전자통신동향분석, 41권 6호 2026, 15-28

VLM/VLA 기반 피지컬 AI 로봇 행동·이동지능 기술 동향

Technology Trends in Action and Navigation Intelligence for VLM/VLA-Based Physical AI Robots

장지호 (J. Chang, changjh@etri.re.kr) 필드로보틱스연구실 책임연구원
정용섭 (Y. Jeong, yongseop@etri.re.kr) 필드로보틱스연구실 선임연구원
이우주 (W. Lee, dnwn24@etri.re.kr) 필드로보틱스연구실 연구원
최승민 (S.M. Choi, ccsmm@etri.re.kr) 필드로보틱스연구실 책임연구원/실장, UST 인공지능학과 교수

ABSTRACT

Recent advances in foundation models have extended AI from digital environments to the physical world, thereby accelerating the emergence of Physical AI. In robotics, action intelligence based on vision-language-action and navigation intelligence based on vision-language navigation are evolving rapidly. Recent approaches have explored the separation between high-level reasoning and real-time execution. This paper reviews these developments and their integration into mobile manipulation and whole-body coordination. We also discuss data, simulations, and on-device execution as implementation foundations for integrated Physical AI.

KEYWORDS Physical AI, Robot Foundation Model, VLA, VLM, VLN

I. 서론

대규모 언어 모델(LLM: Large Language Model)과 비전-언어 모델(VLM: Vision-Language Model)을 중심으로 발전해 온 생성형 인공지능은 최근 디지털 공간을 넘어 물리 세계에서 직접 행동하는 피지컬 AI(Physical AI)로 확장되고 있다. 피지컬 AI는 로봇과 자율주행차 등의 물리적 개체가 센서를 통해 환경을 인지하고 상황을 이해 · 추론하여 그 결과를 실제 행동으로 연결하는 인공지능 기술이다.

그림 1에서 보듯이 피지컬 AI와 기존 디지털 AI의 중요한 차이는 AI의 출력이 물리적인 결과를 발생시킨다는 점이다. 생성형 AI의 잘못된 답변은 다시 생성할 수 있지만, 로봇의 잘못된 판단은 충돌, 파손 또는 안전사고로 이어질 수 있다. 입력 역시 텍스트와 이미지에서 카메라, LiDAR, IMU 등의 센서 정보로 확장되고, 출력은 텍스트에서 속도, 토크, 관절각 등의 실제 모터 명령으로 변화한다. 따라서 피지

* DOI: https://doi.org/10.22648/ETRI.2026.J.410602
* 공동 제1저자 장지호, 최승민
* This work is the result of research conducted under the support of the Ministry of Science and ICT (MSIT) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) through the "Development of Core Technologies for Application Support of Autonomous Agents and Devices" program (RS-2026-25520169).

본 저작물은 공공누리 제4유형
출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.
©2026 한국전자통신연구원

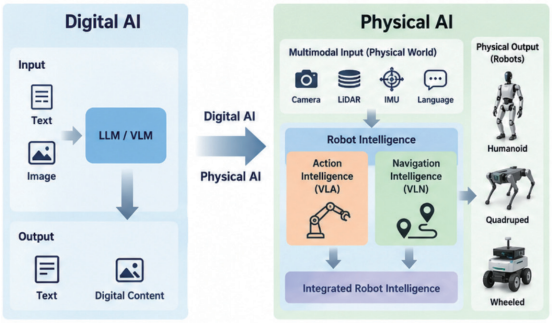


그림 1 디지털 AI에서 피지컬 AI로의 확장과 로봇 행동·이동지능

컬 AI에는 추론 능력뿐만 아니라 실시간성, 안전성 및 환경 적응성이 동시에 요구된다[1].

본고에서는 로봇의 피지컬 AI를 행동지능(Action Intelligence)과 이동지능(Navigation Intelligence)을 중심으로 살펴본다. 행동지능은 물체를 집고 옮기거나 도구를 사용하는 등 주변 환경과 물리적으로 상호작용을 하는 능력으로, 최근에는 시각과 언어 명령으로부터 직접 행동을 생성하는 VLA(Vision-Language-Action)가 핵심 기술로 부상하고 있다. 한편, 이동지능은 목적지를 이해하고 주변 환경을 인지하여 안전하게 이동하는 능력으로, VLM과 VLN(Vision-Language Navigation)의 발전에 따라 기존 지도 기반 이동에 시각 · 언어 의미의 이해와 추론을 결합하는 방향으로 확장되고 있다[2,3].

실제 로봇 임무에서는 행동과 이동을 명확히 분리하기 어렵다. 예를 들어, “창고로 이동하여 상자를 가져오라”라는 명령은 장거리 이동, 장애물 회피, 대상 탐색, 접근, 물체 조작 및 복귀를 연속적으로 요구한다. 따라서 향후 로봇지능은 행동과 이동을 독립적으로 해결하는 수준을 넘어 인지-추론-이동-조작을 하나의 장기 임무로 연결하는 통합 지능으로 발전할 필요가 있다.

본고에서는 VLA 기반 행동지능과 VLM · VLN 기반 이동지능의 최근 발전을 살펴보고, 추론과 실행의 연결 방식 및 행동 · 이동 통합을 중심으로 기술 흐름을 정리한다. 또한, 시뮬레이션과 온디바이스 실행 등 실제 구현 조건을 논의한다.

II. 로봇 파운데이션 모델의 개념

1. 파운데이션 모델과 로봇 지능

파운데이션 모델(Foundation Model)은 대규모 데이터로 사전학습되어 다양한 후속 작업에 활용되는 범용 AI 모델이다[4]. 로봇 파운데이션 모델(Robot Foundation Model)은 시각 · 언어 이해를 실제 로봇 행동으로 연결한다. 그림 2는 2022~2026년의 연구 흐름을 작업 조작(Manipulation)과 내비게이션(Navigation)으로 나누어 보여준다.

규칙 기반 로봇 제어는 강화학습(RL: Reinforcement Learning)과 행동복제(BC: Behavior Cloning) 등의 모방학습(IL: Imitation Learning)으로 발전했으나, 정책은 대개 특정 로봇 · 작업에 한정되었다. 로봇 파운데이션 모델은 공동 학습으로 다양한 몸체(Embodiment)를 제어하고, VLM의 세계 지식을 활용해 새로운 지시 · 장면에 대응하며 적은 데이터로 적응(Fine-Tuning)하는 것을 목표로 한다. 다음에서는 데이터, 추론 · 실행 구조, VLA를 살펴본다. 개별 원문 대신 서베이를 인용한 항목은 해당 문헌의 정리를 따른다.

2. 로봇 학습 데이터의 확장

로봇 학습에는 영상 · 상태 · 행동 · 결과가 시간상으로 정렬된 데이터가 필요하며, 원격 조작 수집의 비용과 다양성 확보가 과제이다. Open X-Embodiment [5]는 21개 기관 · 22종 로봇의 100만 개 이상 에피소드를, DROID[4]는 564개 장면 · 86개 작업의 약 7.6만 개 에피소드를 공개하였다.

