전자통신동향분석, 41권 6호 2026, 41-53

피지컬 AI가 여는 자율제조: 데이터 중심 AI에서 정밀 조작까지

Physical AI-Driven Autonomous Manufacturing: From Data-Centric AI to Precision Manipulation

한효녕 (H.N. Han, hyonyoung.han@etri.re.kr) 자율제조행동지능연구실 책임연구원
손영성 (Y.S. Son, ysson@etri.re.kr) 자율제조행동지능연구실 책임연구원
김록원 (R.W. Kim, rwkim@etri.re.kr) 자율제조행동지능연구실 책임연구원
안혜영 (H.Y. An, hya@etri.re.kr) 자율제조행동지능연구실 박사후연구원
이정인 (J.I. Lee, jilee@etri.re.kr) 자율제조행동지능연구실 선임연구원
손유호 (Y.H. Son, yhson@etri.re.kr) 자율제조행동지능연구실 선임연구원
배유성 (Y.S. Bae, bys724@etri.re.kr) 자율제조행동지능연구실 선임연구원
윤준완 (J.W. Yun, junwan.yun@etri.re.kr) 자율제조행동지능연구실 연구원
이은서 (E.S. Lee, eslee@etri.re.kr) 자율제조행동지능연구실 책임연구원

ABSTRACT

Manufacturing is moving from fixed automation to physical artificial intelligence (AI) that adapts to changing products, tasks, and environments. This article reviews autonomous manufacturing through two connected axes: data-centric AI and precision manipulation. It first examines industrial digital twins and simulation-based data generation for manufacturing. It then discusses data collection devices, cross-embodiment dataset platforms, and dataset quality assessment, followed by vision-language-action models, video-based motion learning, sim-to-real transfer, and force- and tactile-aware control for contact-rich tasks. This review shows that deployment depends less on a single model than on a closed loop that links data collection, curation, training, validation, and execution. The Key research needs include common action representations, physical and multimodal quality metrics, continuous-action policies, hierarchical real-time control, and lifecycle safety assurance.

KEYWORDS Autonomous Manufacturing, Data-Centric AI, Physical AI, Precision Manipulation, Robot Foundation Models, Sim2Real Transfer, Vision-Language-Action Models

* DOI: https://doi.org/10.22648/ETRI.2026.J.410604
* 본 연구 논문은 한국전자통신연구원 연구운영지원사업 [25ZR1100, 자율적으로 연결·제어·진화하는 초연결 지능화 기술 연구] 및 2024년도 산업통상자원부 및 산업기술기획평가원(KEIT) 연구비 지원 사업(RS-2024-00419641)의 일환으로 수행되었음.

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

I. 서론

제조업은 숙련 인력 감소, 다품종 소량생산, 짧아진 제품 수명주기, 공급망 변동에 동시 대응해야 하며, 이에 따라 로봇 도입도 빠르게 늘고 있다. 국제로봇연맹(IFR)에 따르면 2024년 한 해 동안 전 세계 공장에 새로 설치된 산업용 로봇은 약 54만 2천 대로 10년 전의 두 배 이상이고, 가동 중인 산업용 로봇은 약 466만 대에 이른다[1]. 기존 산업용 로봇은 정형 환경에서 높은 속도와 정밀도를 제공하지만, 작업물과 공정이 바뀔 때마다 티칭과 재프로그래밍이 필요하다. 수학적 모델에 기반한 제어는 안정성과 설명 가능성이 높지만 미지의 변화에 대한 적응이 제한적이고, 데이터 기반 학습은 적응성을 높이는 대신 대규모 데이터와 연산을 요구한다.

파운데이션 모델은 이 두 접근 사이의 역할 분담을 바꾸고 있다. 대규모 언어 모델과 비전-언어 모델이 자연어 지시, 장면 의미, 장기 계획을 담당하고, 비전-언어-행동(VLA: Vision-Language-Action) 모델은 행동 출력을 결합해 지각 · 추론 · 제어를 하나의 학습 체계로 연결한다. 물리 세계에서 이러한 모델이 센서와 구동기를 통해 판단하고 행동하는 체계를 피지컬 AI(Physical AI)라고 한다[2].

본고에서 자율제조는 로봇이 작업과 환경의 변화를 스스로 인식하고 판단하여 작업 방식을 조정하며, 실행 결과를 다시 학습에 반영하는 제조 체계를 뜻한다. 피지컬 AI는 티칭과 재프로그래밍에 의존하던 기존 자동화를 언어 지시와 소량의 시연, 현장 데이터로 새로운 작업에 대응하는 방식으로 바꿀 수 있어, 다품종 소량생산의 셋업 부담을 줄이고 정형화하기 어려운 조립 · 삽입 공정까지 자동화의 범위를 넓힐 가능성을 열어 준다.

자율제조의 병목은 모델 크기보다 데이터와 실행의 연결에 있다. 로봇 데이터는 수집 비용이 많이 들고 로봇마다 자유도, 센서, 행동 공간이 다르다. 조립 · 삽입 · 배선과 같은 접촉 작업은 시각만으로 상태를 판단하기 어렵고, 행동을 이산 토큰으로 표현하면 필요한 정밀도가 손실될 수 있다. 따라서 본고는 자율제조 기술을 데이터의 수집 · 통합 · 품질관리와 접촉 기반 정밀 조작이라는 두 축으로 살펴본다. 전자는 모델 구조보다 학습 데이터의 품질과 구성을 체계적으로 개선하여 성능을 높이는 데이터 중심 AI(Data-Centric AI)의 관점에 해당한다. 아울러 두 축을 폐루프로 연결하기 위한 기술 과제를 제시한다. Ⅱ장에서는 제조 현장의 로봇 기술과 가상 플랫폼 동향을, Ⅲ장에서는 데이터 수집 · 통합 · 품질관리를, Ⅳ장에서는 VLA 모델과 영상 기반 동작 학습, AI 기반 정밀 제어를 다루며, Ⅴ장에서는 결론과 향후 과제를 제시한다.

II. 자율제조 기술 동향

이 장에서는 자율제조가 적용되는 제조 현장의 로봇 기술 현황과, 학습 데이터 생성과 정책 검증의 무대가 되는 가상 플랫폼 동향을 살펴본다.

