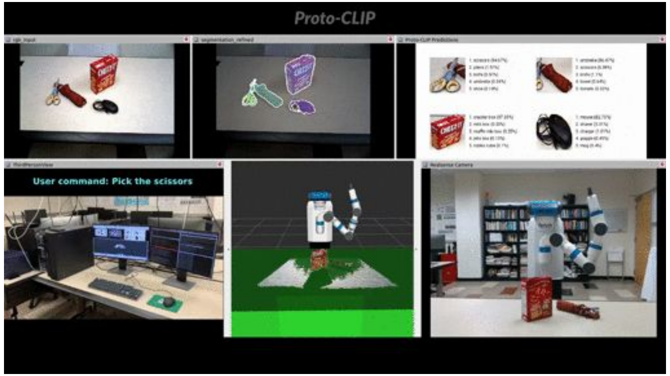
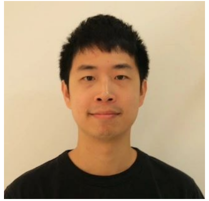
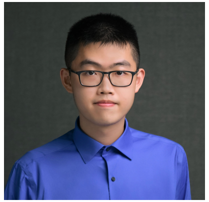
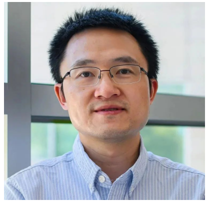
Proto-CLIP: Vision-Language Prototypical Network for Few-Shot Learning



Jishnu Jaykumar P — UT Dallas
Kamalesh Palanisamy — UT Dallas
Yu-Wei Chao — NVIDIA
Xinya Du — UT Dallas
Yu Xiang — UT Dallas



