상호작용 가능한 세계를 생성하다: 월드 모델 기반 미디어 생성 기술의 동향

ㅁ **정재석, 어영정** / 연세대학교

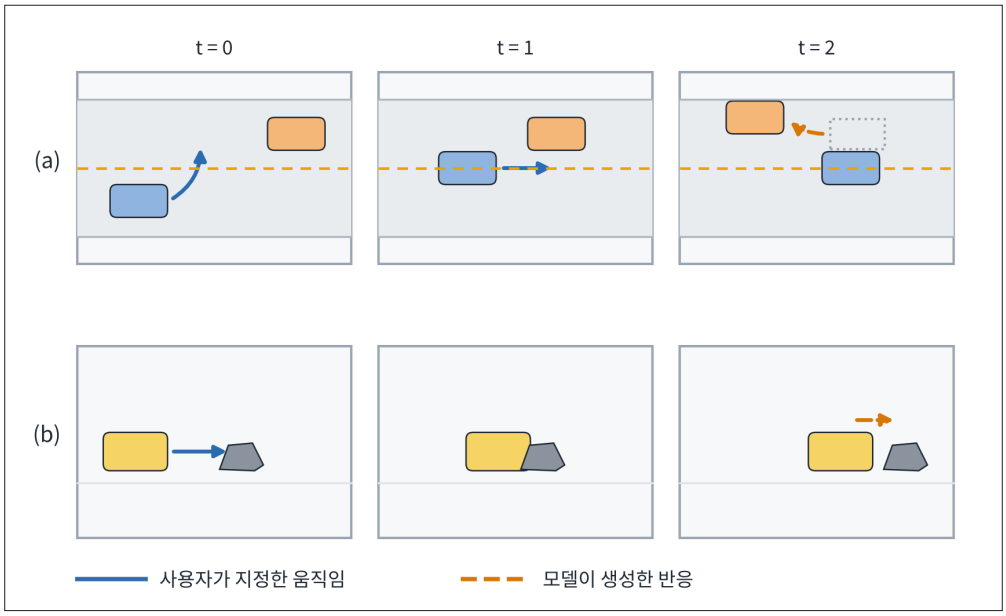
요 약

비디오 생성 모델의 품질이 높아지면서, 생성 모델을 한 편의 영상을 만드는 도구가 아니라 사용자가 들어가 시점을 바꾸고 개입할 수 있는 세계의 시뮬레이터로 쓰려는 연구가 빠르게 늘고 있다. 본 기고문은 이러한 흐름을 월드 모델의 관점에서 정리한다. 먼저 미디어 생성에 쓰이는 월드 모델이 갖춰야 할 조건을 시공간 일관성, 자유 시점 관찰, 물리적 개입의 세 가지로 나누고, 이 기준으로 행동 조건 월드 모델, 카메라 제어, 공간 기억, 3차원 기하 정보의 결합, 객체 제어 연구를 살펴본다. 이어서 공간 기억과 사용자 개입을 하나의 3차원 월드 상태로 묶으려는 본 연구실의 연구 방향을 소개한다. 끝으로 버추얼 프로덕션, 인터랙티브 콘텐츠, 시뮬레이션 데이터 생성 등 방송과 미디어 분야의 응용 가능성과 남은 과제를 논의한다.

I. 서 론

최근 2~3년 사이 비디오 생성 모델은 짧은 클립을 만드는 수준을 넘어, 한 장의 영상과 짧은 설명만으로 수 초에서 수십 초 길이의 사실적인 영상을 만들어 내는 단계에 이르렀다[1,2]. 품질이 일정 수준을 넘자 연구의 관심도 바뀌었다. 생성된 영상을 감상하는 대상으로 보는 대신, 그 안에서 카메라를 움직이고 물체를 옮기며 결과를 관찰할 수 있는 하나의 세계로 보려는 시도가 늘어난 것이다. 이렇게 행동이나 조건이 주어졌을 때 세계가 어떻게 바뀌는지를 예측하고 생성하는 모델을 월드 모델(world model)이라 부른다[3].

월드 모델은 원래 로봇과 자율주행 에이전트를 학습시키는 시뮬레이터로 주목받았지만[4,5,6], 미디어 분야에서도 의미가 크다. 제작자는 촬영 전에 카메라 동선과 피사체의 움직임을 바꿔 가며 장면을 미리 확인할 수 있고, 시



<그림 1> 사용자 개입과 생성된 반응의 개념도. (a) 파란 차를 중앙선을 따라 주황 차 쪽으로 움직이면 주황 차가 위쪽으로 살짝 비켜 지나간다. (b) 불도저를 밀면 바위가 밀려난다. 실선은 사용자가 지정한 움직임, 점선은 모델이 생성한 반응이다

청자는 정해진 화면 대신 원하는 시점에서 장면을 보거나 장면에 직접 개입할 수 있다. 이는 방송 콘텐츠가 일방향 영상에서 상호작용 가능한 경험으로 넓어질 수 있음을 뜻한다.

