지속성 있는 비디오 월드 모델 (Persistent Video World Model) 연구 동향

ㅁ **김승룡** / KAIST

요 약

최근 대규모 비디오 생성 모델의 발전과 함께, 이미지·텍스트·행동(action)을 입력으로 받아 환경의 미래 상태를 영상으로 예측하는 비디오 월드 모델(video world model)이 차세대 시뮬레이션 기술로 주목받고 있다. 비디오 월드 모델이 로보틱스, 자율주행, 게임, 디지털 트윈 등에 실질적으로 활용되기 위해서는 행동에 대한 정밀한 제어성과 실시간 상호작용을 위한 효율성뿐 아니라, 긴 시간 동안 생성된 세계가 일관되게 유지되는 지속성(persistence)이 필수적이다. 본고에서는 지속성 있는 비디오 월드 모델링(persistent video world modeling)을 중심으로 최근 연구 동향을 살펴본다. 행동 조건부 생성과 자기회귀 생성 기법을 시작으로, KV 캐시 기반의 암묵적 월드 메모리, 3차원 표현 기반의 명시적 월드 메모리, 실세계에 접지된(grounded) 월드 모델, 동적 객체 및 다중 에이전트로의 확장 연구를 차례로 다루고, 월드 액션 모델(world action model)로의 확장과 향후 연구 방향을 논의한다.

Ⅰ. 서 론

월드 모델(world model)은 에이전트가 환경에 대한 내부 표현을 학습하고, 주어진 행동에 따라 환경이 어떻게 변화할지를 예측하는 모델을 의미한다[1]. 오랫동안 강화학습과 로보틱스 분야에서 논의되어 온 이 개념은 최근 대규모 비디오 생성 모델의 등장으로 새로운 국면을 맞이하였다. 확산 모델[2]과 Diffusion Transformer(DiT)[3]를 기반으로 한 Sora[4] 등의 모델은 물체의 운동, 조명, 가려짐과 같은 물리 세계의 규칙성을 상당 부분 습득한 것으로 평가받으며, Google DeepMind의 Genie[5]는 사용자 입력에 반응하여 실시간으로 탐험 가능한 세계를, NVIDIA의 Cosmos[6]는 로보틱스와 자율주행을 위한 월드 파운데이션 모델을 제시하였다.

그러나 비디오 월드 모델이 실제 시뮬레이터로 기능하기 위해서는 짧은 클립 생성에서는 드러나지 않던 일관성(consistency) 문제를 해결해야 한다. 예를 들어 카메라를 돌려 어떤 장소를 벗어났다가 돌아왔을 때 건물이나 가구의 배치가 바뀌어 있거나, 시야 밖으로 나간 사람이 사라지거나 엉뚱한 위치에 다시 나타나는 현상이 빈번하게 관찰된다. 본고에서는 생성된 세계가 시간 경과와 시점 변화에도 정적 구조, 동적 객체의 상태, 상호작용의 결과를 일관되게 유지하는 성질을 지속성(persistence)이라 부르고, 이를 중심으로 최근 연구 동향을 정리한다.

Ⅱ. 비디오 월드 모델의 기초

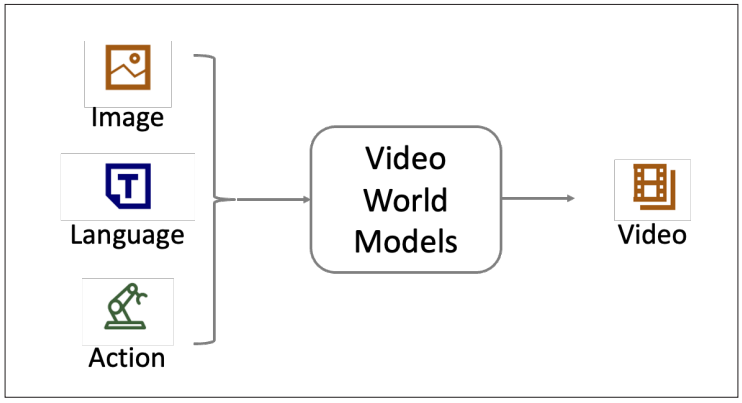
현재 대부분의 비디오 월드 모델은 확산 모델에 기반한다. 확산 모델은 데이터에 점진적으로 잡음을 더하는 전방 과정과 잡음으로부터 데이터를 복원하는 역방향 과정으로 구성되며[2], 비디오 생성에서는 3차원 VAE로 압축한 시공간 잠재 표현 위에서 DiT[3]가 잡음을 예측하는 구조가 사실상 표준으로 자리 잡았다. 비디오 월드 모델은 이미지, 텍스트 프롬프트, 행동을 조건으로 영상을 생성하며, 생성되는 각 프레임은 시뮬레이션 세계의 예측된 상태로 간주된다(〈그림 1〉).

비디오 월드 모델은 세 가지 요구 조건을 만족해야 한다. 첫째, 카메라 이동, 키보드 입력, 로봇 행동 등에 정확히 반응하는 행동 조건부(action-conditioned) 생성이다. 둘째, 이전에 생성된 장면과 객체의 상태를 기억하고 유지하는 일관성으로, 이는 메모리 및 KV 캐시 설계와 직결된다. 셋째, 실시간 상호작용을 위한 효율성으로, 영상 전체를 한 번에 생성하는 대신 청크(chunk) 단위로 순차 생성하는 자기회귀(autoregressive) 생성이 요구된다. 세 조건은 서로 얽혀 있어, 과거 정보를 모두 저장하면 일관성은 높아지지만 효율성이 떨어지고, 행동 제어가 부정확하면 같은 장소로 돌아오는 것 자체가 불가능해진다.

