Appendix

8. Experimental Details

In this section, we go through the hyperparameter details for all the experiments for reproducibility.

Basic settings: We follow the original CLIP [81] to L2-normalize the features after the encoder before sending them into the linear layer. We also use the L2-normalized text features to initialize the final linear layer weight following WiSE-FT [100]. For all cross-modal adaptation experiments, half of the batch is image samples and the other half is text samples. For all experiments, we use AdamW optimizer following WiSE-FT [100] and tune the hyperparameters including initial learning rate, weight decay, and batch size on the few-shot validation set. We perform a learning rate warmup with 50 iterations, during which the learning rate goes up linearly from 0.00001 to the initial value. We then perform a cosine annealing learning rate scheduling over the course of 12800 iterations. We do early stopping based on the few-shot validation set performance evaluated every 100 iterations. Furthermore, because the logit scale (inverse of softmax temperature) is a learnable weight clipped at 100 during CLIP-pretraining [81], we reuse the given logit scale of 100 for all experiments except for partial finetuning, where we find lowering it to 50 can improve validation performance. Future work may choose to set the logit scale as a learnable parameter instead.

We now report the range of hyperparameter search for each method. Note that the search range is kept the same for all 11 target datasets.

Linear Probing: For all linear probing experiments, we perform a grid search of learning rate in $[0.001, 0.0001]$, weight decay in $[0.0, 0.01, 0.0001]$, and batch size in $[8, 32]$.

WiSE-FT: To compare with linear probing, we adopt the same procedure above to train the linear classifier and then perform post-hoc ensembling with the text-based classifier with a fixed ratio of 0.5.

Partial Finetuning: For all partial finetuning experiments, we perform a grid search of learning rate in $[0.00001, 0.000001, 0.0000001]$, weight decay in $[0.0, 0.001, 0.00001]$, and batch size is set to 8. CLIP [81] adopts a modified version of ResNet-50 image encoder, in which the final average pooling layer is replaced by an attentional pooling layer. We thus choose this layer as the finetuning target for all ResNet-50 experiments. For ViT-B/16 encoder, we simply finetune the last transformer layer. In the next section, we also show that finetuning the text encoder is not as effective.

Cross-modal Prompting: We follow the same setup and hyperparameters used in CoOp [113]. We use the ResNet-50 backbone with 16 learnable tokens, and append the class name to the end of the tokens. Following CoOp, we use SGD with a learning rate of 0.002, decayed using the cosine annealing rule. We train for 200 epochs for 8 and 16 shots, 100 epochs for 2 and 4 shots, and 50 epochs for 1 shot (except ImageNet which is fixed at 50 epochs). The learning rate for the first epoch is fixed at 0.00001. We also use the same random resized crop transformations as CoOp.

Included Dataset	ESC-50 [77] Class	ImageNet [15] Class
ImageNet-ESC-19	rooster	rooster
	hen	hen
	chirping-birds	chickadee
	frog	tree frog
	dog	otterhound
	cat	egyptian cat
	insects	fly
	crickets	cricket
	pig	pig
	sheep	big-horn sheep
	airplane	airliner
	train	high-speed train
	chainsaw	chainsaw
	keyboard-typing	computer keyboard
	clock-alarm	digital clock
	mouse-click	computer mouse
	vacuum-cleaner	vacuum cleaner
	clock-tick	wall clock
	washing-machine	washing machine
ImageNet-ESC-27	can-opening	can opener
	church-bells	church bells
	crackling-fire	fire screen
	toilet-flush	toilet seat
	water-drops	sink
	drinking-sipping	water bottle
	pouring-water	water jug
	sea-waves	sandbar

Table 9. **ImageNet-ESC dataset class matchings.**

Cross-modal Adapter: We follow the same 2-layer MLPs architecture in CLIP-Adapter [21] with a residual ratio of 0.2. Specifically, the first linear layer downsizes the input feature to $\frac{1}{4}$ of the original dimension and the second linear layer transforms it back to the original dimension. Each linear layer is followed by a ReLU function. Finally, the transformed features are multiplied by 0.2 and added with 0.8 * the original feature. We use a single adapter for both image and text features. We perform a grid search of learning rate in $[0.0001, 0.00001, 0.000001, 0.0000001]$, weight decay in $[0.0, 0.001, 0.00001]$, and batch size is set to 8. We do not adopt the cache-modal and training-free initialization proposed in the follow-up Tip-Adapter [111] method. Also, we notice that Tip-Adapter uses test set to perform early stopping; we however strictly follow the CoOp protocol to use the few-shot validation set for all hyperparameter searching.