〈그림 1〉은 이러한 생성을 단순화해 나타낸 것이다. 사용자는 파란 차를 중앙선을 넘어 앞으로 움직이거나 불도저를 바위 쪽으로 미는 등 일부 객체의 움직임만 지정하고, 주황 차가 살짝 비켜 지나가거나 바위가 밀려나는 반응은 모델이 생성한다. 본 기고문은 II장에서 월드 모델의 개념과 미디어 생성 관점의 요구 조건을 정리하고, III장에서 이 조건에 따라 최근 연구 동향을 분류한다. IV장에서는 본 연구실의 연구 방향을 소개하고, V장에서 응용 가능성과 남은 과제를 논의한 뒤 VI장에서 마무리한다.

II. 월드 모델의 개념과 요구 조건

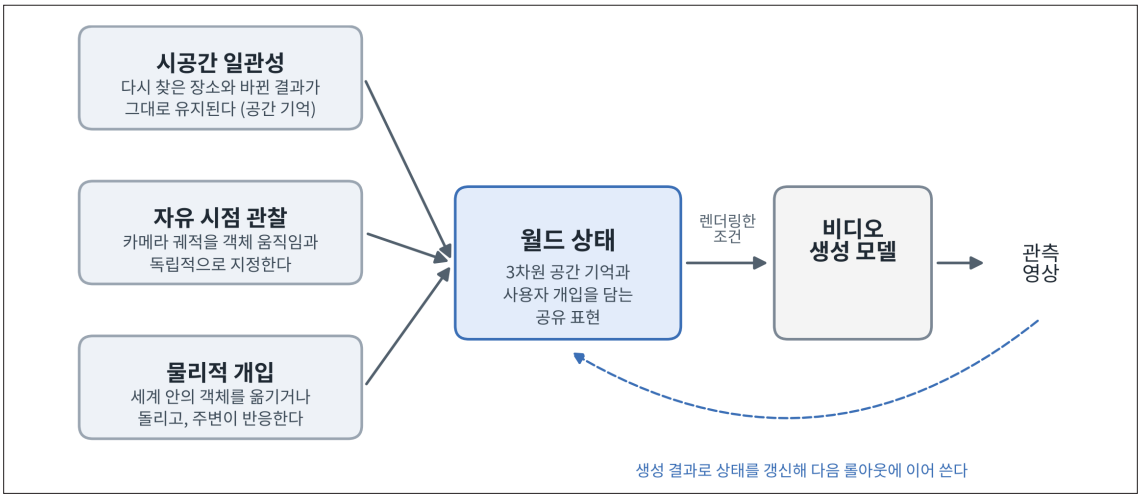
1. 월드 모델의 정의

월드 모델은 현재 세계의 상태와 행동이 주어졌을 때 다음 상태를 예측하는 모델이다. Ha와 Schmidhuber는 환경을 잠재 공간으로 압축하고 그 안에서 미래를 예측하는 모델을 학습해, 에이전트가 실제 환경 대신 모델이 만든 환경 안에서 정책을 학습할 수 있음을 보였다[3]. 이후 월드 모델은 크게 두 방향으로 발전했다. 하나는 미래를 픽셀이나 영상 잠재 공간에서 직접 생성하는 생성형 월드 모델이고, 다른 하나는 관측을 생성하지 않고 추상적인 표현 공간에서 미래를 예측하는 방향이다[7]. 미디어 생성과 직접 맞닿은 쪽은 앞의 생성형 월드 모델이므로, 본 기고문은 이를 중심으로 다룬다.

2. 미디어 생성 관점의 세 가지 요구 조건

생성형 월드 모델이 미디어 제작과 소비에 쓰이려면 사실적인 영상을 만드는 것만으로는 부족하다. 본 기고문은 다음 세 가지 조건을 다룬다.

첫째는 시공간 일관성이다. 카메라가 어떤 장소를 떠났



<그림 2> 미디어 생성을 위한 월드 모델의 세 가지 요구 조건과 이를 하나의 월드 상태로 묶는 구조

다가 다시 돌아왔을 때 그 장소는 이전과 같은 모습이어야 하고, 사용자의 개입으로 바뀐 결과도 유지되어야 한다. 이를 위해서는 모델이 지금까지 생성한 세계를 기억하는 공간 기억이 필요하다.

둘째는 자유 시점 관찰이다. 사용자는 세계를 원하는 시점에서 볼 수 있어야 한다. 특히 카메라의 움직임은 장면 속 객체의 움직임과 독립적으로 지정할 수 있어야 한다. 화면 안에서 물체가 왼쪽으로 이동한 것이 물체가 움직여서인지 카메라가 움직여서인지 모호하면, 제작자가 의도한 장면을 정확히 만들기 어렵다.

