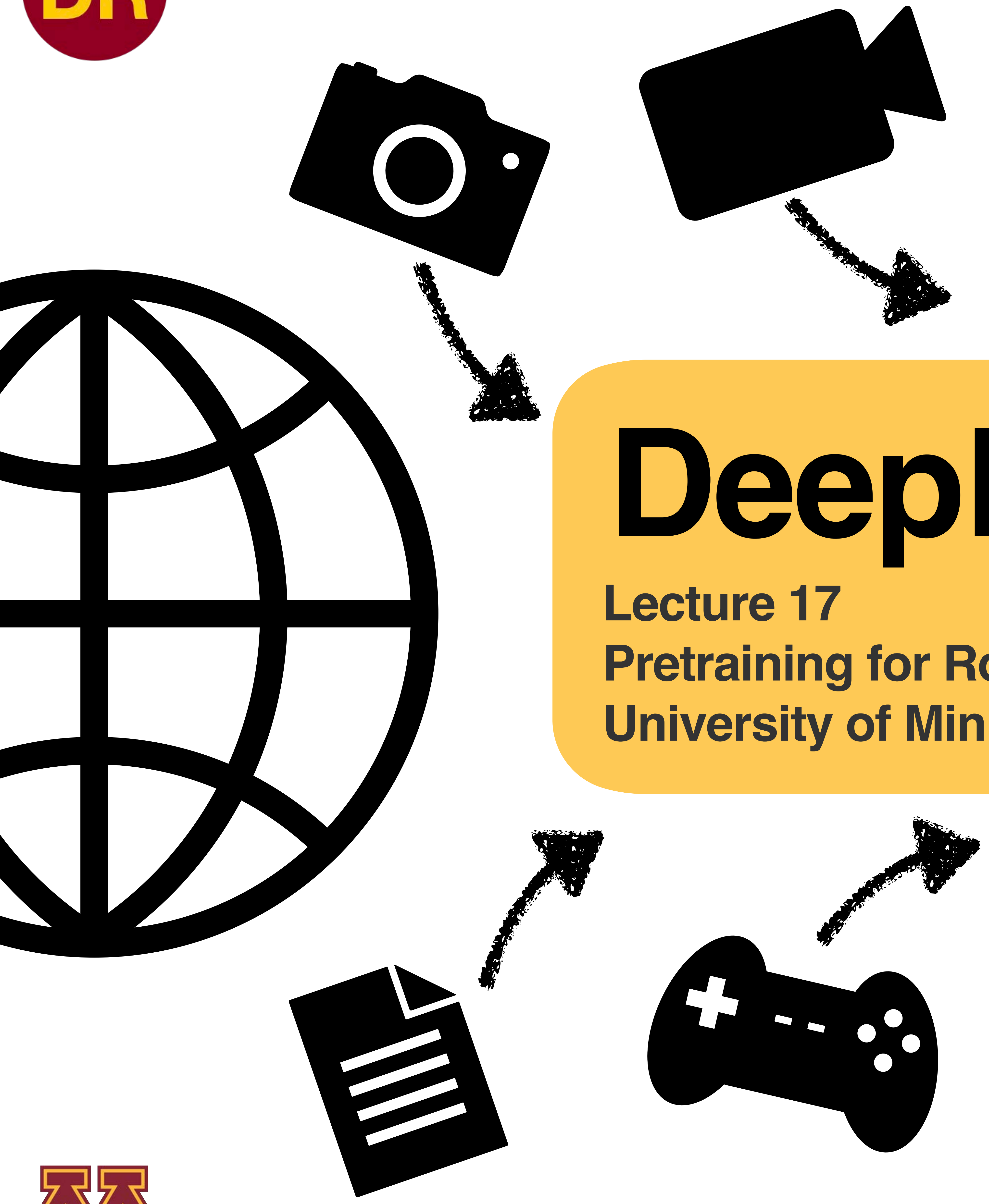
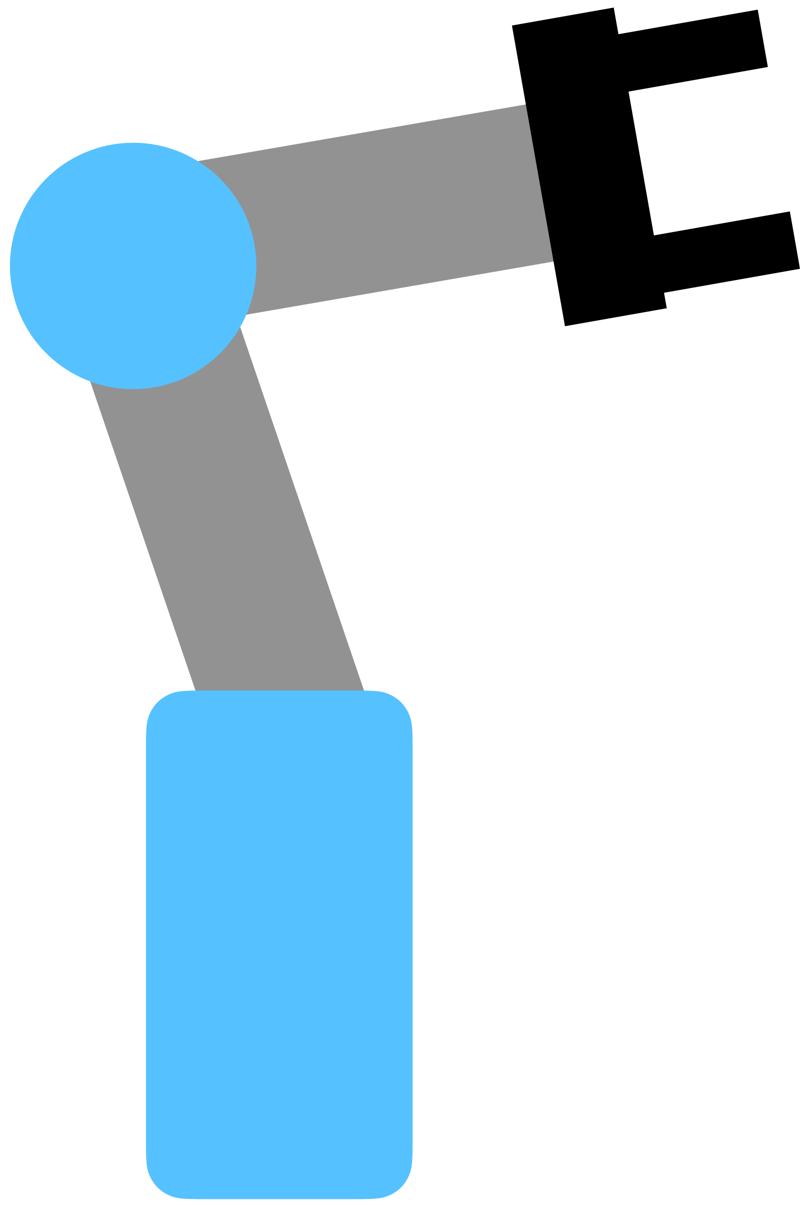


DeepRob

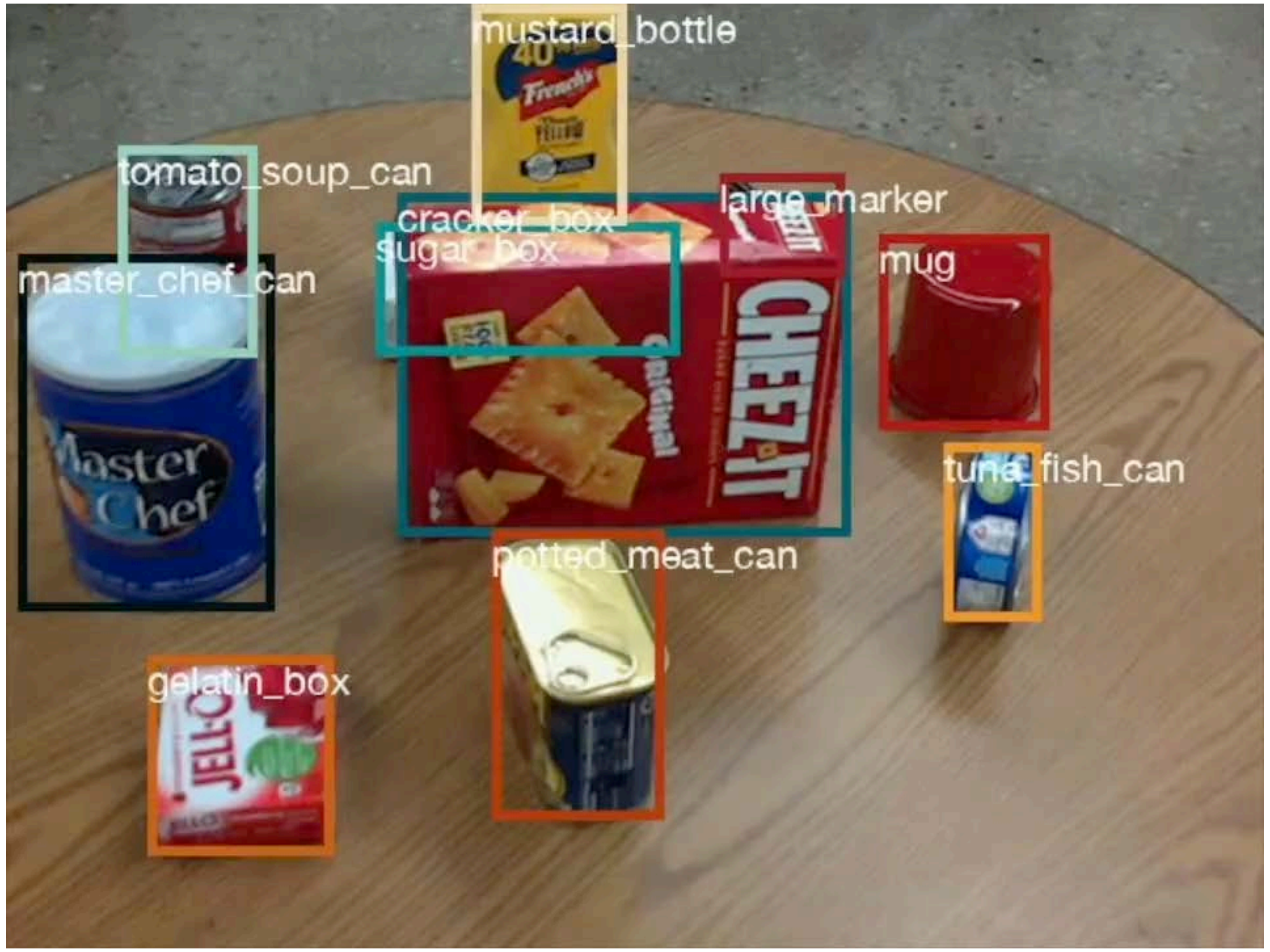
Lecture 17
Pretraining for Robot Manipulation
University of Minnesota



Slides prepared by Guanang (Shirley) Su and Karthik Desingh

Project 3 - *deadline extended*

- Instructions available on the website
  - Here: https://rpm-lab.github.io/CSCI5980-F24-DeepRob/projects/project3/
  - Uses [PROPS Detection dataset]()
- Implement CNN for classification and Faster R-CNN for detection
- Autograder will be available soon!
- **Due Monday, November 1st 11:59 PM CT**



What is representation?

Cognitive Science:

Symbolic View:

Thinking through abstract symbols.

Embodied View:

Thinking shaped by physical interactions and senses.

Computer Science:

Explicit Representations:

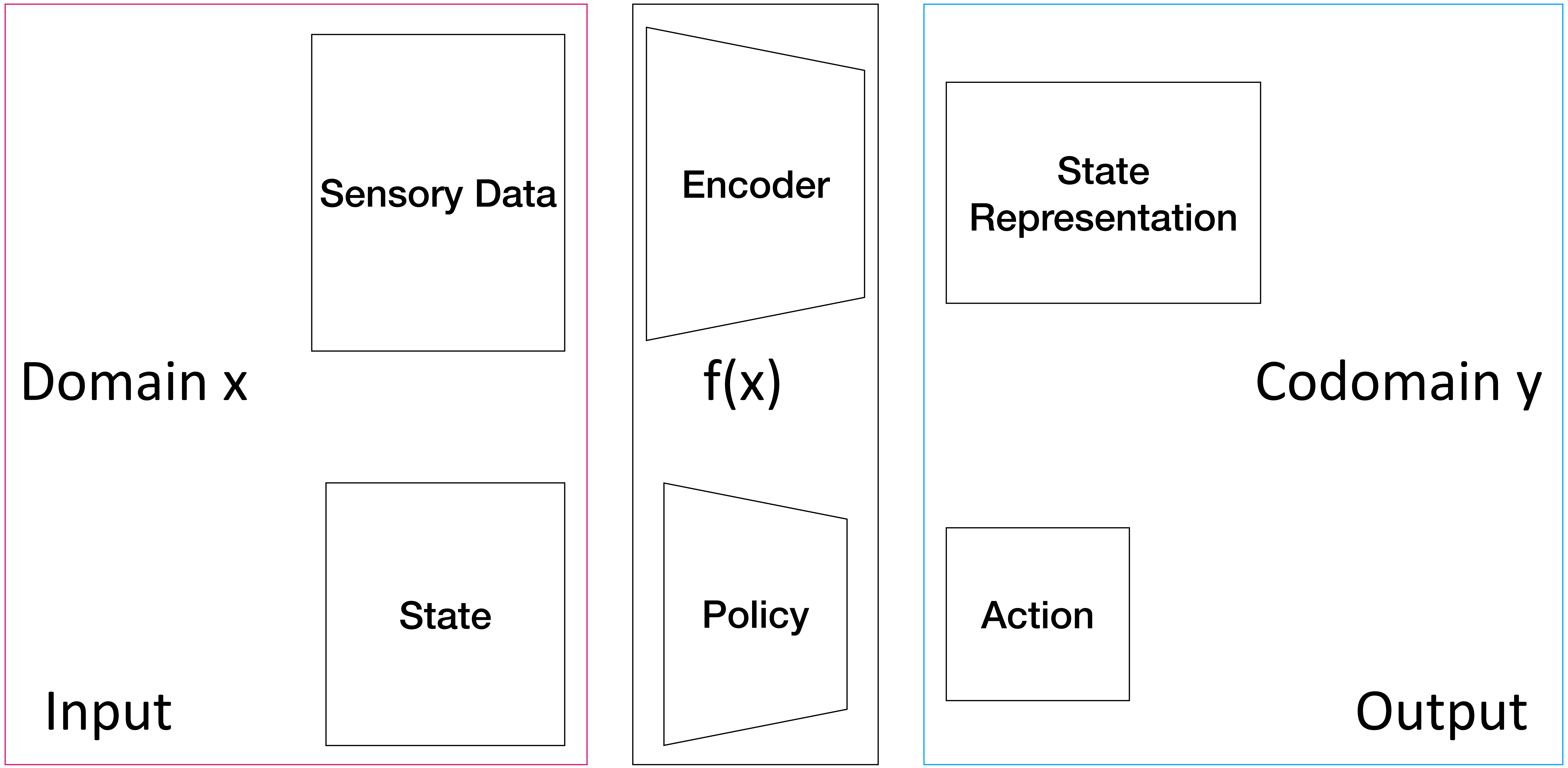
Clear, human-understandable forms like actions or labels.

