JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 11, pp. 3441-3447, Nov. 2024

Check for updates

비디오 향상/조화 기법을 통한 텍스트 기반 고품질 비디오 편집

김 한 영[1] · 서 가 연[2] · 정 아 영[2] · 김 승 원[3] · 정 희 용[3] · 조 영 준[3*]
[1]전남대학교 인공지능융합학과 석사과정
[2]전남대학교 인공지능학부 학사과정
[3]전남대학교 인공지능융합학과 교수

Text-Based High Quality Video Editing via Video Enhancement and Harmonization

Han-young Kim[1] · Gayun Suh[2] · Ayeong Jeong[2] · Seung-Won Kim[3] · Hieyong Jeong[3] · Yeong-Jun Cho[3*]

[1]Master's Course, Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, Korea
[2]Bachelor's Course, Department of Artificial Intelligence, Chonnam National University, Gwangju 61186, Korea
[3]Professor, Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, Korea

[요 약]

본 연구는 텍스트 기반 비디오 편집과 비디오 조화/향상 기법을 결합한 새로운 프레임워크를 제안한다. 이는 기존 모델의 품질 문제와 편집의 어려움을 해결하기 위한 것으로, 전문 지식 없이도 고품질 비디오 편집이 가능하게 한다. 텍스트 기반 비디오 편집 모델에 비디오 조화 및 향상 기법을 추가하여, 편집된 비디오의 전체적인 품질과 자연스러움을 크게 개선하는 프레임워크를 구성하였다. 이 과정에서 텍스트 기반 의미론적 이미지 분할 기술을 활용하여 프롬프트 내의 핵심 객체와 배경을 효과적으로 구분하여 추가적인 입력 없이 비디오 조화 기법을 적용 가능하게 하였다. 또한, Gradio를 통해 전체 프레임워크를 통합하여 쉽게 배포 및 이용 가능하도록 구현하였다. 이 접근 방식은 비디오 편집 분야에서 사용자 경험을 크게 개선하고 고품질 편집을 더욱 접근 가능하게 만들었다. 사용자는 복잡한 편집 도구를 배우지 않고도 간단한 텍스트 명령으로 원하는 비디오 편집을 수행할 수 있게 하였다. 본 연구는 텍스트 기반 비디오 편집 기술의 발전에 기여하며, 향후 관련 연구의 기초가 될 것으로 기대된다.

[Abstract]

This study proposes a novel framework combining text-based video editing with video harmonization and enhancement techniques. It addresses quality issues and editing difficulties in existing models, enabling high-quality video editing without expert knowledge. The integration of video harmonization and enhancement techniques into a text-based video editing model significantly improves the overall quality and naturalness of edited videos. The process utilizes text-based semantic image segmentation to effectively distinguish between key objects and backgrounds mentioned in the prompt, enabling the application of video harmonization techniques without additional input. Furthermore, the entire framework is implemented through Gradio, making it easily deployable and accessible. This approach greatly enhances the user experience in video editing, making high-quality editing more accessible. Users can perform desired video edits using simple text commands without learning complex editing tools. This research contributes to the development of text-based video editing technology and can serve as a foundation for future related studies.

색인어 : 텍스트 기반 비디오 편집, 조건부 생성 모델, 디퓨전, 비디오 조화, 비디오 향상

Keyword : Text-Based Video Editing, Conditional Generative Model, Diffusion, Video Harmonization, Video Enhancement

http://dx.doi.org/10.9728/dcs.2024.25.11.3441

Received 29 August 2024; **Revised** 26 September 2024
Accepted 07 October 2024

***Corresponding Author; Yeong-Jun Cho**

Tel: +82-62-530-3432
E-mail: yj.cho@jnu.ac.kr

Ⅰ. 서 론

최근, 텍스트-비디오 생성 기술이 발전하면서, 사용자가 텍스트 프롬프트를 정의하는 단순한 과정으로 고품질의 비디오를 생성할 수 있게 되었다. 예를 들어, Diffusion[1] 기반의 SORA[2], Lumiere[3] 등의 모델은 반사, 투영 등의 물리 법칙을 이해하고 생성한다거나, 눈동자의 비친 모습까지 추정하는 등의 고성능의 비디오 생성이 가능하다.

