8584	0.8620
Gemini	0.2782	0.2947	0.8655	0.8678
Grok	0.2689	0.2549	0.8598	0.8567

반면 Grok-2는 상반된 결과를 보였다. Text Only 방식에서 ROUGE-L 0.2689를 기록하였으나, Text+Visuals 방식에서는 0.2549로 오히려 5.21% 감소하였다. BERTScore에서도 0.8598에서 0.8567로 0.36% 하락하였다. 절대값으로는 ROUGE-L에서 0.0140포인트(약 4개 단어), BERTScore에서 0.0031포인트의 성능 저하가 발생했다. 이러한 성능 저하의 원인은 Grok-2의 아키텍처 특성에서 찾을 수 있다. Diao et al.(2025)에 따르면, Grok 시리즈는 텍스트 토큰에 대해서는 인과적(causal) 어텐션을, 이미지 토큰에 대해서는 양방향(bidirectional) 어텐션을 적용하는 혼합 어텐션 구조를 사용한다. 그러나 이러한 구조에서 이미지 임베딩과 텍스트 임베딩 간의 정렬이 불완전할 경우, 두 모달리티의 정보가 효과적으로 통합되지 못하고 오히려 상호 간섭을 일으킬 수 있다. 특히 Grok-2는 텍스트 정보와 시각 정보의 효과적인 통합에 어려움을 겪거나, 시각적 정보가 오히려 노이즈로 작용하여 텍스트 기반 성능을 저하시키는 것으로 분석된다[19]. 이는 Grok-2의 멀티모달 임베딩 정렬 메커니즘이 다른 모델들에 비해 상대적으로 불안정할 가능성을 시사한다.

모델별 절대 성능을 비교하면, Text Only 방식에서는 Gemini-2.5가 ROUGE-L 0.2782로 가장 높은 성능을 보였으며, Grok-2(0.2689), GPT-4o(0.2443) 순이었다. Gemini-2.5와 Grok-2의 차이는 0.0093포인트로 약 3개 단어에 해당하며, Grok-2와 GPT-4o의 차이는 0.0246포인트로 약 7개 단어에 해당한다. 그러나 Text+Visuals 방식에서는 Gemini-2.5가 0.2947로 여전히 1위를 유지하였고, GPT-4o(0.2608)가 Grok-2(0.2549)를 추월하였다. 이는 멀티모달 정보 활용 능력이 모델 순위를 변화시킬 수 있음을 보여준다. 특히 GPT-4o는 Text Only 방식에서 3위였으나 멀티모달 정보 추가로 2위로 상승하여, 시각 정보 처리 능력이 우수함을 입증하였다.

BERTScore 측면에서는 모든 모델이 0.85~0.87 범위에서 비교적 유사한 성능을 보였다. Text Only 방식 기준으로 Grok-2가 0.8598로 가장 높았고, Gemini-2.5(0.8655), GPT-4o(0.8584) 순이었으나, 최대 격차가 0.0071포인트에 불과하여 의미적 유사도 측면에서는 모델 간 차이가 미미하다. Text+Visuals 방

식 기준으로는 Gemini-2.5가 0.8678로 가장 높았고, GPT-4o(0.8620), Grok-2(0.8567) 순이었다. 이는 의미적 유사도 측면에서는 모델 간 격차가 크지 않으며, 멀티모달 정보 추가의 영향도 제한적임을 시사한다. BERTScore의 경우 모든 모델에서 향상폭이 0.0023~0.0036포인트로 1% 미만에 그쳐, 의미적 임베딩 공간에서는 멀티모달 정보의 기여가 표면적 표현 개선에 비해 상대적으로 작음을 재확인할 수 있다.

