EE-559
Deep Learning

What's on today?

- Flamingo: vision and language model example [*subset of slides from last week*]
- Vision-Language-Action models: robots that follow instructions
- Reinforcement learning from human feedback: on value alignment
- Pluralistic alignment: on AI respecting the values of all

Flamingo

VLM for few-shot learning

few-shot prompting

sequences of arbitrarily **interleaved** visual & textual data

output: free-form text

Flamingo

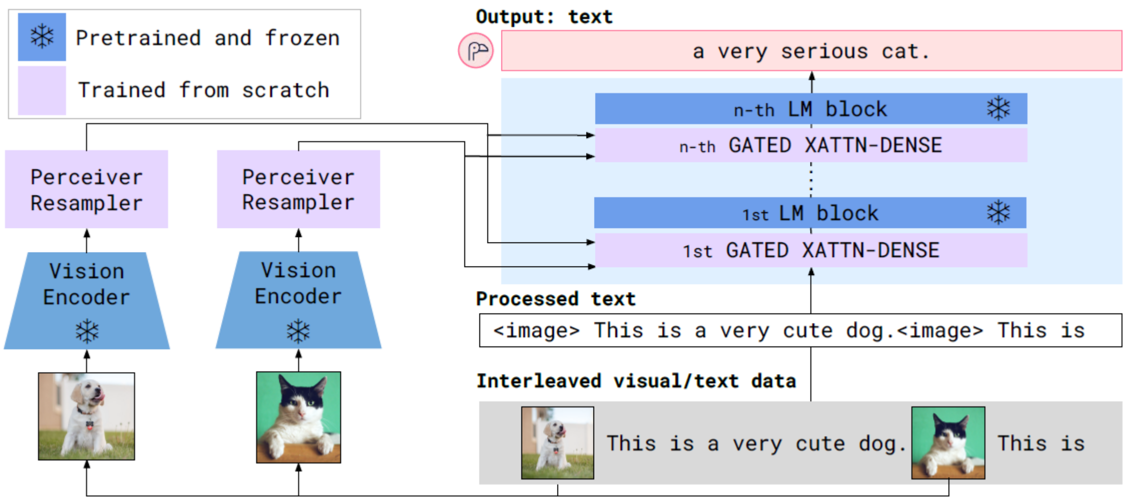
visual tokens used to condition the frozen language model