1. 제조 현장 적용을 위한 로봇 기술

제조 로봇의 적용 범위는 고정 자동화에서 유연 자동화로 확대되고 있다. 협동로봇과 자율이동로봇은 작업자와 공간을 공유하며 머신텐딩(Machine Tending, 가공설비에 소재를 넣고 꺼내는 작업), 물류 이송, 팔레타이징을 수행한다. 그러나 제품 변경에 따른 티칭과 치공구 교체 부담은 여전히 크다. VLA 기반 로봇은 대규모 사전학습을 통해 새로운 물체와 자연어 지시에 대응하고, 현장 데이터로 미세조정해 라인별 프로그래밍 부담을 줄이는 방향으로 발전하고 있다[2].

표 1 제조 공정별 요구 특성과 관련 기술

공정 유형	요구 특성	필요한 데이터	핵심 기술(본고 절)
자재 운반, 키팅, 부품 안착	실패 위험이 낮고 시각 중심으로 수행 가능하며 작업 종류가 다양함	시각 시연, 작업 단계 레이블	VLA, 영상 기반 동작 학습 (Ⅳ장 1절)
육안 검사	조명·차폐 조건이 변하고 불량 사례가 드묾	다양한 조건의 영상, 실패·불량 사례	합성 데이터, 데이터 품질평가 (Ⅱ장 2절, Ⅲ장 3절)
부품 삽입·체결	밀리미터 이하의 정밀도, 접촉 상태의 급격한 변화	힘/토크·촉각이 동기화된 시연, 접촉 사건 정렬	힘·촉각 통합 VLA, 계층형 정밀 제어(Ⅳ장 1절, 2절)
와이어 하네스, 유연물 조작	물체의 변형과 부분 관측, 복잡한 접촉	촉각·양손 시연, 변형체 시뮬레이션	촉각 VLA, 유연체 모델링 (Ⅱ장 2절, Ⅳ장 1절)

휴머노이드 로봇은 사람을 기준으로 설계된 작업 공간을 크게 바꾸지 않고 투입할 수 있다는 장점이 있다. 현재 산업 파일럿은 자재 운반, 키팅(Kitting, 부품을 작업 단위로 묶어 공급하는 작업), 부품 안착, 육안 검사처럼 실패 위험이 비교적 낮은 작업에 집중된다. 와이어 하네스, 유연물 조작, 협소 공간 체결과 같이 접촉 상태가 복잡한 공정은 힘 · 촉각 센싱, 실시간 제어, 안전 검증이 함께 필요해 적용 범위가 제한적이다. 제조 공정별 요구 특성과 이를 뒷받침하는 데이터 · 제어 기술은 표 1과 같이 정리할 수 있다.

현장 적용의 핵심은 로봇 한 대의 성능보다 운영 폐루프에 있다. 원격조작으로 시연을 수집하고, 성공 · 실패와 작업 단계를 정렬해 학습 데이터로 만들며, 배치 후 발생한 예외를 다시 수집하는 체계가 필요하다. 자율제조는 범용 모델을 도입하는 일회성 프로젝트가 아니라 데이터 수집, 정책 갱신, 공정 검증을 반복하는 운영 과정으로 보아야 한다.

2. 제조 가상 플랫폼과 시뮬레이터

제조 가상 플랫폼은 공장 레이아웃과 사이클타임을 검증하는 3차원 도구에서 실제 설비 데이터와 연계되는 디지털 트윈(Digital Twin)으로 확장되고 있다[3,4]. Siemens Process Simulate와 Dassault Systèmes DELMIA는 로봇 셀 모델링, 도달성 · 충돌 · 사이클타임 분석, 오프라인 프로그래밍, 가상 시운전을 지원한다[5,6]. 이를 이용하면 설비를 구축하거나 변경하기 전에 공정과 제어 로직을 검증할 수 있다.

피지컬 AI의 확산으로 시뮬레이터는 학습 데이터 생성과 정책 검증 환경의 역할도 맡는다. NVIDIA Factory는 체결 · 삽입 · 기어 조립처럼 접촉이 중요한 작업을 GPU 물리 시뮬레이션으로 구현하고(그림 1)[7], Isaac Lab은 병렬 물리 시뮬레이션, 센서 모델, 데이터 수집, 도메인 랜덤화를 결합한다[8]. 실제 현장에서 반복하기 어렵거나 위험한 실패 조건을 가상환경에서 생성하면 데이터의 범위와 안전성을 동시에 높일 수 있다.

향후 플랫폼은 가상 시운전, 로봇 · 센서 물리 모델, 합성데이터 생성, 정책 평가를 하나의 흐름으로

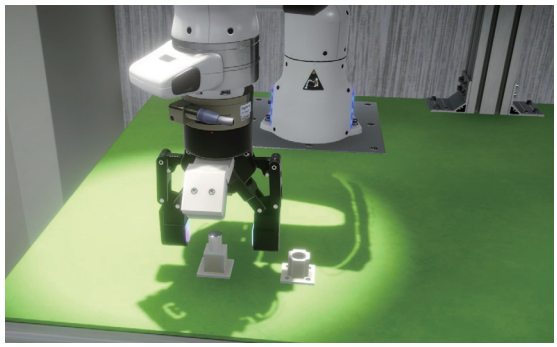


그림 1 정밀 로봇 조립 시뮬레이션 환경

연결해야 한다. 이를 위해서는 실제 설비와 가상 모델의 실시간 동기화, 접촉 · 마찰 · 변형의 고충실도 모델링, 제조 데이터 모델 간 상호운용성이 필요하다. 시뮬레이션은 실제 데이터를 대체하기보다 부족한 조건을 보완하고 시뮬레이션-실제 전이(Sim2Real)의 오차를 체계적으로 관리하는 수단으로 활용해야 한다.

III. 데이터 수집과 관리

이 장에서는 조작 학습 데이터를 수집하는 장치, 서로 다른 데이터를 통합하는 플랫폼, 데이터의 품질을 평가하는 방법의 순서로 데이터 수집과 관리 기술을 살펴본다.

1. 조작 데이터 수집 장치와 기술

로봇 조작 학습 데이터는 작업 장면의 관측, 로봇 행동, 내부 상태가 동기화된 시계열이다. 접촉 중심 제조 작업에서는 RGB 또는 RGB-D 영상, 관절과 엔드이펙터(End-Effector) 자세, 그리퍼 상태, 힘/토크(Force/Torque), 촉각을 같은 시간 축에 맞추어야 한다. 수집 장치는 작업자의 자연스러운 시연과 정확한 행동 레이블을 함께 제공해야 한다.