Implicit Representations:

Internal data structures, often numeric, such as matrices or vectors, that encode patterns, features, or properties extracted from data.

What is representation learning?

A process of discovering features or representations from data that capture essential information for a task, such as shapes, textures, or patterns.



Types of Learning Features

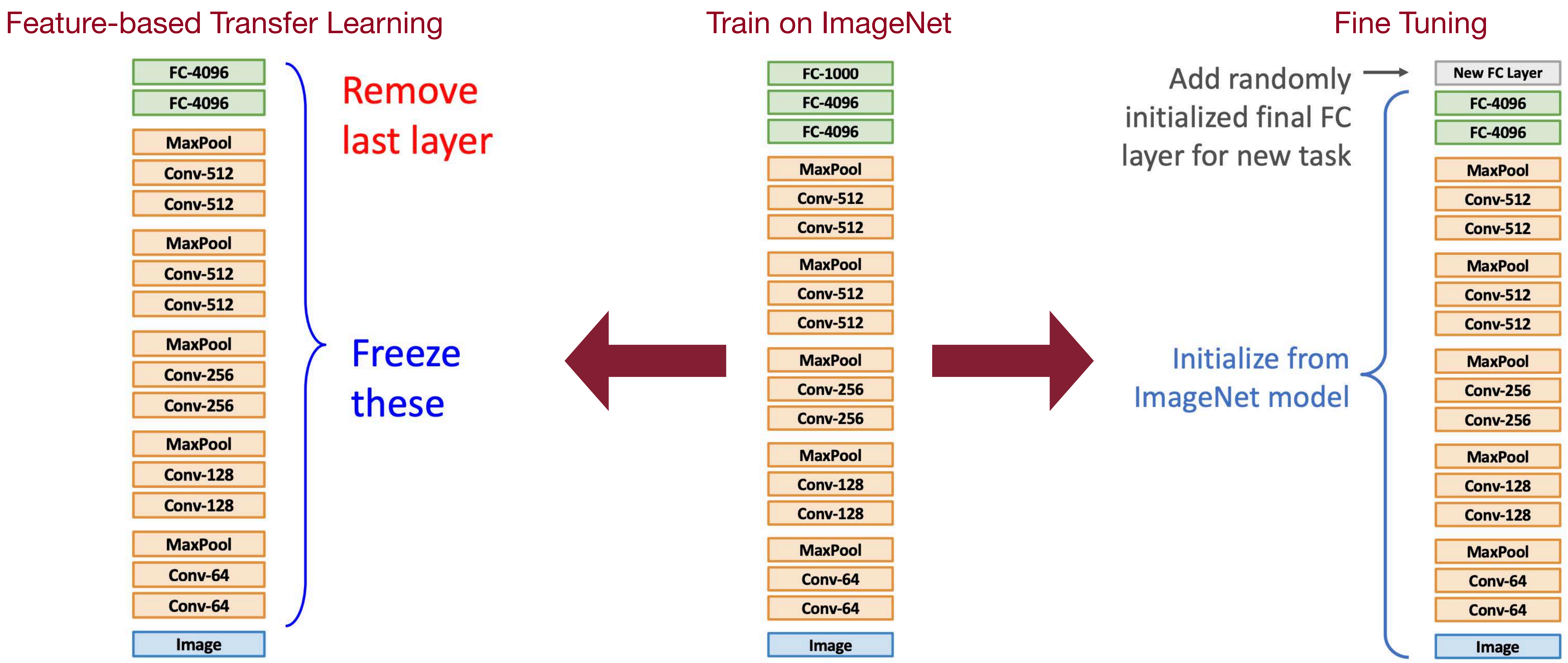
Low-Level Features (edges, textures, colors) build the base for recognizing complex objects.

High-Level Features (objects, shapes) aid in scene understanding and object segmentation.

Temporal Features capture sequences and actions, essential for video or action-based tasks.

Spatial-Relational Features help understand 3D spaces, critical for robotics.

How Transfer Learning Work?



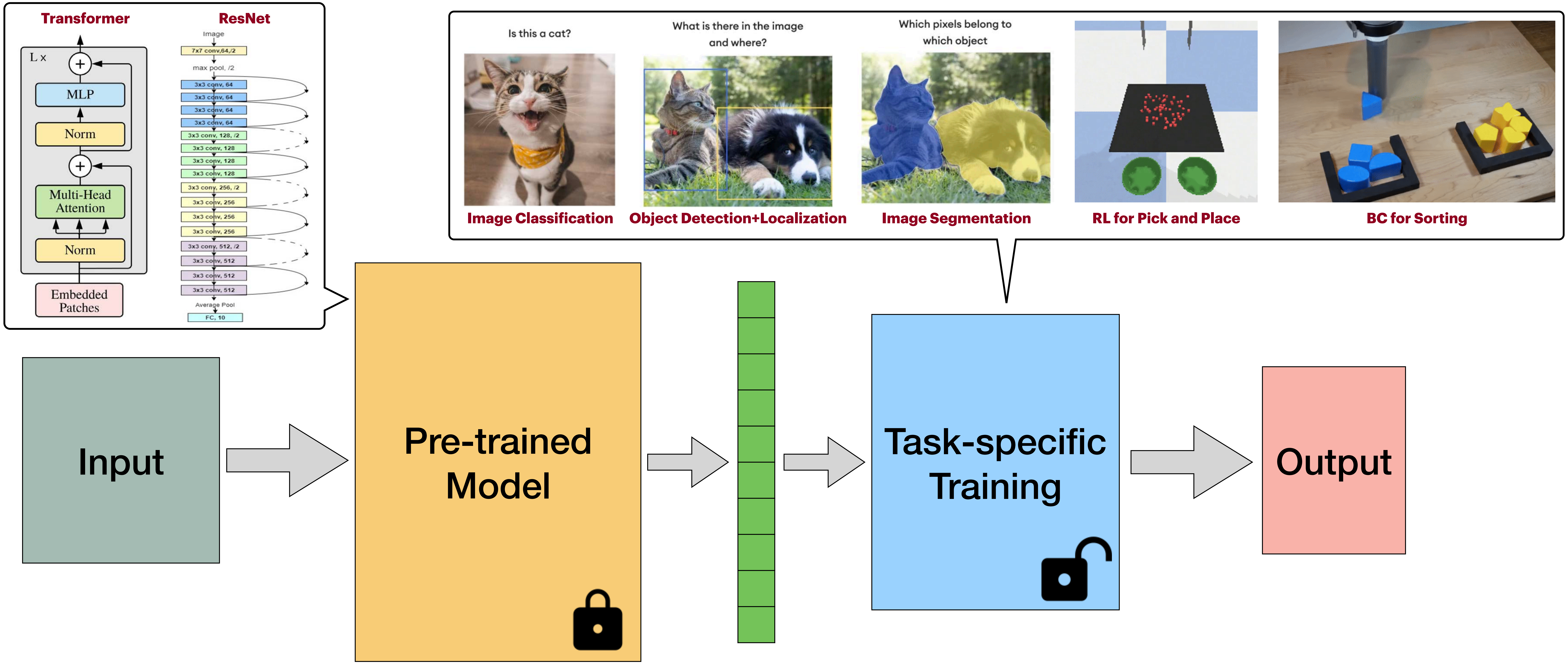
Use CNN as a feature extractor

Captured from Lec11 in CSCI5980 Fall 2024, https://rpm-lab.github.io/CSCI5980-F24-DeepRob

What is Pretraining ?

- A process of initializing a model with pre-existing knowledge before fine-tuning it on specific tasks or datasets.
- **Pretraining** leverages representation learning on large, general datasets, preparing a model to recognize these features without task-specific training.

How Does Pretrain Work?



Related work and progression of using “pretraining”

2017

2018

- GPT (Generative Pretrained Transformer)
- BERT (Bidirectional Encoder Representations from Transformers)
- Image Transformer

2019

- T5 (Text-To-Text Transfer Transformer)

2020

- SimCLR (Simple framework for contrastive learning of visual representations)
- MoCo (Momentum Contrast)
- ViT (Vision transformer)

- CURL (Contrastive Unsupervised Representations for Reinforcement Learning)

2021

- SORNET

- MAE (Masked Autoencoder)
- CLIP (Contrastive Language Image Pretraining)
- DALL-E
- DINO

2022

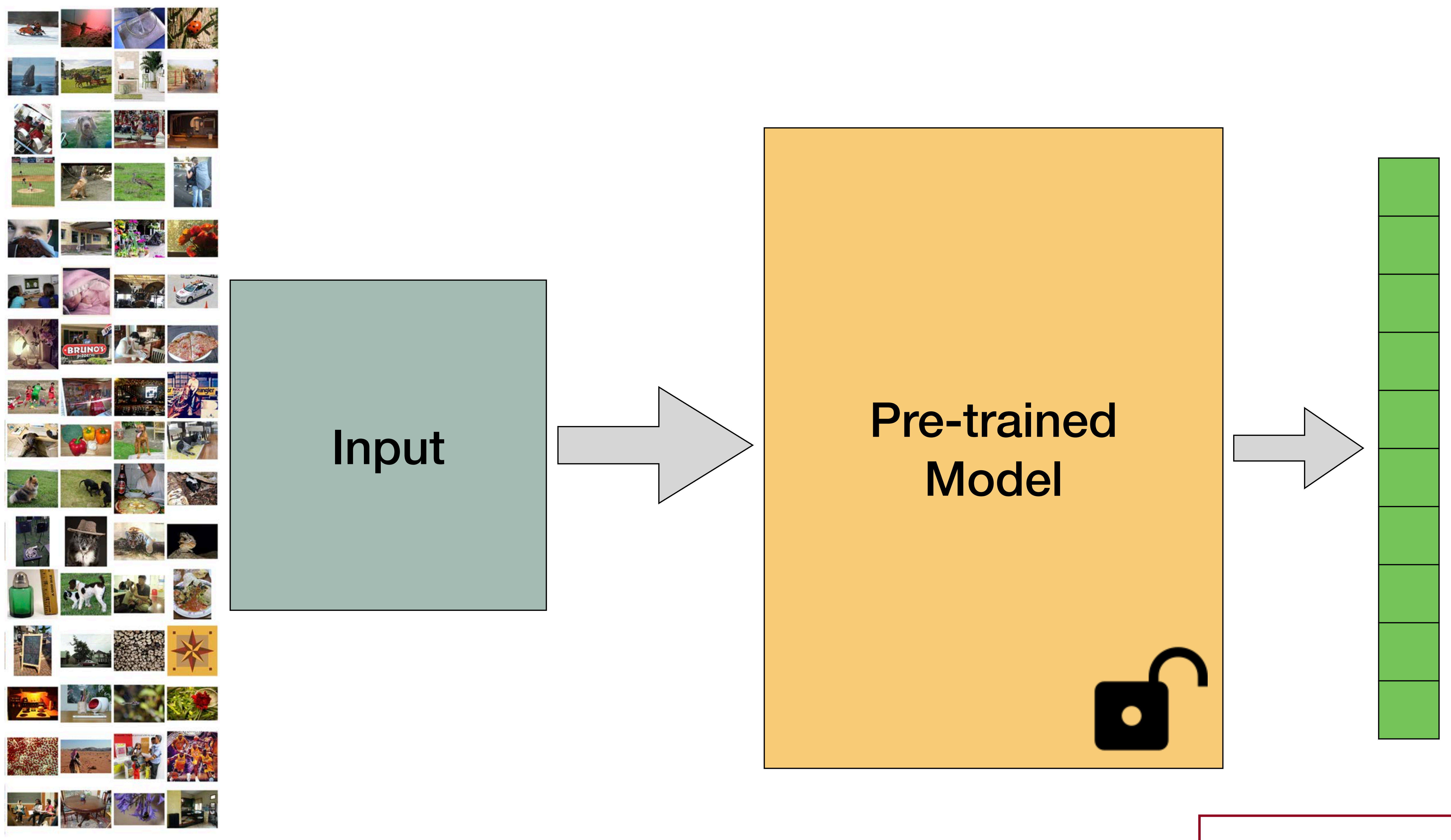
- PerAct: Multi-task transformer
- R3M (Robust Representations for Robotic Manipulation)

2023

- SUGAR (3D Pre-training for Robotics)
- 3D-MVP (3D Multiview Pretraining)