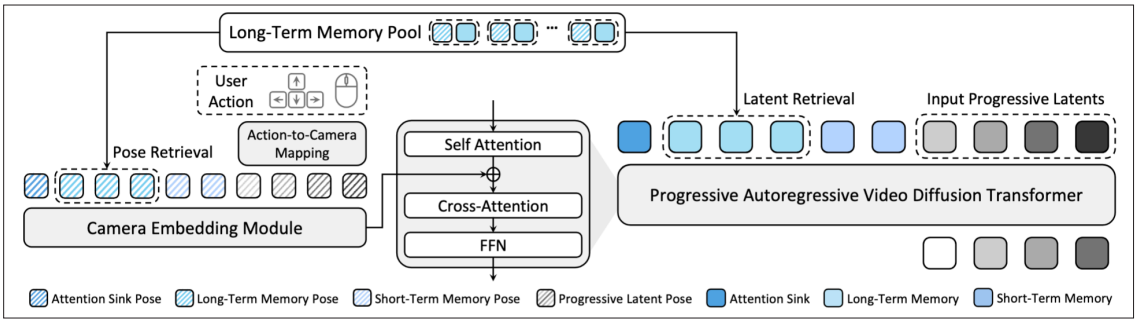
Ⅲ. 행동 조건부 생성과 실시간 자기회귀 생성

1. 카메라 제어 기반 행동 조건부 생성

비디오 월드 모델에서 가장 기본적인 행동은 카메라의 움직임이다. 초기 연구들은 카메라 포즈를 임베딩하여 사전학습된 비디오 생성 모델에 주입하였으며, GEN3C[7]



<그림 1> 비디오 월드 모델의 개념

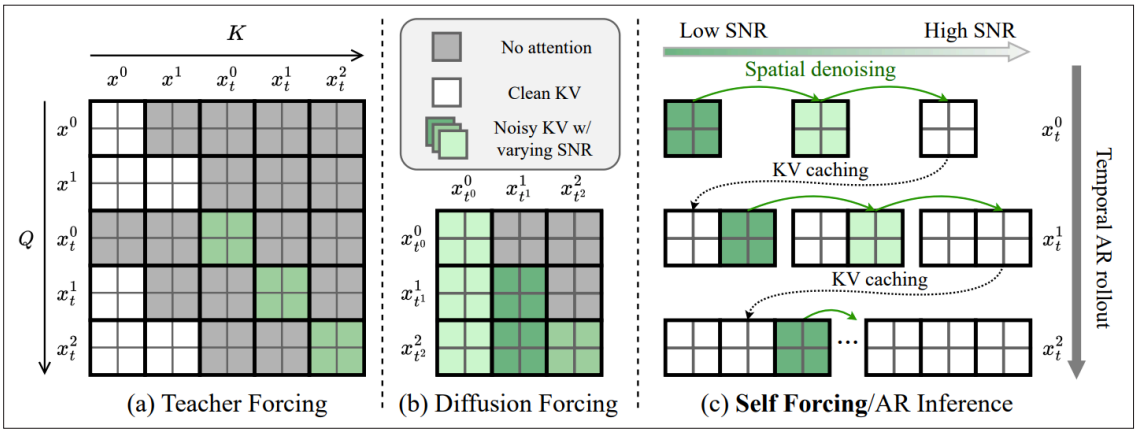


<그림 2> WorldCam의 전체 구조

는 입력 영상의 깊이로부터 3차원 캐시를 구성하고 이를 목표 궤적으로 렌더링한 결과를 조건으로 사용하여 카메라 제어 정확도와 3차원 일관성을 함께 향상시켰다. 게임 환경에서는 Matrix-Game 2.0[8]처럼 키보드·마우스 입력에 따라 실시간으로 세계를 생성하는 모델이 활발히 연구되고 있다. WorldCam[9]은 사용자 행동을 리 대수(Lie algebra) 상의 카메라 포즈로 변환하여 점진적 자기회귀 비디오 확산 트랜스포머의 조건으로 사용하고(〈그림 2〉), 과거 잠재 표현과 포즈를 장기 메모리 풀에 저장한 뒤 현재 포즈와 가까운 메모리를 검색하여 3차원 일관성을 확보한다. 약 3,000분 분량의 게임 플레이 데이터로 학습된 WorldCam은 정밀한 행동 제어, 장시간 추론, 3차원 일관성 측면에서 기존 모델 대비 우수한 결과를 보였다.

2. 자기회귀 비디오 확산과 오차 누적 문제

자기회귀 생성의 가장 단순한 학습 방식인 Teacher Forcing(TF)과 프레임마다 다른 잡음 수준을 부여하는 Diffusion Forcing(DF)[10]은 학습 시에는 깨끗한 문맥을, 추론 시에는 모델이 생성한 불완전한 문맥을 사용하므로 학습-추론 불일치가 발생하고, 생성이 길어질수록 오차가 누적된다. Self Forcing(SF)[11]은 학습 단계에서도 KV 캐시를 사용해 실제 추론과 동일한 자기회귀 롤아웃을 수행하고 생성된 비디오 전체에 손실을 적용함으로써 이 불



