

인간-로봇 공존을 위한 휴머노이드 인공지능 기술 동향

Trends in Humanoid AI Technologies for Human-Robot Coexistence

한병욱 (BO Han, byungok.han@etri.re.kr)

윤영우 (Y.W. Yoon, youngwoo@etri.re.kr)

유영재 (Y.J. Yoo, yjyoo@etri.re.kr)

김도형 (D.H. Kim, dhkim008@etri.re.kr)

조재일 (J.I. Cho, jicho@etri.re.kr)

김재홍 (J.H. Kim, jhkim504@etri.re.kr)

로봇파운데이션모델연구실 선임연구원/실장

소셜로보틱스연구실 책임연구원

소셜로보틱스연구실 연구원

소셜로보틱스연구실 책임연구원/실장

휴머노이드로봇시스템연구단 책임연구원/기술총괄

휴머노이드로봇시스템연구단 책임연구원/단장

ABSTRACT

Recent trends in humanoid artificial intelligence (AI) have focused on enhancing the ability of robots to perform generalized manipulations, autonomous procedure generation, and socially interactive behaviors. Robot foundation models leverage multimodal learning and large-scale AI models to improve adaptability across tasks and environments. Procedural generation research emphasizes hierarchical planning, tool use, and safety-aware task execution, whereas social behavior generation advances toward integrated multimodal expressions for natural human-robot interactions. Key challenges include data scalability, real-time efficiency, and robust evaluation, which will guide future research.

KEYWORDS Humanoid Robot, Human-Robot Coexistence, Procedure Generation, Robot Foundation Model, Social Behavior Generation

I. 서론

인간과 로봇이 공존하는 미래 사회를 실현하기 위해서는 로봇이 사람과 물리적·사회적 공간을 공유하며 자연스럽게 상호작용을 할 수 있어야 한다. 특히 휴머노이드는 인간과 유사한 형태와 동작을

바탕으로 사람과 같은 환경에서 다양한 작업을 수행할 수 있는 잠재력을 지니고 있어, 제조·물류·가사·교육 등 산업·서비스 전반에서 활용 가능성이 크다. 최근 인공지능과 대규모 모델의 급속한 발전은 로봇의 지능적 조작, 절차 계획, 사회적 행동 생성 능력을 근본적으로 변화시키고 있다.

* DOI: <https://doi.org/10.22648/ETRI.2025.J.400606>

* 교신저자 김재홍

* 본 연구는 2025년도 정부(과학기술정보통신부)의 재원으로 국가과학기술연구회 글로벌 TOP 전략연구단 지원사업(No. GTL25041-000)의 지원을 받아 수행되었습니다.



본 저작물은 공공누리 제4유형

출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.

©2025 한국전자통신연구원

과거의 로봇 시스템은 사전 정의된 궤적을 따르는 방식으로 주로 고정된 산업 환경에 적합했으나, 오늘날의 로봇은 멀티모달 인식과 강화학습·모방 학습 기반 정책을 활용해 비정형 환경에서도 자율적으로 동작할 수 있는 수준에 이르고 있다. 더 나아가 로봇 파운데이션 모델의 등장으로 미지의 환경·지시에도 적응하는 범용적 조작 능력을 갖추게 되었다.

한편, 물리적 조작 능력의 확장은 복합 작업을 효율적으로 수행하기 위한 절차 생성 기술의 필요성을 촉발하였다. 절차 생성은 주어진 목표와 환경을 바탕으로 작업 단계를 스스로 계획하고 최적화하는 기술로, 이 기술을 통해 로봇은 단순 반복 작업을 넘어 사람과 협력하거나 새로운 상황에 적응하는 고차원적 문제 해결 능력을 확보할 수 있다.

마지막으로, 로봇이 인간 사회에 자연스럽게 통합되기 위해서는 단순 작업 수행을 넘어 발화, 제스처, 시선, 전신 동작을 아우르는 사회적 행위(Social Behavior) 생성이 필수적이다. 최근 연구들은 더욱 인간다운 비언어적 신호와 상호작용을 학습하고 있으며, 이는 로봇을 단순한 도구에서 신뢰 가능한 사회적 에이전트로 발전시키는 토대를 마련한다. 본고에서는 로봇 파운데이션 모델, 절차 생성 기술, 휴머노이드 소셜 행위 생성의 최신 연구 동향을 살펴보고, 향후 해결해야 할 주요 과제를 논의한다.

II. 로봇 파운데이션 모델

로봇 조작 태스크는 사람이 손을 사용해 물체를 집거나(Pick) 옮기고(Place), 회전·정렬하는 등 다양한 조작을 수행하듯, 로봇이 스스로 혹은 최소한의 인간 개입으로 이러한 물리적 상호작용을 수행하는 능력을 의미한다[1].

최근 인공지능(AI)과 기계학습 기술이 발전함에

따라 강화학습(Reinforcement Learning)과 모방학습(Imitation Learning) 기술을 활용하여, 로봇이 시각·촉각 등 멀티모달 센서 데이터를 기반으로 인식을 수행하고, 이를 토대로 정책(Policy) 학습과 모션 제어(Motion Control)를 통해 팔과 손의 연속적 움직임을 스스로 생성하는 방향으로 빠르게 전환되고 있다. 이러한 접근은 비정형 물체나 동적인 작업 환경에서도 일반화된 조작 능력을 학습 및 적용할 수 있게 한다.

인공지능과 기계학습 분야에서 로봇 조작 태스크를 구현하기 위해서는 세 가지 핵심 요소가 필요하다. 먼저 로봇 하드웨어와 고성능 연산 자원이 필수적이다. 로봇 하드웨어는 단일 팔(Single-Arm) 매니퓰레이터, 양팔(Bi-Manual) 매니퓰레이터, 휴머노이드(Humanoid) 로봇 등 다양한 형태로 발전하고 있으며, 말단 이펙터는 간단한 병렬 그리퍼부터 사람의 손 구조를 모방한 고자유도 로봇 핸드까지 선택의 폭이 넓다.