Pretraining in Computer Vision

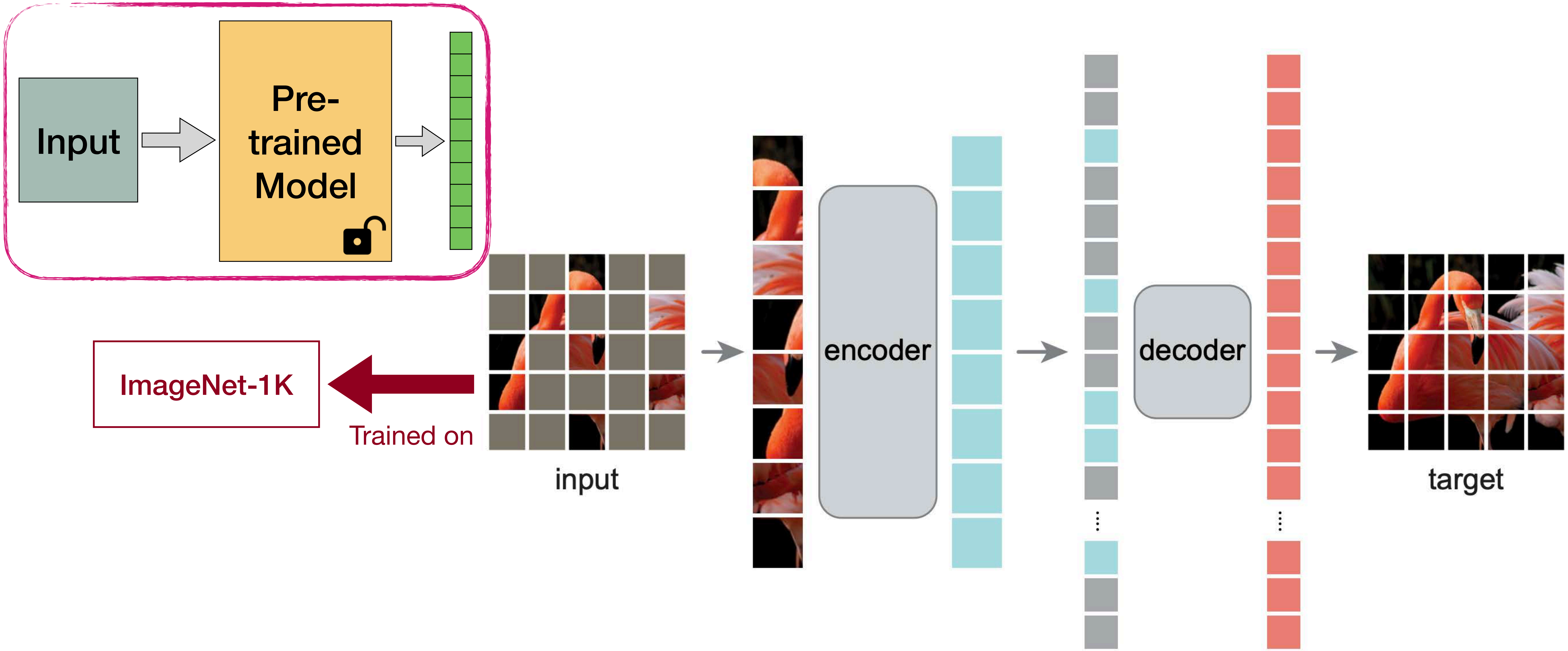
Pretraining Process



Visual Input

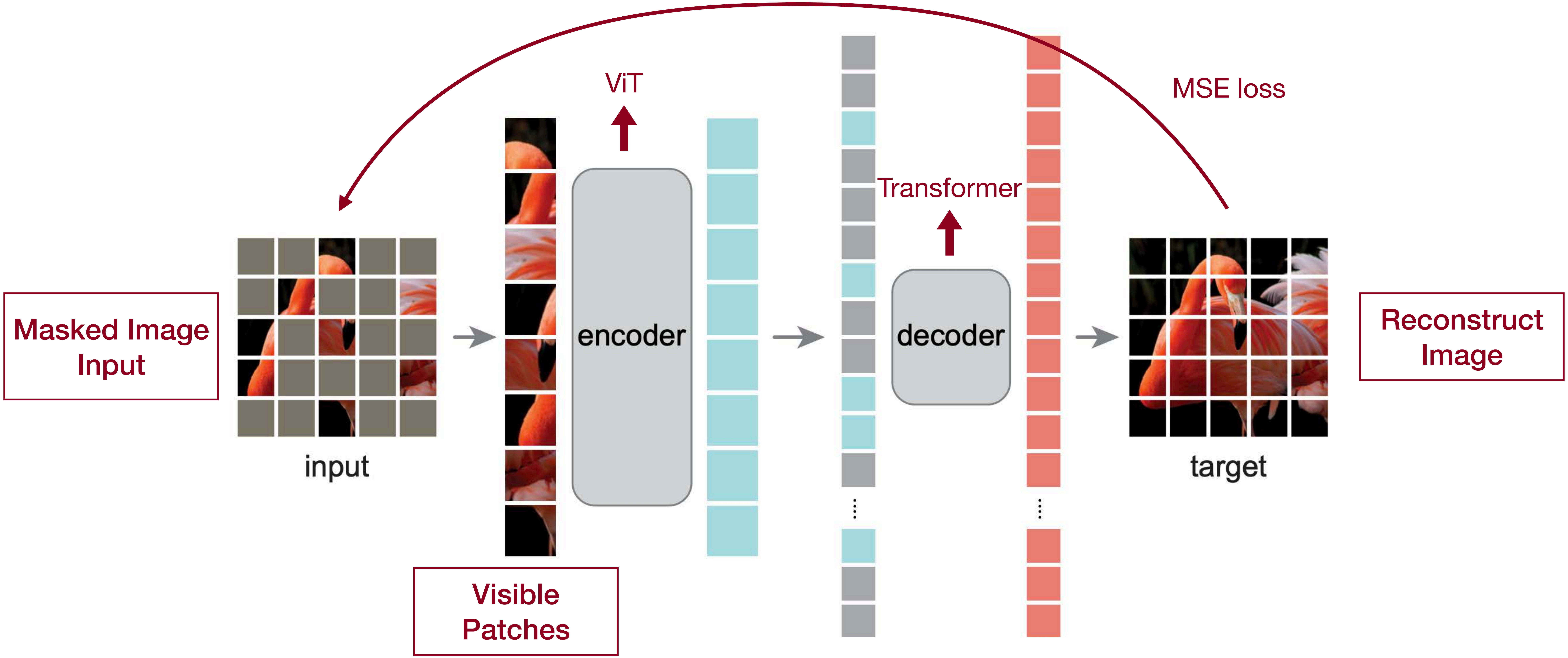
Goal: learn good representations

MAE (Masked Autoencoders)



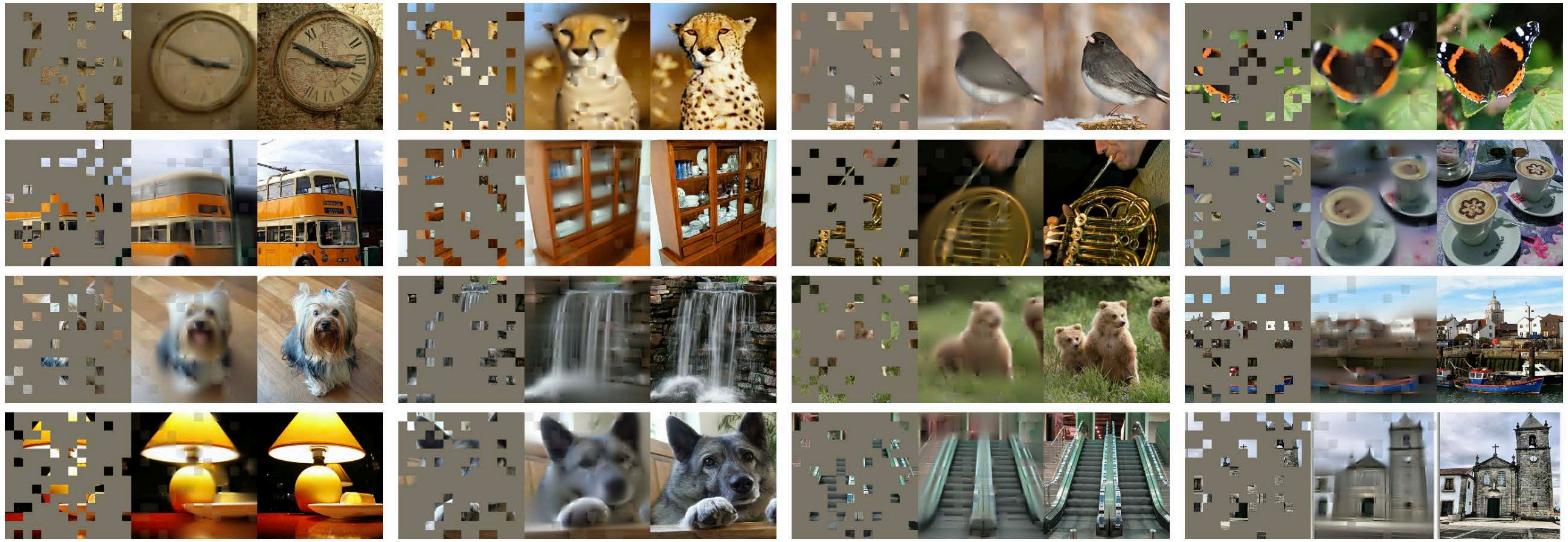
Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 16000-16009).

MAE Architecture



He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 16000-16009).

MAE Results



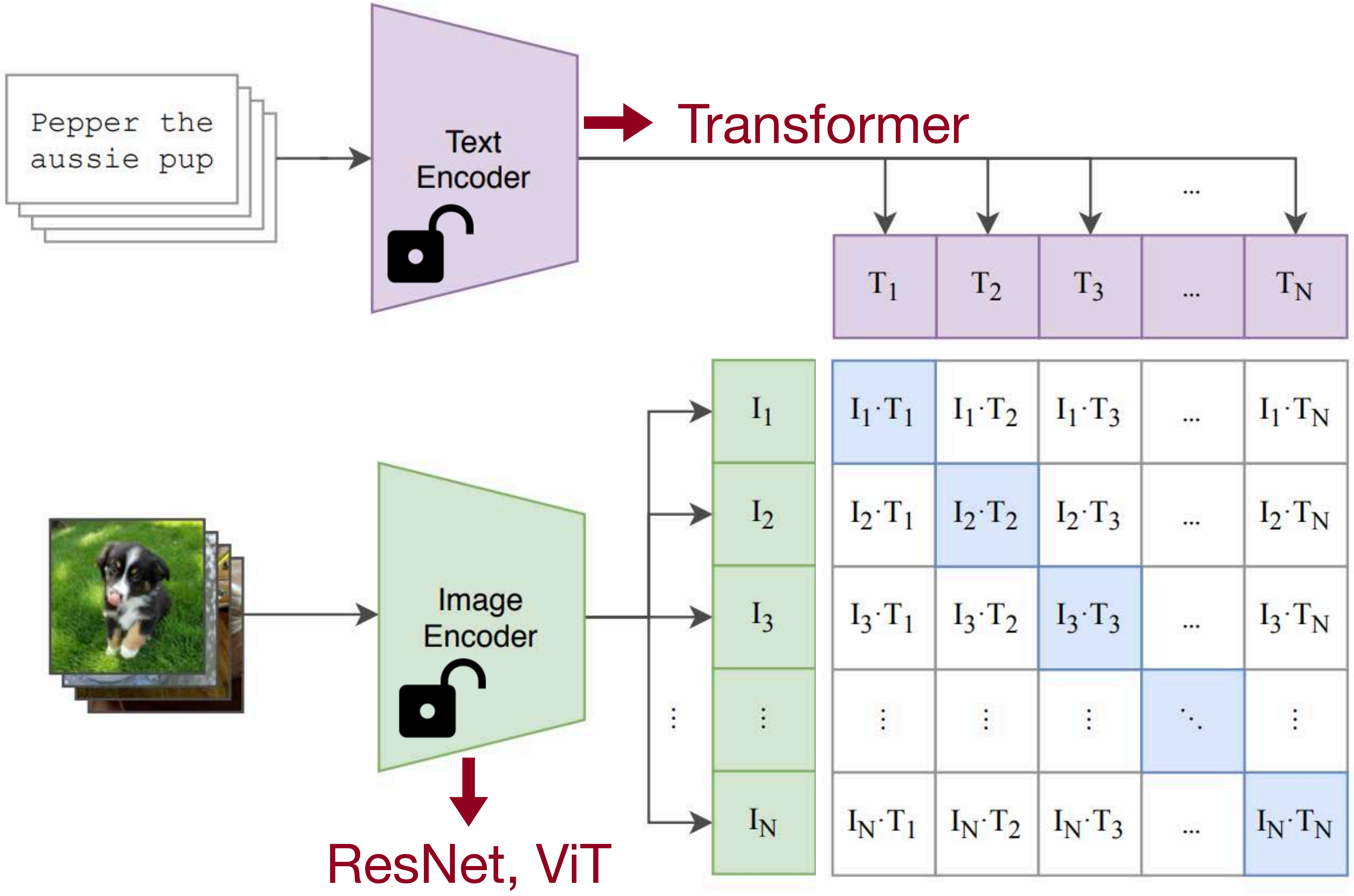
Example results on ImageNet validation dataset - 80%

Example results on COCO dataset

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 16000-16009).

CLIP (Contrastive Language-Image Pre-Training)

- **Contrastive Pre-training**



Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In International Conference on Machine Learning (ICML) 2021, Vol. 139. 8748–8763.

