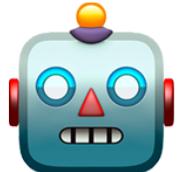


No Training? Towards Adapting Zero-Shot Models to Robotics Tasks

Samir Yitzhak Gadre
Columbia Artificial Intelligence and Robotics (CAIR) Lab

CVPR 2022 tutorial on vision-based robot learning

Motivation

- Large, pre-trained models have made a splash in:
 -  NLP
 -  Vision
- Many such models show *zero-shot* capabilities
- Ideally we can leverage these existing models in other domains:
 -  Robotics!

Overview (According to GPT-3)

create a script for a CVPR tutorial on using large, pre-trained models for robotics

Overview (According to GPT-3)

create a script for a CVPR tutorial on using large, pre-trained models for robotics

This CVPR tutorial will focus on using large, pre-trained models for robotics. We will firstly go through the basics of what a pre-trained model is and how it can be used for robotics. We will then explore how to use these models in different robotics applications, such as object detection and classification, navigation, and control. Finally, we will discuss some of the challenges associated with using large, pre-trained models in robotics and how to overcome them.

Overview (According to GPT-3)

create a script for a CVPR tutorial on using large, pre-trained models for robotics

This CVPR tutorial will focus on using large, pre-trained models for robotics. We will firstly go through the basics of what a pre-trained model is and how it can be used for robotics. We will then explore how to use these models in different robotics applications, such as object detection and classification, navigation, and control. Finally, we will discuss some of the challenges associated with using large, pre-trained models in robotics and how to overcome them.

- ✓ Pre-training at a high-level and success stories
- ✓ Language Modeling
- ✓ Vision-Language Modeling
- ✗ How to pre-train a robot with robot data

Overview (According to GPT-3)

create a script for a CVPR tutorial on using large, pre-trained models for robotics

This CVPR tutorial will focus on using large, pre-trained models for robotics. We will firstly go through the basics of what a pre-trained model is and how it can be used for robotics. We will then explore how to use these models in different robotics applications, such as object detection and classification, navigation, and control. Finally, we will discuss some of the challenges associated with using large, pre-trained models in robotics and how to overcome them.

-  Manipulation
-  Planning
-  Navigation
-  Object detection

Overview (According to GPT-3)

create a script for a CVPR tutorial on using large, pre-trained models for robotics

This CVPR tutorial will focus on using large, pre-trained models for robotics. We will firstly go through the basics of what a pre-trained model is and how it can be used for robotics. We will then explore how to use these models in different robotics applications, such as object detection and classification, navigation, and control. Finally, we will discuss some of the challenges associated with using large, pre-trained models in robotics and how to overcome them.

