



2. PaLM-E: An Embodied Multimodal Language Model

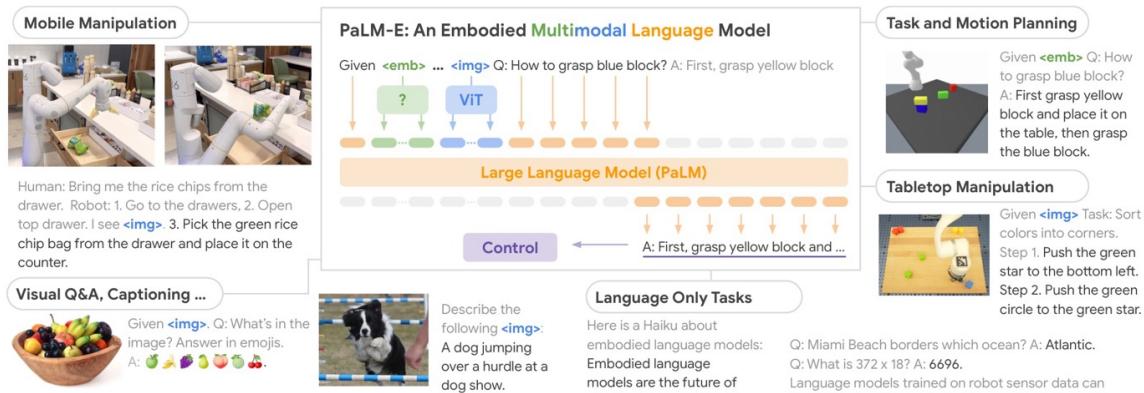


Figure 1: PaLM-E is a single general-purpose multimodal language model for embodied reasoning tasks, visual-language tasks, and language tasks. PaLM-E transfers knowledge from **visual-language** domains into **embodied reasoning** – from robot planning in environments with complex dynamics and physical constraints, to answering questions about the observable world. PaLM-E operates on **multimodal sentences**, i.e. sequences of tokens where inputs from arbitrary modalities (e.g. images, neural 3D representations, or states, in green and blue) are inserted alongside text tokens (in orange) as input to an LLM, trained end-to-end.

0. Abstract

- 대형 언어 모델은 복잡한 작업을 수행하는 것으로 입증됨
- 단어와 지각 간의 연결을 확립하기 위한 도전 과제가 제기됨
 - 실제 세계의 연속적인 센서 모달리티를 언어 모델에 직접 통합하여 시각적, 연속적인 상태 추정 및 텍스트 입력 인코딩을 교차하는 **멀티모달 문장**을 입력으로 사용

▼ 멀티모달(Multi-modal)

- 다양한 센서 모드 또는 정보 유형을 포함하는 시스템 또는 모델을 나타내는 용어 → 다양한 감각을 사용하여 정보를 수집하고 처리
- 여러 가지 입력 유형을 처리하고 이러한 다양한 입력 유형을 통합하여 효과적인 결과를 생성하는 데 사용
ex) 이미지, 텍스트, 음성, 동영상 등 다양한 형식의 데이터를 동시에 처리하고 정보를 추출하거나 작업을 수행하는 데 사용
- 다양한 종류의 정보를 통합하여 보다 정확하고 효과적인 결정을 내리고 작업을 수행하는 데 도움을 줄 수 있으며, 이는 다양한 현실 세계 응용 분야에

서 중요한 역할을 함

- 이러한 인코딩을 대규모 언어 모델과 함께 다양한 실제 과제를 위해 end-to-end로 훈련
- 순차적인 로봇 조작 계획, 시각적 질문 응답, 캡션 작성 등이 과제에 포함됨
- 평가 결과, 단일 대규모 멀티모달 모델인 **PaLM-E**는 다양한 관찰 모달리티에서 다양한 결합체에 대한 다양한 실제 추론 과제를 처리할 수 있으며, 더불어 긍정적인 전이를 보임
 - 인터넷 규모의 언어, 비전 및 시각 언어 도메인에서 다양한 공동 훈련을 통해 이점을 얻으며, 562B 파라미터로 구성된 최대 모델인 PaLM-E-562B는 로봇 과제를 훈련한 것 외에도 OK-VQA에서 최첨단 성능을 보이며 증가하는 규모에서의 일반 언어 기능을 유지

1. Introduction

Visually-conditioned Jokes + Few-shot prompting



I see (image omitted). Description: A cow sitting on a rug. Joke: Why is the cow sitting on a rug? Because it wants to be a cow-ch! ... I see . Description: a donkey is carrying a dog, cat, and rooster. Joke: what do you call a donkey with a rooster on his back. A rooster booster.

Robot Visual Perception, Dialogue, and Planning



Robot: I am a robot operating in a kitchen. Given , what do I see? a bowl of eggs, a bowl of flour, and a bowl of sugar.. and a spatula.. and a spoon.. and a bag of flour.. and a bag of sugar. When a human asks me to do a task, I will respond with the sequence of actions I would do to accomplish the task with only the items I see. Human: Use all of the ingredients you see to make a cake batter: 1. crack egg. 2. put egg in bowl. 3. put flour in bowl. 4. put sugar in bowl. 5. mix. 6. put in pan. 7. bake. 8. eat. 9. clean up.

Zero-shot Multimodal Chain-of-Thought



Given . Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step. 1. do not enter. 2. except bicycles. 3. do not entry except bicycles. 4. yes.

Robot Initiative



Given . Q: If a robot wanted to be useful here, what steps should it take? A: clean table.. pick up trash.. pick up chairs.. wipe chairs.. put chairs down.

Physical Prediction



Given . Q: What will the robot do next? A: fall.

Spatial Grounding



Given . Q: What flavor is the doughnut on the right? Strawberry. Q: What flavor is the doughnut on the left? A blueberry.

Zero-shot: Multi-image Relationships



Q: Where is in this scene ? A: a stove top with a pot on it.

OCR-free Math Reasoning



Given . Q: Who are the two teams playing in this photo, and which was the last to win a championship, and which year did they win, and who was their star player that year? A: Let's think step by step. 1. team in white is new york knick. 2. team in green is boston celtic. 3. last to win championship was boston celtic. 4. year was 2008. 5. star player was paul pierce. 6. i am not sure if this is correct. i am not basketball fan. i just google it.

