졸업논문
소개

실제 서울 거리를 따라 영상을 생성하는 서울 월드 모델

서준영 / KAIST 인공지능대학원 컴퓨터비전연구실

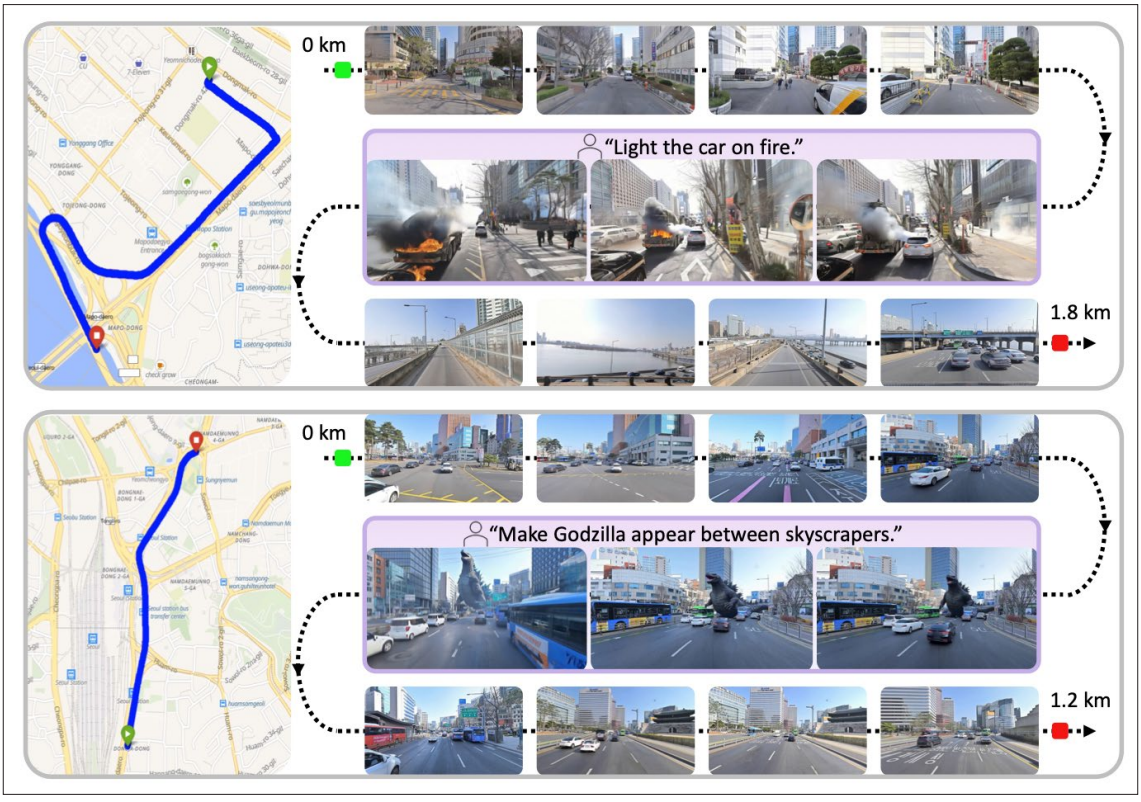
영상 생성 인공지능은 몇 장의 사진과 문장만으로 그럴듯한 도시 풍경을 만들어 낼 수 있다. 하지만 화면이 사실적으로 보인다고 해서 실제 거리의 건물과 도로가 그 자리에 있는 것은 아니다. 출발 지점의 모습이 주어져도 다음 골목의 형태는 모델이 추측하며, 이동 거리가 길어질수록 같은 도시를 계속 걷고 있다는 감각이 약해질 수 있다. 실제 장소를 탐색하거나 특정 지역에서 일어날 상황을 시각화하려면, 영상의 매 순간이 현실의 위치와 연결되어야 한다.

