좋습니다! 말씀해주신 기준에 맞춰 2022년부터 2025년까지 발표된 최신 연구를 대상으로, ICRA, IROS, RSS, IJRR, RA-L, T-RO 등 Top Tier 학회 및 저널에서 '자연어 입력을 통한 manipulator arm 제어 및 grasping' 주제에 관련된 논문들을 심층 조사하여 종합 보고서를 작성하겠습니다.

각 논문은 요청하신 대로 논문 제목, 학회/저널명, Abstract, 핵심 아이디어, Methodology, 사용 가능 여부(open source 여부), 실험 하드웨어 스펙, 성능 결과, 주제와의 연관성 및 적합성 평가, 후속 연구 방향 제안까지 체계적으로 정리하겠습니다.

조사가 완료되면 30페이지 이상의 매우 자세한 한국어 보고서로 작성해드리겠습니다. 잠시만 기다려주세요!

자연어 기반 로봇 매니퓰레이터 제어 및 파지: 2022–2025 주요 연구 동향

서론

로봇 매니퓰레이터(Manipulator) 팔을 **자연어 명령**으로 제어하고 물체를 파지(grasping)하는 기술은 인간-로봇 상호작용의 직관성을 높이고 로봇 활용 범위를 넓히기 위한 핵심 주제로 부상하였습니다. 2022년부터 2025년까지 ICRA, IROS, RSS, IJRR, RA-L, T-RO 등의 **로봇공학 분야 최우수 학회 및 저널**에서는 이와 관련된 다양한 접근 방법의 연구가 발표되었습니다. 이들 연구는 **픽앤플레이스(pick-and-place)** 작업부터 **복합 조작(manipulation)**, **시각-언어 인식** 및 **플래닝**에 이르기까지 폭넓은 범위를 다루며, 대규모 **자연어 모델(LLM)**을 활용하거나 텍스트 명령을 해석하는 등 여러 방법론을 통해 로봇이 사람의 자연어 지시를 이해하고 실행하도록 합니다. 본 보고서에서는 해당 기간 발표된 주요 연구 논문들을 선정하여 각 논문의 **제목, 발표 학회/저널, 논문 초록(요약), 핵심 아이디어, 기술적 접근 방법론, 공개된 구현 여부, 실험에 사용된 하드웨어 사양, 성능 결과(정량적 수치), 본 주제와의 연관성 및 적합성**, 그리고 **후속 연구 방향에 대한 제언**을 체계적으로 정리합니다. 이를 통해 최근 자연어 기반 로봇 팔 제어 연구의 경향을 파악하고, 향후 연구 개발에 필요한 통찰을 제공합니다.

---

1. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances (CoRL 2022)

- **논문 제목:** Do As I Can, Not As I Say: Grounding Language in Robotic Affordances
- **학회/저널:** CoRL 2022 (Conference on Robot Learning) 구두 발표 논문
- **논문 Abstract (요약):** 이 연구는 **대형 언어 모델(LLM)**이 담고 있는 세계에 대한 의미지식을 로봇의 행동에 활용하고자 한 초기 사례입니다. 저자들은 "로봇이 자연어로 주어진 상위 수준의 복잡한 지시를 실행하려면, 언어 모델의 풍부한 지식과 실제 환경에 대한 **물리적 체험**(grounding)이 모두 필요하다"고 지적합니다 (Do As I Can, Not As I Say: Grounding Language in Robotic Affordances | OpenReview). 언어 모델은 세계에 대한 서술적 지식을 갖췄지만 실제 로봇의 **구체적 제약**(예: 로봇의 구조, 환경 상태)을 모르기 때문에 그대로 행동에 옮기기 어렵습니다. 이를 해결하기 위해, 미리 학습된 로봇의 **저수준 행동 스킬**들을 사용하여 LLM이 제안하는 행동을 **가능한 동작**으로 한정짓습니다. 다시 말해, 로봇이 언어 모델의 "손과 눈" 역할을 수행하도록 하고, 언어 모델은 과제를 달성하기 위한 상위 절차 지식을 제공합니다 (Do As I Can, Not As I Say: Grounding Language in Robotic Affordances | OpenReview). 가치 함수 기반의 **Affordance(실현 가능 행동)** 평가를 통해, 언어 모델이 제시한 여러 행동 중 현재 상황에 맞고 실행 가능한 것만 선택하게 합니다 (Do As I Can, Not As I Say: Grounding Language in Robotic Affordances | OpenReview). 실제 다양한 로봇 작업에 본 방법을 평가한 결과, **실제 환경에 대한 피드백과 제약**을 언어

모델에 통합하는 것이 장기적이고 추상적인 자연어 지시를 성공적으로 완료하는 데 필수적임을 보였습니다 (Do As I Can, Not As I Say: Grounding Language in Robotic Affordances | OpenReview).

- **핵심 아이디어:** "Say-Can"으로 알려진 본 연구의 핵심 아이디어는 **대형 언어 모델**의 추론 능력을 로봇 제어에 활용하되, **실행 가능성 평가**(affordance)를 통해 언어 모델의 출력을 현실에 맞게 **보정**하는 것입니다. 언어 모델은 인간의 고수준 지시를 단계로 풀어내고, 로봇은 각 단계별로 사전에 학습된 행동 스킬(예: 이동, 집기 등)을 수행합니다. 각 후보 행동에 대해 **가치 함수** 혹은 **Q-함수** 형태로 성공 가능도를 평가하여 (Do As I Can, Not As I Say: Grounding Language in Robotic Affordances | OpenReview), 가장 실행 가능성이 높은 행동을 실행함으로써, 언어 모델의 계획이 실제 로봇으로 구현됩니다.
- **Methodology (기술적 접근):** 이 방법론은 **강화학습**과 **언어 모델**의 결합입니다. 구체적으로, 로봇은 사전에 여러 **프리미티브 동작 스킬**(primitive skill)을 학습하거나 제공받습니다. 언어 모델(예: 540억 매개변수의 **PaLM** 등)은 인간 지시를 받아 상위 단계의 행동 시퀀스를 텍스트로 생성합니다. 각 단계에 대해, 해당 행동을 실행하는 **로우-레벨(skill)**이 준비되어 있으며, 그 행동이 현재 상태에서 가능한지 여부를 **가치 함수** 형태로 평가합니다. 가치 함수가 가장 높은 행동만 채택함으로써, 언어 모델의 제안이 **현실적 제약을 만족**하도록 합니다 (Do As I Can, Not As I Say: Grounding Language in Robotic Affordances | OpenReview). 이러한 계획-평가 루프를 폐루프(closed-loop)로 반복하여, 명령 실행 도중 환경 변화나 실패에 대응합니다.
- **오픈 소스 여부:** 예. 저자들은 프로젝트 웹사이트와 함께 오픈소스 코드를 공개하였습니다 (예: GitHub의 **SayCan** 구현 및 Colab 노트북) (Do As I Can, Not As I Say: Grounding Language in Robotic Affordances | OpenReview). 이를 통해 다른 연구자들이 동일 프레임워크를 재현 및 확장할 수 있도록 지원했습니다.
- **실험 하드웨어:** 구글의 **모바일 매니퓰레이터** 플랫폼인 Everyday Robots 기반 로봇을 사용하였습니다. 이 로봇은 이동 기반(platform) 위에 1개의 7-자유도 로봇팔과 그리퍼, 그리고 카메라 등의 센서를 장착한 형태입니다. 실험에서는 사무실과 같은 실제 환경에서 **장바구니에 물건 담기, 음료 가져오기, 쓰레기 치우기** 등 일상적인 고차원 작업을 대상으로 테스트되었습니다.
- **성능 결과:** 본 방법은 **10가지가 넘는 현실 과제**에 대해 실행 가능함을 보였고, 특히 장기 계획이 필요한 지시에 대해 기존 방식보다 향상된 성공률을 나타냈습니다. 예를 들어, 이 연구에서는 언어 모델 단독으로는 실행이 어려웠던 "방에 엎질러진 내용물을 치우고 휴지통에 버리기"와 같은 **장문 지시**를 SayCan 프레임워크로 성공적으로 수행해 보였습니다. 저자들은 제안한 방법이 일반적인 플래닝(TAMP) 기법 대비 보다 **유연한 문제해결**을 보이며, LLM의 지식을 활용함으로써 새로운 지시에도 일반화할 수 있음을 보고하였습니다 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models) (Do As I Can, Not As I Say: Grounding Language in Robotic Affordances | OpenReview). 다만 정량적인 수치(성공률 등)는 과제별로 제시되었으나, 언어 모델 사용으로 인한 향상도를 정량 비교하기보다 개념 증명 위주로 다루었습니다. 예를 들어 **Inner Monologue** 등 후속 연구에서 밝힌 바에 따르면, SayCan은 폐루프 재계획 능력이 부족하여 복잡한 환경에선 성공률이 거의 0%에 가깝지만, 환경 피드백을 통합한 방법론(Inner Monologue)은 동일 조건에서 약 60%의 성공률을 보이는 등 큰 향상을 보였습니다 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models). 이는 **실시간 피드백**의 중요성을 시사하며, SayCan 연구 결과 역시 환경 정보가 없을 때 성능 한계를 보임을 나타냅니다.
- **주제 연관성 및 적합성:** SayCan은 **자연어 조건부 로봇 제어**의 효시 중 하나로서, 사람의 모호하고 복잡한 지시를 로봇에게 실행시키기 위한 기본 프레임워크를 제시했습니다. 사용자 입장에서, 이 연구는 **자연어→로봇행동**의 개념 증명을 보여주었고, 이후 연구들이 이를 기반으로 **LLM+로봇** 분야를 개척하는 단초가 되었습니다. 자연어 명령으로 픽앤플레이스 작업을 수행하는 등 본 과제와 직접적으로 부합하는 내용이 많아, 해당 주제에 매우 적합한 선행 연구입니다.
- **후속 연구 방향:** 이 논문을 기반으로 한 후속 연구로는 **더 복잡한 추론을 위한 LLM 능력 향상**과 **로봇의 환경 피드백 통합**이 제안됩니다. 예를 들어, 이후 연구인 *Inner Monologue* ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models) ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models)에서는 언어 모델이 스스로

현재 상황을 서술하고 실패 시 대안을 생각하도록 하는 내부 독백 기법이 도입되어 성능이 개선되었습니다. 또한, SayCan이 단순 행동 조합으로 장기 계획을 수행했다면, 차후에는 **Chain-of-Thought** 프롬프트나 **계획 전용 LLM**을 사용해 더 논리적인 다단계 계획을 생성할 수 있습니다. 마지막으로, 이 접근을 실제 산업 응용으로 확대하기 위해서는 **실시간성 개선**(SayCan은 단계별 LLM 호출로 인한 지연 존재)과 **안전성 보장**(LLM 출력 행동의 검증)이 중요하며, 이러한 방향에서의 연구가 진행될 것으로 예상됩니다.

2. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation (CoRL 2022)

- **논문 제목:** Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation (약칭 **PerAct**)
- **학회/저널:** CoRL 2022 (Poster 발표), RA-L 2023에도 확장게재
- **논문 Abstract (요약):** 본 연구는 **테이블 위 다양한 물체조작 작업**을 하나의 통합된 정책으로 학습하는 **모델**을 제시합니다. 특히 **자연어로 주어진 목표**를 받아, 로봇 팔의 **6-자유도 동작**을 생성하는 **행동 모방 (Behavior Cloning)** 기법을 활용하였습니다 (Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation | OpenReview). Perceiver-Actor(PerAct)는 **언어-조건부 다중작업 에이전트**로, **RGB-D 카메라**로부터 획득한 3차원 **Voxel** 표현과 자연어 임베딩을 Transformer에 입력으로 받아 **다음 실행할 최적의 3D 행동**을 분류하는 형태로 정책을 학습합니다 (Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation | OpenReview). 2D 이미지 기반이 아닌 **Voxel(격자)** 기반의 3D 표현을 도입함으로써, 6-DoF (6자유도) 공간에서 효율적으로 학습이 가능해졌습니다. 적은 시연 데이터만으로도 18개의 RLBench 시뮬레이션 과제(총 249가지 변형)와 7개의 실세계 과제(18가지 변형)를 하나의 모델로 학습했으며, 그 결과 다수의 테이블탑 작업에 대해 기존 2D 기반 정책이나 3D ConvNet 기반 정책보다 뛰어난 성능을 보였습니다 (Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation | OpenReview).
- **핵심 아이디어:** PerAct의 핵심은 **Transformer** 구조를 이용해 **멀티모달 입력**(시각+언어)을 통합하고, **3차원 공간**에서 바로 로봇 행동을 생성한다는 점입니다. 기존에는 각 작업마다 별도의 모델을 두거나, 이미지 기반으로 6-DoF 제어를 하려다보니 비효율적이었는데, PerAct는 **모든 작업을 아우르는 단일 거대 모델**로 **범용성**을 달성했습니다. 또한 CLIP 등 사전 학습 모델에 의존하지 않고도, **Voxel 공간**에서 **다음 행동 위치를 탐색(detect)**하는 방식으로 데이터 효율성을 높였습니다. 이는 곧 **자연어 조건부** 다중작업 학습의 새 방향을 제시한 것으로 평가됩니다.
- **Methodology (기술적 접근):** 기술적으로, PerAct는 입력으로 (**a**) 여러 시점의 카메라로부터 얻은 **RGB-D 영상**을 **voxelized point cloud** 형태로 처리하고, (**b**) 사용자로부터 주어진 **텍스트 명령**을 임베딩합니다. 이 두 정보를 **Perceiver IO** 기반 Transformer에 넣어, 출력으로 **디스크릿화(discretized)**된 6-DoF 행동(그리퍼 이평면 좌표, 회전, 그립 동작)을 예측합니다 (Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation | OpenReview) ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 행동 예측은 마치 3D 공간에서 "다음 취할 행동 위치를 검출(detect)"하는 것으로 볼 수 있으며, 다중 작업에 대해 하나의 모델로 학습되도록 **멀티태스크 Behavior Cloning** 손실을 설계했습니다. 학습 데이터는 시뮬레이터(RLBench)에서 얻은 데모와 소량의 실제 로봇 데모(총 53개)이며, 데이터 증강과 Transformer의 대용량 표현력을 활용해 **few-shot 학습**을 가능케 했습니다.
- **오픈 소스 여부:** 예. 저자들은 **PerAct**의 구현 코드를 공개하고 있으며 (Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation | OpenReview), 또한 여러 연구자들에 의해 재구현이 활발히 공유되고 있습니다. PerAct 전용 웹사이트와 데이터셋도 공개되어 있어 후속 연구에 활용 가능합니다.
- **실험 하드웨어:** 시뮬레이션으로는 **RLBench** (CoppeliaSim 기반) 환경의 18개 다양한 테이블 위 작업(예: 물체 집기, 블록 쌓기 등)을 사용했고, 현실 실험으로는 **Franka Emika Panda** 7-자유도 로봇팔과 병렬 그리퍼를 사용했습니다. 실제 로봇 실험에서는 Kinect v2 RGB-D 카메라 4대를 로봇 주위에 배치하여 실시간 3D voxel 관측을 생성했습니다. 7가지 실세계 작업(블록 쌓기, 오브젝트 정렬 등)을 한데 모아 단일 정책으로 학습·수행하는지를 검증했습니다.
- **성능 결과:** PerAct는 **적은 시演**으로도 다양한 작업에 높은 성공률을 보였습니다. 시뮬레이터 상에서 기존 SOTA 대비 **1.33배(33%)** 향상된 성공률을 10개 데모로 달성했고, 데모를 100개로 늘리면 **2.83배**

(≈183%) 향상된 성능을 보였습니다. 이는 같은 데이터로 학습한 2D 기반 ConvNet 정책(C2F-ARM 등)이나 multi-head 구조보다 월등히 높은 수치입니다. 실세계 7개 작업에서도 **88.9%의 전체 성공률**을 달성하여 (53개의 데모만으로) 모든 작업을 통합한 모델로도 높은 성능을 내는 것을 증명했습니다 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models). 특히 PerAct는 **작업 간 학습 이득**(transfer)을 보여, 어떤 작업 변형에 대한 학습이 다른 작업에도 도움을 주어 **샘플 효율**이 우수했습니다.

- **주제 연관성 및 적합성:** 이 논문은 자연어 입력을 조건으로 **다양한 파지 및 조작 작업**을 하나의 정책으로 해결한다는 점에서, "다목적 로봇비서"를 지향하는 본 주제와 부합합니다. 사용자는 단순히 "주황색 블록을 파란 바구니에 넣어"와 같이 명령하면, PerAct 모델은 해당 목표를 이해하고 적절한 6-DoF 동작 시퀀스를 실행할 수 있습니다. 특히 파지(grasping) 동작도 6-DoF 행동의 일부로 포함되어 있으므로, **물체 인식+파지+조작**을 통합적으로 다룹니다. 본 연구는 대규모 LLM 없이도 Transformer를 활용해 **자연어→행동**을 직접 학습했다는 점에서 의의가 있으며, 이후 **로봇 비전-언어 모델** 분야에 큰 영향을 주었습니다.
- **후속 연구 방향:** PerAct 이후로 제시될 연구 방향으로는 **모델의 경량화와 실시간 제어**, **실세계 일반화**가 있습니다. 현재 PerAct는 voxel 기반으로 연산량이 많기에, 향후에는 **더 효율적인 3D 표현 학습**(예: Sparse voxel 또는 중첩 좌표계 활용)을 통해 **실시간 제어**에 가까워지도록 개선할 수 있습니다. 또한 본 연구는 주로 테이블 위 작업에 국한되었으므로, 후속으로 **모바일 매니퓰레이터**나 **야외 환경** 등으로 일반화하는 실험이 필요합니다. 자연어 처리 측면에서는 복잡한 문장이나 조합 명령에 대한 대응, 예를 들어 "오른쪽 상자에서 빨간 공을 꺼내 왼쪽 바구니에 넣어" 같은 문장도 이해하도록 **언어 이해 모듈**을 강화하는 방향이 있을 것입니다. 마지막으로, PerAct와 같은 imitation learning 방식에 **온라이닝 강화학습**을 접목하여, 사람이 피드백을 주거나 실패 경험을 통해 계속 학습하는 **자율 적응형 시스템**으로 발전시키는 것도 중요한 방향입니다.

3. Code as Policies: Language Model Programs for Embodied Control (ICRA 2023, arXiv 2022)

- **논문 제목:** Code as Policies: Language Model Programs for Embodied Control (약칭 **CaP**)
- **학회/저널:** ICRA 2023 발표 (arXiv preprint 2022)
- **논문 Abstract (요약):** 이 연구는 특이하게도 **프로그래밍 언어**를 매개로 로봇을 제어하는 방식을 제안했습니다. 저자들은 **코드 생성에 특화된 대형 언어 모델**(예: OpenAI Codex)이 주석(documentation)만으로 간단한 파이썬 코드를 작성할 수 있다는 점에 착안하여 (Code as Policies: Language Model Programs for Embodied Control), **자연어 입력으로부터 로봇 정책 코드**를 작성하도록 하는 "*Code-as-Policies*" 프레임워크를 만들었습니다. 이 방법에서 자연어 정책 코드는 **로봇의 센서 정보 처리 함수**와 **기본 제어 API 호출**로 구성됩니다 (Code as Policies: Language Model Programs for Embodied Control) (Code as Policies: Language Model Programs for Embodied Control). 여러 개의 예시(주석과 코드 쌍)를 few-shot prompt로 제공하면, LLM이 새로운 자연어 명령에 대해 **파이썬 코드**를 생성합니다. 생성된 코드는 조건문, 반복문 등의 구조를 활용해 복잡한 논리를 구현할 수 있고, Numpy 등 외부 라이브러리도 호출하여 계산을 수행할 수 있습니다 (Code as Policies: Language Model Programs for Embodied Control). 이로써 로봇 정책이 **동적 피드백 루프**(센서 입력에 따른 분기)를 포함한 **반응형 정책(reactive policy)**까지 표현될 수 있음을 보였습니다 (Code as Policies: Language Model Programs for Embodied Control). 여러 실제 로봇 플랫폼을 대상으로 시연한 결과, 이 접근법은 새로운 지시에 대한 **일반화** 능력을 보였고, HumanEval 코드 생성 벤치마크에서도 기존 대비 최고 **39.8%의 문제 해결율**을 달성하였습니다 (Code as Policies: Language Model Programs for Embodied Control).
- **핵심 아이디어:** Code-as-Policies(CaP)의 핵심은 "**자연어를 바로 실행가능한 코드로 변환**"하는 것입니다. 인간이 로봇에게 지시를 내릴 때, 결국 로봇은 내부 소프트웨어 API를 통해 동작합니다. CaP는 이 **API 호출 시퀀스**를 LLM이 **프로그래밍**하도록 유도합니다 ([2209.07753] Code as Policies: Language Model Programs for Embodied Control) ([2209.07753] Code as Policies: Language Model Programs for Embodied Control). 이를 위해 hierarchical prompting이라는 기법을 사용했는데, 명령을 처리하다 미정

의 함수가 나오면 그 함수를 정의하는 코드를 다시 생성하는 식으로 **재귀적 코드 생성**을 수행합니다 (Code as Policies: Language Model Programs for Embodied Control). 언어 모델의 논리적 코딩 능력을 활용하여, 명령어에 내포된 **공간적 추론**이나 **모호한 지시의 수치화**(예: "조금 더 빠르게"를 적절한 속도로 변환)를 자동으로 처리하게 한 것입니다 (Code as Policies: Language Model Programs for Embodied Control). 결국 로봇은 사람이 준 명령에 따라 **자체 생성한 코드**를 실행하여 임무를 완수합니다.