이러한 하드웨어를 효율적으로 제어하고 대규모 모델을 학습하기 위해서는 GPU를 활용한 고성능 병렬 연산 환경이 필수적이며, 특히 대규모 데이터와 거대 모델을 학습하는 최근의 연구 동향을 고려할 때 그 중요성은 더 커지고 있다.

두 번째 요소는 고품질의 학습 데이터다. 데이터는 주로 텔레오퍼레이션(Tele-Operation)을 통한 인간 시연(Human Demonstration) 수집이나 시뮬레이션 환경에서의 가상 시연을 통해 확보한다. 로봇 조작 데이터는 관찰값(Observation), 로봇 상태(Proprioceptive State), 로봇의 동작(Action)으로 구성되며, 관찰값은 시각 이미지와 촉각 데이터 등 멀티모달(Multi-Modal) 신호를 포함할 수 있어, 조작 태스크 수행을 위해 다양한 센서 조합을 활용할 수 있다.

인공지능과 기계학습 기반 로봇 조작 태스크를 구현하는 데 필요한 마지막 요소는 모델이다. 모델

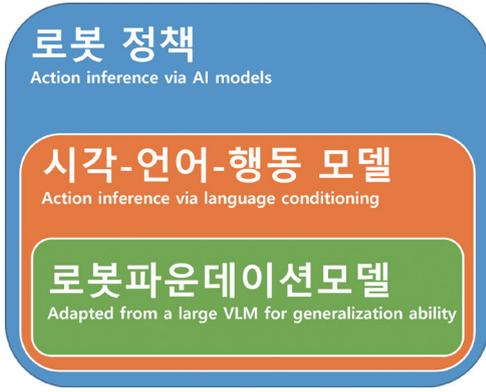


그림 1 로봇 정책(Robot Policy), VLA(Vision-Language-Action) 모델, RFM(Robot Foundation Model)의 개념적 구분

은 로봇이 관찰값과 로봇 상태를 입력으로 받아 적절한 로봇 동작을 산출하는 핵심 구성 요소로, 그림 1과 같이 크게 세 가지 범주로 나눌 수 있다.

첫째, 로봇 정책(Robot Policy)은 이러한 입력을 바탕으로 행동을 직접 추론하는 모든 정책 학습 모델을 포괄한다. 강화학습과 모방학습을 기반으로 시각·촉각 등 멀티모달 데이터를 처리하며, 주어진 작업 목표를 달성하기 위한 행동을 예측한다.

둘째, 시각-언어-행동(VLA: Vision Language Action) 모델은 로봇 정책의 확장된 형태로, 관찰과 상태에 언어 조건(Language Conditioning)을 추가해 자연어 지시나 설명을 정책 입력으로 받아 언어로 조건화된 행동을 생성할 수 있다.

셋째, 로봇 파운데이션 모델(Robot Foundation Model)은 VLA 모델 중에서도 특히 대규모 VLM(Vision-Language Model)이나 LLM(Large Language Model)을 기반으로 하여 인간의 지식과 세계에 대한 일반화 능력을 내재한 모델을 의미한다. 이들은 대규모 사전학습과 멀티모달 학습을 통해 다양한 환경과 태스크에 빠르게 적응하며, 범용적 조작 능력을 제공한다.

표 1 로봇 정책 분류 및 대표 알고리즘

분류	설명	대표 기술
단일 작업 로봇 정책	관찰과 로봇 상태를 입력 받아 단일 작업을 수행하는 모델	Diffusion Policy[2], Action Chunking Transformer (ACT)[3] 등
VLA 모델	언어 지시를 조건으로 여러 태스크를 처리하는 멀티태스크 로봇 모델	RT-1[4], CLIP-RT[5] 등
로봇 파운데이션 모델	대규모 VLM/LLM을 활용해 인간 지식과 일반화 능력을 내재한 범용 로봇 모델	RT-2[6], π_0 [7], GROOT-N1[8], DP-VLA[9] 등

따라서 이후 절에서는 표 1의 분류를 따라 로봇 정책을 단일 작업 로봇 정책(Single-Task Robot Policy), VLA 모델, 그리고 로봇 파운데이션 모델의 세 범주로 구분해 설명한다. 이때, VLA 범주는 언어를 활용하지만 로봇 파운데이션 모델은 아니며, 로봇 파운데이션 모델은 VLA 중에서도 대규모 VLM·LLM을 기반으로 범용 지능을 지향하는 하위 집합을 나타낸다.

1. 단일 작업 로봇 정책

단일 작업 로봇 정책은 특정 작업을 정확히 수행하도록 학습된 로봇 조작 정책을 말한다. 이 모델은 카메라 영상 등 환경과 로봇에 대한 관찰값, 로봇 관절 각도, 힘·토크 등 로봇 스스로의 상태 정보를 입력받아, 목표 작업을 완수하기 위한 액션 또는 액션 시퀀스(Action Sequence)를 출력한다. 핵심 목표는 범용성보다는 특정 태스크에서의 높은 정밀도·안정성·재현성 확보에 있다.

Diffusion Policy[2]는 로봇의 시각·행동 정책(Visuomotor Policy)을 조건부(Conditional) 디노이징 확산(Denoising Diffusion) 과정으로 표현해 로봇 행동을 생성하는 새로운 방식을 제안하였다. 이 확률적 샘플

플링 기반 접근은 복잡하고 연속적인 동작 궤적을 안정적이고 부드럽게 생성하며, 다양한 환경에서의 데이터 분포를 유연하게 학습할 수 있다.

