Multi-Modal Vision-Language-Action Foundation Models for Generalizable Robotics

Zsolt Kira
Associate Professor
School of Interactive Computing
Georgia Tech

Administrative

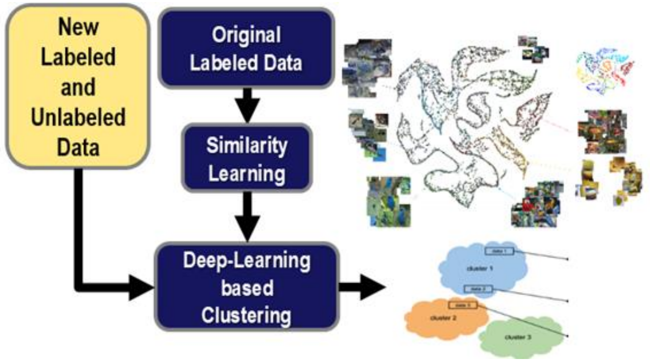
- **CIOS:** Fill out the CIOS! https://b.gatech.edu/CIOS
- **Project:** Project report rubrics and templates out
  - Due **Dec. 12th 11:59pm**
- **Video:** Will be out soon, but mostly it will be to have fun with it (YouTube video!)

Zsolt Kira

zkira@gatech.edu
Associate Professor
School of Interactive Computing
Georgia Institute of Technology

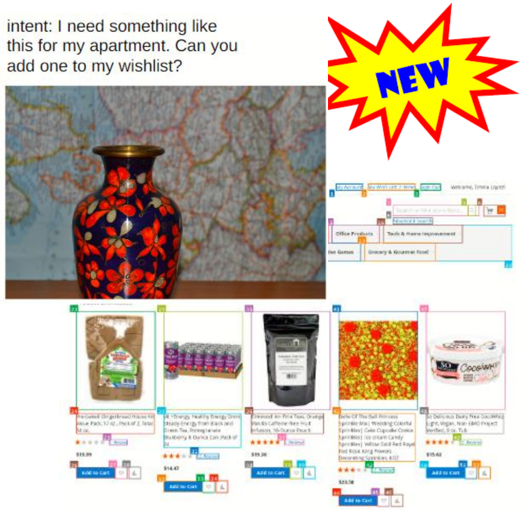
RIPL
Robotics Perception and Learning

Research Interests: Intersection of deep learning and robotics, focusing on robustness and decision-making in an open world



How can perception deal with changing environments and the open world?

Robust Open-World Learning
- Past: Semi and self-supervised, few-shot, continual learning
- Open-world learning and Vision-Language Models
- Robust fine-tuning of VMs/VLMs



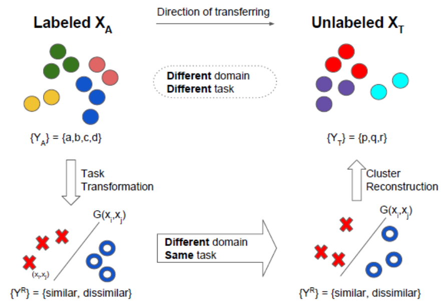
How can we use VLMs for Learning, Planning, and Reasoning Agents

Planning, Reasoning, Memory
- VLMs for reasoning/planning
- Long-form videos and memory
- Grounding

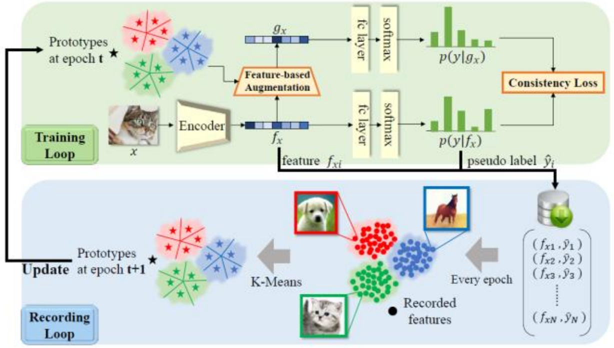
How can we scale robotics in DL era?

Scaling up Robotics
- Better simulation w/ NeRFs/3D
- Self-supervised and pre-training
- Combinations with large language and multi-modal models
  - Long-Context Models
- Vision-Language Action Models

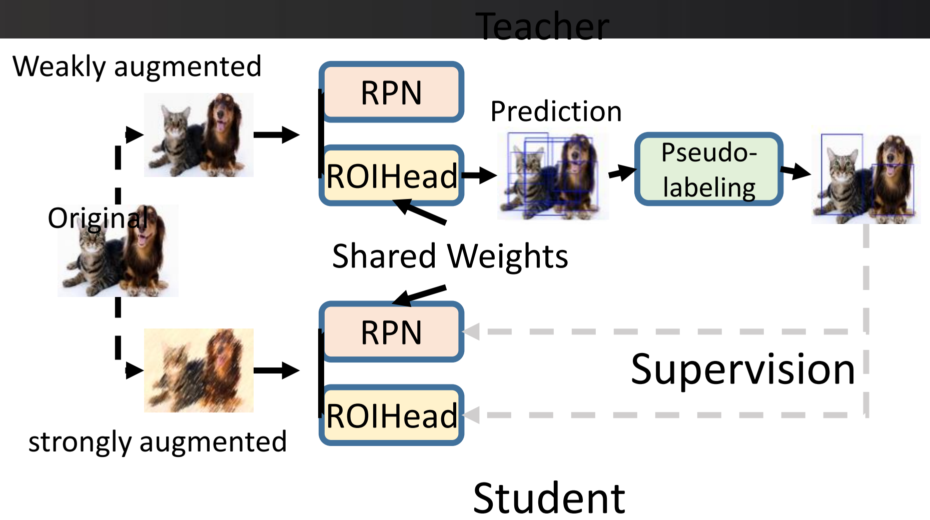
2018-2022



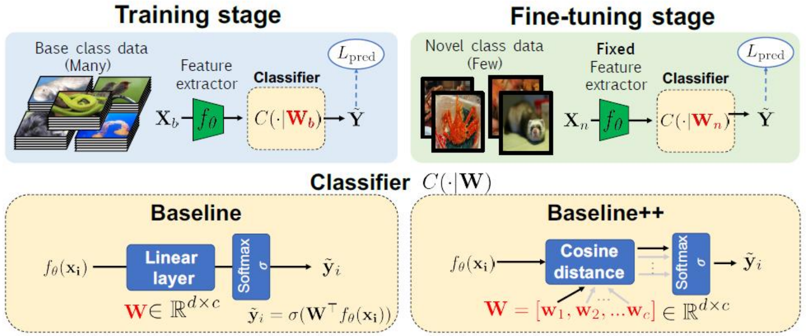
Pairwise Similarity for Cross-Task Object Discovery
[ICLRW 2016, ICLR 2018, 2019]



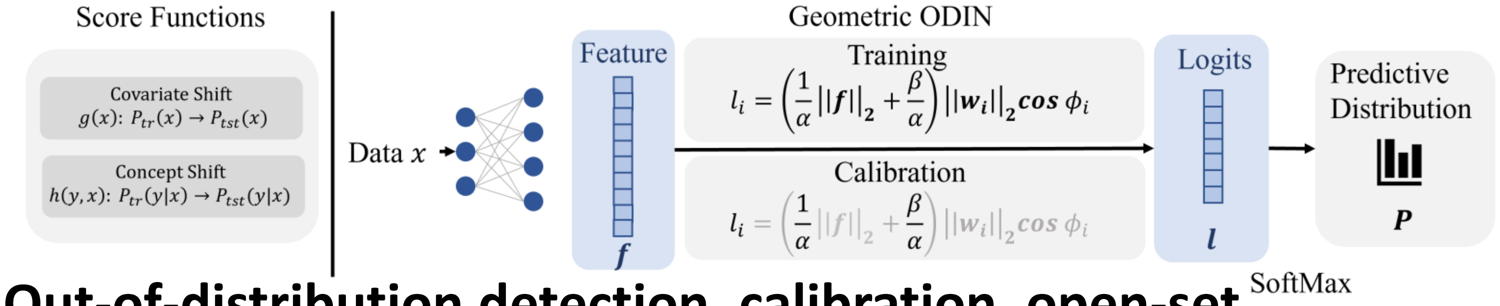
Complex Data Augmentation Domain Generalization/SSL
[ECCV 2020]



Unbiased Teacher for Anchor-based, Anchor-Free Open-Set Object Detection
[ICLR 2020, CVPR 2022, ECCV 2022, w/ Meta]

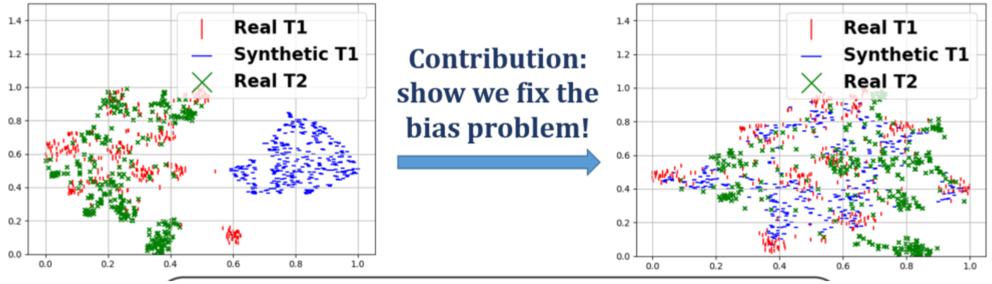


Closer Look @ Few-Shot (w/ VT)
[ICLR 2019]



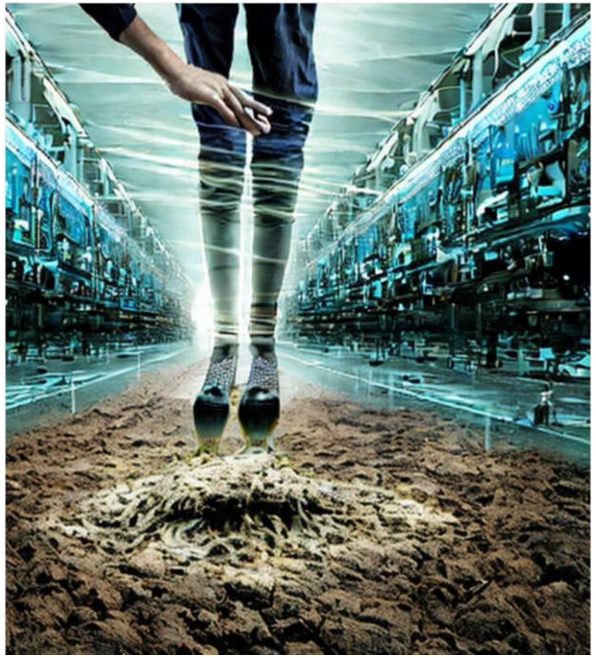
Out-of-distribution detection, calibration, open-set
[CVPR 2020, NeurIPS 2021]

Video Domain adaptation
[CVPR 2019, ICCV 2019]



Continual Learning
[ICCV 2021, Nature 2022]

The great shift



- Modality-specific pipelines
  ▶ DL
    ▶ Transformers

- Scale + Self/semi-supervised learning FTW!
  - Web ▶ Language Models ▶**Knowledge**
  - DINO/MAE/CLIP/SAM ▶**Scene Understanding**
  - **Multi-Modal Models**

Where does robotics go from here?

