한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 31, No. 3, P. 57~66

지능형 혼합현실을 위한 멀티모달 RAG 아키텍처 설계

김한얼° 배종환 정원영 박상훈*

서강대학교 메타버스전문대학원

{skykim, bjh0309, forever, mshpark}@sogang.ac.kr

Design of Multimodal RAG Architecture for Intelligent Mixed Reality

Haneol Kim° Jonghwan Bae Wonyoung Jeong Sanghun Park*

Graduate School of Metaverse, Sogang University

요 약

인공지능(AI)과 혼합현실(MR) 기술의 융합한 사용자 인터랙션 관련 연구가 활발히 진행되고 있으나, 시각 정보가 중요한 MR 환경에서 기존 텍스트 중심 챗봇 방식에는 한계가 명확하다. 본 연구는 이러한 문제를 해결하기 위해 사용자의 음성 명령과 이미지를 동시에 처리할 수 있는 새로운 멀티모달 RAG(검색-증강 생성) 아키텍처를 제안한다. 제안하는 아키텍처는 ColQwen2.5를 검색 엔진으로, GPT-4.1 nano를 비전 및 생성 엔진으로 활용한다. PDF 문서 내용으로부터 실시간으로 구축된 인메모리 벡터 데이터베이스에서 사용자의 멀티모달 쿼리(query)와 가장 관련성 높은 컨텍스트를 검색한 후, 이를 기반으로 GPT-4.1이 최종 답변을 생성하는 구조이다. 본 아키텍처의 유용성을 검증하기 위해 BMW 차량 매뉴얼 기반 챗봇을 Meta Quest 3 환경에서 구현하였다. 실험 결과, 평균 응답 시간 2.46초를 달성했으며, 텍스트만 사용하면서 자동차 부품의 명시적 지칭(예: ”트렁크”, CR 99.3%)을 한 경우와 멀티모달 검색에서 지시대명사(예: ”이것”, CR 94.6%)를 사용한 경우의 성능 차이가 불과 4.7%에 불과해, 시각 정보가 언어적 모호성을 효과적으로 해소함을 명확히 입증하였다. 본 연구는 멀티모달 접근법이 단순히 텍스트의 한계를 보완하는 것을 넘어, 사용자가 정확한 용어를 모르는 상황에서도 시각 정보를 통해 효과적인 정보 검색이 가능함을 보여주어, 실용적인 MR 기반 검색 시스템 구축에 중요한 시사점을 제공한다.

Abstract

While the convergence of Artificial Intelligence (AI) and Mixed Reality (MR) technology is gaining prominence in user interaction research, conventional text-centric chatbots face clear limitations in visually-driven MR environments. This study proposes a novel multimodal Retrieval-Augmented Generation (RAG) architecture that simultaneously processes both voice commands and images to overcome these challenges. The proposed architecture utilizes ColQwen2.5 as its retrieval engine and GPT-4.1 nano as its vision and generation engine. It operates by retrieving the most relevant context for a user’s multimodal query from an in-memory vector database, which is built in real-time from PDF document contents, and subsequently generating a final answer based on this retrieved information. To validate its utility, a chatbot based on a BMW vehicle manual was implemented in a Meta Quest 3 environment. Experimental results showed an average response time of 2.46 seconds. Notably, the performance difference between text-only retrieval with explicit references to car parts (e.g., ”trunk”, CR 99.3%) and multimodal retrieval using demonstrative pronouns (e.g., ”this”, CR 94.6%) was merely 4.7%, clearly demonstrating that visual information effectively resolves linguistic ambiguity. This research shows that the multimodal approach not only complements the limitations of text but also enables effective information retrieval even when users do not know the exact terminology, providing important implications for building practical MR-based retrieval systems.

키워드: 멀티모달 RAG, 혼합현실(MR), 지능형 메뉴얼

Keywords: Multimodal RAG, Mixed Reality(MR), Intelligent Manual

*corresponding author: Sanghun Park / Sogang University (mshpark@sogang.ac.kr)

Received : 2025.06.13./ Review completed : 1st 2025.06.30. 2nd 2025.07.10./ Accepted : 2025.07.14.
DOI : 10.15701/kcgs.2025.31.3.57
ISSN : 1975-7883(Print)/2383-529X(Online)

1 서론

최근 Google에서 AI 모델 Gemini[1]를 탑재한 Google 글래스를 공개하는 등 혼합현실(Mixed Reality) 기술에 AI를 접목하는 시도가 주목받고 있다. Google의 MR 글래스는 실시간 음성 번역, Gemini와의 채팅, 길찾기 등을 보여주는 디스플레이를 탑재하고 있으며, 사진을 촬영하는 것 또한 가능하다고 설명한다. 이는 메타 퀘스트와 같은 HMD(Head Mounted Display) 기반 기기에서 구현되어 온 이미지 및 음성 기반 상호작용을 AI 기술과 융합하여 실생활로 적극 확장하는 계기가 될 것으로 예상된다. 이처럼 MR 기기가 일상으로 들어와 현실 세계의 객체와 상호작용이 늘어날수록 즉각적인 시각 정보 기반의 질의응답(QA)이 중요해지므로, OpenAI의 ChatGPT, Google의 Gemini 등 현재 웹과 모바일에서 대표적으로 사용되는 챗봇 QA 서비스를 넘어 MR 환경의 다양한 정보를 처리할 수 있는 아키텍처 개발이 필요하다.

LLM(Large Language Model) 기반 응답 생성은 환각(hallucination)이나 문맥에 대한 불충분한 이해가 있는 경우 사용자의 이해도를 반감시킬 가능성이 존재한다[2]. 이를 보완하기 위한 방법으로 RAG 구조가 주목받은 바 있다. RAG(Retrieval-Augmented Generation)는 외부 데이터베이스에 저장된 지식과 사용자 질의를 연동하여 보다 정확하고 문맥에 맞는 응답 생성이 가능하게 한다. 이는 다양한 도메인에 활용할 수 있으며, 학습 분야에서도 적극 활용되고 있다. RAG를 활용한 학습용 챗봇은 특정 도메인의 전문성을 높이고, 학습 자원에 대한 접근성을 높이는데 기여한다[3]. 그러나 기존 RAG는 MR 환경에서와 같이 시각적 단서와 음성 등을 활용한 복합적인 질문에 대해 사용자 의도를 정확하게 파악하여 답변을 제공하는 데 어려움이 있다. 기존의 챗봇 시스템은 대부분은 단일 형태의 입력(텍스트 등)에 의존하거나, 정적인 콘텐츠를 기반으로 설계되어 있기 때문이다[4]. 그에 따라 사용자 입력을 실시간으로 이해하고 응답할 수 있는, MR 환경에서 작동하는 RAG 챗봇 아키텍처의 개발을 구상하게 되었다.