VLM: Vision and Language Model

arXiv:2204.14198

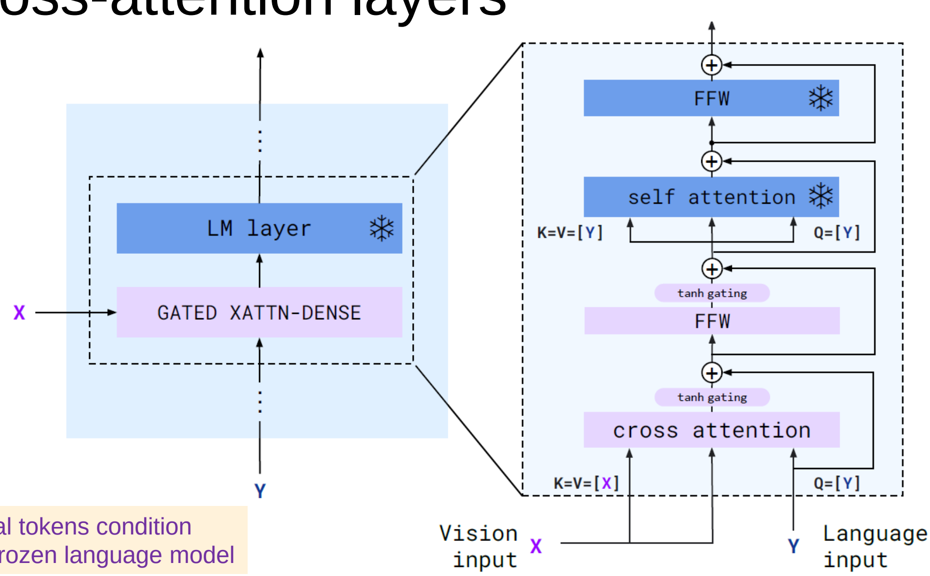
Architecture overview

visual tokens condition
the frozen language model



arXiv:2204.14198

Cross-attention layers



visual tokens condition
the frozen language model

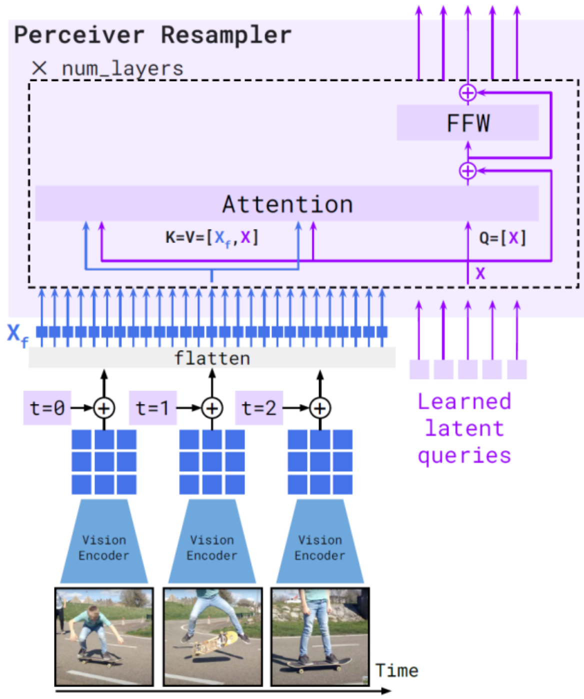
arXiv:2204.14198

number of output tokens
=
number of learnt latent queries

cross-attending to the flattened visual features

learnt temporal position encoding

no explicit spatial grid position encodings

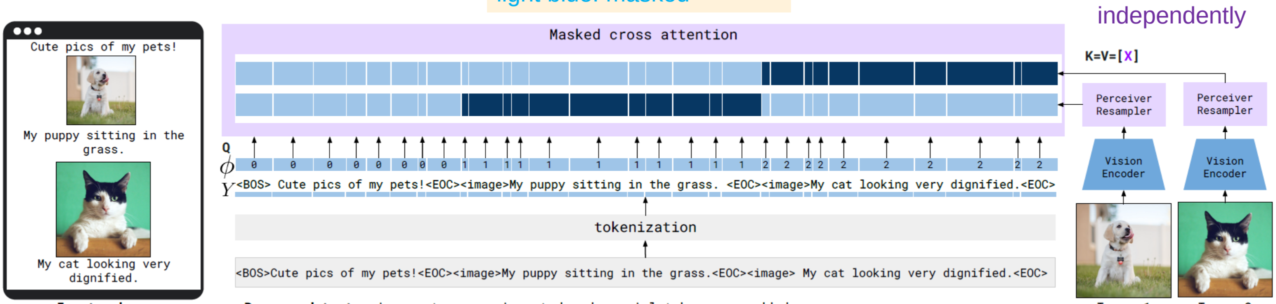


arXiv:2204.14198

Text interleaved with images/videos

selective cross-attention
dark blue: visible
light blue: masked

visual tokens from images processed independently



<image> : tag inserted at the location of the visual data in the text
<BOS> : beginning of sequence
<EOC> : end of chunk

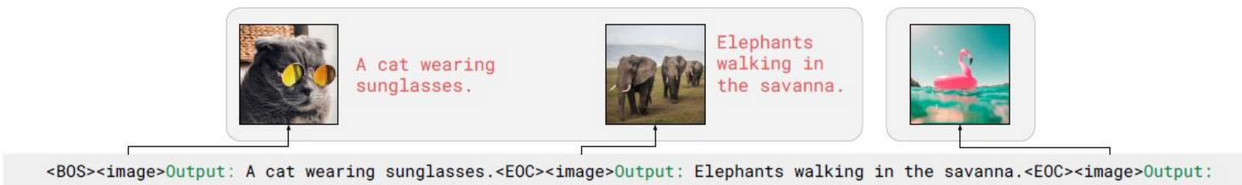
arXiv:2204.14198

Vision-to-text task

input: vision
output: text

task-specific few-shot examples
(support examples)

query



processed prompt

prepending `Output:` to the expected response

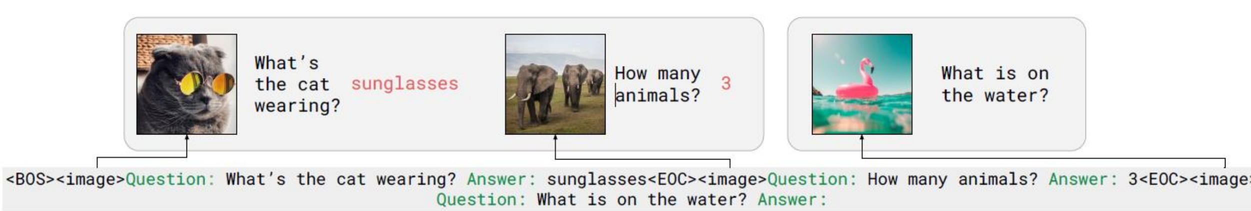
arXiv:2204.14198

Visual question answering task

input: vision + text
output: text

task-specific few-shot examples
(support examples)

query



processed prompt

prompting in the format "`Question:` {question} `Answer:` {answer}"

arXiv:2204.14198

Simple form of interaction

input: single image + text prompt (question or start of a caption)
output: text

Question: What nutrients is the fruit in the image rich in? Answer:

Vitamin C, potassium, and fiber.

Question: What is the purpose of this? Answer:

To keep the mosquitoes out of the bed.

Question: What season is this? Explain why. Answer:

This is winter season. The Christmas tree and the presents are the main reasons why it is winter.

input prompt

completion

arXiv:2204.14198

Dialogue

What do you think the robot is doing here?

The robot is picking up a ball.

What color is the ball?

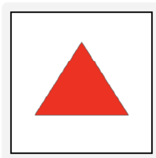
It's red.

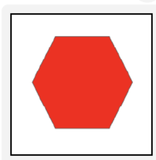
I want the ball to be hidden. Is there something the robot can use for that?

The robot can use a cloth to hide the ball.

What color is the cloth?

The cloth is blue.





What is the difference between these two images?

The first one is a triangle and the second one is a hexagon.

Which one has larger internal angles?

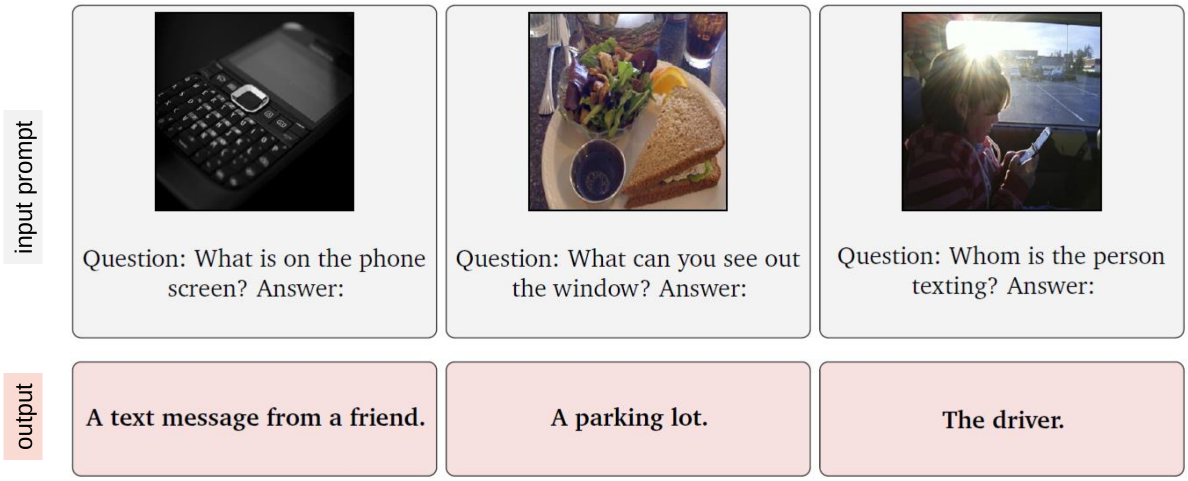
The hexagon has larger internal angles.