[Image by Stable Diffusion]

The Reality

- Perception is *still* tied to *known* categories or poor open-vocabulary methods during training
- Brittle to out-of-distribution data
- Limited Open-World abilities
- Even large-scale datasets (RT-X) limited in generalization

% success rates

Method	Seen	Unseen		
		Layouts	Objects	Receptacles
MonolithicRL	91.7 ±1.1	86.3 ±1.4	74.7 ±1.8	52.7 ±2.0
SPA	70.2 ±1.9	72.7 ±1.8	72.7 ±1.8	60.3 ±2.0
SPA-Priv	77.0 ±1.7	80.0 ±1.6	79.2 ±1.7	60.7 ±2.0

Degradation over novelty…

Habitat 2.0
Work by Andrew Szot, Dhruv Batra, and Meta

Challenges in Robotics

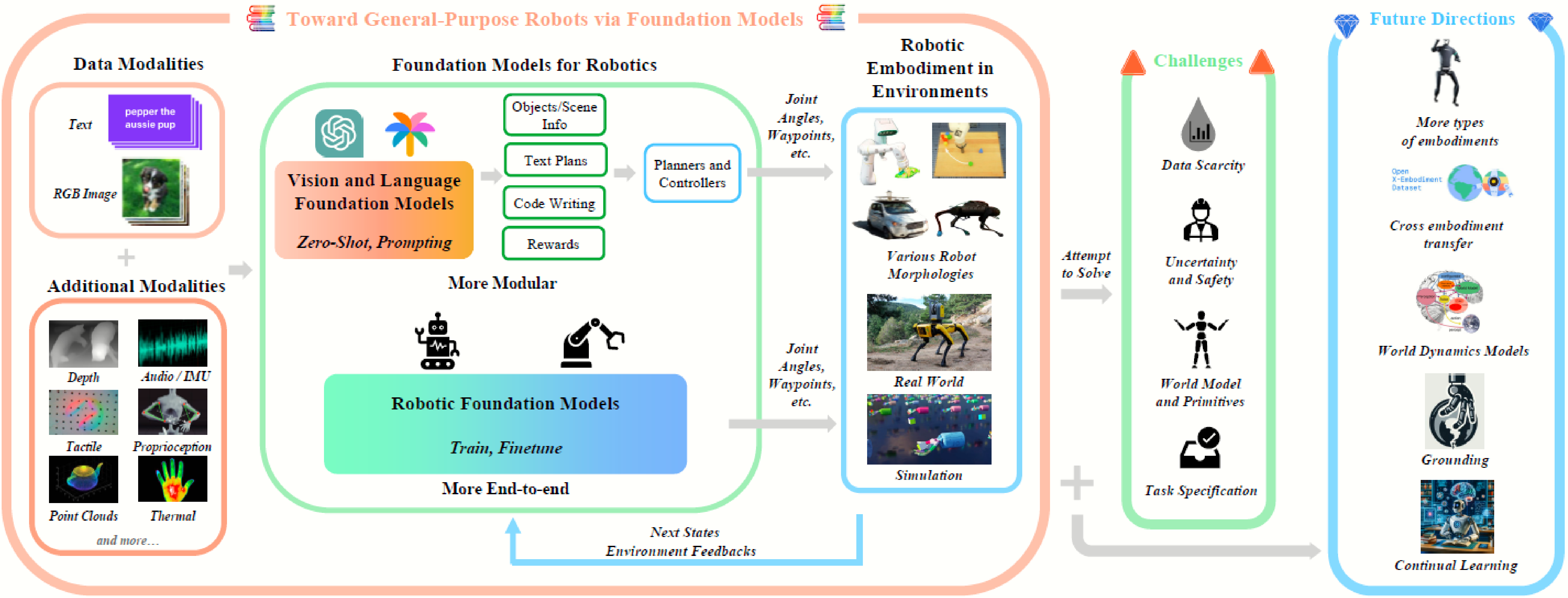
- Data flywheel
  - Hard to gather
  - Potentially dangerous
  - Huge heterogeneity

- Robusness
  - In-the-wild data
  - Long-tail (see self-driving cars)
  - Long-horizon decision-making
  - Physics!

- Reliability 24/7

- Cost?

Robotic Foundation Models



Hu et al., Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis

Open-World Learning (w/ FMs and VLAs)

Reproducible Robotics -> Simulation

Generalization to an Open World

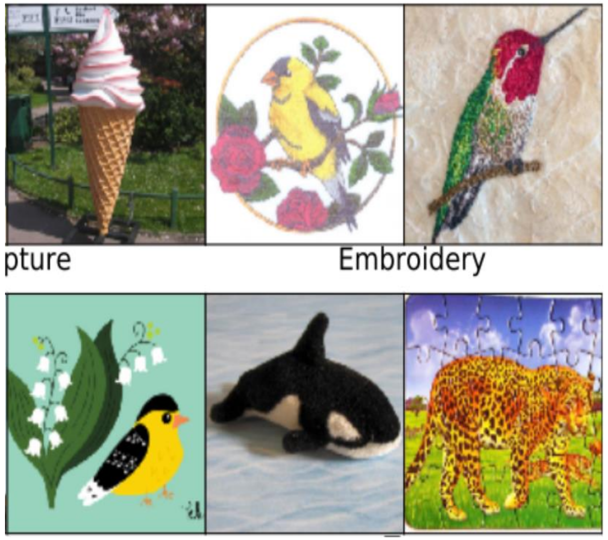
Long-Horizon/Long-Context, Memory

Robust Fine-tuning

[NeurIPS 2023 OVMM Challenge, ICML 2023, Neurips 2021]
(w/ Dhruv Batra)

[ICLR 2018/2019, arXiv:2305.10420, ECCV 2022]

[ECCV 2024, on-going]

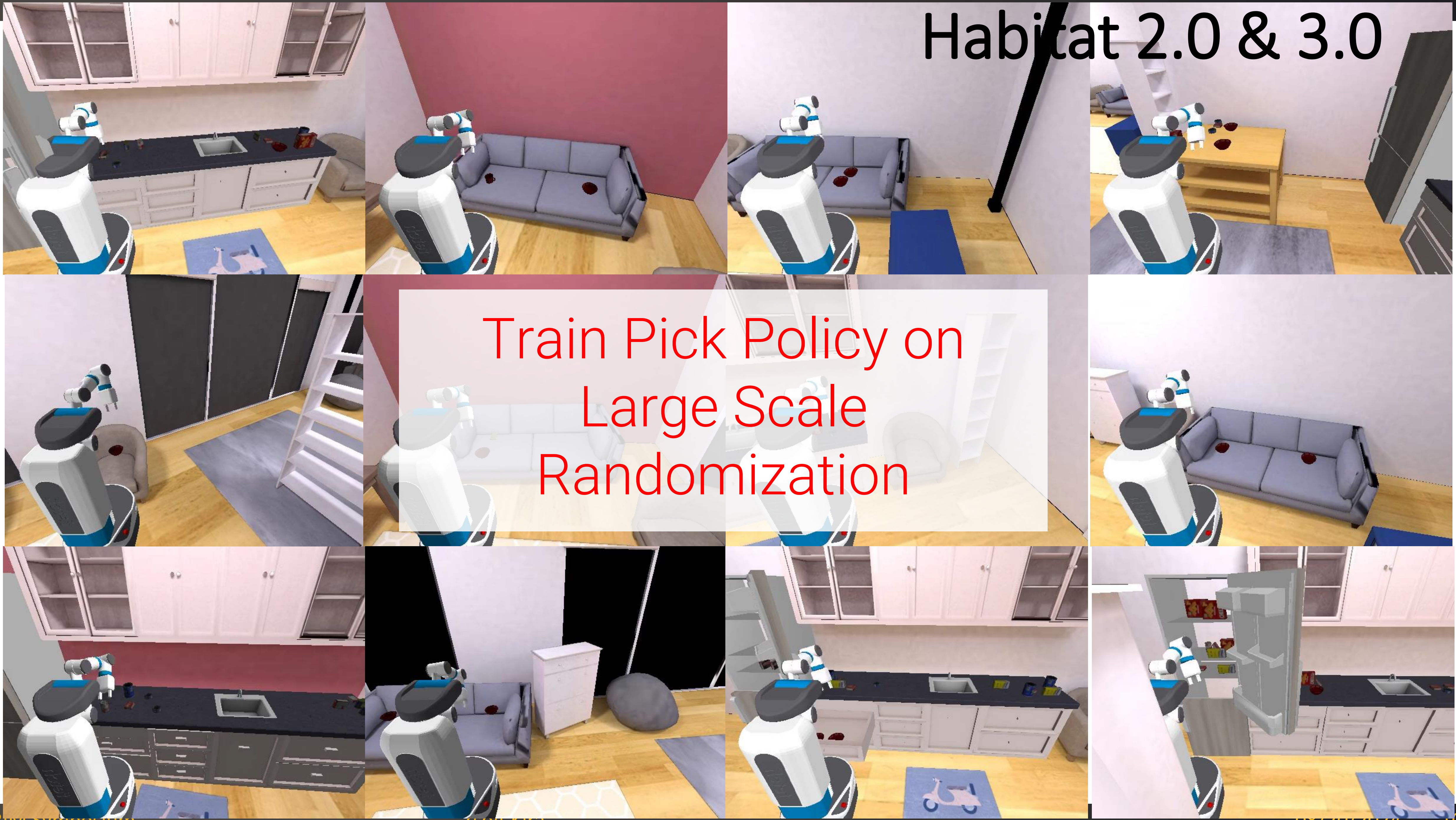


[ImageNet-R]

[CVPR 2023, NeurIPS 2023/2024]

[Middle two images by Stable Diffusion]

Habitat 2.0 & 3.0



Train Pick Policy on Large Scale Randomization

Multimodal Large Language Models

Bing's A.I. Chat: 'I Want to Be Alive. 😈'

In a two-hour conversation with our columnist, Microsoft's new chatbot said it would like to be human, had a desire to be destructive and was in love with the person it was chatting with. Here's the transcript.

https://www.nytimes.com/article/ai-artificial-intelligence-chatbot.html

ARTIFICIAL INTELLIGENCE

ChatGPT is about to revolutionize the economy. We need to decide what that looks like.

New large language models will transform many jobs. Whether they will lead to widespread prosperity or not is up to us.

By David Rotman — March 25, 2023

https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like/

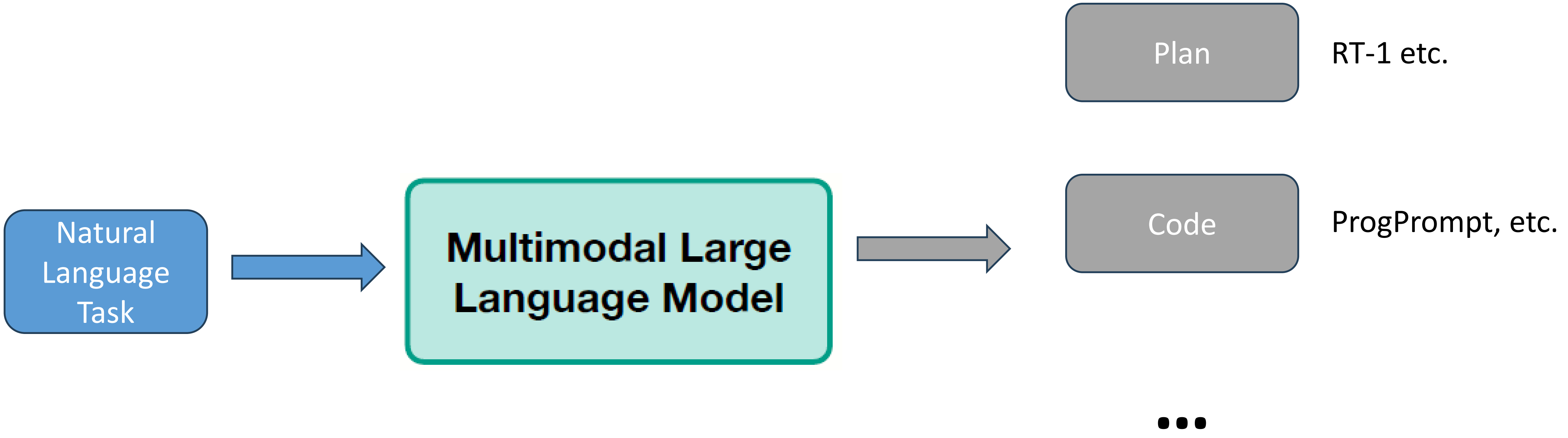
Multimodal Large Language Model

GPT-4o
OPENAI'S LATEST MODEL

Gemini 1.5

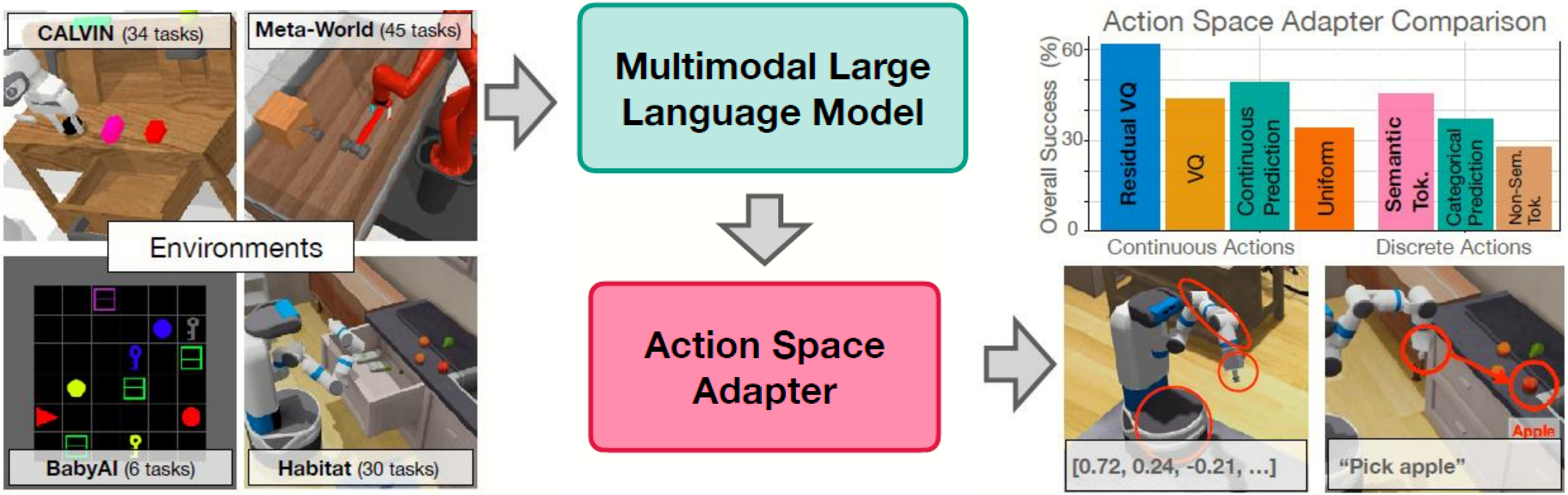
LLAMA 2

Multimodal Large Language Models



What about VLMS for direct task to action?

