www.ksii.or.kr

2026년
추계학술 발표대회

10월 22일(목) – 10월 24일(토)

세인트존스호텔 (강원도 강릉시)

주최 · 주관

사단법인 한국인터넷정보학회

후 원

㈜나루씨큐리티, ㈜피씨엔, ㈜아이티센엔텍, 한화시스템㈜,
㈜LIG시스템, 케이투웹테크㈜, 강원관광재단, 강릉관광개발공사



www.ksii.or.kr
KSII
사단법인 KOREAN SOCIETY FOR INTERNET INFORMATION 한국인터넷정보학회

논문발표

10월 23일(금) 10:40~12:10 (Room 5 / 바부다 4)

세션 I	좌장: 김영철 교수(홍익대)

AI-Metaverse

I-1 가상 융합 쇼핑몰에서의 테스트 에이전트 기반 자동 테스트 메커니즘
정현욱, 김장환, 김영철(홍익대)

I-2 가상 융합 전시관 내 시각·공간 문맥 활용 기반 고도화 RAG 지향 AI 도슨트 구현 사례
김인용, 김장환, 서채연(홍익대), 전병국(강원대), 김영철(홍익대)

I-3 사용자 의도와 생성형 AI 결과의 일치도 평가를 위한 개념 정의 및 평가방법
정연주, 서채연, 김영철(홍익대)

I-4 사용자 선호도 메트릭스 기반 Text-to-Video 생성 결과 평가 방안
정연주, 서채연, 김영철(홍익대)

I-5 LLM기반 AI챗봇 자가 치유 메커니즘
김성회, 서채연, 김영철(홍익대)

10월 23일(금) 10:40~12:10 (Room 6 / 카리브 2)

[워크숍] XR 콘텐츠 자율변형 프레임워크 기술 세션 3	자체 진행

10월 23일(금) 10:40~12:10 (Room 7 / 카리브 1)

육사 AI R&D센터 특별섹션 세미나 1	좌장: 이준 교수(육사)

사용자 선호도 메트릭스 기반 Text-to-Video 생성 결과 평가 방안

정연주[1], 서채연[2], 김영철[3*]
[1]홍익대학교 스마트도시과학경영대학원
[2,3*]홍익대학교 소프트웨어융합학과
e-mail: hello.yeonjujung@gmail.com, bob@hongik.ac.kr

Evaluating Text-to-Video Generation Results Based on User Preference Metrics

Yeonju Jung[1], Chaeyun Seo[2], R. Young Chul Kim[3*]

1. 서 론

텍스트-영상 생성(Text-to-Video)은 사용자의 자연어 입력을 영상으로 변환한다. 하나의 생성 영상에는 대상, 속성, 행동, 움직임, 공간관계, 촬영방식 등 서로 다른 요소가 동시에 포함된다. 따라서 사용자가 원하는 영상이 생성되었는지를 평가하려면 영상 전체의 품질뿐 아니라 사용자가 요구한 조건이 실제 결과에 반영되었는지를 구분하여 확인해야 한다.

생성형 AI는 동일한 입력에서도 서로 다른 결과를 생성하기 때문에, 생성 결과의 품질만으로 사용자가 원하는 결과가 반영되었는지를 판단하기 어렵다. Human Preference Score는 기존 IS와 FID 등의 지표가 실제 사용자 선택과 충분히 일치하지 않는다는 점을 확인하고, 동일한 프롬프트에서 생성된 여러 결과에 대한 사용자 선택을 기반으로 Human Preference Score를 제안하였다[5]. ImageReward는 실제 사용자 프롬프트와 생성 결과를 대상으로 이미지-텍스트 일치도(Image-Text Alignment), 충실도(Fidelity), 무해성(Harmlessness)을 평가한 뒤 생성 결과의 선호 순위를 비교하였다[6]. 이러한 연구는 생성형 AI 결과를 평가할 때 모델의 공통 성능뿐 아니라 사용자가 어떤 결과를 더 적합하다고 판단하는지도 함께 고려할 필요가 있음을 보여준다.

본 연구에서는 사용자 선호도(User Preference)를 사용자가 생성 결과에서 중요하게 판단하는 조건으로 한정한다. 이는 본 연구에서 평가 절차를 구성하기 위한 조작적 정의이며, 사용자가 원하는 결과를 생성 결과에서 확인할 수 있는 사용자 요구사항(User Requirement)으로 구체화한다. 이후 각 요구사항과 관련된 영상 평가요소를 대응시키고, 해당 요소가 실제 생성 결과에 반영되었는지를 비교한다.