ImageNet-ESC Experiments: For all linear probing experiments on ImageNet-ESC, we perform a grid search of learning rate in $[0.1, 0.01, 0.001, 0.0001]$, weight decay in $[0.0, 0.01, 0.0001]$, and batch size is 8.

9. Additional Results

In this section, we present all the results with standard deviation over multiple runs. Here is an overview (please refer to table captions for more discussion):

1. **Per-dataset results for all methods:** We show Figure 6 and Table 10. In particular, we note that cross-modal adaptation consistently outperforms prior methods across a wide variety of visual recognition datasets, further strengthening our claim that our approach should be the de-facto adaptation method for finetuning multimodal models.

2. **Ablation for augmentation techniques:** In Table 11, we show the performance of all combinations of image and text augmentation techniques. Importantly, simple *text* augmentation strategies work very well for *visual* recognition.

3. **Ablation for classifier initialization:** In Table 12, our experiments suggest that (a) text-based initialization is beneficial for both linear and partial finetuning, and (b) cross-modal adaptation can improve the performance regardless of the initialization.

4. **Ablation for partial finetuning:** In Table 13, we confirm that partial finetuning of the image encoder is more effective than finetuning the text encoder.

5. **Complete results for all reported methods:** In Table 14, we show the standard deviation for all methods reported in the main paper and appendix, including ViT-based encoder results.

6. **Complete results on ImageNet-ESC benchmark:** We show the complete results on ImageNet-ESC-19 and ImageNet-ESC-27 for both image-classification in Table 15 and audio-classification in Table 16. We additionally include the results of the text-based classifier and cross-modal linear probing with all three modalities (including text) for reference. Including the text modality seems to be the most performant, which is expected since the benchmark is curated based on textual information, i.e. matching label names. We also note that just adding text modality is better than including all three modalities; we believe this issue can be alleviated with better alignment between the image and audio representations, e.g. scaling the pre-training data for AudioCLIP. Furthermore, the standard deviations of the experiments are higher than those of the vision-language adaptation experiments because the randomly sampled one-shot sample can make a huge difference in the performance. However, cross-modal adaptation is more performant not by chance – in more than 75% of the experiments, adding the one-shot-audio or one-shot-image to the same set of samples can outperform uni-modal linear probing.

7. **Comparison to ProDA [63]:** In Table 17, we compare to ProDA, another promising SOTA method that does automatic prompt ensembling with 36 learned templates. We are told by the authors that they do not follow the dataset split given by CoOp [113], and use the official test split of each dataset whenever possible or sample their own test split from the train set. Therefore, we cannot directly compare to their performance since CoOp [113] use their own test split for most datasets and ProDA does not release the code yet. In particular, official test sets exist for two of the target datasets (Food101 [6] and DTD [14]). We therefore switch to the official test split for these two datasets and use the CoOp's split for the rest of the 9 datasets in Table 17 as our best attempt to compare to ProDA [63]. Note that ProDa also does not report the use of a few-shot validation set. In conclusion, our approach is still more performant than theirs under most scenarios with significantly fewer training resources.

8. **180 templates used for mining:** In Table 18, we show the pool of templates we use when mining based on few-shot validation set performance.