이미지, 오디오, 비디오 등 다양한 데이터 입력을 가능하게 하기 위해, 멀티모달 모델을 활용한다. 멀티모달 모델은 크게 '이해' 모델과 '생성' 모델로 나뉘는데, 그 중 이해 모델(VLU:Video-Language Understanding)은 다중 데이터 입력으로부터 수신하고, 추론 및 생성을 통해 출력이 가능한 모델이다. 최근 VLU의 발전은 일반 이해를 가능하게 하였으며, GPT-4V, Gemini, Qwen 시리즈, InternVL 시리즈 등이 최근 모델에 속한다. 특히 Qwen2-VL은 동적 해상도 처리와 강력한 다중 입력 처리를 위해 Multimodal Rotary Embedding(M-RoPE)을 활용한다[5]. M-RoPE를 통해 1D 텍스트, 2D 이미지, 3D 비디오의 위치 정보를 동시에 이해할 수 있게 설계되었다[6]. 이러한 특성은 MR 환경에서 필요한 다중 입력 처리를 효과적으로 수행할 수 있게 한다.

본 연구는 이러한 기존 챗봇의 한계를 극복하고, MR 환경에서 효율적인 정보 검색 및 답변 생성을 가능하게 하는 새로운 멀티모달 RAG 챗봇 아키텍처를 제안하고자 한다. 특히, 자동차 사용 설명 챗봇 개발 사례를 통해 복잡한 자동차 내부 버튼 식별 및 사용 절차 안내와 같이 시각 정보가 필수적인 분야에서 제안하는 아키텍처의 유용성을 입증하고자 한다. 궁극적으로 본 연구는 MR 환경에서 사용자에게 더욱 직관적이고 몰입감 있는 정보 접근 경험을 제공하고, 실제 작업 환경에서의 생산성 향상에 기여할 수 있는 기반을 마련하는 것을 목표로 한다.

2 관련 연구

2.1 RAG

RAG는 대규모 언어 모델(LLM)의 응답 정확성과 사실성을 향상시키기 위해 검색(retrieval)과 생성(generation)을 결합한 하이브리드 아키텍처이다. 앞서 언급했듯이 이는 LLM이 지닌 대표적인 한계인 환각 문제와 최신 정보 및 도메인 지식 부족을 효과적으로 보완할 수 있는 접근 방식으로 주목받고 있다.

RAG라는 용어는 2020년 발표된 논문 "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"에서 처음 소개되었다[7]. 초기의 RAG는 모델이 학습한 parametric 지식에 외부 문서 기반의 non-parametric 지식을 결합하여 파인튜닝하는 방식에 집중하였다. 그러나 2022년 ChatGPT 등장 이후, RAG는 모델 재학습 없이도 외부 지식을 실시간으로 활용할 수 있는 추론 기반 접근으로 진화하였다. 현재 RAG는 LLM의 추론 능력을 활용하여 지식 보강 및 정보 갱신 수단으로 널리 활용되고 있다.

RAG는 일반적으로 다음의 세 단계로 구성된다. 먼저 인덱싱(indexing) 단계에서는 PDF, HTML 등 다양한 형태의 원본 데이터를 일관된 텍스트 형식으로 변환한 후, 이를 청크(chunk) 단위로 분할하고 임베딩(embedding)하여 벡터 데이터베이스에 저장한다. 검색 단계에서는 사용자 쿼리를 벡터화(vectorize)하고 저장된 벡터들과의 유사도(similarity)를 기반으로 관련성이 높은 청크들을 추출한다. 마지막으로 생성 단계에서는 쿼리와 검색된 문서들을 결합하여 LLM에 입력하고, 최종 응답을 생성한다.

최근 RAG 연구는 크게 세 가지 패러다임으로 구분된다. Naive RAG는 가장 기본적인 형태로, 단순한 인덱싱-검색-생성 흐름을 따른다. 그러나 검색 정확도 저하, 정보 중복 및 문맥 불일치 등의 한계가 존재한다. 이를 보완하기 위한 Advanced RAG는 사전 검색 단계에서 쿼리 재작성 및 데이터 구조 최적화, 사후 검색 단계에서 재랭킹, 문맥 압축, 중요 정보 강조 등의 기법을 포함한다. 또한, Modular RAG는 검색 모듈의 추가 구성, 파인튜닝 가능성 등 적응성을 강화하기 위한 진화된 형태로 연구되고 있다[8].

특정 도메인 특화를 위해 키워드 기반의 정확한 검색이 필요한데, 이에 따라 최근에는 벡터 검색과 전통적인 키워드 검색을 결합하는 하이브리드 검색 방식이 주목받고 있다. 그러한 접근을 구현한 Google Vertex AI 등을 비롯하여, 외부 그래프 구조 데이터베이스를 기반으로 하는 GraphRAG 등 RAG 기법을 활용한 다양한 접근을 통해 검색 정확도와 LLM 활용 효율성을 동시에 향상시키고 있다[9].

2.2 멀티모달 RAG

최근 인공지능 연구는 텍스트뿐 아니라 이미지, 음성, 비디오 등 다양한 형태의 데이터를 동시에 처리할 수 있는 멀티모달 인공지능(multimodal AI)으로 확장되고 있다[10]. 멀티모달 AI는 서로 다른 모달리티 간의 의미적 상관관계를 파악함으로써 단일 모달 기반 모델이 갖는 정보 해석의 한계를 극복할 수 있으며[11], 이는 인간과의 자연스러운 상호작용, 복합적인 의사결정, 실세계 환경 이해가 중요한 응용 분야의 핵심 기술로 자리 잡고 있다. 이러한 흐름에 발맞춰 RAG 또한 텍스트 중심의 구조에서 벗어나 이미지, 음성 등을 함께 처리하는 멀티모달 RAG로 진화하고 있다[12]. 멀티모달 RAG는 텍스트 외에도 이미지, 음성, 센서 데이터 등 다양한 입력을 받아 외부 정보를 검색하고 이를 바탕으로 생성 결과를 도출한다. 예를 들어, 이미지 설명, 도면 해석, 비디오 요약과 같은 복합 정보 요청에 대해 적절한 응답을 생성하는 것이 가능하다. 사용자의 음성 질문에 대해 시각 정보를 기반으로 응답하거나, 특정 이미지를 참조하여 복잡한 문맥에 대응하는 기능도 구현할 수 있다. 멀티모달 RAG 구현에는 CLIP, Flamingo와 같이 텍스트와 이미지 임베딩을 동일 벡터 공간에서 정렬할 수 있는 멀티모달 모델들이 활용된다[13]. 이들은 멀티모달 데이터를 벡터화하여 벡터 데이터베이스에 저장하고, 이를 바탕으로 검색 및 생성을 수행한다. 성공적인 멀티모달 RAG 구현을 위해서는 각 모달리티 간 퓨전 전략, 쿼리 재작성 및 멀티모달 문맥 압축, 모달리티 중요도 조정 등 정교한 설계가 요구된다.