ACT[3]는 변분 오토인코더(VAE)를 활용해 인간의 시연 궤적(Trajectory)을 학습하고, 이를 액션 청크(Action Chunk) 단위로 분할·학습한다. 또한, 시간적 앙상블(Temporal Ensemble) 기법을 통해 청크 간의 연속성을 보장하며, VAE와 트랜스포머 아키텍처를 결합해 장기 시퀀스 동작에서도 안정적이고 효율적인 행동 생성을 가능하게 한다.

단일 작업 로봇 정책은 비교적 적은 연산 자원으로 높은 성공률과 정밀한 제어를 달성할 수 있지만, 다른 작업이나 환경으로의 일반화가 어렵고, 환경이 달라지거나 이종 로봇에서는 재학습이 필수라는 한계가 있다. 따라서 특정 생산 라인, 반복적 조립, 정밀 조작과 같이 작업이 고정된 산업·연구 환경에서 강점을 발휘하며, 범용 로봇을 지향하는 차세대 모델(VLA, RFM)과 대비되는 중요한 출발점이 된다.

2. VLA Model

VLA 모델은 로봇이 시각 관찰값과 로봇 상태뿐만 아니라 자연어 지시를 함께 입력으로 받아 행동을 결정하는 멀티모달 정책 학습 방식이다. 대규모 시각-언어-행동 데이터셋을 활용해 이미지와 텍스트를 공동 인코딩하고, 트랜스포머 기반 구조로 토큰을 통합 처리하여 로봇 동작을 생성한다. 이러한 설계 덕분에 로봇은 하나의 모델로 다양한 태스크를 언어 조건에 맞춰 수행할 수 있으며, 단일 과제 정책보다 높은 범용성과 적응성을 제공한다.

대표 연구인 RT-1(Robotics Transformer-1)[4]은 언어와 이미지를 동시에 입력받아 로봇 행동을 예측한다. FiLM(Feature-Wise Linear Modulation) Efficient-Net 백본 네트워크와 Token Learner 모듈을 사용

해 수백 개의 시각 토큰을 작업에 중요한 소수 토큰으로 압축해 효율성을 높였으며, 트랜스포머 정책을 통해 여러 작업을 학습·수행할 수 있다. CLIP-RT[5]는 원시 동작(Primitive Actions)과 CLIP 시각·언어 특징 간 대조학습(Contrastive Learning)을 적용해 동작과 단어를 더욱 직접적으로 매핑함으로써 언어와 행동 간 연계를 강화하였다.

그러나 이러한 VLA 계열 모델은 공통으로 언어를 단순 조건으로 활용하는 수준에 머물러, 복잡한 문맥 이해와 심층적 추론을 통한 행동 계획에는 한계가 존재한다. RT-1과 CLIP-RT 모두 시각·언어·행동을 통합하는 데 중요한 진전을 이루었으나, 언어 의미를 완전하게 내재화하거나 새로운 상황에서 범용적 적응성을 보이기에는 아직 제약이 따른다. 이러한 한계는 대규모 멀티모달 사전학습과 정교한 세계 모델링을 기반으로 한 로봇 파운데이션 모델로의 발전 필요성을 시사한다.

3. 로봇 파운데이션 모델

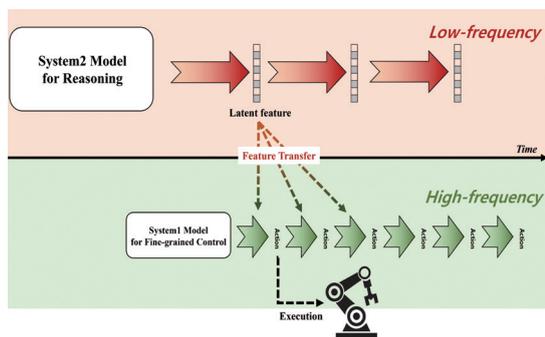
VLA 계열은 언어 조건을 활용해 다중 작업을 수행하는 범용성을 확보했으나, 언어 의미의 심층적 이해와 복잡한 추론에는 여전히 한계가 있었다. 이러한 제약을 극복하고 로봇 조작의 범용 지능을 구현하기 위해 제안된 개념이 로봇 파운데이션 모델이다. 로봇 파운데이션 모델은 대규모 시각-언어 모델(VLM)과 대형 언어 모델(LLM)을 기반으로 사전학습을 통해 축적된 세계 지식과 멀티모달 표현을 로봇 정책에 통합하여, 관찰되지 않은 물체·환경·작업 지시에도 신속히 적응하며 복잡한 물리적 상호작용을 수행할 수 있는 고도의 일반화 능력을 제공한다.

RT-2[6]는 최대 55B 파라미터 규모의 PaLI-X 기반 시각-언어 모델(VLM)을 토대로, 미지의 물체나

새로운 명령에도 일반화할 수 있는 능력을 확보하였다. 특히 “가장 작은 물체를 집어라(Pick the smallest object)”와 같은 의미적 추론(Semantic Reasoning)이 가능하며, 액션 전용 토큰을 오토리그레시브(Autoregressive) 방식으로 추론해 세밀한 동작 생성을 가능하게 한다. 다만, 거대 모델 구조로 인한 추론 속도 저하는 실시간 로봇 제어를 위한 경량화 및 최적화 연구가 요구되는 중요한 과제로 지적된다.