Contrastive Learning

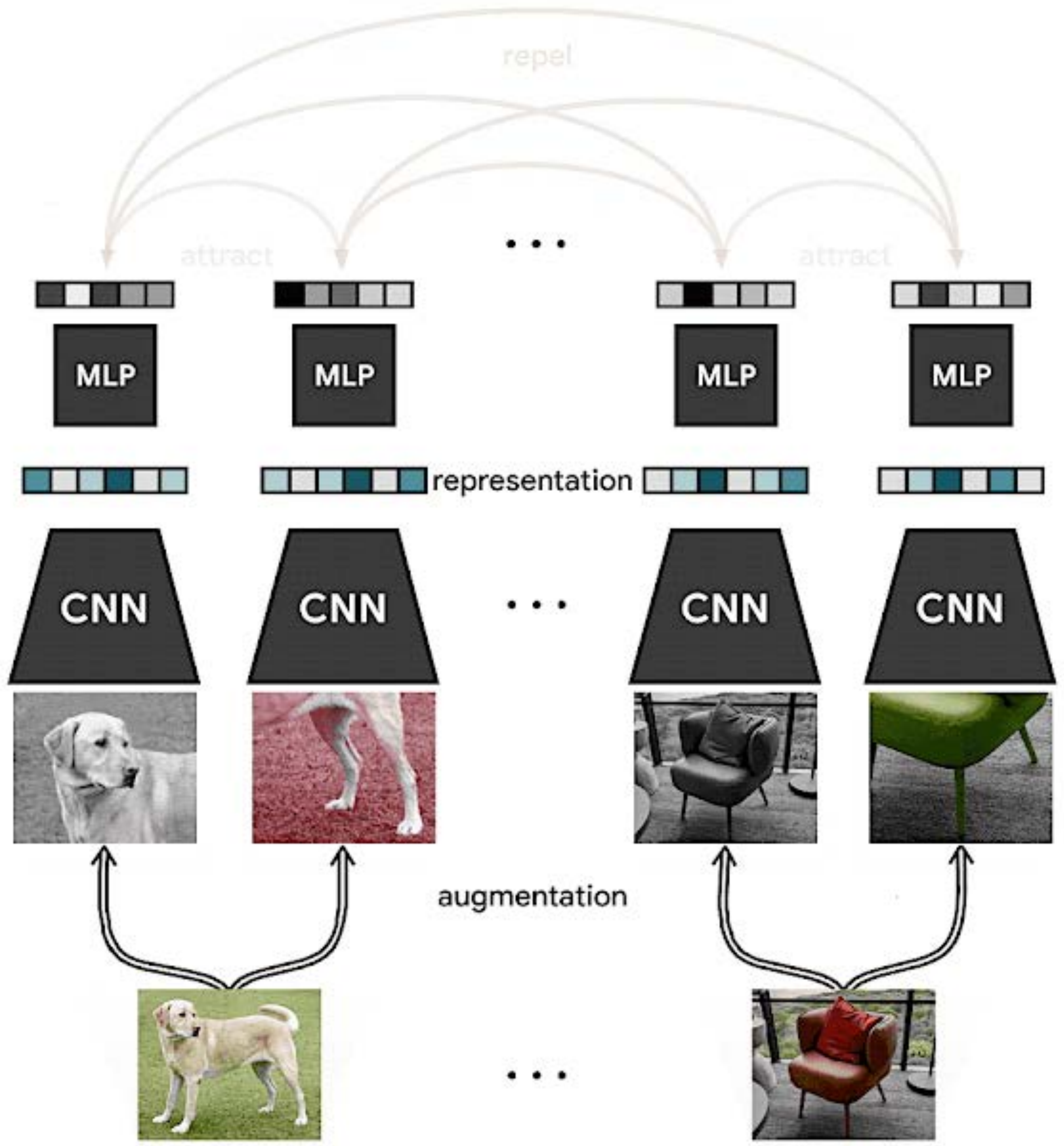
Contrastive learning is an approach that focuses on extracting meaningful representations by contrasting positive and negative pairs of instances.

Input → Encoder Network → S_high → Projection Network → S_low

Encoder Network: "Feature Extractor"

Projection Network: "Refine Feature"

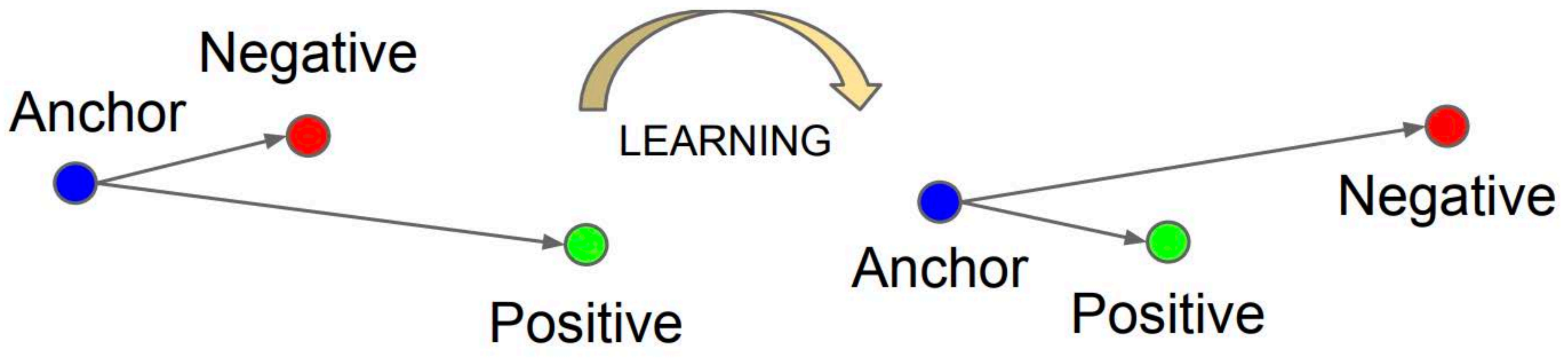
Why?
- Dimensional Reduction
- Normalization and size Control
- Improving Discriminative Power
- Avoiding Collapse



Source from SimCLR - https://github.com/google-research/simclr

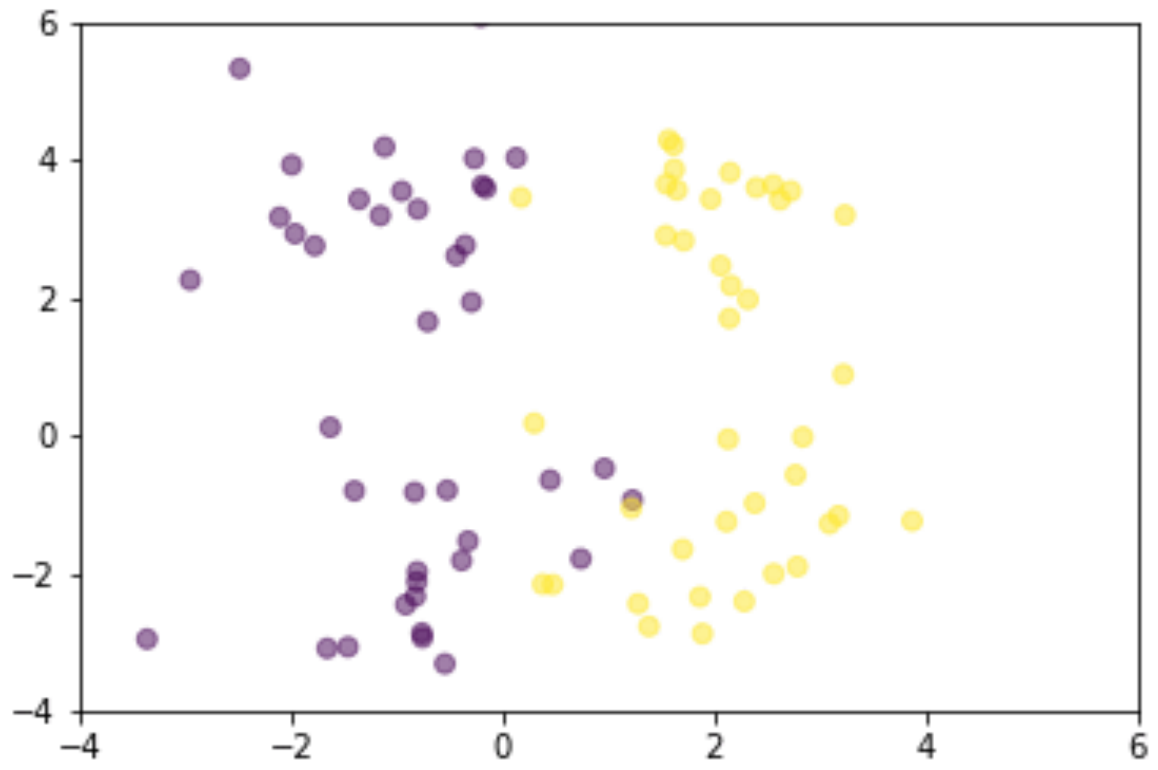
Contrastive Loss

Triplet Loss



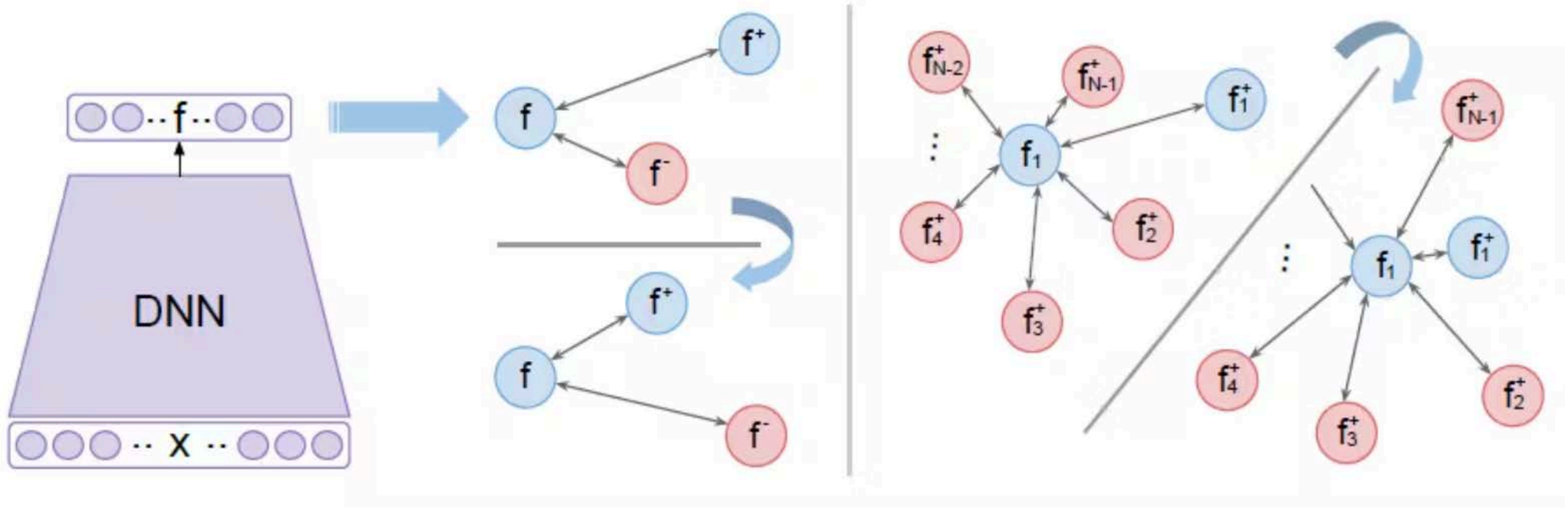
Calculate the squared Euclidean distance matrix based on the following equation:

$$\mathcal{L}_{\text{tri}}^{m}(x, x^{+}, x^{-}; f) = \max\left(0, \|f - f^{+}\|_2^2 - \|f - f^{-}\|_2^2 + m\right)$$



Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 815-823).

N-pair Loss



Multi-Class N-pair loss (Sohn 2016)

N-1 negative example & 1 positive example

$$\mathcal{L}(\{x, x^{+}, \{x_i\}_{i=1}^{N-1}\}; f) = \log\left(1 + \sum^{N-1} \exp(f^{\top} f_i - f^{\top} f^{+})\right)$$

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, *29*.

Contrastive Loss

InfoNCE loss (Information Noise-Contrastive Estimation loss)

Setup:

f_A: The feature vector for the anchor (A)

f_P: The feature vector for the positive sample (P)

f_{N_i}: The feature vector for the i-th negative sample (N)

Steps:

1. Dot Products (Similarities): Compute the similarity between:
   - Anchor and Positive Anchor and Positive: $\text{sim}(A, P) = f_A^\top f_P$
   - Anchor and each Negative Anchor and each Negative: $\text{sim}(A, N_i) = f_A^\top f_{N_i}$ for each N_i
2. InfoNCE Loss Formula: The InfoNCE loss for a single anchor-positive pair is:

   $$L = -\log \frac{\exp(\text{sim}(A, P))}{\exp(\text{sim}(A, P)) + \sum_{i=1}^{N} \exp(\text{sim}(A, N_i))}$$

   This formula maximizes the similarity between the anchor and positive pair while minimizing the similarity between the anchor and all negative pairs.

Anchor-Positive Similarity: $\text{sim}(A, P) = f_A^\top f_P = 2.5$

Anchor-Negative Similarities:

$\text{sim}(A, N_1) = f_A^\top f_{N_1} = 0.5$, $\text{sim}(A, N_2) = f_A^\top f_{N_2} = 1.0$, $\text{sim}(A, N_3) = f_A^\top f_{N_3} = 0.2$

1. Calculating exponentials for each similarity:

$\exp(\text{sim}(A, P)) = \exp(2.5) \approx 12.18$

$\exp(\text{sim}(A, N_1)) = \exp(0.5) \approx 1.65$

$\exp(\text{sim}(A, N_2)) = \exp(1.0) \approx 2.72$

$\exp(\text{sim}(A, N_3)) = \exp(0.2) \approx 1.22$

2. Sum of exponentials:

$\text{Total} = 12.18 + 1.65 + 2.72 + 1.22 = 17.77$

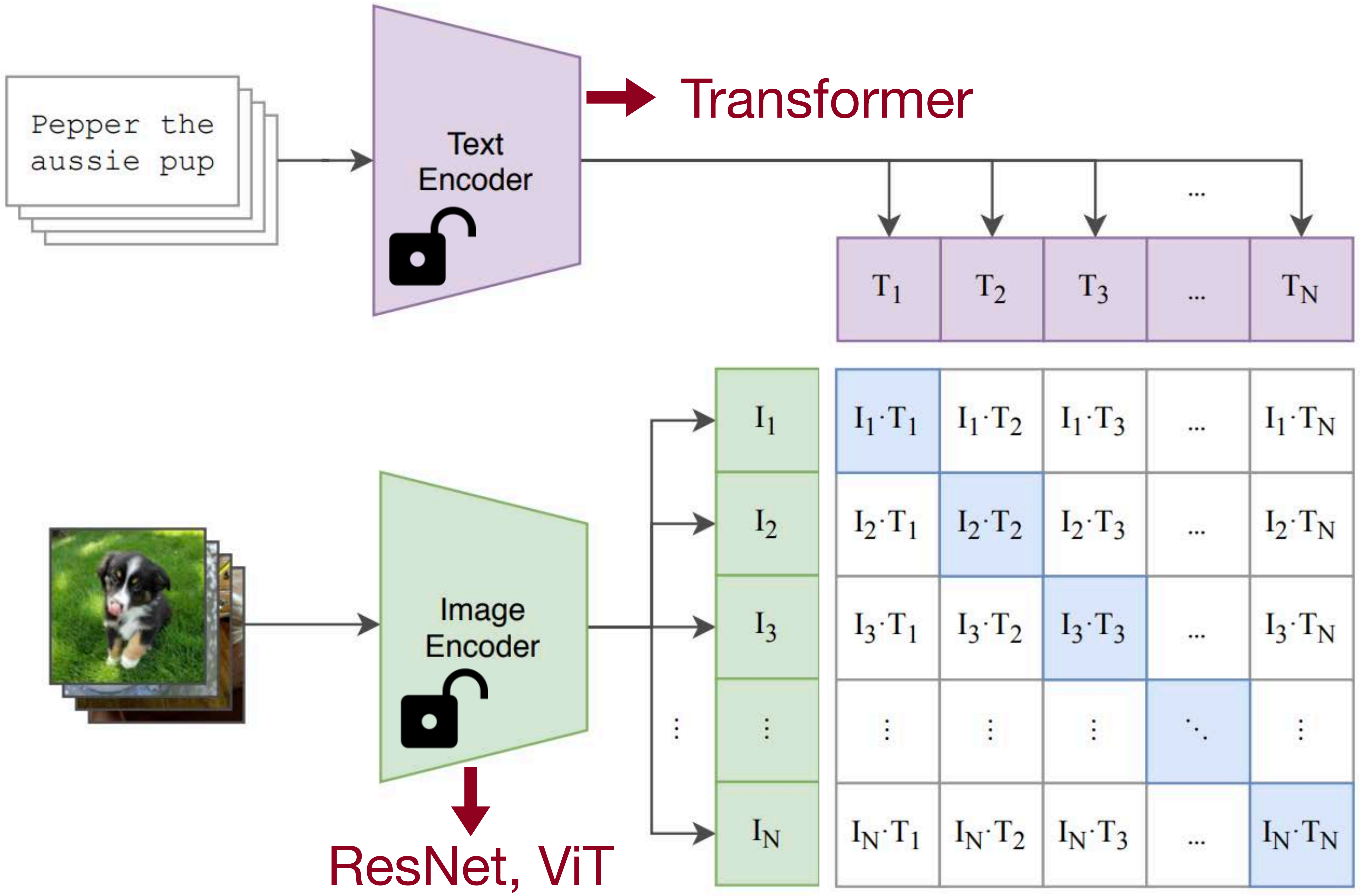
3. The InfoNCE loss calculation:

$$L = -\log\left(\frac{12.18}{17.77}\right)$$

$$L = -\log(0.686) \approx 0.376$$

CLIP - Pretraining

- **Contrastive Pre-training**

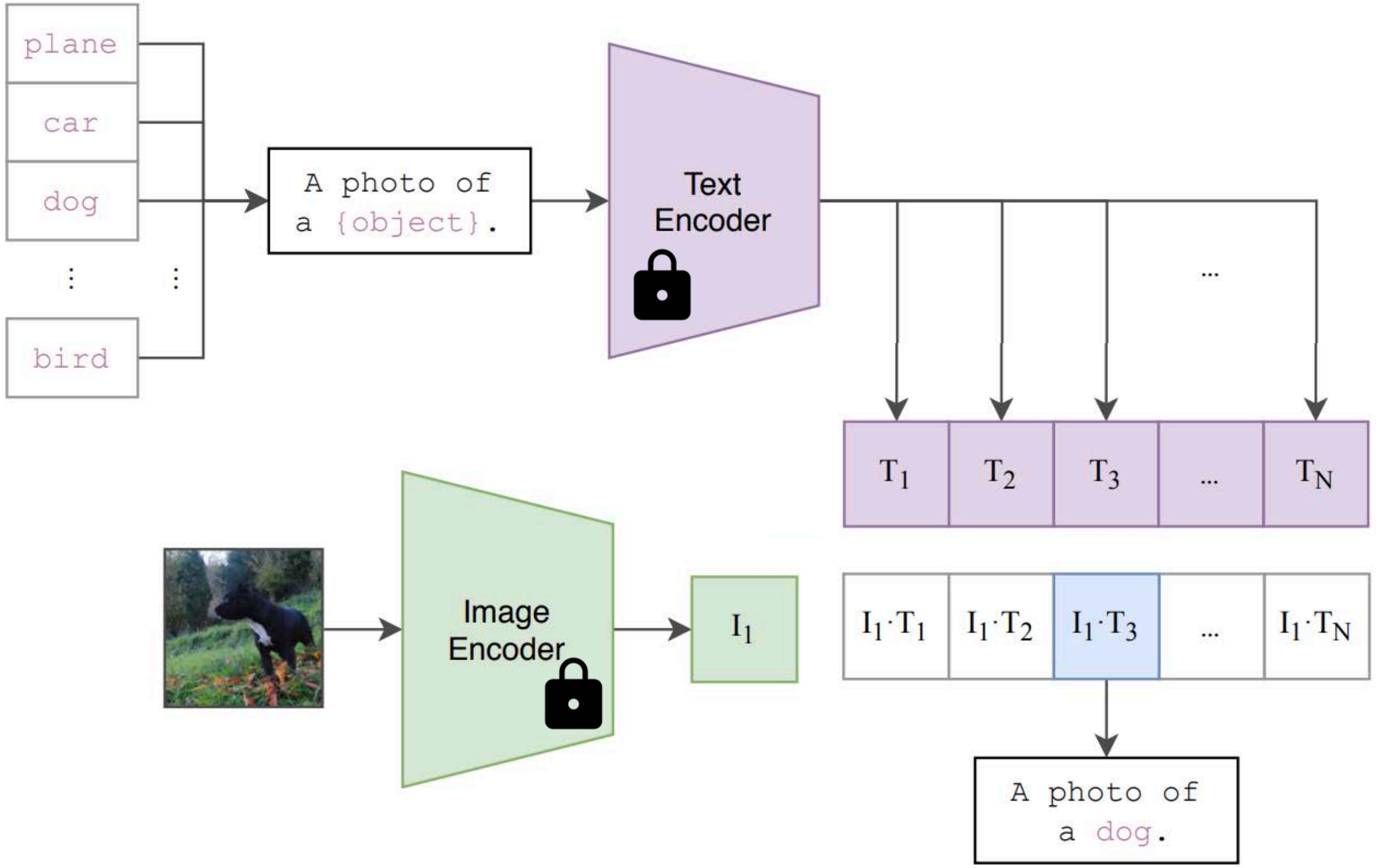


Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In International Conference on Machine Learning (ICML) 2021, Vol. 139. 8748–8763.

CLIP - Usage

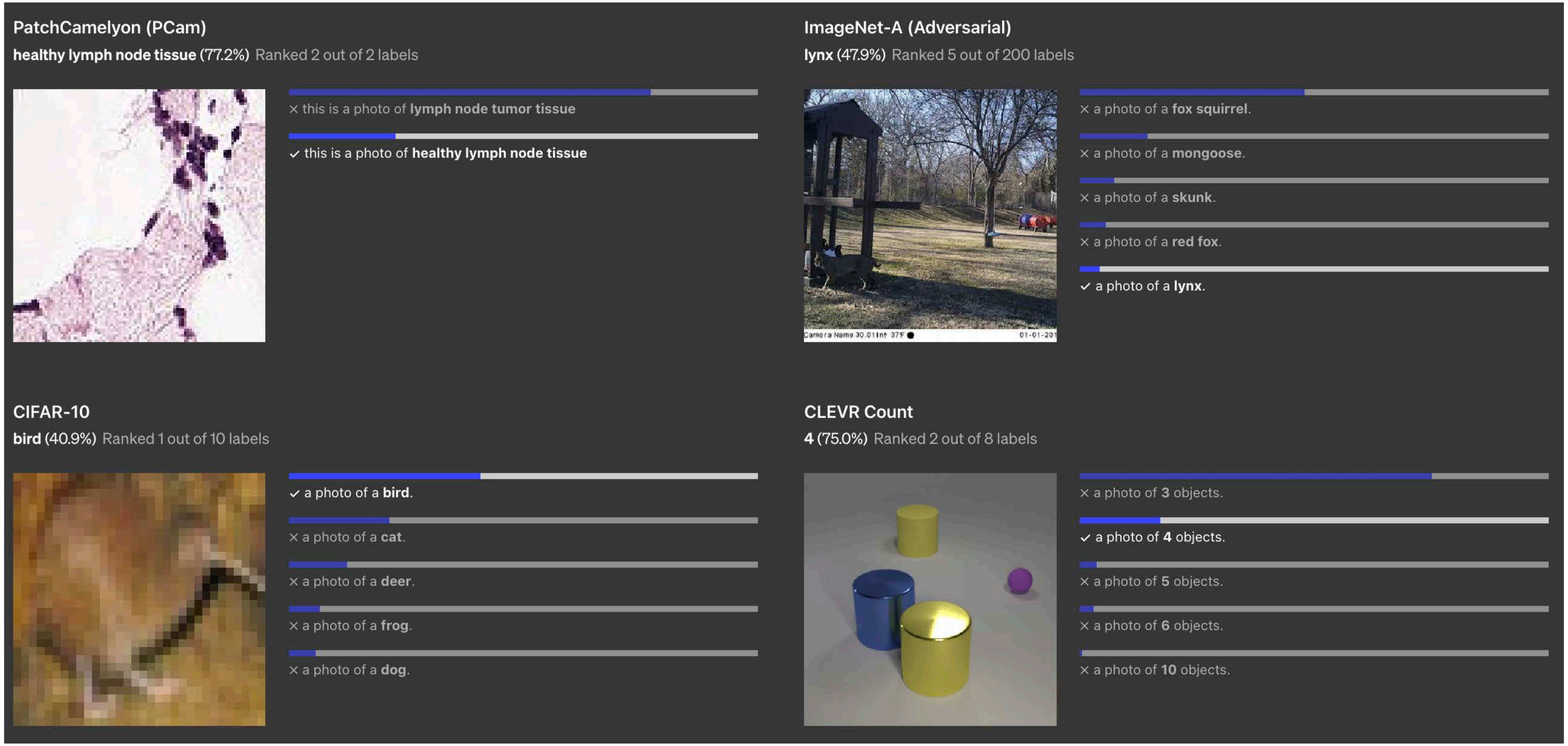
Create dataset classifier from label text

Use for Zero-shot Prediction



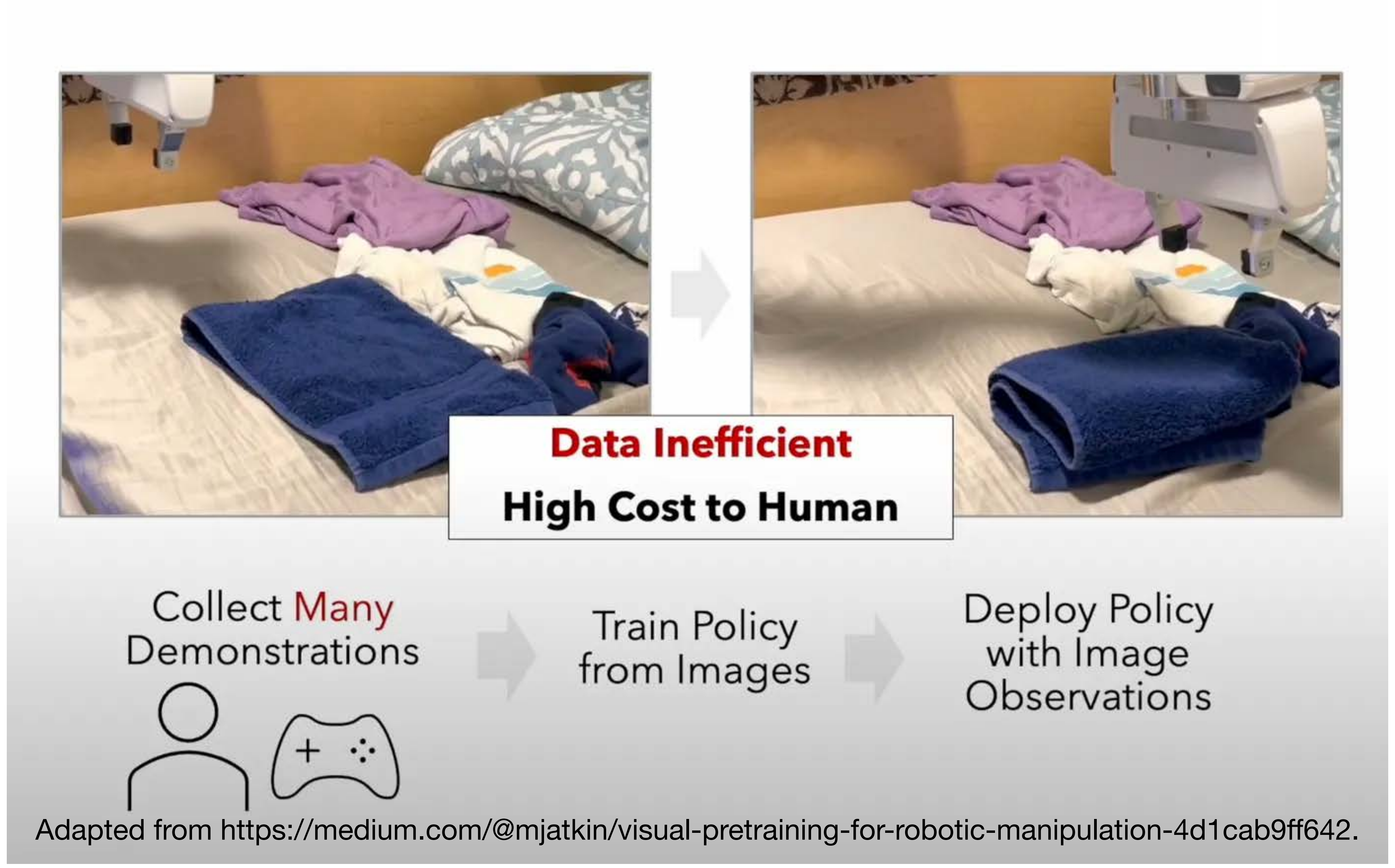
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In International Conference on Machine Learning (ICML) 2021, Vol. 139. 8748–8763.

CLIP - Results



Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In International Conference on Machine Learning (ICML) 2021, Vol. 139. 8748–8763.