이러한 텍스트-비디오 생성 기술의 발전에도 불구하고, Temporal-aware 정보가 고려되지 않아 프레임 사이의 일관성이 유지되지 않는 flickering 문제, Shape Editing의 어려움, Video Inversion 시 에러 누적으로 인한 원본 손상, 새로운 모션이나 객체로의 수정 한계 등의 문제로 항상 사용자가 만족할 만한 결과를 얻을 수 있는 것은 아니다. 생성된 비디오가 사용자의 의도와 다르게 나온다면, 사용자는 1) 비디오를 처음부터 다시 생성하거나, 2) 생성된 비디오를 편집한다.

1) 비디오를 처음부터 다시 생성하기 위해 컴퓨팅 자원이 많이 필요하기 때문에 자원이 풍부한 대형 서버가 필요하고, 많은 이미지를 생성해야 하기 때문에 탄소가 많이 발생한다. 그림 1에서, Image Generation이 다른 작업에 비해 가장 많은 양의 CO_2를 발생시키는데, 연속된 이미지의 집합인 비디오를 생성할 때는 프레임 수에 비례하는 CO_2가 발생할 것이다. 이처럼 비디오를 재생성하는 작업은 이전에 생성된 비디오에 대한 자원이 모두 낭비되어 비효율적이다.

2) 그렇다면, 비디오를 편집하는 작업은 어떨까? 생성된 비디오를 편집하는 작업은 대형 서버가 필요 없이 상대적으로 적은 컴퓨팅 자원으로도 비디오 편집이 가능하다. 또한, 전체의 프레임 수를 모두 편집하지 않고 편집이 필요한 프레임만 편집을 진행하기 때문에, 처음부터 편집하는 것보다 상대적으로 탄소 발생량이 적다. 이는 비디오를 처음부터 다시 생성하는 것보다는 효율적일 수 있지만, 여전히 여러 한계점을 가지고 있다. 비디오를 직접 편집하기 위해서 전문 편집 도구의 사용법을 알고 있어야 하며, 많은 시간과 노력을 쏟아야 한다. 최근 딥러닝 모델을 활용하여 편집을 진행하는 모델이 많이 개발되고 있으나, 이를 사용하기 위해서는 딥러닝 모델을 다루는 전문 지식이 필요하다. 또한, 딥러닝 모델을 활용한 비디오 편집은 편집 내용 구현의 초점이 맞춰져 있기 때문에, 객체와 배경의 조화나 전체적인 비디오 프레임의 품질을 많이 고려하진 않는다. 따라서, 대부분의 비디오 편집 모델은 편집 후 비디오의 품질이 떨어지는 경향이 있다.

이와 같은 문제를 해결하기 위해, 텍스트 기반의 비디오 편집 모델로 전문 지식이 필요하지 않은 비디오 편집을 진행하면서, 비디오 조화/향상 기법을 적용하여 전체적인 비디오의 조화, 품질에 대해서도 좋은 평가를 받을 수 있는 종합적인 프레임워크를 구현하고 평가하고자 한다.

본 논문의 주요 기여점은 다음과 같다.

1) 전문 지식이 필요하지 않는 비디오 편집

학습된 텍스트 기반의 비디오 편집 모델을 사용하여 전문 지식이 없이 비디오 편집을 가능하게 한다.

2) 비디오 프레임 내의 조화/품질 개선

편집된 비디오에 비디오 조화/향상 기법을 적용하여 객체와 배경을 자연스럽게 조화시키고, 비디오 프레임의 전체적인 품질을 개선한다.

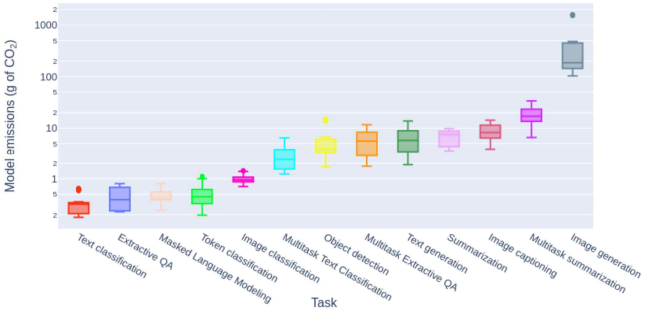


그림 1. 각 작업의 CO_2 발생량 [4]
Fig. 1. CO_2 emissions of each task [4]

