JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 2, pp. 473-484, Feb. 2026

Check for updates

멀티모달 검색-증강 생성과 대규모 언어 모델을 활용한 맞춤형 미술관 도슨트 시스템

강 채 원[1] · 김 예 빈[1] · 유 채 민[1] · 윤 소 은[1] · 이 윤 서[1] · 유 견 아[2*]
[1]덕성여자대학교 컴퓨터공학부 학사과정
[2]덕성여자대학교 컴퓨터공학부 교수

Personalized Museum Docent System Based on Multimodal Retrieval-Augmented Generation and Large Language Models

Chaewon Kang[1] · Yebeen Kim[1] · Chaemin Yu[1] · Soeun Yoon[1] · Yunseo Lee[1] · Kyeonah Yu[2*]

[1]Undergraduate Program, Department of Computer Engineering, Duksung Women's University, Seoul 01369, Korea
[2]Professor, Department of Computer Engineering, Duksung Women's University, Seoul 01369, Korea

[요 약]

본 연구는 시선 추적, 작품 영상, 음성 질의 등을 통합하여 개인화고 맥락 인지적인 해설을 생성하는 멀티모달 RAG 기반 미술관 도슨트 시스템을 제안한다. 제안된 시스템은 YOLOv8 객체 탐지, CLIP 임베딩, FAISS 기반 검색, 그리고 LLM-RAG 융합 구조를 결합하여 관람객의 주목 대상을 식별하고 관심에 부합하는 적응형 해설을 제공한다. 기존의 오디오 가이드나 모바일 앱과 달리, 본 시스템은 관람객의 시각적 주의와 음성 질의에 동적으로 반응하며, STT와 TTS를 통한 자연스럽고 몰입감 있는 상호작용을 구현한다. 또한 시선 및 질의 데이터를 수집·분석하여 전시 기획 및 운영을 위한 데이터 기반 인사이트를 제공한다. 오픈소스 기반의 저비용 구조와 클라우드 연동성을 바탕으로, 실제 미술관 환경에서의 실용성과 개인화 문화 콘텐츠 경험의 새로운 패러다임을 제시한다.

[Abstract]

This study proposed a multimodal retrieval-augmented generation (RAG)-based museum docent system that integrated gaze tracking, artwork imaging, and voice queries to generate personalized and context-aware explanations. The proposed system combined YOLOv8-based object detection, CLIP embeddings, FAISS-based similarity retrieval, and LLM–RAG fusion to identify the focus of visitors and deliver adaptive narration aligned with their interests. Unlike conventional audio guides or mobile applications, the proposed approach dynamically responded to the visual attention and spoken inquiries of visitors, enabling interactive and immersive experiences through seamless STT and TTS integration. Furthermore, the system collected gaze and query data to support data-driven insights for exhibition planning and management. Built on an open-source low-cost architecture with cloud connectivity, it demonstrated practical applicability in real museum environments and introduced a new paradigm for personalized cultural content experiences.

색인어 : 인공지능 도슨트, 시선 추적, 대규모 언어 모델, 멀티모달 상호작용, 검색-증강 생성
Keyword : AI Docent, Gaze Tracking, Large Language Model, Multimodal Interaction, Retrieval-Augmented Generation

http://dx.doi.org/10.9728/dcs.2026.27.2.473

Received 24 December 2025; **Revised** 15 January 2026
Accepted 09 February 2026

***Corresponding Author; Kyeonah Yu**

Tel: +82-2-901-8346
E-mail: kyeonah@duksung.ac.kr

Ⅰ. 서 론

최근 전 세계적으로 문화예술 공간의 디지털 전환이 가속화되면서, 전시 해설(도슨트) 서비스 또한 기존의 오디오 가이드나 모바일 애플리케이션 기반을 넘어, 인공지능 기술과 융합된 차세대 인터페이스로 발전하고 있다[1]. 특히 관람객의 몰입적 경험을 중시하는 현대 전시 환경에서는, 단순한 일방향 정보 제공을 넘어서 관람객 개개인의 관심과 요구를 실시간으로 반영하는 맞춤형 해설 서비스가 요구되고 있다. 그러나 기존의 인공지능 기반 전시 안내 시스템은 특정 키워드 기반 질의응답에 국한되거나, 사전에 정의된 설명을 단순히 출력하는 수준에 머무르는 경우가 많아, 관람객의 실제 시선·행동·질의 맥락을 반영하는 몰입형 해설을 제공하는 데 한계가 존재한다[2]. 또한 대규모 언어 모델(Large Language Model, LLM)을 단독으로 활용하는 경우, 학습 데이터에 의존한 환각(hallucination) 문제로 인해 전시 작품 설명의 정확성과 신뢰성을 보장하기 어렵다[3].

본 연구에서는 이러한 한계를 극복하기 위해, 멀티모달 검색-증강 생성(Retrieval-Augmented Generation, RAG) 기반 미술관 도슨트 시스템을 제안한다. 제안된 시스템은 LLM의 자연스러운 언어 생성 능력에, 외부 지식 검색을 결합하여 신뢰도 높은 설명을 제공한다. 더 나아가 텍스트에 한정된 기존 RAG와 달리, 시선 추적(gaze tracking), 음성 질의(STT), 작품 이미지(객체 인식 및 CLIP 임베딩) 등 다양한 입력 모달리티를 통합함으로써 사용자의 실제 행동과 맥락에 최적화된 해설을 실시간으로 생성한다. YOLOv8 기반 객체 탐지와 CLIP-FAISS 유사도 검색을 통해 작품 속 세부 객체를 인식하고, 관람객의 시선 데이터와 결합함으로써 “어떤 작품을 보는가”를 넘어 “작품의 어느 부분에 주목하는가”까지 반영한 맞춤형 해설을 구현한다.

이와 같은 접근은 기존의 오디오 가이드나 앱 기반 안내 서비스와 근본적으로 차별화되며, 관람객은 별도의 화면 조작 없이 시선과 음성만으로 자연스러운 상호작용을 경험할 수 있다. 나아가 본 연구는 단순히 전시 안내를 넘어, 관람객의 시선 데이터와 질문 로그를 분석하여 전시 기획 및 운영에도 활용 가능성을 제시한다. 따라서 본 연구는 문화예술 감상 방식의 패러다임 전환을 이끌 뿐만 아니라, 멀티모달 RAG라는 새로운 기술적 프레임워크를 통해 차세대 스마트 전시 해설 시스템의 학문적·실용적 기초를 제공할 것으로 기대된다.

본 논문은 다음과 같이 구성된다. 2장에서는 관련 연구를 정리하고 기존 도슨트 시스템과의 차별점을 분석한다. 3장에서는 제안하는 멀티모달 RAG 기반 미술관 도슨트 시스템의 전체 아키텍처와 주요 모듈을 설명한다. 4장에서는 시스템의 구현 환경과 세부 구현 방법, 그리고 성능 평가 결과를 제시한다. 마지막으로 5장에서는 결론과 향후 연구 방향을 논의한다.

Ⅱ. 관련 연구

박물관과 미술관에서의 해설 서비스는 오랜 기간 오디오 가이드나 모바일 애플리케이션 기반 안내 시스템이 중심을 이루어왔다. 그러나 이러한 기존 시스템은 사용자의 실제 행동이나 관심 지점을 반영하기 어렵고, 일방적 정보 전달에 그친다는 한계가 있었다. 최근에는 인공지능 기술의 발전과 대규모 언어 모델(LLM), 문서 검색 기반 생성(RAG) 기법의 보급으로 인해, 전시 환경에서도 더욱 지능적이고 상호작용적인 도슨트 시스템이 시도되고 있다.

