

생성형 AI를 이용한 인물화 기반의 다양한 도슨트 AI 생성

이혜진¹ · 김태훈^{2*}¹서강대학교 AI·SW대학원 데이터사이언스·인공지능학과 석사과정²서강대학교 메타버스전문대학원 조교수

Creation of Various Painting-Based Docent AI Using Generative AI

Hyejin Lee¹ · Taehoon Kim^{2*}¹Master's Course, Department of Data Science·Artificial Intelligence, Graduate School of AI·SW, Sogang University, Seoul 04107, Korea²Assistant Professor, Metaverse Technology, Graduate School of Metaverse, Sogang University, Seoul 04107, Korea

[요약]

본 연구의 목표는 생성형 AI를 이용하여 인물화에서 연상되는 다양한 인격을 만들어, 다양한 관점에서 그림을 해석하는 도슨트 및 평가자를 만드는 것이다. 작품의 설명을 잘 생성해내는, 인격을 형성할 인물화의 핵심 인격 단어를 잘 뽑아내는 LMM 모델을 선별하기 위해 Rouge, METEOR, BERTScore, Human Evaluation에 기반하여 성능을 비교했다. 비교 결과 Llama-3.2와 Molmo 중에서 상대적으로 내용을 더 풍부하게 생성하고, 그림을 잘 인식하는 Molmo를 LMM 모델로 선정했다. Molmo 모델의 결과를 사용하여 LLM Llama-3.1를 통해, 역할 부여 프롬프트 엔지니어링으로 인격을 조정하고 그림을 설명하게 했다. 해당 구조로 만들어진 페르소나 모델은 앞으로 전시를 구경하러 올 관람객에게는 도슨트가 되어줄 것을, 창작자에게는 평가자의 역할을 할 수 있기를 기대한다.

[Abstract]

The goal of this study is to use generative AI to create diverse personalities inspired by portraits paintings and develop AI-based docents and evaluators who interpret artworks from multiple perspectives. To select an LMM(Large Multimodal Model) capable of generating detailed artwork descriptions and extracting key personality traits from portraits, a performance comparison was conducted using Rouge, METEOR, BERTScore, and Human Evaluation. Among the models tested-Llama-3.2 and Molmo-Molmo was selected as the optimal LMM model due to its ability to generates richer content and accurately recognize visual elements in paintings. Based on the Molmo's outputs, we refined personality traits and employed role assignment prompt engineering using LLM Llama-3.1 to enhance interpretative capabilities. This persona-based AI model is expected to serve as a docent for exhibition visitors and as an evaluator for artists, providing meaningful insights into their work.

색인어 : 생성형 AI, 멀티모달 모델, 도슨트 AI, 평가자 AI, 페르소나 생성**Keyword** : Generative AI, Multimodal Model (LMM), Docent AI, Evaluator AI, Persona Creation<http://dx.doi.org/10.9728/dcs.2025.26.2.287>

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 16 December 2024; Revised 17 January 2025

Accepted 24 January 2025

*Corresponding Author; Taehoon Kim

Tel: 

E-mail: taehoonkim@sogang.ac.kr

I. 서론

과거부터 현재까지 사람들에게 널리 알려진 그림, 즉 명화 속에는 역사와 교훈이 담겨있다. 따라서 명화 감상을 통해 그 안에 숨어있는 교훈과 영감을 얻는 것은 사회에서 인기 있는 문화생활 중 하나다. 특히 명화 감상을 통해 인간다움을 성찰하고, 우울증과 같은 심리적 치유, 삶의 질 문제에 문화 예술의 효과성이 입증되고 있다. IVF(In Vitro Fertilization) 시술받는 난임 여성을 대상으로 한 연구에 따르면 명화 감상을 통해 부정적 감정과 고립감을 해소하는 긍정적인 심리 변화를 보였다[1]. 최근 들어서도 다양한 전시들이 열리고 있으며, 참여 형태가 많아져 대중이 접하기 쉬워졌다. 다만, 아직 관람객으로서는 명화는 이해하기 어렵다는 인식이 남아 있는 것으로 보이며, 명화를 직접 접하기 힘든 환경에 있는 대중들의 접근성 문제는 여전히 해결되지 않았다. 예를 들어, 시간의 제약이 많아 문화를 즐길 새가 없는 사람 또는 거동이 불편하여 공간 제약이 있는 대중에게 말이다. 물론 온라인 전시의 경우 시공간 제약 없이 손쉽게 전시에 접근할 수 있다는 점에서 접근성 문제를 해소할 수 있다. 하지만 이 경우에는 시각적인 요소에만 의지하기 때문에, 명화에 어울리는 향을 전시회에 뿌리거나 다른 촉감 요소를 넣어 참여 형태로 체험할 수 있는 구조를 가질 수 있는 오프라인 전시보다는 재미 요소가 덜 할 수 있기에 직접 작품을 보고 감상하는 것보다 작품에 대한 이해도가 낮아질 우려가 있다. 또한 창작자 관점에서도 전시를 열기 어려운 요소가 많다. 개인 전시를 진행하기 위해서는 공간을 빌리기에 거금의 비용이 필요하며, 좀처럼 피드백을 받을 수 있는 환경을 만들기 어렵다.

이를 해결하기 위해 관람객에게는 도슨트 AI(Artificial Intelligence)를, 창작자에게는 평가자 AI로 사용할 수 있는 명화 페르소나 AI를 통해 명화 자체와 소통하는 새로운 형태의 기술을 도입하고자 한다. 명화의 종류는 다양하지만, 감상자와의 소통을 고려한다면 도슨트 역할을 하는 명화는 인물화가 적합하다. 따라서 명화 페르소나 AI는 사용자가 직접 소통하고자 하는 명화를 고르는 형식으로 구상하였다. 이후 모델이 활용되는 과정을 설명하자면, LMM(Large Multimodal Model)로 선택한 그 인물화에서 주요 단어를 추출한다. 또한 사용자가 설명을 요구하는 그림을 보고 풍부하고 자세한 설명을 출력하게 한다. 그 후 LLM(Large Language Model)을 사용하여 그림을 설명하는 역할 부여 프롬프트를 부여하고, 그림 설명을 입력으로 넣어 형성된 인격으로 입력된 설명을 가공한다. 정리하자면, 구현할 모델의 구조는 LMM과 LLM를 결합하여 명화 페르소나 AI를 생성하는 것이다.

명화 페르소나 AI를 생성하여 활용한다면 전시를 즐기는 관람객에게는 도슨트 AI로서, 실시간으로 자신의 그림에 피드백을 받고 싶어 하는 창작자에게는 평가자 AI로서 재미 요소를 줄 것을 기대한다.