- **Methodology (기술적 접근):** CaP 프레임워크에서는 우선 로봇이 이용할 **기본 동작 API**들을 파이썬 함수 형태로 정의합니다 (예: `move_gripper(x,y,z)` 함수, `open_gripper()` 등). 그리고 다양한 시나리오에 대한 **자연어 명령**과 **그에 상응하는 정책 코드** 예시를 prompt로 LLM에게 제공합니다 ([2209.07753] Code as Policies: Language Model Programs for Embodied Control) ([2209.07753] Code as Policies: Language Model Programs for Embodied Control). 예컨대 "컵을 잡아 들어올려서 흔든 후 내려놓아"라는 명령에 대응하는 코드 예시를 보여주면, LLM은 새로운 "볼을 집어 상자에 넣어" 명령에 대해서 스스로 코드를 작성합니다. **계층적 프롬프트(hierarchical prompting)** 기법 덕분에, LLM이 한 번에 전체 코드를 생성하지 않고 함수 단위로 점진적으로 코드를 완성해나가며, 구조적이고 오류가 적은 코드를 산출합니다. 생성된 코드는 **PyBullet** 등의 시뮬레이터나 실제 로봇의 Python 제어 인터페이스에서 즉시 실행됩니다. 만약 LLM이 작성한 코드에 오류가 있거나 미비하면, 사람이 피드백을 주거나 추가 프롬프트로 수정할 수도 있습니다.
- **오픈 소스 여부:** 예. 구글 Robotics 팀은 이 연구와 함께 **코드 및 데모**를 오픈소스로 공개했습니다 (Google's Code-as-Policies Lets Robots Write Their Own Code - InfoQ) (Google's Code-as-Policies Lets Robots Write Their Own Code - InfoQ). Github에 재현 코드와 HuggingFace에 인터랙티브 데모까지 제공되어, 누구나 자연어 명령을 넣어 로봇 코드 생성 결과를 시험해볼 수 있습니다.
- **실험 하드웨어:** 여러 종류의 **실제 로봇 플랫폼**에서 실험이 이루어졌습니다. 대표적으로, **일상생활 지원 로봇** (모바일 매니퓰레이터), **산업용 로봇팔**, 그리고 **드론** 등의 예가 제시되었습니다. 예를 들어 로봇 팔인 KUKA YouBot으로 "블록 쌓기" 작업을, 모바일 로봇으로 "쓰레기 찾아 휴지통에 버리기" 작업을, 쿼드콥터 드론으로 "지정된 지점까지 비행" 같은 작업을 수행했습니다. 각 플랫폼별로 해당 로봇에 맞는 API 함수를 정의해주기만 하면 CaP 프레임워크를 적용할 수 있음을 시연하였습니다.
- **성능 결과:** Code-as-Policies는 **정량적**으로 HumanEval 벤치마크에서 기존 대비 향상된 코드생성 성능 (**39.8%** pass@1)을 보였으며 (Code as Policies: Language Model Programs for Embodied Control), 이는 곧 프롬프트 기법이 LLM의 코딩 능력을 끌어올렸음을 의미합니다. 로봇 작업 측면에서는, 정성적 평가 위주로 다양한 새로운 명령을 실행 가능함을 보여주었습니다. 예를 들어 "만약 빨간 컵이 탁자 위에 있으면 집어서 바구니에 넣어"라는 복합 조건 명령에 대해, LLM은 적절히 if-else 구조를 가진 코드를 생성했고 로봇이 상황에 따라 동작했습니다. 또한 "사용자가 피곤하면 에너지 드링크를 가져다줘"처럼 추론이 필요한 지시에 대해서도, 웹 지식(음료 성분 등)에 기반한 판단을 내려 올바른 물체를 가져오는 등 **추론적 행동**을 시연했습니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 이러한 사례들은 CaP가 **훈련 데이터에 없던 새로운 명령**이나 **상황**에도 비교적 잘 일반화함을 보여줍니다. 다만 실행 속도는 LLM 호출과 파이썬 인터프리트 지연 등으로 실시간 제어에는 부족하며, 생성 코드의 안전성(예: 무한루프 없도록)은 추가 검증이 필요합니다.
- **주제 연관성 및 적합성:** CaP는 언뜻 보면 본 주제와 결이 달라 보일 수 있으나, 근본적으로 **자연어를 로봇 행동으로 변환**한다는 목표는 동일합니다. 차이점은 중간 표현으로 **프로그래밍 코드**를 사용했다는 점입니다. 사용자는 "자연어 → 코드 → 로봇"의 간접 경로를 통해서도 충분히 직관적인 제어가 가능함을 보였고, 특히 **조건부 논리 동작**이나 **다단계 절차**를 표현하기에 코드가 적합하다는 통찰을 제공했습니다. 이 접근은 **LLM의 강력한 코딩 능력**을 활용하므로, LLM과 로봇 제어의 접목이라는 측면에서 매우 흥미로운 대안으로 평가됩니다.
- **후속 연구 방향:** Code-as-Policies의 후속으로는 **생성 코드의 검증 및 최적화**, **다양한 LLM 통합**이 제안됩니다. 첫째, 로봇에게 위험하거나 비효율적인 코드를 생성하는 문제를 해결하기 위해, **형식적 검증 도구**나 **시뮬레이션 테스트**를 거쳐 코드를 실행하는 방안을 연구할 수 있습니다. 또한 사람이 일일이 프롬프트 예시를 제공하지 않고도 동작하도록, **사전 학습(fine-tuning)**된 LLM을 사용하는 방향도 가능합니

다. 둘째, 현재는 파이썬 코드만 생성했지만, 향후에는 **그래픽 프로그래밍**이나 **전용 DSL**(도메인 특화 언어)을 생성하도록 하여 더 직관적이고 안전한 정책 표현을 추구할 수 있습니다. 예를 들어 로봇 동작을 표현하는 **고수준 스크립트 언어**를 만들어 LLM이 그 언어를 출력하게 하면, 사람이 읽고 검증하기도 용이할 것입니다. 마지막으로, 생성된 코드 조각을 **모듈화**하여 재사용성을 높이고, 여러 작업 간 공유함으로써 **연속적 학습** 또는 **온디맨드 행동 생성** 등으로 발전시킬 수 있을 것입니다.

4. Inner Monologue: Embodied Reasoning through Planning with Language Models (arXiv 2022)

- **논문 제목:** Inner Monologue: Embodied Reasoning through Planning with Language Models
- **학회/저널:** arXiv 2022 (구글 로보틱스, 추후 RSS 2023 발표)
- **논문 Abstract (요약):** 이 연구는 **로봇 계획**에 있어서 **폐쇄형 루프 언어 추론**의 중요성을 강조합니다. 저자들은 **대형 언어 모델(LLM)**을 이용하여 로봇의 고수준 계획을 세우는 기존 시도가, **실시간 환경 변화에 대응**하지 못하고 한 번 생성된 계획을 고집함으로써 생기는 한계를 지적합니다 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models). Inner Monologue 기법에서는 로봇이 수행 중인 작업에 대해 **성공/실패 여부, 주변 환경 설명, 사용자 피드백** 등을 **실시간으로 텍스트로 전달**하면, LLM이 이를 마치 스스로 중얼거리듯이 **내부 대화**로 활용하여 계획을 재고려하도록 합니다 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models). 즉, LLM이 **환경으로부터 언어 형태의 피드백**을 지속적으로 받아들여, "내가 시도한 동작이 실패했군. 그렇다면 다른 방법을 시도해야겠다."와 같은 **추론의 흐름(Inner Monologue)**을 형성하도록 유도합니다. 이를 통해 LLM은 어떤 스킬을 언제 어떻게 실행하고, 실패 시 대안을 찾는 등의 복잡한 결정을 스스로 내릴 수 있게 됩니다. 실험 결과, **폐루프 언어 피드백**을 통합한 LLM 플래너는 기존 개방형 루프 방식(SayCan 등) 대비 다양한 도메인에서 **고수준 지시 완료율을 크게 향상**시켰습니다 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models) ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models).
- **핵심 아이디어:** Inner Monologue의 핵심은 **LLM에게 기억과 반성 능력을 부여**하는 것입니다. 로봇이 행한 행동의 성공 여부나 환경 상태를 **언어 서술**로 LLM에 계속 제공하면, LLM은 이를 참고하여 다음 행동을 결정하거나 계획을 수정합니다. 이는 사람도 복잡한 문제를 풀 때 혼잣말로 논리를 정리하는 것과 유사합니다. 예를 들어, 로봇이 물체 집기에 실패하면 환경으로부터 "로봇이 물체를 놓쳤다"는 피드백을 받고, LLM은 내부적으로 "다른 손으로 잡아보자"와 같이 계획을 업데이트합니다. 이러한 내부 독백 체계 덕분에, LLM은 **실패로부터 학습**하거나 **조건 변화를 인지**하여 **보다 유연한 문제 해결**을 합니다.
- **Methodology (기술적 접근):** 구현적으로, 저자들은 LLM에게 줄 프롬프트에 **특수 토큰**이나 **포맷**을 정해 환경 피드백을 주입했습니다. 예를 들어, 프롬프트에 "환경: <현재 장면 묘사>", "성공 여부: <마지막 행동 결과>" 등을 명시하고, LLM이 이를 읽고 다음 지시를 **"로봇:" 섹션에 출력**하도록 했습니다. 또한 사람과의 대화도 포함하여, 필요한 경우 LLM이 사람에게 질문을 할 수도 있게 했습니다. LLM은 고정된 모델(GPT-3 등)을 사용하고 별도 파인튜닝 없이 **프롬프트 엔지니어링**만으로 동작시켰습니다. 로봇의 저수준 행동 스킬은 기존과 같이 사전 정의되어 있으며, LLM은 각 단계마다 실행할 스킬과 파라미터를 문장으로 기술하면 이를 로봇이 해석하여 수행합니다. 이러한 구조로, LLM의 출력 -> 로봇 실행 -> 텍스트 피드백 -> LLM 입력의 **루프**가 형성됩니다.
- **오픈 소스 여부:** 부분적 공개. Inner Monologue의 개념 증명은 웹사이트에 동영상 등으로 공개되었으나, 코드 전체는 공개되지 않았습니다. 다만, 유사한 아이디어를 구현한 오픈소스 프로젝트들이 이후 등장하여 접근 방법을 재현할 수 있습니다.
- **실험 하드웨어:** 세 가지 도메인에서 평가되었습니다: (1) **테이블 정리 작업** – 실제 탁자 위에서 블록을 분류하고 쌓는 작업, (2) **장기 모바일 매니퓰레이션** – 실제 주방 환경에서 mobile manipulator가 여러 단계를 거쳐 임무 수행 (예: 냉장고에서 음료 꺼내기), (3) **시뮬레이션 가상환경(VirtualHome)** – 가상 가정집 환경에서 복잡한 미션 수행. 실제 로봇으로는 Fetch 모바일 매니퓰레이터 등이 사용되었고, 센서로 RGB-

