한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 32, No. 3, P. 173~183

WebVR180: 양안 일관성 및 시각 품질 정제 기반 VR180 영상 데이터셋

최창수° 정석현 구한슬 이정진*

숭실대학교

{choics2001, jsh19444, haaanseul}@soongsil.ac.kr, jungjinlee@ssu.ac.kr

WebVR180: VR180 Video Dataset with Stereoscopic Consistency and Visual Quality

Changsu Choi° Seokhyun Jeong Hanseul Koo Jungjin Lee*

Soongsil University

요 약

최근 영상 생성 기술의 발전과 몰입형 콘텐츠 수요의 증가는 VR180과 같은 입체 영상 콘텐츠를 생성하기 위한 학습 데이터의 필요성을 높이고 있다. 그러나 VR180 영상 생성 모델의 학습에는 좌·우 시점의 안정적인 대응 관계와 전방 반구 영역의 투영 특성을 반영한 대규모 영상-텍스트 데이터가 요구되며, 기존 데이터셋만으로는 이러한 조건을 충분히 충족하기 어렵다. 본 연구는 VR180 영상 생성 모델 학습을 위한 대규모 고해상도 영상 클립-캡션 데이터셋 WebVR180을 제안한다. 제안 데이터셋은 웹에서 수집한 대규모 VR180 영상을 바탕으로 구축되었다. 이를 대상으로 깊이 범위, 색상 불일치, 기하 정렬 오차, 선명도 불일치, 좌·우 시점 대응 이상의 5가지 지표를 통해 양안 구조 유효성을 엄격히 검증하고, 주변부 비네팅, 밝기 분포, 화면 내 텍스트 포함 여부, 미적 품질을 종합적으로 평가하는 다중 지표 기반의 단계적 정제 절차를 적용하였다. 정제된 클립에는 장면 내용과 시간적 변화를 반영한 고밀도 캡션을 부여하였으며, 최종적으로 2048 × 1024 해상도의 VR180 영상 클립-캡션 쌍 396,597개를 확보하였다. 정제 단계별 학습 데이터를 이용해 영상 생성 모델을 파인튜닝한 비교 실험에서, 최종 정제 데이터로 학습한 모델은 보고된 모든 양안 불일치 지표에서 가장 낮은 오류를 보였다. 또한 WebVR180의 단안 ERP180 표현 데이터를 이용해 최신 대규모 비디오 생성 모델을 파인튜닝하고, 생성 결과를 원근 시점으로 변환하여 분석함으로써 다양한 영상 생성 모델 및 실제 시청 관점의 결과 분석에 대한 활용 가능성을 확인하였다.

Abstract

Recent advances in video generation and growing demand for immersive content highlight the need for VR180-specific training data. However, training VR180 video generation models requires large-scale video-text pairs that reflect stable binocular correspondence and hemispherical projection characteristics, which existing datasets do not sufficiently provide. We introduce WebVR180, a large-scale, high-resolution video clip-caption dataset for training VR180 video generation models. Sourced from VR180 videos collected from the web, the dataset is refined through a staged multi-metric curation pipeline. This pipeline rigorously evaluates stereoscopic structural validity across five key metrics—depth range, color mismatch, geometric alignment error, sharpness mismatch, and left–right view correspondence anomalies—while further assessing peripheral vignetting, brightness distribution, text overlays, and aesthetic quality. The curated clips are annotated with dense captions describing scene content and temporal changes, resulting in 396,597 VR180 clip-caption pairs at 2048×1024 resolution. Curation-stage ablations show that the model trained on the fully curated data achieves the lowest errors across all reported stereoscopic mismatch metrics. Furthermore, fine-tuning large-scale video generation models with the monocular ERP180 representations of WebVR180 and evaluating perspective-reprojected outputs demonstrate its applicability across different models and practical viewing perspectives.

키워드: VR180, 영상 생성, 영상 데이터셋
Keywords: VR180, Video Generation, Video Dataset

*corresponding author: Jungjin Lee / Soongsil University (jungjinlee@ssu.ac.kr)

Received : 2026.05.29./ Review completed : 1st 2026.06.07. 2nd 2026.06.25./ Accepted : 2026.06.29.
DOI : 10.15701/kcgs.2026.32.3.173
ISSN : 1975-7883(Print)/2383-529X(Online)

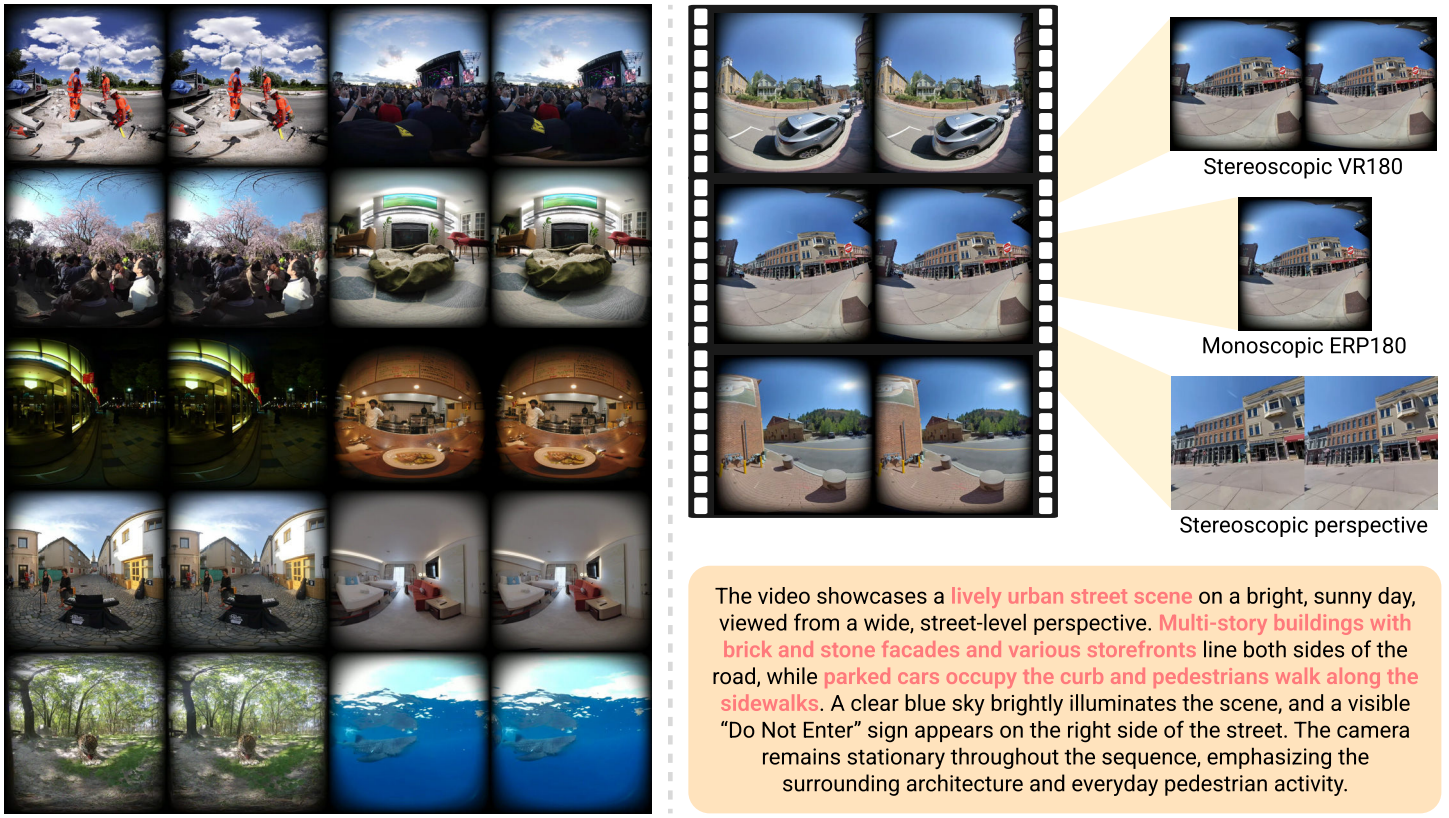


Figure 1: WebVR180 provides diverse stereoscopic VR180 video clips paired with dense captions describing scene content and temporal camera motion. The left panel presents representative samples from the dataset, while the right panel illustrates a video clip in stereoscopic VR180 representation, its monoscopic ERP180 view, and stereoscopic perspective views, along with its text prompt.