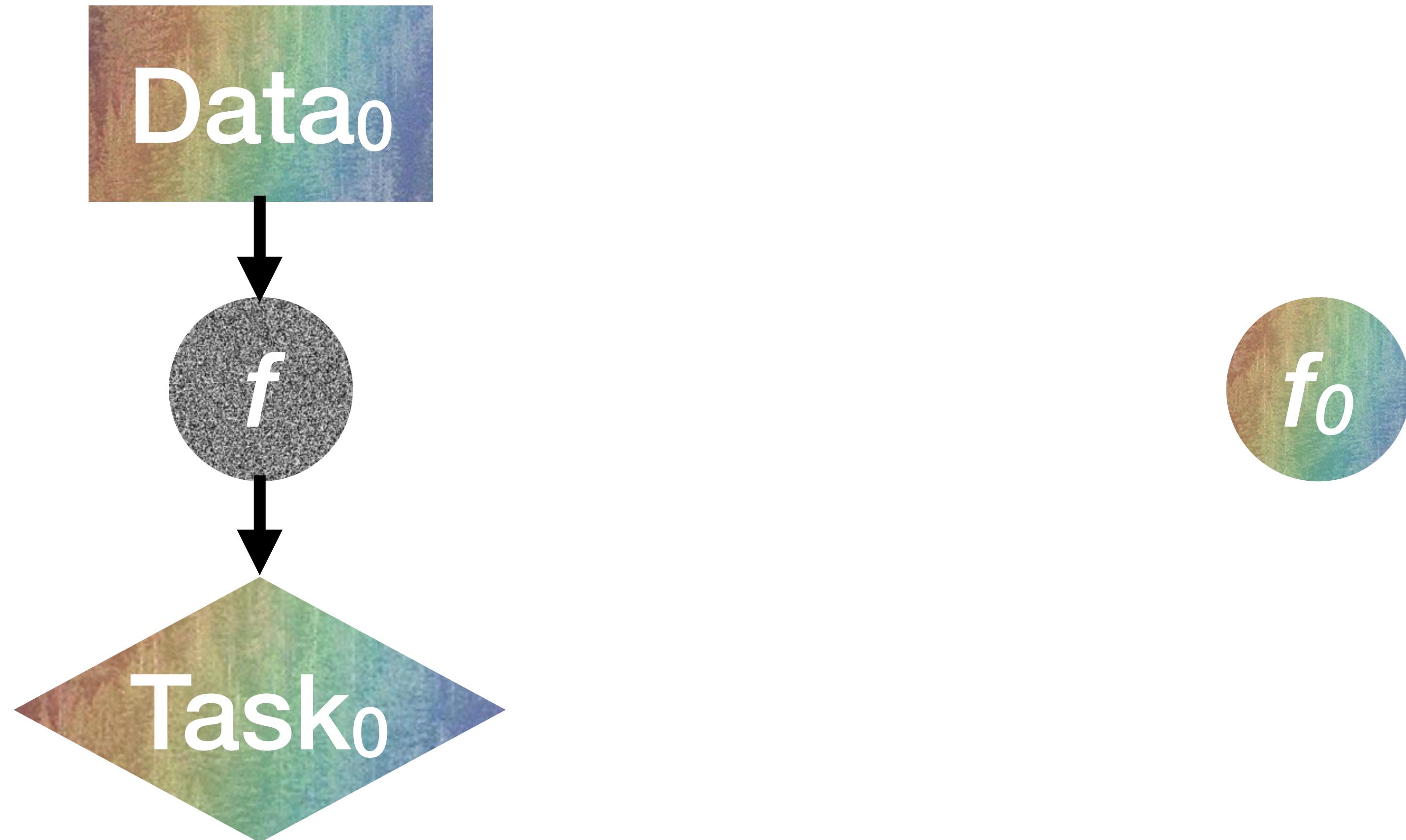


Limitations

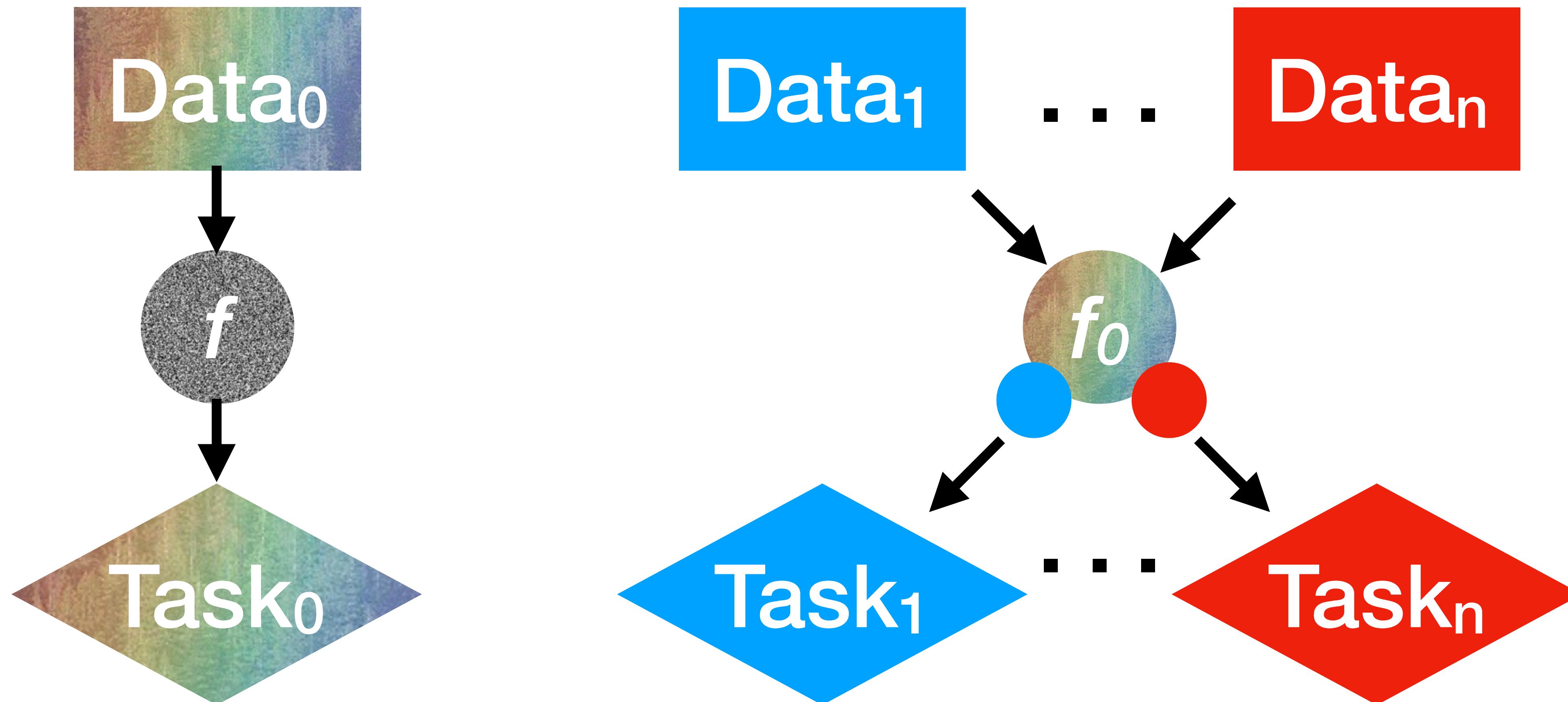


Exciting future directions!