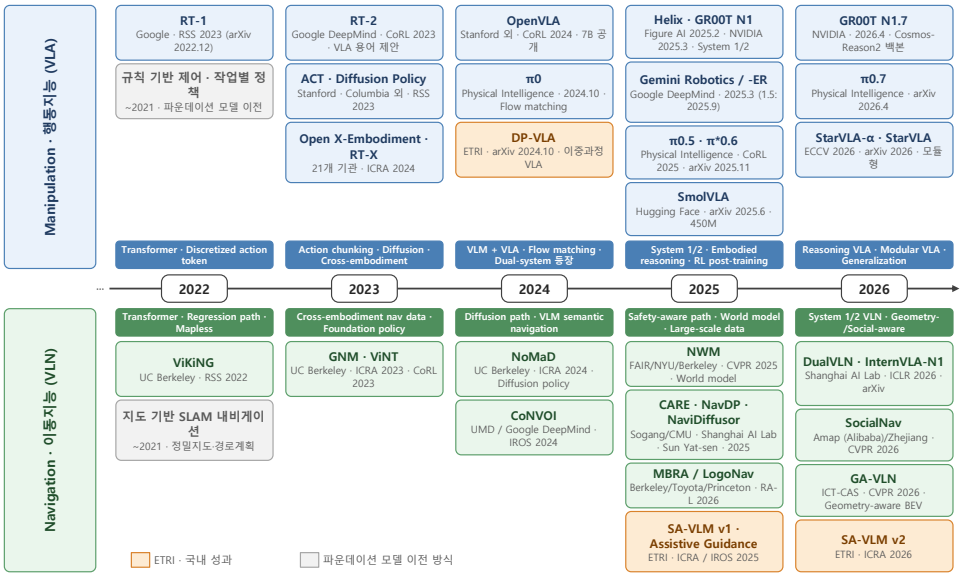


그림 2 VLA와 VLN의 발전 타임라인

학습 데이터는 인간 영상 · 잠재 행동 · 합성 데이터로 확장되고 있다. 첫째, 손 자세를 기록한 인간 1인칭 데이터셋 EgoDex(829시간)가 공개되었다[6]. NVIDIA는 인간 1인칭 사전학습 · 인간-로봇 정렬을 다룬 EgoScale을 GR00T N1.7의 인간 영상 활용 기반으로 소개하며, 2만 시간 이상의 데이터를 활용한 평가에서 규모 확대에 따른 조작 성능 개선을 보고하였다[7]. 둘째, 행동 라벨 없는 영상의 잠재 행동(Latent Action) 학습이다. LAPA[6]는 프레임 변화를 이산 잠재 행동으로 부호화해 VLA를 사전학습하고, 행동 주석이 있는 로봇 데이터로 실제 행동 공간에 적응시킨다. UniVLA[6]도 이를 활용하며, WholeBodyVLA는 이동 · 조작 영상의 잠재 행동으로 휴머노이드를 학습한다. 셋째, 생성 모델과 시뮬레이션의 합성 데이터이다. DreamGen[8]은 수행 영상을 생성하고 잠재 행동 모델 또는 역동역학 모델로 의사 행동을 복원하여 정책을 학습한다. 이는 Ⅵ장 1절의 물리 시뮬레이터 기반 생성과 상보적이나 생성 · 검증 절차는 다르다. 이들 데이터는 사전학습 · 로봇별 적응 · 실환경 평가 · 경험 기반 후속학습에 상보적으로 활용된다.

3. 추론과 실행의 분리: System 1/2

고수준 의미 판단과 저수준 실행을 나누는 연구가 확대되고 있다. System1/System2는 빠른 직관과 느린 숙고를 구분하는 심리학의 이중과정 이론을 차용한 비유이다[9]. ETRI의 DP-VLA(Dual Process VLA, 2024.10)[9]는 대형 모델의 특징으로 소형 행동 정책을 조건화하여 RoboCasa에서 추론 효율과 성공률의 개선 가능성을 보였다. Helix[10]와 GR00T N1[11]도 의미 처리와 빠른 행동 생성을 연결하나 학습 · 실행 방식은 다르다.

그림 3의 두 역할은 잠재 표현, 영상상의 목표 · 경유점, 서브태스크 언어 등으로 연결된다. 표시 주파수는 예시이며, 모델 · 장치별 갱신 · 추론 · 실행 주기는 다르다.

DP-VLA는 OpenVLA 특징을 조건으로 소형

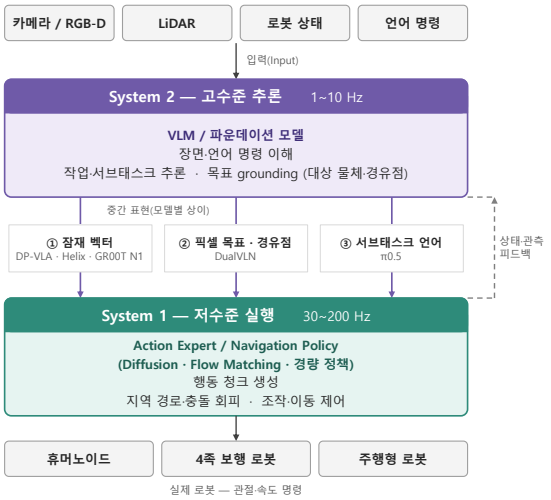


그림 3 System 1/2 기반 피지컬 AI 구조

BC-Transformer를 별도 학습하고, 지시 시작 시 얻은 특징을 실행 중 재사용한다[9]. DualVLN[3]은 픽셀 · 잠재 목표를 전달하며, π0.5[12]는 서브태스크 언어에 이어 행동을 생성한다. 명시적 목표는 해석 · 오류 분석에 유리하고, 잠재 표현은 학습된 정보를 전달하나 완전한 보존을 보장하지 않는다. 최신 관측 · 상태 입력과 제어 · 안전 검증은 모델별로 고려한다.

의미 판단과 실행은 계산량과 응답성 요구가 다르다. 의미 표현 갱신, 청크 추론, 청크 내 행동, 제어 명령의 주기도 구분해야 한다. 행동 시간 해상도 50Hz는 초당 50회 청크 추론을 뜻하지 않으므로, 실행 성능은 장치 · 관측 및 행동 주기 · 청크 길이와 함께 해석해야 한다.

두 독립 모델만이 해법은 아니다. π0.5는 단일 모델에서 추론과 행동을 생성하고, FiS-VLA[6]는 내부에 빠른 실행 경로를 구성한다.

SmolVLA[13]는 현재 행동을 실행하는 동안 다음 행동 청크를 계산하는 비동기 추론을 사용한다. 이는 고수준 추론과 저수준 행동 생성의 모듈 구분과는 별개로, 행동 예측과 실행을 병행하여 대기 시간을 줄이는 운용 방식이다.

III. VLA 기반 로봇 행동지능

1. VLA의 주요 발전 흐름

VLA는 시각(Vision) · 언어(Language) 이해를 로봇 행동(Action)으로 확장한다. 표 1은 대표 모델의 구조와 공개 현황이다. 이하 분류는 시기별이 아닌 행동 표현과 추론 · 실행 구조에 따른 것으로, 한 모델이 여러 흐름에 걸칠 수 있다.

1.1 이산 행동 토큰 기반 접근

VLA라는 용어를 처음 제안한 RT-2(2023)[14]는 웹 규모 시각-언어 · 로봇 데이터를 공동 학습(Co-Fine-Tuning)하고 행동을 이산 토큰으로 표현하여 VLM을 로봇 제어로 확장하였다. Open X-Embodiment[5]로 학습한 OpenVLA(2024)[6]는 여러 로봇을 제어하는 7B 공개 모델로 기준점을 제시하였다. 매 시점의 행동 토큰을 자기회귀 출력하는 초기 방식에서는 생성 지연과 정밀한 연속 제어를 위한 행동 표현이 개선 과제였다.