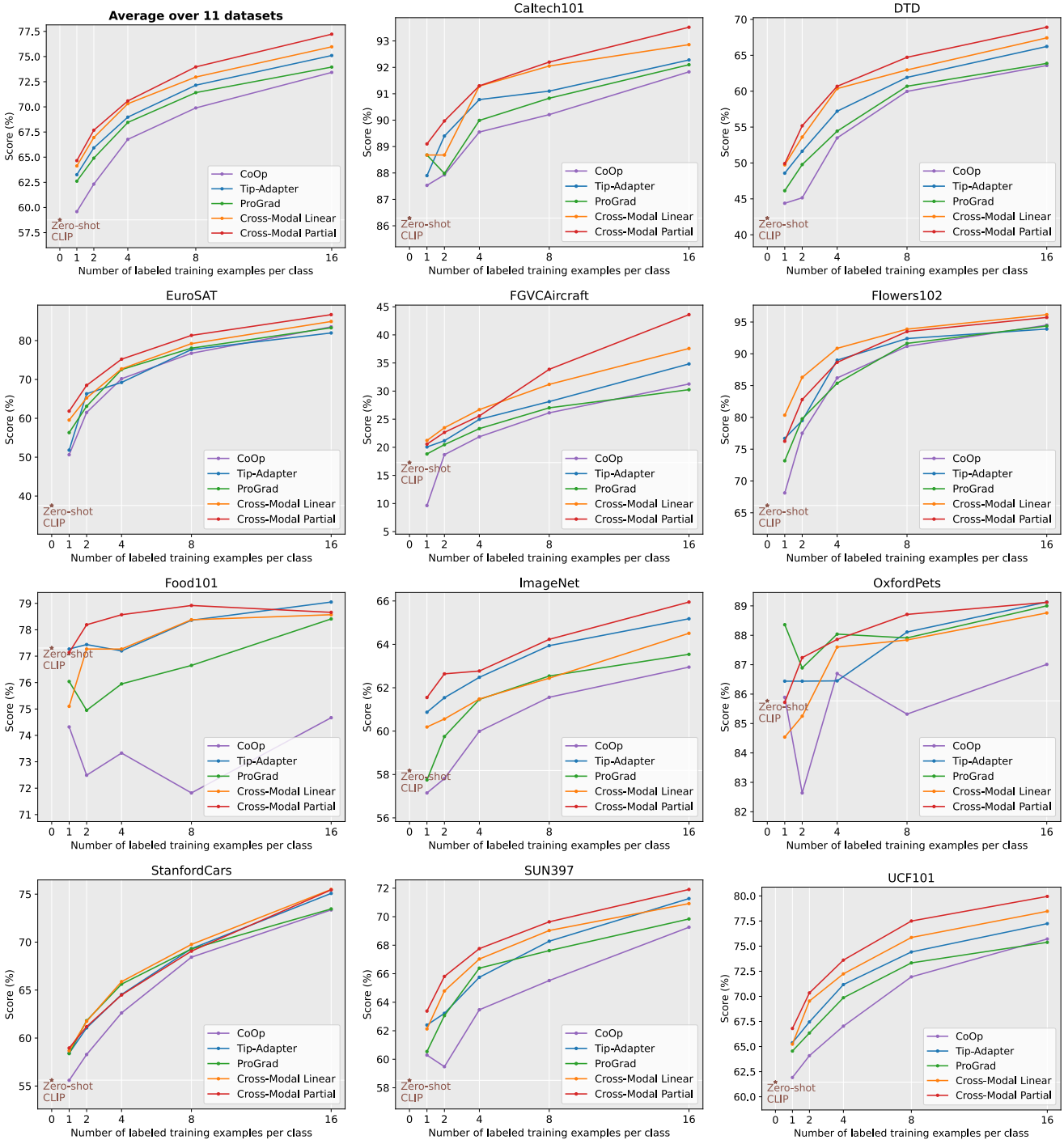

Figure 6. **Comparison of few-shot learning results across 11 datasets.** We show our main methods (cross-modal linear probing and partial finetuning) and compare them with prior works. We note that the Tip-Adapter [111] numbers shown are our own re-run of the method, where we replace their early-stopping on the test set with early stopping on the few-shot validation set for a fair comparison. As seen in the plots, cross-modal partial finetuning consistently outperforms prior works across the datasets, and cross-modal linear probing is also generally more performant.

