일반논문 (Regular Paper)
방송공학회논문지 제31권 제1호, 2026년 1월 (JBE Vol.31, No.1, January 2026)
https://doi.org/10.5909/JBE.2026.31.1.152
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

대규모 언어 모델 기반 애니메이션 숏폼 리프레이밍

이 강 희[a)], 양 해 준[a)], 배 재 형[a)], 김 탁 훈[a)‡], 최 종 원[a)‡]

LLM-based Animation Reframing for Short-Form Video

Kanghee Lee[a)], HaeJun Yang[a)], Jaehyeong Bae[a)], Takhoon Kim[a)‡], and Jongwon Choi[a)‡]

요 약

멀티모달 대형 언어 모델(Multimodal Large Language Models, MLLMs)은 실사 기반 데이터로 학습되어 과장된 기하 형태 및 단순화된 셰이딩 등 애니메이션 영상을 정확히 이해하지 못한다. 이로 인해 영상 리프레이밍 과정에서 주요 캐릭터가 프레임 밖으로 벗어나거나, 사전 학습에 포함되지 않은 캐릭터에 대한 인식 성능이 저하되는 문제가 발생한다. 이를 해결하기 위해 본 논문에서는 애니메이션 영상과 대본 기반 텍스트 프롬프트를 해석하고, 캐릭터 이미지를 쿼리로 활용하는 트레이닝 프리 숏폼 변환 파이프라인을 제안한다. 제안 방식은 MLLM 기반 장면 추출 후 객체 검출을 수행하고, 적응형 줌 제어(Adaptive Zooming)를 통해 리프레이밍 과정에서 발생하는 객체 손실 및 크롭 오류를 해결하고자 한다.

Abstract

As most multimodal large language models (MLLMs) are trained on real-world data, MLLMs face challenges in accurately interpreting animated videos that feature stylized visual characteristics such as exaggerated geometry and simplified shading. As a result, video reframing often places key characters outside the frame, and recognition performance degrades when characters are not included in the pre-training data. To address this issue, this paper proposes a training-free short-form transformation pipeline that jointly interprets animated video content and script-based text prompts while utilizing character images as visual queries. The proposed approach first performs scene extraction using an MLLM, followed by object detection based on visual queries, and then applies Adaptive Zooming to mitigate object loss and cropping errors that may occur during the reframing process.

Keyword : Large Language Model, Segment Retrieval, Visual Query Localization, Animation

a) 중앙대학교 첨단영상대학원(Department of Advanced Imaging, GSAIM, Chung-Ang University, Seoul)
‡ Corresponding Author : 김탁훈(Takhoon Kim), 최종원(Jongwon Choi)
E-mail: takhoonkim@cau.ac.kr, choijw@cau.ac.kr
Tel: +82-2-820-6962, +82-2-820-5870
ORCID: https://orcid.org/0009-0009-4081-3907, https://orcid.org/0000-0001-9753-8760
※This research was supported by the “Regional Innovation System & Education (RISE)” through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government. (2025-RISE-01-024-04)
· Manuscript December 1, 2025; Revised January 7, 2026; Accepted January 7, 2026.

Ⅰ. 서 론

대부분의 멀티모달 대형 언어 모델(Multimodal Large Language Models, MLLMs)은 주로 실사 이미지 및 비디오 데이터셋을 기반으로 사전 학습되어 있다[1][2][3][4][5][6]. 이에 반해 애니메이션은 양식화된 선, 단조로운 채색, 과장된 기하학적 비율 등 실사 도메인과는 확연히 구분되는 시각적 추상성을 지닌다. 이러한 도메인 격차로 인해 기존 MLLM은 애니메이션의 장면 구조를 해석하는 데 어려움을 겪는다. 이는 애니메이션을 소비가 편리한 형태인 숏폼(Short-form) 콘텐츠로 변환하는 비디오 리프레이밍(Reframing) 작업에서 두드러진다. 리프레이밍은 원본 영상에서 의미적으로 중요한 객체를 중심으로 시야를 재구성하여 화면 비율 변경 또는 해상도 변환하는 과정을 의미한다. 기존 모델은 범용 지식에 의존하므로, 학습 데이터에 존재하지 않는 특정 애니메이션 내 고유 캐릭터를 식별하는 데 근본적인 한계가 있다. 핵심 캐릭터가 소실되거나 서사적으로 중요한 요소가 프레임 밖으로 벗어나는 부정확한 크롭(Crop) 오류가 빈번히 발생한다.

이러한 문제를 해결하기 위한 기존 연구들은 데이터를 재구성하여 대규모 파라미터를 재학습시키는 경우가 많다. 하지만 애니메이션 비디오의 특성상, 리프레이밍 작업을 위해 연속된 프레임에 등장하는 모든 인물의 묘사와 위치 등의 라벨링을 얻는 것은 현실적으로 매우 어렵다.

