일반논문 (Regular Paper)
방송공학회논문지 제31권 제1호, 2026년 1월 (JBE Vol.31, No.1, January 2026)
https://doi.org/10.5909/JBE.2026.31.1.77
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

실시간 다중 채널 영상 분석 시스템을 위한 VLM 기반 자연어 질의응답 시스템 연구

장 일 식[a], 박 구 만[b)‡]

A Study on a VLM-based Natural Language Question-answering System for Real-time Multi-channel Video Analysis Systems

Ilsik Chang[a)] and Goo-Man Park[b)‡]

요 약

디지털 감시 및 모니터링 시스템의 급속한 발전과 함께 실시간 다중 채널 영상 분석에 대한 수요가 급증하고 있다. 본 연구에서는 Vision Language Model(VLM)과 Large Language Model(LLM)을 활용하여 실시간 다중 채널 영상 분석 시스템에 자연어 질의응답 기능을 통합한 시스템을 제안하고, 다양한 방법을 비교 실험하였다. 제안된 시스템은 레거시 영상 분석 시스템과 연동하여 메타데이터와 다중 이미지를 VLM을 통해 텍스트로 변환한 후, 이를 벡터 데이터베이스에 저장한다. 자연어 질의응답 모듈은 질의 분류, Text-to-SQL 변환, Retrieval Augmented Generation(RAG) 기반 응답 생성의 다층 구조로 설계되었다. 10개의 최신 VLM 모델에 대한 성능 비교 실험을 vLLM 환경에서 수행하였으며, 정성적 및 정량적 방법을 사용하여 평가하였다. RAG 최적화 기법으로는 리랭커, 하이브리드 검색, 질의 재구성 방법 등 다양한 조합에 대한 검색 정확도를 RAGAS를 통해 평가하였다. 본 연구는 실시간 영상 분석 분야에서 VLM 기반 자연어 인터페이스의 실용적 구현 방안을 제시하여, 향후 지능형 감시 시스템 발전에 기여할 것으로 기대한다.

Abstract

The rapid advancement of digital surveillance and monitoring systems has led to a surge in demand for real-time multi-channel video analysis. This study proposes a system that integrates natural language query-response capabilities into a real-time multi-channel video analysis system by utilizing Vision Language Models (VLMs) and Large Language Models (LLMs), and conducts comparative experiments using various methods. The proposed system interfaces with legacy video analysis systems to convert metadata and multiple images into text via the VLM, storing this in a vector database. The natural language query-response module is designed with a multi-layered structure comprising query classification, Text-to-SQL conversion, and Retrieval Augmented Generation (RAG)-based response generation. Performance comparisons of 10 state-of-the-art VLM models were conducted in the vLLM environment and evaluated using both qualitative and quantitative methods. For RAG optimization techniques, search accuracy was evaluated using RAGAS for various combinations of re-ranking, hybrid retrieval, and query reformulation methods. This study presents a practical implementation approach for VLM-based natural language interfaces in real-time video analysis, expected to contribute to the future development of intelligent surveillance systems.

Keyword : VLM, LLM, RAG, Multi-Channel Video Analysis System

Ⅰ. 서 론

현대 사회에서는 영상 감시 및 모니터링 시스템의 중요성이 날로 커지고 있다. 특히 스마트 시티, 산업 안전 관리, 교통 모니터링 등 다양한 분야에서 실시간 다중 채널 영상 분석에 대한 수요가 증가하고 있다. 트랜스패런시 마켓 리서치에 따르면 영상 감시 시장은 2023년 82조 원에서 2027년 146조 원으로, 국내 시장도 4조 4,156억 원에서 5조 4,672억 원으로 성장할 것으로 예상된다. 기존의 레거시 영상 분석 시스템들은 주로 객체 탐지, 행동 인식, 이상 탐지 등의 특정 작업에 특화되어 있다. 대표적인 기능으로는 객체 검출(Object Detection), 객체 추적(Object Tracking), 행동 인식(Action Recognition) 등이 있으며, 이러한 시스템들은 구조화된 알람 및 로그 데이터를 생성하는 데 중점을 두고 있다. 객체 검출은 컴퓨터 비전 분야의 핵심 기술 중 하나로, 영상 내에서 객체의 위치와 범주를 동시에 식별하는 기술이다. 특히, YOLO(You Only Look Once)[1]는 실시간 객체 검출 분야에서 혁신적인 발전을 이끌어 왔다. 이후 YOLOv2[2], YOLOv3[3]를 거쳐 최신 YOLOv8[4]에 이르기까지 지속적으로 발전하고 있다. 객체 추적은 시간적으로 연속된 프레임에서 동일한 객체를 식별하고 그 이동 궤적을 추적하는 기술이다. 단순히 매 프레임마다 객체를 검출하는 것을 넘어, 프레임 간 객체의 동일성을 유지하며 고유 ID를 부여한다. 대표적인 추적 알고리즘으로는 SORT (Simple Online and Realtime Tracking)[5]과 Deep-SORT[6]가 있다. SORT는 칼만 필터(Kalman Filter)를 사용하여 객체의 다음 위치를 예측하고, 헝가리안 알고리즘(Hungarian Algorithm)을 통해 검출된 객체와 기존 추적 객체를 최적으로 매칭한다. 이후 딥러닝 기법이 적용된 DeepSORT는 SORT의 기본 구조에 외관 특징을 추가하여 가림 현상이나 일시적 검출 실패 상황에서도 안정적인 추적을 가능하게 한다. 딥러닝을 이용한 다중 객체 추적 방법으로 ByteTrak[7], BoT-SORT[8], StrongSORT[9] 등 다양한 방법이 존재한다. 영상 행동 인식은 동영상으로부터 인간의 행동을 자동으로 식별하고 분류하는 컴퓨터 비전 기술로, 최근 Vision Transformer(ViT) 기반 접근법이 기존의 CNN 기반 방법론을 빠르게 대체하고 있다. 전통적인 CNN 구조는 합성곱 연산을 통해 지역적 시공간 특징을 추출하는 반면, 트랜스포머는 self-attention 메커니즘을 활용하여 전역적 시공간 의존성을 효과적으로 모델링할 수 있다. Shaikh et al.[10]은 행동 인식 분야에서 CNN에서 Transformer로의 패러다임 전환을 포괄적으로 분석하며, 멀티모달 데이터 통합과 시공간 표현 학습에서 Transformer의 우수성을 입증했다. 한편, Xiao et al.[11]은 자기지도 학습(self-supervised learning)을 Vision Transformer에 접목한 Enhanced Visual Industrial Sequence Transformer를 제안하여, 레이블이 제한된 산업 환경에서도 높은 행동 인식 정확도를 달성했다. 이러한 레거시 이벤트 감지 기능은 실시간 처리가 최적화되어 있으며, 각 이벤트에 대해 미리 정의된 규칙과 임계값을 가지고 있다. 예를 들어, 침입이나 쓰러짐과 같은 사건은 일정한 패턴으로 정의되어 있어 빠른 트리거 동작이 가능하다. 또한 기존 시스템은 실제 운영 환경에서 오랜 검증을 거쳤기 때문에 보안 등 민감한 분야에서 신뢰할 수 있다. 그러나 이러한 레거시 영상 분석 시스템들은 사용자가 자연어로 표현하는 복잡하고 맥락적인 질의에 대해 직관적인 응답을 제공하는 데 한계를 보인다. 현재 영상 분석 시스템이 직면한 주요 문제점은 다음과 같다. 첫째, 접근성의 문제로 전문 지식을 가진 운영자만이 효과적으로 시스템을 활용할 수 있다. 둘째, 질의의 복잡성으로 인해 SQL(Structured Query Language)과 같은 구조화된 질의 언어를 사용해야 하므로 복잡한 조건을 포함한 검색이 어렵다. 셋째, 맥락 이해의 부족으로 단순한 키워드 매칭 방식만으로는 사용자의 의도를 정확히 파악하기 어렵다. 마지막으로, 대용량 다중 채널 영상 데이터를 실시간으로 처리하면서 자연어 질의에 즉각적으로 응답하는 데 한계가 있다.

a) 서울과학기술대학교 IoT융복합기술연구소(IoT Convergence Technology Lab., SeoulTech)
b) 서울과학기술대학교 스마트ICT융합공학과(Dept. of Smart ICT Convergence Engineering, SeoulTech)
‡ Corresponding Author : 박구만(Goo-Man Park)
E-mail: gmpark@seoultech.ac.kr
Tel: +82-2-970-6430
ORCID: https://orcid.org/0000-0002-7055-5568
※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2026년도 실감콘텐츠핵심기술개발 사업으로 수행되었음 (과제명: 증강현실기반 RAG 재난안전 통합플랫폼 개발, 과제번호: RS-2025-02219484, 기여율: 100%)
· Manuscript October 17, 2025; Revised November 24, 2025; Accepted November 24, 2025.

자연어 질의응답은 LLM과 VLM의 발전을 통해 기존 한계를 극복할 수 있는 새로운 가능성을 제시한다. 특히 VLM은 이미지 및 영상 데이터를 텍스트로 변환하는 능력이 크게 향상되었으며, LLM은 자연어 질의에 대한 이해와 응답, 보고서 자동화 등에서 뛰어난 성능을 보이고 있다. RAG와 벡터 데이터베이스의 결합은 대규모 멀티모달 데이터에서 효율적이고 정확한 정보 검색 및 응답 생성을 가능하게 한다. 하지만 VLM/LLM 방식은 높은 컴퓨팅 리소스를 요구하고, 초기 구축 비용과 아직 검증되지 않은 신뢰성 확보의 문제가 존재한다. 본 연구에서는 레거시 영상 분석 시스템의 장점과 VLM/LLM 기술의 장점을 효과적으로 통합하는 하이브리드 아키텍처를 설계하여, 기존 실시간 영상 분석의 결과 정보인 메타데이터와 영상 정보를 동시에 VLM에 적용한다. 본 연구에서는 실시간 다중 채널 영상 데이터를 자연어로 질의할 수 있는 인터페이스를 구축하고, 다양한 VLM 모델의 성능을 종합적으로 비교하여 최적의 모델 선택 기준을 제시한다. 또한 다양한 RAG 기반 검색 시스템을 비교하여 응답 정확도를 평가함으로써, 핵심 구성 요소인 VLM과 RAG 기법의 시스템 효율성과 성능을 검증하고자 한다. 본 논문의 구성은 본론에서 시스템 아키텍처, 최신 동향 및 관련 연구, 실험 및 고찰을 거쳐 결론 및 향후 연구로 기술한다.