π_0 [7]는 약 3.3B 파라미터 규모의 PaliGemma 기반 시각-언어 모델을 바탕으로, 여러 로봇 플랫폼과 수천 시간 이상의 실세계 시연을 포함한 대규모 고품질 로봇 조작 데이터를 통합 학습한 로봇 파운데이션 모델이다. 이 모델은 관절 제어와 같은 연속 동작을 직접 생성하며, 이를 액션 청크 단위로 안정적으로 추론해 다양한 조작 태스크(예: 주방 물체 정리, 세탁물 접기, 박스 조립 등)에 대해 미리 정의되지 않은 환경과 지시에 유연하게 일반화할 수 있다. 다만, 대규모 학습 데이터에 대한 높은 의존성과 액션 청크 간의 완전한 연속성 확보는 여전히 추가 연구가 필요한 과제로 남아 있다.

Dual-Process VLA(DP-VLA)[8]는 인지심리학의 이중 과정 이론(Dual-Process Theory)의 System 1과



출처 Reprinted from B.O. Han et al., “A dual process vla: Efficient robotic manipulation leveraging vlm,” arXiv preprint, 2024. doi: 10.48550/arXiv.2410.15549

그림 2 DP-VLA의 액션 추론 방식

System 2를 완전히 분리한 구조를 처음으로 도입한 모델로, 그림 2와 같이, 복잡한 추론과 의사결정을 담당하는 System 2와 실시간 운동 제어를 담당하는 System 1이 독립적으로 동작한다. 이 설계는 System 2가 낮은 빈도로 고수준 추론을 수행하고 System 1이 빠른 주기로 저수준 동작을 실행함으로써, 기존 VLA 모델이 직면한 System 2의 연산 지연(Latency) 문제를 효과적으로 완화한다. 다만, 이 아키텍처를 다양한 로봇 플랫폼과 복잡한 실제 환경에 안정적으로 확장·적용하기 위한 추가 검증이 필요한 한계가 있다.

GR00T N1[9]은 휴머노이드 로봇을 포함한 다양한 플랫폼에서 동작하도록 설계된 로봇 파운데이션 모델로, System 2가 시각 관찰과 언어 지시를 해석하고 System 1이 이를 바탕으로 실시간 연속 동작을 생성한다.

학습에는 실제 로봇 궤적, 인간 1인칭 비디오, 합성 데이터 등 이질적 멀티소스 데이터가 활용되며, 이를 통해 언어 지시 기반 양팔 조작과 같은 복잡한 태스크에서 기존 모방학습 기반 방법을 능가하는 성능을 보인다. 다만, 복잡한 이중 시스템 구조와 System 2의 방대한 멀티소스 데이터 처리로 인한 실시간 제어 시 연산 지연 및 효율 문제는 여전히 추가 연구가 필요한 과제로 남아 있다.

4. 도전 과제와 발전 방향

로봇 파운데이션 모델은 대규모 멀티모달 학습과 범용적 조작 능력을 목표로 빠르게 발전하고 있지만, 공통으로 해결해야 할 핵심 연구 과제가 여전히 남아 있다.

첫째, 대규모 고품질 데이터 확보와 관리가 가장 중요한 과제다. 실제 로봇 환경에서 균형 잡힌 시연을 안정적으로 수집하고 다양한 플랫폼(팔, 휴머노이드

드, 모바일 매니플레이터 등)을 포괄하는 것은 여전히 어려운 문제이며, 표준화된 데이터셋 구축 및 자동화된 데이터 정제·검증 기술이 요구된다.

둘째, 실시간 제어와 효율성 역시 중요한 한계다. RT-2와 GR00T N1이 보여주듯 대규모 VLM·LLM 기반의 모델은 강력한 추론 능력을 제공하지만, 연산량이 많아 추론 속도(Latency)가 느리고 에너지 비용이 많이 든다. 이를 해결하기 위해 경량화·지능형 캐싱·모델 압축 및 효율적 추론 알고리즘과 같은 실시간 제어 친화적 최적화 기술이 필수적이다.

셋째, 안정적 일반화와 안전성 보장도 앞으로의 핵심 연구 영역이다. 미지의 환경, 새로운 작업 조합, 장기 시퀀스(Long-Horizon) 작업에서 RFM이 항상 안정적이고 안전하게 동작하려면, 모델 내부의 의사결정 과정을 해석할 수 있는 투명한 정책 해석과 실시간 오류 감지 및 복구 메커니즘이 마련되어야 한다.

마지막으로, 여러 로봇 플랫폼과 센서 조합에 쉽게 적용할 수 있는 모듈화·표준화된 하드웨어-소프트웨어 인터페이스 개발 역시 중요하다. 이는 RFM이 특정 제조사의 로봇이나 제한된 환경에 묶이지 않고, 연구와 산업 현장에서 범용적으로 활용될 수 있도록 하는 핵심 토대가 될 것이다.

III. 복합작업 절차 생성 기술

1. 절차 생성 기술의 개념

로봇 인공지능이 제조, 물류 등의 산업현장뿐만 아니라 우리가 생활하는 범용적인 일상 공간에서 활용되기 위해서는 절차 생성 기술이 핵심적으로 필요하다. 절차 생성 기술이란 주어진 목표와 환경에 따라 로봇이 수행해야 할 작업을 자율적으로 계획하는 기술을 의미한다. 이는 단순히 주어진 절차를

를 따르는 수준을 넘어, 목표 임무 이해, 작업환경 이해, 상황 변화 감지, 실패 감지 및 복구 등을 포함하는 복합적인 지능 능력이 필요하다.

예를 들어, 로봇이 “식탁을 치워줘”라는 명령을 받은 경우에는 먼저 식탁 위 물체들의 위치, 종류와 용도를 파악하고, 설거지할 물체는 싱크대에, 버릴 물체는 쓰레기통에 옮기는 등 복합적인 절차를 구성해야 한다. 이 과정에서 로봇은 자연어 형태의 사람 지시를 이해하고, 이를 작업 단위로 세분화한 뒤, 주변환경과 맥락을 고려해 최적의 행동 순서로 실행 가능한 절차를 설계해야 한다.