1. 서론

최근 몰입형 미디어와 실감형 콘텐츠 수요가 확대되면서, 몰입감 있는 영상 표현 방식의 중요성이 높아지고 있다. 넓은 시야와 입체감을 제공하는 형식인 VR180은 공연, 여행, 전시 및 가상 체험처럼 시야 범위와 입체 표현이 중요한 콘텐츠에 활용될 수 있다. 최근 영상 생성 모델의 발전은 원근 영상 중심의 제작 범위를 넘어, VR180과 같은 몰입형 영상 콘텐츠 제작에도 활용될 수 있는 기술적 기반을 마련하고 있다. 그러나 VR180 영상 생성 모델을 학습하려면 일반 2D 영상이나 단안 전방위 영상과 달리, 좌·우 시점 간의 안정적인 대응 관계와 전방 180° 광시야각 표현이 함께 유지된 데이터가 요구된다.

기존 대규모 영상-텍스트 데이터셋은 주로 단안 원근 영상을 대상으로 구축되어 VR180에서 요구되는 양안 구조와 전방 반구 시야각 표현을 반영하지 못한다. 양안 및 파노라마 영상 데이터셋 또한 각각 원근 투영 기반의 양안 표현과 360° 전방위 표현에 초점을 두기 때문에, 등장방형도법 기반의 전방 180° 양안 영상과 텍스트 조건을 함께 요구하는 VR180 영상 생성 모델 학습에 직접 활용하기에는 한계가 있다 [1, 2, 3, 4, 5, 6]. 한편 웹 수집 VR180 영상은 촬영 및 후처리 조건이 일정하지 않아 좌·우 시점의 정렬, 색상, 선명도, 주변부 표현 차이가 발생할 수 있으며, VR180 형식만으로 학습 데이터에 적합한 양안 구조와 시각 품질이 보장되지는 않는다 [7]. 이러한 이유로 텍스트 조건 기반 VR180 영상 생성 학습을 위해서는 도메인 특성과 품질 요건을 반영한 대규모 영상-텍스트 데이터 수집 및 정제 과정이 필요하다.

이에 본 연구는 VR180 영상 생성 모델 학습을 위한 대규모 · 고해상도 영상 클립-캡션 데이터셋 WebVR180을 제안한다 (Fig. 1). WebVR180은 YouTube에서 수집한 VR180 영상 후보군 31,625개를 장면 단위 클립으로 구성하고, 각 클립에 고밀도 캡션을 생성하여 구축하였다. 데이터 정제 과정에서는 대규모 영상 생성 모델에서 활용된 학습 데이터 구성 및 시각 품질 선별 절차와 VR180 양안 품질 평가 지표를 각각 참고하였다 [7, 8, 9]. 이를 바탕으로 양안 유효성 판별, 주변부 비네팅 검증, 밝기 분포, 텍스트 검출, 미적 품질 평가를 정제 항목으로 구성하고, VR180 맞춤형 정제 파이프라인으로 통합하였다. 그 결과 2048×1024 해상도의 VR180 영상 클립 396,597개와 이에 대응하는 캡션을 확보하였다. 또한 정제 단계별 파인튜닝과 다양한 영상 생성 모델 적용 실험을 통해 제안 데이터셋의 정제 효과와 모델 적용성을 확인하였다.

본 연구의 주요 기여는 다음과 같다. (1) 텍스트 조건 기반 VR180 영상 생성 모델 학습을 위해, 등장방형도법으로 표현된 전방 180° 고해상도 양안 영상 클립과 고밀도 캡션으로 구성된 대규모 영상 클립-캡션 데이터셋 WebVR180을 구축하였다. (2) VR180 영상의 학습 적합성을 확보하기 위해, 양안 구조 유효성, 입체 깊이 관계, 주변부 비네팅 및 장면 단위 시각 품질을 함께 고려하는 단계적 데이터 정제 절차를 제안하였다.

Table 1: Comparison of representative stereoscopic, panoramic, and VR180-related video datasets. Our dataset provides 396K high-resolution stereoscopic 180° ERP(Equirectangular Projection) clips paired with captions for VR180 video generation.

Dataset	Stereo	Projection	Resolution	Caption	Scale
WSVD [1]	✓	Perspective	Mixed	–	10.8K
UniStereo [3]	✓	Perspective	832×480	✓	103K
WEB360 [4]	–	360° ERP	1024×512	✓	2.1K
HELVIPAD [5]	✓	360° ERP	1920×512	–	40K frames
Stereo4D [6]	✓	180° ERP	Mixed	–	100K+
Ours	✓	**180° ERP**	**2048×1024**	✓	**396K**

2. 관련 사례 및 연구

2.1 양안 및 VR180 영상 품질 평가

입체 영상에서는 좌 · 우 시점 간 시차가 일관된 깊이 정보로 해석될 때 입체감이 형성되므로, 두 시점 사이의 기하 정렬이 어긋나거나 색상과 선명도가 불균형하면 시청 품질이 저하될 수 있다. 전방 반구 영역을 양안으로 기록하는 VR180 영상에서는 이러한 양안 품질 문제와 함께 넓은 시야각 및 투영 형식에서 비롯되는 특성도 고려해야 한다.

Lavrushkin et al.은 YouTube VR180 영상 1,000개에 8가지 객관 품질 지표를 적용하여, 기하 정렬, 색상, 선명도, 채널 대응 등 양안 품질 문제가 관찰됨을 보였다 [7]. 이는 플랫폼상에서 VR180 형식으로 제공되더라도, 양안 구조와 시각 품질이 학습 데이터에 적합한지는 별도로 검증해야 함을 시사한다. 본 연구는 이러한 문제의식과 양안 품질 평가 항목을 참고하여 VR180 영상의 구조적 특성을 반영한 정제 기준을 구성하고, 이를 대규모 영상-텍스트 데이터셋 구축을 위한 정제 파이프라인으로 통합한다.

2.2 양안 및 광시야각 영상 데이터셋

VR180 영상은 좌 · 우 시점 정보와 전방 반구 시야각을 함께 포함하므로, 기존의 양안 영상 데이터셋 및 광시야각 영상 데이터셋 연구와 관련된다. Table 1은 관련 데이터셋을 양안 정보 제공 여부, 투영 형식, 해상도, 캡션 제공 여부, 규모 및 활용 목적에 따라 비교한 결과를 보여준다.

양안 영상 분야에서는 WSVD [1], SVD [2], UniStereo [3] 등이 깊이 추정, 3차원 영상 처리, 입체 변환 연구에 활용되어 왔다. 광시야각 영상 분야에서는 360° 등장방형도법 파노라마 영상 생성을 위한 WEB360 [4]과 전방위 입체 깊이 추정을 위한 HELVIPAD [5]가 제안되었다. 이러한 데이터셋들은 양안 시점 정보 또는 넓은 시야각 표현을 제공하지만, 일반 양안 영상이나 360° 전방위 영상을 중심으로 구성되어 VR180 형식의 영상 생성 학습과는 차이가 있다.