본 연구에서 제안한 서울 월드 모델(Seoul World Model, SWM)은 이 문제를 거리뷰 사진의 검색과 영상 생성의 결합으로 풀었다. 사용자가 지도에서 출발 위치와 카메라의 이동 경로를 정하고 원하는 장면을 문장으로 입력하면, 모델은 경로 주변의 거리뷰를 찾아 그 지역의 모습에 맞춰 다음 영상을 만든다. 생성한 영상을 다시 다음 구간의 입력으로 사용하는 방식이므로 여러 구간을 이어 긴 이동 경로를 표현할 수 있다. 도로의 구조와 건물의 외형은 실제 서울의 자료를 참조하면서, 날씨나 등장 물체 등은 문장으로 바꿀 수 있다.

핵심 자료는 서울의 주요 도로를 촬영한 거리뷰 파노라마 약 120만 장이다. 연구에서는 번호판과 보행자를 흐리게 처리한 뒤, 전처리를 거친 약 44만 장을 학습에 사용했다. 거리뷰는 위치를 알려 주는 풍부한 자료지만, 한 장의 사진은 한 시점의 기록이다. 같은 거리라도 촬영 날짜에 따라 자동차와 사람, 간판이나 조명이 달라진다. 촬영 당시 지나가던 차를 그대로 복사하면 새 영상의 움직임과 충돌하고, 실제 장소의 고정된 구조도 제대로 구분하지 못한다.

이를 줄이기 위해 학습할 때 참고 거리뷰와 정답 영상에 서로 다른 시점에 촬영된 자료를 짝지었다. 모델은 두 자료에 공통으로 남는 도로와 건물 같은 구조를 더 신뢰하고, 일시적으로 등장한 차량이나 보행자는 구별하도록 유도된다. 실제 사용 때 검색된 사진이 현재의 거리와 다른 시기에 촬영되더라도, 장소의 뼈대를 참조할 수 있게 하는 학습 방식이다.

거리뷰는 연속 동영상처럼 촘촘히 찍히지 않는다는 문제도 있다. 사진 사이 간격이 보통 5~20m이므로 그대로 이어 붙이면 카메라가 순간이동한 듯 보인다. 연구에서는 듬성듬성한 거리뷰를 중간 프레임이 있는 영상으로 바꾸는 보간 과정을 설계했다. 특히 비디오 압축 표현이 네 프



<그림 1> 실제 서울의 경로를 따라 생성한 영상과 문장으로 바꾼 장면. 원 논문의 <그림 1>을 발췌했다.

레임을 한 단위로 다루는 점에 맞춰 주요 사진을 짧게 반복하는 방식으로, 학습 중 사진의 위치를 안정적으로 지키면서 부드러운 이동 영상을 얻었다.

실제 거리뷰만으로는 도로 위 차량처럼 앞으로 달리는 시점에 학습이 치우치기 쉽다. 보도에서 걷거나 횡단보도를 건너는 움직임, 건물 사이를 자유롭게 바라보는 카메라까지 다루기 위해 도시 시뮬레이터 CARLA에서 약 1만 2,700개의 영상을 추가로 만들었다. 보행자, 차량, 자유 카메라 경로를 함께 학습해 사용자가 입력할 수 있는 이동 방식의 폭을 넓혔다.

영상 생성 단계에서는 먼저 목표 경로와 가까운 거리뷰를 찾고, 실제 카메라가 바라볼 방향에서 보이는 영역이 충분한지 깊이 정보를 이용해 확인한다. 선택된 사진은 두 경로로 모델에 전달된다. 하나는 사진을 목표 시점으로 옮겨 도로와 건물의 배치를 알려 주는 기하학적 참조이고, 다른 하나는 원본 사진을 직접 제공해 색과 외관의 세부 정보를 보완하는 의미적 참조이다. 옮겨 그린 사진만 사용하면 가려진 부분과 깊이 추정 오차가 남고, 원본만 사용하면 위치 관계가 흐려질 수 있어 두 정보를 함께 쓴다.