Q: I am getting just two custom pizza for me and my friend. How much should I pay in total? A: Let's think step by step. 1st step: 2 custom pizzas. 2nd step: 2 x \$9.99. 3rd step: \$19.98.

Photo 1: <img1>. Photo 2: <img2>. Photo 3: <img3>.

Photo 1 at 10:30 am: <img1>. Photo 2, at 12:45 pm: <img2>. Photo 3, at 3:45 pm: <img3>. Q: I forgot, what did I have for lunch, and what time was it? A: Let's think step by step. 1. you had a sandwich for lunch. 2. it was 12:45 pm.

End-to-End Egocentric Q&A, Dialogue



- **대형 언어 모델(Large Language Models, LLMs)**은 대화, 단계별 추론, 수학 문제 해결 및 코드 작성과 같은 다양한 도메인에서 강력한 추론 능력을 나타냄
 - 그러나 실제 세계에서 추론을 위해서는 **grounding(기초화)** 문제를 해결해야 함
 - 대규모 텍스트 데이터에서 LLMs를 훈련하는 것은 물리적 세계와 관련된 표현을 만들어낼 수 있지만, 이러한 표현을 실제 세계의 시각적/물리적 센서 모달리티와 연결하는 것이 중요

시각적/물리적 자료 ↔ 언어적 자료

- LLM의 출력을 학습된 로봇 정책 및 affordance 함수와 연결하여 결정을 내리도록 하는 시도가 있었음

▼ affordance 함수

- 물체나 환경이 제공하는 기능 또는 사용 가능성을 나타내는 개념
- 주로 로봇 공학 및 컴퓨터 비전 분야에서 사용
 - 환경에서 로봇이 수행할 수 있는 작업 또는 행동을 모델링하는 데 사용
- 환경의 다양한 요소와 상호작용하여 특정 작업이나 동작을 수행하는 데 얼마나 적합한지를 나타내는 함수
- LLM 자체에는 텍스트 입력만 제공되는 한계가 있었음
 - 장면의 기하학적 구성이 중요한 많은 작업에는 충분하지 않았음
- 더불어, 현재 최첨단의 시각 언어 모델이 시각-언어 작업(ex> 시각적 질문 응답)을 직접적으로 해결할 수 없다는 것이 입증됨
- 해당 논문에서는 연속 입력을 직접적으로 통합하는 **물리 세계에 내재된 언어 모델 (embodied language models)**을 제안
 - 이미지 및 상태 추정과 같은 입력은 언어 토큰과 동일한 임베딩에 포함되며 텍스트와 동일한 방식으로 Transformer 기반 LLM의 self-attention 레이어에 의해 처리

▼ self-attention 레이어

- 트랜스포머(Transformer) 딥러닝 모델 아키텍처에서 사용되는 중요한 구성 요소 중 하나
- 주요 목적: 입력 시퀀스의 각 요소 간의 상호 작용 모델링
- 특징

1. 상호 작용 모델링

- 입력 시퀀스의 각 요소가 다른 요소에 얼마나 많은 중요성을 둘지를 결정 → 문장 내의 단어 간의 의미적 관계를 파악할 수 있음

2. 병렬 처리 가능

- Self-attention은 모든 입력 요소 간의 상호 작용을 병렬로 처리할 수 있음 → 다른 종류의 RNN (순환 신경망) 기반 아키텍처보다 계산 효율성이 높음

3. 문맥 고려

- Self-attention은 문장 내의 모든 단어를 고려하여 각 단어의 임베딩 벡터를 조절 → 단어의 문맥을 고려한 표현을 얻을 수 있음
- 연속 입력을 미리 훈련된 LLM에서 시작하여 encoder를 통해 연속 입력을 주입
 - encoder는 자연어 텍스트로 순차적인 결정을 출력하도록 end-to-end로 훈련되며, 이 결정은 물리적 에이전트에 의해 해석될 수 있도록 낮은 수준의 정책을 조건화하거나 물리적 질문에 답변을 제공
- 다양한 설정에서 해당 접근법을 평가하고, 시각 입력에 대한 표준 vs 객체 중심 ViT 인코딩과 같은 다양한 입력 표현을 비교하며, 언어 모델을 훈련하는 동안 encoder를 동결하거나 fine-tuning하는 것, 여러 작업에 대한 공동 훈련이 전이를 가능하게 하는지 등을 조사
- 접근법의 폭을 조사하기 위해 로봇 조작 영역(현실 세계에서의 닫힌 루프를 포함하는 두 가지)에서 세 가지 로봇 조작 영역, VQA 및 이미지 캡션과 같은 표준 시각-언어 작업, 그리고 언어 작업에서 평가
 - 여러 작업에 대한 모델 훈련이 개별 작업에 대한 모델 훈련보다 성능을 향상시킨다는 것을 확인
 - 작업 간 전이가 로봇 과제에 대한 데이터 효율성을 향상시킬 수 있음을 확인
 - 학습 예제 수를 줄여 학습 성공률을 크게 높일 수 있으며, 새로운 객체 or 볼 수 없는 객체의 새로운 조합에 대한 one-shot 또는 zero-shot 일반화를 보여줌
- **PaLM-E-562B**
 - 540B PaLM LLM과 22B Vision Transformer(ViT)을 통합
 - 현재로서는 가장 큰 시각-언어 모델
 - 작업별 특수한 fine-tuning을 필요로 하지 않고, OK-VQA 벤치마크에서 최고 수준의 성능을 달성함
 - 단일 이미지 예제만을 훈련한 경우에도 zero-shot 멀티모달 chain-of-thought(CoT) 추론, few-shot 프롬프팅, OCR-free 수학 추론 및 멀티 이미지 추론 등 다양한 작업을 수행할 수 있는 것을 발견
 - 원래 언어만을 다루는 개념이었던 zero-shot CoT는 작업별 프로그램을 통한 멀티 모달 데이터에서 이미 나타났지만, end-to-end 모델을 통한 것은 처음임
- **주요 기여**
 - 일반적인 전이 학습된 multi-embodiment 결정 에이전트를 제안하고 이를 다중 모달 대규모 언어 모델의 훈련에 물리적 데이터를 혼합하여 훈련할 수 있다는 것을 보여줌

