Physical AI의 최근 연구 동향과 발전 방향에 대한 연구

김종훈[1], 김의직[2], 김동완[3*]
[1]동아대학교 전자공학과 석사과정, [2]한림대학교 소프트웨어학부 교수, [3]동아대학교 전자공학과 부교수

A Study on Recent Research Trends and Future Directions in Physical AI

JongHoon Kim[1], Eui-Jik Kim[2], Dongwan Kim[3*]
[1]M.S. Course, Department of Electronic Engineering, Dong-A University
[2]Professor, Division of Software, Hallym University
[3]Associate Professor, Department of Electronics Engineering, Dong-A University

요 약 Physical AI는 로봇이나 드론과 같은 기계 시스템의 작동 방식을 학습하고 지능을 내장하여 상호작용하는 기술이다. Physical AI는 특히 자율 로봇 시스템 공학 분야에서 주목을 받고 있으며, 기계가 실제 환경에서 복잡한 작업을 독립적으로 학습하고 실행하도록 발전되고 있다. 본 논문은 자율 로봇 시스템에 적용되는 Physical AI의 대표적 연구 사례로 QT-Opt, Dreamer, Gato, RobotCat에 대해 알고리즘, 요구되는 데이터, 연산 복잡도, Sim-to-Real, 한계를 비교하고, Sim-to-Real 기반 연구로는 Domain Randomization, World Model 기반 자기 지도 학습(SGF), 의료 분야 사례를 검토하였다. 또한, 최근 연구 사례인 V-JEPA 2와 Cosmos에 대해 데이터 규모, 활용 사례, 한계점 측면에서 분석하였다. 마지막으로, 최근 연구 사례의 한계점을 바탕으로 향후 발전 방안에 대해 논의한다.

주제어 : 인공지능(AI), 물리적 인공지능, 자율 로봇 시스템, AI 융합, 시뮬레이션 - 실제 환경 전이

Abstract Physical AI refers to technologies enabling mechanical systems—such as robots and drones—to learn their operational dynamics and embed intelligence. This field has attracted significant attention in autonomous robotic systems engineering, where machines learn and execute complex tasks in real-world environments. In this paper, we compare QT-Opt, Dreamer, Gato, and RoboCat across algorithm, data requirements, computational complexity, Sim-to-Real, and limitations. We also review domain randomization and world-model-based self-supervised learning (SGF), and a surgical-robot case. Furthermore, we assess V-JEPA 2 and NVIDIA Cosmos in terms of data scale, use cases, and constraints. Finally, we discuss future development strategies based on the limitations identified in recent studies.

Key Words : Artificial Intelligence(AI), Physical AI, Autonomous Robotics Systems, AI Convergence, Sim-to-Real

이 논문은 2024년도 동아대학교 연구년 지원에 의하여 연구되었음.
*교신저자 : 김동완(dongwankim@dau.ac.kr)
접수일 2025년 07월 18일 수정일 2025년 08월 10일 심사완료일 2025년 08월 15일

1. 서론

최근 인공지능 기술의 눈부신 발전은 다양한 산업과 학문 분야에 걸쳐 새로운 응용 가능성을 열어주고 있다. 특히 사물인터넷 기술의 발전으로 인한 풍부한 실데이터 학습 기반 컴퓨터 비전, 대형 언어 모델 기술은 인공지능 기술 고도화를 가능하게 만들었다[1]. 하지만, 산업 현장에서 실제 물리적 동작을 수행하는 사람을 대체하는 로봇 동작과 연계된 인공지능 기술에 대한 고도화는 로봇을 학습시키기 위한 데이터 부족으로 인하여 컴퓨터 비전 및 대형 언어 모델 기술 대비 상대적으로 더딘 기술 발전이 이루어졌다[2]. 이에, 산업 현장에서 물리적 동작을 수행하는 로봇에 지능을 부여하고 로봇 수행 동작을 고도화하기 위한 Physical AI 기술에 대한 관심이 증대되기 시작하였다[3].

Physical AI는 실제 물리 환경에서 지각, 판단, 운동 제어를 통합하여 로봇이 지능적이고 적응적인 행동을 수행하는 기술이다[4]. 이러한 Physical AI는 다양한 기술의 융합을 요구한다. 특히 자율 로보틱스 기술과 인공지능 기술과의 결합은 물리적 환경에서 학습 기반으로 복잡한 작업을 수행하는 Physical AI의 핵심 기술로 일컬어지고 있다. 본 논문에서는 Physical AI 기술의 중 자율 로보틱스 기술과 연계된 사례 중심으로 대표적 연구 사례와 최근 동향을 분석하여 기술의 한계점을 도출하고, 이를 바탕으로 향후 연구 방안에 대해 제안한다.

2. Physical AI 대표적 연구 사례

Physical AI는 지능적 정보 처리를 넘어, 물리적 환경에서 자율적이고 적응적인 행동을 수행하기 위한 기술을 의미한다. Physical AI는 소프트웨어 연산으로 제한되지 않으며, 사물의 인식, 작업 계획, 제어 동작을 통합한 지능적 로봇 동작을 목표로 한다. 이러한 Physical AI의 대표 기술은 자율 로보틱스와 강화학습을 융합한 기술이다. 이러한 기술은 로봇이 실제 시행착오를 통해 최적의 행동 전략을 학습할 수 있도록 한다. 대표적 연구 사례는 DeepMind의 QT-Opt, Dreamer, Gato, 및 RoboCat 등이며, 이들은 현실 세계에서 물체 조작, 이동, 조립 등의 작업 고도화를 지향한다[5,6,7,8,9].