Method	Shots	Dataset											
		Caltech [55]	ImageNet [15]	DTD [13]	EuroSAT [34]	Aircraft [65]	Food [7]	Flowers [72]	Pets [76]	Cars [51]	SUN397 [103]	UCF101 [93]	Average
Zero-Shot CLIP	0	86.29	58.18	42.32	37.56	17.28	77.31	66.14	85.77	55.61	58.52	61.46	58.77
CoOp	1	87.53	57.15	44.39	50.63	9.64	74.32	68.12	85.89	55.59	60.29	61.92	59.77
	2	87.93	57.81	45.15	61.50	18.68	72.49	77.51	82.64	58.28	59.48	64.09	62.32
	4	89.55	59.99	53.49	70.18	21.87	73.33	86.20	86.70	62.62	63.47	67.03	66.77
	8	90.21	61.56	59.97	76.73	26.13	71.82	91.18	85.32	68.43	65.52	71.94	69.89
	16	91.83	62.95	63.58	83.53	31.26	74.67	94.51	87.01	73.36	69.26	75.71	73.42
Tip-Adapter	1	87.90 ± 0.75	60.87 ± 0.04	48.58 ± 0.63	51.81 ± 2.45	20.06 ± 0.39	**77.27 ± 0.39**	76.70 ± 0.28	86.44 ± 1.35	58.42 ± 0.47	62.40 ± 0.27	65.38 ± 0.29	63.26 ± 0.68
	2	89.40 ± 0.22	61.54 ± 0.01	51.64 ± 0.58	66.32 ± 2.06	21.17 ± 0.62	77.44 ± 0.07	79.50 ± 1.07	86.44 ± 0.44	61.06 ± 0.41	63.22 ± 0.62	67.45 ± 1.77	65.93 ± 0.72
	4	90.78 ± 0.16	62.48 ± 0.01	57.21 ± 0.33	69.23 ± 2.85	24.97 ± 0.84	77.20 ± 0.43	89.00 ± 0.44	86.45 ± 0.71	64.54 ± 0.38	65.75 ± 0.15	71.17 ± 0.36	68.98 ± 0.61
	8	91.10 ± 0.18	63.94 ± 0.16	61.92 ± 0.83	77.69 ± 2.45	28.13 ± 1.06	78.36 ± 0.12	92.40 ± 0.24	88.11 ± 0.42	69.32 ± 0.08	68.28 ± 0.34	74.42 ± 0.72	72.15 ± 0.60
	16	92.28 ± 0.66	65.18 ± 0.15	66.23 ± 0.79	81.96 ± 2.26	34.83 ± 0.92	79.05 ± 0.26	93.90 ± 0.68	89.13 ± 0.28	75.08 ± 0.23	71.27 ± 0.13	77.24 ± 0.3	75.10 ± 0.61
ProGrad	1	88.68 ± 0.34	57.75 ± 0.24	46.14 ± 1.74	56.32 ± 3.04	18.81 ± 0.50	76.04 ± 0.54	73.18 ± 0.73	**88.36 ± 0.73**	58.38 ± 0.23	60.54 ± 0.24	64.55 ± 0.50	62.61 ± 0.80
	2	87.98 ± 0.69	59.75 ± 0.33	49.78 ± 1.37	63.10 ± 3.77	20.47 ± 0.90	74.95 ± 0.57	79.77 ± 0.65	86.89 ± 0.42	61.81 ± 0.45	63.06 ± 0.11	66.35 ± 0.18	64.90 ± 0.86
	4	89.99 ± 0.26	61.46 ± 0.07	54.43 ± 0.86	72.53 ± 1.29	23.32 ± 0.36	75.95 ± 0.27	85.37 ± 0.96	**88.04 ± 0.50**	65.62 ± 0.43	66.39 ± 0.43	69.86 ± 0.30	68.45 ± 0.52
	8	90.83 ± 0.07	62.54 ± 0.03	60.69 ± 0.10	78.04 ± 2.45	27.02 ± 0.67	76.65 ± 0.23	91.64 ± 0.24	87.91 ± 0.54	69.29 ± 0.11	67.62 ± 0.28	73.33 ± 0.65	71.41 ± 0.49
	16	92.10 ± 0.39	63.54 ± 0.08	63.87 ± 0.99	83.29 ± 0.85	30.25 ± 1.09	78.41 ± 0.08	94.37 ± 0.24	89.00 ± 0.32	73.46 ± 0.29	69.84 ± 0.18	75.38 ± 0.10	73.96 ± 0.42