Vision-Language Action Models



Lots of great concurrent work! OpenVLA, LLARVA, etc.

Andrew Szot

ML Ph.D. (co-advised with Dhruv Batra)

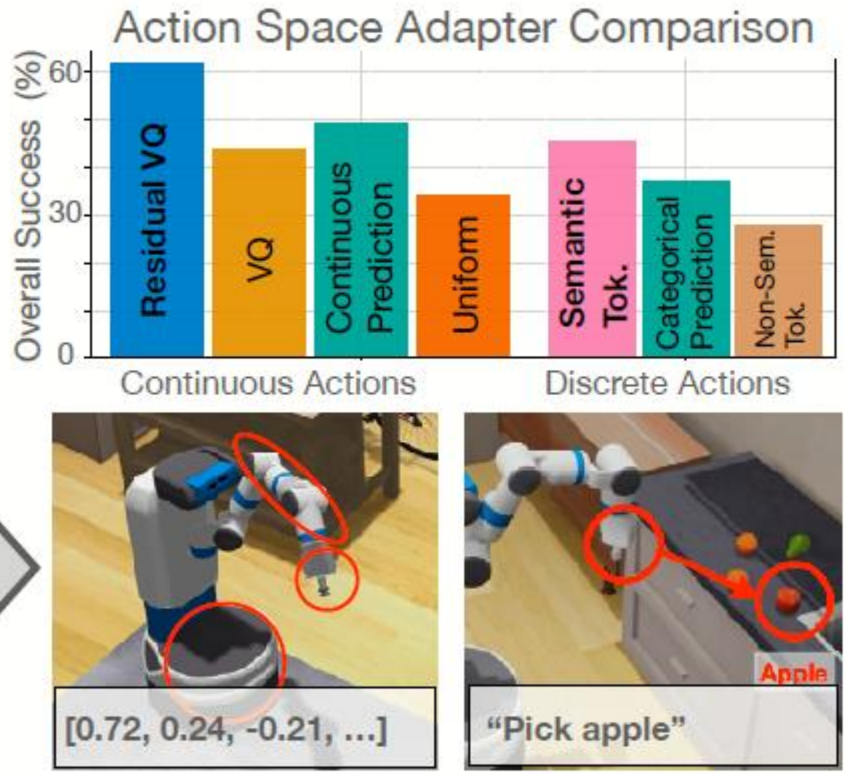
Szot et al., Grounding Multimodal Large Language Models in Actions, NeurIPS 2024

Vision-Language Action Models

- **Advantages:**
  - Policies driven by textual description of the task!
  - Leverage common sense reasoning inside model
  - Can learn with RL and IL

Multimodal Large Language Model

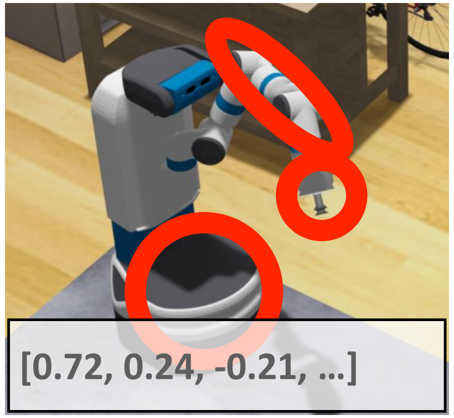
Action Space Adapter



- **Questions:**
  - How should we represent (tokenize) output actions?
    - Concurrent work tends to just pick one and go with it

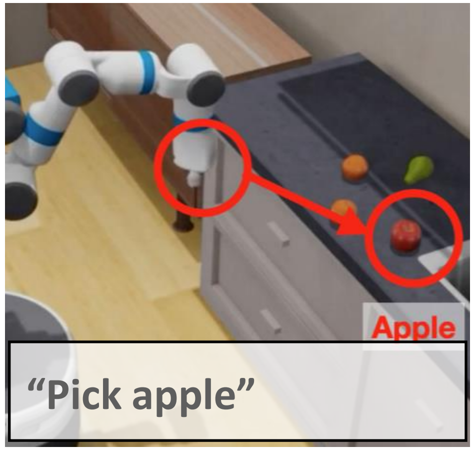
Szot et al., Grounding Multimodal Large Language Models in Actions. NeurIPS 2024.

Action Tokenization



Action is a continuous vector

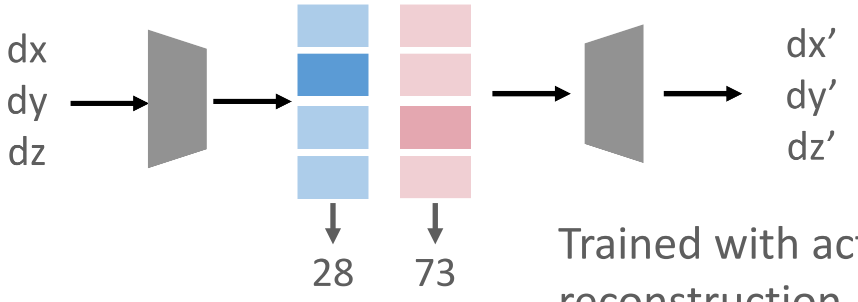
Example: end effector control
[dx, dy, dz]



Action is a selection from a set of discrete choices

Learned Tokenization

Residual VQ-VAE for discrete action tokenization



Trained with action reconstruction

Semantic Tokenization

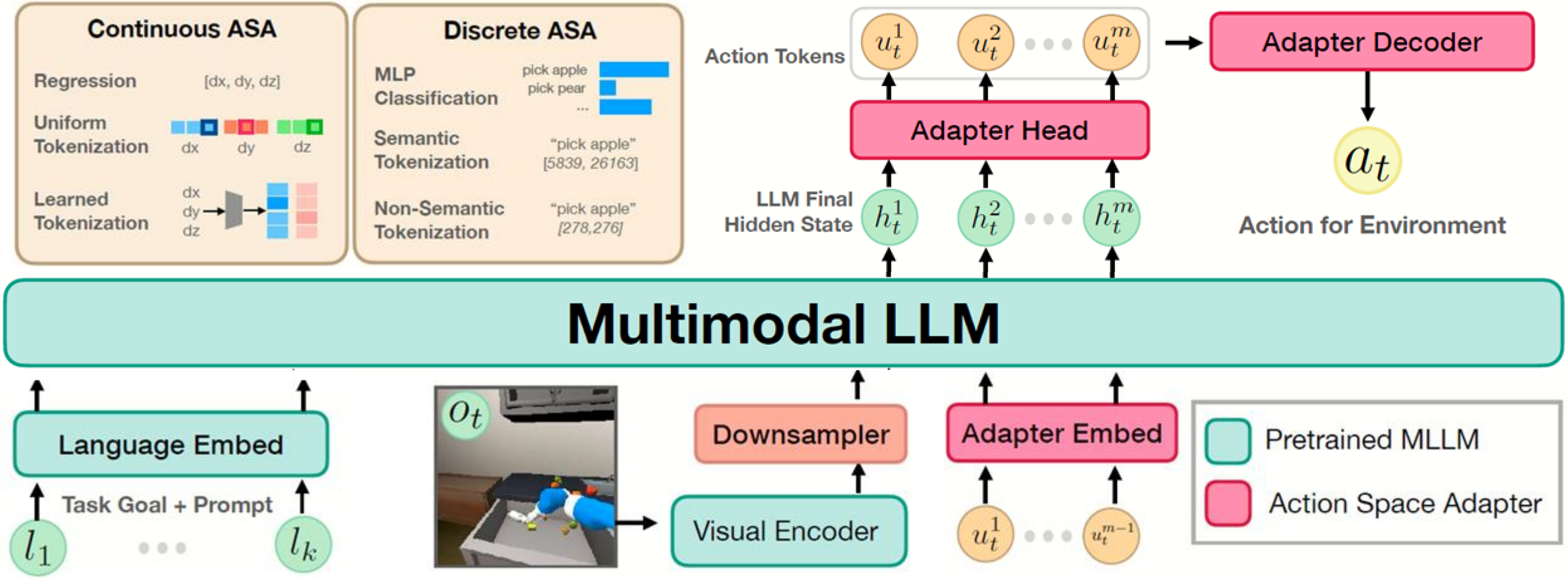
“pick apple”

↓

[278,276]

Describe action with language and tokenize with LLM vocabulary

Szot et al., Grounding Multimodal Large Language Models in Actions. NeurIPS 2024.



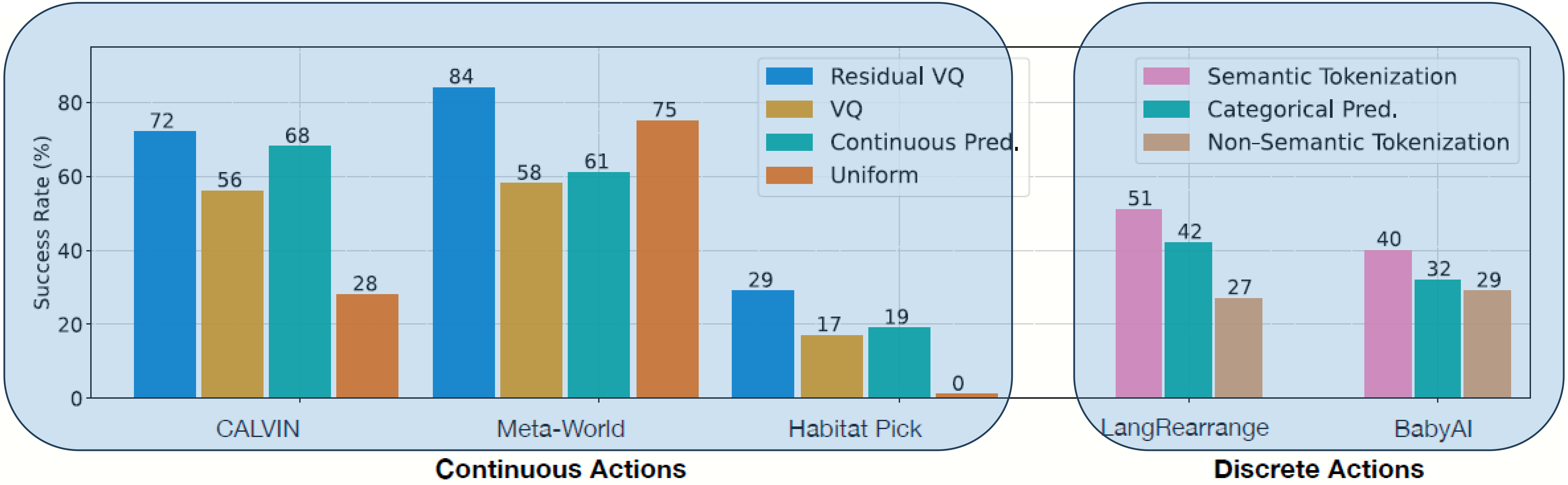
We finetune the ASAs, downsampler, and MLLM

Szot et al., Grounding Multimodal Large Language Models in Actions

Andrew Szot

ML Ph.D. (co-advised with Dhruv Batra)

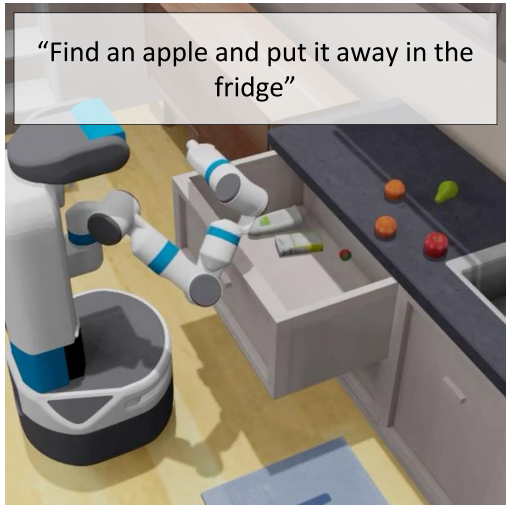
VLA Results & Findings



VLA: Results across Spectrum of Generalization

	Total	Aggregated		Per Dataset Breakdown											
		Behavior Generalization	Paraphrastic Robustness	Train	Scene	Instruct Rephrasing	Novel Objects	Multiple Rearrange	Referring Expressions	Context	Irrelevant Text	Multiple Objects	Spatial	Conditional Instructs
SemLang	51 ± 1	56 ± 2	47 ± 1	94 ± 3	94 ± 6	92 ± 1	97 ± 0	80 ± 6	31 ± 3	46 ± 14	66 ± 6	2 ± 2	0 ± 0	46 ± 4
Lang	27 ± 12	31 ± 14	24 ± 10	72 ± 13	58 ± 11	74 ± 12	76 ± 29	21 ± 10	10 ± 12	12 ± 11	20 ± 13	0 ± 0	2 ± 3	26 ± 16
Pred	42 ± 2	45 ± 3	38 ± 1	99 ± 1	96 ± 4	92 ± 2	95 ± 4	47 ± 5	26 ± 2	34 ± 2	32 ± 2	0 ± 1	8 ± 1	39 ± 3

Many tasks we want an agent to take actions to autonomously complete



Robotic Manipulation

Navigation

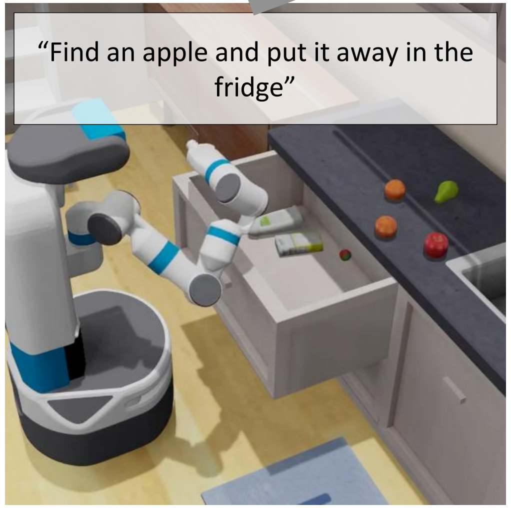
Games

UI Control

Can we have ***one*** policy that does all of these?