1.2 행동 청킹과 연속 행동 생성

VLA와 별도로 행동 시퀀스를 생성하는 학습 기반 정책도 발전하였다. Diffusion Policy(2023)[4]는 관측을 조건으로 행동 시퀀스의 잡음을 반복 제거(Denoising)하여 다양한(Multi-Modal) 연속 행동 분포를 표현한다. ACT(2023)[4]는 미래 행동 청크를 예측하고 시간상으로 겹치는 예측을 결합(Temporal Ensembling)한다. π0(2024)[15]는 PaliGemma 3B와 Action Expert(300M)를 결합하고 Flow Matching으로 행동 시간 해상도 최대 50Hz의 청크를 생성한다. 7종 로봇 · 68개 작업을 공동 학습해 Cross-Embodiment 범용 정책을 지향했으며, GR00T N1[11]과

SmolVLA[13]도 VLM과 생성형 행동 모듈을 결합하였다. π0-FAST(2025)[16]는 행동 청크를 주파수 영역에서 압축한 FAST 토큰으로 자기회귀 생성하여, 보고된 평가에서 확산 방식에 견줄 성능을 더 적은 학습 비용으로 얻었다.

OpenVLA-OFT[6]는 병렬 디코딩과 연속 행동 회귀로 처리량을 수십 배 높였다. Real-Time Chunking [17]은 실행 중인 행동을 고정하고 나머지를 인페인팅(Inpainting)하여 재학습 없이 청크 전환 지연의 영향을 줄였다. 행동 표현과 계산 · 실행의 시간적 정합이 응답성 향상의 주요 방향이다.

1.3 추론과 실행의 결합 방식

DP-VLA[9]에 이어 Helix[10]는 휴머노이드 상체 35자유도의 200Hz 제어를 보고하였고, GR00T N1[11]은 VLM(Eagle-2)의 시각 · 언어 정보를 확산 트랜스포머(DiT: Diffusion Transformer)가 참조해 Flow Matching으로 행동을 생성하는 공개 모델을 제시하였다. Google DeepMind는 Gemini 2.0 기반 VLA인 Gemini Robotics와 공간 이해 · 파지점 · 궤적을 다루는 체화추론 모델 Gemini Robotics-ER을 제공하고[18], 행동 전에 사고 과정을 생성하는 Gemini Robotics 1.5[19]로 확장하였다. StarVLA[20]

표 1 대표 VLA의 구조 및 공개 현황

모델	기관	발표	VLM 백본	행동 생성 방식	규모	공개
RT-2	Google DeepMind	2023. 7.	PaLI-X / PaLM-E	이산 행동 토큰(자기회귀)	55B / 12B	비공개
OpenVLA	Stanford 외	2024. 6.	Llama-2 7B + DINOv2·SigLIP	이산 행동 토큰	7B	공개
π0	Physical Intelligence	2024. 10.	PaliGemma 3B	Action Expert + Flow Matching 기반 행동 청크	≈3.3B	공개(openpi, 2025. 2.)
π0-FAST	Physical Intelligence	2025. 1.	PaliGemma 3B	FAST 주파수 토큰(자기회귀)	≈3B	공개
SmolVLA	Hugging Face	2025. 6.	SmolVLM-2	Flow Matching Action Expert, 비동기 추론	450M	공개
DP-VLA	ETRI(한국)	2024. 10.	OpenVLA (실험 백본)	특징 조건부 BC-Transformer, 별도 학습	—	비공개
Helix	Figure AI	2025. 2.	7B VLM(S2)	80M 정책(S1), 잠재 표현 조건부	7B + 80M	비공개
Gemini Robotics / -ER	Google DeepMind	2025. 3. (1.5: 2025. 9.)	Gemini 2.0 / 2.5	VLA와 체화추론(ER) 모델 분리	비공개	제한공개
GR00T N1	NVIDIA	2025. 3.	Eagle-2(S2)	DiT + Flow Matching(S1)	2B	공개
π0.5	Physical Intelligence	2025. 4.	PaliGemma 3B	서브태스크 언어 예측 → Flow Matching	≈3B	공개(openpi, 2025. 9.)
π*0.6	Physical Intelligence	2025. 11.	Gemma 3 4B	≈860M Action Expert, Recap(경험·교정 RL) 후속학습	≈5B	비공개
GR00T N1.7	NVIDIA	2026. 4.	Cosmos-Reason2-2B	DiT + Flow Matching, 서브태스크 추론	3B	코드·가중치 공개
π0.7	Physical Intelligence	2026. 4.	Gemma 3 4B	≈860M Action Expert, 다중 조건 프롬프트	≈5B	비공개
StarVLA-α[21]	—	2026	Qwen3-VL-4B-Instruct	공유 MLP 연속 행동 헤드	4B+MLP	공개

는 백본과 행동 헤드를 교체하여 학습 · 평가하는 개발 코드베이스이다. 이와 구분되는 모델인 StarVLA-α[21]는 Qwen3-VL-4B-Instruct와 공유 MLP 연속 행동 헤드를 결합한다. 명시적 계층화, 단일 모델 내 추론 · 행동 결합 및 직접 행동 정책이 병존하는 사례들이다.

2. 최근 흐름

2.1 추론형 VLA

최근 연구는 행동 생성에 체화 추론(Embodied Reasoning)과 시간적 문맥을 결합한다. ECoT[6]는 계획 · 서브태스크 · 물체 경계상자 · 그리퍼 위치 등 시각적으로 접지된 추론을 행동하기 전에 생성하도록 학습하여, 해당 일반화 평가에서 추가 로봇 데이터 없이 OpenVLA 대비 성공률을 28%p 높였다. CoT-VLA[6]는 미래 프레임을 시각적 서브골로 생성한 뒤 행동을 출력한다. 물리적 상식 · 체화 추론 VLM인 Cosmos-Reason[22]의 후속 버전은 GR00T N1.7의 백본으로 채택되었다[7]. π0.5는 서브태스크 언어에 이어 행동을 생성하고[12], π0.7은 과거 관측을 압축하는 기억과 언어 · 시각적 하위목표 등의 조건으로 수행을 조절하며 새로운 작업 조합으로의 일반화를 탐색한다[23]. 추론 이득은 추가 계산 지연과 함께 평가해야 한다.

2.2 강화학습 기반 후속학습

시연 데이터의 품질 · 분포에 영향받는 VLA의 개선을 위해 운용 경험과 사람의 교정을 이용하는 강화학습 기반 후속학습(Post-Training)이 확대되고 있다. $\pi^*0.6$[24]의 Recap은 시연 · 자율 수행 · 원격 조작 교정 데이터를 이점(Advantage) 조건부 정책 학습에 활용하여 일부 평가 작업에서 처리량 증가와 실패율 감소를 보고하였다. VLA-RL[6] 등도 강화학습을 통한 정책 개선을 탐색한다. 이는 시뮬레이션 학습 · 검증과 결합할 수 있으나 경험 기반 강화학습과 Real-to-Sim, Sim-to-Real 순환은 구분된다.