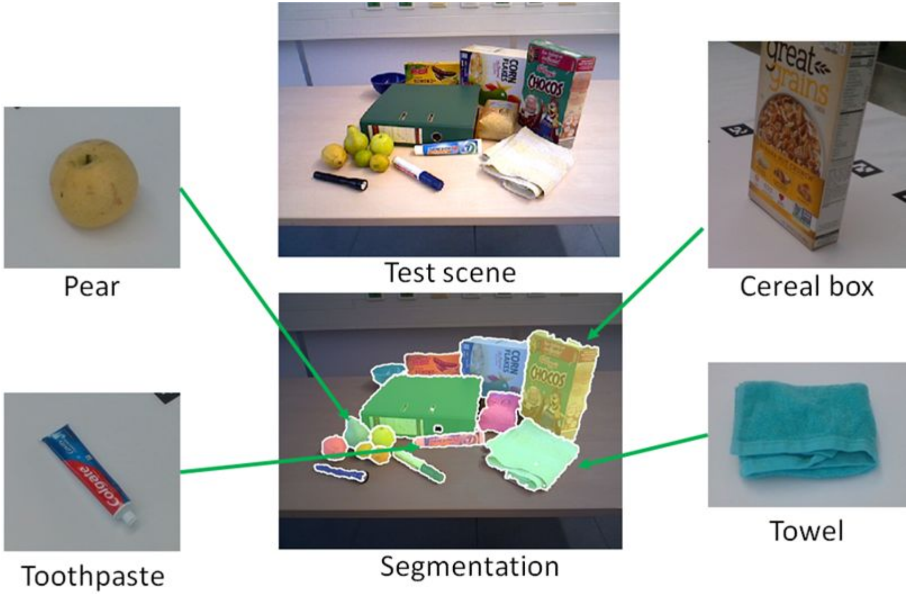




IROS '24 ABU DHABI
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems

Motivation: Few-shot object classification in cluttered robotic environments

A sample robotics environment

Clutter Scene



Goal: A robot should identity various (daily) objects in clutter scenes

Our approach: Object Classification using Few-Shot Learning

Motivation: Few-shot object classification in cluttered robotic environments

Method	OCID (Real) [10]					
	Use GT segmentation (#classes, #objects)			Use segmentation from [31] (#classes, #objects)		
	All (52, 2300) Clean S	Unseen (41, 1598) Clean S	Seen (11, 702) Clean S	All (52, 2300) Clean S	Unseen (41, 1598) Clean S	Seen (11, 702) Clean S
Training setting: clean support set with pre-training (top-1, top-5)						
k-NN [22]	14.65, 25.22	15.33, 24.41	41.03, 72.65	12.70, 23.22	13.70, 22.59	36.75, 67.95
Finetune [22]	22.26, 50.17	**26.41**, 58.20	31.62, 80.34	21.30, 48.57	**24.34**, 53.94	35.47, 67.38
ProtoNet [27]	**25.17**, **57.30**	25.22, **58.45**	51.99, **94.73**	**22.96**, **51.96**	22.65, **54.32**	49.86, **87.75**
MatchingNet [12]	17.39, 48.35	14.64, 50.06	51.85, 90.31	15.78, 45.13	13.08, 46.93	49.15, 84.47
fo-MAML [9]	11.43, 31.48	11.58, 34.73	36.89, 69.94	10.91, 29.17	10.01, 32.35	31.77, 63.68
fo-Proto-MAML [22]	14.35, 28.96	5.63, 40.61	45.58, 71.51	13.39, 26.96	5.51, 37.73	41.74, 67.24
CTX [29]	17.48, 46.57	18.21, 49.81	51.85, 87.75	15.70, 43.83	16.90, 46.31	47.86, 81.34
CTX+SimCLR [29]	18.57, 50.30	20.46, 51.06	**57.55**, 93.16	16.48, 46.17	17.71, 47.12	**52.14**, 85.75
Training setting: cluttered support set with pre-training (top-1, top-5)						
k-NN [22]	13.70, 23.83	15.33, 24.28	47.72, 72.79	13.26, 23.22	14.14, 22.90	44.73, 68.66
Finetune [22]	22.17, 53.35	24.34, 55.63	31.91, 71.51	18.26, 44.22	20.65, 52.00	36.04, 69.52
ProtoNet [27]	21.35, 50.57	22.34, 51.31	51.99, 90.46	18.61, 47.22	18.21, 48.12	45.44, 85.33
MatchingNet [12]	17.52, 50.96	17.77, 52.32	49.43, 88.18	16.52, 46.52	15.58, 48.81	43.45, 82.76
fo-MAML [9]	16.48, 38.52	13.70, 39.49	37.46, 77.07	15.35, 35.04	11.08, 34.36	40.31, 69.94
fo-Proto-MAML [22]	11.04, 28.70	4.01, 38.67	43.73, 72.65	9.91, 26.35	3.57, 35.79	40.46, 68.09
CTX [29]	19.00, 45.48	17.71, 44.74	51.85, 88.75	17.13, 42.22	16.08, 42.12	47.15, 83.19
CTX+SimCLR [29]	**24.61**, **62.39**	**25.16**, **63.52**	**65.81**, **96.30**	**22.17**, **57.43**	**23.28**, **57.57**	**59.12**, **88.32**
Using pre-trained CLIP models [35]						
Few-shot Tip-Adapter ViT-L/14-Finetune [36]	**60.17**, 83.04	**59.64**, 85.17	85.75, **99.00**	**54.87**, **78.91**	**56.07**, 80.29	79.20, 91.88
Few-shot Tip-Adapter ViT-L/14 [36]	56.78, 83.22	55.38, 84.86	**86.89**, 98.58	52.35, 76.26	51.69, 79.04	**80.06**, **92.45**
Zero-shot CLIP ViT-L/14 [35]	54.57, **84.74**	55.94, **87.92**	83.62, 98.58	50.43, 78.52	52.07, **81.54**	75.07, 92.17
Zero-shot CLIP ViT-B/32 [35]	41.87, 75.26	41.30, 77.91	78.06, 97.58	39.83, 69.43	39.17, 72.09	70.66, 90.88
Zero-shot CLIP ViT-B/16 [35]	40.70, 73.96	40.24, 76.03	76.50, 95.73	39.35, 68.83	38.61, 70.15	70.66, 88.89
Zero-shot CLIP RN50x64 [35]	42.96, 75.83	43.62, 77.41	76.64, 96.01	40.04, 70.87	41.74, 72.22	69.94, 90.46
Zero-shot CLIP RN50x16 [35]	38.52, 73.04	40.11, 75.72	79.49, 96.30	35.65, 67.30	37.30, 69.77	70.94, 89.74
Zero-shot CLIP RN50x4 [35]	35.96, 68.52	34.42, 70.03	73.93, 95.73	34.00, 63.78	32.48, 65.46	67.95, 88.60
Zero-shot CLIP ResNet-101 [35]	32.96, 68.30	32.67, 69.52	77.49, 96.87	31.09, 63.87	31.85, 65.96	69.66, 89.74
Zero-shot CLIP ResNet-50 [35]	25.91, 58.43	29.04, 64.39	61.40, 93.16	24.70, 55.61	28.04, 61.20	57.69, 86.47

V

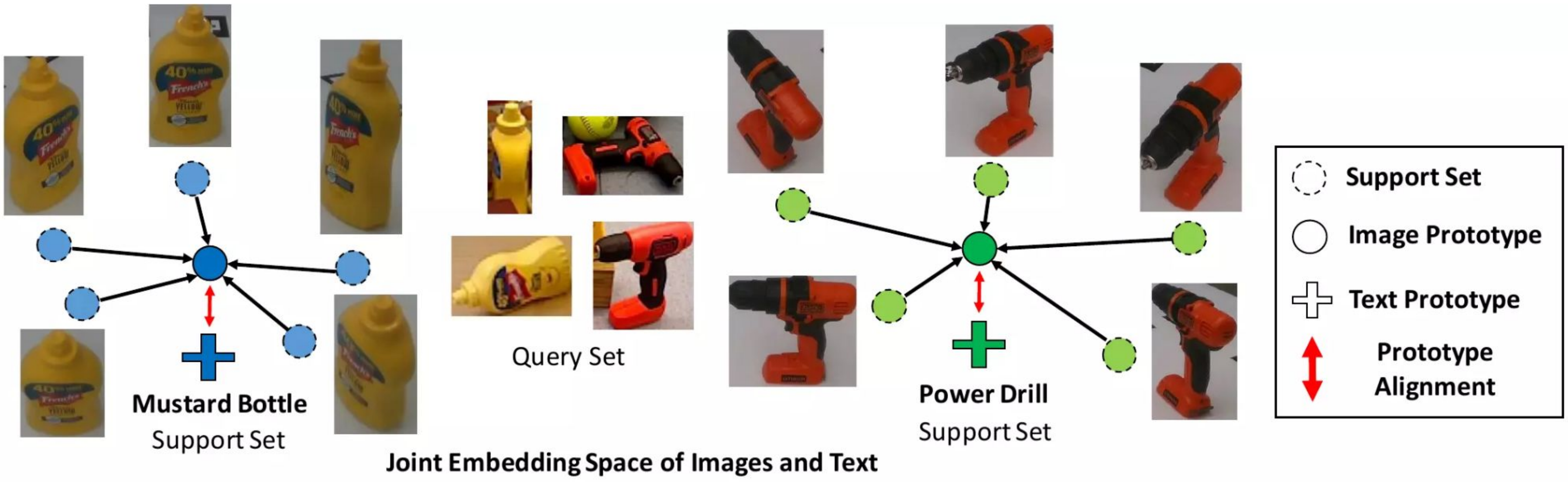
V

V + L

Observation: Vision+Language models (CLIP and it's related work) outperform the existing few shot vision only methods

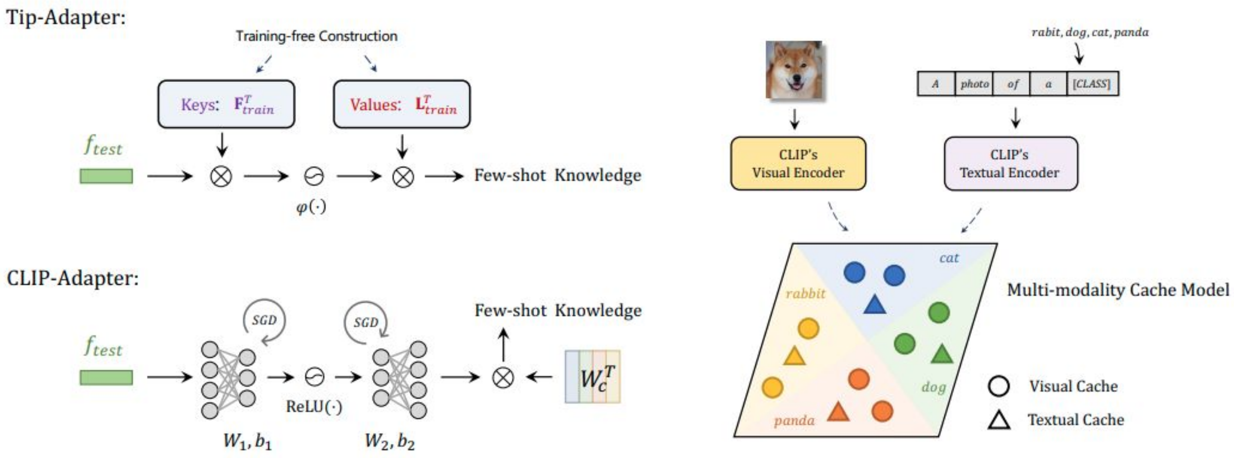
J. J. P. et al., "FewSOL: A Dataset for Few-Shot Object Learning in Robotic Environments," ICRA, London, 2023.