LLM을 전시 안내/추천 시스템으로 활용해 관람 맥락과 선호를 반영하는 접근을 시도하여 지식 응답의 품질과 상황 적합성을 높이고자 하는 시도들이 있다[1],[4]. [1]에서는 LLM을 추천 시스템으로 활용해 관람객 맥락과 선호를 반영한 맞춤형 추천을 시도하였으며, 맥락 인식(Context-aware) 모델의 가능성을 제시하였으며 [4]에서는 LLM을 이용해 사용자의 선호·질의·맥락 데이터를 실시간 반영하여 전시 경로를 추천하고 작품 설명을 생성하는 개인화된 대화형 투어 가이드를 구현했다. 대규모 전시 설명 데이터와 질문-응답 데이터를 구축하여 RAG 기반 소형 LLM(sLLM)을 활용한 지능형 도슨트 시스템이 제안된 바 있다[3]. 이 연구는 실제 박물관 환경에 적용되어 사용자 만족도를 평가하였으며, 생성형 응답과 신뢰성 있는 해설 제공 가능성을 입증하였다. 그러나 이 연구는 주로 텍스트 및 음성 질의응답에 집중되어 있으며, 사용자의 시선 추적이나 시각적 관심 영역을 반영하는 기능은 포함하지 않았다. 미술작품 질의응답을 위해 RAG 방식의 적용하기 위해 대형 컨텍스트 입력 방식을 비교하여, 예술 분야의 멀티모달 데이터(이미지 + 텍스트) 처리 성능을 평가한 연구도 있다[5]. 연구 결과, RAG 기법이 검색 정확도와 응답 일관성 측면에서 우수한 성능을 보였고, 데이터 규모나 문맥 길이에 따라 대형 컨텍스트 입력이 유리한 경우도 있어 예술 QA 환경에서는 두 가지 방식의 하이브리드 구조를 제안하였다. 챗봇 시스템을 이용하여 관람객과 자연스럽게 질의응답하는 도슨트 구현을 제안한 연구도 있다[6],[7]. [6]에서는 박물관 방문객의 자연어 질의를 바탕으로 지식베이스 기반 챗봇 시스템을 구현하여 사용자의 질의 흐름을 분석해 적합한 미술품 정보를 검색하고 응답하는 인터랙티브 안내 환경을 제안하였으면 [7]에서는 관람객의 자연어 질의에 기반하여 미술관 안내를 수행하는 AI 챗봇 시스템을 제안하여 다양한 질의패턴을 지원하고 지식베이스 검색을 통해 응답을 제공하는 방식을 적용했다. 그러나 두 가지 연구 모두 시선 추적이나 객체 인식까지 통합한 실시간 멀티모달 입력을 활용하지는 않았다.

전시장 내의 위치 인식 및 보조 안내를 위한 기술들도 발전해왔다. 시각장애인을 위한 모바일 기반 도슨트 시스템을 개발하여, BLE 비콘과 객체 탐지 기반 위치 인식을 활용하여 시각장애인을 위한 내비게이션 및 안내 서비스가 구현되어

사용자 이동이나 위치 변화에 따라 안내 내용이 동적으로 조정 가능해졌으나, 관람객의 시선이나 작품 내부의 세부 영역까지 반영하지는 못했다[8]. 또한 증강현실 기반 도슨트 시스템(CARDS)이 제안되어, 전시물 주변에 AR 핀을 배치하고 사용자의 위치와 방향에 따라 증강된 시각 정보를 제공하였다[9]. 이러한 연구는 공간적 맥락에 맞춘 시각적 보조 기능을 강조하였으나, 설명 생성보다는 시각적 인터페이스 제공에 초점을 맞추고 있으며, LLM이나 RAG 기반의 동적 설명 생성 기능은 포함되지 않았다.

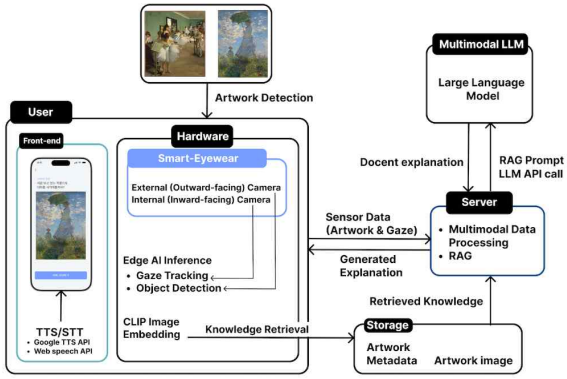
관람객의 시선 추적 기술을 전시 안내에 적용한 연구도 존재한다. [10]에서는 미술관 모바일 앱 사용 과정에서 아동 관람객의 시선 움직임 및 행동을 분석하고, 시선 데이터를 기반으로 사용자 중심 모바일 안내 앱 설계 방향을 제시하였다. 한편, 시선 기반 상호작용에 대한 체계적 검토 연구는 박물관·전시 맥락에서 방문객이 무엇을 보는가(what) 뿐 아니라, 어디에 주목하는가(where)를 파악해 인터페이스 설계 고도화의 필요성을 제기했다[2]. 더 나아가 모바일 시선추적을 통해 관람객의 전시 참여도와 학습 경험을 정량 분석한 연구는, 시선 데이터가 관람 경험 평가 및 맞춤형 안내 설계의 기초 자료로 활용될 가능성을 제시하였다[11]. 이러한 연구들은 주로 시선 데이터의 분석 및 활용 가능성에 초점을 맞추고 있으며 실제 시선 추정 방법이 다루어진 것은 컴퓨터 비전 분야에서이다[12]. 이 연구에서는 합성기반 학습(Learning by synthesis)의 시선 추정 방법을 제안하여 별도의 장비나 머리 고정 없이도 시선 추정이 가능함을 보였으나 실시간 도메인의 적용을 제한적이었다. 이처럼, 시선 기반 관심 영역 분석은 가능해졌지만, 객체 인식 및 해설 생성까지 이어진 종합 시스템으로 발전된 연구는 아직 제한적이다.

기존 연구들을 종합하면, 최근 도슨트 시스템은 LLM과 RAG를 활용한 지식 응답 정확성 개선, 위치 및 이동 기반 안내, AR 인터페이스, 챗봇 기반 대화형 안내 등으로 발전해 왔다. 그러나 대부분의 연구는 관람객의 시선과 주목 대상을 실시간으로 추적하여 해설에 반영하는 기능이 미흡하며, 경량 하드웨어 환경에서의 실현 가능성 또한 충분히 검토되지 않았다. 따라서 본 연구가 제안하는 멀티모달 RAG 기반 도슨트 시스템은 시선 추적, 객체 인식, 음성 질의 등 다양한 입력 모달리티를 통합하여 기존 연구의 한계를 보완하고, 관람객 맞춤형 몰입형 경험을 제공한다는 점에서 의의가 있다.

Ⅲ. 제안된 멀티모달 RAG 기반 미술관 도슨트 시스템

본 연구에서는 관람객의 시선 정보와 음성 질의, 그리고 작품 객체 인식 결과를 통합하여, 실시간으로 개인화된 해설을 제공하는 멀티모달 RAG 기반 미술관 도슨트 시스템을 제안한다.

3-1 시스템 개요



*This figure includes a screenshot of the application in Korean.

그림 1. 전체 시스템 아키텍처

Fig. 1. Overall system architecture

그림 1은 본 연구에서 제안하는 멀티모달 RAG 기반 미술관 도슨트 시스템의 전체 시스템 개요를 보여준다. 제안 시스템은 스마트 아이웨어가 위치한 엣지 하드웨어 계층, 관람 화면 및 음성 인터페이스를 제공하는 프론트엔드 계층, 데이터 처리를 담당하는 백엔드 계층, 멀티모달 대규모 언어 모델(LLM)을 활용한 AI 서버, 그리고 작품 관련 데이터를 저장·관리하는 스토리지 계층으로 구성된 분산형 아키텍처를 따른다. 전체 시스템은 관람객의 시선 및 음성 입력을 실시간으로 수집하고, 이를 분석하여 생성된 해설을 다시 사용자에게 제공하는 흐름으로 설계되어 있다.

관람객이 스마트 아이웨어를 착용하고 전시 작품을 감상하면, 외부 환경을 촬영하는 카메라와 시선 추적 센서, 마이크 모듈을 통해 시선 정보, 영상 데이터, 음성 질의가 실시간으로 수집된다. 엣지 디바이스에서는 수집된 영상과 시선 정보를 기반으로 사용자가 주목하고 있는 작품 또는 작품 내 영역을 식별하며, 해당 결과는 이후 해설 생성을 위한 핵심 입력으로 활용된다.