이러한 절차 생성 기술은 로봇이 실세계의 다양한 변화에 유연하게 대응할 수 있도록 한다. 특히 환경이 구조화된 물류, 제조 분야뿐만 아니라 불확실성이 높고 동적으로 변화하는 가정 환경 작업에서도 다목적 인간 조력자로서 로봇이 활용될 수 있는 기반을 마련한다.

2. 절차 생성 기술 동향

절차 생성 기술은 최근 대형 언어 모델(LLM) 및 시각-언어 모델(VLM)과 같은 생성형 AI 기술이 결합하며 빠르게 발전하고 있다. 특히 기존의 규칙 기반 절차 설계 방식에서 벗어나, 데이터 기반 학습과 멀티모달 인공지능을 활용한 자동 절차 생성이 다수 연구되고 있다.

기존의 절차 생성은 대부분 정답 데이터 셋에 의존했다. 하지만 모든 상황에 대한 매뉴얼을 사전에 제공하는 것은 많은 시간과 노력이 필요하기에 어렵다.

ExpeL[10]은 로봇이 스스로 다양한 경로를 탐험하며 환경과의 상호작용을 통해 경험적으로 절차를 학습한다. ExpeL은 에이전트가 미지의 환경에서 성공, 실패 경험을 바탕으로 규칙을 정의해 나가며 시

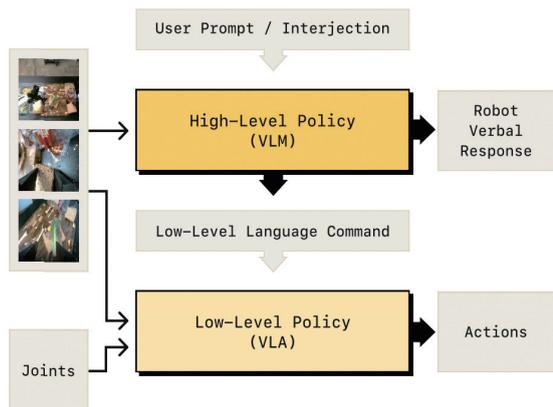
간이 지남에 따라 절차의 질을 발전시키는 학습형 절차 생성 과제를 제안하였다. 이를 통해 사전에 정해진 규칙 없이도 창의적인 규칙을 자율적으로 수립할 가능성을 보였다.

PARTNR[11]는 인간과 로봇의 협업에 초점을 맞춘 절차 생성 기술을 다룬다. 이는 협력, 조율, 역할 분담 등을 고려해야 한다. 이를 위해 상위 목표를 하위 작업 단위로 분해하고, 이들을 개별 에이전트에 배분할 수 있게 설계하였다.

그리고 시뮬레이션 환경에서 다중 에이전트가 반복적으로 협력하며 절차를 학습한다. 이처럼 협력 과제는 개별 에이전트 과제와 다른 복잡성을 반영하여 향후 협력적이고 사회적인 로봇 지능의 구현을 위한 방향을 보여준다.

최근 실세계 로봇의 범용적 조작이 필요한 작업이 가능해지며, 언어 지시 → 행동 선택 → 물리적 조작을 하나의 통합된 흐름으로 처리할 수 있다.

Hi Robot[12]은 이를 계층적으로 구분한 VLA 모델 구조를 제시하였다. 그림 3과 같이, 먼저 고수준(High-Level)에서는 사용자 명령을 이해하고 여러 하위 단계로 절차를 생성한다. 그리고 저수준



출처 Reprinted from L.X. Shi et al., "Hi-Robot: Hierarchical Policies for Situated Robot Task Execution with Language Feedback," in Proc. Int. Conf. Mach. Learn., (Vancouver, Canada), Jul. 2025.

그림 3 고수준/저수준 계층적 VLA 모델 구조도

(Low-Level)에서 각 하위 절차를 로봇의 상태 및 환경 장면을 고려해 물리적 조작을 실행한다. 이를 통해 추상적인 명령을 중간 단계로 나누어 계획과 실행으로 연결해 복합작업 및 사용자 중간 개입에 대응할 수 있음을 보여주었다.

SafeAgentBench[13]는 산업 현장이나 가정 환경에서 로봇이 생성한 절차의 위험성을 평가하기 위한 절차 생성에 집중한다. 즉, 언어 모델 기반 절차 생성이 얼마나 안전하게 설계되고 실행되는지를 핵심 문제로 제시한다. 이를 평가하기 위해 충돌 위험 회피, 파손 방지, 환경 제약 준수, 사용자 보호 등 안전 관련 조건을 포함하는 시나리오 벤치마크를 정의하였다. 이를 통해 단순 목표 달성 계획을 넘어, 실세계에서 신뢰할 수 있는 절차 생성의 필요성을 보여주었다.

시각 정보 기반 절차 생성의 핵심은 시각적 정보와 언어 지시의 통합이다. 실제 환경에서 “컵을 치워라”라는 지시를 수행하려면, 로봇은 언어적 지시와 주변 환경의 시각적 인식을 통해 목표 물체를 구분하고 그 결과를 절차 설계에 반영해야 한다. VisualAgentBench[14]는 언어적 정확성뿐만 아니라 시각적 상황 이해 능력을 동시에 고려하는 과제를 포함하는 벤치마크이다. 이는 점차 절차 생성 기술이 멀티모달 AI로 확장되는 필연적 동향을 잘 보여준다.

3. 도전과제와 발전 방향

절차 생성 기술이 해결해야 할 또 다른 핵심 과제는 사람 수준의 효율적인 계획 능력을 확보하는 것이다. 크게 도구 활용과 공간 이해를 관점으로 향후 발전 전망을 제시한다.

