JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 6, pp. 109-118, June 2025
https://doi.org/10.9708/jksci.2025.30.06.109

An Inclusive AI Docent System for Accessible and Interactive Art Appreciation using Vision and Language Models

Min-Su Kim*, Min Kim**, Hyeon-jung Kwak***, Ye-jun Choi****, Hyung-rok Lee*****, Chi-wook Ahn******, Won Joo Lee†, Young-Bok Cho‡

*Graduated Student, Dept. of Computer Engineering, Daejin University, Pocheon, Korea
**Student, Dept. of Convergence Software, Myongji University, Seoul, Korea
***Student, Dept. of Computer Science and Engineering, Ewha Womans University, Seoul, Korea
****Student, Dept. of Geoinformatics, Inha University, Incheon, Korea
*****Graduated Student, Dept. of Sport Industry Studies, Yonsei University Graduate School, Seoul, Korea
******Student, Dept. of Military Digital Convergence, Ajou University, Suwon, Korea
†Professor, Dept. of Computer Science & Engineering, Inha Technical College, Incheon, Korea
‡Professor, Dept. of Computer Education, Gyeongkok National University, Andong, Korea

[Abstract]

In this paper, we propose an interactive art appreciation system that integrates computer vision and large language models to enable users—including those with visual impairments—to actively engage with visual art. The system recognizes artworks using YOLO-based object detection and VGG16 classification, and applies HSV-based color correction to enhance the reliability of emotional analysis. Subsequently, the Qwen2.5-VL-3B model summarizes visual content, while the Qwen2.5-32B model generates emotionally enriched descriptions and facilitates interactive dialogues. Additionally, a retrieval-augmented generation (RAG) framework is implemented to answer user questions, and prompts inspired by Visual Thinking Strategies (VTS) are used to elicit emotional responses and foster meaningful engagement. The proposed system demonstrates the potential of artificial intelligence to enhance both the accessibility and immersiveness of art appreciation, offering a new model for inclusive interaction with visual culture.

▸ **Key words**: Artwork recognition, Computer vision, Vision-Language Model, Large Language Model, Visual Thinking Strategies, Human-AI interaction

• First Author: Min-Su Kim, Corresponding Author: Young-Bok Cho, Won Joo Lee
*Min-Su Kim (minkim0526@gmail.com), Dept. of Computer Engineering, Daejin University
**Min Kim (kim89@mju.ac.kr), Dept. of Convergence Software, Myongji University
***Hyeon-jung Kwak (2171003@ewhain.net), Dept. of Computer Science and Engineering, Ewha Womans University
****Ye-jun Choi (12201309@inha.edu), Dept. of Geoinformatics, Inha University
*****Hyung-rok Lee (hrl1220@yonsei.ac.kr), Dept. of Sport Industry Studies, Yonsei University Graduate School
******Chi-wook Ahn (ahnchiwook@ajou.ac.kr), Dept. of Military Digital Convergence, Ajou University
†Won Joo Lee (wonjoo2@gmail.com), Dept. of Computer Science & Engineering, Inha Technical College
‡Young-Bok Cho (ybcho@gknu.ac.kr), Dept. of Computer Education, Gyeongkuk National University
• Received: 2025. 04. 24, Revised: 2025. 05. 30, Accepted: 2025. 06. 19.

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

[요 약]

본 논문은 시각장애인을 포함한 다양한 사용자가 시각 예술을 주체적으로 감상할 수 있도록, 컴퓨터 비전과 대규모 언어 모델을 결합한 대화형 미술 감상 시스템을 제안한다. 제안 시스템은 YOLO 기반 객체 탐지와 VGG16 분류기를 통해 작품을 인식하고, HSV 색공간 기반 색상 보정을 적용하여 감정 분석의 신뢰도를 향상시킨다. 이어서 Qwen2.5-VL-3B 모델은 시각 정보를 요약하고, Qwen2.5-32B 모델은 감성적 설명과 상호작용형 대화를 생성한다. 또한 RAG 기반 질문 응답 구조와 Visual Thinking Strategies(VTS)를 활용한 프롬프트 설계를 통해 감상자의 감정 표현을 유도하고, 의미 있는 상호작용을 가능하게 한다. 본 시스템은 미술 감상의 몰입성과 접근성을 동시에 향상시키는 인공지능 응용의 가능성을 보여준다.

▸ **주제어**: 미술 감상, 시각장애인, 대화형 시스템, 컴퓨터 비전, 대규모 언어 모델, 감정 기반 AI

I. Introduction

현대 사회에서 인공지능(AI)의 영향력은 기술적 영역을 넘어 예술과 문화 전반으로 확장되고 있다. 특히 최근 등장한 대규모 언어 모델(Large Language Model, LLM)은 탁월한 자연어 이해 및 생성 능력을 바탕으로, 사람과 컴퓨터 간의 상호작용 방식을 근본적으로 변화시키고 있다[1]. 이러한 기술 발전은 기존 예술 감상의 일방향적인 정보 전달 방식을 넘어, 사용자와의 상호작용 중심으로 전환되는 계기를 제공하고 있다. 기존의 미술관이나 온라인 전시에서 제공되는 감상 방법은 주로 작품에 대한 텍스트 설명이나 오디오 가이드를 사용자가 수동적으로 수용하는 방식이었다. 이 방식은 기본적인 작품 정보 제공에는 효과적이지만, 사용자의 개인적 해석이나 정서적 반응을 반영하는 데 한계가 있으며, 사용자의 참여와 몰입을 제한하는 문제가 있었다. 본 연구에서는 이러한 문제를 극복하고, 시각장애인을 포함한 다양한 사용자들이 감상의 주체로서 능동적으로 참여할 수 있도록 하는 대화형 미술 감상 시스템을 제안한다. 특히 국내 사용자 환경에 적합하도록 한국어 기반의 인터페이스를 설계하였으며, 시각 중심의 전통적인 감상 방식에서 벗어나, 언어와 감성 중심의 새로운 예술 경험을 제공하고자 하였다. 이를 위해 본 연구에서는 오픈소스 기반의 대규모 언어 모델인 Qwen2.5를 활용하였다. Qwen2.5 모델은 프롬프트 설계의 유연성과 커스터마이징이 용이하고, GPT-4 또는 Claude 등 기존 상용 API 모델과 달리 로컬 환경에서 독립적으로 구동이 가능하여, 비용 효율성 및 시스템의 제어 측면에서도 유리하다[2]. 이는 본 시스템의 기술적 자율성과 향후 확장 가능성을 확보하는 데 중요한 기반이 되었다. 감상자의 질문이나 감정 표현에 따라 해석이 유연하게 변화하는 구조를 설계하여, 기존의 정답 중심 해설을 넘어 열린 해석과 심층적 상호작용을 지원하고자 하였다.