2.3 MR 기술과 RAG 활용

MR 기술은 몰입형 상호작용 환경을 제공한다. 이러한 특성을 통해 MR 기술을 활용해 고급 작업을 위한 시뮬레이션, 프로토타이핑, 공동 설계, 상황 인식, 즉각적 피드백 제공 등이 가능하다[14]. 산업 현장이나 의료 시뮬레이션과 같은 몰입형 MR 애플리케이션에서는 사용자의 음성이나 제스처 기반 질의에 즉각적이고 정확한 도메인 지식을 제공해야 한다. 이때 RAG를 함께 활용하면 외부 지식을 효과적으로 검색하고 적시에 사용자에게 제공할 수 있다[15]. RAG는 관련 문서를 사전 인덱싱하고, 실시간 쿼리에 따라 의미 기반으로 정보를 검색 및 조합하여 LLM을 통해 응답을 생성함으로써, 사용자 경험을 향상시킬 수 있다. Despina 등은 연구를 통해, 산업 환경에서 지식 전달 문제를 해결하기 위해 강화된 RAG를 활용한 LLM과 MR을 통합하는 시도를 한 바 있다[16]. 이들은 특정 도메인에 특화된 지식을 MR 환경에서 작업자들이 두손을 자유롭게 하며 AI Agent를 통해 피드백 받을 수 있도록 하는 아키텍처를 설계했다. MR 환경은 시각적 · 공간적 정보가 풍부하기 때문에, 이를 이해하고 활용하는 데 있어 멀티모달 RAG의 활용 가능성이 높다.

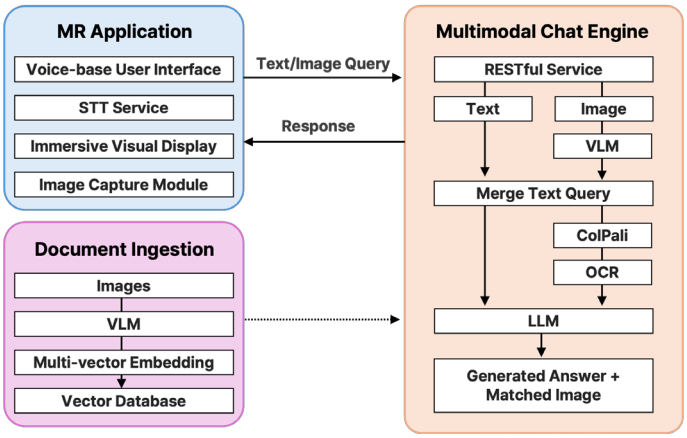

Figure 1: Overall Pipeline

3 챗봇 시스템 개발

3.1 프레임워크 개요

본 연구에서는 멀티모달 RAG 아키텍처를 MR 환경에 적용함으로써, 사용자와 시스템 간의 몰입형 정보 검색 및 질의응답 상호작용을 실현하는 프레임워크를 제안한다. 제안된 프레임워크는 Figure 1과 같이 MR 어플리케이션, 문서처리 시스템, 그리고 멀티모달 챗봇 엔진 세 가지 주요 구성 요소로 이루어진다. 이들 각 단계는 데이터 흐름을 기반으로 상호연계된다.

MR 애플리케이션은 사용자와 시스템 간의 인터페이스 역할을 수행하며, 음성 기반 사용자 인터페이스, 디스플레이, 이미지 캡처 모듈 등으로 구성된다. 사용자 질의는 음성 또는 이미지의 형태로 입력된다. 이 중 음성 입력은 사용자의 자연어 질의를 수용하기 위한 핵심 입력으로, Google의 YAMNet으로 사용자의 음성발화 여부를 판단하고, OpenAI의 Whisper 모델을 통해 음성 신호가 텍스트로 변환되며, 이미지는 캡처되어 멀티모달 챗 엔진으로 전송된다[17].

문서 처리는 PDF와 같은 정형 문서에서 정보를 추출하고 벡터화하여 데이터베이스를 구성하는 백엔드 프로세스이다. 이미지는 VLM(Vision-Language Model)을 통해 문서 내 시각적 및 언어적 정보를 추출하여 Vector 데이터베이스에 저장되며, 이는 이후 ColPali 아키텍처인 ColQwen2.5 모델을 이용하여 텍스트 쿼리를 입력 받아 가장 관련성이 높은 이미지(페이지)를 검색하고, 그 관련성을 나타내는 유사도 점수를 계산하여 순위화할 수 있다[18].

멀티모달 챗봇 엔진은 위 과정을 통해 입력된 질의에 대해 적절한 검색 경로를 결정하고, 검색된 문맥을 바탕으로 최종 응답을 생성하는 핵심 처리 모듈이다. RESTful API Service를 통해 MR 애플리케이션으로부터 질의를 수신한 후, 이미지 정보는 VLM을 통해 전처리된 후 텍스트 질의와 병합된다. 이후 관련 문맥을 검색하고, 이를 통합하여 최종적으로 LLM을 통해 응답을 생성한다. 응답은 생성된 답변과 함께 참조 이미지, 페이지 정보 등으로 구성되며 MR 애플리케이션으로 반환된다.

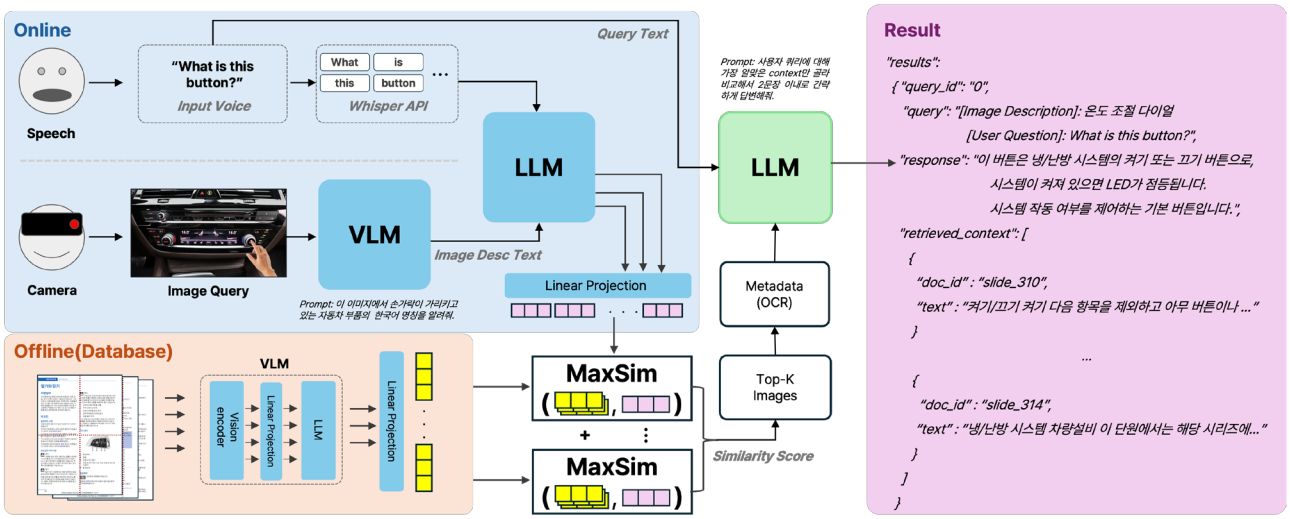

Figure 2: Multimodal RAG Architecture

3.2 데이터베이스 구축

RAG 아키텍처 구현을 위해, PDF 기반의 입력 문서를 통해 시각적 정보 기반의 이미지 데이터베이스를 구축한다. 특히, PDF 문서의 각 페이지는 개별 슬라이드 이미지(PNG)로 추출되어 Image DB에 저장된다. 이는 MR 환경에서 슬라이드를 가시화하거나, 이미지 유사도 기반의 검색 시스템 구현에 활용될 수 있다. 슬라이드 이미지 자체가 인터페이스 요소로 작동할 수 있기 때문에, 시각 정보의 보존과 빠른 접근성이 중요하다. 이렇게 구축된 데이터베이스는 MR 환경에서 음성 또는 자연어 질의에 대한 의미 기반 검색 및 멀티모달 응답 생성을 위한 핵심 요소로 작동하며, 이미지와 텍스트를 통합적으로 참조하고 응답을 생성하는 RAG 아키텍처의 기반이 된다. Figure 3는 이러한 전체적인 흐름을 보여준다.