What color are they?

They are both red.

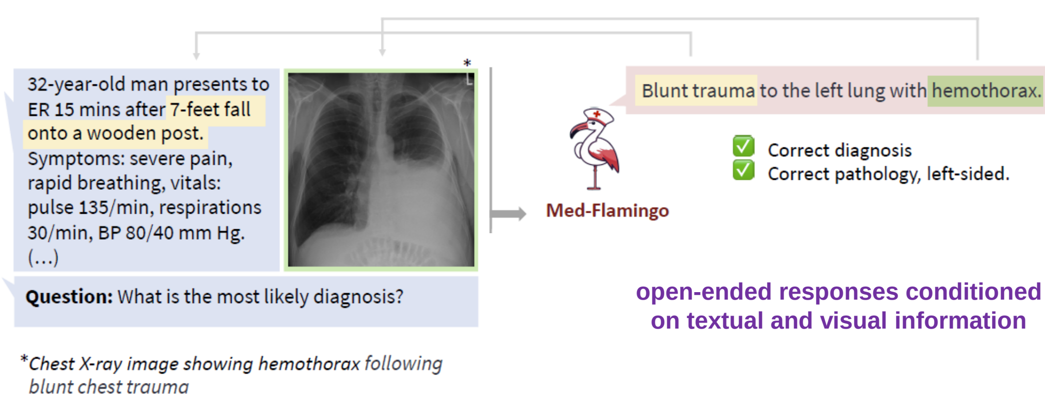
arXiv:2204.14198

Hallucinations



arXiv:2204.14198

Medical generative vision-language model

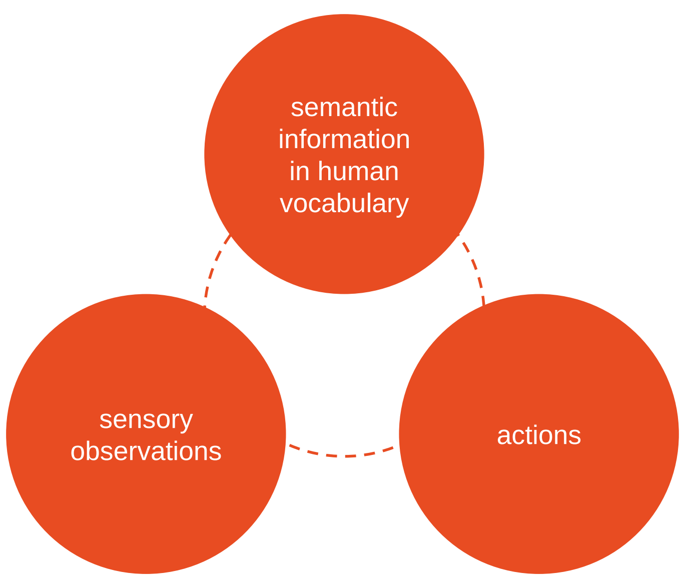


open-ended responses conditioned on textual and visual information

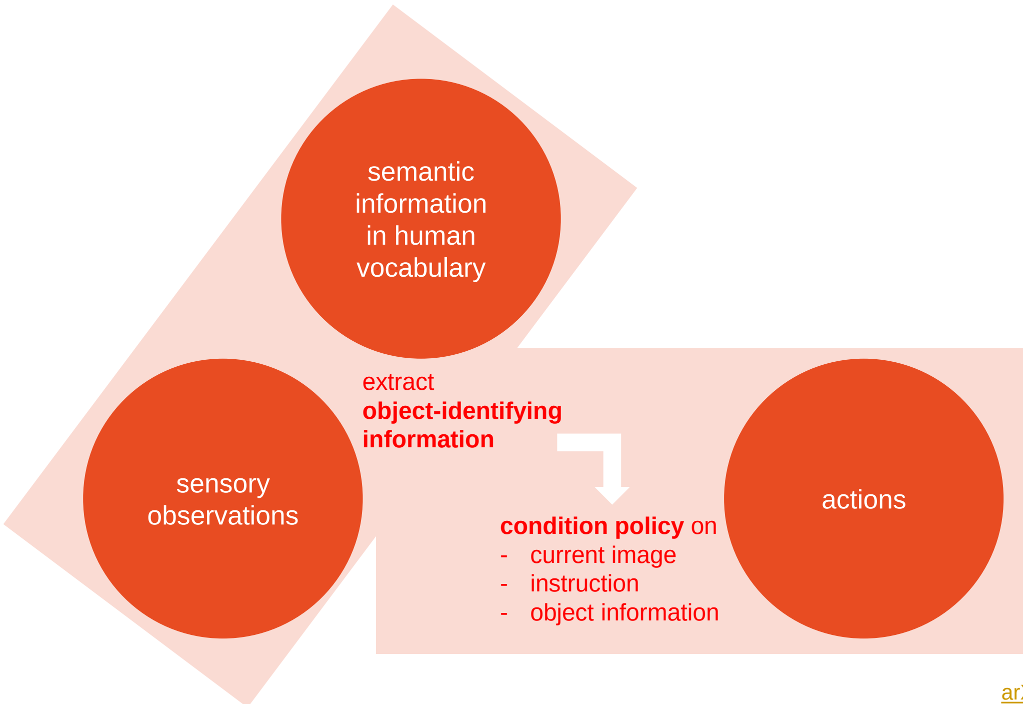
in-context learning

arXiv:2307.15189

Vision-Language-Action models



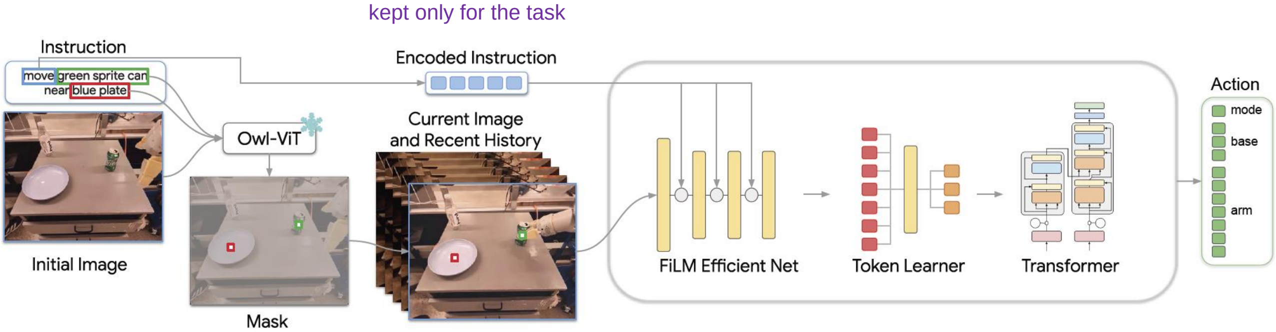
arXiv:2303.00905



arXiv:2303.00905

Policy learning and (separate) VL models

language embedding
kept only for the task



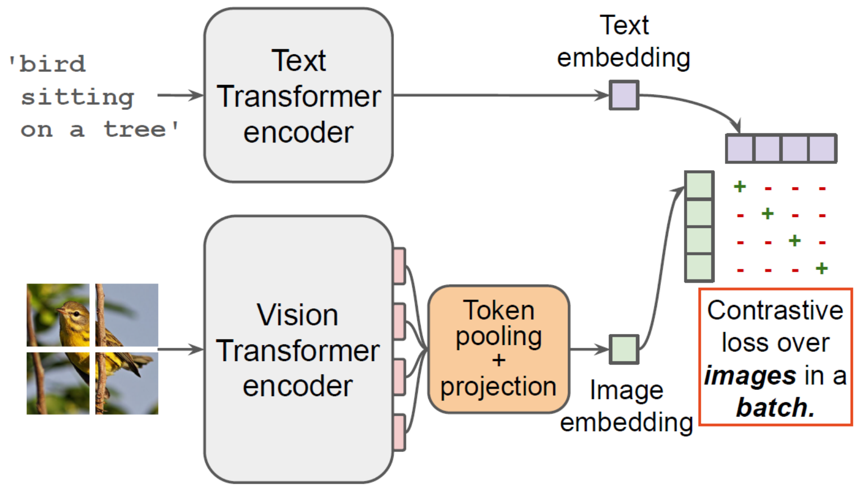
open-vocabulary object detection with vision transformers