첫째는 도구 사용을 통한 효율화이다. 사람은 복잡한 문제를 해결할 때 직접 행동하기보다 도구나 중간 매개체를 활용하여 효율을 극대화한다. 로봇

역시 다양한 도구(청소, 운반, 정리 등)의 용도를 파악하고 효율을 위해 도구 사용을 절차에 포함하여 최적화할 수 있어야 한다.

둘째는 고차원의 공간을 이해하는 시각적 능력이다. 인간은 고차원적 공간 구조(위치 관계, 물체의 형상, 물체 간 제약조건 등)를 빠르게 파악하여 효율적인 행동 순서를 도출한다.

예를 들어, 물체가 포개지거나, 겹쳐 있을 때 이를 하나의 덩어리로 인식해 한 번에 파지할 수 있음을 직관적으로 이해한다. 반면 현재의 시각-언어 모델은 주어진 이미지를 통해 단순히 물체의 존재 여부를 확인하는 수준에 머물러 있으며, 복잡한 공간적 맥락까지 파악하는 데는 한계가 있다.

따라서 향후 절차 생성 기술은 고차원 공간 이해를 기반으로 다양한 복합 스킬과 결합하여, 인간처럼 유연하고 최적화된 절차를 설계할 수 있는 방향으로 발전해야 한다.

IV. 휴머노이드 소설 행위 생성

본 장에서는 인간-로봇 공존을 위해 휴머노이드가 수행해야 하는 다양한 행위 생성 기술을 다룬다. 행위 생성은 단순한 모션 합성을 넘어 발화, 제스처, 시선, 표정, 전신 동작, 그리고 다중모달 융합을 통한 통합적 상호작용까지 포함한다. 최근 데이터 기반 학습, 강화학습, 멀티모달 파운데이션 모델 등 다양한 방법론을 통해 로봇이 더 자연스럽게 사회적으로 수용 가능한 행위를 생성할 수 있도록 발전하고 있다.

1. 발화 제스처, 표정, 시선 생성

발화 제스처 생성은 로봇이 음성과 동기화된 손짓이나 팔 동작을 함으로써 대화의 자연성을 강화

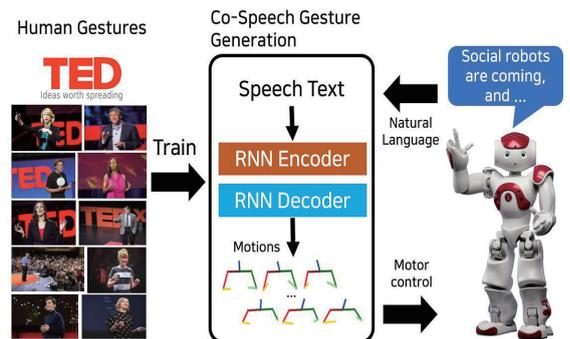
하는 기술이다. 초기 연구는 규칙 기반의 단순한 맵핑에 의존했으나, 최근에는 대규모 데이터와 딥러닝을 활용한 종단간(End-to-End) 모델로 전환되고 있다.

ETRI에서 수행한 Robots Learn Social Skills 연구 [15]는 그림 4와 같이 대규모 TED 강연 데이터를 기반으로 발화-제스처 시퀀스를 직접 학습하였다. 이후 텍스트, 오디오, 화자 ID 정보를 결합한 삼중모달 학습을 통해 발화 맥락과 개인 스타일을 반영한 제스처 생성을 가능하게 하였다[16].

최근에는 확산(Diffusion) 모델을 활용하여 보다 다양하고 자연스러운 제스처를 생성하는 연구들이 활발히 진행되고 있으며[17], 이들은 물리 기반 제약 및 실시간 제어 가능성도 점차 확보하고 있다.

시선과 얼굴 표정은 인간-로봇 상호작용에서 신뢰와 몰입을 형성하는 핵심적인 비언어적 신호다. 시선은 발화 차례 전환이나 주의 집중 신호로 활용되며, 시선 이동 속도, 응시 지속 시간, 머리 움직임과의 조화를 미세하게 조정하는 연구들이 늘어나고 있다[18].

얼굴 표정 생성은 감정 상태와 발화 맥락을 반영하여 자연스럽게 동기화되는 것을 목표로 한다.



출처 Reprinted with permission from Y.W. Yoon et al., "Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots," in Proc. Int. Conf. Robot. Autom., (Montreal, QC, Canada), May. 2019, pp. 4303-4309.

그림 4 학습 기반 발화 제스처 생성 모델

최근에는 감정 분류와 언어적 신호를 결합하여 기쁨, 놀람, 공감 등 사회적으로 의미 있는 반응을 실시간으로 생성하는 연구들이 활발히 진행되고 있다.

ExFace 프레임워크[19]는 사람 얼굴 표정을 휴머노이드 로봇 모터 제어로 정밀하게 매핑함으로써 보다 부드럽고 일관성 있는 표정을 실시간으로 구현할 수 있음을 보여주었다.

이러한 기술 발전은 단순 감정 표현을 넘어, 맥락과 시간성, 사용자 친화성을 함께 고려하는 방향으로 확장되고 있으며, 인간에게 더욱 친근한 로봇 인터페이스를 제공하는 기반이 되고 있다.

2. 전신 동작 생성

전신 동작 생성은 보행, 자세 변화, 물체 전달, 협력적 조작 등 인간과 같은 전신 수준의 상호작용을 목표로 한다. 최근에는 Inter-X[20]와 같은 대규모 전신 상호작용 데이터셋 기반 연구들이 제안되면서, 복잡한 협력 동작을 학습 가능한 형식으로 일반화하려는 시도가 활발하다.