Our Idea



Our proposed Proto-CLIP model learns a *joint embedding space of images and text*, where *image prototypes* and *text prototypes* are learned using *support sets* for few-shot classification.

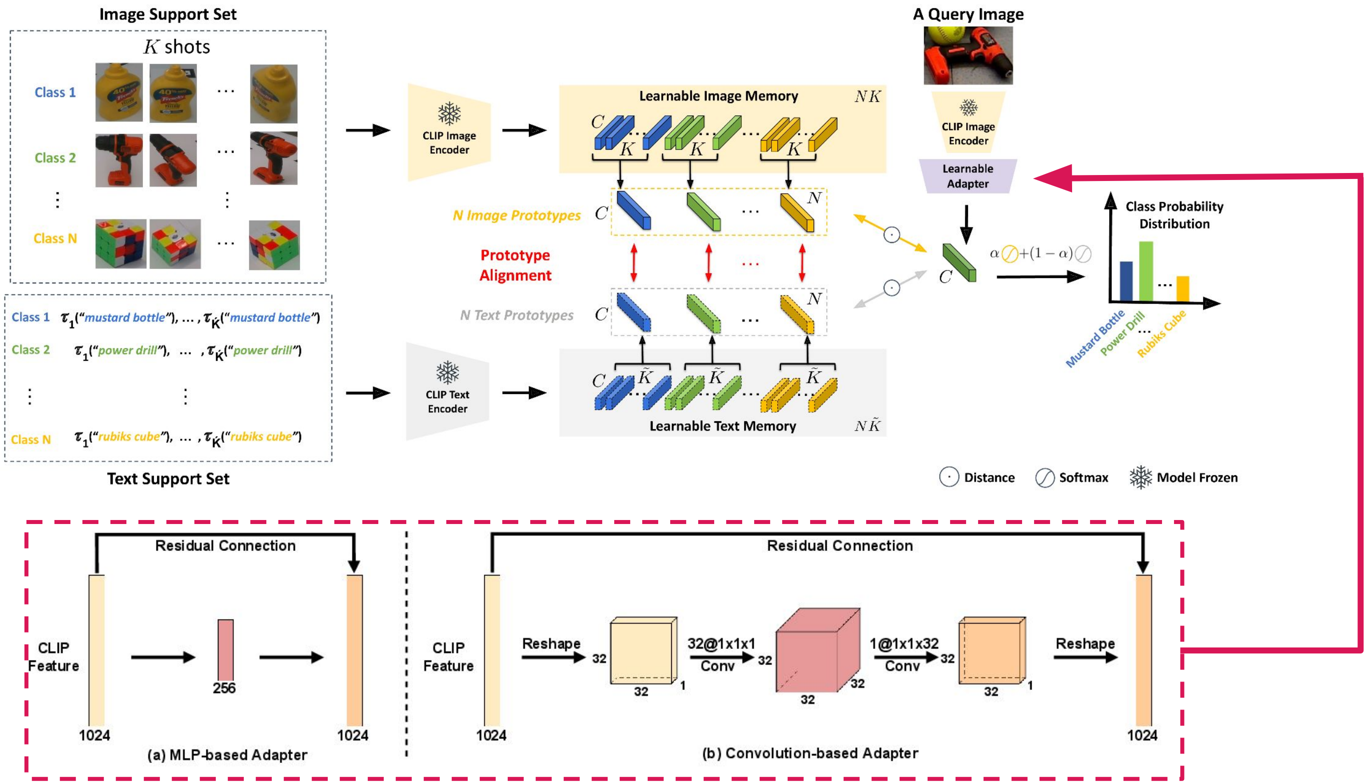
Related Vs Ours



Zhang et. al. Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification.
In Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel

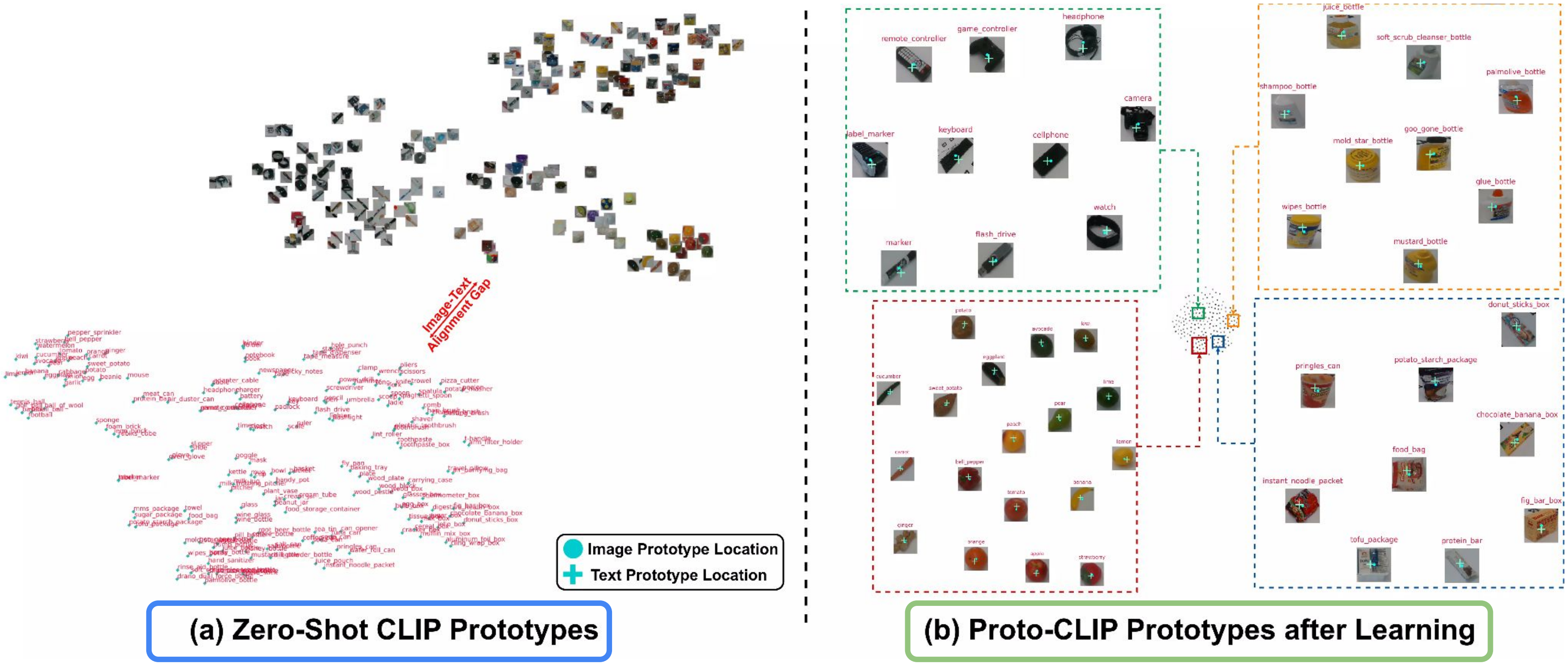
Method	Use Support Sets	Adapt Image Embedding	Adapt Text Embedding	Align Image and Text
Zero-shot CLIP [1]	✗	✗	✗	✓
Linear-probe CLIP [1]	✓	✓	✗	✗
CoOp [8]	✓	✗	✓	✗
CLIP-Adapter [9]	✓	✓	✓	✗
Tip-Adapter [10]	✓	✓	✗	✗
PROTO-CLIP (Ours)	✓	✓	✓	✓

Model Overview



Overview of our proposed Proto-CLIP model. The CLIP image encoder and text encoder are frozen during training ❄️. The image memory, the text memory and the adapter network are learned 🔥.

Barnes-Hut t-SNE visualization using the FewSOL dataset



(a) Image and text prototypes from zero-shot CLIP, which are not aligned.

(b) Aligned image and text prototypes from Proto-CLIP-F.

Few-shot classification results on different datasets using the ResNet50 backbone

Dataset # classes	ImageNet 1,000	FGVC 100	Pets 37	Cars 196	EuroSAT 10	Caltech101 100	SUN397 397	DTD 47	Flowers 102	Food101 101	UCF101 101	FewSOL 52
Zero-shot CLIP [1]	60.33	17.10	85.83	55.74	37.52	85.92	58.52	42.20	66.02	77.32	61.35	25.91
1 shots												
Linear-Probe CLIP [1]	22.07	12.89	30.14	24.64	51.00	70.62	32.80	29.59	58.07	30.13	41.43	-
CoOp [8]	57.15	9.64	85.89	55.59	50.63	87.53	60.29	44.39	68.12	74.32	61.92	-
CLIP-A [9]	61.20	17.49	85.99	55.13	61.40	88.60	61.30	45.80	73.49	76.82	62.20	-
Tip [10]	60.70	19.05	86.10	57.54	54.38	87.18	61.30	46.22	73.12	77.42	62.60	27.30
Tip-F [10]	**61.13**	**20.22**	**87.00**	**58.86**	59.53	**89.33**	**62.50**	**49.65**	**79.98**	**77.51**	**64.87**	**27.91**
Proto-CLIP	60.31	19.59	86.10	57.29	55.53	87.99	60.81	46.04	76.98	77.36	63.15	27.09
Proto-CLIP-*F*	60.32	19.50	85.72	57.34	54.93	88.07	60.83	35.64	77.47	77.34	63.07	22.22
Proto-CLIP-F-Q^T	59.12	16.26	83.62	52.77	**61.95**	88.48	61.43	32.27	68.53	75.16	62.44	21.65
2 shots												
Linear-Probe CLIP [1]	31.95	17.85	43.47	36.53	61.58	78.72	44.44	39.48	73.35	42.79	53.55	-
CoOp [8]	57.81	18.68	82.64	58.28	61.50	87.93	59.48	45.15	77.51	72.49	64.09	-
CLIP-A [9]	61.52	20.10	86.73	58.74	63.90	89.37	63.29	51.48	81.61	77.22	67.12	-
Tip [10]	60.96	21.21	87.03	57.93	61.68	88.44	62.70	49.47	79.13	77.52	64.74	26.22
Tip-F [10]	**61.69**	**23.19**	87.03	**61.50**	**66.15**	**89.74**	63.64	**53.72**	82.30	**77.81**	66.43	27.43
Proto-CLIP	60.64	22.14	**87.38**	60.01	64.89	89.05	63.12	51.06	83.39	77.34	67.46	**28.35**
Proto-CLIP-*F*	60.64	22.14	**87.38**	60.04	64.86	89.09	63.20	49.88	**83.52**	77.34	67.49	26.17
Proto-CLIP-F-Q^T	60.48	20.01	85.28	60.02	63.59	89.49	**65.46**	45.69	81.20	76.15	**68.83**	25.91
4 shots												
Linear-Probe CLIP [1]	41.29	23.57	56.35	48.42	68.27	84.34	54.59	50.06	84.80	55.15	62.23	-
CoOp [8]	59.99	21.87	86.70	62.62	70.18	89.55	63.47	53.49	86.20	73.33	67.03	-
CLIP-A [9]	61.84	22.59	87.46	62.45	73.38	89.98	65.96	56.86	87.17	77.92	69.05	-
Tip [10]	60.98	22.41	86.45	61.45	65.32	89.39	64.15	53.96	83.80	77.54	66.46	28.70
Tip-F [10]	**62.52**	25.80	**87.54**	64.57	74.12	90.56	66.21	**57.39**	88.83	**78.24**	**70.55**	29.13
Proto-CLIP	61.30	23.25	87.19	63.33	68.67	89.57	65.51	55.91	88.23	77.58	69.50	29.13
Proto-CLIP-*F*	61.30	23.31	86.95	63.34	68.52	89.62	65.57	57.21	88.27	77.58	69.55	27.09
Proto-CLIP-F-Q^T	61.80	**27.63**	87.11	**66.24**	**80.64**	**91.81**	**68.09**	56.86	**89.85**	76.94	70.16	**30.30**
8 shots												
Linear-Probe CLIP [1]	49.55	29.55	65.94	60.82	76.93	87.78	62.17	56.56	92.00	63.82	69.64	-
CoOp [8]	61.56	26.13	85.32	68.43	76.73	90.21	65.52	59.97	91.18	71.82	71.94	-
CLIP-A [9]	62.68	26.25	87.65	67.89	77.93	91.40	67.50	61.00	91.72	78.04	73.30	-
Tip [10]	61.45	25.59	87.03	62.93	67.95	89.83	65.62	58.63	87.98	77.76	68.68	29.22
Tip-F [10]	64.00	30.21	88.09	69.25	77.93	91.44	68.87	62.71	91.51	**78.64**	74.25	32.43
Proto-CLIP	62.12	27.63	88.04	64.93	69.42	90.22	67.37	59.34	92.08	77.90	71.08	29.83
Proto-CLIP-*F*	63.92	31.32	**88.55**	70.35	78.94	92.54	69.59	62.35	93.79	78.29	74.81	**33.26**
Proto-CLIP-F-Q^T	**64.03**	**35.82**	87.46	**71.50**	**81.89**	**92.62**	**70.02**	**64.01**	**94.28**	78.61	**75.34**	32.70
16 shots												
Linear-Probe CLIP [1]	55.87	36.39	76.42	70.08	82.76	90.63	67.15	63.97	94.95	70.17	73.72	-
CoOp [8]	62.95	31.26	87.01	73.36	83.53	91.83	69.26	63.58	94.51	74.67	75.71	-
CLIP-A [9]	63.59	32.10	87.84	74.01	84.43	92.49	69.55	65.96	93.90	78.25	76.76	-
Tip [10]	62.02	29.76	88.14	66.77	70.54	90.18	66.85	60.93	89.89	77.83	70.58	28.87
Tip-F [10]	65.51	35.55	**89.70**	75.74	84.54	92.86	71.47	66.55	94.80	**79.43**	78.03	34.04
Proto-CLIP	62.77	29.67	88.61	68.11	72.95	91.08	68.09	61.64	92.94	78.11	73.35	29.96
Proto-CLIP-*F*	65.75	37.56	89.62	75.25	83.53	93.43	**71.94**	**68.56**	95.78	79.09	77.50	**35.22**
Proto-CLIP-F-Q^T	**65.91**	**40.65**	89.34	**76.76**	**86.59**	**93.59**	72.19	68.50	**96.35**	79.34	**78.11**	34.70