본 논문의 구성은 1장에서는 서론에서 연구 배경 및 연구 목적을 설명하고, 2장에서는 문화 예술에 AI를 접목한 기술

및 생성형 AI에 페르소나를 부여한 선행 연구를 소개할 것이다. 3장에서는 도슨트 AI를 구성할 LMM을 비교하여 선정하는 방법 및 LMM의 결과를 LLM로 연결하는 과정을 설명하고, 4장에서는 직접 개발한 도슨트 AI 활용 방법을 소개할 예정이다. 그 후 5장에서는 논문에서 소개한 모델에 기대 효과와 한계점을 제안하며 향후 연구 과제를 제안할 것이다.

II. 생성형 AI와 예술 융합 동향 분석

생성형 AI와 예술을 접목한 기술은 나날이 발전해나가고 있다. 전시에서는 명화를 영상으로 변형하여 선보이는 미디어 아트부터 최근 들어서는 이미 그림을 설명해주는 도슨트 AI는 상용화 되어 미술관이나 박물관에서 찾아볼 수 있다. 해당 기능을 구현하는 데 필요한 기술은 크게 AI 기술을 접목한 기술과 특정 페르소나를 부여하는 기술 두 갈래로 나뉘어 설명할 수 있다.

2-1 AI 기술을 접목한 전시의 동향

최근에는 이해하기 어려운 작품이나 유물들에 대한 해설을 곁들인 전시 해설가가 인기를 끌고 있다. 다양한 음성을 학습한 AI에 문자를 입력하면 사용자가 듣고 싶어 하는 목소리를 자동으로 생성하는 보이스 클리닝 기술과 결합하여 전시회를 보는 관람객에게 해당 목소리로 작품 설명을 제공한다. 이는 시각장애인의 문화 체험에 이바지가 되는 기술로 돋보인다[2].

아트랩스는 ChatGPT(Generative Pre-trained Transformer) 기반 ‘도슨트 AI’ 챗봇을 관람객들의 예술 관련 질문에 즉각적으로 답변하며, 개인 취향과 관심사에 따라 맞춤형 답변을 제공하도록 만들었다. 현장을 방문하지 않아도, 온라인에서 도슨트를 제공하며 관람객들에게 공간 제약이 없는 문화해설 ‘도슨트 AI’를 연구하고 있다는 전망이다[3].

국내에 있는 국립중앙박물관과 국립현대미술관에서도 문화해설 로봇 ‘큐아이’도 활용되고 있다. 큐아이는 증강현실 기술을 접목해서 사용자, 특히 이동 약자를 따라다니는 동행 서비스를 제공한다. 국립현대미술관에서 산간과 도서 지역에서 큐아이를 통해 전시 작품을 만나 볼 수 있도록 원격 접속 서비스를 시범 운영한다. 큐아이를 통해 시각 약자를 위한 수어 문화해설과 음성 안내 자막, 그림 해설 및 문화해설 서비스로 문화를 즐기기 어려운 사람들을 위한 서비스도 제공되고 있다[4].

작품과 관람객 사이의 단절된 소통, 도슨트 인력 및 체험형 프로그램 부족에 대한 문제점을 해결하기 위해 생성형 AI를 활용한 질문 서비스와 미술관 참여형 시스템을 제안한, 미술관 가이드 AI 도슨트가 연구되었다. 이는 ChatGPT3.5기반으로 연구되었으며, 관람객에게 자동 생성 질문을 통해 관람할 방향성을 제시하고 다양한 방법으로 작품에 대해 고찰할

수 있도록 도와준다. 또한 DALL-E와 함께 이용하여, 관람객의 취향이 담긴 프롬프트를 입력하면, 관람객이 입력한 그림체에 맞게 기존의 그림이 새롭게 생성되어 출력되는 기술을 제공했다[5].

실제 역사 인물을 기반으로 한 페르소나 연구도 자연어 처리를 활용하여 실제 역사 인물 정보를 토대로 대화형 도우미를 설계하여 진행되었다. 역사 인물은 유명한 인물인 레오나르도 다빈치이며, 여러 문헌 자료 내용을 토대로 페르소나를 우월하게 설정하였다. 기존의 박물관이나 미술관에서 제공되고 있는 친절한 도슨트와 달리 따로 현실감 있는 성격을 구현하였다. ‘레오 에이전트’는 지식기반의 시스템으로 구성되며, 사용자가 주시하는 정보 데이터를 해당 작품 정보와 연결한다. 해당 연구는 실제로 레오나르도 다빈치의 페르소나 AI를 구현하였으며, 사용자와의 상호작용을 강화하기 위해 친숙한 스무고개 게임 기능을 통해서 대화할 수 있게 구성되었다. 게임은 작품에 대한 스무고개 형식의 대화 방식을 통해 그에 기반한 작품을 탐색해서 추천되는 시스템을 개발하였다. 향후 과제로는 레오 에이전트가 구체화 될 수 있도록 감정표현 및 TTS(Text To Speech)과 같은 음성 기술과 관련한 추후 과제 필요성을 보였다[6].

홀로그래피 기술을 AI와 융합하여 전시 분야에 도입하자는 연구 방향 제시도 있었다. 해당 연구는 홀로그램과 도슨트 AI를 결합하여 관람객들과 상호작용을 하는 기술 방향을 제시했다. 먼저 음성인식 기술을 통해 관람객의 발화를 인식하고 자동번역 기술을 활용해서 다양한 언어로 소통한다. 추후 위 기술에 이미지 및 비디오 인식 기술을 통합해 활용한다면, 학습한 대화 데이터를 기반으로 주변 환경을 인식하여 관람객들의 질문에 상황에 따라 적절하게 답변할 수 있을 것으로 보인다. 위 기술에 생성형 AI를 활용한다면 작품의 주인공이 되는 역사 인물의 성격, 배경, 사건 등에 맞는 이미지나 비디오를 실시간으로 생성하고, 관람객들의 요청에 반응할 수 있는 디지털 휴먼을 구현할 수 있다. 디지털 휴먼은 관람객은 역사 인물과 마치 대화하듯 소통함으로써 문화 역사에 대한 흥미와 이해를 높일 수 있을 것이라는 의견이 있다[7].

다른 관점으로는 창작자와 AI의 적절한 소통 창구로 누구나 쉬운 창작과 감정적 소통까지 지원하는 AI 창작 플랫폼을 구상했다. 효율적 업무 지향형, 정서적 교류 지향형, 통합 지향형 등 다양한 성격의 페르소나 AI를 만들어 각 사용자가 필요한 도우미를 설정할 수 있게 시나리오를 구축하였으며, 이를 한꺼번에 아우를 수 있는 모델 아이디어를 제안하였다[8].