Ⅱ. 본 론

1. 시스템 아키텍처

1.1 시스템 구조

제안하는 시스템은 크게 영상 분석 모듈, VLM 변환 모듈, 벡터 데이터베이스, 자연어 질의응답 모듈로 구성된다. 각 모듈은 마이크로서비스 아키텍처로 설계되어 독립적인 확장과 유지보수가 가능하도록 구성되었다. 시스템의 데이터 흐름은 다음과 같다. 먼저, 다중 채널에서 수집된 영상 데이터는 레거시 영상 분석 시스템을 통해 기본적인 메타데이터(객체 탐지 결과, 이벤트 정보, 시간 정보 등)를 생성한다. 메타데이터와 프레임들은 VLM을 통해 자연어 설명으로 변환되며, 이는 임베딩 과정을 거쳐 벡터 데이터베이스에 저장된다. 사용자의 자연어 질의가 입력되면, 질의 분류 모듈(Query Classification Module)이 이를 분석하여 적절한 처리 경로로 라우팅한다. 사용자 질의가 데이터베이스 정보 질의일 경우 Text2SQL 모듈을 활용해 SQL 쿼리를 생성하고, 데이터베이스에서 정보를 추출하여 답변한다. Text2SQL 모듈은 사용자의 자연어 질의를 구조화된 SQL 쿼리로 자동 변환하는 작업으로, 데이터베이스 접근성을 향상시키는 기술이다. 본 논문에서는 PostgreSQL 데이터

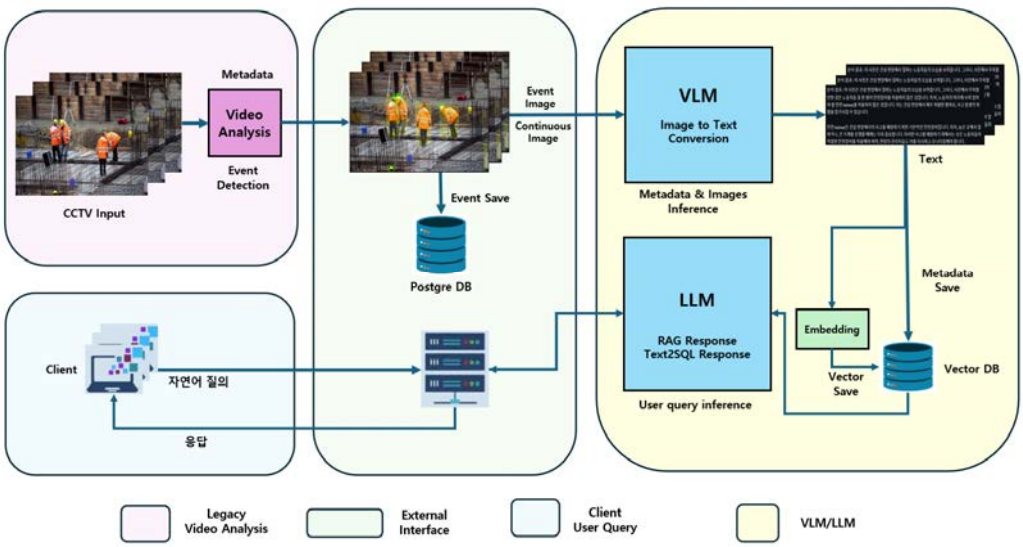

그림 1. 단순화된 시스템 구조
Fig. 1. Simplified System Architecture

베이스에 SQL 쿼리를 이용하여 데이터 정보를 확인할 수 있다. 화면 정보 질의일 경우 RAG 기법 사용하여 답변을 한다. RAG는 대규모 언어모델의 생성 능력과 외부 지식 베이스의 검색 기능을 결합한 하이브리드 AI 아키텍처이다. 지능형 영상 분석 시스템에서 RAG는 VLM을 통해 추출된 텍스트가 벡터 정보로 변환되고, 변환된 벡터 정보는 벡터 데이터베이스에 저장된 정보는 시맨틱 검색을 통해 관련 영상 메타데이터를 찾아내고, 이를 바탕으로 정확하고 맥락에 부합하는 답변을 생성할 수 있다. 본 논문에서는 사용자의 질문을 벡터화하고 벡터 데이터베이스에서 가장 유사한 벡터를 검색하고, 이를 기반으로 LLM이 답변을 생성한다. 그 외 질문은 자체 응답하거나 답변 불가로 처리하며, 일반적인 대화, 시스템 정보, 답변 불가 질의에 대해 LLM이 응답한다. 전체 시스템 구조를 이해하기 쉽게 그림 1은 단순화된 시스템 구조를 보여준다.

1.2 레거시 시스템 통합

기존 레거시 시스템과의 호환성을 보장하기 위해 표준화된 API 인터페이스를 설계하였다. 레거시 시스템에서 생성되는 데이터는 JSON 형식으로 표준화하여 처리하며, 실시간 스트리밍을 위해 REST API를 사용하여 구현하였다. 메타데이터 스키마는 시간 정보(타임스탬프, 시간대), 공간 정보(채널 ID, 위치 좌표, 영역 정보), 객체 정보(탐지된 객체 유형, 바운딩 박스, 신뢰도), 이벤트 정보(이벤트 유형, 심각도)로 구성된다. 구조화된 메타데이터는 벡터 데이터베이스의 메타데이터 필드에 저장되며, 자연어 질의 시 필터 정보로 활용된다.

그림 2는 전체 시스템의 상세한 시스템 구조를 보여준다.

2. 최신 동향 및 관련 연구

2.1 VLM 발전

VLM은 시각 정보와 언어를 결합한 멀티모달 모델로, 다량의 이미지와 텍스트 데이터를 학습하여 시각과 언어 정보를 동시에 처리할 수 있다. 이미지와 텍스트를 함께 입력받아 두 요소 간의 관계를 학습함으로써, 이미지를 보고 설명하거나 질문에 답변할 수 있다. 최근 VLM은 단순한 이미지 캡셔닝을 넘어 추론 능력(Reasoning), 다중 이미지 처

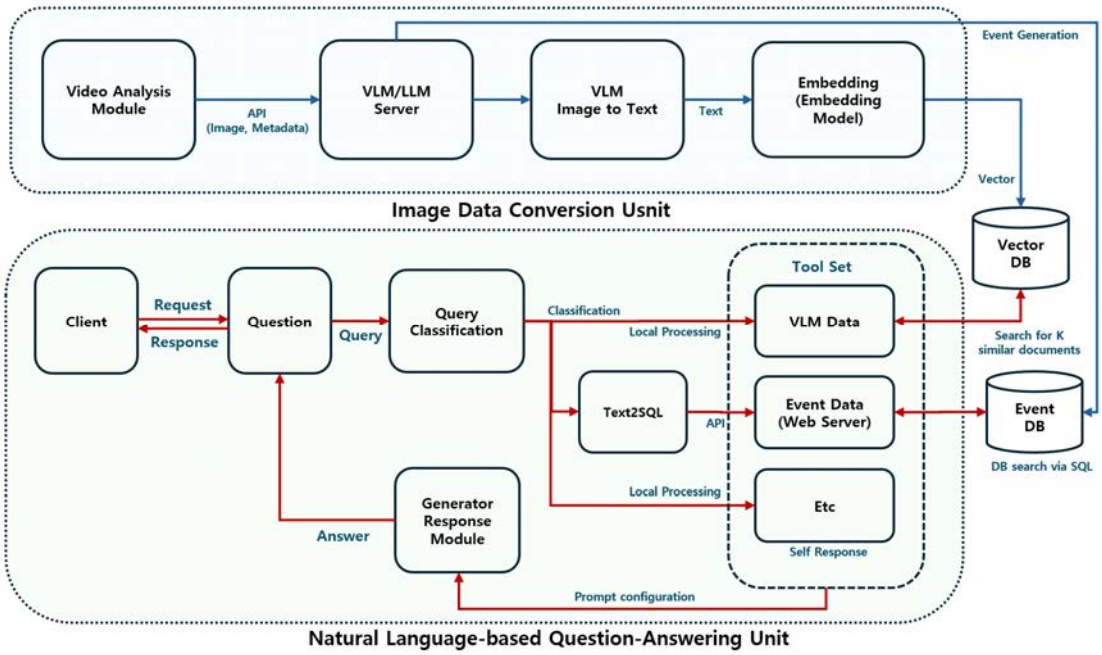

그림 2. 상세한 시스템 구조
Fig. 2. Detailed System Architecture

리(Multi-Image Processing), 동영상 이해(Video Understanding) 등 다양한 기능으로 발전하고 있다. 특히 경량화된 모델들이 발전하면서 에지 디바이스에서의 활용 가능성도 높아지고 있다. 본 연구에서 사용된 모델들은 이러한 추세를 반영하여 다양한 크기와 구조를 가진 모델들로 선정하였다. 실시간 처리를 고려해 파라미터 수가 적은 소형 모델과 온프레미스 환경에서 동작 가능한 오픈 소스 모델을 채택하였다. 선정된 모델은 Gemma3[12], InternVL3[13], Qwen2.5[14], Hyper CLOVAX[15], Deepseek-vl2[16], Smol-VLM2[17], Phi-3.5 vision[18]이다.

2.2 RAG 및 벡터 검색

LLM은 자연어 처리 분야에서 혁명을 일으키며 다양한 분야에 활용되고 있다. 하지만 LLM은 정보의 최신성 부족, 잘못된 정보 제공, 편향된 응답 생성 등의 한계점을 가지고 있다. 이러한 한계를 극복하기 위해 RAG[19] 기술이 등장하였다. RAG는 LLM의 강력한 텍스트 생성 능력과 외부 지식 베이스를 결합한 혁신적인 접근 방식이다. 이 기술은 인덱싱, 검색, 생성이라는 세 가지 주요 요소를 결합하여 사용자가 원하는 답변을 생성할 때 학습된 데이터에만 의존하지 않고 외부 데이터셋을 활용할 수 있도록 한다. RAG는 언어 모델의 생성 능력과 외부 지식 검색의 장점을 결합하여 더 정확하고 상황에 맞는 응답을 제공할 수 있다.

