일반논문 (Regular Paper)
방송공학회논문지 제31권 제2호, 2026년 3월 (JBE Vol.31, No.2, March 2026)
https://doi.org/10.5909/JBE.2026.31.2.333
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

생성형 AI 영화 제작을 위한 노드 기반 다중 시점 이미지 생성 파이프라인 설계 및 일관성 분석

권 오 주[a], 구 민 서[b], 서 제 웅[b], 김 에스더[c], 김 기 범[c], 박 구 만[d], 이 영 화[e], 최 영 희[c]‡

Design and Consistency Analysis of a Node-Based Multi-View Image Generation Pipeline for Generative AI Movie Production

O ju Kwon[a], Minseo Koo[b], Jeung Seo[b], Esther Kim[c], Kibeom Kim[c], Gooman Park[d], Younghwa Lee[e], and Younghee Choi[c]‡

요 약

최근 생성형 인공지능은 영화 제작 전 공정에 걸쳐 혁신을 일으키고 있으나, 기존 상용 SaaS 기반 도구는 단일 이미지의 품질 대비 다중 시점에서의 인물 및 객체 일관성 유지에 구조적 한계를 보인다. 본 연구는 이러한 한계를 극복하기 위해 Qwen 계열 텍스트-투-이미지 모델을 기반으로 LoRA 적응 기법과 노드 기반 워크플로우를 결합한 다중 시점 생성 파이프라인을 제안한다. 제안된 파이프라인은 카메라의 이동과 회전을 파라미터화하여 명시적으로 제어하며, Multi-Angle LoRA를 통해 인물의 정체성을 보존하면서 다양한 구도 변환이 가능하도록 최적화되었다. 실험 결과, 정성적 평가에서 다양한 앵글 변화에도 시각적 특성이 안정적으로 유지됨을 확인하였으며, CLIP 이미지 인코더 기반 정량 분석에서 평균 코사인 유사도 0.8764를 기록하여 높은 의미적 일관성을 검증하였다. 본 연구는 정밀한 장면 설계가 요구되는 제작 환경에서 노드 기반의 통제 중심 생성 구조가 상용 도구의 보완적으로 대안이 될 수 있음을 시사한다.

Abstract

Recent generative AI technologies are rapidly transforming film production across pre-production, production, and post-production stages. However, while existing commercial SaaS-based generation tools provide high quality for single images, they exhibit structural limitations in maintaining consistency of characters and objects across diverse camera perspectives. This study proposes a multi-angle generation pipeline that integrates LoRA (Low-Rank Adaptation) techniques and node-based workflows based on the Qwen series text-to-image model to overcome these constraints. The proposed architecture is designed to explicitly control camera movement and rotation through parameterization and is optimized for various composition changes while preserving character identity via Multi-Angle LoRA. Experimental results confirmed through qualitative evaluation that core visual characteristics remain stable despite various camera angle changes. Furthermore, quantitative analysis using a CLIP image encoder-based approach recorded an average cosine similarity of 0.8764, verifying semantic consistency. These results suggest that in production environments requiring precise scene design and multi-perspective control, a control-oriented generation structure such as a node-based workflow is necessary as a complementary alternative to commercial SaaS tools.

Keyword : Generative AI, LoRA, ComfyUI, Staying Consistent, Cosine similarity

Ⅰ. 서 론

최근 생성형 인공지능 기술의 급격한 발전은 시나리오 창작, 콘셉트 아트 제작, 프리비주얼라이제이션에 이르기까지 영화 제작 전반의 공정 효율화를 가속화하고 있다. 특히 텍스트-투-이미지 및 텍스트-투-비디오 기반 생성 모델은 초기 기획 단계에서 창작자의 아이디어를 신속하게 시각화할 수 있는 도구로 활용되며, 콘텐츠 제작 환경에 실질적인 변화를 가져오고 있다. 이러한 흐름 속에서 시각적인 스타일과 세계관의 일관성이 중요한 영화 제작 분야에서는 생성형 AI의 활용 가능성이 더욱 주목받고 있다[1].

현재 미드저니(Midjourney), 나노 바나나 프로(Nano Banana Pro)와 같은 SaaS(Software as a Service) 기반 이미지 생성형 AI(Generative AI) 모델들은 직관적인 사용자 인터페이스와 자동화된 파이프라인을 통해 고품질 이미지를 손쉽게 생성할 수 있도록 지원한다. 그러나 영화 제작 환경에서는 단일 이미지의 품질 확보만으로는 충분하지 않다. 실제 제작 과정에서는 동일 인물과 객체가 다양한 카메라 시점과 장면 변화 속에서도 일관되게 유지되어야 하며, 시간적 흐름에 따른 서사 전개 속에서도 시각적 정합성이 확보되어야 한다. 또한 특정 객체의 색상, 재질, 위치와 같은 속성을 선택적으로 수정하는 정밀 제어 기능은 반복적인 장면 구성 과정에서 필수적으로 요구된다[2].

그러나 SaaS 기반 이미지 생성형 AI 모델들은 내부 구조 및 파라미터 접근이 제한되는 구조를 갖는 경우가 많아, 동일 조건에서의 재현성 확보나 세부 연출 요소에 대한 정밀 제어에 제약이 존재한다. 이로 인해 장면 간 인물의 형태가 미세하게 변화하거나, 특정 속성만을 수정하고자 할 때 전체 장면이 재구성되는 문제가 발생할 수 있다.

이에 본 연구는 생성형 AI 영화 제작을 위한 노드 기반 다중 시점 이미지 생성 파이프라인을 ComfyUI로 설계하고, 이를 통해 통제 가능한 생성 구조를 제안한다[3]. 제안하는 파이프라인은 모듈화된 노드 구조를 기반으로 파라미터 및 조건을 반복적으로 조정할 수 있도록 구성되며, Qwen 계열의 텍스트-투-이미지 생성 모델을 기반 모델로 활용하여 다중 시점 환경에서의 구조적 일관성 확보를 목표로 한다. 또한 LoRA(Low-Rank Adaptation) 기반 적응 기법을 결합함으로써 스타일 및 캐릭터 특성을 유지한 상태에서 다양한 카메라 시점 간 시각적 정합성을 강화하도록 하고 빠르게 이미지 추론이 가능하도록 한다. 이를 위해 본 연구는 다음과 같은 핵심 문제를 설정한다.

첫째, 기존 SaaS 기반 이미지 생성 방식은 다중 시점 변화 상황에서 동일 인물과 장면 요소의 시각적 일관성을 구조적으로 유지할 수 있는가?

둘째, 노드 기반 워크플로우와 LoRA 적응 기법을 결합한 통제 중심 생성 구조는 카메라 시점 변화를 명시적으로 제어하면서도 인물 정체성과 장면 분위기의 일관성을 보다 안정적으로 확보할 수 있는가의 질문이다.

따라서 본 논문은 제안된 파이프라인의 다중 시점 일관성을 정성적 평가 및 정량적 지표를 통해 분석함으로써, 영화 제작 환경에서의 적용 가능성을 검증하고자 한다.