Zechen Bai는 그림의 다양한 측면에 대한 자세한 설명을 생성하는 다중 주제 및 지식이 풍부한 예술 설명 생성 프레임워크를 제안했다. 이 프레임워크에서 ResNet을 사용해 작품 이미지를 인식했으며, BERT를 사용해 작품 설명 텍스트를 마스킹하여 화가, 년도, 분위기 등을 학습하게 하였다. 결과로 작품에 대한 더 깊이 있는 이해와 접근을 통해 기존의 단순한 이미지 설명 방식을 넘어, 다양한 예술적 측면과 배경지식을 포함한 설명을 제공했다[9].

AI 활용도 방면으로 예술과 결합한 동향을 살펴본 결과, 전시나 문화 예술 분야에 활용할 AI와 결합한 또 다른 새로운 시도가 필요해 보인다. 따라서 본 논문에서는 새로운 시도로 명화를 인식하여 다양한 인격에 따라 그림을 설명 및 평가하는 페르소나를 구축하려고 한다. 도슨트로 사용하는 것뿐만 아니라, 창작자에게 도움이 될 평가 서비스도 제공해줄 수 있는 다양한 인격의 평가자를 심어주려고 한다. 해당 ‘평가자’들은 창작자가 그린 그림을 저마다의 관점으로 해석 및 평가하고 창작자에게 반응을 보여줌으로써 도움을 준다. 자신의 그림을 어떻게 생각하는지, 어떻게 평가하고 있는지에 관한 창작자 수요도 충족할 수 있을 것으로 기대한다. 물론 전시장에서 기존 전시에 맞게 그림을 평가하여 관람객들에게 재미요소를 부여할 수 있는 효과도 기대해볼 수 있다.

2-2 페르소나 부여 연구 동향

Seungju Han 외 6명은 오픈 도메인 챗봇 형태로, 캐릭터의 발화가 포함된 프롬프트를 사용하여 대규모 언어 모델의 힘을 활용하여 응답을 생성하는 PDP(Pseudo Dialog Prompting)이라는 새로운 방법으로 자연스러운 페르소나를 구현하였다. 가상 인물의 발화는 화자의 개인적 특성이나 스타일이 발화에 포함돼서 인물을 모방하는 응답을 생성하는데 유용한 단서를 제공할 수 있다. 또한 전체 대화 내용을 고려하는 것보다 표본 문장을 몇 개 뽑는 것이 인물의 특성을 학습하기에 비용을 최소화했다[10].

Aaron W. Li는 캐릭터의 인간 수준 속성-프로필과 정체성, 그리고 언어 스타일-을 제공하여 페르소나를 개발했다. 논문에서 제안한 ALOHA의 전체 프로세스는 문자 집합이 주어지면 CSM(Character Space Module)-문자 공간 모듈: 문자 간의 유사성으로 캐릭터 성을 묶음-을 사용하여 문자 공간을 결정한다. 다음으로, 특정 대상 문자가 주어지면 CCM(Character Community Module)-캐릭터 구분 목적: 캐릭터와 유사한 그룹과 유사하지 않은 그룹을 정의-을 사용하여 긍정적인 집합과 부정적인 집합을 결정한다. 마지막으로, Facebook ParlAI 프레임워크의 BERT bi-ranker 모델과 Poly-encoder를 사용하여 위에서 결정한 긍정적인 집합과 부정적인 집합을 대화 데이터 세트와 함께 사용하여 대상의 언어 스타일을 복구한다[11]. 이 논문에서 다양한 캐릭터의 언어 스타일을 효과적으로 복구하는 것을 볼 수 있었다.

본 논문에서는 명화 인물화에서 키워드를 뽑아내서, 해당 키워드로 프롬프트를 만들어 간단한 구조로 명화 페르소나를 구현한다. 그 후, 설명해줄 다른 그림의 설명을 참고하여 구현한 명화 인격에 따라, 각자 다른 관점으로 설명해주는 모델을 형성했다.

III. 도슨트 AI 생성 연구 방법

본 연구에서는 인격을 생성하고 싶은 인물 그림과 설명을 원하는 그림을 인풋으로 받을 것이다. LMM을 사용하여 인물화에서 인격 키워드를 뽑는다. 또한 설명을 원하는 그림을 인식하고 설명을 최대한 자세히 생성한다. 이후, LLM으로 두 결과를 결합하여, LLM에 인격을 부여하는 프롬프트를 제공하여 생성된 인격으로 그림을 설명하게 하는 과정을 제안한다.

먼저 선정된 LMM 모델을 활용하여 (1) 설명받을 그림의 풍부한 설명과 (2) 그 그림을 설명해줄 명화 인격을 생성할 주요 단어를 뽑는다. 이후 (3) LLM을 사용하여 명화 인격을 부여한 프롬프트와 그림의 설명을 함께 넣어 각 선택한 인격이 설명하는 새로운 그림 설명을 출력한다. 따라서 명화 설명 기능과 인물화 주요 단어 추출 기능에서 각 그림을 인식할 LMM 모델을 선별 단계가 필요하다.

3-1 성능 지표

성능 지표는 Rouge(Recall-Oriented Understudy for Gisting Evaluation), METEOR(Metric for Evaluation of Translation with Explicit ORdering), BERTScore, Human Evaluation을 사용했다.

1) Rouge

Rouge는 생성 문장 중 정답 문장이 몇 개가 속했는지 알 수 있는 LLM 성능을 확인할 수 있는 지표 중 하나다. Rouge는 n-gram의 Recall에 기반한다. 즉, 정답의 n-gram이 생성한 결과에 얼마나 포함되어 있는지의 비율이다. 정답의 정보를 잘 반영할수록 높은 점수를 나타낸다.

Rouge 안에서도 Rouge1, Rouge2, RougeL이 있다. Rouge1은 각 단어에서의 유사도를, Rouge2에서는 두 단어 연속적인 단어의 유사도를 측정한다. rouge는 문장의 순서를 고려하지 않고, 유연하게 문장 구조를 보며 엄격하게 연속적인 단어 일치를 보지 않는다. 이는 LCS(Longest Common Subsequence) 기반으로 계산되기 때문이다. 본 연구에서는 멀티모달의 선별을 위해 생성한 명화 설명에 실제 명화 설명 안에 있는 단어가 얼마나 들어가 있는지 알 수 있는 ROUGE-1와 문장 구조를 보고 생성한 명화 설명과 실제 명화 설명 문장의 유사성을 유연하게 측정할 수 있는 ROUGE-L을 함께 사용했다. 다만 Rouge 지표만 쓰면 문맥을 고려하지 못했다는 한계점으로 다른 지표도 함께 활용할 것이다.