Wise-FT	1	85.49 ± 0.81	58.30 ± 0.24	44.17 ± 0.72	52.30 ± 2.00	18.61 ± 0.54	71.88 ± 0.02	65.83 ± 0.54	81.73 ± 1.15	55.66 ± 0.15	56.59 ± 0.10	59.39 ± 1.33	59.09 ± 0.69
	2	87.00 ± 0.68	59.08 ± 0.34	46.95 ± 0.27	57.07 ± 4.26	20.88 ± 0.36	73.54 ± 0.11	71.02 ± 0.94	82.75 ± 0.62	58.67 ± 0.15	60.15 ± 0.10	62.74 ± 0.67	61.80 ± 0.77
	4	89.03 ± 0.17	60.48 ± 0.11	52.23 ± 0.70	62.45 ± 4.09	23.33 ± 0.38	76.17 ± 0.33	77.10 ± 0.50	85.95 ± 0.52	62.09 ± 0.35	63.18 ± 0.22	66.14 ± 0.46	65.29 ± 0.71
	8	90.07 ± 0.34	61.85 ± 0.22	55.56 ± 0.50	71.40 ± 2.80	26.97 ± 0.28	76.72 ± 0.31	82.54 ± 0.34	86.52 ± 0.45	66.00 ± 0.47	65.25 ± 0.48	69.84 ± 0.33	68.43 ± 0.59
	16	90.79 ± 0.15	62.84 ± 0.11	61.74 ± 0.61	77.79 ± 0.52	31.75 ± 0.46	77.80 ± 0.04	86.91 ± 0.71	87.50 ± 0.30	71.28 ± 0.20	67.46 ± 0.17	72.20 ± 0.03	71.64 ± 0.30
Cross-Modal Linear Probe	1	88.68 ± 0.17	60.19 ± 0.14	49.74 ± 0.24	59.54 ± 5.28	**21.21 ± 1.37**	75.10 ± 1.81	80.35 ± 0.22	84.54 ± 1.92	58.68 ± 0.17	62.13 ± 0.30	65.24 ± 0.36	64.13 ± 1.09
	2	88.68 ± 2.04	60.56 ± 0.10	53.61 ± 2.36	65.23 ± 2.42	23.48 ± 0.56	77.27 ± 0.07	**86.30 ± 0.94**	85.25 ± 2.46	61.75 ± 0.29	64.79 ± 0.13	69.53 ± 0.74	66.95 ± 1.10
	4	91.29 ± 0.51	61.48 ± 0.15	60.36 ± 0.46	72.72 ± 2.00	26.70 ± 0.48	77.27 ± 0.66	**90.86 ± 0.15**	87.60 ± 0.22	65.88 ± 0.06	67.03 ± 0.43	72.24 ± 0.35	70.31 ± 0.50
	8	92.05 ± 0.09	62.44 ± 0.08	62.96 ± 0.12	79.21 ± 2.13	31.19 ± 1.45	78.38 ± 0.19	**93.88 ± 0.50**	87.84 ± 0.65	69.76 ± 0.63	69.03 ± 0.16	75.86 ± 0.37	72.96 ± 0.58
	16	92.86 ± 0.20	64.51 ± 0.05	67.43 ± 1.51	84.91 ± 0.27	37.58 ± 0.82	78.57 ± 0.54	**96.16 ± 0.19**	88.76 ± 0.32	75.49 ± 0.36	70.92 ± 0.03	78.47 ± 0.12	75.97 ± 0.40
Cross-Modal Wise-FT	1	88.61 ± 0.15	60.90 ± 0.22	48.17 ± 0.17	55.09 ± 7.22	20.62 ± 0.44	77.05 ± 0.19	77.18 ± 1.70	86.54 ± 0.56	**59.10 ± 0.40**	62.47 ± 0.32	65.65 ± 0.55	63.76 ± 1.08
	2	88.56 ± 1.95	61.77 ± 0.16	51.83 ± 0.66	64.33 ± 3.76	21.88 ± 0.30	77.62 ± 0.21	81.84 ± 0.19	87.01 ± 0.12	**62.24 ± 0.33**	64.19 ± 0.63	69.11 ± 0.92	66.40 ± 0.84
	4	89.94 ± 0.23	62.45 ± 0.13	56.23 ± 0.98	72.22 ± 2.18	24.11 ± 0.14	78.25 ± 0.09	85.46 ± 0.99	87.99 ± 0.22	65.31 ± 0.87	65.61 ± 0.57	70.88 ± 0.20	68.95 ± 0.60