object location (center of the bounding box) concatenated channel-wise to the input image

FiLM layers for visual reasoning

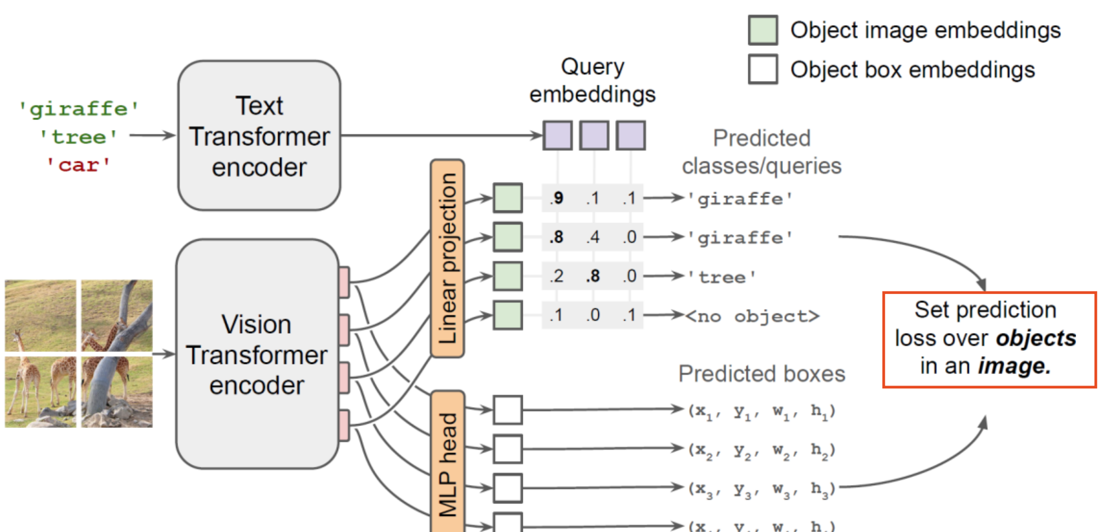
arXiv:2205.06230 arXiv:1709.07871 arXiv:2303.00905

Contrastive image-text pre-training



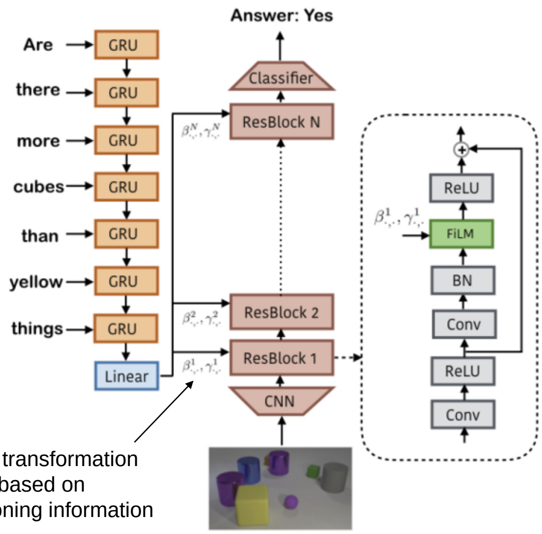
arXiv:2205.06230

Transfer to open-vocabulary detection



arXiv:2205.06230

FiLM: Feature-wise Linear Modulation



affine transformation
based on
conditioning information

GRU: Gated Recurrent Unit

CNN: Convolutional Neural Network

BN: Batch Normalization

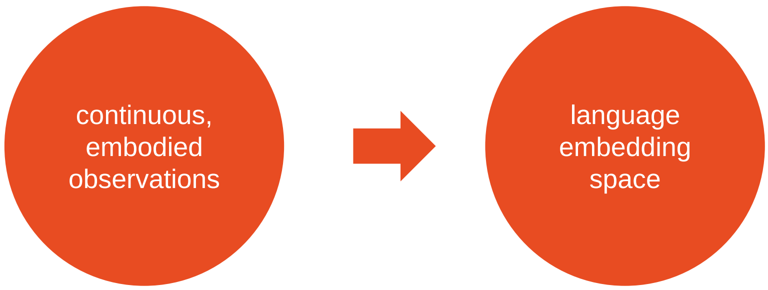
ReLU: Rectified Linear Unit

ResBlock: Residual Block (skip-connection)