본 논문은 이러한 연구 배경과 필요성을 바탕으로, 인공지능 기술을 활용한 대화형 미술 감상 시스템의 기획, 설계, 구현 및 실험 결과를 종합적으로 기술한다. 논문의 구성은 다음과 같다. Ⅱ장에서는 관련 연구와 기술적 배경을 정리하며, Ⅲ장에서는 시스템의 구성 요소와 핵심 기술을 상세히 설명한다. Ⅳ장에서는 구현된 시스템의 실험과 평가 결과를 분석하고 논의한다. 마지막으로 Ⅴ장에서는 연구의 결론과 향후 발전 방향을 제시한다.

II. Preliminaries

1. Related works

최근 멀티모달 대형 언어 모델(Multimodal Large Language Model, MLLM)의 발전은 이미지와 텍스트 정보를 통합적으로 이해하고 생성할 수 있는 인공지능 응용의 가능성을 크게 확장하고 있다. 대표적인 사례로 LLaVA(Large Language and Vision Assistant)[3]와 같은 통합형 MLLM이 등장하였으며, 이미지 기반 질의응답, 시각적 설명 생성, 대화형 감상 지원 등 다양한 기능을 통해 예술 감상의 새로운 접근 방식을 제시하고 있다. 특히 LLaVA 모델을 기반으로 개발된 LLaVA-Docent 시스템은 실시간으로 감상자의 질문에 대한 응답을 생성하여 AI 도슨트의 대표적인 적용 사례로 평가받고 있다. 최근 버전인 LLaVA-Docent-V2에서는 데이터 품질 개선과 사용자 질문 흐름에 따른 구조적 응답 설계가 더욱 강화되었으며,

이러한 발전은 미술 교육 현장에서도 주목받고 있다[3]. 그러나 본 연구에서는 기존의 MLLM 기반 시스템과 차별화되는 새로운 구조를 제안한다. 기존의 단일 통합 모델 방식을 탈피하고, 감상의 흐름을 명확히 모듈화하여 설계하였다. 즉, 시각 정보 분석을 위한 Vision-Language Model(VLM)과 감성적 설명 생성을 위한 대규모 언어 모델(Large Language Model, LLM)을 독립적으로 구성함으로써, 시스템의 유연성과 설명 품질을 동시에 확보하고자 하였다. 특히 한국어 환경과 시각장애인, 어린이, 외국인 등 다양한 사용자층의 특성을 고려하여, 보다 효과적이고 포용적인 구조를 마련하였다.

2. Visual Art Enjoyment for the Visually Impaired

최근 시각장애인의 문화예술 향유에 대한 사회적 관심과 기술적 지원의 필요성이 지속적으로 증가하고 있지만, 실제 이들의 예술 감상 기회를 확대할 수 있는 기술적 장치는 여전히 부족한 상황이다. 박혜영·김혜랑(2024)의 연구[4]에 따르면, 시각장애인을 위한 문화예술 지원 기술은 주로 내비게이션 안내, 오디오 설명, 단순 이미지 묘사 등 제한적 기능에 집중되어 있으며, 심층적인 감상이나 정서적 교감을 촉진하는 방식은 부족하다는 점을 기존 연구들의 경향을 통해 시사하고 있다. 또한 시각장애인의 미술 감상 경험을 다룬 실증 연구[5]에서는, 시각장애인들이 단순한 사물 식별 정보를 넘어 작품 내 객체의 형태, 배치, 색상 등 시각적 상세내용에 대한 설명을 강하게 요구하고 있음이 나타났다. 특히 전맹 참여자조차도 배치 위치와 구체적인 색감 표현에 대한 정보 요청이 높았으며, 이는 시각 정보를 대체하거나 보완할 수 있는 심화된 해설의 필요성을 시사한다. 따라서 시각장애인을 위한 미술 감상 시스템은 단순 정보 전달에 그치지 않고, 감정적 몰입과 주체적 해석을 유도할 수 있도록 설계되어야 한다. 특히 작품의 분위기, 색상 조합이 주는 정서적 인상, 인물이나 사물 간의 관계 구조 등을 구체적이고 풍부하게 전달하는 것이 단순 정보 전달을 넘어 감상의 '경험화'를 가능하게 하는 핵심 요소로 작용할 수 있다. 이를 위해 본 연구는 시각 및 언어 처리 기술을 통합한 멀티모달 접근을 기반으로 한다. 먼저 YOLOv5를 활용하여 작품 이미지 내 주요 객체를 실시간으로 탐지하고, VGG16을 통해 탐지된 영역을 분류함으로써 작품 속 요소의 정체와 구조적 배치를 정밀하게 파악한다. 이후 Qwen2.5-VL과 같은 고성능 시각-언어 통합 모델을 통해 이러한 시각적 정보를 서술형 언어로 변환하고, Qwen2.5-32B 언어 모델은 작품의 분위기, 색감, 상징성 등 감정적 인상을 풍부하게 설명하는 역할을 수행한다. 이와 같은 기술적 설계를 통해 시각장애인은 단순한 객체 나열을 넘어, 작품의 구성과 정서에 대한 입체적이고 몰입감 있는 감상 경험을 체험할 수 있게 된다.

