Supplementary Materials: Caption-based MultiModal Adapter in Zero-Shot Classification

Anonymous Authors

1 OVERVIEW

In this supplementary material, we present detailed results and analyses across various aspects of our research:

- Detailed results for each of the five CLIP backbone networks across multiple datasets are thoroughly documented in **Figure** 1, 2, 3, 4, 5.
- Assessments of CLIP similarity between support and target distribution (**Table** 6), optimal support set sizes (**Table** 7), and scalability of performance as support set sizes increase are presented (**Figure** 1).
- Performance comparisons of our training-free few-shot classification method (**Figure** 2).
- The prompt used in caption generation (**Listing** 1).

2 DETAILED RESULTS FOR EACH BACKBONE NETWORKS

We conducted our experiments in using five CLIP[3] backbone networks as encoders: ResNet-50, ResNet-101, ViT-B/32, ViT-B/16, and ViT-L/14 [2]. We reported the average results across these five backbone networks for each dataset in the main text. The detailed results for each backbone network are shown in **Table** 1, 2, 3, 4, 5

The ***CapS-Adapter*** performed better on five backbones—ResNet-50, ResNet-101, ViT-B/32, ViT-B/16, and ViT-L/14—compared to the better performer between SuS-X-SD-CuPL and SuS-X-SD-Photo, with average improvements of 2.53%, 2.82%, 2.06%, 1.57%, and 1.62%, respectively. It also outperformed the zero-shot CLIP by 6.37%, 4.98%, 5.17%, 4.60%, and 4.70 % respectively.

3 DETAILS ABOUT CLIP SIMILARITY RESULTS

To evaluate whether the image distribution of the support sets closely resembles the target data distribution, we adopt the method of calculating the average CLIP similarity between the images in the support set and the test set of the target dataset. All results on 19 datasets are shown in **Table** 6.

The CLIP similarity score of *CapS* is on average 1.50% and 2.71% higher than SuS-SD-Photo and SuS-SD-CuPL, respectively, and achieved the highest value among the three methods in 10 out of 19 datasets.

4 BEST SUPPORT SET SIZES

In our main results, for the support set-based methods SuS-X-SD-CuPL, SuS-X-SD-Photo[4], and ***Caps-Adapter*** (Ours), we compared performances across 5, 10, 25, 50, 75, and 100 support set images per class, selecting a specific size of support set to achieve great performance. The number of images in the *Caps* for each dataset is listed in **Table** 7.

5 COMPARISON AS SUPPORT SET SIZE INCREASE

We visualized the changes in classification accuracy for SuS-X-SD-CuPL, SuS-X-SD-Photo, and ***Caps-Adapter*** datasets as the size of the support set increased (image numbers = 5, 10, 25, 50, 75, 100) in **Figure** 1.

In some datasets where SuS-X-SD-CuPL and SuS-X-SD-Photo exhibited a trend of decreasing accuracy as the size of the support set increased, ***Caps-Adapter*** (depicted by the blue line in **Figure** 1) showed a trend of increasing accuracy with the growth of the support set size. Even in cases where all three methods showed a declining trend, the decrease in ***Caps-Adapter*** was more gradual, primarily due to the images in the caps being closer to the target distribution.

6 TRAING-FREE FEW SHOT CLASSIFICATION WITH M-ADAPTER

We adapt *M-Adapter* method to the training-free few-shot adaptation regime and compared it with the current state-of-the-art (SOTA) model, TIP-X. We conducted this experiment using 1, 2, 4, 8, 16 shots. The results across 8 datasets and the average result are presented in **Figure** 2.

In these 8 datasets, when using exactly the same few-shot image features, M-Adapter outperforms TIP-X by an average of 0.57% across all shots. In these datasets, *M-Adapter* (represented by the blue line in **Figure** 2) consistently outperformed TIP-X. We believe this is due to M-Adapter effectively balancing inter-modal and intra-modal correlations by incorporating text features from caption-based prompts into inference, aligning with our analysis in our **Ablation Study**.