Proto-CLIP performs poorly in low shots setting but as shots increase the performance improves w.r.t. to other baseline models.

Ablation Study: Adapter vs Dataset

Adapter	Train-Text-Memory	ImageNet	FGVC	Pets	Cars	EuroSAT	Caltech101	SUN397	DTD	Flowers	Food101	UCF101	FewSOL
MLP	✗	61.06	35.31	85.61	72.19	83.47	92.58	68.54	63.89	95.01	74.05	76.16	28.65
MLP	✓	61.06	**37.56**	85.72	73.61	**83.53**	92.13	69.71	63.89	**96.06**	74.05	76.16	32.87
2xConv	✗	**65.75**	34.38	**89.62**	**75.25**	81.85	93.40	**71.94**	67.85	94.76	**79.09**	**77.50**	27.13
2xConv	✓	58.60	35.82	89.21	74.34	81.78	93.02	69.79	67.32	95.82	78.06	76.37	27.13
3xConv	✗	65.37	34.41	88.74	**75.25**	82.21	**93.43**	71.63	67.67	94.40	79.11	**77.50**	29.78
3xConv	✓	59.63	36.15	87.93	72.68	81.57	92.74	68.64	**68.56**	95.78	78.61	77.03	**35.22**

Observation: 🌐 **Different datasets** 📊 behave differently on various **adapters**🔌

Ablation Study: Loss vs Dataset

Loss	ImageNet	FGVC	Pets	Cars	EuroSAT	Caltech101	SUN397	DTD	Flowers	Food101	UCF101	FewSOL
$\mathcal{L}_1$	62.67	20.34	73.21	73.77	78.98	92.25	68.34	66.49	**96.14**	77.39	76.66	34.57
$\mathcal{L}_2$	62.29	4.71	0.00	0.00	38.95	0.28	66.93	67.38	10.31	77.71	57.41	32.70
$\mathcal{L}_3$	62.27	4.14	0.00	0.00	38.09	0.24	64.86	67.38	10.27	77.69	57.55	20.22
$\mathcal{L}_1 + \mathcal{L}_2$	65.39	36.24	88.58	75.39	82.78	**93.71**	71.65	68.09	96.06	78.69	77.29	33.48
$\mathcal{L}_2 + \mathcal{L}_3$	62.33	3.87	0.00	0.00	36.86	0.24	64.84	68.32	8.20	77.35	57.52	19.61
$\mathcal{L}_1 + \mathcal{L}_3$	65.43	36.84	88.58	**75.51**	82.84	93.35	71.44	68.32	**96.14**	78.80	**77.53**	33.43
$\boxed{\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3}$	**65.75**	**37.56**	**89.62**	75.25	**83.53**	93.43	**71.94**	**68.56**	96.06	**79.09**	77.50	**35.22**

L1 := Classification Loss

$$\left.\begin{array}{l}\text{L2 := Image Prototype to Text Prototype Distance Loss}\\ \text{L3 := Text Prototype to Image Prototype Distance Loss}\end{array}\right\}\ \text{InfoNCE Loss}$$

Observation: 🔄 Overall, all three losses ⚖️ are required to achieve better performance 🚀.