수집 장치는 로봇을 직접 조작하는 방식과 로봇 없이 사람의 시연을 기록하는 방식으로 나눌 수 있으며, 후자는 다시 휴대형, 착용형, 전신 추적형으로 세분된다(그림 2). 티치 펜던트, 직접교시, 조이스틱, 3차원 마우스, 가상현실 조작기는 범용성이 높지만 입력 공간과 로봇 관절 공간의 변환, 조작 피로, 지연이 데이터 품질에 영향을 준다. 리더-팔로워형은 조작기와 로봇의 구조를 맞추어 이 부담을 줄인다. ALOHA와 ALOHA 2는 정밀 양손 시연을 수집하고[9,10], GELLO는 대상 로봇과 유사한 운동학 구조의 저비용 조작기로 직관적인 제어를 제공한다[11].

휴대형과 착용형 장치는 특정 로봇과 작업셀에 대한 종속성을 낮춘다. UMI는 카메라가 장착된 휴대형 그리퍼로 현장 시연을 기록한다[12]. 한국전자통신연구원도 실제 로봇 그리퍼와 유사한 착용형 장치 EVO를 개발하여, 손목 RGB-D 영상과 양 손끝의 촉각 정보를 동기화하고 접촉이 많은 양손 조작 시연을 수집 · 전이하는 방법을 제시하였다[13]. DEXOP은 손 외골격과 촉각을 이용해 손가락 단위의 접촉 조작을 수집한다[14]. 전신 추적형으로는 TWIST2가 가상현실 기반 전신 추적과 1인칭 영상



출처 참고문헌 [9-15]를 바탕으로 재구성.

그림 2 로봇 조작 데이터 수집 장치와 발전 방향

을 결합해 이동과 양손 조작을 포함한 휴머노이드 시연으로 범위를 넓힌다[15].

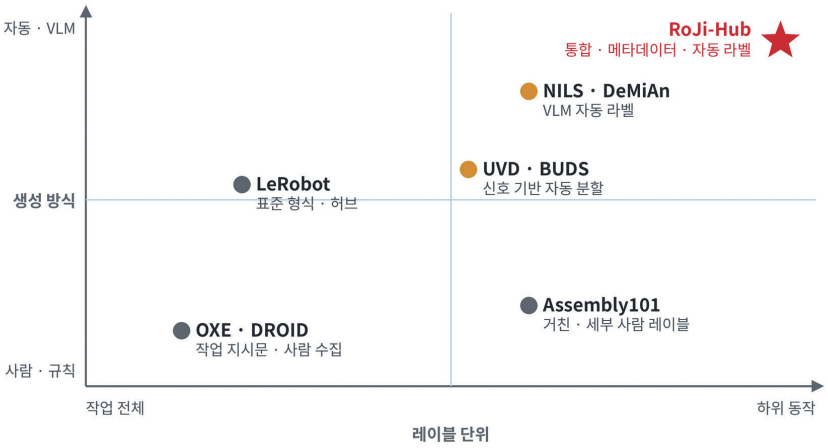
제조 현장에서는 데이터 규모와 함께 재현성과 정합성을 관리해야 한다. 카메라-로봇 좌표계 보정, 센서 간 타임스탬프 동기화, 접촉 사건 정렬, 성공 여부와 작업 단계의 레이블을 수집 시점부터 기록해야 한다. 이 정보가 갖추어져야 서로 다른 장치에서 얻은 데이터를 공통 행동 표현으로 변환하고 VLA 모델과 정밀 조작 정책에 재사용할 수 있다.

2. 로봇 조작 데이터셋 플랫폼

대규모 로봇 학습에서는 데이터 형식과 의미 구조가 새로운 병목이다. Open X-Embodiment는 여러 기관과 로봇의 데이터를 공통 형식으로 통합해 교차 임바디먼트 학습의 가능성을 보였고[16], DROID는 다양한 실내 환경에서 수집한 7만 6천 개의 시연 궤적을 공개했다[17]. 그러나 로봇, 센서, 행동 공간, 저장 구조가 서로 다르고, 많은 에피소드에 작업 전체를 설명하는 짧은 지시문만 있어 필요한 데이터를 검색하거나 하위 동작 단위로 학습하기 어렵다.

LeRobot은 수집 · 저장 · 스트리밍 · 학습을 하나의 오픈소스 스택으로 연결하고 표준 데이터셋 형식을 제공한다[18]. 형식 통합 다음 단계는 장면, 로봇 형태, 행동 공간, 작업 유형, 제어 주기, 제약 조건을 메타데이터로 관리하는 것이다. 긴 시연은 학습 가능한 하위 동작으로 나누어야 한다. BUDS는 다중감각 신호를 군집화해 반복 스킬을 찾고[19], UVD는 사전학습 시각 표현의 변화로 하위 목표 경계를 검출한다[20]. NILS는 파운데이션 모델로 경계와 자연어 레이블을 함께 생성하며[21], DeMiAn은 동작 · 장면 · 자세 · 추론 관점의 밀집 주석이 정책 학습에 주는 효과를 다룬다[22]. Assembly101은 거친 단계와 세부 동작을 함께 제공하는 다층 레이블의 선례이다[23]. 데이터셋 플랫폼과 레이블링 연구의 위치는 그림 3과 같이 정리할 수 있다.

이러한 한계를 해결하기 위해 한국전자통신연구원에서는 이질적인 로봇 데이터를 공통 스키마와 메타데이터로 정규화하고, 에피소드를 하위 동작 단위로 구조화하여 데이터 간 검색 · 결합 · 학습을 지원하는 통합 데이터 플랫폼을 개발하고 있다. 이



출처 참고문헌 [16-23]을 바탕으로 재구성.