- 현재의 최첨단 일반적 목적의 시각-언어 모델은 기본 상태에서(zero-shot) 실제 세계 문제를 해결하지 못하지만, 훈련된 일반 목적 시각-언어 모델이 효과적인 실제 세계 문제를 해결할 수 있음을 보여줌
- 신경망 장면 표현 및 객체 레이블링 멀티모달 토큰과 같은 새로운 아이디어를 도입
- **PaLM-E**가 양적으로 뛰어난 비전 및 언어 일반 목적 모델임을 보여줌
- 언어 모델의 크기를 확장하면 치명적인 기억 손실이 적은 multi-modal fine-tuning 이 가능함을 입증함

2. 관련 연구

- 일반적인 vision-language 모델

- 이전에는 이미지와 텍스트를 동시에 이해하고 처리할 수 있는 대형 비전-언어 모델 (VLMs)가 등장하지 않았지만, 이제 VLMS는 이미지와 텍스트를 동시에 처리할 수 있으며 시각적 질문 응답, 캡션 생성, 광학 문자 인식 및 객체 감지와 같은 작업에 적용될 수 있음
- 이미지 통합 방법
 - **PaLM-E**는 이미지와 텍스트를 잠재 벡터의 다중 모달 문장으로 표현하여 문장의 어느 부분에서는 유연하게 여러 이미지를 처리할 수 있음
 - **Frozen**: 비전 인코더 매개 변수가 언어 모델을 고정하고 역전파를 통해 최적화
⇒ **PaLM-E**는 대체 입력 모달리티(ex. neural 장면 표현)를 도입 → 성과

- 액션 출력 모델

- 이전 작업은 실제 세계에서 로봇 또는 에이전트가 주어진 시각적 정보와 언어적 정보를 기반으로 어떤 동작을 취할지를 예측하는 것에 초점
 - **VIMA**: 다중 모달 프롬프트, 언어의 역할은 주로 작업 명세를 설명하는 것임
 - **PaLM-E**: 텍스트로 고수준 명령을 생성 → 자체 예측에 대한 조건을 자연스럽게 설정하고 매개 변수에 내장된 세계 지식을 직접 활용할 수 있음 → 실제 세계에서의 단순 추론뿐만 아니라 질문에 대한 답도 할 수 있음
 - **Gato**: 일반적인 다중 모드 에이전트

- 내재된 작업 계획

- 임베디드 도메인에서 LLMs(대형 언어 모델)를 활용하기 위한 여러 방법들이 제안되었지만, 계획을 위한 표현으로 자연어를 고려하는 시도는 적음

- **PaLM-E**는 타 모델과 달리 접지(grounding)에 대한 보조 모델에 의존하지 않고 직접적으로 계획을 생성하도록 훈련
→ pre-trained LLMs에 저장된 풍부한 의미 지식을 계획 프로세스에 직접 통합할 수 있도록 함

3. PaLM-E: 내제된 멀티모달 언어 모델

- **PaLM-E의 주요 아키텍처 아이디어**
 - 이미지, 상태 추정 또는 기타 센서 모달리티와 같은 연속적인, 실제 현장 관측을 사전 훈련된 언어 모델의 언어 임베딩 공간으로 주입하는 것
 - 연속적인 관측을 언어 토큰의 임베딩 공간과 동일한 차원의 벡터 시퀀스로 인코딩함으로써 실현
 - 연속적인 정보는 언어 모델에 언어 토큰과 유사한 방식으로 주입됨
 - PaLM-E는 **디코더 전용** 언어 모델
 - 접두사 또는 프롬프트가 주어진 경우 자동 회귀적으로 텍스트 완성을 생성
 - 사전 훈련된 언어 모델로 **PaLM**을 사용하고, 이를 **Embodied(구체화된)**로 만듦 → PaLM-E
 - PaLM-E의 **입력**
 - 텍스트 + (다중) 연속 관측
 - 관측에 해당하는 multi-modal 토큰은 텍스트와 교차하여 multi-modal 문장을 형성
 - ex) "Q: 와 사이에 무슨 일이 일어났나요?"
 → 이미지의 임베딩
 - PaLM-E의 **출력**
 - 모델에 의해 자동 회귀적으로 생성된 텍스트
→ 질문에 대한 답변 or PaLM-E가 텍스트 형식으로 생성한 결정의 시퀀스
 - PaLM-E가 결정이나 계획을 생성하는 작업을 수행할 때, 저수준 정책 또는 플래너가 이러한 결정을 저수준 동작으로 변환할 수 있다고 가정
- **Decoder-only LLMs**
 - 토큰의 시퀀스 $w_i \in W$ 의 텍스트 조각인 $w_{1:L} = (w_1, \dots, w_L)$ 의 확률 $p(w_{1:L})$ 을 예측하기 위해 훈련된 생성 모델

$$p(w_{1:L}) = \prod_{l=1}^L p_{\text{LM}}(w_l | w_{1:l-1}),$$

- 전형적인 신경 아키텍처는 이를 p_{LM} 이라는 대형 트랜스포머 네트워크로 분해하여 구현

- **접두사 decoder-only LLMs**

- LLM은 자기 회귀적(autoregressive)
 - 사전 훈련된 모델은 아키텍처를 변경하지 않고도 접두사 $w_{1:n}$ 에 조건을 걸 수 있음
- 접두사(or 프롬프트) $w_{1:n}$ 은 LLM이 이어지는 토큰 $w_{n+1:L}$ 을 예측할 때 기반이 되는 맥락을 제공
 - 모델의 예측을 조절하기 위한 추론에 자주 사용됨

- **토큰 임베딩 공간**

- 토큰 w_i : 자연어의 (부)단어에 해당하는 이산적이고 유한한 고정 어휘 집합 W 의 요소
- 내부적으로 LLM은 w_i 를 단어 토큰 임베딩 공간 $\chi \subset \mathbb{R}^k$ 로 임베딩
 - 임베딩 매핑은 일반적으로 크기가 큰 임베딩 행렬로 표현됨
 - end-to-end로 훈련됨
 - 연구에서 어휘 집합 W 의 크기는 25600으로 정의되어 있음

