

# DeepRob

[Student] Lecture 19

by Naga Hemachand Chinta, Sai Tarun Inaganti, Shashank Sharma

Visual Pretraining and Robot Manipulation

University of Michigan and University of Minnesota



# Contents

---

- Pre-Training in NLP:
  - BERT
  - GPT
  - ChatGPT
- Pre-Training in CV:
  - MAE
  - CLIP
  - DALL-E
- Pre-Training in Robotics:
  - R3M
  - MVP
  - SORNet
  - DALL-E Bot

# Pretraining???



Image Source:

<https://dreme.stanford.edu/news/expand-mathematical-thinking-during-block-and-pretend-play>



Image Source: <https://lovevery.eu/community/blog/child-development/when-should-my-child-be-able-to-stack-6-building-blocks/>

# Foundation Models???

- Models trained on broad data.
- Using self-supervision
- Can be adapted to a wide range of downstream tasks.
- Eg: BERT, ChatGPT, GPT-3, DALL-E



Image Source :

<https://hai.stanford.edu/news/reflections-foundation-models#:~:text=We%20define%20foundation%20models%20as%20wide%20orange%20of%20downstream%20tasks.>



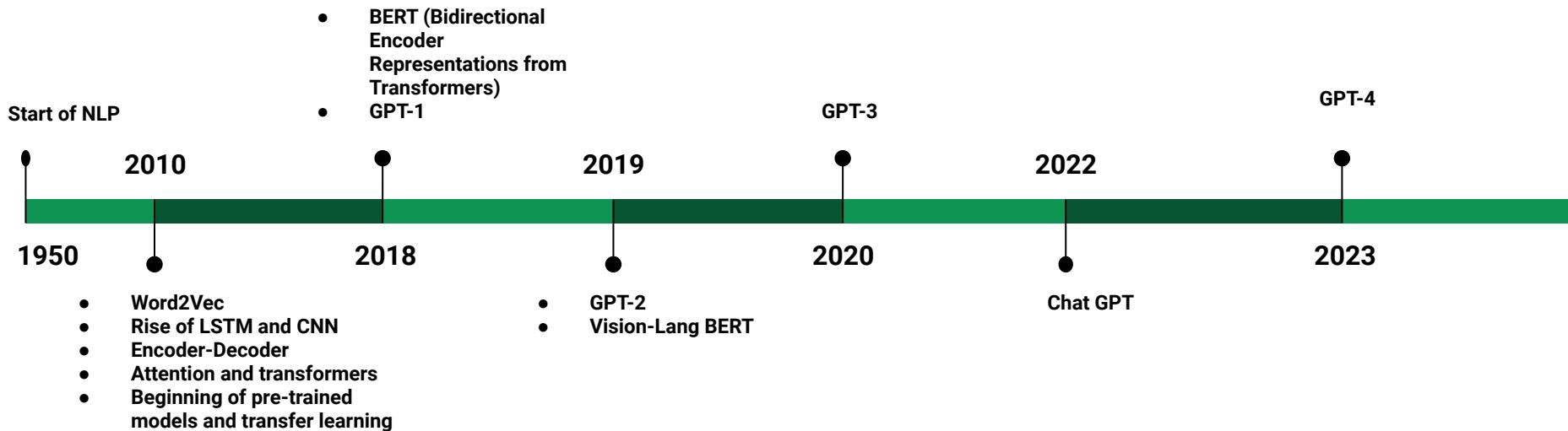
Pre-Training

Foundation Model

Fine Line

# Examples of pre-training in NLP

General timeline:





# BERT

BERT: Bidirectional Encoder Representations from Transformers.



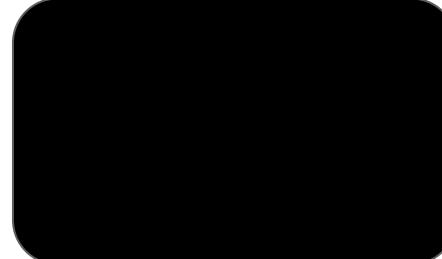
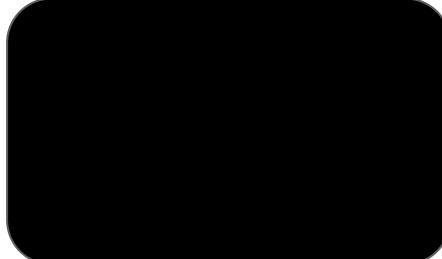
Gif Sources: <https://medium.com/mlearning-ai/getting-contextualized-word-embeddings-with-bert-20798d8b43a4>  
<https://giphy.com/explore/google-bert>

# BERT

---

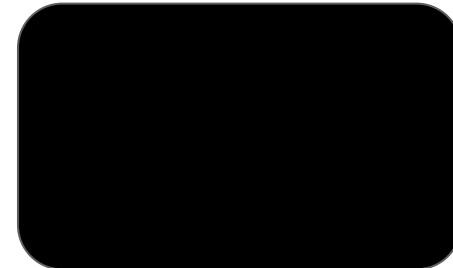
Architecture

Pretraining



# BERT

---



# BERT

---

Architecture

Pretraining

Fine Tuning

# BERT

---

Architecture

Pretraining

Fine Tuning

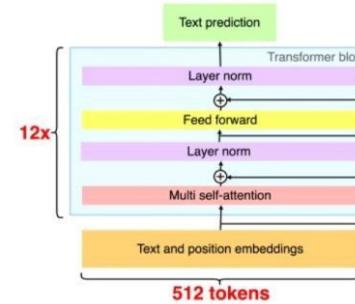
Applications

# Examples of pre-training in NLP

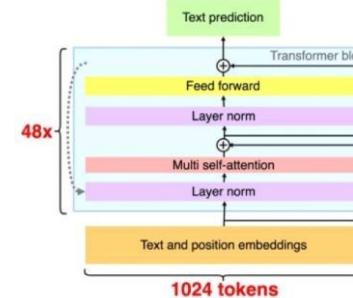
## GPT: Generative Pre-training



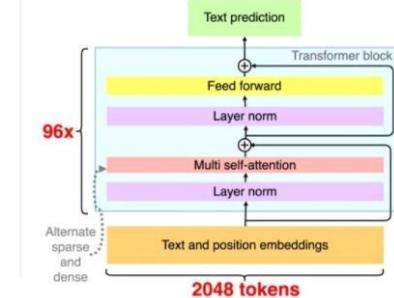
**GPT-1**



**GPT-2**



**GPT-3**



# Examples of pre-training in NLP

GPT 1 vs GPT 2 vs GPT 3 vs GPT 4:

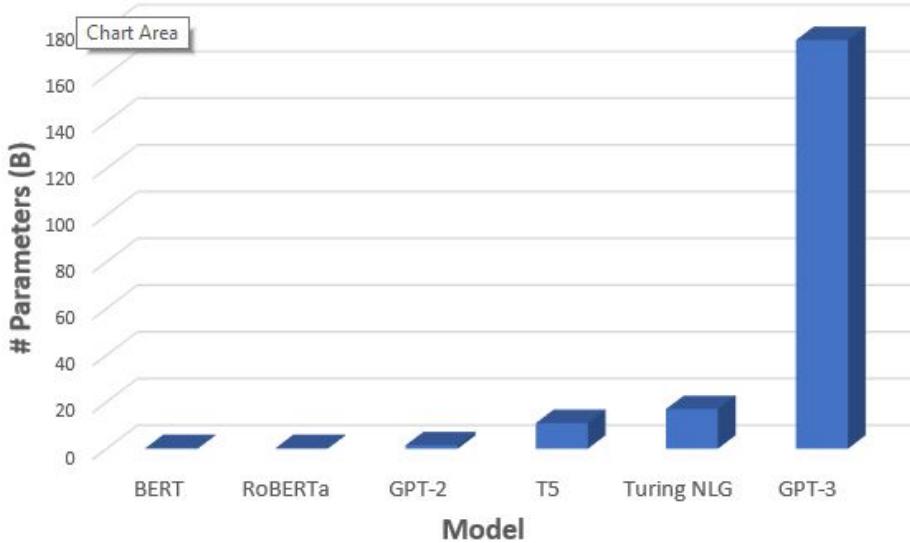


Image Source: <https://towardsdatascience.com/gpt-3-the-new-mighty-language-model-from-openai-a74ff35346fc>

# Examples of pre-training in NLP

GPT 1 vs GPT 2 vs GPT 3 vs GPT 4:

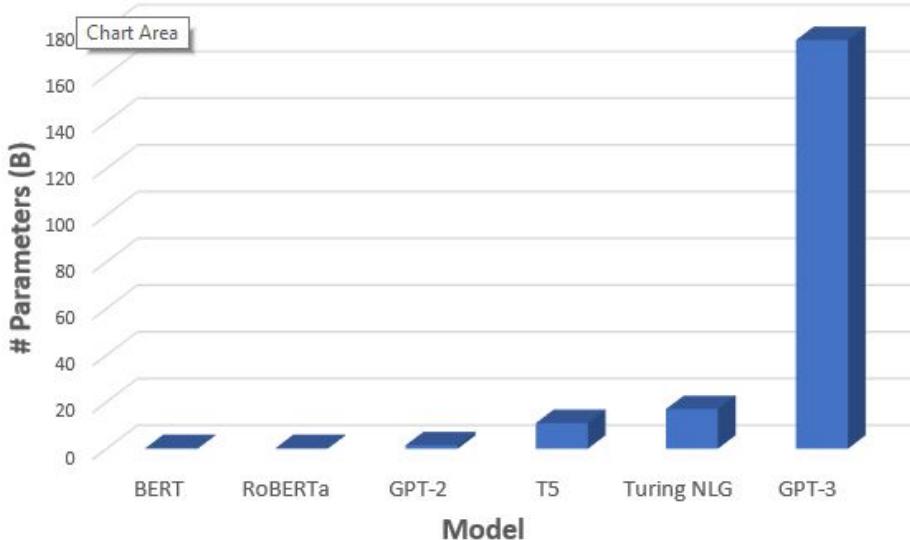
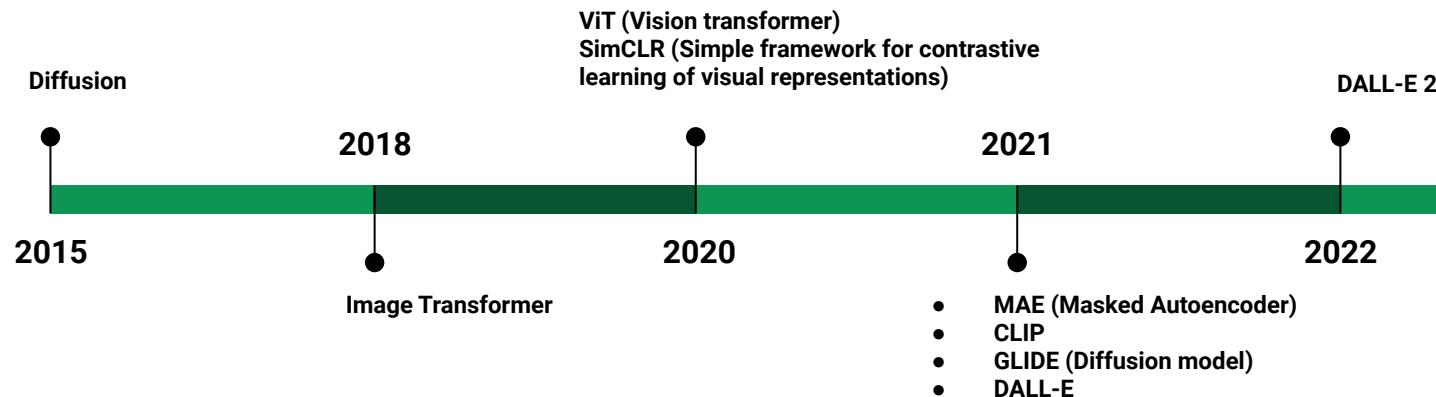


Image Source: <https://towardsdatascience.com/gpt-3-the-new-mighty-language-model-from-openai-a74ff35346fc>  
<https://ai/plainenglish.io/embracing-language-model-evolution-gpt-2-gpt-3-and-gpt-4-in-the-ai-landscape-e3e340dc5693>

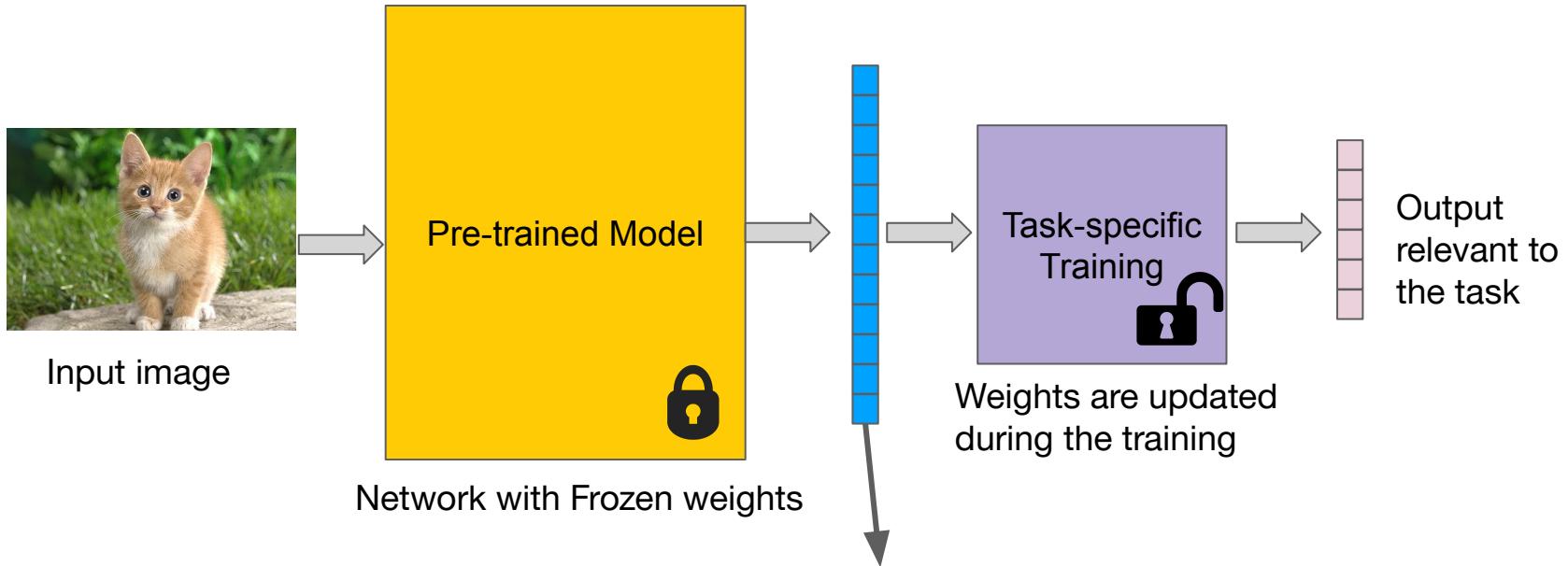
# Examples of pre-training in CV

General Timeline:



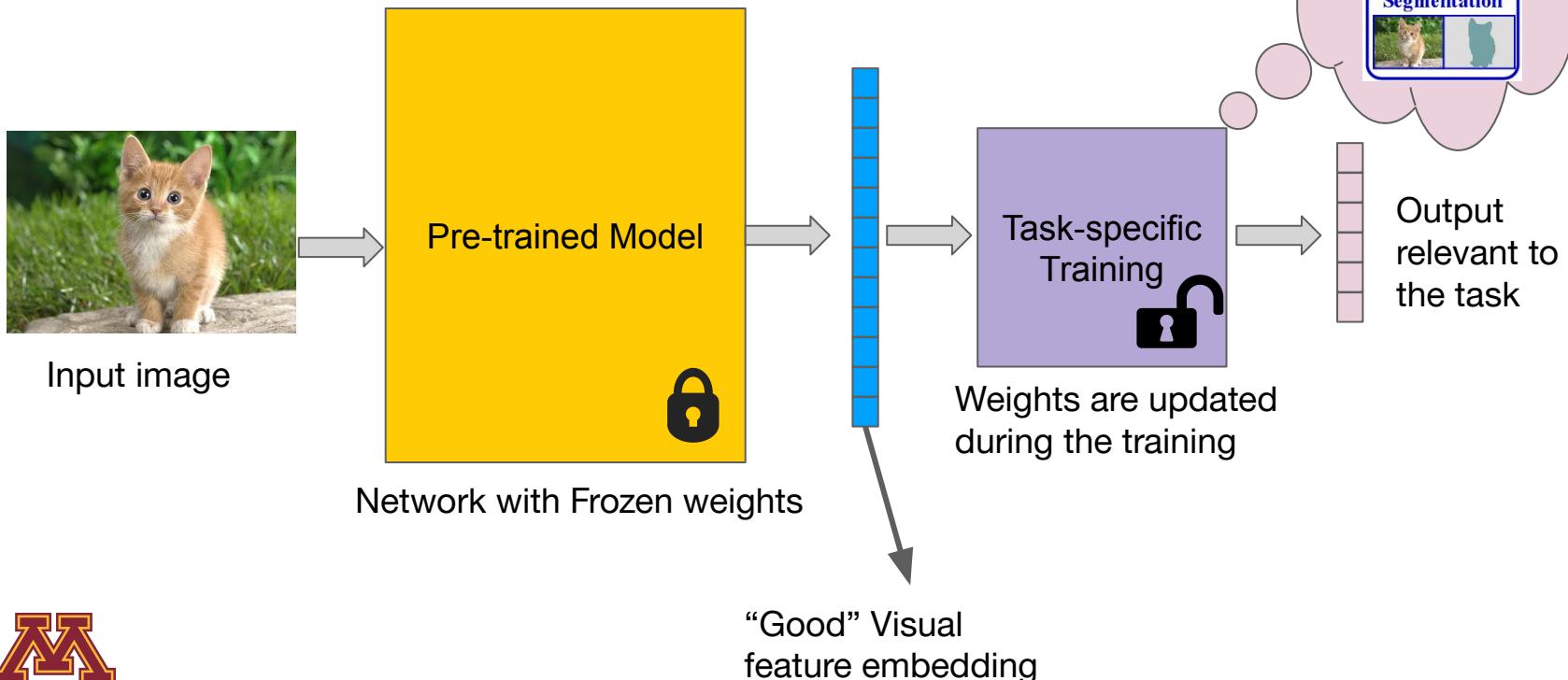
# Pre-Training in CV???

Task-specific training using pre-trained model



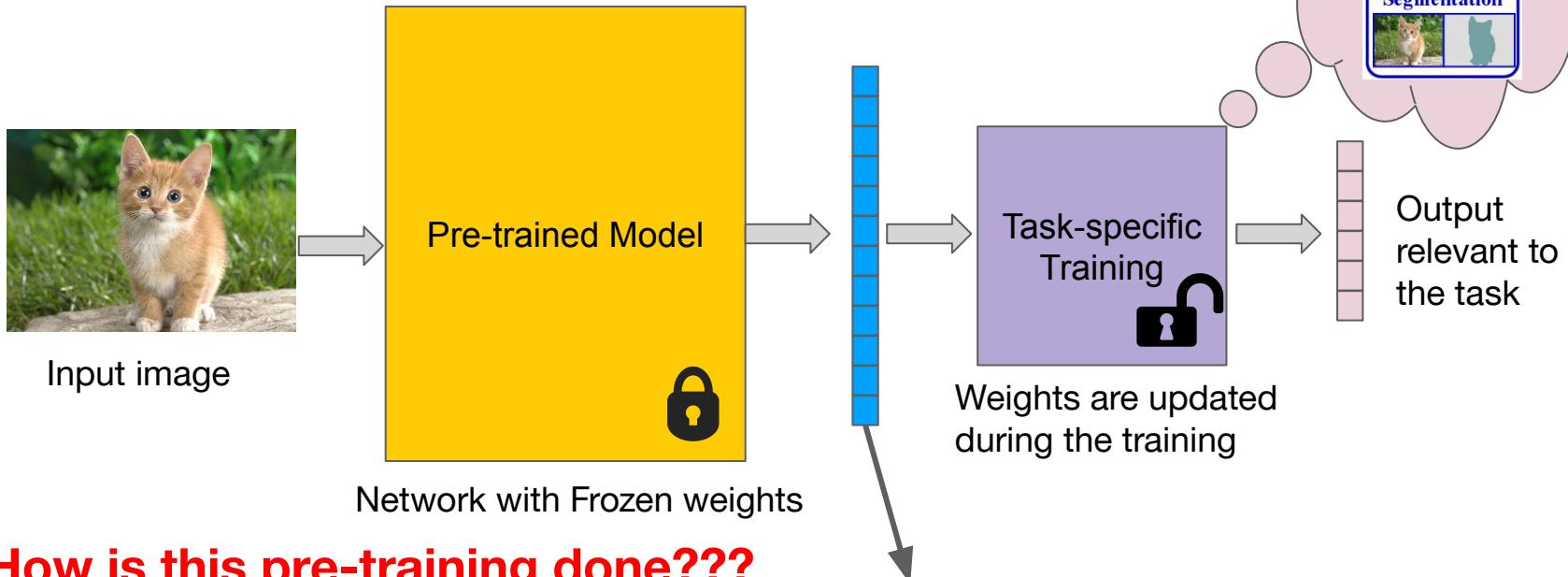
# Pre-Training in CV???

Task-specific training using pre-trained model



# Pre-Training in CV???

Task-specific training using pre-trained model

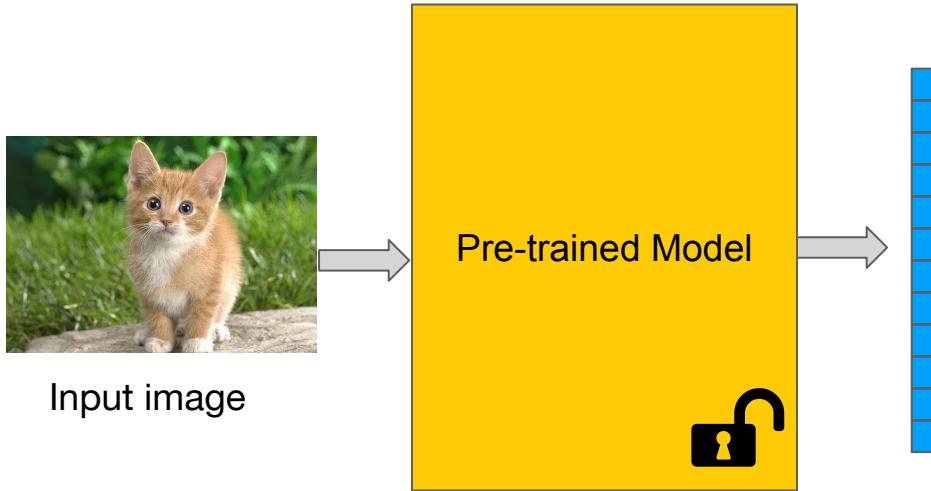


**How is this pre-training done???**



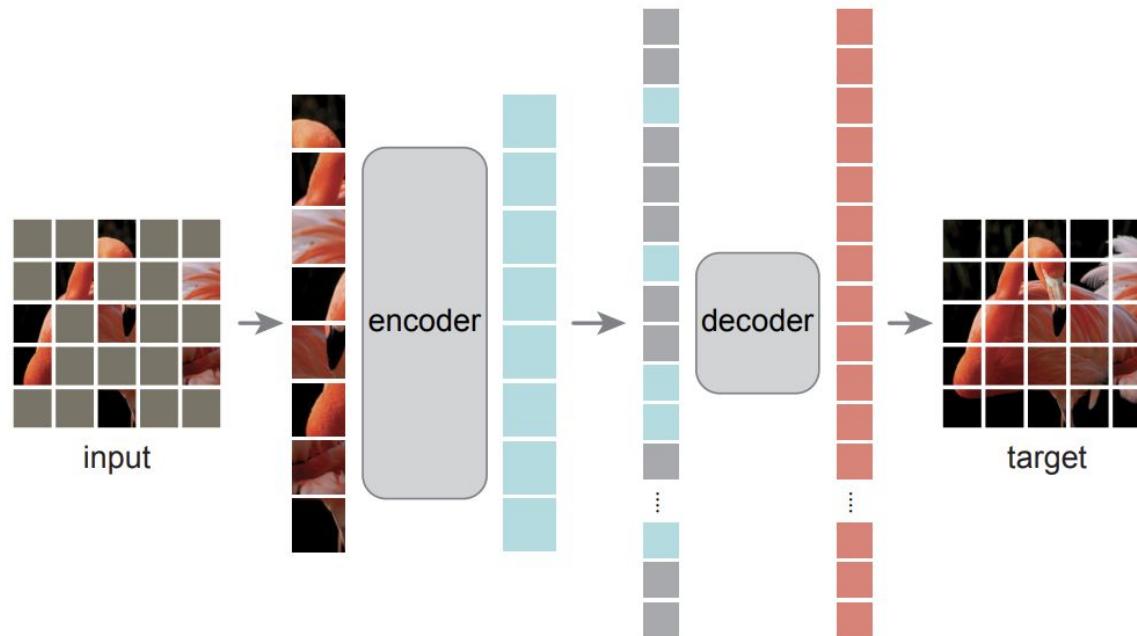
“Good” Visual feature embedding

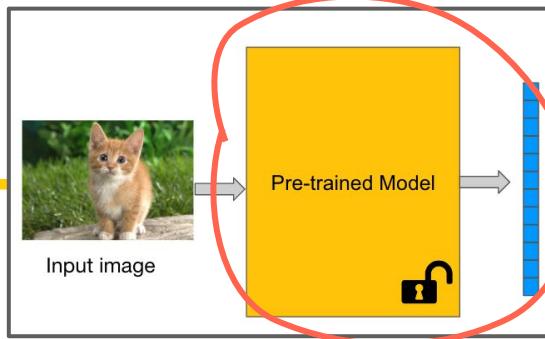
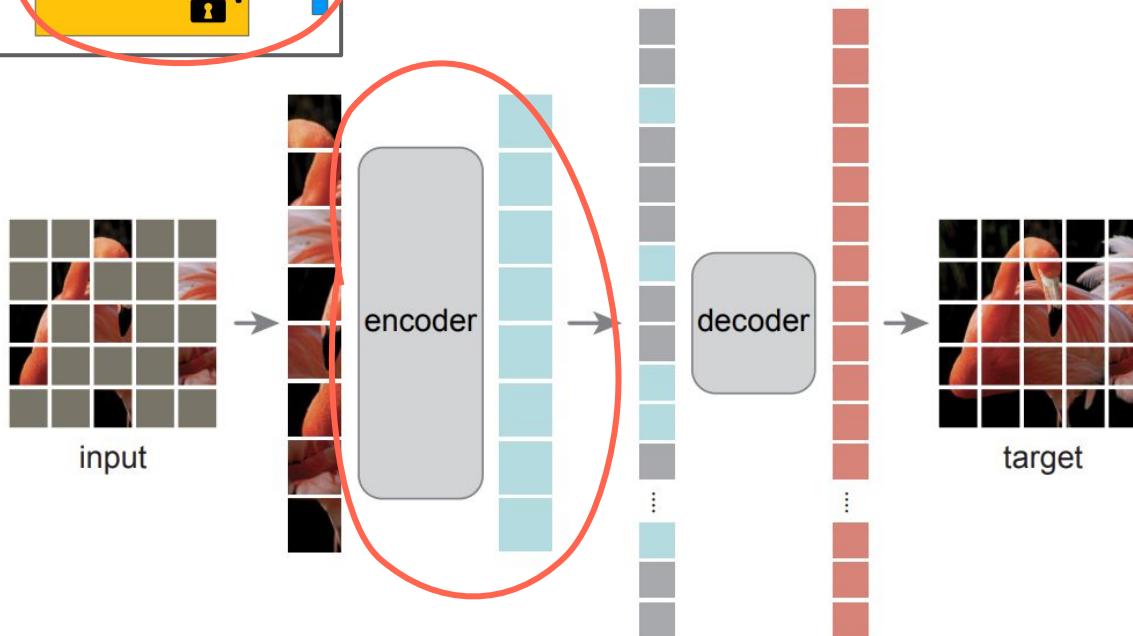
# Let us look into some pre-training models



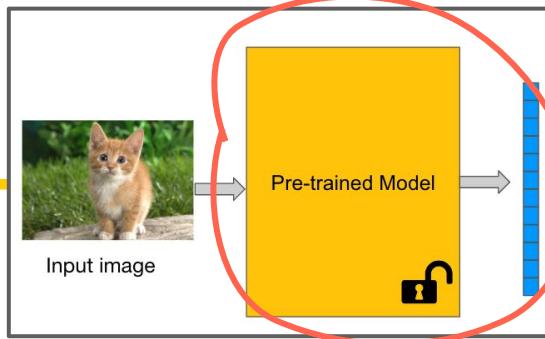
We will look into:  
• MAE  
• CLIP  
and their training objectives