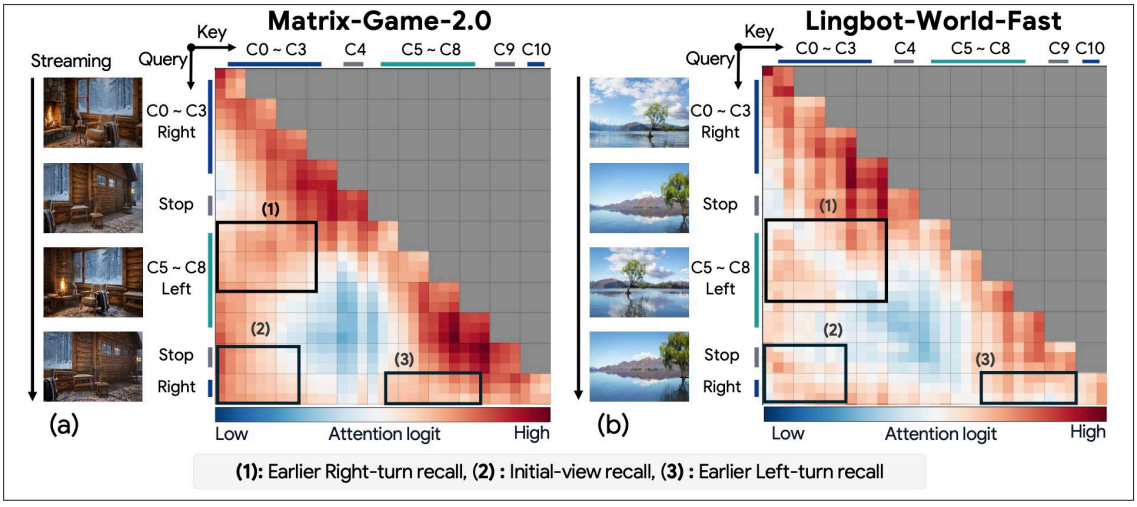
<그림 3> 자기회귀 비디오 확산 모델의 학습 방식 비교(TF, DF, SF)

일치를 해소하고 실시간 생성을 가능하게 하였다(〈그림 3〉). Rolling Forcing[12]은 잡음 수준이 점진적으로 증가하는 연속 프레임을 함께 잡음 제거하는 롤링 윈도우와, 초기 프레임의 KV를 전역 기준점으로 보존하는 어텐션 싱크(attention sink)를 결합하였다. Deep Forcing[13]은 추가 학습 없이 캐시의 약 50%에 달하는 확대된 어텐션 싱크를 유지하고, 최근 프레임의 어텐션 점수를 기준으로 중요한 상위 토큰만 보존하는 참여적 압축(participative compression)을 적용하였다. 이는 KV 캐시를 단순한 버퍼가 아닌, 무엇을 기억하고 잊을지를 결정하는 메모리로 바라보는 관점의 전환이라는 점에서 이후의 월드 메모리 연구와 맞닿아 있다.

Ⅳ. 암묵적 월드 메모리: KV 캐시 기반 접근

지속성을 확보하는 첫 번째 접근은 트랜스포머의 KV 캐시 자체를 세계에 대한 메모리로 활용하는 것이다. 대규모 언어 모델의 문맥 내 학습 능력과 유사하게, 최근 월드 모델들은 명시적인 메모리 학습이나 긴 영상 학습 없이도 모든 과거 프레임의 KV 캐시(full KV cache)를 유지하면 이전에 방문한 장소를 상당히 일관되게 재현하는 월드 메모리를 보인다[8]. 그러나 전체 KV 캐시는 생성 길이에 비례하여 GPU 메모리와 연산량이 증가하므로 실용적이지 않고, 슬라이딩 윈도우는 윈도우를 벗어난 장면을 완전히 잊어버린다.

월드 모델의 어텐션 맵을 분석하면, 카메라가 이전에 보았던 방향으로 돌아왔을 때 현재 쿼리가 해당 시점의 과거 KV에 강하게 집중하는 어텐션 희소성이 관찰된다(〈그림 4〉). 즉, 모델은 이미 어떤 과거 KV가 현재 시점과 대응되는지 알고 있다. 이에 기반한 월드 검색(world retrieval)은 카메라 포즈, 시야각(FOV) 중첩, 쿼리-어텐션 점수 등의 유사도로 필요한 KV만을 선택적으로 불러온다. 또한 비디오 확산 모델의 키-키 매칭만으로 점 추적이 가능하다는 최근 연구[14]에 착안하여, 새 프레임의 키가 기존 키와 유사하면 이미 저장된 영역으로 보고 제거하고 유사도가 낮으면 새로 드러난 공간으로 보고 저장하는 월드 압축(world compression)을 수행할 수 있다. 검색과 압축을 결합한 WorldKV[15]는 전체 KV 캐시보다 훨씬 적은 메



<그림 4> 월드 모델의 어텐션 희소성과 시점 대응 KV 캐시

모리로 유사한 수준의 재방문 일관성을 달성하였다. 암묵적 메모리 접근은 사전학습된 모델의 능력을 그대로 활용하며 별도의 3차원 복원 모듈이 필요 없다는 장점이 있다.

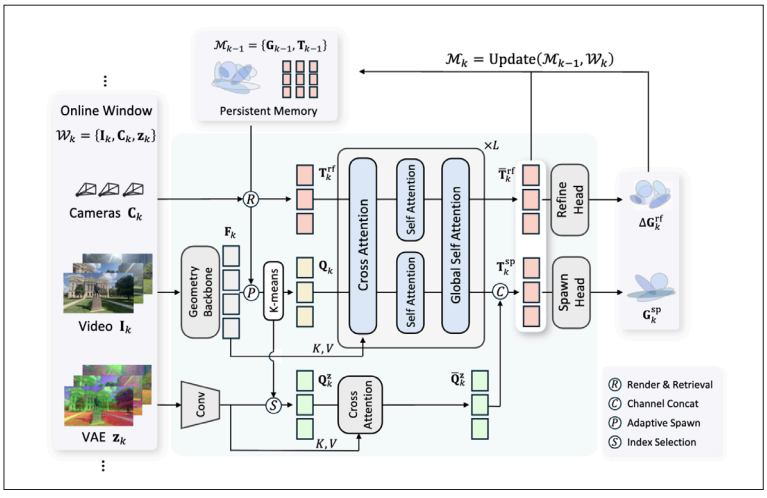
V. 명시적 월드 메모리: 3차원 표현 기반 접근

두 번째 접근은 생성된 세계를 명시적인 3차원 표현으로 저장하는 것이다. 생성된 영상으로부터 3차원 장면을 복원하여 메모리에 저장하고, 다음 청크 생성 시 미래 카메라 궤적으로 메모리를 렌더링하여 조건으로 제공하면, 이전에 방문한 영역이 메모리로부터 렌더링되므로 일관성이 자연스럽게 유지된다. 3D Scene Prompting[16]은 동적 SLAM으로 입력 영상의 정적 영역만을 포인트 클라우드로 복원하여 목표 궤적으로 워핑한 뒤 다음 청크를 생성함으로써 기존 카메라 제어 방법 대비 장면 일관성을 크게 향상시켰고, Spatia[17]는 장면 포인트 클라우드를 생성 과정에서 반복적으로 갱신하는 공간 메모리를 제안하였다.