- **멀티모달 문장: 연속적인 관찰 주입**

- 이미지 관측과 같은 멀티모달 정보는 이산형 토큰 수준을 건너뛰고 연속적인 관측 값을 직접 언어 임베딩 공간 χ 로 매핑하여 LLM에 주입할 수 있음
 - 연속적인 관측 공간 O 를 χ 의 q 개의 벡터 시퀀스로 매핑하는 인코더 ϕ 를 훈련
- 벡터들은 일반적인 텍스트 토큰과 교차하여 LLM의 접두사를 형성
 - 접두사의 각 벡터 x_i 가 단어 토큰 γ (\rightarrow 텍스트 토큰인 경우) 또는 인코더 ϕ_i (\rightarrow 관측 공간 내의 관측인 경우) 중 하나에서 형성된다는 것을 의미
- 단일 관측값 O_j 는 일반적으로 여러 임베딩 벡터로 인코딩 됨
 - 접두사의 다른 위치에서 서로 다른 인코더 ϕ_i 를 교차하여 서로 다른 관측 공간에서의 정보를 결합하는 것과 같이 다양한 인코더를 교차할 수 있음
 - 이러한 방식으로 연속적인 정보를 LLM에 주입하면서 기존의 위치 인코딩을 재사용

→ VLM 접근 방식과는 달리 고정 위치에 삽입되는 것이 아닌 주변 텍스트 내에서 동적으로 배치됨

- 결과 **embodying**: 로봇 제어 loop에서의 PaLM-E

- PaLM-E: 멀티모달 문장을 입력으로 사용하여 텍스트를 생성하는 생성 모델

- 모델의 출력을 구체화에 연결하기 위해 두 가지 경우를 구별

- 1. 작업이 텍스트만 출력하여 수행될 수 있는 경우

- ex) 구체화된 질문 응답 또는 장면 설명 작업과 같은 경우

- 모델의 출력은 작업의 직접적인 해결책으로 간주됨

- 2. 구체화된 계획 또는 제어 작업을 수행하는 데 사용될 때

- 텍스트를 생성하여 저수준 명령을 조건으로 함

- 스킬이라고 불리는 저수준 동작들의 시퀀스를 생성

- 스킬은 언어로 조건이 되지만 복잡한 지시를 수행하지는 않음

- PaLM-E는 로봇에 의해 저수준 정책을 통해 예측된 결정이 실행되어 새로운 관측을 생성하며, 필요한 경우 재계획 함

- ⇒ 고수준 정책으로서 낮은 수준의 정책을 조절/제어

4. 다른 센서 모달리티에 대한 입력 & 장면 표현

- PaLM-E에 통합되는 다양한 모달리티와 각각의 인코더 설정을 설명

- 2D 이미지 특징을 위한 Vision Transformers (ViTs), 3D-aware Object Scene Representation Transformer (OSRT)와 같은 다양한 인코더 아키텍처를 제안

- 입력 장면을 전역적으로 나타내는 것 뿐만 아니라 개별 객체로 분해하는 객체 중심 표현도 고려

- 상태 추정 벡터

- 로봇 또는 물체의 상태 추정 등

- 장면 내 물체의 상태를 설명하는 벡터 → 언어 임베딩 공간

- 시각 transformer(ViT)

- 이미지를 토큰 임베딩으로 변환하는 트랜스포머 아키텍처

- 다양한 변형을 고려

- 4억 개와 220억 개 파라미터 모델인 ViT-4B와 ViT-22B가 있음

- 처음부터 끝까지 훈련되는 ViT 토큰 러너 아키텍처도 고려
 - 언어 모델과 차원을 일치시키기 위해 각 임베딩을 적절하게 변환
- 물체 중심 표현
 - 시각적 입력은 의미 있는 개체와 관계로 사전 구조화되어 있지 않음 + ViT는 시각적 정보를 정적 그리드 형태로 표현

⇒ 사전 훈련된 기호 위에 동작하는 LLM과 물리적 객체와 상호 작용이 필요한 구체화된 추론에 어려움을 제기
 - 해결) 시각 입력을 LLM에 주입하기 전에 시각 입력을 개별 객체로 분리하는 구조화된 인코더를 고려

→ 물체 인스턴스 마스크를 사용하여 ViT의 표현을 개별 객체로 분해
- 물체 장면 표현 트랜스포머(OSRT)
 - 지상 실측 분할 정보가 필요하지 않는 대안으로 OSRT (Object Scene Representation Transformer)가 있음
 - OSRT는 아키텍처 내의 귀납적 바이어스를 통해 비지도 학습 방식으로 물체를 발견하며 3D 중심의 신경망 장면 표현을 학습

⇒ 물체 슬롯으로 구성되며, 각 슬롯은 MLP를 사용하여 특정 토큰 임베딩으로 변환
- 개체 참조
 - PaLM-E는 계획 작업에서 생성된 계획에서 개체를 참조해야 함
 - 대부분의 경우에는 물체는 고유한 특성을 통해 자연어로 식별됨
 - 몇몇 상황에서는 객체를 짧은 언어로 식별하기 어려울 수 있음
 - 이런 경우를 위해 OSRT와 같은 객체 중심 표현을 사용하여 입력 프롬프트에 대응하는 다중 모달 토큰을 레이블링

⇒ PaLM-E가 생성된 출력 문장에서 특수 토큰을 사용하여 객체를 참조할 수 있으며, 낮은 수준의 정책도 이러한 토큰을 사용하여 작동할 수 있음

5. 훈련 방법

- PaLM-E는 다음과 같은 데이터 셋에서 훈련된 모델임
 - 각 예제는 연속적인 관측값, 텍스트, 그리고 인덱스로 구성
 - PaLM-E는 디코더 전용 모델이며, 텍스트는 다중 모달 문장의 일부와 텍스트 토큰만으로 구성됨