7 PROMPT USED WHEN GENERATING CAPTIONS

```
prompt =
"""
<|User|>:
    Generate a concise and accurate description for the
    following image. Please ensure to include key
    elements and any details.

<|Bot|>:

"""
```

Listing 1: Prompt used when generating captions

We provided the manually crafted prompt we use for generating image captions through multimodal large language models in **Listing**1. Due to extensive fine-tuning aimed at enhancing captioning capabilities, *ShareCaptioner*[1] is relatively insensitive to variations in prompts. Consequently, the quality of the captions it generates is minimally impacted by changes in the prompt, allowing us to utilize simpler prompts.

Table 1: Detailed results for RN50. *Avarage is calculated across 19 datasets.

	Birdsnap	CIFAR-10	CIFAR-100	CUB	Caltech101	Caltech256	Country211	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	ImageNet	ImageNet-R	ImageNet-Sketch	OxfordPets	SUN397	StanfordCars	UCF101	Average*
ZS-CLIP	30.60	73.04	40.58	41.23	85.96	78.97	13.45	41.02	26.84	16.77	62.89	74.07	60.33	59.54	35.45	81.85	59.24	56.31	55.49	52.30
CuPL	35.63	74.18	42.87	48.90	89.21	80.29	13.26	48.70	38.35	19.59	65.57	76.95	57.52	61.17	35.07	84.87	62.72	57.22	59.00	55.32
CuPL+e	35.79	74.62	43.40	48.80	89.37	80.55	**14.26**	47.51	37.15	19.29	66.01	77.51	58.98	61.15	**35.86**	85.09	63.08	57.19	61.22	55.62
SUS-X-SD-Photo	36.52	74.67	43.99	49.00	89.21	**80.67**	14.11	49.35	41.25	19.08	**66.75**	77.59	**60.50**	**61.27**	35.41	85.66	63.02	**57.24**	61.46	56.14
SUS-X-SD-CuPL	37.21	74.67	44.75	49.15	89.33	80.60	14.15	50.41	37.84	19.62	66.67	77.52	**60.50**	61.23	35.43	85.17	63.08	57.21	61.30	56.10
CapS-Adapter (Ours)	**38.77**	**75.44**	**45.95**	**49.19**	**89.45**	80.65	**14.26**	**59.93**	**54.81**	**24.54**	66.63	**78.58**	60.34	61.26	35.46	**87.79**	**64.72**	57.06	**69.81**	**58.67**

Table 2: Detailed results for RN101. *Avarage is calculated across 19 datasets.

	Birdsnap	CIFAR-10	CIFAR-100	CUB	Caltech101	Caltech256	Country211	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	ImageNet	ImageNet-R	ImageNet-Sketch	OxfordPets	SUN397	StanfordCars	UCF101	Average*
ZS-CLIP	30.64	**79.13**	48.83	41.78	89.45	82.98	**15.32**	41.55	24.80	17.61	**62.08**	77.83	62.52	66.35	40.62	83.65	59.95	**62.65**	57.92	55.03
CuPL	31.32	74.10	37.45	45.25	92.09	82.06	11.41	49.29	28.88	20.55	59.20	77.18	55.93	68.04	35.86	85.53	61.72	61.33	54.56	54.30
CuPL+e	32.29	74.55	40.58	46.79	92.12	83.29	12.41	49.11	26.46	19.38	60.98	78.60	58.71	68.34	38.16	86.73	62.53	61.67	59.24	55.37
SUS-X-SD-Photo	35.26	74.64	46.61	47.12	**92.25**	83.26	12.40	51.41	35.77	20.76	61.27	79.18	**62.77**	68.35	40.71	87.38	62.45	61.31	61.59	57.08
SUS-X-SD-CuPL	35.81	76.05	47.35	47.07	92.17	83.27	12.46	51.60	36.35	21.27	61.19	79.06	62.64	**68.36**	**40.76**	86.84	62.48	61.51	60.38	57.19
CapS-Adapter (Ours)	**40.14**	75.16	**50.04**	**47.17**	92.21	**83.30**	12.56	**60.99**	**50.77**	**26.28**	61.14	**81.85**	62.53	68.34	40.74	**89.48**	**65.00**	61.50	**70.90**	**60.01**

Table 3: Detailed results for ViT-B/32. *Avarage is calculated across 19 datasets.