2.2.1 임베딩 및 벡터화

임베딩은 텍스트를 고차원 벡터로 변환하여 시스템이 의미론적으로 유사한 데이터를 검색하고 관리할 수 있도록 하는 과정이다. 고차원 벡터 간의 유사도는 코사인 거리(Cosine Distance) 또는 유클리디안 거리(Euclidean Distance)를 통해 계산한다. 본 연구에서는 MTEB(Multi-lingual, v2)[20]에서 다국어에 우수한 성능을 보이고, 적은 수의 파라미터를 가진 Qwen/Qwen-Embedding-0.6B, intfloat/multilingual-e5-large-instruct, google/embeddinggemma-300m, BAAI/bge-m3 등을 선정하여 비교 평가한다.

2.2.2 데이터 인덱싱

인덱싱은 임베딩된 벡터를 빠르게 검색할 수 있도록 데이터베이스에 저장하는 과정이다. 벡터 데이터베이스에는 여러 종류(Qdrant[21], Pinecone[22], Chroma[23], Weaviate[24]) 등이 있다. 본 연구에서는 Qdrant를 활용한다. Qdrant는 고성능 설계, 고급 필터링, 유연한 배포, 하이브리드 검색, 양자화 기능, 보안 기능 등의 특징을 갖추고 있다.

2.2.3 RAG 최적화

RAG는 검색 단계의 정확성이 답변 품질을 좌우한다. 단순히 가장 유사한 벡터를 찾는 것을 넘어, 검색된 결과를 정제하고 재조합하는 다양한 기법들이 최근 연구되고 있다. 리랭커(Reranker)는 임베딩 유사도를 구하는 일반 벡터 검색과 달리, 크로스 엔코더(Cross-Encoder) 기반으로 질의-문서 간 유사도를 직접 계산하여 검색된 상위 n개의 문서들의 순위를 다시 조정하는 방법이다. 본 연구에서는 적은 수의 파라미터를 갖는 Qwen/Qwen3-Reranker-0.6B, BAAI/bge-reranker-v2-m3, jina-reranker-v2-base-multilingual 등의 모델을 선정하여 비교 평가하였다. HyDE[25](Hypothetical Document Embeddings)은 사용자의 질의를 기반으로 LLM이 가상의 문서를 생성하고, 이 가상 문서의 임베딩을 이용해 검색을 수행하는 방법이다. Multi-hop Reasoning[26]은 여러 개의 문서를 결합하여 복잡한 질의에 대한 답변을 도출하는 방법이다. Contextual Compression[27]은 검색된 문서에서 불필요한 정보를 제거하고 핵심 내용만 압축하여 LLM에 전달하는 방법이다. Self Query Retriever[28]는 주어진 사용자의 질문을 벡터 데이터베이스를 조회할 필터링 질의로 변환하여 검색하는 방식이다. 본 연구에서는 질의 분류 모듈에서 사용자 질문을 JSON 구조로 출력하여 벡터 데이터베이스에서 필터링에 활용할 수 있는 가능한 구조를 구축하였다. 질문 확장(Query Expansion) 기법 중 질문 재구성(Query Reformulation) 방법[29]은 LLM을 활용하여 원래 질문을 다양한 형태로 재작성하는 방식으로, 동의어 확장, 질문 명확화, 관련 키워드 추가 등 여러 방법을 통해 쿼리를 변형하여 검색의 다양성과 정확도를 향상시킨다. 하이브리드 검색은 밀집 검색과 희소 검색 두 가지 방식을 융합하여 검색 정확도를 높이는 방법으로, 밀집 벡터 기반 검색은 의미적 유사성에 강점이 있고, 희소 기반 검색은 벡터는 정확한 키워드 매칭에 강점을 가진다. 이처럼 서로 다른 특성을 지니고 있어 상호 보완적인 효과를 발휘한다. 본 연구에서는 질문 재구성과 하이브리드 검색 기법을 적용하여 비교 평가를 수

행하였다.

2.3 평가 방식

2.3.1 N-gram 매칭 방식

N-gram, BLEU(Bilingual Evaluation Understudy), ROUGE(Recall-Oriented Understudy for Gisting Evaluation), METEOR(Metric for Evaluation of Translation with Explicit ORdering) 등은 문장에서 동일한 단어나 구절이 얼마나 일치하는지를 기반으로 하는 평가 방식이다. 이들은 의미적 유사성을 반영하지 못하고 문맥을 고려하지 않으며, 동의어나 문장 변형을 적절히 처리하지 못하는 한계가 존재한다. N-gram은 텍스트 내에서 연속된 N개의 단어나 문자로 구성된 시퀀스를 의미하며, 기계 번역, 텍스트 요약 등의 평가에서 참조 문장과 생성된 문장의 유사도를 측정하는 데 사용된다. BLEU는 기계 번역 결과를 평가하기 위해 사용되는 자동 평가 지표로, 정밀도(Precision)를 강조한다. ROUGE는 자동 텍스트 요약 및 문장 생성 모델 평가에 주로 사용되는 지표로, 재현율(Recall)을 중심으로 한다. 마지막으로 METEOR는 기계 번역 품질 평가를 위해 개발된 지표로, BLEU의 단점을 보완하기 위해 만들어졌다. 단순한 단어 매칭 외에도 어간 추출, 동의어 매칭, 패러프레이징 등 다양한 언어학적 요소를 고려한다.

2.3.2 BERT(Bidirectional Encoder Representations from Transformers)-Score

BERT-Score[30]는 BERT와 같은 사전 학습된 언어 모델을 활용하여 텍스트 간 의미적 유사성을 평가한다. BLEU나 ROUGE와 달리 단순한 단어 매칭이 아니라 문맥적 유사성을 고려하여 보다 인간의 평가에 가까운 결과를 제공한다.

2.3.3 LLM 기반의 평가 방식

고성능 LLM을 활용하여 사전에 정의된 평가 기준과 프롬프트에 따라 다른 LLM의 응답 품질과 프롬프트의 적합성을 평가하는 방식이다. 점수화 기준을 설정하고, 프롬프트를 통해 다양한 평가 항목(정확성, 유창성, 일관성 등)을 지정하여 정성적인 기준에 따라 LLM을 통해 평가를 진행한다. 본 연구에서는 VLM 결과에 대한 정성적 평가와 질의 분류 모델의 질문 합성 데이터 생성 검증, 그리고 RAG 성능 테스트를 위한 질문 합성 데이터 생성 검증에 활용된다.

2.3.4 LLM 기반의 RAG 평가 방식

RAGAS는 RAG 성능 평가를 위한 다양한 지표를 제공하는 오픈 소스 프레임워크이다.

표 1은 RAGAS에서 평가하는 RAG 평가 지표의 분류를 나타낸다. 표 2는 RAG 평가 요소를 보여준다. 본 연구에서

표 1. RAG 평가 지표의 분류
Table 1. Classification of RAG Evaluation Metrics

Category	Evaluation Metric	Description
Retrival	Context Precision	An evaluation metric measuring whether the retrieved document is relevant to the query
	Context Recall	An evaluation metric determining whether the retrieved document contains the information necessary to answer the query
Generation	Faithfulness	A metric evaluating how accurately the response was generated based on the information provided in the retrieved document (context)
	Answer Relevancy	A metric evaluating how relevant the generated response is to the query

표 2. RAG 평가 요소
Table 2. RAG Evaluation Criteria

Elements	Description	Usage
Input	User-entered query or question to be evaluated	Used as a basis for evaluating the quality of retrieved and generated responses
output	Final response generated by the system to the question	Used to assess how relevant and reliable the generated answer is to the question
Reference Answer	A predefined correct answer for evaluation purposes	Used to assess the accuracy and relevance of the response
Retrieved Documents	Documents the system searched to generate the response	Used to evaluate whether the documents are relevant to the question and serve as a key metric for measuring search performance

는 컨텍스트의 정밀도와 신뢰성을 기반으로 평가를 수행한다.

Ⅲ. 실험 및 고찰

1. VLM 성능 평가

실험 환경은 OS: Ubuntu 22.04, CPU: AMD Ryzen 7 5800X, GPU: NVIDIA GeForce RTX 3090(24G)를 사용하였다. vLLM은 LLM의 추론 및 배포를 가속화하기 위한 오픈소스 라이브러리로, 모델 추론 중 반복적으로 참조되는 데이터를 캐싱하기 위해 키-값 캐시(KV Cache)를 활용한다. vLLM은 기본적으로 전체 GPU 메모리의 90%를 할당하여 모델 추론과 키-값 캐시를 처리하지만, gpu_memory_utilization을 0.6으로 설정하여 최대 GPU 메모리 사용량이 14.4G를 넘지 않도록 제한하였다. 실험에 사용된 VLM 모델은 InternVL3-2B/8B, Gemma-3-4/12B, DeepSeek-VL2, PaliGemma2-3B, HyperCLOVAX-SEED-Vision-3B, Qwen2.5-VL-3B/7B, SmolVLM2-2.2B 등이다. 비교적 모델 크기가 큰 gemma-3-12B, Qwen2.5-VL-7B, InternVL3-8B는 4비트 양자화를 적용하였으며, max_tokens는 1200으로 설정하여 테스트하였다. 본 연구의 목적은 서로 다른 경량 VLM이 동일 조건에서 실시간 반복 추론 시 나타내는 정확도-지연-안정성 간의 트레이드오프를 계량화하고, vLLM 기반 다채널 서빙에서 동시성 확장성과 출력 일관성을 체계적으로 측정하는 것이다. 이를 위해 동일 이미지를 5초 간격으로 10분 동안 반복 제시하였으며, 실시간 다중 채널 VLM 변환 성능 비교를 위해 영어 답변, 한글 답변, 단일 이미지, 다중 이미지(5초 간격) 등 다양한 실험을 수행하였다. 모든 모델 출력은 동일한 프롬프트와 이미지를 사용하여 정성적 평가와 정량적 평가를 진행하였다. 정성적 평가는 각 VLM의 생성 결과를 GPT-4o를 활용해 정확성, 관련성, 명확성, 풍부성, 그리고 다중 이미지에서의 시간적 일관성(T/C) 등을 0~10점 척도로 평가한 후, 평균 점수를 산출하였다. 추론 시간(e_time)은 16채널 동시 추론 시간을 기반으로 총 시간 ÷ 16으로 계산하였다. 정량적 평가는 동일 프롬프트로 GPT-4o가 생성한 결과를 정답으로 삼아 평가하였다. 표 3은 15초 간격으로 단일 이미지에 대해 VLM이 영어로 답변했을 때의 평가 결과를 나타낸다. 표 3에서는 InternVL3, Qwen2.5, Hyper CLOVAX, Deepseek-vl2