상위 성능 조합의 모델 분포를 추가로 분석한 결과, Gemini-2.5가 ROUGE-L 상위 10개 조합 중 7개를 차지하여 가장 안정적인 성능을 보였다. 이는 표 7에서 확인된 Gemini-2.5의 우수한 평균 성능과 일치하는 결과이다. Gemini-2.5는 다양한 섹션 조합과 논문 특성에서 일관되게 높은 성능을 유지하여, 다양한 논문 유형을 처리해야 하는 학술 논문 요약 시스템 구축에 가장 적합한 모델로 평가된다. 특히 Text Only와 Text+Visuals 방식 모두에서 1위를 유지한 점은, Gemini-2.5의 전반적인 텍스트 이해 능력과 멀티모달 통합 능력이 균형 잡혀 있음을 의미한다.

GPT-4o는 평균 성능에서는 중위권이었으나, 다양한 섹션 조합에서 일관된 성능을 유지하는 특성을 보였다. 특히 멀티모달 정보 추가 시 가장 큰 향상폭(ROUGE-L 6.75%, BERTScore 0.42%)을 보인 점은, GPT-4o가 시각적 정보 통합에 강점을 가지고 있음을 나타낸다. 이는 이미지가 많은 논문이나 복잡한 도표를 포함한 논문 요약에 GPT-4o를 활용하는 것이 효과적일 수 있음을 시사한다. 또한 Text Only 방식에서의 낮은 기준 성능(0.2443)에도 불구하고 멀티모달 정보로 상당한 성능 향상을 달성한 것은, GPT-4o의 시각 정보 처리 모듈이 특히 효과적으로 작동함을 보여준다.

Grok-2는 평균 성능에서는 멀티모달 정보 추가 시 성능이 하락하였으나, 특정 조합에서는 예외적으로 우수한 성능을 보였다. 4.3절에서 확인한 바와 같이, Grok-2는 텍스트가 많은 논문의 Introduction+Proposal Experimental setup & results+Conclusion 조합에서 BERTScore 0.8767로 1위를 달성하였다. 이는 Grok-2가 특정 조건에서 최고 성능을 발휘할 수 있으나, 전반적인 안정성은 떨어진다라는 것을 의미한다. 평균적으로는 멀티모달 정보

가 성능을 저하시키지만, 텍스트 정보가 풍부하고 포괄적인 섹션 조합을 사용할 경우 오히려 최고 성능을 달성할 수 있는 특이한 패턴을 보인다. 따라서 Grok-2는 특정 논문 유형이나 섹션 조합에 특화된 시스템 개발에 활용할 수 있으나, 다양한 논문을 처리해야 하는 일반적인 시스템에는 적합하지 않을 수 있다.

종합하면, 멀티모달 요약 시스템 개발 시 모델 선택은 시스템의 목적과 대상 논문 특성을 고려하여 이루어져야 한다. 안정적이고 일관된 성능을 추구한다면 Gemini-2.5가 최적의 선택이다. Gemini-2.5는 Text Only와 Text+Visuals 방식 모두에서 최고 성능을 유지하며, 평균 약 6%의 멀티모달 향상률을 보여 균형 잡힌 성능을 제공한다. 멀티모달 정보 활용의 극대화가 목표라면 GPT-4o가 적합하다. GPT-4o는 멀티모달 정보 추가 시 6.75%의 최고 향상률을 보이며, 특히 시각 정보가 풍부한 논문에서 강점을 발휘할 것으로 예상된다. 특정 조건에서의 최고 성능이 필요하고 대상 논문의 특성이 명확하다면 Grok-2를 선택하는 전략적 접근이 가능하다. 다만 Grok-2는 평균적으로 멀티모달 정보가 성능을 저하시키므로, 사용 시 논문 특성과 섹션 조합을 신중히 검토해야 한다.

4.6 Abstract 기준 평가의 한계

본 실험에서 관찰된 멀티모달 방식의 평균 성능 저하 현상은 평가 방법론 자체의 한계를 시사한다. Abstract는 텍스트로만 구성되어 있어 도표, 그래프, 수식 이미지 등의 시각적 정보를 포함하지 않는다. 따라서 멀티모달 요약문이 "Figure 3의 실험 결과에서 제안 방법은 baseline 대비 15.3% 향상을 보였다"와 같이 구체적 정보를 포함하더라도, Abstract는 "제안 방법은 우수한 성능을 달성했다"와 같이 추상적으로 서술되어 ROUGE 점수가 낮아지는 구조적 문제가 존재한다.