식별된 작품 정보와 사용자의 시선 방향, 질의 맥락은 백엔드 서버로 전달되며, 백엔드 서버는 검색 증강 생성(RAG) 구조를 기반으로 스토리지에 저장된 작품 메타데이터 및 설명 정보를 결합하여 멀티모달 LLM을 통한 도슨트 해설 생성을 수행한다. 생성된 해설은 관람 맥락을 반영한 자연어 텍스트 형태로 출력되며, 프론트엔드 인터페이스를 통해 사용자에게 전달된다.

스토리지 계층은 작품 이미지, 메타데이터, 설명 텍스트 등 시스템 운영에 필요한 정적·동적 데이터를 관리하며, 백엔드 및 AI 서버의 요청에 따라 효율적인 검색과 갱신을 지원한다. 이러한 엣지–서버–AI 계층 간 역할 분리는 실시간 상호작용을 유지하면서도 시스템 확장성과 유연성을 확보할 수 있도록 한다.

3-2 주요 모듈 설계 및 기능

본 절에서는 제안하는 멀티모달 RAG 기반 미술관 도슨트 시스템의 기능적 구성 요소를 네 개의 모듈, 즉 시선 기반 객체 선택 모듈, 멀티모달 검색 및 지식 결합 모듈, LLM 기반 해설 생성 모듈, 실시간 음성 및 인터페이스 모듈로 구분하여 설명한다.

앞서 제시한 시스템 아키텍처(하드웨어, 백엔드 서버, AI 서버, 스토리지)는 이러한 기능 모듈들이 실제로 실행·배치되는 물리적 및 소프트웨어적 기반 구조를 나타내며, 본 절의 모듈 구분은 시스템의 논리적 처리 흐름과 기능적 역할에 초점을 둔다.

특히, 본 시스템은 시각 인지 정확도를 극대화하기 위해 범용 YOLOv8 모델을 베이스로 전이 학습을 수행한 두 종류의 커스텀 모델을 운용한다. 시선 방향 추적을 위해 안구 및 동공 데이터셋을 학습시킨 모델을 Eyedia-Gaze로, 복잡한 전시장 환경 내에서 작품과 비작품(인물, 가구, 배경 등)을 구분하여 예술 작품만을 정밀하게 탐지하도록 학습 시킨 모델을 Eyedia-Art로 명명한다. 표 1은 본 연구에서 제안하는 시선 추적 및 작품 인식을 위해 설계된 두 개의 커스텀 비전 모델(Eyedia-Gaze 및 Eyedia-Art)의 주요 특성을 요약한 것이다.

표 1. 제안하는 시선·작품 인지용 커스텀 비전 모델
Table 1. Proposed custom vision models for gaze tracking and artwork recognition

Model name	Eyedia-Gaze	Eyedia-Art
Backbone	YOLOv8	YOLOv8
Primary Purpose	Eye-gaze estimation	Artwork detection
Input	Eye images	Artwork images
Output	Gaze Quadrant	Artwork bounding boxes

구체적으로, 시선 기반 객체 선택 모듈은 스마트 아이웨어를 포함한 하드웨어 계층에서 구현되며, 멀티모달 검색 및 지식 결합 모듈은 백엔드 서버와 스토리지 계층을 중심으로 동작한다. 또한 LLM 기반 해설 생성 모듈은 AI 서버에서 수행되며, 실시간 음성 및 인터페이스 모듈은 사용자 단말과 백엔드 서버 간의 상호작용을 담당한다.

이와 같이 각 모듈은 기능적으로 분리되어 설계되었으나, 백엔드 서버와 하드웨어를 중심으로 긴밀하게 연동되어 단일한 도슨트 서비스 파이프라인을 구성한다.

1) 시선 기반 객체 선택 모듈

본 모듈은 관람객의 시선과 음성 입력을 기반으로, 작품 내 주목 대상 객체를 식별하기 위한 핵심 기능을 담당한다. 스마트 아이웨어에 장착된 전방 카메라는 관람객의 시야 전반을 촬영하며, 시선 추적 카메라는 사용자의 눈동자 움직임을 촬영한다(그림 2). 이후 시선 추적 전용 모델인 Eyedia-Gaze와 객체 탐지 전용 모델인 Eyedia-Art를 병렬로 운용하여 입력 프레임 내의 핵심 정보를 식별한다. 특히 Eyedia-Art는 복잡한 전시장 환경 내에서 관람객, 시설물 등 작품이 아닌 물체가 혼재되어 있어도 이를 배제하고 오직 예술 작품만을 정밀하게 구분하여 인식하도록 학습되었다. 이를 통해 탐지된 작품의 위치 정보와 시선 방향을 결합하여 관람객에 실제로 주목하는 영역을 결정하며, 이를 개인화된 설명 생성을 위한 입력으로 활용한다.

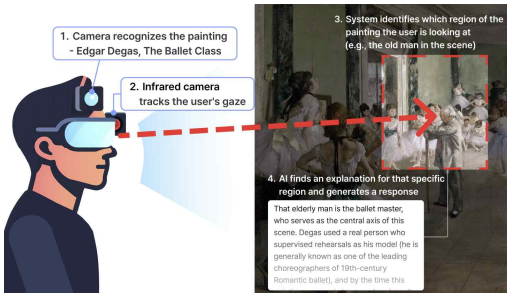


그림 2. 시선 기반 객체 선택
Fig. 2. Gaze-based object selection

시선 추적은 단순히 동공 중심 좌표를 계산하는 방식이 아니라, YOLOv8 기반의 커스텀 객체 검출 모델인 Eyedia-Gaze를 학습하여 눈 영역과 동공 위치를 안정적으로 인식하는 방식을 적용하였다. 전시·체험형 서비스 환경에서는 조명 변화나 사용자의 이동이 잦기 때문에, YOLO의 실시간 검출 성능과 강인한 특징 분리 능력을 활용해 다양한 각도에서 촬영된 눈 영상을 처리할 수 있도록 최적화하였다.

Eyedia-Gaze 모델은 좌·우 눈, 동공, 눈꼬리 등의 특징 포인트를 프레임 단위로 추출하며, 이를 기반으로 사용자의 머리 움직임에 따라 변하는 상대적 위치 변화를 보정하여 시선 벡터를 산출한다. 이 방식은 별도의 복잡한 캘리브레이션 없이도 높은 정확도의 시선 추정이 가능하며, 기존 고정형 시선 추적 시스템 대비 휴대형·착용형 서비스 환경에서 훨씬 유연하게 활용될 수 있다는 장점이 있다.

예술 작품 인식 및 선별에는 Eyedia-Art 모델을 사용하였으며, 이는 Ultralytics YOLOv8 모델을 기반으로 한 전이학습 방식을 적용한 것이다. Eyedia-Art는 단순히 다양한 사물을 인식하는 범용 모델과 달리, 전시장 내의 인물이나 주변 집기 등 비작품 요소가 프레임에 들어오더라도 이를 노이즈로 간주하고 전시 작품만을 선택적으로 탐지하도록 최적화되었다. Roboflow에서 라벨링된 미술 작품 데이터셋을 활용하여 모델을 학습하였고, 데이터 증강을 통해 작은 객체와 복잡한 배경에서도 인식 성능을 향상시켰다. 학습 결과, Eyedia-Art 모델은 정확도 91.2%, 리콜 81.4%, mAP@50 90.2%의 성능을 달성하였다. 이는 전시 환경의 다양한 조명과 시점 변화에서도 안정적인 탐지가 가능함을 의미한다.

2) 지식베이스 기반 멀티모달 검색 모듈

본 시스템의 핵심 구성요소는 작품 설명 지식베이스에 있다. 지식베이스는 Metropolitan Museum of Art Open Access API를 활용하여 구축하였으며, 미술관 소장품의 기본 메타데이터(작품명, 작가, 제작 시기 등)뿐 아니라 작품 해설문, 평론, 예술사적 맥락 등 다양한 텍스트 자료를 포함한다. 이러한 데이터는 관람객 맞춤형 해설을 위한 핵심 기반 자료로 활용된다.