2.1 Q-function Targets via Optimization(QT-Opt)

대표적인 가치 기반 강화학습인 QT-Opt는 Closed-Loop Vision Based 제어를 통해 로봇 팔 파지 정책을 학습하는 분산 Q-Learning 기법이다[5]. 연속적인 행동 공간에서 Cross Entropy Method(CEM)를 활용하여 안정적으로 최적 행동을 선택한다. CEM은 매 반복마다 N개의 값을 샘플링하고, 이 중 상위 M 〈 N개 샘플에 가우시안 분포를 갱신하며, 다시 N개의 샘플을 생성하는 과정을 반복한다. 또한, Clipped Double Q-Learning을 통해 과대 추정 편향을 효과적으로 완화한다. 580,000회 이상의 실제 파지 시도 데이터를 기반으로 Off-Policy 학습과 28,000회의 추가 On-Policy 데이터를 통해 Joint Fine-Tuning을 수행하여, 미지의 객체에 대해 96%의 파지 성공률을 달성하였다.

$$\pi_{\bar{\theta}_i}(s) = \arg\max_a Q_{\bar{\theta}_i}(s,a) \quad (1)$$

수식 (1)은 Greedy Policy를 나타낸 것이며, $\pi_{\bar{\theta}_i}(\mathrm{s})$는 상태 s에서 선택되는 최적의 정책을 의미한다. $\bar{\theta}_i$는 Q-Function의 Target 네트워크 파라미터를 의미하며, 상태 s에서 Q-Value가 가장 높은 행동을 선택하는 것을 의미한다. 연속적인 공간에서는 $\arg\max_a$ 연산의 어려움으로, CEM을 통해 근사적으로 최적 행동을 탐색한다.

$$V_{\bar{\theta}_1,\bar{\theta}_2}(s') = \min_{i=1,2} Q_{\bar{\theta}_i}(s', \arg\max_{a'} Q_{\bar{\theta}_1}(s',a')) \quad (2)$$

수식 (2)는 Clipped Double Q-Learning의 Target Value를 나타낸 것이며, $V_{\bar{\theta}_1,\bar{\theta}_2}(s')$는 다음 상태 s'에서의 Target Value를 의미한다. 두 Target 네트워크 중 더 작은 값을 선택하여 과대 추정 편향을 완화할 수 있다. QT-Opt는 미지의 객체에 대해 96% 파지 성공률을 달성하였지만, 초기 Off-Policy 학습 단계를 위해 대규모 실제 데이터가 필요한 제한점과 CEM으로 인한 많은 연산을 필요하게 된다.

2.2 Dreamer

Dreamer는 고차원 이미지 관측값을 저차원 잠재(Latent) 공간으로 인코딩하고, 궤적을 예측하여 장기간 행동을 학습하는 강화학습 에이전트이다[6]. 이 에이전트는 Vision 제어 문제를 Partially Observable Markov Decision Process(POMDP)로 공식화하며, 총 5개의 모델로 아래의 식과 같이 구성된다.

$$p_\theta(s_t|s_{t-1}, a_{t-1}, o_t) \quad (3)$$
$$q_\theta(s_t|s_{t-1}, a_{t-1}) \quad (4)$$
$$q_\theta(r_t|s_t) \quad (5)$$
$$q_\phi(a_t|s_t) \quad (6)$$
$$v_\psi(s_t) \quad (7)$$

수식 (3)은 표현 모델, 수식 (4)는 전이 모델, 수식 (5)는 보상 모델, 수식 (6)은 행동 모델, 수식 (7)은 가치 모델을 나타낸다. 표현 모델은 관측 o_t와 행동 a_t를 인코딩하여 Markov 전이를 가지는 모델 상태 s_t를 생성한다. 전이 모델은 관측에 상관없이 미래의 상태를 예측하며, 보상 모델은 주어진 모델의 상태에서 보상을 예측한다. 행동 모델은 정책을 구현하여 행동을 예측하고, 가치 모델은 상태 가치를 추정한다. 이 모델들을 통해 Dynamic 학습, 행동 학습, 환경과 상호작용의 3단계로 학습한다.

〈Table 1〉 Dreamer Learning Pipeline

Steps	Dynamic Learning	Behavior Learning	Environment Interaction
1	Data extraction	Generating virtual	Reset environment
2	Encoding observation and action	Predict reward and values	Compute state
3	Update model parameter	Compute return	Predict action
4		Update model	Exploration
5			Execution
6			Add new experience

〈Table 1〉은 Dreamer의 3 단계 학습을 나타낸 표로, Dynamic 학습은 과거 경험을 통해 World 모델을 학습하여 미래 상태와 보상을 예측할 수 있게 한다. 다음 단계인 행동 학습은 학습된 잠재 Dynamic 모델을 사용하여 최적의 행동과 가치를 학습한다. 마지막으로 환경과 상호작용은 학습된 행동 모델을 실제 환경에 적용하여 새로운 경험을 수집하는 단계이다. 위와 같은 학습 과정을 통해 Dreamer는 잠재 공간으로 인코딩하여 데이터 효율성을 높이고, 궤적 내에서 보상을 Backpropagation 하여 행동 및 가치 모델을 최적화함으로써, 장기적인 의사결정을 가능하게 한다. 그러나, 고정된 길이 내에서 궤적을 생성하고 보상을 고려하여 일부 장기 보상을 고려하지 못하는 한계를 가진다.

2.3 Gato

Gato는 범용 트랜스포머 기반 에이전트로, Multi-Modal, Multi-Task, Multi-Embodiment를 지원하는 범용 정책으로 작동한다[7]. 1.2B 파라미터, 24 계층, 2,048 크기의 임베딩, 8,196 차원의 Post-Attention Feedforward를 가진 디코더 전용 트랜스포머로 구성되어 있다. 텍스트, 이미지, 로봇 관절, 이산 및 연속 관측과 행동 등 Multi-Modal 데이터를 처리할 수 있도록, 모든 데이터를 flat 토큰 시퀀스로 직렬화하여 입력으로 사용한다. 데이터를 토큰으로 변환하기 위해 데이터 형식에 따라 아래와 같이 토큰화한다.

- Text: 32,000개의 Subword로 인코딩된다.
- Image: Raster 순서로 겹치지 않는 16 x 16 patch 시퀀스로 변환된다.
- Discrete Value: 행 방향을 우선 순서로 정수 시퀀스로 변환된다.
- Continuous Value: 행을 우선 순서로 부동소수점 시퀀스 변환되지만, 값이 [-1, 1]에서 벗어나면, μ-Law 인코딩하고, 1,024개의 균등 분할로 양자화된다.