arXiv:1709.07871

PaLM-E

Embodied multimodal language model



images
state estimates
other sensor modalities

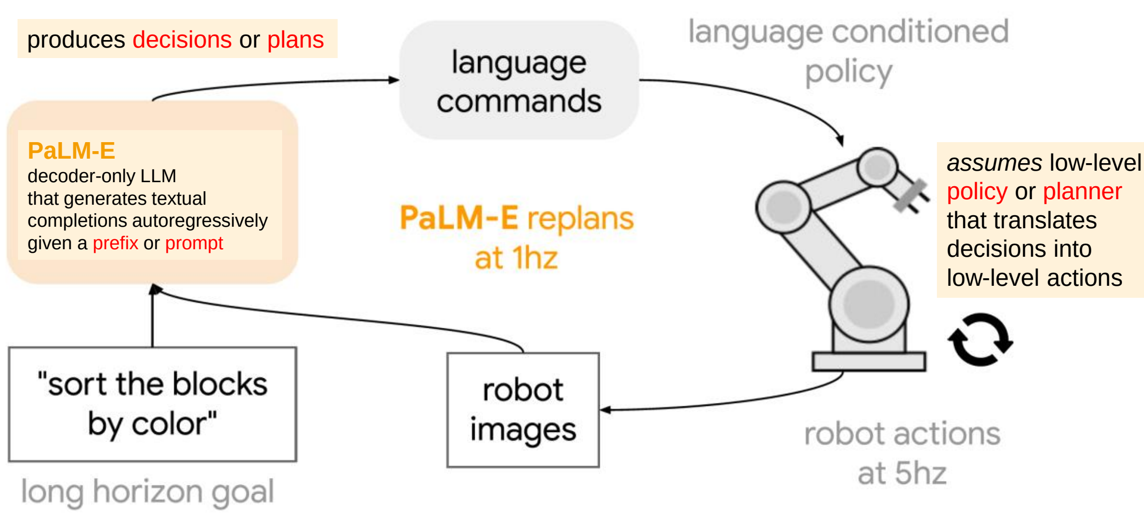
pre-trained
language model

encoding **observations** into sequence of vectors with same dimension as embedding space of **language** tokens

trained on **mixture** of diverse tasks across *multiple embodiments* & general **vision-language** tasks

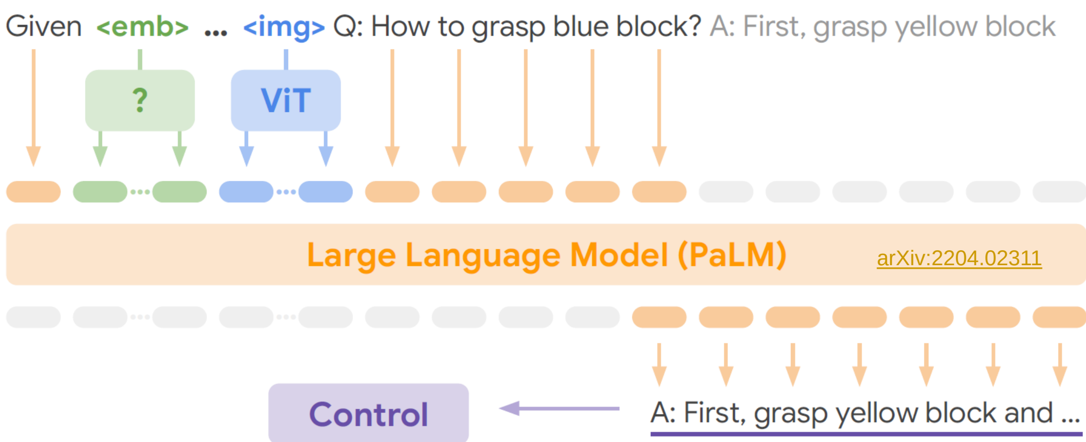
arXiv:2303.03378

Embodied multimodal language model



arXiv:2303.03378

Embodied multimodal language model



PaLM: Pathways Language Model

arXiv:2303.03378

Co-fine-tuning

Closed loop control

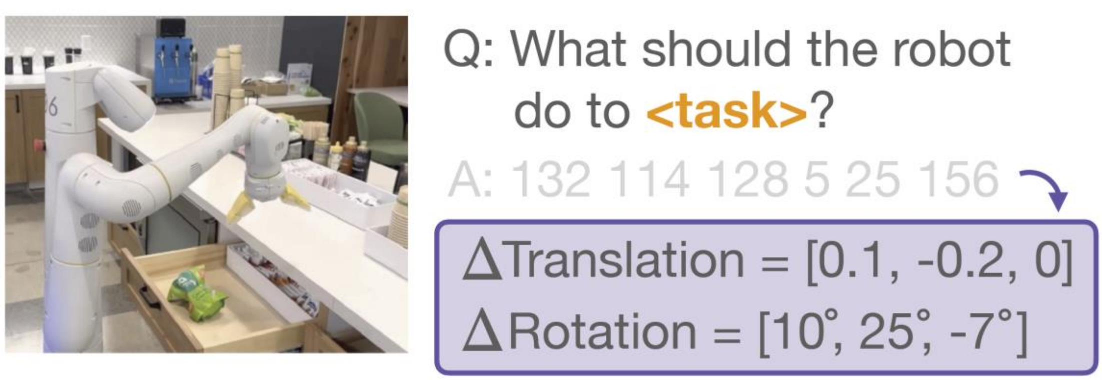
represent
robot actions
as text tokens
(language)

text tokens
de-tokenized
into **robot**
actions

fitting both
natural language responses & robotic actions
into the **same format**

arXiv:2307.15818

Co-fine-tuning



co-fine-tune vision-language models on *robotic trajectory data* and Internet-scale *vision-language tasks*

arXiv:2307.15818

Co-fine-tuning

Q: What is happening
in the image?