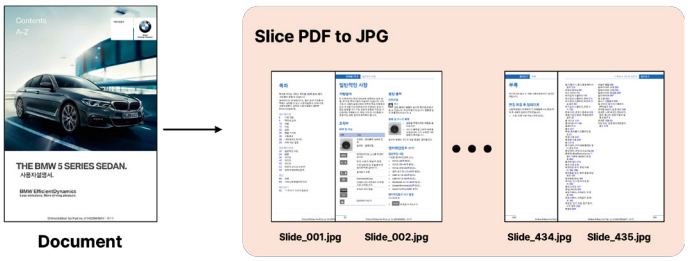

Figure 3: Database Construction Workflow

3.3 멀티모달 RAG 아키텍처

RAG 처리를 위해, 텍스트 및 이미지 입력을 동시에 처리하여, 복합적인 멀티모달 입력에 대해 정확하고 문맥에 맞는 응답을 생성하는 아키텍처를 제안한다. 제안하는 아키텍처는 PDF 문서 처리 및 인덱싱 단계와 쿼리 처리 및 응답 생성 단계로 구성되며, 각 구성요소는 효율적인 검색 및 추론을 위해 상호 유기적으로 동작한다. 전체 구조는 Figure 2에 제시되어 있다.

본 시스템은 사전에 정의된 단일 PDF 문서를 기반으로 동작한다. 시스템 시작 시, 이 PDF 문서의 모든 페이지는 먼저 개별 이미지 파일로 변환되어 임시 캐시 폴더에 저장된다. 이후, 검색 시스템에서는 텍스트 기반 쿼리 모델이 캐시된 각 페이지 이미지를 입력받아 다중 벡터 표현(multi-vector representation)을 생성한다. 이렇게 생성된 페이지 임베딩과 각 이미지 파일 경로는 검색을 위해 인메모리(in-memory) 데이터베이스로 구축되며 이 과정은 서버 시작 시 단 한 번만 수행(offline)된다.

사용자는 텍스트 단독, 이미지 단독, 또는 텍스트와 이미지의 조합, 세 가지 형태로 질의할 수 있으며, 각 입력 유형에 따른 처리 경로는 다음과 같다.

- 입력 처리: 쿼리에 이미지가 포함될 경우(이미지 단독 또는 이미지와 텍스트), 먼저 GPT-4.1 nano 모델이 이미지를 분석하여 상세한 텍스트 설명(caption)을 생성한다.
- 최종 쿼리 구성: 생성된 이미지 캡션은 사용자의 텍스트 쿼리와 결합되어, 검색을 위한 최종 컨텍스트 쿼리를 구성한다. 만약 텍스트 쿼리만 있는 경우, 해당 텍스트 쿼리가 최종 쿼리가 된다. 이는 Figure 2에서 Merge Text에서 이루어진다.
- 관련 페이지 검색(retrieval): 구성된 최종 쿼리는 ColQwen2.5 모델에 전달된다. ColQwen2.5는 텍스트 쿼리를 임베딩한 후, 인메모리 DB에 구축된 페이지 임베딩과 비교하여 가장 관련성이 높은 Top-K개의 페이지 파일 경로를 반환한다.
- 답변 생성(generation): 검색을 통해 얻은 Top-K 페이지의 텍스트 설명과 사용자의 원본 질문은 최종적으로 GPT-4.1 nano 모델에 전달된다. GPT-4.1 nano는 주어진 질문과 검색된 컨텍스트를 종합적으로 이해하여 최종 답변을 생성한다.

결론적으로, 제안된 RAG 아키텍처는 ColQwen2.5 모델을 핵심 검색 엔진으로, GPT-4.1 nano을 핵심 인식(vision) 및 생성(generation) 엔진으로 사용하여 다양한 형태의 멀티모달 입력을 효과적으로 처리한다. 이 구조는 텍스트와 이미지 정보의 유기적인 결합을 통해, 사용자의 복합적인 질의 의도를 깊이 이해하고 정확한 정보를 제공하는 것을 목표로 한다.

3.4 MR 구현

앞서 제안한 멀티모달 RAG 아키텍처를 MR 환경에 적용하기 위해서는 사용자 입력 처리, 정보 검색, 그리고 결과 제시의 각 단계에서 MR 특유의 상호작용 및 시각화 방식을 고려한 구현이 필요하다. 사용자 인터페이스 설계, 센서 데이터 활용, 그리고 정보 시각화 기술의 통합 등의 기술이 이에 해당한다.

MR 환경에서의 사용자 입력은 단순한 텍스트 입력을 넘어, 이미지 입력, 음성 입력과 같은 모달리티(modality)를 포함한다. 이미지 입력으로는 MR 기기의 패스쓰루(passthrough) 카메라 API를 통한 입력이 가능하며, 화면 캡처를 통한 방식이 주요 방식이며 손가락으로 일부 영역을 특정하게 할 수도 있다. 사용자는 특정 객체나 영역을 손가락으로 가리키는 행위를 포함하여 시각적 입력을 전달할 수 있으며(프롬프트 예시:“이 이미지에서 손가락이 가리키고 있는 자동차 부품의 한국어 명칭을 알려주세요.”), 이는 실시간으로 캡처되어 객체 인식 또는 이미지 분석 기술로 처리된다. 또한, MR 디바이스의 마이크로부터 얻어지는 음성 신호는 YAMNet을 통해 521종류의 소리 중에서 실제 사람의 음성 여부를 판단하고, 이후 Whisper 모델을 통해 MelSpectogram 신호로 변환 후, 자연어 텍스트로 변환(speech-to-text)된다. 이처럼 음성만으로도 상호작용이 가능하며, 이렇게 수집된 멀티모달 입력은 RAG 아키텍처의 입력 모듈로 전달되어 텍스트 쿼리 형태로 통합 처리된다. 검색 단계에서는 사용자의 질의에 대응하여 의미적으로 관련 있는 문맥을 이미지 데이터베이스로부터 연관된 슬라이드 이미지나 PDF 페이지를 검색한다.

정보 제시는 MR 환경의 시각적 특성을 고려하여 몰입감 있고 직관적인 방식으로 이루어진다. 검색된 문서 정보나 생성된 답변은 MR 기기의 패스쓰루 모드에서 실제 환경 위에 증강 현실 오버레이 형태로 가상 UI 패널 또는 3D 오브젝트로 시각화된다. 복잡한 정보를 탐색할 필요가 있는 경우, 가상 공간 내에서 상호작용 가능한 정보 인터페이스를 통해 사용자가 직접 조작하거나 탐색을 확장할 수 있다. 특히, 생성된 텍스트와 함께 해당 내용의 근거가 되는 PDF 이미지를 함께 제시함으로써 정보의 신뢰성과 이해도를 높인다. 또한, 사용자 시야각이나 이동 경로의 변화에 따라 정보의 크기, 위치, 투명도 등이 동적으로 조정되는 레이아웃 설계를 통해 시각적 피로를 최소화하고 사용자 경험을 극대화한다.