4.1절에서 확인한 바와 같이, 이미지가 많은 논문과 토큰이 많은 논문 모두에서 멀티모달 방식의 평균 ROUGE-L이 각각 8.45%, 6.93% 하락하였다. 이는 멀티모달 요약이 실제로는 더 풍부하고 구체적

인 정보를 제공함에도 불구하고 Abstract를 정답 데이터로 사용하는 현재 평가 방식에서는 낮은 점수를 받는 역설을 초래한다.

이러한 문제는 멀티모달 요약의 진정한 가치를 평가하는 데 근본적인 장애물로 작용한다. 멀티모달 요약문이 시각적 정보를 텍스트로 효과적으로 통합하여 더 완전하고 유용한 요약을 제공하더라도, Abstract와의 표면적 유사도만을 측정하는 ROUGE나 BERTScore로는 이러한 개선을 포착할 수 없다. 오히려 시각적 정보를 상세히 기술할수록 Abstract의 간결한 서술 방식과 불일치하여 페널티를 받게 된다.

V. 결론 및 향후연구

본 연구는 멀티모달 학술 논문 요약에서 최적 성능을 달성하기 위한 섹션 조합과 시각 정보 활용 전략을 체계적인 Ablation Study를 통해 규명하였다. EMNLP 2024 논문 104편을 대상으로 GPT-4o, Grok-2, Gemini-2.5 세 개의 최신 멀티모달 대규모 언어 모델을 활용하여 15개의 섹션 조합과 3가지 시각 정보 유형에 대한 포괄적 실험을 수행하였다.

첫째, Abstract 기준 평가의 근본적 한계를 확인하였다. 멀티모달 요약문이 구체적 정보를 포함하더라도 Abstract의 추상적 서술과 비교되어 낮은 점수를 받는 구조적 문제가 존재한다. ROUGE-2가 Figure만 사용 시 60.93%, Table만 사용 시 60.63% 하락한 것은 현재 평가 방법론이 멀티모달 요약의 진정한 가치를 측정하는 데 한계가 있음을 보여준다.

둘째, 멀티모달 방식의 제한적 효과가 이미지 임베딩과 텍스트 임베딩 간의 정렬 문제에서 기인함을 발견하였다. Figure와 Table에서 추출된 시각 정보가 LVLM 입력으로 변환되는 과정에서 의미적 일관성이 손실되었으며, 이는 ROUGE-L 기준 Figure만 사용 시 평균 21.19%, Table만 사용 시 평균 18.10%의 성능 저하로 나타났다.

셋째, 멀티모달 거대 언어 모델 간 멀티모달 정보 처리 능력에서 뚜렷한 성능 차이를 확인하였다. Gemini-2.5는 Text+Visuals 방식에서 5.9% 증가로 가장 안정적인 성능을, GPT-4o는 6.8% 증가로 가장

높은 향상률을 보인 반면, Grok-2는 5.2% 감소하여 멀티모달 정보 통합에 어려움을 겪었다.

넷째, Figure와 Table의 통합적 활용이 단독 사용 대비 월등히 우수한 성능을 입증하였다. Figure+Table 조합은 ROUGE-L 평균 0.4137로, Figure만 사용(0.3250) 대비 27.29%, Table만 사용(0.3382) 대비 22.33%의 성능 향상을 보였다. 특히 Introduction+Conclusion 조합에서는 43.66%의 향상으로 가장 큰 개선 효과를 나타냈다. Table이 Figure보다 우수한 성능을 보인 이유는 구조화된 수치 데이터가 텍스트로 변환하기 용이하고 Abstract의 서술 방식과 더 유사하기 때문이며, Figure와 Table을 함께 사용할 때 시각적 트렌드와 정량적 수치가 결합되어 완전한 정보를 제공하기 때문이다.

다섯째, 섹션 조합별 성능 분석에서 정보의 양보다 질적 선택이 더 중요함을 확인하였다. 2개 섹션 조합이 ROUGE-L 평균 0.2856으로 4개 섹션 전체 조합(0.2701)보다 5.4%포인트 우수하였으며, Introduction+Conclusion 조합이 요약문 생성에 최적화된 정보 조합임을 실증하였다.