Pretraining in Robotics



Robotics Pretrain Dataset

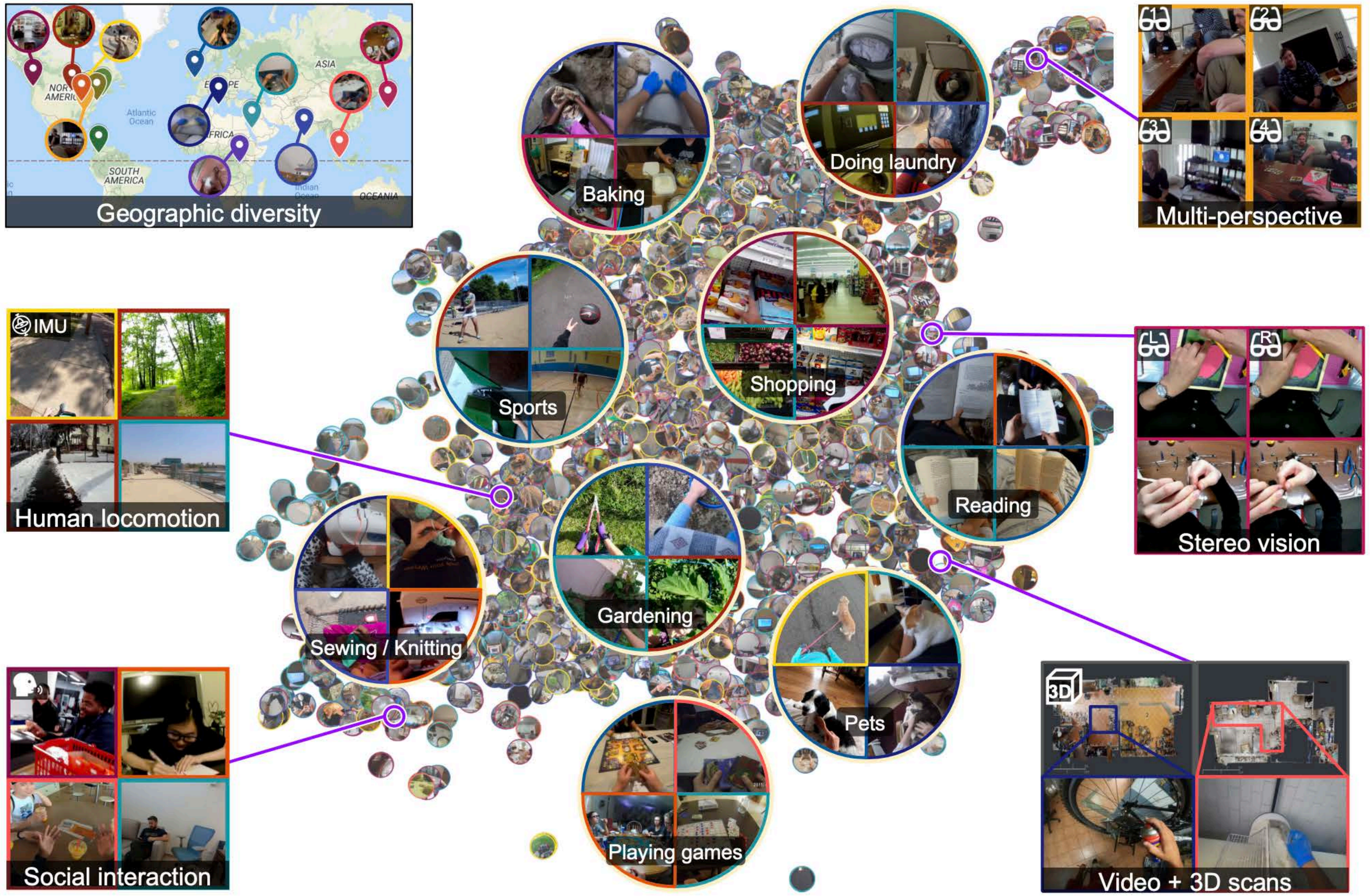
Ego4D

Ego = egocentric

4D = 3D spatial + temporal information

3,670 hours of daily life activity video

hundreds of scenarios



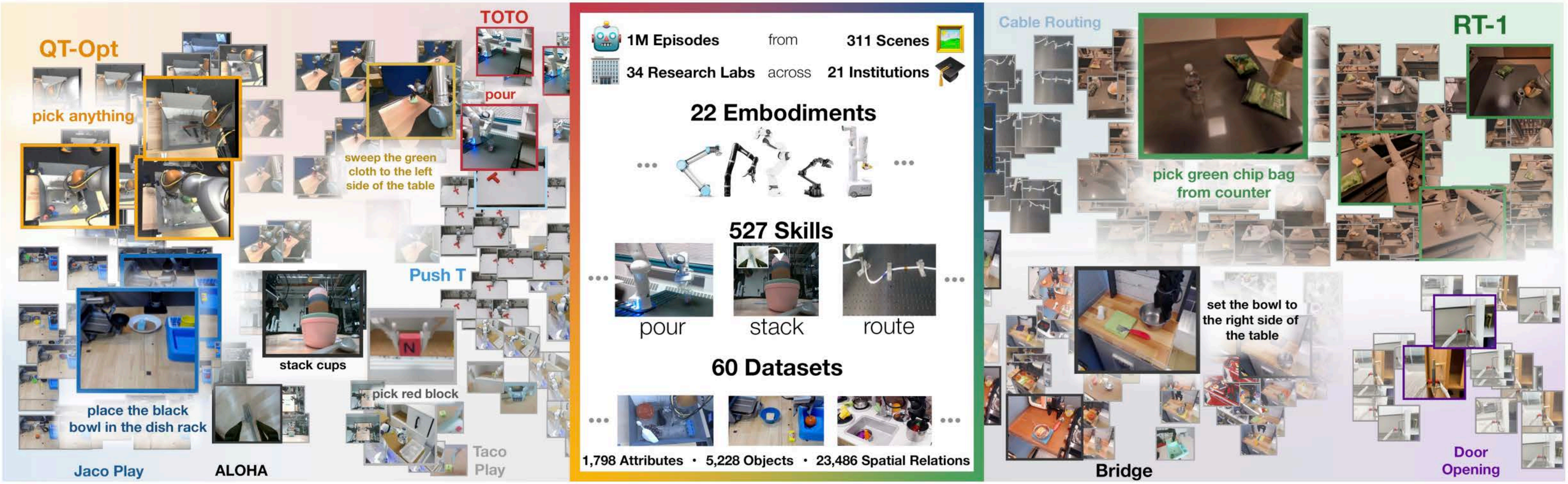
Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., ... & Memisevic, R. (2017). The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 5842-5850).

Robotics Pretrain Dataset

Open X-Embodiment

22 different robots

527 skills (160266 tasks)



O'Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., ... & Fei-Fei, L. (2023). Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*.

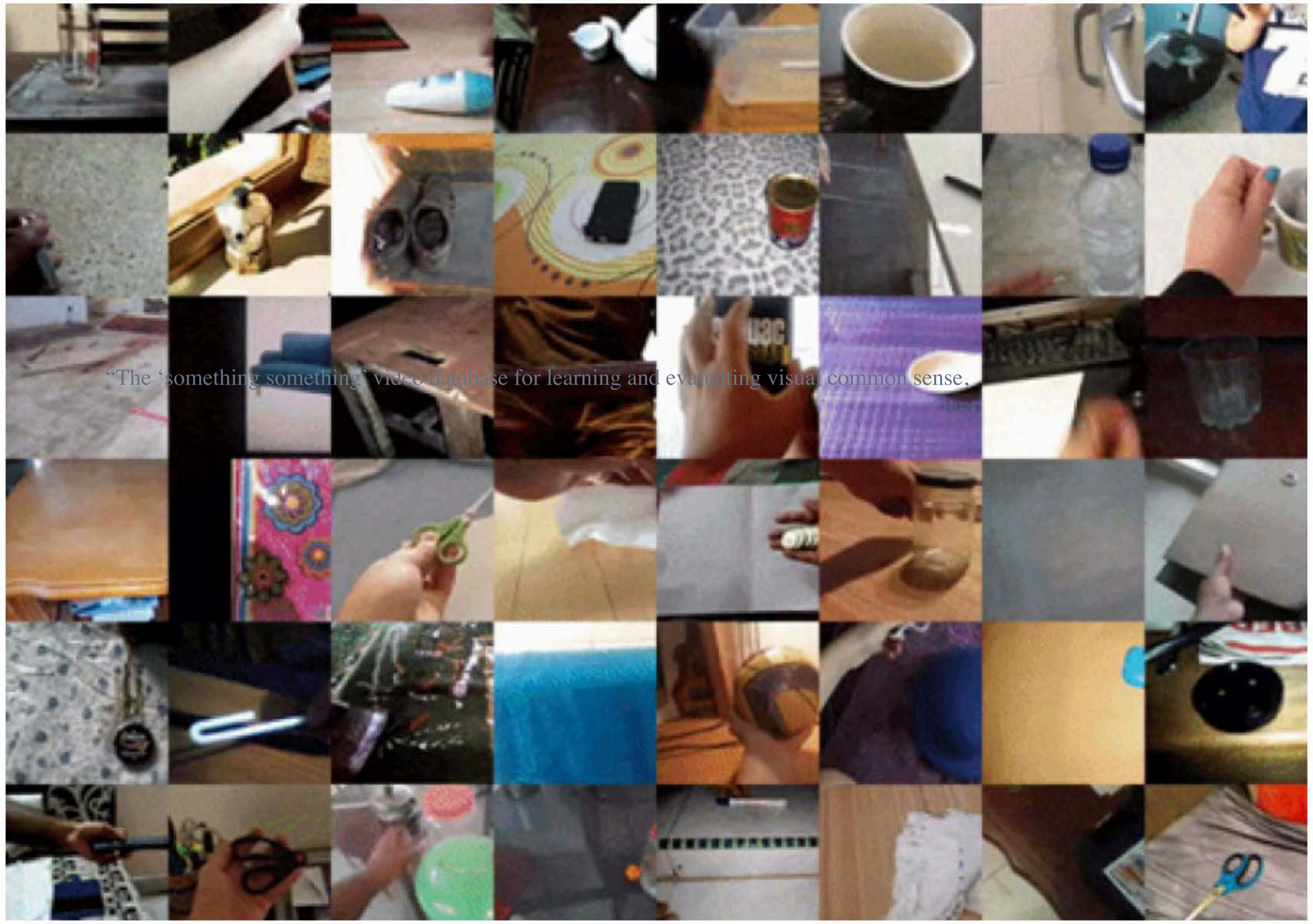
Robotics Pretrain Dataset

Something-something-v2

220,847 short video clips

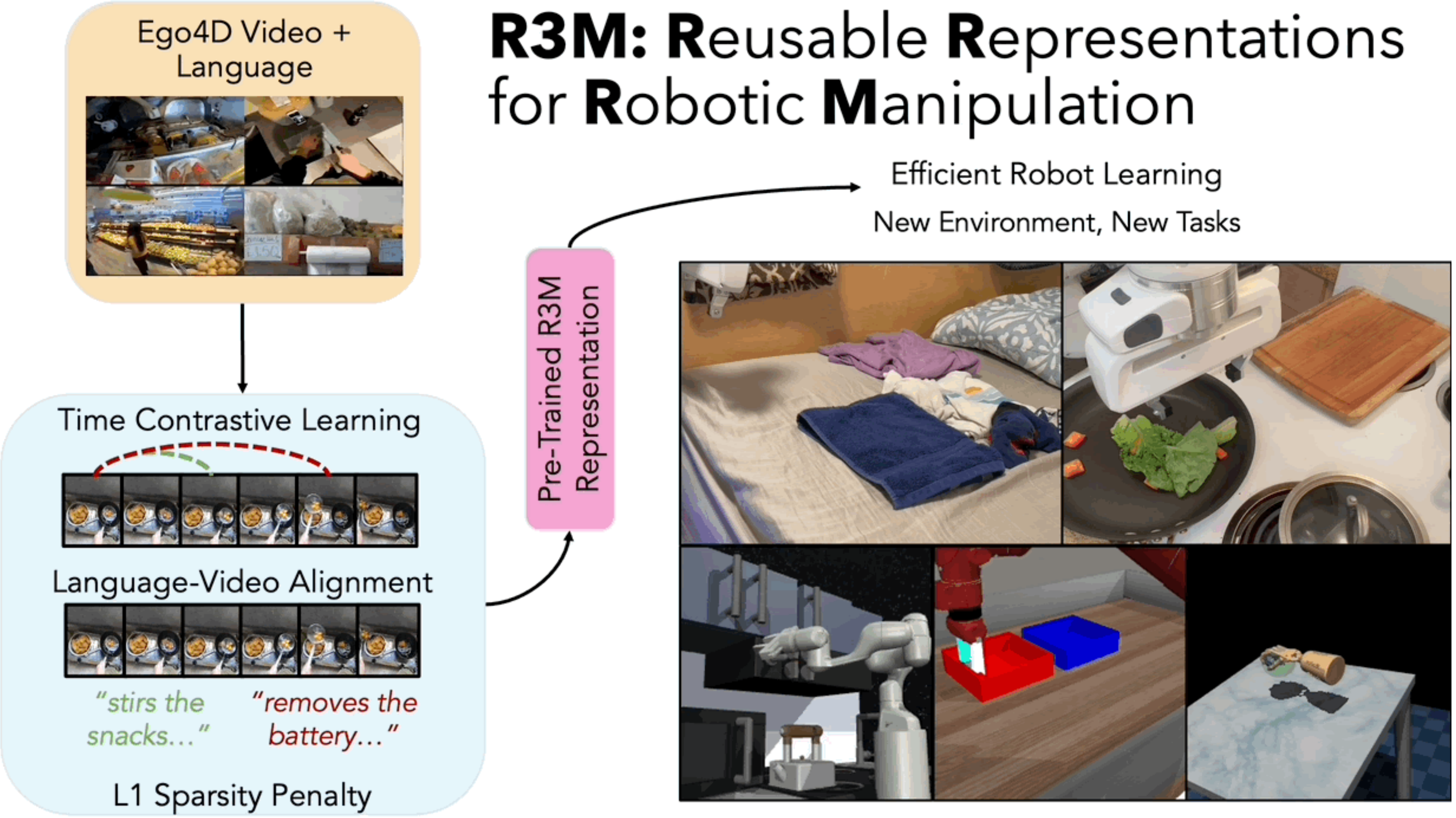
humans perform simple actions with everyday objects