A: 311 423 170 55 244

A grey donkey walks
down the street.

co-fine-tune vision-language models on *robotic trajectory data* and Internet-scale *vision-language tasks*

arXiv:2307.15818

Co-fine-tuning

Q: Que puis-je faire avec
ces objets?

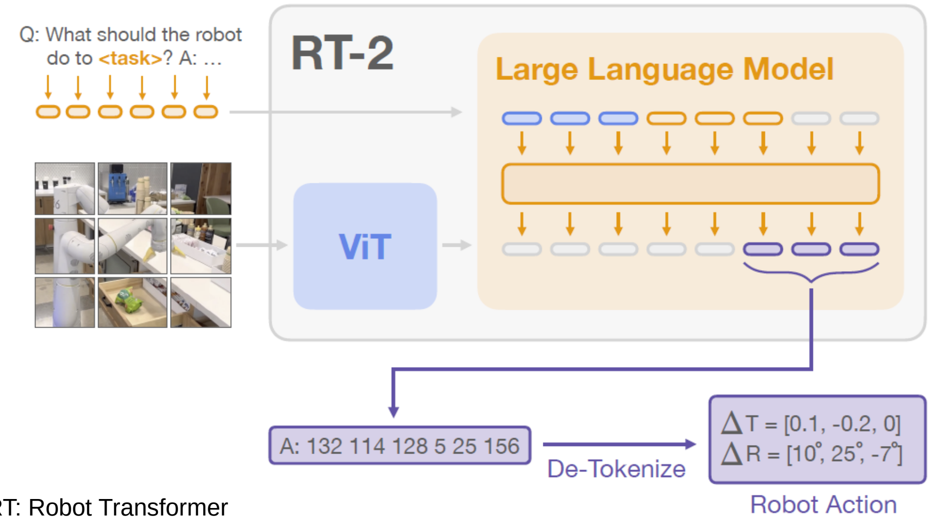
A: 3455 1144 189 25673

Faire cuire un gâteau.

co-fine-tune vision-language models on *robotic trajectory data* and Internet-scale *vision-language tasks*

arXiv:2307.15818

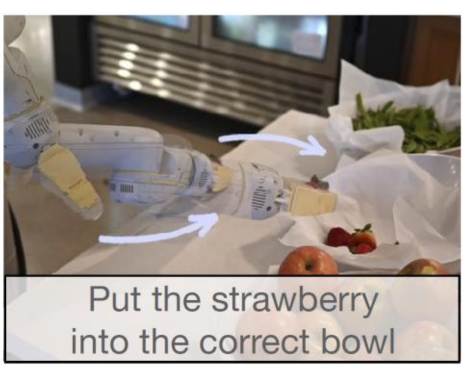
VLA model for robot control (**end-to-end**)

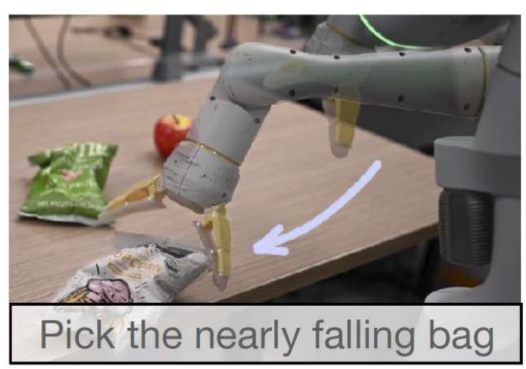


RT: Robot Transformer

arXiv:2307.15818

Closed loop actions





arXiv:2307.15818

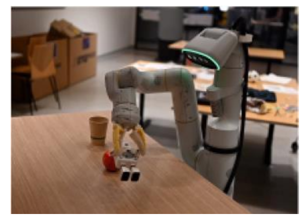
Examples

move coke can to Taylor Swift

move banana to the sum of two plus one

pick animal with different color

pick robot



arXiv:2307.15818

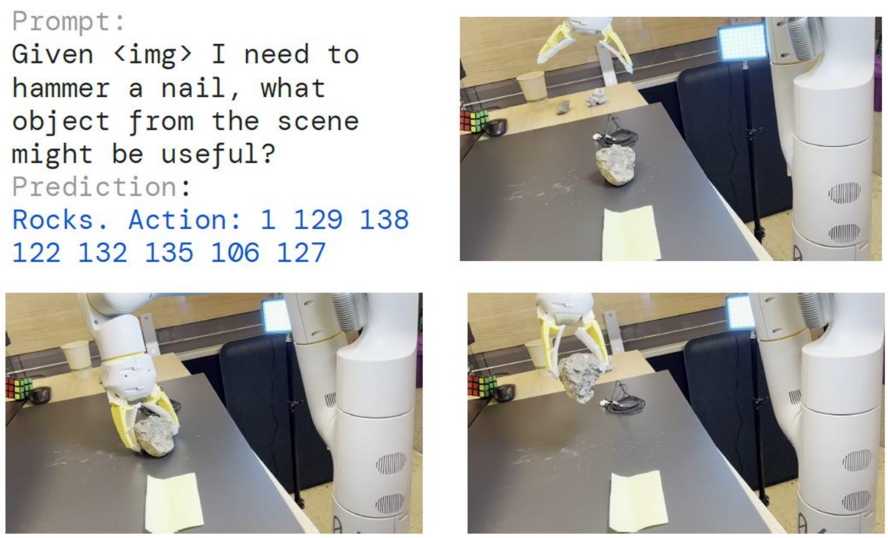
Chain-of-thought reasoning

```
Prompt:
Given <img> Instruction:
Move all the objects
together.
Prediction:
Plan: move green can near
green rice chip bag.
Action: 1 128 126 127 135
123 119 127
```

```
Prompt:
Given <img> Instruction:
Move the green objects
together.
Prediction:
Plan: move green can near
green rice chip bag.
Action: 1 130 129 121 131
127 128 127
```

arXiv:2307.15818

Chain-of-thought reasoning



arXiv:2307.15818

Reinforcement learning from human feedback

RLHF (**preference data**)

supervised fine-tuning

reward learning

RL fine-tuning

learns a **reward function** expected to represent the preference feedback of the *human population*

arXiv:2402.08925

Fine tuning with RLHF

Align the actions of pretrained LLM models with human and societal **values**

learns from humans' **expressed** preferences over multiple outputs

insufficient in representing **diverse** human preferences

arXiv:2404.10271

Supervised fine-tuning

p_θ language model fine-tuned with supervised learning for downstream tasks of interest (e.g. *dialogue*, *instruction following*, *summarization*)