	8	91.36 ± 0.27	63.44 ± 0.14	60.15 ± 2.36	76.92 ± 3.75	28.59 ± 2.21	78.60 ± 0.17	90.72 ± 0.97	88.53 ± 0.22	68.57 ± 1.41	67.42 ± 0.61	74.83 ± 1.18	71.74 ± 1.21
	16	92.48 ± 0.32	65.15 ± 0.05	63.87 ± 2.27	79.96 ± 1.76	33.86 ± 2.14	78.94 ± 0.38	91.65 ± 0.26	**89.38 ± 0.21**	73.64 ± 0.66	68.92 ± 0.57	77.12 ± 0.56	74.09 ± 0.83
Cross-Modal Adapter	1	89.03 ± 0.36	61.23 ± 0.12	47.24 ± 0.91	60.50 ± 4.04	21.04 ± 1.30	75.90 ± 1.66	**80.63 ± 0.28**	85.62 ± 0.71	59.00 ± 0.20	62.86 ± 0.24	65.30 ± 0.38	64.40 ± 0.93
	2	89.36 ± 1.20	61.85 ± 0.01	54.51 ± 1.55	66.08 ± 1.67	**23.58 ± 0.62**	77.53 ± 0.20	85.69 ± 0.22	86.89 ± 0.23	62.22 ± 0.53	65.46 ± 0.26	70.12 ± 0.68	67.57 ± 0.65
	4	**91.33 ± 0.23**	**62.98 ± 0.10**	60.03 ± 0.53	73.46 ± 2.67	**27.55 ± 0.47**	77.92 ± 0.63	90.81 ± 0.28	87.76 ± 0.12	**66.40 ± 0.87**	67.63 ± 0.37	72.67 ± 0.04	**70.78 ± 0.57**
	8	92.08 ± 0.02	63.71 ± 0.06	64.11 ± 0.91	78.83 ± 2.66	32.75 ± 0.14	78.83 ± 0.14	93.57 ± 0.19	87.79 ± 0.11	**70.29 ± 0.45**	68.61 ± 0.52	76.34 ± 0.49	73.35 ± 0.52
	16	92.98 ± 0.14	64.72 ± 0.19	67.51 ± 1.32	82.15 ± 1.92	38.80 ± 1.06	**79.14 ± 0.44**	95.57 ± 0.11	88.64 ± 0.16	**75.96 ± 0.62**	70.91 ± 0.33	78.91 ± 0.14	75.94 ± 0.58
Cross-Modal Partial Finetuning	1	**89.10 ± 0.36**	**61.55 ± 0.45**	**49.92 ± 0.76**	**61.84 ± 5.16**	20.56 ± 0.21	77.14 ± 0.70	76.25 ± 0.42	85.72 ± 0.72	58.96 ± 0.15	**63.38 ± 0.27**	**66.80 ± 0.18**	**64.66 ± 0.85**
	2	**89.97 ± 0.28**	**62.64 ± 0.12**	**55.18 ± 1.77**	**68.48 ± 1.75**	22.65 ± 0.72	**78.19 ± 0.18**	82.80 ± 0.34	**87.24 ± 0.99**	61.19 ± 0.36	**65.81 ± 0.34**	**70.34 ± 0.06**	**67.68 ± 0.63**
	4	91.30 ± 0.75	62.77 ± 0.47	**60.68 ± 0.36**	**75.21 ± 2.10**	25.58 ± 0.61	**78.57 ± 0.15**	88.66 ± 0.28	87.86 ± 0.73	64.49 ± 0.08	**67.76 ± 0.51**	**73.61 ± 0.09**	70.59 ± 0.56
	8	**92.20 ± 0.19**	**64.23 ± 0.11**	**64.72 ± 0.54**	**81.33 ± 1.61**	**33.87 ± 0.70**	**78.92 ± 0.21**	93.50 ± 0.24	**88.71 ± 0.34**	69.06 ± 0.40	**69.64 ± 0.08**	**77.50 ± 1.04**	**73.97 ± 0.50**
	16	**93.52 ± 0.20**	**65.95 ± 0.04**	**68.91 ± 0.49**	**86.67 ± 0.72**	**43.60 ± 0.31**	78.66 ± 0.85	95.72 ± 0.22	89.12 ± 0.32	75.45 ± 0.49	**71.91 ± 0.05**	**79.95 ± 0.46**	**77.22 ± 0.38**