3. The Need for Visual Thinking Strategies (VTS)

시각적 사고 전략(Visual Thinking Strategies, VTS)은 어린이, 외국인, 예술 비전공자 등 미술 감상 경험이 적거나 감상 능력이 낮은 사람들을 위한 교육 방법으로 널리 활용되고 있다[6]. VTS는 단순 정보 습득이나 작품 설명 위주의 교육이 아니라, 감상자가 스스로 '생각하기'와 '말하기'를 수행하며 적극적으로 참여하도록 유도하는 상호작용 중심의 교육 전략이다[6]. 이 경우 VTS가 질문 중심의 접근 방식을 통해 사용자의 적극적인 발화를 유도하고, 감상자의 사고 능력과 작품 이해 능력을 심화하는 효과가 있음을 분석하였다. 또한 「Theory and Practice of Visual Thinking Strategies in Upper Secondary Education」 논문[7]에서도 VTS가 고등학생의 시각적 사고와 비판적 감상 능력을 강화하는 데 효과적인 교육 방법으로 자리 잡고 있으며, 최근 미술 교육 현장에서 더욱 빈번하게 활용되고 있다고 보고하였다. VTS는 작품에 대해 개방적이고 비판적인 질문을 던지고, 감상자의 의견을 표현하며 타인의 의견을 재해석하는 과정을 통해 감상 능력을 자연스럽게 향상시킨다는 점에서 의의가 있다.

본 연구에서는 이러한 VTS의 기본 철학과 방법론을 인공지능 기반 대화 시스템 설계에 적극적으로 도입하여, 사용자의 감성적 해석과 상호작용을 촉진하고자 하였다.

III. The Proposed Scheme

본 연구에서 제안된 시스템은 사용자가 미술 작품에 대한 감정 표현을 입력하면, AI가 이를 바탕으로 질문과 해석을 제시하여 감정 중심의 상호작용을 지원하는 구조를 가진다. 이러한 방식은 앞서 언급한 VTS의 철학을 인공지능을 활용하여 기술적으로 구현한 대표적 사례이다.

1. System Development Environment

개발환경은 [표 1]과 같이 구성하였다. 먼저 사용자가 모바일 앱에서 이미지를 촬영하거나 업로드하면, AWS 서버 환경에서 실행되는 백엔드가 이미지를 처리하고 분석을 수행한다. 주요 언어 모델과 비전 모델은 로컬 GPU 서버 환경에서 구동되며, AWS의 클라우드 서비스가 이미지 관리 및 데이터베이스를 제공한다.

Table 1. Development Environment

Division	Complement
OS	Windows 11 Pro
Library	Python, Dart, TensorFlow, Pytorch
Front End	Flutter (Dart 3.2.3), VS Code
Back End	FastAPI (Python 3.10.13), VS Code, Jupyter Notebook, Google Colaboratory
DB	MySQL, AWS RDS AWS S3
Language Model(LLM)	Qwen2.5, Qwen2.5-Coder model via Groq API
MultiModal model(VLM)	HuggingFace-based Qwen2.5-VL model (RTX 4090 GPU local environment)
API	ChatGPT, Groq, Google Cloud Platform

2. Application Workflow Overview

본 시스템은 [그림 1]과 같이 사용자가 작품 이미지를 입력하는 단계에서부터 작품에 대한 설명 및 대화형 감상까지 전 과정을 자동화된 흐름으로 처리한다.

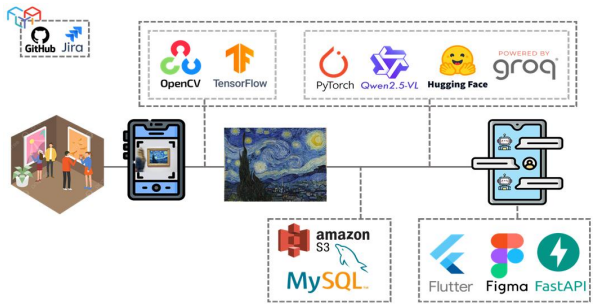


Fig. 1. Application Workflow

사용자가 모바일 환경에서 작품 이미지를 촬영하거나 업로드하면, 백엔드는 YOLO 모델[8]로 작품의 위치를 실시간으로 탐지하고 최적의 프레임 정렬을 지원한다. 또한 적응형 임계값 설정과 캐니 에지 감지를 통해 외곽선을 강조하여 이미지를 정면 시점으로 보정하고, K-means 클러스터링을 통해 주요 색상을 추출해 HSV 색공간 기반 보정을 통해 원작의 분위기와 감성적 느낌을 더 명확히 전달할 수 있도록 색상 표현을 조정하였다. 작품의 정확한 인식은 VGG16 모델과 전이학습을 통해 이루어지며, 작품 메타데이터는 자동으로 분류되어 데이터베이스에 저장된다. 이렇게 정제된 이미지 정보는 Qwen2.5-VL 모델에 입력되어 작품 장면에 대한 간략한 요약 설명을 생성하고, 그 결과는 다시 LLM(Qwen2.5-32B) 모델을 통해 더욱 풍부하고 감성적인 문장으로 재구성된다. 생성된 설명은 텍스트와 음성 형태로 사용자에게 제공되며, 사용자의 추가적인 질문과 감상 표현에 따라 AI가 새로운 질문이나 해석을 제시함으로써 상호작용이 이루어진다.