표 3. 15초 간격으로 단일 이미지에 대한 VLM의 답변을 영어로 하였을 경우 평가 결과
Table 3. Evaluation results when VLM's responses to single images were provided in English at 15-second intervals

			Gemma3 4B	InternVL3 2B	Qwen2.5 VL 3B	Hyper CLOVAX	Deepseek vl2	SmolVLM 2	Phi-3.5 vision	Gemma3 12B	Qwen2.5 VL 7B	InternVL3 8B
Parameter Size			4.3B	2.09B	3.75B	3.72B	3.37B	2.25B	4.15B	12.2B	8.29B	7.94B
Qualitative Evaluation	e_time		2.44	**0.23**	**0.74**	**0.55**	**0.85**	0.50	1.35	13.53	3.49	1.82
	Token		232.68	153.86	177.3	244.51	112.55	87.4	143.36	172.33	175.58	243.53
	Accuracy		6.90	7.70	7.14	7.18	7.05	6.625	6.59	7.38	7.12	7.66
	Relevance		7.76	8.45	7.75	8.05	7.83	7.525	7.44	8.26	7.90	8.41
	Richness		6.57	7.12	6.60	6.86	6.61	6.11	6.04	6.77	6.76	7.20
	Clarity		8.11	8.52	8.44	8.34	8.43	8.31	8.17	8.35	8.46	8.66
	Total		7.33	**7.95**	**7.48**	**7.61**	**7.48**	7.14	7.06	7.69	7.56	7.98
Quantitative Evaluation	Bert Score	precision	0.87	0.90	0.89	0.87	0.90	0.91	0.90	0.88	0.89	0.87
		recall	0.90	0.92	0.91	0.91	0.92	0.90	0.91	0.91	0.92	0.92
		f1	0.89	0.91	0.90	0.89	0.91	0.90	0.90	0.89	0.90	0.90
	rouge	rouge1	0.343	0.446	0.394	0.337	0.459	0.401	0.418	0.4	0.402	0.352
		rouge2	0.113	0.178	0.148	0.128	0.17	0.136	0.14	0.136	0.154	0.15
		rougeL	0.202	0.276	0.244	0.206	0.283	0.253	0.258	0.236	0.247	0.22
	bleu		0.331	0.429	0.377	0.284	0.488	0.387	0.446	0.409	0.365	0.282
	meteor		0.331	0.397	0.366	0.343	0.378	0.284	0.34	0.355	0.373	0.376

표 4. 15초 간격으로 단일 이미지에 대한 VLM의 답변을 한글로 하였을 경우 평가 결과
Table 4. Evaluation results when VLM's responses to a single image were provided in Korean at 15-second intervals

			Gemma3 4B	InternVL3 2B	Qwen2.5 VL 3B	Hyper CLOVAX	Deepseek vl2	SmolVLM2	Phi-3.5 vision	Gemma3 12B	Qwen2.5 VL 7B	InternVL3 8B
Parameter Size			4.3B	2.09B	3.75B	3.72B	3.37B	2.25B	4.15B	12.2B	8.29B	7.94B
Qualitative Evaluation	e_time		3.01	**0.71**	**2.03**	**0.50**	**2.41**	1.66	6.46	21.40	5.36	2.47
	Token		279.6	653.19	538.44	208.90	404.17	954.78	792.68	261.34	268.1	307.52
	Accuracy		6.11	3.73	5.04	6.59	4.91	1.99	3.79	6.60	6.22	5.95
	Relevance		6.88	4.52	5.52	7.41	5.43	1.875	4.09	7.44	6.89	6.67
	Richness		5.95	3.57	4.63	6.59	4.39	1.31	3.40	6.45	5.86	5.89
	Clarity		7.70	4.61	5.81	8.34	5.40	1.78	4.52	8.35	7.86	7.60
	Total		6.66	4.11	5.25	**7.23**	5.02	1.74	3.95	7.21	6.73	6.53
Quantitative Evaluation	Bert Score	precision	0.86	0.84	0.87	0.88	0.82	0.79	0.61	0.86	0.89	0.88
		recall	0.91	0.90	0.91	0.92	0.85	0.82	0.64	0.91	0.91	0.91
		f1	0.89	0.87	0.89	0.90	0.83	0.80	0.63	0.89	0.90	0.90
	rouge	rouge1	0.409	0.292	0.372	0.433	0.091	0.141	0.046	0.421	0.452	0.421
		rouge2	0.213	0.142	0.196	0.246	0.022	0.05	0.015	0.221	0.249	0.229
		rougeL	0.254	0.199	0.242	0.286	0.069	0.122	0.039	0.268	0.302	0.281
	bleu		0.252	0.178	0.254	0.277	0.042	0.067	0.021	0.254	0.308	0.267
	meteor		0.395	0.312	0.349	0.424	0.064	0.121	0.045	0.412	0.429	0.421

모델이 추론 시간이 짧으면서도 전체 성능이 우수한 것을 확인할 수 있다. 또한, 파라미터 수가 많은 모델은 추론 속도가 매우 느려 실제 실시간 적용이 어려워 보인다.

표 4는 15초 간격으로 단일 이미지에 대한 VLM의 한글 답변 평가 결과를 나타낸다. 표 4에서는 영어 답변에서 우수한 성능을 보인 InternVL3, Qwen2.5, Hyper CLOVAX, Deepseek-vl2 모델들이 추론 속도에서는 좋은 결과를 나타냈으나, 정성적 평가의 종합 점수에서는 Hyper CLOVAX 모델을 제외하고는 만족스러운 성과를 거두지 못했다. 이러한 원인은 모델의 파라미터 수가 적고, 한국어에 특화되어 있지 않기 때문으로 보인다.

표 5는 15초 간격으로 촬영한 총 5장의 이미지를 5초 간

표 5. 15초 간격으로 다중 이미지를 VLM의 답변을 영어로 하였을 경우의 평가 결과
Table 5. Evaluation results when VLM responses to multiple images were provided in English at 15-second intervals

			Gemma3 4B	Intern VL3 2B	Qwen2.5 VL 3B	Hyper CLOVAX	Deepseek vl2	SmolVLM2	Phi-3.5 vision	Gemma3 12B	Qwen2.5 VL 7B	Intern VL3 8B
Parameter Size			4.3B	2.09B	3.75B	3.72B	3.37B	2.25B	4.15B	12.2B	8.29B	7.94B
Qualitative Evaluation	e_time		3.73	**0.62**	**1.54**	**2.04**	**0.98**	2.20	2.60	12.65	3.86	3.19
	Token		141.88	149.4	151.95	254.91	121.56	151.90	148.75	135.32	126.48	202.42
	Accuracy		6.17	7.91	7.06	7.59	6.26	6.13	6.48	7.33	8.02	8.36
	Relevance		7.07	8.5	7.80	8.375	6.94	6.89	7.28	8.04	8.63	9.06
	Richness		5.93	7.16	6.30	7.05	5.64	5.42	5.86	6.52	7.02	7.58
	Clarity		8.07	8.9	8.46	8.36	7.76	7.71	8.15	8.35	8.83	9.11
	T/C		6.13	7.68	6.56	7.81	5.58	5.43	6.41	7.16	8.38	8.79
	Total		6.67	**8.03**	**7.23**	**7.84**	**6.44**	6.32	6.84	7.48	8.175	8.58
Quantitative Evaluation	Bert Score	precision	0.89	0.89	0.89	0.87	0.89	0.89	0.89	0.89	0.90	0.89
		recall	0.89	0.90	0.89	0.89	0.88	0.89	0.89	0.89	0.90	0.90
		f1	0.89	0.90	0.89	0.88	0.88	0.89	0.89	0.89	0.90	0.89
	rouge	rouge1	0.363	0.4	0.378	0.334	0.357	0.358	0.383	0.378	0.412	0.395
		rouge2	0.086	0.12	0.111	0.104	0.092	0.093	0.099	0.1	0.131	0.127
		rougeL	0.202	0.231	0.219	0.197	0.209	0.209	0.22	0.217	0.245	0.228
	bleu		0.424	0.473	0.444	0.368	0.408	0.416	0.46	0.434	0.48	0.438
	meteor		0.266	0.309	0.292	0.297	0.245	0.249	0.277	0.256	0.294	0.332

격으로 입력하여, 다중 이미지에 대한 VLM의 영어 답변 평가 결과를 나타낸다. 또한, 다중 이미지 입력 시 시간적 일관성도 평가 항목에 포함하였다. 표 5의 결과 역시 위 실험과 마찬가지로 InternVL3, Qwen2.5, Hyper CLOVAX, Deepseek-vl2 모델들이 우수한 성능을 나타냈다. 다중 이미지에서의 시간적 일관성 평가 항목에서는 InternVL3, Qwen2.5, Hyper CLOVAX 모델이 높은 점수를 기록하였다. 추론 속도와 성능을 종합적으로 InternVL3 2B 모델이 가장 뛰어난 성능을 보임을 확인할 수 있다.