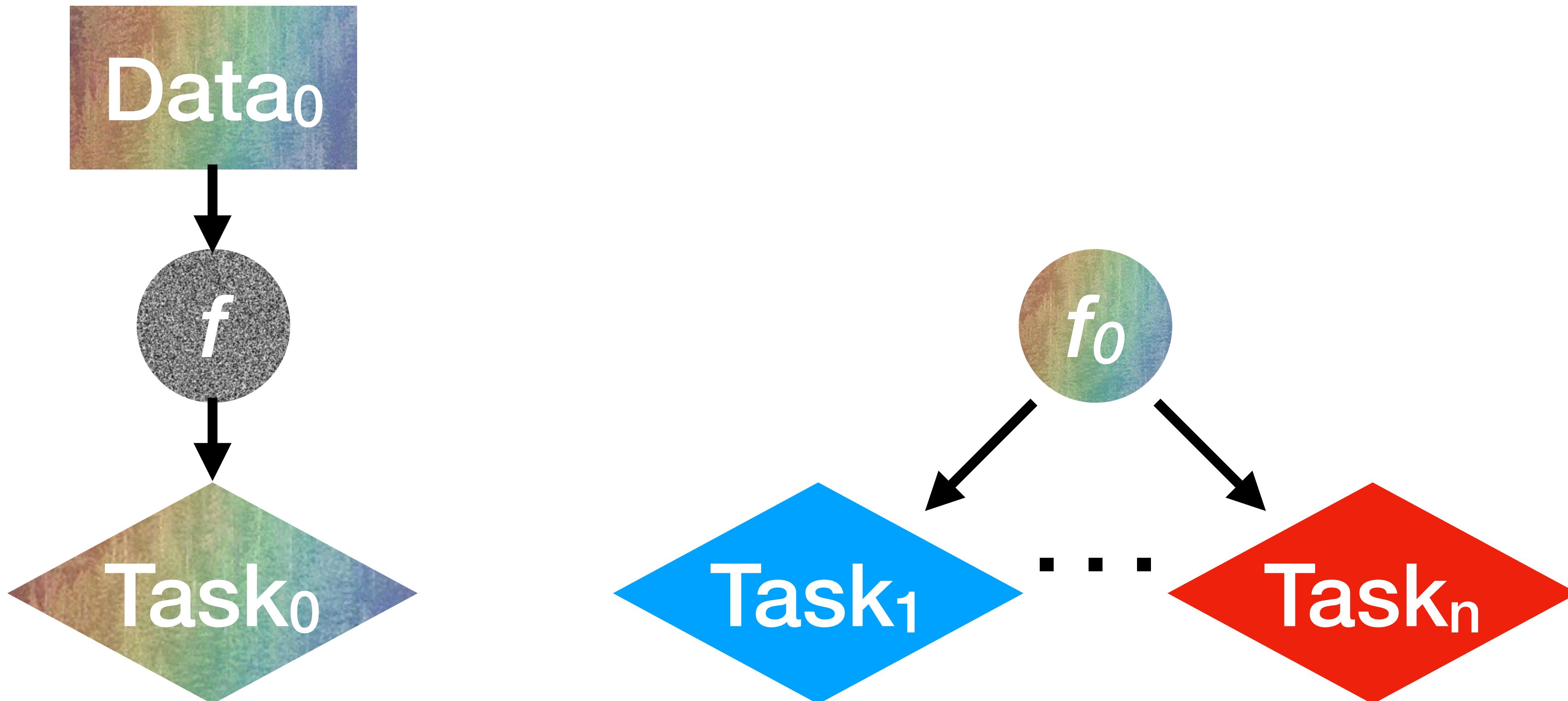
Pre-Training



Fine-Tuning

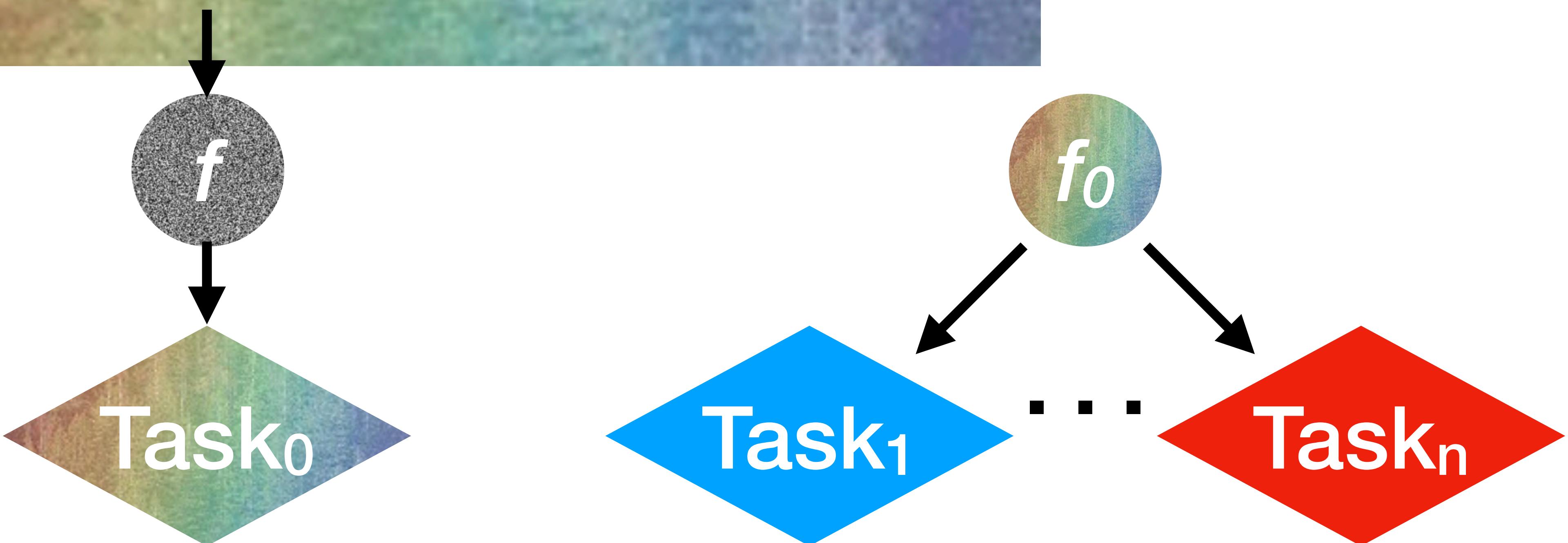


Zero-Shot Inference

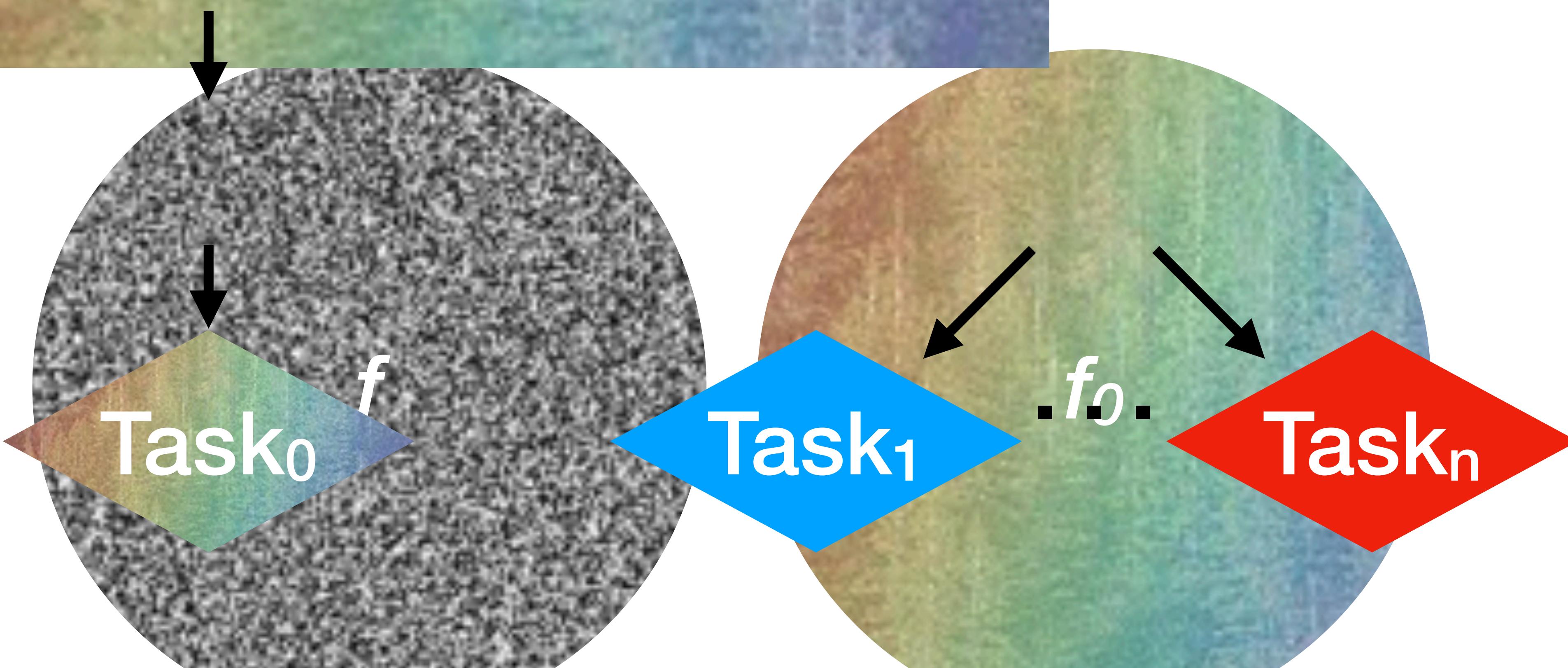


rence

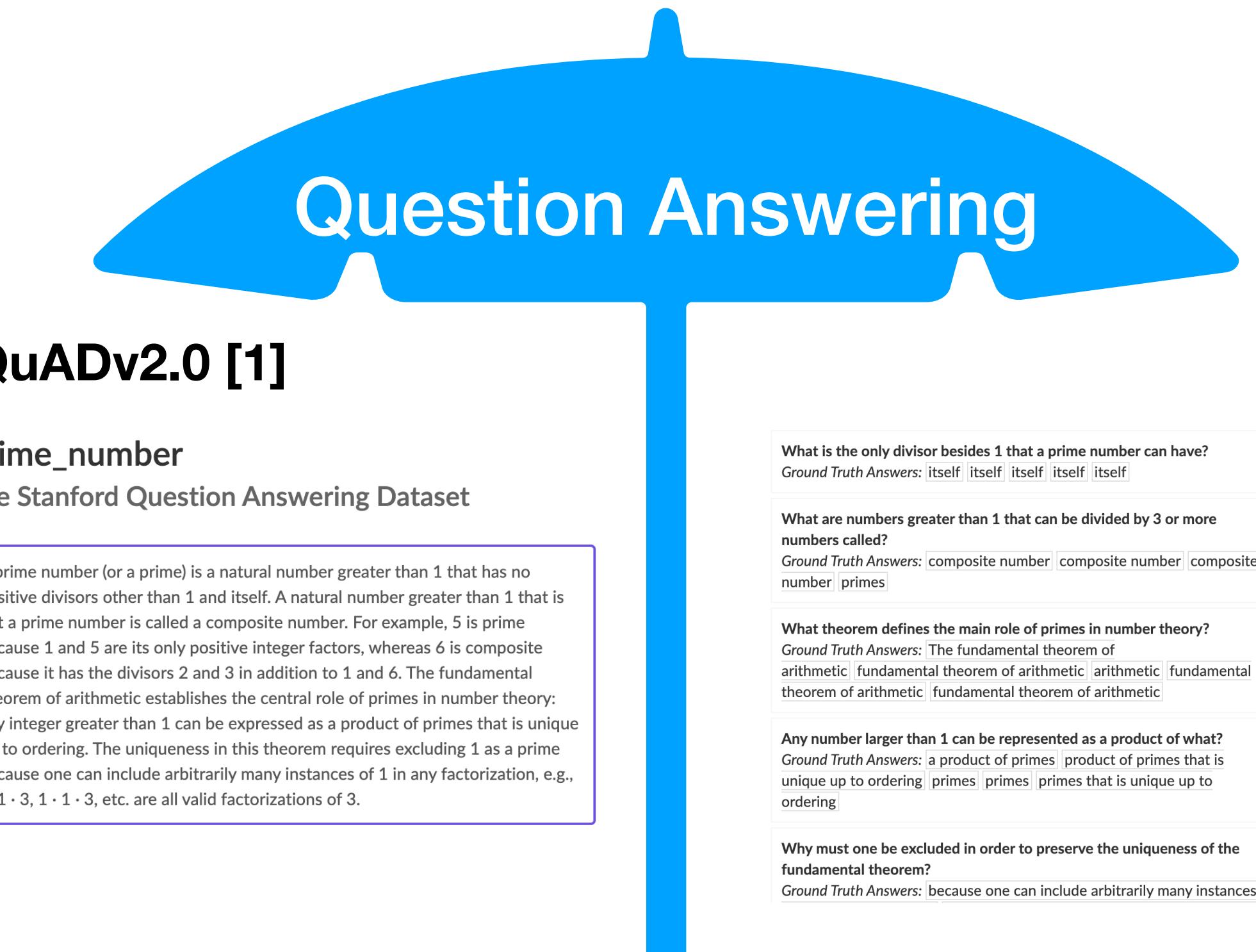
Data₀



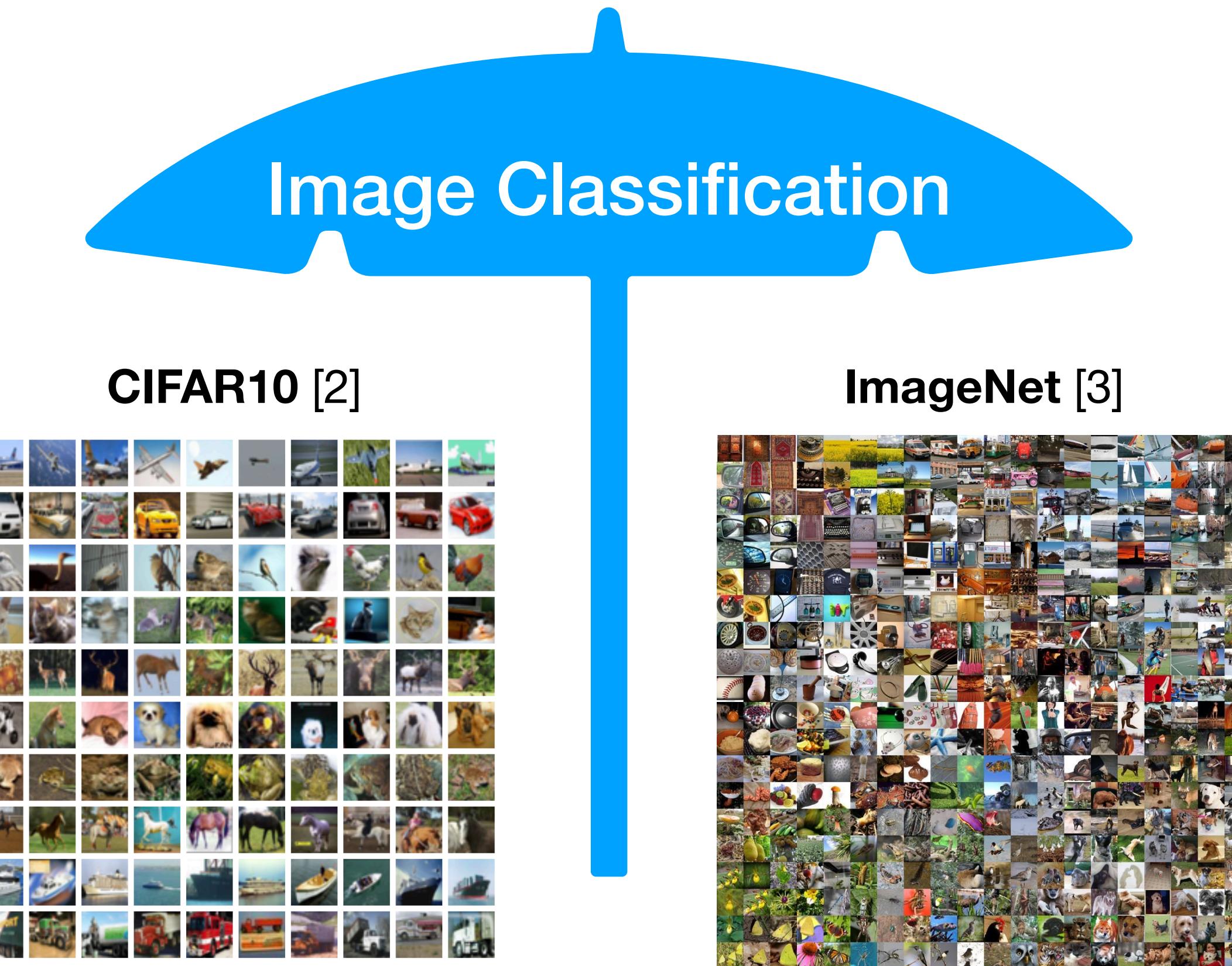
rence



Where is zero-shot doing “ok”?



Where is zero-shot doing “ok”?



[2] Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[3] Russakovsky et al. Imagenet large scale visual recognition challenge. IJCV 2015.

What seems to be getting us there?

- In NLP?
