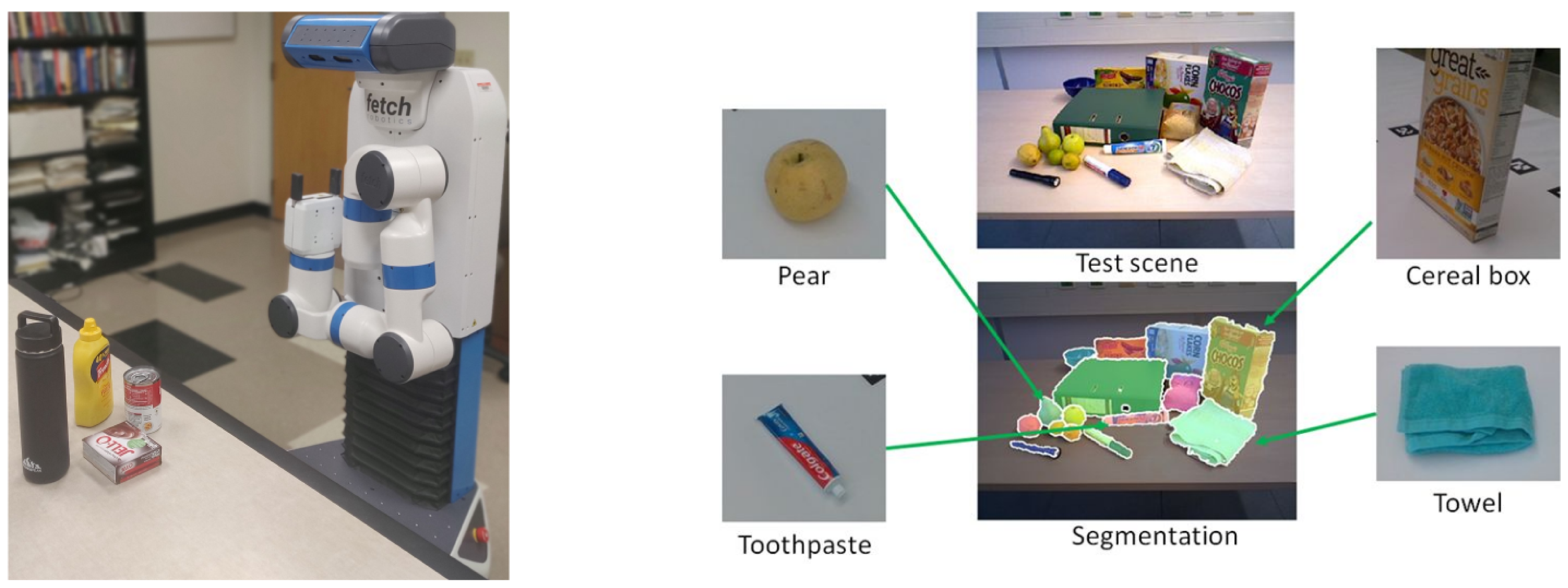
Proto-CLIP: Vision-Language Prototypical Network for Few-Shot Learning

Jishnu Jaykumar Padalunkal [1] Kamalesh Palanisamy [1] Yu-Wei Chao [2] Xinya Du [1] Yu Xiang [1]