Table 10. **Per-dataset results on the ResNet-50 backbone.** We also include results from prior works for easier comparison. The zero-shot CLIP numbers differ from those reported in the original CLIP paper because we use one single prompt per dataset. We **bold** the best result for each shot and each dataset, and underline the second best result. We see that cross-modal adaptation methods consistently produce the best performance across almost all dataset. The Tip-Adapter results are reproduced using only the few-shot validation set for hyperparameter searching and early stopping.

Finetuning	ImageAug	TextAug	Number of shots				
			1	2	4	8	16
Linear	CenterCrop (1 view)	N/A (Uni-Modal Adaptation)	36.58$_{(1.47)}$	48.85$_{(1.43)}$	58.87$_{(0.82)}$	66.46$_{(0.74)}$	71.63$_{(0.50)}$
	+Flip (2 views)		37.51$_{(1.46)}$	**49.43**$_{(1.59)}$	**59.37**$_{(0.74)}$	66.65$_{(0.64)}$	71.83$_{(0.54)}$
	+RandomCrop (2 views)		37.74$_{(1.47)}$	49.21$_{(1.46)}$	59.23$_{(0.82)}$	**66.70**$_{(0.60)}$	**71.94**$_{(0.54)}$
	+RandomCrop (10 views)		**37.76**$_{(1.20)}$	49.25$_{(1.14)}$	59.13$_{(0.92)}$	66.52$_{(0.59)}$	71.89$_{(0.49)}$
	CenterCrop (1 view)	Class name	61.78$_{(1.17)}$	65.34$_{(0.79)}$	68.98$_{(0.67)}$	72.01$_{(0.57)}$	74.91$_{(0.59)}$
		`a photo of a {cls}.`	63.22$_{(1.37)}$	66.18$_{(0.74)}$	69.73$_{(0.53)}$	72.51$_{(0.71)}$	75.29$_{(0.62)}$
		Hand Engineered	**63.66**$_{(1.25)}$	66.67$_{(0.91)}$	**70.33**$_{(0.53)}$	72.92$_{(0.61)}$	75.54$_{(0.53)}$
		Template Mining (21 views)	63.50$_{(1.33)}$	**67.21**$_{(0.80)}$	70.26$_{(0.65)}$	**73.07**$_{(0.63)}$	**75.73**$_{(0.54)}$
	+Flip (2 views)	Class name	61.84$_{(0.79)}$	65.32$_{(1.15)}$	69.25$_{(0.52)}$	72.32$_{(0.56)}$	75.27$_{(0.49)}$
		`a photo of a {cls}.`	63.36$_{(0.84)}$	66.42$_{(1.20)}$	69.88$_{(0.62)}$	72.73$_{(0.71)}$	75.53$_{(0.49)}$
		Hand Engineered	**64.13**$_{(1.09)}$	66.95$_{(1.10)}$	70.31$_{(0.50)}$	72.96$_{(0.58)}$	**75.97**$_{(0.40)}$
		Template Mining (21 views)	63.88$_{(1.21)}$	**67.19**$_{(0.97)}$	**70.32**$_{(0.70)}$	**73.10**$_{(0.57)}$	75.70$_{(0.59)}$
	+RandomCrop (2 views)	Class name	61.47$_{(1.27)}$	65.09$_{(1.20)}$	68.94$_{(0.64)}$	72.06$_{(0.76)}$	75.12$_{(0.59)}$
		`a photo of a {cls}.`	63.32$_{(1.14)}$	66.05$_{(0.92)}$	69.93$_{(0.63)}$	72.91$_{(0.53)}$	75.67$_{(0.50)}$
