

버티컬 AI는 데이터에서 시작한다

: MBC 방송영상 AI 학습데이터 구축 성과 및 사업과 전략

유영진 _ MBC AI전략팀 부장

[목차]

1. 들어가며: 방송사는 정말 ‘데이터 부자’인가?
2. 버티컬 AI의 전제: 고품질 학습데이터와 명확한 태스크 정의
3. MBC 방송영상 AI 학습데이터 구축 사례
4. 구축한 학습데이터는 무엇을 가능하게 했나
5. 방송영상 학습데이터의 가치와 사업화 전략
6. 마치며: 아카이브는 저절로 자산이 되지 않는다



MBC

요약문

인공지능(이하 AI) 시대에 방송사가 보유한 아카이브는 그 자체만으로 자산이 되지 않는다. 아카이브는 원석에 가깝고, 방송사가 인공지능 모델의 태스크(Task)를 정의하고 그에 맞는 학습데이터로 가공할 때 비로소 자산이 된다. MBC는 2025년 <방송영상 AI 학습용 데이터 구축 사업>에서 5,372시간의 학습데이터를 구축해, K-콘텐츠를 이해하는 모델과 배경 이미지 생성 모델의 성능 향상을 확인했다. 방송영상에는 언어·행동·감정·물리 환경과 한국의 공간·문화가 수십 년 치 기록돼 있어, 한국의 실제 세계를 이해하는 차세대 AI의 학습데이터로도 가치가 높다. 본 글은 MBC 사례를 바탕으로 아카이브가 자산이 되기 위한 과정을 정리한다. 나아가 월드 모델 등의 등장으로 영상 기반 학습데이터 수요가 커지는 시장 환경에 맞춰, MBC가 방송사의 새로운 수익원으로 검토 중인 학습데이터 거래 플랫폼을 제안한다.

❶ 들어가며: 방송사는 정말 '데이터 부자'인가?

2025년 K-콘텐츠 수출액은 149억 달러로 역대 최대를 경신했지만, 정작 국내 미디어·콘텐츠 기업은 광고 감소와 유통 정체로 자생력을 잃어가고 있다. 방송사는 회당 평균 30억 원 안팎으로 치솟은 드라마 제작비를 감당할 수 없어, 글로벌 OTT에 유통 권리를 넘겨 제작비를 보전받는 선택을 거듭한다. 수익 기반은 무너지고 비용 부담만 커지는 구조가 고착된 것이다.

이 구조를 바꿀 돌파구는 방송 제작 현장의 AI 전환(AI Transformation, 이하 AX)이다. 2026년 지상파 4사 신년사는 하나같이 AX를 최우선 과제로 꼽았고, 국내 콘텐츠 제작사 세 곳 중 한 곳은 이미 생성형 AI를 활용하고 있다. AX의 필요성은 더 이상 논쟁거리가 아니다. 남은 과



출처: 이유나 (2025. 9. 4), 어도비시 왜 이러나...“한복 그려 달랬더니 기모노 나와”, <YTN>

제는 개별 공정 효율화를 넘어 콘텐츠 제작 전 과정을 다시 설계하는 일이다.

다만, 현재 AX는 대부분 범용 AI 모델을 도입해 솔루션을 개발하는 방식으로 진행된다. 영상 컷 분할, 블러 작업, 자막 생성 같은 단순 반복 작업에는 충분하지만, 방송 제작의 문법과 한국의 문화·역사 맥락을 이해해야 하는 작업에서는 한계를 드러낸다. 2025년 9월 어도비(Adobe) 포토샵의 생성형 AI에 ‘한복’을 프롬프트로 입력하자 일본식 의상 이미지가 다수 출력돼 논란이 됐다. 한복도 제대로 이해하지 못하는 모델에 사극의 배경영상 생성을 맡길 수는 없다. 연출 의도에 맞는 한국 장면을 만들려면 방송 제작 과정과 우리 문화를 학습한 미디어 특화 버티컬 AI(Vertical AI) 모델이 필요하고, 그 모델 개발은 결국 방송영상 기반의 학습데이터에서 출발한다.

그렇다면 방대한 아카이브를 보유한 방송사는 정말 ‘데이터 부자’일까? 그렇지 않다. 아카이브를 보관하는 일과 이를 학습데이터로 전환하는 일은 다르다. 수십 년간 쌓아 온 영상은 권리 검토, 비식별 처리, 장면 단위 분할, 설명문·메타데이터 작성을 거쳐야 비로소 학습데이터가 된다. 방송사가 가진 것은 원석이지, 버티컬 AI 모델 학습에 곧바로 쓸 수 있는 학습데이터가 아니다.

MBC는 2025년 방송미디어통신위원회의 <방송영상 AI 학습용 데이터 구축 사업>에 선정돼, 60년 넘게 쌓아 온 아카이브에서 K-콘텐츠 특화 영상 이해 모델과 배경 이미지 생성 모델 개발을 위한 학습데이터를 구축했다. 본 글은 방송 아카이브가 학습데이터 자산이 되기까지의 과정을 살펴보고, 버티컬 AI 모델 개발을 통한 제작 과정의 혁신, 그리고 방송영상 학습데이터의 가치와 이를 활용한 신규 비즈니스 전략을 짚어본다.

② 버티컬 AI의 전제: 고품질 학습데이터와 명확한 태스크 정의

2-1. 버티컬 AI의 성능은 학습데이터가 좌우한다

범용 모델은 방송 도메인을 모른다. 연출자의 의도나 서사의 전개와 같은 영상 문법을 이해하지 못한다. 한국의 장소·사물·행동을 ‘아시아 어딘가의 것’ 정도로 뭉뚱그리기도 한다. 이런 모델을 현장에 투입하면 일부 제작 공정의 시간은 줄어도 결과물을 수정하는 데 더 많은 시간이 들 수 있다.