그림 3 로봇 조작 데이터 수집 장치와 발전 방향

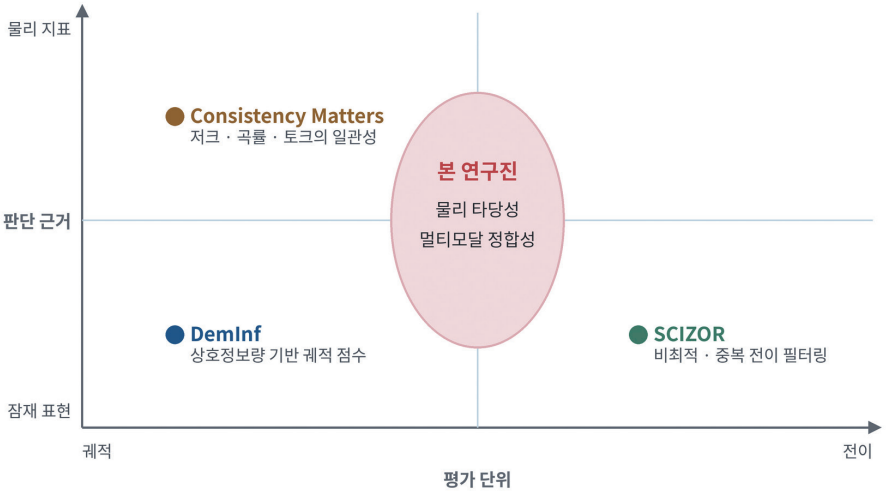
와 같은 데이터 표준화와 의미 단위 구조화는 서로 다른 데이터셋을 동일한 기준으로 활용하고, 학습 결과의 재현성과 로봇 간 전이 가능성을 높이기 위한 기반이 된다.

3. 데이터 품질

로봇 학습 데이터가 커질수록 저품질 시연, 잘못된 레이블, 중복 샘플, 작업 범위의 공백이 함께 늘어난다. 실무자 700여 명을 대상으로 한 산업 설문에서도 모델 성능 미달의 주요 원인으로 데이터 부족과 차폐 · 조명 · 아티팩트 등 품질 문제가 지적되었다[24]. ISO/IEC 5259 시리즈는 분석 · 기계학습 데이터의 품질 측정, 관리 요구사항, 시각화 프레임워크를 제시하며 품질관리를 정량적 절차로 확장하고 있다[25].

최근 연구는 정책을 학습한 뒤 성공률로 품질을 확인하는 방식에서 벗어나 학습 전에 데이터를 평가하며, 평가 방식은 궤적 단위의 정보량, 전이 단위의 자기지도 선별, 물리 지표의 일관성으로 나뉜다(그림 4). DemInf는 궤적 단위의 상태-행동 상호정보량으로 다양성과 행동 예측 가능성을 점수화한다[26]. SCIZOR는 평가 단위를 개별 전이로 낮추어 비최적 구간과 중복 구간을 자기지도 방식으로 제거한다[27]. Consistency Matters는 저크, 곡률, 관절 토크 등 물리적으로 해석 가능한 지표의 일관성이 학습 성능을 예측할 수 있음을 보였다[28].

제조 데이터의 품질은 단일 궤적만으로 판단하기 어렵다. 앞서 소개한 방법은 주로 상태-행동 궤적과 관절 수준의 운동 · 토크 특성을 평가 대상으로 하므로, 비전 · 힘/토크 · 촉각이 함께 구성되는 접촉 중심 제조 데이터의 공정 제약 만족 여부와 이종 센서 간 정합성까지는 충분히 다루지 못한다. 이에 한국전자통신연구원은 비전, 힘/토크, 로봇 궤적이 공정 제약과 물리 법칙을 따르는 정도를 측정하는 물리 타당성 지표와 이종 센서 간 시간적 · 의미적 일치도를 측정하는 멀티모달 정합성 지표를 함께 개발하고 있다. 품질 점수는 폐기 여부만 결정하는 값이 아니라 재수집, 가중 학습, 레이블 수정, 센서 보정으로 이어지는 운영 신호가 되어야 한다.



출처 참고문헌 [26-28]을 바탕으로 재구성.

그림 4 학습 전 로봇 데이터 품질평가 방법의 유형

IV. 인공지능 정밀 조작 제어

Ⅲ장에서 다룬 데이터의 수집 · 통합 · 품질관리는 이 장에서 다루는 모델과 제어기의 학습 기반이 된다. 이 장에서는 VLA 모델과 영상 기반 동작 학습으로 작업의 의도와 목표를 얻는 방법과, 이를 접촉이 많은 정밀 조작으로 실행하기 위한 AI 기반 제어 기술을 순서대로 살펴본다. 전자가 무엇을 할 것인가를 다룬다면, 후자는 어떻게 정밀하게 실행할 것인가를 다룬다.

1. VLA 모델과 영상 기반 로봇 동작 학습

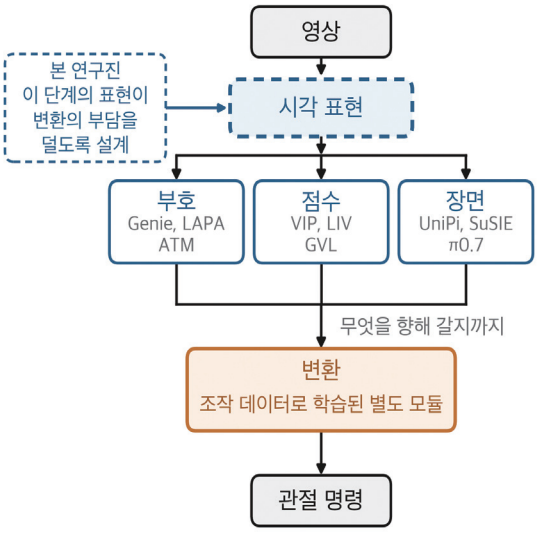
비전-언어-행동(VLA) 모델은 대규모로 사전학습된 비전-언어 모델(VLM: Vision-Language Model)을 기반으로 영상과 언어 지시로부터 로봇의 행동을 직접 생성하며, RT-2와 OpenVLA가 대표적인 초기 모델이다[29,30]. 최근의 대표적인 모델은 다음과 같다.

Physical Intelligence의 $\pi 0$는 사전학습된 VLM 위에 플로 매칭(Flow Matching) 기반 행동 전문가를 결합해 여러 로봇의 데이터로 연속적인 행동 묶음을 생성하는 범용 정책이며[31], 후속 모델인 $\pi 0.7$은 언어 지시뿐만 아니라 수행 방식을 설명하는 메타데이터와 하위 목표 영상 등 다양한 맥락 정보를 조건으로 주어 동작을 조종할 수 있게 하였다[32]. NVIDIA의 GR00T N1은 영상과 언어를 해석하는 System 2와 확산 트랜스포머로 행동을 실시간 생성하는 System 1을 함께 학습하는 이중 시스템 구조의 개방형 휴머노이드 파운데이션 모델로, 실제 로봇 궤적, 인간 영상, 합성 데이터를 혼합하여 학습한다[33]. Google DeepMind의 Gemini Robotics는 Gemini 2.0을 기반으로 개방형 언어 지시에 대응하는 VLA 모델이며, 공간 · 시간 이해를 강화한 Gemini Robotics-ER와 함께 제시되었고 적게는 100개의 시연만으로 특정 작업에 미세조정할 수 있음을 보였다[34].