이와 같은 구현을 통해 사용자는 MR 환경 내에서 시각, 음성, 공간 입력을 결합한 멀티모달 상호작용을 기반으로 필요한 정보를 자연스럽게 질의하고, 해당 정보를 현실 세계와 연동된 형태로 직관적으로 확인할 수 있다. 이는 정보 탐색과 활용의 효율성을 크게 향상시킬 뿐만 아니라, 몰입형 지식 탐색 환경 구축에 있어 핵심적인 기반이 된다.

4 자동차 사용 설명 챗봇 구현

본 장에서는 앞서 설명한 챗봇 시스템 프레임 워크를 자동차 사용 설명 사례에 적용하여 프레임워크의 실현 가능성 및 효율성에 대해 탐색하고자한다.

4.1 개발 환경 및 데이터셋

사례 구현을 위한 개발 환경은 Ubuntu 24.04 LTS 운영체제를 기반으로 하며, AI 모델 학습 및 추론 가속을 위해 NVIDIA RTX 4090 GPU를 활용한다. 핵심적인 MR 인터페이스 구현 및 사용자 상호작용은 MR 기능을 지원하는 Meta Quest 3 디바이스를 통해 이루어진다. 게임엔진으로는 Unity6(6000.0.40f1)의 URP 환경에서 Meta XR Interaction SDK(72.0.0)를 사용하여 MR 환경 내에서 디지털 콘텐츠를 실제 영상 위에 정합 및 오버레이하고, Inference Engine(2.2.1)을 사용하여 사용자 인터페이스 및 음성인식 기능을 구현한다. 멀티모달 RAG 아키텍처의 서버는 Python(3.10.18) 환경에서 개발하였으며, PyTorch(2.7.1), CUDA(11.8), Transformers(4.51.3), OpenAI(1.84.0), FastAPI(0.115.12)를 주요 라이브러리로 사용하였다.

RAG 시스템의 지식 기반(knowledge base)은 BMW의 공식 사용자 매뉴얼인「THE BMW 5 SERIES SEDAN 사용자 설명서」(341페이지)와「내비게이션, 엔터테인먼트, 통신 사용자 설명서」(94페이지)를 기반으로 구축하였다. 이 두 문서를 대상으로, 3장에서 서술한 데이터베이스 구축 절차를 동일하게 적용하였다. 시스템의 성능 평가를 위한 테스트 데이터셋은 다음과 같이 구성하였다. 먼저 Figure 4와 같이 5가지 차량 부품(트렁크, 연료탱크 마개, 디스플레이, 에어컨 온도 조절 다이얼, 계기판)을 손가락으로 가리키는 이미지를 활용하였다. 그리고 각 부품에 대해 앞서 언급한 매뉴얼에서 답변을 찾을 수 있는 질문을 생성하여, 총 50개의 질의응답 세트를 구축하였다.

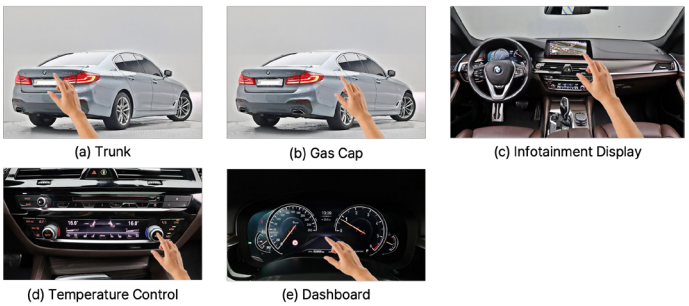

Figure 4: Input Images

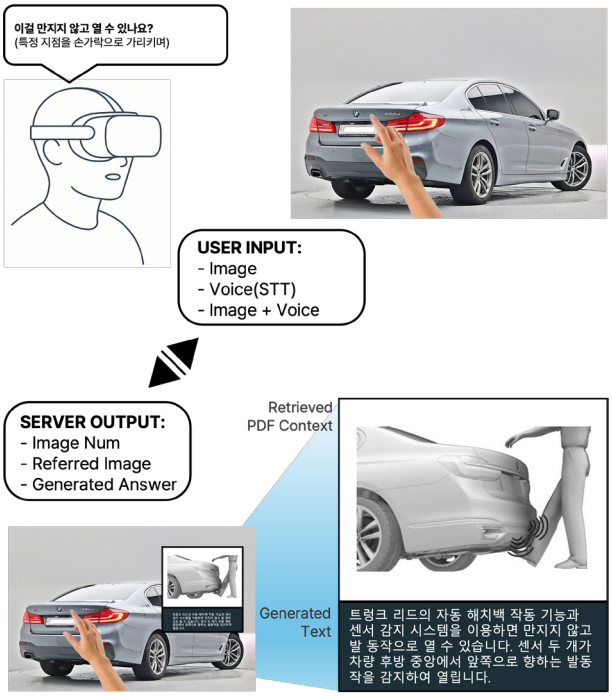

Figure 5: Workflow Visualization

Figure 6: Result Visualization in MR

4.2 구현 결과

개발한 시스템을 기반으로, Figure 5에서 확인할 수 있듯이 BMW 차량의 사용 설명을 지원하기 위한 사용자 상호작용 애플리케이션을 구현하여 실제 적용하였다. 구현된 애플리케이션은 MR 헤드셋을 착용한 사용자가 차량 내부의 특정 지점을 손가락으로 가리키며 음성으로 질문을 입력하는 방식으로 작동한다. 입력된 이미지와 음성 정보를 기반으로 서버에서 관련 정보를 검색하고 자연어 형태의 응답을 생성하는 과정을 포함한다.

사용 사례는 다음과 같다. 사용자가 MR 환경에서 차량의 특정 요소를 지목하고 “이걸 만지지 않고 열 수 있나요?”와 같은 질문을 하면, 이는 이미지 인식 결과인 “트렁크(혹은 트렁크 리드)”와 음성 인식결과를 종합하여 시스템에 입력한다. 입력된 정보는 서버로 전송되며, 서버는 차량 관련 매뉴얼이나 문서를 포함한 PDF 자료에서 관련 문맥을 검색한다. 이 검색 과정에서는 사용자의 질문과 가장 관련성 높은 페이지를 식별하며, 해당 문서 내용을 바탕으로 “트렁크 리드를 만지지 않고 열 수 있습니다. (중략)”와 같이 질의에 대한 답변을 생성한다.

서버는 최종적으로 세 가지 정보를 출력한다. 관련 이미지의 번호, 사용자가 가리킨 대상과 관련된 참조 이미지, 질의에 대한 자연어로 생성된 답변이다. 이러한 결과는 Figure 6에서 볼 수 있듯이 MR 인터페이스 상에서 사용자에게 시각적으로 제공된다.

본 시스템은 직관적인 사용자 경험을 제공할 수 있다. 사용자는 실제 차량 환경에서 MR 기기를 착용하고 자신이 궁금한 부분을 직접 가리키고 질문할 수 있으며, 이에 대한 정보를 시각적 · 언어적으로 동시에 얻을 수 있다. 이 시스템은 향후 차량 매뉴얼의 대체 수단으로 활용될 수 있으며, 운전자가 차량 기능에 대해 필요한 정보를 실시간으로 얻을 수 있게 함으로써 사용자 만족도를 높일 수 있다. 더하여, 차량 고장 시 자가 진단을 통해 서비스 센터 방문을 줄일 수 있으며, 신규 운전자 교육 및 기능 학습 도구로서도 활용 가능성이 있다.