먼저, 예술 작품의 원본 이미지를 입력하면 YOLOv8 객체 탐지 모델이 작품 내 주요 객체를 바운딩 박스 단위로 식별한다. 이후 OpenCV를 사용하여 각 객체를 crop 이미지로 자동 추출 및 저장하며, 파일명과 폴더 구조를 통해 원본 이미지 ID와 연동되도록 관리한다. 탐지 결과로 생성된 라벨링 ID는 사람이 이해할 수 있는 의미적 라벨로 변환되어, 대응되는 작품 설명과 자동으로 매핑된다.

다음 단계에서는 CLIP 모델을 통해 crop 이미지와 관련 텍스트 자료를 동일한 임베딩 공간에 매핑하고, FAISS 기반 벡터 검색을 통해 입력 이미지와 가장 유사한 객체를 고속으로 탐색한다. 이 과정을 통해 지식베이스 내에서 해당 작품 또는 세부 객체와 가장 관련성이 높은 설명을 실시간으로 검색할 수 있으며, 선택된 텍스트 자료는 RAG 구조를 통해 LLM으로 전달되어 자연스럽고 맥락적인 구어체 해설을 생성하는 데 활용된다.

본 모듈의 학술적 기여는 두 가지 측면에서 요약할 수 있다. 첫째, 지식베이스 기반의 멀티모달 검색 구조를 통해 관람객의 시선·이미지 정보를 RAG 프레임워크에 통합함으로써 실시간 맞춤형 정보 검색과 해설 생성을 동시에 달성하였다. 둘째, 기존의 정형화된 오디오 가이드나 사전 정의된 텍스트 설명과 달리, 관람객의 실제 관심 영역과 시선 데이터를 반영하여 개인화된 몰입형 해설 경험을 제공할 수 있다는 점에서 차별성을 가진다.

3) LLM 기반 해설 생성 모듈

지식베이스에서 검색된 관련 자료는 GPT 계열의 대형 언어모델에 입력되어 자연스러운 한국어 구어체 해설로 변환된다. 생성 과정에서는 사용자의 음성 질의 내용과 시선 정보가 함께 반영되어, 관람객이 실제로 주목하거나 질문한 대상에 대해 맥락적으로 적합한 설명을 제공한다. RAG 구조를 활용함으로써, LLM이 생성 과정에서 발생할 수 있는 환각을 최소화하고, 신뢰성 높은 해설을 제공할 수 있다. 이러한 접근은 사전 정의된 오디오 가이드 콘텐츠에 의존하지 않고, 관람객의 관심과 반응에 따라 실시간으로 해설을 생성한다는 점에서 차별화된 몰입형 해설 경험을 제공한다.

4) 실시간 음성 및 인터페이스 모듈

본 모듈은 관람객의 음성 질의 인식과 시스템 해설의 음성 출력을 담당한다. STT와 TTS 기술을 결합하여 관람객이 별도의 입력 장치 없이 음성으로 질문하고, AI가 생성한 해설을 자연스러운 음성으로 들을 수 있도록 한다.

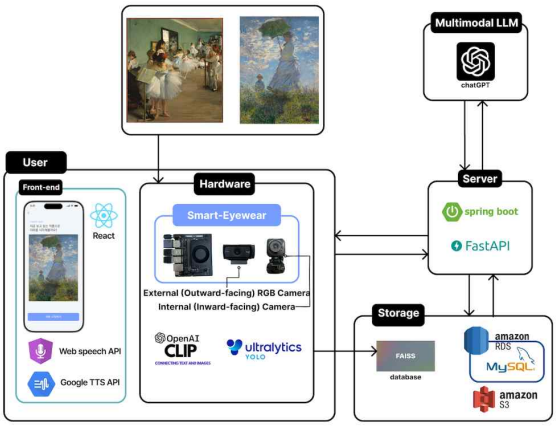
브라우저 환경에서는 Web Speech API를 활용하여 음성 입력을 수집한다. React 환경에서 구현된 useStt 혹은 사용자의 음성 지원 여부를 점검하고, 한국어 인식 모드와 중간 결과 출력을 활성화한 STT 인스턴스를 생성한다. 음성 인식 과정에서 브라우저의 onresult 이벤트는 확정된 문장과 진행 중인 문장을 분리하여 전달하며, 확정된 문장은 FastAPI 기반 서버로 전송되어 LLM 질의응답 프로세스로 연결된다. 응답 생성 후에는 TTS 모듈이 LLM의 텍스트 출력을 자연스러운 음성으로 변환하여, 웹 애플리케이션에 접속한 사용자의 오디오 출력 장치를 통해 즉시 재생된다.

이러한 음성 상호작용 구조를 통해 관람객은 시선 추적 정보와 결합된 질의응답을 실시간으로 경험할 수 있으며, 화면 조작 없이 자연스러운 대화형 해설 서비스를 이용할 수 있다. 하드웨어와 FastAPI–Spring Boot–React 기반의 웹 인터페이스는 실시간 비동기 통신 구조로 설계되어 백엔드와 프론트엔드 간 데이터 지연(latency)을 최소화하였다. 또한 관람객의 시선·음성·질의 로그 데이터를 수집·저장함으로써, 전시 운영 및 향후 관람 행동 분석에 활용할 수 있다. 이와 같은 경량·저비용 구조는 고가의 전용 장비 없이도 실제 전시 환경에서 실시간 멀티모달 인터랙션 구현이 가능하며, 향후 다양한 전시 공간으로 확장할 수 있는 실용적 기반을 제공한다.

이와 같은 단계적 통합 구조는 기존의 정적 오디오 가이드나 모바일 앱 안내와 달리, 관람객의 행동·시선·질의를 중심으로 작동하는 실시간 상호 작용형 도슨트 경험을 제공한다. 또한 각 모듈이 API 기반으로 독립적으로 구성되어 있어, 다른 전시 환경이나 경량 하드웨어로의 이식 및 확장이 용이한 구조적 유연성을 갖춘다.

Ⅳ. 시스템 구현 및 결과

그림 3은 제안된 멀티모달 RAG 기반 미술관 도슨트 시스템의 전체 구현 구조를 보여준다. 본 시스템은 하드웨어(Smart Eyewear), 백엔드 서버(Spring Boot 기반 API), 프론트엔드 웹·모바일 인터페이스(React), 그리고 AI 해설 엔진(LLM·RAG)으로 구성된다. 본 시스템은 Spring Boot와 React, Java를 이용하여 웹 애플리케이션 형태로 구현되었으며, 웹 브라우저 환경에서 직접 사용할 수 있도록 설계하였다. 백엔드 서버는 AWS EC2 환경에서 구동되고, 데이터베이스는 MySQL을 기반으로 구축하였다. 작품 정보는 Metropolitan Museum of Art Open Access API를 통해 수집하였으며, 각 작품의 주요 메타데이터와 설명 정보를 추가하여 확장된 데이터셋을 구성하였다. 이 데이터는 FAISS 데이터베이스에 벡터 형태로 저장되어 LLM에서 실시간 검색 및 증강(RAG)에 활용된다. 해설 생성을 위한 대형 언어 모델은 OpenAI API를 통해 연동하였다.



*This figure includes a screenshot of the application in Korean.
그림 3. 시스템 구현을 위한 소프트웨어 아키텍처
Fig. 3. Software architecture for system implementation

4-1 스마트 아이웨어 제작

본 연구에서는 Jetson Orin Nano 8GB 모듈을 중심으로 동작하는 스마트 아이웨어 프로토타입을 제작하였다. 시스템 전면에는 전시장 환경과 예술 작품 전체를 촬영하기 위한 Camera Module 3을 장착하고, 관람객의 동공 움직임을 정밀하게 추적하기 위해 Camera Module 3을 추가로 배치하였다. 사용자 입력은 음성 기반 상호작용 대신 소형 리모컨을 활용한 직관적 조작 방식으로 구성하여, 전시장처럼 소음이나 주변 환경의 영향을 받기 쉬운 공간에서도 안정적으로 기능을 수행할 수 있도록 설계하였다. 또한 전체 하드웨어는 경량 폼팩터 중심으로 설계해 장시간 착용 시에도 피로감을 최소화하였다(그림 4).