이와 같이 데이터를 토큰으로 변환한 후, 다음과 같은 순서를 따른다.

- Text: 원본 입력 텍스트와 동일한 순서로 나열한다.
- Image: Raster 순서로 나열한다.
- Tensor: 행 방향을 중심으로 나열한다.

데이터를 토큰화하고 순서를 정한 후에는 각 토큰에 대해 임베딩을 수행함으로써, 최종 모델 입력을 생성한다.

$$\log p_\theta(s_1, \ldots, s_L) = \sum_{l=1}^{L} \log p_\theta(s_l|\, s_1, \ldots, s_{l-1}) \quad (8)$$

수식 (8)은 토큰 시퀀스에 대해 Autoregressive Modeling을 나타낸다. $s_{1:L}$는 전체 토큰 시퀀스이며, p_θ는 시퀀스 전체 확률을 의미한다. 확률의 연쇄 법칙을 사용하여 데이터를 모델링하며, 이 값을 기반으로 텍스트와 행동 토큰에만 마스크된(Masked) Cross- Entropy로 훈련 손실을 얻을 수 있다. 또한, 평가 시에 성공적인 시연 시퀀스를 프롬프트로 삽입하여, 별도의 식별자 없이 원하는 목표 행동으로 모델을 유도할 수 있다. 이와 같이 Gato는 서로 다른 데이터를 단일 가중치로 처리할 수 있어 범용성이 뛰어나며, 프롬프트 조건화로 새로운 작업에 빠르게 적응할 수 있다. 그러나, 최대 1,024 토큰으로 문맥 길이의 제한점과 이미지 및 에이전트의 관측 토큰은 처리하지 못하는 한계점이 있다.

2.4 RoboCat

비전 기반 로봇 조작을 위한 자기 개선(Self-Improvement) 범용 에이전트인 RoboCat은 Gato에서 제시한 Autoregressive 트랜스포머 아키텍처를 기반으로 학습한다[8]. 이 에이전트는 현재 관절 상태, 이미지 관측 및 목표 이미지를 입력받아 VQ-GAN 인코더를 사용하여 미래 이미지 토큰을 예측한다. 이 과정에서 100~1,000개의 실제 데이터만으로 Fine-Tuning 하여 새로운 로봇 형태, 물체 및 지각 변형, Sim-to-Real 등에 성공적으로 적응할 수 있다.

$$\pi(a_t|o_t, g_t) = P_\theta(a_t|x_{<t}, I_{<t}, g_{<t}) \quad (9)$$

수식 (9)는 RoboCat의 정책을 나타내며, Auto-regressive 트랜스포머 모델로 모형화된다. g_t는 목표 이미지 토큰 시퀀스, I_t는 이미지 관측, x_t는 관절 상태를 나타낸다. P_θ는 정책 네트워크를 의미하는데, Context로 받은 각 정보를 통해 다음 행동을 예측한다. 이 모델을 기반으로 사전 학습, Generalist 학습, Fine-Tuning, 자기 개선 순으로 최적화가 진행된다. 이로 인해, RoboCat은 VQ-GAN으로 Multi-Modal 입력을 처리하고, 적은 수의 시연으로 빠르게 적응할 수 있으며, 자기 개선을 통해 성능을 향상시켰다. 그러나, 초기 데이터의 양과 품질에 대한 의존도가 높아, 데이터가 부족하거나 낮은 품질일 경우, 학습 수렴 속도가 느려진다.

〈Table 2〉는 Physical AI의 대표적 연구 사례인 QT-Opt, Dreamer, Gato, RoboCat를 사용된 알고리즘, 요구되는 데이터, 연산 복잡도, Sim-to-Real, 한계 순으로 비교한 표를 나타낸다. 연산 복잡도 및 Sim-to-Real의 성능에 대해 ○, △, × 로 비교하였으며, 각각 우수(낮은 복잡도), 보통, 미흡(높은 복잡도)을 의미한다.

2.5 실제 로봇 구현(Optimus)

Tesla가 공개한 Optimus는 무게가 57kg, 키가 1.73m의 인간형 로봇으로, 최대 20kg 리프팅 및 8km/h 이동이 가능하며, 총 40개의 Actuator를 탑재한다[9]. 자동차 제조 환경에서 반복적 및 위험한 작업을 수행하도록 설계되었지만, 현장 적용의 효율성, 제한적인 자율성 등의 한계로 인해 개선하기 위한 연구들이 진행되고 있다.

3. Sim-to-Real 연구 사례

실제 로봇을 통해서 하는 데에는 Physical AI를 학습하기에는 한계가 있기에 시뮬레이션 영역을 통한 학습으로 Physical AI가 확장되고 있다. 대표적으로, 물리 환경에서 직접 학습하는 데 드는 시간과 비용, 안전 문제를 해결하기 위해 시뮬레이션 기반 학습과 실제 환경으로의 전이(Sim-to-Real)가 중요해졌다. 이를 위해, Domain Randomization, World Model 기반 자기 지도 학습(Self-Supervised Learning, SSL), 비디오 기반 학습 등의 기법이 연구되고 있다[10,11,12].

3.1 Domain Randomization

Domain Randomization은 현실과 시뮬레이션 간의 차이를 극복하기 위해 제안된 기법으로, 시뮬레이터를 무작위로 수행하여 모델이 다양한 환경을 경험할 수 있게 한다[10]. 이 기법은 충분한 시뮬레이션 다양성을 제공하여 실제 환경에서도 모델이 일반화될 수 있도록 한

〈Table 2〉 Comparison of Representative Physical AI Research Cases

Representative Cases	Algorithm	Data Requirements	Computational Complexity	Sim-to-Real	Limitation
QT-Opt[5]	Distributed Q-Learning	Over 580k Real Grasping Data	×	×	Requires massive data
Dreamer[6]	Model-based RL	High-Dimensional Images	△	△	Limitations in reflecting long-term reward
Gato[7]	Autoregressive Transformer	Multi-Modal Data	×	△	Context length limitation(1,024 Token)
RoboCat[8]	Gato + Self-Improvement	Small-Scale Real Data	△	○	Sensitive to initial data

다. 먼저 데이터마다 다음과 같은 Domain 요소를 무작위화한다.