 - Large Language Models (LLM) i.a., GPT-3 [4]
- In Vision?
 - Vision-Language Models (VLM) i.a., CLIP [5]

[4] Brown et al. *Language Models are Few-Shot Learners*. NeurIPS 2020.

[5] Radford et al. Learning Transferable Vision Models From Natural Language Supervision. ICML 2021.

Language Model Pre-Training

Predict the next word in the _____

Language Model Pre-Training

Predict the next word in the sentence.

Zero-shot Inference (NLP)

If I have 7 apples and I eat 3 apples how many do I have left?

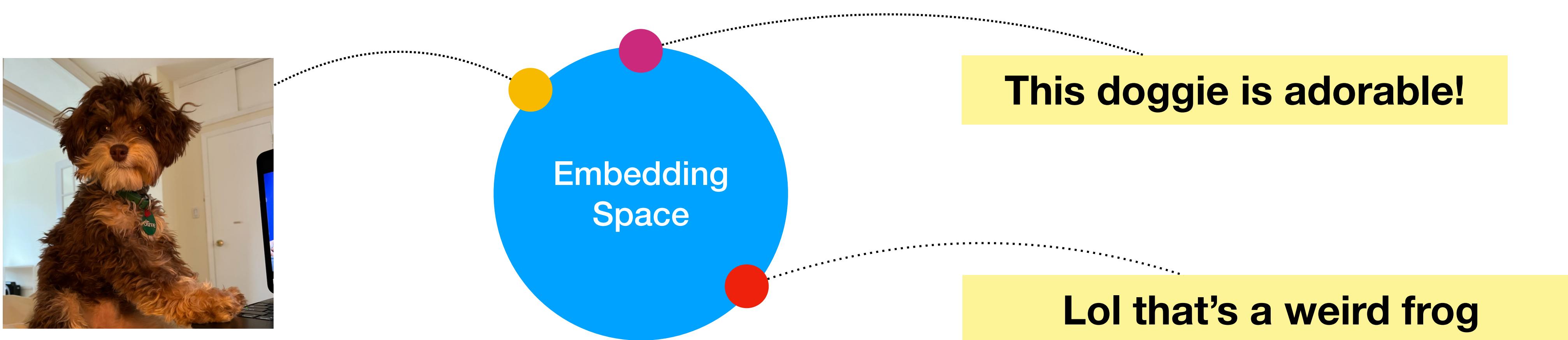
I would have 4 apples left.