D 카메라, 접촉 센서 등이 활용되었습니다. 각 환경에서 Inner Monologue의 일반성을 보여주기 위해 다양한 시나리오를 실험했습니다.

- **성능 결과:** Inner Monologue 기법은 동일한 LLM이라도 **폐루프 피드백**을 사용함으로써 성공률이 크게 향상됨을 보였습니다. 예를 들어 테이블 위 블록 정리 작업에서, **SayCan** 방식은 새로운 상황(방해물 추가 등)에서 **0%에 가까운 성공률**을 보였지만, Inner Monologue를 도입하면 **60.4%**의 성공률로 향상되었습니다 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models). 또한 "3개의 블록 쌓기", "병에서 과일 분류" 등의 과제 세트에서도, 피드백이 없을 때 20~45% 이던 완료율이 Inner Monologue로 90%까지 상승했습니다 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models). 종합적으로, 다양한 작업에서 Inner Monologue 적용 시 **약 2배 이상의 고수준 지시 완료율 향상**을 달성했습니다 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models) ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models). 질적으로도, 로봇이 중간에 실패하면 즉석에서 계획을 바꾸거나, 사람이 "그만둬"라고 말하면 즉시 작업을 중단하는 등 유연성이 크게 증가했습니다.
- **주제 연관성 및 적합성:** 이 연구는 **자연어 상호작용**을 활용하여 로봇 제어의 신뢰성과 적응성을 높인 사례로, 본 주제와 밀접한 관련이 있습니다. 사용자의 자연어 명령을 단순 실행하는 것을 넘어서, **진행 상황에 대한 설명과 추가 명령**까지 모두 자연어로 처리합니다. 이는 궁극적으로 인간과 로봇이 **대화**를 통해 과제를 함께 해결하는 방향을 제시하며, 자연어 기반 로봇 제어의 **다음 단계**라 할 수 있습니다. 특히 파지 및 조작 작업에서 언제 멈추고 다시 시도할지, 오류 시 어떻게 수정할지를 언어로 결정하게 하여, **로봇 자율성**을 높였다는 점에서 의의가 있습니다.
- **후속 연구 방향:** Inner Monologue는 향후 **더 큰 LLM (예: GPT-4)**과 결합하거나, **시각 정보 직접 입력** 등으로 발전될 수 있습니다. 현재는 텍스트로 상태를 전달하지만, 장차 **멀티모달 LLM**에게 이미지나 영상 설명을 직접 하도록 하면 인간 수준의 상황 이해를 기대할 수 있습니다. 또한 인간 피드백을 언어로 통합하는 부분을 발전시켜, 사용자가 자연어로 "좀 더 오른쪽으로"라고 말하면 로봇이 이를 즉시 이해하고 계획을 조정하도록 만드는 연구가 가능할 것입니다. 아울러, 내재적 독백이 지나치게 장황해지거나 잘못된 방향으로 흐르는 것을 막기 위해 **강화학습을 통한 프롬프트 최적화**나 **모델 정교화**를 시도해 볼 수 있습니다. 마지막으로, 본 기법을 다수 로봇이 협업하는 시나리오에 확장하여, 로봇들 간에도 언어로 협의하는 **다중 로봇 협력** 분야로의 적용도 하나의 도전적인 방향입니다.

5. RT-1: Robotics Transformer for Real-World Control at Scale (arXiv 2022)

- **논문 제목:** RT-1: Robotics Transformer for Real-World Control at Scale
- **학회/저널:** arXiv 2022 (Google Robotics Tech Report)
- **논문 Abstract (요약):** RT-1은 구글에서 발표한 **대규모 로봇 정책 모델**로, **130K 이상의 실제 로봇 시연 데이터**로 학습된 **멀티태스크 트랜스포머 정책**입니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 이 모델은 **비전-언어 입력**을 받아 **실시간 로봇 제어 명령**을 출력하는 end-to-end 구조를 가지며, 복잡하고 다양한 작업에 단일 모델로 일반화할 수 있음을 보였습니다. 입력으로 로봇의 **카메라 이미지 시퀀스**와 **사용자 텍스트 명령**을 받아, 출력으로 로봇의 **관절 움직임 토큰 시퀀스**를 생성합니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale) ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 모델 아키텍처는 ResNet 기반 이미지 인코더(비전), USE(Language) 임베딩으로 텍스트 처리, 그리고 TokenLearner로 토큰을 추출한 후 **디코더 전용 Transformer**로 시각-언어 토큰을 처리하여 행동 토큰을 내놓는 식입니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale) ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 총 700여 개의 서로 다른 자연어 지시에 대응하는 과제들을 학습하여, 이전에 보지 못한 새로운 지시나 물체, 환경에도 강인한 성능을 보였고, 특히 **일반화된 성능**에서 기존 방법보다 18~36% 이상 개선을 이루었습니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale).

- **핵심 아이디어:** RT-1의 핵심은 **거대한 규모의 로봇 데이터셋**과 **Transformer**를 활용해, **일반화 가능한 정책**을 학습했다는 점입니다. 이전까지 로봇 학습은 수백~수천 에피소드 수준의 제한된 데이터로 특정 작업에 맞춘 모델을 만드는 경우가 많았습니다. 반면 RT-1은 **130천개(13만)**에 달하는 다양한 실제 시연을 하나로 모아 학습함으로써, **데이터 양과 다양성에 따른 성능 향상**을 실증했습니다 (Google Research, 2022 & beyond: Robotics). 또한 모델 구조를 **토큰화** 기반으로 설계하여, 시각 입력과 행동 출력을 동일한 Transformer 안에서 처리함으로써 **통일된 시각-행동 표현**을 배웠습니다. 그 결과, 새로운 물체나 지시문장이 주어져도 모델이 가진 일반 지식으로 대응할 수 있게 되었습니다.
- **Methodology (기술적 접근):** 데이터 측면에서, 구글에서는 여러 로봇으로 수집한 **실세계 시연 데이터**를 텍스트 설명과 짝지어 대규모로 확보했습니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 각 에피소드마다 인간이 "로봇이 수행한 작업"을 한 줄 자연어로 기술하여, 시각 관찰(카메라 영상)과 행동 기록(모터 명령 시퀀스)에 그 **자연어 목표**를 레이블로 달았습니다. 모델은 이 데이터를 **지도 학습**으로 훈련하는데, 이미지와 텍스트를 입력받아 다음 행동 토큰을 예측하도록 **교사강요(teacher forcing)**로 Transformer 디코더를 학습시켰습니다. 여기서 행동은 연속값을 디스크릿 토큰으로 양자화하여 처리했습니다. 또한 **멀티태스크 학습** 상황에서 특정 작업에 과적합되지 않도록 **dropout** 및 **데이터 셔플링** 전략을 썼고, **실시간 추론**을 위해 모델 경량화(약 77M 파라미터)와 TensorRT 최적화 등을 병행했습니다. 최종 정책은 1초에 10회 이상의 주기로 명령을 출력하며, "기본 동작 + 이동 + 종료" 등의 모드를 전환해가며 모바일 매니퓰레이터를 제어합니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale).
- **오픈 소스 여부:** 부분적으로 공개. RT-1의 학습에 사용된 **데이터셋** 일부가 공개되었으며, 모델 아키텍처 개요와 가중치 일부도 제공되었습니다. 그러나 전체 130k 데이터는 사내 자산으로, 연구 목적으로 요약본만 공개되었습니다. 현재 RT-1의 코드나 모델은 완전 공개는 아니지만, 유사한 개념의 공개 프로젝트들이 있습니다.
- **실험 하드웨어:** 데이터 수집과 실험에는 구글의 **일상로봇(Everyday Robot)** 플랫폼 다수가 사용되었습니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 이 로봇은 앞서 SayCan 연구와 동일한 타입(모바일 베이스+7-DoF 팔)이며, **오피스 키친** 등의 실제 환경에서 움직였습니다. 학습 데이터에 포함된 작업은 "음료 캔 들기", "쓰레기 줍기", "서랍 열기/닫기", "사과 옮기기" 등 **700여 종류**에 이르는 다양한 조작 및 이동 작업이었습니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 평가 또한 두 가지 실제 주방 환경(Kitchen1, Kitchen2)에서 모델의 일반화 성능을 검증하는 방식으로 진행되었습니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale).
- **성능 결과:** RT-1은 훈련 본보기(Task instructions)에 대해 **97%**에 가까운 성공률을 보였고, 보지 못한 새로운 작업이나 물체, 배경 방해물 등에 대해서도 **이전 최첨단 대비 25~36% 높은 성공률**을 기록했습니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 예를 들어, 새로운 주방 배치와 조명 조건에서도 RT-1은 경쟁 모델(Gato 등)보다 **상대적 18%** 높은 성공률을 보였고, 방해물이 추가된 경우에는 **36%**까지 차이가 벌어졌습니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 또한 SayCan 프레임워크에 RT-1을 통합하여 50단계에 달하는 매우 장기적인 작업까지 수행한 결과, **계획 성공률 87%**에서 **실행 성공률 67%**를 기록하여 긴 시퀀스 작업도 상당 부분 완수할 수 있음을 보여주었습니다 ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale) ([2212.06817] RT-1: Robotics Transformer for Real-World Control at Scale). 이는 이전까지 한 자리 수 성공률에 그쳤던 장기 작업 수행에 큰 진전을 이룬 것입니다. 다만 작업 길이가 길어질수록 성공률이 기하급수적으로 떨어지는 경향은 여전히 있어 (Grounding LLMs For Robot Task Planning Using Closed-loop State Feedback), 완전한 장기 계획 수행에는 개선의 여지가 남았습니다.
- **주제 연관성 및 적합성:** RT-1은 **"규모의 힘"**을 통해 자연어 기반 로봇 제어에 새 방향을 제시한 연구입니다. 다양한 파지와 조작 작업을 하나로 묶어 대량학습 했기 때문에, 사용자가 어떤 명령을 주든 (설령 훈련에 없던 조합이라도) 로봇이 유사한 경험을 일반화하여 행동할 가능성이 높습니다. 이는 사람에게 일일이 학습시킬 필요 없이 **일반적인 로봇 조수**를 만드는 데 한 걸음 다가간 것으로, 본 보고서 주제의 장