그러나 장시간 생성으로 확장하면 두 가지 문제가 발생한다. 첫째, 픽셀 단위 포인트 클라우드는 832×480 해상도 81프레임 기준 약 1억 8,950만 개의 값을 필요로 하며, 워핑된 영상을 다시 VAE로 인코딩해야 하므로 실시간 추론이 어렵다. 둘째, 청크 단위로 3차원 캐시를 누적하면 정렬 오차가 쌓이고 중복되거나 잘못 배치된 프리미티브가 메모리를 오염시킨다. 이를 해결하기 위해 소수의 3D 가우시안으로 장면을 표현하는 컴팩트 3D 가우시안 메모리가 제안되었다(〈그림 5〉). 기반이 되는 C3G[18]는 약 2천 개의 가우시안만으로 장면을 표현하여, 픽셀 단위 표현 대비 가우시안 수를 65.5배, 메모리를 15배 줄이면서도 PSNR을 0.57 dB 향상시켰다. 메모리 갱신기가 생성된 영상을 컴팩트 가우시안으로 복원하여 전역 메모리를 자기회귀적으로 갱신하고, 이를 미래 궤적으로 렌더링한 결과를 조건으로 DiT가 다음 청크를 생성한다. 이 방식은 생성 길이가 늘어나도 메모리와 지연 시간이 거의 증가하지 않으며, Spatia 대비 장면 일관성도 우수하였다.

