Distilled Feature Fields Enable Few-Shot Language-Guided Manipulation

CoRL 2023 ✅

Oral presentation ⭐

Best Paper Award 🏅

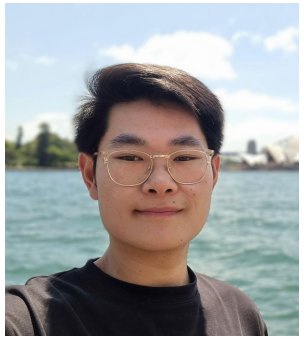
Project: f3rm.github.io

Jishnu P

Reading Group | IRVL

1/26/24

Authors



William Shen*
MIT

Ge Yang*
MIT, IAIFI

Alan Yu
MIT

Jansen Wong
MIT

Leslie Pack Kaelbling
MIT

Phillip Isola
MIT

Coming back to the Title

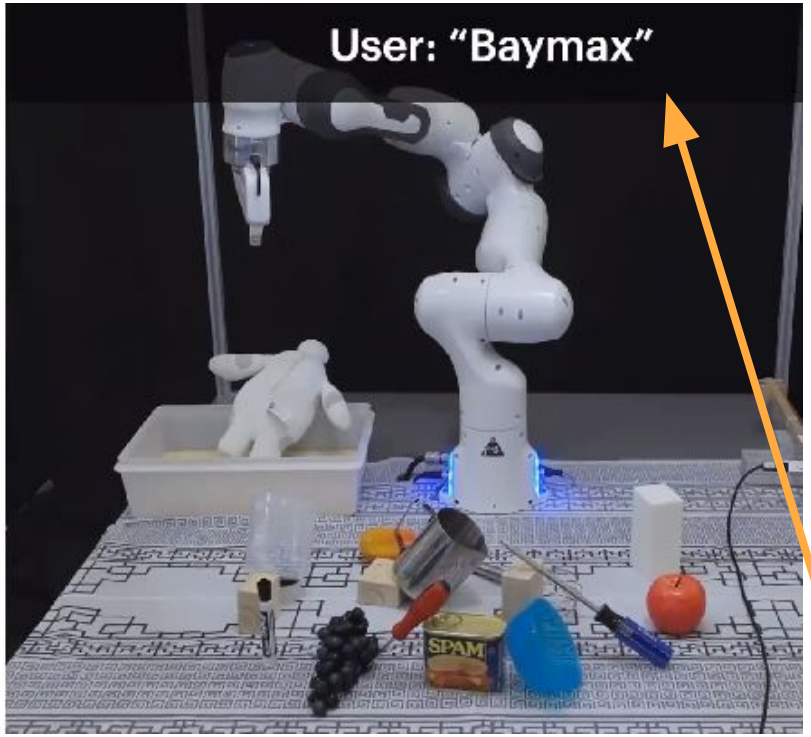
Distilled Feature Fields Enable Few-Shot Language-Guided Manipulation

F3RM: Feature Fields for Robot Manipulation

Motivation

Demonstration

Caterpillar Toy



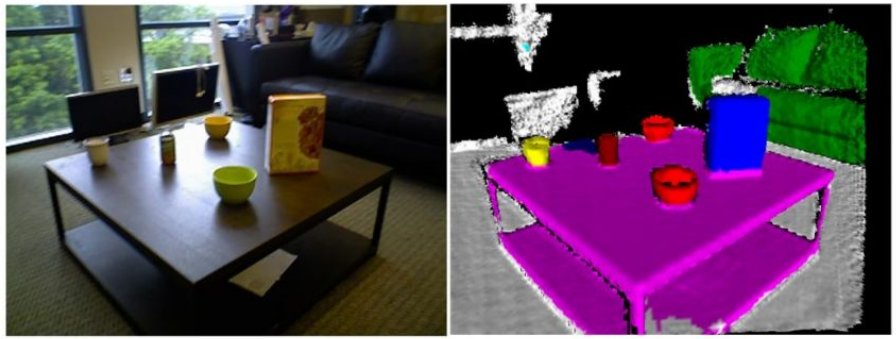
Demo on one toy and **user input asks to pick a different one**

Is it possible that a robot is able to use a given example demo to generalize on given task?