2.3 경량화와 온디바이스 실행

수억~수십억 파라미터의 경량 VLA도 개발되고 있다. SmolVLA[13]는 450M 모델과 비동기 실행을 제시하였다. 백본 · 행동 헤드 최적화와 계산 · 실행 주기 조정은 배포 비용을 낮추지만, 크기만으로 능력이나 실시간성을 판단할 수 없다.

3. 평가 체계와 실환경 검증

LIBERO · SimplerEnv[6]와 DP-VLA의 RoboCasa 기반 평가[9]는 재현 가능한 비교를 지원하나 실환경 신뢰성을 보장하지 않는다.

RoboArena[25]는 여러 기관의 실로봇에서 정책을 쌍대 비교한다. 성공률과 함께 실환경 전이 · 반복 안정성 · 처리량(Throughput) · 사람 개입도 평가해야 한다.

실물 검증에는 학습 · 평가 환경과 성공 기준 · 반복 횟수를 명시하고 관측 · 동역학 · 제어 조건의 영향을 구분한다. 장기 수행의 실패 감지 · 복구, 분포 밖 일반화와 안전성도 평가한다.

4. VLA 모델 실험 예시

ETRI에서는 공개 VLA와 개발 코드베이스로 Pick & Place 적용 실험을 수행하였다. 그림 4와 같이 LeRobot[26]의 Leader-Follower 원격 조작으로 RGB 영상 · 로봇 상태 · 행동을 수집하고, 같은 환경을 Isaac Sim에 구성하여 학습 · 평가에 활용하였다.

약 100개 에피소드로 π0[15], GR00T N1[11] 및 StarVLA[20] 코드베이스를 활용한 구성을

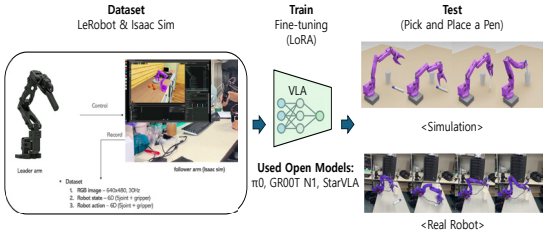


그림 4 공개 VLA와 개발 코드베이스를 활용한 Pick & Place 적용 실험(ETRI 12동)

LoRA(Low-Rank Adaptation) 방식으로 파인튜닝하여 'Pick and Place a Pen'을 확인하였다. 시뮬레이션 검증 후 실로봇에서도 시각 관측 · 작업 명령에 따른 연속 로봇팔 행동으로 작업을 수행하여, 수집-학습-가상 검증-실물 적용의 파이프라인을 구성하였다.

이 가운데 π0의 예비 평가에서 작업 성공률은 시뮬레이션에서 약 80%였으나 실제 환경에서는 30% 대로 감소하였다. 관측 · 동역학 · 제어 인터페이스 및 평가 조건에 대한 추가 분석이 필요하며, 제한된 예비 실험만으로 모델 간 일반적인 우열이나 성능 차이의 원인을 분리하여 판단하지 않는다.

VLA에서는 이산 · 연속 행동 생성, 추론 · 실행의 결합, 경험 기반 후속학습과 실행 효율화가 병행 발전한다. 이를 넓은 작업 공간의 임무로 확장하려면 대상과 작업 위치까지 이동해야 한다. 다음 장에서는 시각 이동정책과 언어 · 공간 의미를 활용하는 이동지능을 살펴본다.

IV. VLN 기반 로봇 이동지능

1. 이동지능 패러다임의 변화

전통적인 로봇 내비게이션은 지도 작성, 위치추정, 전역 · 지역 경로계획 및 제어 모듈로 구성된다. 구조화된 환경에서는 높은 신뢰성을 제공하지만, 정밀지도를 사전에 구축하고 환경 변화에 맞춰 관리해야 한다. 최근에는 이러한 지도 기반 이동에 AI

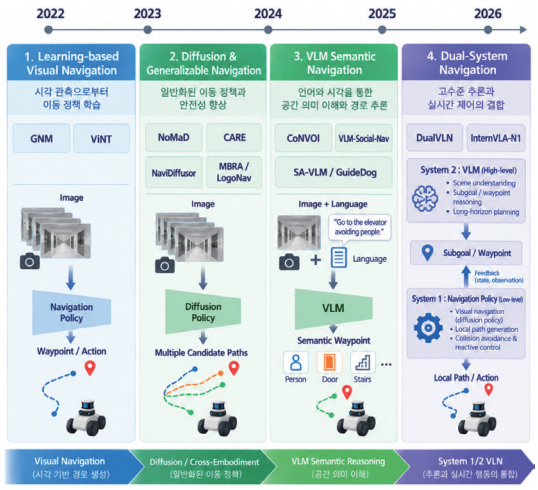


그림 5 접근 유형별 VLN 기술의 발전 동향

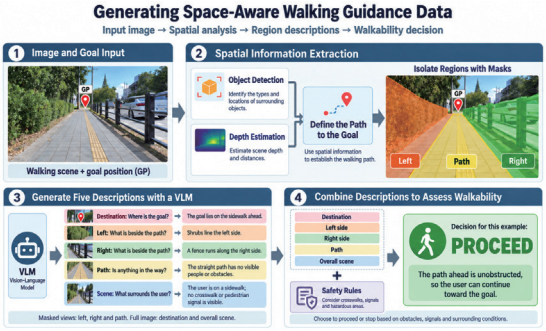
기반 이동지능이 결합하고 있다. VLM의 발전으로 로봇은 장애물의 위치뿐만 아니라 문, 계단, 복도, 엘리베이터와 같은 공간의 의미를 이해하고 자연어 명령과 연결할 수 있게 되었다. 이동지능은 목적지까지의 기하학적 경로 계산을 넘어 공간의 의미와 상황에 맞는 이동 방법을 추론하는 문제로 확장되고 있다.

이 장에서는 그림 5와 같이 VLN의 발전을 이해하기 위한 출발점으로 학습 기반 Visual Navigation을 먼저 살펴본다. 시각 관측으로 이동 행동을 생성하는 정책에 언어 지시와 공간 의미를 결합하는 과정을 검토하고, 고수준 목표 추론과 실시간 이동정책을 연계하는 최근 VLN 연구로 논의를 확장한다.

2. 학습 기반 Visual Navigation

학습 기반 Visual Navigation에서는 카메라 영상으로 이동 경로나 경유점(Waypoint)을 생성한다. GNM은 6종의 로봇에서 수집한 약 60시간의 데이터를 활용하였고, ViNT는 다양한 로봇과 환경의 수백 시간 규모 데이터를 활용하는 내비게이션 파운데이션 모

델을 제시하였다. NoMaD는 목표 지향 이동과 탐색을 함께 다루는 확산 정책이다[2,27,28]. CARE와 NaviDiffusor는 비용 지도나 장애물 정보를 활용하여 학습 기반 정책이 생성하는 경로의 안전성을 보완한다[29,30]. MBRA는 대규모 주행 데이터의 행동 주석을 개선하는 접근으로, 원시 약 2,000시간에서 정제한 약 700시간의 데이터를 활용하여 LogoNav 정책을 학습하였다[31]. 이러한 시각 이동정책은 다음 절에서 다룰 언어 · 공간 의미 기반 목표를 실제 이동으로 연결하는 실행 기반이 된다.