- 모델은 교차 엔트로피 손실을 사용하여 훈련 ⇒ 해당 손실은 텍스트의 일부인 토큰들에 대한 것
- 모델 내에서 다중 모달 문장을 형성하기 위해 특별한 토큰을 사용
⇒ 해당 위치에서 인코더의 임베딩 벡터로 대체됨
- 사전 훈련된 PaLM 모델의 변형을 사용하며, 연속적인 관측값은 입력 인코더를 통해 모델에 주입됨
- 모델은 다양한 파라미터 크기로 사용됨
⇒ PaLM-E-**12B**, PaLM-E-**84B** 및 PaLM-E-**562B**로 나타냄
- **모델 동결(freezing)을 통한 변형**
 - 대부분의 아키텍처는 인코더, 프로젝터 및 LLM_{pLM} 으로 구성됨
 - PaLM-E를 훈련할 때, 세 부분의 매개변수를 모두 업데이트하는 방법이 있지만, LLM은 적절한 프롬프트를 제공하면 강력한 추론 능력을 보임
⇒ LLM을 동결하고 입력 인코더만 훈련할 수 있는지를 조사
 - 이 경우, 인코더는 동결된 LLM이 관찰에 기반하고, 능력에 대한 정보를 LLM에 전달하는 임베딩 벡터를 생성해야 함
⇒ 이러한 인코딩 훈련은 입력 조건부 소프트 프롬프팅의 한 형태로 볼 수 있음
 - 실험에서는 OSRT와의 상호작용을 고려하여 슬롯 표현을 동결하고 작은 프로젝터만 업데이트하는 방식을 활용하였음
- **업무들 간의 공동 훈련(Co-training)**
 - 모델을 다양한 다양한 데이터로 공동 훈련 시의 효과를 조사
 - 다양한 작업에서 가져온 인터넷 규모의 비전 및 언어 데이터로 구성한 데이터 활용
⇒ 전체 중에서 약 8.9%만이 구체화된 데이터이며, 각 구체화에 대해 여러 작업이 존재

6. 실험 설계

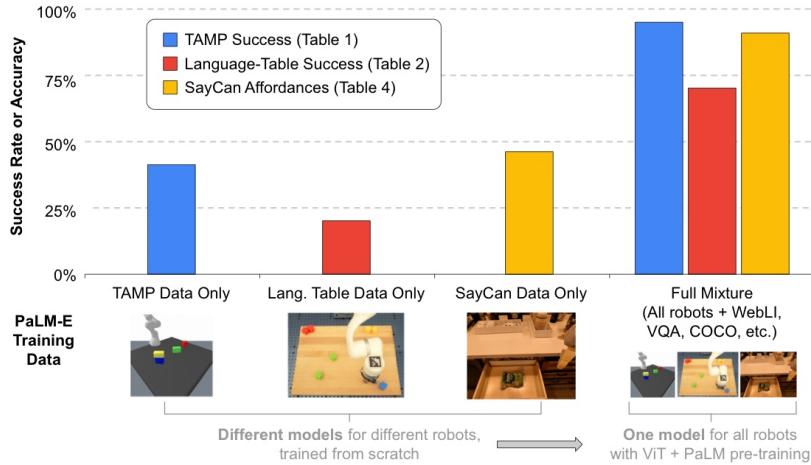


Figure 3: Overview of *transfer learning* demonstrated by PaLM-E: across three different robotics domains, using PaLM and ViT pretraining together with the full mixture of robotics and general visual-language data provides a significant performance increase compared to only training on the respective in-domain data. See Tab. 1, Fig. 4, Tab. 2, Tab. 4 for additional data in each domain.

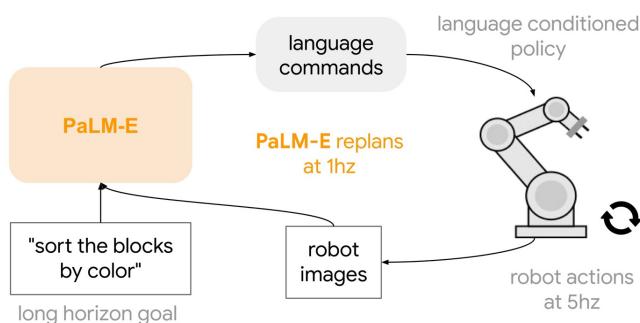
- (이동 가능한) 로봇 조작 작업을 시뮬레이션과 두 가지 다른 실제 로봇에서 다양한 로봇 구현체를 통해 조사
- 시각적 질문 응답, 이미지 캡션 생성 및 언어 모델링과 같은 일반적인 비전-언어 작업에서의 PaLM-E의 성능 또한 평가
- 실험은 크게 **두 가지 범주**로 나뉨
 1. 성능, 일반화 및 데이터 효율성 관련
 - 다양한 입력 표현을 비교
 2. 주요 PaLM-E 버전에 초점
 - 원시 이미지를 처리하는 사전 훈련된 ViT 및 PaLM 언어 모델을 사용하여 다양한 작업, 로봇 구현체 및 데이터셋 혼합물에 대한 단일 모델의 성능을 조사
 - 공동 훈련이 다양한 작업 및 구현체에서의 전송을 가능하게 하는지를 조사
⇒ 공동 훈련 전략과 모델 매개변수 크기에 따른 성능, 일반화 및 데이터 효율성 영향을 연구
 - 마지막으로, LLM을 동결하고 시각 정보를 주입하는 ViT만 훈련하는 것이 가능한지를 실험
- **Baseline**
 - 로봇 구현체 데이터로 훈련되지 않은 최첨단 시각 언어 모델 PaLI

- 오라클 affordances가 제공된 SayCan 알고리즘

6-1. 로봇 환경/업무

- 세 가지 로봇 환경
 1. Task and Motion Planning (TAMP) 도메인
 2. 탁상용 물체 밀기 환경
 3. 모바일 조작 도메인
- 각 도메인에서 PaLM-E는 해당 도메인의 전문 데이터를 기반으로 훈련됨
 - TAMP 작업: 가능한 계획에 대한 복잡한 조합 문제를 포함 → 다단계 계획을 생성해야 함
 - 탁상용 물체 밀기 환경: 다양한 물체, 언어의 다양성 및 복잡한 밀기 동작을 포함
 - 물체의 위치에 대한 추론이 필요한 TAMP 및 Language-Table 환경과 같은 작업을 수행
 - 모바일 조작 도메인: 주방 환경에서 다양한 작업을 수행하며 계획을 실행하고 외부 간섭 또는 저수준 제어 정책의 실패에 대응해야 함
 - 모든 도메인에서 PaLM-E는 계획 및 시각적 질문 응답(VQA) 작업을 수행

6-2. TAMP 환경



- 해당 실험에서 사전 훈련된 LLM (언어 모델)은 동결됨
 - 입력 표현은 TAMP 환경의 훈련 장면만을 사용하여 훈련됨
- 3~5개의 물체가 있는 장면에서 대부분의 입력 표현은 유사한 성능을 보임
 - 물체의 수를 늘릴 때, 사전 훈련된 LLM을 사용하면 특히 entity referrals와 관련하여 성능이 크게 향상되는 것을 확인
 - ▼ **entity referrals**