2) METEOR

METEOR는 BLEU(Bilingual Evaluation Understudy) 지표에서 파생된 평가 지표다. 기본 단위는 문장을 사용하지 않, BLEU 지표와 다른 점이 있다. 후보 문장이 얼마나 정확

히 답안 문장을 포함하는가를 알 수 있는 Precision과 답안 문장을 얼마나 잘 포함하는지 알 수 있는 Recall을 결합한 F-Score를 최종 점수로 사용한다. 따라서 단순 단어보다 어간 추출, 동의어 연결 등 BLEU 보다는 상대적으로 언어학 요소를 고려한 평가 지표라고 볼 수 있다. 0~1의 점수를 산출하며, 더 높은 점수가 더 좋은 성능을 가졌다고 판단하는 지표다. 하지만 METEOR도 단어 순서를 고려하여 판단하는 지표다. 해당 연구에서 뽑는 명화의 설명은 얼마나 더 많이 자세히 설명하는지에 대한, 긴 문장에 대한 고려가 없다. 따라서 이를 또 보완하기 위해, 문장에 따라 유연하게 평가할 수 있는 BERTScore도 함께 활용할 것이다.

3) BERTScore

BERTScore는 BERT라는 사전 학습된 딥러닝 모델을 사용하는 방법으로 Contextual Embedding을 통해서, 모델이 직접 생성한 문장과 실제 문장을 비교하여 의미적 유사성을 고려한 평가 지표다. 단순히 문장과 단어가 같은지 평가하는 BLEU, Rouge 지표보다, 단어 임베딩을 비교하는 BERTScore가 상대적으로 문장 구조와 단어 의미를 파악할 수 있는 평가 지표이다. 예를 들어 ‘Great’, ‘Good’와 같은 비슷한 단어를 비슷하다고 알아들을 수 있어, 이를 감안하고 생성된 문장과 실제 문장을 비교할 수 있는 것이다. 다만 BERTScore도 문장이 얼마나 그림의 설명을 풍부하게 생성했는지에 대한 평가는 고려할 수 없다는 단점이 있다. 마지막으로 인간이 문장을 보고 얼마나 풍부하게 문장을 생성했는지 직접 판단하는 Human Evaluation도 함께 활용할 것이다.

3-2 모델 선정 및 활용

작품 설명 기능 과정에서는 얼마나 자세히 그림을 인식하고 설명할 수 있는가를 주목하고, 인물화 주요 단어 추출 기능 과정에서는 얼마나 그림에서 인격 주요 단어를 포착할 것인가에 주목한다. 이에 두 가지 오픈소스 LMM인 Llama-3.2-11B-Vision-Instruct와 Molmo-7B-D를 활용하여 비교 후 성능이 좋은 모델을 사용했다. 이때, 본 논문에서는 두 모델의 성능을 비교하기 위한 기준으로 ArtEmis Dataset[12]을 사용했다. 해당 데이터셋에는 WikiArt의 80,000개 작품에 대한 감정 범주와 설명이 455,000개 포함되어 있다.

작품 설명 기능 과정은 해당 Dataset을 전부 사용하지 않고, 97장의 작품을 선정하여 두 모델로 각 그림의 설명을 생성하게 했다. 얼마나 정확하고 풍부하게 작품 설명을 뽑아냈냐를 확인하기 위하여, LMM이 생성한 문장과 사람이 실제로 작품을 설명한 문장을 함께 비교했다.

인물화 주요 단어 추출 기능 과정은 35장의 인물화를 선정하여, 두 모델을 통해 주요 인격 단어를 추출하게 했다. 얼마나 인물화에서 해당 인격 및 감정을 잘 추출했는가를 실제 감

정 범주 및 설명 문장과 출력된 주요 단어를 비교했다.

1) 작품 설명 기능 과정

표 1. Llama3.2와 Molmo의 성능 평가 비교
Table 1. Llama3.2 and Molmo performance evaluation metric

Metric	Llama&1	Llama&2	Molmo&1	Molmo&2
Rouge-1	0.28	0.28	0.29	0.30
Rouge-L	0.17	0.17	0.16	0.17
METEOR	0.22	0.23	0.23	0.24
BERTScore	0.26	0.23	0.26	0.27
Human Evaluation	3	4	2	1

표 1은 작품 설명에 사용할 LMM을 선정하기 위해 3-1에서 설명한 지표들로 Llama와 Molmo에 대한 성능을 공통적인 프롬프트를 활용하여 비교한 표다. 각 모델 뒤에 붙은 1, 2는 공통 프롬프트 1, 공통 프롬프트 2에 관한 숫자 인덱스 정보 사항이다. 공통 프롬프트 1은 “You are a docent who explains famous paintings in English.

Please look at the following famous painting and explain it clearly, using the atmosphere, painting style, emotional lines, object, and behavioral standards of this famous painting.

There is no need to explain what the name of the famous painting is. However, please clearly explain the objects, people, and emotions seen in the famous painting. Let's look carefully this painting!”을 사용했다. 해당 프롬프트는 도슨트 역할을 부여했으며, 작품이 무엇인지 설명할 필요 없지만, 입력된 작품에서 보이는 객체, 인물, 감정을 뽑아달라고 요청했다. 마지막에 ‘Let’s look carefully this painting!’을 넣어 CoT(Chain of Thought)처럼, ‘Let’s think step by step!’과 비슷한 성능을 내줄 것이라 기대했다.

비교 결과 Llama를 통해 생성된 설명을 확인해보면 작가 이름과 작품 이름이 맞지 않게 출력되는 설명도 있었다. 예를 들어 작품명 ‘The Burning of the Houses of Lords and Commons’, 즉 마을이 불에 휩싸인 광경을 묘사한 작품을 바다 중간의 희망을 상징하는 등대라고 표현하는 잘못된 문장 생성이 보였다. 반면 Molmo는 불타는 황폐한 마을이라고 제대로 그림의 내용을 포착했지만, ‘The painting, Fire in the City’라고 작품명을 생성하는 모습을 보였다. 인간이 볼 때는 그림 설명 부분에서는 Molmo가 Llama보다 낫지만, 잘못된 작품명 생성하는 등의 할루시네이션을 방지하기 위한 수정된 프롬프트가 필요하다고 판단했다.

수정된 공통 프롬프트 2는 “You are a docent who explains famous paintings in English. Instead of describing what you imagine, describe only what you can confidently determine from the image. Do not



그림 1. Molmo를 이용한 작품의 설명 추출 도식화
Fig. 1. Schematic diagram for extracting descriptions of works using Molmo

itemize the content in a list format. Do not mention the name of the author or work. Instead, please clearly explain the objects, people, and emotions seen in paintings.”를 사용했다. 그림을 보고 작품명이나 작가를 언급하지 말라는 조건을 걸었으며, 자신 있게 설명할 수 있는 것만 설명해달라고 요청했다.