이들 모델은 언어 지시와 소량의 시연으로 새로운 작업에 대응하지만, 기본 구조가 시각과 로봇 상태를 중심으로 하므로 삽입 · 체결처럼 접촉 상태가 결과를 좌우하는 공정에서는 한계가 있다. 이를 보완하기 위해 촉각과 힘 정보를 VLA에 통합하는 연구가 최근 빠르게 늘고 있다. ForceVLA는 6축 힘/토크 신호를 독립된 모달리티로 보고 전문가 혼합(MoE: Mixture of Experts) 방식의 융합 모듈로 VLA에 결합하여, $\pi 0$ 기반 기준 모델보다 성공률을 23.2% 향상시키고 플러그 삽입 등에서 최대 80%의 성공률을 보고하였다[35]. VTLA(Vision-Tactile-Language-Action)는 시각 · 촉각 · 언어를 함께 입력받고 직접 선호 최적화(DPO: Direct Preference Optimization)로 토큰 예측 손실과 연속 로봇 제어를 연결하여, 핀 삽입 작업에서 처음 보는 핀 형상에도 90% 이상의 성공률을 보이고 실제 삽입 실험으로 전이하였다[36]. OmniVTLA는 시각 기반과 힘 기반 촉각 센서를 모두 다루는 이중 경로 촉각 인코더로 촉각의 의미를 언어 · 시각과 정렬하고, 13만 5천 개 샘플의 촉각 데이터셋 ObjTac을 구축하여 물체 파지 · 이동 과제에서 그리퍼 96.9%, 다지 손 100%의 성공률을 보고하였다[37]. 다만 이들 연구는 제한된 과제와 실험실 환경에서 검증되어, 산업 현장 수준의 신뢰성과 반복 정밀도의 검증은 남은 과제이다.

VLA 모델의 학습에는 대규모 로봇 시연이 필요하지만, 로봇별 시연을 반복 수집하는 비용이 많이 든다. 새로운 작업을 수행하려면 정밀한 로봇 제어기와 함께, 현재 장면에서 어떤 변화를 만들어야 하는지 나타내는 동작 신호가 필요하며, 최근 연구는 대규모 영상에서 변화 규칙을 먼저 학습하고 소량의 로봇 데이터로 행동 공간에 연결한다.

영상의 동작 신호는 네 유형으로 정리할 수 있다



출처 참고문헌 [32,38-48]을 바탕으로 재구성.

그림 5 영상에서 로봇 동작으로 이어지는 공통 변환 구조

(그림 5). 잠재 행동 방식은 연속 장면의 변화를 압축한 표현을 학습한 뒤 로봇 명령과 연결한다[38-41]. 가치 · 진행도 방식은 작업이 얼마나 진행되었는지를 점수화해 정책의 목표로 사용하고[42-44], 미래 장면 방식은 생성 모델이 다음 목표 영상을 만들고 저수준 정책이 그 장면에 도달하도록 한다[32,45,46]. 명시적 모션 방식은 점 궤적이나 픽셀 모션처럼 장면 안의 움직임을 직접 예측해 동작 의도로 사용한다[47,48]. 네 방식 모두 영상의 의미를 실제 관절 명령으로 바꾸는 로봇별 어댑터가 필요하다.

이 가운데 명시적 모션 방식은 잠재 표현보다 해석이 쉽고 로봇 동작으로의 변환 경로가 분명하다는 점에서 주목받고 있다. Any-Point Trajectory Modeling은 영상에서 점 궤적을 예측해 정책의 조건으로 사용하며[47], DAWN은 픽셀 모션을 고수준 동작 의도와 저수준 로봇 제어를 연결하는 해석 가능한 중간 표현으로 활용한다[48]. DAWN에 이어 한국전자통신연구원은 Stony Brook University와의 후속 공동 연구를 통해 영상으로부터 물체와 로봇의 모션 방향을 예측하고, 이를 실행 가능한 동작으로 변환하는 모델과 로봇 적용 시스템을 개발하고 있다. 다만 이러한 영상 표현만으로 접촉력과 실행 안정성을 보장하기 어려우므로, 그림 5와 같이 영상 기반 동작 추론, 로봇별 행동 변환, 정밀 제어를 계층적으로 구성하고 각 단계의 오차를 검증해야 한다.

2. 인공지능 기반 정밀 제어

조립, 삽입, 체결과 같은 접촉 집약 작업은 전통적으로 임피던스 제어와 힘 제어로 수행되었다. 이 방식은 안정적이지만 접촉 · 마찰이 모델과 다르거나 공정 조건이 바뀌면 파라미터를 다시 조정해야 한다. 학습 기반 제어는 실제 상호작용에서 접촉전략을 익혀 변화에 적응하지만, 시각과 로봇 상태만으로는 미세한 접촉 상태와 국부 힘을 직접 관측하기 어렵다.

최근 연구는 비전, 로봇 상태, 힘/토크, 촉각을 결합한다. FMB는 다중 형상의 부품 삽입에서 접촉 정보를 활용하는 일반화 가능한 학습 환경을 제시했고[49], PolyTouch는 시각 기반 촉각을 확산 정책과 결합했다[50]. Reactive Diffusion Policy는 저주파 동작 생성과 고주파 촉각 반응을 분리해 빠른 접촉 변화에 대응하며[51], TacDiffusion은 힘/토크 명령을 직접 생성해 정밀 삽입 작업으로 일반화한다[52]. 촉각 · 힘 정보를 VLA 모델과 결합하는 흐름은 앞 절에서 살펴본 바와 같다.