Ⅱ. 이론적 배경

2-1 텍스트 기반 비디오 편집

전문 지식이 없는 일반 사용자가 딥러닝을 활용한 비디오 편집을 사용하기 위해서, 텍스트 기반 비디오 편집을 사용한다. 텍스트 기반 비디오 편집은 텍스트 프롬프트와 비디오 원본을 입력으로, 프롬프트의 내용으로 비디오 원본을 수정하는 과정이다. 이 과정에서 비디오의 핵심 프레임이나 원본 비디오를 설명하는 프롬프트를 입력으로 같이 요구하기도 한다. 예시로, Tune-A-Video[5] 모델처럼, 이미지를 입력받아 미리 학습된 텍스트-이미지 생성 모델을 사용하여 비디오를 편집/생성하는 방법이 있다. 본 논문에서는 Fatezero[6] 모델을 사용한다. Fatezero는 spatial-temporal attention block을 사용한 입력 비디오와 텍스트 프롬프트 처리를 통해 비디오를 편집한다. Fatezero는 제로샷 모델임에도 불구하고, 원샷 모델인 Tune-A-Video 모델보다 편집 성능이 좋음을 확인하였다(4-2 참조).

2-2 비디오 조화/향상

비디오 조화(Video Harmonization)와 향상(Video Enhancement)은 비디오의 품질과 관련된 기술이다. 이 두 기술은 편집된 비디오의 품질을 개선하고 자연스러움을 높이는 데 기여한다. 이미지/비디오 조화 기법은 배경과 객체를 마스크로 구분하여, 둘 중 하나의 속성을 조절하여 다른 하나에

자연스럽게 맞추는 기법이고, 이미지/비디오 향상 기법은 전체적인 프레임의 속성을 조절하여 더 자연스러운 프레임을 만드는 기법이다. 비디오 조화/향상은 주로 이미지 조화/향상의 모델을 가져와 사용하는데, 본 논문의 프레임워크에서 비디오 편집에 소모되는 자원이 많기 때문에, 속도가 빠르고 효율성이 높은 모델을 사용하여야 한다. 본 논문의 프레임워크에서 사용한 모델은 Harmonizer[7]로, 사람이 이미지 편집에서 자주 조절하는 필터를 캐스케이드 방식으로 하나씩 조절하면서 이미지/비디오의 품질을 향상시키는 모델이다. 이를 통해 이미지/비디오에서 전체적인 품질을 향상시킬 수 있다.

2-3 텍스트 기반 의미론적 이미지 분할

비디오 조화 기법에는 객체의 분할 마스크가 필요하다. 분할 마스크를 기반으로 배경과 객체를 구분하여, 객체를 배경과 조화롭게 맞춰간다. 본 논문에서 제안한 프레임워크에서, 어떤 객체로 편집이 진행되는지 예측할 수 없고 다양한 객체들이 편집 결과로 등장하기 때문에, 일반적인 분할 모델을 학습시켜 사용하는 것은 불가능하다. 또한, 이미지 분할 모델에서 파운데이션 모델로 여겨지는 Segment Anything[8]을 사용한다고 하더라도, 생성된 마스크의 객체에서 어떤 객체가 비디오의 핵심 객체인지 파악할 수 없다.

이를 해결하기 위해 텍스트 기반 객체 검출기인 GroundingDINO[9]를 결합한 Grounded SAM[10]을 활용하여 비디오 편집에 입력된 텍스트 프롬프트의 핵심 객체로 분할을 진행하여 분할 마스크를 얻는다. 이를 통해 비디오 조화 기법을 사용할 수 있다.