한편 Stereo4D는 YouTube에서 수집한 VR180 양안 영상을 기반으로 동적 3차원 장면의 4D 재구성 데이터를 구축한 데이터셋이며, 재구성 품질을 확보하기 위한 정제 및 보정 절차를 포함한다 [6]. 그러나 이러한 절차는 카메라 포즈, 깊이, 움직임 궤적 등 4D 재구성 정보의 안정적 추정을 목적으로 하며, 텍스트 조건 VR180 영상 생성 학습에 필요한 장면 단위 영상-텍스트 구성이나 학습용 샘플 선별 관점의 양안 구조 및 시각 품질 검증을 목적으로 하지는 않는다. 본 연구는 이러한 한계에 주목하여, VR180 영상을 장면 단위의 고해상도 클립-캡션 쌍으로 구성하고, 양안 구조 유효성과 시각 품질을 함께 검증하는 정제 절차를 통해 텍스트 조건 VR180 영상 생성 모델 학습에 활용 가능한 대규모 클립-캡션 데이터셋을 구축한다.

2.3 영상 생성 모델

최근 영상 생성 모델은 텍스트, 이미지, 영상 등 다양한 조건 입력을 활용하여 시각적 품질과 시공간적 일관성을 향상시키는 방향으로 발전해왔다. 특히 U-Net 기반 확산 모델을 Transformer 백본으로 대체한 Diffusion Transformer 계열은 공간적 · 시간적 의존관계 학습과 모델 규모 확장에 유리하여, Wan과 HunyuanVideo와 같은 공개 대규모 영상 생성 모델의 기반으로 활용되고 있다 [8, 9]. 본 연구에서는 구축한 VR180 클립-캡션 데이터셋을 이용해 Wan2.1에서 데이터 정제 수준에 따른 학습 효과를 비교하고, Wan2.2 및 HunyuanVideo1.5 모델에 대한 적용 결과를 통해 제안 데이터셋의 모델 간 활용 가능성을 검증한다.

3. VR180 영상 기반 클립-캡션 데이터 구성

3.1 VR180 데이터 수집

본 연구에서는 대규모 VR180 데이터를 확보하기 위해 YouTube에서 영상을 수집하였다. 수집 영상의 주제 편향을 줄이기 위해 YouTube-8M [10]의 영상 라벨 중 빈도 상위 250개 항목을 검색어로 사용하고, 각 검색에는 VR180 검색 필터를 적용하였다. 이를 통해 여행, 공연, 스포츠, 자연환경, 실내 공간 등 다양한 장면과 촬영 환경을 포함하는 VR180 영상 후보군을 확보하였다.

VR180 검색 필터를 적용하더라도 검색 결과에는 2D-to-3D 변환 등으로 생성된 가상 양안 콘텐츠가 포함될 수 있으므로, 제목과 설명 등의 메타데이터를 활용해 이를 초기 제거하였다. 또한 서로 다른 검색 키워드로 반복 수집된 영상은 중복 제거하여, 최종적으로 31,625개의 VR180 영상 후보군을 구성하고 후속 클립 생성 및 정제 과정에 활용하였다.

3.2 장면 단위 영상 클립 생성

수집된 VR180 영상은 길이가 일정하지 않으며, 하나의 영상 안에도 장면 전환이나 촬영 구간 변화가 포함될 수 있다. 이러한 영상을 학습 샘플로 사용하기 위해 PySceneDetect의 AdaptiveDetector [11]로 원본 영상의 샷 경계를 검출하고, 검출된 구간을 장면 단위 클립으로 분리하였다. 각 클립은 영상 생성 모델의 입력

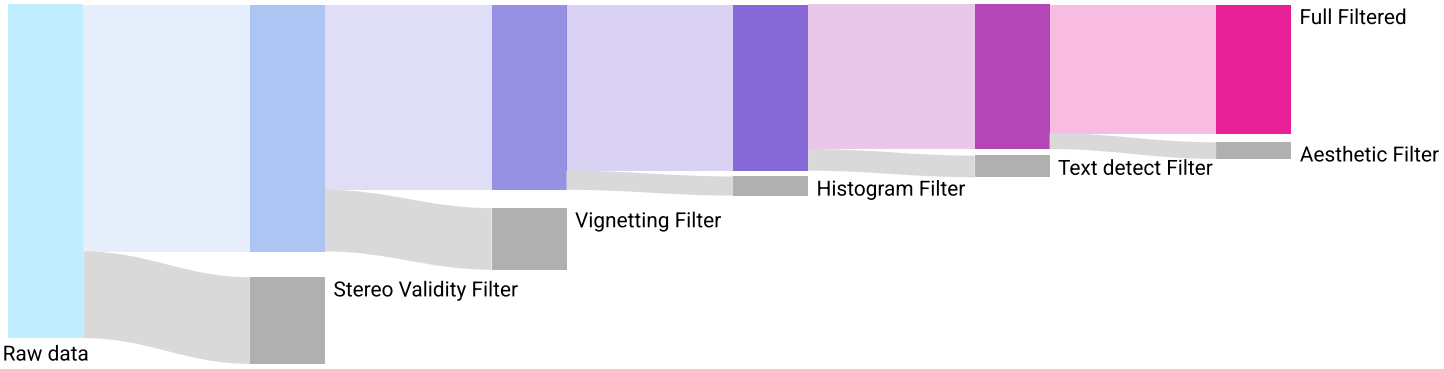


Figure 2: **Overall construction and curation pipeline of WebVR180.** VR180 video candidates collected from YouTube are first filtered using stereoscopic validity and vignetting criteria. The retained videos are then segmented into scene-level clips and further curated using histogram, text detection, and aesthetic quality filters. The remaining clips constitute the final WebVR180 dataset, while gray flows indicate videos or clips discarded at each filtering stage.

규격에 맞추어 초당 16프레임으로 변환한 뒤, 학습 길이 조건을 고려하여 32프레임 이상 161프레임 이하가 되도록 구성하였다.

이후 정제 과정은 평가 항목의 특성에 따라 원본 영상 또는 클립을 기준으로 수행하였다. 영상 전체에서 판단해야 하는 VR180 형식 유효성 검증과 주변부 비네팅 필터링은 원본 영상에 적용하고, 장면별로 달라질 수 있는 텍스트 포함 여부, 조도 특성, 미적 품질 필터링은 클립에 적용하였다. 구성된 클립은 이후 영상 생성 모델 학습을 위한 기본 데이터 단위로 사용하였다.

3.3 영상 클립-캡션 쌍 데이터 생성

영상 생성 모델의 학습 데이터는 각 영상 클립과 해당 내용을 설명하는 텍스트 캡션의 쌍으로 구성된다. 캡션 생성 과정에서는 양안 구성, 주변부 비네팅 및 기하 왜곡과 같은 입력 영상의 형식적 시각 특성이 캡션 내용에 미치는 영향을 줄이기 위해 단안 시점의 중앙 약 80% 영역을 입력으로 사용하였다. 이를 통해 장면 내용에 집중된 캡션을 생성할 수 있도록 하였다.

본 연구에서는 장면의 시각적 내용과 시간적 변화를 충분히 반영하는 고밀도 캡션 생성을 위해 Qwen2-VL-7B-Instruct를 활용하였다 [12]. 캡션 생성 시에는 영상의 스타일과 촬영 구도, 주요 객체와 동작, 배경 및 장면 내 움직임, 간판이나 자막 등 영상에 보이는 글자, 카메라 움직임을 포함하는 하나의 서술형 문단을 생성하도록 하였으며, 관찰되는 시각 정보에 기반한 객관적 설명을 유도하였다. 이를 통해 정제된 영상 클립에 대응하는 캡션을 확보하고, 영상 생성 모델 학습을 위한 영상 클립-캡션 쌍 데이터셋을 구성하였다.