5 실험 및 평가

멀티모달 RAG 아키텍처 기반 챗봇 시스템의 실제 MR 환경 적용 가능성과 효용성을 확인하고자, 구현된 자동차 사용 설명 챗봇에 대한 실험 설계 및 실험 결과를 제시하고자 한다.

5.1 평가 지표

평가는 전체 소요시간 측정과 정량적 검증 방법을 활용한다. 정량적 검증에는 Amazon에서 개발한 RAGChecker를 사용한다[19]. RAGChecker는 Overall Metrics, Retriever Metrics, Generator Metrics를 통해 RAG 시스템의 전반적인 성능, 검색 성능, 생성 성능을 측정하도록 설계되었다. 본 연구에서는 그러한 측정 메트릭 중 전반적인 성능을 평가하는 검색 성능을 측정하는 정확한 검색률(CR: Claim Recall), 내용 정밀도(CP: Context Precision), 신뢰성(F: Faithfulness), 환각 현상(H: Hallucination) 지표를 선택하여 그 결과를 제시한다.

Claim Recall은 검색된 컨텍스트에서 질문과 관련된 모든 중요한 클레임들이 얼마나 잘 포함되어 있는지를 측정하는 지표로, 검색기가 답변 생성에 필요한 정보를 완전하게 찾아내는 능력을 평가한다. Context Precision은 검색된 컨텍스트 중에서 실제로 질문 답변과 관련된 정보의 비율을 측정하여 검색 정보의 품질과 관련성을 평가한다. Faithfulness는 생성된 답변이 제공된 컨텍스트 정보에 얼마나 충실하게 기반하고 있는지를 측정하는 지표로, 답변의 각 문장이 검색된 컨텍스트에서 직접 도출되거나 논리적으로 추론 가능한지를 확인한다. Hallucination은 생성 모델이 제

공된 컨텍스트에 없는 정보를 마치 사실인 것처럼 생성하는 현상의 발생 정도를 측정하며, 낮은 점수가 더 좋은 성능을 나타낸다. 이러한 지표들은 RAG 시스템의 검색 품질과 생성 품질을 종합적으로 평가하여 시스템의 전반적인 성능과 신뢰성을 정량적으로 측정할 수 있게 한다.

$$CR = \frac{|\{c_i^{(gt)} | c_i^{(gt)} \in \{\text{chunk}_j\}\}|}{|C^{(gt)}|} \tag{1}$$

$$CP = \frac{|\{r\text{-chunk}_j\}|}{k} \tag{2}$$

$$F = \frac{|\{c_i^{(m)} | c_i^{(m)} \in \{\text{chunk}_j\}\}|}{|C^{(m)}|} \tag{3}$$

$$H = \frac{|\{c_i^{(m)} | c_i^{(m)} \notin \text{gt and } c_i^{(m)} \notin \{\text{chunk}_j\}\}|}{|C^{(m)}|} \tag{4}$$

5.2 실험 결과

5.2.1 평균 실행 소요 시간 분석

시스템의 실시간성을 평가하기 위해 전체 평균 실행 소요 시간을 측정하였다. 평균 총 소요 시간은 2.46초로, MR 환경에서 사용자와의 원활한 상호작용이 가능한 수준의 응답 속도를 보였다. 세부적으로는 입력 이미지에 대한 캡션 생성(captioning)에 1.56초, 435페이지 분량의 벡터 DB로부터 관련 문서를 검색(retrieval)하는 데 0.07초, 그리고 검색된 내용을 바탕으로 LLM이 최종 답변을 생성(generation)하는 데 0.85초가 소요되었다. 특히 벡터 DB를 활용한 검색 과정이 매우 효율적으로 이루어짐을 확인할 수 있다.

Table 1: Average Execution Time(seconds)

Captioning	Retrieval	LLM	Total
1.56	0.07	0.85	2.46

5.2.2 시나리오별 실험 결과

본 실험에서는 텍스트 전용 검색과 멀티모달 검색 환경에서의 성능 차이를 분석하기 위해 3가지 주요 시나리오를 설계하여 평가하였다. 각 시나리오는 명시적 지칭과 지시대명사 사용, 시각 정보 유무에 따른 검색 효율성을 측정하며, 별도 언급이 없는 경우 상위 5개(Top-5) 검색 결과를 기반으로 진행되었다. 평가 지표로는 5.1에서 언급한 CR, CP, F, H를 사용하였다.

1. **텍스트 기반 질의: 명시적 지칭과 지시대명사 사용**
    - Q1) 리모콘으로 트렁크를 어떻게 여나요? (이미지: 없음)
    - Q2) 리모콘으로 이것을 어떻게 여나요? (이미지: 없음)

    텍스트만 사용하는 경우, 사용자가 부품명을 명확히 지칭(Q1)했을 때 CR 99.3%, CP 70.4%로 매우 높은 검색 성능을 보였다. 하지만 이미지 정보 없이 "이것"이라는 지시대명사를 사용(Q2)하자 CR은 79.2%, CP는 58%로 급감하며 검색 성능이 크게 저하되었다. 이는 텍스트만으로는 지시대명사가 지칭하는 대상을 특정할 수 없어 정보 검색에 한계가 있음을 명확히 보여준다. 그러나 해당 수치가 완전히 낮지 않은 것은 텍스트 내에 부품을 추론할 수 있는 단서가 포함되어 있기 때문으로 해석된다. 예를 들어, "이것이 열리다가 멈추는 경우는 어떤 때인가요?"라는 질문에 대해서는 "트렁크가 열리다가 멈추는 경우는 차량이 움직이거나 (중략) 기능 장애가 있을 때입니다."라는 답변을 제공했다. 이는 "열리다가 멈추는" 동작이 트렁크의 고유한 특성으로 인식되어 해당 부품에 관한 정보가 검색되었기 때문이다. 반면, 본래 연료 탱크 마개를 대상으로 한 "이것을 닫을 때 주의할 점은 무엇인가요?"라는 질문에 대해서는 "트렁크 리드 또는 도어를 닫을 때는 신체 일부가 끼지 않도록 (중략) 조심하시기 바랍니다."라는 부정확한 답변이 제공되었다. 이러한 결과는 텍스트만을 활용하는 RAG 시스템에서 지시대명사 사용 시 의도하지 않은 검색 결과로 인해 시스템의 정확도가 저하될 수 있음을 시사한다.