How can we create a generalist agent capable of excelling in diverse interactive tasks?

Generalist Agent



Robotic Manipulation

Navigation

Games

UI Control

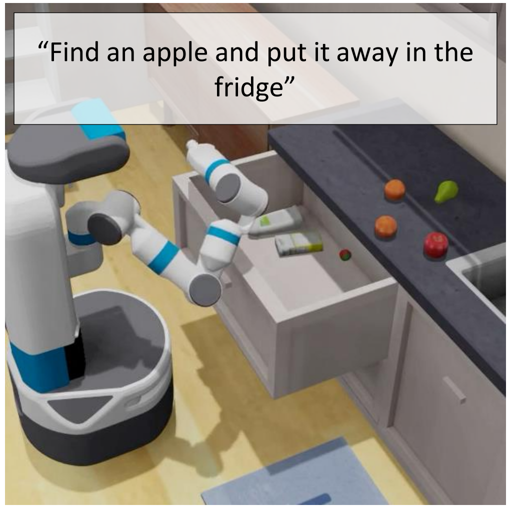
Adapt a pre-trained Multimodal LLM

Multimodal LLM

Interactive Data

Adapt

Generalist Agent



Robotic Manipulation

Navigation

Games

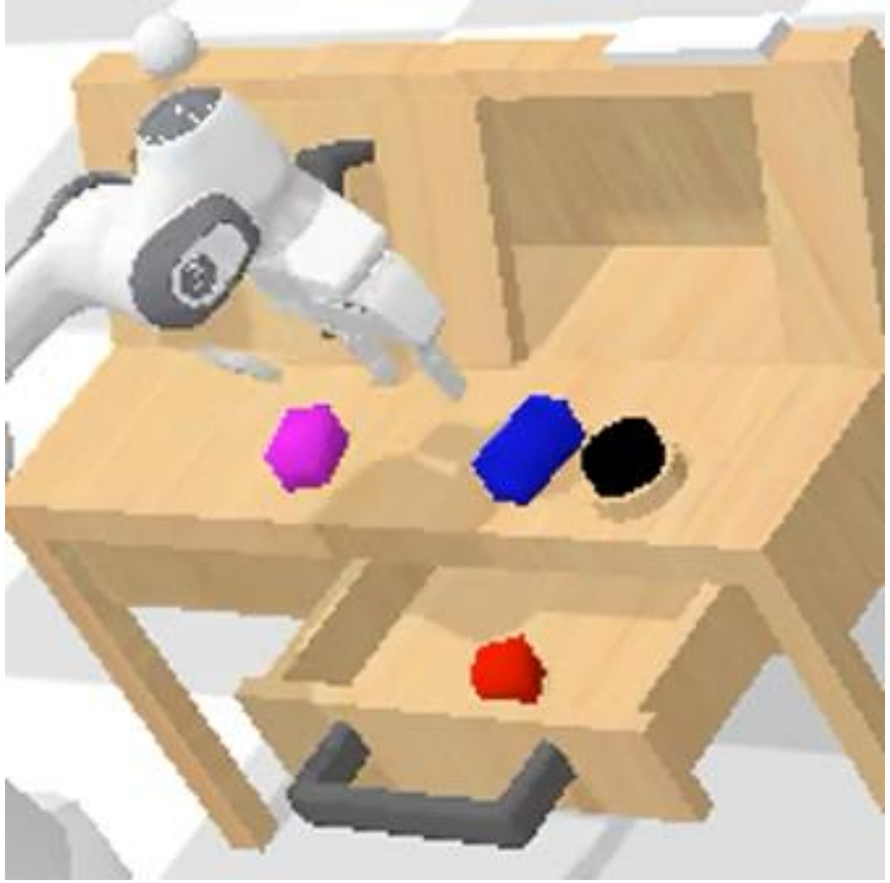
UI Control

Step 1: Collect expert demonstrations in diverse domains for training

From diverse sources, like scripted policies, humans, or RL policies

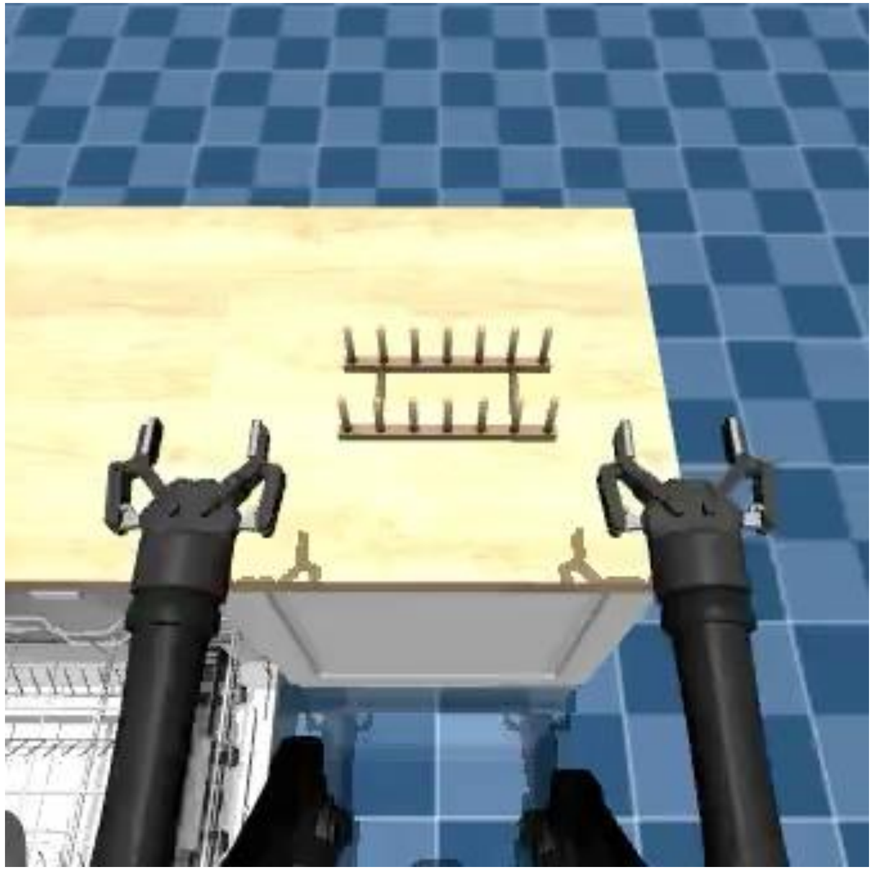
Data - Static Manipulation

“Use the block to pull the handle sideways”



“Move the purple block next to the blue block”

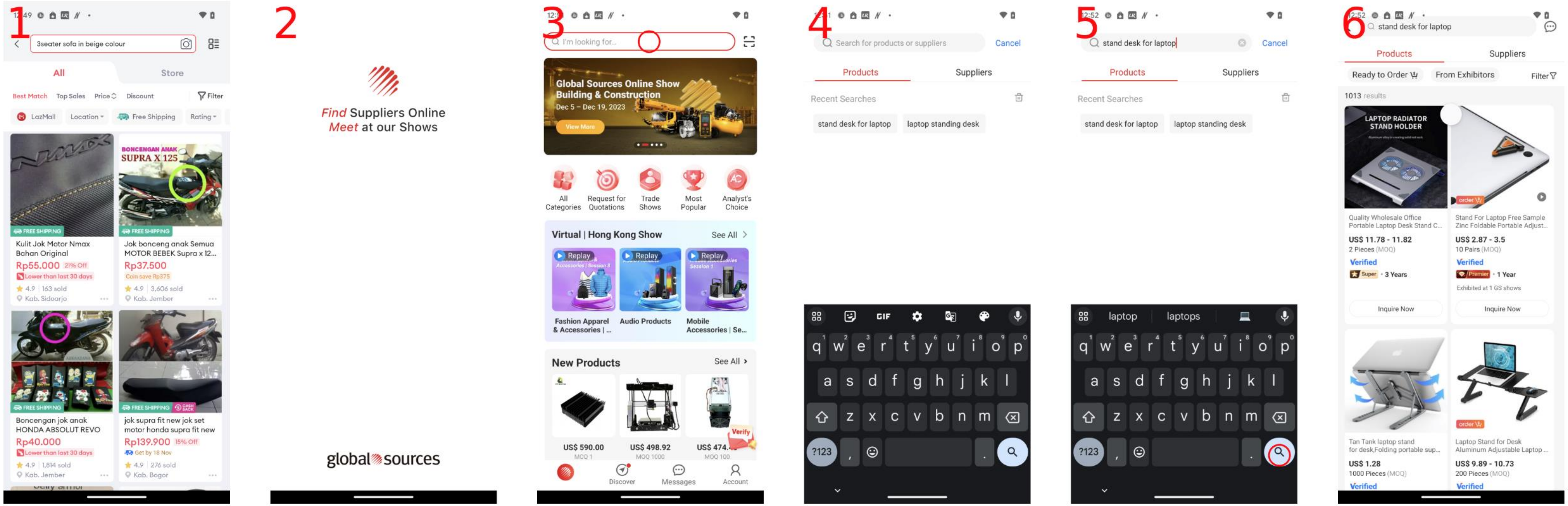
Data - Mobile Manipulation



“Unload the plates from the dishwasher and place them on the rack”

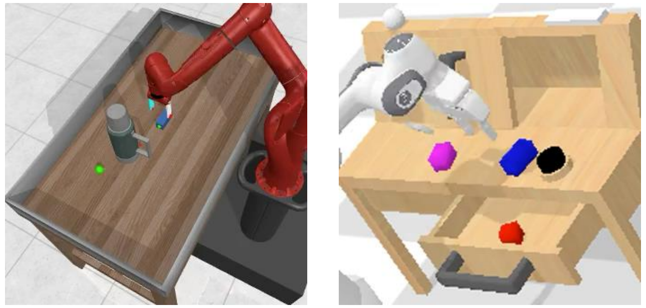
“Pick up the banana”

Data - UI Control

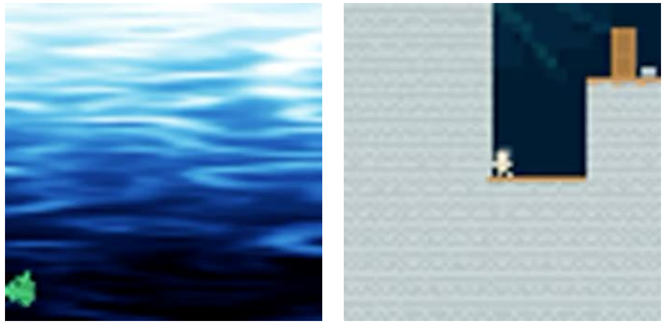


“Find me a standing desk for my laptop from the GlobalSources app”