시선 추적은 전용 모델인 Eyedia-Gaze를 기반으로 하며, GPU 가속을 최적화하였다. 객체 탐지는 Eyedia-Art 모델을 사용하였으며, 실시간성을 확보하기 위해 Ultralytics YOLOv8 구조를 베이스로 학습하여 초당 30프레임 이상의 실시간 탐지 속도를 확보하였으며, PyCUDA를 이용해 탐지 결과를 GPU 메모리 상에서 직접 처리하여 데이터 이동 지연을 최소화하였다. 또한 CLIP 모델을 통해 이미지·텍스트 임베딩을 생성하고, FAISS를 이용해 Jetson 메모리상에 벡터 인덱스를 구축하여 밀리초(ms) 단위의 고속 유사도 검색을 구현하였다.

Eyedia-Art 객체 탐지 모델은 Ultralytics YOLOv8n 구조를 기반으로, 사전 학습된 가중치를 활용한 전이학습(transfer learning) 방식으로 학습되었다. 학습 데이터셋은 Roboflow 플랫폼에서 제공되는 전시 이미지 기반 객체 탐지 데이터로 구성되었으며, 미술작품 영역을 바운딩 박스로 주석화한 이미지를 학습·검증·테스트셋으로 분할하여 활용하였다. 다양한 전시 환경에서의 일반화 성능을 확보하기 위해 회전 및 스케일 변환 등의 데이터 증강 기법을 적용하였다.

객체 탐지 성능 평가는 검증 데이터셋(141장, 334개 객체)을 기준으로 Precision, Recall, mAP@50, mAP@50-95 지표를 사용하여 수행되었으며, 실험 결과 Precision 88.5%, Recall 82.7%, mAP@50 88.4%, mAP@50-95 68.4%의 성능을 보였다. 이는 복잡한 전시 이미지 환경에서도 예술 작품 객체를 안정적으로 탐지할 수 있음을 의미한다.

Eyedia-Gaze 시선 추정 모델은 Ultralytics YOLOv8n-classification 구조를 기반으로 하며, ImageNet으로 사전 학습된 가중치를 활용한 전이학습 방식으로 학습되었다. 입력 데이터는 스마트 아이웨어 환경에서 직접 수집한 관람객의 눈 이미지로 구성되었으며, 기존 공개 시선 데이터셋은 카메라 위치 및 촬영 조건 차이로 인해 적용이 어려워 자체 데이터셋을 구축하였다.

총 141장의 눈 이미지로 구성된 데이터셋은 사분면 기준으로 비교적 균형 있게 분포되어 있으며(Q1-Q4), 데이터 수의 한계를 보완하기 위해 이미지 증강 기법을 적용하였다. 전체 데이터의 80%를 학습, 20%를 검증 세트로 분할하여 학습을 수행하였다. 모델은 각 사분면에 대한 확률 분포를 출력하며, 가장 높은 확률을 갖는 클래스를 시선 방향으로 판단한다.

Eyedia-Gaze 모델은 Jetson Orin Nano 8GB 환경에서 GPU 가속을 통해 실시간으로 동작하도록 최적화되었으며, Eyedia-Art 객체 탐지 결과와 결합되어 사용자의 시선이 가리키는 작품 영역을 실시간으로 판별하는 전처리 모듈로 활용된다.

모든 구성요소는 와이파이 기반 네트워크를 통해 백엔드 서버와 연결되며, 영상·시선·음성 데이터를 실시간으로 전송한다. 이러한 하드웨어 설계를 통해 본 연구의 스마트 아이웨어는 시선 추적, 음성 인식, 객체 탐지 입력을 통합적으로 수집하는 인터랙티브 플랫폼으로 구현되었으며, 후속의 RAG 기반 해설 생성 시스템과 연동되는 실시간 데이터 입력 장치로 활용된다.

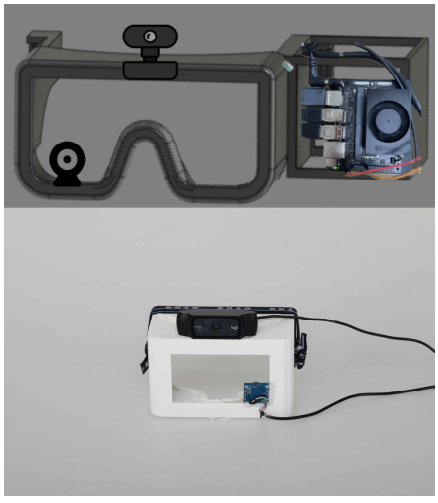


그림 4. 스마트 아이웨어 프로토타입과 완성 모듈
Fig. 4. Prototype and final module of the smart eyewear

4-2 실시간 데이터 수집 및 전송

스마트 아이웨어에서 수집된 시선·영상 정보는 Jetson Orin Nano에서 실시간으로 처리되며, 최종 분석 결과만을 Wi-Fi 네트워크를 통해 AI 백엔드 서버로 전송한다. 두 대의 카메라에서 입력되는 영상은 각각 사용자의 눈 영역과 전방 환경을 촬영하며, Jetson 내부에서 커스텀 모델인 Eyedia-Gaze와 Clip모델을 통해 시선 방향과 작품을 직접 판별한다.

이 과정에서 Eydia-Gaze 모델이 눈·시선 특징 검출 모델이 프레임 단위로 사용자의 시선 방향을 추정한다. 전방 카메라에서는 Eyedia-Art가 작품 또는 환경 요소를 동시에 인식하여 두 카메라 분석 결과를 조합함으로써 사용자가 현재 어떤 작품을 바라보고 있는지를 실시간으로 도출한다.

아이웨어에서는 이러한 연산 후 시선 방향, 응시 대상 ID, Confidence Score 등 핵심 결과값만을 JSON 형태로 백엔드 서버에 전송한다. 서버는 이를 비동기적으로 수신하여 기록·분석 모듈로 전달한다.

실험 결과, 평균 프레임 전송 속도는 약 25–30fps, 네트워크 지연은 약 180~220ms 수준으로 측정되었다. 이는 전시 환경에서 관람객의 시선 이동 및 음성 질의에 대해 거의 실시간에 가까운 응답이 가능함을 의미한다. 또한, ffmpeg 스트림은 네트워크 대역폭 변동에도 안정적으로 유지되며, 프레임 손실률은 1% 이하로 관찰되었다. FastAPI 서버는 Python 기반의 비동기(Asynchronous) 구조로 설계되어, 입력 데이터의 병렬 처리 및 모델 추론 병목 현상을 최소화하였다.

데이터 흐름은 “스마트 안경(입력) → 와이파이 네트워크(전송) → FastAPI 서버(처리)”의 순서로 진행되며, 처리된 결과는 Spring Boot 서버를 거쳐 React 프론트엔드로 전달된다. 이러한 실시간 전송 구조를 통해, 관람객이 시선을 이동하거나 질문을 던질 때 즉시 객체 인식–설명 생성–음성 응답이 순환되며, 몰입감 있는 도슨트 경험을 실시간으로 구현할 수 있었다.

4-3 백엔드 구현

백엔드 서버는 엣지 디바이스에서 수신되고 처리된 영상 및 음성 데이터를 이용하여, 객체 인식부터 해설 생성에 이르는 전체 추론 파이프라인을 수행한다. 본 연구에서는 효율적인 기능 분리를 위해 AI 추론 및 멀티모달 검색을 담당하는 FastAPI 서버와 사용자 인터페이스 및 서비스 로직을 관리하는 Spring Boot 서버로 이원화하여 구현하였다.

백엔드의 처리 파이프라인은 크게 (1) YOLOv8을 활용한 작품 분석 및 데이터베이스화, (2) CLIP 임베딩 및 FAISS 기반의 지식 검색, (3) LLM-RAG 기반의 해설 생성 단계로 구성된다.