- 테이블 위의 장애물 수와 형태
- 테이블 위 모든 물체의 위치와 Texture
- 테이블, 바닥, 박스, 로봇의 Texture
- 카메라의 위치, 방향, 시야각
- 조명의 위치, 방향, 반사 특성
- 이미지에 추가되는 무작위 노이즈의 종류와 양

또한, 무작위 RGB, 두 개의 무작위 RGB 사이의 Gradient, 두 개의 무작위 RGB 값으로 이루어진 체크무늬 패턴 중에서 무작위 노이즈로 선택된다. 이와 같이 무작위화된 시뮬레이터에서 수십만 장의 렌더링 된 이미지와 대응되는 객체 중심 좌표를 라벨링 하여 합성 데이터를 생성한다. 이 데이터를 기반으로 Adam Optimizer를 사용하여 학습한다. 시뮬레이터에서 학습한 모델을 실제 영상에 적용한 결과, Grasping 로봇이 실제 위치와의 평균 감지 오차가 약 1.5cm로 위치를 예측하고, 잡동사니나 부분 가림과 같은 특수한 상황에서도 우수한 성능을 보였다. 그러나, 각 객체마다 3D 모델을 생성하고, 개별의 탐지기를 학습해야 하며, 한 번 학습된 모델은 새로운 환경에서는 적용이 어렵다는 한계점을 가진다.

3.2 Simple, Good, Fast(SGF) World Model

SGF World Model은 자기 지도 표현 학습(Self-Supervised Representation Learning), 프레임 및 행동 Stacking, 데이터 증강을 활용한 World Model이다[11]. 자기 지도 표현 학습으로 이미지 관측을 효과적으로 임베딩하고, 이를 바탕으로 전이, 보상, 종료 모델을 결정론적 예측 방식으로 학습한다. 학습 후 상태 공간에서 예측한 궤적을 활용하여 Policy-Gradient로 정책을 최적화한다. 자기 지도 표현 학습은 시간 일관성을 위해 임베딩 Mean Squared Error(MSE)를 최소화하고, Variance-Invariance-Covariance Regulari-zation(VICReg)[12]에서 제안된 Variance and Con-variance(VC)를 사용하여 정보량을 최대화한다.

$$L_{Repr.}(\theta) \tag{10}$$
$$= \mathbb{E}_t[\overbrace{\frac{\eta}{D} \| h_\theta(\tilde{z},a) - \tilde{z}' \|_2^2}^{Temporal\ Consistency} + \overbrace{VC(\tilde{z}) + VC(\tilde{z}')}^{Imformation\ Maximization}]$$

$$VC(z) = \frac{1}{D}\sum_{j=1}^{D}[VR(z) + CR(z)] \tag{11}$$

$$VR(z) = \rho \max(0, 1 - \sqrt{\mathrm{Cov}(z)_{i,j} + \varepsilon}) \tag{12}$$

$$CR(z) = v\sum_{k \neq j} \mathrm{Cov}(z)_{j,k}^2 \tag{13}$$

수식 (10)은 전체 표현 손실, 수식 (11)은 VC, 수식 (12)는 Variance Regularization(VR), 수식 (13)은 Covariance Regularization(CR)을 나타낸다. 여기서, $\tilde{z}, \tilde{z}'$는 현재와 다음 관측값에 대한 임베딩, $h_\theta(\tilde{z},a)$는 다음 임베딩을 예측기로 예측한 값, η는 가중치, D는 임베딩 차원, $\rho, v > 0$은 각각 VR, CR의 가중치, $\varepsilon = 1 \times 10^{-4}$는 수치적 불안정을 방지하기 위한 값을 의미한다. VC를 통해 정보량을 최대화할 수 있고, 예측한 임베딩과 관측한 임베딩 사이의 MSE를 최소화하여, 일관성을 보장할 수 있다. SGF는 Dreamer와 비슷하게 잠재 World 모델을 학습하지만, 자기 지도 표현 학습, 결정론적 전이 모델 사용, Actor-Critic 방식 적용 등을 통해 트랜스포머, Recurrent Neural Network(RNN), 이미지 재구성 없이도 효과적인 World Model을 구축할 수 있다. 결정론적 Markov Decision Process(MDP)에만 적용이 가능하며, 이미지 관측에만 한정된 제약이 있다.

3.3 Sim-to-Real 기반 의료 분야 연구 동향

Domain Randomization 및 World Model 기반 자기 지도 학습과 같은 기법을 기반으로 Sim-to-Real을 의료 분야에서 적용하는 다음과 같은 연구가 진행되고 있다. Yafei Ou and Mahdi Tavakoli[13]는 로봇 보조 수술(Robot-Assisted Surgery, RAS)에서 Indirect Simultaneous Positioning(ISP) 문제 해결을 위해 Sim-to-Real 기반 접근을 제안한다. Finite Element Modeling(FEM) 기반 시뮬레이션 환경에서 심층 강화학습을 활용해 정책을 학습하며, State Augmented MDP를 통해 시뮬레이션의 불규칙성을 완화한다. 또한, Bayesian Optimization(BO)을 사용함으로써, Grasping Point를 통해 수술 전 계획된 Grasping Point가 설정된다. 먼저 FEM 기반 Soft Object 시뮬레이션 프레임워크인 SOFA[14]를 사용하여 환경 구축한다.

$$\mathbf{M}\ddot{x}(\tau) = F_{int}(\tau) + F_{ext}(\tau) \tag{14}$$

수식 (14)는 FEM 기반 시뮬레이션에서 사용되는 뉴턴 법칙을 보여준다. $\mathbf{M}$은 질량 행렬, x는 시스템의 Degrees of Freedom(DOF)이며, $\mathbf{M}\ddot{x}$는 질량 x 가속도인 관성력을 의미한다. 또한, F_{int}와 F_{ext}는 각각 내부