Ⅲ. 제안 방법 및 구현

3-1 제안 방법론

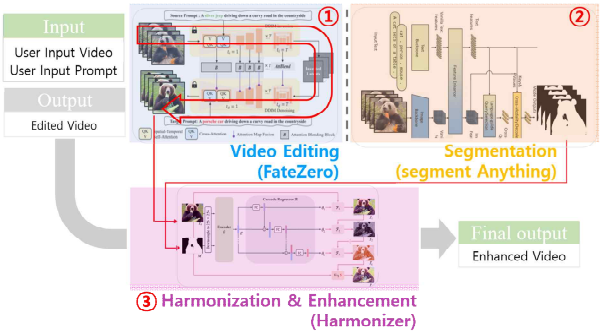


그림 2. 텍스트 기반 비디오 고품질 편집 과정
Fig. 2. Text-based video high-quality editing process

본 연구에서 제안하는 종합적인 방법론은 텍스트 기반 비디오 편집, 텍스트 기반 의미론적 이미지 분할, 그리고 비디오 조화 및 향상 기법을 순차적으로 적용하는 프로세스로 구성된다(그림 2 참조). 우선, 사용자로부터 원본 비디오와 편집을 위한 텍스트 프롬프트를 입력받아 FateZero 모델을 통해 비디오 편집을 수행한다. FateZero는 spatial-temporal attention block을 활용하여 입력 비디오와 텍스트 프롬프트를 효과적으로 처리하며, 이 과정에서 텍스트 프롬프트에 따라 비디오의 내용이 변경된다(그림 2-①, FateZero를 사용한 Video Editing). 다음으로, 편집된 비디오의 각 프레임에 대해 텍스트 기반 의미론적 이미지 분할을 수행한다. 이를 위해 GroundingDINO와 Segment Anything Model을 결합한 Grounded SAM을 사용한다. 본 단계에서는 비디오 편집에 사용된 텍스트 프롬프트의 핵심 객체를 기반으로 각 프레임의 주요 객체에 대한 분할 마스크를 생성하며, 이는 후속 단계인 비디오 조화에 필수적인 입력을 제공한다(그림 2-②, Segment Anything을 활용한 Segmentation). 최종적으로, 비디오 조화 및 향상 기법을 적용한다. 먼저 Harmonizer 모델을 사용하여 비디오 조화를 수행하는데, 이 과정에서는 편집된 비디오와 이전 단계에서 생성된 분할 마스크를 입력으로 사용한다. Harmonizer는 사람이 이미지 편집에서 자주 사용하는 필터들을 캐스케이드 방식으로 순차적으로 적용하여 객체와 배경 간의 조화를 개선한다. 이후 비디오 향상 기법을 적용하여 전체적인 비디오 품질을 개선하는데, 이 단계에서는 밝기, 채도, 명암 등 전반적인 시각적 품질을 향상시킨다.

이러한 방법론은 비디오 편집 기술의 접근성을 크게 높이며, 전문가가 아닌 일반 사용자도 고품질의 비디오 편집 결과를 쉽게 얻을 수 있게 한다.

3-2 구현

배포 및 실험을 위한 종합 프레임워크의 구현에는 Gradio[11]를 사용하였다. Gradio는 머신러닝 모델을 위한 사용자 인터페이스를 신속하게 생성할 수 있는 오픈 소스 Python 라이브러리이다. 이 도구는 복잡한 모델을 직관적인 웹 인터페이스로 변환하여, 개발자와 비개발자 모두가 AI 모델과 쉽게 상호작용할 수 있게 한다. Gradio는 다양한 입력 유형을 지원하며, 코드 몇 줄만으로 데모를 구축할 수 있어 프로토타이핑과 모델 공유를 간소화한다. 3-1의 내용을 총 3단계로 구분하여 순차적으로 진행할 수 있도록 하였다. 알맞은 프롬프트를 입력하고 버튼을 누르는 식으로 각 파트의 결과물을 간단하게 얻을 수 있도록 구현하였다. 각 파트가 종료되면, 사용자가 결과물을 확인하고 다음 파트로 넘어갈지를 결정할 수 있도록 다중 탭 구조로 구현하였다.

먼저 3-2-1)에서 텍스트 기반 비디오 편집에 대한 구현 내용을 설명하고, 3-2-2)에서 텍스트 기반 비디오 분할 구현에 대해 설명하겠다. 마지막으로 3-2-3)에서 비디오 조화/향상 기법을 적용하고 확인하여 최종 결과를 얻어내는 구현에 대해 설명하겠다.