4. VR180 영상 필터링 파이프라인

웹에서 수집한 VR180 영상에는 양안 구조가 불안정한 샘플, 유효 시야가 제한된 샘플, 시각 품질이 낮은 샘플이 포함될 수 있다. 본 절에서는 수집된 VR180 영상 후보군에서 이러한 샘플을 선별하기 위한 단계적 정제 파이프라인을 설명한다 (Fig. 2). 영상 후보군은 영상 단위에서 양안 구조 유효성과 주변부 비네팅을 평가하고, 클립 단위에서 밝기 분포, 텍스트 검출, 미적 품질을 기준으로 추가 정제하였다. 이 과정에서는 처리 효율성과 영상 전반의 품질 반영을 위해 각 영상 또는 클립에서 시작과 끝 구간을 제외한 나머지 구간 내에 9개 평가 시점을 균등하게 설정하고, 해당 프레임을 기준으로 정제 기준을 설정하였다. 각 필터링 지표의 임계값은 전체 후보군에서 산출된 측정값의 분포 경향을 분석하고 값이 극단적으로 치우치거나 샘플 품질 변화가 두드러지는 구간을 후보 임계 대역으로 설정한 뒤, 해당 구간의 샘플을 시각적으로 교차 검증하여 결정하였다. 단계별 정제 과정을 통해 최종적으로 396,597개의 고품질 영상 클립을 확보하였다.

4.1 양안 구조 유효성 필터링

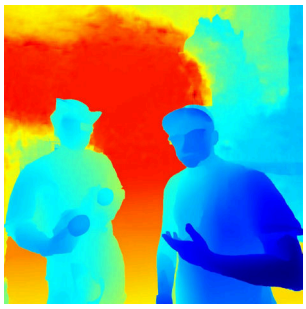
실제 수집 영상에는 시차 범위가 비정상적이거나 양안 정렬과 시각 품질이 일치하지 않는 사례가 포함될 수 있으며, 좌 · 우 시점이 뒤바뀐 경우도 관찰된다. 이러한 양안 품질 문제를 선별하기 위해 Lavrushkin et al.의 VR180 입체 품질 평가 항목을 참고하여 원본 영상의 양안 구조 유효성을 영상 단위로 검증하였다 [7].

4.1.1 신뢰도 가중 시차 분포 기반 깊이 범위 산정

양안 구조 유효성 검증에서는 먼저 좌 · 우 시점의 시차 분포를 이용해 각 영상이 표현하는 깊이 범위를 평가하였다. 등장방형도법으로 표현된 VR180 원본 영상은 주변부로 갈수록 기하 왜곡이 커지므로, 전체 영역을 그대로 사용하면 시차 분석 결과가 주변부 투영 왜곡의 영향을 크게 받아 양안 구조의 이상을 안정적으로 판단하기 어렵다. 이를 줄이기 위해 좌안과 우안을 분리한 뒤, 각 시점에서 전방 중심을 기준으로 가로 · 세로 시야각 90°의 원근 투영 영상을 추출하여 분석 입력으로 사용하였다. 전방 중심 영역은 주변부보다 기하 왜곡이 작고, 주요 장면의 양안 대응 관계를 상대적이고 안정적으로 관찰할 수 있어 깊이 범위 평가에 적합하다.

추출된 좌 · 우 영상 쌍에 대해서는 RAFT-Stereo [13]를 이용해

(a) Stereoscopic Video Data



(b) Disparity Map

(c) Confidence Map

Figure 3: **Disparity and confidence estimation.** (a) A stereoscopic VR180 video sample. (b) Disparity map estimated using RAFT-Stereo [13]. (c) Confidence map computed using the disparity plane sweep-based method of Lee et al. [14].

시차 지도를 추정하였다 (Fig. 3(b)). 이때 가림 경계, 반복 패턴, 질감이 부족한 영역에서는 좌·우 영상 간 대응점이 불안정하여 시차 추정값의 신뢰도가 낮아질 수 있다. 이러한 영역의 영향을 줄이기 위해 Lee et al.의 disparity plane sweep 기반 신뢰도 산정 방법을 적용하였다 [14]. 신뢰도 산정은 우안 영상을 일정한 간격으로 수평 이동했을 때, 신뢰할 수 있는 시차 값이라면 이동량 변화에 맞게 예측 결과도 일관되게 변해야 한다는 가정에 기반한다. 실제 계산에서는 우안 영상을 여러 수평 위치로 이동시켜 시차 지도를 반복 추정하고, 이동량 변화가 추정 시차에 얼마나 일관되게 반영되는지를 픽셀별 신뢰도로 사용하여 신뢰도 지도를 구성하였다 (Fig. 3(c)). 산출된 신뢰도 지도는 깊이 범위 산정과 색상 불일치 등 양안 품질 지표 계산에 반영하였다.

4.1.2 양안 구조 품질 평가 지표

양안 구조 유효성은 깊이 범위, 색상 불일치, 기하 정렬 오차, 선명도 불일치, 좌·우 시점 대응 이상의 5개 항목으로 평가하였다. 이 중 깊이 범위와 색상 불일치는 Lavrushkin et al.의 VR180 입체 품질 평가 방법에서 제시된 시차 및 색상 불일치 계산 절차를 기반으로 하였다 [7]. 기하 정렬 오차, 선명도 불일치, 좌·우 시점 대응 이상은 정제 목적에 맞게 산출 방식과 판정 기준을 정의하였다. 허용 기준을 벗어난 항목이 있는 영상은 안정적인 VR180 양안 구조를 갖추지 못한 것으로 판단하고, 다음 단계에서 제외하였다.

Parallax error. VR180 영상의 시차 분포는 장면이 표현하는 입체 깊이의 범위를 나타낸다. 깊이 범위가 지나치게 작으면 좌·우 시점 간 깊이 단서가 부족해 장면이 평면적으로 보일 수 있고, 과도한 양의 시차는 VR 환경에서 부자연스러운 입체감을 유발할 수 있다. 깊이 범위 산정에는 앞서 구성한 신뢰도 지도를 함께 사용하였다. 추정된 시차 지도에서 픽셀별 시차 값을 구간별로 누적하여 시차 분포를 구성하고, 각 픽셀의 신뢰도 값을 해당 시차 구간에 대한 기여도로 반영하였다. 이를 통해 신뢰도가 낮은 시차 값은 분포에서 차지하는 유효 비중이 작아지므로, 폐색 영역이나 저텍스처 영역 등에서 발생할 수 있는 불안정한 극단 시차 값이 최종 시차 경계 추정에 과도하게 반영되는 것을 완화하였다. 이후 신뢰도 가중 시차 분포의 누적 히스토그램에서 하위 1%와 상위 1%에 해당하는 양 끝단을 제외하고, 남은 유효 구간의 하한과 상한을 각각 음의 시차 경계와 양의 시차 경계로 추정하였다. 이를 기준으로 깊이 범위가 지나치게 작거나 과도한 양의 시차가 나타나는 영상을 제외 대상으로 분류하였다.

Color mismatch. 좌·우 영상 사이의 색상 차이가 크면 동일한 장면이 양안에서 서로 다르게 표현되어 안정적인 양안 구조 학습을 방해할 수 있다. 이때 같은 물체나 영역의 위치가 시차에 의해 수평 방향으로 달라지므로, 색상을 비교하기 전에 추정된 시차 지도를 이용해 우안 영상을 좌안 기준으로 정렬하였다. 색상 차이는 정렬된 대응 위치의 픽셀별 색상 차이로 계산하고, 앞서 산출한 시차 신뢰도를 가중치로 적용하여 영상 단위의 평균 색상 차이를 산출하였다. 이를 통해 대응 관계가 불확실한 영역의 색상 차이가 최종 지표에 과도하게 반영되지 않도록 하였으며, 평균 색상 차이가 허용 범위를 초과한 영상은 제외 대상으로 분류하였다.