본 연구에서는 추가적인 학습 없이 애니메이션 영상과 해당 대본을 함께 활용하여, 사용자의 질의에 부합하는 숏폼 형태의 요약 영상을 자동 생성하는 파이프라인을 제안한다. 제안 방식은 사전 학습된 MLLM을 활용하여 대본 기반 장면 추출을 수행한 뒤, 선택된 구간 내 캐릭터를 비주얼 쿼리 기반 객체 검출(Visual Query Localization) 방식으로 탐색한 결과를 보정 정보로 활용한다. 이를 기존 가상 카메라 제어 기법에 결합시킴으로 크롭 오류를 해소한다. 비주얼 쿼리 로컬라이제이션을 통해 얻은 특정 캐릭터들의 바운딩 박스 정보를 활용하여 장면에 따라 적응형 줌 제어(Adaptive Zooming)를 수행한다. 이를 통해 해상도 변환 과정에서 발생하는 주요 객체 소실이나 잘못된 크롭 문제를 완화하고 장면 간 시각적 안정성을 확보한다. 본 논문의 기여점은 다음과 같다.

애니메이션 대본과 사용자 질의를 함께 입력하여, 특정 캐릭터·관계·상황에 부합하는 에피소드 및 시간 구간을 효과적으로 검색하는 텍스트 기반 Retrieval 절차를 제안한다. 이를 통해 서사적 맥락을 반영한 핵심 장면을 자동으로 선별할 수 있다.

선택된 장면에 대해 레퍼런스 캐릭터 이미지를 쿼리로 활용하여, 사전 학습된 MLLM이 추가 학습 없이 프레임 내 특정 캐릭터를 직접 탐색하고 바운딩 박스를 추출하는 비주얼 쿼리 기반 객체 검출 절차를 제안한다.

비주얼 쿼리를 통해 얻은 캐릭터 위치·크기 정보를 기반으로, 기존 가상 카메라 기반 프레이밍에 장면마다 적응형 줌 기능을 적용하여 해상도 변환 과정에서 발생하는 주요 객체 소실, 크롭 오류, 장면 간 불연속성을 보정한다.

본 논문은 Ⅱ장 관련 연구로 시작하여 Ⅲ장 제안 방법 및 학습 방식의 소개, Ⅳ장 실험을 통한 제안 방법의 성능 검증으로 이어진다. 그리고 마지막 Ⅴ장에서는 결론 및 향후 연구 방향에 대하여 논의한다.

Ⅱ. 관련 연구

1. 멀티모달 언어 시각 모델(Multi-modal Language-Vision Models)

최근 멀티모달 대형 언어 모델은 텍스트-이미지 및 텍스트-비디오 간의 추론 능력을 기반으로 장면 요약, 질의응답, 묘사 등 다양한 비전-언어 작업 수행에서 우수한 성능을 보여왔다.

BLIP-2[2]는 Cross-attention 구조를 사용하여 피처 이미지를 LLM과 잘 이어질 수 있도록 변환해 주는 Q-former 기반으로 캡셔닝 및 이미지-텍스트 매칭 작업에서 우수한 성능을 보였다. 이후 InstructBLIP[3], LLaVA[4] 등은 CLIP/BLIP-2의 Vision Encoder를 LLM에 결합하여 캡셔닝 성능을 개선하였으며, 비전-언어 인스트럭션 튜닝을 하여 이미지에 대한 질의응답, 세밀한 묘사 등의 고수준 추론을

가능하게 했다. 이러한 고수준 MLLMs은 단일 이미지에서 문맥을 추론하고, 복잡한 지시문이 주어졌을 때 다양하고 세밀한 답변들을 제공한다.

그러나 단일 이미지 기반 MLLM은 시간적 정보를 포함하지 않아 연속적인 장면 맥락 이해에는 한계가 있으며, Video-ChatGPT[5], LLaVA-Video[6] 등이 이를 개선하기 위해 시간적 정보를 포함하여 비디오 전체를 추론할 수 있도록 하였다.

하지만 이러한 MLLM의 비약적인 발전에도 불구하고, 이를 애니메이션 도메인의 숏폼 자동 변환 작업에 직접적으로 접목한 연구는 아직 충분히 탐구되지 않았다. 이에 본 연구는 MLLM의 방대한 사전 지식과 추론 능력을 해당 분야에 도입하고자 한다. 구체적으로 별도의 추가 없이도 서사적 맥락을 파악하여 핵심 구간을 추출하고, 제로샷 기반의 캐릭터 쿼리 객체 검출을 수행함으로써 기존 MLLM의 잠재력을 실용적인 콘텐츠 제작 파이프라인으로 확장하였다.