174 unique action labels with a specific type of interaction



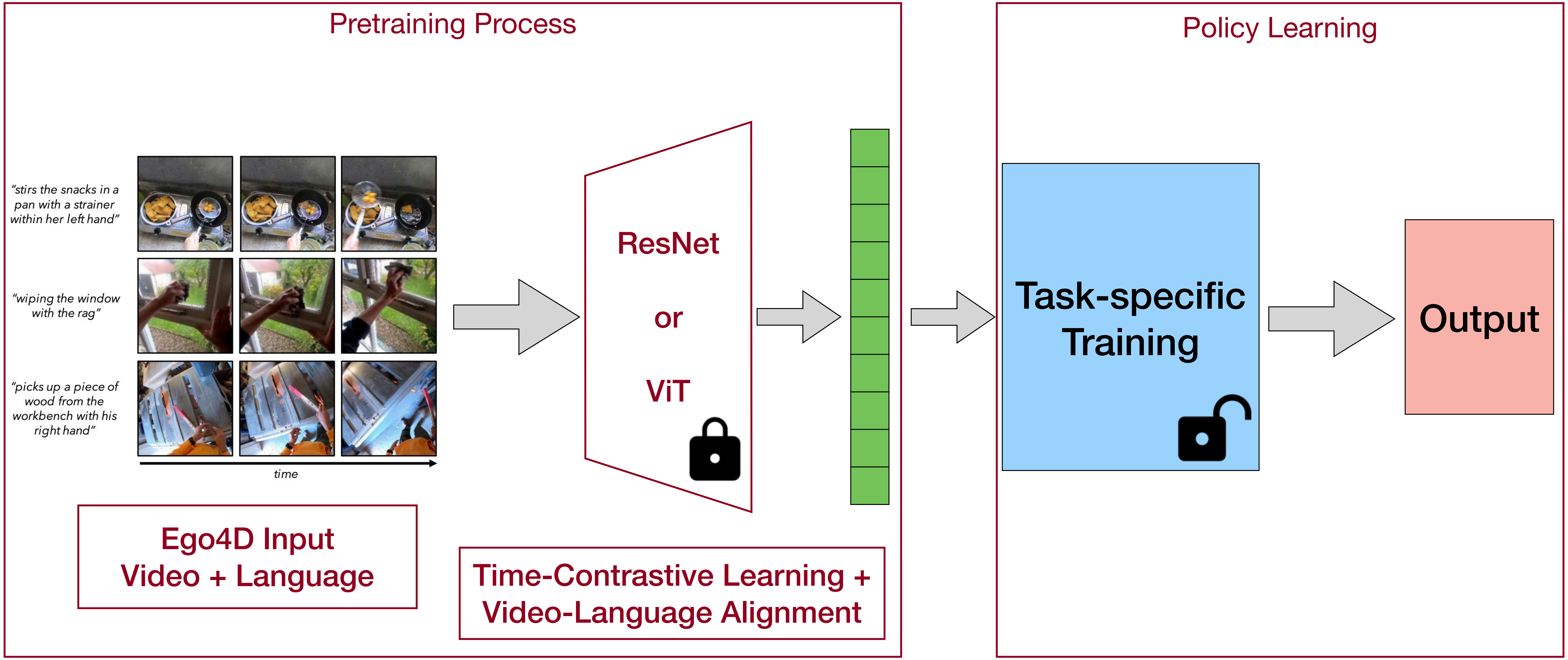
Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., ... & Memisevic, R. (2017). The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision* (pp. 5842-5850).

R3M: A Universal Visual Representation for Robot Manipulation



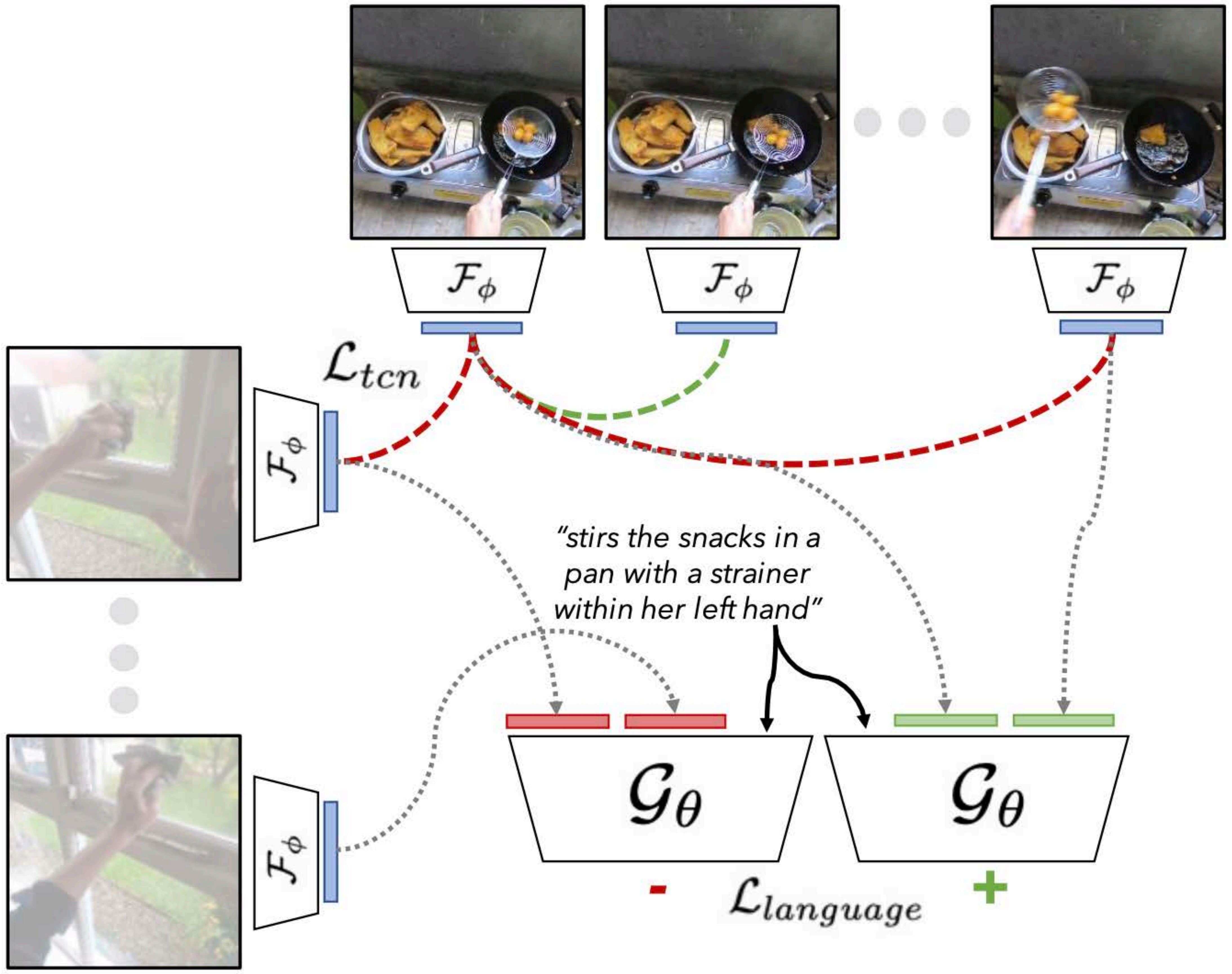
Nair, S., Rajeswaran, A., Kumar, V., Finn, C., & Gupta, A. R3M: A universal visual representation for robot manipulation. In Conference on Robot Learning (CoRL) 2022.

R3M - Pipeline



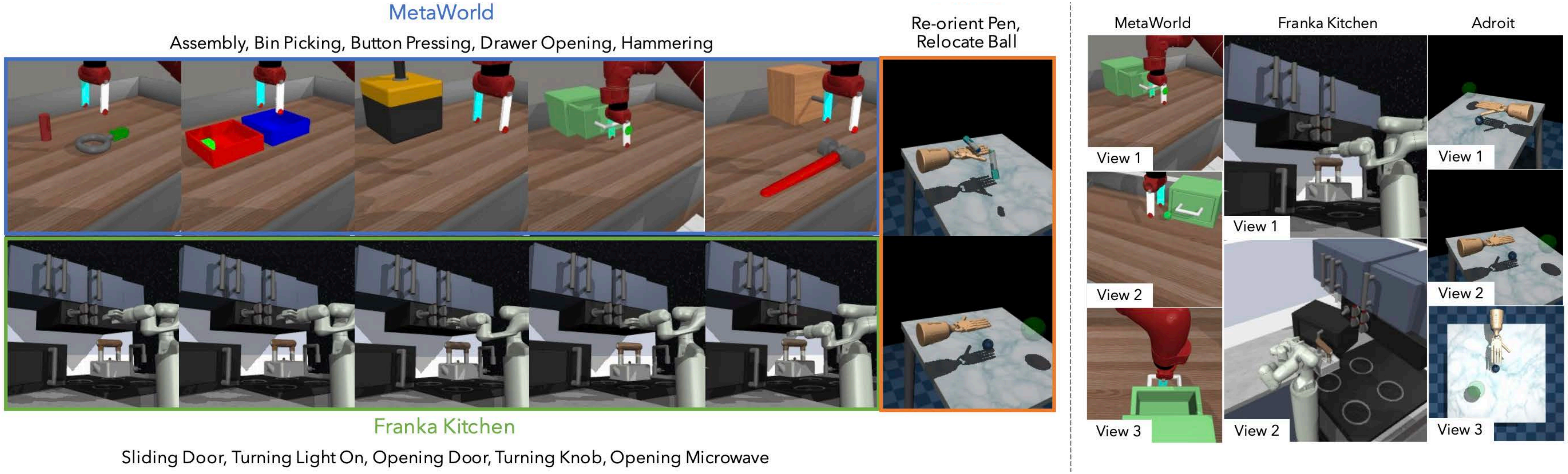
Nair, S., Rajeswaran, A., Kumar, V., Finn, C., & Gupta, A. R3M: A universal visual representation for robot manipulation. In Conference on Robot Learning (CoRL) 2022.

R3M Training



Nair, S., Rajeswaran, A., Kumar, V., Finn, C., & Gupta, A. R3M: A universal visual representation for robot manipulation. In Conference on Robot Learning (CoRL) 2022.

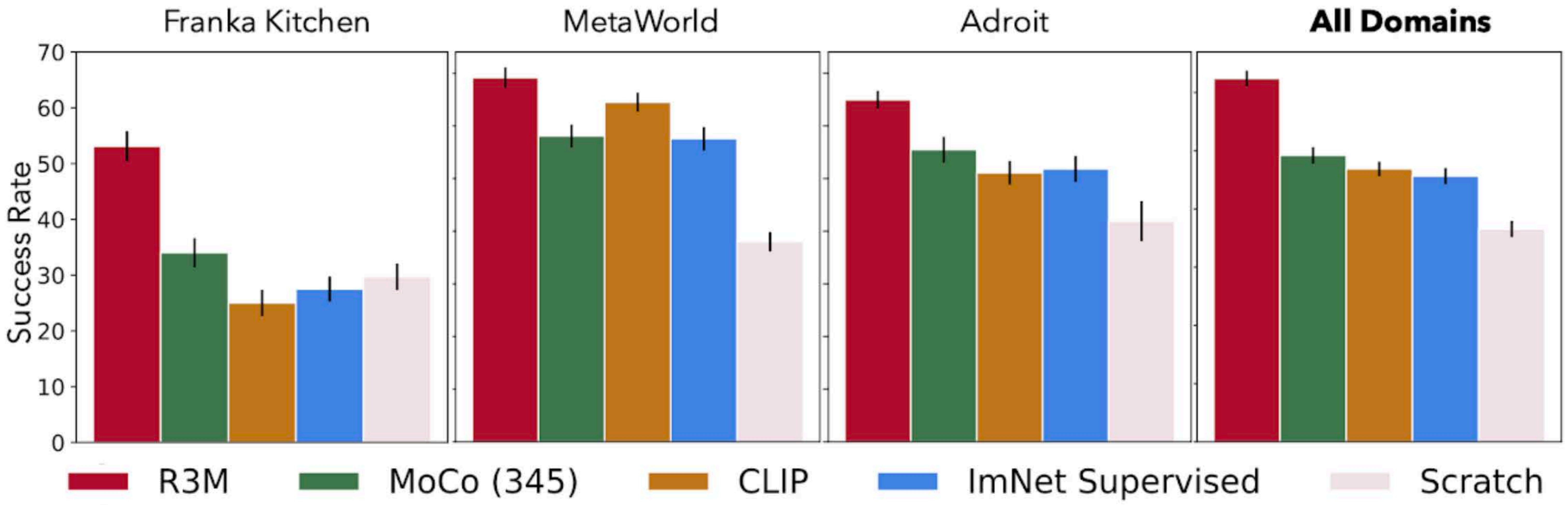
R3M - Evaluations



Nair, S., Rajeswaran, A., Kumar, V., Finn, C., & Gupta, A. R3M: A universal visual representation for robot manipulation. In Conference on Robot Learning (CoRL) 2022.