와 외부 힘의 벡터를 의미한다. 이 식을 기반으로 시간 단위로 수치 해석하여 FEM 시뮬레이션에서 조직의 움직임을 예측할 수 있다. 이 기법은 각 포인트를 파악하고 이동시키기 위해 아래와 같은 Displacement Constraint를 적용한다.

$$F_a = k_s(q_{t+1} - q_t) \tag{15}$$

수식 (15)는 시뮬레이션에서 Grasping Point가 움직일 때 조직 Mesh에 작용하는 스프링 힘을 계산하는 것을 보여준다. k_s는 스프링 강성 계수를 나타내며, q_t와 q_{t+1}은 각각 잡힌 노드의 조작 전과 조작 후의 위치를 의미한다. $q_{t+1} - q_t$를 계산하여 실제로 이동한 벡터를 계산할 수 있다. 이 스프링 모델 기반 식을 통해 Grasping Point가 이동함에 따라 조직 Mesh에 작용하는 변형력을 계산하고, 특정 경로로 유도할 수 있다.

$$s_t = p_t - p_{des} \in \mathbb{R}^{N \cdot D} \tag{16}$$

$$a_t = \Delta q = q_{t+1} - q_t \in [-0.2, 0.2]^{M \cdot D} \tag{17}$$

$$r(s_t, a_t) = \lambda\left(1 - \sqrt{\frac{\| p_t - p_{des} \|}{\| p_0 - p_{des} \|}}\right) \tag{18}$$

수식 (16), 수식 (17), 수식 (18)은 각각 상태 공간, 행동 공간, 보상 함수를 보여준다. N은 Controlled Point의 수, M은 Grasping Point의 수, D는 공간 차원, p_t는 Controlled Point의 위치, p_{des}는 Controlled Point의 목표 위치를 의미한다. 또한, q_t와 q_{t+1}은 현재와 다음 시간의 Grasping Point의 위치, $\| p_t - p_{des} \|$와 $\| p_0 - p_{des} \|$는 각각 현재 오차 크기 및 초기 오차 크기, λ는 Scaling Factor를 나타낸다. 이와 같이 정의된 상태, 행동, 및 보상을 기반으로 MDP를 정의한다. 또한, State Augmented MDP를 도입하기 위해 상태 공간을 아래와 같이 확장한다.

$$\begin{aligned} J &:= S \times A^{K-1} \\ j_t &:= (s_t, a_{t-1}, a_{t-2}, \cdots, a_{t-K+1}) \end{aligned} \tag{19}$$

수식 (19)는 증강된 상태 공간을 보여준다. J는 증강된 상태 공간, K는 최대 길이를 의미한다. 이 식을 기반으로 상태 공간을 확장하여 FEM 기반 시뮬레이터에서의 POMDP가 MDP로 근사하여 불규칙성을 완화할 수 있다.

$$\begin{aligned} &\pi^* \\ &= \arg\max_{\pi} \sum_{t=0}^{T} \mathbb{E}_{(s_t, a_t) \sim \rho}[\gamma^t r(s_t, a_t) + \alpha H(\pi(\cdot|s_t))] \end{aligned} \tag{20}$$

수식 (20)은 SAC 알고리즘의 최적 정책을 보여준다. $H(\pi(\cdot|s_t))$는 정책의 엔트로피, α는 온도 파라미터를 나타낸다. SAC는 최대 엔트로피 강화학습 프레임워크를 기반으로 하는 Off-Policy 알고리즘으로, 과거 경험을 저장하여 Off-Policy 학습을 가능하게 한다.

$$\begin{aligned} q_0^* &= \arg\min_{q_0 \in U} f(q_0) \\ &= \arg\min_{q_0 \in U} [q_e(q_0, p_0, p_{des}, \pi) - q_0] \end{aligned} \tag{21}$$

수식 (21)은 BO를 위한 목적함수를 의미한다. q_0와 p_0는 각각 초기 Grasping Point와 Controlled Point, p_{des}는 Desired Position, U는 가능한 초기 Grasping Location의 집합을 의미한다. 또한, q_e는 학습된 정책 π로 시뮬레이션을 실행하였을 때 얻어지는 Grasping Point의 최종 위치를 나타내는 함수를 의미한다. 목적함수를 최소화하기 위해 [Fig. 1]의 알고리즘을 기반으로 시뮬레이션 환경에서 최적의 Grasping Point를 탐색한다. Acquisition Function으로 Expected Improvement(EI)가 사용되었으며, EI는 계산된 목적함수보다 새로운 Grasping Point에서 개선될 확률과 그 차이를 기반으로 기댓값을 도출하는 함수를 의미한다. 약 4.1시간의 학습과 377초의 계획 시간을 통해, Controlled Point를 약 1.3mm 이내의 오차로 배치하였다. 그러나, 2D 환경을 기반으로 실험되어, 실제 수술 환경에서 적용하기 위해서는 3D 환경으로의 확장, 복잡한 생체 조직 모델링, 실시간 제어에 대한 추가 연구 등이 필요하다.

Algorithm 1 Grasping Point Optimization

Input: p_0, p_{des}, policy π, U, acquisition function $a(q_0)$,
Gaussian Process Estimator(GP),
initial number of sample n_0, total number of evaluations N
Output: Optimal grasping point q_0^*

```
Register p_0 and p_des in the simulator;
Sample q_0 ~ Uniform(U);
Evaluate f(q_0^*) using π;
n ← n_0;
while n ≤ N do
    Update the posterior distribution of GP ← q_0^n;
    Optimize q_0^n = arg max a(q_0);
    Evaluate f(q_0^*) using π;
    n ← n + 1;
end while
Choose q_0^* = arg min f(q_0);
```

[Fig. 1] Grasping Point Optimization[13]

〈Table 3〉 Comparison of Sim-to-Real Cases

Sim-to-Real Cases	Algorithm	Data Requirements	Computational Complexity	Sim-to-Real	Limitations
Domain Randomization[10]	Randomization + Adam	Hundreds of Thousands of Rendered RGB Images	△	○	Requires a separate 3D model and detector per object
SGF [11]	Self-Supervised Representation Learning	64x64 RGB	○	△	Applicable only to deterministic MDPs and relies on image-based observation
Sim-to-Real Surgical Robot Learning[12]	FEM+BO	4 Vertices of Phantom Tissue and 2 Controlled Points	×	△	Limitations of the 2D environment-based approach

〈Table 3〉은 Domain Randomization, SGF, Sim-to-Real Surgical Robot Learning을 사용된 알고리즘, 요구되는 데이터, 연산 복잡도, Sim-to-Real, 한계 순으로 비교한 표를 나타낸다.