2. 쿼리 기반 객체 검출

쿼리 기반 로컬라이제이션 연구들은 텍스트 또는 이미지 쿼리를 기반으로 이미지 내의 특정 대상의 위치를 찾는 데 초점을 맞춘다. Grounding DINO[7]는 텍스트를 기반으로 객체의 위치를 바운딩 박스로 탐지하고, MDETR[8]은 이미지와 텍스트 쿼리를 입력으로 받아, 텍스트에 언급된 대상을 검출 및 위치를 추정하며, AutoFlip[9]은 Machine Learning 기반 영상 리프레이밍 구조로서, 객체 검출 및 트래킹을 활용한 리프레이밍 구조를 제안하였다. TubeDETR[10], VITA[11] 등은 이를 시간적으로 확장하여 비디오 내 해당 객체의 등장 구간을 Sequence 단위로 추적한다. 그러나 이러한 모델들은 대부분 현실 세계 영상 데이터로 학습되어 있어 애니메이션의 비현실적인 스타일 영상에서는 캐릭터 인식 안정성이 떨어지는 문제가 있다.

본 논문은 MLLM을 활용한 대본 기반 장면 추출로 핵심 장면을 선택하고, 이미지 쿼리 기반 객체 검출을 활용해 캐릭터 위치를 추정한 후, 적응형 줌 제어 기술을 통해 리프레이밍 시 발생하는 문제를 해결하고자 한다.

Ⅲ. 제안하는 방법

그림 1은 본 연구에서 제안하는 파이프라인의 전체 개요를 나타낸다. 해당 파이프라인은 서사 기반 에피소드 매칭, 대본 기반 구간 추출과 캐릭터 쿼리 객체 검출 과정을 나타낸다. 이후 적응형 줌 제어를 통해 최적의 숏폼 영상을 생성하게 된다.

1. 서사 기반 에피소드 매칭 (Episode Matching based on synopsis)

본 단계에서는 대규모 멀티모달 언어모델 Gemma-3-27B-IT[12] 모델을 이용하여 사용자 프롬프트 Q에 가장 적합한 에피소드를 선택한다. 각 에피소드 인덱스 $\{e_i\}_{i=1}^{N}$에 대해 사전에 요약된 시놉시스 $\{S\}_{i=1}^{N}$가 구축되어 있다. 여기서 시놉시스는 핵심 서사와 주요 사건을 간략히 정리한 텍스트 요약을 의미한다. 대형 언어모델의 입력은 다음의 수식 (1)과 같다. f는 대형 언어모델, e_*는 선택된 최적의 에피소드를 의미한다.

$$e_* = f([Q; S_1; S_2; \dots ; S_N] \tag{1}$$

프롬프트와 모든 에피소드 시놉시스를 순차적으로 연결하여 입력되고 모델을 프롬프트와 의미적으로 가장 적합한 에피소드를 선택한다.

2. 대본 기반 장면 추출 (Segment Retrieval based on script)

선정된 에피소드 내에서 구체적인 장면을 특정하기 위해, 앞서 사용한 Gemma 모델을 동일하게 활용한다. 각 에피소드마다 대본 $\{D_i\}_{i=1}^{N}$가 구축되어 있다. 모델의 입력은 수식 (2)와 같이 사용자 프롬프트와 해당 에피소드의 대본 정보를 결합한 형태로 구성된다. 이때 f는 대형 언어모델 D^*는 수식 (1)로부터 얻은 에피소드 e_*에 해당하는 대본을 의미한다.