출처 참고문헌 [34]를 참고하여 저자 작성

그림 6 공간 정보와 영역별 장면 설명을 활용한 SAIT 학습데이터 생성 절차

3. VLM 의미·추론 기반 이동지능

VLM을 이동지능에 적용하면 로봇은 주변 환경의 의미를 해석하고 언어 명령과 현재 상황을 함께 고려하여 이동할 수 있다. 예를 들어, "사람이 많은 곳을 피해 출구로 이동하라"라는 명령은 단순한 최단경로 계획만으로 해결하기 어렵다. 로봇은 사람과 통로를 인식하고 혼잡도 및 사회적 규칙을 이해해야 한다. VLM-Social-Nav, CoNVOI 등의 연구는 이러한 의미적 · 사회적 정보를 경로계획에 반영하는 방향을 제시하고 있다[32,33].

그림 6과 관련된 ETRI의 SA-VLM 연구는 영상과 목표 위치에 대한 공간 이해를 시각장애인 보행 안내에 활용한다[34]. SAIT(Space-Aware Instruction Tuning) 데이터 생성 단계에서는 가상의 경로와 주변 영역을 구분하여 공간 관계 및 장면 설명을 구성하며, 이를 학습한 모델은 목적지와 경로 주변 상황을 해석하여 안내 정보를 생성한다.

후속 연구로 SA-VLM v2도 발표되었다[35]. 이러한 안내 추론은 로봇의 실제 경로 제어와 구분된다. GuideDog 관련 연구에서는 사용자가 직진 · 좌회전 · 우회전 등의 방향을 지시하면, 안내로봇이 온라인으로 보행 경로의 구조를 인식하여 해당 방향으로 이동 가능한 경로를 생성한다[36]. 이들 연구는 공간 의미 및 사용자 의도의 이해를 로봇 이동과 연결하는 접근을 보여준다.

VLN에서도 고수준 의미 추론과 실시간 실행을 연계하는 System 2/System 1 구조가 연구되고 있다. InternVLA-N1 계열의 DualVLN은 VLM이 시각 관측과 언어 지시로 영상상의 중간 목표와 잠재 목표 표현을 생성하고, 경량 이동정책이 이를 최신 시각 관측과 결합하여 연속 경로를 생성한다[3]. 이는 의미적 목표 판단과 주변 환경에 대한 빠른 대응을 연결하는 접근이다.

이러한 흐름은 VLA와 VLN이 의미적 목표와 물리적 실행을 효율적으로 연결해야 한다는 공통 과제를 보여준다. 다만 조작은 접촉을 포함한 관절 · 말단 행동을, 이동은 공간 이력과 주변 장애물을 고려한 경로나 속도 명령을 주로 다룬다. 조작과 이동을 하나의 임무로 연결하려면 각 실행 정책의 출력과 물리적 제약에 맞는 목표 전달 방식이 중요하다.

4. 공개 Visual Navigation 모델의 적용 예시

VLN의 실행 기반이 되는 시각 이동정책의 적용

가능성을 살펴보기 위해 ETRI에서는 공개 Visual Navigation 모델로 이동 실험을 수행하였다.

그림 7은 Real2Sim 방식으로 구축한 빌딩, 공장 및 지하철의 Digital Twin 환경에 NoMaD와 MBRA/LogoNav 관련 공개 구성을 추가 학습 없이 적용한 결과이다. 각 환경은 편도 약 200m 이내의 정적 경로로 구성하고, 동일한 시작점과 목표점에서 주행 양상을 확인하였다.

실험 결과, 대규모 글로벌 주행 데이터로 학습된 LogoNav(dense)는 세 환경 모두에서 목표 지점까지 장거리 주행에 성공하였다. 반면 NoMaD, MBRA와 LogoNav(sparse)는 일부 구간을 이동하였으나 전체 경로를 완주하지 못하였다. 다만, 이 차이를 데이터 규모만의 효과나 일반적인 모델 우열로 해석하기는 어렵다.

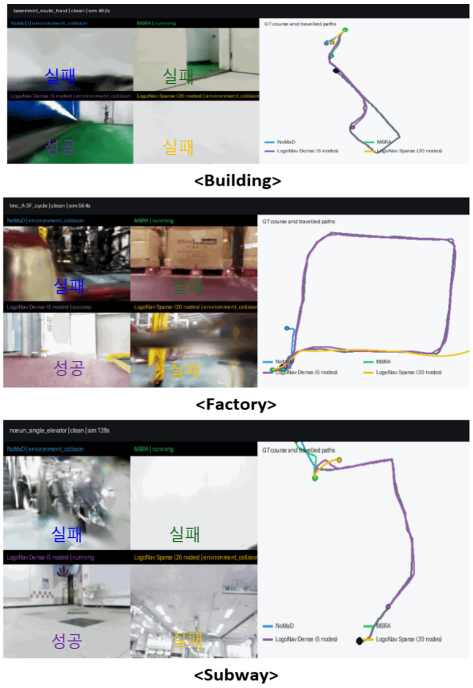


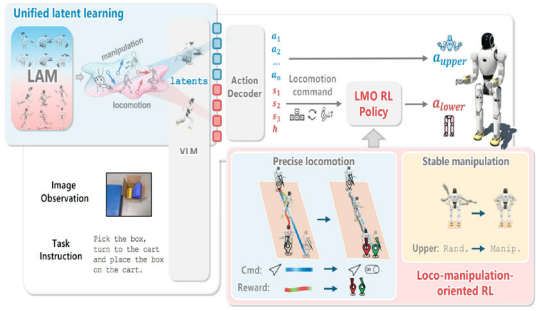
그림 7 빌딩, 공장 및 지하철 Digital Twin 환경에서 공개 Visual Navigation 모델의 Zero-shot 장거리 이동 실험 예시: NoMaD/MBRA/LogoNav (dense topo.)/LogoNav(sparse topo.)

본 실험은 공개 Visual Navigation 모델을 현장 추가 학습 없이 새로운 Digital Twin에 적용하는 절차와 장거리 이동 가능성을 확인한 사례이다. 동적 장애물과 복잡한 이동 상황을 포함한 실환경으로 확장하려면 다양한 로봇 · 환경 데이터와 함께 장거리 누적 오차, 환경 변화 및 Out-of-Distribution(OOD) 일반화에 대한 검증이 필요하다. 실제 임무는 목적지 도달에 이어 대상 접근과 조작을 요구하므로, 다음 장에서는 행동과 이동의 연결을 살펴본다.

V. 행동·이동 통합 로봇지능

1. 행동과 이동의 통합

조작지능과 이동지능은 실제 임무에서 긴밀하게 연결된다. 물건을 가져오는 임무는 대상 탐색, 접근, 파지와 운반을 요구하며, 이동은 조작 가능한 위치와 자세를 만드는 과정이기도 하다. 통합은 이동 · 조작을 순차적으로 연결하는 수준과 조작 목적에 맞춰 몸체와 팔을 긴밀하게 협응하는 수준으로 나누어 볼 수 있다. 핵심은 모든 동작의 동시 수행이 아니라 임무에 필요한 행동을 적절한 상태에서 연결하는 데 있다.



출처 Reprinted from H. Jiang et al., "WholeBodyVLA: Towards Unified Latent VLA for Whole-Body Loco-Manipulation Control," in Proc. Int. Conf. Learn. Represent., (Rio de Janeiro, Brazil), 2026. [37].