Good Scene Representation is the key to
Open-Ended Generalization

Scene Representation

Spring 2024: CS 4391 Introduction to Computer Vision



Understand the 3D world from 2D images

Large Vision-Lang Models

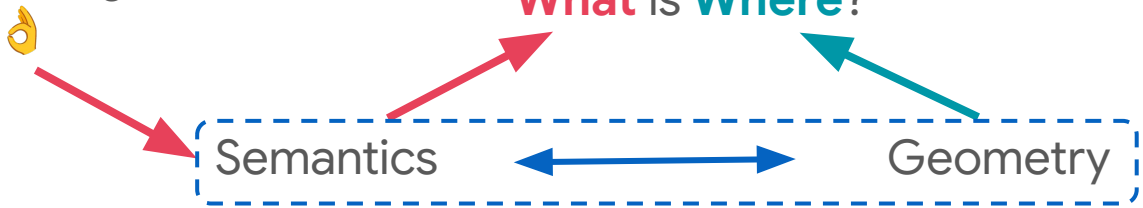


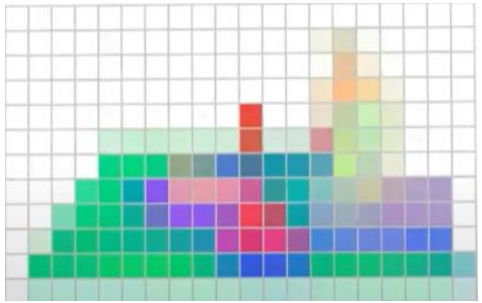
Image: https://labs.utdallas.edu/irvl/courses/cs4391_spring24/

2D Foundation Models

2D Foundation Models

Images

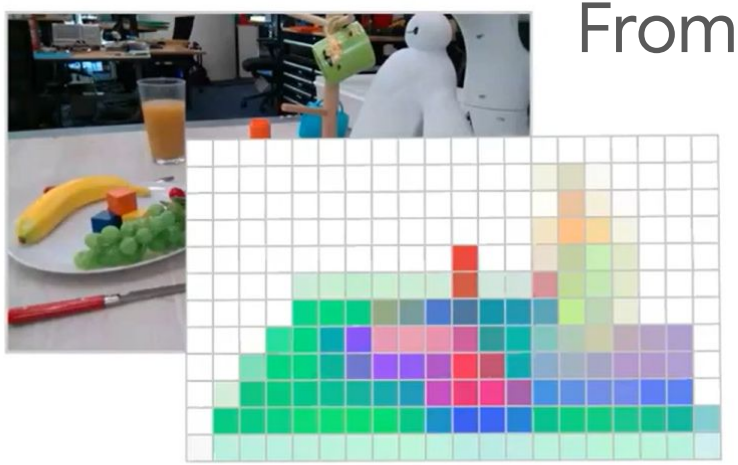
VLM



2D Feature Maps

2D Foundation Models

2D Image



feature map

From where to Pick?