4. Physical AI 최근 연구 사례 비교

본 절에서는 최근 Physical AI 분야에서 주목받고 있는 Meta AI의 V-JEPA 2[15]와 NVIDIA의 Cosmos [16]를 비교 및 분석한다.

먼저, Meta V-JEPA 2는 대규모 비디오 관찰 데이터로부터 자기 지도 학습을 통해 World Model을 학습하고, 소량의 로봇 동작 데이터로 Fine-Tuning 하는 단계적 학습 구조를 갖는다[15]. 학습 절차는 다음과 같다.

1. Video Pre-Training
2. Action-Conditioned World Model
3. Planning: Zero-Shot Robot Control

Video Pre-Training 단계에서는 100만 시간 이상의 인터넷 비디오와 100만 장의 이미지를 포함하는 데이터셋을 사용한다.

$$\min imize_{\theta,\phi,\Delta_y} \| P_\phi(\Delta_y, E_\theta(s)) - sg(E_{\bar{\theta}}(y)) \|_1 \quad (22)$$

수식 (22)는 사전 학습 단계에서 사용되는 목표 함수를 보여준다. θ는 인코더 E의 파라미터, ϕ는 예측기 P의 파라미터, Δ_y는 마스크된 위치인 학습 가능한 마스크 토큰, $sg(\cdot)$는 Stop-Gradient 함수를 의미한다. 또한, x는 입력 비디오 y에서 무작위로 마스크되거나 제거된 값을 나타낸다. 인코더로 마스크된 표현을 추출하고, 예측기로 표현을 예측하며, Stop-Gradient로 Gradient Backpropagation을 차단한다. 이 식 기반으로 예측기와 인코더를 동시에 업데이트하며, 손실은 마스크된 값만으로 계산된다. 다음 단계인 Action-Conditioned World Model은 사전 학습된 인코더를 고정하고, 소량의 데이터만으로 학습하여 Action-Conditioned 예측이 가능한 Latent World Model을 도출하는 단계이다. 4 fps, 256x256, 16 프레임인 총 62시간 분량의 영상을 Tokenizer로 잠재 표현으로 인코딩하고, 행동 및 상태를 시공간 순서대로 입력한다. 24 계층 트랜스포머 예측기가 다음 잠재 표현을 예측하도록 학습하여 V-JEPA 2-AC를 얻을 수 있다. Planning: Zero-Shot Robot Control 단계에서는 V-JEPA 2-AC를 기반으로 목표 이미지 하나만 사용하여 대응하는 행동 시퀀스를 실시간 탐색 및 실행한다.

$$\varepsilon(\hat{a}_{1:T}; z_k, s_k, z_g) := \| P(\hat{a}_{1:T}; s_k, z_k) - z_g \|_1 \quad (23)$$

수식 (23)은 V-JEPA 2-AC를 사용하여 계획할 때 최소화하는 에너지 함수를 나타낸다. z_k는 현재 로봇 카메라 영상을 인코딩한 특징 맵, s_k는 현재 로봇 팔의 End-Effector 상태, z_g는 목표 이미지를 인코딩한 특징 맵을 나타내며, P는 미래 상태 표현 함수를 의미한다. 이 에너지 함수를 CEM을 통해 최소화하여 행동 시퀀스를 찾는다. 이 행동 시퀀스 중 첫 행동만 로봇에 실행하고, 로봇 상태가 변하면 다시 반복하여 계획하는 Receding Horizon Control 기법을 통해 별도의 학습 없이 목표 이미지만으로 Zero-Shot 로봇 조작을 가능하게 한다. V-JEPA 2는 대규모 자기 지도 비디오 학습을 통해 물리적 세계를 이해하고, Zero-Shot 로봇 조작 계획을 가능하게 하였다. 그러나,

카메라 위치에 대한 민감성, 목표 이미지 기반의 계획으로 언어 및 다른 데이터 형식의 한계 등을 가지고 있다.

NVIDIA Cosmos는 특정 Physical AI 환경에 맞춰 World Foundation Model(WFM)을 미세 조정하여 맞춤형 World Model을 구축할 수 있도록 공개되었다[16]. Cosmos는 Video Curator, Tokenizers, Pre-Trained World Foundation Models, World Foundation Model Post-Training Samples, Guardrail의 다섯 가지 주요 구성 요소로 이루어져 있다. 먼저, Video Curator는 Split, Filtering, Annotation, Deduplication, Sharding의 5단계의 파이프라인으로 이루어져 있다. Split 단계는 긴 비디오를 Shot으로 나누고 클립에 기록하며, Filtering 단계는 World Foundation Model 구축에 가치 없는 클립을 제거한다. Annotation 단계에서는 비디오 설명을 추가하고, Deduplication 단계에서는 의미적 중복을 제거한 후, 비디오 클립을 해상도와 화면 비율에 따라 Sharding 한다. Tokenizer는 대규모 모델의 기본 구성 요소로써, 인코더-디코더 구조로 설계되었다.