방송의 맥락을 이해하는 모델을 개발하려면, 이를 배울 수 있는 학습데이터가 필요하다. 웹에서 수집한 영상은 출처와 품질이 검증되지 않았을 뿐 아니라 대부분 완성된 결과물이다. 완성된 영상에서는 촬영과 편집의 결과만 보일 뿐, 그 장면을 왜 그렇게 찍었고 어디서 잘랐는지는 드러나지 않는다. 그 답은 방송사 아카이브에 있다. 원본·방송본·대본·자막이 함께 남아 있어 영상 문법과 편집 판단이 이뤄진 과정도 학습할 수 있다.

모델 성능은 학습데이터의 양과 질에 달려 있다. 특히 버티컬 AI에서는 잘 정제된 소량의 고품질 데이터가 검증되지 않은 대량의 데이터보다 더 큰 학습 효과를 발휘한다. 이것이 방송사의 AX가 아카이브를 활용한 고품질 특화 학습데이터 구축에서 시작해야 하는 이유다.

2-2. 무엇을 학습시킬 것인가: 태스크 정의는 방송사의 몫이다

학습데이터를 구축할 때 데이터의 양과 라벨링 품질만큼 중요한 요소가 있다. 버티컬 AI 모델이 담당할 태스크가 무엇인지 정하는 일이다. 태스크를 어떻게 정의하느냐에 따라 필요한 데이터의 형태도 달라진다. 장면 검색에는 장면 단위 클립과 설명문이, 하이라이트 추출에는 촬영 원본 가운데 방송본에 쓰인 구간을 표시한 라벨이, 사극 배경 생성에는 시대·장소·소품이 검증된

이미지와 설명문이 각각 필요하다. 실제로 미디어 특화 모델 미디어GPT(MediaGPT)를 개발한 연구진은 이 모델이 범용 모델을 앞선 두 요인으로 도메인 데이터와 ‘도메인 전문가가 직접 정의한 과제’를 꼽았다.

태스크를 정의하는 일은 곧 AI로 해결할 문제를 정하는 일이다. 이 작업은 제작 현장의 요구를 가장 잘 아는 방송사의 몫이다. 성공적인 AX를 이루기 위해서는 어느 공정에서 사람의 공수가 가장 많이 들어가는지, 어떤 작업이 반복되는지를 파악해 이를 해결하기 위한 태스크를 정의하는 작업이 선행돼야 한다. 버티컬 AI 모델 개발을 외부에 맡기더라도, 방송사가 직접 AX 수요를 파악해 해결하고자 하는 문제를 정의하고, 그에 맞는 학습데이터를 구축해야 한다.

③ MBC 방송영상 AI 학습데이터 구축 사례

3-1. 왜 시작했나: K-콘텐츠 특화 영상 생성 모델을 통한 제작 과정 혁신

MBC는 방송 제작 전 과정에 AI를 도입해 제작 방식을 혁신하는 일이 지상파 방송사가 처한 위기를 돌파할 생존 전략이라고 판단했다. 개별 공정만 자동화해서는 수익은 줄고 비용은 치솟는 구조를 바꿀 수 없기 때문이다. 이를 위해 한국 고유의 문화와 시대 배경을 반영한 ‘K-콘텐츠 특화 영상 생성형 AI 모델 및 솔루션’(이하 ‘K-콘텐츠 특화 모델’)을 개발해 제작 비용과 시간을 크게 줄이고, K-콘텐츠의 수익성을 개선한다는 중장기 목표를 세웠다.

K-콘텐츠 특화 모델의 첫 적용 대상으로는 사극·시대극을 택했다. 사극·시대극은 세트·미술비가 전체 제작비의 약 35%를 차지한다. 세트 대신 AI가 생성한 배경영상을 활용한 버추얼 스튜디오 제작 방식을 도입하면 이 비용 구조를 바꿀 수 있다. MBC가 K-콘텐츠 특화 모델을 개발하기 위한 중장기 계획은 그림 2와 같다.

최종 목표인 K-콘텐츠 특화 모델을 개발하려면 한국의 언어·문화·영상 문법이 반영된 대규모 고품질 멀티모달 학습데이터가 필요하다. 그러나 수십만 건에 이르는 데이터를 매번 수작업으로 가공하는 방식은 시간과 비용 면에서 지속하기 어렵다. 아카이브에서 학습에 적합한 영상을 골라내고 설명문을 자동으로 붙이는 데이터 구축 파이프라인이 필요한 이유다. 이 파이프라인이 한국 맥락을 정확히 읽어내려면 한국 콘텐츠를 이해하고 설명하는 ‘영상 이해 모델’이 먼저 마련돼야 한다.

〈그림 2〉 콘텐츠 제작 전반의 혁신을 위한 MBC의 중장기 계획



출처: MBC

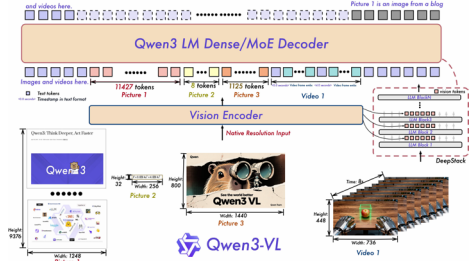
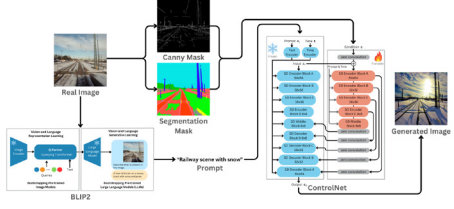
2025년에 수행한 사업은 중장기 계획의 1·2단계에 해당한다. 고품질 학습데이터 5,000시간을 구축하고 이 데이터를 바탕으로 영상 이해 모델과 배경 이미지 생성 모델(이하 ‘배경 생성 모델’)을 개발하는 것이 목표였다. MBC가 사업의 주관기관을 맡았으며 아이엠비씨(iMBC), 클라우드웍스, 데이터메이커, LG AI연구원이 참여했다.