2.1 Detection of Painting Regions

사용자가 입력한 이미지에서 미술 작품 영역을 정확히 탐지하기 위해, 본 시스템은 단계별 탐지 과정을 수행한다. 우선 YOLO 모델을 이용해 그림의 위치를 탐지하며, 이때 Roboflow에서 제공한 약 4,000장의 미술 작품 데이터셋[9]으로 학습된 모델을 사용하여 높은 인식 정확도를 확보하였다. 탐지된 이미지에서는 적응형 임계값 설정과 캐니 에지 감지를 사용하여 외곽 윤곽선을 검출한 뒤, 원근 변환으로 정면 시점 보정을 수행한다. 탐지된 영역이 충분히 신뢰할 만한 크기가 아닐 경우 원본 이미지를 그대로 활용하며, 최종적으로 정제된 관심 영역을 추출하여 이후 단계에서 사용한다.

```
# Extracting artwork area based on perspective transform
1: function EXTRACT_DOMINANT_COLORS(image, k)
2:     Reshape image to 1D array of pixels
3:     Convert pixel values to float32
4:     Define K-means criteria (epsilon + max_iter)
5:     Run K-means clustering with k clusters
6:     Count number of pixels in each cluster
7:     Sort cluster centers by frequency
8:     return sorted color palette
9: end function
```

Code. 1. ROI-based color analysis and artwork region detection

2.2 Artwork Metadata Retrieval

작품 정보 분석 단계는 탐지된 작품 이미지를 바탕으로 작품의 주요 특징을 시각적, 맥락적, 언어적 관점에서 심층적으로 분석한다. 이 과정은 크게 색상 추출, 제목 분류, VLM 특징 추출의 세부 단계로 구성된다.

2.2.1 Dominant Color Extraction

작품의 색상 정보는 작품의 분위기와 정서를 분석하는 데 필수적이다. 본 연구에서는 [그림 2] 단계로 색상 정보를 추출하고 분석한다.

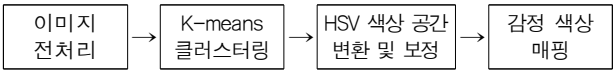


Fig. 2. Color Extraction Step

첫 번째로 이미지 전처리는 ROI로 분리된 작품 이미지를 적절한 크기를 224x224 픽셀로 리사이징하고 픽셀 데이터를 2차원 배열로 변환한다. 두 번째로 K-means 클러스터링은 OpenCV 기반 K-means 클러스터링을 적용해 대표 색상을 5개 클러스터로 추출하고, 각 클러스터의 픽셀 점유율을 계산하여 주요 색상 팔레트를 생성한다. 세

번째로 HSV 색상 공간 변환 및 보정은 추출된 RGB 색상은 HSV(Hue, Saturation, Value) 색공간으로 변환된 후, 명도와 채도를 인간 시각의 특성과 유사하게 보정한다[10].여기서 HSV 색공간 기반 색상 보정을 적용한 감정분석에서는 RGB 대비 최대 6.3%의 정확도 향상[11]이 보고되었고 이는 채도 및 명도의 정규화가 감정 인식 모델의 신뢰도 향상에 실질적으로 기여함을 의미한다. 또한 이는 명암 대비가 강한 작품이나 조명 환경의 영향을 최소화하여, LLM이 생성하는 설명의 일관성과 정확도를 높이기 위한 전처리 과정이다. 마지막으로 보정된 색상 정보는 명칭 변환 과정을 통해 '푸른색', '갈색', '회색' 등 직관적인 색상 설명으로 변환되며, 이는 LLM 프롬프트 구성 시 작품의 분위기를 묘사하는 데 중요한 입력으로 활용된다.

```
# K-means based primary color extraction
function EXTRACT_DOMINANT_COLORS(image, k)
    Flatten image into a 1D array of pixels
    Convert pixels to float format
    Apply K-means clustering with k clusters
    Count the number of pixels in each cluster
    Sort cluster centers by frequency
    return sorted dominant colors
end function

# Color correction based on the HSV color space
function ADJUST_HSV(rgb, lightness = 1.4, saturation = 1.3)
    Convert RGB to HSV → (h, s, v)
    if v < 150 then
        v ← min(v × lightness, 255)
    s ← min(s × saturation, 255)
    Convert adjusted HSV back to RGB
    return adjusted RGB
end function

# Mapping of emotional color categories
function GET_COLOR_NAME(rgb)
    Convert RGB to HSV → (h, s, v)
    Normalize hue: h ← h × 2
    if s < 40 then
        if v > 200 then return "Bright"
        else if v < 80 then return "Dark"
        else return "Gray"
    else if 85 ≤ h < 165 then return "Blue"
        else if 45 ≤ h < 90 then return "Green"
        else if 20 ≤ h < 45 then return "Yellow"
        else return "Other"
end function
```

Code. 2. Dominant color extraction and HSV-based correction

2.2.2 Artwork Title Classification

작품의 제목은 감상에 필수적인 역사적·문화적 맥락을 제공하기위해 [그림 3] 과정으로 작품의 제목과 관련 정보를 식별한다.

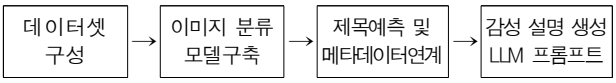


Fig. 3. Work Information Title Classification Step

데이터셋 구성은 Metropolitan Museum of Art(The Met)의 작품 데이터 약 300여 종을 Google Cloud Platform의 BigQuery 서비스를 통해 자동 수집하였다[12]. 작품 이미지와 메타데이터(JSON 형식)는 Selenium 및 urllib을 통해 수집 및 정제되었으며, 작가명, 제작 연도, 문화적 배경 등 상세 정보가 포함되어 있다. 다음은 이미지 분류 모델을 구축하기 위해 VGG16 모델을 ImageNet 사전 학습된 가중치로 초기화한 뒤, 수집된 데이터를 활용해 전이학습을 진행하여 작품 분류기를 구축하였다.

```
# Classification of works using CNN model
function PREDICT_IMAGE(image_path, model, classes, threshold)
    Load image and resize to 224×224
    Normalize image and expand dimensions
    pred ← model.predict(image)
    prob ← Maximum prediction probability
    idx ← Index of predicted class
    if prob < threshold then
        return "Unknown Title"
    else
        return classes[idx]
    end if
end function
```

Code. 3. Artwork title classification and metadata linkage

모델 입력을 위한 전처리를 수행하여 분류 정확성을 높였다. 세 번째로 작품 제목 예측 및 메타데이터 연계를 위해 이미지 입력 시 VGG16 모델이 300여 클래스 중 하나의 작품 제목을 예측하며, 이를 바탕으로 JSON 기반의 메타데이터와 연동하여 추가적인 맥락 정보를 제공한다. 마지막으로, 분류 결과와 메타데이터는 Qwen2.5-32B의 프롬프트에 함께 삽입되어, 생성되는 설명의 방향을 구체화하고, 내용의 신뢰성과 맥락적 정합성을 높이는 데 활용된다.