	Birdsnap	CIFAR-10	CIFAR-100	CUB	Caltech101	Caltech256	Country211	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	ImageNet	ImageNet-R	ImageNet-Sketch	OxfordPets	SUN397	StanfordCars	UCF101	Average*
ZS-CLIP	32.10	**89.87**	64.07	45.06	91.68	82.79	15.48	43.09	43.04	18.60	63.87	78.86	63.80	68.62	42.17	80.92	62.79	60.10	61.14	58.32
CuPL	37.72	88.38	64.34	53.49	**93.35**	84.60	14.89	50.00	51.35	20.40	66.91	80.15	60.89	69.91	41.73	86.40	65.49	61.22	63.15	60.76
CuPL+e	37.69	88.66	64.95	53.50	93.06	84.86	15.63	49.53	50.48	20.49	66.83	80.61	62.42	69.99	**42.48**	86.94	65.91	**61.35**	65.00	61.07
SUS-X-SD-Photo	38.38	88.63	65.25	53.47	93.06	84.91	15.63	51.42	55.78	**26.58**	**68.05**	80.73	**63.96**	**70.02**	42.16	88.06	65.83	61.21	65.29	62.02
SUS-X-SD-CuPL	38.81	88.71	65.21	**53.61**	92.92	84.87	15.64	51.65	51.37	20.55	67.80	80.62	63.86	69.97	42.19	87.19	65.68	61.31	65.16	61.43
CapS-Adapter (Ours)	**40.54**	88.65	**65.95**	53.40	93.06	**84.93**	**15.66**	**60.99**	**61.89**	25.77	67.28	**80.99**	63.77	70.00	42.20	**89.62**	**67.27**	61.11	**73.20**	**63.49**

Table 4: Detailed results for ViT-B/16. *Avarage is calculated across 19 datasets.

	Birdsnap	CIFAR-10	CIFAR-100	CUB	Caltech101	Caltech256	Country211	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	ImageNet	ImageNet-R	ImageNet-Sketch	OxfordPets	SUN397	StanfordCars	UCF101	Average*
ZS-CLIP	38.46	**91.12**	67.25	49.34	93.51	86.05	19.44	45.04	50.34	23.13	66.95	84.43	68.83	76.93	48.38	86.94	65.63	66.16	65.16	63.19
CuPL	43.38	89.82	68.01	56.77	94.12	87.38	19.24	53.25	55.64	27.27	72.80	85.79	66.71	77.71	47.88	90.30	67.89	66.12	66.38	65.36
CuPL+e	43.53	89.82	68.47	57.30	**94.20**	87.42	20.14	52.83	55.27	27.42	72.84	86.24	67.98	77.83	**48.45**	90.49	68.10	**66.35**	68.57	65.76
SUS-X-SD-Photo	44.98	90.02	68.76	**57.59**	93.96	87.37	20.10	53.84	59.28	27.51	72.55	86.37	**69.09**	77.83	**48.45**	91.66	67.98	66.21	68.81	66.22
SUS-X-SD-CuPL	45.24	90.31	68.65	57.51	93.83	87.44	20.10	54.20	56.62	28.26	**73.04**	86.47	68.92	77.85	**48.45**	90.43	67.87	66.19	68.52	66.10
CapS-Adapter (Ours)	**47.37**	90.29	**69.50**	56.75	94.08	**87.49**	**20.15**	**63.53**	**65.96**	**33.30**	72.84	**86.87**	68.90	**77.92**	48.44	**92.40**	**69.36**	66.29	**76.55**	**67.79**

Table 5: Detailed results for ViT-L/14. *Avarage is calculated across 19 datasets.