Ⅱ. 선행연구 및 기술 배경

1. Reference-based Generation(IP-Adapter)

텍스트 기반 확산 모델의 한계를 보완하기 위해, 참조 이미지를 조건으로 활용하는 Reference-based Generation 기법이 제안되었다. 대표적으로 IP-Adapter(Image Prompt Adapter)가 있으며, 이는 기존 텍스트 조건 외에 이미지 임베딩을 추가적으로 주입함으로써 생성 이미지의 시각적 일

a) 서울과학기술대학교 정보통신미디어공학과(Department of Information and Communication Media Design Engineering of Seoul National University of Science and Technology)
b) 서울과학기술대학교 전자IT미디어공학과(Department of Electronic IT Media Engineering, Seoul National University of Science and Technology)
c) 서울과학기술대학교 문예창작학과(Department of Creative Writing of Seoul National University of Science and Technology)
d) 서울과학기술대학교 스마트ICT융합공학과(Department of Smart ICT Convergence Engineering, of Seoul National University of Science and Technology)
e) 시그마케이주식회사(SigmaK Co., Ltd.)
‡ Corresponding Author : 최영희(Younghee Choi)
E-mail: chanwch@seoultech.ac.kr
Tel: +82-2-970-6273
ORCID: https://orcid.org/0009-0004-2286-5029
※ 이 연구는 서울과학기술대학교 교내 일반과제 연구비 지원으로 수행되었습니다.
· Manuscript February 24, 2026; Revised March 12, 2026; Accepted March 13, 2026.

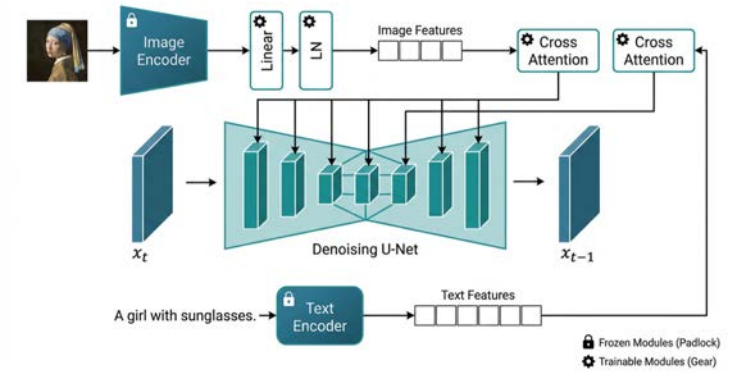


그림 1. IP-Adapter 전체 구조
Fig. 1. Overall Architecture of IP-Adapter

관성을 향상시킨다.

특히, IP-Adapter는 사전 학습된 확산 모델의 가중치를 직접 수정하지 않고, 별도의 어댑터 모듈을 통해 이미지 특징을 Cross-Attention 계층에 주입하는 구조를 가진다[4]. 이로 인해 기존 모델의 표현력을 유지하면서도 특정 인물, 객체, 스타일 등의 시각적 특성을 안정적으로 반영할 수 있다.

그림 1에서는 기존 어댑터 방식인 서로 다른 모달리티의 특징을 단일 어텐션 연산으로 통합하는 것과는 달리, 텍스트와 이미지 조건을 동시에 반영하기 위해 분리된 Cross-Attention 구조를 지닌다. 최종 출력은 다음과 같이 정의된다.

$$Znew = = softmax(\frac{QK^T}{\sqrt{d}})V + softmax(\frac{QK'^T}{\sqrt{d}})V' \quad (1)$$

여기서 Q는 쿼리 벡터, K, V는 텍스트 특징 벡터로부터 도출된 키와 값이며 K', V'는 이미지 특징으로부터 도출된 키와 값을 의미한다. 이러한 구조는 각 모달리티에 특화된 어텐션 가중치를 학습하도록 유도함으로써, 텍스트와 이미지 조건 간 표현력을 향상시킨다.

그림 2는 수많은 SaaS 기반 이미지 생성형 AI 모델 중 하나인 도구인 미드저니에서의 IP-Adapter 활용 예시다. 원하는 그림체를 지닌 이미지와 본 연구에서 설계한 프롬프트 “A man walking a load, full shot”과 같이 조건화(Conditioning)하여 생성된 결과 이미지를 보여주고 있다[5].

2. LoRA Fine-Tuning

LoRA는 LLM 분야에서 제안된 가중치 미세조정 기법으로, 사전 학습된 대규모 모델의 전체 파라미터를 재학습하

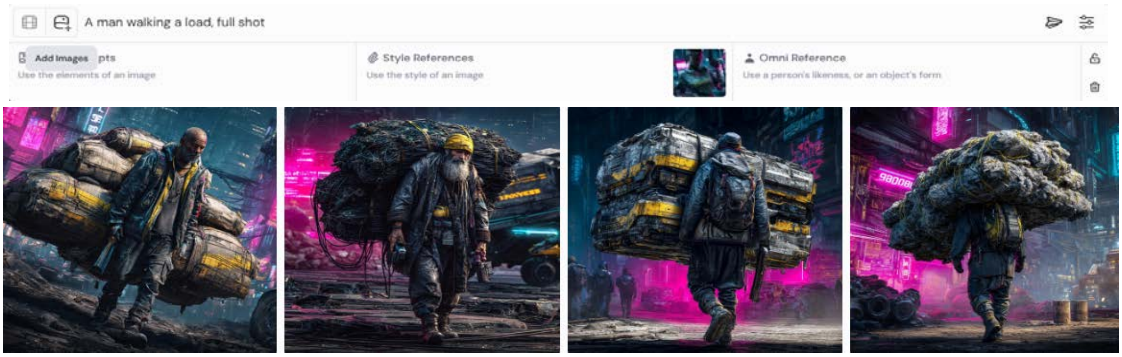


그림 2. Midjourney(SaaS)에서 IP-Adapter 기반 이미지 프롬프트 생성 사례
Fig. 2. Example of Image-Prompt-Based Generation Using an IP-Adapter Framework in Midjourney(Saas)

지 않고 저랭크 행렬을 추가하여 효율적으로 모델을 학습시키는 방법이다. 전체 파라미터를 재학습하는 방법은 높은 계산 비용과 메모리 자원을 요구하는 반면, LoRA는 기존 가중치를 고정한 상태에서 일부 선형 계층에 저차원 행렬을 삽입하여 학습함으로써 연산량을 크게 줄인다[6]. 연산은 다음과 같다.

$$W' = W + BA \tag{2}$$

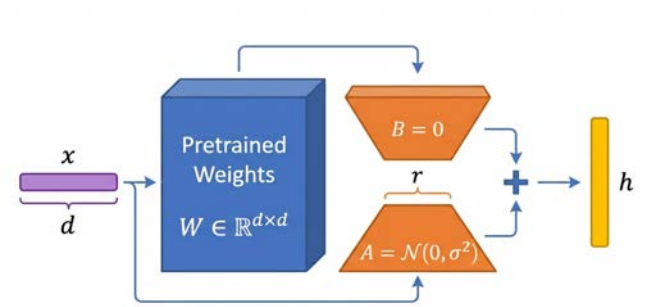


그림 3. LoRA 아키텍처
Fig. 3. Architecture of Low-Rank Adaptation(LoRA)

그림 3은 LoRA의 구조를 시각적으로 나타낸 것이다. 기존의 사전 학습된 가중치 W는 고정된 상태로 유지되며, 학습 과정에서는 두 개의 저랭크 행렬 A와 B가 추가로 학습된다. 이때 rank r은 가중치 업데이트가 이루어지는 저차원 부분공간의 차원을 나타낸다.