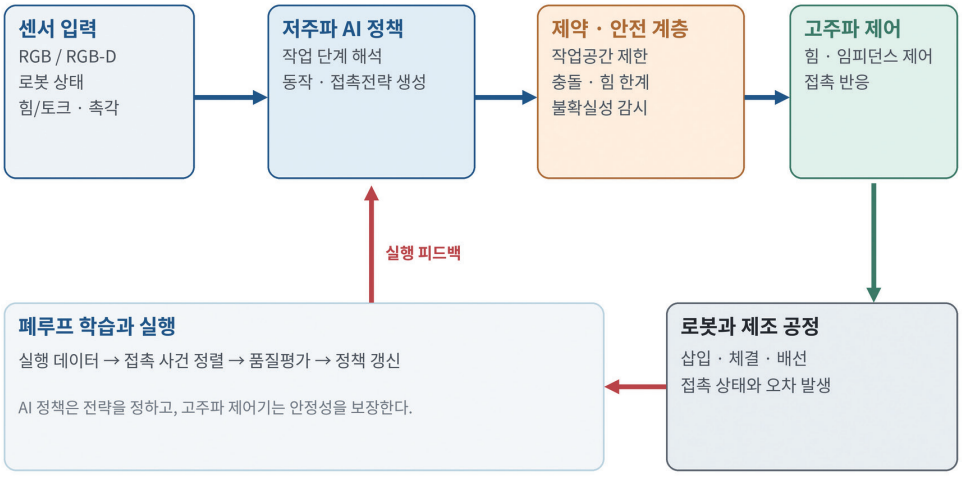
이와 달리 시뮬레이션에서 정책을 학습해 실제 로봇으로 전이하는 접근도 조립 공정에서 활발하다. IndustReal은 시뮬레이션 인식 정책 갱신, 부호거리장(Signed Distance Function) 기반 보상, 샘플링 기반 커리큘럼, 정책 수준의 행동 적분기를 결합하여

삽입 등 접촉이 많은 조립 작업을 시뮬레이션에서 실제 로봇으로 전이하였다[53]. 이는 Ⅱ장 2절에서 소개한 접촉 중심 시뮬레이션과 같은 흐름에서 정밀 조립 정책의 시뮬레이션-실제 전이를 다룬 사례이다.

정책이 출력하는 행동의 표현 방식도 정밀도를 좌우한다. RT-2와 OpenVLA는 행동을 이산 토큰으로 바꾸어 언어 모델의 다음 토큰 예측으로 생성하므로 사전학습된 비전-언어 모델을 그대로 활용할 수 있다[29,30]. 그러나 단순한 이산화는 고주파 · 고정밀 동작에서 성능이 떨어지는 문제가 있어, FAST는 행동 궤적을 주파수 영역에서 압축해 토큰화하는 방법을 제안했다[54]. 반면 π0는 플로 매칭 기반의 행동 전문가로 연속적인 행동 묶음(Action Chunk)을 생성해 고주파 제어를 지원한다[31]. 미세한 조정이 필요한 접촉 공정에서는 이산화에 따른 정밀도 손실을 피할 수 있는 연속 행동 정책이 유리할 수 있으며, 앞서 소개한 확산 기반 촉각 · 힘 정책도 같은 흐름에 있다.

현장 적용에는 학습 정책과 고전 제어의 계층화가 적합하다고 판단하며, 본고에서는 이를 그림 6과 같은 구조로 정리한다. AI 정책은 작업 단계와 접촉전략을 저주파로 결정하고, 고주파 힘 · 위치 · 임피던스 제어기는 안정성과 정밀도를 보장한다. 제약 · 안전 계층은 작업공간, 충돌, 허용 접촉력, 불확실성을 감시한다. 실행 중 수집한 접촉 사건과 오차는 데이터 품질평가를 거쳐 정책 갱신에 사용한다. 이러한 구조는 데이터와 제어를 하나의 폐루프로 연결한다.

제조 현장에 배치하려면 성능뿐만 아니라 안전과 검증 체계도 갖추어야 한다. 학습 정책은 평균 성공률만으로 승인할 수 없으므로, 디지털 트윈에서 정상 · 고장 · 교란 조건을 체계적으로 시험하고 실제 셀에서는 힘 한계, 충돌 회피, 작업공간 제한을 학습 정책과 독립된 안전 계층으로 보장해야 한다. 산업용 로봇의 안전 요구사항은 ISO 10218 시리즈로 규정되어 있으며[55], 분포 밖 입력이나 언어 지시의 모호성을 감지하면 정지, 재계획, 작업자 인계로 전환하는 기준도 필요하다. 배치 후에는 데이터 분포와 성능의 변화를 감시하고 모델 · 데이터 · 설정의



출처 참고문헌 [49-52]를 바탕으로 재구성.

그림 6 멀티모달 계층형 정밀 조작 제어 구조

버전을 함께 관리하며, 정책이 갱신될 때마다 회귀 시험을 시행하여 실패 기록을 다음 학습 주기에 반영해야 한다.

V. 결론 및 향후 과제

본고에서는 피지컬 AI 기반 자율제조를 데이터의 수집 · 통합 · 품질관리와 접촉 기반 정밀 조작이라는 두 축으로 나누어 최근 기술 동향을 분석하였다. 제조 로봇은 고정 자동화에서 VLA 기반 유연 자동화로 확대되고 있으며, 가상 플랫폼은 설비 검증 도구에서 학습 데이터 생성과 정책 검증 환경으로 역할이 넓어지고 있다. 데이터 측면에서는 특정 로봇에 대한 종속성을 낮춘 수집 장치, 형식 통합을 넘어 의미 구조화를 지향하는 데이터셋 플랫폼, 학습 전에 데이터를 평가하는 품질평가 기술이 등장하였다. 정밀 조작 측면에서는 VLA 모델이 언어 지시와 소량의 시연으로 새로운 작업에 대응하는 범용성을 보이고, 영상에서 동작 신호를 얻는 방식은 잠재 행동, 가치 · 진행도, 미래 장면, 명시적 모션의 네 유형으로 정리된다. 접촉 작업에서는 촉각과 힘 정보를 VLA에 통합하는 연구가 늘고 있으며, AI 정책과 고주파 제어기를 계층적으로 구성하는 방향이 제시되고 있다.

이러한 기술을 제조 현장에 적용하기 위해서는 다음 과제가 해결되어야 한다. 첫째, 수집 장치와 로봇이 달라도 데이터를 비교하고 결합할 수 있는 공통 스키마와 행동 표현, 그리고 물리 타당성과 멀티모달 정합성을 함께 평가하는 품질 지표가 필요하며, 품질 점수가 재수집과 가중 학습으로 이어지는 데이터 폐루프로 운영되어야 한다. 둘째, 접촉 작업이 요구하는 밀리미터 이하의 정밀성과 실시간성을 위해 과도한 이산화로 정밀도를 잃지 않는 연속 행동 정책과, 저주파 AI 정책과 고주파 제어기를 결합한 계층형 제어 구조가 필요하다. 셋째, 제조용 피지컬 AI는 평균 성공률만으로 승인할 수 없으므로 디지털 트윈 기반 시험, 독립적인 안전 계층, 배치 후 성능 감시와 회귀시험을 포함하는 수명주기 검증 체계를 갖추어야 한다.