1) YOLOv8을 활용한 작품 분석 및 데이터베이스화

이 단계는 고도화된 해설 서비스 제공을 위해 전시 작품을 사전에 분석하고 검색 가능한 형태로 저장하는 과정이다. FastAPI 서버는 객체 탐지 모델인 YOLOv8을 사용하여 등록할 작품 이미지를 분석한다.

1-1) 객체 식별 및 선별: YOLOv8는 작품 내에서 해설이 필요한 핵심 객체들을 바운딩 박스 단위로 식별한다.

1-2) Crop 이미지 생성 및 ID 매핑: 식별된 객체들은 OpenCV를 통해 개별 crop 이미지로 자동 추출된다. 각 crop 이미지는 원본 작품 ID와 연동된 json, 인덱스 파일 구조로 관리되어 데이터 일관성을 유지한다.

1-3) 메타데이터 저장: 탐지된 객체의 라벨은 사람이 이해할 수 있는 텍스트 라벨로 변환되어 데이터베이스에 저장되며, 이후 설명 자동 매핑 과정에서 사용된다.

이러한 전처리 과정을 통해 단일 작품 이미지를 다수의 세부 객체 단위로 분해하여, 추후 실시간 서비스 단계에서 관람객의 시선 지점에 최적화된 정보를 즉각적으로 검색할 수 있는 기반을 마련한다.

2) CLIP 임베딩 및 FAISS 기반의 지식 검색

사전 구축된 객체 데이터베이스를 바탕으로, 실시간 관람 상황에서 최적의 지식을 찾아내는 과정이다. crop 이미지가 생성된 후, OpenAI의 CLIP 모델을 이용하여 각 이미지의 특징을 고차원 벡터(512차원 임베딩)로 변환하였다. 이 벡터는 이미지의 시각적 패턴과 의미적 정보를 동시에 포함하며, 동일한 임베딩 공간에 텍스트 설명을 매핑함으로써 이미지–텍스트 간의 의미적 비교가 가능해진다. 이후 FAISS 라이브러리를 활용하여 모든 이미지 벡터를 인덱싱하고, 벡터 기반 유사도 검색 구조를 구축하였다. 관람객의 시선이 특정 영역의 객체를 가리키면, 해당 crop의 벡터가 FAISS 데이터베이스와 비교되어 가장 유사한 설명 데이터(Top-K)가 검색된다. 이 단계는 RAG 구조의 Retrieval 부분에 해당하며, 지식베이스 내에서 관련성이 높은 예술 작품 설명을 빠르게 검색하는 역할을 수행한다.

3) LLM-RAG 기반의 해설 생성

검색된 설명 문서는 LLM로 전달되어, 자연스럽고 감성적인 구어체 해설로 변환된다. 본 연구에서는 OpenAI GPT-4o API를 사용하였으며, 한국어 도슨트 스타일에 맞게 프롬프트를 설계하였다. 프롬프트는 작품명, 작가, 제작 시기, 검색된 설명 문장 등을 입력받아, “감상자에게 작품의 맥락과 의미를 대화하듯 전달하는 설명문”을 생성하도록 구성되었다. 예를 들어, “양산을 쓴 여인”의 여인이 포함된 영역을 응시한 경우, 모델은 그림 5와 같은 형태의 해설을 생성한다. 영어 버전의 해설도 가능하다.

이렇게 생성된 해설은 FastAPI 서버를 통해 Spring Boot 서버로 전달되며, TTS 모듈에 의해 음성으로 변환되어 아이웨어를 착용한 관람객에게 실시간으로 제공된다.

*These are screenshot of the app in Korean.
그림 5. 아이웨어 착용 시연과 LLM 생성 응답 예시
Fig. 5. Eyewear demonstration and LLM response examples

4-4 시연 및 성능 평가

1) 실험 환경 및 구성

본 연구에서 제안하는 실시간 대화형 미술관 가이드 시스템의 성능 평가는 NVIDIA Jetson Orin Nano Developer Kit 플랫폼을 기반으로 하는 엣지 디바이스 환경에서 수행했다. 실험 환경은 실내 미술관 조건을 모사한 통제된 환경으로 조성되었으며, 주요 환경 및 시스템 구성 요소의 세부 사항은 표 2와 같다.

표 2. 주요 환경 및 시스템 구성 요소
Table 2. System environment and core components

Component Category	Specifications	Notes
Hardware (Edge Device)	NVIDIA Jetson Orin Nano Developer Kit (8GB RAM, 1024-core NVIDIA Ampere GPU)	GPU-accelerated env.
OS/JetPack	Ubuntu 20.04 LTS (JetPack 5.1.2)	
Object Detection Model	Eyedia-Art(Fine-tuned YOLOv8n), YOLOv8	Artwork/Object distinction
Gaze Tracking Model	Eyedia-Gaze(Fine-tuned YOLOv8n)	Pupil/Gaze quadrant estimation
Vector Search Library	FAISS (faiss-cpu 1.7.4)	
Embedding Model	CLIP (ViT-B/32)	
Language Model(LLM)	GPT-4 Turbo API (Azure OpenAI Service)	
Network Connectivuty	Wi-Fi 5 (802.11ac)	Average latency: 50ms

2) 평가 절차 및 지표 정의

본 연구에서는 제안 시스템의 성능을 검증하기 위해 응답 지연(Response Time), 검색 정확도(Retrieval Accuracy), 시선-객체 매칭 정확도(Gaze-Object Matching Accuracy)를 중심으로 정량 평가를 수행하고, 사용자 설문을 통한 정성 평가를 병행하였다. 모든 실험은 동일한 하드웨어(Jetson Orin Nano) 및 네트워크 환경에서 수행되었으며, 조건 간 공정한 비교를 위해 동일한 입력 질의 세트와 평가 절차를 적용하였다.

정답 데이터(Ground-truth)는 평가 대상 작품별로 설명 대상 객체(object_id)와 정답 설명 문서(doc_id)를 사전에 매핑하여 구축하였다. 하나의 객체에 대해 복수의 정답 문서를 허용함으로써 Top-K 기반 검색 정확도 평가가 가능하도록 설계하였다.

질의 세트는 (i) 작품 또는 객체를 직접 지칭하는 1차 질의와 (ii) 의미·맥락을 확장하는 2차 질의로 구성하였으며, 모든 실험 조건에서 동일한 질의를 사용하여 변동성을 최소화하였다.

구성 요소별 기여도를 분석하기 위해 'Full system'을 기준으로 시선 추적 또는 RAG 구성 요소를 제거한 ablation 조건을 포함하였다. 비교 조건은 시선 기반 객체 선택을 제거한 'No gaze', 검색 및 RAG를 제거한 'No RAG', 텍스트 기반 검색만 수행하는 'Text-only RAG'로 정의하였다. 각 조건에서 LLM에 전달되는 입력 정보는 명시적으로 고정하여 조건 간 비교 가능성을 확보하였다.

시선 기반 객체 선택은 프레임 단위 시선 좌표를 입력으로 하며, 시간 윈도우 기반 안정화 후 시선점이 포함되는 객체 또는 시선점-객체 중심 거리 최소 규칙을 통해 대상 객체를 결정하였다. 정확도는 시스템이 선택한 객체가 정답 객체와 일치하는 비율로 산출하였다.

검색 정확도는 시선 기반으로 선택된 객체에 대해 시스템이 반환한 상위 K개 설명 후보에 정답 문서가 포함되는지를 기준으로 Recall@1 및 Recall@5로 평가하였다. 응답 속도는 사용자 입력 완료 시점부터 TTS 출력 시작까지의 종단 간 지연으로 정의하였으며, 평균, 표준편차, 95th percentile을 산출하였다. 초기 캐시 및 모델 준비 시간의 영향을 최소화하기 위해 워밍업 후 측정을 수행하였다.

3) 성능 평가 결과

시스템의 성능은 응답 속도, 검색 정확도, 시선 추적 정확도의 세 가지 핵심 측면에서 정량적으로 평가하였다(표 3).

또한 제안된 멀티모달 구성 요소 각각의 기여도를 분석하기 위해, 시선 추적 및 RAG 모듈을 단계적으로 제거한 ablation study를 추가로 수행하였다.