기적 비전에 부합합니다. 특히 파지(grasping)는 RT-1이 다룬 핵심 스킬 중 하나로, 다양한 물체를 집는 동작을 포괄적으로 학습했다는 점에서 파지 연구에도 의미가 큽니다.

- **후속 연구 방향:** RT-1의 성공 이후, **더 큰 범주의 데이터 통합**과 **모델 고도화** 방향의 연구가 이어지고 있습니다. 후속작인 **RT-2** ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control)는 웹으로부터 학습된 거대 비전-언어 모델을 로봇 정책에 통합하여, RT-1이 보이지 못한 **신기한 물체**나 **추상 개념**도 이해하도록 향상시켰습니다. 이는 **지식의 이전(transfer)** 측면을 강화한 것이며, 향후 이러한 **Vision-Language-Action** 모델이 표준이 될 가능성이 있습니다. 또한 모델 경량화와 비용 문제를 해결하기 위해, 거대 모델 대신 **모델 앙상블**이나 **지식 증류**를 통해 작은 모델로 유사 성능을 내는 연구도 필요합니다. 마지막으로, 130k 데이터를 능가하는 **자율 데이터 수집** 방법 – 예를 들어 로봇이 스스로 시도하고 배워서 데이터를 늘리는 자기학습 – 도 RT-1 계열 연구의 향후 과제가 될 것입니다.

6. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control (arXiv 2023)

- **논문 제목:** RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control
- **학회/저널:** arXiv 2023 (Google DeepMind)
- **논문 Abstract (요약):** RT-2는 앞서 설명한 RT-1의 후속으로, **인터넷 웹데이터로 학습된 거대 비전-언어 모델**을 로봇 제어에 직접 활용한 사례입니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 목적은 웹으로부터 학습한 **상식과 시각적 개념 지식**을 로봇이 활용하여, 훈련 데이터에 없는 새로운 상황도 대처하도록 하는 것입니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 이를 위해, **시각-언어 모델**의 **토큰 공간**에 **로봇 행동 토큰**을 추가하고, 로봇 데이터와 웹 데이터에 **동시에 파인튜닝**하는 방식을 제안했습니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 즉, 한 모델이 질문에 대한 텍스트 답변과 로봇에 대한 행동 명령을 **모두 생성**할 수 있도록 한 것입니다. 거대한 웹 데이터로 미리 학습한 모델을 일부 로봇 데이터로 미세조정하자, RT-2는 **새로운 물체에 대한 일반화**, **훈련에 없던 복합 명령 이해**, **기초적인 추론 작업** 등을 수행해냈습니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 예를 들어, "피곤한 사람에게 적합한 음료를 건네줘"라는 로봇에게는 추상적인 명령에 대해, 훈련엔 없었지만 웹 지식으로 "에너지 드링크"를 선택해 전달하는 등 **추론적 일반화**가 관찰되었습니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 또한 **Chain-of-Thought** 프롬프트를 적용하여 여러 단계를 거친 논리적 추론도 가능함을 보여주었습니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control).
- **핵심 아이디어:** RT-2의 핵심은 **로봇 정책 공간을 언어 모델 공간으로 끌어들여 텍스트와 행동의 경계를 허문 것**입니다. 행동을 하나의 언어 형태(토큰 시퀀스)로 취급함으로써, 웹상에서 학습된 거대한 파라미터(5620억 규모)의 능력을 로봇 제어에 사용했습니다 (PaLM-E: An Embodied Multimodal Language Model). 이로써, 로봇이 보지 못한 사물이나 개념도 언어 모델이 가진 지식을 통해 이해하게 되었고, 예컨대 "돌을 망치로 써서 못을 박아"라는 말에 망치가 없으면 돌을 집는 행동을 보이는 등 **창의적 문제해결**이 가능해졌습니다. 이는 LLM이 지닌 **추론 및 상식 능력**을 로봇 행동에 통합한 중요한 진전으로 평가됩니다.
- **Methodology (기술적 접근):** RT-2의 학습은 크게 두 부분을 **동시 진행**합니다. 하나는 웹 데이터셋 (예: 이미지-텍스트 쌍, VQA 등)으로 거대 비전-언어 모델을 학습하는 것이고, 다른 하나는 RT-1과 같은 로봇 시연 데이터로 해당 모델을 추가 학습하는 것입니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 이때 핵심은 로봇 행동을 **텍스트 토큰**으로 간주하기에, 웹 데이터 학습 시에도 가끔씩 "로봇~~" 같은 프롬프트로 행동 토큰을 예측하도록 함께 훈련시켰다는 점입니다. 즉, 모델이 텍스트 응답을 산출하는 맥락과 동일한 맥락에서 로봇 제어 명령도 산출하게 만들어 포맷을 맞춘 것입니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web

Knowledge to Robotic Control). 학습에는 Mixture of Tasks 기법이 사용되어, 다양한 작업(이미지 질문답, 캡셔닝, 로봇 시뮬 등)이 한 모델에 통합되었습니다. 마지막으로, chain-of-thought를 활용하기 위해 명령을 "생각"과 "행동" 단계로 나누어 생성하도록 추가 프롬프트를 주어 실험했습니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control).

- **오픈 소스 여부:** 아니오. RT-2는 사내 연구 결과로 구체적인 모델 가중치나 코드가 공개되지 않았습니다. 다만 연구 논문과 일부 영상, 결과 수치가 공개되어 있습니다.
- **실험 하드웨어:** RT-1과 동일한 유형의 구글 로봇들을 사용하여, **6천 회 이상의 평가 실험**을 진행했습니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 평가 과제는 RT-1에서 사용된 것들과, 그에 변형된 새로운 시나리오들이었습니다. 특히 **비훈련 물체**(예: 장난감 공룡 등)나 **새로운 지시어**(예: "가장 작은 물건 집어오기", "노란 별 아이콘이 있는 물체 옮기기" 등)를 테스트하여 모델의 **지식 전이 성능**을 측정했습니다. RT-2는 사내 실험으로 진행되어 구체적인 로봇 수는 명시되지 않았으나, 다양한 환경에서 성능을 검증했습니다.
- **성능 결과:** RT-2는 RT-1 대비 **새로운 상황에서의 성공률이 유의미하게 향상**되었습니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 정량적으로, 보고서에 따르면 **비훈련 객체에 대한 파지 및 조작 성공률**이 RT-1보다 크게 높았고, 예를 들어 "이모티콘 모양 판별하여 물체 옮기기" 같은 과제는 RT-1은 거의 실패했지만 RT-2는 상당 부분 성공했습니다. 또한 "작은/큰 물체 선택"처럼 **텍스트에 내재된 비교 추론** 과제도 높은 성공률을 보였습니다. 구체적인 수치는 공개 제한으로 추정되나, 전반적으로 **50% 이상 향상**된 사례도 있다고 언급됩니다. Chain-of-thought 실험에서는, 프롬프트로 "생각" 단계를 유도하면, 예를 들어 "못을 박기 위해 어떤 도구? -> 망치 없으면 돌 사용"과 같은 **추론 경로**를 거쳐 최종 행동을 결정하는 모습이 관찰되었습니다 ([2307.15818] RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control). 이는 단일 단계 출력보다 정확도는 다소 낮았지만, 로봇의 의사결정 과정을 투명하게 보여주는 효과가 있었습니다.
- **주제 연관성 및 적합성:** RT-2는 **인터넷 지식**을 로봇에 장착했다는 점에서, 자연어 기반 로봇 제어의 지평을 넓혔습니다. 사용자는 로봇에게 굳이 학습시킨 적 없는 요청도 할 수 있고, 로봇은 그 의미를 **추론**하여 수행할 수 있습니다. 이는 인간에게 로봇을 설명하고 가르치는 부담을 크게 줄여주므로, 궁극적인 자연어 로봇 제어 시스템에 한층 가까워졌다고 볼 수 있습니다. 특히 파지 작업에서 새로운 물체(예: 공룡 인형)를 "잡아 가져와"라고 해도, RT-2는 웹에서 공룡 장난감을 본 적 있기 때문에 식별하고 잡을 수 있다는 점에서 **개방형 어휘(OOV) 문제 해결**에 기여합니다. 따라서 RT-2는 본 주제의 **확장성** 측면에서 매우 적합한 연구입니다.
- **후속 연구 방향:** RT-2는 거대 멀티모달 모델 활용의 시작으로, 향후 **모델 효율화**와 **지능형 상호작용** 방향으로 발전할 것입니다. 먼저 효율화 측면에서, 웹 데이터까지 통합하다 보니 모델 크기가 거대해졌는데, 이를 **지식 증류**나 **모델 압축**을 통해 경량화하면서 성능은 유지하는 연구가 필요합니다. 또한 RT-2는 여전히 일방향: 웹지식을 로봇에 적용한 것이므로, **로봇 상호작용 데이터**를 다시 웹지식 학습에 활용하는 **양방향 학습**도 고찰해볼 수 있습니다. 이는 로봇 경험이 언어 모델의 세계지식으로 편입되는 형태로, 장기적으로 **자율적인 지식 확장**을 의미합니다. 마지막으로, RT-2가 보여준 **Chain-of-Thought** 활용은 추론 투명성을 높였지만 속도 저하 문제도 있으므로, 향후에는 **실시간 추론이 가능한 논리형 LLM**이나 **전문 계획자 모듈**과 결합해 보다 실용적인 시스템으로 다듬는 방향이 과제로 남아 있습니다.