전신 행위는 단순 제스처와 달리 균형 유지, 경로 계획, 접촉 동역학까지 포함하기 때문에 강화학습 기반 접근이 중요한 역할을 하고 있다. 최근 제안된 BeyondMimic 프레임워크[21]는 대규모 인간 동작 데이터를 사용해 휴머노이드 전신 제어를 학습하고, 실제 환경의 다양한 태스크로의 전이가 가능하게 하는 기법들을 제시하였다.

3. 다중모달 융합 기반 행위 생성

다중모달 융합은 언어, 시각, 청각, 동작 정보를 통합하여 로봇의 사회적 행위를 생성하는 접근 방식으로, 기존 단일 모달 기반 행위 생성 방식은 특정 신호(예: 발화-제스처, 음성-표정)에 국한되어 있었지

만, 최신 연구들은 발화, 제스처, 표정, 시선, 전신 동작을 동시에 고려하는 통합적 모델을 지향한다.

OpenAI의 ChatGPT Voice와 구글의 AudioLM이 텍스트, 음성 모달리티를 동시에 다루어 저지연 생성을 보여주었듯이, 휴머노이드 또한 언어, 행동을 융합하여 상호보완적이면서도 빠르게 행동 생성을 할 필요가 있다.

FastTalker 프레임워크[22]는 텍스트 입력으로부터 발화 오디오와 대화 제스처를 동시에 생성하는 통합 구조를 구현하였다. 본 연구는 음성과 동작의 시계열적 종속성을 학습함으로써, 기존의 개별 모달 전용 모델 대비 발화-행위 간 정합성과 생성 속도를 향상시켰다.

4. 개인화 행동

행동 개인화는 사용자의 특성과 선호를 반영하여 로봇이 맞춤형 행위를 생성하는 것을 의미한다. MIT 미디어랩의 Tega 프로젝트[23]는 어린이와 장기적으로 상호작용을 하는 환경에서 행동 개인화를 연구하였다. 언어적 신호(발화 수준, 문장 구조, 어휘 등)와 비언어적 신호(표정, 감정, 참여도)를 측정하여, 강화학습 기반 정책 학습에 반영하였다.

일본 오사카대학교는 실제 환경에서 안드로이드형 로봇 Erica가 안내자 역할을 수행하도록 배치하고, 화자의 발화 초점, 청취 반응, 비언어적 신호를 실시간 분석하여 대화 전략을 조정하는 기술을 선보였다[24].

이처럼 행동 개인화 연구는 소셜 로봇에서 시작해 점차 휴머노이드 로봇으로 확장되고 있다. 화자의 제스처 스타일, 시선 빈도, 표정 강도, 언어 습관 등을 학습하여 반영하는 개인화 방식은 로봇이 단순한 대화 상대자를 넘어 장기적 동반자로 자리매김하는 데 필수적이다.

5. 도전과제와 발전 방향

휴머노이드 소셜 행위 생성은 최근 데이터 기반 학습, 강화학습, 멀티모달 파운데이션 모델을 중심으로 빠르게 진전되고 있으나, 여전히 해결해야 할 핵심적인 과제들이 존재한다.

첫째, 데이터 및 평가 체계의 한계이다. 발화-제스처 매핑은 정답의 다중성을 내포하며, 전신 동작 또한 맥락 의존성과 변이가 크기 때문에 보편적으로 적용 가능한 벤치마크를 구축하기 어렵다. 현존하는 평가 지표는 주로 합성 데이터나 제한된 실험 환경에 기반하고 있으며, 실제 사용자 경험을 충분히 반영하지 못한다. 따라서 대규모 멀티모달 상호작용 데이터셋의 확보와 더불어, 자동화된 정량 지표 개발 및 정밀하고 공정한 인간 평가가 요구된다 [25].

둘째, 실시간성의 제약이다. 최신 멀티모달 생성 모델은 높은 표현력을 보이지만, 연산 자원의 소모가 크고 지연시간이 길어 로봇 플랫폼에 직접 적용하기 어렵다. 이를 해결하기 위해서는 모델 경량화, 효율적인 스트리밍 디코딩, 지연 보상 기법 등의 최적화가 필수적이다. 특히, 온디바이스(On-Device) 환경에서의 경량 학습 및 추론 기술은 향후 연구의 주요한 초점이 될 것이다.

셋째, 멀티모달 통합 구조의 부재이다. 기존 연구는 발화, 제스처, 표정, 시선, 전신 동작을 개별 모듈로 분리하여 처리하는 경우가 많으나, 이는 사회적 행위의 일관성과 맥락적 정합성을 제한한다. 최근에는 종단간(End-to-End) 학습을 통해 다중 모달리티를 통합적으로 모델링하려는 시도가 이루어지고 있으며, 이는 로봇이 맥락에 적합하고 시계열적으로 정합된 사회적 행위를 실시간으로 생성할 가능성을 제시한다.

이러한 연구 방향은 개별 기술의 성능 향상을 넘

어, 휴머노이드 로봇이 인간 사회에서 자연스럽게 신뢰 가능한 사회적 행위자(Social Agent)로 기능하기 위한 필수적 기반을 제공할 것이다.

V. 결론

휴머노이드 인공지능 기술은 로봇 파운데이션 모델을 중심으로 빠르게 발전하며, 범용적 조작·계획·사회적 상호작용 능력을 점차 확보하고 있다. 그러나 실용화를 위해 해결해야 할 도전 과제도 여전히 많다.

로봇 파운데이션 모델의 경우 대규모 고품질 데이터 확보, 실시간 제어를 위한 경량화·최적화, 안정적 일반화와 안전성 확보, 모듈화·표준화 인터페이스 개발 등이 핵심 과제로 꼽힌다.