3-2. 무엇을 만들었나: 영상 이해와 생성을 위한 학습데이터 구축

본 사업에서도 학습데이터의 세부 구성에 앞서 AI 태스크를 정의했다. 첫 번째는 한국 고유의 언어·문화·영상 문법을 이해하고 상세한 설명문을 생성하는 ‘K-콘텐츠 영상 이해 고도화’다. 두 번째는 한국 고유의 자연 배경 이미지를 입력한 구도에 맞춰 생성하는 ‘자연 배경 키프레임 이미지 생성 고도화’다. 영상이 아닌 키프레임(배경영상 생성의 기준이 되는 대표 프레임) 이미지 생성을 대상으로 삼은 것은, 키프레임을 먼저 만들고 이를 자연어 프롬프트와 함께 입력하면 연출자의 의도를 더 정확히 반영한 배경영상을 생성할 수 있기 때문이다. 이렇게 정의한 AI 태스크에 맞춰 구축할 학습데이터를 ‘K-콘텐츠 영상 이해 데이터셋’(이하 ‘영상 이해 데이터셋’) 3,000시간, ‘배경영상 상세 설명·생성 데이터셋’(이하 ‘배경영상 데이터셋’) 2,000시간으로 기획했다(표 1).

‘영상 이해 데이터셋’은 MBC 보도·시사교양·드라마·예능 프로그램 영상에서 인물의 행동이 두

〈표 1〉 구축한 학습데이터의 구조와 학습 모델

구분	K-콘텐츠 영상 이해 데이터셋	배경영상 상세 설명·생성 데이터셋
AI 태스크	K-콘텐츠 영상 이해 고도화: 한국 고유의 언어·문화·영상 문법을 반영한 상세 설명문 생성 및 영상 이해	자연 배경 키프레임 이미지 생성 고도화: 입력한 세그멘테이션 맵을 조건으로 한국 자연 배경 키프레임 이미지 생성
원천 영상	- 방송 송출 영상 - 장면(scene) 단위 클립(60초 전후) - 구성 비율: 보도 15%, 시사교양 30%, 드라마 30%, 예능 25%	- 방송 제작 단계의 촬영 원본 영상 - 컷(cut) 단위 클립(20초 전후) - 구성 비율: 자연 35%, 도시 30%, 전통 20%, 배경 예셋 15%
라벨링 구성	- 객체 데이터: 클립 대표 이미지에 주요 객체 바운딩 박스 탐지 - 행동 데이터: 인물의 주요 행동에 대한 상세 설명문(주체·동작·방향·맥락) - 장면 데이터: 영상 전체 맥락 설명문 - VQA 데이터: 행동 인지·의도 추론·시공간 추론을 위한 질의응답 설명문	- 객체 설명 데이터: 영상 내 객체 및 장면 전반 설명문 - 장소 설명 데이터: 시간대·계절·날씨·분위기 등 장소 설명문 - 촬영 정보 데이터: 카메라 움직임·구도·연출 등 촬영 기법 설명문 - 세그멘테이션 데이터: 자연 배경에 대한 시맨틱 세그멘테이션 라벨링
구축 목표	3,000시간	2,000시간
학습 모델	Qwen3-VL(Vision-Language) 모델 	SDXL(Stable Diffusion XL) 기반 컨트롤넷(ControlNet) 
모델 지표	영상 이해 벤치마크 MVBench 64.0 이상	이미지 품질 지표 FID(Fr�chet Inception Distance) 53 이하

출처: MBC

드러지는 약 1분 길이 장면(scene) 단위 클립으로 구성했다. 인물과 상호작용을 하는 객체를 탐지하고, 인물의 행동과 영상 전체 맥락을 각각 행동 설명문과 장면 설명문으로 작성했다. 핵심은 마지막 단계인 시각 질의응답(Visual Question Answering, 이하 VQA) 데이터다. 앞선 정보

를 바탕으로 ‘인물이 왜 그런 행동을 했는지’, ‘다음엔 어떤 일이 일어날지’와 같은 의도와 인과 관계를 묻고 답하는 데이터를 구축했다. 학습데이터를 객체·행동·장면·VQA의 네 층위로 구성한 것은, 모델이 영상 속 대상을 인식하는 데 그치지 않고 행동의 의도와 인과관계까지 추론하도록 학습시키기 위해서다.

‘배경영상 데이터셋’은 한국 고유의 자연·도시·전통 건축물 영상을 대상으로 객체 설명·장소 설명·촬영 정보의 세 가지 설명문을 작성하고(표 1), 여기에 배경영상에서 추출한 이미지를 하늘·산·강·건물 등 19개 클래스로 영역 분할(시맨틱 세그멘테이션, semantic segmentation)한 ‘세그멘테이션 데이터’ 약 3만 건을 더했다. 장소와 촬영 정보까지 설명문으로 작성한 것은 ‘늦가을 오후, 흐린 하늘, 드론 하이앵글’처럼 연출자가 지시하는 조건을 모델이 배우게 하기 위해서다. 이렇게 구축한 학습데이터로 두 태스크의 모델을 각각 학습시켰다. 영상 이해 모델은 알리바바가 공개한 Qwen3-VL을 ‘영상 이해 데이터셋’과 ‘배경영상 데이터셋’의 설명문 데이터로 추가 학습(파인튜닝)해, 한국 영상의 맥락을 설명하고 질문에 답하는 능력을 향상시켰다. 배경 생성 모델은 이미지 생성 모델 SDXL에, 주어진 구도대로 그리도록 안내하는 컨트롤넷을 붙인 구조다. SDXL은 그대로 두고 컨트롤넷만 세그멘테이션 데이터로 학습시켜, 세그멘테이션 맵과 텍스트 프롬프트를 입력하면 그에 맞는 한국 자연 배경 키프레임이 생성되도록 했다.