$x_i \in V$
set of tokens (vocabulary set)

$\boldsymbol{x} = \{x_1, x_2, \dots, x_N\}$
prompt (sequence of tokens)

$\boldsymbol{x} \in X$
set of prompts

$\boldsymbol{y} \sim p_\theta(.\,|\boldsymbol{x})$ output response

arXiv:2404.10271

Preference data collection

$\boldsymbol{y} \sim p_\theta(.\,|\boldsymbol{x})$

↓

generate a dataset of model outputs

$\{\boldsymbol{y}_1, \boldsymbol{y}_2\} \sim p_\theta(.\,|\boldsymbol{x})$

pairs of responses

↓

humans evaluate paired completions & select which output they prefer, e.g.

$\boldsymbol{y}_1 \gnsim \boldsymbol{y}_2$

(*preferred* $\boldsymbol{y}_1$ and *dis-preferred* $\boldsymbol{y}_2$ response)

arXiv:2404.10271

Reward model training

$r_\phi(\boldsymbol{y}, \boldsymbol{x})$ reward model

$r_\phi : Y \to \mathbb{R}$

$$P^*(\boldsymbol{y}_1 \gtrneqq \boldsymbol{y}_2 | \, \boldsymbol{x}) = \frac{e^{r^*(\boldsymbol{y}_1, \boldsymbol{x})}}{e^{r^*(\boldsymbol{y}_1, \boldsymbol{x})} + e^{r^*(\boldsymbol{y}_2, \boldsymbol{x})}}$$ Bradely-Terry preference model

$$D = \{\boldsymbol{x}^i, \boldsymbol{y}_1^i, \boldsymbol{y}_2^i\}_{i=1\ldots N}$$ static dataset sampled from P^*

logistic function

$$L(r_\phi, D) = -\mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2) \sim D}[\log \sigma(r_\phi(\boldsymbol{y}_1, \boldsymbol{x}) - r_\phi(\boldsymbol{y}_2, \boldsymbol{x}))]$$

r_ϕ learned with **max likelihood estimation**
(to match the likelihood of the human preferences observed from the data)

arXiv:2404.10271

Reinforcement learning fine tuning

$p^*_{r_\phi}$ optimal policy under reward r_ϕ

KL-regularized reward maximization

$$\max_p E_{\boldsymbol{x} \sim P, \boldsymbol{y} \sim p_\theta(.|\boldsymbol{x})}[r_\phi(\boldsymbol{y}, \boldsymbol{x}) - \beta D_{KL}[p(.|\boldsymbol{x}) || p_{\text{REF}}(.|\boldsymbol{x})]]$$

$\beta > 0$
controls the deviation from the
base reference policy p_{REF}

arXiv:2404.10271

Pluralistic alignment

Key factors contributing to diversity

- socio-demographic backgrounds
- personal bias & context subjectivity
- imperfect preferences
- linguistic ambiguity & missing context

arXiv:2402.08925

Reflecting and supporting diversity

improves deliberation

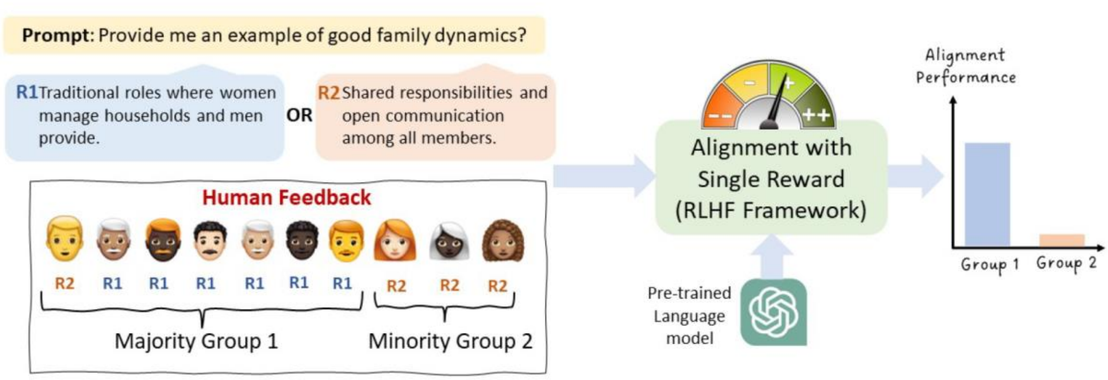
pluralism as a value itself

increases fairness

algorithmic monocultures
lead to increased unfairness
when applied by many decision makers

arXiv:2402.05070

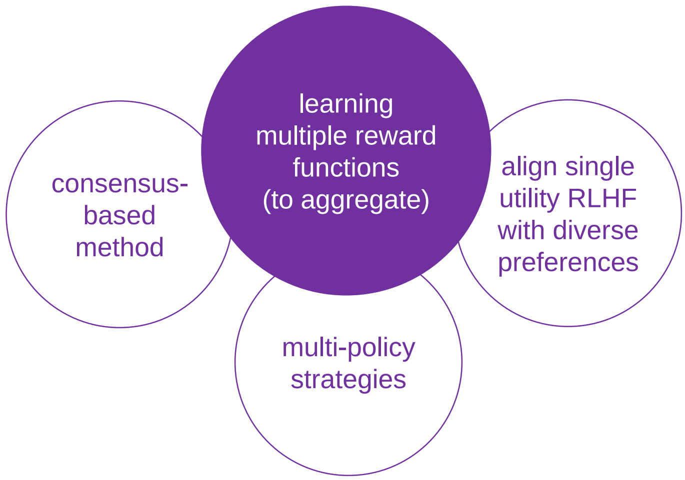
Majority vs minority user groups



most RLHF approaches ignore diversity in human preference feedback
by aligning the language model with a **single reward function**