## MAE

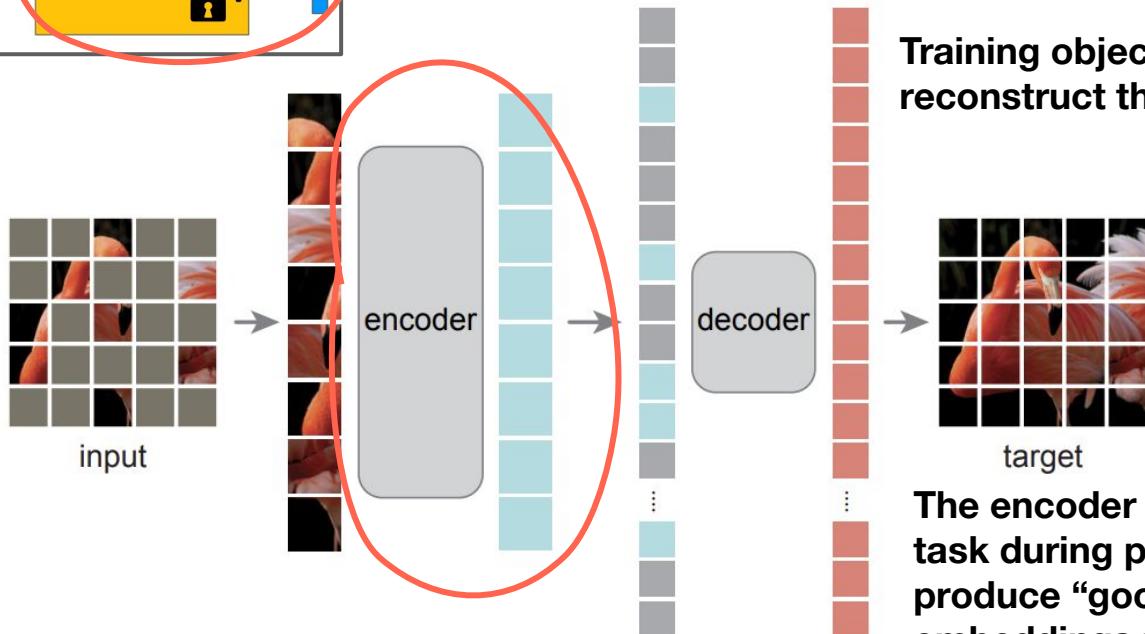


**DR****MAE**

DR



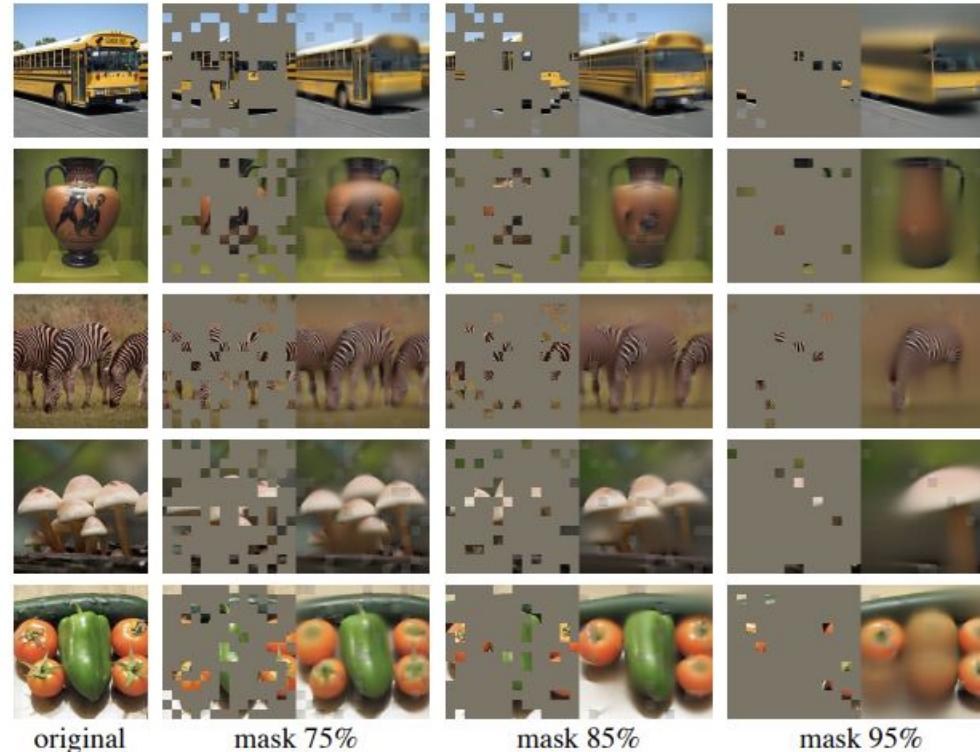
MAE



Training object is to  
reconstruct the image.

The encoder is given the hard task during pretraining to produce “good” visual embeddings to aid the decoder to reconstruct the unmasked original image

# MAE has challenging task at hand!



## Contrastive Language-Image Pretraining (CLIP)

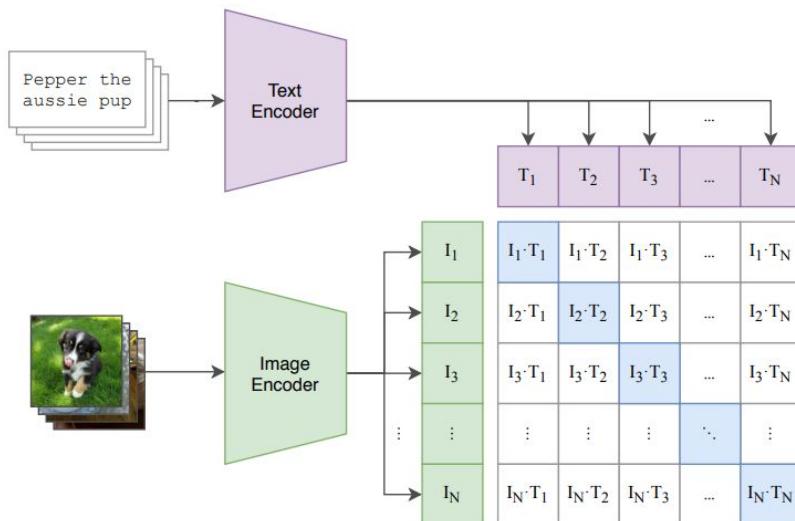
---

Turns the input (image or text) into embeddings/features  
(fixed length unit vector)

The angle between the unit vectors represents how different  
the inputs are.

# CLIP

(1) Contrastive pre-training



Training:

$I_1 \dots I_N$  - Image embeddings

$T_1 \dots T_N$  - Text embeddings

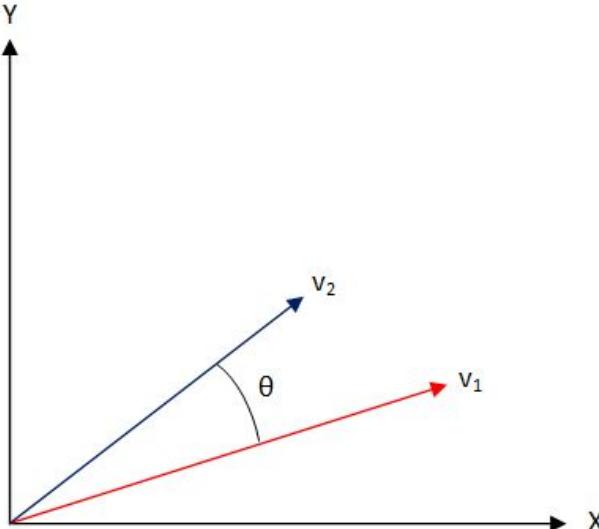
Embeddings are features vectors of fixed length and magnitude 1.

$I_x \cdot T_y$  - correlation score

(Source - Learning Transferable Visual Models From Natural Language Supervision, Radford et al., 2021)



# CLIP



**Correlation score:**

The degree of alignment between two vectors

AKA cosine similarity

$$\cos \theta = (\mathbf{v}_1 \cdot \mathbf{v}_2) / \|\mathbf{v}_1\| \|\mathbf{v}_2\| = \mathbf{v}_1 \cdot \mathbf{v}_2$$

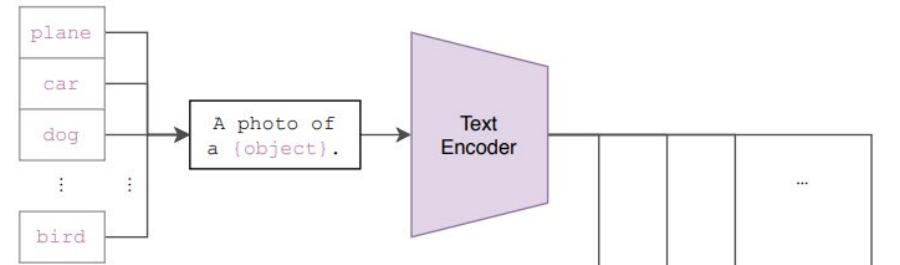
$\mathbf{v}_1$  and  $\mathbf{v}_2$  are unit vectors which are feature embeddings corresponding to two images.

# CLIP

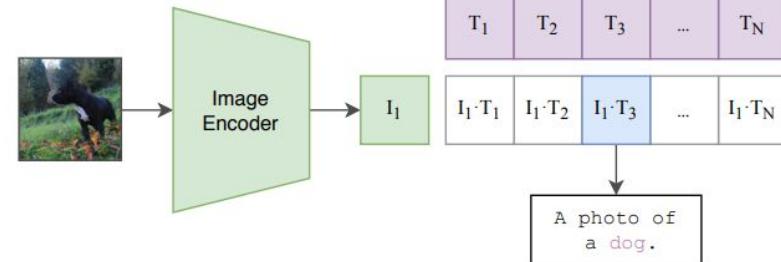
## Applications:

1. zero-shot image classification
2. Providing image and language representations for downstream tasks

(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



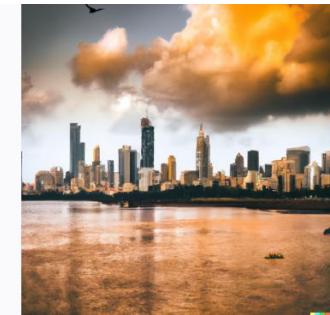
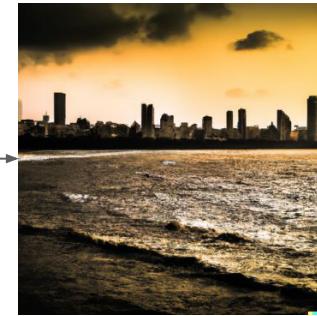
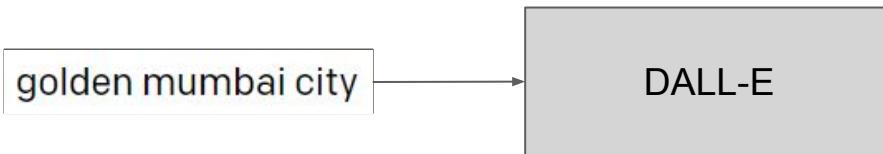
DR

# Pre-trained CLIP made DALL-E possible!

---

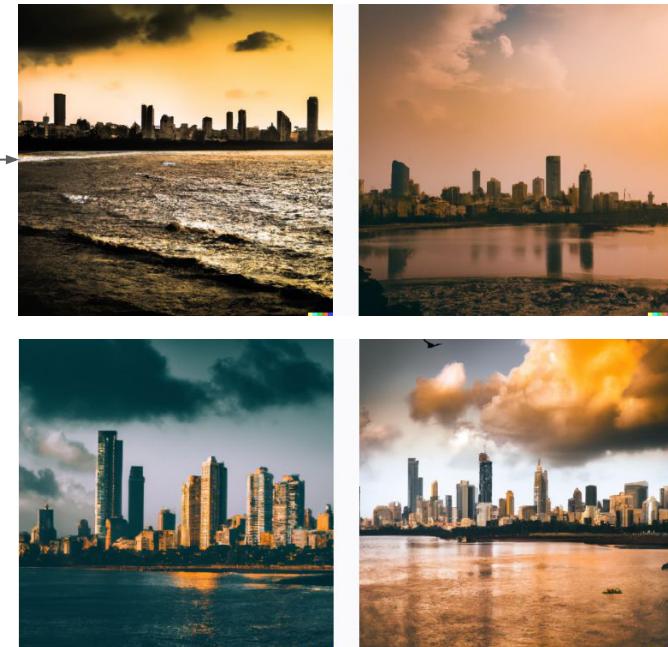
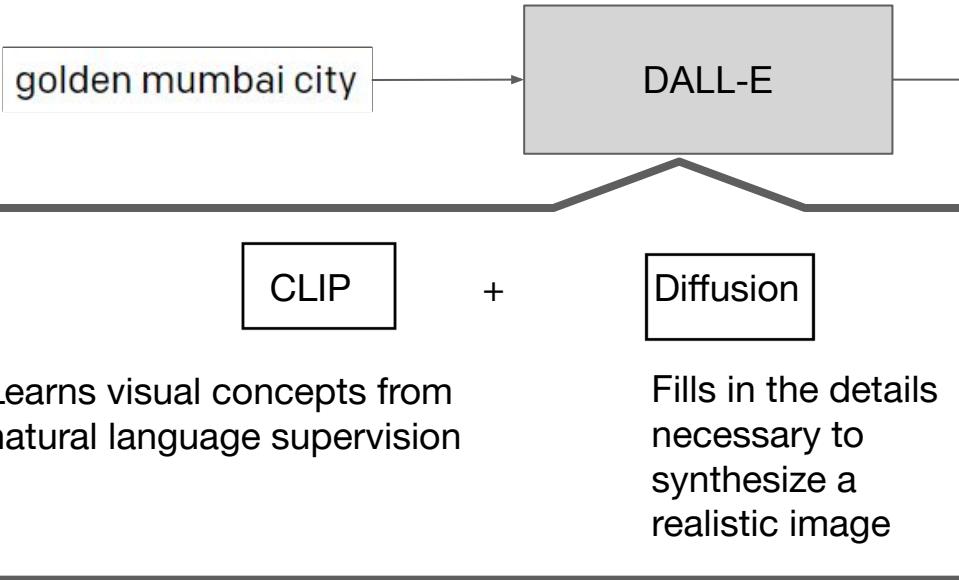
M M

# Pre-trained CLIP made DALL-E possible!



# Examples of pre-training in CV

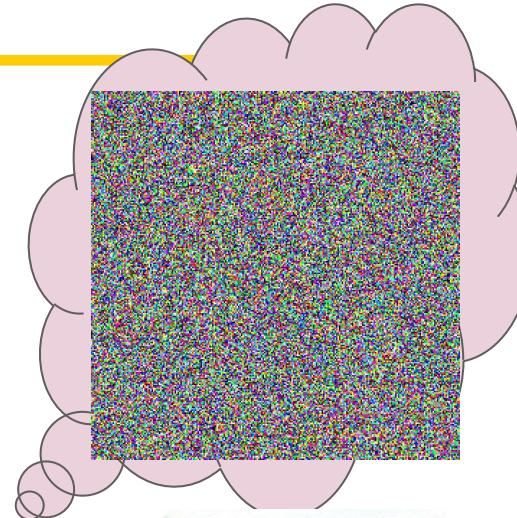
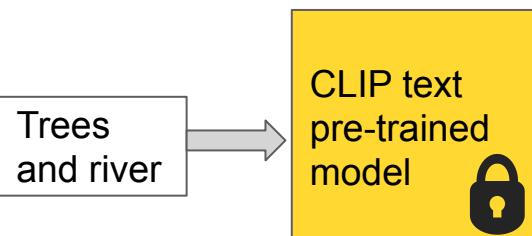
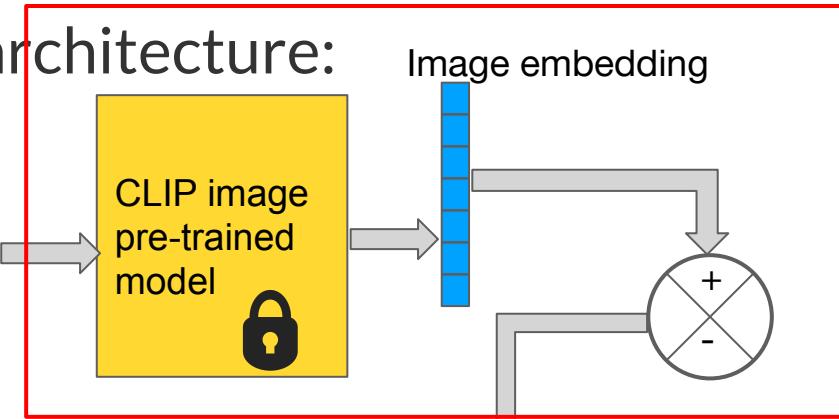
DALL-E 2:



# Examples of pre-training in CV

## DALL-E 2 architecture:

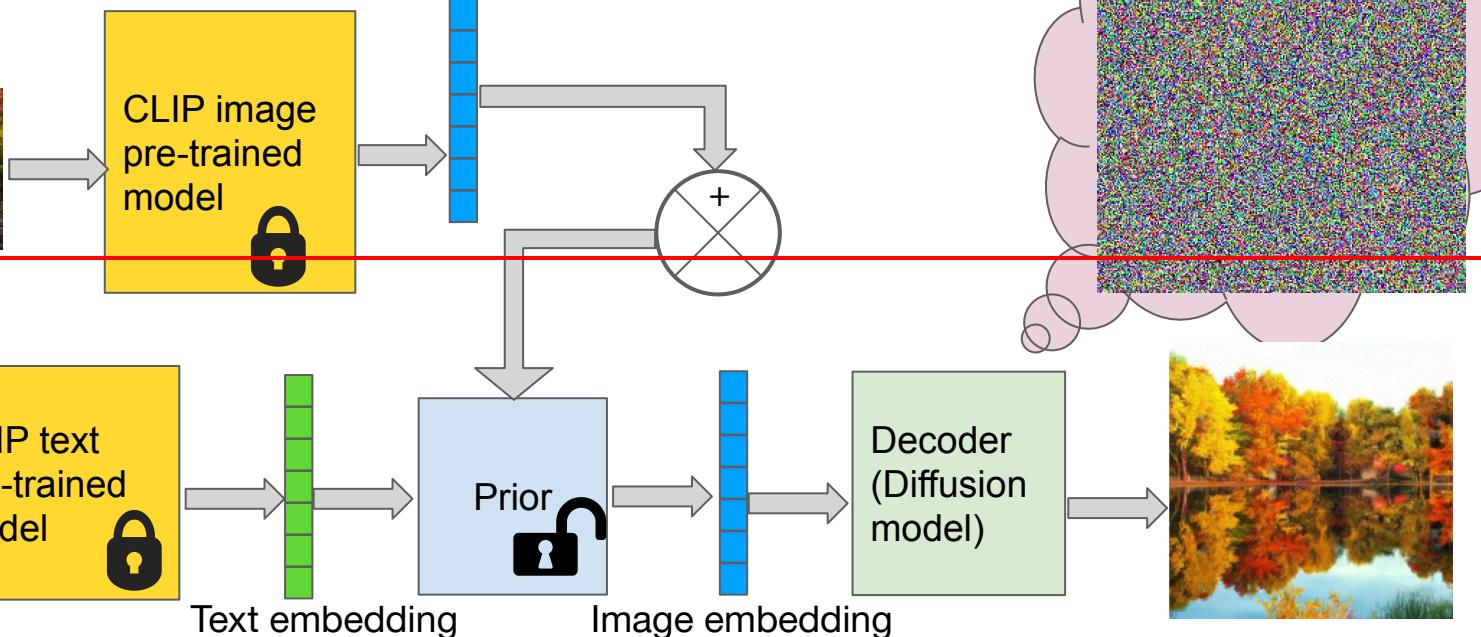
Training:



# Examples of pre-training in CV

## DALL-E 2 architecture:

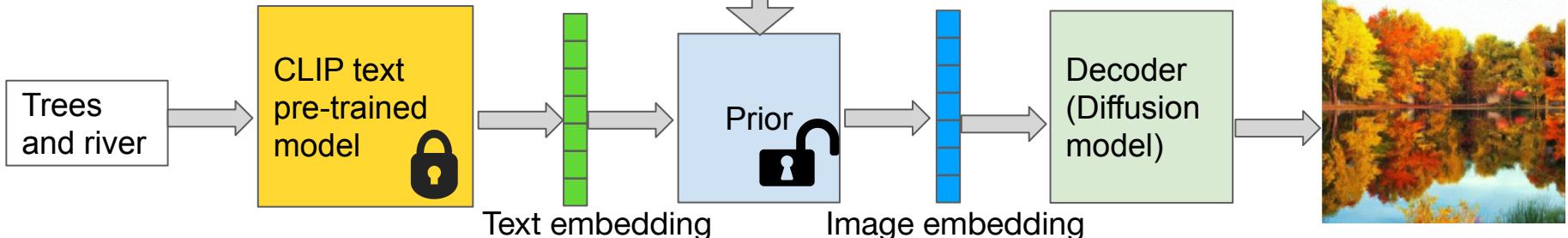
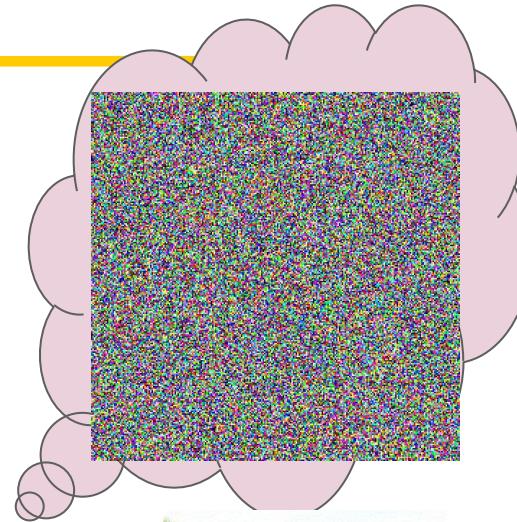
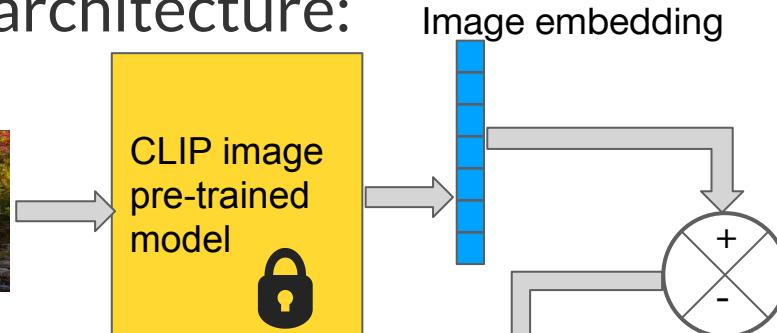
Training:



# Examples of pre-training in CV

## DALL-E 2 architecture:

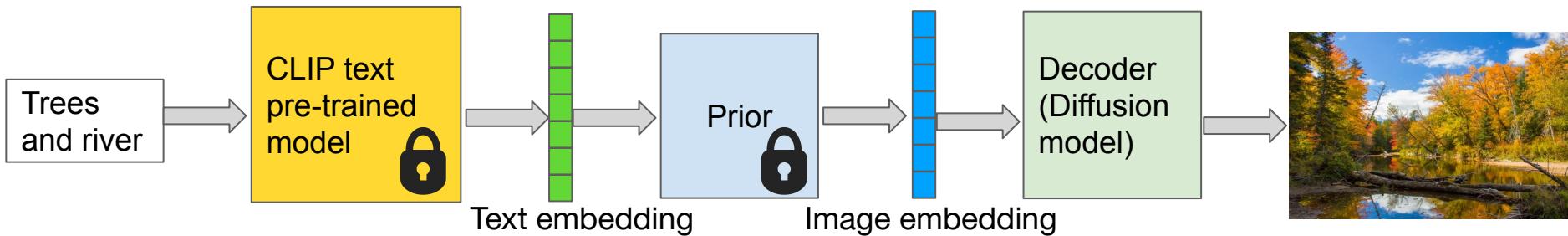
Training:



# Examples of pre-training in CV

DALL-E 2 architecture:

Testing:



# Examples of pre-training in CV

DALL-E 1 vs DALL-E 2:

Fox in a farm



DALL-E 1

Image source: <https://openai.com/product/dall-e-2>

# Examples of pre-training in CV

DALL-E 1 vs DALL-E 2:

Fox in a farm



DALL-E 2

Image source: <https://openai.com/product/dall-e-2>

# Examples of pre-training in CV

## DALL-E 1 vs DALL-E 2:

- More accurate caption matching
- More photorealism

Fox in a farm



DALL-E 1



DALL-E 2

Image source: <https://openai.com/product/dall-e-2>

# Examples of pre-training in CV

---

What made DALL-E 2 better than DALL-E 1:

- DALL-E 1 uses **discrete variational autoencoder (dVAE)**, **next token prediction** and **CLIP model re-ranking**.

# Examples of pre-training in CV

---

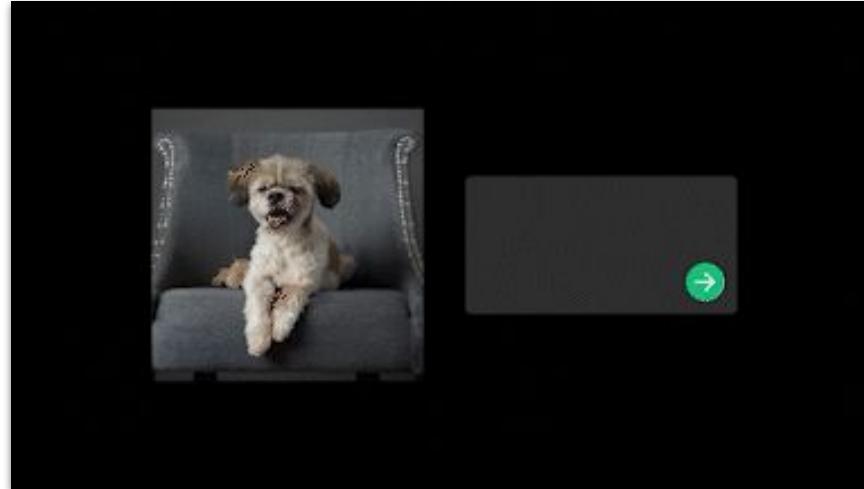
What made DALL-E 2 better than DALL-E 1:

- DALL-E 1 uses discrete variational autoencoder (dVAE), next token prediction and CLIP model re-ranking.
- DALL-E 2 uses **CLIP embedding directly** and decodes image via **diffusion** similar to GLIDE (a text guided diffusion model).

# Examples of pre-training in CV

DALL-E 2 additional:

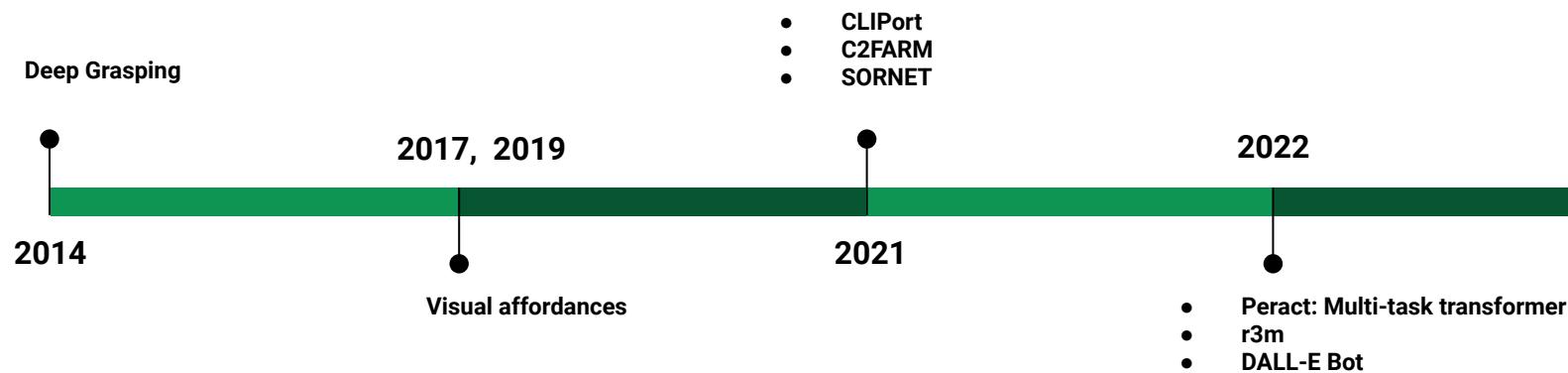
- Text based image editing



Gif source: <https://openai.com/product/dall-e-2>

# Examples of pre-training in Robotics

General timeline:



# R3M

**R3M:** Reusable Representation for Robotic Manipulation.

**Universal Visual Representation:** A universal visual representation refers to a visual encoding of data that can be used across multiple tasks or domains.

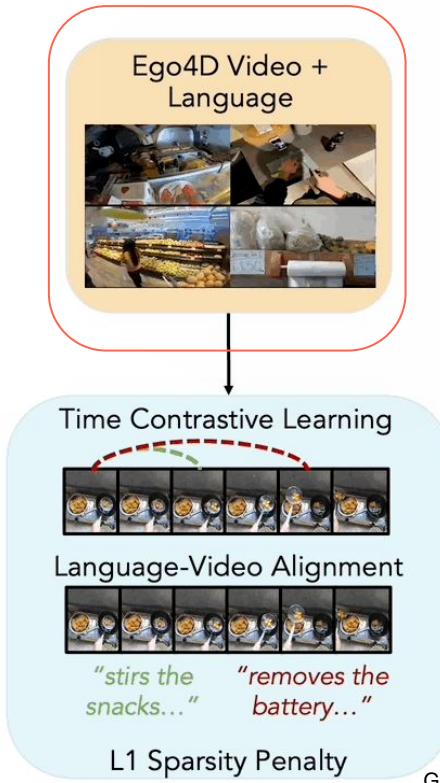
**Manipulation:** Ability of a robot to interact and physically manipulate objects in its environment. For example: grasping, picking up, moving, and placing objects.

**Application:** Anything that needs manipulation



Gif Source: <https://sites.google.com/view/robot-r3m/>

# Data Set



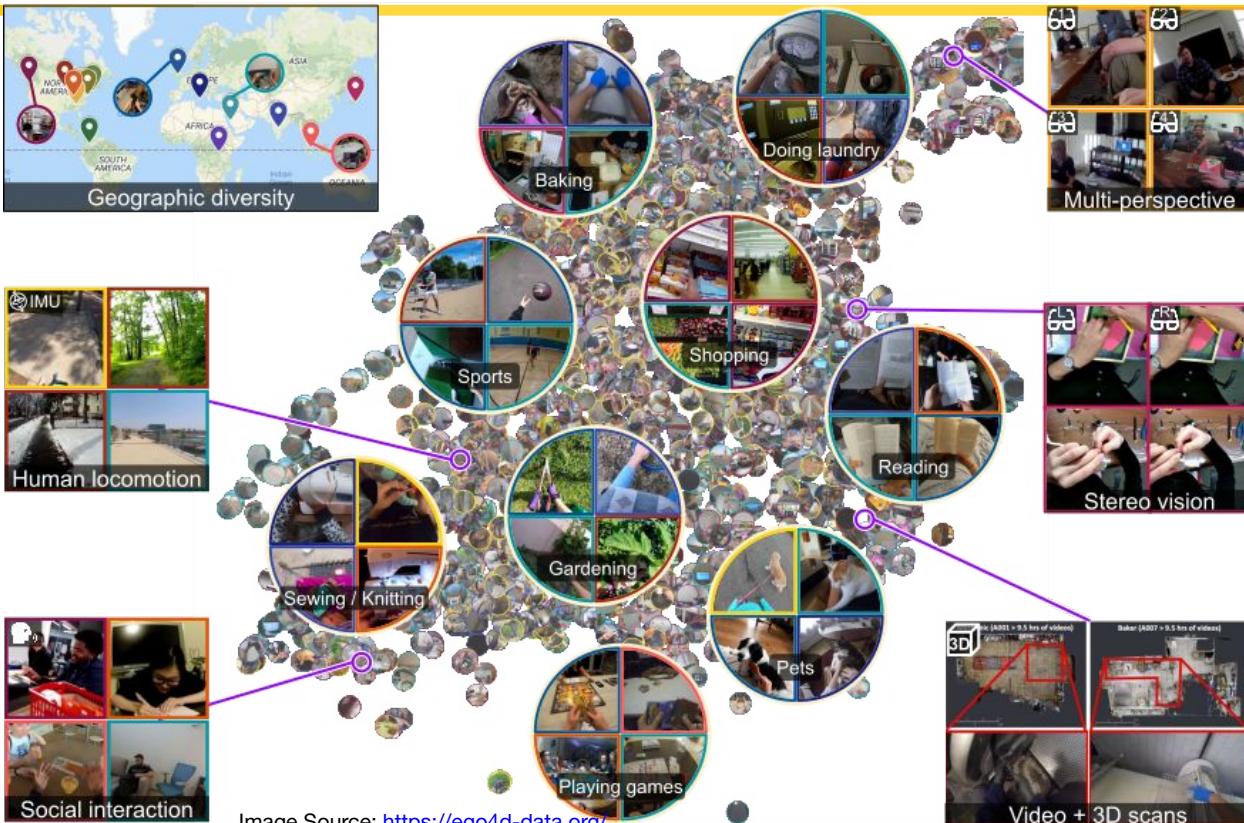
## R3M: Reusable Representations for Robotic Manipulation