3-3. 어떻게 만들었나: 단계별 구축 과정

이 글에서는 아카이브에서 수집한 가공 전 영상을 ‘원시데이터’, 정제를 거쳐 클리핑한 영상을 ‘원천데이터’, 여기에 라벨링 데이터를 더해 가공한 결과물을 ‘학습데이터’로 구분했다.

원시데이터 수집: 원시데이터는 MBC가 1961년부터 축적해 온 약 50만 시간의 아카이브에서 수집했다. 영상 이해 데이터셋에는 사람의 행동이 두드러지고 VQA 라벨링에 적합한 원시데이터가, 배경영상 데이터셋에는 생성하려는 배경영상에 맞는 구도로 촬영된 원시데이터가 필요하다. 좋은 원천데이터를 골라내려면 충분한 양의 원시데이터가 필요하므로, MBC는 사업의 목표치인 10,000시간을 채운 뒤에도 수집을 계속해 총 18,600시간(목표의 186%)을 확보했다.

원천데이터 정제: 원시데이터 가운데 AI 태스크에 적합한 부분만 골라내는 단계다. 영상 이해 데이터셋은 맥락이 이어져야 하므로 장면 단위 클립으로, 배경영상 데이터셋은 생성 모델이 배

경영상을 컷 단위로 만들기 때문에 컷 단위 클립으로 분할했다. 장면 전환 탐지 모델(Trans-NetV2)로 컷을 자동 분할한 뒤 작업자가 검수하는 방식으로 방대한 분량을 효율적으로 처리했다.

비식별 처리: 학습데이터에 포함된 개인정보는 비식별 처리가 필수이며, 그 방식은 데이터의 목적에 따라 달랐다. 표정과 행동의 맥락 자체가 학습 대상인 영상 이해 데이터셋은 얼굴을 가리는 대신 생성형 AI 기반 페이스 스왑으로 신원만 바꾸고 표정은 남겼다. 표정이 필요 없는 배경영상은 탐지와 추적을 동시에 수행하는 객체 추적 모델 YOLO11-JDE로 얼굴을 프레임 누락 없이 찾아내 블러 처리했다.

라벨링 가공: 라벨링 가공에서 MBC가 고수한 원칙은 AI 모델로 초안을 만들지 않고 사람이 직접 작성하는 것이었다. 모델이 생성한 초안을 사람이 고치는 방식은 빠르지만, 그 데이터로는 모델이 이미 아는 것 이상을 배우기 어렵고 사람이 영상을 이해하는 독창적인 방식도 익히기 어렵다. 사업 기간이 짧은 상황에서 비용과 시간 부담이 큰 원칙이었지만 끝까지 지켜 구축을 완료했다.

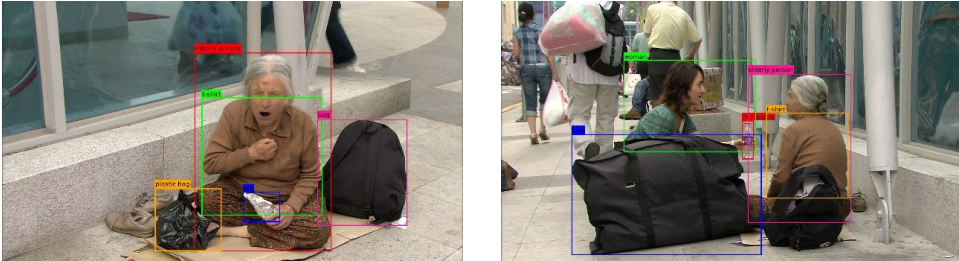
검수·품질 검증: 라벨링 데이터 검수는 일부만 샘플링하지 않고 전수 검수로 진행했다. 작성자와 검수자를 분리하고 숙련된 작업자가 검수를 담당하도록 했으며, 검수자가 반려한 데이터는 다시 수정하는 과정을 거쳤다. 구축한 데이터가 모델 성능을 실제로 높이는지도 학습을 통해 검증했다. 마지막으로 공식 품질 검증 기관에서 형식성·다양성·구문 정확성·의미 정확성·유효성의 다섯 항목을 검증받았으며, 확인된 오류를 모두 보완했다.

3-4. 무엇을 얻었나: 결과물과 모델 성능, 그리고 학습데이터의 의의

사업 수행 결과 총 5,372시간의 학습데이터를 구축했다(목표 5,000시간의 107%). ‘영상 이해 데이터셋’은 3,317시간으로, 객체 데이터 278,958건과 행동·장면·VQA 데이터 총 129,479건이다. ‘배경영상 데이터셋’은 2,055시간으로, 객체 설명 118,165건, 장소 설명 85,534건, 촬영 정보 86,650건, 세그멘테이션 30,273건이다. 표 2와 표 3은 그 예시다.