2.2.3 VLM Feature Extraction

VLM을 이용해 작품 이미지에서 자연어로 된 특징을 추출하는 단계로 [그림 4]과 같이 작품의 시각적 내용을 요약하여 이후 감성적 상호작용에 활용된다.

프롬프트 구성은 사용자가 실제로 작품을 보고 설명하는 듯한 문장을 얻기 위해 프롬프트에 "이 그림은 무엇을 묘사하고 있으며, 어떤 분위기를 가지고 있나요?" 등의 질문을 포함한다.

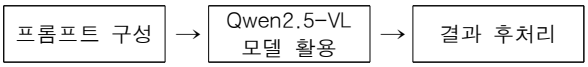


Fig. 4. VLM Feature Extraction Step

두 번째로 HuggingFace를 통해 로컬 환경에서 배포된 경량 모델인 Qwen2.5-VL을 활용하여 입력된 이미지와 프롬프트를 바탕으로 간결한 자연어 설명(2~3문장)을 생성한다. 마지막으로 생성된 설명에서 한국어 및 영어 이외의 문자(한자 등)는 제거하고, 최종 설명의 일관성과 간결성을 확보하여 후속 LLM 단계의 입력으로 제공한다.

```
# Generating explanations based on visual language models (Qwen)
function GENERATE_VLM_DESCRIPTION(image)
    if image is array then
        Convert to RGB PIL image
    else
        Open image from file path and convert to RGB
    end if
    Resize image to (512 × 512)
    prompt ← "Describe this painting in 2-3 natural sentences, as if in conversation."
    messages ← [
        { "role": "user", "content": [
            { "type": "image", "image": image },
            { "type": "text",  "text": prompt }
        ]}
    ]
    chat_input ← Apply Qwen chat template to messages
    inputs ← Preprocess chat_input and image for model
    output ← Generate description with Qwen-VL (max 256 tokens)
    return Postprocessed description (with prompt removed)
    description ← Decode model output and remove special tokens
end function
```

Code. 4. Feature extraction using vision-language models (VLMs)

2.3 Automatic Generation of Artwork Descriptions

작품 설명 생성 단계는 사용자가 입력한 작품 이미지에 대해 전 단계에서 분석된 시각 정보(색상, 제목, 장면 요약 등)와 메타데이터를 바탕으로 자연스러운 문장을 구성하는 과정이다. 앞선 과정에서 수집한 정보를 기반으로 설명을 생성하여 생성된 설명의 신뢰도를 보장하도록 한다.

작품 설명 생성은 VLM 특징 추출의 결과와 함께 색상 분석에서 도출된 감정 키워드, 작품 제목 및 메타데이터 정보를 Qwen2.5-32B 모델에 입력하여 풍부하고 감성적인 해설 문장을 생성한다. 해당 모델은 Groq API를 통해 실시간으로 추론되며, 시각장애인 등 설명의 정밀도와 일관성이 중요한 사용자 환경을 고려하여 선택되었다. 작품 제목이 인식된 경우, The Met의 JSON 메타데이터를 참조하여 작가, 제작 시기, 문화적 맥락 등의 정보를 함께 반영하며, 데이터가 존재하지 않거나 불확실한 경우에는 주관적 감상에 기반한 시각적 묘사 중심의 설명을 생성하도록 설계되었다. 이러한 생성 구조는 단순한 장면 설명을 넘어서, 작품에 대한 정보 전달과 감성적 해석을 동시에 만족시키는 사용자 맞춤형 해설을 제공하며, 이후 대화형 감상 서비스의 기반 데이터로 활용된다.

```
# Configuring LLM prompts to generate work descriptions
function GENERATE_RICH_DESCRIPTION(title, vlm_desc, colors, edges)
    info ← Search artwork metadata by title
    if info is not found then
        color_labels ← Get emotional labels for top 5 colors
        prompt ← Compose descriptive prompt using vlm_desc and color_labels
    else
        artist ← info["artist"] or "Unidentified Artist"
        color_labels ← Get unique emotional labels for top 5 colors
        prompt ← Compose prompt using title, artist, vlm_desc, and color_labels
    end if
    response ← Call Qwen LLM with prompt and generation settings
    return Generated description (stripped and cleaned)
end function
```

Code. 5. Artwork description generation using a general-purpose large language model.

2.4 Conversational Art Appreciation Interface

본 시스템은 감상자가 예술 작품과 능동적으로 상호작용할 수 있도록, 대화형 감상 기능을 핵심으로 설계되었다. 특히 시각장애인, 예술 비전공자, 외국인 등 감상 접근성이 낮은 사용자도 능동적으로 작품을 해석하고 표현할 수 있도록 유도하는 상호작용 구조를 채택하였다. 대화형 감상은 사용자의 발화를 정보형 질문과 감성 표현 유형으로 구분하여 작동한다. 초기에는 발화 유형 분류를 위한 NLP 기반 모델 도입을 검토하였으나, 경량화 및 실시간 응답성을 고려하여 '~요?', '~인가요?' 등 의문 표현과 작품 관련 키워드를 포함한 정보형 발화와 감정적 표현 및 주관적 인상을 포함한 감성형 발화로 간단한 규칙 기반 분석 방식으로 대체하였다.