종합하면, 자율제조의 실용화는 범용 모델의 성능 경쟁보다 데이터 폐루프, 계층형 정밀 제어, 수명주기 검증을 하나의 시스템으로 통합할 때 가능하다. 이러한 과제가 해결될 때 피지컬 AI는 티칭과 재프로그래밍에 의존하던 자동화를 넘어, 다품종 소량생산 환경에서도 스스로 작업을 조정하고 학습하는 자율제조의 기반이 될 것이다.

용어해설

피지컬 AI 센서로 물리 환경을 인식하고 로봇·기계의 행동을 생성하며 실행 결과를 다시 학습에 반영하는 인공지능 체계

교차 임바디먼트 서로 다른 로봇의 자유도, 센서, 행동 공간 차이를 넘어 지식이나 정책을 전이하는 학습 개념

시뮬레이션-실제 전이 시뮬레이션에서 학습한 정책을 실제 로봇에 적용할 때 물리 모델과 센서 차이를 줄이는 기술

비전-언어-행동(VLA) 모델 카메라 영상과 자연어 지시를 입력받아 로봇의 행동을 직접 출력하는 파운데이션 모델

디지털 트윈 실제 설비와 공정을 가상 공간에 동기화된 모델로 구현하여 검증, 분석, 제어에 활용하는 기술

데이터 폐루프 데이터 수집, 정제, 학습, 검증, 배치, 재수집이 순환하며 데이터와 정책을 함께 갱신하는 운영 체계

참고문헌

[1] International Federation of Robotics, "World Robotics 2025 Report – Industrial Robots – released by IFR," IFR Press Release, 2025. 9. 25. https://ifr.org/ifr-press-releases/news/global-robot-demand-in-factories-doubles-over-10-years

[2] A. Psiris et al., "Foundation Models in Robotics: A Comprehensive Review of Methods, Models, Datasets, Challenges and Future Research Directions," arXiv preprint, 2026, doi: 10.48550/arXiv.2604.15395

[3] C. Cimino et al., "Review of Digital Twin Applications in Manufacturing," Comput. Ind., vol. 113, 2019, doi: 10.1016/j.compind.2019.103130

[4] Q. Qin et al., "Robot Digital Twin Systems in Manufacturing: Technologies, Applications, Trends and Challenges," Robot. Comput. Integr. Manuf., vol. 97, 2026, doi: 10.1016/j.rcim.2025.103103

[5] Siemens, "Process Simulate," Siemens Digital Industries Software. https://www.siemens.com/en-us/products/tecnomatix/process-simulate-software/ [Accessed Sep. 14, 2026.]

[6] Dassault Syst mes, "Virtual Twin for Manufacturing," DELMIA. https://www.3ds.com/products/delmia/virtual-twin-manufacturing [Accessed Sep. 14, 2026.]

[7] O. Omotuyi, "Advancing Robotic Assembly with a Novel Simulation Approach Using NVIDIA Isaac," NVIDIA Technical Blog, 2022. 7. 8. https://developer.nvidia.com/blog/advancing-robotic-assembly-with-a-novel-simulation-approach-using-nvidia-isaac/

[8] M. Mittal et al., "Isaac Lab: A GPU-Accelerated Simulation Framework for Multi-Modal Robot Learning," arXiv preprint, 2025, doi: 10.48550/arXiv.2511.04831

[9] T.Z. Zhao et al., "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in Proc. Robot.: Sci. Syst., (Daegu, Republic of Korea), July 2023, doi: 10.15607/RSS.2023.XIX.016

[10] ALOHA 2 Team et al., "ALOHA 2: An Enhanced Low-Cost Hardware for Bimanual Teleoperation," arXiv preprint, 2024, doi: 10.48550/arXiv.2405.02292

[11] P. Wu et al., "GELLO: A General, Low-Cost, and Intuitive Teleoperation Framework for Robot Manipulators," in Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst., (Abu Dhabi, United Arab Emirates), 2024, pp. 12156-12163, doi: 10.1109/IROS58592.2024.10801581

[12] C. Chi et al., "Universal Manipulation Interface: In-the-Wild Robot Teaching without In-the-Wild Robots," in Proc. Robot.: Sci. Syst., (Delft, Netherlands), July 2024, doi: 10.15607/RSS.2024.XX.045

[13] H. Han et al., "Ergonomic Multimodal Data Collection System for Versatile Robot Manipulation Tasks," in Proc. Int. Conf. Inf. Commun. Technol. Converg., (Jeju, Republic of Korea), 2025, pp. 1141-1143.

[14] H.S. Fang et al., "DEXOP: A Device for Robotic Transfer of Dexterous Human Manipulation," arXiv preprint, 2025, doi: 10.48550/arXiv.2509.04441

[15] Y. Ze et al., "TWIST2: Scalable, Portable, and Holistic Humanoid Data Collection System," arXiv preprint, 2025, doi: 10.48550/arXiv.2511.02832

[16] 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.

[17] A. Khazatsky et al., "DROID: A Large-Scale In-the-Wild Robot Manipulation Dataset," in Proc. Robot.: Sci. Syst., (Delft, Netherlands), July 2024.

[18] R. Cadene et al., "LeRobot: An Open-Source Library for End-to-End Robot Learning," in Proc. Int. Conf. Learn. Represent., (Rio de Janeiro, Brazil), 2026.

[19] Y. Zhu et al., "Bottom-Up Skill Discovery from Unsegmented Demonstrations for Long-Horizon Robot Manipulation," IEEE Robot. Autom. Lett., vol. 7, no. 2, 2022, pp. 4126-4133, doi: 10.1109/LRA.2022.3146589

[20] Z. Zhang et al., "Universal Visual Decomposer: Long-Horizon Manipulation Made Easy," in Proc. IEEE Int. Conf. Robot. Autom., (Yokohama, Japan), 2024, pp. 6973-6980, doi: 10.1109/ICRA57147.2024.10611125

[21] N. Blank et al., "Scaling Robot Policy Learning via Zero-Shot Labeling with Foundation Models," in Proc. 8th Conf. Robot Learn., (Munich, Germany), 2025, pp. 4158-4187.