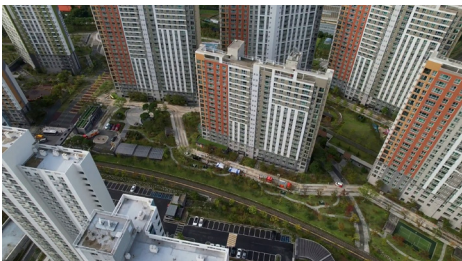
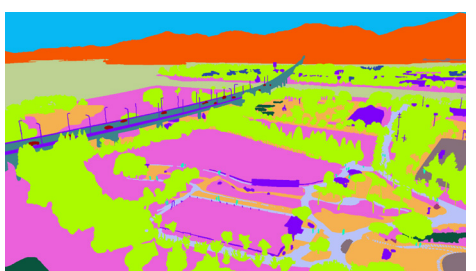
공식 품질 검증 결과, 구축 데이터는 다섯 항목 모두에서 목표치를 충족했다. 데이터 품질에서

〈표 2〉 K-콘텐츠 영상 이해 데이터셋 구축 예시

구분	구축 예시
객체 데이터	
행동 데이터	20초~24초: 초록색 원피스를 입은 여성이 냉장고로 걸어간다. 여성은 오른손으로 냉장고 문을 열고 왼손으로 음료가 든 페트병을 꺼낸다. 여성은 오른손으로 냉장고 문을 닫고 하얀색 가방에서 붉은색 지갑을 꺼낸다. 29초~35초: 초록색 원피스를 입은 여성이 갈색 티셔츠를 입은 노인에게 걸어간다. 여성은 노인 앞에 검은색 가방을 내려놓는다. 여성은 노인 앞에 쪼그려 앉아 왼손으로 페트병을 잡고 오른손으로 뚜껑을 연다. 여성은 페트병을 든 왼손을 앞으로 내민다.
장면 데이터	낮 시간, 사람이 붐비는 활기찬 인도에 있는 에스컬레이터를 초록색 원피스를 입고 검은색과 하얀색 가방을 든 여성이 타고 올라온다. 갈색 티셔츠를 입은 노인이 인도에 양반다리 자세로 앉아 오른손으로 가슴을 두드린다. 여성은 상점 냉장고 문을 열고 음료가 든 페트병을 꺼낸다. 여성은 붉은색 지갑을 하얀색 가방에서 꺼낸다. 열쇠가 바닥의 맨홀 뚜껑 위로 떨어진다. 여성이 노인에게 다가가 페트병 뚜껑을 따서 노인에게 건넨다. 노인은 페트병을 건네받는다. 노인은 여성에게 고개 숙여 인사하고 물을 마신다. 한적한 주택가를 붉은색 자동차가 달린다. 자동차는 주택 앞에 멈춘다.
VQA 데이터	Q1. 초록색 원피스를 입은 여성의 어떤 행동이 계기가 되어 맨홀 뚜껑 위에 열쇠를 떨어뜨렸나요? (1) 냉장고에서 페트병을 꺼냈다. (2) 노인에게 페트병을 건넸다. (3) 가방에서 지갑을 꺼냈다. (4) 에스컬레이터를 타고 올라왔다. Q2. 다음 중 초록색 원피스를 입은 여성이 페트병을 든 왼손을 앞으로 내민 이유는 무엇인지 고르시오. (1) 여성 자신이 페트병 안의 음료를 마시기 위해 (2) 갈색 티셔츠를 입은 노인에게 페트병을 건네기 위해 (3) 갈색 티셔츠를 입은 노인에게 페트병에 담긴 음료를 뿌리기 위해 (4) 페트병 안의 음료를 바닥에 붓기 위해

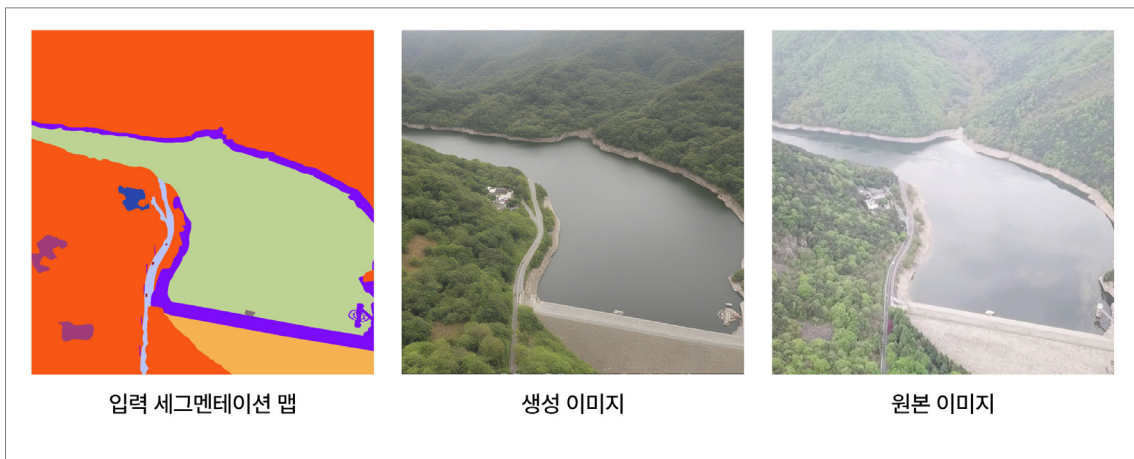
출처: MBC

〈표 3〉 배경영상 상세 설명·생성 데이터셋 구축 예시

구분	구축 예시	
객체 설명 데이터		영상에는 강의 후면에 고층 빌딩들이 밀집해 있는 도시의 스카이라인이 보입니다. 영상 중앙에는 가장 눈에 띄는 붉은색 외관의 빌딩이 우뚝 솟아 있으며, 그 오른쪽으로는 유리 건물이 있습니다. 붉은색 외관 빌딩의 좌측 전면 강 위에는 반구 형태의 유리 돔 구조물이 떠 있습니다. 유리 돔 구조물의 후면에 위치한 강변에는 앙상한 가지의 나무들이 길게 이어져 있습니다.
장소 설명 데이터		촬영 장소는 대규모 전통 가옥 단지입니다. 낮 시간대에 촬영되었고 그림자가 희미한 것을 보아 흐린 날씨입니다. 전통 가옥 단지는 산과 나무에 둘러싸여 있습니다. 나무들은 울창하고 녹음이 짙은 모양입니다.
촬영 정보 데이터		상공에서 하이 레벨 구도로 촬영하던 카메라는 영상이 시작되면 시계 방향으로 회전합니다. 영상 39초 시점 카메라는 화면 상부에 있는 운동 코트를 향하여 앞으로 나아간 후 다시 시계 반대 방향으로 회전합니다. 해당 영상은 드론 등의 항공 촬영 기법을 활용하여 촬영된 영상입니다.
세그멘테이션 데이터	 〈원천 데이터〉	 〈세그멘테이션 데이터〉