Grasp in 3D

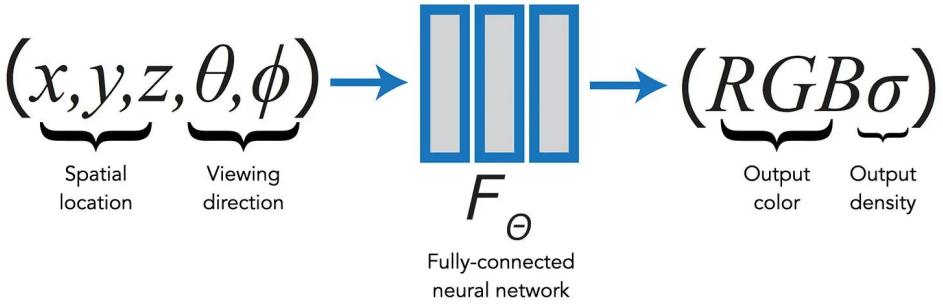
Semantics ✅

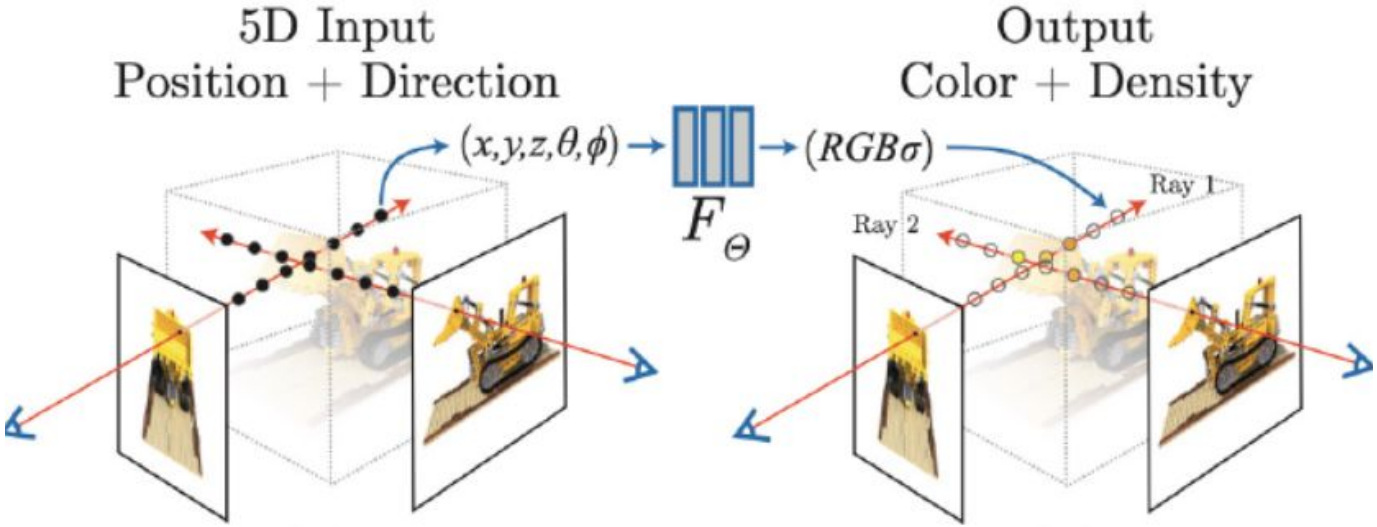
Geometry ❌

Semantics

3D Geometry

NeRF: Quick Intro





2D Foundation Models

+

NeRFs 

=

3D

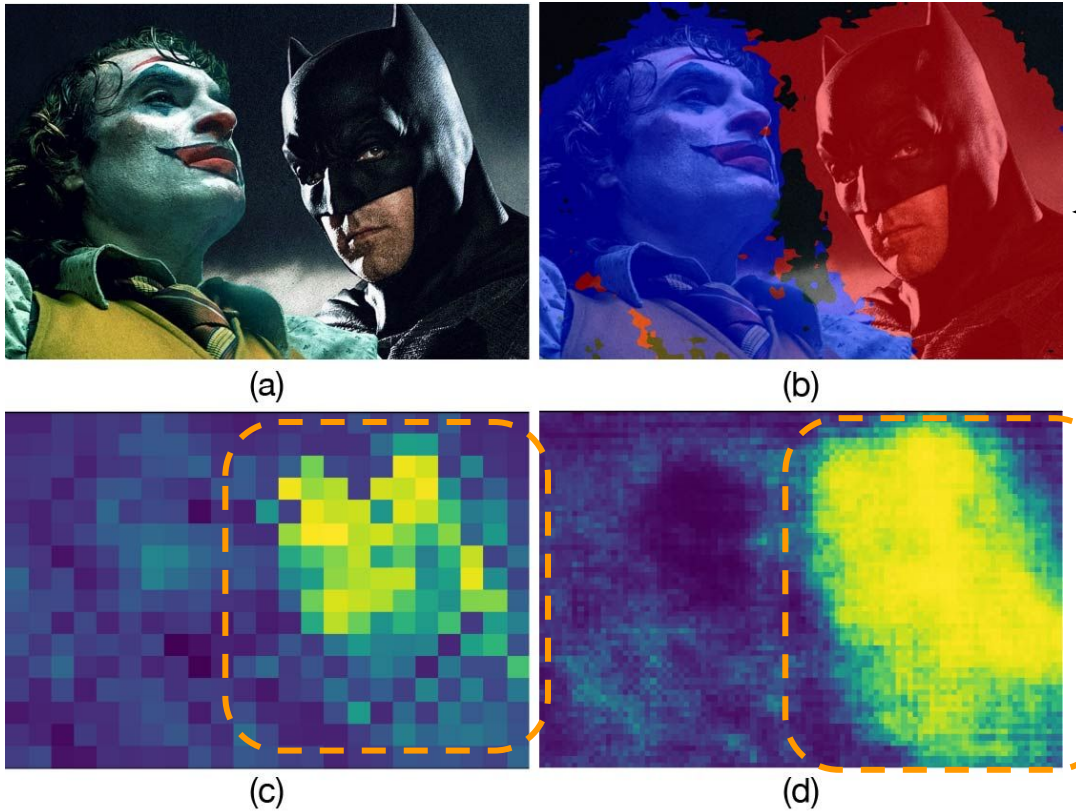
Feature Fields

F3RM has 2 components

Scene Representation to enable Few Shot Language-guided **Manipulation**