Geometric mismatch. 좌·우 영상에서는 깊이에 따른 수평 시차가 자연스럽게 나타나지만, 수직 방향의 위치 차이나 회전, 크기 차이는 정상적인 양안 정렬 관계를 벗어난 기하 정렬 오차에 해당한다. 이를 평가하기 위해 각 시점에서 SIFT 특징점을 추출하고, 대응 특징점을 기반으로 RANSAC을 적용하여 아핀 변환 행렬을 추정하였다 [15, 16]. 추정된 아핀 변환에서 수평 이동 성분을 제외한 뒤, 수직 방향 병진 성분을 수직 위치 오차로 사용하고 선형부에서 좌표축의 방향 변화와 길이 변화를 계산하여 각각 회전 및 크기 차이 지표로 사용하였다. 이 중 하나라도 허용 범위를 벗어나는 영상은 양안 정렬이 유효하지 않은 샘플로 판단하여 제외 대상으로 분류하였다.

Sharpness mismatch. 좌·우 영상 사이의 선명도 차이가 크면 동일한 장면의 경계와 질감이 양안에서 서로 다르게 인식되어, 시청 시 입체감이 불안정해지거나 눈의 피로감이 증가할 수 있다. 선명도 불일치는 각 시점의 라플라시안 분산을 계산한 뒤, 두 값의 비율을 로그 변환한 값으로 평가하였다. 산출된 선명도 차이가 허용 범위를 벗어나는 영상은 한쪽 시점의 화질 저하가 두드러지는 샘플로 판단하여 제외 대상으로 분류하였다.

Channel mismatch. 좌·우 채널이 서로 뒤바뀐 경우 또는 동일 영상이나 서로 대응하지 않는 시점 영상으로 구성된 경우에는 양안 시차가 장면의 깊이 순서와 일관되지 않게 나타날 수 있다. 이를 평가하기 위해 MiDaS [17] 기반 단안 깊이 추정 결과와 양안 시차 지도로부터 각각 상대적 깊이 순서를 산출하고, 두 순서가

Figure 4: **Vignetting filtering.** Samples with excessive dark boundaries are identified and removed to exclude severe vignetting and invalid projection patterns from the VR180 dataset.

Figure 5: **Text detect filtering.** Samples containing prominent embedded text or graphic overlays are detected and removed to improve the visual quality of the VR180 training data.

얼마나 일치하는지 비교하였다. 두 깊이 순서의 일치 정도는 스피어만 순위 상관계수로 측정하였으며, 상관관계가 허용 기준을 벗어나는 영상은 채널 뒤바뀜 또는 비정상적인 시차 구조가 존재하는 샘플로 판단하여 제외 대상으로 분류하였다.

이상의 5개 양안 구조 품질 평가 지표를 적용하여, 깊이 표현이나 좌·우 시점 간 대응 관계가 불안정한 영상을 제외하였다. 전체 원본 영상의 73.37%가 양안 구조 유효성 기준을 통과하여 후속 시각 품질 평가 대상으로 유지되었다.

4.2 주변부 비네팅 필터링

웹에서 수집한 VR180 후보 영상 중에는 등장방형도법으로 변환되지 않은 원형 어안 렌즈 화면이 그대로 남아 있거나, 주변부가 검정 픽셀로 넓게 채워진 과도한 비네팅 사례가 포함될 수 있다. 이러한 영상은 등장방형도법 기반 VR180의 공간 표현과 일치하지 않거나, 학습에 활용 가능한 유효 시야가 제한되므로 영상 단위에서 제외하였다. 이를 위해 각 평가 프레임의 상·하·좌·우 가장자리에서 내부 방향으로 연속된 검정 픽셀 영역의 길이를 측정하고, 이를 해당 방향 길이에 대한 비율로 정규화하였다 (Fig. 4). 프레임별 측정값은 방향별로 모은 뒤 극단값을 제외하고 영상 단위로 평균 집계하여, 방향별 비네팅 지표로 사용하였다. 해당 정제 단계를 적용한 결과, 직전 단계에서 선별된 영상의 74.99%를 후속 정제 대상으로 유지하였다.

4.3 밝기 분포 기반 필터링

장면 단위로 분할된 클립 중에는 장면 전체가 지나치게 어둡거나 명암 대비가 낮아 주요 객체와 배경 구조, 장면 내 움직임이 충분히 드러나지 않는 경우가 있다 [18]. 이러한 클립을 선별하기 위해 각 클립의 평가 프레임을 회색조로 변환하고, 픽셀 밝기 값의 평균과 표준편차를 계산하였다.

프레임별 밝기 평균과 표준편차는 클립 단위로 평균 집계하여 밝기 분포 지표로 사용하였다. 평균 밝기가 낮은 클립은 시각 구조가 드러나지 않는 극저조도 샘플로, 밝기 표준편차가 낮은 클립은 경계와 질감 구분을 위한 명암 단서가 부족한 샘플로 판단하였다. 해당 정제 단계를 적용한 결과, 직전 단계에서 구성된 클립의 93.41%를 후속 정제 대상으로 유지하였다.

Figure 6: **Word cloud of frequent concepts.** Concepts extracted from dense captions illustrate the diverse scenes and visual contexts covered by the dataset.

4.4 텍스트 검출 기반 필터링

자막, 워터마크, 안내 문구 등 화면에 포함된 텍스트는 객체와 배경의 시각 정보를 가릴 뿐 아니라, 모델이 장면 내용과 무관한 문자 패턴을 학습하게 할 수 있다. 따라서 화면 내 텍스트 비중이 높은 클립을 제외하기 위해 텍스트 검출 기반 정제 단계를 수행하였다 [19].

구체적으로 각 클립의 평가 프레임에 CRAFT [20] 텍스트 검출기를 적용하여 텍스트 영역을 검출하고, 검출 영역의 면적이 전체 프레임에서 차지하는 비율을 텍스트 면적 비율로 산출하였다 (Fig. 5). 프레임별 텍스트 면적 비율은 클립 단위로 평균 집계하여 대표값으로 사용하였으며, 대표값이 높은 클립은 후속 정제 대상에서 제외하였다. 해당 정제 단계를 적용한 결과, 직전 단계에서 구성된 클립의 87.86%를 후속 정제 대상으로 유지하였다.

4.5 미적 품질 기반 필터링

영상 생성 모델의 학습 데이터는 장면 내용뿐만 아니라 전체적인 구도와 시각적 완성도 측면에서도 일정 수준 이상의 품질을 갖추어야 한다. 이를 위해 LAION-Aesthetics [21] 기반의 미적 품질 평가를 적용하여 학습에 부적합한 저품질 클립을 추가로 선별하였다. 각 클립의 평가 프레임에서 산출된 미적 품질 점수는 클립 단위로 평균 집계하여 대표 점수로 사용하였다. 평가 결과 시각적 품질이 낮은 클립을 제외하였으며, 직전 단계에서 선별된 입력 클립의 89.15%를 최종 데이터셋 구성 대상으로 유지하였다.