셋째는 물리적 개입이다. 시점을 바꾸며 세계를 둘러보는 것을 넘어, 세계 안의 객체를 옮기거나 회전시키고 그 결과로 주변이 어떻게 반응하는지 관찰할 수 있어야 한다. 이때 사용자가 모든 움직임을 일일이 지정할 필요는 없다. 객체 하나의 경로를 정해 주면 부딪힌 물체가 밀려나거나 걷는 동물의 다리가 움직이는 것과 같은 나머지 반응은 모델이 그럴듯하게 채워 주는 것이 바람직하다.

〈그림 2〉는 이 세 조건의 관계를 정리한 것이다. 세 조건을 각각 다른 입력으로 처리하면 서로 충돌하기 쉽다. 반면 공간 기억과 사용자 개입을 하나의 월드 상태에 담고, 그 상태를 원하는 카메라 궤적으로 렌더링해 비디오 생성 모델의 조건으로 쓰면 세 조건을 일관되게 다룰 수 있다. Ⅳ장에서 소개할 본 연구실의 연구가 이 구조를 따른다.

Ⅲ. 연구 동향

1. 행동 조건 비디오 월드 모델

게임과 같이 행동이 명확히 정의된 환경에서는 키보드와 마우스 입력을 조건으로 다음 프레임을 생성하는 연구가 먼저 성과를 냈다. DIAMOND는 확산 모델로 Atari 게임 환경을 모사해 그 안에서 강화학습 에이전트를 학습시켰고[8], GameNGen은 확산 모델만으로 게임 DOOM을 실시간에 가깝게 구동할 수 있음을 보였다[9]. Genie는 인터넷 영상에서 행동 레이블 없이 잠재 행동을 학습해, 한 장의 영상으로부터 조작 가능한 2D 환경을 만들어 냈다[10]. Genie 3는 이를 사실적인 3차원 환경으로 확장해 실시간 상호작용을 지원하고, 텍스트로 날씨 변화나 새로운 물체의 등장 같은 사건을 일으킬 수 있게 했다[11].

공개 모델 진영에서도 상호작용형 월드 모델이 빠르게

늘고 있다. Matrix-Game 2.0은 실시간 스트리밍 생성을 오픈소스로 공개했고[12], Hunyuan-GameCraft는 과거 프레임 기록을 조건으로 함께 사용해 움직임이 큰 게임 영상을 키보드와 마우스 입력에 맞춰 생성한다[13]. Yume-1.5는 행동 입력에 더해 텍스트로 세계에 일어날 일을 지시할 수 있게 했다[14]. 한편 이미 학습된 비디오 생성 모델을 행동 조건 월드 모델로 바꾸는 연구도 이어지고 있다. AVID는 사전학습된 비디오 확산 모델에 어댑터를 붙여 행동에 따라 영상을 생성하게 했고[15], Vid2World는 비디오 확산 모델을 자기회귀적이고 행동 조건을 받는 형태로 바꾸는 방법을 제시했다[16].

대규모 기반 모델도 월드 모델을 표방하고 있다. Wan[1]과 같은 공개 비디오 생성 모델이 다양한 월드 모델 연구의 출발점으로 쓰이고 있으며, NVIDIA의 Cosmos 계열은 로봇과 자율주행 같은 물리 AI를 위한 월드 모델을 목표로 한다[2]. 이들 모델은 사실적인 영상 품질을 제공하지만, 그 자체로는 카메라와 객체를 세밀하게 제어하거나 긴 시간 동안 장면을 기억하는 기능을 충분히 갖추고 있지 않다.

2. 카메라 제어와 장면 탐험

자유 시점 관찰을 위해서는 생성 영상의 카메라를 정확히 제어해야 한다. CameraCtrl은 카메라 궤적을 광선(ray) 기반 표현으로 바꾸어 사전학습된 비디오 생성 모델에 조건으로 주입했고[17], CameraCtrl II는 이를 동적인 장면의 탐험으로 확장했다[18]. GEN3C는 추정한 깊이로 3차원 점군 캐시를 만들고, 이를 목표 카메라로 렌더링한 영상을 조건으로 사용해 카메라 제어의 정확도와 장면의 일관성을 함께 높였다[19].

탐험의 대상을 실제 세계로 넓힌 연구도 등장했다. Seoul World Model은 서울 시내 120만 장의 거리뷰 파노라마를 데이터베이스로 두고, 주어진 위치와 카메라 궤적에 맞는 거리뷰를 검색해 생성의 근거로 삼는다[20]. 그 결과 실제 도시의 거리 구조를 따르는 수백 미터 길이의 주행 영상을 생성할 수 있다. 실제 장소에 기반한 생성은 지역 정보 서비스나 도시 단위의 가상 콘텐츠로 이어질 수 있다는 점에서 미디어 분야에서도 주목할 만하다.