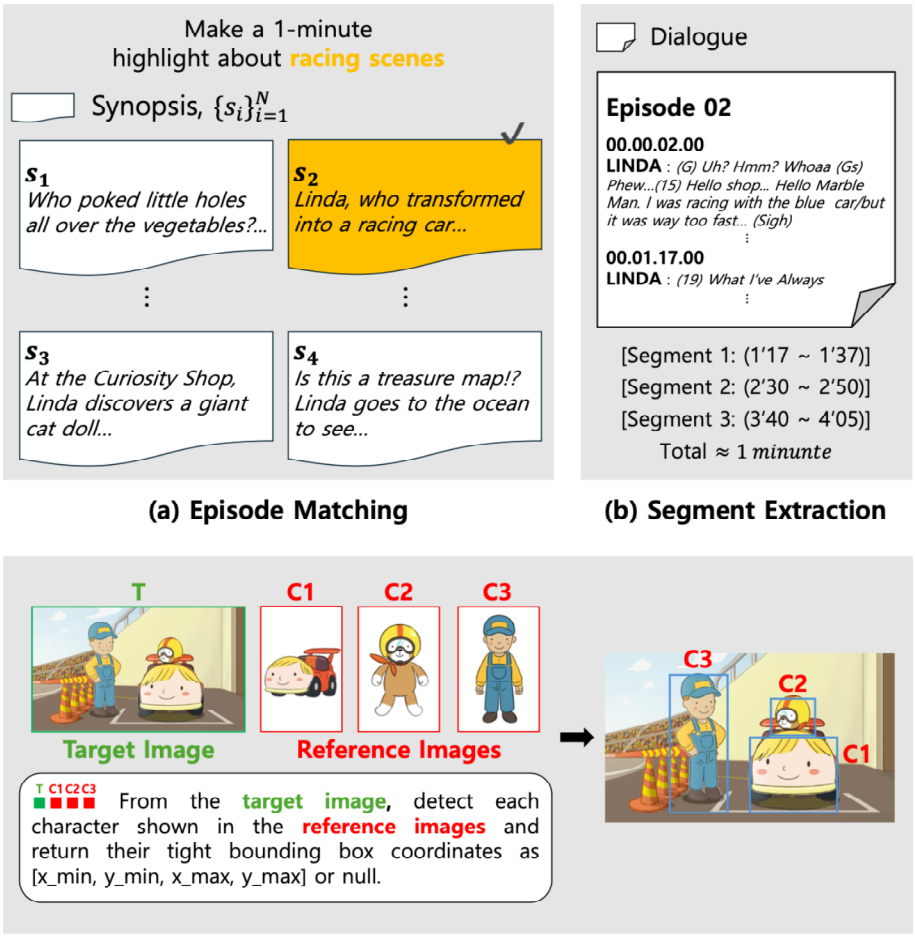

그림 1. 제안하는 프레임워크의 전체 구조
Fig. 1. Overall architecture of the proposed framework

$$Time\ Segments = f([Q; D_*]) \quad (2)$$

이때 대본 전체의 길이가 모델의 입력 허용량(Context Window)을 초과할 경우, 질의와 의미적 연관성이 높은 문맥만을 선별하여 대본을 재구성함으로써 입력을 최적화한다. 모델은 이 축약된 입력을 바탕으로 사용자가 요구한 분량에 맞춰 핵심 장면의 시간 구간을 산출한다. 사용자의 설정에 따라 생성 길이를 유연하게 조절할 수 있다. 이를 통해 사용자는 영상을 수동으로 탐색하는 과정 없이, 질의 의도에 부합하는 에피소드 내 핵심 구간을 자동으로 확보하게 된다.

3. 캐릭터 쿼리 기반 객체 검출

선택된 시간 구간에 속하는 프레임에 대해서는 앞서 활용한 Gemma 모델을 비전 태스크로 확장하여 캐릭터 위치를 추정한다. 그림 1 (c)와 같이, 모델은 해당 시점의 실제 프레임(Target)과 레퍼런스 갤러리에 저장된 캐릭터별 대표 이미지(Reference)를 입력으로 받아 시각적 매칭을 수행한다.

모델은 입력된 정보를 바탕으로 캐릭터별 바운딩 박스와 신뢰도를 출력한다. 검출의 강건성을 확보하기 위해 신뢰도가 임계값보다 낮거나 영역의 크기가 지나치게 작은 후

보는 오검출로 간주하여 제외한다. 이후 중복된 영역을 제거하기 위해 비최대 억제(Non-maximum Suppression)를 적용하여 프레임별 최종 검출 결과를 산출하며, 이 데이터는 후속 가상 카메라 모듈에서 화면 내 주요 관심 영역(Region of Interest)을 결정하는 핵심 지표로 활용된다.

선택된 시간 구간 내의 모든 프레임에 대해 대규모 모델을 적용하는 것은 연산 비용 측면에서 매우 비효율적이다. 이를 해결하기 위해 본 시스템은 해당 구간 내부에서 스트라이드(Stride) 간격으로 프레임을 샘플링하여 연산량을 절감하도록 설계하였다. 아울러 사용자가 강조하고 싶은 특정 캐릭터가 있을 경우, 해당 캐릭터의 이미지만을 쿼리로 구성할 수 있도록 하여 사용자 의도에 따른 높은 자유도와 유연성을 제공한다.

4. 최종 영상 생성

최종적으로 원본 영상을 숏폼 포맷으로 재구성하기 위해 가상 카메라 시스템을 적용한다. 전반적인 카메라 경로 계획 및 스무딩 알고리즘은 AutoFlip[11] 프레임워크를 기반으로 하되, 본 연구에서는 앞서 추출한 캐릭터 바운딩 박스 정보를 줌(Zoom) 제어의 핵심 변수로 활용하여 기존 방식과 차별화하였다.