	Birdsnap	CIFAR-10	CIFAR-100	CUB	Caltech101	Caltech256	Country211	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	ImageNet	ImageNet-R	ImageNet-Sketch	OxfordPets	SUN397	StanfordCars	UCF101	Average*
ZS-CLIP	46.46	**95.84**	76.20	56.42	94.16	89.04	28.23	56.26	50.89	30.42	77.30	90.35	75.95	87.29	59.70	92.64	69.56	77.90	72.64	69.86
CuPL	50.61	95.47	77.00	62.60	96.75	90.32	28.00	61.93	62.67	35.94	79.33	91.00	73.38	87.87	58.87	93.27	71.76	78.35	71.72	71.94
CuPL+e	50.91	95.57	77.45	**62.99**	96.59	90.22	**29.07**	62.17	62.42	34.95	**80.51**	91.29	74.58	88.11	59.32	93.76	72.30	**78.52**	74.81	72.40
SUS-X-SD-Photo	52.54	95.66	78.05	62.94	96.80	90.30	29.00	62.12	66.73	36.15	80.39	91.40	**76.04**	88.09	**59.71**	94.63	72.28	78.44	74.68	72.94
SUS-X-SD-CuPL	52.52	95.67	78.09	62.75	**96.84**	**90.33**	29.00	61.82	65.95	35.55	80.35	91.37	76.00	**88.12**	59.69	93.92	72.32	78.37	74.46	72.80
CapS-Adapter (Ours)	**54.17**	95.74	**79.21**	62.91	96.75	**90.33**	29.04	**70.33**	**72.17**	**42.81**	**80.51**	**91.72**	75.98	88.09	59.69	**95.26**	**73.52**	78.45	**80.02**	**74.56**

Table 6: Comparison of CLIP similarity(%) between images in support set and target test set. *Avarage is calculated across 19 datasets.

	Birdsnap	CIFAR-10	CIFAR-100	CUB	Caltech101	Caltech256	Country211	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	ImageNet	ImageNet-R	ImageNet-Sketch	OxfordPets	SUN397	StanfordCars	UCF101	Average*
SUS-SD-CuPL	67.77	62.47	65.83	68.52	66.88	81.87	55.92	93.55	77.11	**66.08**	94.06	64.93	54.35	61.92	77.67	84.97	57.35	72.54	54.83	69.93
SUS-SD-Photo	68.19	63.62	67.86	68.97	67.76	**84.03**	**58.11**	92.98	**80.42**	64.47	**95.41**	66.10	**56.45**	58.62	**81.13**	88.08	**58.41**	**73.67**	57.43	71.14
CapS (Ours)	**79.95**	**64.85**	**69.56**	**76.77**	**84.46**	79.95	51.74	**93.60**	73.98	63.30	86.69	**79.12**	55.26	**72.29**	66.83	**94.66**	55.71	60.52	**70.86**	**72.64**

Table 7: Best support set size of *Caps-Adapter* under 5 CLIP backbones.

Dataset	**CLIP Backbone**				
	RN50	**RN101**	**ViT-B/32**	**ViT-B/16**	**ViT-L/14**
Birdsnap	37500	50000	25000	37500	37500
CIFAR-10	50	1000	50	100	250
CIFAR-100	5000	10000	500	5000	10000
CUB	10000	5000	2000	5000	2000
Caltech101	505	505	505	505	2525
Caltech256	2570	12850	19275	1285	25700
Country211	2110	5275	5275	1055	2110
DTD	4700	4700	4700	4700	3525
EuroSAT	500	750	500	250	250
FGVCAircraft	7500	5000	10000	10000	10000
Flowers102	2550	7650	2550	2550	510
Food101	1010	7575	5050	5050	7575
Imagenet	10000	10000	5000	10000	10000
Imagenet-R	1000	2000	2000	2000	2000
Imagenet-Sketch	5000	5000	25000	5000	5000
OxfordPets	2775	1850	1850	2775	1850
SUN397	19850	29775	19850	29775	19850
StanfordCars	4900	1960	980	980	1960
UCF101	10100	7575	5050	7575	7575