향후 연구도 다음과 같은 연구를 진행할 계획이다. 첫째, 학술 논문에 멀티모달 요약에 적합한 정답 요약문에 대한 대규모 벤치마크 데이터셋을 개발할 예정이다. 둘째, 멀티모달 데이터를 단일 임베딩 공간으로 정렬(Alignment)하는 새로운 방안을 연구할 계획이다. 셋째, 논문 특성에 따라 최적 조합을 자동으로 판단하는 능동적 요약 시스템을 연구할 계획이다. 넷째, 의학, 공학 등 다른 분야 논문으로의 확장 연구를 통해 일반화 가능성을 검증할 계획이다. 생성된 요약문의 정보 완전성, 일관성, 가독성 등을 학술 논문 작성 경험이 풍부한 연구자들이 직접 평가하고, 멀티모달 요약문에 포함된 시각 정보의 적절성과 유용성에 대한 사용자 만족도 조사를 병행하여 ROUGE와 BERTScore가 포착하지 못하는 요약 품질의 다양한 측면을 규명해야 한다.

본 연구는 멀티모달 학술 논문 요약 분야에서 체계적인 Ablation Study를 통해 Figure+Table의 통합적 활용 필요성과 평가 방법론의 한계를 실증적으로 규명하였다. 2개 섹션 조합과 Figure+Table 통합 활용이 최적 성능을 달성하는 핵심 전략임을 입증하

였으며, 새로운 평가 프레임워크의 필요성을 제시하였다. 본 연구 결과는 향후 멀티모달 요약 시스템 개발과 평가 방법론 개선에 중요한 기초 자료로 활용될 것으로 기대되며, 학술 논문 자동 요약 시스템의 실용화를 앞당기는 데 기여할 것으로 판단된다.

References

- [1] X. Zhong, Z. Tan, and S. Gao, "SMSMO: Learning to generate multimodal summary for scientific papers", *Knowledge-Based Systems*, Vol. 310, Article No. 112908, Feb. 2025. <https://doi.org/10.1016/j.knosys.2024.112908>.
- [2] Z. Tan, X. Zhong, and B. Chiu, "Multimodal paper summarization with hierarchical fusion", *2024 International Conference on Engineering and Emerging Technologies (ICEET)*, Dubai, United Arab Emirates, pp. 1-6, Dec. 2024. <https://doi.org/10.1109/iceet65156.2024.10913899>.
- [3] X. Chen, H. Alamro, and M. Li, "Target-aware abstractive related work generation with contrastive learning", *Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, pp. 373-383, Jul. 2022. <https://doi.org/10.1145/3477495.3532065>.
- [4] Z. Shi, S. Gao, and Z. Zhang, "Towards a unified framework for reference retrieval and related work generation", *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, pp. 5785-5799, Dec. 2023. <https://doi.org/10.18653/v1/2023.findings-emnlp.385>.
- [5] S. Syed, K. Al Khatib, and M. Potthast, "TI; dr progress: Multi-faceted literature exploration in text summarization", *Proc. of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, St. Julian's, Malta, pp. 195-206, Mar. 2024. <https://doi.org/10.18653/v1/2024.eacl-demo.21>.
- [6] J. Zhu, H. Li, and T. Liu, "MSMO: Multimodal