MaskCLIP and MaskCLIP+

Input Image

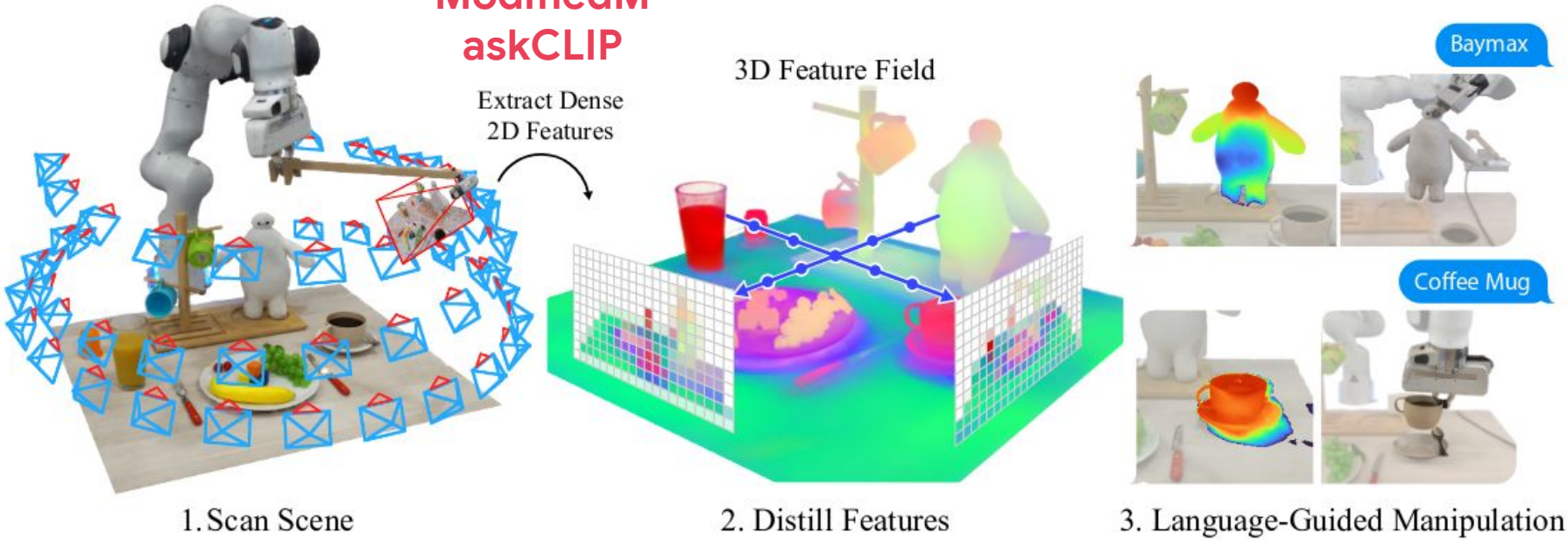


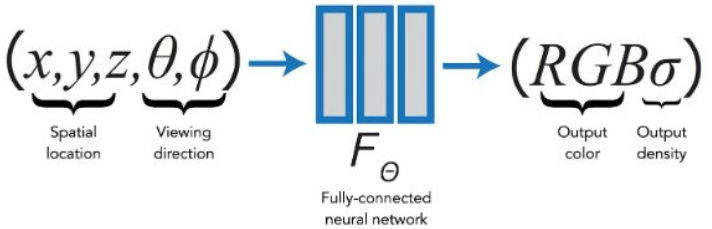
SegMask

For Text Query: 'Batman' (c) MaskCLIP and (d) MaskCLIP+ show confidence maps

Source: https://www.mmlab-ntu.com/project/maskclip

ModifiedM
askCLIP





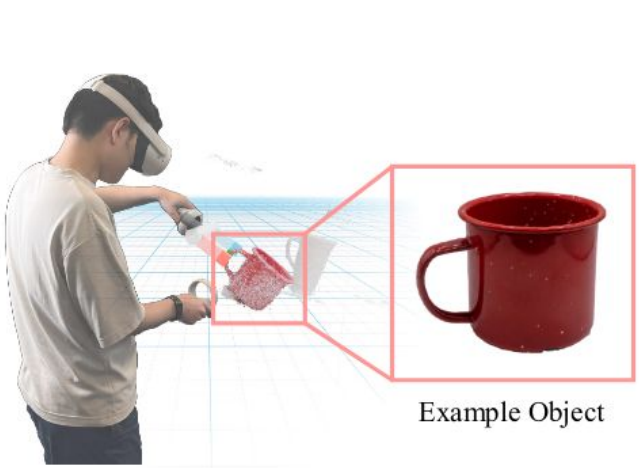