VI. 실세계에 접지된 월드 모델

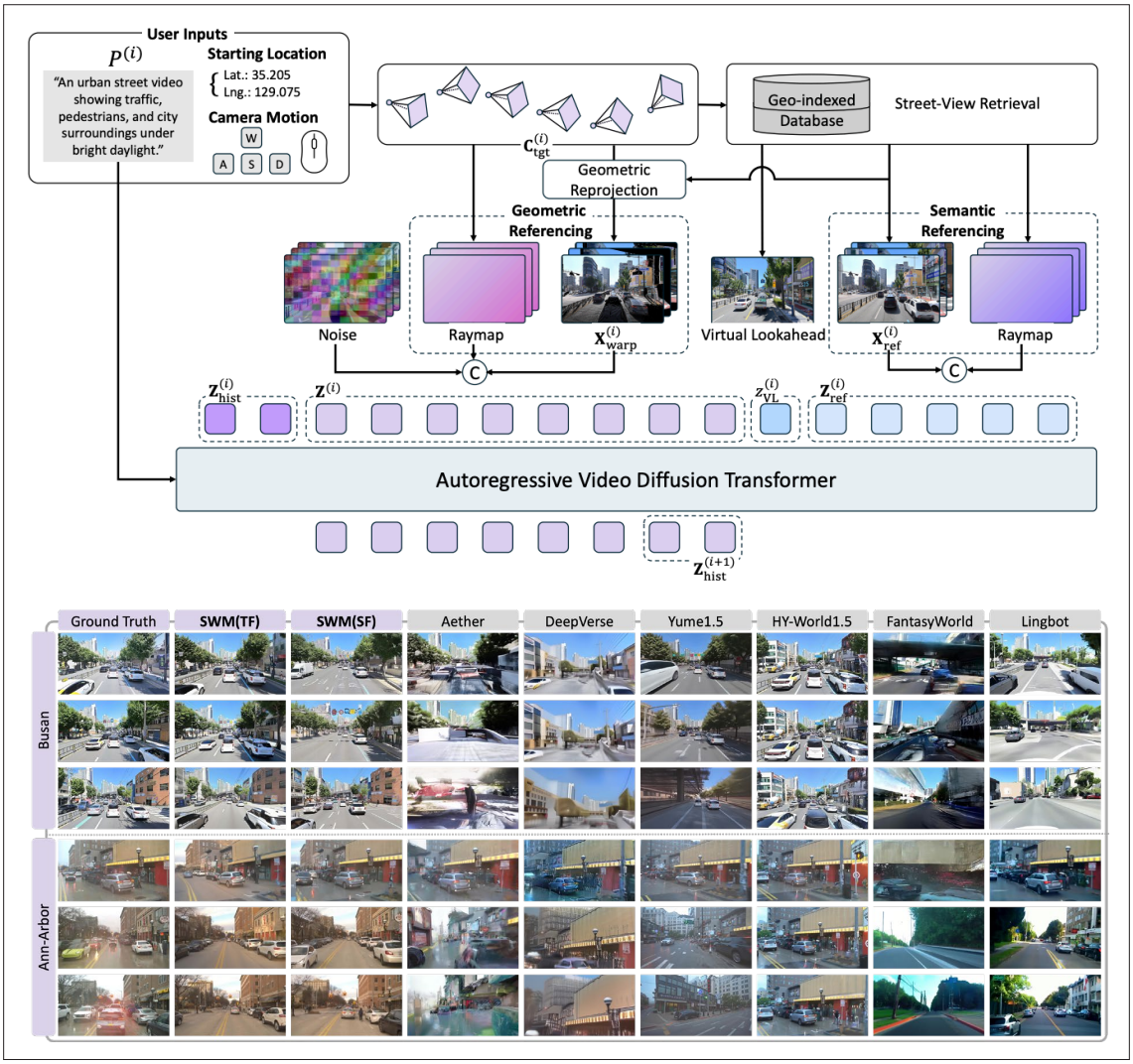
지금까지의 월드 모델은 시작 이미지 너머의 모든 것을



<그림 5> 컴팩트 3D 가우시안 메모리 기반 월드 모델의 구조

모델이 상상하여 만들어 낸다. 만약 월드 모델이 실재하는 세계 위에서 동작한다면, 사용자는 익숙한 도시를 탐험하며 가상의 시나리오를 체험할 수 있고, 이는 도시 계획, 지역 특화 자율주행 시나리오 생성, 도시 단위 내비게이션 등으로 이어질 수 있다. Seoul World Model(SWM)[19]은 서울시 거리뷰 데이터베이스를 검색하여 생성에 활용하는 검색 증강 생성(retrieval-augmented generation, RAG) 기반 월드 시뮬레이션을 제안하였다(〈그림 6〉). 사용자가 시작 위치(위도·경도), 카메라 궤적, 텍스트 프롬프트를 입력하면 모델은 주변의 실제 거리뷰를 참조하여 실제 도시와 부합하는 영상을 생성한다. 이를 위해 약 44.8 km × 31.0 km 범위에서 수집한 120만 장의 파노라마를 핀홀 영상으로 렌더링하고, 미터 단위 깊이와 6자유도 포즈를 추정하여 GPS 좌표와 정렬하였다.

실세계 접지에는 두 가지 난제가 있다. 첫째, 거리뷰는 특정 순간을 담고 있지만 시뮬레이션 세계는 동적이어야



<그림 6> Seoul World Model의 전체 구조와 결과

한다. SWM은 참조 영상을 목표 시퀀스와 다른 시점에 촬영된 것으로 구성하는 교차 시간 페어링(cross-temporal pairing)을 도입하여, 모델이 일시적 객체 대신 변하지 않는 공간 구조에 의존하도록 학습하였다. 둘째, 거리뷰는 약 5~20 m 간격으로 희소하게 촬영되므로, 뷰 보간으로 연속적인 학습 영상을 구성하고 CARLA 시뮬레이터로 보행자, 차량, 자유 카메라 궤적의 합성 데이터를 추가하였다. 검색된 참조 영상은 깊이 기반으로 목표 시점에 워핑하여 공간 배치 단서를 제공하는 기하적 참조와, 원본 잠재 표현을 시간 축으로 결합하여 외관 세부 정보를 제공하는 의미적 참조로 함께 활용된다. 또한 카메라가 수백 미터를 이동하면 초기 프레임 기반 어텐션 싱크가 무의미해지므로, 궤적 종점 근처의 미래 거리뷰를 싱크로 사용하는 가상 전방 싱크(virtual lookahead sink)를 제안하여 드리프트를 억제하였다. Self Forcing으로 학습된 SWM은 단일 H100 GPU에서 15.2 fps로 동작하며, 부산과 미국 앤아버 벤치마크에서 기존 월드 모델 대비 우수한 영상 품질, 카메라 추종 정확도, 3차원 정합도를 보였고 수 킬로미터 궤적에서도 오차 누적 없이 영상을 생성하였다. 다만 뷰 보간 기반 의사 정답이 품질의 상한을 제한하고 동적 객체를 명시적으로 모델링하지 않는다는 점은 향후 과제로 남아 있다.