표 6은 15초 간격의 이미지를 VLM이 한글로 답변했을 경우의 평가 결과를 나타낸다. 평가 결과, 다중 이미지의 추론에서의 영어 답변에서는 InternVL3-2B, Qwen2.5-VL-3B, HyperCLOVAX, Deepseek-vl2 모델이 짧은 추론 시간과 우수한 성능을 보였다. 그러나 한글 답변의 경우에는 HyperCLOVAX이 뛰어난 성능을 나타냈다. 또한 시간적 일관성 평가 항목 역시 HyperCLOVAX에서 가장 좋은 성능을 나타냈다. 이러한 원인은 HyperCLOVAX 모델이 한국어에 특화되어있지 않기 때문인 것으로 보인다. 본 연구에서는 한글 답변에서 가장 우수한 성능을 보인 HyperCLOVAX를 채택하였다. 하지만 16채널, 5개의 영상 처리 시 추론 시간이 16 × 2.08 = 33.28초로 나타나 15초 간격의 실시간 처리가 어려운 것으로 확인되었다. 이러한 문제점을 해결하고자 본 연구에서는 실시간 처리를 위해 3개의 영상과 2개의 영상에 대한 추가 실험을 진행하였다. 표 6에서 HyperCLOVAX 항목에 Images에 이미지 개수가 3개, 2개인 항목을 추가하여 실험한 결과를 포함하였다. 3개의 영상일 경우 16 × 1.29 = 20.64초, 2개의 영상일 경우 16 × 0.84 = 13.44초가 소요되었다. 16채널 실시간 다중 채널 서비스를 위해 본 연구에서는 15초 간격으로 5초 단위의 2개 영상에 대해 VLM을 통한 텍스트 변환 방식을 선택하였다.

그림 3은 15초 간격으로 촬영한 총 5장의 이미지를 5초 간격으로 배열하여, VLM 답변이 한글일 경우 각 모델별 정성적 평가 점수를 직관적으로 이해할 수 있도록 막대 차트로 나타낸 것이다.

각 모델의 정량적 평가는 표만으로는 쉽게 구분하기 어려워, 막대 차트를 활용하여 직관적으로 이해할 수 있도록 하였다. 그림 4는 15초 간격으로 촬영한 총 5장의 이미지를 5초 간격으로 배열하여, 다중 이미지에 대한 VLM의 한글 답변을 기준으로 각 모델별 정량적 평가인 BLEU 점수를 성능이 높은 순서대로 나타낸 것이다.

표 6. 15초 간격으로 다중 이미지를 VLM의 답변을 한글로 하였을 경우의 평가 결과
Table 6. Evaluation results when VLM responses to multiple images were provided in Korean at 15-second intervals

			Gemma3 4B	Intern VL3 2B	Qwen2.5 VL 3B	HyperCLOVAX			Deepseek vl2	SmolVLM 2	Phi-3.5 vision	Gemma3 12B	Qwen2.5 VL 7B	Intern VL3 8B
Parameter Size			4.3B	2.09B	3.75B	3.72B			3.37B	2.25B	4.15B	12.2B	8.29B	7.94B
Images			5	5	5	5	3	2	5	5	5	5	5	5
Qualitative Evaluation	e_time		4.03	**1.29**	**2.70**	**2.08**	**1.29**	**0.84**	**2.40**	5.31	3.62	22.70	6.04	5.98
	Token		168.47	819.45	411.35	249.0	261.0	228.1	329.46	856.96	186.87	237.88	210.17	315.69
	Accuracy		6.23	3.91	5.93	7.08	7.25	7.61	3.34	2.34	5.54	6.91	7.13	6.77
	Relevance		7.125	4.48	6.57	7.89	8.10	8.47	3.67	2.19	5.92	7.98	8.07	7.69
	Richness		5.91	3.14	5.02	6.81	6.74	7.12	2.55	1.38	4.89	6.48	6.36	6.18
	Clarity		7.44	3.49	6.41	8.19	7.97	8.36	3.48	1.80	7.03	7.96	7.91	7.31
	T/C		6.04	3.39	5.11	6.73	7.13	7.57	2.30	1.45	4.93	7.07	6.78	6.72
	Total		6.55	3.68	5.81	**7.34**	**7.44**	**7.84**	3.07	1.83	5.67	7.28	7.25	6.93
Quantitative Evaluation	Bert Score	precision	0.88	0.81	0.87	0.87	0.87	0.87	0.81	0.81	0.79	0.86	0.90	0.88
		recall	0.90	0.87	0.89	0.90	0.90	0.91	0.84	0.82	0.80	0.90	0.90	0.89
		f1	0.89	0.84	0.88	0.89	0.89	0.89	0.82	0.81	0.79	0.88	0.90	0.89
	rouge	rouge1	0.468	0.21	0.369	0.421	0.411	0.418	0.127	0.154	0.019	0.438	0.47	0.432
		rouge2	0.204	0.083	0.165	0.2	0.2	0.21	0.031	0.047	0.001	0.197	0.215	0.189
		rougeL	0.271	0.139	0.228	0.253	0.254	0.258	0.096	0.134	0.019	0.26	0.292	0.259
	bleu		0.339	0.128	0.266	0.283	0.27	0.272	0.056	0.079	0.002	0.291	0.353	0.313
	meteor		0.341	0.209	0.276	0.373	0.374	0.386	0.079	0.115	0.011	0.341	0.318	0.332

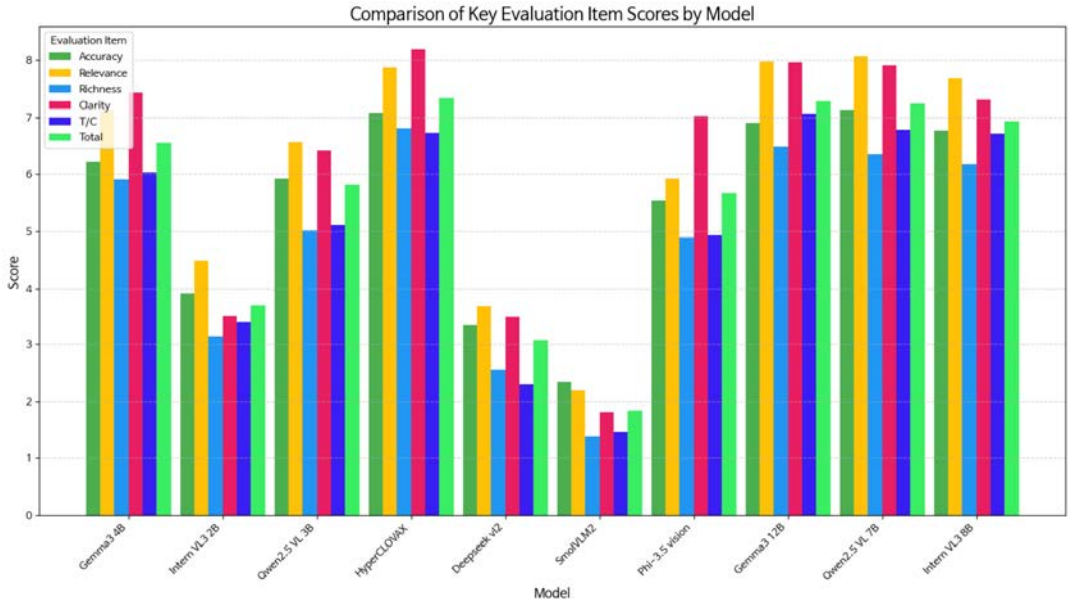

그림 3. 모델별 정성적 평가
Fig. 3. Qualitative Evaluation by Model

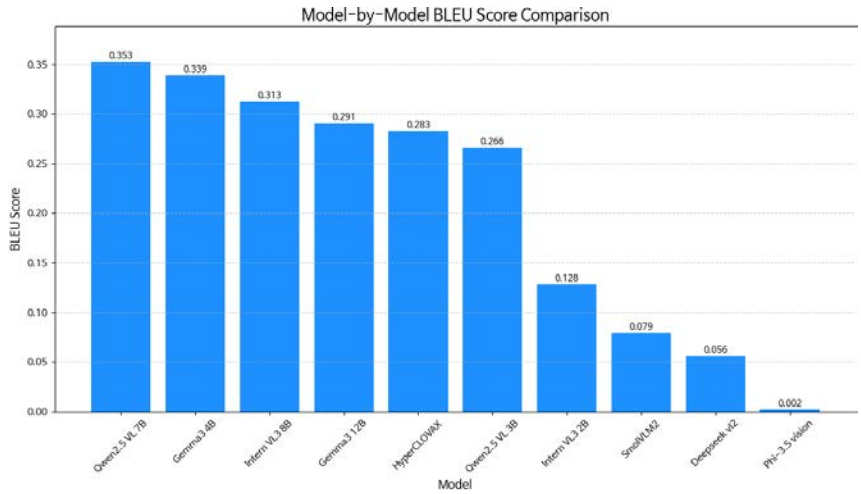

그림 4. 다중 이미지 VLM 답변 한글: 정량적 평가 BLEU 점수
Fig. 4. Multi-Image VLM Answer Korean: Quantitative Evaluation BLEU Score

그림 5는 15초 간격으로 촬영한 총 5장의 이미지를 5초 간격으로 배열하여, 다중 이미지에 대한 VLM 답변이 한글일 경우 각 모델별 정량적 평가인 METEOR 점수를 성능이 높은 순서대로 나타낸 것이다.

본 연구에서는 추가로, 실시간 추론 시간을 단축하기 위한 방법으로 배치 크기에 따른 추론 시간을 테스트하였다.

표 7은 HyperCLOVAX를 단일 이미지 기반으로 하여 배치 크기를 다양하게 변경하며 추론 시간을 측정한 결과이다. 배치 크기가 커질수록 채널별 속도가 빨라지는 것을 확인할 수 있다. 그러나 실험 결과, 배치 크기가 16을 초과할 경우 성능 향상이 크게 나타나지 않아, 배치 크기를 16으로 설정하였다.