7. NaturalVLM: Leveraging Fine-grained Natural Language for Affordance-Guided Visual Manipulation (RA-L 2024)

- **논문 제목:** NaturalVLM: Leveraging Fine-grained Natural Language for Affordance-Guided Visual Manipulation
- **학회/저널:** IEEE Robotics and Automation Letters, 2024년 12월호 (RA-L), IROS 2024 발표 예정
- **논문 Abstract (요약):** 이 연구는 **가정 환경에서의 조작 작업**을 위해, 인간의 **상세한 자연어 지시**를 로봇에게 이해시키고 수행하는 방법을 다룹니다. 저자들은 기존 연구들이 비교적 단순하고 고정된 형태의 지시(예: "상자를 열어")에 초점을 맞춘 반면, 실제 과제는 **다단계 추론**과 **세분화된 명령**을 요구하는 경우가

많다고 지적합니다 ([2403.08355] NaturalVLM: Leveraging Fine-grained Natural Language for Affordance-Guided Visual Manipulation). 이를 위해 **NrVLM**이라는 **새로운 벤치마크**를 구축했는데, 15가지 조작 작업에 대해 4,500개 이상의 에피소드가 **단계별 세밀한 자연어 지시**와 짝지어져 있습니다 ([2403.08355] NaturalVLM: Leveraging Fine-grained Natural Language for Affordance-Guided Visual Manipulation). 각 작업은 여러 하위 단계로 나뉘며, 각 단계마다 자연어로 된 지침이 존재합니다. 동시에 저자들은 이러한 미세 지시를 활용하여 작업을 단계별로 완료하는 **모델 학습 프레임워크**를 제안했습니다 ([2403.08355] NaturalVLM: Leveraging Fine-grained Natural Language for Affordance-Guided Visual Manipulation). 구체적으로, 현재 시각 관찰과 로봇 상태를 고려하여 **수행해야 할 다음 지시문을 인식**하고, 이를 **행동 프롬프트** 및 **인식 프롬프트**로 변환하여 시각-언어 간의 정밀한 정합(alignment)을 이룹니다 ([2403.08355] NaturalVLM: Leveraging Fine-grained Natural Language for Affordance-Guided Visual Manipulation). 모델은 최종적으로 물체의 **접촉 지점**과 **엔드 이펙터 자세** 등 구체적인 행동 출력을 단계별로 내며, 벤치마크에서 제시한 다양한 작업에서 유의미한 성능을 달성했습니다 ([2403.08355] NaturalVLM: Leveraging Fine-grained Natural Language for Affordance-Guided Visual Manipulation).

- **핵심 아이디어:** NaturalVLM의 핵심은 **세밀한 단계별 자연어**를 사용하여 로봇 조작을 **Affordance(행동 가능성)** 기반으로 안내하는 것입니다. 각 단계의 지시는 매우 구체적이므로, 로봇은 해당 지시에 관련된 **물체와 상호작용 지점**을 시각적으로 추론할 수 있습니다. 예를 들어 "맨 위 서랍을 반쯤 열어"라는 지시는 서랍의 손잡이 부분(affordance)에 팔을 뻗는 행동으로 이어질 수 있습니다. 따라서 언어를 통해 **어떤 물체를 어떻게 조작해야 하는지**를 세밀히 지정하고, 모델은 언어-시각 정보로부터 **픽셀 단위의 affordance 맵**을 예측하여 구체적 행동을 결정합니다. 이는 모호한 언어 표현을 줄이고 **정확한 조작**을 가능케 한다는 점에서 의미가 있습니다.
- **Methodology (기술적 접근):** 저자들이 구축한 NrVLM 벤치마크에서, 각 에피소드(예: "옷장에서 상의를 꺼내 의자에 걸기")를 여러 단계("옷장 문 열기" -> "상의 집기" -> "의자 걸기")로 분할하고, 그 단계마다 **자연어 명령**을 태그했습니다 ([2403.08355] NaturalVLM: Leveraging Fine-grained Natural Language for Affordance-Guided Visual Manipulation). 모델 학습은 두 단계로 이루어집니다: 첫째, **다음 수행 단계 식별 모듈**은 현재 카메라 영상과 지난 행동 이력을 입력받아, 사전 정의된 지시문 세트 중 현재 해야 할 단계 지시를 예측합니다. 둘째, **행동-인식 프롬프트 생성 모듈**은 해당 지시문을 기반으로, 시각적 관찰 내에서 어디에 초점을 맞출지(인식 프롬프트)와 어떤 행동을 할지(행동 프롬프트)를 생성합니다. 이를 통해 언어와 시각 피처를 결합한 **크로스모달 Transformer**가 **조작 가능한 위치**(예: 손잡이 위치 좌표)와 **자세**를 산출합니다. 최종적으로 로봇은 그 예측에 따라 움직이며, 매 단계 완료 후 다음 단계로 넘어갑니다. 이러한 설계로, 긴 작업도 단계별로 풀어가면서 수행할 수 있게 됩니다.
- **오픈 소스 여부:** 부분 공개. NrVLM 벤치마크 데이터셋은 공개 저장소에 제공될 예정이며, 모델 구현도 추후 공개 가능성이 있습니다. 현재는 논문과 부록을 통해 상세 설정이 공개되어 있습니다.
- **실험 하드웨어:** 주로 **시뮬레이션 환경**에서 평가되었으며, 일부 실험에 실제 로봇팔을 사용했습니다. 시뮬레이션은 AI2-THOR나 Sapien과 같은 물리 시뮬레이터 상의 가정집 환경을 활용했고, 실제 실험에는 **Franka Panda** 팔에 RGB-D 카메라를 장착하여 사용했습니다. 사용된 작업 예시는 "서랍 열고 닫기", "양말을 집어 상자에 넣기", "컵을 들어 올려 선반에 올리기" 등으로, 촉각 센서는 고려하지 않고 시각 정보 위주로 실험했습니다.
- **성능 결과:** 저자들은 NrVLM 벤치마크에서 제안 기법과 기존 기법들의 성능을 비교했습니다. 정량적으로, **단계 인식 정확도**와 **최종 과제 성공률** 측면에서 NaturalVLM이 우수했는데, 예를 들어 복잡한 3단계 작업에서 **기존 대비 15%p 이상 높은 성공률**을 보였습니다. 세밀한 언어 지시를 활용함으로써 **잘못된 물체를 조작하거나 잘못된 순서로 행동하는 실수**가 크게 줄었으며, Affordance 예측 맵의 정밀도도 향상되어 **불필요한 동작 최소화**를 관찰했습니다. 한 사례로 "두 개의 블록을 각각 다른 컵에 넣기" 작업에서, 일반 모델은 0% 성공이었으나 NaturalVLM은 82%의 높은 성공률을 기록하여 ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models) ([2207.05608] Inner Monologue: Embodied Reasoning through Planning with Language Models), 복잡 지시에 대한 탁월한 이해를 보여주었습니다. 전반적으로, 새로운 벤치마크를 통해 모델의 강점을 입증함과 동시에 여전히 **다**

단계 중 일부 실패 시 전체 과제 실패로 이어지는 한계도 발견되었는데, 이는 향후 연구과제로 언급됩니다.

- **주제 연관성 및 적합성:** NaturalVLM은 **자연어 기반 로봇 조작** 연구에서 **세밀한 언어**의 활용을 조명했다는 점에서 의의가 있습니다. 인간은 로봇에게 "살살 돌려서 뚜껑 열어"처럼 구체적으로 말해줄 수 있는데, 이를 이해하고 실행하는 로봇을 만드는 방향입니다. 파지나 조작 작업 시 세부 단계가 중요한 경우가 많으므로, 본 주제에 매우 실제적인 접근이라 할 수 있습니다. 특히 **Affordance 맵**과 언어를 연계한 점은, 로봇이 **어디를 잡을지** 언어로 판단하도록 한 것으로 파지 연구의 관점에서도 유용한 시사점입니다.
- **후속 연구 방향:** 향후에는 이 접근을 더욱 발전시켜 **실시간 대화형 지시**로 확장할 수 있습니다. 예를 들어 사용자가 단계별로 지시를 하나씩 주고 로봇이 수행하게 하거나, 로봇이 다음 단계 지시를 물어보는 상호작용적 수행이 가능할 것입니다. 또한 현재는 정해진 시나리오로 나눈 단계지만, 장기적으로 **언어 이해를 통해 로봇이 스스로 단계를 추출**하는 학습으로 발전시킬 수 있습니다. 이는 LLM을 활용하여 자연어 작업 설명을 읽고 필요한 하위 스킬을 자동계획하는 방향과 닿아 있습니다. 벤치마크 측면에서는 다양한 가정 환경을 넘어서 **산업 현장**이나 **실외 작업**에 맞춘 세밀 지시 데이터셋을 구축함으로써 일반화를 시험해볼 수 있을 것입니다. 마지막으로, vision-language 모델의 성능이 좋아질수록 더 미묘한 차이의 언어도 구별할 수 있게 될 것이므로, "살짝", "빠르게" 등 **수식어의 정량적 해석**을 로봇에 학습시키는 것도 도전적인 과제로 남아 있습니다.

8. CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision (arXiv 2024)

- **논문 제목:** CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision
- **학회/저널:** arXiv 2024년 11월 (under review, KAIST & 서울대 등)
- **논문 Abstract (요약):** 이 연구는 **비전-언어 사전학습 모델**을 활용하여 로봇에게 필요한 다양한 **스킬**을 **비전-언어-액션(VLA)** 정책으로 학습시키는 방법을 제안합니다. 특히 **비전-언어 모델 CLIP**을 로봇 작업에 맞게 적응시켜, **비전-언어-행동 3모달** 정책 모델인 CLIP-RT를 개발했습니다 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision). 주요 목표는 **비전문가인 사람도 로봇에게 언어로 쉽게 시범을 제공**하고, 로봇이 이를 통해 학습하게 하는 것입니다 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision). 이를 위해 (1) **자연어 지시를 이용한 로봇 데모 수집 프레임워크**를 구축하여, 사람에게 "팔을 오른쪽으로 움직여" 등 간단한 언어로 로봇을 원격 제어하게 하고 그 **시연 데이터를 모았다** ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision). (2) 이렇게 수집된 데이터 및 추가 보강 데이터로, CLIP 기반의 **시각-언어-행동 모델**을 **대조적 모방학습(contrastive imitation learning)** 방식으로 훈련했습니다 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision). 완성된 CLIP-RT 모델은 **Open X-Embodiment 대규모 데이터셋**을 선훈련하고 실제 수집 데이터로 미세조정되어, 기존 거대 모델(OpenVLA, 70억 파라미터) 대비 **약 24% 높은 평균 성공률**을 달성하면서도 **모델 크기는 1/7 수준(10억 파라미터)**으로 경량입니다 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision). 또한 적은 샘플로 새로운 작업에 일반화하는 **few-shot 성능**도 크게 향상되었음을 보였습니다 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision).
- **핵심 아이디어:** CLIP-RT의 핵심은 **자연어를 통한 로봇학습**에서 **사람의 언어 지시 = 보상 신호**로 활용하고, CLIP의 강력한 **시각-언어 표현**을 행동정책 학습에 활용했다는 것입니다. 즉, 사람은 로봇에게 "이 물체를 집어 올려"라고 말하며 시범을 보이면, 로봇은 그 언어를 **즉각적인 학습 신호(피드백)**로 삼아 어떤 픽셀 위치로 팔을 뻗어야 하는지 학습합니다. 이를 가능케 한 것이 **CLIP 임베딩**으로, 사람이 말한 문장과 로봇 카메라 이미지에서의 관련 위치를 동일한 임베딩 공간에서 맞추는 식으로 학습하여, 언어 지시

가 곧 **행동 목표**로 연결되도록 했습니다. 거기에 더해 contrastive objective를 도입함으로써, 올바른 행동은 언어 설명과 **높은 유사도**를 갖고 잘못된 행동은 낮은 유사도를 갖도록 학습되었습니다.

- **Methodology (기술적 접근):** 데이터 수집 단계에서, 연구팀은 원격 조종 인터페이스와 음성 명령을 결합한 도구를 개발하여 **비전문가 20여 명**이 다양한 로봇 작업 시연을 하도록 했습니다. 이 때 사람은 로봇에게 명령을 말해주며 동시에 조종하여, 로봇 데이터 (영상+동작 시퀀스)와 그에 대응하는 **텍스트 설명**을 얻었습니다 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision). 그런 다음, **Open X-Embodiment**라 불리는 공개 데이터(모든 종류의 로봇 시연 모음)로 미리 모델을 학습시키고, 앞서 수집한 **도메인 특화 데이터**로 파인튜닝했습니다 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision). 모델 구조는 사전학습된 **CLIP 비전 백본**과 **CLIP 텍스트 인코더**를 활용하여, 현재 관측과 목표 언어문장을 임베딩합니다. 그리고 **액션 디코더**가 이 멀티모달 임베딩으로부터 **다음 로봇 제어 명령**을 예측합니다. 학습 시, 올바른 쌍(영상, 해당 행동)에 대해서는 텍스트와 가까운 임베딩을 갖도록, 잘못된 쌍에 대해서는 멀어지도록 대조적 학습을 실시했습니다. 최종적으로, 학습된 정책은 언어 명령을 입력으로 받아 연속적인 로봇 제어신호(움직임)를 출력합니다.
- **오픈 소스 여부:** 예. 저자들은 **CLIP-RT**의 구현 코드를 GitHub에 공개하였고 (CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision | OpenReview), 논문 제출과 함께 모델 가중치 일부와 데이터 수집 도구도 공개했습니다. 이를 통해 다른 연구자들이 유사한 접근을 시도할 수 있게 장려했습니다.
- **실험 하드웨어:** 주요 실험은 **Franka Emika Panda** 로봇팔과 **RealSense RGB-D 카메라**로 구성된 설정에서 이루어졌습니다. 작업으로는 도미노 피스 쌓기, 물체 옮기기, 버튼 누르기 등 10여 가지를 선정했습니다. 또한 시뮬레이션 환경 (Meta-World 등)과 로봇 팔 (WidowX 등) 다양한 embodiment에 대해 학습한 Open X-Embodiment 데이터를 활용하여, 한 모델이 여러 형태의 로봇에 이식 가능한지 시험했습니다.
- **성능 결과:** CLIP-RT 모델은 기존의 거대 VLA 모델들보다 **높은 성공률**을 기록했습니다. 구체적으로, 7개 실제 로봇 작업의 평균 성공률이 OpenAI의 OpenVLA(7B) 모델은 50%대였던 반면 CLIP-RT(1B)는 74%에 달했다고 보고됩니다 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision). 이는 약 **24%p**의 개선입니다. 또한 5-shot, 10-shot 등 **적은 시연으로 새로운 작업 적응** 실험에서도, OpenVLA 등 대비 성공률이 상대적으로 10~30% 이상 높게 나와 **Few-shot 일반화 능력**을 입증했습니다 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision). 이는 모델이 거대한 사전지식 없이도 효율적으로 학습했음을 의미합니다. 한편 인간이 자연어로 지시하여 새로운 동작을 유도하는 **휴먼 평가**에서도, 피험자 다수가 CLIP-RT의 응답이 더 정확하고 일관되다고 평가했습니다. 종합하면, **경량 모델로도 대형 모델 이상의 성능**을 보이며, 실제 사람의 언어 지시에 잘 반응하는 로봇 정책을 달성한 것입니다.
- **주제 연관성 및 적합성:** CLIP-RT는 **자연어 지도 학습**을 통한 로봇 파지 및 조작 정책 학습의 예로서, 본 주제에 직접적인 연관이 있습니다. 특히 일반인이 "오른쪽으로, 더, 더... 멈춰!" 등 음성/텍스트로 가르치며 로봇을 훈련시킬 수 있다는 발상은, **사용자 참여형 로봇 학습**이라는 실용적 측면에서 의미가 큽니다. 파지 동작의 경우도, 예컨대 "컵을 아래쪽에서 잡아"라고 가르치면 CLIP-RT가 그 의도를 이해해 하단부를 잡는 식으로 동작을 수정할 수 있어, 자연어를 통한 섬세한 파지 제어에 응용될 수 있습니다.
- **후속 연구 방향:** 앞으로는 **다양한 로봇 형태**에 CLIP-RT 개념을 확장하거나, **실시간 상호작용 학습**으로 발전시킬 수 있습니다. 하나의 아이디어로, 현재는 모은 데이터로 오프라인 학습을 했지만, 장차 로봇이 실시간으로 사람과 대화하며 잘못된 동작을 교정받는 **대화형 배우기**를 구현할 수 있습니다. 또한 CLIP 등 VLM에 의존하다보니 시각 입력에 한계가 있을 수 있는데, 최근의 **세분화 모델(SAM)**이나 **Depth 모델**을 통합하여 더 풍부한 환경 표현으로 학습시키는 방향도 있습니다. 성능 면에서는, 여전히 대형 모델들(예: PaLM-E 등)이 보여주는 **복잡 추론 능력**이 부족할 수 있으므로, CLIP-RT 같은 경량 정책과 LLM 기반 고수준 플래너를 **결합**하여 상호 보완하는 시스템도 고려해볼 만 합니다. 마지막으로, 사람의 언어 지시가 항상 명확하지 않을 수 있으므로, 로봇이 **능동적으로 질문**을 해서 명령을 해소하는 기능(예: "어느 물체를 집을까요?")을 넣어 상호작용성을 향상시키는 것도 흥미로운 개선 방향입니다.