기존 Text-to-Video 평가연구는 사전에 정의된 공통 평가 차원을 중심으로 생성 결과를 평가한다. 본 연구는 이러한 평가요소를 모든 생성 결과에 동일하게 적용하는 대신, 사용자가 원하는 결과를 요구사항 단위로 구분하고 각 요구사항과 직접 관련된 영상 평가요소를 선택하여 생성 결과와 대응시키는 평가 절차를 제안한다. 이를 통해 기존의 모델 및 영상 품질 중심 평가와 구분하여, 사용자 요구사항을 기준으로 생성 결과의 반영 여부를 평가할 수 있는 방법을 구체화했다는 점에서 차이가 있다. 또한 하나의 요구사항에 여러 평가요소가 포함된 경우 이를 구분하여 비교함으로써, 요구사항의 일부 조건만 생성 결과에 반영된 경우도 확인할 수 있도록 하였다. 이러한 절차는 생성 영상 전체에 하나의 평가값을 부여하는 방식과 달리, 사용자 요구사항별로 결과의 반영 여부를 확인할 수 있도록 한다.

2. 관련 연구

VBench는 영상 생성 품질을 Video-Condition와 Video Quality, Consistency로 구분하고, 대상·배경 일관성, 움직임 부드러움, 객체, 행동, 색상, 공간관계, 장면 등 16개의 평가 차원으로 세분화하였다[1]. EvalCrafter는 실제 사용자 프롬프트에서 메타정보를 추출하고, 영상 품질, 텍스트-영상 정렬, 움직임 품질, 시간적 일관성을 평가한 뒤 객관적 평가결과와 사용자 평가를 비교하였다[2].

T2V-CompBench는 복합 프롬프트를 속성, 행동, 움직임, 공간관계, 객체 상호작용, 개수 등의 단위로 구분하여 각 요소를 개별적으로 평가한다[3].

CineTechBench는 촬영 크기, 촬영 각도, 구도, 카메라 이동, 조명, 색상, 초점거리의 7개 촬영기법 차원을 정의하였다[4].

Human Preference Score는 동일한 프롬프트에서 생성된 여러 결과에 대한 실제 사용자 선택과 기존 자동 평가 지표의 차이를 분석하였고[5], ImageReward는 이미지-텍스트 일치도, 충실도, 무해성 등을 평가한 뒤 생성 결과의 선호 순위를 비교하였다[6].

본 연구는 모든 평가요소를 일괄 적용하지 않고, 사용자가 원하는 결과를 요구사항으로 구분한 뒤 각 요구사항과 직접 관련된 평가요소만 선택하여 생성 결과와 비교하는 절차로 재구성한다.

3. 사용자 선호도 메트릭스 기반 영상 평가방법

(표 1) Text-to-Video 관련 선행연구의 영상 평가요소

구분	평가요소
내용·관계	대상, 속성, 공간관계, 행동, 객체 상호작용, 개수
움직임·시간	움직임, 부드러움, 대상·배경 일관성, 시간적 변화
촬영·표현	화면 크기, 촬영구도, 카메라이동, 조명, 색상, 초점거리

본 연구는 먼저 Text-to-Video 관련 선행연구에서 사용된 영상 평가요소를 추출하였다. 선행연구에서 서로 다른 용어를 사용하더라도 실제 평가대상과 평가내용이 동일하거나 유사한 경우 같은 범주로 구분하였다. VBench와 T2V-CompBench는 대상, 속성, 행동, 공간관계 및 움직임 등을 세부 평가단위로 구분하고 있으며[1][3], CineTechBench는 촬영과 표현에 관한 요소를 촬영기법의 세부 차원으로 제시한다[4]. 본 연구에서는 이러한 평가요소를 사용자가 원하는 생성 결과와 대응시키기 위해 그림 1과 같이 네 단계로 평가 절차를 구성하였다.



(그림 1) 사용자 선호도 메트릭스 기반 Text-to-Video 평가 절차

3.1 사용자 요구 및 원하는 결과 정의

첫 번째 단계에서는 사용자가 생성 결과에서 원하는 조건을 구분한다. 하나의 자연어 입력에는 대상, 속성, 장면, 행동, 움직임, 촬영 및 표현 조건이 함께 포함될 수 있다. 따라서 이를 생성 결과에서 각각 확인할 수 있는 요구사항 단위로 나눈다. 본 연구의 예시에서는 사용자 요구사항을 아래와 같이 구분한다. 이 단계에서는 평가값을 부여하지 않고 사용자가 결과에서 확인하려는 조건을 정의한다.