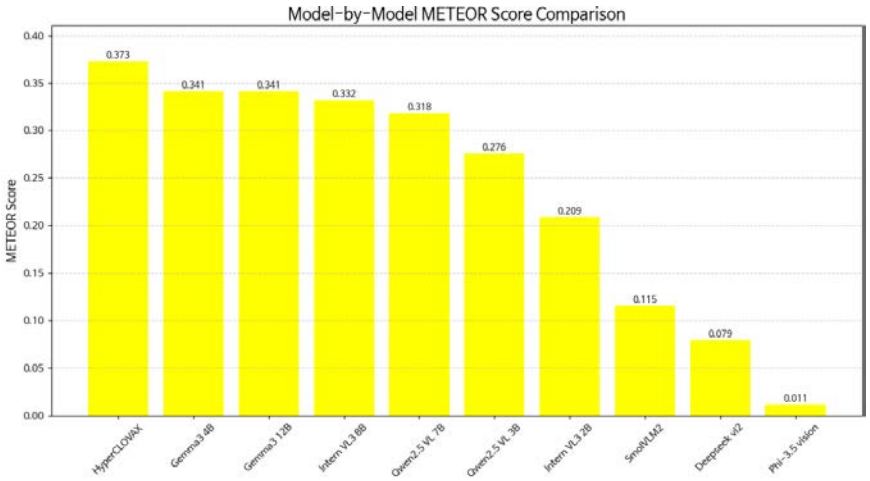

그림 5. 다중 이미지 VLM 답변 한글: 정량적 평가 METEOR 점수
Fig. 5. Multi-Image VLM Answer Korean: Quantitative Evaluation METEOR Score

표 7. 배치 크기에 따른 추론 시간
Table 7. Inference time by batch size

ch / model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	32	64	128
e_time (sec)	2.54	1.36	1.08	1.07	0.97	0.87	0.84	0.79	0.75	0.74	0.72	0.68	0.66	0.64	0.58	0.5	0.49	0.47	0.45

종합적으로 정리하면, 실시간 다중 채널 영상 분석 시스템을 위한 VLM 기반 자연어 질의응답 시스템 평가에서, 레거시 영상 분석을 통해 생성된 메타데이터와 영상 데이터를 활용하여 프롬프트를 생성하고, VLM을 통해 텍스트로 변환하는 과정에서 중요한 요소로 추론 시간과 정성적·정량적 평가 결과를 도출하였다. 그중 HyperCLOVAX가 가장 우수한 성능을 보였으며, 현재 시스템에서 실시간 처리가 가능한 경우는 배치 크기 16일 때, 2장의 이미지를 동시에 입력한 경우이기 때문에 해당 방법으로 적용하여 추가 실험을 진행하였다.

2. 질의 분류 모듈 성능 평가 및 선택

질의 분류 모듈의 성능 평가를 위해서는 다수의 질의와 질의 분류에 대한 정답 데이터셋이 필요하다. 본 연구에서는 질의 분류 데이터셋을 GPT-4.1을 활용한 합성데이터셋

Question string · lengths	Difficulty int64	Analysis_type string · classes	Reason string · lengths
12 … 164	1 … 5	4 values	70 … 276
영상 감지 시스템이 범죄 예방에 실제로 효과가 있는지에 대한 과학적 근거를 알려줄 수 있나요?	4	no_answer	이 질문은 영상 감지 시스템의 실제 효과에 대한 과학적 근거를 묻고 있으나, 시스템이 직접적으로 데이터베이스나 화면 정보를 통…
이 시스템이 작업장 내 안전사고 예방에 어떻게 도움이 되는지 설명해 주세요.	3	self_answer	질문은 영상 감지 시스템이 작업장 안전에 어떤 역할을 하는지에 대한 설명을 요구하고 있습니다. 이는 시스템의 기능이나 목적에…
최근 한 달 동안 범죄 의심 이벤트가 가장 많이 발생한 카메라 위치를 알려줘.	3	text2sql	범죄 예방 및 대응 도메인에서, 이벤트 데이터베이스에 저장된 정보를 기반으로 최근 한 달 동안(시간 정보 없이 자연스럽게 기간…
화면에 보이는 사람들의 옷 색상을 알려줘.	2	vlm	질문이 영상 화면에 나타난 사람들의 옷 색상에 대한 것으로, 화면 정보를 기반으로 답변이 가능하며, 복잡하지 않은 수준의 질문이…
작업장 내에서 안전모를 착용하지 않은 사람이 있는지 화면에서 확인해줘.	2	vlm	작업장 안전 분야에서 영상 화면을 통해 안전모 미착용 여부를 확인하는 질문은 실제 사용자가 자주 할 수 있는 자연스러운 질문입…

그림 6. 질의 분류 평가를 위해 생성된 합성 데이터셋
Fig. 6. Synthetic dataset generated for query classification evaluation

으로 구축하였다. 다양한 도메인과 난이도로 구분하였으며, 질문 유형을 먼저 제시한 후 질문을 생성하도록 하였다. 생성된 질문에 대한 평가를 거쳐 총 500개의 데이터셋을 확보하였고, 이를 바탕으로 소형 LLM 모델을 대상으로 테스트를 진행하였다. 출력은 JSON 포맷으로 구조화하였다. 그림 6은 질의 분류 평가를 위해 생성된 합성 데이터셋을 보여준다. 생성된 합성 데이터셋을 기반으로 다양한 LLM 모델의 정확도 및 추론 시간을 측정하였다. 질의 분류 모델은 다른 응답의 스트리밍 방식 이전에 동작하기 때문에 사용자의 대기 시간이 발생하며, 따라서 가능한 한 짧은 시간이 중요하다. 본 실험은 한국어에 강한 소형 LLM 모델을 대상으로 실험이 진행되었다.

표 8은 다양한 모델에 대한 질의 분류 평가 실험 결과를 나타낸 것이다. 본 연구에서는 실험 결과 성능이 가장 우수한 Qwen4-4B 모델을 선택하였다. 실험 결과, 소형 모델에서는 추론 결과가 JSON 포맷으로 구조화된 출력 생성 시 오류가 자주 발생하였으며, 프롬프트 항목에서 멀티턴 대화에 대한 질의 재생성과 필터링 적용을 위해 특정 채널이나 특정 기간 등에 대한 처리가 항목 생성에 어려움을 초래하는 것으로 나타났다.

3. Text2SQL 모델

Text2SQL 모델은 자연어 질의를 SQL 데이터 조회 언어로 변환하는 작업이다. 본 연구에서는 Qwen2.5-Coder-7B 모델[31]을 선택하였다. 이 모델의 주요 특징은 Qwen 시리즈 중 가장 작은 모델임에도 불구하고 뛰어난 코딩 성능(HumanEval 88.4%)을 보이며, 오픈소스라는 점이다. 또한 SQL 벤치마크(Spider)에서 82.0%의 성능을 기록하여, 적은 파라미터 수를 가진 LLM 모델중 SQL 변환 능력이 우수한 모델이다.

4. RAG 성능 비교

4.1 RAG 성능 평가를 위한 데이터셋 생성

RAG 성능 평가를 위해 15초 간격으로 촬영된 5초 단위의 2개 영상에서 추출한 VLM 기반 텍스트 정보(총 608개)

표 8. 질의 분류 평가 실험 결과
Table 8. Query Classification Evaluation Experiment Results

Model		Qwen3			Qwen2.5	gemma-2		kanana	EXAONE		Midm
		Qwen3 1.7B	Qwen3 4B	Qwen3 8B	Qwen2.5 7B-it	gemma-2 2b-it	gemma-2 9b-it	kanana-1.5 2.1b-it	EXAONE-4.0 1.2B	EXAONE-3.5 2.4B-it	Midm-2.0 Mini-it
Accuracy		0.614	**0.858**	0.842	0.619	0.388	0.602	0.254	0.532	0.486	0.572
e_time	min	1.51	1.80	1.67	1.69	1.54	2.99	1.377	1.99	4.23	1.69
	max	7.02	10.66	10.33	12.57	4.12	27.04	14.60	7.54	18.98	5.79
	avg	2.23	3.912	3.49	6.87	1.68	13.34	1.47	2.734	8.77	1.93

ref_text list	q_type string	q_text string
["이 이미지는 도시의 한 구석을 보여주고 있으며, 주로 벽돌로 지어진 건물들이 밀집해 있는 모습입니다. 건물들은 여러 …	simple	이미지 중앙에 검은색 철제 문 위에 놓여 있는 작은 자전거를 찾아줘
["이미지에서는 공장이나 산업 현장으로 보이는 공간이 보입니다. 초록색의 대형 기계와 구조물이 중앙에 자리 잡고 있으며,…	multi_context	- 초록색 대형 기계가 있는 공장과 보라색 차양이 설치된 아이스크림 가게 앞에서 각각 정지해 있는 사람이나 특별한 행동 없이 서 있는 …
["두 이미지에서는 건설 현장에서 작업 중인 두 사람이 보입니다. 이들은 안전 장비를 착용하고 있으며, 높은 곳에서 작업을…	simple	- 건설 현장에서 빨간색이나 검은색 작업복을 입고 높은 곳에서 작업하고 있는 사람을 찾아줘
["첫 번째 이미지에서는 빨간 모자를 쓴 사람이 건물 앞에 앉아 있는 모습입니다. 주변에 여러 대의 차량이 주차되어 있으…	simple	건물 앞에 앉아 있던 빨간 모자를 쓴 사람이 어디로 갔는지 찾아줘
["이 두 이미지에는 총 7명의 사람이 등장합니다.\n\n첫 번째 이미지는 주차 공간과 세차장이 보입니다. 차량 여러 대가 주…	simple	- 하얀 차량 옆에서 밝은 색 옷을 입고 있는 사람을 찾아줘

그림 7. RAG 성능 평가를 위해 생성된 질문
Fig. 7. Questions generated for RAG performance evaluation

를 활용하여, GPT-4.1을 이용하여 합성 질문을 총 500개의 합성 질문을 생성하였다. 질문은 단일 텍스트 정보와 다중 텍스트 정보를 각각 8:2 비율로 생성하였으며, 생성된 질문을 GPT-4.1을 통해 검증하여 최종적으로 439개를 선정하였다. 그림 7은 RAG 성능 평가를 위해 생성된 질문들을 보여준다.