- 언어 모델이 특정 객체나 개체를 가리키거나 참조하는 것

ex) 이미지 캡션 생성 작업

"빨간색 사과"라는 이미지 캡션에서 "빨간색 사과"가 "사과"라는 개체를 가리키는 entity referral의 한 예

⇒ 언어 모델이 문맥 내에서 특정 객체나 개체를 가리키거나 참조하여 작업을 수행하는 것을 나타내는 용어

- 2B LLM은 8B 변형과 비교하여 분포 밖의 일반화 능력이 더 뛰어나며, 비사전 훈련 LLM은 분포 밖의 일반화 능력이 거의 없음을 확인

- SayCan Baseline

- 오라클 affordance 함수를 사용
- TAMP 환경에서 장기 계획을 구성하기에는 정보가 부족 → 성능 저하

3~5개 물체 훈련 결과

	Object-centric	LLM pre-train	Embodied VQA				Planning	
			q ₁	q ₂	q ₃	q ₄	p ₁	p ₂
SayCan (oracle afford.) (Ahn et al., 2022)	✓	-	-	-	-	-	38.7	33.3
PaLI (zero-shot) (Chen et al., 2022)	✓	-	0.0	0.0	-	-	-	-
PaLM-E (ours) w/ input enc:								
① (State State)	✓(GT)	✗	99.4	89.8	90.3	88.3	45.0	46.1
② (ViT + TL ViT-4B single robot)	✓(GT)	✓	100.0	96.3	95.1	93.1	55.9	49.7
③ (ViT-4B full mixture)	✗	✓	34.7	54.6	74.6	91.6	24.0	14.7
④ (OSRT (no VQA))	✗	✓	-	45.9	78.4	92.2	30.6	32.9
⑤ (OSRT)	✓	✓	-	-	-	-	71.9	75.1
	✓	✓	99.7	98.2	100.0	93.7	82.5	76.2

Table 1: Comparison of different input representations on TAMP environment (in terms of success rates), where data from TAMP constitutes only 1% (i.e., 320 samples for p₁, p₂ each) of total training data size. PaLM-E outperforms both PaLI and SayCan on embodied VQA and planning tasks. Cross-domain transfer is observed, since the PaLM-E with ViT-4B trained on our full data mixture improves planning performance. OSRT, despite using no large-scale data, provides the most effective input encodings for learning. (GT) means ground-truth object-centric information provided. In all experiments, the LLM is frozen. The non-object centric ViT-4B variant utilizes color to reference objects, hence q₁ cannot be evaluated here. The LLM is frozen in these experiments (except for the case where it is not pre-trained). Sec. B.1 describes the tasks q₁-q₄, p₁, q₂.

- 입력 표현 간에 계획 작업에 대한 중요한 차이가 있음을 볼 수 있음
 - 상태 입력: 낮은 데이터 범위에서 LLM의 사전 훈련이 유익
 - 두 ViT 변형 (ViT+TL, ViT-4B):
 - 모두 작은 데이터에 대한 계획 작업 해결 능력이 떨어짐
 - 그러나 다른 모든 로봇 환경과 일반 vision-language 데이터셋을 공동 훈련(⇒ ViT-4B generalist) 시 ViT-4B의 성능이 두 배 이상 향상됨

→ 서로 다른 로봇 구현체와 작업 간의 중요한 전이 효과

3. 입력 표현으로 OSRT를 사용 시 가장 뛰어난 성능을 보였음

⇒ 3D를 고려한 객체 표현의 강점

- TAMP VQA 데이터를 제거하고 640개의 계획 작업 예제만을 훈련하는 경우 성능이 (약간) 감소
 - 로봇 데이터로 훈련되지 않은 최첨단 비전-언어 모델인 PaLI은 이러한 작업을 해결 할 수 x
 - 일반적인 VQA 작업과 유사한 q2 (테이블 위에 물체가 왼쪽/오른쪽/가운데에 있는지) 및 q3 (물체의 수직 관계)를 평가

6-3. 언어 테이블 환경

Zero-shot Baselines				Task 1			Task 2			Task 3					
				0.0			-			-					
				0.0			-			-					
PaLM-E-	trained on	from scratch	LLM+ViT pretrain	LLM frozen	Task finetune	# Demos	10	20	40	10	20	40	10	20	80
12B	Single robot	✓	✗	n/a	✓	20.0	30.0	50.0	2.5	6.3	2.5	11.3	16.9	28.3	
12B	Full mixture	✗	✓	✓	✗	-	-	20.0	-	-	36.3	-	-	29.4	
12B	Full mixture	✗	✓	✗	✗	-	-	80.0	-	-	57.5	-	-	50.0	
12B	Full mixture	✗	✓	✗	✓	70.0	80.0	80.0	31.3	58.8	58.8	57.5	54.4	56.3	
84B	Full mixture	✗	✓	✗	✗	-	-	90.0	-	-	53.8	-	-	64.4	

Table 2: Results on planning tasks in the simulated environment from Lynch et al. (2022).

Task 1. Q: There is a block that is closest to {i.e., top right corner}. Push that block to the other block of the same color.

Task 2. Q: How to sort the blocks by colors into corners?

Task 3. Q: How to push all the blocks that are on the {left/right} side together, without bringing over any of the blocks that are on the {right/left} side?

Table 3: Task prompts for Tab. 2.