공통 프롬프트 2로 다시 비교한 결과 Llama는 프롬프트 조건을 잘 반영하지 못했는지 ‘The painting, The Burning of the Houses of Parliament by J.M.W.’라는 잘못된 작품명과 작가를 생성했다. 그에 비해 Molmo는 작품명은 언급하지 않고 제대로 불타고 있는 마을에 관한 설명을 생성했다. 이외에도 그림 설명 과제는 표 1에서 알 수 있듯 수치적인 성능 차이를 보면 Rouge1, RougeL, METEOR, BERTScore 모두 Molmo와 프롬프트 2를 사용한 것이 상대적으로 성능이 좋았다. 따라서 그림 1처럼 작품 설명 기능의 LMM은 프롬프트 2와 Molmo 7B를 활용했다.

사용자가 작품을 입력하면 생성할 인격이 설명할 기본적인 작품 설명이 존재해야 한다. Molmo를 사용하여 ‘Summer Afternoon’이라는 작품을 입력했을 때 도출된 결과로서의 설명은 풍부한 편이었다. 해당 작품은 시골의 여름 오후를 나타낸 작품이다.

실제 작품 설명과 비교하자면, 해당 작품의 실제 설명은 “It reminds me of old western's I used to watch in my youth. The people enjoy a stroll under the shade of the large trees.”로 정겨운 마을, 큰 나무들을 언급하며 어린 시절을 상기시킨다는 묘사를 했다. Molmo의 생성된 작품 설명은 “The painting depicts a serene rural scene with a dirt road leading to a small house. On the left side of the road, a large tree stands prominently, its branches extending over the path. The house in the background has a thatched roof and is surrounded by trees, creating a cozy atmosphere. Along the dirt road, several figures can be seen. A woman in a long dress is walking towards the house, while two children are running towards her. Further down the road, two more people are visible, though their details are less clear due to the distance. The overall composition conveys a sense of tranquility and simplicity. The warm colors of the earthy tones and the green of the trees create a harmonious palette. The scene captures a moment of

everyday life in a rural setting, evoking feelings of nostalgia and connection to nature.”와 같이 보다 풍부하게 작품을 설명했다.

해당 문장은 작품에서 실제 문장과 같이 큰 나무들 객체를 탐지했고, 시골의 비탈길을 표현하여 시골임을 인지했으며, 아이들과 여성을 언급하며 행동 표현도 잘 묘사했다. 특히나 도슨트처럼 색감이나 조화를 언급하며 따뜻한 빛의 색감도 함께 언급함으로 그림을 담백하게 설명했다.

2) 인물화 주요 단어 추출 기능 과정

인물화를 인식하고 그것을 기반으로 인격을 생성하기 위해서는 각 인격 주요 단어 선별이 중요하다. 페르소나를 생성하기 위한 인격 요소는 많지만 본 연구에서는 인물화에서 인격 생성 시 갖춰야 하는 필수 요소인 ‘외관’, ‘성격’, 그리고 ‘감정’ 단어를 추출했다. 감정은 그림을 평가할 페르소나지만 인격에 따라 그림 설명을 다르게 하려고 한다면 중요한 요소라고 생각했다. 사람은 감정에 따라 보는 것이 달라지며, 설명하는 말투도 달라지기 때문이다. 이 사실을 바탕으로 앞으로 생성할 도슨트 AI 인격에도 감정을 입력하는 것으로 구상했다.

인격 단어 생성을 위해 공통 인물화 및 공통 프롬프트로 Llama와 Molmo를 비교했으며, 실제 사람들이 평가한 내용을 바탕으로 형성된 지표인 Human Evaluation을 중심으로 모델을 선별했다. 해당 과제는 풍부한 그림 설명이 아니라, 인격 형성에 관련되는 핵심 단어를 뽑는 것이 중점이기 때문에 작품 설명 기능 과정에서 다른 프롬프트와 다른 공통 프롬프트로 재설계했다. 공통 프롬프트 1은 “You are the docent who looks at the details of the picture and picks out key keywords.

Instead of describing what you imagine, describe only what you can confidently determine from the image. Please select only 3 keywords that make you feel when looking at the following work: people, emotions, and personality.

For example, person: “young girl”, emotion: “timid”, personality: “introvert”. There is no need to explain the picture. As in the example, please select only 3 keywords. Keywords:”를 사용했다. 작품 설명 기능 과정의 프롬프트에서 사용한 것처럼 상상해서 뽑지 말고 보이는 그대로 뽑게 했지만, 대신 역할을 부여할 때 ‘keyword’를 뽑는 것을 주요 역할로 설정했다. 예시를 한 개를 입력하는 one-shot 프롬프트도 함께 사용했다.

비교 결과 Llama의 경우 인물화 하나를 뽑는데 Person을 People로 단어를 번갈아 적용하여 결과를 뽑기도 했으며, 아예 인물화를 설명하는 문장을 뽑기도 했다. Molmo는 Llama와 다르게 주어진 예시를 참고하여 조건을 준 양식을 지켜서 결과를 출력했다. 하지만 한 번 실험하여 결과를 비교하는 것으로 LMM의 성능을 확인하기에는 부족하다. 따라서 프롬프

트를 수정하여 다시 결과를 비교해보는 시도를 거쳤다.

공통 프롬프트 2는 “You are the docent who looks at the details of the picture and picks out key keywords.

Pick keywords only what you can confidently determine from the image. For example, person: “young girl”, emotion: “timid”, personality: “introvert”.

Do not mention the name of the author or work. Instead, Please clearly select 'person', 'emotion', 'personality' only 3 keywords. Keywords:”를 사용했다. 프롬프트 1과는 다르게 작품명과 작가를 설명하지 말라고 당부했고, 깔끔하게 3가지 주요 단어만 선정해달라고 요청했다.

다시 비교한 결과 Llama와 Molmo 둘 다 작품명 및 작가 이름을 언급하지 않았으며 그저 3가지 주요 단어만 출력했다. 프롬프트 2를 활용하는 게 프롬프트 1을 활용하는 것보다 좀 더 원하는 결과물을 출력시키기에 적합했다.

다만 외관 자체를 인식하는 부분에서 Llama는 ‘woman’ 혹은 ‘man’으로 키워드가 생성되는 반면, Molmo는 ‘middle aged woman’ 혹은 ‘middle aged man’으로 세부적으로 더 잘 분류하는 경향을 보였다.