Zero-shot Inference (NLP)

Is "I love computer vision!!!" a positive phrase?

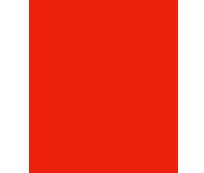
Yes, "I love computer vision!!!" is a positive phrase.

CLIP Pre-Training



Zero-shot Inference (Vision)

- With vision-language features we can create *arbitrary* image classifiers.

Input image	Prompts to create classifier	Similarity scores give class label
	A photo of a dog.	
	A photo of a frog.	

We've got some zero-shot capabilities

Lights (language), camera (vision),

What's missing for robotics?

Lights (language), camera (vision),

What's missing for robotics?

Lights (language), camera (vision),
action! (motor control)

Pre-train like we do in NLP and Vision?



I want pre-training too!

Pre-train like we do in NLP and Vision?

- People have definitely been up to cool things! For example, with datasets [7,8,9,10]
- Maybe, but it seems pretty hard at the present moment
 - Data scale?
 - Data diversity?
 - Scaling Reinforcement Learning (RL)?
 - Pre-training objective?

[7] Sharma et al. *Multiple Interactions Made Easy (MIME): Large Scale Demonstrations Data for Imitation*. CoRL 2018.

[8] Dasari et al. *RoboNet: Large-Scale Multi-Robot Learning*. arXiv 2019.

[9] Song et al. *Grasping in the Wild: Learning 6DoF Closed-Loop Grasping from Low-Cost Demonstrations*. RA-L 2020.

[10] Yen-Chen et al. *Learning to See before Learning to Act: Visual Pre-training for Manipulation*. ICRA 2020.

Pre-train like we do in NLP and Vision?



Hi GPT, I am a robot. Can I pre-train like you were pre-trained?

Pre-train like we do in NLP and Vision?



Hi GPT, I am a robot. Can I pre-train like you were pre-trained?

No, you cannot.

What now?



Hi GPT, I am a robot. Can I pre-train like you were pre-trained?

No, you cannot.

Before giving up, ask for help!



Hi GPT, I am a robot. Can I pre-train like you were pre-trained?

No, you cannot.

Before giving up, ask for help!



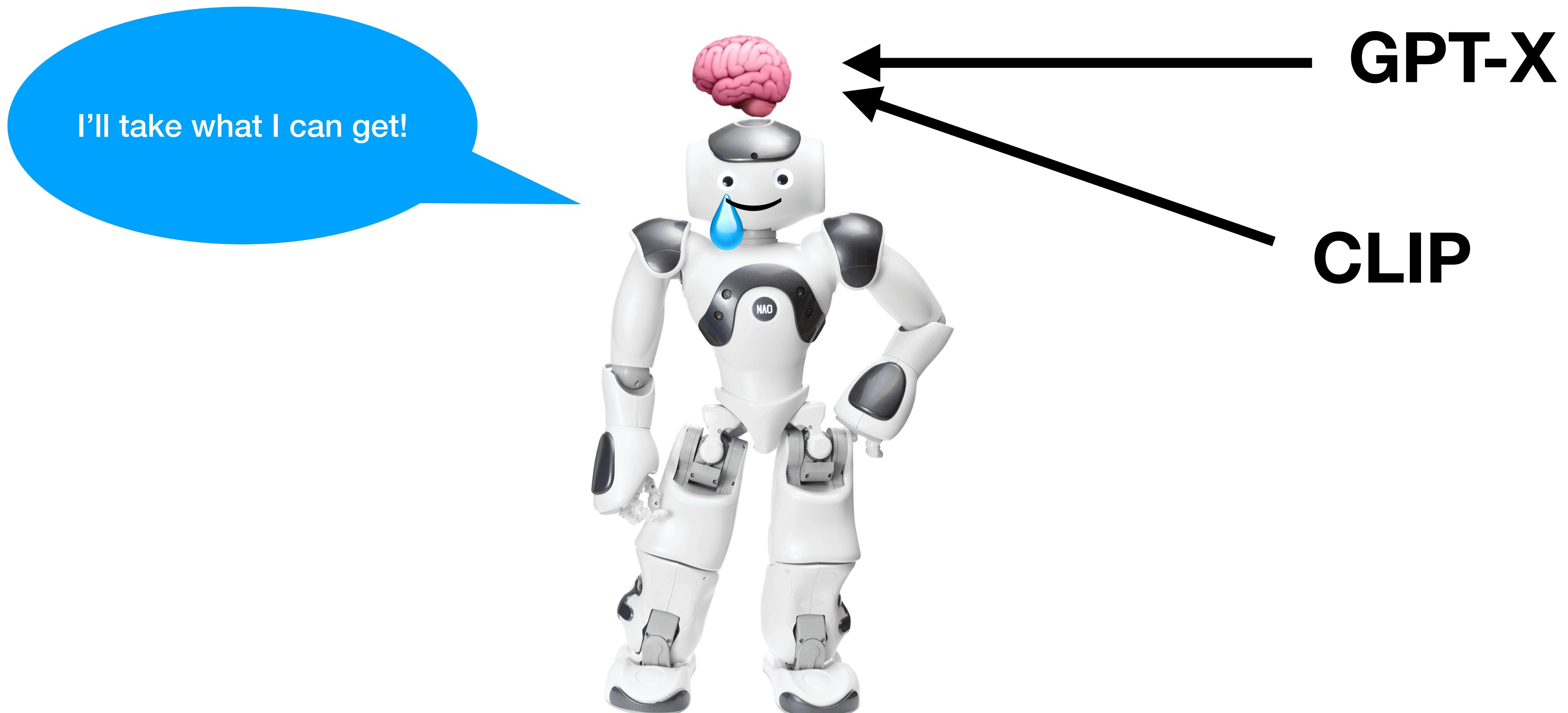
Hi GPT, I am a robot. Can I pre-train like you were pre-trained?

No, you cannot.

But can you help me think?

Yes, I can help you think, but you will not be able to pre-train like I was pre-trained.

Adapting pre-trained models for robotics



What does *adapting* mean?

- **Fine-tuning** a pre-trained model
 - [11] Shridhar et al. *CLIPort: What and Where Pathways for Robotic Manipulation*. CoRL 2021.
 - [12] Khandelwal et al. Simple but Effective: CLIP Embeddings for Embodied AI. CVRP 2022.
- **Zero-shot inference** directly with pre-trained models
 - [13] Huang et al. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. ICML 2022.
 - [14] Ahn et al. *Do As I Can, Not As I Say: Grounding Language in Robotic Affordances*. arXiv 2022.
 - [15] Zeng et al. *Socratic Models. Composing Zero-Shot Multimodal Reasoning with Language*. arXiv 2022.
 - [16] Gadre et al. *CLIP on Wheels: Open-Vocabulary Models are (Almost) Zero-Shot Object Navigators*. arXiv 2022.

What does *adapting* mean?