$$\hat{x}_{0:T} = D(\varepsilon(x_{0:T})) \tag{24}$$

수식 (24)는 인코더-디코더 구조를 간단히 나타낸 것이다. $x_{0:T} \in \mathbb{R}^{(T+1)\times H\times W\times 3}$은 원본 비디오 시퀀스로, $H\times W$크기의 RGB 이미지 프레임을 나타낸다. ε은 인코더 함수로, 원본 비디오를 잠재 토큰 $z = \varepsilon(x_{0:T})$로 압축한다. D는 디코더 함수로 z를 원본 비디오와 같은 형태로 복원한다. 이 식을 통해 재구성 영상 $\hat{x}_{0:T}$를 얻을 수 있으며, 이는 원본에 얼마나 동일하게 복원하였는지의 평가값이 된다. World Foundation Model(WFM) Pre-Training은 Diffusion 기반 WFM과 Autoregressive 기반 WFM으로 나뉜다. Diffusion 기반 WFM은 인코딩된 잠재 공간에서 동작하는 잠재 Diffusion Model로, 노이즈 제거 과정을 통해 패턴을 학습하고, 텍스트 조건을 반영한 고품질 비디오 생성을 가능하게 한다. 또한, Autoregressive 기반 WFM은 과거 출력 시퀀스를 조건으로 활용하여 예측하는 구조로, 누적된 Context를 반영해 일관성과 디테일을 유지한 비디오 생성을 가능하게 한다. World Model Post-Training에서는 Pre-Trained WFM을 특정 작업에 맞춰 Fine-Tuning 하는 단계이다. 마지막으로, Guardrail은 Pre-Guard와 Post-Guard로 구성되어 유해 프롬프트 및 출력을 차단할 수 있다. Pre-Guard는 텍스트 영역에서 동작하며, Large Language Model(LLM) 기반 Filter와 Keyword 차단 목록을 사용한다. 이를 통해 명시적인 유해 단어를 포함하는 프롬프트를 차단하고, 유해한 프롬프트로 판단할 경우, 출력하지 않고 오류 메시지를 전송한다. Post-Guard는 생성된 비디오 출력을 Face Blur Filter와 비디오 콘텐츠 안전 Filter로 검사한다. Face Blur Filter는 얼굴 인식 신경망을 통해 얼굴 영역의 크기가 일정 픽셀 이상이면 픽셀화하여 흐리게 만든다. 또한, 비디오 콘텐츠 안전 Filter는 비디오를 안전과 유해로 분류하여 일정 프레임이 안전하지 않은 것으로 분류되면 전체 비디오를 유해한 것으로 표시한다. Cosmos는 카메라 제어, 로봇 조작, 자율 주행 등 다양한 작업 환경에서 범용적으로 사용될 수 있는 WFM을 제안하고, Guardrail을 통해 유해 출력 및 입력을 차단할 수 있다. 그러나, 객체를 일관되게 유지하지 못하는 문제, 실제 물리 법칙을 준수하지 못하는 경우, 명령어와 실제 행동 간의 불일치 등의 제한점을 가지고 있다.

〈Table 4〉 Comparison of V-JEPA 2 and Cosmos

Recent Trend Researches on Physical AI	Algorithm	Data Scale	Use Cases	Advantages	Limitations
V-JEPA 2[14]	Large-Scale Self-Supervised Video Pre-training+ Receding Horizon Control	Over 1 Million Hours of Internet Videos and 1 Million Images	Probe-Based Classification, Video Q&A, and Zero-Shot Robot Control	Zero-Shot	Sensitive to camera viewpoint and limitations of language and other Type data
Cosmos[15]	WFM+LLM	About 100M Clips of Videos	Camera Control, Robot Manipulation, and Autonomous Driving	Generality and Safety	Object permanence and consistency issues, physical violations, and discrepancy between commands and actions

〈Table 4〉는 V-JEPA 2와 Cosmos를 알고리즘, 데이터 규모, 활용 사례, 이점, 한계점에 대해 비교한 표이다.

5. 결론

Physical AI는 지능이 물리적 구현을 통해 실현된다는 관점에서, 전통적인 인공지능과는 근본적인 차이를 지닌다. 자율 로보틱스와 강화학습 기술과의 융합은 이러한 지능을 실제로 구현하기 위한 핵심 기술이다. 본 논문은 DeepMind의 QT-Opt, Dreamer, Gato, RoboCat을 비교·분석을 통해, 대규모 데이터, 높은 연산 자원 필요성 등이 실제 환경에서의 적용성 및 일반화에 대한 한계점을 분석하였다. Sim-to-Real 측면에서는 Domain Randomization과 World Model 기반 자기 지도 학습(SGF)이 일반 로봇 팔 파지뿐만 아니라 의료 환경과 같은 특수한 환경에서의 적용 가능성을 확인하였으나, 여전히 한정된 적용성과 3D 환경의 한계 등 개선점이 남아있다. 최근 제안된 연구 사례인 V-JEPA 2는 인터넷 비디오와 로봇 시연 데이터를 통한 자기 지도 사전 학습을 통해 소량의 데이터만으로 성능을 향상시키는 데이터 효율성 문제를 완화하였다. 또한, Cosmos는 대규모 Multi-Modal을 학습하여 다양한 물리적 환경 과제에서 범용적으로 적용 가능성과 안전 및 윤리 문제의 해결 방안을 보여주었다. 그러나, 여전히 카메라 위치의 한계, 물리 법칙 불일치, 객체 유지 문제와 같은 한계가 남아있다. 향후 연구에서는 실제 물리 법칙의 이해를 기반으로 객체 일관성 보장, 모델 규모 확장 등의 개선 방안이 필요하다.

REFERENCES

[1] I.Kaur and A.J.Jadhav, "Survey on Computer Vision Techniques for Internet-of-Things Devices", 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology(IAICT), Bali, Indonesia, pp.244-250, 2023.

[2] J.Arents and M.Greitans, "Smart Industrial Robot Control Trends, Challenges and Opportunities within Manufacturing", Applied Sciences, Vol.12, No.2, pp.937, 2022.

[3] A.Miriyev and M.Kovac, "Skill for Physical Artificial Intelligence", Nature Machine Intelligence, Vol.2, No.11, pp.658-660, 2020.