긴 영상에서 특히 어려운 부분은 오차의 누적이다. 모델은 앞서 만든 장면을 참고해 다음 장면을 생성하므로 작은 실수가 뒤로 전해진다. 출발 장면을 계속 고정된 기준으로 삼으면 카메라가 수백 미터를 이동한 뒤에는 그 기준이 현재 장소와 멀어진다. SWM의 가상 전망 앵커(Virtual

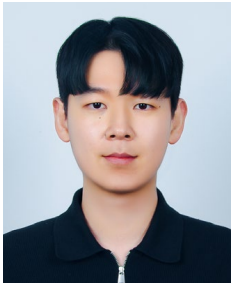
Lookahead Sink)는 매 구간마다 진행 방향 근처의 새 거리뷰를 찾아 앞으로 도착할 장면처럼 제시한다. 앞으로 갈 위치에 깨끗한 기준 영상을 계속 새로 놓아 주는 셈이며, 이를 통해 생성 결과가 실제 경로에서 벗어나는 현상을 줄인다.

연구팀은 서울에서 학습한 모델을 부산과 미국 앤아버의 도시 벤치마크에서도 평가했다. 두 도시는 학습에 포함되지 않았으며, 각 벤치마크는 약 100m 경로를 다루는 30개 영상으로 구성된다. 비교에서는 화면의 자연스러움과 시간적 일관성, 입력한 카메라 경로를 따르는 정도, 실제 장소의 고정된 구조를 재현하는 정도를 함께 살폈다. 서울 월드 모델은 비교 대상 모델들보다 전반적으로 좋은 성능을 보였고, 부산 평가의 영상 품질 지표 FID는 28.43으로 비교 대상 중 가장 낮은 49.63보다 낮았다. FID는 값이 낮을수록 기준 영상의 분포에 더 가깝다는 뜻이다.

구성 요소를 하나씩 빼는 실험도 설계의 이유를 확인해 준다. 촬영 시점을 달리한 짝짓기를 없애자 부산 평가의 FID가 28.43에서 44.74로 나빠졌다. 합성 도시 영상을 제외했을 때는 다양한 카메라 경로를 따르는 능력과 장소 구조의 일치도가 낮아졌다. 가상 전망 앵커를 제거하거나 출발 장면만 고정 기준으로 사용할 때에는 시간이 흐르면서 화질 저하가 더 빨리 나타났다. 이는 검색된 사진을 많이 주는 것만으로 충분하지 않고, 어떤 자료를 어떻게 학습하고 언제 다시 참조하는지가 중요함을 보여 준다.

서울의 장거리 시연에서는 수 킬로미터 경로를 따라 이어지는 영상을 만들었고, 도심에 큰 파도가 밀려오거나 빌딩 사이에 거대한 등장물이 나타나는 장면도 문장으로 제어했다. 이러한 결과는 실제 도시의 배치를 유지하면서 가상의 상황을 시각화할 가능성을 보여 준다. 다만 생성 영상은 현실을 측정한 기록이나 교통·재난 결과의 물리적 예측이 아니다. 지도에 연결된 시각적 시뮬레이션이라는 성격을 분명히 하고 활용해야 한다.

이번 연구는 대규모 거리뷰가 도시 전체를 위한 외부 기억으로 작동할 수 있음을 보여 준다. 촬영 시점의 차이와 드문 공간 표본을 학습 자료로 다루고, 매 구간에서 주변 장소를 다시 찾아 장거리 생성의 흔들림을 줄였다는 점이 핵심이다. 앞으로 더 정확한 위치·깊이 정보와 다양한 실제 이동 영상을 결합한다면, 도시 탐색과 지역 기반 콘텐츠, 계획안의 시각화에 활용할 수 있는 기반이 될 것이다. (프로젝트 페이지: https://seoul-world-model.github.io/)



서 준 영

- 2016년 ~ 2022년 : 고려대학교 전기전자공학부 학사
- 2023년 ~ 2024년 : Sony AI 연구 인턴
- 2024년 ~ 현재 : KAIST AI 석박사통합과정
- 2025년 : Meta 연구 인턴
- 2025년 ~ 2026년 : NAVER AI Lab 협업
- 2026년 ~ 현재 : NVIDIA 연구 인턴
- 주관심분야 : 비디오 월드 모델, 생성형 AI, 3D 컴퓨터비전