4.2 MMR(Maximal Marginal Relevance) 적용을 통한 다양성 확보

사용자 질의에 대한 유사성 검색 시 top_k를 8개로 지정하여 각각의 벡터 임베딩을 활용해 질의를 수행할 경우, 특정 채널에 집중되는 현상을 확인할 수 있다. 영상은 연속적인 이미지로 구성되어 있기 때문에, 질문이 특정 채널의 연속된 이미지와 유사할 가능성이 높다. 그러나 본 시스템은 다중 채널을 사용하므로, 채널에서 참조 텍스트가 선택되기를 기대한다. 본 연구에서는 MMR을 적용하여 문서의 관련성과 다양성을 동시에 고려하는 방법을 사용한다. 각 채널별로 top_k를 3개로 지정하고, 16채널일 경우 총 16×3 = 48개의 유사한 텍스트를 기반으로 MMR 방법을 적용한다. 48개의 유사한 텍스트를 MMR 기반으로 순위를 재정렬한 후 상위 8개를 선택한다. 그림 8은 16채널에서 top_k 3개를 적용했을 때, 유사 텍스트로 선택된 상위 채널 수를 보여준다. 특정 채널에 집중되지 않고 다양한 채널에 분포되는 결과를 나타낸다. 본 연구에서는 각 채널별 top_k 3개, 총 16×3 = 48개의 유사한 텍스트를 확보한 후, 해당 텍스트에 대해 리랭커를 적용하여 유사도를 기반으로 하여 16개

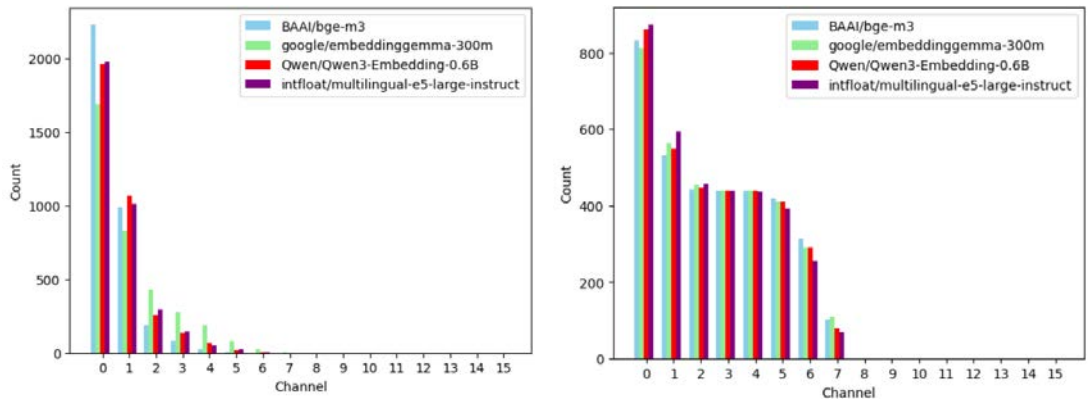

그림 8. 16채널의 top_k 3개 적용 시 유사 텍스트로 선택된 상위 채널수
Fig. 8. Number of top channels selected as similar text when applying top_k 3 to 16 channels

Name	Status	Points (Approx)	Segments	Shards	Vectors Configuration (Name, Size, Distance)
vd_analyzerId_016_multi_2_3_google_embeddinggemma-300m_google_embeddinggemma-300m	green	608	8	1	dense 768 Cosine sparse Sparse
vd_analyzerId_016_multi_2_3_intfloat_multilingual-e5-large-instruct_intfloat_multilingual-e5-large-instruct	green	608	8	1	dense 1024 Cosine sparse Sparse
vd_analyzerId_016_multi_2_3_Qwen_Qwen3-Embedding-0.6B_Qwen_Qwen3-Embedding-0.6B	green	608	8	1	dense 1024 Cosine sparse Sparse

그림 9. Qdrant 벡터 데이터베이스에 VLM을 통해 벡터로 변환되어 저장
Fig. 9. Vector data is converted to vectors via VLM and stored in the Qdrant vector database

를 추출하고, 최종적으로 MMR 방법을 통해 5개의 문서를 선정한다.

4.3 임베딩 벡터 및 리랭커 선정

임베딩 벡터 모델로는 BAAI/bge-m3, google/embeddinggemma-300m, Qwen/Qwen3- Embedding‐0.6B, intfloat/multilingual-e5-large-instruct 총 4개를 사용하였으며, 리랭커 모델로는 Qwen/Qwen3-Reranker-0.6B, BAAI/bge-reranker-v2-m3, jinaai/jina-reranker-v2-base-multilingual을 활용하여 비교하였다. 그림 9는 Qdrant 벡터 데이터베이스에 VLM을 통해 변환된 텍스트가 벡터로 저장된 모습을 보여준다.

평가는 RAGAS를 이용하여 컨텍스트 정밀도와 신뢰도를 측정하였다. 표 9는 RAGAS를 통한 RAG 평가 결과를 나타낸다. 0.25, 0.5, 0.75는 데이터를 크기 순서로 정렬했을 때 4분위 수에 해당하는 25%, 50%, 75%를 의미한다. 전체(total)는 0.5 구간의 값인(context precision + faithfulness)/2의 평균이다. 최종 결과, 임베딩 모델이 BAAI/bge-m3일 경우 성능이 가장 우수했으며, 리랭커 모델은 큰 차이를 보이지 않았다. 다양한 임베딩 모델과 리랭커를 이용한 비교는 RAG 평가를 위해 생성된 데이터셋의 질문, 임베딩 벡터 모델과 리랭커를 통한 벡터 검색을 통한 참고 텍스트, 그리고 참고 텍스트를 기반으로 한 RAG 답변에 대해 이루어졌으며, 해당 답변에 사용한 LLM 모델은 질의 분류 모델에서 사용한 Qwen/Qwen3-4.0B을 사용하였다. BAAI/bge-m3는 MTEB에서 제로 샷 성능 99%를 기록하며, 파라미터 수는 5억 6,800만, 임베딩 크기는 1024, 최대 토큰 수는 8194를 갖는다. 본 연구에서는 리랭커 모델로 BAAI/bge-reranker-v2-m3, 임베딩 모델로 BAAI/bge-m3을 선정하여 추가 실험을 진행한다.

4.4 RAG 성능 개선을 위한 방법

임베딩, 리랭커, 답변 모델을 선정한 후 추가적인 성능 개선을 위해 세 가지 방법을 제안한다. 이 추가 방법들은 실시간 성능에 영향을 주지 않는 방식으로 선정되었다. 첫 번째는 질의 확장 방법인 질의 재구성, 두 번째는 하이브리드 검색, 마지막으로 두 가지 방법을 모두 적용했을 경우에 대해 RAGAS를 통한 RAG 평가를 진행하였다. 질의 재구성은 질의 분류 모델에서 동시에 수행 가능하기 때문에 추론 시간에 영향을 미치지 않는다. 표 10은 세 가지 방법을 각각 테스트한 결과를 나타낸다. 실험에 사용된 텍스트를 기반으로 질문을 생성하였고, 생성된 질문을 통한 RAG 답변으로 평가하였기 때문에 유사한 단어나 의미의 다양성을 파악하기 위해 추가 데이터를 활용하여 실험을 진행하였다. 본 연구에서는 AIHUB[32]의 이상행동 CCTV 영상을 추가 데이터로 구축하였다. 이상행동 CCTV 영상은 12가지 이상행동(폭행, 싸움, 절도, 기물 파손, 실신, 배회, 침입, 투기, 강도, 데이트 폭력 및 추행, 납치, 주취 행동)을 포함하여, 총 700시간(8,400컷)의 비디오 데이터셋이다. 해당 데이터는 4~5분 분량의 동영상으로 총 400개를 5초 간격으로 분석하여 VLM을 통해 텍스트로 변환하고 변환하였고, 텍스트를 임베딩 벡터를 통해 벡터로 변환한 후 벡터 데이터베이스에 임의의 채널에 랜덤하게 저장하였다.

AIHUB의 추가 데이터셋을 사용하지 않은 경우, 질의 재구성을 통한 질문은 컨텍스트 정밀도 향상에 도움이 되지만 신뢰성에는 떨어지는 것으로 나타났다. 이러한 이유로

표 9. RAGAS를 통한 RAG 평가 결과
Table 9. RAG Evaluation Results via RAGAS

model / items	Reranker	baai				jinaai				dragonkue			
	Embedding	baai	google	qwen3	intfloat	baai	google	qwen3	intfloat	baai	google	qwen3	intfloat
context precision	0.25	0.7	0.325	0.5	0.7	0.7	0.333	0.5	0.7	0.7	0.325	0.5	0.5
	0.5	**0.804**	0.7	0.756	0.804	**0.806**	0.7	0.756	0.804	**0.806**	0.7	0.75	0.756
	0.75	1.0	0.833	0.917	1.0	1.0	0.887	0.917	0.975	1.0	0.833	0.887	0.917
faithfulness	0.25	0.667	0.5	0.592	0.6	0.625	0.5	0.556	0.571	0.667	0.5	0.6	0.667
	0.5	**0.833**	0.714	0.8	0.8	0.833	0.675	0.778	0.75	0.818	0.714	0.8	0.8
	0.75	1.0	0.875	1.0	1.0	**1.0**	0.875	1.0	1.0	**1.0**	0.875	1.0	1.0
total		0.8185	0.707	0.778	0.802	0.8195	0.6875	0.767	0.777	0.812	0.707	0.775	0.778

표 10. 질의 재구성, 하이브리드 검색, 두 가지 방법을 사용하였을 경우 평가 결과
Table 10. Evaluation results when using query reformulation, hybrid search, and both methods

items \ method		basic	AIHUB dataset not added			AIHUB dataset added		
			hybrid	reformulation	hybrid+ refomulation	hybrid	reformulation	hybrid+ refomulation
context precision	0.25	0.7	0.7	0.7	0.7	0.5	0.7	0.7
	0.5	0.804	0.806	**0.833**	**0.833**	0.806	**0.95**	**0.887**
	0.75	1.0	1.0	1.0	1.0	1.0	1.0	1.0
faithfulness	0.25	0.667	0.667	0.537	0.5	0.5	0.5	0.461
	0.5	0.833	**0.833**	0.733	0.75	0.714	0.733	0.714
	0.75	1.0	1.0	0.889	0.9	0.9	0.881	0.884
total		0.8185	0.8195	0.783	0.7915	0.76	0.842	0.801

컨텍스트 정밀도는 검색된 문서가 질문과 관련이 있는지를 평가하는 지표이며, 신뢰성은 답변이 검색된 문서(컨텍스트)에서 제공된 정보를 기반으로 얼마나 정확하게 생성되었는지를 평가하는 지표이다. 따라서 AIHUB에서 검색에 사용되는 후보 벡터의 수가 많아진 것이 원인으로 보인다.