[22] B. Kim et al., "How to Instruct Your Robot: Dense Language Annotations Power Robot Policy Learning," arXiv preprint, 2026, doi: 10.48550/arXiv.2605.17077

[23] F. Sener et al., "Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (New Orleans, LA, USA), June 2022, pp. 21096-21106.

[24] Voxel51, "2026 State of Visual and Physical AI Survey," Voxel51, 2026. 5. 27. https://voxel51.com/whitepapers/state-of-physical-ai-2026

[25] ISO/IEC, "Artificial Intelligence – Data Quality for Analytics and Machine Learning (ML)," ISO/IEC 5259-2:2024 (Part 2: Data

Quality Measures); ISO/IEC 5259-3:2024 (Part 3: Data Quality Management Requirements and Guidelines); ISO/IEC TR 5259-6:2026 (Part 6: Visualization Framework for Data Quality), 2024-2026.

[26] J. Hejna et al., "Robot Data Curation with Mutual Information Estimators," in Proc. Robot.: Sci. Syst., (Los Angeles, CA, USA), June 2025, arXiv:2502.08623.

[27] Y. Zhang et al., "SCIZOR: A Self-Supervised Approach to Data Curation for Large-Scale Imitation Learning," arXiv preprint, 2025, doi: 10.48550/arXiv.2505.22626

[28] M. Sakr et al., "Consistency Matters: Defining Demonstration Data Quality Metrics in Robot Learning from Demonstration," ACM Trans. Hum.-Robot Interact., vol. 15, no. 2, 2026, pp. 1-31.

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

[30] M.J. Kim et al., "OpenVLA: An Open-Source Vision-Language-Action Model," in Proc. 8th Conf. Robot Learning, (Munich, Germany), 2025, pp. 2679-2713.

[31] K. Black et al., "π0: A Vision-Language-Action Flow Model for General Robot Control," arXiv preprint, 2024, doi: 10.48550/arXiv.2410.24164

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

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

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

[35] J. Yu et al., "ForceVLA: Enhancing VLA Models with a Force-Aware MoE for Contact-Rich Manipulation," in Proc. 39th Conf. Neural Inf. Process. Syst., (San Diego, CA, USA), 2025, pp. 103609-103639.

[36] C. Zhang et al., "VTLA: Vision-Tactile-Language-Action Model with Preference Learning for Insertion Manipulation," Biomim. Intell. Robot., 2026, doi: 10.1016/j.birob.2026.100333

[37] Z. Cheng et al., "OmniVTLA: Vision-Tactile-Language-Action Models with Semantic-Aligned Tactile Sensing," arXiv preprint, 2025, doi: 10.48550/arXiv.2508.08706

[38] J. Bruce et al., "Genie: Generative Interactive Environments," in Proc. Int. Conf. Mach. Learn., (Vienna, Austria), July 2024, pp. 4603-4623.

[39] S. Ye et al., "Latent Action Pretraining from Videos," in Proc. Int. Conf. Learn. Represent., (Singapore, Singapore), Apr. 2025.

[40] H. Wei et al., "Learning Additively Compositional Latent Actions for Embodied AI," arXiv preprint, 2026, doi: 10.48550/arXiv.2604.03340

[41] Z. Tang et al., "ALAM: Algebraically Consistent Latent Action Model for Vision-Language-Action Models," arXiv preprint, 2026, doi: 10.48550/arXiv.2605.10819

[42] Y.J. Ma et al., "VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training," in Proc. Int. Conf. Learn. Represent., (Kigali, Rwanda), 2023.

[43] Y.J. Ma et al., "LIV: Language-Image Representations and Rewards for Robotic Control," in Proc. Int. Conf. Mach. Learn., (Honolulu, HI, USA), July 2023, pp. 23301-23320.

[44] Y.J. Ma et al., "Vision Language Models Are In-Context Value Learners," in Proc. Int. Conf. Learn. Represent., (Singapore, Singapore), 2025.

[45] Y. Du et al., "Learning Universal Policies via Text-Guided Video Generation," in Proc. Advances in Neural Information Processing Systems, vol. 36, (New Orleans, LA, USA), 2023, pp. 9156-9172.

[46] K. Black et al., "Zero-Shot Robotic Manipulation with Pre-Trained Image-Editing Diffusion Models," in Proc. 37th Int. Conf. Neural Inf. Process. Syst., (Vienna, Austria), 2024.

[47] C. Wen et al., "Any-Point Trajectory Modeling for Policy Learning," in Proc. Robot.: Sci. Syst., (Delft, Netherlands), July 2024.

[48] E.-R. Nguyen et al., "Pixel Motion Diffusion is What We Need for Robot Control," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Denver, FL, USA), 2026, pp. 23663-23672.

[49] J. Luo et al., "FMB: A Functional Manipulation Benchmark for Generalizable Robotic Learning," Int. J. Robot. Res., vol. 44, no. 4, 2025, pp. 592-606, doi: 10.1177/02783649241276017

[50] J. Zhao et al., "PolyTouch: A Robust Multi-Modal Tactile Sensor for Contact-Rich Manipulation Using Tactile-Diffusion Policies," in Proc. IEEE Int. Conf. Robot. Autom., (Atlanta, GA, USA), 2025, pp. 104-110, doi: 10.1109/ICRA55743.2025.11128816

[51] H. Xue et al., "Reactive Diffusion Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation," in Proc. Robot.: Sci. Syst., (Los Angeles, CA, USA), June 2025, doi: 10.15607/RSS.2025.XXI.052

[52] Y. Wu et al., “TacDiffusion: Force-Domain Diffusion Policy for Precise Tactile Manipulation,” in Proc. IEEE Int. Conf. Robot. Autom., (Atlanta, GA, USA), 2025, pp. 11831–11837, doi: 10.1109/ICRA55743.2025.11127334

[53] B. Tang et al., “IndustReal: Transferring Contact-Rich Assembly Tasks from Simulation to Reality,” in Proc. Robot.: Sci. Syst., (Daegu, Republic of Korea), July 2023, doi: 10.15607/RSS.2023.XIX.039

[54] K. Pertsch et al., “FAST: Efficient Action Tokenization for Vision-Language-Action Models,” arXiv preprint, 2025, doi: 10.48550/arXiv.2501.09747

[55] ISO, “Robotics – Safety Requirements,” ISO 10218-1:2025 (Part 1: Industrial Robots); ISO 10218-2:2025 (Part 2: Industrial Robot Applications and Robot Cells), 2025.