Efficient Robot Learning  
New Environment, New Tasks



Gif Source : <https://sites.google.com/view/robot-r3m/>

# Ego 4D



DR

# Ego 4D



Gif Source: [https://www.seas.upenn.edu/~shzhou2/projects/eos\\_dataset/](https://www.seas.upenn.edu/~shzhou2/projects/eos_dataset/)



# Simulation Environments

MetaWorld

Assembly, Bin Picking, Button Pressing, Drawer Opening, Hammering



Adroit

Re-orient Pen,  
Relocate Ball



Franka Kitchen

Sliding Door, Turning Light On, Opening Door, Turning Knob, Opening Microwave

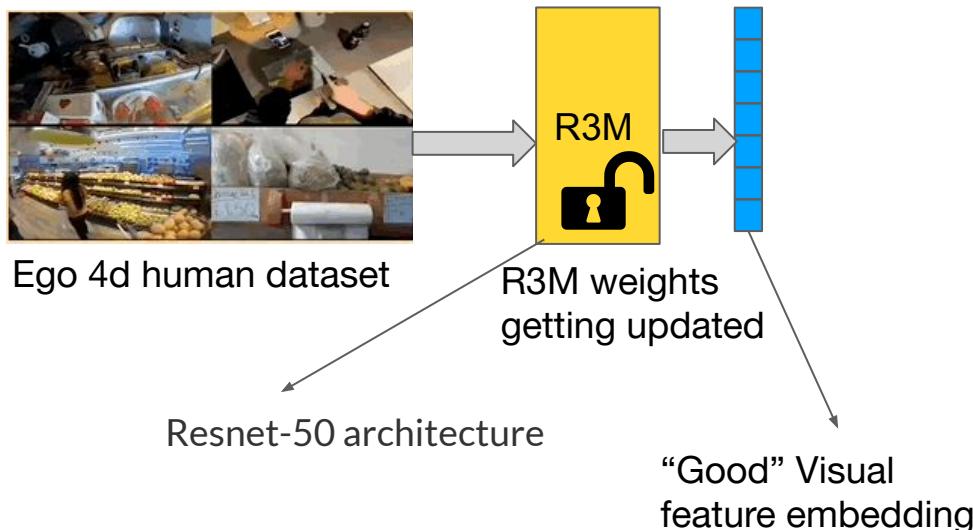


Image Source : <https://sites.google.com/view/robot-r3m/>

# Examples of pre-training in Robotics

Network architecture:

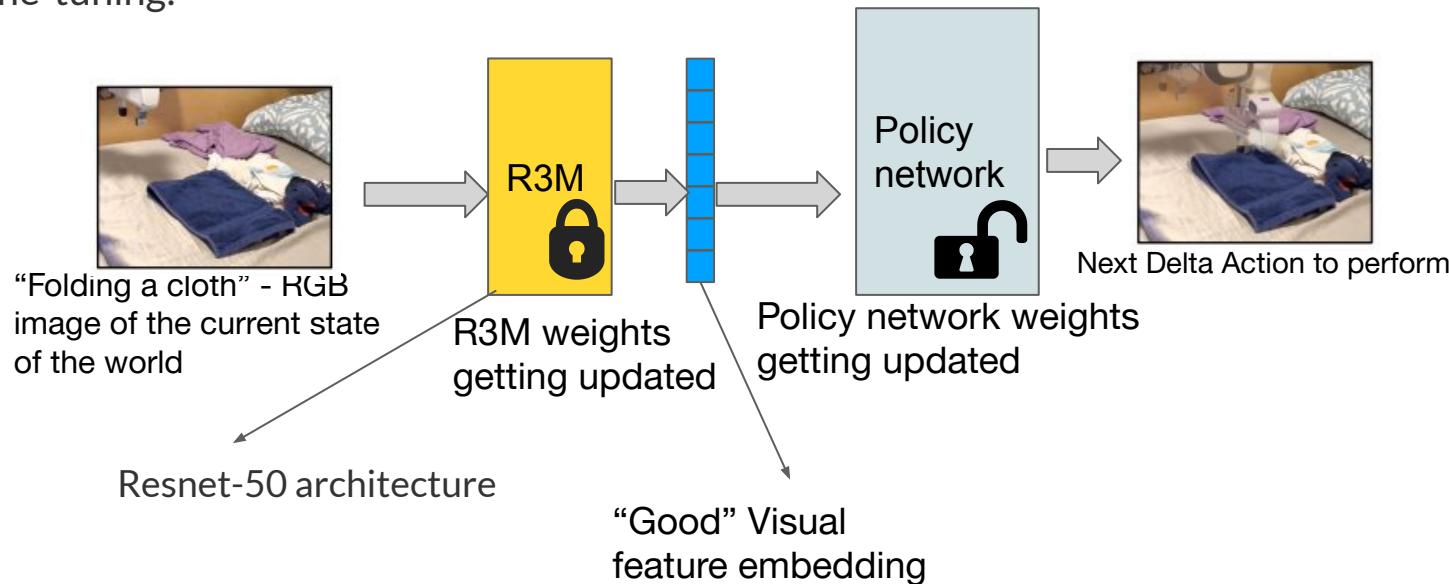
Pre-training:



# Examples of pre-training in Robotics

Network architecture:

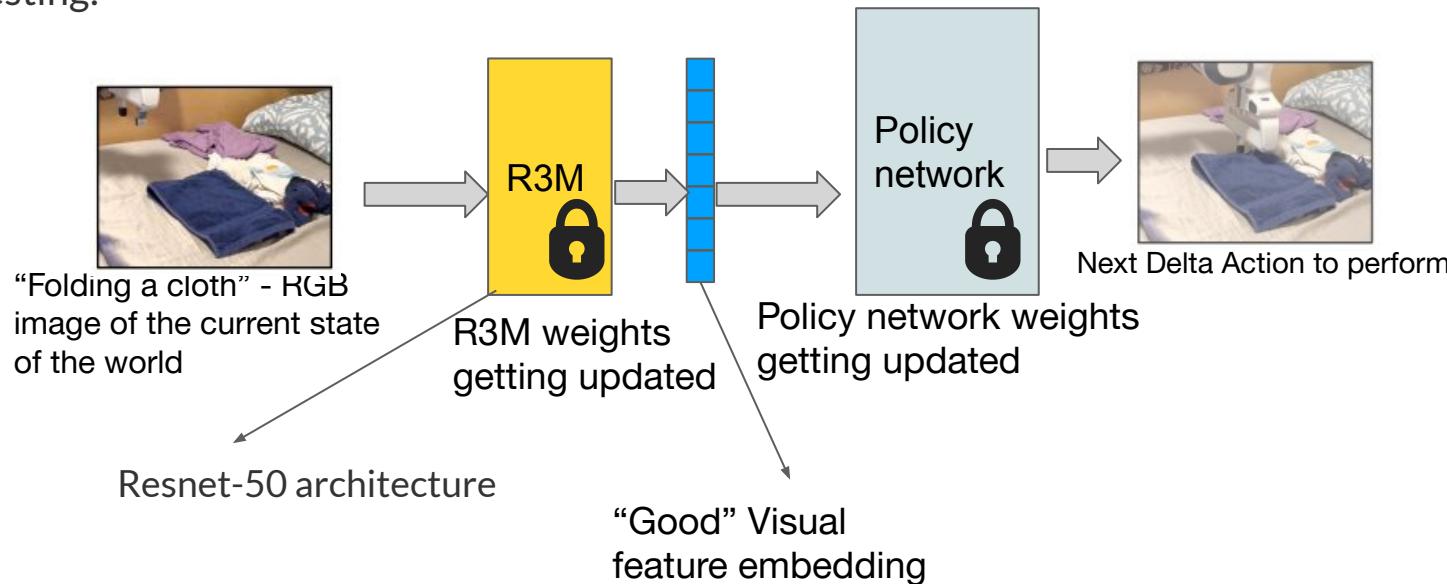
Fine-tuning:



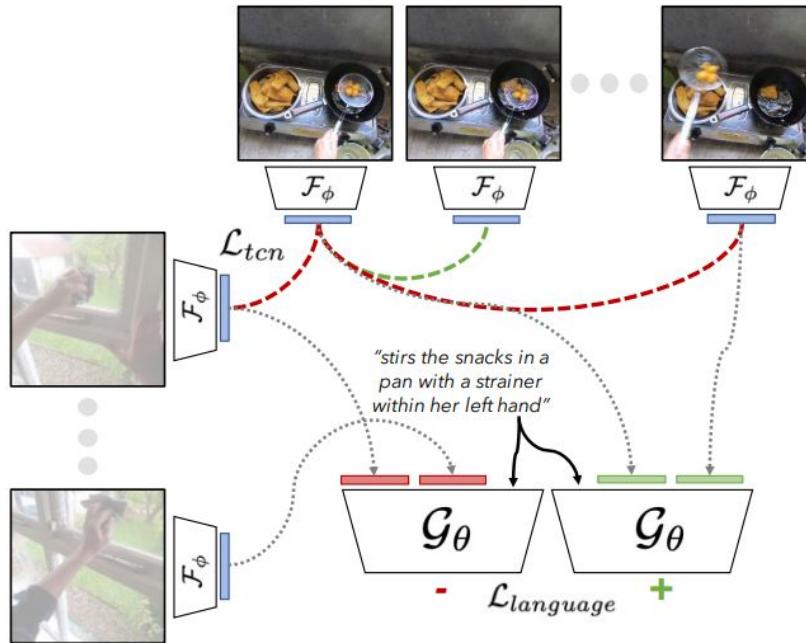
# Examples of pre-training in Robotics

Network architecture:

Testing:



# RESNET – Pretraining Objectives

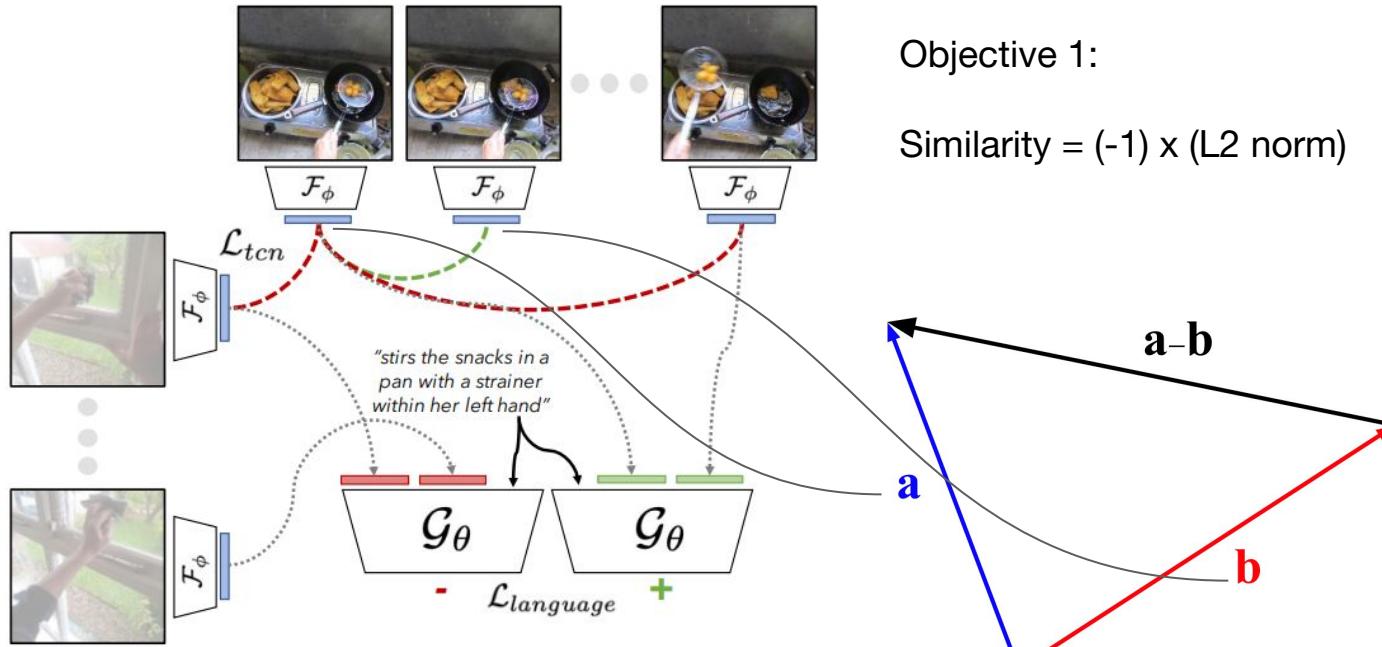


Objective 1:

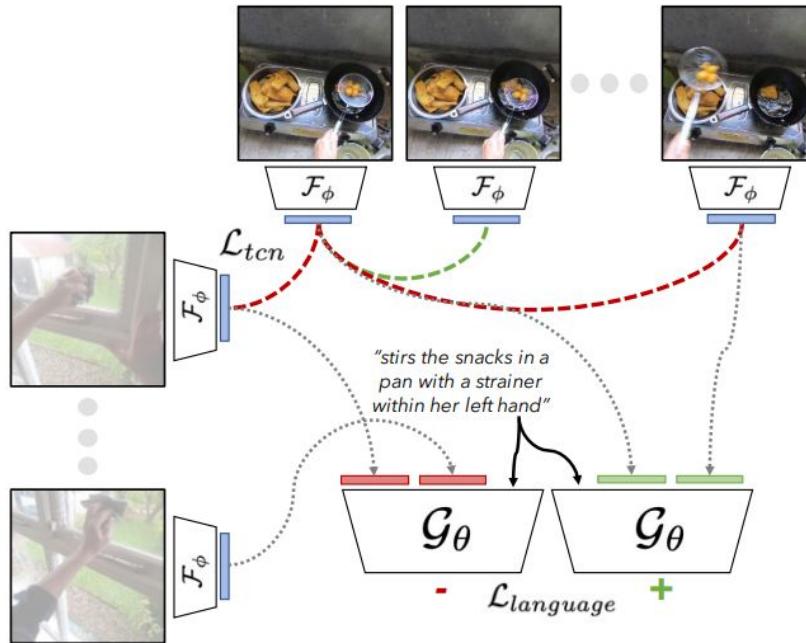
Capture the temporal dynamics

Closer frames must be more similar

# RESNET – Pretraining Objectives



# RESNET – Pretraining Objectives



Objective 1:

$$\text{Similarity} = (-1) \times (\text{L2 norm})$$

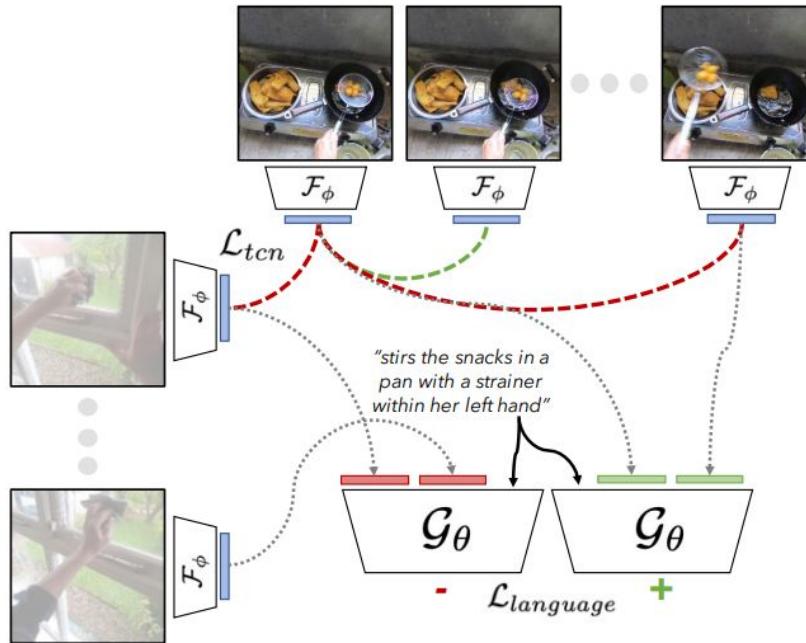
$$\mathcal{L}_{tcn} = - \sum_{b \in B} \log \frac{e^{\mathcal{S}(z_i^b, z_j^b)}}{e^{\mathcal{S}(z_i^b, z_j^b)} + e^{\mathcal{S}(z_i^b, z_k^b)} + e^{\mathcal{S}(z_i^b, z_i^{\neq b})}}$$

$z_j^b$  - image representation

$S$  - similarity score between two frames

$i, j, \dots$  are randomly sampled for each video sequence

# RESNET – Pretraining Objectives



Objective 2:

Capture semantically relevant features

AKA video-language alignment (similar to CLIP)

# RESNET – Pretraining Objectives

Objective 2: Video alignment

$$\mathcal{L}_{language} = - \sum_{b \in B} \log \frac{e^{\mathcal{G}_\theta(z_0^b, z_{j>i}^b, l^b)}}{e^{\mathcal{G}_\theta(z_0^b, z_{j>i}^b, l^b)} + e^{\mathcal{G}_\theta(z_0^b, z_i^b, l^b)} + e^{\mathcal{G}_\theta(z_0^{\neq b}, z_{j>i}^{\neq b}, l^b)}}$$

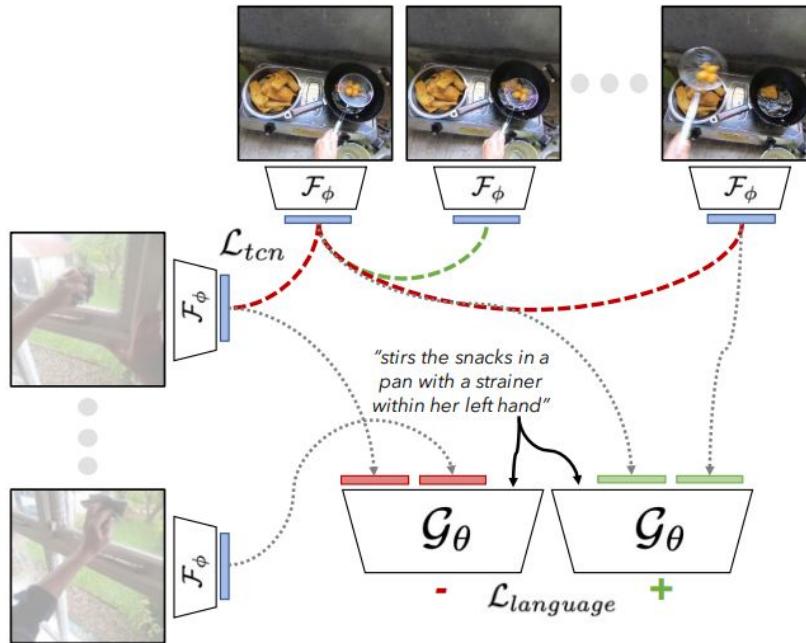
$z_j^b$  - image representation for the  $j^{\text{th}}$  frame in the  $b^{\text{th}}$  frame sequence (video)

$l^b$  - language representation for the text description corresponding to the  $b^{\text{th}}$  video

$\mathcal{G}_\theta$  - Transition score correlating the initial and final frames to the text label (Nair et al.)

$i, j, \dots$  are randomly sampled for each video sequence (NCE)

# RESNET – Pretraining Objectives



Objective 3:

Representations must be compact/sparse

L1 + L2 Regularization

# RESNET – Pretraining Objectives

---

Overall objective - minimize the following loss function

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{I_{0,i,j,k}^{1:B} \sim \mathcal{D}} [\lambda_1 \mathcal{L}_{tcn} + \lambda_2 \mathcal{L}_{language} + \lambda_3 \|\mathcal{F}_\phi(I_i)\|_1 + \lambda_4 \|\mathcal{F}_\phi(I_i)\|_2]$$

1.  $\mathcal{L}_{tcn}$  - Time contrastive network loss
2.  $\mathcal{L}_{language}$  - Video-language alignment loss
3.  $\|\mathcal{F}_\phi\|_1$  - L1 regularization loss
4.  $\|\mathcal{F}_\phi\|_2$  - L2 regularization loss

I, j and k are randomly sampled, then the mean loss is calculated over the samples.



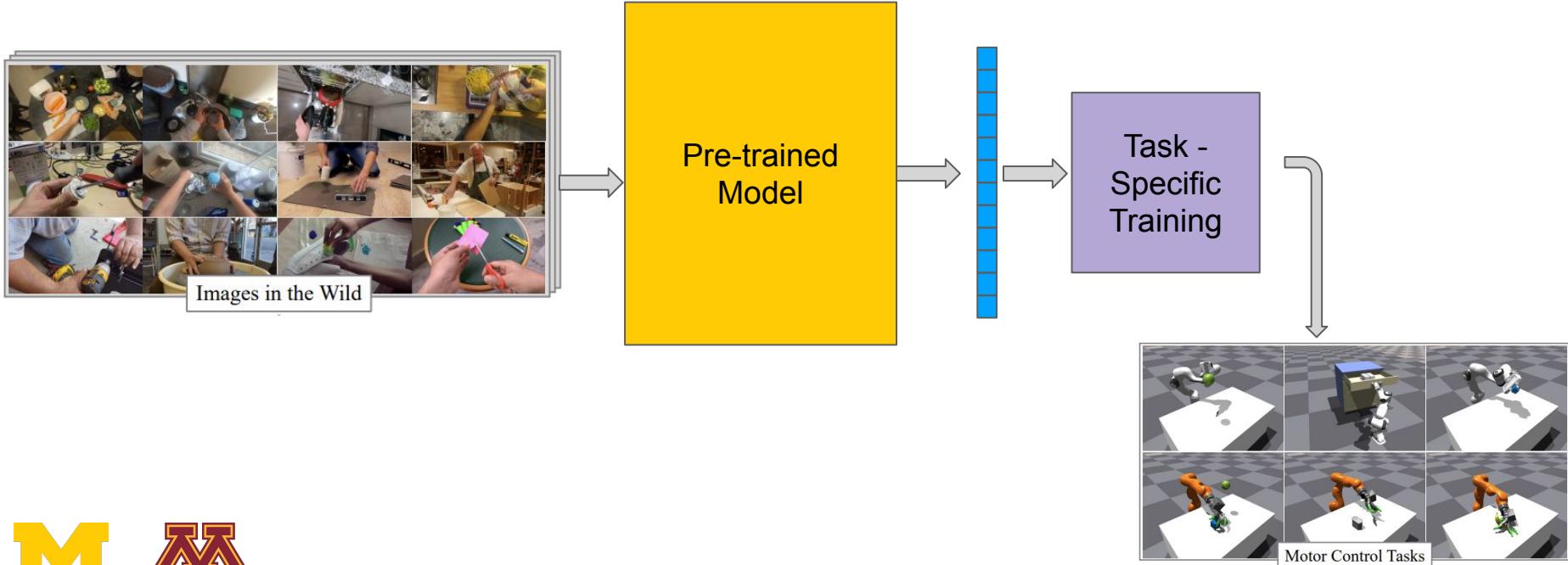
# RESNET – Performance

Success out of 10 trials	R3M	CLIP
Closing Drawer	<b>80%</b>	70%
Putting Mask in Dresser	<b>30%</b>	10%
Putting Lettuce in Pan	<b>60%</b>	0%
Pushing Mug to Goal	<b>70%</b>	40%
Folding Towel	<b>40%</b>	0%
Average	<b>56%</b>	24%

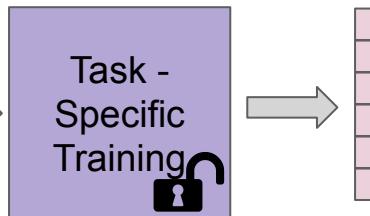
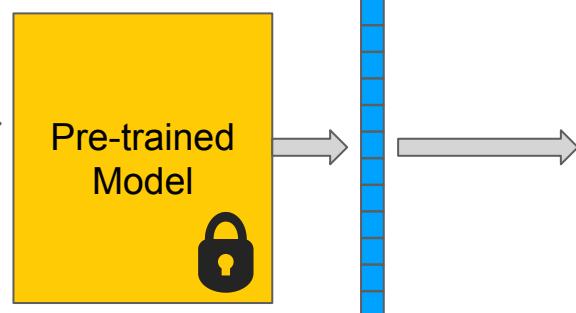
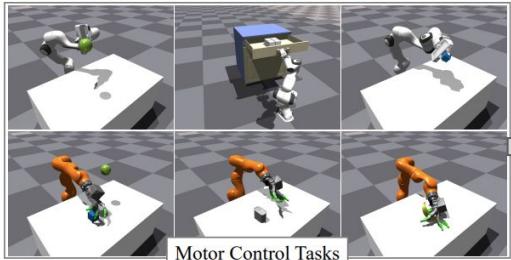
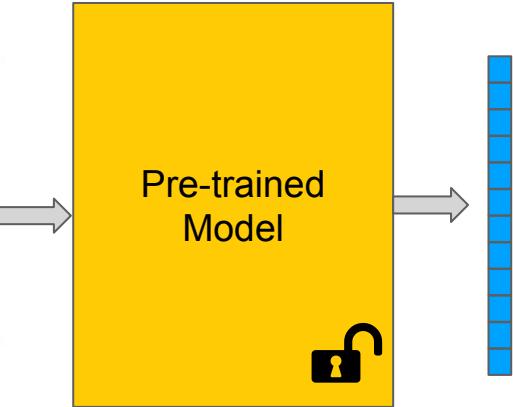
Experiment derived from Parisi et al.

(additional details go [here](#))

# MVP (Masked Visual Pretraining)



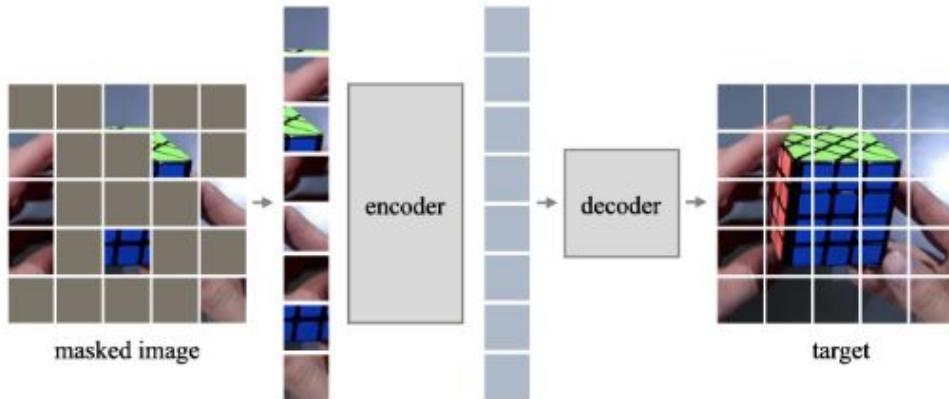
# MVP (Masked Visual Pretraining)



Motor  
commands or  
equivalent  
action  
commands



# MVP (Masked Visual Pretraining)

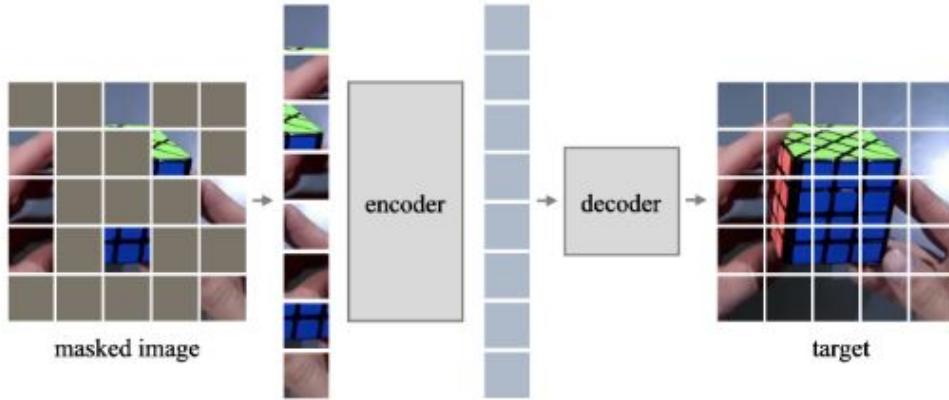


(a) masked visual pretraining

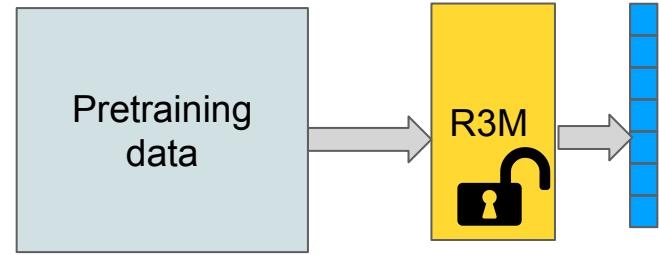
Remember MAE!