1) 텍스트 기반 비디오 편집

텍스트 기반 비디오 편집을 구현한 모습은 그림 3에서 볼 수 있다. 입력 비디오에서 편집하고자 하는 부분을 단어 형태로 입력하고, 편집하고자 하는 방향도 단어로 입력하면 자동으로 프롬프트를 생성하여 동작하는 방식으로 구현하였다(Prompt Configuration). 비디오를 업로드하고 프롬프트를 통해 설정 파일을 만들면 그 설정을 기반으로 비디오를 편집할 수 있고 그 결과도 확인할 수 있게 구현하였다(Generation, Result).

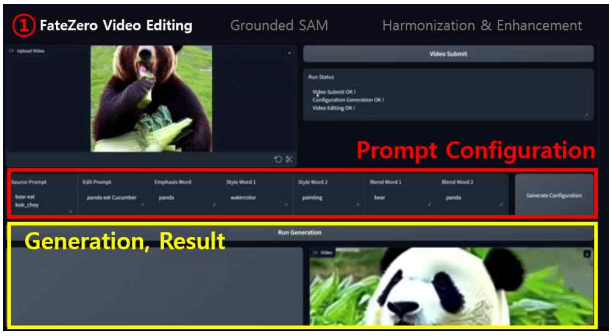


그림 3. 텍스트 기반 비디오 편집 구현
Fig. 3. Implementation of text-based video editing

2) 텍스트 기반 비디오 분할

텍스트 기반의 이미지 분할을 구현은 그림 4에서 확인할 수 있다. 편집된 비디오의 프레임을 추출하여 각 이미지를 텍스트 기반 이미지 분할 모델에 입력, 마스크로 분할하고자 하는 객체를 프롬프트로 입력받는다(Prompt Configuration). 분할 과정이 완료되면 결과를 확인할 수 있다. 이를 통해 비디오 조화 기법에 사용할 분할 마스크 이미지를 얻을 수 있다(Result).

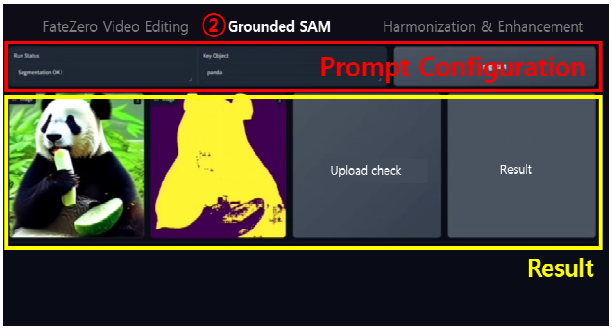


그림 4. 텍스트 기반 이미지 분할 구현
Fig. 4. Implementation of text-based image segmentation

3) 이미지 조화/향상

3-2-1), 3-2-2)에서 얻은 결과로 비디오 조화/향상 기법을 차례대로 적용할 수 있도록 구현하였다(그림 5 참조). 이미지 분할 파트에서 얻은 마스크 이미지들과 비디오로 비디오 조화를 진행한다. 이를 통해 편집된 비디오의 프레임 내에서 객체와 배경의 조화가 더 자연스럽게 조정된다. 또한, 조화 기법을 적용한 비디오에 비디오 향상 기법을 적용하여 프레임 전체의 품질 향상을 달성할 수 있다.

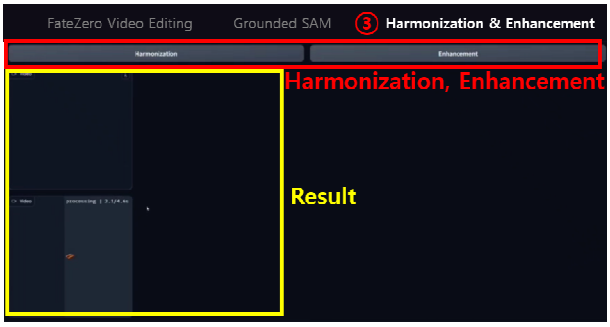


그림 5. 비디오 조화 향상 구현
Fig. 5. Implementation of video harmonization and enhancement

전체가 한 번에 동작하도록 구현하지 않는 이유는, 각 모델의 성능이 불안정할 수 있기 때문이다. 중간 과정의 확인 없이 처음부터 끝까지 한 번의 클릭으로 구현된다면, 서론에서 언급한 자원의 낭비가 비디오 편집과정에서도 발생할 수 있기 때문이다.