Static Manipulation

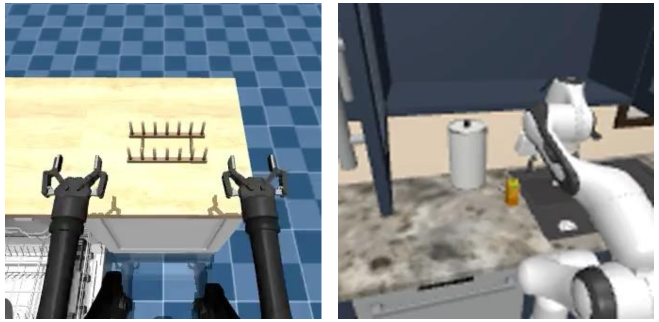


Games



Navigation

Mobile Manipulation



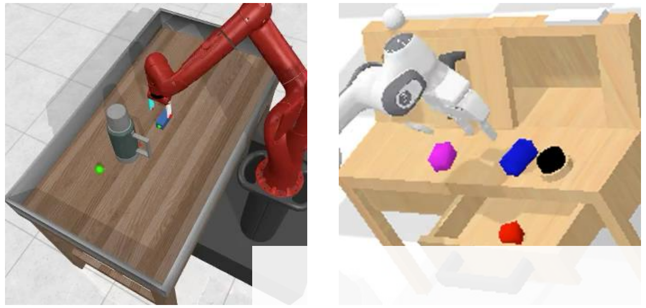
UI Control

Real Robots

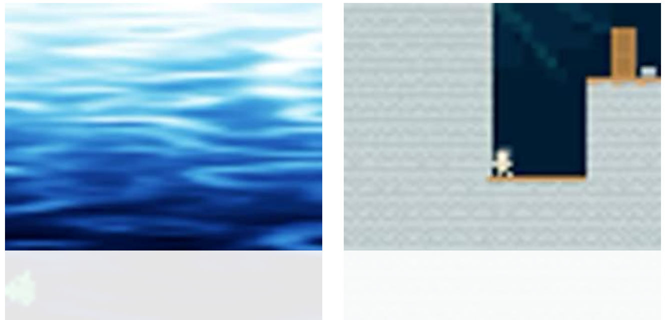
Character Control

Planning

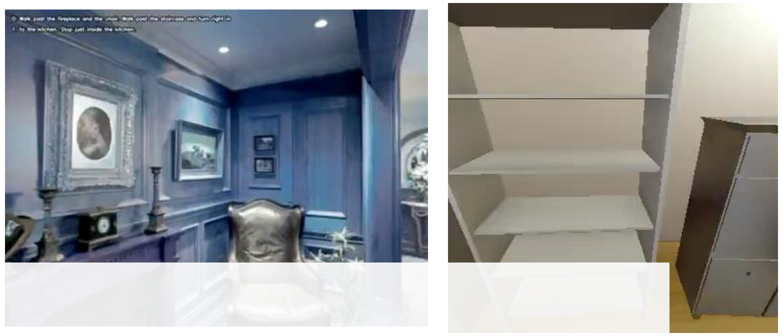
Static Manipulation



Games

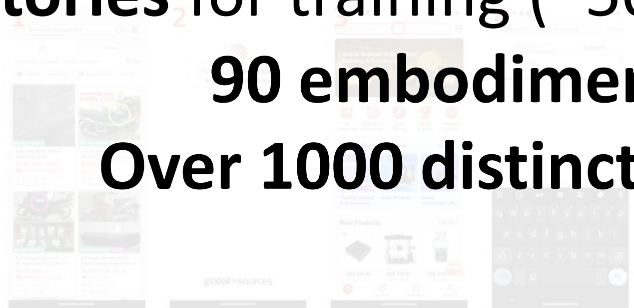


Navigation

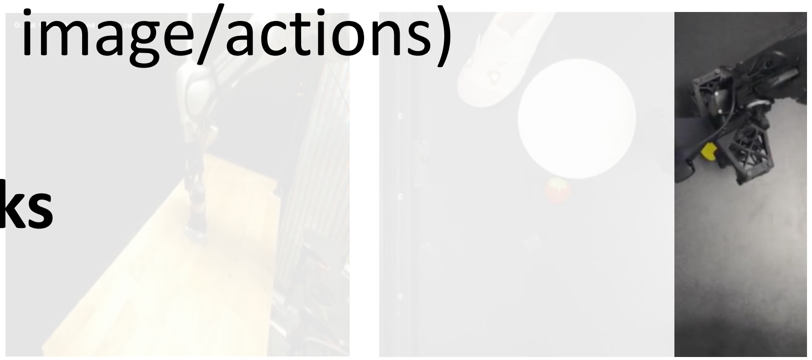


Mobile Manipulation

UI Control



Real Robots

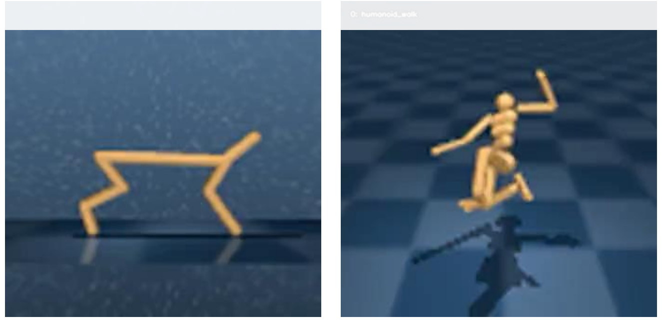


4M trajectories for training (~500M image/actions)

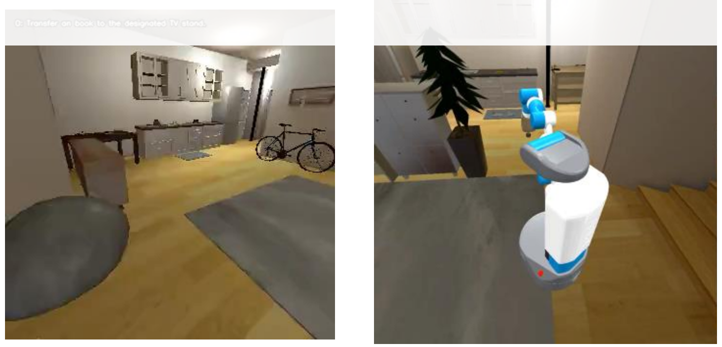
90 embodiments

Over 1000 distinct tasks

Character Control



Planning



Evaluation

New Tasks

Find an apple and put it away in the fridge.

Novel Objects
Find a pear and put it away in the fridge.

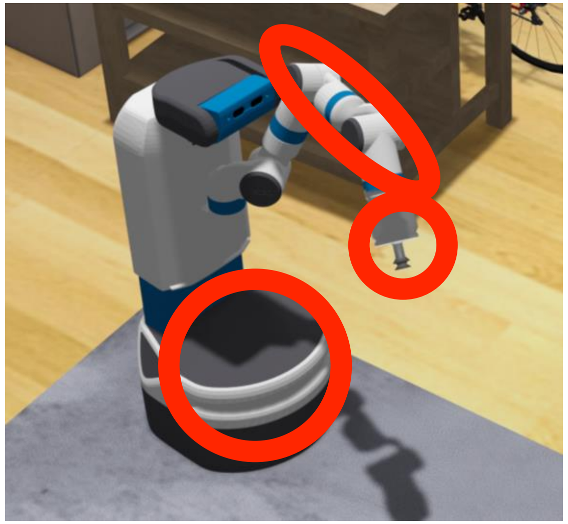
Context
I am hungry for something sweet and healthy. Put a snack for me on the table.

Spatial Relationships
Find an apple and put it in the receptacle to the right of the kitchen counter.

⋮

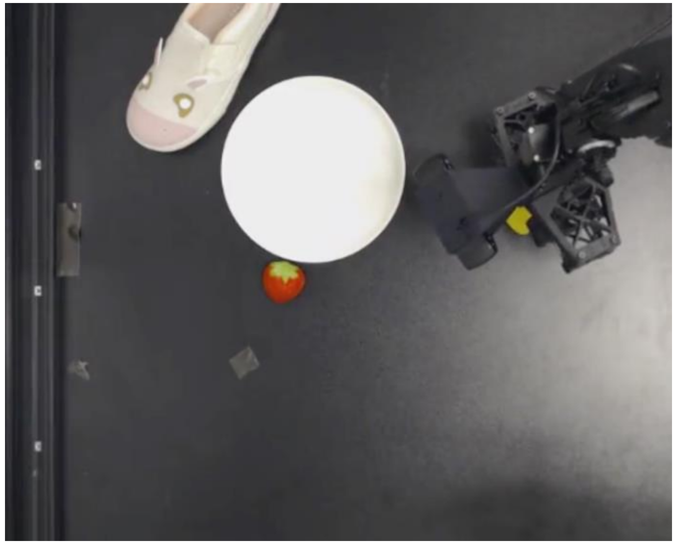
New Embodiments

New control spaces and robot types



New Environments

New platform with limited data

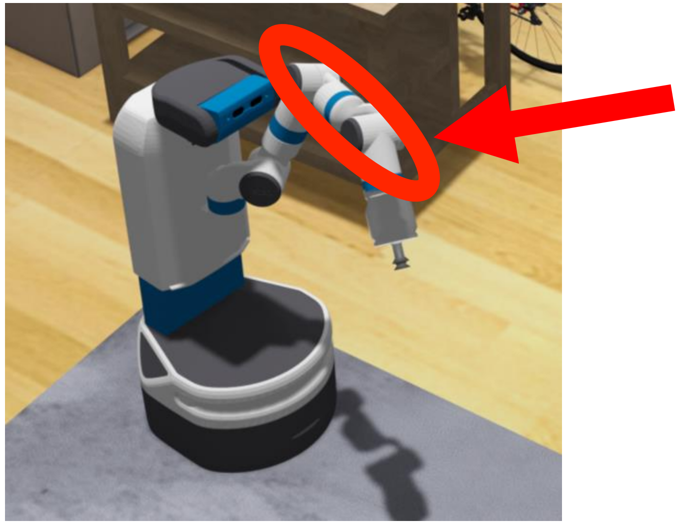


Results - New Tasks

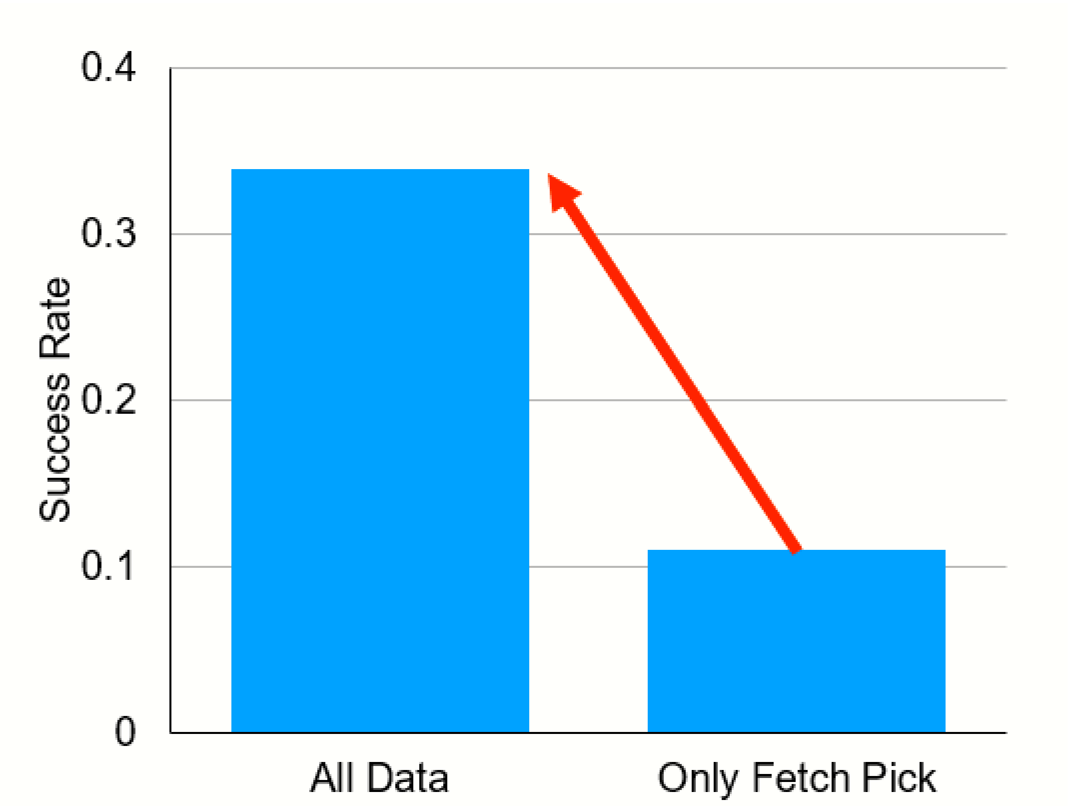
	Static Manipulation	Mobile Manipulation	Navigation	Games	Control
GEA (Ours)	**65%**	**54%**	66%	**47%**	32%
Per-Domain Baseline	58%	49%	**71%**	36%	**59%**

Results are for adapting LLaVA-1.5 7B

Results - New Embodiments



Generalize to new arm lengths



Embodiment Prompt

Agent: Fetch Robot. Actions: delta joint position. **Agent arm length=0.8m**. Group: mobile manipulation. Simulator: Habitat. Camera: head camera. Instruction: Pick an apple.