그림 8 WholeBodyVLA의 이동·조작 잠재 행동 학습과 상체 행동·하체 이동 실행을 결합한 구조

이동과 조작을 긴밀하게 협응하는 이동조작(Loco-Manipulation) 연구도 확대되고 있다. 그림 8의 WholeBodyVLA는 이동 · 조작 영상에서 얻은 잠재 행동을 VLA 학습에 활용하고 로봇 데이터로 실행 명령에 연결한다. 상체 행동과 하체 이동 명령을 생성하되, 하체는 별도의 LMO(Loco-Manipulation-Oriented) 강화학습 정책으로 실행한다[37]. 이는 행동 학습의 통합과 실행 제어의 역할 분담을 함께 활용하는 사례이다. 연속 · 장기 임무 수행 사례도 제시하였다.

2. 통합 학습과 모듈형 실행

행동 · 이동 통합은 학습과 실행의 관점으로 나누어 볼 수 있다. 공동 학습은 이동과 조작의 관계를 표현과 정책 학습에 함께 반영하고, 모듈형 실행은 계획 · 정책 · 제어의 역할을 나누어 행동을 수행한다. 두 방식은 상호 배타적이지 않으며, 이동과 조작을 통합 학습한 뒤 로봇별 제어 모듈로 실행할 수도 있다.

모듈형 구성에서는 고수준 계획이 이동 · 조작 목표를 전달하고, 각 정책이 명령을 생성하여 로봇별 제어기가 실행한다. Being-0는 고수준 계획을 모듈형 기술 실행과 연결하는 사례이고[38], Visual Whole-Body Control은 시각 관측을 활용한 물리적 이동 · 조작 협응의 사례이다[39]. 두 연구는 계획의 연결과 전신 제어라는 서로 다른 통합 수준을 보여준다.

조작과 이동을 연결할 때는 상위 목표를 어떤 표현으로 전달하는지, 하위 정책이 어떤 행동을 출력하는지, 이전 관측과 작업 이력을 어디에 유지하는지를 함께 고려해야 한다. 조작에서는 대상과의 접촉 및 도달 가능한 자세가 중요하고, 이동에서는 공간 이력과 충돌 회피가 중요하다. 전신 이동조작에서는 이러한 조건에 몸체의 균형과 안정성까지 더해지므로, 의미적 임무 연결과 물리적 협응을 구분하여 설계 · 평가할 필요가 있다.

이러한 통합을 실제 로봇에서 구현하려면 모델 구조와 함께 데이터 확보, 실가상 검증, 제한된 탑재 자원에서의 실행 및 안전 제어를 고려해야 한다. 다음 장에서는 이러한 구현 기반과 국내 추진 방향을 살펴본다.

VI. 피지컬 AI의 학습·구현 기반과 발전 방향

1. 로봇 데이터와 시뮬레이션의 결합

앞서 살펴본 데이터 확장과 경험 기반 학습을 실제 로봇에 적용하려면 데이터 수집부터 학습 · 검증 · 배포까지 이어지는 절차가 필요하다. 실로봇 데이터는 사전학습과 로봇별 적응뿐만 아니라 실환경 평가와 실패 분석에도 사용된다. 시뮬레이션은 환경과 로봇 조건을 통제한 데이터 생성 및 사전 검증을 지원하며, 실제 데이터와 상보적으로 활용할 수 있다.

그림 9는 NVIDIA GTC 자료로 제시된 도식의 구성 개념을 참고하여 재구성한 것이다. 피지컬 AI

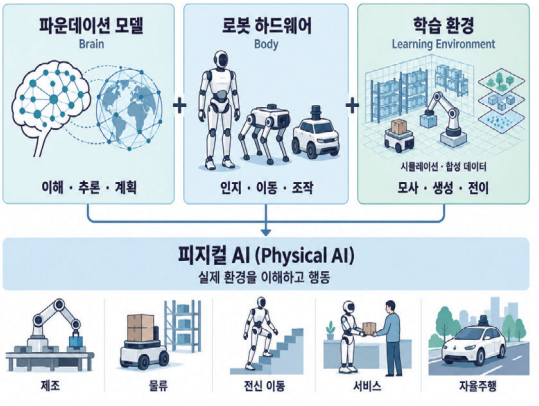


그림 9 피지컬 AI의 개념 및 구성 요소

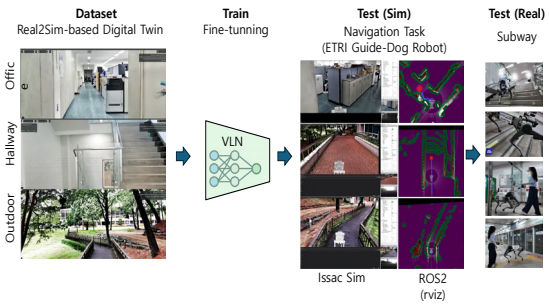


그림 10 ETRI에서 연구 중인 로봇 이동지능을 위한 Digital Twin 기반 데이터셋 및 학습·검증 파이프라인

는 Brain(파운데이션 모델), Body(로봇 H/W), Learning Environment(시뮬레이션 및 합성 데이터)의 결합으로 구성될 수 있다. Isaac Sim 등의 로봇 시뮬레이터와 로봇 모델, 강화학습 기반 제어 모델 및 Real2Sim 환경 모델링을 결합하면 실제 환경을 디지털 공간에 재현하고 다양한 행동 · 이동 데이터를 생성할 수 있다.

ETRI에서도 그림 10과 같이 실환경 기반 Digital Twin 데이터셋과 시뮬레이션을 연계하여 이동지능 모델을 학습 · 검증하는 파이프라인을 연구하고 있다. 실제 환경의 영상과 센서 정보로 Real2Sim 환경을 구성하고 생성한 주행 데이터를 학습에 활용하며, Isaac Sim · ROS2 기반 시뮬레이션 검증을 실로봇 평가와 연결한다. 이때 실가상 간 관측 · 동역학 · 제어 인터페이스의 차이를 함께 점검해야 한다.

실환경 재현(Real2Sim), 가상 학습 · 검증, 학습된 정책의 실환경 전이와 평가(Sim2Real)를 연결하려면 각 단계의 데이터와 평가 기준을 연계해야 한다. 실제 로봇의 실패나 사람의 개입이 발생한 구간을 수집하고, 관련 장면과 조건을 가상 환경에서 재현하여 정책을 개선한 뒤 실물에서 다시 검증하는 순환을 고려할 수 있다. 이러한 실가상 순환은 경험 데이터를 정책 개선에 사용하는 후속 학습과도 결합할 수 있다.

2. World Model 기반 예측 지능

학습 데이터 생성과 실행 전 검증을 보완하는 월드 모델도 연구되고 있다. 물리적 상황을 해석하는 추론 VLM과 미래 장면을 생성 · 예측하는 월드 모델은 구분할 필요가 있다. 전자는 관측과 임무의 의미를 판단하고, 후자는 조건에 따른 미래 전개를 예측하여 데이터 생성이나 계획 평가에 활용될 수 있다.

NVIDIA Cosmos 제품군에서도 Cosmos-Reason은 물리적 추론 VLM이며, Cosmos-Predict 계열은 미래 영상 생성 · 예측을 담당한다[22]. 행동 조건부 월드 모델은 현재 상태와 후보 행동에 따른 미래 변화를 예측하여 실행 전 평가에 활용할 수 있다. 실제 로봇에 활용하려면 행동 조건에 따른 예측 정확도와 물리적 일관성을 검증해야 한다.

