

김 선 옥
한국항공대학교

‘월드 모델 기반의 차세대 미디어 생성 기술’ 특집호를 내며

이번 특집호에서는 인공지능 기반의 차세대 시뮬레이션 및 콘텐츠 제작의 패러다임을 바꿀 월드 모델 기반의 미디어 생성 기술들을 소개하고자 합니다.

최근 몇 년간 인공지능 기술의 급격한 발전은 비디오 생성 모델의 품질을 눈부시게 향상시켰습니다. 이제 생성 인공지능은 단순히 한 편의 시각적 영상을 일방향적으로 만들어내는 도구를 넘어섰습니다. 사용자의 카메라 조작, 행동 입력, 언어적 프롬프트에 실시간으로 반응하며 내부 환경을 시뮬레이션하고 일관되게 갱신하는 ‘탐험 가능한 세계(World)’의 구축으로 진화하고 있습니다. 월드 모델(World Model)은 에이전트의 행동에 따라 환경이 어떻게 변화할지 내부적으로 예측하고 동역학을 학습하는 기술입니다. 과거 로보틱스나 강화학습의 효율적 정책 학습을 위해 주로 논의되던 이 개념은 최신 대규모 생성형 AI 모델들과 결합하며 미디어 생성 분야의 핵심 응용 기술로 급부상했습니다. 버추얼 프로덕션, 실시간 인터랙티브 게임, 자율주행 Closed-loop 시뮬레이션 등에 이르기까지 물리 세계를 이해하고 상호작용하는 범용 인공지능(Physical AI)으로 나아가는 징검다리가 되고 있습니다. 본 특집호에서는 월드 모델 기술의 기초부터 시공간 지속성 메모리, 옴니모달 효율화, 3차원 기하 통합 및 로봇 학습 시뮬레이션 연계까지, 차세대 미디어 생성 기술을 다각도로 조명한 전문가들의 원고 5편을 준비하였습니다.

첫 번째 기고문(“지속성 있는 비디오 월드 모델 연구 동향”)에서는 비디오 월드 모델이 게임, 디지털 트윈 등 실세계 시뮬레이터로 기능하기 위해 필수적인 ‘지속성(Persistence)’ 확보 기법을 중점적으로 다룹니다. 카메라가 시야를 벗어났다 돌아와도 구조와 객체 상태를 일관되게 유지하기 위한 최근의 연구 흐름을 소개합니다. 구체적으로는 자기회귀 생성에서의 오차 누적을 해소하는 Self Forcing 기법과 함께, KV 캐시를 활용한 암묵적 월드 메모리(WorldKV), 컴팩트 3D 가우시안 표현을 활용한 명시적 월드 메모리(C3G) 기술을 분석합니다. 나아가 서울시 거리뷰 RAG 데이터베이스를 결합해 수 킬로미터 궤적을 오차 없이 생성해내는 실세계 접지 모델(Seoul World Model)의 최신 성과를 소개합니다.

두 번째 기고문(“효율적인 옴니모달 월드 모델을 위한 핵심 기술과 연구 동향”)에서는 영상, 음성, 언어, 행동 신호를 단일 시간축에서 처리하는 옴니모달(Omnimodal) 체계로의 전환과 이에 따른 계산 및 메모리 병목 해결을 위한 효율화 계층을 정리합니다. 고해상도 시공간 토큰 폭증 문제를 극복하기 위해 문맥 압축, 모델 압축, 캐시 및 기억, 적응적 연산, 시스템 서빙 등 5개 계층의 경량화 프레임워크를 제안합니다. 특히 저자 연구진이 개발한 순전파 기반의 영차 민감도 토큰 가지치기 기법인 ‘ZOO-Prune’의 메커니즘을 상세히 소개하며, 성능을 유지하면서도 종단 간 추론 속도를 획기적으로 가속하는 시각 유도형 옴니모달 설계 방향을 제시합니다.

세 번째 기고문(“상호작용 가능한 세계를 생성하다: 월드 모델 기반 미디어 생성 기술의 동향”)에서는 생성형 월드 모델이 미디어 제작과 소비에 실질적으로 활용되기 위해 갖춰야 할 3대 조건인 시공간 일관성, 자유 시점 관찰, 물리적 개입을 정의합니다. 기존 연구들이 공간 기억과 사용자 개입 표현을 분리하여 처리할 때 발생하던 문제(배경 밀림, 가림 오차 등)를 분석하고, 이를 해결하기 위해 공간 기억과 사용자 개입을 ‘하나의 통합된 3차원 월드 상태’의 기하 변환으로 묶어 다루는 연구 방향을 소개합니다. 사용자가 물체를 조작한 결과가 카메라 시점 변화 이후에도 온전히 유지되는 고정밀 인터랙티브 미디어의 응용 가능성을 보여줍니다.

네 번째 기고문(“월드 모델 기술 동향”)에서는 모델 기반 강화학습(MBRL)의 역사적 흐름 속에서 월드 모델이 어떻게 진화해 왔는지 기본 개념과 주요 기술적 이정표를 체계적으로 정리합니다. 잠재 동역학 공간에서 가상의 미래 궤적을 상상하여 대규모 과제를 해결한 전통적 연구(PlaNet, Dreamer 계열)부터 시작하여, 픽셀을 직접 복원하지 않고 추상 표현을 예측하는 I-JEPA/V-JEPA 계열을 비교 분석합니다. 그리고 최근 자율주행 분야에서 활발히 연구되고 있는 GAIA-1, Drive-Dreamer 및 범용 사전학습 물리 파운데이션 모델인 NVIDIA Cosmos에 이르기까지 범용 월드 모델의 발전 현황과 시각적 품질 및 물리적 인과 관계를 정량화할 수 있는 새로운 평가 기법들을 제안합니다.

다섯 번째 기고문(“From Visual Worlds to Actionable Worlds: Real-to-Sim for Robot Learning”)에서는 시각적인 모습만 흉내 내는 ‘보는 세계’에서 물체의 충돌, 접촉 상태, 관절 구동 및 물리 법칙이 작용하는 ‘행동하는 로봇 학습 세계’로 나아가는 Real-to-Sim 기술을 논의합니다. 단순한 픽셀 공간 예측이나 단일 가우시안 렌더링이 지닌 물리 시뮬레이션으로서의 한계를 극복하기 위해, 장면을 개별 객체 단위 자산(Asset)으로 분리하는 기법을 설명합니다. 3DGS의 시각 표현력과 삼각형 메시의 물리적 계산력을 결합한 하이브리드 표현(SplatMesh) 및 능동적 상호작용 기반의 물리 파라미터 추정법을 검토하고, 로봇 하드웨어에 종속되지 않는 계층적 월드 액션 모델(WAM)의 closed-loop 확장성을 제시합니다.

이번 “월드 모델 기반의 차세대 미디어 생성 기술” 특집호 출간을 위해서, 바쁘신 와중에서도 원고를 준비해 주신 저자분들께 깊이 감사드립니다. 본 특집호가 학계와 산업계 독자 여러분께 월드 모델 기술에 대한 동향 파악과 관련 연구에 조금이나마 도움이 되기를 기대합니다.