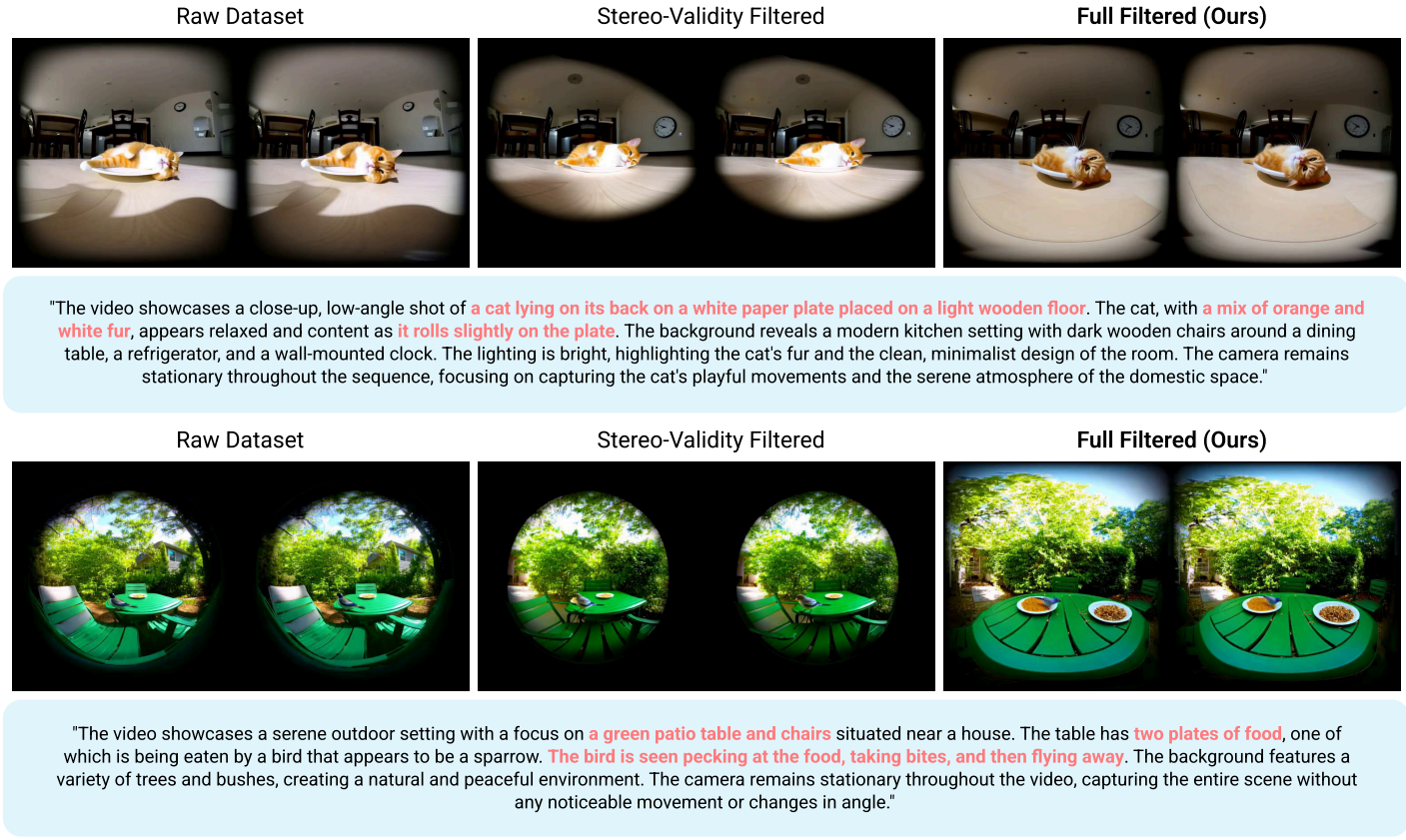


Figure 7: **Qualitative comparison across curation stages.** Models fine-tuned on the Raw Dataset exhibit unstable stereoscopic structure or peripheral appearance, whereas Stereo-Validity Filtered improves left–right consistency. Full Filtered (Ours) produces more natural VR180 videos with stable stereoscopic structure and improved peripheral visual quality.

Table 2: **Stereoscopic mismatch comparison.** Full Filtered (Ours) minimizes errors across all metrics, ensuring stereoscopic stability.

Metric	Raw Dataset	Stereo-Validity	Full Filtered (Ours)
Color	0.0317	0.0277	**0.0257**
Vertical	0.7505	0.6541	**0.6442**
Rotation	0.4589	0.4634	**0.4332**
Scale	1.1239	1.0331	**0.9577**
Sharpness	0.0559	0.0542	**0.0525**
Channel	0.2959	0.2870	**0.2784**

Table 3: **VBench evaluation across dataset configurations.** Full Filtered (Ours) maintains comparable visual quality while achieving the highest motion consistency score.

Metric	Raw Dataset	Stereo-Validity	Full Filtered (Ours)
Subject	**0.9857**	0.9803	0.9830
Background	**0.9842**	0.9802	0.9820
Motion	0.9961	0.9961	**0.9964**
Dynamic	0.040	**0.072**	0.060
Aesthetic	**0.5464**	0.5328	0.5370
Imaging	**62.12**	60.35	61.01

5. 데이터셋 분석 및 유효성 검증

5.1 캡션 기반 데이터셋 의미 다양성 분석

구축된 데이터셋의 의미적 다양성과 주제 분포를 확인하기 위해, 약 396K개의 생성 캡션을 기반으로 군집 분석을 수행하고 단어 빈도 분포를 워드 클라우드로 시각화하였다 (Fig. 6). 분석 결과, 제안 데이터셋은 특정 소수 주제에 편중되지 않았으며, 가장 큰 메타 범주도 전체의 약 6.5% 수준에 머물러 특정 장면 유형이 데이터셋 전체를 지배하지 않음을 확인하였다. 주요 메타 범주에는 인물 동작, 공연 · 예술, 도로 및 이동 환경, 실내 공간, 자연 배경 등이 포함되었으며, 워드 클라우드에서도 dancers, players 등 인물 활동과 room, beach 등 공간 배경, car, aircraft 등 이동 객체 관련 단어가 함께 나타났다. 이는 제안 데이터셋이 특정 시각 패턴이나 제한된 문맥에 치우치지 않고, 다양한 환경과 객체 맥락을 포함하고 있음을 보여준다.

5.2 데이터 정제 단계에 따른 파인튜닝 성능 평가

데이터 정제 단계의 효과를 평가하기 위해 필터링을 적용하지 않은 Raw Dataset, 양안 구조 유효성 필터링만 적용한 Stereo-Validity Filtered, 전체 정제 과정을 적용한 Full Filtered (Ours)를 비교하였다. 실험에서는 세 데이터셋 각각에서 무작위로

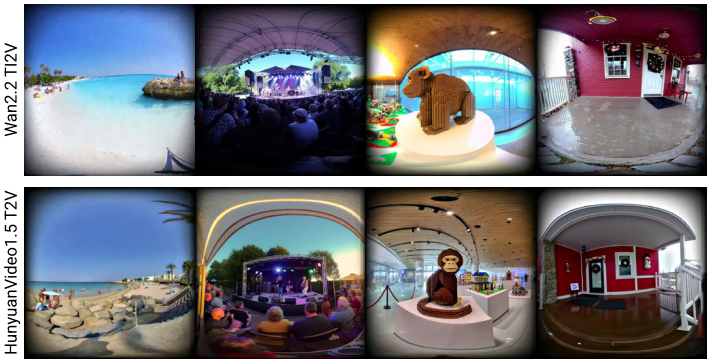


Figure 8: **Applicability across video generation models.** Qualitative VR180 results from Wan2.2-TI2V and HunyuanVideo1.5-T2V fine-tuned on the proposed dataset, demonstrating its applicability to different video generation models.

Table 4: **VBench evaluation across projection formats.** Models fine-tuned on the proposed dataset are evaluated in the original ERP 180° format and the projected 90° perspective view.