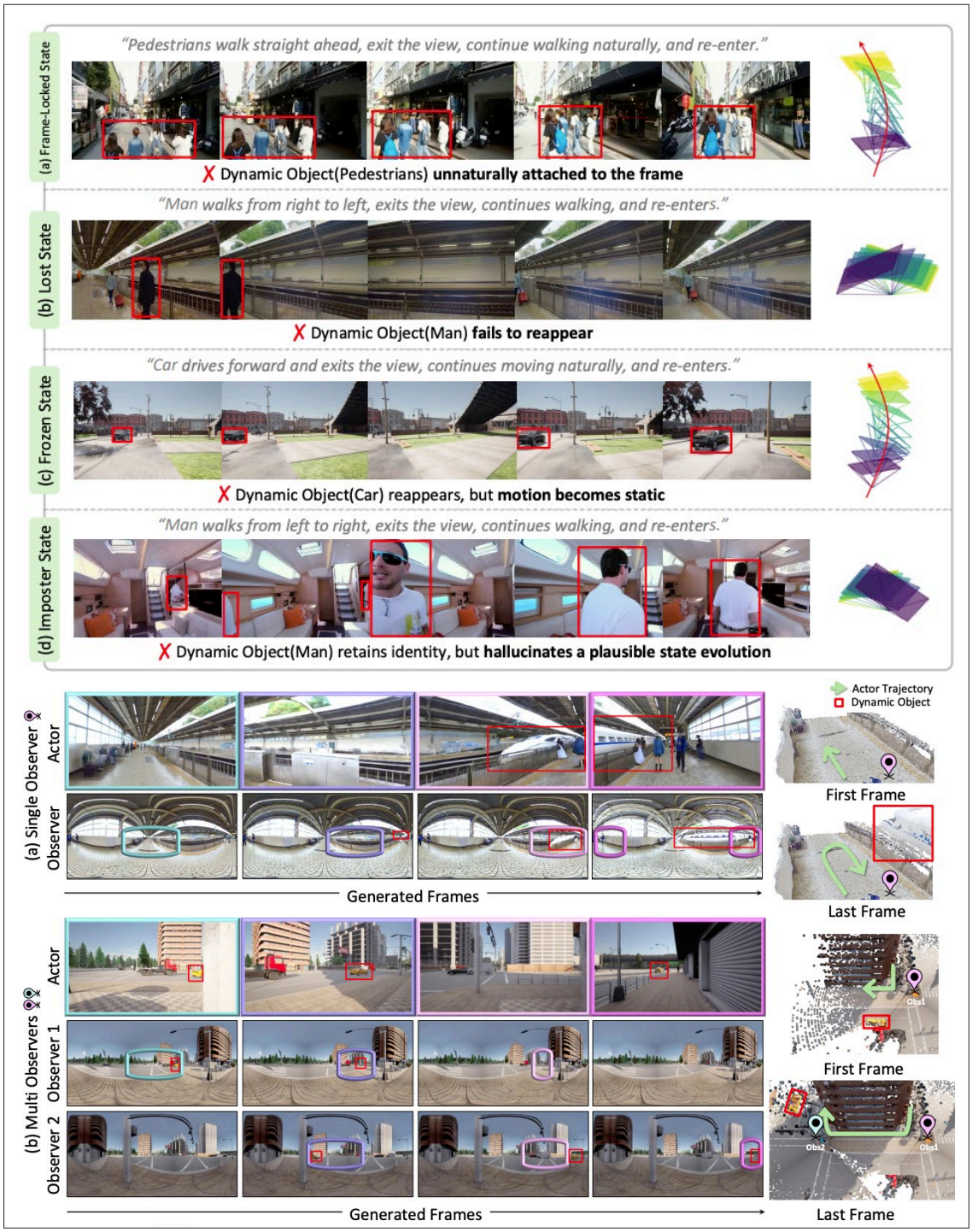
Ⅶ. 동적 객체와 다중 에이전트로의 확장

1. 동적 객체의 지속성

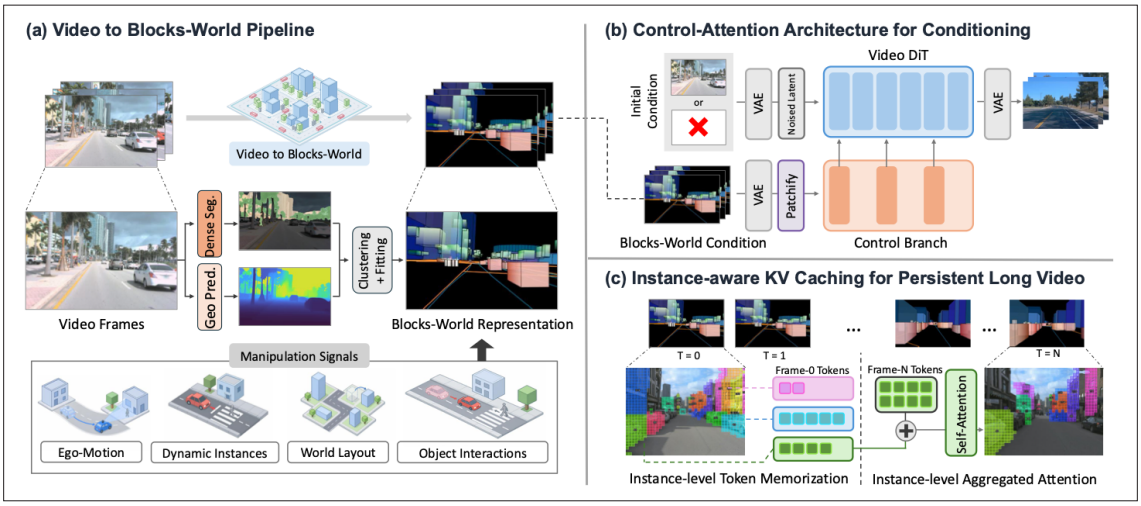
앞선 메모리 기반 접근들은 대부분 정적인 장면 구조에 초점을 맞춘다. 그러나 실제 세계에는 스스로 움직이는 동적 객체가 존재하며, 이들의 상태 역시 지속되어야 한다. 최근 연구[20]에 따르면 '고양이가 침대 위로 뛰어오른다'는 상황에서 카메라가 고양이로부터 멀어졌다 돌아오면, 현재의 월드 모델들은 고양이가 카메라를 따라 끌려오거나, 멈춰 있거나, 사라지거나, 위치가 갑자기 바뀌거나, 분열하는 등의 실패를 보인다(〈그림 7〉). 이는 관찰되지 않는 동안에도 세계의 상태를 유지·갱신하는 지속적 상태 핵심(persistent state core)이 결여되어 있음을 의미한다. 이에 대해 사용자 시점(actor)과 함께 세계 전체를 관찰하는 관찰자(observer) 파노라마 시점을 공동 생성하는 방법[21]이 제안되었다. 행위자의 시야를 벗어난 객체도 관찰자 시점에서는 계속 생성되므로, 다시 그 방향을 바라볼 때 객체의 위치와 상태가 일관되게 유지된다.