출처: MBC

〈그림 3〉 SDXL 기반 컨트롤넷의 배경 이미지 생성 결과



출처: MBC

가장 중요한 의미 정확성은 VQA를 제외한 데이터 유형에서 96.8~99.6%였고, 구축 난도가 가장 높았던 VQA도 93.4%로 목표(90%)를 넘겼다. 학습한 모델의 성능도 모두 목표치를 초과 달성했다. 영상 이해 모델은 MVBench 68.8(목표 64.0 이상), 배경 생성 모델은 FID 24.7(목표 53 이하, 낮을수록 우수)을 기록했다.

학습 전후 영상 이해 모델의 설명문을 비교하면 성능 향상이 뚜렷하다. 눈 덮인 길에 앉아 있는 소년에게 남자가 다가가는 장면에서 학습 전 모델은 ‘음식 한 그릇을 건네 먹도록 돕는다’라고 잘못 서술했지만, 학습 후 모델은 ‘소년이 일어서도록 돕기 위해 몸을 굽힌다’라고 정확히 서술했다. ‘전통적인 주거지’라고만 설명하던 집도 ‘초가지붕’과 ‘꽃무늬 담요’ 같은 한국 고유의 객체를 언급해 묘사했고, 냉장고 앞에 선 소년의 ‘호기심 가득한 표정’처럼 인물의 감정과 상태까지 읽어냈다.

배경 생성 모델 역시 입력한 세그멘테이션 맵과 텍스트 프롬프트에 맞춰 자연스러운 배경 키프레임을 만들어 냈다(그림 3). 특히 강·산·바다처럼 데이터 비중이 큰 클래스는 실제와 유사하게 생성됐다. 다만 구조물처럼 클래스 정의가 세밀하지 않거나 데이터가 적은 영역에서는 어색한 부분이 남았다. 이는 다음 단계에서 개선해야 할 과제이기도 하다.

현재 VLM(Vision-Language Model) 연구의 주요 관심사는 영상 속 객체의 시간적 흐름과 인과 관계를 텍스트로 충실히 옮기는 일이다. 그런데 최근 공개된 학습데이터조차 자율주행 같은 특

정 도메인에 치우쳐 있거나 대량 구축을 위해 자동화 스크립트에 의존하는 경우가 대부분이다. 또한, 시공간 복합 추론·정밀한 인과 추론·현실 세계의 물리 법칙 라벨링을 모두 갖춘 학습데이터는 찾아보기 어렵다. 본 사업에서 구축한 학습데이터는 하나의 영상을 객체·행동·장면·VQA의 네 층위로 심층 분석해 시공간 및 인과 추론까지 갖췄다는 점에서 의의가 크다. 쌓여만 있었던 아카이브가 의미 있는 학습데이터로 바뀌기 시작한 것이다.

④ 구축한 학습데이터는 무엇을 가능하게 했나

학습데이터 구축 사업의 성과는 데이터 구축량이나 모델의 벤치마크 점수 향상에 머물러서는 안 된다. 학습데이터로 개발한 모델을 다시 업무에 활용해 일하는 방식을 바꿀 때 비로소 가치가 생긴다. 사업을 마친 뒤 MBC가 영상 이해 모델을 활용한 두 사례를 소개한다.

먼저 AI 기반 아카이브 검색 시스템의 고도화다. 그동안 아카이브 검색은 제목, 방송 일자, 출연진 등 기본적인 메타데이터에 의존해 왔고, 장면 안에서 누가 무엇을 하는지는 검색할 수 없었다. 사업 결과물인 영상 이해 모델은 장면마다 객체와 행동, 배경을 서술하는 설명문을 자동으로 생성하며, 이 설명문이 곧 검색할 수 있는 메타데이터가 된다. 표 2의 예처럼 ‘노인에게 음료를 건네는 장면’이라는 문장만으로도 영상을 찾을 수 있다. 제작진이 자료를 찾는 데 들이던 시간은 줄고, 묻혀 있던 영상의 활용도 역시 높아진다.

다음으로 영상 이해 모델을 활용한 학습데이터 구축 공정의 효율화다. 학습데이터 구축은 원시데이터 수집부터 설명문 작성·객체 탐지·세그멘테이션 같은 라벨링까지 사람의 손을 거쳐야 하는 노동 집약형 공정이다. MBC는 본 사업으로 개발한 영상 이해 모델을 이 공정에 투입해 원시데이터의 추출·정제와 라벨링 초안 작성을 자동화하고 있다. 이는 3-3의 ‘사람이 직접 작성한다’라는 원칙을 뒤집는 것이 아니라, 사람이 직접 라벨링한 데이터로 모델에 사람의 시각을 가르쳤기에 가능해진 다음 단계다. 다만 모델이 초안을 작성하더라도 사람의 검수와 수정을 필수 단계로 두고, 사람이 직접 작성한 데이터의 비율도 일정 수준 이상으로 유지할 방침이다. 두 공정을 효율화한 핵심은 영상의 맥락과 흐름을 사람에 가깝게 이해하는 영상 이해 성능이며, 그 출발점이 방송 아카이브로 구축한 5,372시간의 학습데이터다. 방송사의 AX는 완성된 AI 도구를 들여오는 데서 끝나지 않는다. 방송사는 자사 데이터로 모델을 만들고 그 모델이 다

시 데이터와 업무를 효율화하는 순환 구조를 갖추는 데 집중해야 한다.