+ **Features**

2. Feature Distillation

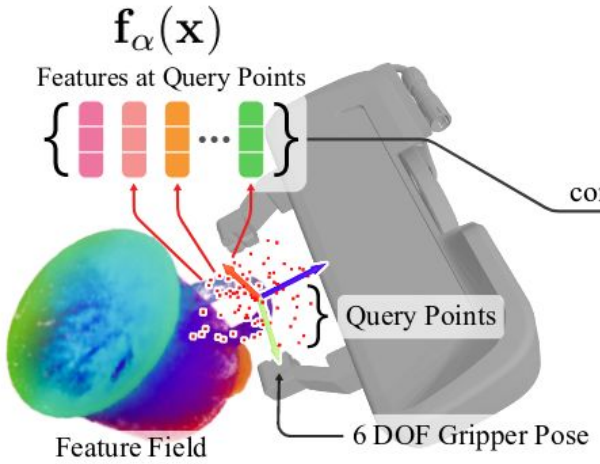
set of N 2D feature maps $\{\mathbf{I}_i^f\}_{i=1}^N$, where $\mathbf{I}^f = \mathbf{f}_{\text{vis}}(\mathbf{I})$

quadratic loss $\mathcal{L}_{\text{feat}} = \sum_{\mathbf{r}\in\mathcal{R}} \left\|\hat{\mathbf{F}}(\mathbf{r}) - \mathbf{I}^f(\mathbf{r})\right\|_2^2$