- summarization with multimodal output", Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 4154-4164, Oct.-Nov. 2018. <https://doi.org/10.18653/v1/d18-1448>.
- [7] H. Shakil, A. Farooq, and J. Kalita, "Abstractive text summarization: State of the art, challenges, and improvements", *Neurocomputing*, Vol. 603, Article No. 128255, Oct. 2024. <https://doi.org/10.1016/j.neucom.2024.128255>.
- [8] G. H. Lee, Y.-H. Park, and L. K. Joo, "Building a Korean Text Summarization Dataset Using News Articles of Social Media", *KIPS Transactions on Software and Data Engineering*, Vol. 9, No. 8, pp. 251-258, Aug. 2020. <https://doi.org/10.3745/KTSDE.2020.9.8.251>.
- [9] Y. Lee, J. Si, and S. Kim, "Design of a TextRank-Based Summarization Method Specialized for Korean Documents", Proc. of KIIT Conference, Jeju, Korea, pp. 970-971, Jun. 2025.
- [10] A. Cohan and N. Goharian, "Scientific article summarization using citation-context and article's discourse structure", Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 390-400, Sep. 2015. <https://doi.org/10.18653/v1/d15-1045>.
- [11] G. Sharma, A. Paretkar, and D. Sharma, "Synthesizing scientific summaries: An extractive and abstractive approach", arXiv preprint arXiv:2407.19779, Jul. 2024. <https://doi.org/10.48550/arXiv.2407.19779>.
- [12] E. Ko, N. Kim, and K. Kim, "Automatic quality evaluation with completeness and succinctness for text summarization", *Journal of Intelligence and Information Systems*, Vol. 24, No. 2, pp. 125-148, Jun. 2018. <https://doi.org/10.13088/jiis.2018.24.2.125>.
- [13] R. C. Belwal, S. Rai, and A. Gupta, "Text summarization using topic-based vector space model and semantic measure", *Information Processing & Management*, Vol. 58, No. 3, Article No. 102536, May 2021. <https://doi.org/10.1016/j.ipm.2021.102536>.
- [14] T. Wang, C. Yang, and M. Zou, "A study of extractive summarization of long documents incorporating local topic and hierarchical information", *Scientific Reports*, Vol. 14, No. 1, Article No. 10140, May 2024. <https://doi.org/10.1038/s41598-024-60779-z>.
- [15] T. Gigant, C. Guinaudeau, and F. Dufaux, "Summarization of multimodal presentations with vision-language models: Study of the effect of modalities and structure", arXiv preprint arXiv:2504.10049, Apr. 2025. <https://doi.org/10.48550/arXiv.2504.10049>.
- [16] P. Janjani, M. Palan, and S. Shirude, "Converging dimensions: Information extraction and summarization through multisource, multimodal, and multilingual fusion", *Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2024)*, Varna, Bulgaria, Vol. 15462, pp. 168-183, Sep. 2024. https://doi.org/10.1007/978-3-031-81542-3_14.
- [17] A. Saha, A. Tiwari, and S. Ruthvik, "Two eyes, two views, and finally, one summary! Towards multi-modal multi-tasking knowledge-infused medical dialogue summarization", arXiv preprint arXiv:2407.15237, Jul. 2024. <https://doi.org/10.48550/arXiv.2407.15237>.
- [18] N. Deas, E. Turcan, I. E. P. Mejia, and K. McKeown, "MASIVE: Open-ended affective state identification in English and Spanish", Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Miami, Florida, USA, pp. 20467-20485, Nov. 2024. <https://doi.org/10.18653/v1/2024.emnlp-main.1139>.
- [19] H. Diao, M. Li, and S. Wu, "From pixels to words -- towards native vision-language primitives at scale", arXiv preprint arXiv:2510.14979, Oct. 2025. <https://doi.org/10.48550/arXiv.2510.14979>.

저자소개

주 민 선 (Min Sun Ju)



2022년 3월 ~ 현재 :
국립군산대학교 소프트웨어학과
학사과정
관심분야 : 자연어처리, 기계학습,
인공지능, 강화학습

온 병 원 (Byung-Won On)



2007년 8월 :
펜실베이니아주립대학교
컴퓨터공학과(박사)
2008년 2월 ~ 2009년 5월 :
브리티시컬럼비아대학교
박사후연구원
2010년 9월 ~ 2011년 8월 :

일리노이대학교 ADSC센터 선임연구원
2011년 9월 ~ 2014년 3월 : 서울대학교
차세대융합기술연구원 연구교수
2014년 4월 ~ 현재 : 국립군산대학교 소프트웨어학과
교수
관심분야 : 자연어처리, 기계학습, 인공지능, 강화학습