기존의 비디오 리프레이밍 기술이 단순히 영상의 중요도 맵(Saliency Map) 중심을 따라가는 데 그쳤다면, 제안하는 시스템은 적응형 줌 아웃(Adaptive Zoom-out)을 통해 캐릭터의 스케일을 최적화한다. 이를 위해 시스템은 숏폼 가로 폭의 약 84%(좌우 안전 여백 16% 제외)를 목표 점유 너비로 설정하고, 이를 감지된 바운딩 박스의 합산 너비로 나누어 목표 줌 비율을 산출한다. 이때 과도한 변형을 방지하기 위해 0.6과 1.0으로 각각 하한과 상한을 설정한다. 예를 들어 캐릭터 바운딩 박스의 너비의 합이 300px인 경우, 목표 너비를 현재 너비로 나눈 값이 약 3.0이 된다. 이는 상한선인 1.0을 초과하는 값으로, 캐릭터가 이미 화면 내에 여유 있게 포함되어 잘리지 않고 온전히 보존된 상태임을 의미한다. 시스템은 굳이 줌 아웃을 실행하지 않고 원본 해상도 비율을 유지하여 최적의 화질을 보장한다. 반대로 합산 너비가 1300px에 달하는 경우, 계산된 비율은 약 0.7이 된다. 시스템은 화면 세로 비율을 0.7까지 낮추는 조정을 수행한다. 이 과정에서 캐릭터의 표시 크기가 화면 세로의 0.7배 수준으로 축소됨에 따라 시각적인 줌 아웃(Zoom-out) 효과가 발생하며, 이를 통해 거대한 피사체를 화면 밖으로 잘라내지 않고 온전히 담아낸다. 단, 캐릭터가 이보다 더 거대하여 계산된 비율이 지나치게 낮아질 경우, 시인성 저하를 방지하기 위해 하한선인 0.6으로 강제 고정한다.

수평 카메라 이동의 경우 기존 AutoFlip[11]의 로직을 채택하여 영상의 모션, 에지, 채도 정보를 결합한 중요도 맵을 기반으로 중심점을 산출한다. 이때 카메라가 중심점을 즉각적으로 따라갈 때 발생하는 기계적인 떨림을 방지하기 위해, 현재 위치와 목표 지점 간의 거리에 비례하여 이동 속도를 조절하는 1차 동역학 모델을 적용하였다. 또한, 인접 프레임 간 색상 히스토그램 차이가 임계치를 초과할 경우 이를 새로운 장면으로 판단하여 카메라 상태를 초기화함으로써, 이전 샷의 관성이 다음 샷에 영향을 주지 않도록 처리하였다.

Ⅳ. 실험 결과

1. 정량적 평가

표 1은 입력 프롬프트에 따른 서사 기반 에피소드 매칭의 결과와 해당 에피소드 내에서의 올바른 장면 추출의 정량적 결과이다. 해당 실험을 위해 Gemma-3-27B-IT[12]를 활용하여 데이터셋을 생성하였다. 그림 1과 같이 각 대본 스크립트들은 타임스탬프를 가지는 장면 구간이 존재한다. 각 장면의 핵심 사건과 등장인물을 요약하여 장면별 프롬프트를 생성하였다. 총 10개 에피소드에 대해 에피소드당 10개의 장면별 프롬프트 쿼리를 생성하였다. 이렇게 구성된 프롬프트 Q는 수식 (1)에서 최적의 에피소드를 구함과

표 1. 서사 기반 에피소드 매칭 및 장면 추출 정량적 결과
Table 1. Result of synopsis-based episode matching and script-based scene retrieval

Model	Episode Matching	Segment Extraction
Qwen2-7b-IT[13]	48%	12%
Gemma-3-27B-IT[12]	77%	84%

동시에 수식 (2)에서 장면 추출을 수행하기 위한 입력으로 활용된다. 각 에피소드의 번호와 프롬프트가 생성된 타임라인을 정답 라벨로 활용하였다. 장면 추출의 결과는 예측된 장면 구간이 GT 타임라인을 포함하면 정답으로 예측하였다.

본 실험에서는 상대적으로 경량인 7B급 LLM과 대규모 27B급 LLM을 비교하여, 모델 규모에 따른 에피소드 매칭 및 장면 추출 성능 차이를 정량적으로 제시하였다. 표 1에서 확인할 수 있듯이, Gemma-3-27B-IT[12]는 에피소드 매칭(77%)과 장면 추출(84%) 모두에서 Qwen2-7b-IT[13] 대비 일관된 성능 향상을 보였다. 특히 장면 추출에서는 두 모델 간 격차가 크게 나타났으며(12% → 84%), 이는 컨텍스트 윈도우 길이가 상대적으로 더 큰 모델이 대본 내 장면 단서(사건·인물·대사)를 더 넓은 범위에서 통합적으로 활용할 수 있어, 장면 구간 식별 정확도 향상으로 직접 이어짐을 시사한다.

2. 정성적 평가

그림 2는 제안하는 캐릭터 쿼리 기반 객체 검출과 가상 카메라 시스템을 결합하여 생성한 최종 리프레이밍 결과를 보여준다. 그림 하단에 명시된 텍스트는 각 결과물 생성에 활용된 입력 프롬프트를 나타낸다. 두 영상 결과 모두 9:16 비율의 숏폼 포맷으로 리프레이밍된 이후에도, 주요 캐릭터들이 프레임 밖으로 잘려 나가는 크롭 현상 없이 화면 내에 온전히 보존된 모습을 확인할 수 있다. 또한, 입력된 텍스트 프롬프트의 의미적 맥락에 부합하는 에피소드 구간이 전체 영상에서 정확하게 검색 및 선정되었음을 보여준다.