⑤ 방송영상 학습데이터의 가치와 사업화 전략

AI 학습데이터 시장에서 권리가 정리된 고품질 영상의 수요가 커지면서 그 가치도 높아지고 있다. 최근 모델 개발의 경쟁은 텍스트를 얼마나 잘 다루느냐가 아니라, 영상·음성·이미지까지 하나의 모델로 학습해 실제 세계를 얼마나 잘 이해하느냐로 옮겨 가고 있다.

엔비디아(NVIDIA)가 CES 2025에서 피지컬 AI 시대를 선언한 이래 학습데이터 수요는 월드 모델(World Model)과 시각-언어-행동 모델(Vision-Language-Action Model, 이하 VLA)로까지 확대됐고, 이 모델들은 장면 전환과 물리적 변화, 행동과 결과의 인과, 여러 사람의 상호작용을 담은 대규모 실사 영상을 요구한다(표 4). 방송영상은 뉴스부터 드라마·스포츠까지 언어·행동·감정·물리 환경을 고품질로 담고 있다. 여기에 한국의 공간과 사람, 문화가 수십 년의 시간 축 위에 쌓여 있어, 이 요구에 가장 가까운 데이터다.

〈표 4〉 AI 모델 유형별 학습 목적과 학습데이터 형태		
AI 모델	학습 목적	필요한 학습데이터 형태
LLM	언어 이해·추론·생성	자막·대본 텍스트, 발화 전사, 요약·질의응답 쌍, 화자 분리 데이터
VLM	시각-언어 정합 이해, 장면 추론	영상-텍스트 정렬 데이터, 장면 설명(캡셔닝), 장면 분할, 감정·표정·맥락 태깅
월드 모델	실세계 물리 법칙·환경 변화 학습 및 시뮬레이션	시간 축 기반 장면 전환·물리적 변화 시퀀스, 연속 프레임 환경 변화 데이터, 다 시점·공간 구조 데이터
VLA	인지·추론·행동 통합 학습	행동-결과 인과 구조 데이터, 다인 상호작용·객체 추적 데이터, 의도·행동 시퀀스 데이터

출처: MBC

영상 데이터의 수요와 가치가 이처럼 커지면서, 그동안 광고와 콘텐츠 유통에 의존해 온 방송사에도 학습데이터 공급이라는 새로운 수익원이 열리고 있다. 라이온스게이트(Lionsgate)는 2024년 9월 자사 영화·TV 카탈로그를 학습시킨 맞춤형 영상 생성 모델을 런웨이(Runway)와 함께 개발하기로 했고, 뉴욕타임스는 2025년 5월 아마존과 AI 학습 이용 허락(라이선스) 계약을 맺어 연 2,000만~2,500만 달러를 받는 것으로 알려졌다. 글로벌 AI 학습데이터 시장은 2026년 약 44억 달러에서 2034년 약 232억 달러로 연평균 22.9% 성장할 것으로 전망된다. 관건은 방송사가 이 시장에 어떤 방식으로 참여하느냐다. 아카이브를 원시데이터 그대로 판매하는 것은 일회성 수익에 그친다. 하나의 원시데이터에서 장면 분할·감정 분석·행동 인식 등 태스크별 학습데이터를 구축할 수 있는데, 원시데이터를 통째로 넘기면 그 파생 가치는 구매 기업에 돌아가기 때문이다. 방송사의 역할은 원시데이터 판매자가 아니라 가공까지 맡는 학습데이터 큐레이터여야 한다.

이 큐레이터 역할을 사업으로 구체화한 것이 MBC가 구상하는 ‘학습데이터 거래 플랫폼’이다. 아카이브 원본을 파는 대신, 원본은 방송사 안에 두고 수요 기업이 원하는 학습데이터를 주문에 맞춰 만들어 라이선스 방식으로 공급하는 플랫폼이다. 플랫폼은 크게 세 단계로 구성된다. 데이터 레이크(Data Lake)가 아카이브 영상과 메타데이터를 한곳에 모으고, 큐레이션 파이프라인이 요청에 맞춰 데이터를 골라 권리·개인정보 검증과 비식별 처리를 거쳐 라벨링·패키징하며, 마켓플레이스가 검색부터 라이선스 계약과 제공까지 처리한다. 수익원은 목적과 기간을 한정된 라이선스 대가와 큐레이션 서비스가 기본이고, 학습데이터로 개발한 편집·다국어 더빙·하면 해설·심의 보조 같은 버티컬 AI 솔루션을 외부에 서비스형 소프트웨어(SaaS)로 제공하는 사업으로 확장할 수 있다.

학습데이터 거래 플랫폼의 가장 큰 걸림돌은 영상에 얹힌 권리관계다. 방송영상에는 방송사의 저작권 외에도 출연자의 초상권·음성권, 실연자의 저작인접권, 작가와 외주 제작사의 저작권, 음악 저작권 등 여러 권리가 겹겹이 얹혀 있다. 게다가 이 권리들을 다룬 기존 계약은 AI 학습에 쓰이는 상황을 상정하지 않았다. 한 편의 프로그램에 얹힌 권리자만 수십 명에 이른다. 건별 협의와 수작업 정산으로는 학습용 거래의 규모와 속도를 감당할 수 없다.

MBC가 검토하는 해법은 권리의 토큰화(Tokenization)와 자동 배분이다. 영상 단위의 권리자와 지분을 실물 연계 자산(Real World Asset) 토큰으로 등록하고, 목적·기간·재학습 허용 여부 같은 이용 조건은 프로그램 가능한 IP 라이선스(Programmable IP License)에 담아 둔다. 그러면

학습용 이용이 발생할 때마다 스마트 계약(Smart Contract)이 수익을 권리자별 지분대로 자동 배분한다. 잊힌 과거 작품의 출연자와 창작자에게도 수익이 돌아가는, 권리자 누구도 소외되지 않는 거래가 학습데이터 시장의 전제다.