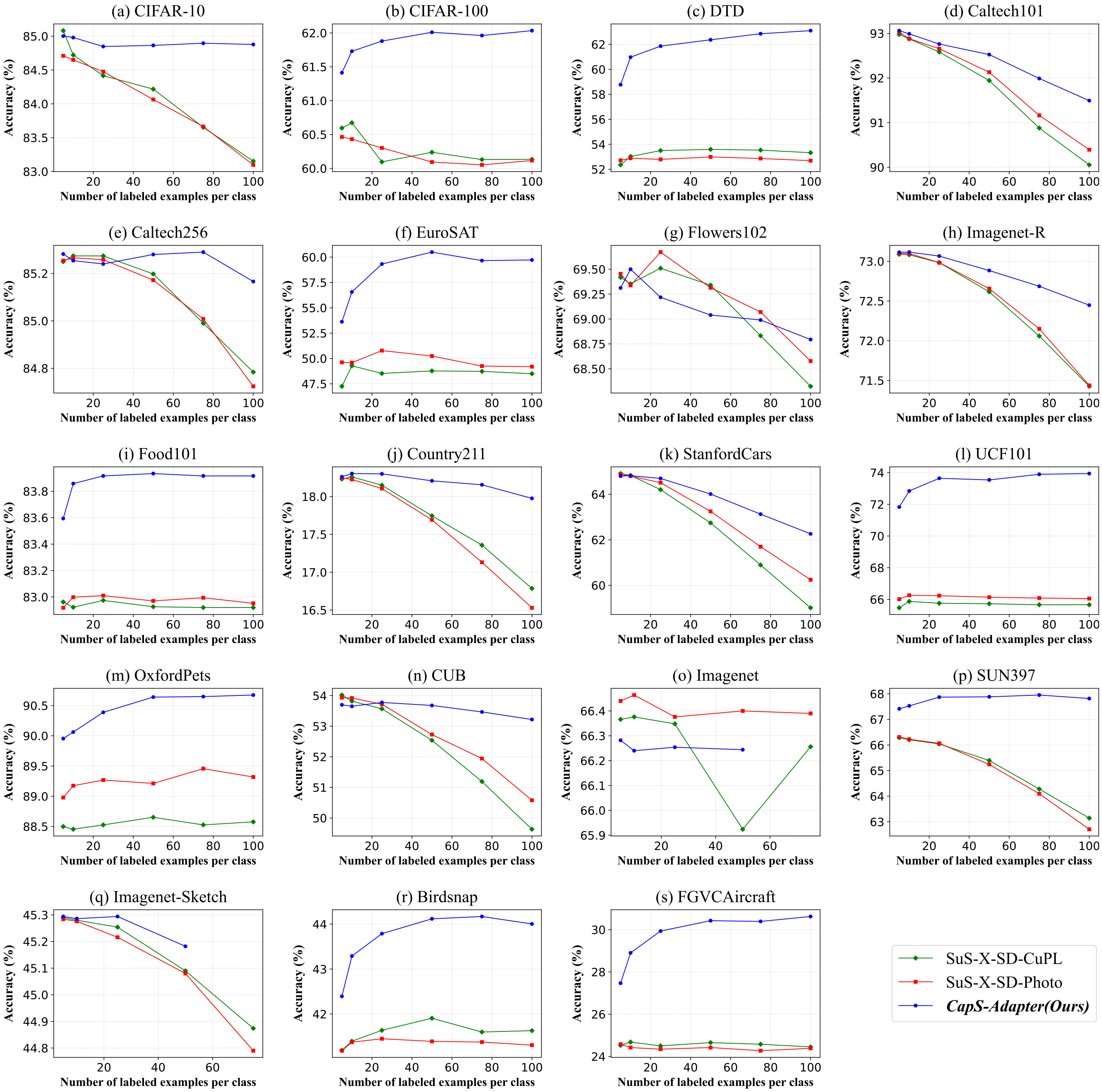

Figure 1: Changes in classification accuracy with the size of the support set, comparing SuS-X-SD-CuPL, SuS-X-SD-Photo, and *Caps-Adapter*.