3. 공간 기억

비디오 생성 모델은 한 번에 처리할 수 있는 프레임 수가 제한되어 있어, 카메라가 오래 움직이면 앞에서 본 장면을 잊어버린다. 이를 해결하는 방법은 크게 세 가지다. 첫째는 과거에 생성한 프레임 가운데 현재 시점과 관련 있는 것을 검색해 문맥으로 넣는 방식이다[21,22]. 둘째는 과거 문맥을 압축해 더 긴 기록을 담는 방식이다[23,24]. 셋째는 장면을 점군이나 서펠(surfel) 같은 3차원 표현으로 저장하고, 필요한 시점에서 렌더링해 쓰는 방식이다[19,25]. 최근에는 공간 기억을 생성 도중에 갱신하거나[26], 여러 형태의 기억을 함께 쓰는 연구[27], 실시간 상호작용에서 긴 시간의 기하 일관성을 유지하려는 연구[28]도 나왔다.

기억해야 할 대상은 정적인 배경만이 아니다. 카메라가 다른 곳을 보는 동안에도 움직이던 사람이나 차는 계속 움직여야 하므로, 시야 밖에서 일어나는 변화를 이어서 모사하려는 연구가 등장했다[29,30]. 이는 영상으로 관찰되지 않는 시간에도 세계의 상태를 유지해야 한다는 점에서, 월드 모델이 화면이 아니라 세계 자체를 표현해야 한다는 요구를 잘 보여준다.

4. 3차원 기하 정보의 결합

공간 기억의 세 번째 방식은 3차원 장면 표현 연구와 자연스럽게 연결된다. 3D Gaussian Splatting[31]처럼 명시적인 3차원 표현은 임의의 시점에서 빠르게 렌더링할 수 있고 기하 연산을 직접 적용할 수 있다. 이러한 장점을 월드 모델에 들여오려는 연구가 늘고 있다. Aether는 영상과 함께 깊이와 카메라 자세 같은 기하 정보를 하나의 모델에서 예측해 재구성, 예측, 계획을 함께 다뤘고[32],

<표 1> 월드 모델 연구의 공간 기억, 카메라 제어, 객체 개입 표현 비교

연구	공간 기억	카메라 제어	객체 개입
Genie 3 [11]	생성 프레임 문맥	이동 행동 입력	텍스트로 사건 지정
GEN3C [19]	3차원 점군 캐시	카메라 궤적	없음
Seoul World Model [20]	실제 거리뷰 검색	카메라 궤적	없음
MotionCanvas [42]	영상 잠재 특징	카메라 궤적	2차원 상자, 점 궤적
VerseCrafter [43]	RGB 점군	카메라 궤적	3차원 Gaussian
WorldDirector [44]	비디오 문맥 토큰	카메라 궤적	2차원 상자 궤적
SymphoMotion [45]	점군 렌더링	카메라 궤적	3차원 점 궤적
Real2SAM2Real [46]	비디오 잠재 특징	카메라 궤적	3차원 메시

DeepVerse는 4차원 기하 정보를 자기회귀적으로 함께 생성하는 월드 모델을 제안했다[33]. 게임 환경에서는 픽셀 기록 대신 지속적인 3차원 상태를 유지하는 월드 모델도 연구되었다[34].

다만 지금까지의 3차원 공간 기억은 주로 정적인 배경을 기억하는 데 쓰였고, 그 안의 객체를 사용자가 옮기는 개입까지 같은 표현으로 다루는 경우는 드물었다. 3차원 상태가 시점 이동에는 잘 쓰이지만, 세계를 바꾸는 개입의 수단으로는 아직 충분히 쓰이지 않고 있는 셈이다.