이와 같은 LoRA 구조는 이후 이미지 생성 분야로 확장되었으며, 특히 확산 모델 기반 T2I(Text to Image) 생성 모델에서 활용되고 있다. Stable Diffusion과 같은 구조에서는 U-Net의 어텐션 계층에 LoRA를 삽입하여 특정 스타일, 객체, 인물 속성 등을 효율적으로 학습한다. 이를 통해 전체 모델을 재학습하지 않고 소규모의 파라미터를 추가하는 것만으로도 효율적인 추론이 가능하다[7].

이러한 특성은 특정 인물의 얼굴 정체성을 유지하면서 다양한 시점에서 일관된 이미지를 생성하는 데에도 활용될 수 있다. 이에 본 연구에서는 먼저 인물 얼굴의 일관성을 유지하는 LoRA 학습 파일을 구축하기 위하여 학습용 데이터셋 이미지들을 생성하였다.

그림 4는 Qwen Image Edit 모델을 활용하여 생성한 20장의 데이터셋 이미지이다. Qwen 모델은 대규모 사전 학습을 통해 입력 이미지의 시각적 특징을 보존하면서도 텍스트 조건에 따른 구조적 변형이 가능하다는 장점이 있다. 본 연구에서는 이러한 특성을 활용하여 인물의 정체성을 유지한 상태에서 다양한 시점 이미지를 생성하였다. 프롬프트는 profile view, three-quarter view, high-angle, low-angle 등

그림 4. LoRA 학습을 위한 이미지 데이터셋
Fig. 4. Image Dataset for LoRA Training

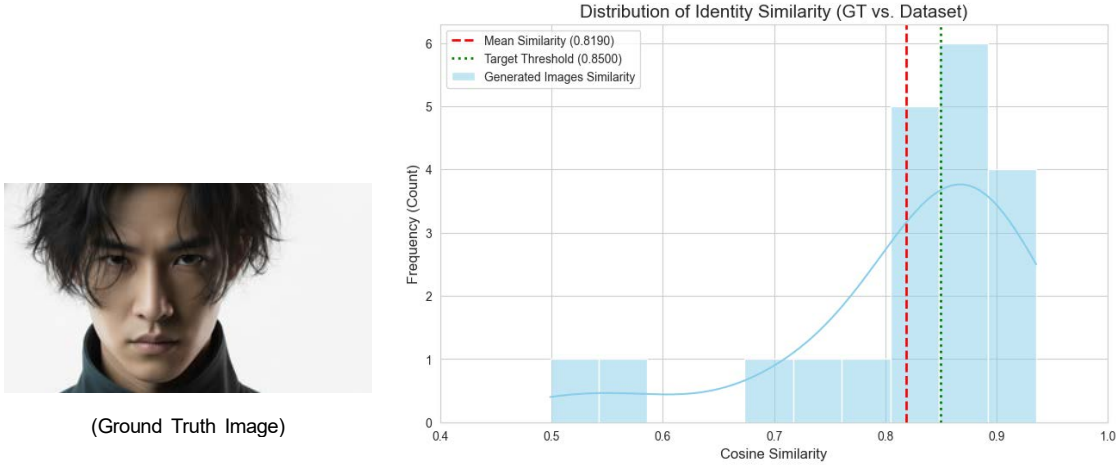


그림 5. 데이터셋 이미지와 기준 이미지 간 유사도 비교
Fig. 5. Similarity Between Dataset Images and Ground Truth Images

실제 촬영 구도를 중심으로 설계되었으며, 모든 생성 과정에서 동일한 Seed 값을 고정하여 확률적 변동을 최소화함으로써 인물의 얼굴 특징과 스타일이 안정적으로 유지되도록 하였다.

그림 5는 구축된 20장의 데이터셋과 원본 이미지(Ground Truth) 간의 정체성 일관성을 검증하기 위해 FaceNet 기반의 안면 유사도 분석을 실시하였다[8]. 검증을 위한 임계값은 안면 인식 시스템에서 일치성을 보장하는 기준인 0.85로 설정하였다. 이는 생성된 이미지가 단순한 시각적 유사성을 넘어 LoRA 학습을 위한 정교한 참조 데이터로서의 품질을 확보했는지 판별하기 위함이다.

분석 결과, X축에서 확인할 수 있듯이 대부분의 이미지가 0.85 근처 및 그 이상의 고유사도 구간에 집중되어 있음을 확인하였다. 전체 이미지의 평균 유사도는 0.8190으로 측정되었으나, 이는 High-angle, Low-angle, Apply Rembrandt lighting 등의 이미지와 같이 다양한 시점과 조건이 부여된 상황을 위해 의도적으로 포함시킨 극단적인 촬영 구도 이미지들로 인해 나타난 통계적 결과이다. 만약 모든 이미지의 유사도가 0.85를 상회하였다면, 이는 정면 위주의 단조로운 데이터셋으로 구성되어 다양한 공간적인 맥락을 학습하기 어려웠을 것이다. 따라서 본 연구의 데이터셋은 인물의 정체성을 유지하는 '일관성 그룹'과 시점에 따른 대응력을 높이는 '다양성 그룹'이 전략적으로 조합되었으며, 이러한 분포 특성은 모델이 인물의 고유 특징을 보존함과 동시에 입체적인 공간 학습을 가능하게 하는 핵심적인 근거가 된다.

표 1. 학습 데이터셋과 LoRA 미세조정 설정
Table 1. Training Dataset and LoRA Fine-tuning Configuration

Traning model method	Using Flux-dev-lora-trainer
Trigger_word	jaehyuk
Steps	1000
Learning_rate	0.0004
Batch_size	1
Resolution	512,768,1024
Lora_rank	16
Optimizer	Adamw8bit

이후 표 1에서는 본 작업 예시에서 구축한 학습용 데이터셋 구성과 LoRA 파인튜닝에 사용된 주요 하이퍼파라미터 설정을 정리하여 제시한다. 각 이미지는 1024x1024 크기로 resize하여 전처리하였으며 사용된 LoRA 학습 하이퍼파라미터는 기존 LoRA 기반 인물 학습 사례와 공개 커뮤니티 실험 결과를 참고하여 설정하였다. Rank는 모델의 표현력을 확보하면서도 과적합을 방지하기 위해 16으로 설정하였으며, 학습 스텝은 데이터 수를 고려하여 1000 step으로 제한하였다. Learning rate는 배경과 그림체 학습에서 10^{-6}

근사값을 권장하는데 여기서는 캐릭터 얼굴의 일관성을 목표로 함으로써 권장된 값이 10^{-4}근사값이므로 0.0004로 설정하여 안정적인 수렴을 유도하였고, Batch size는 GPU 메모리 제약과 소규모 데이터 특성을 고려하여 1로 설정하였다.