[4] F.Bousetouane, "Physical AI Agents: Integrating Cogni- tive Intelligence with Real-World Action", arXiv preprint arXiv:2501.08944, pp.1-27, 2025.

[5] D.Kalashnikov, A.Irpan, J.Ibarz, A.Herzog, E.Jang, D. Quillen, E.Holly, M.Kalakrishnan, V.Vanhoucke, and S.Levine, "QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation", in Proceedings of The 2nd Conference on Robot Learning (COLR), Vol.87, pp.651-673, 2018.

[6] D.Hafner, T.Lillicrap, J.Ba, and M.Norouzi, "Dream to Control: Learning Behaviors by Latent Imagination", in Proceedings of the International Conference on Learning Representations(ICLR), pp.1-20, 2020.

[7] S.Reed, K.Zolna, E.Parisotto, S.G.Colmenarejo, A.Novikov, G.B.Maron, M.Gimenez, Y.Sulsky, J.Kay, J.T. Springenberg, T.Eccles, J.Bruce, A.Razavi, A.Edwards, N.Heess, Y.Chen, R.Hadsell, O.Vinyals, M.Bordbar and N.D.Freitas, "A Generalist Agent", Transactions on Machine Learning Research(TMLR), No.371, pp.2835-8856, 2022.

[8] K.Bousmalis, G.Vezzani, D.Rao, C.Devin, A.X.Lee, M. Bauza, T.Davchev, Y.Zhou, A.Gupta, A.Raju, A.Laurens, C.Fantacci, V.Dalibard, M.Zambelli, M.F.Martins, R. Pevceviciute, M.Blokzijl, M.Denil, N.Batchelor, T.Lampe, E.Parisotto, K.Zolna, S.Reed, S. G. Colmenarejo, J. Scholz, A. Abdolmaleki, O. Groth, J.B.Regli, O.Sushkov, T.Rothorl, J.E.Chen, Y.Aytar, D.Barker, J.Ortiz, M. Riedmiller, J.T.Springenberg, R.Hadsell, F.Nori, and N. Heess, "RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation", Transactions on Machine Learning Research(TMLR), No.1540, pp.2845-8856, 2024.

[9] A.A.Malik, T.Masood, and A.Brem, "Intelligent Humanoids in Manufacturing to Address Worker Shortage and Skill Gaps: Case of Tesla Optimus", arXiv preprint arXiv:202304.04949, pp.1-20, 2023.

[10] J.Tobin, R.Fong, A.Ray, H.Schneider, W.Zaremba, and P.Abbeel, "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World", in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS), Vancouver, BC, Canada, pp.23-30, 2017.

[11] J.Robine, M.Hoftmann, and S.Harmeling, "Simple, Good, Fast: Self-Supervised World Models Free of Baggage", in Proceedings of the International Conference on Lear- ning Representation(ICLR), Nairobi, Kenya, No.10365, pp.1-21, 2025.

[12] A.Bardes, J.Ponce, and Y.LeCun, "VICReg: Variance-Invariance-Convariance Regularization for Self-Supervised Learning", in Proceedings of the International Conference on Learning Representations(ICLR), pp.1-22, 2022.

[13] Y.Ou and M.Tavakoli, "Sim-to-Real Surgical Robot Learning and Autonomous Planning for Internal Tissue Points Manipulation Using Reinforcement Learning", IEEE Robotics and Automation Letters,

Vol.8, No.5, pp.2502-2509, 2023.

[14] F.Faure, C.Duriez, H.Delingette, J.Allard, B.Gilles, S. Marchesseau, H.Talbot, H.Courtecuisse, G.Bousquet, I.Peterlik, and S.Cotin, "SOFA: A Multi-Model Framework for Interactive Physical Simulation", Soft Tissue Biomechanical Modeling for Computer Assisted Surgery, Berlin: Springer, pp.283-321, 2012.

[15] M.Assran, A.Bardes, D.Fan, Q.Garrido, R.Howes, M. Komeili, M.Muckley, A.Rizvi, C.Roberts, K.Sinha, A.Zholus, S.Arnaud, A.Gejji, A.Martin, F.R.Hogan, D.Dugas, P.Bojanowski, V.Khalidov, P.Labatut, F.Massa, M.Szafraniec, K.Krishnakumar, Y.Li, X.Ma, S.Chandar, F.Meier, Y.LeCun, M.Rabbat, and N.Ballas, "V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning", arXiv preprint arXiv: 2506.09985, pp.1-48, 2025.

[16] NVIDIA, "Cosmos World Foundation Model Platform for Physical AI", arXiv preprint arXiv:2501.03575, pp.1-75, 2025.

김 종 훈(JongHoon Kim) [준회원]

- 2025년 2월 : 동아대학교 전자공학과 (공학사)
- 2025년 3월 ~ 현재 : 동아대학교 전자공학과 석사과정

〈관심분야〉
전기자동차 표준, Edge 컴퓨팅, Physical AI

김 의 직(Eui-Jik Kim) [정회원]

- 2004년 2월 : 고려대학교 전기전자전파공학부 (공학사)
- 2006년 2월 : 고려대학교 전자컴퓨터공학과 (공학석사)
- 2013년 2월 : 고려대학교 전기전자전파공학과 (공학박사)
- 2006년 2월 ~ 2009년 7월 : 삼성전자 DMC연구소 선임연구원
- 2009년 8월 ~ 2013년 8월 : KT 융합기술원 선임연구원
- 2013년 9월 ~ 현재 : 한림대학교 소프트웨어학부 교수

〈관심분야〉
사물인터넷, 무선센서네트워크, 무선전력전송, 해상무선통신, 머신러닝, 블록체인

김 동 완(Dongwan Kim) [정회원]

- 2003년 8월 : 고려대학교 전자공학과 (공학사)
- 2006년 2월 : 포항공과대학교 정보통신공학과 (공학석사)
- 2015년 2월 : 고려대학교 전기전자공학과 (공학박사)
- 2017년 3월 ~ 현재 : 동아대학교 전자공학과 부교수

〈관심분야〉
전기자동차 표준, Edge 컴퓨팅, Physical AI