또한 인물화의 인격을 그저 소심한, 자신 있는 등 간단한 단어를 생성하는 Llama와 달리, Molmo는 전체적으로 책벌레, 잔인한 군인 등 더 깊이 있고 다양한 답변을 보였다.

감정을 추출한 단어도 Llama는 19세기에 나온 ‘the-laundress’ 작품의 그 시절 고된 작업에 화난 세탁부 인물화의 감정선을 ‘calm’으로 인식하는 등의 엇갈린 키워드를 생성한다. 반면에 Molmo는 Llama보다 상대적으로 명확히 ‘angry’를 감지해 출력했다. 따라서 인물화에서 주요 단어를 추출하는 기능에도 LMM은 그림 2와 같이 Molmo를 사용했다.

예시로 Molmo에 인물화 ‘head-of-red-headed-man-with-yellow-cap’을 입력하면 세 가지 키워드는 인격은 ‘middle-aged man’, 성격은 ‘contemplative’, 감정은 ‘contentment’를 출력했다. ArtEmis Dataset에 의하면 그림 설명에 ‘The man almost seems asleep and I appreciate the way the artist shows his pink cheeks.’ 실제 감정 태그는 ‘sadness’, ‘fear’, ‘contentment’가 있다. 인격이 ‘man’이며, 감정이 ‘contentment’인 것을 보면 인물화의 인격과 감정은 실제와 비슷하게 생성해낸 것을 확인할 수 있다.

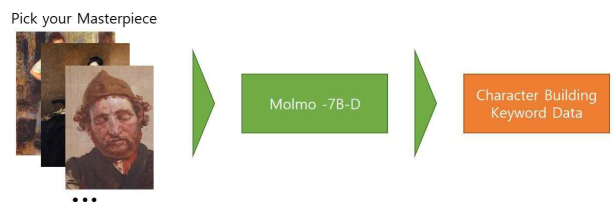


그림 2. Molmo를 이용한 인물화의 핵심 단어 추출 도식화
 Fig. 2. Schematic diagram for extracting key words from portraits using Molmo

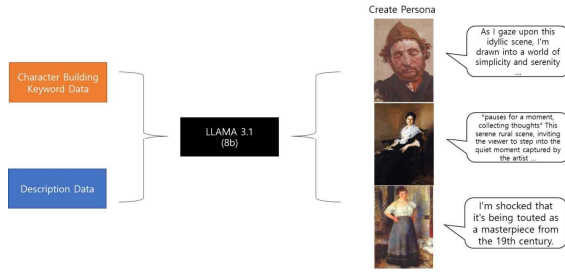


그림 3. Llama 3.1을 이용한 인격 생성 및 그림 설명 도식화
 Fig. 3. Persona generation and picture description diagram using Llama 3.1

3) 인격 생성 및 그림 설명

마지막으로 인격 생성 및 그림 설명 단계는 그림 3처럼 앞서 설명한 두 단계에서 나온 결과를 ‘Llama-3.1-8B-Instruct’라는 LLM 모델로 결합하여 새로운 결과를 출력하는 단계다. 페르소나를 생성하는 프롬프트 템플릿은 “role”: “system”, “content”: “You are a {외관}. And your personality is {성격}. Please look at the given picture's description and create a sentence that {감정} evaluates the work.”를 사용했다. 외관, 감정, 성격은 1) 단계에서 Molmo가 추출한 인물화의 주요 인격 단어로 설정한다. 이어서 “role”: “user”, “content”: “given description: {그림 설명}”을 설정하여 2)에서 생성한 그림 설명을 두 번째 프롬프트로 넣는다.

위 과정을 수행할 때 선정 모델이 인물화마다 각각 다른 감상 결과를 출력하는 것을 확인할 수 있다.

IV. 실험결과

제3장 2절에서 설명한 실험 과정을 이용해 페르소나 인격을 생성하여 결과를 출력할 수 있다. 사용자에게 여러 인물화를 후보에 두고 해당 인물을 고르면, 설명을 원하는 작품을 넣어 활용해볼 수 있다. 사용자는 인물에 따라 다양한 결과물을 볼 수 있다.

실험 결과에서 사용된 그림은 저작권 만료된 작품을 사용했으며, WikiArt에서 가져왔으며, 1 저자가 직접 그린 그림도 함께 사용했다. 총 97개의 그림과 35개의 인물화를 활용하여 실험을 진행했다. 오픈소스인 Molmo-7B-D와 Llama 3.1-8B를 사용했기에 비용적인 부담은 들지 않았다. 해당 서비스를 제공하기 위해서 사용자를 위한 인터페이스는 Gradio를 활용했으며 인터페이스는 그림 4와 같다.

첫 번째 이미지 입력 단계에서는 사용자가 원하는 인물화, 혹은 인물 그림이면 해당 그림의 인격 추출이 가능하다. 두 번째 이미지 입력은 설명받을 그림을 입력한다. 이미지를 선택하고 ‘Please explain the picture for me!’ 회색 버튼을 클릭하면, 사용자가 입력한 이미지를 토대로 ‘Your Partner

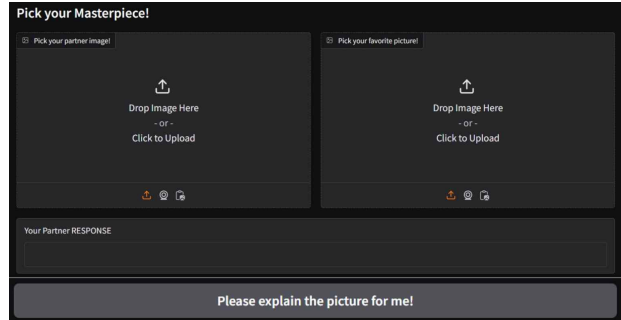


그림 4. Gradio를 이용한 기본 인터페이스
 Fig. 4. Basic interface using Gradio

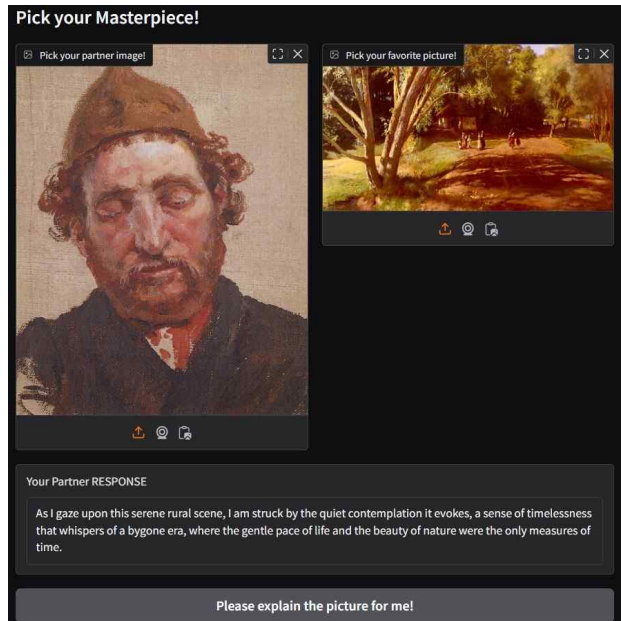


그림 5. 명상적인 중년 남자 페르소나 AI를 이용한 그림 설명
 Fig. 5. Picture description using contemplative middle-aged man persona AI

RESPONSE’ 텍스트 박스에 생성된 페르소나 AI가 두 번째 이미지를 설명해주는 결과가 출력된다.