표 2는 FLUX 기반 LoRA 파인튜닝을 적용하고 텍스트 프롬프트를 통해 상황을 부여하여 생성된 결과 이미지를 나타낸다[9]. 실내 미래 도시 건물 장면과 도심 골목 장면과 같이 프롬프트에 따른 공간적 맥락은 변화하였으나, 인물의 얼굴 특징, 헤어스타일 및 전반적인 외형적 정체성은 일관되게 유지되는 경향이 확인된다. 이는 LoRA가 장면 변화와 무관하게 인물 고유의 특성을 안정적으로 학습하고 재현함을 보여준다.

그림 6은 LoRA 학습 데이터 중 하나의 샘플 이미지와

표 2. FLUX 기반 LoRA 파인튜닝 결과 이미지
Table 2. Generated Images Using FLUX-Based LoRA Fine-Tuning

Prompt	Output Image
He is talking to someone inside a future city building	
He is standing in an alley in the city of Seoul, bust shot	

그림 6. 학습 데이터와 LoRA 생성 이미지 간 정량적 평가
Fig. 6. Quantitative Evaluation Between Training Data and LoRA-Generated Image

이를 Flux 기반으로 학습한 LoRA를 적용한 후 생성된 이미지 간의 코사인 유사도 및 거리를 ArcFace 기반 임베딩으로 정량 평가한 결과를 나타낸다[10]. 측정된 코사인 거리 값은 0.4839로, 동일 인물 판별 임계값 0.68보다 낮게 나타났으며 두 이미지는 동일 인물로 분류되었다. 본 연구에서는 동일 인물 판별 기준으로 DeepFace 구현에서 ArcFace 모델에 대해 제시한 권장 임계값인 0.68을 적용하였으며, 해당 값은 LFW 벤치마크 기반 실험을 통해 검증된 값이다[11].

Ⅲ. 다중 시점 인물 일관성 생성을 위한 이미지 생성 방법

기존의 AI 기반 이미지 생성 모델은 단일 이미지 단위에서 높은 시각적 완성도를 보여주며, 영화 제작의 콘셉트 아트나 초기 비주얼 탐색 단계에서 널리 활용되고 있다. 특히 미드저니와 같은 SaaS 기반 이미지 생성형 AI 모델은 복잡한 프롬프트 입력과 참조 이미지 기능만으로도 스타일화된 고품질 이미지를 생성할 수 있어, 비전문가도 손쉽게 시각적 결과물을 얻을 수 있다는 장점을 가진다.

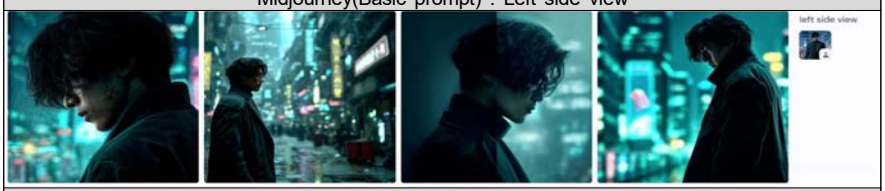
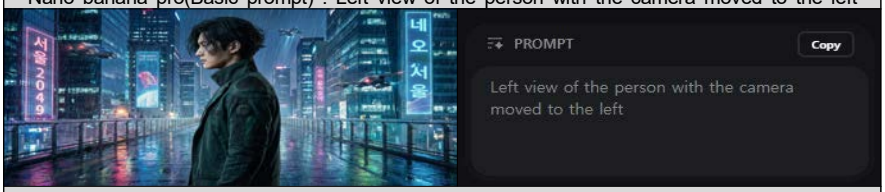
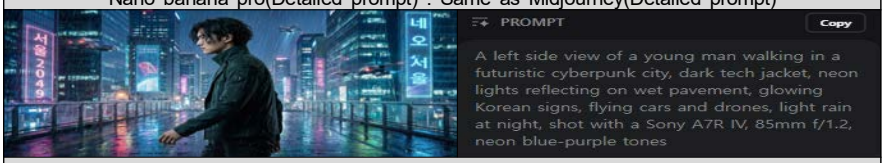
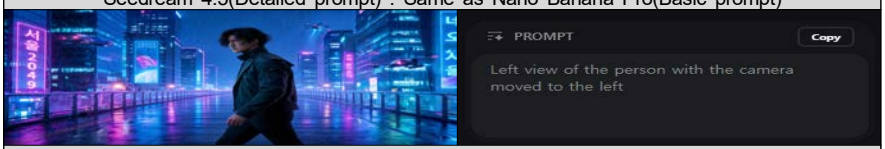
그러나 이러한 SaaS 기반 이미지 생성형 AI 모델들은 영화 제작 과정에서 필수적인 요소인 '동일 인물의 일관성' 측면에서는 구조적인 한계를 지닌다.

표 3은 Scene 1의 첫 번째 Cut 이미지에서 두 번째 컷인 왼쪽 측면 시점(Left side view)을 생성하기 위해 상용 모델들을 활용한 실험 사례를 나타낸다. 위 모델들을 이용할 때 참조 이미지로 'Scene 1-1 Cut' 이미지와 여러 단순한 프롬프트 혹은 복잡한 프롬프트를 입력하여 생성된 결과이다. 첫 번째로, 미드저니에서는 참조 이미지의 스타일을 약간 반영하였지만, 이미지의 톤, 인물의 속성, 배경이 크게 변화된 결과가 생성되었으며 두 번째로, 현재 이미지 생성 모델 중에서 SOTA 모델인 나노 바나나 프로 모델은 배경과 인물의 속성, 톤 모두 참조 이미지와 일관성을 잘 유지하지만 본 연구의 목표는 인물의 단순한 측면 이미지가 아니라, 카메라가 인물을 중심으로 왼쪽으로 이동하며 촬영된 시점 이미지를 생성하는 것이다. 따라서, 위 프롬프트 외에도 다양한 변형 프롬프트를 사용하여 실험을 수행하였으나 목표로 하는 이미지를 안정적으로 생성하기 어려웠다. 마지막으로 시드림(Seedream) 모델 또한 색감이 크게 달라졌으며, 인물의 스타일은 비슷하게 출력되었지만, 나노 바나나 프로 모델과 마찬가지로 인물의 옆모습이 나왔으며 인물의 피부톤이 부자연스럽게 생성되었다. 이러한 결과를 통해 일관된 다중 시점의 이미지를 생성하고자 할 경우 SaaS 기반 이미지 생성형 AI 모델의 텍스트 프롬프트 및 참조 이미지 기반 생성 방식은 인물의 전반적인 분위기나 스타일은 유지할 수 있으나, 배경, 장신구, 의상 색감 등 세부적인 시각적 특징을 모든 샷에서 일관되게 재현하는 데에는 어려움이 있다. 특히 프롬프트가 복잡해질수록 세부 요소 간의 우선순위가 불안정해지며, 생성 결과마다 미세한 변형이 누적되는 문제가 발생한다.