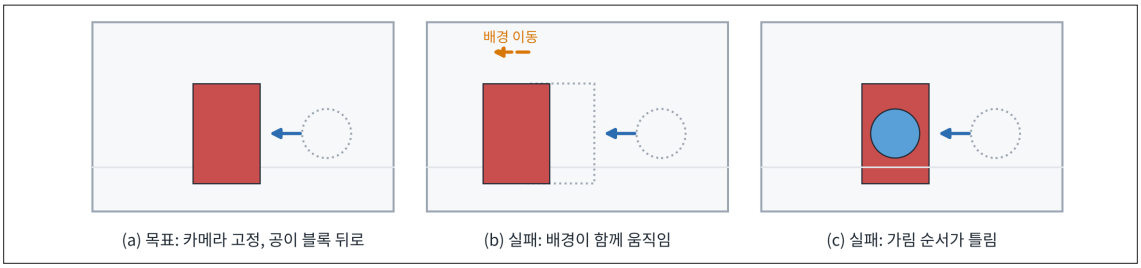
5. 객체 제어

객체의 움직임을 지정하는 연구는 2차원 조건에서 3차원 제어로 발전하고 있다. Boximator는 상자로 객체의 위치와 움직임을 지정했고[35], MotionCtrl은 카메라 움직임과 객체 움직임을 분리된 조건으로 함께 다뤘다[36]. Tora는 궤적 정보를 확산 트랜스포머에 주입했고[37], Motion Prompting은 점 궤적을 프롬프트처럼 사용해 영상 속 움직임을 제어했으며[38], Wan-Move는 잠재 특징을 궤적을 따라 옮기는 방식으로 움직임을 제어했다[39]. LeviTor는 2차원 궤적에 깊이 정보를 더해 3차원 궤적으로 영상을 제어했고[40], Diffusion as Shader는 3차원 추적 영상을 조건으로 사용해 카메라와 객체 제어를 하나의 틀에서 다뤘다[41]. MotionCanvas는 영화의 샷 설계처럼 카메라 움직임과 객체 움직임을 함께 지정할 수 있게 했다[42].

최근에는 월드 모델 관점에서 객체와 카메라를 함께 제어하는 연구가 이어지고 있다. VerseCrafter는 3차원 Gaussian으로 객체의 움직임을 지정하고[43], WorldDirector는 동적 객체를 지속적으로 기억하는 메모리를 두었으며[44], SymphoMotion은 3차원 점 궤적으로 카메라와 객체의 움직임을 함께 다룬다[45]. Real2SAM2Real은 3차원 메시 캐시를 비디오 생성의 보조 문맥으로 사용한다[46].

6. 정리

〈표 1〉은 대표적인 연구가 공간 기억, 카메라 제어, 객체 개입을 각각 어떻게 표현하는지 정리한 것이다. 대부분의 연구는 공간 기억과 사용자 개입을 서로 다른 표현으로 다룬다. 예를 들어 배경은 점군으로 기억하면서 객체의 움직임은 2차원 상자나 궤적으로 지정하는 식이다. 이렇게 두 표현이 분리되어 있으면 배경이 조금씩 틀어지거나, 화면 속 움직임이 객체의 이동인지 시점의 변화인지 모호해지는 문제가 생긴다.



<그림 3> 객체와 카메라를 함께 제어할 때 생길 수 있는 실패 유형. (a) 목표: 카메라를 고정하고 공을 블록 뒤로 보낸다. (b) 객체의 이동이 시점 변화로 흡수되어 배경이 함께 움직인다. (c) 앞뒤 관계를 반영하지 못해 공이 블록 앞에 그려진다

Ⅳ. 본 연구실의 연구 방향: 공간 기억과 사용자 개입의 통합

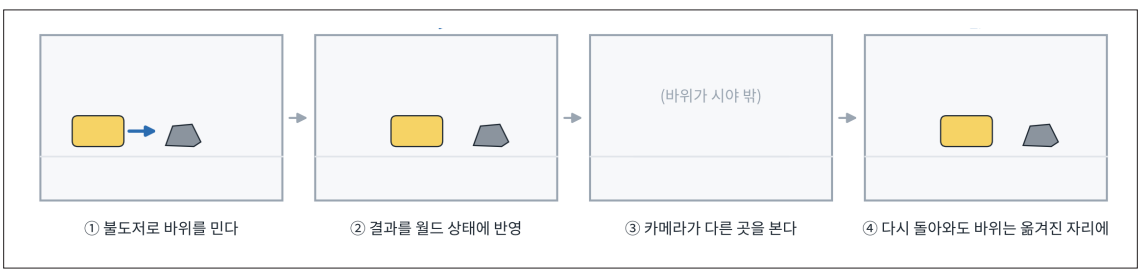
1. 문제 정의

본 연구실은 앞에서 정리한 세 가지 조건, 즉 지속적인 세계 안에서 자유로운 시점으로 관찰하면서 물리적으로 개입하는 것을 함께 만족하는 월드 모델을 연구하고 있다. 핵심 문제는 객체의 움직임과 카메라의 움직임이 화면에서는 모두 픽셀의 이동으로 나타난다는 점이다. 같은 화면 궤적이라도 객체가 움직인 결과일 수도, 카메라가 움직인 결과일 수도 있다. 객체의 움직임을 화면 좌표의 궤적으로 지정하면 이 모호성을 피하기 어렵다.

〈그림 3〉은 이 모호성 때문에 생기는 실패를 단순화해 나타낸 것이다. 카메라를 고정한 채 공을 블록 뒤로 보내려 할 때, 객체와 카메라의 움직임이 분리되지 않으면 공의 이동이 시점 변화로 흡수되어 배경이 함께 움직인다. 반대로 카메라는 고정되더라도 장면의 3차원 앞뒤 관계를 모르면 공이 블록 앞에 그려진다.