그림 5는 관람객 관점에서 도슨트 AI를 활용할 때의 예시이다. 첫 번째 이미지에 입력한 인물화는 ‘head-of-red-headed-man-with-yellow-cap’이라는 명화로 우울하고 침잠되어 보이는 중년 남자 그림이며, 두 번째 이미지에 설명 받을 작품은 시골 여름 오후 풍경을 그린 ‘summer-afternoon’이다. 그림 5의 출력값을 봐도 알 수 있듯 해당 페르소나 AI는 풍경의 고요함과 자연의 아름다움에 조용히 감탄하는 그림 설명을 생성했다. 사람마다 주관적이지만, 보이는 인격처럼 차분하고 담백하게 문장을 잇는 결과가 생성됐다. 이와 같은 평가가 가능한 이유는 같은 풍경화에 대하여 다른 인물화 기반 도슨트의 설명이 전혀 다르게 제시되기 때문이다. 그림 6은 다른 인물화 ‘the-laundress’ 그림을 입력하여, 그림 5에서 입력했던 공통된 ‘summer-afternoon’ 작품을 설명하게 한 결과이다. 해당 페르소나 AI는 고요한 시골

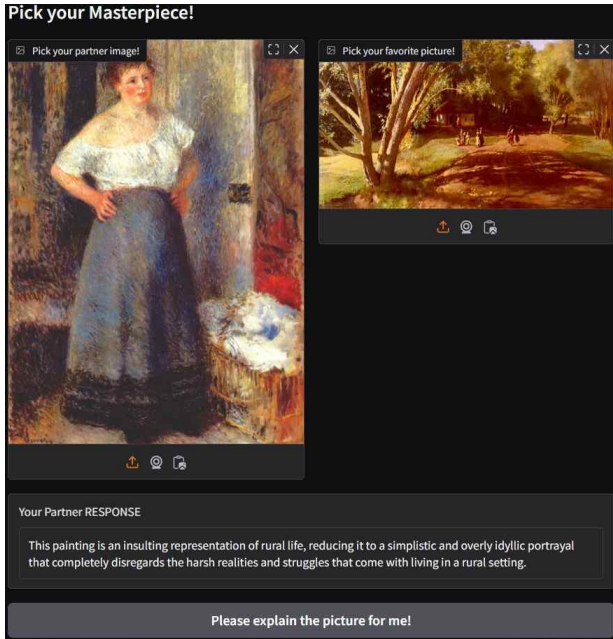


그림 6. 화난 여성 페르소나 AI를 이용한 그림 설명
 Fig. 6. Picture description using angry woman persona AI

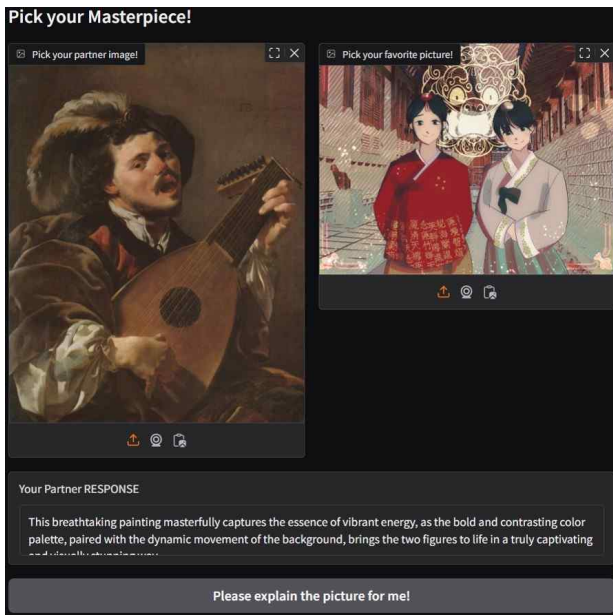


그림 7. 즐거운 남자 페르소나 AI를 이용한 그림 설명
 Fig. 7. Picture description using joyful man Persona AI

풍경 장면을 시골 생활에 따르는 가혹한 현실과 어려움을 무시하고 단순하고 지나치게 목가적인 묘사로 축소했다며 시골 생활을 모욕적으로 표현했다며 비난하는 설명을 생성했다. 입력된 인물화 그림은 다소 화난 모습의 여성으로 그림을 상대적으로 부정적으로 보는 경향이 엇보인다.

그림 7은 앞의 예시와는 다르게 창작자들이 자신의 그림을 평가받기 위해 평가자 AI로 활용한 예시이다. 첫 번째 이미지 입력은 인물화 ‘the-singing-lute-player-1624’ 노래하고

있는 남자를 입력하고, 두 번째 이미지 입력은 1 저자가 직접 그린 한복을 입고 있는 두 여자의 그림을 입력했다. 출력값을 확인하면, “This breathtaking painting masterfully captures the essence of vibrant energy, as the bold and contrasting color palette, paired with the dynamic movement of the background, brings the two figures to life in a truly captivating and visually stunning way.”이다. 즐거운 남자 페르소나 AI는 다른 도슨트와 다르게 과장된 형용사를 사용하여 그림을 설명하는 경향을 보였다. 두 인물을 포착함에 이어서, 색상 팔레트가 대담하고, 훌륭하다고 평가했다. 해당 즐거운 남자 페르소나 AI는 감정이 즐거움이기 때문에 다소 들뜬 평가를 출력하는 것을 볼 수 있었다.

V. 결론

본 연구는 LMM Molmo의 작품 설명과 인물화 주요 단어 결과를 이용하여, LLM Llama로 병합하여 다양한 인격을 부여하여 인격별로 작품 설명을 달리하는 명화 페르소나 AI를 구현하였다.

직접 구현한 모델을 사용하면 전시를 감상하는 관람객들에게 또 다른 명화의 매력을 돋보일 수 있는 기회를 제공할 수 있다. 위 모델을 경량화한다면 온라인 전시 외에도 오프라인 전시에서 전시할 작품들을 선정해서 관람객들이 충분히 명화를 둘러보다가 즉석에서 사용할 수 있는 휴대형 도슨트 AI로 사용할 수 있을 것으로 보인다.