이를 통해 비디오를 재생성하는 것보다 자원의 낭비를 줄일 수 있는 텍스트 기반 고품질 편집을 달성하였고, 배포하여 많은 사용자가 간편하게 이용할 수 있도록 구현하였다.

Ⅳ. 실험 및 결과

4-1 정성적 평가

1) 비디오 편집

그림 6. 비디오 편집 적용 결과
Fig. 6. Results of applying video editing

비디오 편집의 결과는 그림 6에서 확인할 수 있다. 그림 6은 14프레임의 원본 비디오와 이를 편집한 비디오의 3, 7, 9, 11, 13번째 프레임을 가져와 비교하는 그림이다. 제로샷 모델임에도 불구하고, 원본 비디오(Source Video)와 원본 비디오의 프롬프트(Source Video Prompt)와 편집 프롬프트(Editing Prompt)를 입력으로 편집을 진행한 결과, 두 편집 결과 모두 편집 프롬프트의 내용대로 충실히 편집이 진행됨을 확인할 수 있다.

2) 비디오 조화/향상

비디오 조화/향상의 결과는 그림 7에서 확인할 수 있다. 그림 7의 각 열은 원본 비디오의 프레임(Source), 편집 결과 비디오의 프레임(Edited output), 비디오 조화/향상을 모두 적용한 비디오의 프레임(Final output)이다. 각 열의 편집 내용은(dog, zebra), (man, wonder woman), (chair, turtle)이다. 원본 비디오에 비해 편집 결과는 육안으로 확인했을 때, 조도 값이 낮아지거나(2행, 3행) 배경과 객체의 조화가 부자연스러운 결과(1행)를 확인할 수 있다. 이를 개선하기 위해, 편집 결과에 비디오 조화/향상 기법을 적용하여 적용 전 결과보다 자연스러운 품질 향상을 달성할 수 있다. 편집 결과(2열)와 조화/향상 기법을 적용한 최종 결과(3열)을 비교해보면, 비디오 조화 기법을 통해 육안으로 보기에 전경(zebra, wonder woman, turtle)과 후경(background)이 더 자연스럽게 구성되는 것을 확인할 수 있다. 또한, 비디오 향상 기법을 통해, 편집 결과에 비해 최종 결과가 밝기, 채도, 명암 등이 육안으로 보기에 더 자연스럽게 조정된 것을 볼 수 있다. 추가적으로, 이는 4-2에서 사용자 연구를 통해 정량적으로 평가하였다.

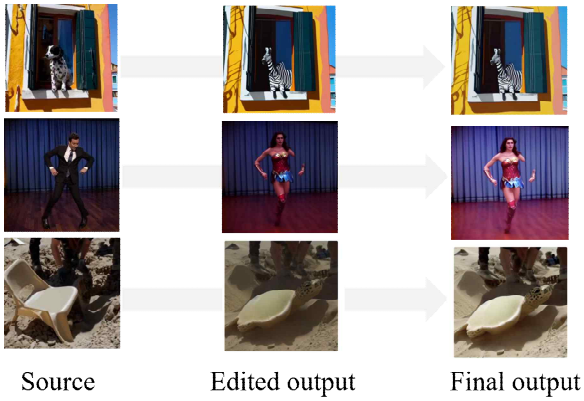


그림 7. 비디오 조화/향상 적용 결과
Fig. 7. Results of applying video harmonization and enhancement

4-2 정량적 평가

평가 척도는 총 6가지로, Tune-A-Video, Fatezero와 비교하였다. Fatezero+H는 비디오 조화 기법까지 적용, Fatezero+H+E는 비디오 조화/향상 기법을 모두 적용한 결과이다. 실험은 총 9개의 비디오를 같은 텍스트 프롬프트로 편집/후처리한 결과로 평가하였다(표 1, 표 2 참조).