3. 비교 실험

그림 3은 제안하는 적응형 줌 아웃 기법과 줌 제어가 없는 베이스라인 AutoFlip[11]의 리프레이밍 결과를 비교하여 보여준다. 우선 줌 아웃을 적용하지 않은 베이스라인 결과를 살펴보면, 객체의 크기를 고려하지 않고 단순히 가상 카메라의 중심점 이동에만 의존함에 따라 캐릭터의 일부가

그림 2. 캐릭터 쿼리 로컬라이제이션 기반 최종 리프레이밍 결과
Fig. 2. Final reframing results based on Character Query Localization

숏폼 프레임 밖으로 잘려 나가는 크롭 현상이 발생함을 명확히 확인할 수 있다. 이는 고정된 배율로는 화면 구도가 급격히 변하거나 피사체의 크기가 달라지는 다양한 에피소드 상황에 유연하게 대처할 수 없음을 시사한다.

반면, 적응형 줌 아웃을 적용한 결과는 각 장면의 객체 정보에 맞춰 줌 배율이 자동으로 최적화되므로, 장면 변화

그림 3. 적응형 줌 아웃(Adaptive Zoom-out)과 고정 줌 비율(Fixed Zoom Ratio) 방식의 리프레이밍 결과 비교
Fig. 3. Comparison of reframing results between the proposed Adaptive Zoom-out and a fixed zoom ratio approach

에도 캐릭터를 온전히 보존하는 강건한 성능을 보인다. 실제로 그림 3의 좌측 결과는 캐릭터가 화면을 많이 차지하여 0.6의 강한 줌 아웃이 적용된 반면, 우측 결과는 상대적 여유가 있어 0.699의 약한 줌 아웃이 적용되었다. 이는 시스템이 각 장면의 특성을 분석하여 서로 다른 줌 비율을 산출했음을 보여주며, 결과적으로 두 경우 모두 캐릭터가 숏폼 프레임 내부에 안정적으로 위치하고 있음을 확인할 수 있다.

그림 4는 동일한 비디오와 프롬프트에서 서로 다른 비주얼 쿼리 입력을 사용했을 때의 숏폼 결과를 비교한다. 그림 4는 좌측에서 우측으로 갈수록 더 많은 비주얼 쿼리 입력을 점진적으로 추가한 설정을 나타낸다. 비주얼 쿼리를 사용하지 않는 경우에는 적응형 줌 아웃이 적용되지 않는다. 동일한 시간 축에서 결과를 비교하면, 비주얼 쿼리에 해당하는 객체가 점차 선명하게 유지·보존되는 경향을 확인할 수 있다.

그림 5는 프레임 시퀀스를 시간(T)으로 쌓아 구성한 Spatio-temporal 정육면체에서 가로 $W = 400$을 기준으로 단면을 시간 축으로 쌓아 시각화한 결과이다. 캐릭터의 바운딩 박스의 중심점을 사용하여 리프레이밍 영역을 선택하는 방식은 프레임 간 객체 위치가 흔들리며 시각적 안정성이 떨어진다. 이는 바운딩 박스 좌표의 특성상 조금이라도 좌표가 흔들리거나, 다수의 객체를 포착되어 크롭 된 영역이 이전 및 다음 프레임과 차이가 커질 때 안정성이 떨

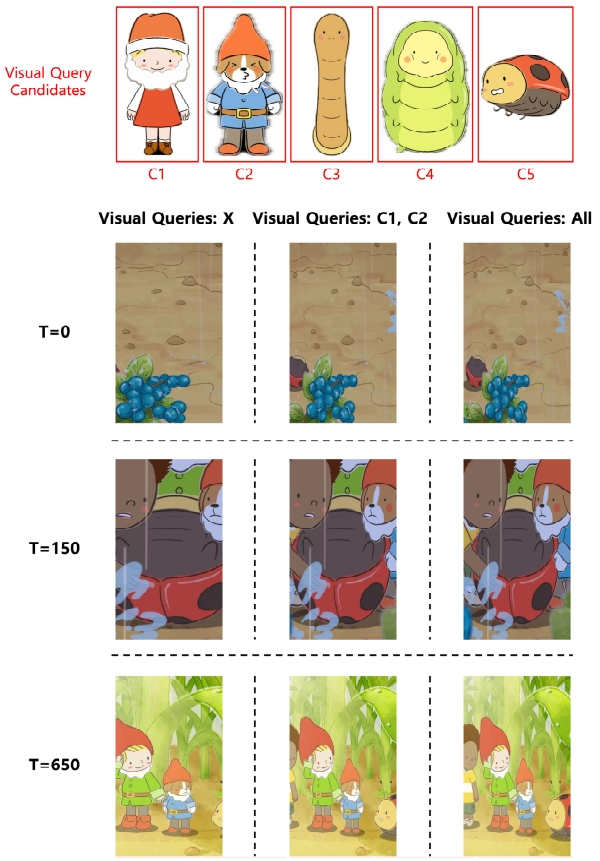

그림 4. 점진적 비주얼 쿼리 입력을 사용하여 생성한 숏폼 결과 비교
Fig. 4. Comparison of short-form videos generated with with progressively increasing visual query inputs

어지기 때문이다. 반면 가상 카메라의 중심점과 바운딩 박스 정보를 함께 사용한 리프레이밍 방식은 시각적 안정성이 높게 유지되는 것을 확인할 수 있다.