향후에는 VLM의 의미적 추론과 World Model의 물리적 미래 예측을 결합하여, 현재 상황의 이해에서 행동 결과를 예측하고 계획하는 방향으로 발전할 것으로 예상된다.

3. 온디바이스 피지컬 AI

클라우드 기반 AI는 높은 연산성능을 활용할 수 있지만, 로봇 제어에서는 네트워크 지연과 통신 단절 문제가 발생할 수 있다. 특히 장애물 회피, 보행 및 긴급정지와 같이 높은 실시간성과 안전성이 요구되는 기능을 클라우드에 의존하는 데에는 한계가 있다.

따라서 로봇 자체의 AI 반도체에서 VLM 및 행동 · 이동 모델을 실행하는 On-device Physical AI가 중요하다. 이를 위해서는 모델 경량화와 양자화뿐만 아니라 제한된 연산자원에서 효율적으로 실행할 수 있는 하드웨어 친화적 모델 구조가 필요하다.

탑재 실행에서는 모델 크기뿐만 아니라 메모리

사용량, 관측부터 명령 출력까지의 지연, 행동 청크 계산과 실행의 주기를 함께 고려해야 한다. 고수준 의미 판단과 저수준 실행의 역할 분담이나 청크의 비동기 처리는 활용 가능한 방법이지만, 실제 장치에서 지연 예산과 제어 안정성을 확인해야 한다.

4. 생성형 AI와 안전 제어의 결합

VLM이나 생성형 정책이 의미상으로 타당한 행동을 생성하더라도 로봇의 크기, 동역학 및 주변 장애물을 고려하면 실제로 실행하기 어렵거나 위험할 수 있다.

따라서 생성형 AI가 생성한 행동이나 경로를 그대로 실행하기보다 Cost Map, Collision Checker 및 Safety Controller 등의 안전 검증 계층과 결합할 필요가 있다. 특히 행동 · 이동지능에서는 생성된 Action이나 Path가 로봇의 물리적 제약과 주변 환경의 안전 조건을 만족하는지를 실시간으로 검증하는 것이 중요하다.

이는 기존 로봇공학이 생성형 AI로 대체된다는 의미가 아니다. 오히려 생성형 AI의 일반화 · 추론 능력과 기존 로봇공학의 실시간성 · 안전성을 결합하는 것이 실제 환경에서 동작하는 피지컬 AI 구현의 핵심이 될 것으로 예상된다.

5. 통합 구현 방향과 국내 연구 사례

실제 임무 수행을 위해서는 의미 추론 · 행동 생성과 학습 데이터, 실가상 평가, 탑재 실행 및 안전 제어를 함께 설계해야 한다. 이러한 구성은 특정 모델 구조에 한정되지 않으며, 적용 로봇의 센서 · 연산자원 · 제어 제약과 목표 임무에 맞춰 연결할 필요가 있다.

이러한 공통 구현 과제를 이동지능에 적용하는 국내 추진 사례로, ETRI에서는 보행 로봇 등 자율행동체의 이동 임무 수행을 위한 온디바이스 VLM 기반 이동지능 기술을 개발하고 있다. 공공정보와 오픈지식을 활용한 전역경로 생성, 이동공간 인지 · 추론을 위한 VLM 학습, 인지와 제어의 역할 분담을 고려한 실시간 안전경로 생성, K-NPU(국산 NPU) 기반 온디바이스 VLM 경량화, 정밀지도 의존성을 낮춘 관제 및 비평탄 환경 이동제어를 하나의 파이프라인으로 통합하는 것이 핵심이다. 이는 단일 VLM 모델 개발에 그치지 않고, 고수준 의미 추론에서 지역경로 생성, 실시간 제어, 온디바이스 실행 및 실제 로봇 실증까지 연결하는 Full-stack Physical AI 이동지능을 지향한다.

VII. 결론

본고에서는 VLA 기반 행동지능과 시각 이동정책에서 언어 · 공간 의미를 결합하는 VLN으로 이어지는 이동지능의 발전을 함께 살펴보았다. 조작과 이동은 시각 · 언어로 표현된 목표를 실제 행동과 연결한다는 과제를 공유하지만, 출력 행동과 공간 이력, 접촉 · 안정성 조건은 서로 다르다. 최근에는 생성형 행동정책, 추론과 실행의 역할 분담, 경험 기반 학습을 활용하여 이러한 차이를 연결하는 연구가 진행되고 있다.

향후 실제 임무 수행 능력을 높이기 위해서는 이동과 조작의 순차적 연결 및 전신 협응 수준을 구분하고, 데이터 · 실가상 검증 · 탑재 실행을 함께 설계해야 한다. 또한, 장기 수행 중 실패 감지와 복구, 사람의 개입 및 안전성을 평가하여 제한된 실험의 성공을 실제 운용 신뢰성으로 확장하는 과정이 필요하다.

용어해설

피지컬 AI(Physical AI) 센서로 물리 환경을 인지·이해하고, 판단 결과를 로봇 등의 실제 행동으로 연결하는 인공지능 기술

로봇 파운데이션 모델(Robot Foundation Model) 대규모 로봇·시각·언어 데이터로 사전학습하여 다양한 로봇과 작업에 활용하는 기반 모델

월드 모델(World Model) 현재 관측이나 상태, 행동 등의 조건으로 환경의 미래 변화를 예측하여 데이터 생성과 계획·검증 등에 활용하는 모델

VLM(Vision-Language Model) 영상과 언어 정보를 함께 처리하여 장면과 지시의 의미를 이해·추론하는 모델

VLA(Vision-Language-Action) 시각 관측과 언어 지시를 로봇의 실행 행동으로 연결하는 모델

VLN(Vision-Language Navigation) 시각 관측과 자연어 지시를 바탕으로 목적지까지 이동하는 기술

참고문헌

[1] L.H.K. Wong et al., "A Survey of Robotic Navigation and Manipulation with Physics Simulators in the Era of Embodied AI," arXiv preprint, 2026, doi: 10.48550/arXiv.2505.01458

[2] A. Sridhar et al., "NoMaD: Goal Masked Diffusion Policies for Navigation and Exploration," in Proc. IEEE Int. Conf. Robot. Autom., (Yokohama, Japan), 2024, pp. 63-70. doi: 10.1109/ICRA57147.2024.10610665

[3] M. Wei et al., "Ground Slow, Move Fast: A Dual-System Foundation Model for Generalizable Vision-and-Language Navigation," in Proc. Int. Conf. Learn. Represent., (Rio de Janeiro, Brazil), 2026.

[4] Y. Ma et al., "A Survey on Vision-Language-Action Models for Embodied AI," IEEE Trans. Neural Netw. Learn. Syst., vol. 37, no. 7, 2026, pp. 3031-3051. doi: 10.1109/TNNLS.2025.3650584

[5] A. O'Neill et al., "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," in Proc. IEEE Int. Conf. Robot. Autom., (Yokohama, Japan), 2024, pp. 6892-6903. doi: 10.1109/ICRA57147.2024.10611477

[6] R. Shao et al., "Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey," arXiv preprint, 2026, doi: 10.48550/arXiv.2508.13073

[7] E. Llontop and K. Vadrevu, "NVIDIA Isaac GR00T N1.7: Open Reasoning VLA Model for Humanoid Robots," Hugging Face Blog, 2026. 4. 17. https://huggingface.co/blog/nvidia/gr00t-n1-7

[8] J. Jang et al., "DreamGen: Unlocking Generalization in Robot Learning through Video World Models," in Proc. 9th Conf. Robot Learn., (Seoul, Republic of Korea), 2025, pp. 5170-5194.