	ERP 180°		Persp. 90°	
Metric	**Hunyuan**	**Wan**	**Hunyuan**	**Wan**
Subject	0.9467	**0.9809**	0.9316	**0.9768**
Background	0.9537	**0.9804**	0.9504	**0.9742**
Motion	0.9844	**0.9932**	0.9785	**0.9918**
Dynamic	**0.3480**	0.0710	**0.5274**	0.1594
Aesthetic	0.5170	**0.5390**	0.5247	**0.5406**
Imaging	0.6782	**0.6783**	0.5682	**0.5895**

약 150K개의 샘플을 추출하여 Wan2.1-T2V-1.3B 모델에 LoRA를 적용하고, 동일한 조건에서 파인튜닝하였다 [22]. 학습은 NVIDIA B200 GPU 8장 환경에서 약 20시간 동안 수행하였으며, 평가용 텍스트 프롬프트 1,000개로부터 생성 샘플을 산출하였다. 생성 결과는 정성적 비교, 양안 불일치 지표, VBench 기반 영상 생성 품질의 세 관점에서 평가하였다.

Qualitative comparison. Fig. 7은 각 데이터 구성으로 학습한 모델의 생성 결과 예시를 비교한다. Raw Dataset 결과에서는 일부 예시에서 좌·우 영상이 동일한 장면을 안정적으로 공유하지 못하고, 주요 객체의 위치나 형태가 서로 다르게 생성되는 문제가 관찰되었다. 또한 원형 어안 렌즈 형식의 검정 주변부가 크게 남아 있어 VR180 영상으로 활용 가능한 유효 시야가 제한된다. Stereo-Validity Filtered에서는 좌·우 시점 간 구조적 대응은 상대적으로 개선되지만, 원형 시야 경계와 주변부 비네팅이 여전히 두드러진다. 반면 Full Filtered (Ours)는 양안 구조와 주변부 표현을 보다 안정적으로 유지하며, 전반적으로 더 자연스러운 VR180 생성 결과를 보였다.

Stereoscopic consistency evaluation. Table 2는 생성 결과에서 측정한 양안 불일치 지표를 비교한다. 각 지표는 값이 낮을수록 두 시점 간 불일치가 작음을 의미하며, Full Filtered (Ours)는 Color, Vertical, Rotation, Scale, Sharpness, Channel mismatch의 모든 항목에서 가장 낮은 값을 기록하였다. 이는 양안 구조 유효성 필터링과 후속 시각 품질 필터링을 함께 적용한 정제 과정이 VR180 생성 결과의 양안 구조를 더 안정적으로 만드는 데 기여함을 보여준다.

General visual quality evaluation. Table 3은 데이터 정제 단계에 따른 VBench 평가 결과를 나타낸다 [23]. 세 데이터 구성의 VBench 점수는 전반적으로 유사한 수준을 보였다. 이는 초기 수집 데이터가 VBench가 측정하는 영상 생성 품질 측면에서 이미 일정 수준을 확보하고 있었기 때문으로 해석된다. 따라서 이 결과는 제안한 정제 과정이 VR180 학습 데이터로서의 적합성을 높이는 과정에서도, 기본적인 영상 생성 품질을 크게 저하시키지 않았음을 시사한다.

5.3 다양한 영상 생성 모델에 대한 적용성 평가

제안 데이터셋이 특정 모델에 한정되지 않고 활용될 수 있는지 확인하기 위해, Wan2.2-5B-TI2V와 HunyuanVideo1.5를 대상으로 추가 실험을 수행하였다. 두 모델은 VR180 영상의 단안 시점으로 구성한 180° 등장방형도법 데이터를 이용해 LoRA를 적용하여 파인튜닝하였다 [22]. 학습은 NVIDIA H100 GPU 8장 환경에서 각각 약 300시간 동안 수행하였으며, 평가용 텍스트 프롬프트 1,000개로부터 생성한 샘플을 정성적 비교와 VBench 기반 영상 생성 품질 평가에 활용하였다.

Qualitative comparison. Fig. 8은 두 모델의 생성 결과 예시를 보여준다. Wan2.2-5B-TI2V와 HunyuanVideo1.5 모두 180° 등장방형도법 형식에 대응하는 초광각 장면 구성과 주변부 표현을 생성 결과에 반영하였다. 이는 제안 데이터셋을 이용한 파인튜닝이 다양한 영상 생성 모델에서도 VR180의 단안 등장방형도법 표현을 학습하는 데 활용될 수 있음을 시사한다.

General visual quality evaluation. Table 4는 180° 등장방형도법 생성 결과와 이를 시야각 90°의 원근 시점으로 변환한 결과의 VBench 평가를 제시한다 [23]. 제안 데이터셋으로 파인튜닝된 두 모델은 180° 등장방형도법 형식에서 전반적으로 양호한 VBench 점수를 보였으며, 원근 시점으로 변환한 뒤에도 장면 일관성과 시각 품질을 유지하였다. 이는 제안 데이터셋이 다양한 영상 생성 모델의 VR180 단안 등장방형도법 학습에 활용될 수 있고, 생성 결과 또한 실제 시청에 가까운 원근 시점으로 변환해 활용할 수 있음을 보여준다.

5.4 원근 시점 변환 기반 생성 결과 분석

VR180 영상은 전방 180° 반구 영역을 등장방형도법 영상으로 저장하지만, 실제 시청 과정에서는 HMD의 머리 회전이나 데스크톱·모바일 환경의 드래그 조작에 따라 선택된 일부 영역이 일정 시야각의 원근 화면으로 표시된다. 따라서 생성 결과가 실제 시청 화면에서도 자연스럽게 유지되는지 확인하기 위해, 생성된 양안 VR180 영상을 정면, 좌측, 우측 방향에서 각각 시야각 90°의 원근 시점으로 변환하였다. Fig. 9는 변환 결과의 예시를 보여준다. 각 원근 시점에서 주요 객체의 형태와 장면 구성이 일관되게 유지

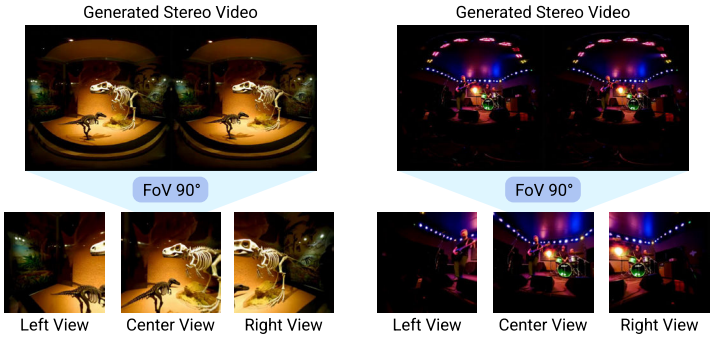


Figure 9: **Perspective-views of generated VR180 videos.** Generated stereoscopic results are projected into left, center, and right views with a 90° field of view, illustrating their appearance under practical VR180 viewing perspectives.

되었으며, 등장방형도법 영상의 주변부에 해당하는 좌측 및 우측 방향에서도 두드러진 구조 왜곡이나 객체 형태의 붕괴가 관찰되지 않았다. 이는 생성 결과의 품질이 정면 영역에만 집중되지 않고, 사용자의 시선 이동에 따라 관찰되는 주변 방향에서도 장면 구조와 객체 형태를 유지함을 시사한다.

6. 한계점 및 향후 연구

본 데이터셋은 YouTube에 공개된 VR180 영상을 기반으로 하므로, 촬영 장비, 렌즈 특성, 실제 유효 화각, 해상도, 후처리 방식이 영상마다 일정하지 않다. 이러한 차이는 다양한 환경에서 제작된 VR180 콘텐츠를 포함한다는 장점이 있지만, 특정 카메라 또는 광학 조건을 엄밀히 통제해야 하는 분석, 평가, 생성 모델 학습에서는 결과 해석을 어렵게 할 수 있다. 향후에는 수집 영상에서 추정 가능한 촬영 및 투영 특성을 기준으로 데이터를 세분화하고, 연구 목적에 따라 선택적으로 활용할 수 있도록 데이터셋을 보완할 필요가 있다.