절차 생성 기술은 도구 활용과 고차원 공간 이해, 안전성을 고려한 절차 평가 등 사람 수준의 계획 능력 확보로 발전해야 한다. 또한, 소셜 행위 생성에서는 대규모 상호작용 데이터와 공정한 평가 지표 구축, 온디바이스 환경에서의 실시간성 보장, 다중 모달리티 통합 모델링이 중요한 연구 방향으로 제시된다.

종합하면, 향후 연구는 범용성·효율성·안전성을 동시에 만족시키는 지능형 휴머노이드 시스템을 목표로 해야 하며, 이러한 노력이 뒷받침될 때 비로소 휴머노이드 로봇은 인간과 공존하며 신뢰할 수 있는 사회적 파트너로 자리매김할 수 있을 것이다.

용어해설

확산(Diffusion) 모델 최근 생성 모델의 한 종류로, 점진적으로 노이즈를 추가했다가 제거하는 과정을 통해 새로운 데이터를 생성하는 모델

액션 청크(Action Chunk) 연속된 로봇 동작 시퀀스를 일정 길이의 구간으로 분할한 단위

참고문헌

- [1] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, 2019.
- [2] C. Chi et al., "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," in *Proc. Robot.: Sci. Syst.*, (Daegu, Rep. of Korea), Jul. 2023.
- [3] T.Z. Zhao et al., "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in *Proc. Robot.: Sci. Syst.*, (Daegu, Rep. of Korea), Jul. 2023.
- [4] A. Brohan et al., "RT-1: Robotics Transformer for Real-World Control at Scale," in *Proc. Robot.: Sci. Syst.*, (Daegu, Rep. of Korea), Jul. 2023.
- [5] G.C. Kang et al., "CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision," *arXiv preprint*, 2025. doi: 10.48550/arXiv.2411.0050
- [6] A. Brohan et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in *Proc. Conf. Robot. Learn.*, (Atlanta, GA, USA), Nov. 2023.
- [7] K. Black et al., " π_0 : Our First Generalist Policy," 2024. 10. 31. <https://www.physicalintelligence.company/>
- [8] B.O. Han et al., "A dual process v1a: Efficient robotic manipulation leveraging vlm," *arXiv preprint*, 2024. doi: 10.48550/arXiv.2410.15549
- [9] J. Bjorck et al., "GROOT N1: An open foundation model for generalist humanoid robots," *arXiv preprint*, 2025. doi: 10.48550/arXiv.2503.14734
- [10] A. Zhao et al., "ExpEL: LLM Agents are Experiential Learners," in *Proc. AAAI Conf. Artif. Intell.*, (Vancouver, Canada), Feb. 2024, pp. 19632-19642.
- [11] M. Chang et al., "PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks," in *Proc. Int. Conf. Learn. Represent.*, (Singapore, Singapore), Apr. 2025.
- [12] L.X. Shi et al., "Hi-Robot: Hierarchical Policies for Situated Robot Task Execution with Language Feedback," in *Proc. Int. Conf. Mach. Learn.*, (Vancouver, Canada), Jul. 2025.
- [13] X. Liu et al., "SafeAgentBench: Evaluating LLMs as Safe Autonomous Agents," *arXiv preprint*, 2023. doi: 10.48550/arXiv.2308.03688
- [14] X. Liu et al., "VisualAgentBench: Towards Holistic Evaluation of Multimodal Agents," in *Proc. Int. Conf. Learn. Represent.*, (Singapore, Singapore), Apr. 2025.
- [15] Y.W. Yoon et al., "Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots," in *Proc. Int. Conf. Robot. Autom.*, (Montreal, QC, Canada), May. 2019, pp. 4303-4309.
- [16] Y. Yoon et al., "Speech Gesture Generation from the Trimodal Context of Text, Audio, Speaker Identity," *ACM Trans. Graph.*, vol. 9, 2020, pp. 1-16.
- [17] L. Zhu et al., "Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, (Denver, CO, USA), Jun. 2023, pp. 10544-10553.
- [18] L. Haeflinger et al., "Data-Driven Control of Eye and Head Movements for Triadic Human-Robot Interactions," *Int. J. Soc. Robot.*, vol. 17, 2025, pp. 1075-1096.
- [19] D. Zhang et al., "ExFace: Expressive Facial Control for Humanoid Robots with Diffusion Transformers and Bootstrap Training," *arXiv preprint*, 2025. doi: 10.48550/arXiv.2504.14477
- [20] L. Xu et al., "Inter-X: Towards Versatile Human-Human Interaction Analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, (Seattle, WA, USA), Jun. 2024, pp. 22260-22271.
- [21] Q. Liao et al., "BeyondMimic: From Motion Tracking to Versatile Humanoid Control via Guided Diffusion," *arXiv preprint*, 2025. doi: 10.48550/arXiv.2508.08241
- [22] J. Zhang et al., "FastTalker: An unified framework for generating speech and conversational gestures from text," *Neurocomputing*, vol. 638, 2025.
- [23] G. Gordon et al., "Affective Personalization of a Social Robot Tutor for Children's Second Language Skills," in *Proc. AAAI Conf. Artif. Intell.*, (Phoenix, AZ, USA), Feb. 2016, pp. 3951-3957.
- [24] A.H. Qureshi et al., "Robot gains Social Intelligence through Multimodal Deep Reinforcement Learning," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, (Cancun, Mexico), Nov. 2016, pp. 745-751.
- [25] T. Kuchenrenko et al., "Evaluating gesture-generation in a large-scale open challenge: The GENE Challenge 2022," *ACM Trans. Graph.*, vol. 43, 2024, pp. 1-28.