(1) 응답 속도(Response Time)는 사용자가 음성 질의를 입력한 시점부터 TTS 응답이 출력되기까지의 전체 지연 시간을 측정 기준으로 설정했다. 동일 네트워크 환경에서 50회의 독립 질의를 수행하여 표 3과 같이 평균 1.839초의 응답 시간이 소요되어 실시간 대화형 시스템 설계에서 널리 인용되는 휴리스틱인 최대 반응 지연 시간 2초를 충족하였다[13],[14]. 세부 지연 분포는 평균, 표준편차, 95th percentile 기준으로 표 3에 요약하였다.

(2) 검색 정확도는 시선 기반으로 선택된 객체에 대해 반환된 설명 후보 중 정답 문서가 포함되는지를 기준으로 Recall@K 지표로 평가하였다. 그 결과, Top-1 Recall은 83.5%, Top-5 Recall은 96.2%로 나타나, 사용자의 관심 객체에 대해 높은 검색 신뢰성을 확인하였다. 이는 기존 CLIP-FAISS 기반 이미지-텍스트 검색 및 박물관 도메인 연구에서 보고된 실용 기준(Top-5 Recall ≥90%)을 충족하는 수준이다[5],[15].

(3) 시선 추적 정확도는 Eyedia-Gaze와 Eyedia-Art 모델을 결합한 시선-객체 매칭 결과를 기준으로 평가하였다. 20명의 참여자를 대상으로 120회의 응시 시나리오를 수행한 결과, 표준 조도 조건에서 91.4%의 평균 정확도를 달성하였다. 저조도(300-400 lux) 및 측면 시점 조건에서는 각각 87.2%, 85.9%로 정확도가 감소하였는데, 이는 RGB 기반 카메라 입력에서 조도 저하에 따른 눈 영역 대비 감소로 특징 검출 신뢰도가 낮아지는 현상에 기인한 것으로 분석된다. 이러한 경향은 기존 모바일 및 웹캠 기반 eye-tracking 연구에서 보고된 80-90% 정확도 범위와 유사하다[2],[11]. 본 연구의 목적은 사용자별 캘리브레이션 없이 실제 전시 환경에서의 실시간 적용 가능성을 검증하는 데 있으며, 해당 조건에서도 실용적인 정확도를 달성하였다.

표 3. 성능 평가 요약
Table 3. Summary of performance evaluations

Category	Metric	Results
Response Time	Average latency	1,839ms
	Standard Deviation	449ms
	95th Percentile	2,710ms
Retrieval Accuracy	Top-1 Recall	83.5%
	Top-5 Recall	96.2%
Gaze-Tracking Accuracy	Standard condition	91.4%
	Low-light (300-400 lux)	87.2%
	Side-angle (±20-30°)	85.9%

표 4는 제안된 시스템의 주요 구성 요소를 단계적으로 제거한 절제 연구(ablation study) 결과를 요약한다. 실험 결과, 시선 기반 객체 선택과 멀티모달 RAG는 시스템의 응답 지연과 검색 신뢰성에 상호 보완적으로 기여함을 확인하였다. 시선 정보를 제거한 경우, 객체 범위가 확장되면서 지연 시간이 증가하였으며, RAG를 제거한 설정에서는 응답 속도는 개선되었으나 검색 정확도가 크게 저하되었다. 텍스트 기반 RAG는 일정 수준의 정확도 향상을 제공하였으나, 이미지 기반 단서를 함께 사용하는 전체 시스템 대비 성능이 제한적이었다. 이러한 결과는 제안 시스템에서 시선, 시각 정보, 검색 증강이 결합된 멀티모달 구조의 효과를 정량적으로 뒷받침한다.

표 4. 제안 시스템의 절제 연구 결과
Table 4. Ablation study results of the proposed system

Condition	Latency Avg(ms)	p95(ms)	Top-1 Recall(%)	Top-5 Recall(%)
Full system	1,839	2,710	83.5	96.2
No gaze	2,273	4,388	n/a	n/a
No RAG	1,631	2,729	51.9	51.9
Text-only RAG	2,518	3,938	74.0	81.5

4) 사용자 만족도 조사 및 결과 분석

정량 성능 평가와 함께, 시연 직후 사용자 설문을 통해 LLM이 제공하는 답변의 품질 및 체감 만족도를 추가로 평가하였다. 설문은 5점 리커트 척도(1 : 매우 낮음, 5 : 매우 높음)을 사용하였으며, 총 19명이 참여하였다.

답변 품질과 관련하여, 설명·해설의 신뢰도는 평균 4.58±0.51로 높게 나타났고, 제공 정보의 사실 기반 인식은 19명 전원(100%)이 '네'로 응답하였다. 또한 응답 어투의 적절성은 평균 4.79±0.42로 감성·자연스러움 측면에서도 긍정적으로 평가되었다. 정보 충족 측면에서는 원하던 정보를 충분히 얻었는지가 4.79±0.54, 시스템이 질문 의도를 잘 이해했는지가 4.89±0.32, 전체적으로 해설의 깊이나 폭은 적절했는지가 4.84±0.37로 나타나, 사용자 관점에서 답변의 내용 품질(충분성·정확성·맥락성)이 전반적으로 높게 수용되었음을 확인하였다.

한편, 답변 만족도에 영향을 줄 수 있는 오류·지연 경험은 제한적으로 보고되었다. 시스템이 잘못된 객체를 설명한 경험은 1명(5.3%)에서만 나타났으며, 전체적으로 응답 지연을 체감했다는 응답은 2명(10.5%)으로 주로 "그림 인식 지연"으로 보고되었다. 응답 속도에 대한 체감 평가는 시선 인식 후 설명 속도 4.47±0.77, 추가 질문 시 설명 속도 4.68±0.67, 음성 질문 시 설명 속도 4.68±0.58로 나타났고, 실시간 반응형 경험 역시 4.79±0.54로 평가되어, 시스템의 실시간 처리 성능이 사용자 경험 측면에서도 대체로 일관되게 긍정적으로 인식되었음을 확인하였다.

시스템의 병목 지점을 식별하기 위해 모듈별 평균 처리 시간을 측정한 결과, 전체 파이프라인 처리 시간(1,850 ms)의

약 50.2%인 928.9 ms를 FAISS 벡터 검색이 차지하는 것으로 확인했다. 반면, TensorRT 최적화 및 GPU 가속이 적용된 객체 탐지와 시선 추적은 각각 45.2 ms와 28.7 msLLM API 호출은 평균 193.2 ms가 소요되었으나 표준편차(78.5 ms)가 상대적으로 높아 네트워크 상태에 따른 변동성이 존재했다. 분석 결과, 객체 탐지 및 시선 추적의 낮은 지연 시간은 TensorRT 최적화 및 GPU 가속의 효과를 뒷받침한 반면, FAISS 벡터 검색이 전체 지연 시간의 절반을 차지하는 주된 병목구간임을 확인하여 향후 Product Quantization 또는 GPU 기반 FAISS 활용이 향후 연구 방향으로 제시되었다.

종합적으로 실험 결과는 제안 시스템이 실시간성(1,850ms), 검색 신뢰성(Top-5 Recall 96.2%), 시선 추적 안정성(91.4%)이라는 핵심 요구사항을 충족했다. 또한 LLM을 기반으로 사용자의 질문 의도와 맥락을 반영한 동적 설명 생성이 가능해짐으로써 기존 시스템의 정적·일방향적 안내의 한계를 효과적으로 극복하였다. 시선 추적과 음성 대화를 결합한 멀티모달 상호작용 구조는 관람 과정에 자연스럽게 통합되는 몰입형 사용자 경험을 제공하였으며, Jetson 기반 엣지 컴퓨팅 플랫폼을 활용한 실험을 통해 실시간 처리 성능과 현장 적용 가능성 역시 확인되었다. 이러한 분석은 제안 시스템이 전시 환경에서 요구되는 반응성·정확성·사용성 기준을 전반적으로 충족하며, 향후 확장 가능한 지능형 도슨트 플랫폼으로서의 잠재력을 가진다는 것을 보여준다.