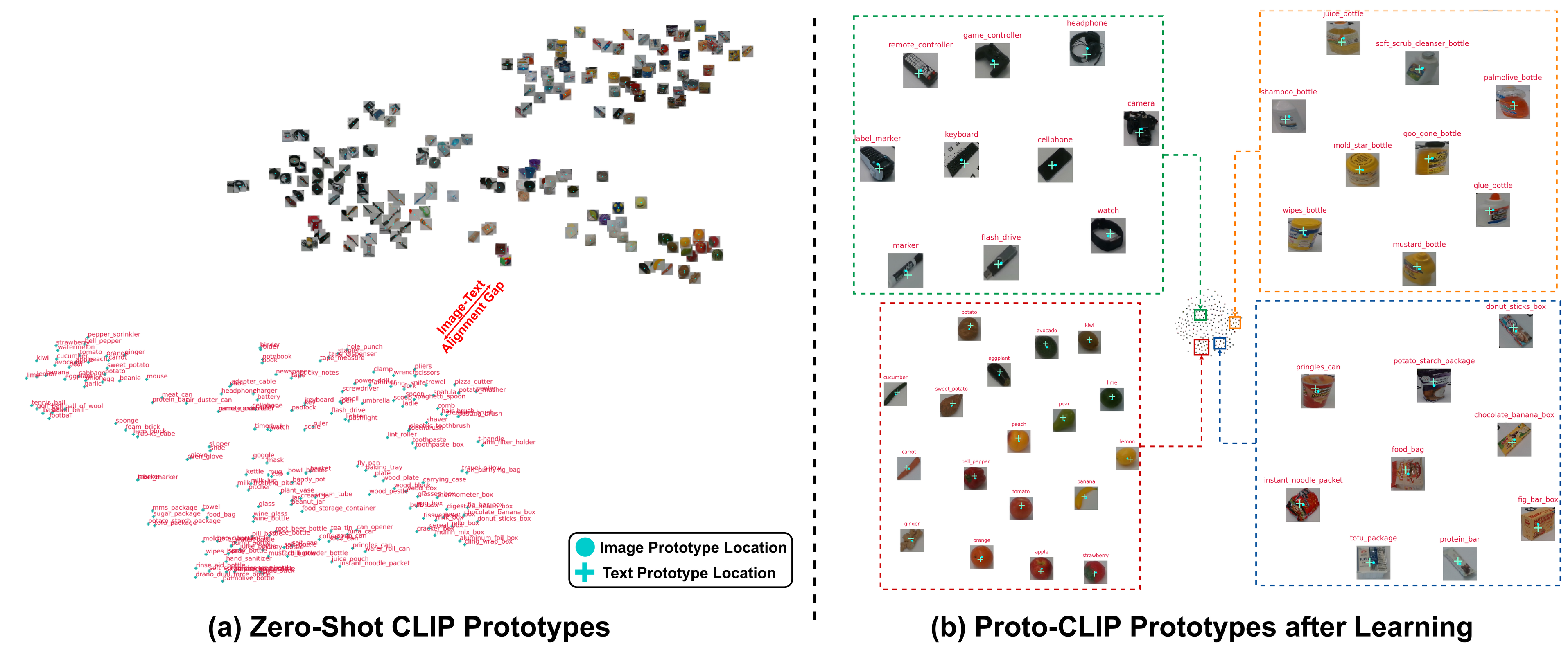
[1]The University of Texas at Dallas [2]NVIDIA

Motivation



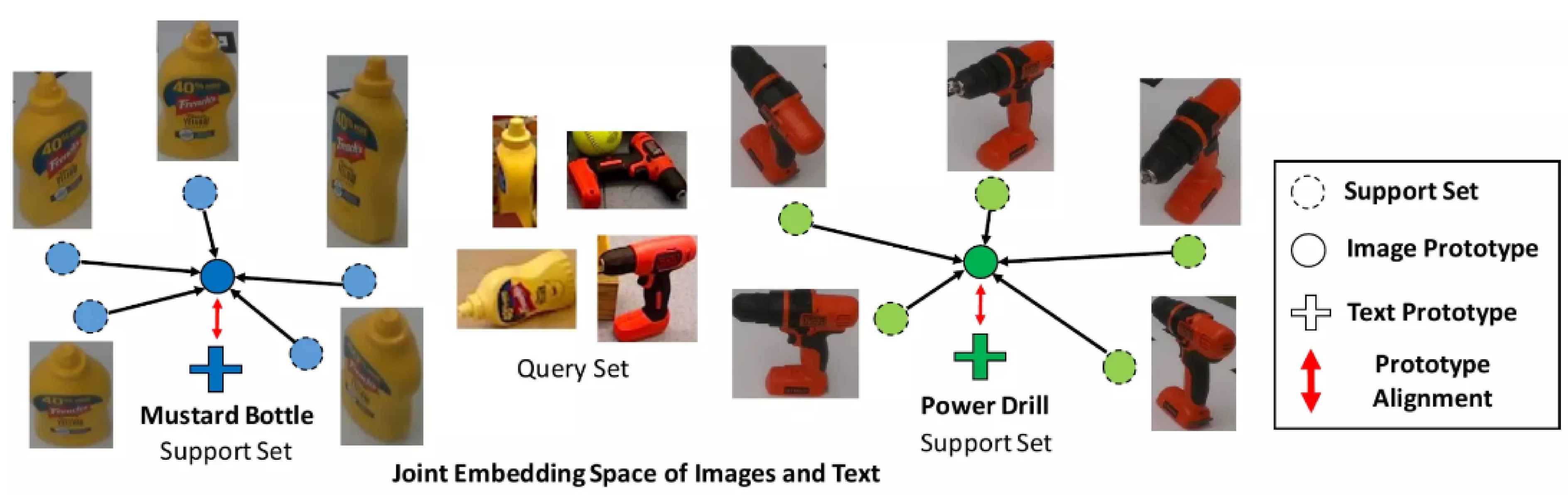
Goal: A robot should identity various (daily) objects in clutter scenes
Our approach: Object Classification using Few-Shot Learning