Ablation Study: Different CLIP Backbones

Model	Adapter	TextM	Backbone				
			RN50	RN101	ViT-B/16	ViT-B/32	ViT-L/14
Zero-Shot-CLIP [1]	-	-	25.91	32.96	40.70	41.87	54.57
Tip [10]	-	-	29.74	37.43	47.00	41.48	56.78
Tip-F [10]	-	-	32.52	41.43	50.17	45.48	60.17
PROTO-CLIP-F	MLP	✗	33.48	39.04	47.96	41.91	58.65
PROTO-CLIP-F	MLP	✓	34.83	40.74	47.43	42.13	58.91
PROTO-CLIP-F	2xConv	✗	35.04	41.04	50.83	46.52	**63.74**
PROTO-CLIP-F	2xConv	✓	35.04	42.52	49.26	43.43	61.61
PROTO-CLIP-F	3xConv	✗	34.13	42.83	**51.91**	**46.87**	62.35
PROTO-CLIP-F	3xConv	✓	**35.22**	**44.09**	50.39	46.57	60.39

Observation: 🚀 **Bigger Vision Transformers deliver superior performance** 🌟

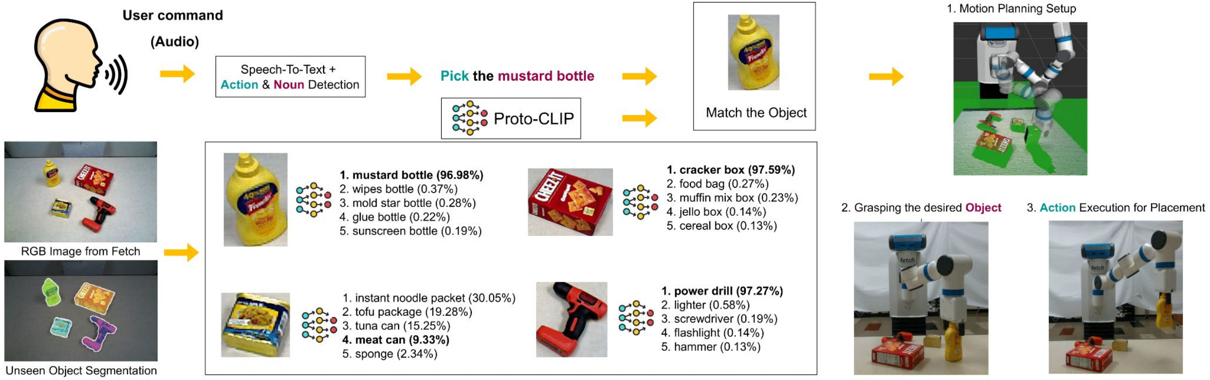
Ablation Study: Out of Distribution (OOD)

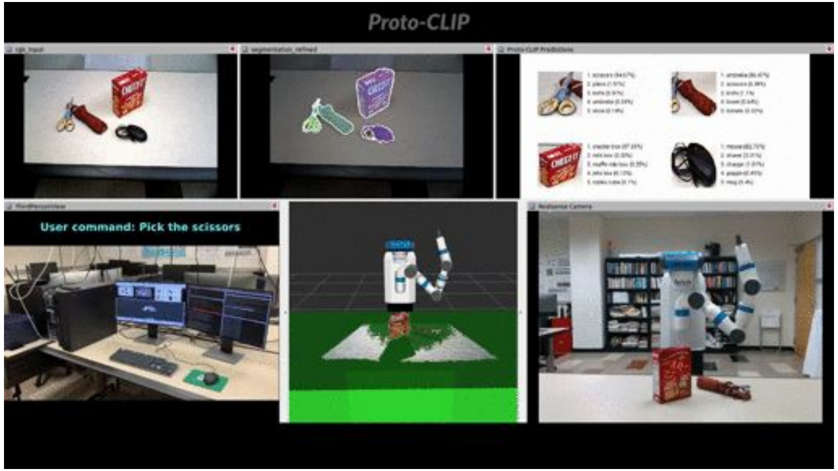
Datasets	**Source**	**Target**	
	ImageNet	-V2 [5]	-Sketch [6]
Zero-Shot-CLIP	60.33	53.27	35.44
Linear Probe CLIP	56.13	45.61	19.13
CoOp	62.95	54.58	31.04
CLIP-Adapter	63.59	55.69	35.68
Tip	62.03	54.60	35.90
Tip-F	65.51	57.11	**36.00**
Proto-CLIP	62.77	55.23	35.62
Proto-CLIP-F	65.75	56.84	35.29
Proto-CLIP-F-Q^T	**65.91**	**57.32**	35.99

Observation: 🏆 **Performs on par with the previous best Tip-A for out-of-distribution (OOD) datasets** 🌍.

Real World Use Case

Joint Object Segmentation and Few-Shot Classification (JOS+FSC) with Object Grasping





🤖 The Fetch robot picks up the object commanded by a user, using classification results from Proto-CLIP 🧠🔍

Real world: 8 sets, each containing 4 different real world objects

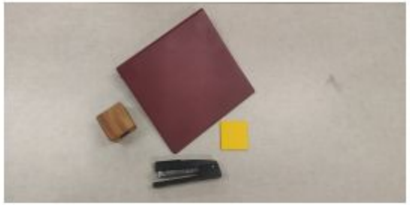
(a) Set-1; mustard_bottle, water_bottle, jello_box, soup_can