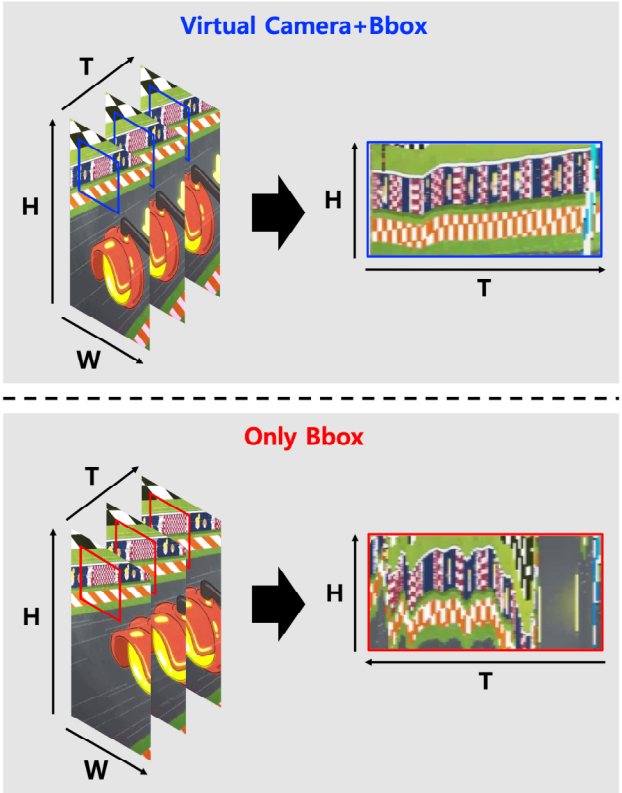

그림 5. 가상 카메라와 바운딩 박스 정보를 사용하여 생성한 숏폼과 바운딩 박스 정보만을 사용하여 생성한 숏폼 비교
Fig. 5. Comparison of short form generated using Virtual Camera and Bounding Box and short form generated using only Bounding Box

V. 결 론

본 연구에서는 애니메이션의 영상의 특유한 시각적 표현 방식으로 인해 기존 MLLMs이 겪던 인식 오류 및 리프레이밍 불안정성 문제를 해결하기 위해 대본 기반 장면 추출과 이미지 기반 객체 검출을 통합한 트레이닝 프리 숏폼 변환 파이프라인을 제안하였다. 제안 방식은 대본을 합친 텍스트 기반 장면 탐색을 통해 의미적으로 핵심적인 구간을 먼저 선정하고 캐릭터 이미지를 비주얼 쿼리로 제공하여 주요 등장 인물의 위치를 추정한 뒤, 적응형 줌 기능을 적용하여 변환 과정에서 캐릭터가 안정적으로 프레임 내에 유지되도록 설계하였다. 제안 방식은 별도의 추가 학습 없이 다양한 애니메이션 콘텐츠에 즉시 적용 가능하며, 입력 텍스트만 변경하여 서로 다른 캐릭터 중심 숏폼 구성을 생성할 수 있다는 장점이 있다. 하지만 사용자 프롬프트에 부합하는 객체에 대한 정답 레이블을 모든 구간·모든 프레임에 대해 구축하는 작업은, 프롬프트 - 객체 대응 관계의 다의성 및 주관성과 애니메이션 영상의 장면별 표현 변동으로 인해 일관된 판정 기준의 정립이 어렵다는 특성상, 대규모 정밀 데이터 구축에 큰 비용이 요구된다. 이로 인해 본 연구에서는 적응적 줌 아웃 과정 전반에 대한 포괄적 정량 정확도 평가를 충분히 제시하지 못하였으며, 이는 본 연구의 한계로 남는다. 향후 연구에서는 프롬프트 조건 하의 객체 정합성에 대한 평가 프로토콜을 정교화하는 한편, 해당 기준을 반영한 프롬프트 숏폼 생성 데이터셋을 구축·공개하는 방향으로 연구 범위를 확장할 예정이다.