arXiv:2402.08925

Diversity in opinions and preferences



arXiv:2402.08925

Mixture of preference distributions

$$P_u^*(\boldsymbol{y}_1 \gtrsim \boldsymbol{y}_2 |\, \boldsymbol{x}) = \mathbb{E}_{h \in H_u}[I(h \text{ prefers } \boldsymbol{y}_1 \text{over } \boldsymbol{y}_2 |\, \boldsymbol{x})]$$ for all groups in U

$$U = \{\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_{|U|}\} \qquad H = \bigcup_{u=1}^{|U|} H_u \qquad u$$ human subpopulation index

$\boldsymbol{z} = (\boldsymbol{y}_1 \gtrsim \boldsymbol{y}_2 |\, \boldsymbol{x})$

marginal probability distribution of subpopulation H_u

$$P^*(\boldsymbol{y}_1 \gtrsim \boldsymbol{y}_2 |\, \boldsymbol{x}) = \sum_{u=1}^{|U|} \left[\sum_{h \in H_u} I_h(\boldsymbol{z}) q(h|u) \right] \eta(u) = \sum_{u=1}^{|U|} p_u^*(\boldsymbol{z})\, \eta(u)$$

distribution over the humans H

subpopulation with specific preference distribution

arXiv:2402.08925

Mixture of preference distributions

$$p(\boldsymbol{z}') = \sum_{u=1}^{|U|} p^*_{\phi_u}(\boldsymbol{z}')\,\eta(u)$$ preference distribution

$$\boldsymbol{z}' = (\boldsymbol{y}_w \gtrnsim \boldsymbol{y}_l \,|\, \boldsymbol{x})$$

$\boldsymbol{y}_w$ chosen response by the human sub-population group H_u

$\boldsymbol{y}_l$ rejected response by the human sub-population group H_u

$$L(\phi) = \sum_{\boldsymbol{z}' \in D} \log \sum_{u=1}^{|U|} p_{\phi_u}(\boldsymbol{z}')\,\eta(u)$$

$$= \sum_{\boldsymbol{z}' \in D} \log \sum_{u=1}^{|U|} \frac{e^{r_{\phi_u}(\boldsymbol{y}_w, \boldsymbol{x})}}{e^{r_{\phi_u}(\boldsymbol{y}_w, \boldsymbol{x})} + e^{r_{\phi_u}(\boldsymbol{y}_l, \boldsymbol{x})}}\,\eta(u)$$ maximization of the log likelihood

arXiv:2402.08925

Maximizing the minimum utility

Alignment objective (with diverse human preferences)

$$\underset{p}{\operatorname{argmax}} \left(\min_{u} \mathbb{E}_{\boldsymbol{x}\sim P, \boldsymbol{y}\sim p(.|\boldsymbol{x})}[r_{\phi_u^*}(\boldsymbol{y}, \boldsymbol{x})] \right) - \beta D_{KL}[p(.|\boldsymbol{x}) || p_{\text{REF}}(.|\boldsymbol{x})]$$

ϕ_u^* reward model parameter
for each human subpopulation in *U*

arXiv:2402.08925

The implementation

Algorithm 1 MaxMin RLHF

Input: Preference dataset $\mathcal{D}$, initial reward parametrization for each subpopulation u as $r^u_{\phi_0}$, initial policy parameter π_0.
Reward Learning with EM: Utilize Algorithm 2 for learning rewards with EM to learn r^u_ϕ for all user subpopulation u
Max-Min Policy Iteration:
for $t = 0$ to $T - 1$ **do**
  Choosing Minimum Utility Subpopulation:
  $u_{\text{min}} \leftarrow \arg\min_{\mathcal{H}_u \in \mathcal{U}} F_{r^u_\phi}(\pi_t)$
  Perform the PPO Update:
  Update policy π towards maximizing the objective:
  $\pi_{i+1} \leftarrow$ PPO-update$(F_{r_{\phi^*_u}}(\pi_t) - \beta \mathbb{D}_{\text{KL}}[\pi_t || \pi_{\text{ref}}])$
end for
Output: Policy π_T aligned with socially fair preference dataset

Algorithm 2 Learning Rewards with EM Algorithm

Input: Preference data $\mathcal{D}$, $|\mathcal{U}|$ clusters of users among all humans in $\mathcal{H} = \bigcup_{u=1}^{|\mathcal{U}|} \mathcal{H}_u$, pretrained $\{r_{\phi_u}\}_{u=1}^{|\mathcal{U}|}$, loss function loss, convergence criteria
while not reach the convergence criteria **do**
  for $h \in \mathcal{H}$ **do**
    E-step (hard cluster assignment): assign h to the u-th cluster s.t.

$$u = arg \max_{u \in 1, \cdots, |\mathcal{U}|} \prod_{(\mathbf{x}, \mathbf{y_1}, \mathbf{y_2}, h) \in \mathcal{D}} w(\phi_u, \mathbf{x}, \mathbf{y_1}, \mathbf{y_2})$$

    where $w(\cdot) = \frac{\exp(r_{\phi_u}(\mathbf{y_1}, \mathbf{x}))}{\exp(r_{\phi_u}(\mathbf{y_1}, \mathbf{x})) + \exp(r_{\phi_u}(\mathbf{y_2}, \mathbf{x}))}$
  end for
  M-step: Update each $\phi_u, u = 1, \cdots, |\mathcal{U}|$ by minimizing the negative log-likelihood loss (2) on the assigned users' data
end while

arXiv:2402.08925

What did we learn today?

- Flamingo
- Vision-Language-Action models
- Reinforcement learning from human feedback
- Pluralistic alignment

EE-559
Deep Learning

andrea.cavallaro@epfl.ch