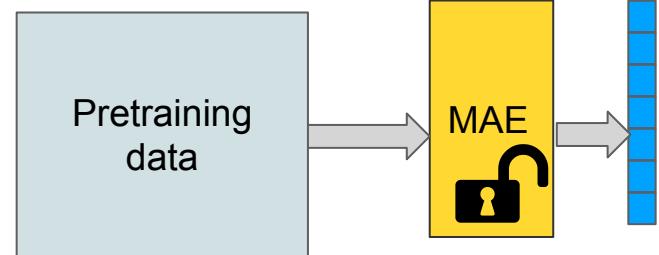
# MVP (Masked Visual Pretraining)



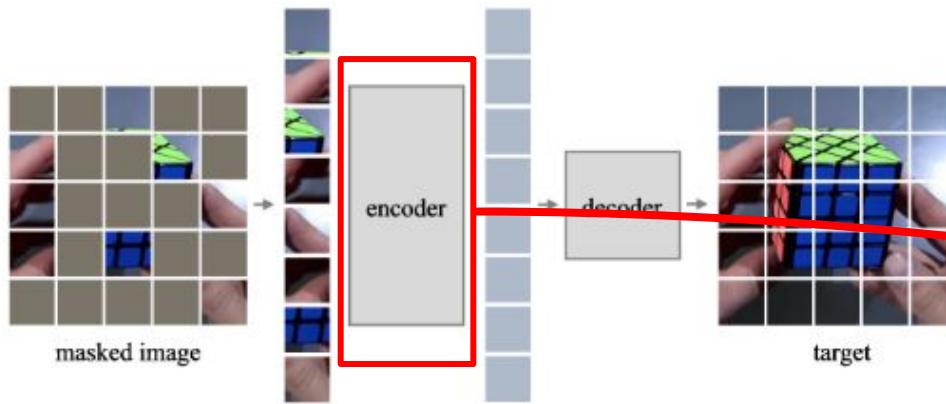
(a) masked visual pretraining



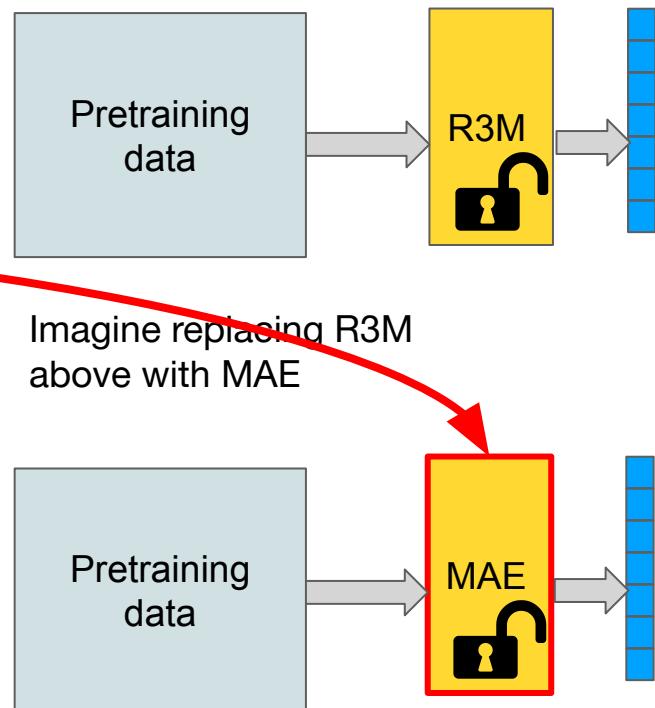
Imagine replacing R3M above with MAE



# MVP (Masked Visual Pretraining)



(a) masked visual pretraining





# MVP - Data Set

---

Egocentric Epic Kitchens dataset +

the YouTube 100 Days of Hands dataset +

the crowd-sourced Something-Something dataset =

Human-Object Interaction dataset (HOI) (~700k Images)



# MAE Reconstructions



Masked

Reconstructed

Ground-Truth

Masked

Reconstructed

Ground-Truth

Masked

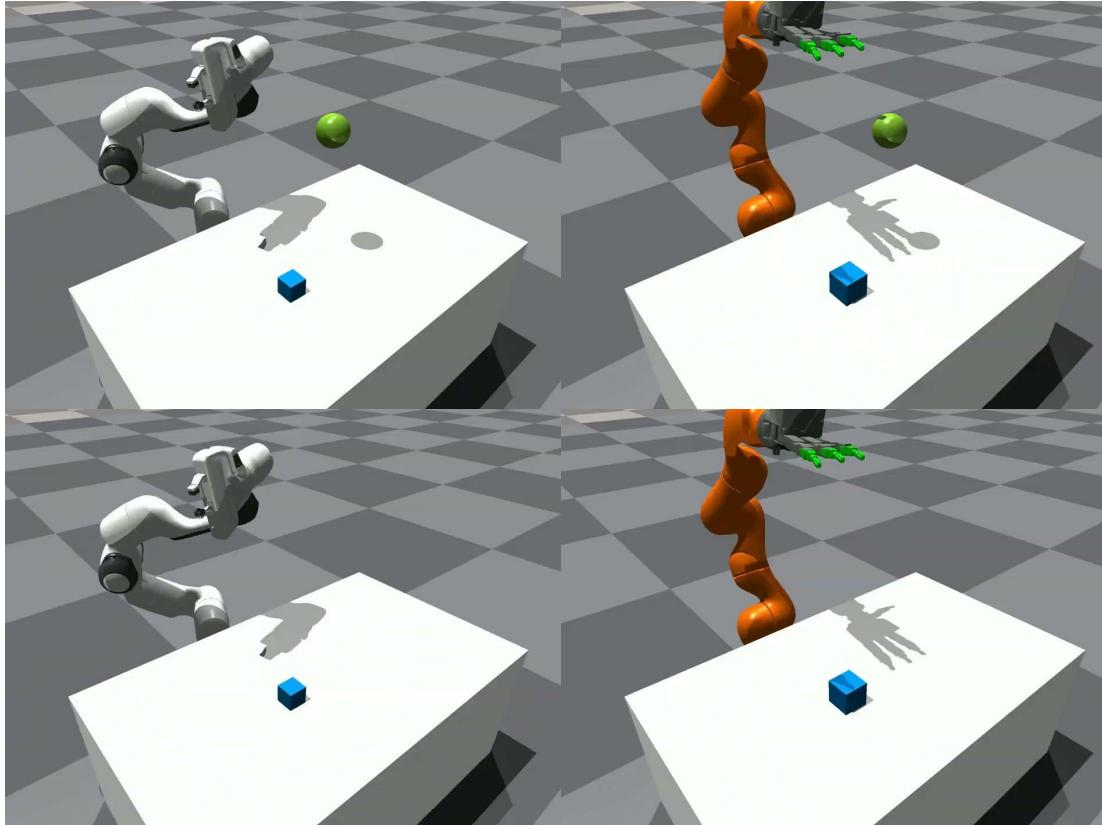
Reconstructed

Ground-Truth



# MVP (Manipulation tasks)

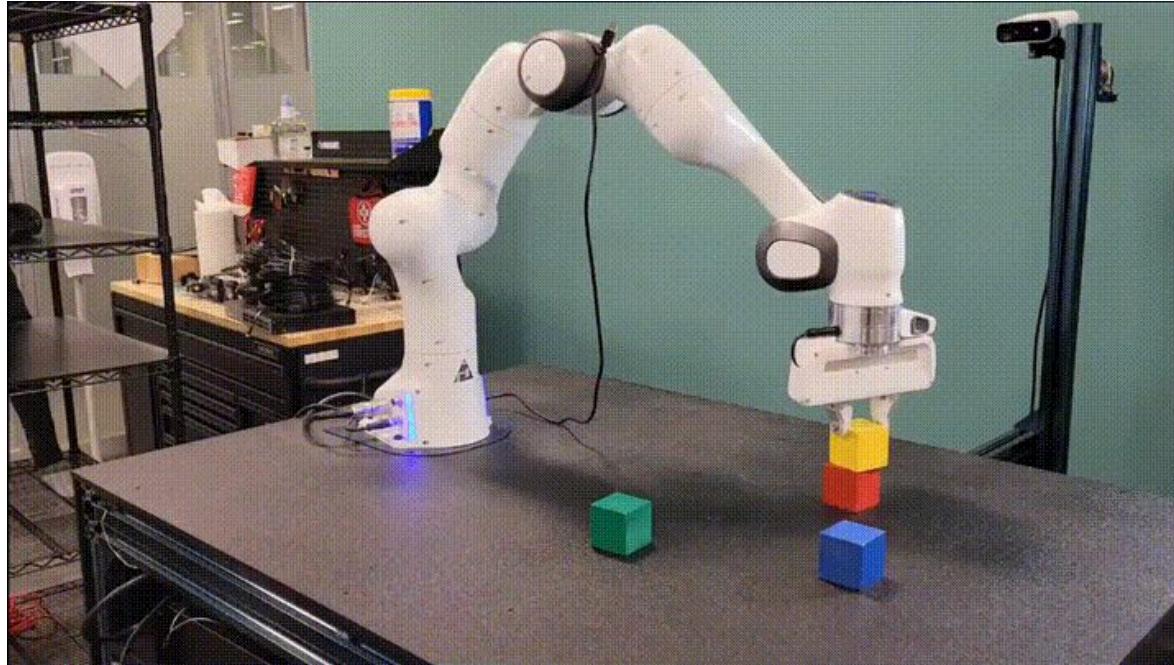
Franka



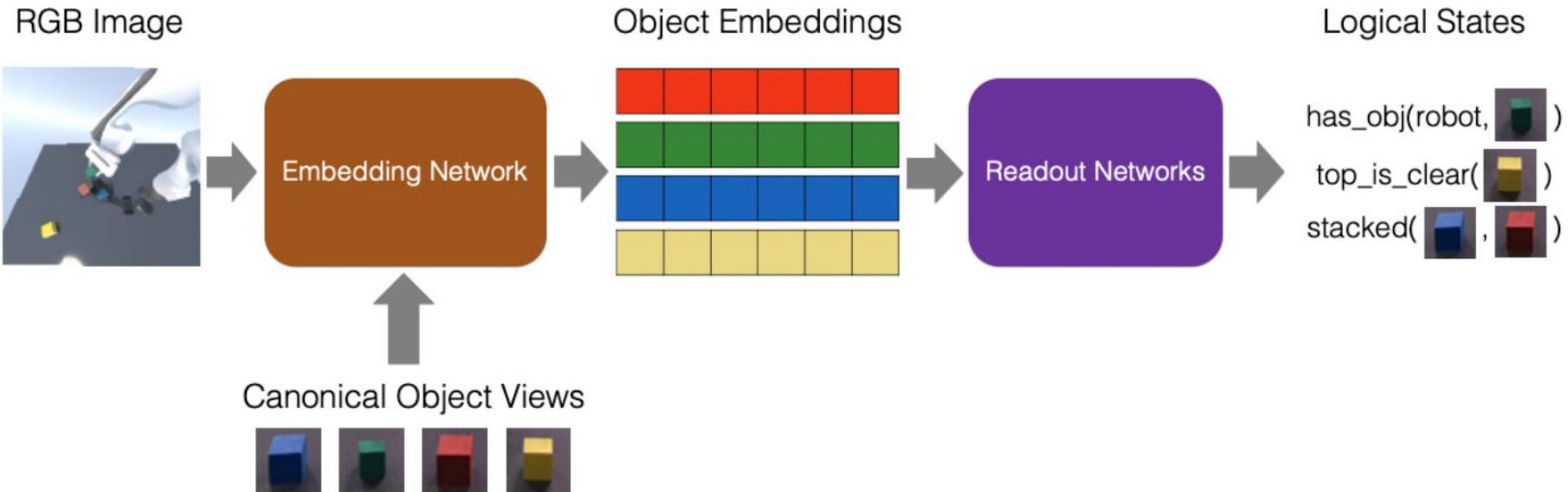
Kuka

**DR**

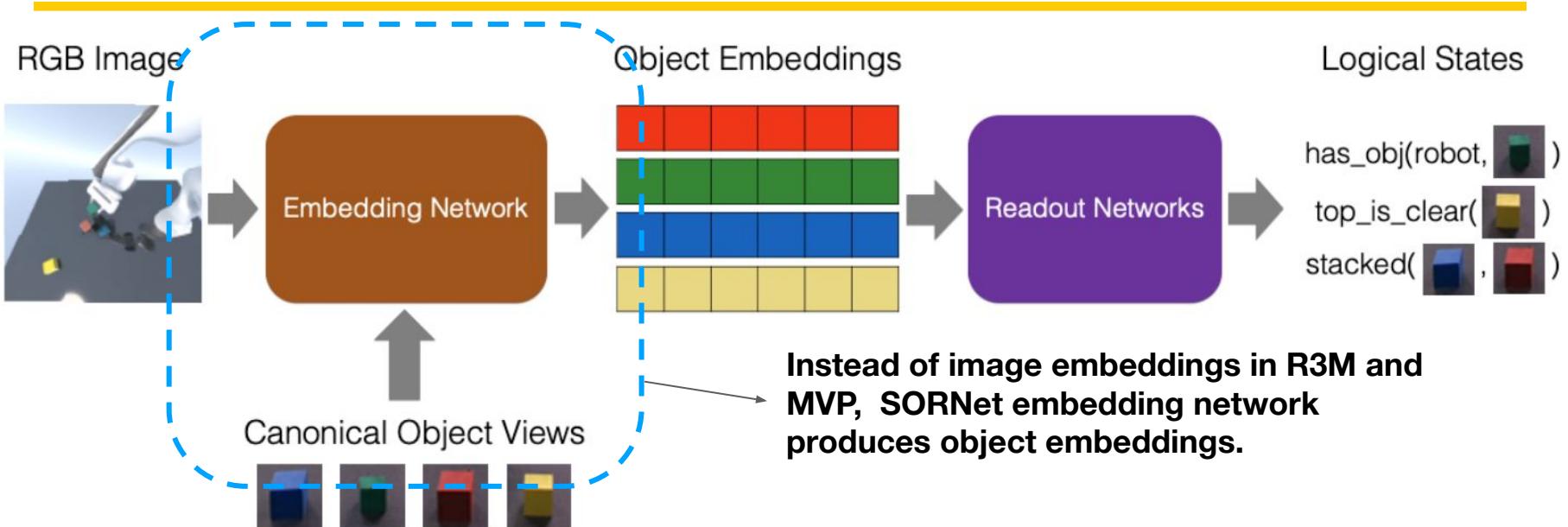
# SORNet: Spatial Object-centric Representation Network