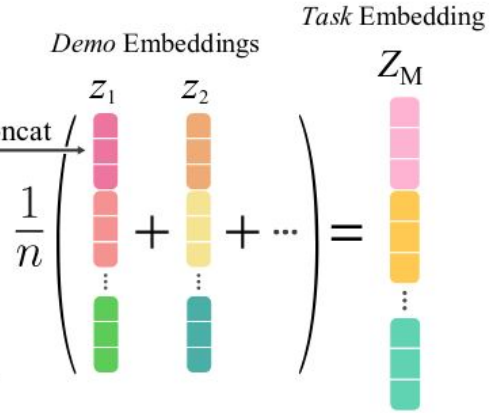
Manipulation: Representing 6-DOF Poses



(a) Collect Demonstrations in VR



(b) Sample Feature Vectors



(c) Average Over n Demos

$n = 2$

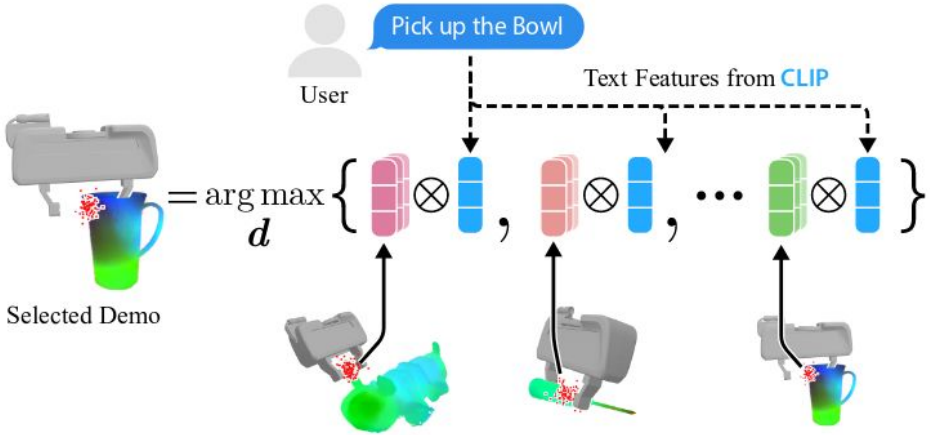
- Idea: Encode Gripper Pose in terms of 3D feature field
- Sample a fixed set of 100 points using 3D Gaussian Sampling in the Gripper Frame
- Mean and Variance are manually adjusted based on important context cues (obj. Body, free space)

$$\mathbf{f}_\alpha(\mathbf{x}) = \alpha(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}), \text{ where } \alpha(\mathbf{x}) = 1 - \exp(-\sigma(\mathbf{x}) \cdot \delta) \in (0, 1)$$

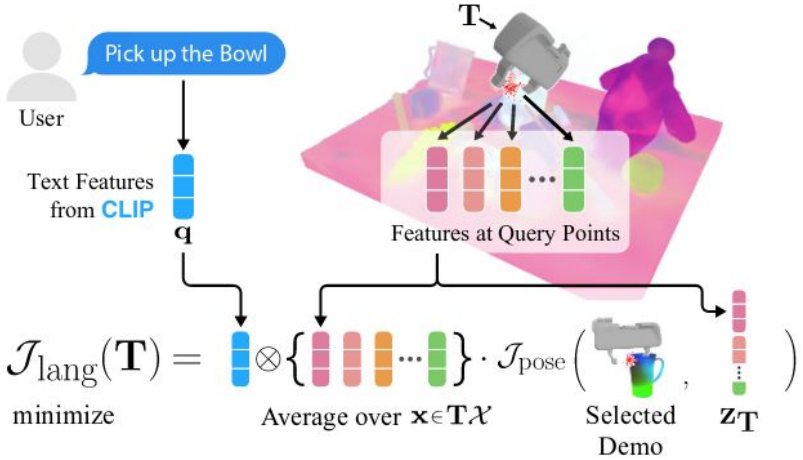
$$\{\mathbf{f}_\alpha(\mathbf{x}) \mid \mathbf{x} \in \mathbf{T}\mathcal{X}\} \quad \mathbf{T} \in SE(3)$$