정보형 발화에 대해서는 사전에 수집한 약 20~30권 분량의 명화 해설 자료를 기반으로 구축한 텍스트 데이터셋을 활용하여 RAG기반 응답을 제공한다. 전체 문장을 문장 단위로 분리하여 임베딩한 후, 사용자의 질의와 60% 이상 유사한 상위 20개의 문장을 검색하고, 이를 기반으로 Qwen2.5-Coder 모델이 최종 응답을 생성한다. 이 구조는 설명의 정확도, 문맥성, 응답 속도 면에서 우수한 성능을 나타내며, 특히 사용자가 미술 감상에 필요한 배경 정보를 빠르게 획득할 수 있도록 돕는다. 감성 표현 발화에 대해서는 VTS의 이론에 기반한 질문 프롬프트 구조를 활

용한다. 사용자의 최근 3개 발화를 히스토리로 저장하여 문맥을 유지하며, VTS 연구에서 제시된 20가지 언어 활동 유형에 기반[6]해 구성한 180개의 질문 예시 중에서 적절한 질문을 선택하거나 RAG 방식으로 생성하였다. 최종 대화 흐름은 먼저 발화 유형을 분석한 뒤, 유형에 따른 응답을 처리하고, 이어서 감상과 연계된 질문을 제시하는 구조로 구성된다. 시스템은 사용자의 감상 맥락에 따라 질문의 방향을 유동적으로 조정하며 대화를 유도하고, 이를 통해 사용자가 단순한 설명의 수용자가 아닌 감상의 주체로서 참여할 수 있도록 지원한다. [그림 5]는 사용자의 입력에 따라 정보의 질문과 감상 표현에 대한 분류 과정을 도식화한 것이다. 이러한 흐름은 예술 감상의 몰입도와 만족도를 높이는 데 기여한다.

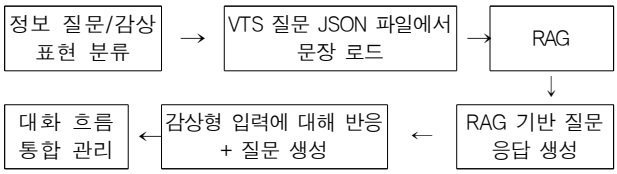


Fig. 5. Classify information questions/expressions based on user input

IV. Experimentation and Evaluation

본 장에서는 제안된 대화형 미술 감상 애플리케이션의 실제 작동 과정을 중심으로 기능별 실험을 수행하고, 이를 통해 사용자 경험 및 접근성 향상 측면에서의 효과를 평가하였다. 특히 시각장애인을 포함한 다양한 사용자 환경을 고려하여, 작품 인식, 설명 제공, 상호작용 대화, 감상 이력 관리 등 주요 기능별 성능과 사용성을 분석하였다.

1. Guidance for Artwork Recognition and Photography

[그림 6]는 앱의 실시간 촬영 인터페이스와 음성 안내 기능의 동작 화면을 시각화 한 것이다.

(가) 실시간 작품 인식

사용자가 '작품 감상하기' 버튼을 선택하면, 카메라가 자동으로 실행되며 YOLO 기반 객체 탐지 알고리즘이 활성화된다. 시스템은 입력 이미지 내 미술 작품을 실시간으로 탐지하고, 작품과 프레임의 상대적 위치 차이를 판단하여 적절한 촬영 각도 및 거리를 안내한다.

(나) 음성 피드백 시스템

시각장애인을 위한 주요 설계 요소로, 화면을 확인할 수 없는 사용자의 촬영을 지원하기 위해 음성 안내 시스템이 도입되었다. 작품이 중심에 위치했을 경우, 안내 음성과 함께 “객체가 중앙에 위치했습니다.” 등의 메시지가 제공되어 음성 피드백을 구현하였다.

(다) 음악 및 대기 안내음

작품 분석 중에는 백색소음 또는 간단한 클래식 음악을 재생하여 시스템이 정상 작동 중임을 청각적으로 인식할 수 있도록 구성하였으며, 이는 사용자의 불안감을 낮추고 앱 신뢰성을 높이는 요소로 작용한다.

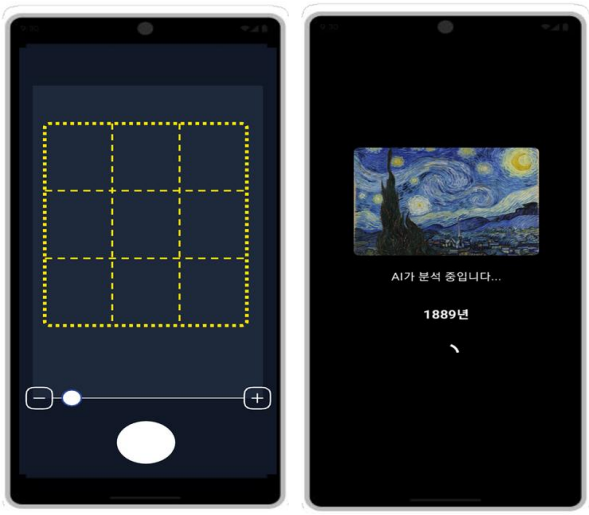


Fig. 6. User Interface for Camera and Voice Feedback

2. Artwork Description and Reflective Dialogue

[그림 7]는 작품 촬영 후 설명 제공 및 감상 대화 흐름을 시각화한 것이다.

(가) 시각적 요약 설명 제공

작품 이미지가 입력되면, Qwen2.5-VL-3B 모델이 시각 정보를 요약한 후 Qwen2.5-32B 모델이 이를 바탕으로 감성적이고 직관적인 해설 문장을 생성한다. 생성된 설명은 텍스트와 음성을 통해 제공되며, 색상, 분위기, 등장 인물의 행동 등 감상의 핵심 요소를 포함한다. 이는 시각정보가 없는 사용자도 작품의 정서를 효과적으로 체감할 수 있도록 설계하였다.