2. 통합 월드 상태

본 연구실은 공간 기억과 사용자 개입을 하나의 3차원 월드 상태에 담는 방향으로 이 문제에 접근하고 있다. 초기 영상으로부터 장면을 3차원 상태로 구성하고, 사용자의 개입은 이 상태 안에서 선택한 객체를 옮기거나 회전시키는 변환으로 표현한다. 카메라 궤적은 이와 별개로 지정하며, 바뀐 상태를 그 궤적으로 렌더링한 결과를 비디오 생성 모델의 조건으로 사용한다(〈그림 2〉). 객체의 움직임은 세계 좌표에서, 카메라는 그와 독립적으로 지정되므로 화면상 움직임의 모호성이 표현 단계에서 줄어들고, 3차원 상태가 앞뒤 관계를 담고 있어 가림도 자연스럽게 반



<그림 4> 개입 결과의 유지. 개입으로 바뀐 장면을 월드 상태에 반영하면, 카메라가 다른 곳을 보았다가 돌아와도 결과가 그대로 남는다

영된다. 현재 이러한 방식으로 객체와 카메라를 함께 제어하는 정확도가 개선되는 것을 확인하고 있다.

이 상태가 영상의 모든 픽셀을 결정하지는 않는다. 렌더링된 조건은 장면의 일부만 알려 주고, 나머지 장면 내용과 사용자가 지정하지 않은 반응은 비디오 생성 모델이 채운다. 〈그림 1〉에서처럼 사용자는 일부 객체의 움직임만 정하고, 주변의 회피나 밀림 같은 반응은 생성 모델의 사전 지식에 맡기는 구조다.

월드 모델이 세계를 기억한다면, 사용자가 바꾼 결과도 기억해야 한다. 상호작용이 일어난 뒤 생성된 결과로부터 바뀐 영역을 다시 얻어 월드 상태를 갱신하면, 다음 생성은 갱신된 상태에서 시작한다. 〈그림 4〉처럼 불도저가 민 바위는 카메라가 다른 곳을 보았다가 돌아와도 옮겨진 자리에 남는다. 사용자는 객체를 조작하고, 다른 시점에서 결과를 확인한 뒤, 다시 개입하는 과정을 이어갈 수 있다.

3. 남은 문제

이 접근에도 풀어야 할 문제가 남아 있다. 주변 물체와의 상호작용은 비디오 생성 모델의 사전 지식에 의존하므로, 카메라 시야 밖에서 일어나는 상호작용은 명시적으로 모사하기 어렵다. 작은 물체를 정밀하게 다루거나 빠른 움직임을 표현하는 데에도 한계가 있으며, 물리적으로 불가능한 개입을 어떻게 다룰지도 정해야 한다.

V. 미디어 응용과 남은 과제

1. 응용 분야

가장 가까운 응용은 버추얼 프로덕션과 사전 시각화다. 현재 가상 스튜디오는 미리 제작한 3차원 그래픽 자산이 있어야 하지만, 월드 모델을 쓰면 한 장의 콘셉트 이미지나 현장 사진에서 출발해 카메라 동선과 소품의 배치를 바꿔 가며 장면을 미리 만들어 볼 수 있다. 촬영 전에 여러 안을 빠르게 비교할 수 있다는 점에서 제작 비용과 시간을 줄일 수 있다.

두 번째는 시청자가 시점을 고르거나 장면에 개입하는 인터랙티브 콘텐츠다. 스포츠 중계나 공연 영상에서 시청자가 원하는 각도로 장면을 다시 보거나, 교육용 콘텐츠에서 학습자가 물체를 직접 움직여 결과를 확인하는 형태를 생각할 수 있다. 게임처럼 상호작용 자체가 핵심인 콘텐츠도 자연스러운 응용 분야다[11,12].

세 번째는 학습용 데이터와 시뮬레이션 환경 생성이다. 자율주행과 로봇 분야에서는 실제로 수집하기 어려운 상황을 월드 모델로 만들어 에이전트를 학습시키고 평가하려는 수요가 크다[2,20]. 특히 객체에 직접 개입하고 그 결과를 여러 시점에서 관찰하는 능력은 시점 탐색만 가능한 시뮬레이터로는 학습하기 어려운 조작 과제에 필요하다.

2. 남은 과제

먼저 긴 시간 동안의 일관성이다. 현재 월드 모델은 수십 초에서 수 분 정도의 상호작용에서는 일관성을 유지하지만, 방송 프로그램 한 편 길이에서 세계를 일관되게 유지하는 것은 아직 어렵다. 다음은 물리적 타당성이다. 생성 모델은 그럴듯한 반응을 만들지만 실제 물리 법칙을 보장하지 않으며, 시야 밖의 사건처럼 영상으로 관찰되지 않는 변화를 다루는 방법도 더 필요하다.