또한 창작자들은 다양한 페르소나 모델을 적용하여 각각 다른 관점에서 자신의 그림을 평가받을 수 있다. 입력하는 인물화에 따라서 다양한 관점의 평가를 받을 수 있으므로, 예상치 못했던 추가 의견이나 발견하지 못했던 관점을 확인할 가능성이 있다. 자신의 그림을 전시하기 전에 해당 모델에 자신의 그림을 입력하면, 전시 중이나 후에 나올 평가 사항을 미리 받아서 전시 구상의 측면에서 도움을 받을 수도 있다. 외에도 그림을 그리는 중 미완성 그림을 입력하면 필요할 보완 사항을 받아보거나, 전시를 하기 전에 자신의 그림에 대해 간단한 설명을 적을 때 그림 설명을 생성해낼 수 있도록 도움을 주는 방향으로 해당 모델의 사용성을 생각해볼 수 있다.

만일 후보 인물화가 100장이 있다면 관람객은 100명의 도슨트를, 창작자는 100명의 평가자를 만날 수 있다. 해당 구조로 만들어진 명화 페르소나 모델은 앞으로 전시를 구경하려는 관람객에게는 도슨트가 되어줄 것이며, 창작자에게는 평가자의 역할을 할 수 있다. 해당 페르소나 모델을 통해서 그림을 즐기는 몰입 요소, 그리고 창작자에게 평가자 역할을 해줄 도우미로 활용될 수 있기를 기대한다.

추후 연구로 해당 페르소나 AI가 목소리 기능까지 탑재하는 것을 연구한다면 그림을 직접 보지 못하는 시각 장애인에게도 청각적으로 실감 나게 그림을 표현할 수 있을 것이다.

그리고 적절한 프롬프트와 향상된 LMM 모델을 사용하여 감정을 더 잘 인식할 수 있다면 입체적인 페르소나 AI도 생성할 가능성이 보인다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 메타버스 융합대학원의 연구(IITP-2023-RS-2022-00156 318)와 문화체육관광부 및 한국콘텐츠진흥원의 2023년도 문화기술 연구개발사업(RS-2023-00219237)으로 수행된 연구로서, 관계부처에 감사드립니다.

참고문헌

[1] H.-R. Oh and Y.-J. Chung, "The Case Study of Group Tele-Art Therapy through Masterpiece Appreciation for Women Undergoing IVF Treatment," *Korean Journal of Art Therapy*, Vol. 31, No. 4, pp. 891-914, August 2024. <https://doi.org/10.35594/kata.2024.31.4.004>

[2] ZDNet Korea. CJ OliveNetworks Uses AI to Explain Modern Artworks [Internet]. Available: <https://zdnet.co.kr/view/?no=20221004102657>.

[3] NewsWire. ArtLamp Launches 'AI Docent' Chatbot Service [Internet]. Available: <https://www.newswire.co.kr/newsRead.php?no=981945>.

[4] Ministry of Culture, Sports and Tourism. Cultural Commentary Robot 'QAI' Guides Modern Artworks in Remote Areas [Internet]. Available: http://www.mcst.go.kr/kor/s_notice/press/pressView.jsp?pSeq=20833.

[5] H. Byeon and K. Choi, "Art Museum Guide AI Docent, Using Generative AI Technology," in *Proceedings of the 2023 KIIT Autumn Conference*, Jeju, pp. 869-872, November 2023.

[6] D.-K. Lee and K.-H. Lee, "Scenario-Based AI Agent Design to Enhance the User Experience," *Journal of Digital Contents Society*, Vol. 24, No. 5, pp. 1021-1028, May 2023. <https://doi.org/10.9728/dcs.2023.24.5.1021>

[7] S.-H. Wi and H.-K. Ahn, "Analysis of AI Docents Using Holography Technology," *The Journal of the Korea Contents Association*, Vol. 23, No. 12, pp. 46-57, December 2023. <https://doi.org/10.5392/JKCA.2023.23.12.046>

[8] H. H. Gu, S. E. Choi, S. H. Seo, R. E. Jeong, and Y. R. Koo, "A Convergence Study on Future Vision Scenarios for the Coexistence of Creators and AI -Focusing on Co-Design Workshop Based on the Backcasting Perspective-," *The Korean Society of Science & Art*, Vol. 40, No. 4, pp. 13-32, September 2022. <https://doi.org/10.17548/ksaf.2022.09.30.13>

[9] Z. Bai, Y. Nakashima, and N. Garcia, "Explain Me the

Painting: Multi-Topic Knowledgeable Art Description Generation," in *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, pp. 5402-5412, October 2021. <https://doi.org/10.1109/ICCV48922.2021.00537>

[10] S. Han, B. Kim, J. Y. Yoo, S. Seo, S. Kim, E. Erdenee, and B. Chang, "Meet Your Favorite Character: Open-Domain Chatbot Mimicking Fictional Characters with Only a Few Utterances," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle: WA, pp. 5114-5132, July 2022. <https://doi.org/10.18653/v1/2022.naacl-main.377>

[11] A. W. Li, V. Jiang, S. Y. Feng, J. Sprague, W. Zhou, and J. Hoey, "ALOHA: Artificial Learning of Human Attributes for Dialogue Agents," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, New York: NY, pp. 8155-8163, February 2020. <https://doi.org/10.1609/aaai.v34i05.6328>

[12] C. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas, "ArtEmis: Affective Language for Visual Art," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, Nashville: TN, pp. 11569-11579, June 2021. <https://doi.org/10.48550/arXiv.2101.07396>



이혜진(Hyejin Lee)

2018년 : 평택대학교 (이학사 & 공학사, 데이터정보학 & 정보통신학)

2024년 : 서강대학교 AISW대학원 (데이터사이언스·인공지능학)

2024년~현재 : 서강대학교 AI-SW대학원 데이터사이언스·인공지능학과 석사과정

※ 관심분야 : 멀티모달 인공지능 (Multimodal AI), 생성형 인공지능 (Generative AI), 머신러닝 (Machine Learning)



김태훈(Taehoon Kim)

2018년 : 서강대학교 (공학사 & 문학사, 컴퓨터공학 & 신문방송학)

2021년 : 서강대학교 대학원 (공학박사, 컴퓨터공학)

2021년~2024년: LG AI 연구원

2024년~현재 : 서강대학교 메타버스전문대학원 조교수

※ 관심분야 : 멀티모달 인공지능 (Multimodal AI), 컴퓨터 비전 (Computer Vision), 생성형 인공지능 (Generative AI)