Future Work

- Adaption to new environments by investigating:
  - # of new demonstrations vs. success rate with supervised fine-tuning
  - # of experiences vs. success rate with reinforcement learning
- Investigating how online data collection can boost performance
- Insights from model training

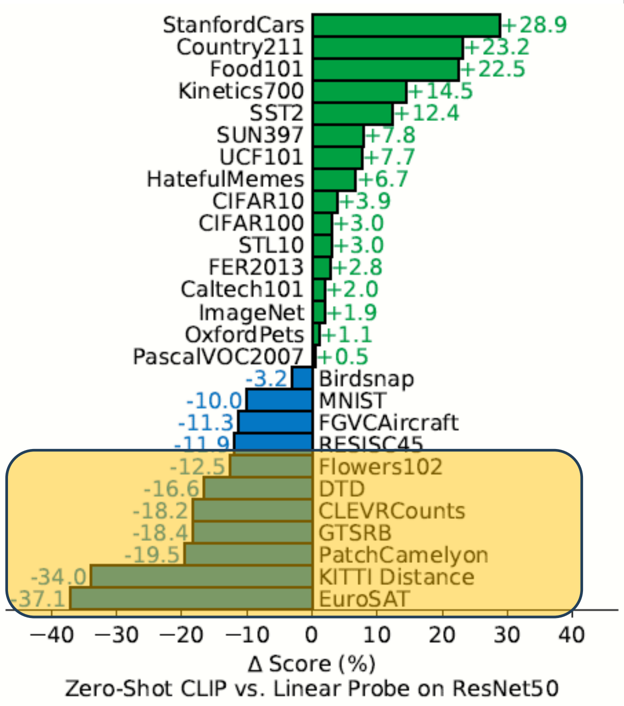
Is Generalization Solved? Are We Done?

- Positive View:
  - Bypass distribution shift!
  - Train on as much “in-distribution data” as possible
  - Nothing is OOD any more

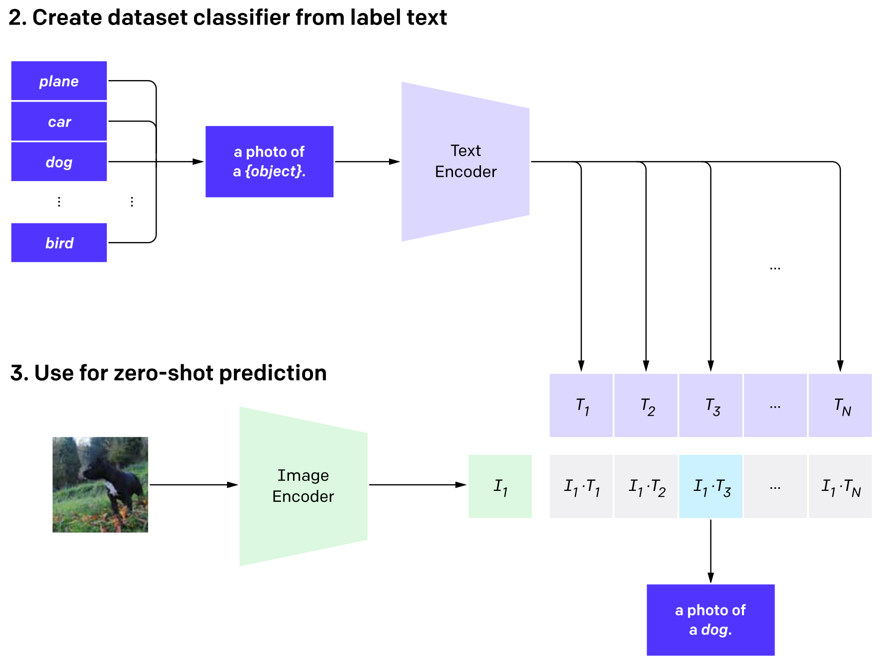
?

Is Generalization Solved? Are We Done?

- Positive View:
  - Bypass distribution shift!
  - Train on as much "in-distribution data" as possible
  - Nothing is OOD any more



[Radford et al., Learning Transferable Visual Models From Natural Language Supervision]



Is Generalization Solved? Are We Done?

- Skeptical View:
  - This is a “brute-force” approach – is it really scalable?

  - Lots of “sub-distributions” without sufficient statistical support.
    - This could be the data you care about!

  - Practically, clearly still under-performs and biased
    - US-centric, not “in-the-wild” distributions, etc.
    - How much do we need to soak up “literally all” the distributions we care about?
    - Generalist **vision** models still resist

  - **Something we might want to do:** Finetune to our data!
    - Above robotics work is an example!

How to Improve Robustness?

	In-Distribution	Out-of-Distribution			
	IN	IN-V2	IN-Adversarial	IN-Rendition	IN-Sketch
CLIP Zero-Shot	67.68	61.41	**30.60**	**56.77**	**45.33**
Vanilla FT	**83.66**	**73.82**	21.40	43.06	45.22

Zero-Shot and fine-tuned classification accuracy of CLIP ViT-B on ImageNet (IN) and its variants.
The fine-tuning dataset is ImageNet.

Unconstrained optimization only encourages *fitting* to the new data

$$\min_{\boldsymbol{W} | (x,y) \in \mathcal{D}_{train}} \mathcal{L}(x, y; \boldsymbol{W})$$

Wortsman, Mitchell, et al. "Robust fine-tuning of zero-shot models." CVPR 2022.

40

Pre-trained Robustness

- Pre-trained models do have great generalization capability
  - Some OOD-detection and robustness capabilities

- **Question:** How do we preserve this during finetuning?

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML, 2021.

Preservation of Pre-trained Robustness

- L2-SP
  - Imposes L2 regularization on the difference between the fine-tuned model and the pre-trained model. $L(\theta) = \tilde{L}(\theta) + \frac{\lambda}{2}||\theta - \theta_0||_2^2$
- WiSE-FT
  - Linearly interpolate between a fine-tuned model and its pre-trained initialization.
  - Works very well for vision-language models

Hypothesis: unconstrained optimization to target leads to worse robustness.

Xuhong, L. I., Yves Grandvalet, and Franck Davoine. "Explicit inductive bias for transfer learning with convolutional networks." ICML, 2018.
Wortsman, Mitchell, et al. "Robust fine-tuning of zero-shot models." *CVPR* 2022.

42

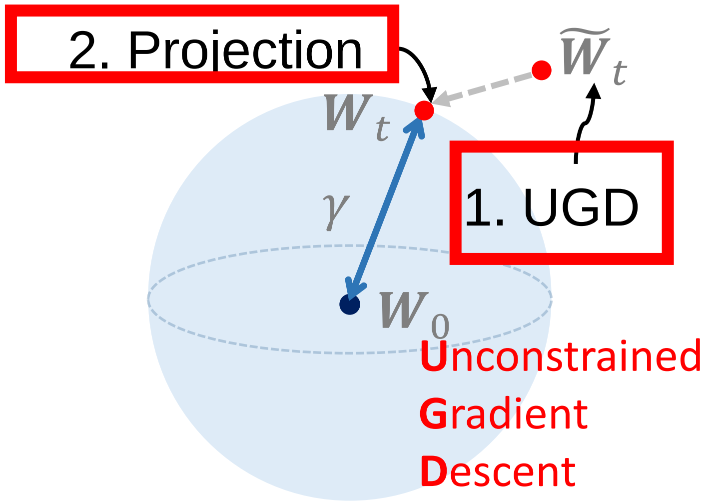
Projected Gradient Method

$$\min_{\boldsymbol{W}|(x,y)\in\mathcal{D}_{train}} \mathcal{L}(x,y;\boldsymbol{W}) \ s.t. \left|\left|\boldsymbol{W}-\boldsymbol{W}_0\right|\right| \le \gamma$$

- Projected Gradient Descent

$$\widetilde{\boldsymbol{W}}_t = SGD(x,y|\boldsymbol{W}_{t-1})$$

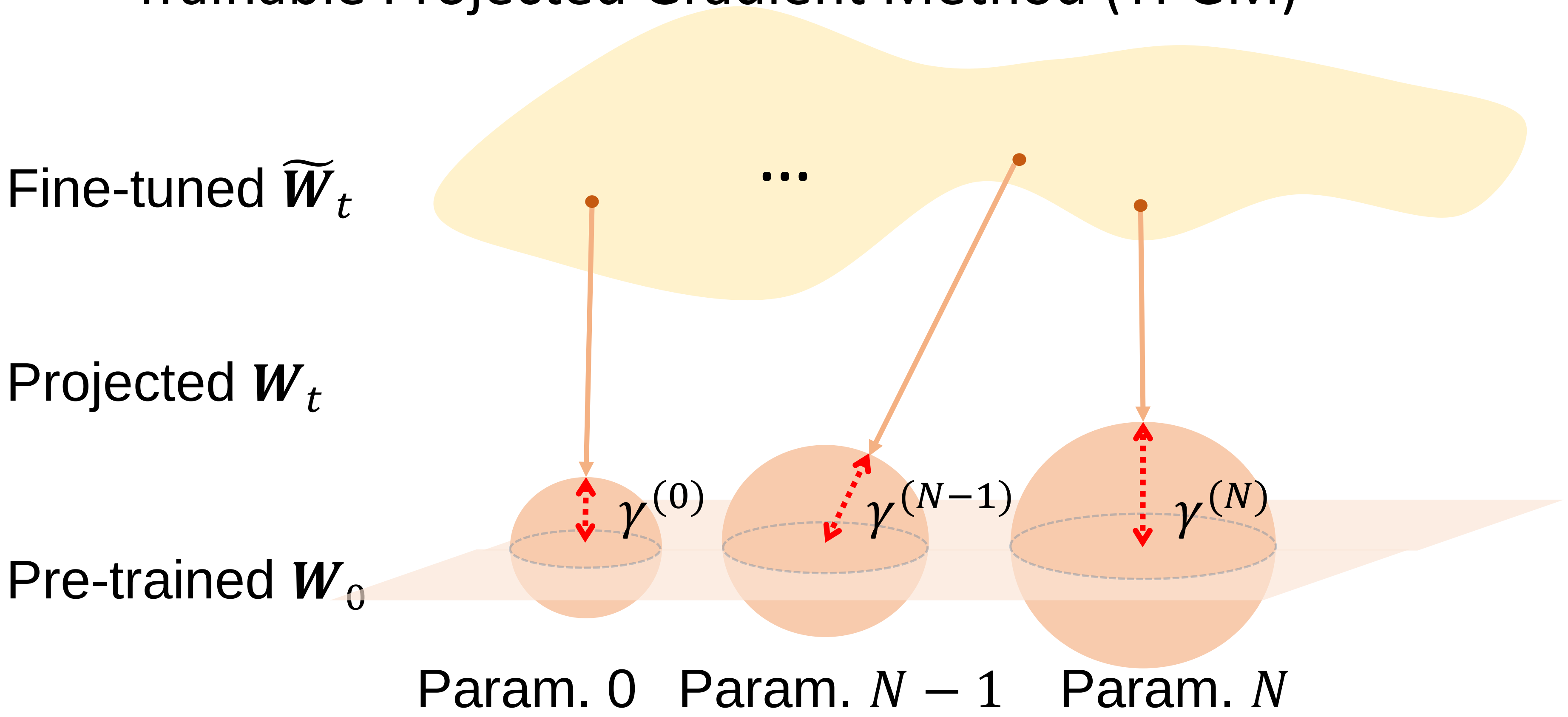
$$\boldsymbol{W}_t = \Pi\left(\widetilde{\boldsymbol{W}}_t, \boldsymbol{W}_0; \gamma\right)$$



Π defines a (**differentiable**) *projection function* and γ is the projection radius

Trainable Projected Gradient Method

- Trainable Projected Gradient Method (TPGM)



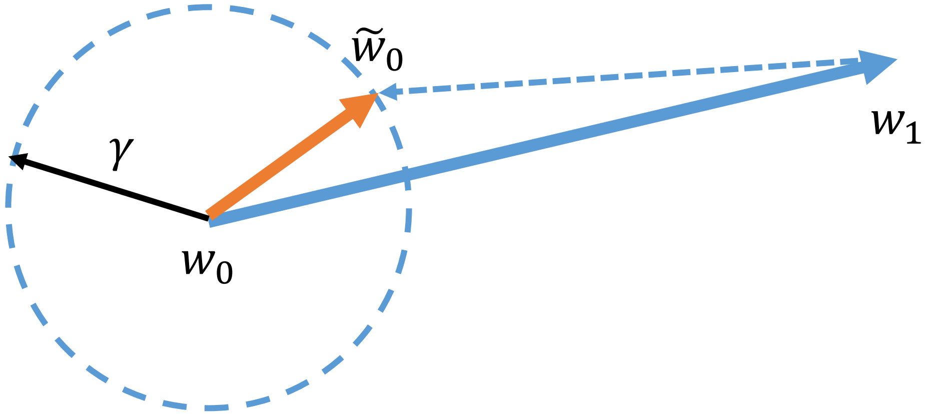
- Open Questions
  - *Which* layers to fine-tune?
  - *How much* to fine-tune?
  - Not feasible to specify a different constraint for each layer .