2. 다중 에이전트와 추상 상태 공간

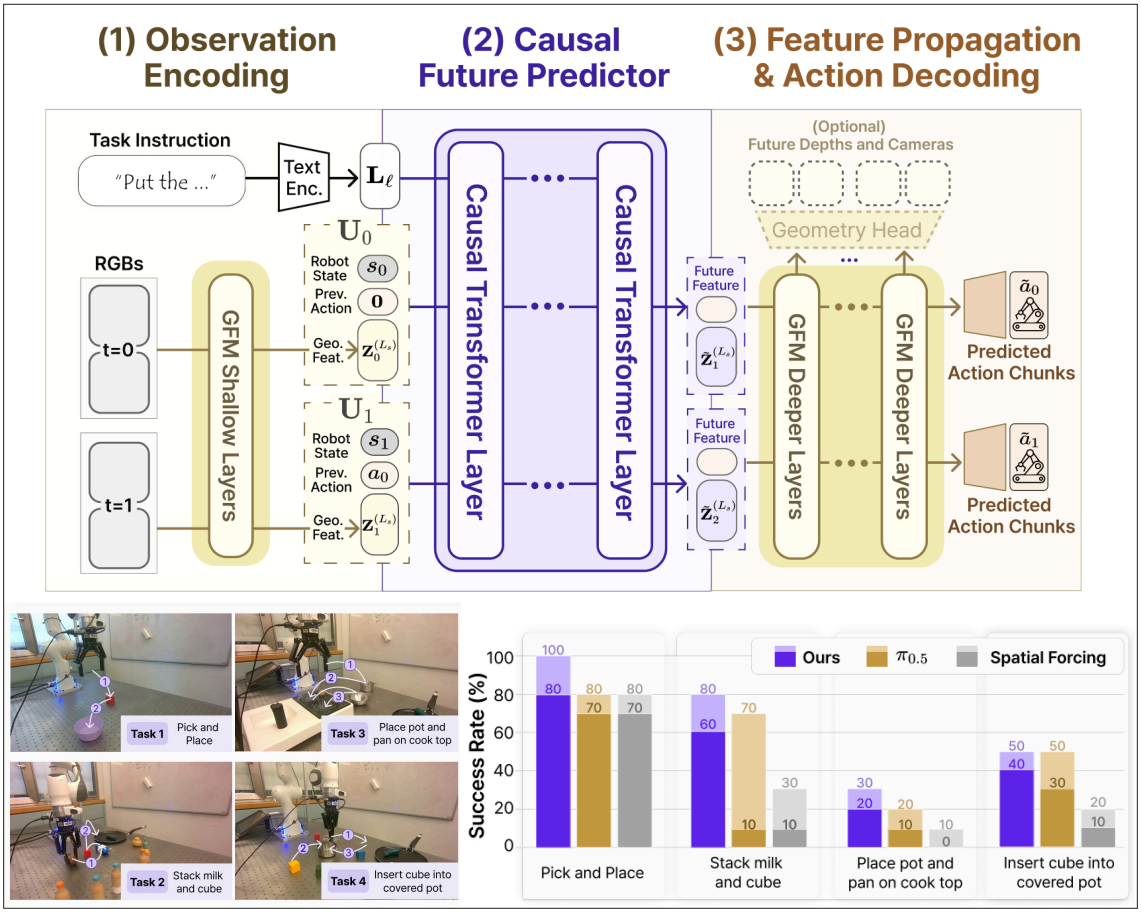
다중 에이전트 월드 모델[22]은 여러 사용자의 머리 포즈와 신체 동작을 입력으로 받아 공유된 세계 안에서 각 사용자의 시점 영상을 동시에 생성한다. 사용자별 독립 생성 시 발생하는 장면·행동·상태 갱신 불일치를 해결하기 위해, 공유 세계를 명시적 기하 메모리로 유지하고 이를 각 사용자 시점으로 재투영하여 조건으로 제공한다. 자율주행과 같이 다수 객체가 상호작용하는 환경에서는 객체 단위의 제어가 필요하다. NVIDIA의 OmniDreams[23]는 지도와 주변 에이전트의 미래 궤적 등 시뮬레이터의 추상화된 장면 상태와 주행 행동을 조건으로 실시간 자기회귀 생성을 수행하므로, 기존 교통 시뮬레이터와 쉽게 결합할 수 있다. 이를 확장하여 기성 모델만으로 영상에서 3차원 블록 세계(Blocks-World) 표현을 추출해 생성 조건으로 사용하고, 인스턴스 단위 KV 캐싱으로 장시간 객체 지속성을 유지하는 연구도 진행되고 있다(〈그림 8〉). 그러나 이러한 방식에서 월드 모델은 주어진 궤적에 외관을 입히는 신경 렌더러로 축소되고, 충돌해야 할 차량이 서로 통과하는 등 물리적으로 일관되지 않은 동역학이 나타난다. 이에 대한 대안으로 RGB 영상과 조감도(BEV) 배치를 공동 생성하는 연구가 진행되고 있다. 가려짐이 없고 외관 정보가 없는 BEV 브랜치가 모델의 용량을 객체의 행동에 집중시켜, 다른 차량에 반응하는 반응적 행동이 생성된다.



<그림 7> 동적 객체의 지속성 실패 유형 및 이의 해결책인 World Observer 결과



<그림 8> Blocks-World 기반 비디오 생성 구조



<그림 9> 기하 행동 모델(GAM)의 구조

Ⅷ. 월드 액션 모델로의 확장

비디오 월드 모델의 궁극적인 응용 중 하나는 로봇 조작이다. 현재 주류인 시각-언어-행동(VLA) 모델은 사전학습된 시각-언어 모델을 행동 예측에 맞게 미세조정한 것으로 언어 접지에 강하지만, 시간적·3차원적 이해가 부족하다. 반면 월드 액션 모델(WAM)은 사전학습된 비디오 모델이 미래 영상과 행동을 함께 예측하도록 적응시킨 것으로, 처음 보는 과제와 환경에 대한 제로샷 일반화 가능성을 보였으나[24], 행동마다 미래 영상을 생성해야 하므로 추론 지연이 VLA 대비 수 배에 달한다. 또한 두 접근 모두 2차원 데이터로 학습되어 카메라 시점 변화와 같은 환경 변화에 취약하다.

이에 대해 기하 파운데이션 모델(geometric foundation model)을 행동 모델로 확장한 기하 행동 모델(GAM)[25]이 제안되었다(〈그림 9〉). GAM은 기하 파운데이션 모델의 얕은 레이어로 관측을 인코딩하고, 인과적 미래 예측기로 미래의 잠재 기하 토큰과 행동 토큰을 예측한 뒤, 나머지 레이어와 깊이 헤드로 미래 깊이를, 경량 행동 헤드로 행동 청크를 출력한다. 14억 개의 파라미터만으로 LIBERO-Plus 벤치마크에서 최고 수준의 일반화 성능을 보였으며, 특히 카메라 교란 조건에서 9.7%p 향상되었다. 이는 미래를 영상이 아닌 기하 표현으로 예측하는 것이 효율성과 강건성 모두에서 유리할 수 있음을 시사한다.