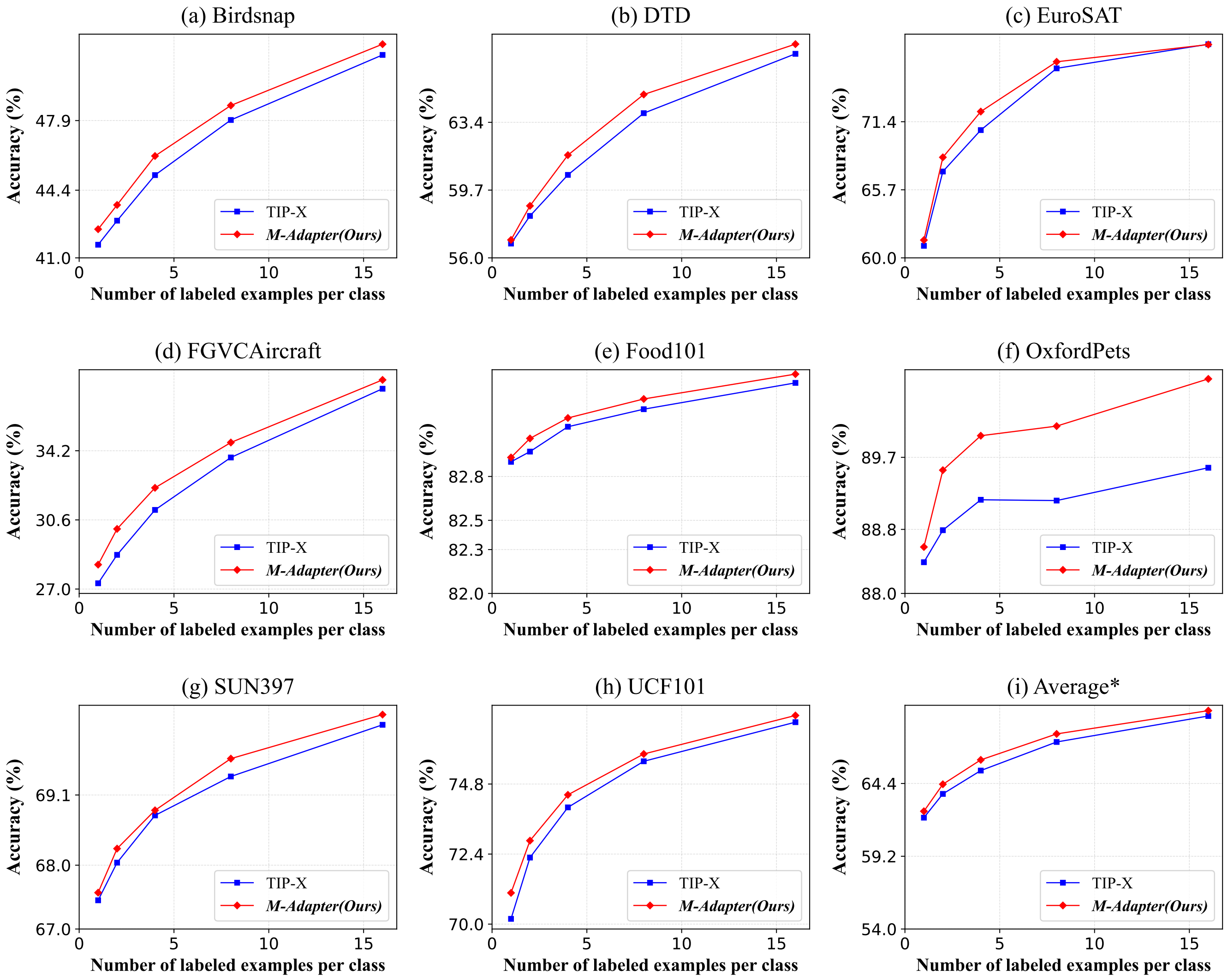

Figure 2: Comparison of TIP-X and *M-Adapter*'s performance under trainig-free few-shot experiment setting. *Avarage is calculated across 8 datasets.

REFERENCES

[1] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793* (2023).

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. 8748–8763.

[4] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. 2022. Sus-x: Training-free name-only transfer of vision-language models. *arXiv preprint arXiv:2211.16198* (2022).