이러한 문제는 단일 이미지 생성에서는 비교적 허용 가능한 수준일 수 있으나, 동일 캐릭터가 반복적으로 등장하는 영화 제작 환경에서는 치명적인 제약으로 작용한다. 예를 들어 동일 인물을 정면, 측면, 후면 등 다양한 시점에서 생성할 경우, 얼굴 윤곽, 체형 비율, 복장 구조가 시점에 따라 달라지거나 형태적 왜곡이 발생하는 현상이 빈번히 관찰된다. 이는 3D 모델링을 기반으로 접근한 것이 아닌 2D 확산 모델만을 활용하는 상용 도구의 근본적인 한계로, 다중 시점 이미지 생성에서 객체 정체성을 안정적으로 유지하지 못하는 원인으로 작용한다. 또한 미드저니와 같은 SaaS 기반 이미지 생성형 AI 모델은 내부 모델 구조와 학습 방식이 공개되지 않은 형태로 제공되기 때문에, 사용자가 특정 캐릭터의 시각적 정체성을 체계적으로 학습시키거나 제어하는 데 한계가 있다. 이는 동일 캐릭터의 반복적인 생성이 필요한 경우에도 매번 프롬프트 수정과 결과 선별에 의존해야 하며, 결과적으로 제작자의 작업 부담과 비용 부담을 증가시키는 요인이 된다.

본 논문에서는 이러한 SaaS 기반 이미지 생성형 AI 모델 사용 시 발생하는 문제를 분석하기 위하여 미드저니, 나노 바나나 프로, 시드림 모델을 이용하여 동일한 참조 이미지를 기반으로 다양한 프롬프트 실험을 수행하였다. 또한 실험 중에 발생한 한계를 극복하기 위해 동일한 캐릭터의 시각적 정체성을 유지하면서 다중 시점 이미지를 생성할 수 있는 일관된 이미지 생성 파이프라인을 구성하고 이를 기반으로 실험을 수행하였다.

표 3. SaaS 이미지 생성 도구의 장면 속 컷 간의 일관성 유지 한계
Table 3. Limitations in Maintaining Inter-Shot Consistency Within a Scene in SaaS Image Generation Tools

<table>
<tr><td colspan="2">Scene 1-1 Cut</td></tr>
<tr><td>  </td><td>For the next generated image (Shot 1-2), the background, character style, and clothing attributes are preserved, while the camera position is adjusted to generate a left-view perspective.</td></tr>
<tr><td colspan="2">Midjourney(Basic prompt) : Left side view</td></tr>
<tr><td colspan="2">  </td></tr>
<tr><td colspan="2">Midjourney(Detailed prompt) : A left side view of a young man walking in a futuristic cyberpunk city, dark tech jacket, neon lights reflecting on wet pavement, glowing Korean signs, flying cars and drones, light rain at night, shot with a Sony A7R IV, 85mm f/1.2, neon blue-purple tones</td></tr>
<tr><td colspan="2">  </td></tr>
<tr><td colspan="2">Nano banana pro(Basic prompt) : Left view of the person with the camera moved to the left</td></tr>
<tr><td colspan="2">  </td></tr>
<tr><td colspan="2">Nano banana pro(Detailed prompt) : Same as Midjourney(Detailed prompt)</td></tr>
<tr><td colspan="2">  </td></tr>
<tr><td colspan="2">Seedream 4.5(Detailed prompt) : Same as Nano Banana Pro(Basic prompt)</td></tr>
<tr><td colspan="2">  </td></tr>
<tr><td colspan="2">Seedream 4.5(Detailed prompt) : Same as Midjourney(Detailed prompt)</td></tr>
<tr><td colspan="2">  </td></tr>
</table>

Ⅳ. 실험 및 결과 분석

앞서 III장에서 분석한 바와 같이, SaaS 기반 이미지 생성 모델은 장면 내 컷 간 시각적 일관성을 구조적으로 보장하기 어렵다. 이에 본 장에서는 GitHub 및 Hugging Face와 같은 공개 커뮤니티에서 공유되는 커스텀 노드를 기반으로 노드형 워크플로우를 설계하고, 이를 통해 다중 시점에서도 일관된 인물 표현을 확보할 수 있는 생성 파이프라인을 제안한 연구를 다음과 같이 하였다.

그림 7의 워크플로우는 Qwen 계열 모델을 중심으로 일관된 구조로 설계되었다. 기본 Diffusion Model로는 이미지 편집 성능이 우수한 qwen_image_edit_2509 모델을 적용하였다[12]. 모델 간 호환성과 표현 일관성을 확보하기 위해 LoRA, VAE, CLIP 역시 Qwen 기반 모델로 통일하였다.

다중 시점 적응 단계에서는 qwen_multi_angle LoRA를 적용하여 카메라 시점 변화에 특화된 특징을 학습하도록 하였다. 이는 단순 스타일 유지가 아니라, 시점 변화 상황에서도 인물 구조와 장면 배치를 안정적으로 유지하도록 보정하는 역할을 수행한다. 또한 추론 효율을 고려하여 Qwen-Image-Lightning-8steps LoRA를 추가 결합하였으며, 이를 통해 8단계 샘플링만으로도 안정적인 시점 변환 결과를 도출할 수 있도록 최적화하였다[13].