또한 VR180 도메인에서는 영상 생성 모델 학습용 데이터셋의 품질과 구성 차이가 모델 성능에 미치는 영향을 일관된 기준으로 비교하기 위한 평가 체계가 아직 충분히 마련되어 있지 않다. 따라서 본 연구의 정제 결과 및 생성 모델 기반 평가는 구축 데이터셋의 특성을 분석하는 데에는 유효하지만, 다른 데이터셋 대비 성능 차이를 직접 판단하는 기준으로 보기는 어렵다. 이를 위해서는 공통 벤치마크 데이터셋과 평가 프로토콜을 마련하여, VR180 데이터셋이 생성 결과의 양안 구조 안정성과 생성 품질에 미치는 영향을 보다 객관적으로 비교할 수 있는 평가 기반을 갖추어야 한다.

7. 결론

본 연구는 VR180 영상 생성 연구를 위한 대규모 영상 클립-캡션 데이터셋 WebVR180을 구축하고, 양안 구조와 시각 품질을 고려한 정제 및 검증 절차를 제안하였다. WebVR180은 YouTube에서 수집한 31,625개의 VR180 영상 후보군을 장면 단위 클립으로 구성하고, 각 클립에 장면의 시각적 내용과 시간적 변화를 반영한 고밀도 캡션을 생성하여 구축하였다. 특히 웹에서 수집한 VR180 영상에 포함될 수 있는 양안 구조 불안정성, 제한된 유효 시야, 낮은 시각 품질 문제를 극복하고자, 원본 영상 단위와 클립 단위에서 도메인 맞춤형 정제 파이프라인을 도입하였다. 그 결과, 최종적으로 396,597개의 정제된 영상 클립-캡션 쌍을 확보하였다.

데이터 특성 분석에서는 WebVR180이 다양한 장면과 객체 맥락을 포함함을 확인하였으며, 정제 단계별 파인튜닝 실험에서는 전체 정제 데이터를 사용한 모델이 더 안정적인 양안 구조와 주변부 표현을 보였다. 또한 Wan2.2-5B-TI2V와 HunyuanVideo1.5-T2V를 이용한 실험 및 원근 시점 변환 분석을 통해, WebVR180이 다양한 영상 생성 모델의 VR180 표현 학습과 실제 시청 화면에 가까운 생성 결과 분석에 활용될 수 있음을 확인하였다. 나아가 WebVR180은 전방 180°의 광시야각과 양안 영상 구조를 함께 포함하므로, VR180 영상 생성뿐 아니라 입체 영상 이해, 광시야 장면 분석, 몰입형 콘텐츠 품질 평가와 같은 후속 연구에도 활용될 수 있다. 이러한 결과는 본 연구가 대규모 학습 데이터가 부족한 VR180 영상 생성 분야에서, 양안 구조와 시각 품질을 함께 고려한 데이터셋 구축 및 검증의 필요성과 유효성을 보였다는 점에서 의의가 있다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2026년도 문화체육관광 연구개발사업(연구개발과제명 : 공연 콘텐츠의 고해상도(8K/16K) 서비스를 위한 AI 기반 영상확장 및 서비스 기술 개발, 연구개발과제번호 : RS-2024-00395886; 연구개발과제명 : 한국형 돔 미디어 구현을 위한 AI 영상 및 음향 통합 오케스트레이션 기술 개발, 연구개발과제번호 : RS-2026-25525791) 과 정부(과학기술정보통신부)의 재원으로 “첨단 GPU 활용 지원 사업”의 지원을 받아 수행된 연구임.

References

[1] C. Wang, S. Lucey, F. Perazzi, and O. Wang, “Web stereo video supervision for depth prediction from dynamic scenes,” in *Proceedings of the International Conference on 3D Vision (3DV)*, 2019, pp. 348–357.

[2] M. Izadimehr, M. Ghanbari, G. Chen, W. Zhou, X. Hao, M. Dasari, C. Timmerer, and H. Amirpour, “Svd: Spatial video dataset,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 12 988–12 994.

[3] G. Shen, Y. Du, W. Ge, J. He, C. Chang, D. Zhou, Z. Yang, L. Wang, X. Tao, and Y.-C. Chen, “Stereopilot: Learning uni-

fied and efficient stereo conversion via generative priors," *arXiv preprint arXiv:2512.16915*, 2025.

[4] Q. Wang, W. Li, C. Mou, X. Cheng, and J. Zhang, "360dvd: Controllable panorama video generation with 360-degree video diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6913–6923.

[5] M. Zayene, J. Endres, A. Havolli, C. Corbière, S. Cherkaoui, A. Kontouli, and A. Alahi, "Helvipad: A real-world dataset for omnidirectional stereo depth estimation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26 975–26 984.

[6] L. Jin, R. Tucker, Z. Li, D. Fouhey, N. Snavely, and A. Holynski, "Stereo4d: Learning how things move in 3d from internet stereo videos," *arXiv preprint arXiv:2412.09621*, 2024.

[7] S. Lavrushkin, I. Molodetskikh, K. Kozhemyakov, and D. Vatolin, "Stereoscopic quality assessment of 1,000 vr180 videos using 8 metrics," *Electronic Imaging*, vol. 2021, no. 2, pp. 350–1, 2021.

[8] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[9] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[10] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[11] B. Castellano, "PySceneDetect." [Online]. Available: https://github.com/Breakthrough/PySceneDetect

[12] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[13] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International conference on 3D vision (3DV)*. IEEE, 2021, pp. 218–227.

[14] J. Y. Lee, W. Ka, J. Choi, and J. Kim, "Modeling stereo-confidence out of the end-to-end stereo-matching network via disparity plane sweep," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 2901–2910.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[17] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[18] X. Ju, Y. Gao, Z. Zhang, Z. Yuan, X. Wang, A. Zeng, Y. Xiong, Q. Xu, and Y. Shan, "Miradata: A large-scale video dataset with long durations and structured captions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 48 955–48 970, 2024.

[19] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[20] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9365–9374.

[21] LAION-AI, "aesthetic-predictor," https://github.com/LAION-AI/aesthetic-predictor, 2022.

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." *Iclr*, vol. 1, no. 2, p. 3, 2022.

[23] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 807–21 818.

〈저자소개〉

최창수
- 2026년 숭실대학교 글로벌미디어학부 학사
- 2026년 ~ 현재 : 숭실대학교 메타버스 · 문화콘텐츠학과 석사과정
- 관심분야: Generative AI, Computer Graphics
- https://orcid.org/0009-0003-0862-087X

정석현
- 2024년 숭실대학교 글로벌미디어학부 학사
- 2024년 ~ 현재 : 숭실대학교 메타버스 · 문화콘텐츠학과 석박통합과정
- 관심분야: Generative AI, Machine Learning, Computer Graphics
- https://orcid.org/0009-0001-4729-0805

구한슬
- 2022년 ~ 현재 : 숭실대학교 AI소프트웨어학부 학사
- 관심분야: Generative AI, Computer Graphics
- https://orcid.org/0009-0001-7551-3017

이정진
- 2010년 숭실대학교 미디어학부 학사
- 2012년 KAIST 문화기술대학원 석사
- 2017년 KAIST 문화기술대학원 박사
- 2016년~2020년 (주)카이 연구이사
- 2020년~2025년 (주)라이브커넥트 CTO 사외이사
- 2020년~현재 숭실대학교 글로벌미디어학부 조교수
- 관심분야: 컴퓨터 그래픽스, VR/AR, 몰입형 시각 미디어, 이미지/비디오 응용, HCI
- https://orcid.org/0000-0003-3471-4848