참 고 문 헌 (References)

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, “Learning transferable visual models from natural language supervision,” in Proceedings of the International Conference on Machine Learning (ICML), Vol. 139, pp. 8748-8763, Jul. 2021.
doi: https://doi.org/10.48550/arXiv.2103.00020

[2] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” Proceedings of the International Conference on Machine Learning (ICML), Vol.202, pp. 19730 - 19742, Jul. 2023.
doi: https://dl.acm.org/doi/10.5555/3618408.3619222

[3] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, and S. Hoi, “InstructBLIP: Towards general-purpose vision-language models with instruction tuning,” Advances in Neural Information Processing Systems (NeurIPS), Vol. 36, pp. 49250 - 49267, Dec. 2023.
doi: https://dl.acm.org/doi/10.5555/3666122.3668264

[4] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” Advances in Neural Information Processing Systems (NeurIPS), vol. 36, pp. 34892 - 34916, Dec. 2023.
doi: https://doi.org/10.48550/arXiv.2304.08485

[5] M. Maaz, H. Rasheed, S. Khan, and F. Khan, “Video-ChatGPT: Towards detailed video understanding via large vision and language

models," Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Vol.1 : Long Papers, pp. 12585 - 12602, Aug. 2024.
doi: https://doi.org/10.18653/v1/2024.acl-long.679

[6] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," arXiv preprint arXiv:2410.02713, Oct. 2024.
doi: https://doi.org/10.48550/arXiv.2410.02713

[7] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," European Conference on Computer Vision (ECCV), Cham, Switzerland: Springer Nature Switzerland, Vol. 15105, pp. 38 - 55, Sep. 2024.
doi: https://doi.org/10.1007/978-3-031-72970-6_3

[8] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "MDETR: Modulated detection for end-to-end multi-modal understanding," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1780 - 1790, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00180

[9] N. Frey and Z. Sun, "AutoFlip: An Open Source Framework for Intelligent Video Reframing," Google Research Blog, https://research.google/blog/autoflip-an-open-source-framework-for-intelligent-video-reframing/, Feb. 2025.

[10] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "TubeDETR: Spatio-temporal video grounding with transformers," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16442 - 16453, Jun. 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01595

[11] C. Fu, H. Lin, Z. Long, Y. Shen, Y. Dai, M. Zhao, and X. Sun, "VITA: Towards open-source interactive omni multimodal LLM," arXiv preprint arXiv:2408.05211, Aug. 2024.
doi: https://doi.org/10.48550/arXiv.2408.05211

[12] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, and S. Iqbal, "Gemma 3 technical report," arXiv preprint arXiv:2503.19786, Mar. 2025.
doi: https://doi.org/10.48550/arXiv.2503.19786

[13] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. "Qwen technical report," arXiv preprint arXiv:2309.16609, Sep. 2023.
doi: https://doi.org/10.48550/arXiv.2309.16609

저 자 소 개

이 강 희

- 2024년 7월 : Vrije Universiteit Amsterdam BSc in Business Analytics
- 2025년 8월 : 중앙대학교 첨단영상대학원 영상공학 석사
- 2025년 9월 ~ 현재 : 중앙대학교 첨단영상대학원 영상공학 박사과정
- ORCID : https://orcid.org/0009-0000-8931-5583
- 주관심분야 : 컴퓨터 비전, 대형언어모델, 멀티모달 언어 모델

양 해 준

- 2022년 8월 : 단국대학교 체육교육과 학사
- 2024년 8월 : 단국대학교 스포츠사이언스융합학과 석사
- 2025년 3월 ~ 현재 : 중앙대학교 첨단영상대학원 영상학과 AI Imaging 석사과정
- ORCID : https://orcid.org/0009-0002-4609-6872
- 주관심분야 : 컴퓨터 비전, 로봇 비전, 로봇 내비게이션, 강화 학습

저 자 소 개

배 재 형

- 2023년 9월 ~ 2025년 8월 : 중앙대학교 영상학과 석사과정
- 2025년 9월 ~ 현재 : 중앙대학교 영상학과 박사과정
- ORCID : https://orcid.org/0009-0002-4809-327X
- 주관심분야 : 컴퓨터 비전, 의료영상처리, 멀티모달 학습

김 탁 훈

- 2005년 : School of Visual Arts(SVA) 컴퓨터아트학과(미술학석사)
- 2002년 ~ 2007년 : School of Visual Arts(SVA) 애니메이션학과 교수
- 2007년 ~ 현재 : 중앙대학교 첨단영상대학원 영상학과 교수
- 2008년 ~ 현재 : (주)탁툰엔터프라이즈 프로듀서/대표
- ORCID: https://orcid.org/0009-0009-4081-3907
- 주관심분야: 애니메이션, 인공지능(AI), 지식재산권(IP), 프로듀싱(Producing)

최 종 원

- 2008년 3월 ~ 2012년 2월 : KAIST 전기전자공학과 학사
- 2012년 3월 ~ 2014년 2월 : KAIST 전기전자공학과 석사
- 2014년 9월 ~ 2018년 8월 : 서울대학교 전기정보공학과 박사
- 2018년 6월 ~ 2020년 2월 : 삼성SDS 인공지능연구센터
- 2020년 3월 ~ 현재 : 중앙대학교 첨단영상대학원 부교수
- ORCID : https://orcid.org/0000-0001-9753-8760
- 주관심분야 : 컴퓨터 비전, 딥러닝, 딥페이크 탐지, 영상 분석, 로봇 비전