(b) Set-2; soft_scrub_cleanser_bottle, tennis_ball, ball, cracker_box

(c) Set-3; cup, jello_box, meat_can, clock

(d) Set-4; tuna_can, air_duster_can, marker, knife

(e) Set-5; keyboard, game_controller, hand_sanitizer, mouse

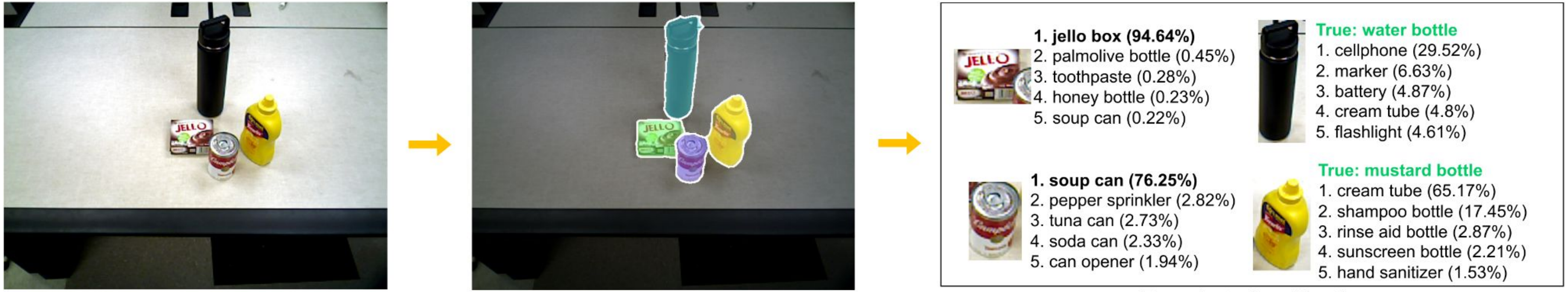


(f) Set-6; wood_block, folder, sticky_notes, stapler

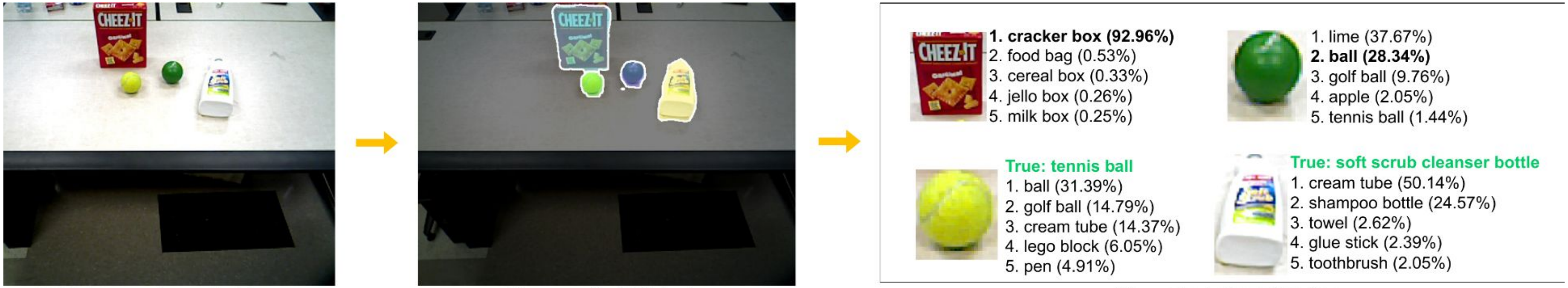
(g) Set-7; key, pen, book, headphone

(h) Set-8; mug, charger, cellphone, spoon

JOS+FSC (Proto-CLIP-F | FewSOL-198)

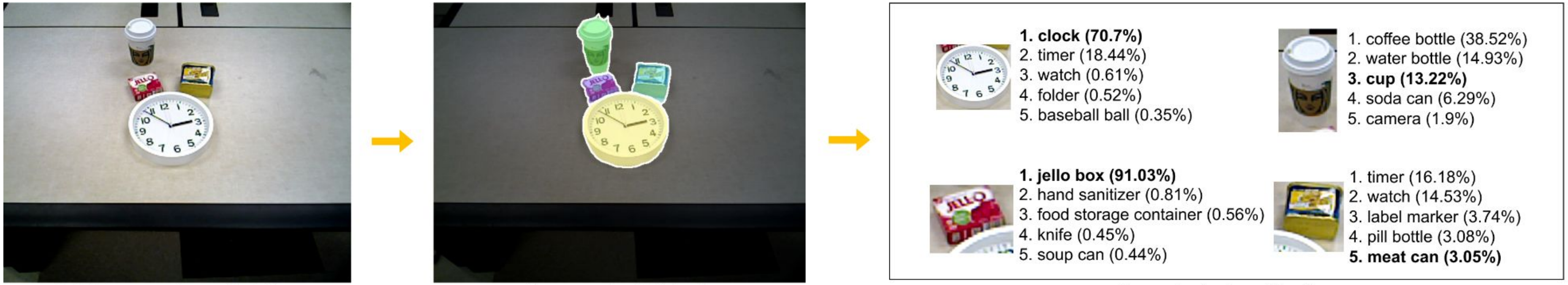


RGB Image from Fetch *Segmented Objects* *Few-shot-classification*

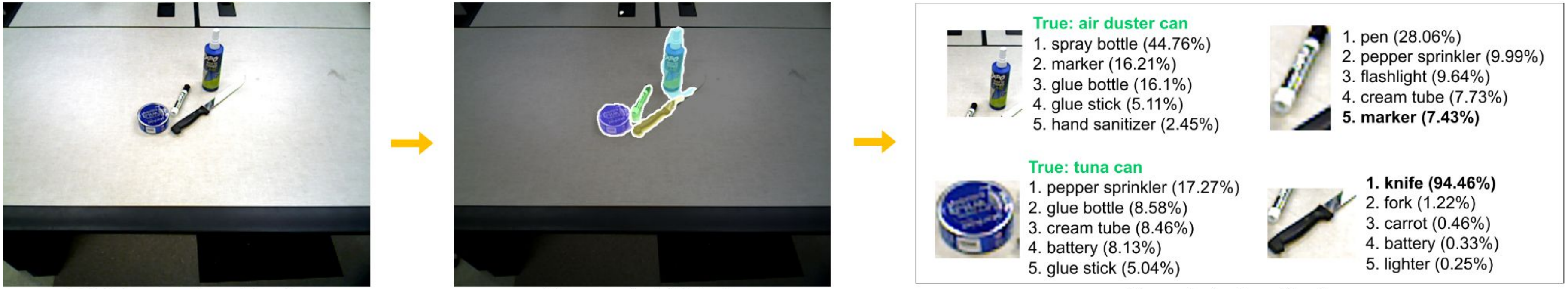


RGB Image from Fetch *Segmented Objects* *Few-shot-classification*

JOS+FSC (Proto-CLIP-F | FewSOL-198)