표 1. 정량적 평가 결과
Table 1. Quantitative evaluation results

Methods	Temporal Consistency	Text Alignment (CLIP Metrics ↑)	Sucess Rate
Tune-A-Video [5]	0.969	0.86	85.71
FateZero[3]	0.975	0.964	96.42
FateZero[3] + H	0.968	0.984	98.28
FateZero[3] + H + E (Ours)	0.967	0.99	100

표 2. 사용자 연구 결과
Table 2. User study results

Methods	Text Alignment	Structure	Preference
Tune-A-Video [5]	40.7375 %	21.6667 %	48.4125 %
FateZero[3] + H + E (Ours)	59.2623 %	78.3333 %	51.5875 %

Temporal Consistency는 비디오에서 프레임 사이의 일관성을 평가하는 지표로, 모든 모델에서 좋은 성능을 보인다. 비디오 편집에 조화/향상 기법을 추가하는 것이 Temporal Consistency를 해치지 않음을 확인할 수 있다.

Text Alignment는 비디오의 프레임과 텍스트 프롬프트 간의 일치/정렬을 확인하는 지표로, CLIP[12]을 이용하여 평가한다. 본 논문에서 제안한 방법이 가장 좋은 Text Alignment를 얻음을 확인할 수 있다.

Success Rate는 편집 의도를 얼마나 잘 반영했는지에 대한 평가 지표이다. 편집 비디오가 원본 프롬프트에 비해 편집 프롬프트의 내용을 얼마나 충실히 반영하고 있는지에 대한 평가이다. 이 역시 본 논문에서 제안한 방법이 가장 좋음을 확인할 수 있다.

User Study는 사용자에게 3가지 지표를 가지고 평가하는 방법으로 실험을 진행하였다. 평가는 Tune-A-Video로 편집된 비디오와 논문에서 제안한 편집/후처리된 비디오 중에서 각 평가 지표에서 더 좋은 비디오를 고르는 방식으로 진행하였다.

Text Alignment는 두 비디오 중에서 텍스트 프롬프트와 더 가까운 비디오를 고르는 평가로, 본 논문에서 제안한 방식이 더 좋은 결과를 얻었다.

Structure는 원본과 비교하여 구조를 더 잘 유지하고 있는지에 대한 평가이다. 본 논문에서 제안한 방법이 압도적인 차이로 우세함을 알 수 있다.

Preference는 두 비디오 중에서 어느 비디오가 더 선호되는지에 대한 평가로, 본 논문에서 제안한 방법이 근소 우세한 결과를 얻을 수 있었다.

V. 결론 및 후속 연구

본 연구는 텍스트 기반 비디오 편집 기술과 비디오 조화/향상 기법을 결합한 새로운 프레임워크를 제안하였다. 이 프레임워크의 주요 목적은 전문적인 지식 없이도 고품질의 비디오 편집을 가능하게 하는 것이다. 결과적으로, 제안된 프레임워크는 복잡한 편집 도구나 딥러닝 모델에 대한 전문 지식 없이도 텍스트 프롬프트 작성만으로 비디오 편집이 가능하게 함으로써 일반 사용자의 접근성을 크게 향상시켰다. Fatezero 모델을 기반으로 한 텍스트 기반 비디오 편집에 비디오 조화 및 향상 기법을 추가함으로써, 편집된 비디오의 전체적인 품질과 자연스러움을 개선했다. 정량적 평가 결과, 제안된 방법(Fatezero+H+E)은 Temporal Consistency, Text Alignment, Success Rate 등 주요 지표에서 기존 모델들보다 우수한 성능을 보였다. User Study를 통해 본 연구에서 제안한 방법이 Text Alignment, Structure 유지, 그리고 전반적인 Preference 측면에서 기존 모델(Tune-A-Video)보다 우수함을 확인했다.제안된 프레임워크는 Gradio를 통해 구현되어 쉽게 배포 가능하며, 다양한 사용자들이 활용할 수 있는 잠재력을 보여주었다.

후속 연구로, 편집된 비디오의 품질을 자동으로 평가하고 개선점을 제안하는 AI 기반 시스템을 개발하는 것이 유용할 것이다. 이러한 시스템은 사용자의 편집 능력 향상을 돕고, 최종 결과물의 전반적인 품질을 보장하는 데 기여할 수 있을 것이다. 이러한 후속 연구를 통해 본 프레임워크는 더욱 강력하고 사용자 친화적인 도구로 발전하여, 비디오 편집 분야에 긍정적 변화를 가져올 수 있을 것으로 기대된다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(No. RS-2022-00165919) 이며 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2023-RS-2023-00256629).

참고문헌

[1] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS ’20)*, Vancouver, Canada, pp. 6840-6851, December 2020.