실시간성과 계산 비용도 중요한 과제다. 시청자와 실시간으로 상호작용하려면 지연 시간이 짧아야 하는데, 고품질 비디오 생성 모델은 여전히 계산량이 크다. 적은 단계로 생성하는 가속 기법과 경량화가 함께 발전해야 한다. 평가 방법도 정립되어야 한다. 영상 품질 지표만으로는 지정한 개입이 세계 안에서 제대로 일어났는지 판단하기 어려우며, 지정한 객체의 위치와 카메라 궤적을 월드 좌표계 기준으로 함께 확인하는 지표가 더 필요하다. 끝으로 생성 콘텐츠의 진위성 문제가 있다. 실제 장소에 기반한 사실적인 영상을 누구나 만들 수 있게 되는 만큼, 생성 사실을 표시하고 출처를 추적하는 기술과 제도가 함께 마련되어야 한다.

VI. 결 론

본 기고문은 비디오 생성 모델이 상호작용 가능한 세계의 시뮬레이터로 확장되는 흐름을 시공간 일관성, 자유 시점 관찰, 물리적 개입의 세 조건으로 정리하고, 세 조건을 하나의 월드 상태 표현으로 묶으려는 본 연구실의 연구 방향을 소개했다. 공간 기억과 사용자 개입을 같은 좌표계에 두면 객체와 카메라를 독립적으로 제어할 수 있고, 사용자가 바꾼 세계를 다음 생성에서도 이어 쓸 수 있다. 월드 모델은 아직 긴 시간의 일관성, 물리적 타당성, 실시간성 등에서 갈 길이 남아 있지만, 제작자와 시청자가 영상 속 세계에 직접 들어가 시점을 고르고 개입하는 새로운 미디어 경험의 기반 기술이 될 것으로 기대한다.

참 고 문 헌

[1] Team Wan et al., “Wan: Open and advanced large-scale video generative models,” arXiv preprint arXiv:2503.20314, 2025.
[2] NVIDIA, “Cosmos 3: Omnimodal world models for physical AI,” arXiv preprint arXiv:2606.02800, 2026.
[3] D. Ha and J. Schmidhuber, “Recurrent world models facilitate policy evolution,” Advances in Neural Information Processing Systems (NeurIPS), 2018.
[4] Y. Du et al., “Learning universal policies via text-guided video generation,” Advances in Neural Information Processing Systems (NeurIPS), 2023.
[5] S. Yang et al., “Learning interactive real-world simulators,” International Conference on Learning Representations (ICLR), 2024.
[6] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
[7] Y. LeCun, “A path towards autonomous machine intelligence,” OpenReview preprint, 2022.
[8] E. Alonso et al., “Diffusion for world modeling: Visual details matter in Atari,” Advances in Neural Information Processing Systems (NeurIPS), 2024.
[9] D. Valevski, Y. Leviathan, M. Arar, and S. Fruchter, “Diffusion models are real-time game engines,” International Conference on Learning Representations (ICLR), 2025.
[10] J. Bruce et al., “Genie: Generative interactive environments,” International Conference on Machine Learning (ICML), 2024.
[11] J. Parker-Holder and S. Fruchter, “Genie 3: A new frontier for world models,” Google DeepMind Blog, Aug. 2025.
[12] X. He et al., “Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model,” arXiv preprint arXiv:2508.13009, 2025.
[13] J. Li et al., “Hunyuan-GameCraft: High-dynamic interactive game video generation with hybrid history condition,” arXiv preprint arXiv:2506.17201, 2025.
[14] X. Mao et al., “Yume-1.5: A text-controlled interactive world generation model,” arXiv preprint arXiv:2512.22096, 2025.
[15] M. Rigter, T. Gupta, A. Hilmkil, and C. Ma, “AVID: Adapting video diffusion models to world models,” Reinforcement Learning Conference (RLC), 2025.
[16] S. Huang, J. Wu, Q. Zhou, S. Miao, and M. Long, “Vid2World: Crafting video diffusion models to interactive world models,” International Conference on Learning Representations (ICLR), 2026.
[17] H. He et al., “CameraCtrl: Enabling camera control for video diffusion models,” International Conference on Learning Representations (ICLR), 2025.
[18] H. He et al., “CameraCtrl II: Dynamic scene exploration via camera-controlled video diffusion models,” IEEE/CVF International Conference on Computer Vision (ICCV), 2025.