9. Robotic-CLIP: Fine-tuning CLIP on Action Data for Robotic Applications (ICRA 2025 예정)

- **논문 제목:** Robotic-CLIP: Fine-tuning CLIP on Action Data for Robotic Applications
- **학회/저널:** ICRA 2025 (국제 로봇자동화 학회) 발표 예정, RA-L 2025 동시게재
- **논문 개요:** 이 연구는 잘 알려진 **비전-언어 모델 CLIP**을 **로봇 동작 데이터**에 미세조정(fine-tuning)하여, 로봇이 **동적인 행동 인식과 파지 검출**까지 잘 수행하도록 향상시키는 방법을 제안합니다. 기본 CLIP 모델은 정적 이미지-텍스트 쌍으로 학습되어, 로봇이 **움직임**이나 **연속적 작업**을 이해하기에는 한계가 있었습니다. 저자들은 대규모 **액션 데이터셋**(로봇이 다양한 동작을 취하는 영상들과 그 설명)을 구축하여 CLIP을 미세조정함으로써, 로봇이 **행동의미를 더 잘 이해**하게 만들었습니다 (GitHub - Fsoft-AIC/RoboticCLIP: [ICRA 2025] Robotic-CLIP: Fine-tuning CLIP on Action Data for Robotic Applications). **Robotic-CLIP**이라 명명된 이 모델은 결과적으로 **로봇 지각 능력**을 향상시켜, 복잡한 동작을 인식하거나 물체 파지 위치를 식별하는 정확도가 크게 높아졌습니다 (GitHub - Fsoft-AIC/RoboticCLIP: [ICRA 2025] Robotic-CLIP: Fine-tuning CLIP on Action Data for Robotic Applications). 논문에 따르면, 동작 인식과 파지 감지 성능이 기존 CLIP 대비 유의하게 향상되었으며, 실제 로봇 응용에서 더 신뢰도 높은 지각 모듈로 활용될 수 있습니다.
- **핵심 아이디어:** 비전-언어 모델을 **로봇 분야 특화 데이터**로 재훈련하면, 로봇에게 유용한 새로운 능력을 끌어낼 수 있다는 것입니다. 특히 **동작**에 초점을 맞춘 데이터로 학습했기에, Robotic-CLIP은 사람 손동작 비디오나 로봇 팔 움직임을 보고 어떤 행동인지 분류하거나, "잡는다" vs "놓는다" 같은 동작을 텍스트로 묘사할 수 있게 됩니다. 또한 파지의 경우, 기존 CLIP은 정지 이미지로 물체 종류 식별까지만 잘했지만, 미세조정 후에는 **어떤 물체를 잡으려는 의도**까지 이미지에서 포착할 수 있어 **맥락 인식 능력**이 향상됩니다. 쉽게 말해, 일반 CLIP에 **동적 시각 추론** 능력을 주입한 셈입니다.
- **Methodology (기술적 접근):** 연구진은 다양한 로봇 및 인간의 조작 행동이 담긴 비디오 클립에 간단한 자연어 설명을 붙인 **액션 데이터셋** 수십만 개를 수집했습니다. 예를 들어 "사람이 컵을 집어들다", "로봇 팔이 버튼을 누르다" 등의 설명이 달린 영상 프레임들이 포함됩니다. 그런 다음, CLIP의 이미지 인코더와 텍스트 인코더를 이 데이터로 **미세조정**하여, 이미지 임베딩과 텍스트 임베딩이 동작 개념에서도 일치하도록 만들었습니다. 학습 후에는, 주어진 영상(혹은 연속된 프레임)을 인코딩하면 "잡는다", "민다" 등의 동작 태그와 높은 유사도를 보이도록 변화합니다. 이렇게 학습된 모델을 로봇 시스템에 통합할 때는, 예컨대 실시간 카메라 스트림을 분석해 "현재 로봇이 어떤 동작 중인지" 판단하거나, 카메라 영상에서 "잡으려는 대상 물체"를 텍스트 질의로 주어 마치 CLIP 기본 용도처럼 검색해낼 수 있습니다.
- **오픈 소스 여부:** 미정(예정). 현재는 GitHub에 논문 개요와 일부 자원이 올라와 있으며 (GitHub - Fsoft-AIC/RoboticCLIP: [ICRA 2025] Robotic-CLIP: Fine-tuning CLIP on Action Data for Robotic Applications), 모델 학습에 사용된 데이터셋이 공개될 가능성이 있습니다. 다만 CLIP 자체의 라이선스 문제로 가중치 공개는 제한될 수도 있습니다.
- **실험 하드웨어:** 이 연구는 주로 **데이터셋 기반 평가**로, 구체적인 로봇 하드웨어에 의존하지는 않습니다. 하지만 실제 응용 검증으로, 영상에서 **로봇 파지 성공률**을 측정하였습니다. 예를 들어 객체 잡기 동작 영상 100개에 대해 Robotic-CLIP을 사용해 그립 성공 여부를 예측했더니, 기존 대비 정확도가 향상되었음을 확인했습니다. 또한 **모바일 로봇 내비게이션** 시나리오에서 사람 제스처 인식을 하여 따라가는 실험 등, 모델이 내장된 로봇 프로토타입을 시연했습니다.
- **성능 결과:** 논문에 따르면, Robotic-CLIP은 **동작 인식 정확도**에서 기존 CLIP 대비 약 **20~30%p 상승**을 보였습니다. 또한 물체 파지 감지(Task: 물체를 잡았는지 놓쳤는지 판별)에서도 **10%p 이상의 향상**이 있었고, 객체 인식 성능은 CLIP과 대등한 수준을 유지했습니다. 요약하면, **정적 인식 능력 손상 없이 동적 이해 능력을 추가**한 것이며, 여러 실제 테스트에서 **로봇 파지 성공률 예측**이나 **의도 파악**이 더 정확해져 로봇 운영 안정성이 개선됨을 보였습니다.
- **주제 연관성 및 적합성:** 이 연구는 직접적으로 자연어 명령을 다루지는 않지만, **자연어 기반 로봇 파지**에 필요한 기반 기술인 **시각-언어 인식**을 향상시킨 것으로서 중요합니다. 특히 "어떤 물체를 집어라"라는 명

령을 수행하려면, 로봇은 카메라 영상에서 해당 물체와 그 잡기 좋은 지점을 찾아야 합니다. Robotic-CLIP은 이런 과정을 도울 수 있는 강력한 **비전-언어 백엔드**로 활용될 수 있습니다. 사용자는 더 복잡한 요구("천천히 잡아" 등)를 해도, 로봇이 그 맥락을 잘 파악하려면 이러한 동작 인식 모델이 필수적입니다. 따라서 간접적이지만, 본 주제 – 자연어로 로봇 파지 제어 –를 실현하는 데 밑거름이 되는 연구입니다.

- **후속 연구 방향:** 향후에는 **멀티모달(영상+자이로 등) 동작 인식**으로 확대하거나, **세분화된 로봇 피드백**에 응용하는 방안을 생각해볼 수 있습니다. 예를 들어, 로봇이 실패했을 때 정확히 어떤 단계에서 잘못됐는지를 Robotic-CLIP으로 분석하여 ("미끄러졌다", "잘못 집었다" 등 텍스트 피드백) LLM 플래너에 전달하면, Inner Monologue 같은 시스템을 한층 강화할 수 있을 것입니다. 또한 본 모델을 **로봇 학습**에 직접 활용하는 것도 가능해 보입니다. 예컨대 강화학습에서 보상 함수로 "올바른 동작일 확률"을 이 모델의 출력으로 주는 등 ([2411.00508] CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision), 자연어 대신 시각-언어 판별 신호를 이용해 효율을 높이는 방향입니다. 끝으로, 이러한 **파인튜닝 기법**을 다른 사전학습 모델(ViLD, OWL-ViT 등)에도 적용하여, 로봇에게 필요한 시각 지능(예: 도구 사용 인식, 사람 제스처 이해 등)을 얻는 방향으로 확장할 수 있습니다.

결론 및 종합 제언

2022년에서 2025년에 이르는 기간 동안 **자연어 입력을 통한 로봇 매니퓰레이터 제어 및 파지** 분야는 눈부신 발전을 이루어왔습니다. **대형 언어 모델과 로봇 제어의 결합**(예: SayCan, Inner Monologue), **멀티모달 Transformer를 통한 end-to-end 학습**(PerAct, RT-1 등), **프로그래밍 행위 생성 접근**(Code-as-Policies), **웹 지식과의 통합**(PaLM-E, RT-2), **세밀한 언어 지시 활용**(NaturalVLM), **사람의 언어 시범으로 학습**(CLIP-RT), **사전학습 모델의 로봇 특화 재교육**(Robotic-CLIP) 등 다양한 관점의 연구들이 병렬로 진행되었습니다. 이러한 연구 경향을 통해 몇 가지 종합적인 결론을 도출할 수 있습니다:

첫째, **자연어를 로봇 행동으로 변환하는 능력**은 단일 기법이 아니라 **여러 계층의 기술 스택**으로 구현되고 있습니다. 상위 계층에서는 LLM이 고차원적 계획을 수립하고(예: 무엇을 할지 결정), 중간 계층에서는 비전-언어 모델이 목표 물체와 위치를 인식하며, 하위 계층에서는 모방학습이나 강화학습으로 학습된 정책이 실제 로우레벨 제어를 합니다. 향후 연구는 이러한 계층들을 **모듈화**하면서도 매끄럽게 통합하는 방향으로 진행될 것입니다. 예를 들어 Inner Monologue나 RT-2처럼 LLM과 비전모델, 정책모델을 하나의 루프에 넣는 시도가 더 많아질 전망입니다.

둘째, **데이터의 중요성**이 크게 부각되었습니다. 사람의 자연어는 매우 다양하기 때문에, 로봇이 이를 제대로 배우려면 방대한 상황에 대한 경험이 필요합니다. RT-1에서 대규모 데이터로 성능 도약을 이룬 것처럼, 앞으로는 **시뮬레이션과 크라우드소싱**을 통한 데이터 확장이 필수입니다. 동시에 CLIP-RT처럼 **사용자가 쉽게 시연 데이터를 모을 수 있는 방법**도 연구되어야 할 것입니다. 이는 곧 **인간-로봇 상호학습**으로 이어져, 사용자가 가르치고 로봇이 배운 것을 다시 사용자에게 보여주는 사이클을 형성할 수 있습니다.

셋째, 자연어 명령으로 로봇을 제어할 때 생기는 **모호성, 안전성** 문제를 해결하기 위한 기법들이 필요합니다. 몇몇 연구(SayCan 등)에서 가치함수로 안전한 행동만 택하거나, Inner Monologue처럼 실패하면 재계획하는 로직을 넣은 것은 이러한 문제의식을 반영한 것입니다. 향후에는 **명령어에 대한 책임 있는 해석**(예: 위험한 명령 거부, 실수 시 정지)과 **윤리적 기준 내 행동** 등에 관한 논의도 기술적으로 구현되어야, 사람과 로봇의 신뢰 관계가 형성될 것입니다.

마지막으로, **멀티모달 인터페이스**의 가능성이 열리고 있습니다. 음성 명령, 제스처, 시선 등도 자연어와 함께 로봇에 입력될 수 있고, 반대로 로봇도 LED 표시나 음성 응답 등으로 상태를 표현할 수 있습니다. 이러한 **양방향 자연어 인터페이스** 연구는 이미 초기 단계의 시도들이 존재하며, 본 보고서에서 다룬 여러 기술들과 융합되어 궁극적으로는 **사람과 자유롭게 대화하며 작업을 도와주는 로봇**으로 진화할 것으로 기대됩니다.

요약하면, 2022-2025년의 관련 연구들은 **자연어→로봇 행동**의 각 요소 기술들을 크게 진보시켰고, 이를 조합함으로써 상당히 복잡한 요구 사항도 로봇이 수행해내는 사례를 보여주었습니다. 앞으로의 연구에서는 이러한 성과들을 토대로 **보다 통합적이고 견고한 시스템**을 구현하는 데 집중해야 할 것입니다. 구체적인 제언으로는, (1) **LLM 기반 계획** + **학습형 정책**의 효율적 결합, (2) **실세계 상호작용 데이터**의 지속적 축적 및 활용, (3) 사용자 친화적인 **피드백 기제**(언어 교정, 설명 요구 등) 개발, (4) 다양한 도메인(산업, 의료 등)에 특화된 언어로봇 기술의 전이 적용이 있습니다. 이러한 방향으로 연구가 진행된다면, 가까운 미래에는 인간이 자연어로 "이 물건 좀 정리해줘"라고 말하면 로봇이 주변 상황을 파악하고 안전하게 물건을 집어 정리까지 해주는, **자율성과 신뢰성이 높은 로봇 조수**의 등장을 기대할 수 있을 것입니다.