조건 제어 단계에서는 텍스트 기반 조건화(conditioning)를 보다 명시적으로 제어하기 위해 CameraControlPromptNode를

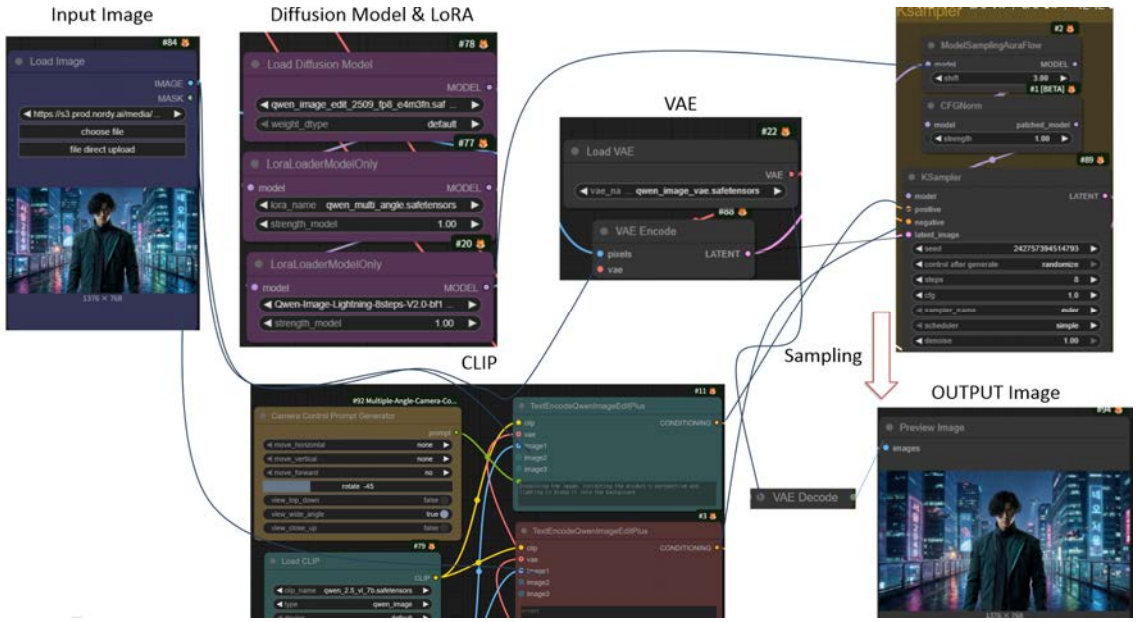


그림 7. Qwen 이미지 모델을 활용한 노드 기반 워크플로우 아키텍처
Fig. 7. Node-Based Workflow Architecture Using Qwen Image Model

표 4. 사용된 모델
Table 4. Used Model

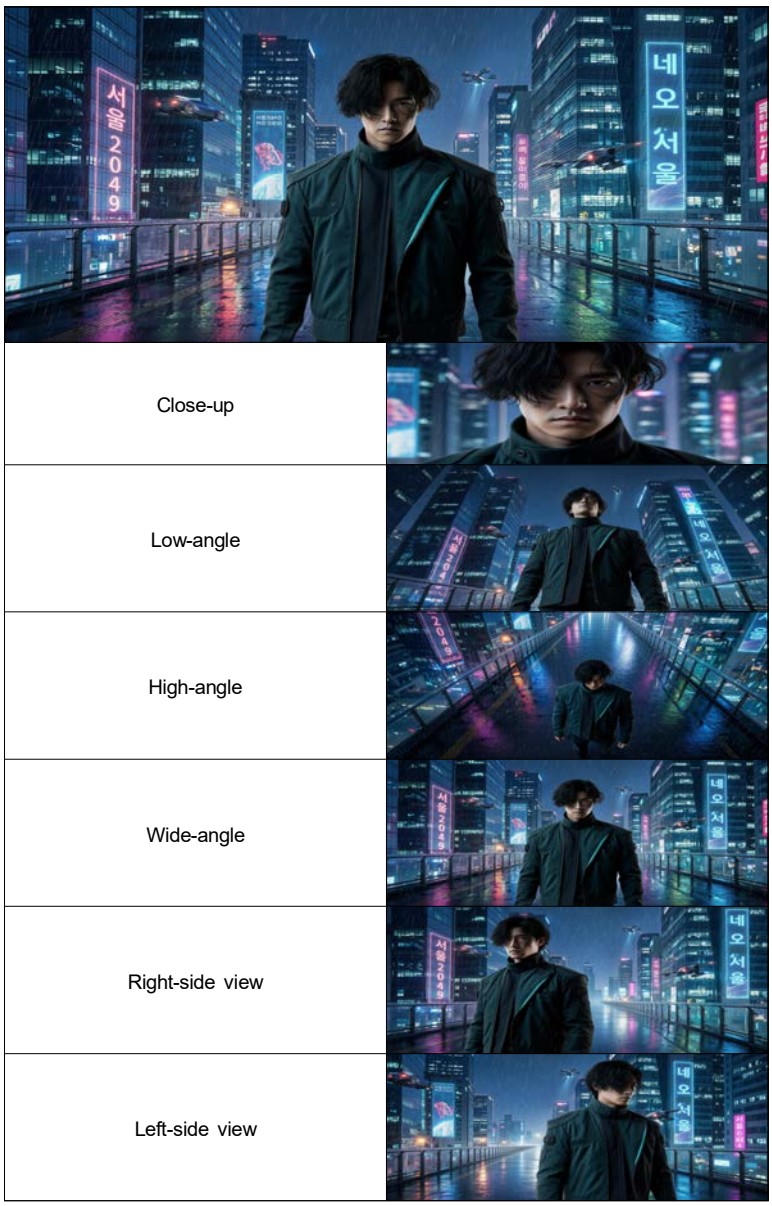
models	Diffusion_model	Qwen_Image_Edit_2509_fp8.safetensors
	Lora 1	Qwen_Multi_Angle.safetensors
	Lora 2	Qwen_Image_Lightning-8steps-v2.0-bf16.saftensors
	VAE	Qwen_image_vae.safetensors
	CLIP	Qwen_2.5_vl_7b.saftensors

추가하였다. 해당 노드는 수평 이동, 수직 이동, 전후 이동 및 회전 파라미터를 조합하여 카메라 움직임을 구조화된 텍스트 프롬프트로 자동 생성한다. 생성된 프롬프트는 긍정 프롬프트 조건화에 반영되어 시점 관련 정보가 모델 내부 표현에 일관되게 전달되도록 한다[14].

이후 잠재 공간 처리 단계에서는 Qwen 기반 VAE(Variational Autoencoder)를 통해 이미지를 Latent 공간으로 인코딩하고, 이를 KSampler의 초기 Latent로 입력하여 확산 과정을 수행한다. 이 과정에서 시점 조건과 적응 모듈이 결합되어 최종 이미지를 생성한다[15].

표 5는 제안한 Qwen 기반 워크플로우를 통해 생성된 다중 시점 이미지 결과를 나타낸다. 생성된 다중 시점 이미지

표 5. Qwen 기반 ComfyUI 워크플로우를 통한 멀티뷰 생성 결과
Table 5. Multi-View Generation Results Using the Qwen-Based ComfyUI Workflow



들은 카메라 앵글과 샷 사이즈가 변화함에도 불구하고 인물의 정체성과 장면의 분위기가 일관되게 유지되는 현상이 보인다. 배경의 공간 구조 또한 일관된 형태를 유지한다. 이는 제안한 워크플로우가 컷 간 시각적 연속성을 효과적으로 확보함을 정성적으로 보여준다.

그림 8은 CLIP 이미지 인코더를 활용하여 기준 이미지와 각 시점별 생성 이미지 간의 코사인 유사도를 정량적으로 분석한 결과를 나타낸다. 전체 평균 유사도는 0.8764로 나타났으며, 이는 다중 시점 변화에도 불구하고 의미적 일관성이 전반적으로 유지됨을 의미한다. 좌, 우 측면 및 와이드 시점에서는 0.91~0.95 수준의 높은 유사도가 관찰되었으며, 극단적인 상·하 앵글(High-angle, Top-down view)에서는 상대적으로 낮은 값을 보였다. 이는 카메라 각도의 변화에 따라 시각 정보의 분포가 달라지면서 임베딩 공간에서의 표현 차이가 증가한 결과로 해석된다.