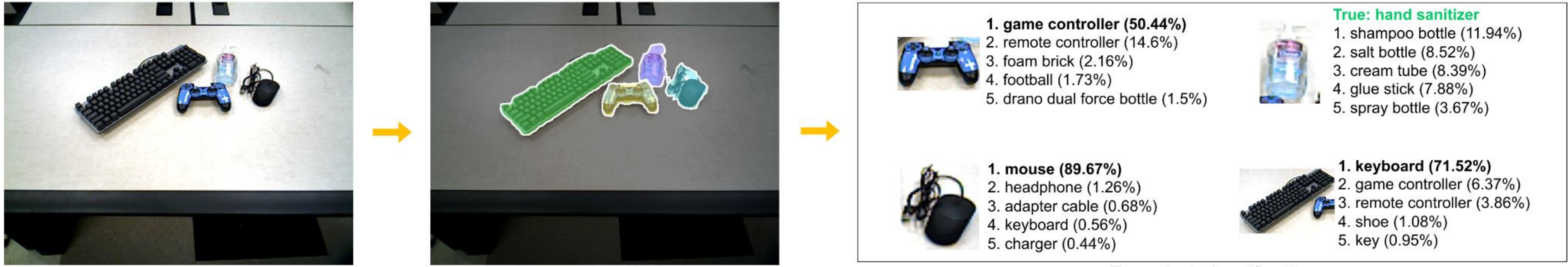


RGB Image from Fetch — *Segmented Objects* — *Few-shot-classification*

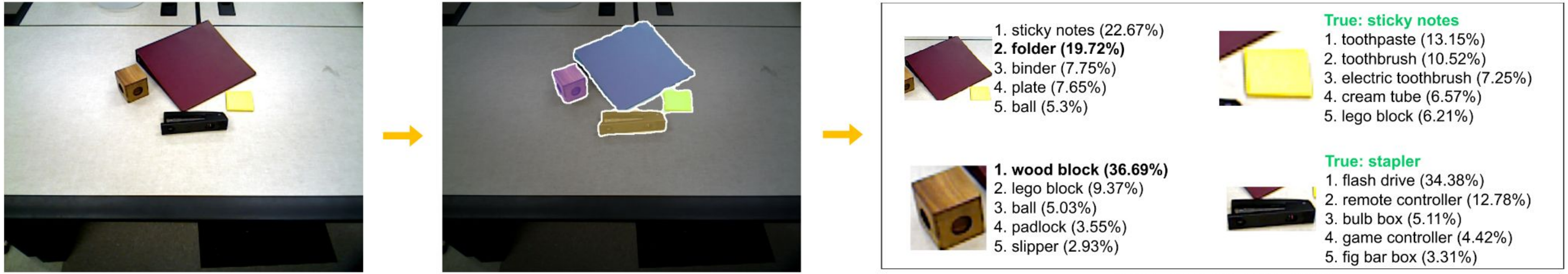


RGB Image from Fetch — *Segmented Objects* — *Few-shot-classification*

JOS+FSC (Proto-CLIP-F | FewSOL-198)

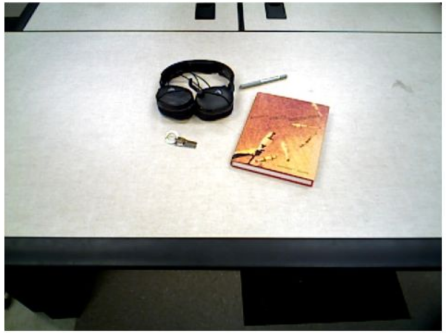


RGB Image from Fetch *Segmented Objects* *Few-shot-classification*



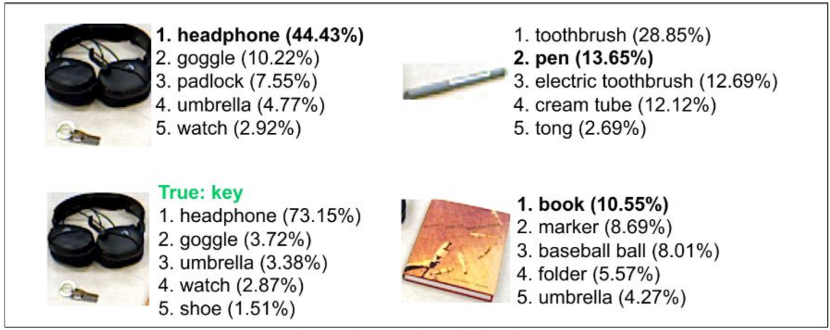
RGB Image from Fetch *Segmented Objects* *Few-shot-classification*

JOS+FSC (Proto-CLIP-F | FewSOL-198)

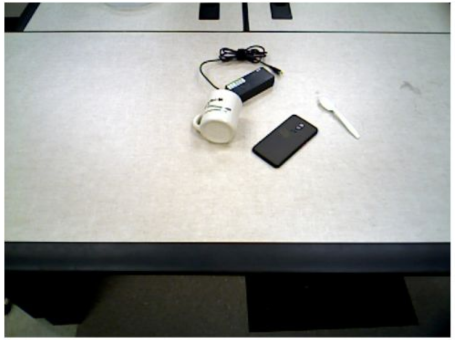


RGB Image from Fetch

Segmented Objects

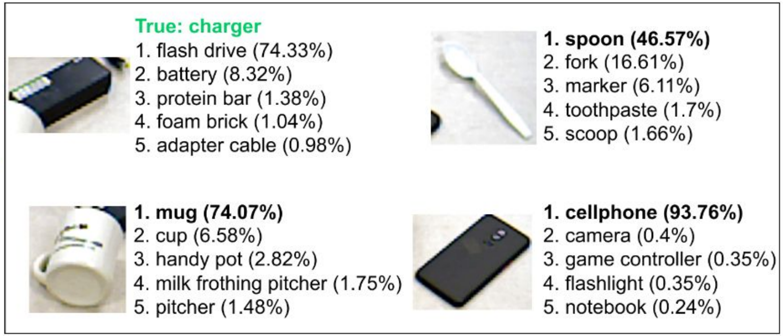


Few-shot-classification



RGB Image from Fetch

Segmented Objects



Few-shot-classification

Contributions

- We introduce Proto-CLIP, a new prototypical network that leverages large-scale vision-language models like CLIP 🖼️📝.
- We've reported its performance across 12 diverse datasets 🌍📊 and conducted real-world testing on a Fetch mobile manipulator 🤖, where Proto-CLIP identifies and grasps objects in cluttered scenes 🧩✋.
- Overall, Proto-CLIP excels in few-shot recognition 🔍 compared to existing methods.

🙏 See you at Poster 4.05!