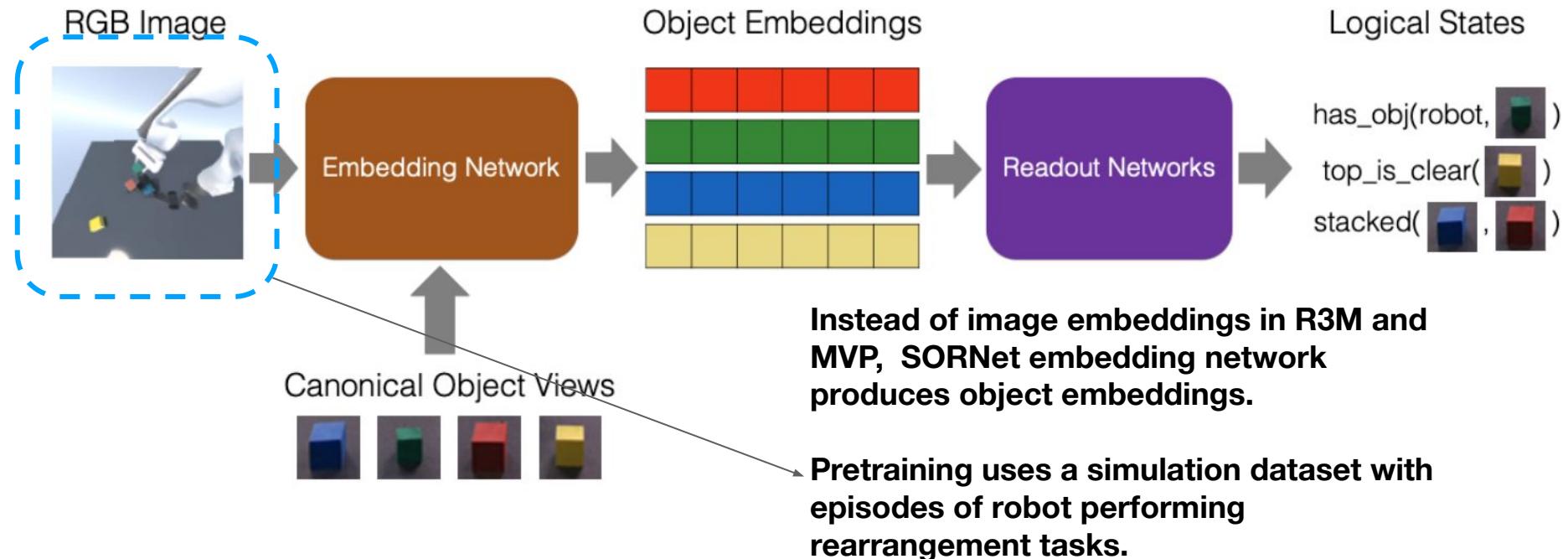
# SORNET



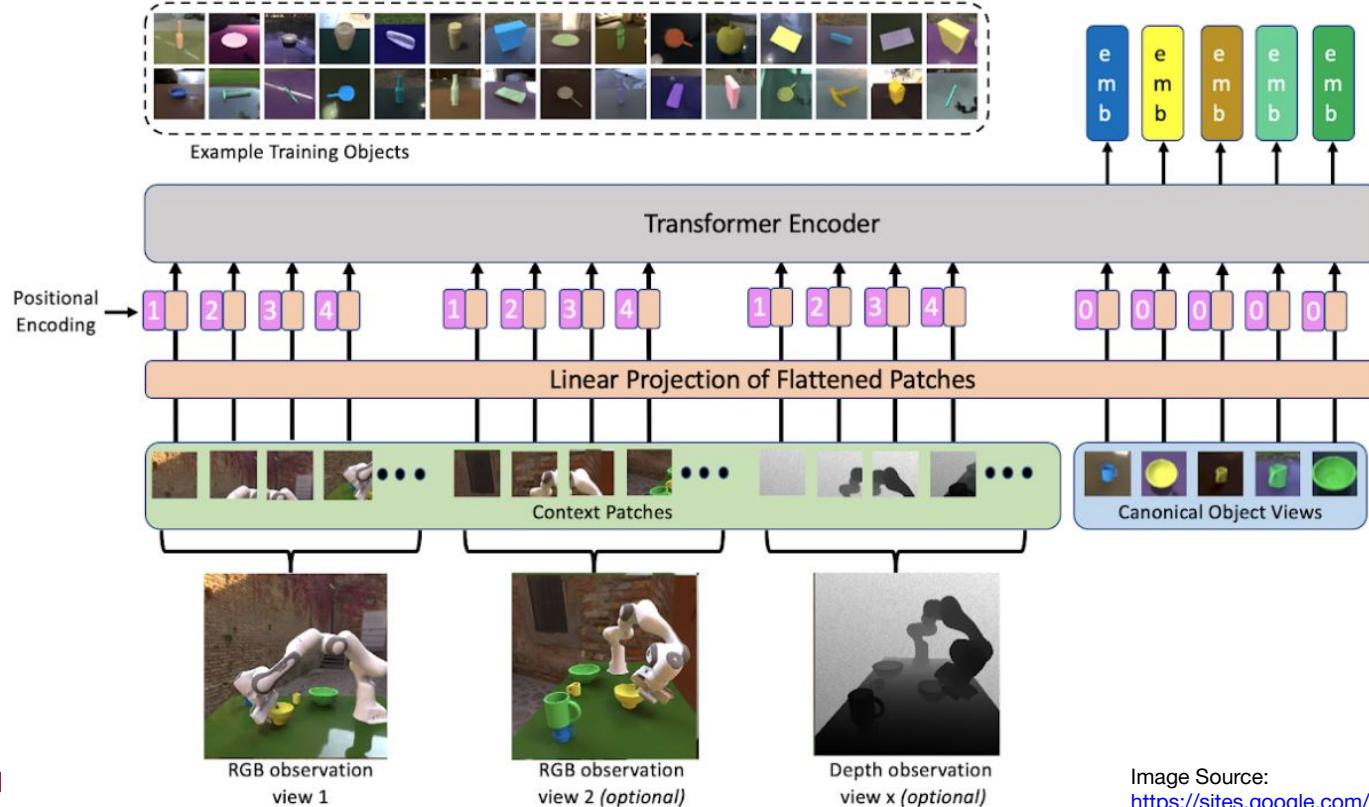
# SORNET



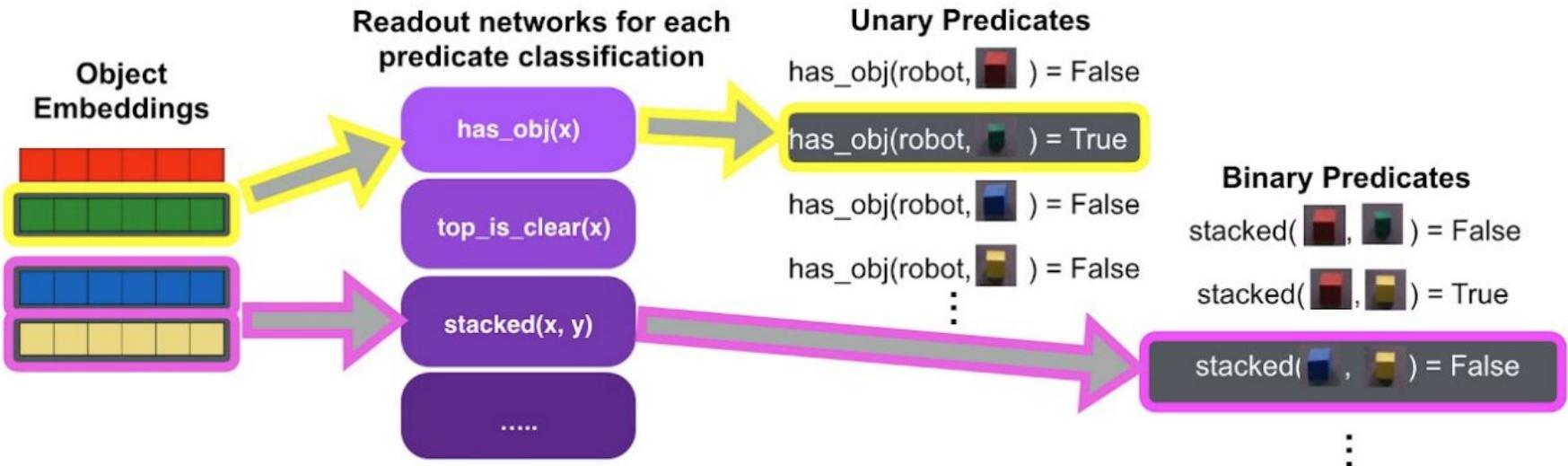
# SORNET



# SORNET - Embeddings



# SORNET - Readout Networks



# SORNET - DATA SET

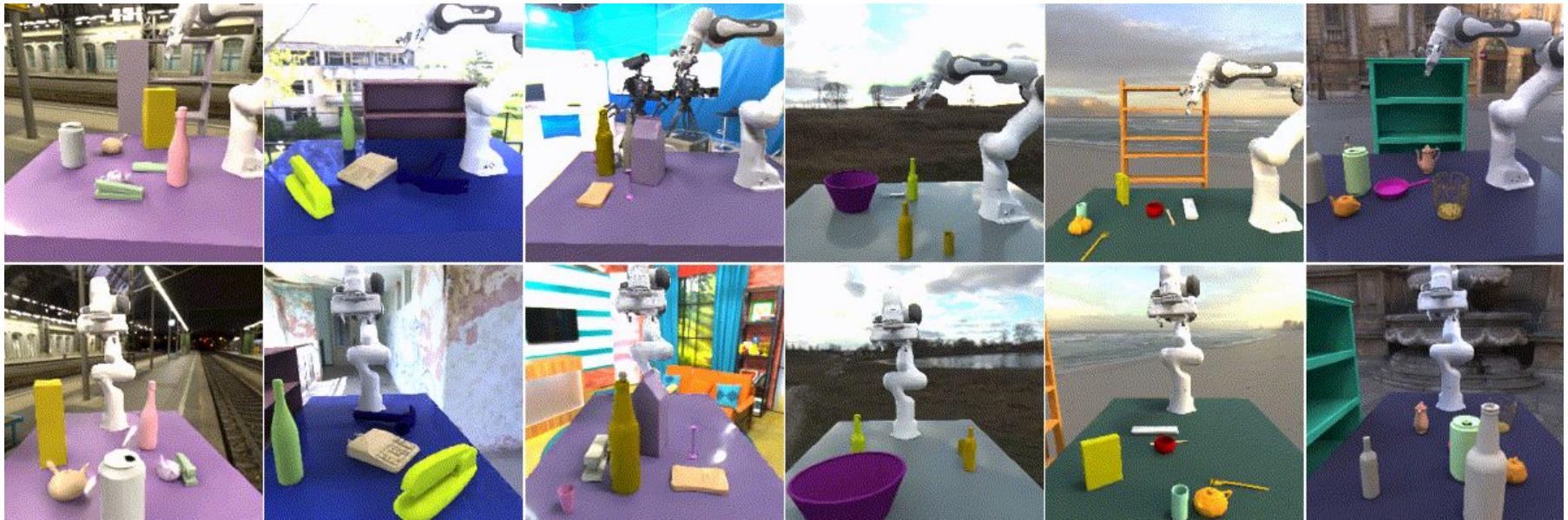


Image Source: <https://sites.google.com/view/sornet-extended>

# Examples of pre-training in Robotics

DALL-E Bot:



Initial image observation



Robot Action

This model is a robotic imaginative engine where it creates the image of the goal state which the robot will try to achieve.



Image Source: <https://www.robot-learning.uk/dall-e-bot>

# Examples of pre-training in Robotics

## DALL-E Bot architecture:

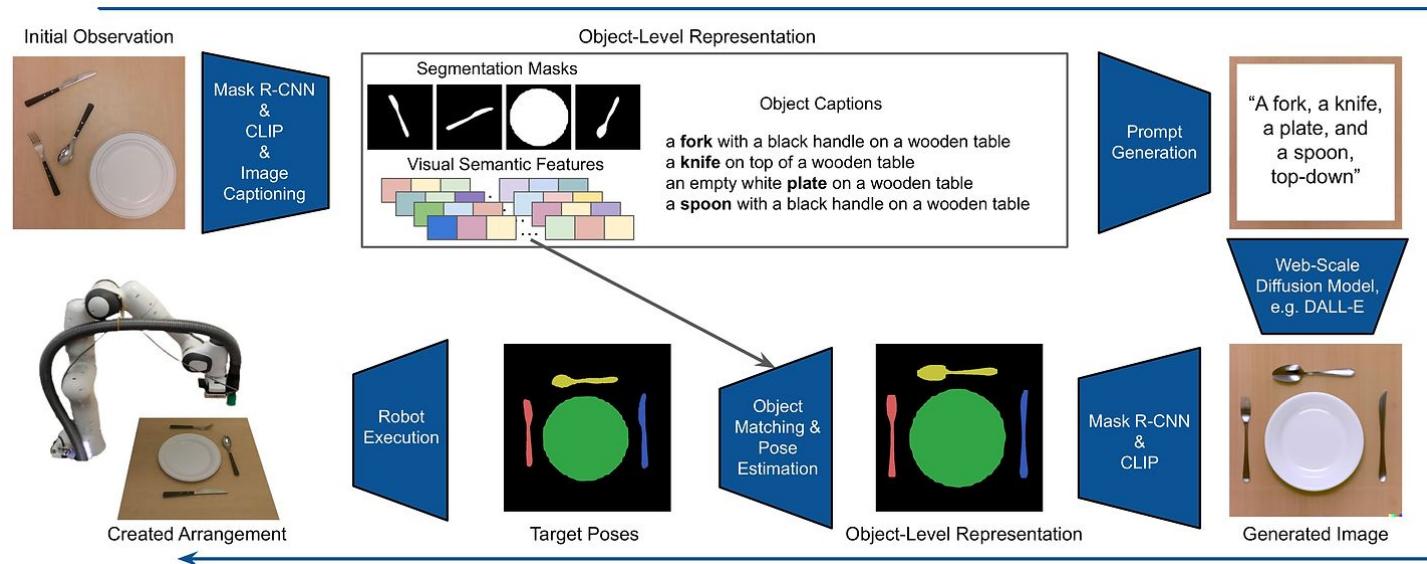


Image Source: <https://www.robot-learning.uk/dall-e-bot>

# Examples of pre-training in Robotics

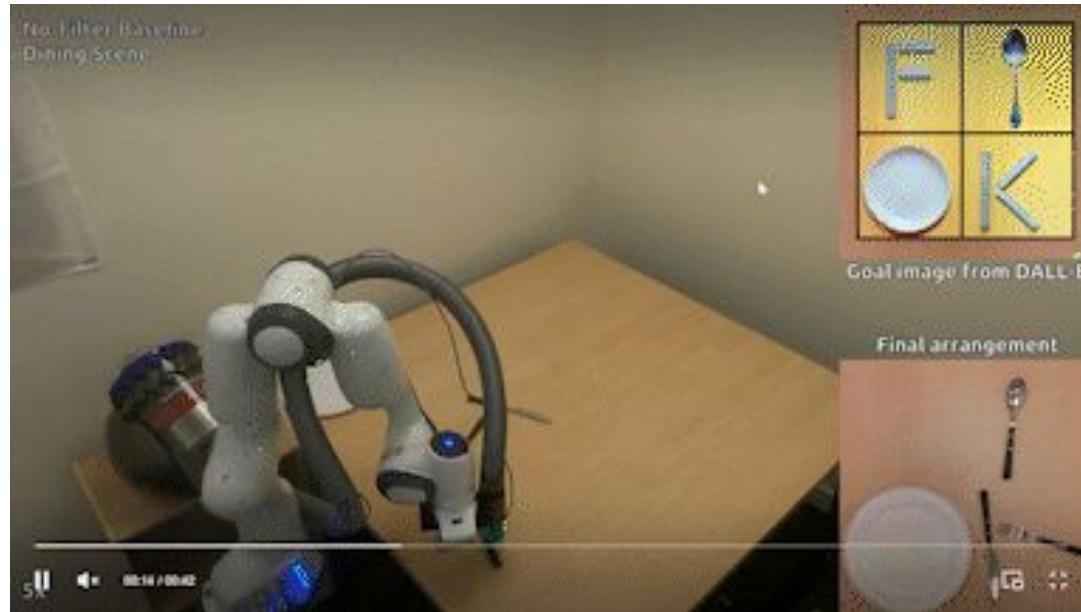
DALL-E Bot demonstration:



Gif source: <https://www.robot-learning.uk/dall-e-bot>

# Examples of pre-training in Robotics

DALL-E Bot limitations :



Gif source: <https://www.robot-learning.uk/dall-e-bot>

# Summary

---

- What are pre-trained models and foundation models
- Difference between pre-trained models and foundation models
- Pre-Training examples in NLP
  - BERT
  - GPT
  - Chat GPT

# Summary

---

- What are pre-trained models and foundation models
- Difference between pre-trained models and foundation models
- Pre-Training examples in NLP
- Pre-Training examples in CV
  - MAE (Masked Auto-encoders)
  - CLIP
  - DALL-E 1 and DALL-E 2

# Summary

---

- What are pre-trained models and foundation models
- Difference between pre-trained models and foundation models
- Pre-Training in NLP
- Pre-Training in CV
- Pre-Training in Robotics
  - R3M
  - MVP
  - SORNET
  - DALL-E Bot

DR



M M