그림 9는 제안하는 워크플로우의 결과 이미지와 주요 SaaS 모델 간의 정량적 비교 분석 결과를 나타낸다. 좌측의 지표를 통해 알 수 있듯이, 제안하는 워크플로우는 Style, Cam Move, Semantic의 세 가지 지표 모두에서 가장 균형 잡힌 높은 수치를 기록하고 있다.

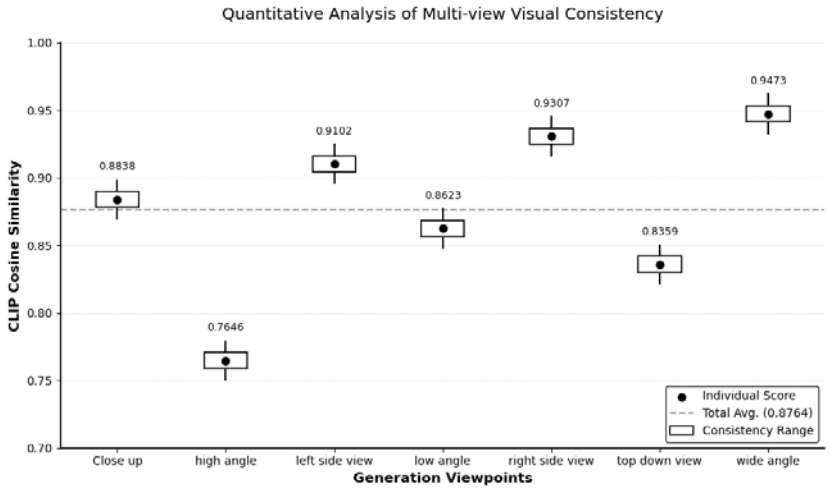


그림 8. CLIP 이미지 인코더를 이용한 다중 시점 이미지 간 코사인 유사도
Fig. 8. Cosine Similarity Across Multi-View Images Using CLIP Image Encoder

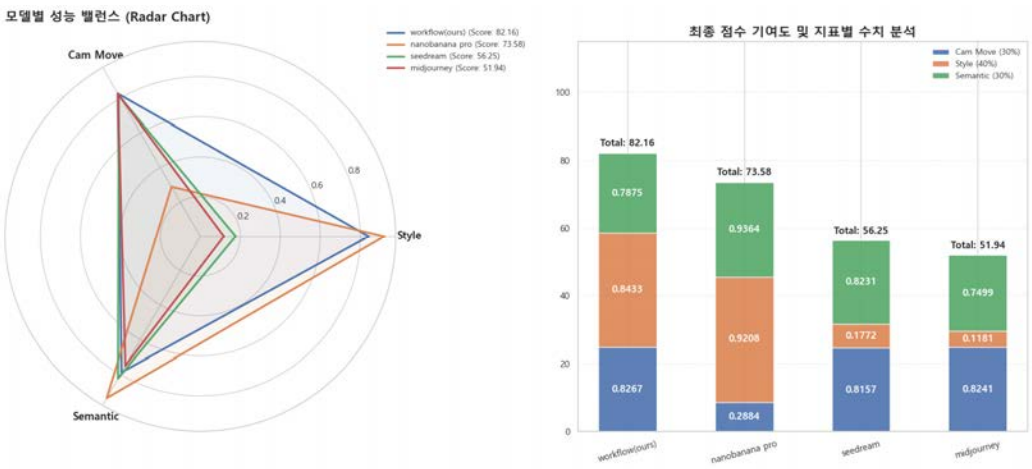


그림 9. 제안하는 워크플로우와 SaaS 모델 간의 정량적 비교 분석
Fig. 9. Quantitative Comparative Analysis between the Proposed Workflow and SaaS-based Generative Models

일반적으로 이미지 유사도 평가에 사용되는 SSIM(Structural Similarity)은 픽셀의 휘도, 대비 및 공간적 좌표 기반의 구조적 유사성을 측정하는 특성을 가진다[16]. 그러나 카메라가 물리적으로 이동하여 배경 구도가 전면적으로 교체되는 연출에서는 픽셀의 물리적 위치가 이동하게 되며, 이로 인해 실제 영상의 연속성과는 무관하게 수치상으로는 낮은 유사도를 기록하는 픽셀 좌표 의존성에 따른 평가 왜곡이 발생한다.

본 연구에서는 이러한 왜곡을 바로잡기 위해 낮은 SSIM을 오히려 역동적인 카메라 워킹의 증거로 삼는 Cam Move(1-SSIM) 지표를 도입하였다. 실험 결과, 제안하는 워크플로우는 Cam Move 수치에서 0.8267을 기록하며, 0.2884에 그친 나노 바나나 프로 대비 실제 영화 촬영과 유사한 3차원적 공간 변화를 성공적으로 구현하였음을 정량적으로 증명하였다.

비록 색채 일관성을 나타내는 Style과 의미적 구성을 뜻하는 Semantic 지표에서 나노 바나나 프로가 각각 0.9208과 0.9364를 기록하여 본 연구의 결과보다 높게 나타났으나, 이는 해당 모델이 카메라의 궤적 이동 없이 동일 구도 내에서 피사체만 회전시키는 방식을 취함으로써 얻은 결과에 불과하다. 즉, 나노 바나나 프로는 픽셀의 공간적인 배치가 크게 변하지 않았기에 높은 유사도 점수를 획득한 것이며, 반면 본 연구의 워크플로우는 급격한 시점 변화라는 실험 조건 속에서도 Style과 Semantic 점수를 높은 수준으로 유지하며 연속성을 확보하였다.

결과적으로 우측의 최종 점수 기여도 분석에서 확인할 수 있듯이, 제안하는 워크플로우는 종합 점수 82.16점으로 전체 모델 중 1위를 기록하였다. 이는 본 연구가 인물의 정체성을 보존하는 동시에, SaaS 기반 이미지 생성 모델 대비 역동적인 카메라 무빙이 가능한 독보적인 노드 기반 제어 능력을 갖추었음을 시사한다.

V. 결 론

본 연구는 영화 제작 환경에서 요구되는 다중 시점 간 시각적 일관성 확보를 위해 기존 SaaS 기반 이미지 생성 모델의 한계를 분석하고, 이를 보완할 수 있는 노드 기반 생성 파이프라인을 제안하였다. 실험 결과, 현재 나온 SaaS 기반 이미지 생성 모델들은 단일 이미지의 완성도는 높으나 시점 변화 시 배경 구도를 고정하거나 인물의 정체성을 왜곡하는 등 연속성 제어에 구조적 제약이 있음을 확인하였다.

이에 본 연구는 Qwen 계열 모델과 Multi-Angle LoRA, Lightning LoRA를 결합한 모듈화된 워크플로우를 설계하여 경량화된 환경에서도 정밀한 시점 제어가 가능하도록 하였다. 특히 본 연구에서 제안한 Cam Move(1-SSIM) 지표 분석 결과, 본 워크플로우는 0.8267의 높은 수치를 기록하며 단순 피사체 회전에 그치는 기존 모델들과 달리 실제 영화 촬영과 유사한 3차원적 공간 변화를 성공적으로 구현하였다.