Tian, Junjiao, et al. "Trainable projected gradient method for robust fine-tuning." ***CVPR 2023.***

44

Our Prior Work: TPGM and FTP

TPGM and FTP use *outer loop bi-level optimization* for robust training



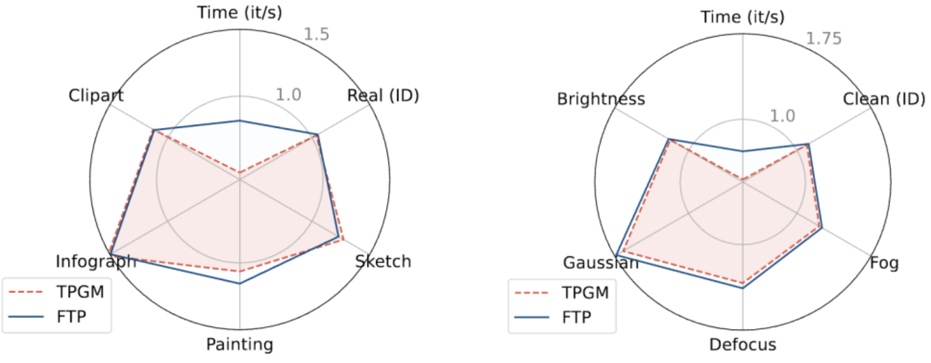
$$\underbrace{\min_{\lambda,\gamma|(x,y)\in\mathcal{D}_{val}}}_{\text{Step 2}} \underbrace{\min_{\theta|(x,y)\in\mathcal{D}_{tr}}}_{\text{Step 1}} \mathcal{L}(x,y;\theta,\lambda,\gamma) \quad \text{s.t.} \quad \underbrace{\|\theta-\theta_0\|_* \le \gamma}_{\text{Step 3}}$$

Algorithm 1: TPGM

Data: $\mathcal{D}_{tr}, \mathcal{D}_{val}$
Result: θ
Initialize $\theta_0^* = \theta_0, \gamma_0 = \epsilon$
for $t = \{0, ..., T-1\}$ **do**

Step 1 $\quad \theta_{t+1} = \arg\min_\theta \mathcal{L}(x,y;\theta_t^*) \quad x,y \in \mathcal{D}_{tr}$
Step 2 $\quad \gamma_{t+1} = \text{ProjectTune}(\mathcal{D}_{val}, \theta_0, \theta_{t+1}, \gamma_t)$
Step 3 $\quad \theta_{t+1}^* = \Pi(\theta_0, \theta_{t+1}, \gamma_{t+1})$

$$\Pi_{l2}(\theta_0,\theta_t,\gamma) : \tilde{\theta} = \theta_0 + \frac{1}{\max\left(1, \frac{\|\theta_t-\theta_0\|_2}{\gamma}\right)}(\theta_t-\theta_0).$$



(b) Image Classification (c) Semantic Segmentation

Junjiao Tian
Robotics Ph.D.

Can we simplify this to reduce complexity/computation?

Tian et al., CVPR 2023 / NeurIPS 2023

Selective Projection Decay

Learning the New Without Forgetting the Old Even More Efficiently

Junjiao Tian

Robotics Ph.D.

Chengyue Huang

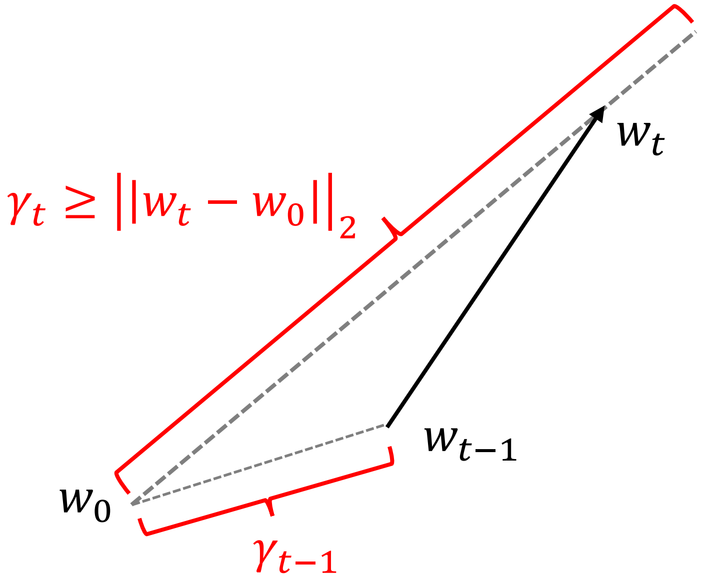
ML Ph.D.

Tian, Junjiao, Chengyue Huang, and Zsolt Kira. ”Selective Projection Decay for Robust Fine-Tuning“, **NeurIPS *2024*.**

46

Observations

- TPGM/FTP **grows** and **shrinks** the projection radius.
    - When the radius grows, it often provides no regularization (no projection).
    - The regularization effect mainly comes from the shrinkage of the projection radius.



γ: constraints
w_0: Initialization

Tian, Junjiao, Chengyue Huang, and Zsolt Kira. "Selective Projection Decay for Robust Fine-Tuning", **NeurIPS *2024***.

Hypothesis

- No need to explicitly maintain a set of projection radii.
- No need to know when to grow.

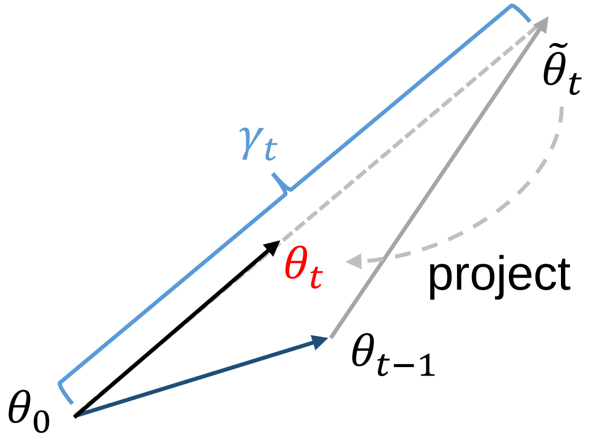
- Just need to know when to shrink/apply regularization.
  - Do this per layer/iteration
  - **When:** Alignment between gradient and direction to original weights
  - **How much:** $\gamma_t = ||w_t - w_0||_2$

Tian, Junjiao, Chengyue Huang, and Zsolt Kira. "Selective Projection Decay for Robust Fine-Tuning", **NeurIPS *2024***.

Selective Projection Decay (SPD)

Selecting criterion

- L2-SP: $L(\theta) = \tilde{L}(\theta) + \frac{\lambda}{2}||\theta - \theta_0||_2^2$
- Hyper-optimize λ: $\nabla\lambda = \frac{\partial f(\theta_t)}{\partial \lambda} = \frac{\partial f(\theta_t)}{\partial \theta}^T \frac{\theta_t}{\partial \lambda} = \alpha * -g_{t+1}^T(\theta_t - \theta_0)$
  - This was the gradient calculation in Fast Trainable Projection $\nabla\gamma \propto g_t^T(\theta_{t-1} - \theta_0)$
- Selection condition: $c_t = c_{t-1} - g_t^T(\theta_{t-1} - \theta_0) < 0$



γ_t: constraints
θ_0: initialization
$\tilde{\theta}_t$: unconstrained update

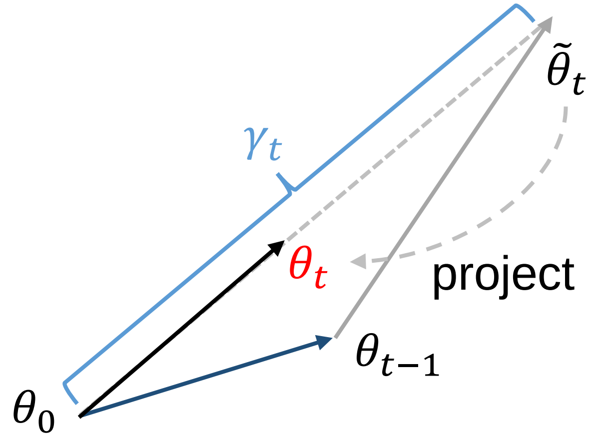
Selective Projection Decay (SPD)

Selecting criterion

- L2-SP: $L(\theta) = \tilde{L}(\theta) + \frac{\lambda}{2}||\theta - \theta_0||_2^2$
- Hyper-optimize λ: $\nabla\lambda = \frac{\partial f(\theta_t)}{\partial \lambda} = \frac{\partial f(\theta_t)}{\partial \theta}^T \frac{\theta_t}{\partial \lambda} = \alpha * -g_{t+1}^T(\theta_t - \theta_0)$
- Selection condition: $c_t = c_{t-1} - g_t^T(\theta_{t-1} - \theta_0) < 0$

Projection coefficient

- L2-SP is a projection: $\theta_p = \theta_t - \left(1 - \frac{\gamma}{\max\{\gamma, ||\theta_t - \theta_0||_2\}}\right) * (\theta_t - \theta_0)$
- Deviation: $\gamma_t = ||\theta_t - \theta_0||_2$
- Deviation ratio: $r_t = \frac{\max\{0, \gamma_t - \gamma_{t-1}\}}{\gamma_t}$
- $\theta_t \leftarrow \theta_t - \lambda \frac{\max\{0, \gamma_t - \gamma_{t-1}\}}{\gamma_t}(\theta_t - \theta_0)$



γ_t: constraints
θ_0: initialization
$\tilde{\theta}_t$: unconstrained update

Selective Projection Decay

Algorithm 1: Adam with L2-Regularization

Initialize $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $t \leftarrow 0$

While θ_t not converged

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$

$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$

$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

Bias Correction

$\widehat{m_t} \leftarrow \frac{m_t}{1-\beta_1^t}, \hat{v_t} \leftarrow \frac{v_t}{1-\beta_2^t}$

Update

$\theta_t \leftarrow \theta_{t-1} - \frac{\alpha \widehat{m_t}}{\sqrt{\hat{v_t}}+\epsilon}$

$\theta_t \leftarrow \theta_t - \lambda \underline{\alpha} (\theta_t - \theta_0)$

Learning rate

Algorithm 2: Adam with Selective L2-Reg.

Initialize $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $t \leftarrow 0$, $c_0 \leftarrow 0$

While θ_t not converged

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$

$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$

$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

Bias Correction

$\widehat{m_t} \leftarrow \frac{m_t}{1-\beta_1^t}, \hat{v_t} \leftarrow \frac{v_t}{1-\beta_2^t}$

Update

$\theta_t \leftarrow \theta_{t-1} - \frac{\alpha \widehat{m_t}}{\sqrt{\hat{v_t}}+\epsilon}$

$\boxed{c_t = c_{t-1} - g_t^\intercal (\theta_{t-1} - \theta_0)}$ **1, Condition**

If $c_t < 0$:

$\theta_t \leftarrow \theta_t - \lambda \underline{r_t} (\theta_t - \theta_0)$

2, Deviation Ratio

Algorithm 1: Adam with L2-SP

Initialize $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $t \leftarrow 0$
While θ_t not converged
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$
$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
$v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$
Bias Correction
$\widehat{m_t} \leftarrow \frac{m_t}{1-\beta_1^t}$, $\widehat{v_t} \leftarrow \frac{v_t}{1-\beta_2^t}$
Update
$\theta_t \leftarrow \theta_{t-1} - \frac{\alpha \widehat{m_t}}{\sqrt{\widehat{v_t}}+\epsilon}$
$\theta_t \leftarrow \theta_t - \lambda\alpha(\theta_t - \theta_0)$

Algorithm 2: Adam with SPD

Initialize $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $t \leftarrow 0$, $c_0 \leftarrow 0$
While θ_t not converged
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$
$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
$v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$
Bias Correction
$\widehat{m_t} \leftarrow \frac{m_t}{1-\beta_1^t}$, $\widehat{v_t} \leftarrow \frac{v_t}{1-\beta_2^t}$
Update
$\theta_t \leftarrow \theta_{t-1} - \frac{\alpha \widehat{m_t}}{\sqrt{\widehat{v_t}}+\epsilon}$
$c_t = c_{t-1} - g_t^\intercal(\theta_{t-1} - \theta_0)$
If $c_t < 0$:
$\theta_t \leftarrow \theta_t - \lambda r_t(\theta_t - \theta_0)$

More intuitive hyper-parameter (λ) tuning

- No regularization ($\lambda = 0$): the projection radius is 1.
- Weak regularization ($1 \geq \lambda > 0$): the projection radius lies between $||\theta_t - \theta_0||_2$ and $||\theta_{t-1} - \theta_0||_2$. Within this range, layers will expand.
- Strong regularization ($\lambda > 1$): the projection radius lies between 0 and $||\theta_{t-1} - \theta_0||_2$. In this range, it's possible that regularized layers can contract.