t-SNE Visualization



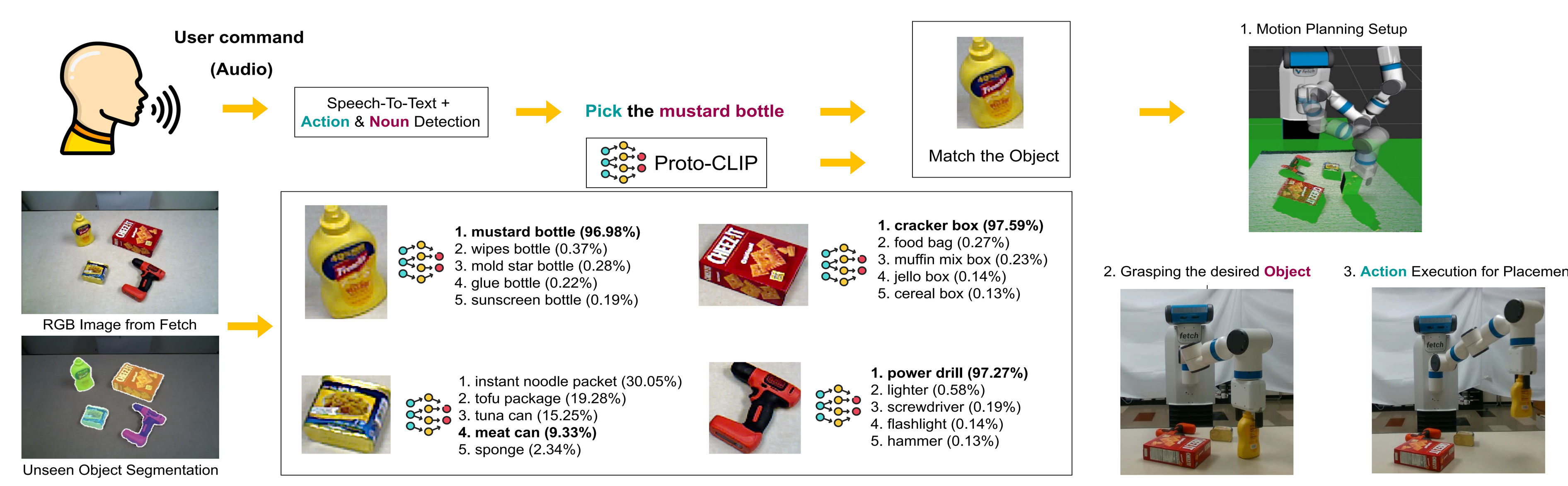
Barnes-Hut t-SNE visualization using the FewSOL-198 dataset. (a) Image and text prototypes from zero-shot CLIP, which are not aligned. (b) Aligned image and text prototypes from **Proto-CLIP**-*F*.

Our Proposal



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.

Real World Use Case



A real world use case of user command oriented grasping. Here, top-5 predictions from the **Proto-CLIP**-*F* (ViT-L/14) model trained on FewSOL-198 are shown. The Speech-To-Text is performed via OpenAI Whisper.

Comparison with related works

Method	"Use Support Sets"	"Adapt Image Embedding"	"Adapt Text Embedding"	"Align Image & Text"
Zero-shot CLIP	✗	✗	✗	✓
Linear-probe CLIP	✓	✓	✗	✗
CoOp	✓	✗	✓	✗
CLIP-Adapter	✓	✓	✓	✗
Tip-Adapter	✓	✓	✓	✗
Sus-X	✓	✓	✗	✗
Proto-CLIP (Ours)	✓	✓	✓	✓

Comparison of our proposed method with the existing CLIP-based few-shot learning methods. "Use Support Sets" indicates if a method uses support training sets for fine-tuning. "Adapt Image/Text Embedding" indicates if a method adapts the image/text embeddings obtained from CLIP. "Align Image and Text" indicates if a method *specifically* aligns images and text in the feature space.