또한, 급격한 시점 변화라는 가혹한 조건 속에서도 Style 및 Semantic 분석을 통해 장면의 핵심 무드와 공간 구성의 의미적 연속성을 증명하였다. 이는 이미지를 저해상도로 리사이징하여 세부 노이즈를 제거하고 전체적인 배치 패턴의 방향성을 측정하는 코사인 유사도 기법을 적용함으로써, 픽셀 위치의 변화와 무관하게 동일한 서사적 공간임을 정량적으로 입증한 결과이다. 이러한 성과는 생성형 AI가 단순한 이미지 생성을 넘어 영화 제작에 필수적인 시각적인 일관성을 정밀하게 통제할 수 있는 실무적 대안임을 시사한다. 본 연구는 제작 목적에 따라 정밀한 제어가 요구되는 상황에서 활용 가능한 워크플로우의 방향을 제시했다는 점에서 의의를 갖는다. 다만, 현재의 실험은 단일 인물을 중심으로 진행되었기에 향후 연구에서는 복잡한 군중 장면에서도 개별 캐릭터의 독립성이 완벽히 보존되는 기법을 연구할 예정이다.

참 고 문 헌 (References)

[1] Byeon-won Jeon and Min-cheol Cha, “HCI Paradigm Shifts in Film and Video Production Following the Introduction of Generative AI and the Convergence of Production Efficiency and Cultural Technology,” Journal of Culture and Technology (JCT), Vol. 12, No. 1, pp. 31-47, January 2026.

[2] Ruihan Zhang, et al., Generative AI for Film Creation: A Survey of Recent Advances, arXiv, arXiv:2504.08296v1 [cs.CV], April 2025. doi: https://doi.org/10.48550/arXiv.2504.08296

[3] Chung-young Lee, “A Study on Generative AI-Based Animation Production Process Using ComfyUI,” Journal of Animation Studies, Vol.21, No. 2, pp. 178-196, June 2025. doi: https://doi.org/10.51467/asko.2025.6.21.2.178

[4] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang, IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion

Models, arXiv preprint, arXiv:2308.06721, August, 2023.
doi: https://doi.org/10.48550/arXiv.2308.06721
[5] Midjourney, Midjourney Official Website, Available: https://www.midjourney.com
[6] Edward Hu, et al., LoRA: Low-Rank Adaptation of Large Language Models, arXiv, arXiv:2106.09685v2 [cs.CL], October 2021.
doi: https://doi.org/10.48550/arXiv.2106.09685
[7] Nupur Kumari, et al., Multi-Concept Customization of Text-to-Image Diffusion, arXiv, arXiv:2212.04488v2 [cs.CV], June 2023.
doi: https://doi.org/10.1109/cvpr52729.2023.00192
[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin, “FaceNet: A Unified Embedding for Face Recognition and Clustering,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815-823, June, 2015.
doi: https://doi.org/10.1109/CVPR.2015.7298682
[9] Likun Li, et al., Block-wise LoRA: Revisiting Fine-grained LoRA for Effective Personalization and Stylization in Text-to-Image Generation, arXiv, arXiv:2403.07500v1 [cs.CV], March 2024.
doi: https://doi.org/10.48550/arXiv.2403.07500
[10] JJiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, ArcFace: Additive Angular Margin Loss for Deep Face Recognition, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4690-4699, June, 2019.
doi: https://doi.org/10.1109/CVPR.2019.00482
[11] https://github.com/serengil/deepface/blob/master/deepface/config/threshold.py
[12] Qwen Team, Qwen-Image Technical Report, arXiv, arXiv:2508.02324v1 [cs.CV], August 2025.
[13] Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R., Adversarial Diffusion Distillation, arXiv preprint arXiv:2311.17042, Nov, 2023.
doi: https://doi.org/10.1007/978-3-031-73016-0_6
[14] Alec Radford, Jong Wook Kim et al., "Learning Transferable Visual Models From Natural Language Supervision," arXiv preprint arXiv:2103.00020v1, February, 2021.
doi: https://doi.org/10.48550/arXiv.2103.00020
[15] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," Foundations and Trends® in Machine Learning, Vol. 12, No. 4, pp. 307-392, 2019.
doi: https://doi.org/10.1561/2200000056
[16] Nilsson, J., and Akenine-Möller, T., “Understanding SSIM,” arXiv:2006.13846v2 [eess.IV], pp. 1-11, June 2020.
doi: https://doi.org/10.48550/arXiv.2006.13846

저 자 소 개

권 오 주

- 2023년 2월 ~ 2025년 2월 : DIMA OBVAN 중계기술팀
- 2025년 2월 : 동아방송예술대학교 방송기술학과 예술학사
- 현재 : 서울과학기술대학교 정보통신미디어공학과 석사생
- ORCID : https://orcid.org/0009-0005-1059-7722
- 주관심분야 : 생성형 AI 영화 제작, 컴퓨터 비전, 방송 시스템 구축, 방송기술

구 민 서

- 현재 : 서울과학기술대학교 전자IT미디어공학과 학부생
- ORCID : https://orcid.org/0009-0007-5746-2520
- 주관심분야 : 5G/6G network for AI, Media computiation, AI communication

서 제 웅

- 현재 : 서울과학기술대학교 전자IT미디어공학과 학부생
- ORCID : https://orcid.org/0009-0007-8861-5884
- 주관심분야 : Artificial Intelligence, Deep Learning, Computer Vision / NLP

저 자 소 개

김 에스더

- 현재 : 서울과학기술대학교 문예창작학과 학부생
- ORCID : https://orcid.org/0009-0002-7840-2926
- 주관심분야 : Computer Vision & Graphics, Generatve AI, Video-Language pre-training

김 기 범

- 2024년 8월 : 서울과학기술대학교 문예창작학과 문학사
- 현재 : 서울과학기술대학교 문예창작학과 석사생
- ORCID : https://orcid.org/0009-0000-7666-4055
- 주관심분야 : 생성형 AI 영화 제작, 영화 연출

박 구 만

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

이 영 화

- 2001년 8월 ~ 2021년 12월 : 현대홈쇼핑 방송그래픽팀 NLE 감독
- 2026년 2월 : 서울과학기술대학교 정보통신미디어공학과 박사
- 2023년 8월 ~ 현재 : 시그마케이주식회사 사내이사
- ORCID : https://orcid.org/0000-0002-2427-7667
- 주관심분야 : Computer Vision & Graphics, Generative AI, Video-Language pre-training

최 영 희

- 현재 : 서울과학기술대학교 문예창작학과 부교수
- ORCID : https://orcid.org/0009-0004-2286-5029
- 주관심분야 : 생성형 AI 영화 제작, 영화 시나리오, 동아시아 영화 및 TV 드라마, 미디어와 영상문학