Language-Table 환경에서의 장기 과제에 대한 성공률

- **PaLM-E**는 장기 과제와 현재 이미지를 입력으로 받아 저수준 정책을 위한 명령을 출력하는 제어 루프에 통합되어 있음
- 인터넷 규모의 시각 및 언어 학습을 통합적으로 진행한 결과, 특히 과제 당 데모가 10개 만 있는 few-shot (= 초기 학습 데이터가 적은) 상황에서 로봇 계획에 더 효과적인 모델이 나오는 것을 확인할 수 있음
(표랑 말이랑 다른데..?)
- 12B 모델을 84B 모델로 확장하는 것은 3개의 과제 중 2개에서 개선을 가져옴
- TAMP 환경과 마찬가지로 SayCan이나 zero-shot PaLI는 가장 쉬운 테스트 과제를 해결할 수 없어 효과적이지 않음
- **실제 로봇 결과와 few-shot 일반화**

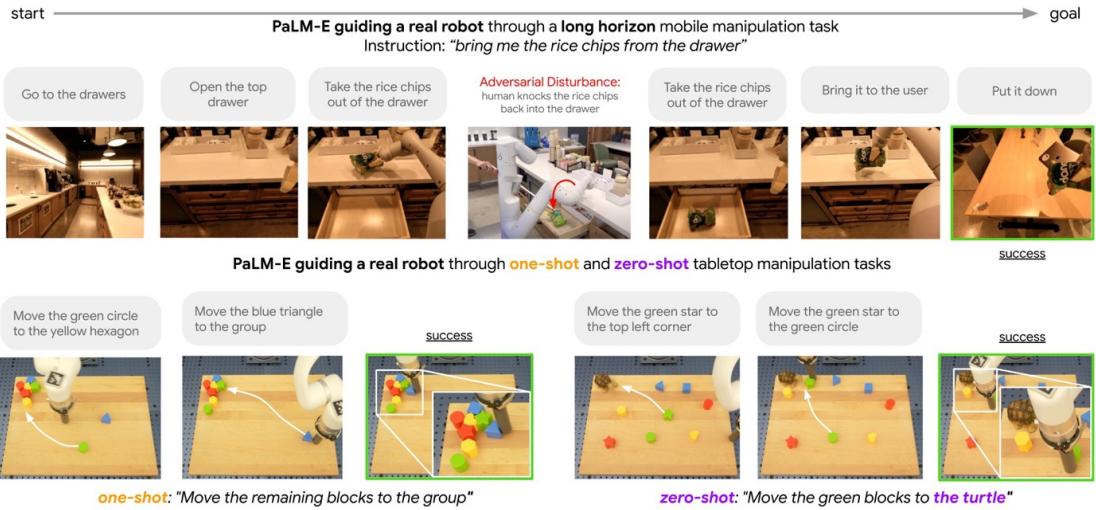


Figure 5: A single PaLM-E model directs the low-level policies of two real robots. Shown is a long-horizon mobile manipulation task in a kitchen, and one-shot / zero-shot generalization with a tabletop manipulation robot.

- PaLM-E는 실제 로봇을 안내하면서 복잡한 테이블 조작 작업을 수행하고, 적대적인 방해에도 견고한 성능을 보임
 - 이미지와 장기 목표를 입력으로 받아 언어 하위 목표를 생성하고 이를 로봇 동작으로 변환하여 사용
- 단일 교육 예제로 다양한 장기 과제를 학습하며, 새로운 객체 조합이나 이전 데이터에서 볼 수 없었던 객체를 포함하는 작업에 대해서도 zero-shot 일반화가 가능

⇒ 로봇 계획 및 학습 분야에서 유망한 모델

6-4. 모바일 조작 환경

- PaLM-E의 성능을 다양하고 어려운 이동식 조작 작업에서 시연
 - 인간의 지시에 따라 내비게이션 및 조작 동작을 계획하는 것을 포함
 - ex) "음료를 쓸았어요, 청소 도구를 가져다 줄래요?"와 같은 명령에 대한 계획 수행

⇒ PaLM-E의 실제 상황 추론 능력을 테스트하기 위한 **3가지 사용 사례** 개발

 1. affordance 예측
 2. 실패 감지
 3. 장기 계획
 - 로봇의 저수준 정책: RGB 이미지와 자연어 명령을 입력으로 받아 end-effector 제어 명령을 출력하는 RT-1 모델을 사용
 - ▼ end-effector 제어 명령

- 로봇이 맨 끝에 위치한 도구 또는 장치를 조작하거나 움직이도록 하는 명령
- 로봇이 특정 작업을 수행하거나 특정 위치로 이동하도록 지시하는 데 사용

Baselines		Failure det.	Affordance
PaLI (Zero-shot) (Chen et al., 2022)		0.73	0.62
CLIP-FT (Xiao et al., 2022)		0.65	-
CLIP-FT-hindsight (Xiao et al., 2022)		0.89	-
QT-OPT (Kalashnikov et al., 2018)		-	0.63
PaLM-E-12B		LLM+ViT frozen	
trained on	from scratch		
Single robot	✓	✗	n/a
Single robot	✗	✓	✓
Full mixture	✗	✓	✓
Full mixture	✗	✓	✗
			0.54
			0.91
			0.78
			0.91
			0.87
			0.77
			0.91

Table 4: Mobile manipulation environment: failure detection and affordance prediction (F1 score).

- 가능성 예측
 - 현재 환경에서 저수준 정책의 기술을 실행할 수 있는지 여부를 예측
 ⇒ VQA(Vision Question Answering) 문제로 정의될 수 있음
 ex) Q: 여기서 <skill>을(를) 실행할 수 있을까요?
 - PaLM-E는 PaLI (제로샷) 및 QT-OPT로 훈련된 가치 함수에 대한 임계값보다 우수한 성능을 보임
- 실패 감지
 - PaLM-E가 PaLI(zero-shot) 및 CLIP의 fine-tuning 버전을 능가
 - hindsight relabeled 데이터로 훈련된 두 개의 CLIP 모델을 활용하는 Xiao 등의 알고리즘을 능가
 ⇒ 데이터셋에서 실패 감지만을 해결하기 위해 특별히 설계됨
- 실제 로봇 결과: 장기 계획
 - PaLM-E는 모바일 조작 작업을 위한 end-to-end 계획에 사용됨
 ⇒ 인간의 명령과 로봇의 이전 단계 이력을 사용하는 프롬프트 구조를 활용
 - PaLM-E는 현재 장면의 이미지 관측과 이전 단계 이력을 고려하여 다음 계획 단계를 생성하도록 훈련됨
 - 각 단계는 저수준 정책에 매핑되며, 이 프로세스는 PaLM-E가 "종료"를 출력할 때까지 자기 회귀적으로 진행됨

- 모델은 2912개의 시퀀스를 기반으로 훈련됨
 - 해당 시퀀스에 대해 모델을 질적으로 평가하고, 모델이 적대적인 방해에도 장기 계획을 성공적으로 수행할 수 있음을 확인