- **Fine-tuning** a pre-trained model
 - [11] Shridhar et al. *CLIPort: What and Where Pathways for Robotic Manipulation*. CoRL 2021.
 - [12] Khandelwal et al. Simple but Effective: CLIP Embeddings for Embodied AI. CVRP 2022.
- **Zero-shot inference** directly with pre-trained models
 - [13] Huang et al. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. ICML 2022.
 - [14] Ahn et al. *Do As I Can, Not As I Say: Grounding Language in Robotic Affordances*. arXiv 2022.
 - [15] Zeng et al. *Socratic Models. Composing Zero-Shot Multimodal Reasoning with Language*. arXiv 2022.
 - [16] Gadre et al. *CLIP on Wheels: Open-Vocabulary Models are (Almost) Zero-Shot Object Navigators*. arXiv 2022.

CLIPort: Big Ideas

- Separate **spatial** and **semantic** information streams
 - CLIP vision encoder to process visual input for semantics
 - CLIP language encoder to process input textual description of the task
- Learn the other components we might need to do pick-and-place by adding added parameters and **collecting human demos**

CLIPort: Task

- Inputs:

Top down RGB + D

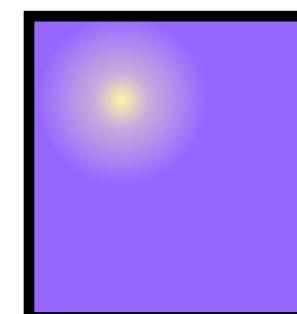


Language description of the task

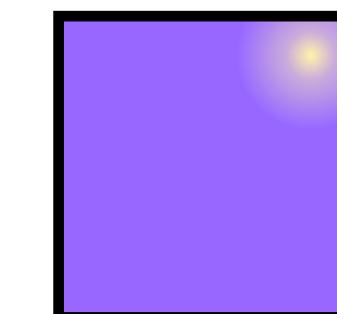
“pack the white tape
in the brown box”

- Output:

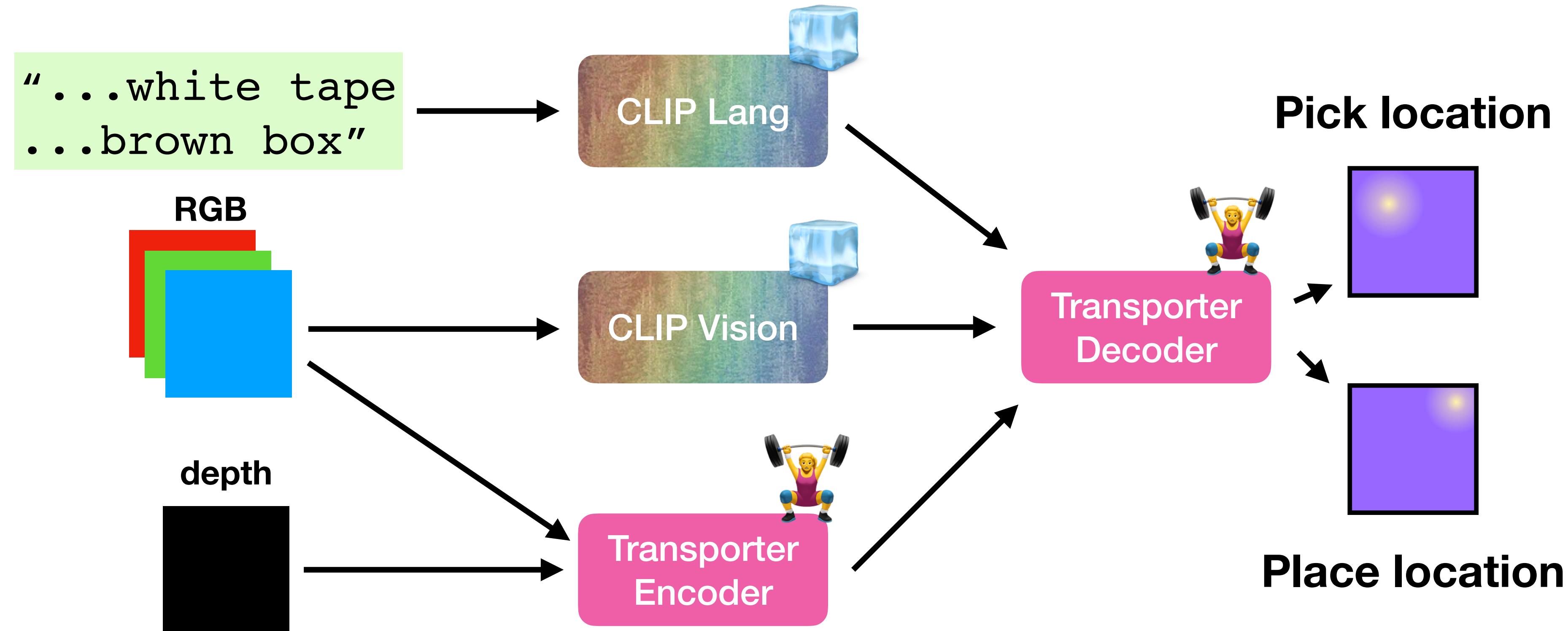
Pick location



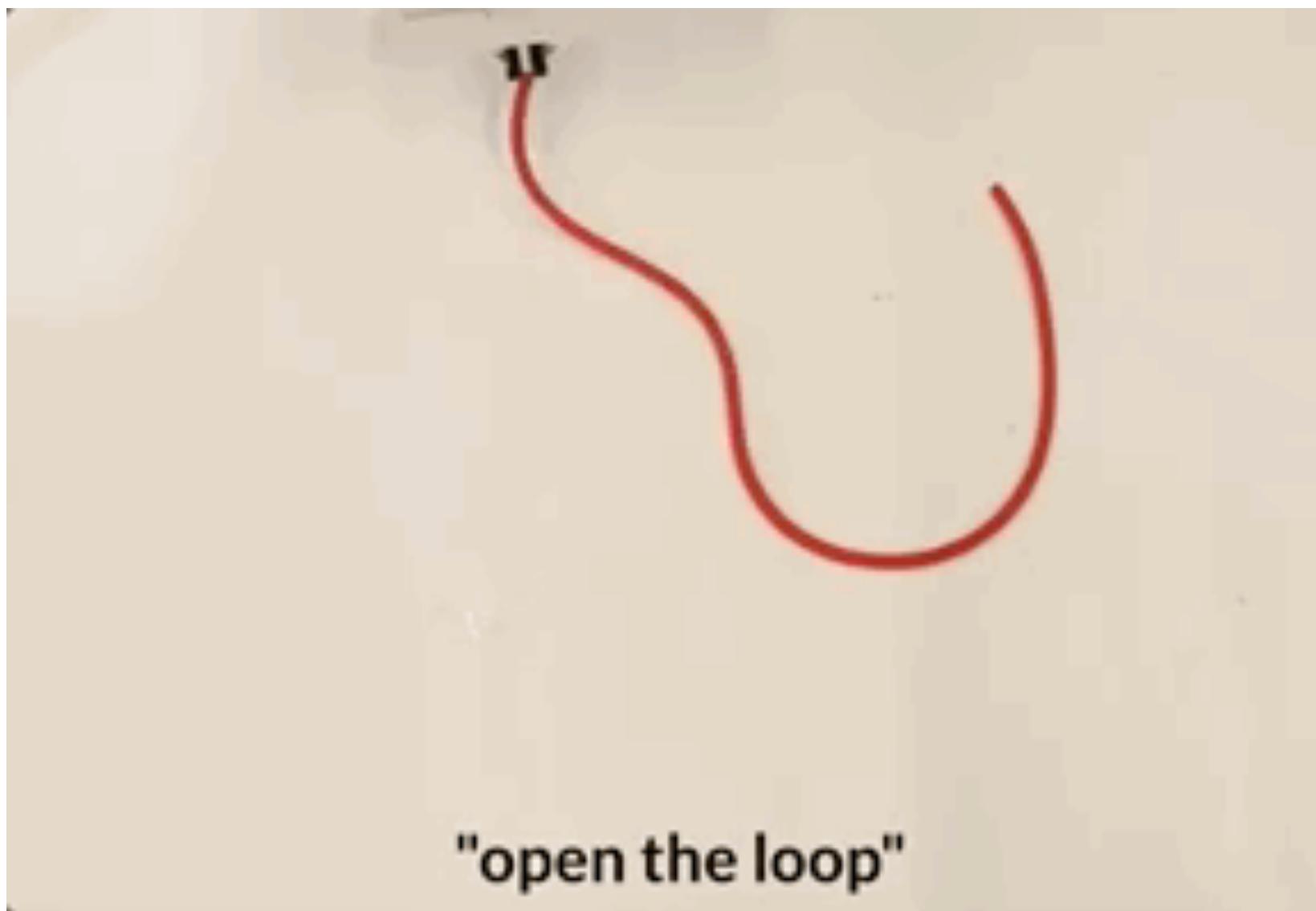
Place location



CLIPort: Model



CLIPort: Demo



"open the loop"



"pick all the cherries and put them in the box"

Possible Drawbacks of Fine-Tuning

- When we fine-tune what capabilities are we losing in the original model?
- Do we lose:
 - Robustness to distribution shift?
 - Zero-shot capabilities?
 - Generality?

LM as Planners: Big Ideas

- **LM can do high-level planning** (given the proper prompting and admissibility checks)

LM as Planners: Task

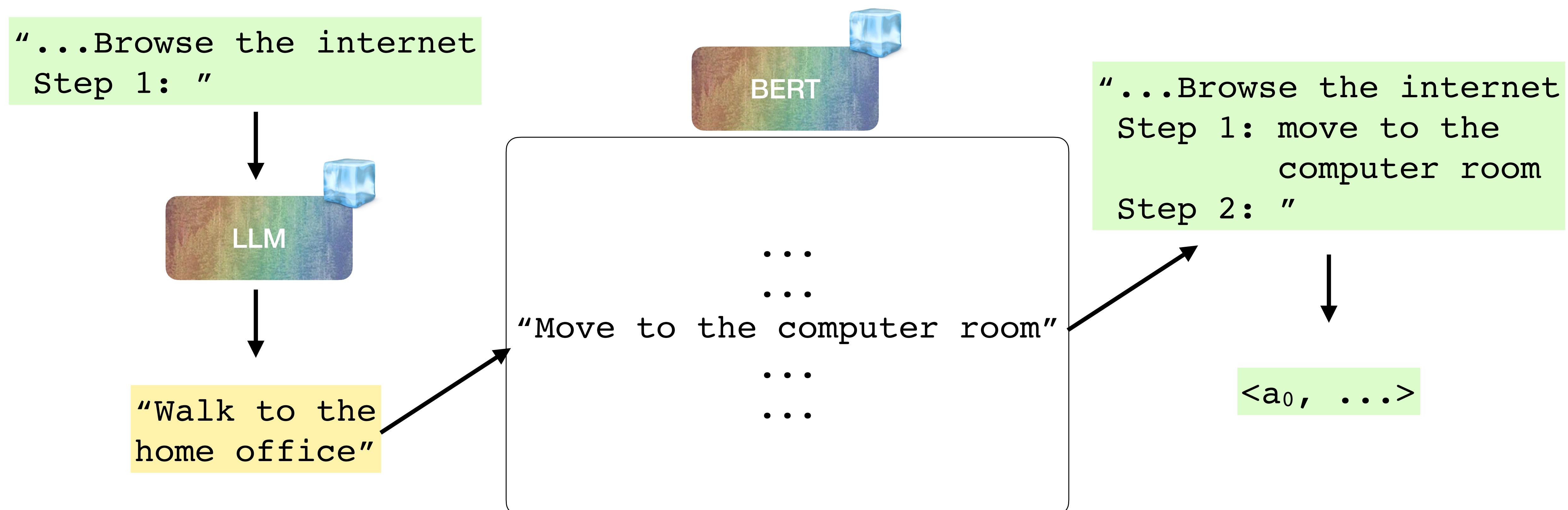
- Inputs:

“...Browse the internet...”

- Output:

$\langle a_0, a_1, \dots, a_n \rangle$

LM as Planners: Model



LM as Planners: Demo

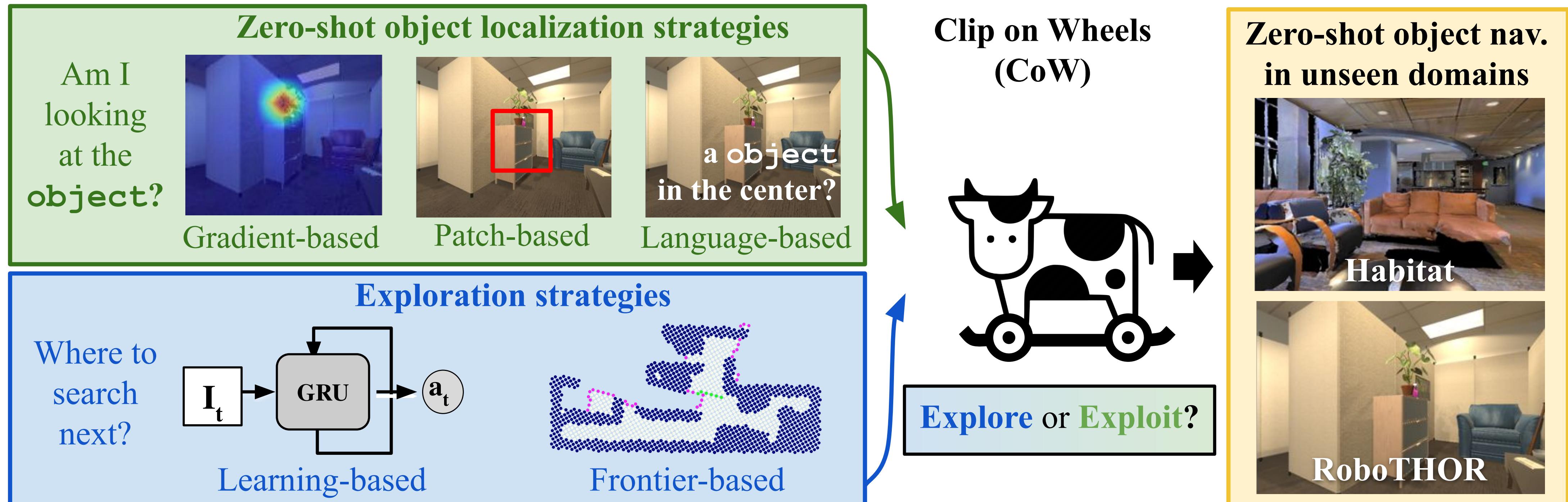


Browse Internet

We ideally want vision in the mix

- Language models seem to be capable high level planners
- Can we use vision-language models to translate perception into action?