(나) 대화형 감상 흐름 구성

사용자가 작품에 대해 질문하거나 감상을 입력하면, 시스템은 발화 내용을 분석하여 질문 유형을 판별한다. 정보 중심 질문의 경우 RAG 기반 응답 구조를 통해 명확한 정보를 제공하며, 감정 중심 표현의 경우 VTS 기반 질문을 이어가는 구조로 구성된다

Fig. 7. AI-based artwork analysis and Interactive appreciation system

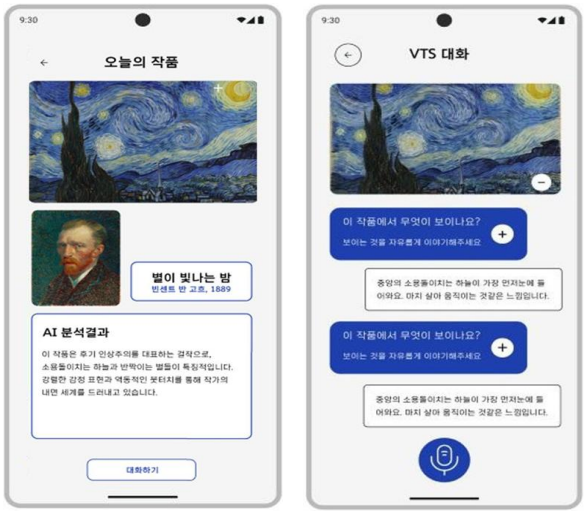


Fig. 8. Artwork Recommendation and Interactive Appreciation System

(다) 다중 대화 히스토리 저장

자연스러운 대화 흐름 유지를 위해 사용자의 최근 발화 내역을 저장하고, 이를 후속 응답 생성 시 참조하도록 설계하였다. 이를 통해 맥락을 고려한 연속적 감상 대화가 가능하다.

3. Artwork Recommendation and Reflective Diary

[그림 8]은 작품 추천 및 감상 이력 관리를 시각화한 것이다.

(가) 오늘의 추천 작품

앱 초기 화면에서는 '오늘의 추천 작품' 기능을 통해 사용자에게 명화 한 점을 제시하며, 해당 작품에 대한 설명 및 감상 기능을 동일하게 제공한다. 추천 대상은 시대별, 장르별로 다양하게 구성된다

(나) 감상 일기 기능

사용자는 작품 감상 이후 느낀 점을 기록할 수 있으며, 시스템은 이전 대화 데이터를 기반으로 감상 이력을 확인함으로써 감상 일기처럼 활용할 수 있다. 작성된 감상 일기는 개인별 감상 기록으로 저장된다.

(다) 감상 기록 열람

감상 이력은 달력 형태로 확인 가능하며, 날짜별로 감상한 작품 이미지, 생성된 설명, 대화 내역을 모두 다시 조회할 수 있다. 추후, 이는 반복 감상 학습과 취향 기반 콘텐츠 추천에 활용이 가능하다.

4. Accessibility-Oriented Design and User Feedback

본 연구에서는 개발 앱의 시각장애인 접근성을 테스트하기위해 직접 사용 및 피드백을 실시하였다.

(가) 시각 정보 대체 설계

시각장애인을 위한 설계의 핵심은 비시각적 정보 대체 제공에 있다. 앱은 자체 TTS/STT 기능을 탑재하고 있으며, 음성 안내를 통해 모든 기능을 조작할 수 있다. 또한 iOS의 VoiceOver 및 Android의 TalkBack과의 호환성을 확보하여 전반적인 접근성을 높였다.

(나) 명도 대비 및 폰트 조정

명도 대비 기준에 따라 텍스트 색상을 자동 조정하고, 사용자 디바이스의 설정에 따라 글꼴 크기를 동적으로 변경하는 기능을 제공하였다. 이는 저시력 사용자에게 특히 유효하였다.

(다) 직접 시연을 통한 피드백 반영

시각장애인 안마사 사용자 3명을 대상으로 시연 테스트를 진행한 결과, 대화형 감상 기능은 기존 오디오 도슨트보다 정서적 몰입감과 자기 표현 기회 면에서 우수하다는 평가를 받았다. 특히 "작품과 대화를 나누는 느낌", "어떻게 감상할지 모르겠는 상황에서 좋은 가이드를 주는 것 같다."와 같은 의견이 반복적으로 나타났다.

V. Conclusions

본 연구는 시각장애인을 포함한 다양한 사용자가 시각 예술을 보다 주체적으로 감상할 수 있도록, 컴퓨터 비전과 대형 언어 모델(LLM)을 결합한 대화형 미술 감상 애플리케이션을 설계하고 구현하였다. 제안된 시스템은 작품 자동 인식(YOLO 기반), 감성 설명 생성(Qwen2.5-VL/32B 활용), 음성 안내(STT/TTS), VTS 기반 질문 응답 등의 요소를 통합하여, 기존의 일방향 도슨트 방식에서 벗어난 몰입형 감상 경험을 제공한다. 실험 결과, 시각 정보 접근이 어려운 사용자에게도 의미 있는 감상 경험을 제공하며, 감상의 주체로서 참여를 유도하는 데 효과적인 것으로 나타났다. 또한 아동, 외국인 등 다양한 사용자군에도 적용 가능성이 확인되었다. 다만, 실제 시각장애인 대상 대규모 실험 및 정량적 효과 분석은 향후 과제로 남아 있으며, 감성 서술의 품질은 언어 모델 성능에 크게 의존하고, 작품의 다양성이나 감상의 주관성에 따라 해설 결과가 달라질 수 있는 한계가 존재한다. 향후 연구는 박물관 전시물, 자연 풍경 등으로의 적용 확대와 함께 감상 이력 기반 맞춤형 서비스, 일상 사물 해설을 포함한 실용적 시각 보조 기능 개발로 확장될 수 있으며, 이를 통해 기술 기반 예술 감상의 새로운 가능성을 제시하고 포용적 문화 접근성 향상에 기여할 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

This work was supported by the SK Telecom's FLY AI Challenger program, conducted in collaboration with the Ministry of Employment and Labor and the Korean Skills Quality Authority as part of the 2024 K-Digital Training.