[9] B.O. Han et al., "A Dual Process VLA: Efficient Robotic Manipulation Leveraging VLM," arXiv preprint, 2024, doi: 10.48550/arXiv.2410.15549

[10] Figure AI, "Helix: A Vision-Language-Action Model for Generalist Humanoid Control," Figure AI News, 2025. 2. 20. https://www.figure.ai/news/helix

[11] J. Bjorck et al., "GR00T N1: An Open Foundation Model for Generalist Humanoid Robots," arXiv preprint, 2025, doi: 10.48550/arXiv.2503.14734

[12] K. Black et al., "π0.5 : A Vision-Language-Action Model with Open-World Generalization," in Proc. 9th Conf. Robot Learn., (Seoul, Republic of Korea), 2025, pp. 17-40.

[13] M. Shukor et al., "SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics," arXiv preprint, 2025, doi: 10.48550/arXiv.2506.01844

[14] B. Zitkovich et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in Proc. 7th Conf. Robot Learn., (Atlanta, GA, USA), 2023, PMLR, vol. 229, pp. 2165-2183.

[15] K. Black et al., "π0 : A Vision-Language-Action Flow Model for General Robot Control," in Proc. Robot.: Sci. Syst., (Los Angeles, CA, USA), 2025. doi: 10.15607/RSS.2025.XXI.010

[16] K. Pertsch et al., "FAST: Efficient Action Tokenization for Vision-Language-Action Models," in Proc. Robot.: Sci. Syst., (Los Angeles, CA, USA), 2025. doi: 10.15607/RSS.2025.XXI.012

[17] K. Black et al., "Real-Time Execution of Action Chunking Flow Policies," Adv. Neural Inf. Process. Syst., vol. 38, 2025, pp. 37596-37620. doi: 10.52202/085713-1122

[18] S. Abeyruwan et al., "Gemini Robotics: Bringing AI into the Physical World," arXiv preprint, 2025, doi: 10.48550/arXiv.2503.20020

[19] A. Abdolmaleki et al., "Gemini Robotics 1.5: Pushing the Frontier of Generalist Robots with Advanced Embodied Reasoning, Thinking, and Motion Transfer," arXiv preprint, 2025, doi: 10.48550/arXiv.2510.03342

[20] StarVLA Community, "StarVLA: A Lego-like Codebase for Vision-Language-Action Model Developing," arXiv preprint, 2026, doi: 10.48550/arXiv.2604.05014

[21] J. Ye et al., "StarVLA-α: Reducing Complexity in Vision-Language-Action Systems," in Proc. 19th Eur. Conf. Comput. Vis., (Malmö, Sweden), 2026, LNCS, vol. 17037, pp. 351-369.

[22] NVIDIA, "Introduction," NVIDIA Cosmos Documentation. [Accessed 2026. 9. 21.] https://docs.nvidia.com/cosmos/latest/introduction.html

[23] Physical Intelligence et al., "π0.7: A Steerable Generalist Robotic Foundation Model with Emergent Capabilities," arXiv preprint, 2026, doi: 10.48550/arXiv.2604.15483

[24] Physical Intelligence et al., "π*0.6 . : A VLA That Learns From Experience," arXiv preprint, 2025, doi: 10.48550/arXiv.2511.14759

[25] P. Atreya et al., "RoboArena: Distributed Real-World Evaluation of Generalist Robot Policies," in Proc. 9th Conf. Robot Learn., (Seoul, Republic of Korea), 2025, pp. 336-364.

[26] R. Cadene et al., "LeRobot: State-of-the-art Machine Learning for Real-World Robotics in PyTorch," GitHub, 2024. https://github.com/huggingface/lerobot

[27] D. Shah et al., "GNM: A General Navigation Model to Drive Any Robot," in Proc. IEEE Int. Conf. Robot. Autom., (London, UK), 2023, pp. 7226-7233.

[28] D. Shah et al., "ViNT: A Foundation Model for Visual Navigation," in Proc. 7th Conf. Robot Learn., (Atlanta, GA, USA), 2023, PMLR, vol. 229, pp. 711-733.

[29] J. Kim et al., "CARE: Enhancing Safety of Visual Navigation through Collision Avoidance via Repulsive Estimation," in Proc. 9th Conf. Robot Learn., (Seoul, Republic of Korea), 2025, pp. 3704-3719.

[30] Y. Zeng et al., "NaviDiffusor: Cost-Guided Diffusion Model for Visual Navigation," in Proc. IEEE Int. Conf. Robot. Autom., (Atlanta, GA, USA), 2025, pp. 11994-12001. doi: 10.1109/ICRA55743.2025.11127466

[31] N. Hirose et al., "Learning to Drive Anywhere With Model-Based Reannotation," IEEE Robot. Autom. Lett., vol. 11, no. 2, 2026, pp. 1242-1249. doi: 10.1109/LRA.2025.3640424

[32] D. Song et al., "VLM-Social-Nav: Socially Aware Robot Navigation Through Scoring Using Vision-Language Models," IEEE Robot. Autom. Lett., vol. 10, no. 1, 2025, pp. 508-515. doi: 10.1109/LRA.2024.3511409

[33] A.J. Sathyamoorthy et al., "CoNVOI: Context-aware Navigation using Vision Language Models in Outdoor and Indoor Environments," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., (Abu Dhabi, United Arab Emirates), 2024, pp. 13837-13844. doi: 10.1109/IROS58592.2024.10802716

[34] B.O. Han et al., "Space-Aware Instruction Tuning: Dataset and Benchmark for Guide Dog Robots Assisting the Visually Impaired," in Proc. IEEE Int. Conf. Robot. Autom., (Atlanta, GA, USA), 2025, pp. 12430-12436. doi: 10.1109/ICRA55743.2025.11128120

[35] W.H. Yun et al., "SA-VLM V2: Useful, Comprehensive, and Concise Guidance for Guide-Dog Robots Assisting the Visually Impaired," in Proc. IEEE Int. Conf. Robot. Autom., (Vienna, Austria), 2026.

[36] J.Y. Lee et al., "Assistive Guidance System Based on Online Path Structure Recognition for the Visually Impaired," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., (Hangzhou, China), 2025, pp. 17178-17185. doi: 10.1109/IROS60139.2025.11247100

[37] H. Jiang et al., "WholeBodyVLA: Towards Unified Latent VLA for Whole-Body Loco-Manipulation Control," in Proc. Int. Conf. Learn. Represent., (Rio de Janeiro, Brazil), 2026.

[38] H. Yuan et al., "Being-0: A Humanoid Robotic Agent with Vision-Language Models and Modular Skills," arXiv preprint, 2025, doi: 10.48550/arXiv.2503.12533

[39] M. Liu et al., "Visual Whole-Body Control for Legged Loco-Manipulation," in Proc. 8th Conf. Robot Learn., (Munich, Germany), 2024, PMLR, vol. 270, pp. 234-257.