2. **텍스트 기반과 멀티모달 접근법 비교**
    - Q3) 컨트롤러로 디스플레이를 어떻게 조작하나요? (이미지: 없음)
    - Q4) 컨트롤러로 이것을 어떻게 조작하나요? (이미지: 디스플레이)

    동일한 질의 상황에서 텍스트 기반 접근법(Q3)과 멀티모달 접근법(Q4)을 직접 비교한 결과, 텍스트만 사용했을 때 CR 99.3%, CP 70.4%를 기록했으며, 멀티모달을 사용했을 때 CR 94.6%, CP 66.4%를 보였다. 흥미롭게도 명시적 지칭을 사용한 텍스트 기반 접근법이 약간 더 높은 성능을 보였으나, 멀티모달 접근법도 여전히 높은 성능을 유지했다. 이는 멀티모달 접근법이 사용자의 부품명 지식 부족을 시각 정보로 효과적으로 보완하여 정보 검색의 한계를 극복할 수 있음을 보여준다. 다만 멀티모달 접근법의 경우 VLM의 부품 인식 결과에 의존하여 검색을 수행하기 때문에, 정확한 명시적 지칭 대비 성능 저하가 발생할 수 있다. 예를 들어, "이 이미지에서 손가락이 가리키고 있는 자동차 부품의 한국어 명칭을 알려줘"라는 프롬프트에서 VLM은 트렁크를 "트렁크(또는 트렁크 리드)"로, 연료 탱크 마개를 "연료 주입구(주유구)" 또는 "주유구"로 인식하는 등 일관성 있는 명칭을 제공하지 못해 검색 품질에 직접적인 영향을 미쳤다. 향후 VLM의 프롬프트 엔지니어링 개선이나 사전 정의된 부품명 목록에서 선택하도록 하는 방식을 통해 보다 일관성 있는 결과를 얻을 수 있을 것으로 기대된다.

3. **검색 결과 수(K)에 따른 성능 변화**

   - Q5) 주행하지 않을 때도 이것을 사용할 수 있나요? (이미지: 에어컨 온도 조절 다이얼) - Top-1
   - Q6) 주행하지 않을 때도 이것을 사용할 수 있나요? (이미지: 에어컨 온도 조절 다이얼) - Top-3
   - Q7) 주행하지 않을 때도 이것을 사용할 수 있나요? (이미지: 에어컨 온도 조절 다이얼) - Top-5

   검색 문서의 수(K)를 변경하는 경우, K=1로 제한했을 때 CP 88.0%로 높은 정밀도를 보였다. 하지만 K=5로 증가하자 CP는 66.4%로 하락하며 정밀도가 저하되었다. 반면 CR은 70.8%에서 94.6%로, F는 63.6%에서 95.2%로 크게 향상되었고, H는 19.4%에서 3.7%로 급격히 감소했다. 이는 더 많은 검색 문서를 참고할수록 CR과 신뢰성은 향상되지만, 관련 없는 정보 포함으로 인해 CP는 자연스럽게 저하되는 트레이드오프 관계를 명확히 보여준다. 특히 H의 급격한 감소는 단일 문서 의존 시 정보 부족으로 인한 허위 정보 생성 위험이 높아지며, 다수 문서 활용을 통해 근거 기반의 신뢰할 수 있는 답변 생성이 가능함을 시사한다.

Table 2에서 확인할 수 있듯이, 텍스트 기반 검색에서는 명시적 지칭(Q1: CR 99.3%)이 지시대명사 사용(Q2: CR 79.2%)보다 현저히 우수한 성능을 보였다. 반면 멀티모달 검색에서는 이러한 성능 격차가 크게 줄어들어(Q3: CR 99.3% vs Q4: CR 94.6%), 시각 정보가 언어적 모호성을 효과적으로 해결함을 확인할 수 있다. 검색 범위를 Top-1에서 Top-5로 확장 시 CP는 88.0%에서 66.4%로 감소했지만 CR은 70.8%에서 94.6%로, F는 63.6%에서 95.2%로 향상되고 H는 19.4%에서 3.7%로 감소하여 정보 검색 시스템 설계에서 정확성과 완전성 간 균형의 중요성을 보여준다.

전체적으로 본 실험은 멀티모달 접근법이 단순히 텍스트의 한계를 보완하는 것을 넘어, 사용자 경험을 개선하고 검색 시스템의 견고성을 높이는 핵심 요소임을 입증하였다. 특히 사용자가 정확한 용어를 모르는 상황에서도 시각 정보를 통해 효과적인 정보 검색이 가능함을 보여주어, 실용적인 검색 시스템 구축에 중요한 시사점을 제공한다.

Table 2: Experiment Results(%)

	CR	**CP**	**F**	**H**
Q1	99.3	70.4	95.5	2.9
Q2	79.2	58	95	3.8
Q3	99.3	70.4	95.5	2.9
Q4	94.6	66.4	95.2	3.7
Q5	70.8	88.0	63.6	19.4
Q6	91.2	74.0	85.4	11.6
Q7	94.6	66.4	95.2	3.7

6 결론

본 연구는 MR 환경에서 사용자가 실제 작업을 수행할 때 실시간으로 도움을 받을 수 있는 멀티모달 RAG를 설계하였으며, 자동차 사용 설명서를 활용한 실험을 통해 이 기술의 실용성과 효과를 확인하였다. 시각 · 언어 · 공간 정보를 통합하는 멀티모달 인코딩과 실환경 데이터를 동적으로 참조하는 RAG를 결합함으로써, 기존 단일모달 또는 오프라인 접근법이 직면해 온 정보 단절과 문맥 불일치 문제를 효과적으로 완화할 수 있음을 보여주었다. 이를 통해 사용자는 MR 헤드셋을 착용한 채 차선 이탈 경고 관련 계기판 확인, 주유소 급유기 기호 확인, 환원제 보충 등 복잡한 절차적 작업을 자연스러운 음성 인터랙션만으로 수행할 수 있었으며, 실험 평가를 통해 작업의 정확성을 확인할 수 있었다.

학술적으로는 실시간 MR 환경에 멀티모달 RAG 아키텍처를 적용한 사례로서, 가상 · 현실 정보가 실시간으로 교차할 때 발생하는 동기화 지연, 문맥 불일치, 인터페이스 혼선 등 복합적인 문제를 포괄적으로 다루는 참조 모델을 제시한다. 또한 객체 인식, 음성 질의 등 이질적 입력을 하나의 파이프라인에서 통합 처리하는 방법론을 공개함으로써, 복잡한 현실 과업을 지원하는 AI-MR 시스템의 새로운 패러다임을 제안한다. 산업적으로는 장비 매뉴얼, 공정 절차, 교육 자료가 여전히 2D 문서나 동영상 형태에 머물러 있는 현장의 한계를 극복할 수 있는 차세대 지능형 증강현실 매뉴얼의 기반 기술을 제시했다는 점에서 의의가 크다. 본 프레임워크는 제조, 의료, 항공, 전문 교육 등 다양한 도메인으로 손쉽게 확장될 수 있는 범용 구조를 갖추고 있으며, 실험을 통해 도입 비용 대비 충분한 가치 창출 가능성을 확인하였다. 이는 AI-MR 융합 기술이 실제 산업 현장의 복잡한 문제 해결에 실질적으로 기여할 수 있음을 보여주는 사례라 할 수 있다.