(표 1) Text-to-Video 관련 선행연구의 영상 평가요소

구분	내용
사용자 요구사항	빨간 우산을 든 여성이 비 오는 골목을 천천히 걸어가고, 뒤에서 따라가며 촬영하며, 여성에게 가까워지며 컬러가 선명해진다
확인된 요구사항	빨간 우산을 든 여성 / 비 오는 골목 / 천천히 걸어감 / 뒤에서 따라감 / 여성에게 가까워지며 컬러가 선명해짐

3.2 사용자 요구사항과 영상 평가요소 매핑



(그림 2) 사용자 요구 사항과 영상 평가요소 매핑 예시

두 번째 단계에서는 각 사용자 요구사항과 직접 관련된 영상 평가요소를 대응시킨다. 하나의 요구사항이 두 개 이상의 평가요소를 포함하는 경우 각 요소를 구분하여 대응시킨다. 예를 들어 '뒤에서 따라감'은 인물을 뒤에서 촬영했는지를 확인하는 촬영 각도와 카메라가 인물을 따라 이동했는지를 확인하는 카메라 이동으로 나누어 평가한다. 이 과정에서 모든 영상 평가요소를 사용하는 것이 아니라, 해당 사용자 요구사항을 확인하는 데 필요한 요소만 선택한다.

3.3 평가요소 단위 생성 결과 비교

세 번째 단계에서는 매핑된 평가요소와 생성 결과를 각각 비교한다. 생성 영상 전체에 대한 평가값을 먼저 부여하지 않고, 사용자 요구사항에 대응되는 각 평가요소가 생성 결과에 반영되었는지를 개별적으로 판단한다. 예를 들어 생성 영상에서 여성을 뒤에서 촬영했지만 카메라가 여성을 따라 이동하지 않았다면, 촬영 각도는 사용자 요구와 대응되지만 카메라 이동은 대응되지 않는 것으로 구분한다. 같은 방식으로 대상, 속성, 장면, 행동, 움직임, 촬영 및 표현 요소를 각각 비교한다. 평가요소를 분리하여 비교하면 하나의 요구사항에 포함된 일부 조건만 충족된 경우도 구분할 수 있다.

3.4 요구사항별 평가 결과 도출

마지막 단계에서는 평가요소별 비교 결과를 다시 사용자 요구사항 단위로 정리한다. 하나의 요구사항에 여러 평가요소가 대응된 경우 각 요소의 결과를 함께 확인하여 해당 요구사항의 반영 여부를 구분한다. 이를 통해 생성 영상 전체에 하나의 평가값을 부여하기보다, 사용자가 원하는 조건 가운데 어떤 내용이 생성 결과에 반영되었고 어떤 내용이 반영되지 않았는지를 요구사항별로 확인할 수 있다. 본 연구에서는 이 단계까지의 평가 절차를 제안하며, 요구사항 간 중요도나 가중치를 적용한 종합점수는 산정하지 않는다.

4. 결론 및 향후 연구

본 연구에서는 Text-to-Video 생성 결과에서 사용자가 원하는 조건의 반영 여부를 평가하기 위해 사용자 요구사항과 영상 평가요소를 매핑하고, 이를 생성 결과와 비교하는 방법을 제안하였다. 기존 영상 생성 평가연구에서 사용된 평가요소를 추출하되 모든 요소를 동일하게 적용하지 않고, 사용자 요구사항과 직접 대응하는 평가요소를 선택하도록 평가 절차를 구성하였다.

제안한 절차는 사용자 요구 및 원하는 결과 정의, 영상 평가요소 매핑, 평가요소 단위 생성 결과 비교, 요구사항별 평가 결과 도출의 네 단계로 구성되며, 사용자가 요청한 개별 조건의 반영 여부를 요구사항별로 확인한다.

본 연구의 기여는 기존 Text-to-Video 평가연구의 영상 평가요소를 사용자 요구사항과 대응시키고, 사용자가 요구한 개별 조건을 기준으로 생성 결과를 평가할 수 있도록 평가 단위를 재구성한 데 있다. 기존의 공통 평가요소를 일괄 적용하는 방식과 달리 사용자 요구에 따라 필요한 평가요소를 선택한다.

본 연구에서는 평가요소별 최종 척도, 중요도, 가중치 및 종합점수 계산방법을 확정하지 않았다. 향후 연구에서는 요구사항별 판정 기준을 구체화하고, 동일한 사용자 요구사항과 평가요소 매핑을 여러 평가자에게 적용하여 판정 결과의 일치 정도를 검증할 예정이다.

참고문헌

[1] Z. Huang et al., "VBench: Comprehensive Benchmark Suite for Video Generative Models," in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21807 - 21818, 2024.

[2] Y. Liu et al., "EvalCrafter: Benchmarking and Evaluating Large Video Generation Models," in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22139 - 22149, 2024.

[3] K. Sun et al., "T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-Video Generation," in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8406 - 8416, 2025.

[4] X. Wang et al., "CineTechBench: A Benchmark for Cinematographic Technique Understanding and Generation," Advances in Neural Information Processing Systems, 2025.

[5] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Human Preference Score: Better Aligning Text-to-Image Models with Human Preference," in Proc. of the IEEE/CVF International Conference on Computer Vision, pp. 2096 - 2105, 2023.

[6] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation," Advances in Neural Information Processing Systems, Vol. 36, 2023.