REFERENCES

[1] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J., "Large Language Models: A Survey," arXiv preprint arXiv:2402.06196, 2024.

[2] Qwen Team, "Qwen2.5: Next-Generation Open-Source Large Language Model," Qwen Blog, March 2025. Available: https://qwenlm.github.io/blog/qwen2.5/

[3] Liu, H., Zhang, Z., Xu, Y., et al., “Visual Instruction Tuning,” arXiv preprint arXiv:2304.08485, Apr. 2023. DOI: 10.48550/arXiv.2304.08485

[4] Park, H. Y., & Kim, H. R., “A Scoping Review on Technology-based Research for Cultural Arts Enjoyment by People with Visual Impairments,” Korean Journal of Visual Impairment, vol. 40, no. 2, pp. 129-155, Apr. 2024. DOI: 10.35154/kjvi.2024.40.2.129. (in Korean)

[5] Park, G. B., & Cho, J. D., & Lee, S. W., “An Analysis of Information Needs in Artwork Appreciation for the Visually Impaired,” Korean Journal of Visual Impairment, vol. 36, no. 4, pp. 23-51, Dec. 2020. DOI: 10.35154/kjvi.2020.36.4.23.(in Korean)

[6] Ryu, J. Y., “Communication in Art Appreciation Education: Focusing on the Analysis of ‘Visual Thinking Strategies (VTS),’” Art Education Review, no. 55, pp. 67-97, 2015. UCI: G704-000621.2015..55.010. (in Korean)

[7] Bachmann, C., “Theory and Practice of Visual Thinking Strategies in Upper Secondary Education,” Forum Oświatowe, vol. 34, no. 1(67), pp. 205-225, 2022. DOI: 10.34862/fo.2022.8.

[8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A., “You Only Look Once: Unified, Real-Time Object Detection,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, Las Vegas, USA, Jun. 2016. DOI: 10.1109/CVPR.2016.91

[9] Roboflow, "Painting Recognition Dataset," Roboflow Universe. Available: https://universe.roboflow.com/capstone-lnujd/painting-recognition.

[10] A. Burambekova and P. Shamoi, "Comparative Analysis of Color Models for Human Perception and Visual Color Difference," presented at the SIST 2025 Conference, Astana, Kazakhstan, June 2024. arXiv:2406.19520. DOI: 10.48550/arXiv.2406.19520.

[11] S. M. Lajevardi and Z. M. Hussain, "Emotion recognition from color facial images based on multilinear image analysis and Log-Gabor filters," 2010 25th International Conference of Image and Vision Computing New Zealand, Queenstown, New Zealand, 2010, pp. 1-6, doi: 10.1109/IVCNZ.2010.6148802.

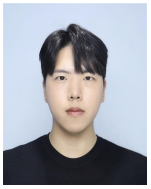
[12] Google Cloud Platform, "The Metropolitan Museum of Art Collection Dataset," BigQuery Public Datasets. Available: https://console.cloud.google.com/marketplace/product/bigquery-public-data/the-met.

Authors

Min-Su Kim received the B.S. degree in Computer Engineering from Daejin University, Korea, in 2024. His research interests include computer vision and large language models, as well as a broad range of interdisciplinary applications.

Min Kim will receive the B.S. degree in Convergence Software from Myongji University, Korea, in 2025. Her research interests include artificial intelligence and data science.

Hyeon-jung Kwak will receive the B.S. degree in Computer Science and Mathematics at Ewha Womans University, Korea, in 2025 Her academic interests lie in the fields of AI and network security. She has participated in multiple research projects exploring the application of deep learning to encrypted traffic analysis and privacy-related challenges.



Ye-jun Choi will receive the B.S. degree in Geoinformatics from Inha University, Korea, in 2026. His research interests include deep learning and its applications. He is currently conducting a graduation project on traffic volume prediction in the context of urban redevelopment.

Hyung-rok Lee received the B.S degree in Sports Science from Korea University and currently a Master's student in the Laboratory of Exercise Medicine, Department Sport Industry Studies, Yonsei University Graduate School.

Chi-wook Ahn will receive the B.S. degree in Military Digital Convergence from Ajou University, Korea, in 2025. His research interests lie in algorithm theory and mathematical artificial intelligence, with a focus on leveraging theoretical approaches to advance AI-driven scientific computing.

Won Joo Lee received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Hanyang University, Korea, in 1989, 1991 and 2004, respectively. Dr. Lee joined the faculty of the Department of Computer Science and Engineering at Inha Technical College, Incheon, Korea, in 2008, where he has served as the Director of the Department of Computer Sc ience and Engineering. He is currently a Professor in the Department of Computer Science and Engineering, Inha Technical College. He has also served as the president of The Korean Society of Computer Information. He is interested in parallel computing, internet and mobile computing, and cloud computing, data science, artificial intelligence.

Young-Bok Cho received the M.S., and Ph.D. degrees in Computer Science from Chungbuk National University, Korea, in 2003 and 2012, respectively. also Dr. Cho received more Ph.D degrees in Medical and Law from Chungbuk National University and Chungnam National University, Korea, in 2019 and 2024, respectively. She has Professor of Information Security at Daejeon University, Daejeon, Korea , in 2018 to 2024, She is currently a Professor in the Computer Education at Gyeongkuk National University, Andong, Korea, in 2024. Her research interests include AI medical image processing, AI security and medical information protection, mobile security.