그럼에도 불구하고 본 연구에는 몇 가지 기술적 · 실무적 한계가 존재한다. 첫째, 이미지 입력과 음성 데이터를 처리하여 답변을 얻기까지 총 3-4초의 지연이 발생해 빠른 반복 작업에서 체감 속도가 떨어질 수 있다. 둘째, 패스스루 카메라 기능의 해상도 한계로 작은 글자체 인식에 정확도 문제가 나타났다. 셋째, 현재 프로토타입은 음성 및 화면 캡처 중심으로 설계되어, 최신 MR 기기에 탑재된 고해상도 손 추적 기반 제스처 인식, 햅틱 피드백, 시선 추적, 생체 신호 등 풍부한 입력 모듈을 충분히 활용하지 못하고 있다. 넷째, 검색 시스템은 빠른 검색 속도를 위해 전체 문서 임베딩을 인메모리 데이터베이스에 구축하므로, 메모리 사용량에 부담을 주는 한계가 있다. 마지막으로, 이미지 캡셔닝 과정에서 동일한 객체가 다양한 형태로 표현되는 비일관성이 관찰되었는데, 이는 검색 성능의 안정성에 영향을 미칠 수 있다.

이러한 한계를 극복하기 위해 다음과 같은 해결책을 제시할 수 있다. 처리 지연 문제는 LLM, VLM 모델의 경량화와 성능 개선으로 응답 생성 시간이 점차 줄어들 것으로 예상된다. 패스쓰루 카메라의 해상도 문제는 향후 출시될 스마트 글래스나 MR 장비

에서 고해상도 카메라 모듈과 이미지 전처리 알고리즘이 개선될 것으로 기대된다. MR 입력 모듈 활용 문제는 고급 제스처 인식, 시선 추적, 햅틱 피드백을 포함하는 인터랙션 디자인 고도화를 통해 개선할 수 있다. 검색 기능의 한계는 이미지 쿼리를 통한 검색 기능을 추가하여 텍스트로 표현하기 어려운 정보에 대한 검색이 가능하도록 확장할 수 있을 것이다. 더불어 VLM과 LLM에 대한 체계적인 프롬프트 엔지니어링은 이미지 캡셔닝의 비일관성 문제를 해결하고 전반적인 성능 향상에 기여할 수 있으며, 이는 멀티모달 검색 시스템의 성능과 신뢰성을 결정짓는 핵심 요소로서 더욱 심층적인 연구가 필요하다.

아울러 스마트 공장, 병원 수술실, 도심 교통 제어, 군사 훈련 등 다양한 현장에 프레임워크를 적용해 도메인별 성능 저하 요인을 체계적으로 분석한다면, 멀티모달 RAG는 MR 기반 지능형 매뉴얼의 사실상 표준 플랫폼으로 자리매김하고, 인간과 AI가 물리적 공간에서 자연스럽게 협업하는 차세대 산업 생태계 구축의 핵심 기술 토대를 제공할 수 있을 것이다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이고(RS-2023-00251681), 정보통신기획평가원의 대학ICT연구센터사업(RS-2023-00259099)과 메타버스융합대학원(RS-2022-00156318)의 지원으로 수행되었음.

References

[1] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023. [Online]. Available: https://doi.org/10.1145/3571730

[3] H.-K. Hong and B. Leem, "Customized gpt-driven educational learning chatbot," *Journal of the Korea Contents Association*, vol. 25, no. 3, pp. 395–407, 2025.

[4] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in *Artificial Intelligence Applications and Innovations*, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds. Cham: Springer International Publishing, 2020, pp. 373–383.

[5] X. Zhang, J. Guo, S. Zhao, M. Fu, L. Duan, G.-H. Wang, Q.-G. Chen, Z. Xu, W. Luo, and K. Zhang, "Unified multimodal understanding and generation models: Advances, challenges, and opportunities," 2025. [Online]. Available: https://arxiv.org/abs/2505.02567

[6] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," 2024. [Online]. Available: https://arxiv.org/abs/2409.12191

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[8] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2312.10997

[9] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, and S. Tang, "Graph retrieval-augmented generation: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2408.08921

[10] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[11] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," 2023. [Online]. Available: https://arxiv.org/abs/2305.05665

[12] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen, "Murag: Multimodal retrieval-augmented generator for open question answering over images and text," 2022. [Online]. Available: https://arxiv.org/abs/2210.02928

[13] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," 2022. [Online]. Available: https://arxiv.org/abs/2204.14198

[14] A. Berni and Y. Borgianni, “Applications of virtual reality in engineering and product design: Why, what, how, when and where,” *Electronics*, vol. 9, no. 7, 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/7/1064

[15] A. Nagy, Y. Spyridis, and V. Argyriou, “Cross-format retrieval-augmented generation in xr with llms for context-aware maintenance assistance,” 2025. [Online]. Available: https://arxiv.org/abs/2502.15604

[16] D. Tomkou, G. Fatouros, A. Andreou, G. Makridis, F. Liarokapis, D. Dardanis, A. Kiourtis, J. Soldatos, and D. Kyriazis, “Bridging industrial expertise and xr with llm-powered conversational agents,” 2025. [Online]. Available: https://arxiv.org/abs/2504.05527

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[18] M. Faysse, H. Sibille, T. Wu, B. Omrani, G. Viaud, C. Hudelot, and P. Colombo, “Colpali: Efficient document retrieval with vision language models,” 2025. [Online]. Available: https://arxiv.org/abs/2407.01449

[19] D. Ru, L. Qiu, X. Hu, T. Zhang, P. Shi, S. Chang, C. Jiayang, C. Wang, S. Sun, H. Li, Z. Zhang, B. Wang, J. Jiang, T. He, Z. Wang, P. Liu, Y. Zhang, and Z. Zhang, “Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation,” 2024. [Online]. Available: https://arxiv.org/abs/2408.08067

〈저자소개〉

김한얼
- 2009 서울시립대학교 전자전기컴퓨터공학부 학사
- 2011 광주과학기술원 기전공학부 석사
- 2011 ~ 2012 서울대학교 자동화연구소 제어계측신기술연구센터 연구원
- 2013 ~ 2022 삼성메디슨 의료영상개발그룹 책임연구원
- 2022 ~ 현재 Unity Technologies Senior Software Engineer
- 관심분야: 인공지능, 확장현실, 접근성
- https://orcid.org/0009-0009-5222-4210

배종환
- 2016 경희대학교 건축학과 건축학사
- 2025 서강대학교 메타버스전문대학원 메타버스테크놀로지 전공 공학석사
- 관심분야: 인공지능, 메타버스, 확장현실, 건축/건설, 디지털트윈
- https://orcid.org/0009-0007-7520-2319

정원영
- 2022 경북대학교 경영학부 경영학사
- 2025 서강대학교 메타버스전문대학원 메타버스테크놀로지전공 공학석사
- 관심분야: 인공지능, 메타버스, 확장현실, 디지털 컨텐츠, 경영전략
- https://orcid.org/0009-0000-4669-1028

박상훈
- 1993 서강대학교 수학과 학사
- 1995 서강대학교 컴퓨터학과 석사
- 2000 서강대학교 컴퓨터학과 박사
- 2022 ~ 2005 대구가톨릭대학교 컴퓨터정보통신공학부조교수
- 2001 University of California, Davis 방문 연구원
- 2005 ~ 2023 동국대학교 멀티미디어학과 교수
- 2023 ~ 현재 서강대학교 메타버스전문대학원 교수
- 관심분야 : 실시간 렌더링, 사실적 렌더링, 과학적 가시화, 고성능 컴퓨팅 등
- https://orcid.org/0000-0002-0015-8305