Results

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**

Results of the ablation study of various query adapters and textual memory bank training using the CLIP ResNet50 backbone with $K = 16$ on **Proto-CLIP**-*F*. In case of a tie, the underlined setup was selected randomly.

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
$\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$	**65.75**	**37.56**	**89.62**	75.25	**83.53**	93.43	**71.94**	**68.56**	96.06	**79.09**	77.50	**35.22**

Ablation study of various Loss functions using the CLIP ResNet50 backbone and $K = 16$. The best performing model architectures for each dataset from the previous table are used here.

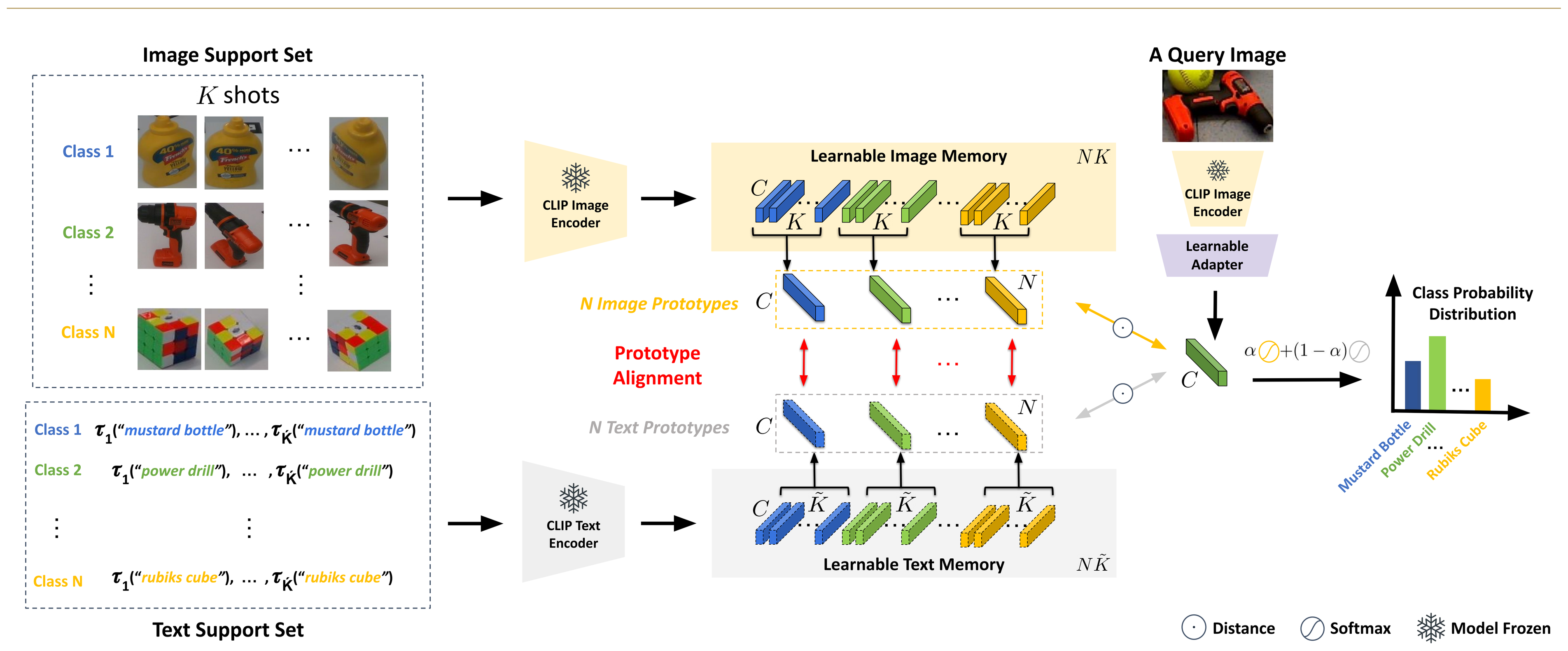
Dataset	Method	1	2	4	8	16	32	64
ImageNet	Tip-Adapter	**60.70**	**60.96**	60.98	61.45	62.01	62.51	62.88
	Proto-CLIP	60.31	60.64	**61.30**	**62.12**	**62.77**	**62.98**	**63.23**
	Tip-Adapter-F	**61.13**	**61.69**	**62.52**	64.00	65.51	66.58	**67.96**
	Proto-CLIP-*F*	60.32	60.64	61.30	63.92	65.75	66.47	65.36
	Proto-CLIP-*F*-Q^T	59.12	60.48	61.80	**64.03**	**65.91**	**66.71**	66.90
FewSOL-52	Tip-Adapter	**27.30**	26.22	28.70	29.22	28.87	✗	✗
	Proto-CLIP	27.09	**28.35**	**29.13**	**29.83**	**29.96**	✗	✗
	Tip-Adapter-F	**27.91**	**27.43**	29.13	32.43	34.04	✗	✗
	Proto-CLIP-*F*	22.22	26.17	27.09	**33.26**	**35.22**	✗	✗
	Proto-CLIP-*F*-Q^T	21.65	25.91	**30.30**	32.70	34.70	✗	✗