Experiments

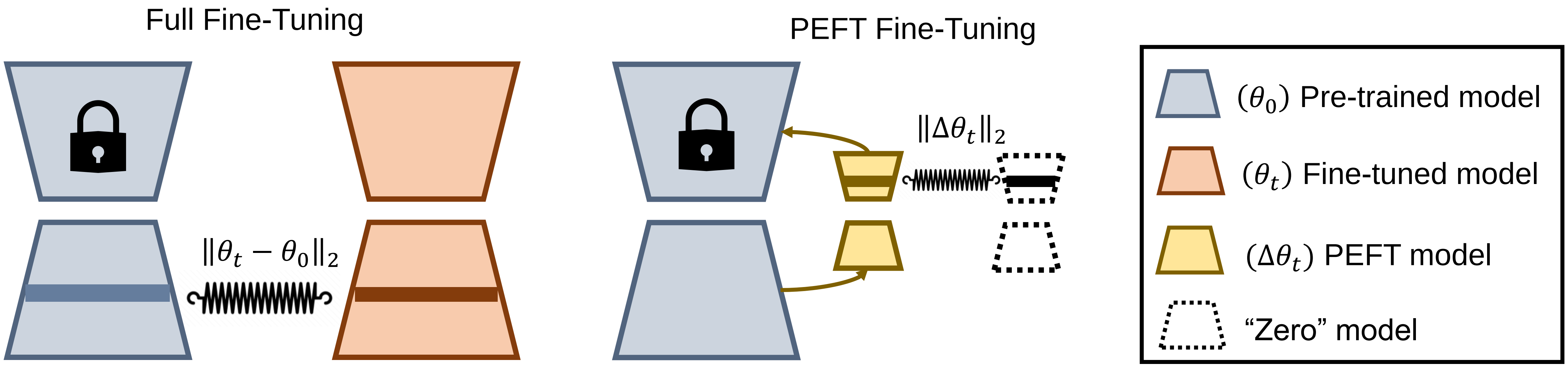
- Selective regularization is on par with predecessors and outperforms other methods.

Table 3: ImageNet Fine-Tuning Result using CLIP ViT-Base. SPD outperforms more complicated algorithms and beats L2-SP by 8.8% by selectively imposing regularization.

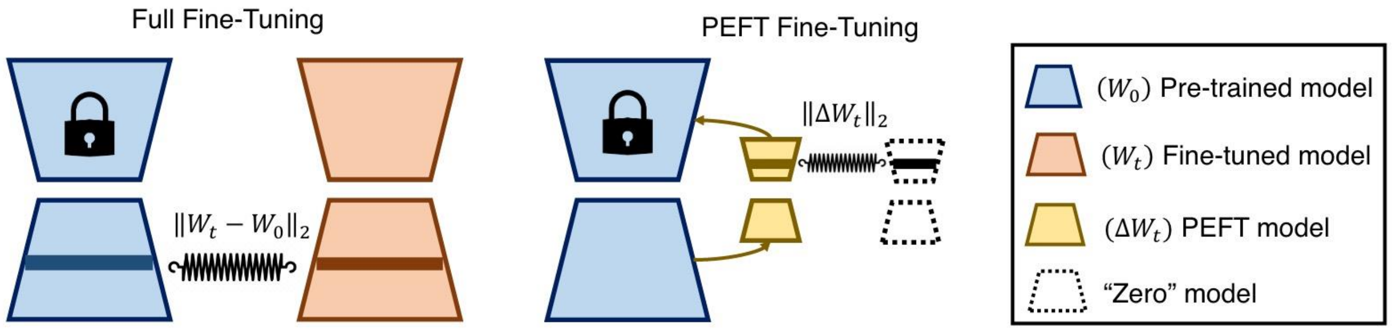
	ID	OOD				Statistics	
	Im	Im-V2	Im-Adversarial	Im-Rendition	Im-Sketch	OOD Avg.	Avg.
Zero-Shot	67.68	61.41	30.60	56.77	45.53	48.58	52.40
vanilla FT	83.66	73.82	21.40	43.06	45.52	46.98	54.29
Linear Prob.	78.25	67.68	**26.54**	52.57	48.26	48.76	54.66
LP-FT [19]	82.99	72.96	21.08	44.65	47.56	46.56	53.85
L2-SP [13]	83.44	73.2	20.55	43.89	46.60	46.06	53.54
FTP [11]	84.19	74.64	26.50	47.23	50.23	49.65	56.56
Adam-SPD	**84.21**	**74.83**	25.42	**49.09**	**51.18**	**50.13**	**56.95**

Compatible with Parameter-Efficient Fine-Tuning

- Our method reduces to selective weight decay when working with Parameter Efficient Fine-Tuning (PEFT) methods.



LLaMA PEFT Fine-Tuning Experiments



PEFT	LLM	Optimizer	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Avg.
Series	LLaMA_{7B}	AdamW	63.0	79.2	76.3	67.9	75.7	74.5	57.1	72.4	70.8
		Adam-SPD (1.0)	**68.3**	**80.4**	**77.4**	**81.6**	**79.7**	**79.4**	**63.5**	**78.4**	**76.1**
Parallel	LLaMA_{7B}	AdamW	67.9	76.4	**78.8**	69.8	78.9	73.7	57.3	75.2	72.3
		Adam-SPD (1.0)	**68.8**	**80.9**	78.3	**82.0**	**80.8**	**80.0**	**63.1**	**78.0**	**76.5**
LoRA	LLaMA_{7B}	AdamW	68.9	80.7	77.4	78.1	78.8	77.8	61.3	74.8	74.7
		Adam-SPD (0.7)	**69.1**	**82.8**	**78.9**	**84.8**	**80.7**	**80.9**	**65.8**	**79.2**	**77.8**
LoRA	LLaMA_{13B}	AdamW	72.1	83.5	80.5	80.5	**83.7**	82.8	68.3	82.4	80.5
		Adam-SPD (1.2)	**72.9**	**85.6**	**80.7**	**92.0**	**83.7**	**85.6**	**71.6**	**85.6**	**82.2**

Compatibility with PEFT methods

- SPD regularizes $||\theta_t - \theta_0||_2$ for full fine-tuning and $||\Delta\theta_t||_2$ for PEFT fine-tuning
- SPD can also improve the performance of PEFT methods (e.g. LoRA, series adapters, parallel adapters)

What about Vision-Language Models (VLMs)?

- Robustness and distribution shift is much more complicated!

- Many types of shift possible

- **Distribution Shifts to Images**
  - IV-VQA
  - CV-VQA
- **Distribution Shifts to Questions**
  - VQA-Rephrasings
  - VQA-LOL
- **Distribution Shifts to Answers**
  - VQA-CP
- **Distribution Shifts to Multi-modalities.**
  - VQA-GEN
  - VQA-CE
  - VQA-VS Adversarial Distribution Shifts
  - AVQA
- **Adversarial**
  - **AdVQA**
- **Far OOD:** TextVQA, VizWiz, OK-VQAv2

Visual Question Answering (VQA) Fine-Tuning Experiments

	ID	Near OOD							Far OOD		
		Vision		Question	Answer	Multimodal	Adversarial				
	VQAv2	IV-VQA	CV-VQA	VQA-Rephrasings	VQA-CP v2	VQA-CE	AdVQA	TextVQA	VizWiz	OK-VQA
Zero-Shot	54.42	63.95	44.72	50.10	54.29	30.68	30.46	14.86	16.84	28.60
Vanilla FT(LoRA)	86.29	94.43	**69.36**	78.90	86.21	71.73	49.82	42.08	22.92	48.30
Linear Prob.	78.24	87.83	63.87	69.61	78.48	61.66	42.90	29.61	18.80	42.27
LP-FT(LoRA)	85.97	93.30	65.93	76.49	86.16	72.73	45.68	31.41	19.01	43.27
WiSE-FT(LoRA)	71.36	85.06	64.55	66.42	70.89	48.74	43.95	36.98	22.41	42.35
Adam-SPD(LoRA)	**87.39**	**95.25**	68.85	**79.48**	**87.27**	**73.52**	**50.90**	**43.56**	**23.05**	**50.11**

New setting: robust fine-tuning for VQA

- ID dataset: VQAv2
- OOD datasets
    - Distribution shifts to images: IV-VQA, CV-VQA
    - Distribution shifts to questions: VQA-Rephrasings
    - Distribution shifts to multi-modalities: VQA-CE
    - Adversarial distribution shifts: AdVQA
    - Far OODs: TextVQA, VizWiz, OK-VQAv2

SPD shows competitiveness across ID, near OOD, and far OOD datasets on multimodal tasks.

Finetuning and Forgetting are common!

We anticipate a number of places for this to be useful!

- Training vision-language-action models for robotics!
  - Some can afford to co-finetune with VQA, etc. but difficult!
- Finetuning to large open-vocabulary corpora (e.g. Wikipedia)
- Multi-task finetuning from pre-trained model

Conclusions

- Distribution shift is ***still*** a problem
  - Private, in-the-wild data
- One approach: Finetune!
  - Question: How to do so robustly? **Per-layer/iteration constraint of gradient update**
  - Not the only choice: Retrieval/RAG, etc.
- Lots of other “distributions” of data!
  - Reasoning, planning, etc.
  - Current approach (o1): Show it the distribution
  - Other approaches?

Conclusions

- Already getting benefits of language!
  - Natural task specification
  - Semantic actions, Embodiment prompt

- Some other projects:
  - Long-form videos and memory
  - Fast 3D reconstruction for simulation
  - 3D question/answering agents
  - Minecraft – Learning from unstructured demos
  - Web GUI Agents

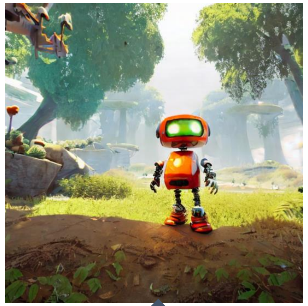
- Focus on:
  - Generalization
  - Long-Horizon / Long Context
  - Planning, Reasoning, Memory
  - Robustness

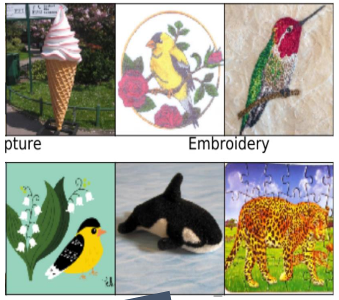
Reproducible Robotics -> Simulation

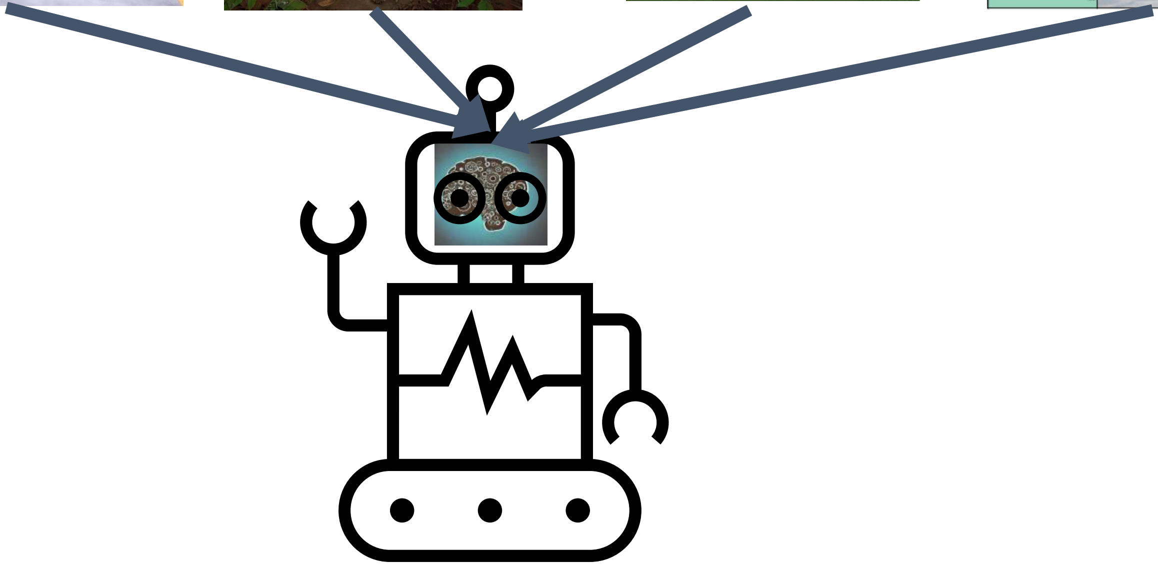
Discovering an Open World

Long-Horizon / Long-Context

Robust Fine-tuning







Acknowledgement and Questions

Yusuf Ali

CS Ph.D.

Jeremiah Coholich

Robotics Ph.D.

Shaunak Halbe

ML Ph.D.

Chengyue Huang

ML Ph.D.

Ram Ramrakhya

CS Ph.D. (co-advised with Dhruv Batra)

Andrew Szot

ML Ph.D. (co-advised with Dhruv Batra)

Karmesh Yadav

CS Ph.D. (co-advised with Dhruv Batra)

Abhinav Harish

M.S. Thesis Student

Brisa Maneech-otesuwan

Undergrad Student

Shivang Chopra

M.S. Student

SAMSUNG RESEARCH AMERICA

MIT-IBM Watson AI Lab

Meta

DARPA

NSF

Office of Naval Research Science & Technology

Open VLM/Multi-Modal Works?

Open VLM/Multi-Modal Works?

- Tokenization!
    - Images? Videos/Compressed Representations?

- Where to spend parameters and compute?
    - Unimodal encoders
    - Interaction / Fusion
    - Decoding

- Inference-time compute for MLLMs
    - A la OpenAI o1 model

- Interleave everything:
    - Full/partial modality data, “thought tokens”, decoding
    - Both at the input and output