6-5. 일반적인 시각-언어 작업에서의 성능

Model	VQAv2		OK-VQA val	COCO Karpathy test
	test-dev	test-std		
<i>Generalist (one model)</i>				
PaLM-E-12B	76.2	-	55.5	135.0
PaLM-E-562B	80.0	-	66.1	138.7
<i>Task-specific finetuned models</i>				
Flamingo (Alayrac et al., 2022)	82.0	82.1	57.8†	138.1
PaLI (Chen et al., 2022)	84.3	84.3	64.5	149.1
PaLM-E-12B	77.7	77.9	60.1	136.0
PaLM-E-66B	-	-	62.9	-
PaLM-E-84B	80.5	-	63.3	138.0
<i>Generalist (one model), with frozen LLM</i>				
(Tsimplakelli et al., 2021)	48.4	-	-	-
PaLM-E-12B frozen	70.3	-	51.5	128.0

Table 5: Results on general visual-language tasks. For the generalist models, they are the same checkpoint across the different evaluations, while task-specific finetuned models use different finetuned models for the different tasks. COCO uses Karpathy splits. † is 32-shot on OK-VQA (not finetuned).

- 일반적인 PaLM-E-562B 모델 하나가 OK-VQA에서 보고된 결과 중 가장 높은 숫자를 달성
 - 특히 OK-VQA에 특별히 파인튜닝된 모델을 능가
- PaLM-E는 VQA v2에서 가장 높은 성능을 달성 ← LLM을 고정시킨 상태임

⇒ 이것은 PaLM-E가 로봇 작업에서의 embodied 리스너 뿐만 아니라 시각-언어 일반화 작업에서도 경쟁력 있는 모델임

6-6. 일반적인 언어 작업에서의 성능

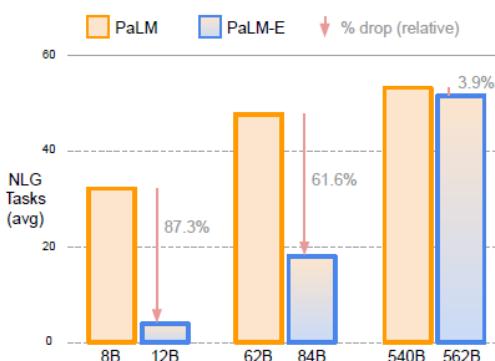


Figure 6: Results on general language tasks (NLG = natural language generation): increasing scale leads to less catastrophic forgetting between a corresponding PaLM-E model and its inherited PaLM model. See full suite of tasks and results in Tab. 8.

- PaLM-E의 평균 성능을 21개의 일반 언어 이해 (NLU) 및 자연어 생성 (NLG) 작업에 대해 보고
- 모델 규모가 증가함에 따라 언어 능력의 치명적인 잊혀짐(catastrophic forgetting) 현상이 상당히 줄어드는 것을 확인할 수 있음
 - 가장 작은 모델인 PaLM-E-12B의 경우, 다중 모달 훈련 중에 NLG 성능의 87.3%가 저하되었지만, 가장 큰 모델인 PaLM-E-562B의 경우, 3.9%만이 저하

7. 실험 요약 & 논의

Generalist vs. Specialist Models - Transfer

- 해당 연구에서 전이의 여러 사례를 보여줌
 - 서로 다른 작업과 데이터셋에서 동시에 훈련된 PaLM-E가 서로 다른 작업에 대해 따로 훈련된 모델에 비해 상당히 향상된 성능을 보임
 - "full mixture"에 대한 공동 훈련은 성능을 두 배 이상 향상시킴
 - LLM/ViT 사전 훈련과 모바일 조작 데이터만 사용하는 대신 전체 혼합물에 대한 훈련을 추가하는 경우 성능이 크게 향상됨
 - Language-Table 실험에서도 유사한 결과를 관찰할 수 있음

데이터 효율성

- 전이 학습 ⇒ 로봇 분야에서 극히 적은 훈련 예제로 로봇 작업을 해결하는 데 PaLM-E를 도움
 - ex)
 - Language-Table: 10에서 80개, TAMP는 320개의 훈련 예제 사이에서 작동
 - OSRT: 기하학적 입력 표현을 사용
 - ⇒ 대규모 시각 데이터를 활용하는 방법과 결합할 수 있음

언어 능력 유지

- 모델의 언어 능력을 유지하는 두 가지 방법이 있음
 1. LLM을 고정하고 입력 인코더만 훈련
 - embodied 언어 모델을 구축하는 데에는 유효한 방법
 - 그러나 때로는 로봇 작업에서 어려움을 겪을 수 있음
 2. 모델 전체를 end-to-end로 훈련
 - 모델은 규모가 커짐에 따라 원래의 언어 성능의 많은 부분을 유지

8. 결론

- 이미지와 같은 다양한 모달 정보를 언어 모델의 임베딩 공간에 통합하여 embodied 언어 모델을 개발
- 실험 결과 최첨단 비전-언어 모델은 엔바디드 추론 작업에는 부족하며, 언어 모델을 affordances를 통해 지지하는 최근 제안도 한계가 있음을 발견 ⇒ **PaLM-E** 제안
- **PaLM-E**
 - 다양한 로봇을 시뮬레이션 및 실제 환경에서 제어할 수 있는 능력 + 일반적인 VQA 및 캡션 작업에 능숙
 - 신경 장면 표현을 모델에 통합 → 대규모 데이터 없이도 효과적으로 작동
 - 여러 로봇 표현 및 일반 비전-언어 작업을 포함한 다양한 작업을 혼합하여 훈련됨
 - 비전-언어 도메인에서 embodied 의사 결정으로의 여러 전이 경로로 이어짐 ⇒ 로봇 계획 작업을 데이터 효율적으로 수행할 수 있음
 - 언어 모델을 고정(→ freezing)시켜도 언어 능력을 유지하면서 일반 목적의 embodied 다중 모달 모델로 나아갈 수 있음을 시사
 - 언어 모델 크기를 확장하면서 치명적인 잊혀짐이 훨씬 적게 발생하는 또 다른 경로도 제시
 - PaLM-E-562B와 같은 가장 큰 모델은 다중 모달 사고 연쇄 추론 및 단일 이미지 프롬프트만을 기반으로 훈련되었음에도 불구하고 여러 이미지에 대한 추론 능력과 같은 신생 능력을 보여줌