CLIP on Wheels (CoW): Big Ideas



CoW: Task

- Inputs:

Egocentric RGB + D



Language for the target object

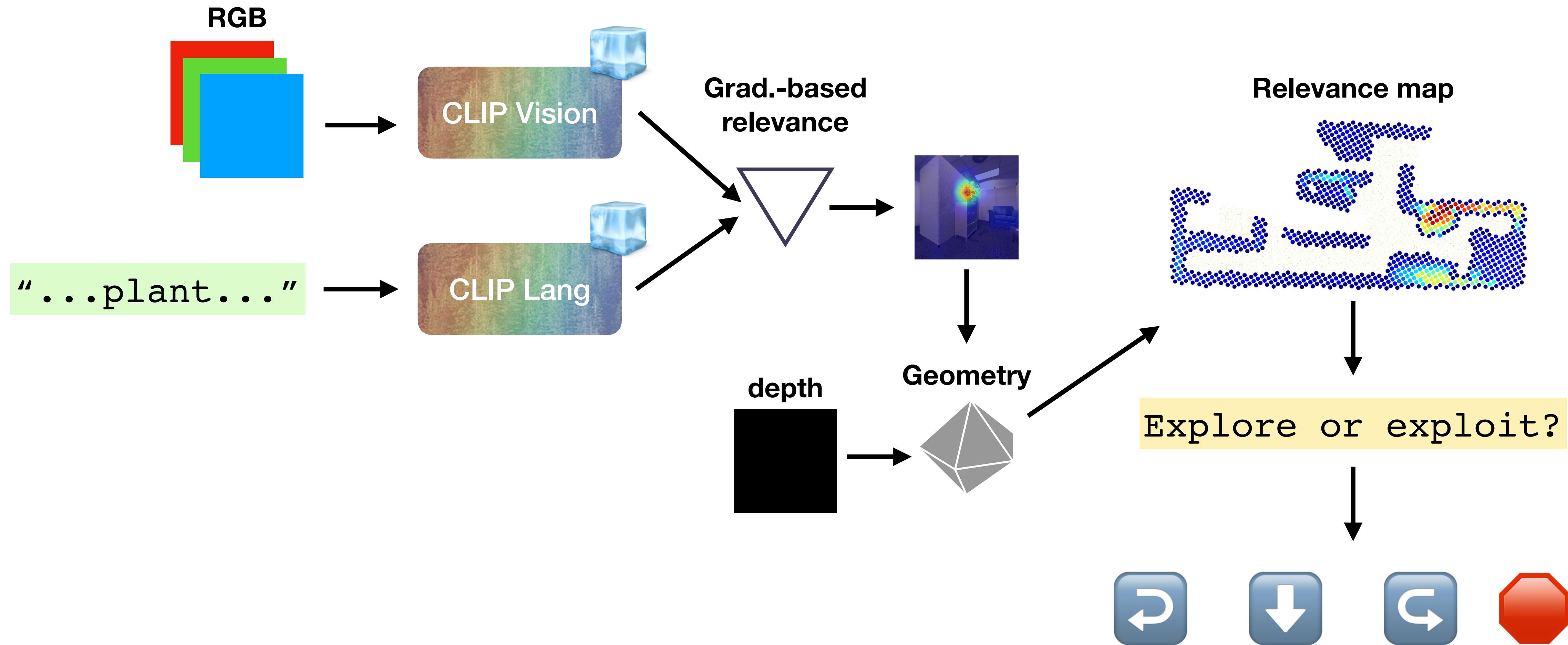
“A photo of a plant.”

- Output:

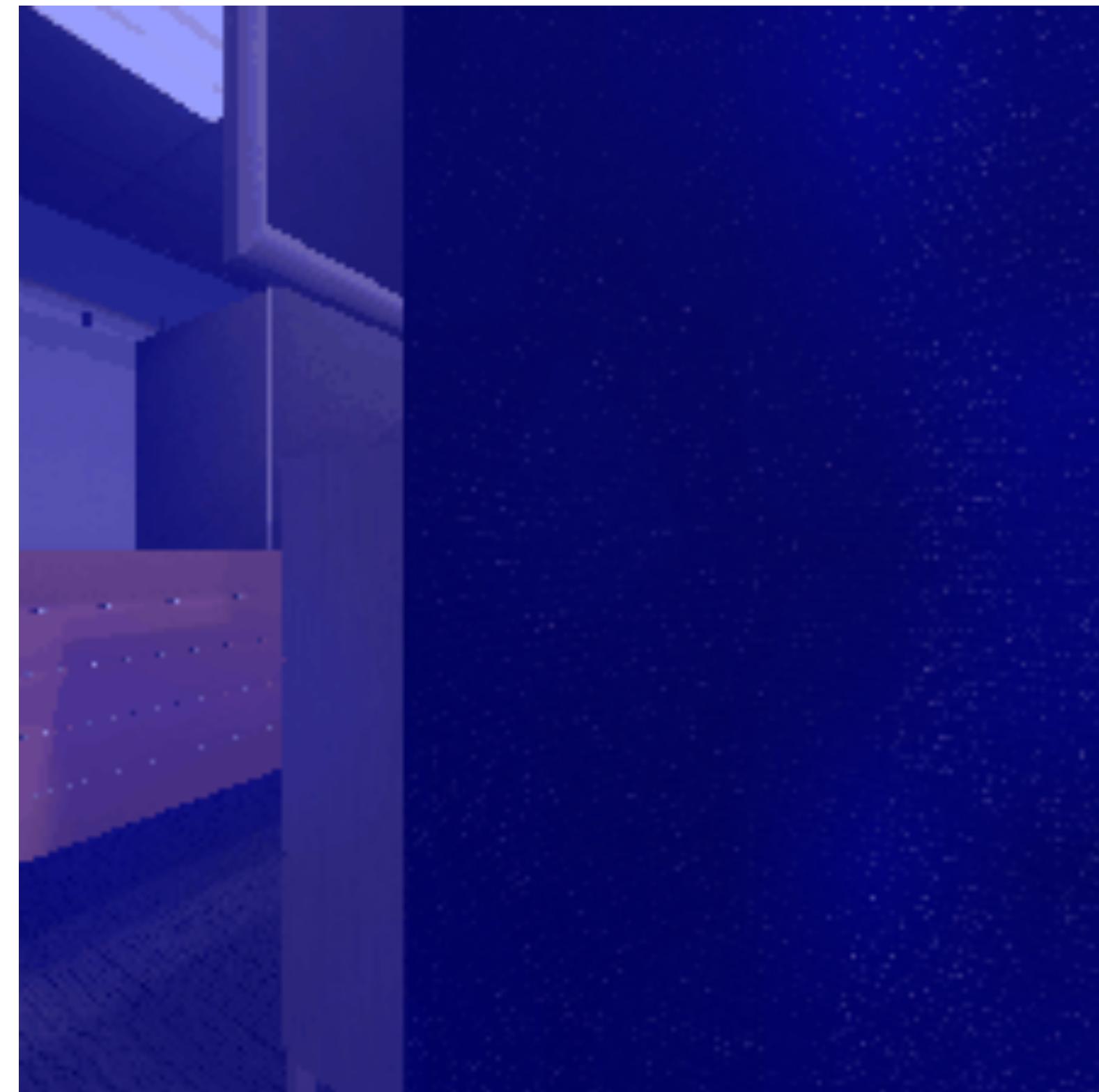
Action: direction to move (or stop)



CLIP on Wheels (CoW): Model



CoW: Demo



Limitations

- We are still beginning to understand the biases and failure modes of large models. Mitigating against these biases should be taken seriously for downstream robotics
- Usually still a gap between zero-shot and fine-tuned performance
- Pay walls

Future Directions

- Pre-trained Vision-Language models are really powerful, we should think of new ways of using them (for robotics) without more training
- How can other vision techniques be used in conjunction with zero-shot models