[2] OpenAI. Video Generation Models as World Simulators [Internet]. Available: https://openai.com/index/video-generation-models-as-world-simulators/.

[3] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, ... and I. Mosseri, “Lumiere: A Space-Time Diffusion Model for Video Generation,” arXiv:2401.12945, 2024. https://doi.org/10.48550/arXiv.2401.12945

[4] A. S. Luccioni, Y. Jernite, and E. Strubell, “Power Hungry Processing: Watts Driving the Cost of AI Deployment?,” in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*, Rio de Janeiro, Brazil, pp. 85-99, June 2024. https://doi.org/10.1145/3630106.3658542

[5] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, ... and M. Z. Shou, “Tune-a-Video: One-shot Tuning of Image Diffusion Models for Text-to-Video Generation,” in *Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 7589-7599, October 2023. https://doi.org/10.1109/ICCV51070.2023.00701

[6] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen, “FateZero: Fusing Attentions for Zero-Shot Text-Based Video Editing,” in *Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 15886-15896, October 2023. https://doi.org/10.1109/ICCV51070.2023.01460

[7] Z. Ke, C. Sun, L. Zhu, K. Xu, and R. W. H. Lau, “Harmonizer: Learning to Perform White-Box Image and Video Harmonization,” in *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022)*, Tel Aviv, Israel, pp. 690-706, October 2022. https://doi.org/10.1007/978-3-031-19784-0_40

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, ... and R. Girshick, “Segment Anything,” in *Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 3992-4003, October 2023. https://doi.org/10.1109/ICCV51070.2023.00371

[9] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, ... L. Zhang, “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” arXiv:2303.05499, March 2023. https://doi.org/10.48550/arXiv.2303.05499

[10] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, ... and L. Zhang, “Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks,” arXiv:2401.14159, January 2024. https://doi.org/10.48550/arXiv.2401.14159

[11] Gradio. Build & Share Delightful Machine Learning Apps [Internet]. Available: https://www.gradio.app/.

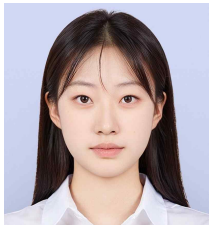
[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, ... and I. Sutskever, “Learning Transferable

Visual Models from Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, Online, pp. 8748-8763, July 2021.

김한영(Han-Young Kim)

2024년 8월 : 전남대학교 인공지능학부 학사(공학)

2024년 9월~현 재: 전남대학교 인공지능융합학과 석사과정
※관심분야 : 이미지 생성 모델



서가연(Gayun Suh)

2021년~현 재: 전남대학교 인공지능학부 학사과정
※관심분야 : 인간-컴퓨터 상호작용

정아영(Ayeong Jeong)

2021년~현 재: 전남대학교 인공지능학부 학사과정
※관심분야 : 텍스트 기반 이미지/비디오 생성

김승원(Seung-Won Kim)

2009년 : University of Tasmania 학사
2011년 : University of Tasmania 석사
2016년 : University of Canterbury 박사

2016년~2020년: University of South Australia 박사 후 연구원
2017년~2020년: Swinburne University of Technology 박사 후 연구원
2020년~2021년: 한국과학기술연구원 박사 후 연구원
2021년~현 재: 전남대학교 인공지능학부 조교수
※관심분야 : 증강/가상현실, 인간-컴퓨터 상호작용, 딥러닝

정희용(Hieyong Jeong)

2009년 : Osaka University 박사(공학)

2009년 4월~2013년 11월: 삼성중공업, 책임연구원
2014년 4월~2019년 8월: Osaka University, 부교수
2019년 9월~현 재: 전남대학교, 교수
※관심분야 : 헬스케어 데이터 및 시스템, 동작분석, 엣지컴퓨팅

조영준(Yeong-jun Cho)

2014년 : 광주과학기술원 정보통신공학과 석사 졸업
2018년 : 광주과학기술원 정보통신공학과 박사 졸업

2018년~2020년: 현대모비스 데이터사이언스팀 책임
2020년 9월~2024년 8월: 전남대학교 소프트웨어공학과 조교수
2024년 9월~현 재: 전남대학교 소프트웨어공학과 부교수
※관심분야 : Computer Vision, Deep Learning