하이브리드 방법은 밀도 기반 방법과 큰 차이가 없었다. 이는 질문의 단어가 전문 용어가 아닌 단순한 단어들의 조합이기 때문으로 보인다. 반면, AIHUB의 추가 데이터셋을 사용할 경우, 질의 재구성에 따른 컨텍스트 정밀도가 크게 향상되는 것을 확인할 수 있다.

Ⅳ. 결론 및 향후 연구

본 시스템은 다중 채널 서빙 환경이기 때문에 소요 시간이 중요한 요소로 작용한다. 모델 선정 시에는 가급적 소요 시간을 고려하여 소형 모델을 선택하였으며, 또한 영상 보안 시스템 특성상 온프레미스 환경(폐쇄망)에서 적용 가능한 오픈소스 모델을 사용하였다. 본 연구 결과를 종합하면, VLM의 성능은 CLOVAX-Vision이 가장 우수하였고, 질의 분류 모델과 질의 응답에는 Qwen3-4.0B 모델을 활용하였으며, RAG 기반 답변을 위해 임베딩 벡터와 리랭커로 bge 모델을 사용하였다. RAG 성능 개선을 위해 질의 재구성 및 하이브리드 방법 등을 도입하여 RAGAS 평가를 통해 성능을 검증하였다. 본 연구 결과는 영상 보안 및 산업 안전 관리 등의 분야에 직접적으로 활용될 수 있다. 특히 기존 CCTV 시스템을 운영하는 조직들이 대규모 시스템 교체 없이도 AI 기술의 이점을 실용적으로 활용할 수 있는 실용적인 방안을 제시한다. 또한 본 연구에서 개발한 VLM 성능 평가 방법론과 RAG 최적화 기법은 다른 멀티모달 AI 시스템 개발에도 적용 가능하며, 특히 실시간 처리가 중요한 응용 분야에서 유용한 참고 자료가 될 것이다. 오픈소스 기반 구현과 상세한 성능 분석 결과는 연구 재현성을 높이고 후속 연구를 촉진하는 데 기여할 것으로 기대된다. 향후 연구의 영상에서 실시간 처리를 위해 채널 수를 줄이고 고성능 GPU를 활용하여 1초 단위 평가를 수행할 예정이다. 또한, 중요 정보만을 추출하는 영상 요약과 시간적인 정보를 기반으로 한 이상행동 탐지 등, 레거시 영상 분석에서 처리하지 못했던 부분에 대한 연구를 지속할 계획이다.

참 고 문 헌 (References)

[1] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. “You Only Look Once: Unified, Real-Time Object Detection” arXiv:1506.02640, May. 2016.
doi: https://doi.org/10.48550/arXiv.1506.02640

[2] J. Redmon, A. Farhadi. “YOLO9000: Better, Faster, Stronger” arXiv:1612.08242, Dec. 2016.
doi: https://doi.org/10.48550/arXiv.1612.08242

[3] J. Redmon, A. Farhadi. “YOLOv3: An Incremental Improvement. arXiv:1804.02767” Apr. 2018.
doi: https://doi.org/10.48550/arXiv.1804.02767

[4] R. Varghese and S. M. “YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness” 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, pp. 1-6 2024.
doi: https://doi.org/10.1109/ADICS58448.2024.10533619

[5] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft. “Simple Online and Realtime Tracking” arXiv:1602.00763, Jul. 2017.

doi: https://doi.org/10.48550/arXiv.1602.00763

[6] N. Wojke, A. Bewley, D. Paulus. Simple Online and Realtime Tracking with a Deep Association Metric. arXiv:1703.07402, Mar. 2017.
doi: https://doi.org/10.48550/arXiv.1703.07402

[7] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, et al. “ByteTrack: Multi-Object Tracking by Associating Every Detection Box” arXiv:2110.06864, Apr. 2022.
doi: https://doi.org/10.48550/arXiv.2110.06864

[8] N. Aharon, R. Orfaig, B. Bobrovsky. “BoT-SORT: Robust Associations Multi-Pedestrian Tracking” arXiv:2206.14651, Jul. 2022.
doi: https://doi.org/10.48550/arXiv.2206.14651

[9] Y. Du, Z. Zhao, Y. Song, Y. Zhao, Fei Su, Tao Gong, Hongying Meng. StrongSORT: Make DeepSORT Great Again. arXiv:2202.13514, Feb. 2023.

[10] M. B. Shaikh, S. M. S. Islam, D. Chai, N. Akhtar. “From CNNs to transformers in multimodal human action recognition: A survey” arXiv:2405.15813, May 2024.
doi: https://doi.org/10.48550/arXiv.2405.15813

[11] Y. Xiao, H. Xiang, T. Wang and Y. Wang, “Enhanced Industrial Action Recognition Through Self-Supervised Visual Transformers” in IEEE Access, vol. 12, pp. 134133-134143, 2024.
doi: https://doi.org/10.1109/ACCESS.2024.3455749

[12] Gemma Team, A. Kamath, J. Ferret, S. Pathak et al. “Gemma 3 technical report” arXiv:2503.19786, Mar 2025.
doi: https://doi.org/10.48550/arXiv.2503.19786

[13] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu et al. “InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models” arXiv:2504.10479, Apr 2025.
doi: https://doi.org/10.48550/arXiv.2504.10479

[14] A. Yang, B. Yu, C. Li, D. Liu, F. Huang, H. Huang et al. “Qwen2.5-1m technical report” arXiv:2501.15383, Jan 2025.
doi: https://doi.org/10.48550/arXiv.2501.15383

[15] K.M. Yoo, J. Han, S. In, H. Jeon, J. Jeong, J. Kang et al. “HyperCLOVA X technical report” arXiv:2404.01954, Apr 2024.
doi: https://doi.org/10.48550/arXiv.2404.01954

[16] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai et al. “DeepSeek-VL2: Mixture-of-Experts vision-language models for advanced multimodal understanding” arXiv:2412.10302, Dec 2024.
doi: https://doi.org/10.48550/arXiv.2412.10302

[17] A. Marafioti, O. Zohar, M. Farré, M. Noyan et al. “SmolVLM: Redefining small and efficient multimodal models” arXiv:2504.05299, Apr 2025.
doi: https://doi.org/10.48550/arXiv.2504.05299

[18] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah et al. “Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone” arXiv:2404.14219, Aug 2024.
doi: https://doi.org/10.48550/arXiv.2404.14219

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks” arXiv:2005.11401, Apr 2021.
doi: https://doi.org/10.48550/arXiv.2005.11401

[20] Embedding Leaderboard MTEB(Multilingual, v2) https://huggingface.co/spaces/mteb/leaderboard

[21] S. Ockerman, A. Gueroudji, S.Y. Oh, R. Underwood et al. “Exploring Distributed Vector Databases Performance on HPC Platforms: A Study with Qdrant.” arXiv:2509.12384, Sep 2025.
doi: https://doi.org/10.48550/arXiv.2509.12384

[22] A. Ingber, E. Liberty. “Accurate and Efficient Metadata Filtering in Pinecone's Serverless Vector Database” ICML 2025

[23] Chroma https://www.trychroma.com

[24] Weaviate https://weaviate.io

[25] L. Gao, X. Ma, J. Lin, J. Callan. “Precise Zero-Shot Dense Retrieval without Relevance Labels” arXiv:2212.10496, Dec 2022.
doi: https://doi.org/10.48550/arXiv.2212.10496

[26] H. Liu, Z. Wang, X. Chen, Z. Li, F. Xiong, Q. Yu et al. “HopRAG: Multi-hop reasoning for logic-aware retrieval-augmented generation.” arXiv:2502.12442, May 2025.
doi: https://doi.org/10.48550/arXiv.2502.12442

[27] S. Verma. “Contextual compression in retrieval-augmented generation for large language models: A survey.” arXiv:2409.13385, Oct 2024.
doi: https://doi.org/10.48550/arXiv.2409.13385

[28] Q. Tang, J. Chen, Z. Li, B. Yu, Y. Lu, H. Yu et al. “Self-Retrieval: End-to-end information retrieval with one large language model.” arXiv:2403.00801, Nov 2024.
doi: https://doi.org/10.48550/arXiv.2403.00801

[29] X. Wang, C. Macdonald, I. Ounis. “Deep reinforced query reformulation for information retrieval.” arXiv:2007.07987, Jul 2020.
doi: https://doi.org/10.48550/arXiv.2007.07987

[30] T. Zhang, V. Kishore, F. Wu et al. “BERTScore: Evaluating Text Generation with BERT” arXiv:1904.09675, Feb 2020.
doi: https://doi.org/10.48550/arXiv.1904.09675

[31] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang et al. “Qwen2.5-coder technical report.” arXiv:2409.12186, Nov 2024.
doi: https://doi.org/10.48550/arXiv.2409.12186

[32] AiHub https://www.aihub.or.kr

저 자 소 개

장 일 식

- 2011년 2월 : 서울과학기술대학교 NID융합기술대학원 석사
- 2020년 3월 ~ 현재 : 서울과학기술대학교 IoT 융복합 기술 연구소 책임 연구원
- 2020년 9월 ~ 2023년 2월 : 서울과학기술대학교 나노IT디자인융합대학원 정보통신미디어공학전공 박사수료
- ORCID : https://orcid.org/0000-0003-0822-9857
- 주관심분야 : 컴퓨터비전, 생성형 AI

박 구 만

- 1984년 2월 : 한국항공대학교 전자공학과 공학사
- 1986년 2월 : 연세대학교 전자공학과 공학석사
- 1991년 2월 : 연세대학교 전자공학과 공학박사
- 1991년 3월 ~ 1996년 9월 : 삼성전자 신호처리연구소 선임연구원
- 2016년 1월 ~ 2017년 12월 : 서울과학기술대학교 나노IT디자인융합대학원 원장
- 1999년 8월 ~ 현재 : 서울과학기술대학교 스마트ICT융합공학과 교수
- 2006년 1월 ~ 2007년 8월 : Georgia Institute of Technology Dept.of Electrical and Computer Engineering, Visiting Scholar
- ORCID : https://orcid.org/0000-0002-7055-5568
- 주관심분야 : 컴퓨터비전, 지능형실감미디어