V. 결 론

본 연구는 관람객의 시선과 음성 입력을 활용하여 실시간으로 개인화된 해설을 제공하는 멀티모달 RAG 기반 미술관 도슨트 시스템을 구현하였다. 제안된 시스템은 YOLOv8 기반 객체 인식, CLIP 임베딩·FAISS 검색, RAG·LLM 기반 해설 생성, STT/TTS 인터페이스를 단일 파이프라인으로 통합하여, 관람객의 주목 대상에 동기화된 맞춤형 설명을 제공한다. 이를 통해 기존 오디오 가이드의 한계를 극복하고, 저비용·개방형 구조로 현장 적용 가능성을 제시하였다.

실험 결과, 본 시스템은 응답 시간, 검색 정확도, 시선 추적 정확도 등 주요 성능 지표에서 실시간 도슨트 시스템의 요구사항을 충족하였으나 모듈별 처리 시간 분석을 통해 전체 지연의 약 절반이 FAISS 벡터 검색에서 발생함을 확인하여 이 부분이 향후 최적화 방향을 제시하는 핵심 결과임을 알 수 있었다. 따라서 향후 연구에서는 Product Quantization(PQ), GPU 기반 FAISS 인덱스 최적화, 벡터 캐싱 전략 등을 적용하여 검색 병목을 완화하고, 전체 파이프라인의 실시간성을 더욱 개선하는 데 초점을 맞출 예정이다.

또한 본 연구는 통제된 전시 환경을 중심으로 시스템의 실시간성 및 정확도를 검증하였으나, 실제 미술관 환경에서는 군중 밀집, 조명 변화, 작품 표면 반사, 유사한 시각적 특징을 갖는 인접 작품 등 보다 복합적인 조건이 존재할 수 있다. 향후 연구에서는 이러한 실제 전시 환경 요소를 체계적으로 반영한 추가 실험을 통해, 시선–객체 매칭의 강인성 및 멀티모달 검색 성능의 일반화 가능성을 검증할 계획이다.

이러한 연구들이 병행된다면, 제안된 시스템은 다양한 전시 환경에서 높은 신뢰성과 응답성을 갖춘 차세대 AI 도슨트 플랫폼으로 확장될 것으로 기대된다.

참고문헌

[1] G. Trichopoulos, M. Konstantakis, G. Alexandridis, and G. Caridakis, M. Konstantakis, G. Alexandridis, and G. Caridakis, “Large Language Models as Recommendation Systems in Museums,” *Electronics*, Vol. 12, No. 18, 3829, 2023. https://doi.org/10.3390/electronics12183829

[2] P. Dondi and M. Porta, “Gaze-Based Human–Computer Interaction for Museums and Exhibitions: Technologies, Applications and Future Perspectives,” *Electronics*, Vol. 12, No. 14, 3064, 2023. https://doi.org/10.3390/electronics12143064

[3] T. Jung and I. Joe, “An Intelligent Docent System with a Small Large Language Model (sLLM) Based on Retrieval-Augmented Generation (RAG),” *Applied Sciences,* Vol. 15, No. 17, 9398, 2025. https://doi.org/10.3390/app15179398

[4] I. Vasic, H.-G. Fill, R. Quattrini, and R. Pierdicca, “LLM-Aided Museum Guide: Personalized Tours Based on User Preferences,” in *Proceedings of the International Conference on Extended Reality*, Lecce: Italy, pp. 249-262, 2024. https://doi.org/10.1007/978-3-031-71710-9_18

[5] M. Zhou, Y. Chen, and T. Li, “Context or Retrieval? Evaluating RAG Methods for Art and Museum Question Answering System,” in *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology* (IWSDS 2025), Bilbao: Spain, pp. 129-136, 2025.

[6] C. Zhou, B. Sinha, and M. Liu, “An AI Chatbot for the Museum Based on User Interaction over a Knowledge Base,” in *Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture*, Manchester: UK, pp. 54-58, 2020. https://doi.org/10.1145/3421766.3421888

[7] S.-W. Jung, E.-S. Choi, S. An, Y. Kang, and S. Jeong, “Implementation of Scenario-Based AI Voice Chatbot System for Museum Guidance,” *The Korea Journal of BigData*, Vol. 7, No. 2, pp. 91-102, 2022. https://doi.org/10.36498/kbigdt.2022.7.2.91

[8] H. An, W. Park, P. Liu, and S. Park, “Mobile-AI-Based Docent System: Navigation and Localization for Visually

Impaired Gallery Visitors," *Applied Sciences*, Vol. 15, No. 9, 5161, 2025. https://doi.org/10.3390/app15095161

[9] S. U. Lee, J. Yun, D. Kim, D. Kim, S. Y. Oh, and S. H. Yoon, "CARDS: Comprehensive AR Docent System," in *Proceedings of the 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct* (ISMAR-Adjunct 2022), Singapore, pp. 739-743, 2022. https://doi.org/10.1109/ISMAR-Adjunct57072.2022.00156

[10] M. Chang, T. Yi, P. Y. Lai, J. H. Lee, and J.-H. Lee, "Analysis of Art Museums' Visitor Behavior and Eye Movements for Mobile Guide App Design," in *Proceedings of the 3rd International Conference on Intelligent Human Systems Integration (IHSI 2020),* Modena: Italy, pp. 1138-1144, 2020. https://doi.org/10.1007/978-3-030-39512-4_173

[11] L. Reitstätter, H. Brinkmann, T. Santini, E. Specker, Z. Dare, F. Bakondi, ... and R. Rosenberg, "The Display Makes a Difference: A Mobile Eye Tracking Study on the Perception of Art before and after a Museum's Rearrangement," *Journal of Eye Movement Research*, Vol. 13, No. 2, pp. 1-29, 2020. https://doi.org/10.16910/jemr.13.2.6

[12] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus: OH, pp. 1821-1828, 2014.

[13] M. Maslych, M. Katebi, C. Lee, Y. Hmaiti, A. Ghasemaghaei, C. Pumarada, ... and J. J. LaViola Jr., "Mitigating Response Delays in Free-Form Conversations with LLM-Powered Intelligent Virtual Agents," in *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, Waterloo: Canada, pp. 1-15, 2025. https://doi.org/10.1145/3719160.3736636

[14] Zilliz, "What Is an Acceptable Latency for a RAG System in an Interactive Setting (e.g., a Chatbot), and How Do We Ensure Both Retrieval and Generation Phases Meet This Target?," Milvus AI Quick Reference, 2025.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, ... and I. Sutskever, "Learning Transferable Visual Models from Natural Language Supervision," in *Proceedigs of the 38th International Conference Machine Learning (ICML)*, pp. 8748-8763, 2021.

강채원(Chaewon Kang)

2021년~현 재: 덕성여자대학교 컴퓨터공학전공 학부생
※관심분야 : 인공지능, 서버 개발, 검색증강생성(RAG)

김예빈(Yebeen Kim)

2021년~현 재: 덕성여자대학교 컴퓨터공학전공 학부생
※관심분야 : 서버 개발, 인공지능, 클라우드, 검색증강생성(RAG)

유채민(Chaemin Yu)

2021년~현 재: 덕성여자대학교 컴퓨터공학전공 학부생
※관심분야 : 클라우드, 서버 개발, 인공지능

윤소은(Soeun Yoon)

2021년~현 재: 덕성여자대학교 컴퓨터공학전공 학부생
※관심분야 : 웹 프론트엔드 개발, 사용자 경험(UX) 설계, 인터랙티브 웹 서비스 디자인

이윤서(Yunseo Lee)

2021년~현 재: 덕성여자대학교 컴퓨터공학전공 학부생
※관심분야 : 클라우드, 인공지능, 검색증강생성(RAG)

유견아(Kyeonah Yu)

1986년 : 서울대학교 제어계측공학과 공학사
1988년 : 서울대학교 제어계측공학과 공학 석사
1995년 : University of Southern California 컴퓨터학과 공학박사
1996년~현 재: 덕성여자대학교 컴퓨터공학전공 교수
※관심분야 : 인공지능과 LLM을 활용한 개인화된 학습, 멀티모달 정보 정달 시스템