Shots ablation results. Backbone='CLIP ResNet50'.

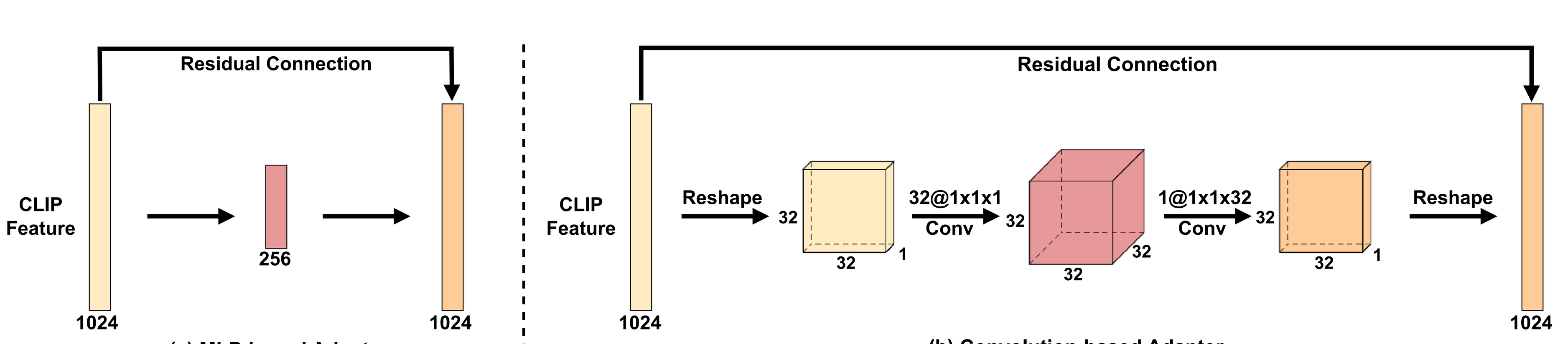
Model	Adapter	TTM	Backbone				
			RN50	RN101	ViT-B/16	ViT-B/32	ViT-L/14
Zero-Shot-CLIP	-	-	25.91	32.96	40.70	41.87	54.57
Tip-Adapter	-	-	29.74	37.43	47.00	41.48	56.78
Tip-Adapter-F	-	-	32.52	41.43	50.17	45.48	60.17
Proto-CLIP-*F*	MLP	✗	33.48	39.04	47.96	41.91	58.65
Proto-CLIP-*F*	MLP	✓	34.83	40.74	47.43	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

Backbone ablation study. Dataset=FewSOL-52. $K = 16$. TTM='Train-Text-Memory'.

Model Overview



Overview of our proposed **Proto-CLIP** model. The image memory, the text memory and the adapter network are learned. Given a class name, τ_i returns the i^{th} out of K predefined text prompts.



Two designs of the adapters.(a) A Multi-layer perceptron-based adapter as in CLIP-Adapter. (b) A convolution-based adapter that we introduce. The feature dimension is for CLIP ResNet50 backbone.

Out of Distribution (OOD)

Datasets	Source	Target	
	ImageNet	-V2	-Sketch
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

OOD accuracy study using Imagenet-V2 and ImageNet-Sketch datasets.

Limitations and Future Work

- Challenges in extreme low-shot scenarios ($K \leq 2$)
- Requires hyperparameter tuning for each specific dataset and backbone.
- Future work will focus on enhancing feature representation learning beyond current CLIP models. One potential avenue is adapting more powerful vision-language models like GPT variants. The FewSOL dataset also offers multiview and depth information about objects, making 3D exploration in few-shot object recognition a promising direction.

Acknowledgments

This work was supported in part by the **DARPA Perceptually-enabled Task Guidance (PTG)** Program under contract number **HR00112220005**.

Scan Me!