$$\mathbf{T} = (\mathbf{R}, \mathbf{t}) \text{ in the world frame}$$

Pipeline for Language-Guided Manipulation



(a) Retrieving Demonstrations



(b) Language-Guided Pose Optimization

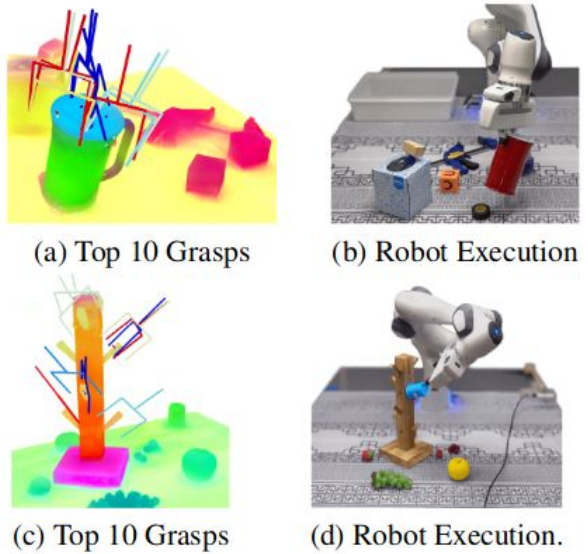
$$\mathcal{J}_{\text{pose}}(\mathbf{T}) = -\cos(\mathbf{z}_{\mathbf{T}}, \mathbf{Z}_M)$$

Details

(a) Grasp mug (lip) (b) Grasp screwdriver

(c) Grasp caterpillar (d) Place cup on rack

Figure 4: **Five Grasping and Place Tasks.** (a) grasping a mug by its lip or handle (Fig.2); (b) a screwdriver by the handle; (c) the caterpillar by its ears; and (d) placing a cup onto a drying rack. Gripper poses indicate one of two demonstrations.



(a) Top 10 Grasps (b) Robot Execution

(c) Top 10 Grasps (d) Robot Execution.

Figure 5: **Generalizing to Novel Objects.** (Top Row) Mug is much bigger than the ones used for demonstration. (Bottom Row) This rack has shorter pegs with a square cross-section. Demo rack is cylindrical (cf. Fig.4d).

Results

	Mug lip	Mug handle	Caterpillar ear	Screwdriver handle	Cup on rack	Total
MIRA [21]	1/10	2/10	6/10	3/10	3/10	15/50
Density	5/10	5/10	**10/10**	2/10	5/10	27/50
Intermediate	2/10	2/10	1/10	3/10	1/10	9/50
RGB	4/10	3/10	9/10	1/10	4/10	21/50
DINO ViT	5/10	4/10	8/10	6/10	**8/10**	31/50
CLIP ViT	7/10	**7/10**	8/10	6/10	6/10	34/50
CLIP ResNet	**9/10**	6/10	9/10	**8/10**	7/10	**39/50**

Table 1: **Success rates on grasping and placing tasks.** We compare the success rates over ten evaluation scenes given two demonstrations for each task. We consider a run successful if the robot grasps or places the correct corresponding object part for the task.

Color	7/10
Material	7/10
Relational	4/10
General	4/10
OOD	9/10
Total	31/50

Table 2: **Success rates of Language-Guided Manipulation.** Language query success rates across semantic categories.