R3M - Results

We also demonstrate that pre-trained R3M representation enables data efficient imitation learning in a comprehensive simulation evaluations across three different benchmarks



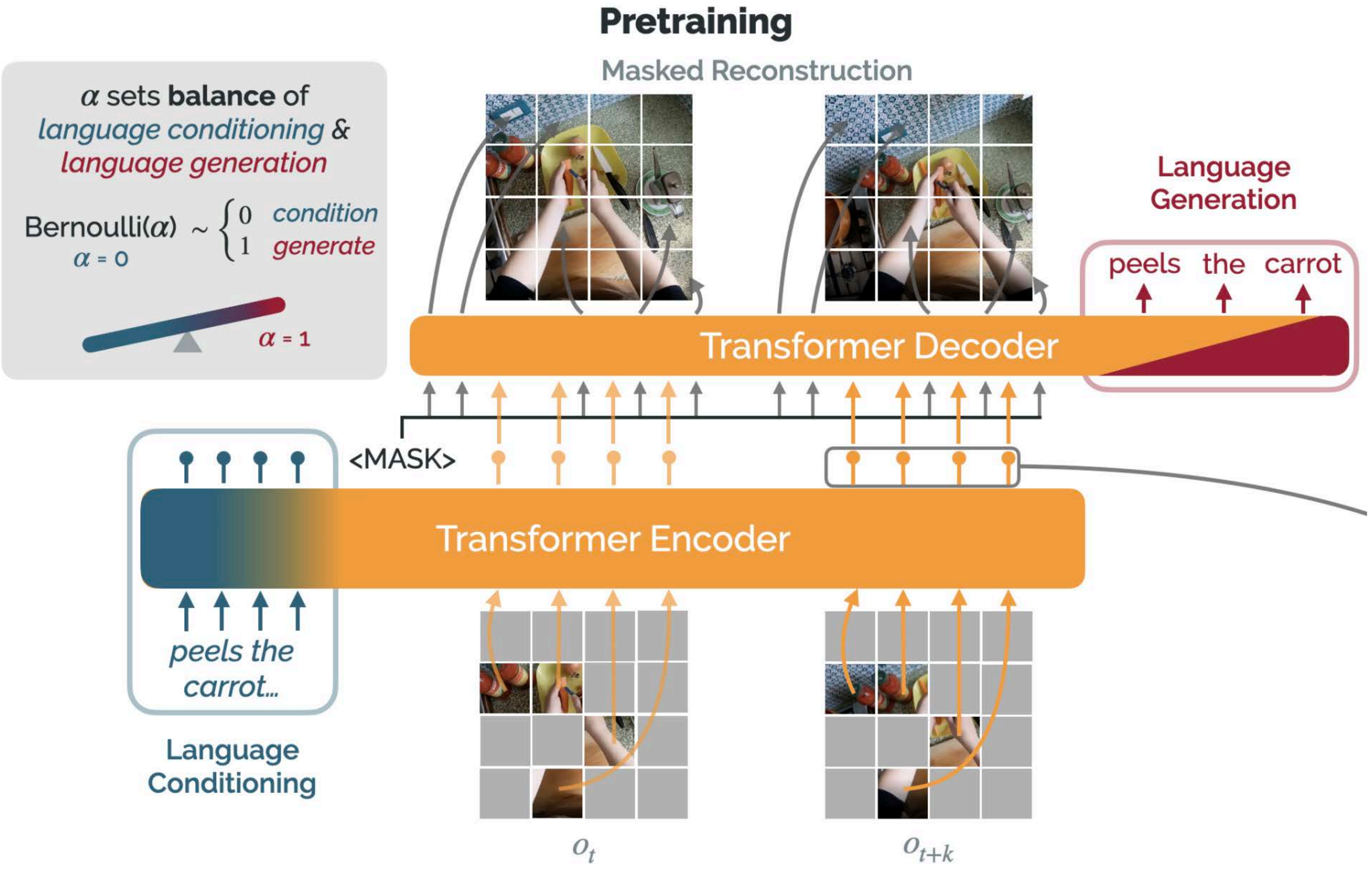
On average, R3M achieves **62%** success rate despite never seeing the environments/tasks before

R3M enables a **>10%** improvement in success rate over existing visual representations CLIP, MoCo(345), and Supervised ImageNet

R3M improves success rate over learning from scratch by **>20%**

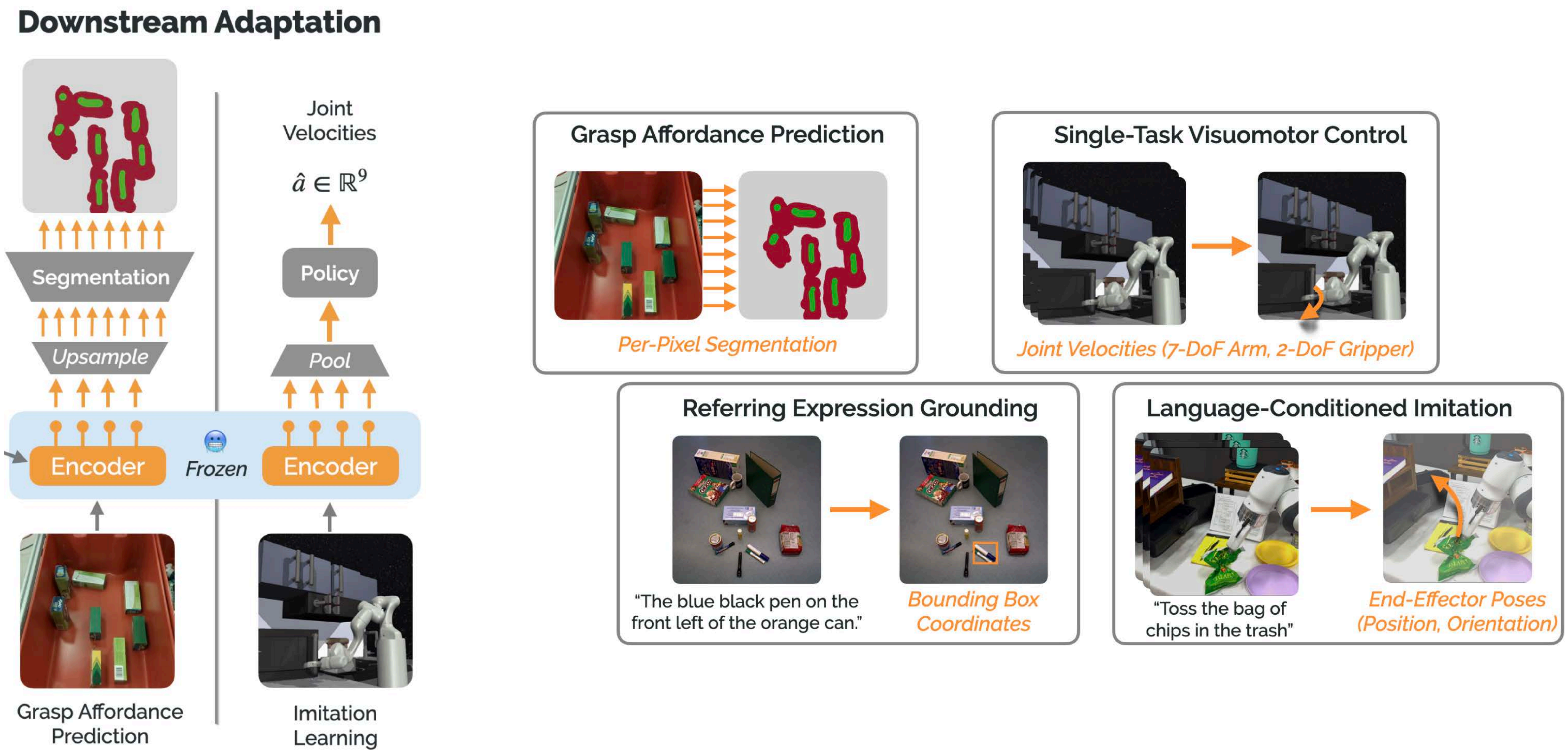
Nair, S., Rajeswaran, A., Kumar, V., Finn, C., & Gupta, A. R3M: A universal visual representation for robot manipulation. In Conference on Robot Learning (CoRL) 2022.

Voltron



Karamcheti, S., Nair, S., Chen, A. S., Kollar, T., Finn, C., Sadigh, D., & Liang, P. (2023). Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*.

Voltron Evaluation



Karamcheti, S., Nair, S., Chen, A. S., Kollar, T., Finn, C., Sadigh, D., & Liang, P. (2023). Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766.*

Voltron Results

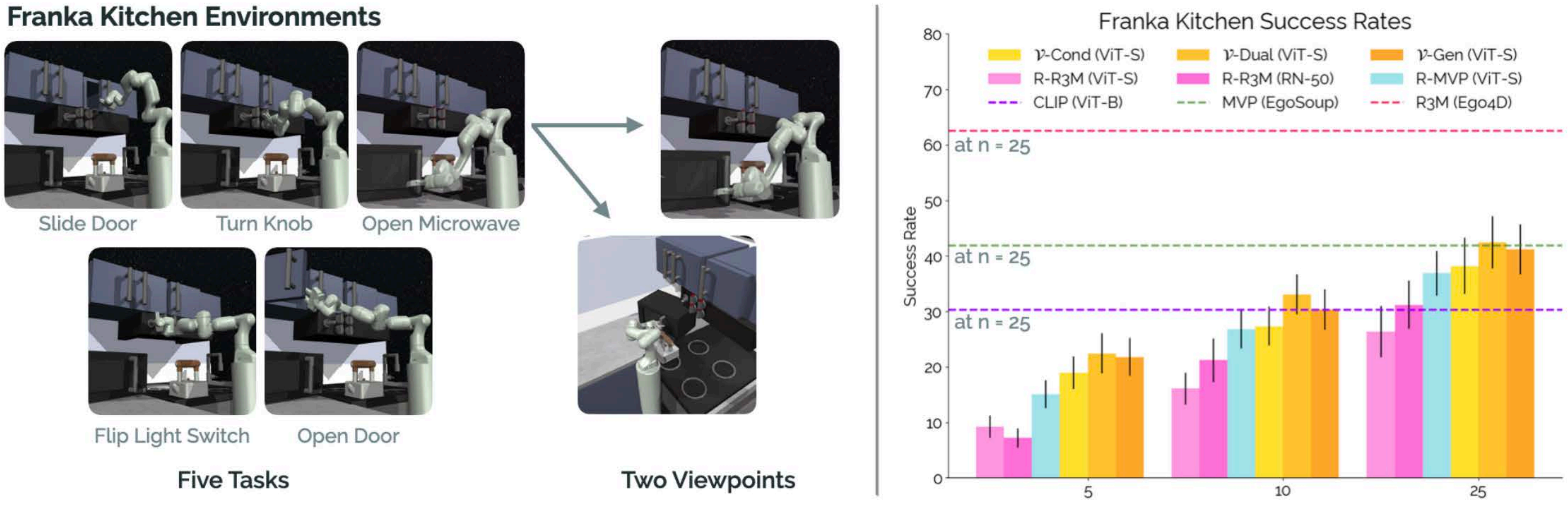


Figure 5: Franka Kitchen – Single-Task Visuomotor Control Results. Visualization of the Franka Kitchen evaluation environments, comprised of five unique tasks, with two camera viewpoints **[Left]**. Results (success rate for each of n demonstrations) for $\mathcal{V}$oltron and baselines, showing the benefit of language-driven learning (over 3 seeds) **[Right]**. In dashed lines (not directly comparable), we plot *CLIP (ViT-B)*, *MVP (EgoSoup)*, and *R3M (Ego4D)* trained with $n = 25$ demonstrations.

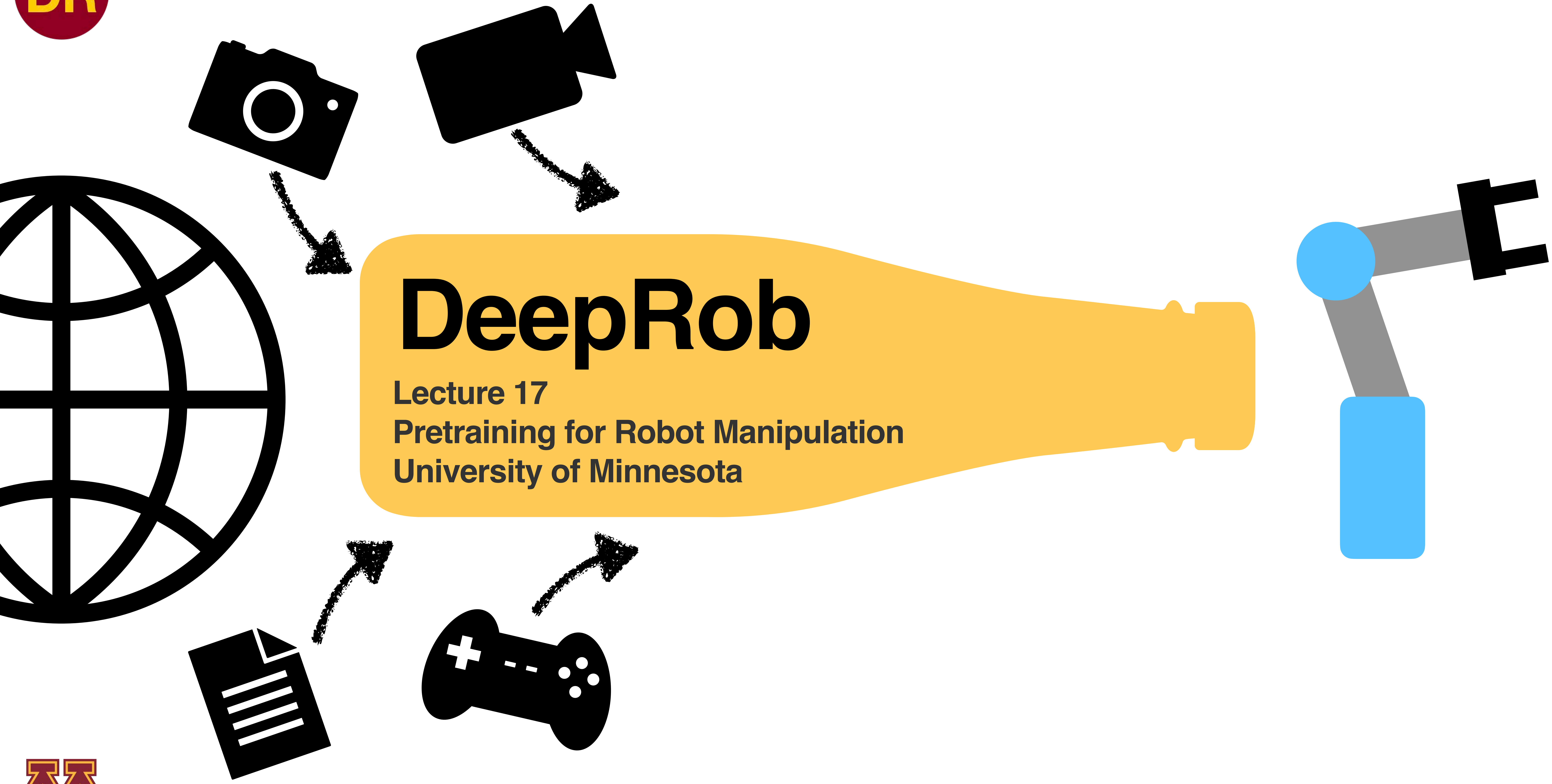
Karamcheti, S., Nair, S., Chen, A. S., Kollar, T., Finn, C., Sadigh, D., & Liang, P. (2023). Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*.

Why we need pretrain?

- Data Efficiency
- Transferability and Faster Learning
- Better Performance
- Generalization

DR

Next Lecture:
Student Lecture
RGB-D Networks and Manipulation



DeepRob

Lecture 17
Pretraining for Robot Manipulation
University of Minnesota