학습데이터 거래 플랫폼은 방송사 한 곳이 추진할 규모의 사업이 아니다. 방송사 연합이 아카이브와 권리 정보를 모으고, AI 모델 개발사와 데이터 가공 업체도 참여해야 한다. 여기에 정부가 학습데이터 권리 처리 기준을 구체화하고 거래 인프라 및 초기 투자를 뒷받침해 줄 때 비로소 가능해진다. 학습데이터 거래 플랫폼이 현실화된다면 방송사가 보유한 아카이브는 K-콘텐츠 산업의 새로운 수익원이 될 것이다. 나아가 한국 시를 한국 데이터로 키우는 데이터 주권의 출발점이 될 수 있다.

⑥ **마치며: 아카이브는 저절로 자산이 되지 않는다**

방송사는 정말 ‘데이터 부자’인가. 지금은 아니지만, 앞으로는 ‘데이터 부자’가 될 수 있다. MBC는 60년 넘게 쌓아 온 아카이브에서 5,372시간의 학습데이터를 구축했고, 그 데이터로 학습한 모델은 한국 영상의 맥락을 읽어내고 한국의 자연 배경을 그려냈다. 나아가 영상 이해 모델은 아카이브 검색과 학습데이터 구축 공정에도 다시 투입됐다. 쌓여만 있던 아카이브가 모델의 성능을 끌어올리고 일하는 방식을 바꾸는 자산으로 변하기 시작한 것이다.

인공지능 시대에 중요한 것은 방송사가 보유한 아카이브의 양이 아니라 이를 얼마나 학습데이터로 가공했느냐다. 어떤 데이터를 만들지는 방송사가 AX 수요를 파악해 정의한 AI 태스크에 달려 있다. 태스크가 데이터의 구조를 정하고, 데이터가 모델의 성능을 정하며, 그 모델이 AX의 성패를 가른다.

방송사의 학습데이터는 일차적으로 AI 모델 성능 향상에 활용되지만, 나아가 방송사의 새로운 수익원이 될 수 있다. AI 모델 경쟁의 무게중심이 멀티모달로 옮겨 가면서 고품질 영상 데이터는 인공지능 시대의 핵심 원자재가 됐다. 방송사가 콘텐츠를 시청자에게 전달하는 역할에 머물지 않고, AI 모델의 성능을 높이는 ‘학습데이터 큐레이터’로 스스로를 다시 정의할 때다.

MBC가 제안하는 학습데이터 거래 플랫폼은 지상파 방송사 중심의 K-콘텐츠 데이터 레이크에서 출발해 AI 데이터를 자유롭게 거래하는 국가대표 플랫폼으로 확장하는 구상이다. 흩어져 있

는 K-콘텐츠를 ‘한국형 특화 AI 데이터’로 묶어 내면, 서구권 데이터 위주로 학습된 글로벌 모델의 문화적 편향에서 벗어나 한국의 문화와 맥락을 제대로 이해하는 AI를 키울 토대가 된다. 국내 AI 사업자는 한국형 특화 모델에 필요한 고품질 데이터를 안정적으로 공급받을 수 있고, 창작자는 잊힌 과거 창작물에서 데이터 수익을 얻어 창작 활동의 기반을 마련할 수 있다. 활용처는 콘텐츠 산업에만 그치지 않는다. 방송영상에 담긴 한국의 지형과 사람의 행동 패턴은 로봇·자율주행 등 피지컬 AI의 학습데이터이며, 게임의 시뮬레이션 데이터와 결합하면 현실과 가상을 오가는 고도화된 학습도 가능해진다.

아카이브는 저절로 자산이 되지 않는다. 광고와 콘텐츠 유통에 기대 온 방송사의 전통적인 수익 구조가 흔들리는 지금, 방송사에 여전히 남아 있는 자원은 아카이브다. 인공지능 시대에 AX를 실행할 의지가 있는 방송사가 아카이브를 새로운 자산으로 전환할 수 있다.

레퍼런스



참고문헌

- 권석원 (2026). Beyond Media with AI: 미디어·콘텐츠 산업의 미래와 생존 전략. <방송문화> 2026년 봄호. 한국방송협회.
- 방송미디어통신위원회 (2026. 6. 19). 2025 회계연도 방송사업자 재산상황 공표.
- 서울신문 (2025. 1. 7). 8년만에 CES 기조연설 나선 젠슨 황 "다음은 '피지컬 AI' 시대".
- 신민경 (2026. 9. 7). "AI가 공부할 데이터, 고갈되나"...불붙는 데이터 확보전. <뉴스1>.
- 유진희 (2024. 8). 제작비 폭등에 따른 국내 드라마 시장의 변화와 개선방안. <미디어 이슈&트렌드> Vol.63. 한국방송통신전파진흥원.
- 이유나 (2025. 9. 4). 어도비시 왜 이러나..."한복 그려 달랬더니 기모노 나와". <YTN>.
- 한국콘텐츠진흥원 (2026. 4). 2025년 4분기 및 연간 콘텐츠산업 동향 분석 보고서.

[국외 문헌]

- Fortune Business Insights (2026. 8). AI Training Dataset Market Size, Share & Industry Analysis, By Type, By Deployment Mode, By End-Users, and Regional Forecast, 2026-2034.
- Gunasekar, S., Zhang, Y., Aneja, J., et al. (2023). Textbooks are all you need. arXiv:2306.11644.
- Qwen Team (2025). Qwen3-VL. <GitHub>
- Shumailov, I., Shumaylov, Z., et al. (2024). AI models collapse when trained on recursively generated data. Nature, 631, 755-759.
- Wang, Z., Lu, Z., Jin, B., & Deng, H. (2023). MediaGPT: A large language model for Chinese media. arXiv:2307.10930.