Results

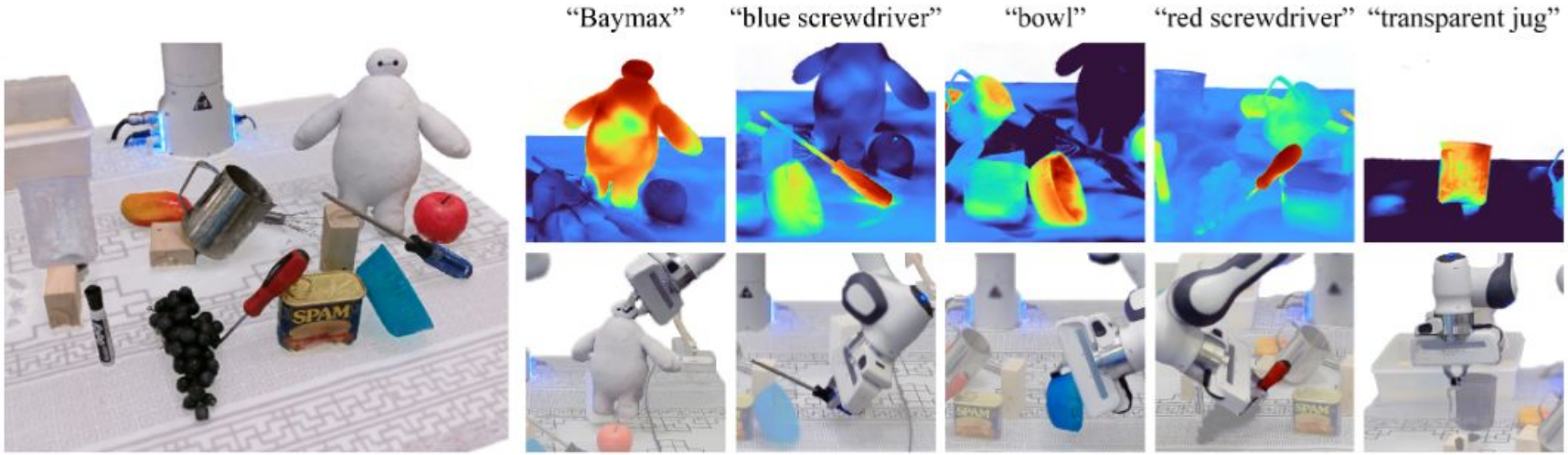
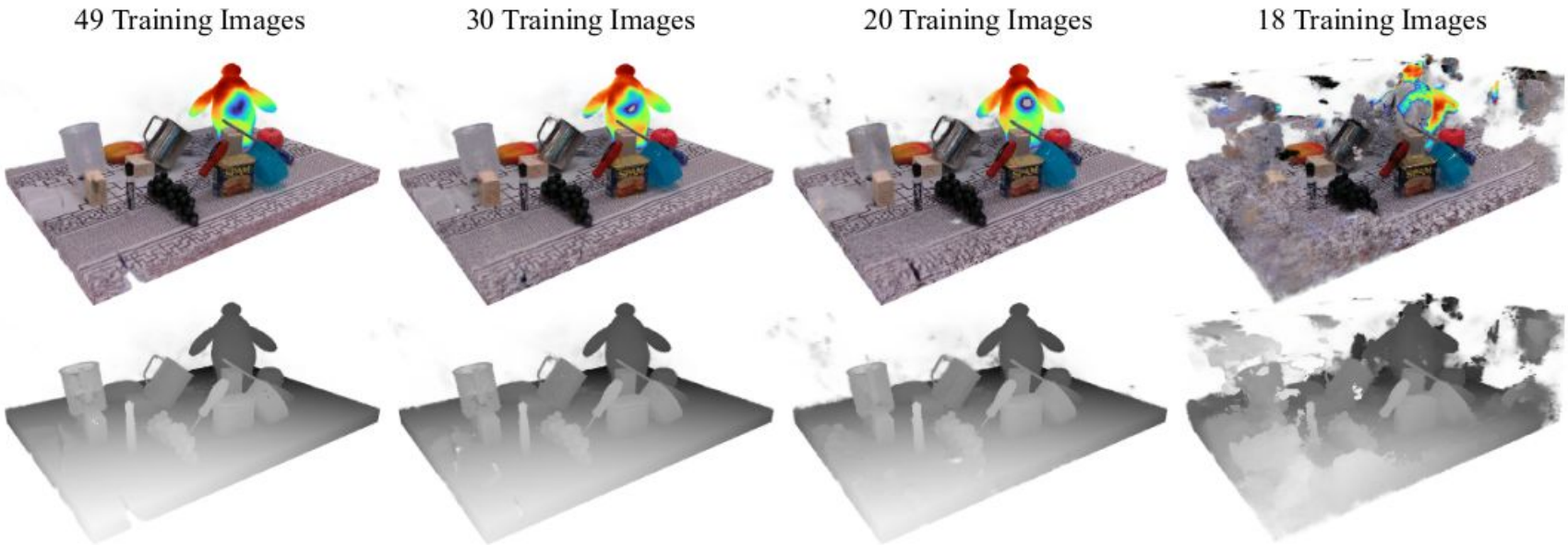


Figure 7: **Language-Guided Manipulation Execution.** (Top Row) Heatmaps given the language queries. (Bottom Row) Robot executing grasps sequentially without rescanning. CLIP can behave like a bag-of-words, as shown by the bleed to the blue bowl for "blue screwdriver."

Ablation Study



Limitations

Requires Dense Views
Optimization Per Scene
Storage

Questions?