Ⅸ. 결론 및 향후 과제

본고에서는 지속성 있는 비디오 월드 모델링을 중심으로 최근 연구 동향을 살펴보았다. 비디오 월드 모델은 행동 제어성, 일관성, 효율성을 요구하며, 이 중 지속성은 장시간 상호작용을 위한 가장 근본적인 과제이다. 이를 위해 KV 캐시를 검색·압축하여 월드 메모리로 활용하는 암묵적 접근과 컴팩트한 3차원 표현을 메모리로 사용하는 명시적 접근이 발전하고 있으며, 실세계 데이터에 접지된 월드 모델, 동적 객체와 다중 에이전트의 상태를 유지하는 연구, 그리고 월드 액션 모델로의 확장이 이어지고 있다.

향후에는 정적 장면과 동적 객체를 하나의 4차원 표현으로 통합하여 효율적으로 갱신하는 월드 메모리, 시야 밖에서도 물리 법칙에 따라 상태가 변화하는 동역학적 지속성, 여러 에이전트가 하나의 세계를 공유하는 다중 에이전트 월드 모델, 그리고 지속성을 정량적으로 평가하기 위한 표준 벤치마크의 구축이 중요한 연구 방향이 될 것이다. 비디오 월드 모델은 물리 세계를 이해하고 상호작용하는 범용 AI로 나아가는 핵심 기술로서, 국내에서도 데이터, 모델, 응용을 아우르는 활발한 연구가 이루어지기를 기대한다.

참 고 문 헌

[1] D. Ha and J. Schmidhuber, “World Models,” arXiv preprint arXiv:1803.10122, 2018.
[2] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” Advances in Neural Information Processing Systems (NeurIPS), 2020.
[3] W. Peebles and S. Xie, “Scalable Diffusion Models with Transformers,” IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
[4] OpenAI, “Video Generation Models as World Simulators,” Technical Report, 2024.
[5] Google DeepMind, “Genie 3,” https://deepmind.google/models/genie/, 2025.
[6] NVIDIA, “Cosmos World Foundation Model Platform for Physical AI,” arXiv preprint arXiv:2501.03575, 2025.
[7] X. Ren et al., “GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

참 고 문 헌

[8] X. He et al., “Matrix-Game 2.0: An Open-Source, Real-Time, and Streaming Interactive World Model,” arXiv preprint arXiv:2508.13009, 2025.
[9] J. Nam et al., “WorldCam: Interactive Autoregressive 3D Gaming Worlds with Camera Pose as a Unifying Geometric Representation,” Workshop on Video World Model: Interaction, Memory, and Efficiency @ CVPR (CVPRW), 2026.
[10] B. Chen et al., “Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion,” Advances in Neural Information Processing Systems (NeurIPS), 2024.
[11] X. Huang et al., “Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion,” Advances in Neural Information Processing Systems (NeurIPS), 2025.
[12] K. Liu et al., “Rolling Forcing: Autoregressive Long Video Diffusion in Real Time,” International Conference on Learning Representations (ICLR), 2026.
[13] J. Yi et al., “Deep Forcing: Training-Free Long Video Generation with Deep Sink and Participative Compression,” International Conference on Machine Learning (ICML), 2026.
[14] Yuan et al., “Denoise to Track: Harnessing Video Diffusion Priors for Robust Correspondence,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026.
[15] J. Yi et al., “WorldKV: Efficient World Memory with World Retrieval and Compression,” arXiv preprint arXiv:2605.22718, 2026.
[16] J. Lee et al., “3D Scene Prompting for Scene-Consistent Camera-Controllable Video Generation,” International Conference on Learning Representations (ICLR), 2026.
[17] J. Zhao et al., “Spatia: Video Generation with Updatable Spatial Memory,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026, pp. 4245-4257.
[18] H. An et al., “C3G: Learning Compact 3D Representations with 2K Gaussians,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026.
[19] J. Seo et al., “Seoul World Model: Grounding World Simulation Models in a Real-World Metropolis,” European Conference on Computer Vision (ECCV), 2026.
[20] J. Lu et al., “Current World Models Lack a Persistent State Core,” arXiv preprint arXiv:2606.20545, 2026.
[21] “World Observer: Joint Actor-Observer Generation,” 2026. (Under Review)
[22] “Multi-Agent World Models,” 2026. (Under Review)
[23] NVIDIA, “NVIDIA OmniDreams: Real-Time Generative World Model for Closed-Loop Autonomous Vehicle Simulation,” arXiv preprint arXiv:2606.03159, 2026.
[24] S. Ye et al., “World Action Models are Zero-shot Policies,” arXiv preprint arXiv:2602.15922, 2026.
[25] J. Han et al., “GAM: Geometric Action Model for Robot Policy Learning,” Workshop on 3D in the Era of World Models @ ECCV (ECCVW), 2026.

저 자 소 개



김승룡

- 2012년 : 연세대학교 전기전자공학과 학사
- 2018년 : 연세대학교 전기전자공학과 박사
- 2018년 ~ 2019년 : 연세대학교 전기전자공학과 박사후 연구원
- 2019년 ~ 2020년 : 스위스 EPFL 박사후 연구원
- 2020년 ~ 2024년 : 고려대학교 조교수
- 2024년 ~ 현재 : 한국과학기술원 김재철AI대학원 부교수
- 주관심분야 : 컴퓨터 비전, 월드 모델, 피지컬 인공지능