참 고 문 헌

[19] 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.
[20] J. Seo et al., "Grounding world simulation models in a real-world metropolis," arXiv preprint arXiv:2603.15583, 2026.
[21] Z. Xiao et al., "WorldMem: Long-term consistent world simulation with memory," Advances in Neural Information Processing Systems (NeurIPS), 2025.
[22] J. Yu et al., "Context as memory: Scene-consistent interactive long video generation with memory retrieval," SIGGRAPH Asia Conference Papers, 2025.
[23] L. Zhang, S. Cai, M. Li, G. Wetzstein, and M. Agrawala, "Frame context packing and drift prevention in next-frame-prediction video diffusion models," Advances in Neural Information Processing Systems (NeurIPS), 2025.
[24] Y. Oshima, Y. Iwasawa, M. Suzuki, Y. Matsuo, and H. Furuta, "WorldPack: Dynamic frame compression for long-context video world modeling," Transactions on Machine Learning Research (TMLR), 2026.
[25] R. Li, P. Torr, A. Vedaldi, and T. Jakab, "VMem: Consistent interactive video scene generation with surfel-indexed view memory," IEEE/CVF International Conference on Computer Vision (ICCV), 2025.
[26] J. Zhao et al., "Spatia: Video generation with updatable spatial memory," arXiv preprint arXiv:2512.15716, 2025.
[27] W. Yu et al., "MosaicMem: Hybrid spatial memory for controllable video world models," arXiv preprint arXiv:2603.17117, 2026.
[28] W. Sun et al., "WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling," arXiv preprint arXiv:2512.14614, 2025.
[29] Z. Duan et al., "LiveWorld: Simulating out-of-sight dynamics in generative video world models," arXiv preprint arXiv:2603.07145, 2026.
[30] K. Chen et al., "Out of sight but not out of mind: Hybrid memory for dynamic video world models," arXiv preprint arXiv:2603.25716, 2026.
[31] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," ACM Transactions on Graphics, vol. 42, no. 4, 2023.
[32] H. Zhu et al., "Aether: Geometric-aware unified world modeling," IEEE/CVF International Conference on Computer Vision (ICCV), 2025.
[33] J. Chen et al., "DeepVerse: 4D autoregressive video generation as a world model," arXiv preprint arXiv:2506.01103, 2025.
[34] S. Garcin et al., "Beyond pixel histories: World models with persistent 3D state," International Conference on Machine Learning (ICML), 2026.
[35] J. Wang et al., "Boximator: Generating rich and controllable motions for video synthesis," International Conference on Machine Learning (ICML), 2024.
[36] Z. Wang et al., "MotionCtrl: A unified and flexible motion controller for video generation," ACM SIGGRAPH Conference Papers, 2024.
[37] Z. Zhang et al., "Tora: Trajectory-oriented diffusion transformer for video generation," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
[38] D. Geng et al., "Motion prompting: Controlling video generation with motion trajectories," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
[39] R. Chu et al., "Wan-Move: Motion-controllable video generation via latent trajectory guidance," Advances in Neural Information Processing Systems (NeurIPS), 2025.
[40] H. Wang et al., "LeviTor: 3D trajectory oriented image-to-video synthesis," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
[41] Z. Gu et al., "Diffusion as shader: 3D-aware video diffusion for versatile video generation control," SIGGRAPH Conference Papers, 2025.
[42] J. Xing et al., "MotionCanvas: Cinematic shot design with controllable image-to-video generation," SIGGRAPH Conference Papers, 2025.
[43] S. Zheng et al., "VerseCrafter: Dynamic realistic video world model with 4D geometric control," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026.
[44] H. Wang et al., "WorldDirector: Building controllable world simulators with persistent dynamic memory," arXiv preprint arXiv:2607.02517, 2026.
[45] G. Zhang et al., "SymphoMotion: Joint control of camera motion and object dynamics for coherent video generation," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026.
[46] J. Wu et al., "Real2SAM2Real: Generative 3D caches as complementary context for video diffusion," arXiv preprint arXiv:2606.00299, 2026.

저 자 소 개



정 재 석

- 2015년 3월 ~ 2022년 2월 : 연세대학교 전기전자공학과 학사
- 2022년 3월 ~ 현재 : 연세대학교 인공지능대학원 석박통합과정
- 2023년 9월 ~ 2024년 2월 : NAVER AI Lab 인턴
- 2026년 6월 ~ 현재 : Familiar Labs Inc.
- 주관심분야 : 비디오 생성, 월드 모델, 확산 모델



어 영 정

- 2006년 3월 ~ 2010년 2월 : 연세대학교 컴퓨터과학과 학사
- 2010년 3월 ~ 2018년 8월 : 연세대학교 컴퓨터과학과 박사
- 2018년 8월 ~ 2020년 8월 : NAVER AI Lab
- 2020년 9월 ~ 2024년 8월 : 연세대학교 글로벌인재학부 응용정보공학전공 조교수, (겸) 인공지능학과
- 2024년 9월 ~ 현재 : 연세대학교 글로벌인재학부 응용정보공학전공 부교수, (겸) 인공지능학과
- 주관심분야 : 비디오 생성 모델, 월드 모델, 확산 모델, 3D/4D Gaussian Splatting, 뉴럴 렌더링