		Hand Engineered	**63.71**$_{(1.50)}$	66.75$_{(0.83)}$	70.19$_{(0.51)}$	72.84$_{(0.60)}$	75.83$_{(0.59)}$
		Template Mining (21 views)	63.68$_{(1.75)}$	**67.14**$_{(0.80)}$	**70.53**$_{(0.53)}$	**72.98**$_{(0.67)}$	**75.75**$_{(0.49)}$
	+RandomCrop (10 views)	Class name	61.52$_{(1.18)}$	65.37$_{(0.82)}$	68.85$_{(0.77)}$	72.12$_{(0.72)}$	75.02$_{(0.63)}$
		`a photo of a {cls}.`	63.35$_{(1.04)}$	66.45$_{(0.73)}$	69.52$_{(0.78)}$	72.69$_{(0.55)}$	75.44$_{(0.72)}$
		Hand Engineered	63.85$_{(1.35)}$	66.87$_{(0.82)}$	**70.19**$_{(0.50)}$	72.98$_{(0.59)}$	75.62$_{(0.51)}$
		Template Mining (21 views)	**63.90**$_{(1.35)}$	**67.00**$_{(0.86)}$	69.94$_{(1.02)}$	**73.04**$_{(0.69)}$	**75.75**$_{(0.54)}$
Partial	CenterCrop (1 view)	N/A (Uni-Modal Adaptation)	29.93$_{(2.37)}$	42.63$_{(0.83)}$	54.27$_{(1.06)}$	64.16$_{(0.81)}$	71.62$_{(0.56)}$
	+Flip (2 views)		**31.68**$_{(1.19)}$	43.61$_{(1.08)}$	55.15$_{(0.77)}$	64.90$_{(0.87)}$	**72.19**$_{(0.44)}$
	+RandomCrop (2 views)		31.01$_{(1.39)}$	**43.78**$_{(1.09)}$	55.16$_{(0.79)}$	**64.91**$_{(0.93)}$	72.03$_{(0.44)}$
	+RandomCrop (10 views)		31.46$_{(1.41)}$	43.76$_{(1.07)}$	**55.23**$_{(0.79)}$	64.74$_{(0.78)}$	72.15$_{(0.41)}$
	CenterCrop (1 view)	Class name	62.50$_{(1.34)}$	65.66$_{(0.84)}$	69.33$_{(0.86)}$	72.93$_{(0.47)}$	76.21$_{(0.41)}$
		`a photo of a {cls}.`	63.78$_{(1.07)}$	66.79$_{(0.68)}$	69.80$_{(0.75)}$	73.40$_{(0.43)}$	76.67$_{(0.35)}$
		Hand Engineered	64.27$_{(0.96)}$	67.14$_{(0.58)}$	**70.26**$_{(0.55)}$	73.53$_{(0.51)}$	76.53$_{(0.48)}$
		Template Mining (21 views)	**64.57**$_{(0.81)}$	**67.21**$_{(0.67)}$	70.24$_{(0.89)}$	**73.71**$_{(0.58)}$	**76.86**$_{(0.32)}$
	+Flip (2 views)	Class name	62.52$_{(1.27)}$	66.02$_{(0.86)}$	69.64$_{(0.65)}$	73.30$_{(0.59)}$	76.44$_{(0.45)}$
		`a photo of a {cls}.`	64.13$_{(0.97)}$	67.16$_{(0.64)}$	69.97$_{(1.22)}$	73.83$_{(0.44)}$	77.03$_{(0.39)}$
		Hand Engineered	**64.66**$_{(0.85)}$	**67.68**$_{(0.63)}$	**70.59**$_{(0.56)}$	73.79$_{(0.50)}$	**77.22**$_{(0.38)}$
		Template Mining (21 views)	64.59$_{(1.02)}$	67.58$_{(0.74)}$	70.58$_{(0.82)}$	**74.00**$_{(0.49)}$	77.16$_{(0.33)}$
	+RandomCrop (2 views)	Class name	62.31$_{(1.78)}$	65.77$_{(0.77)}$	69.52$_{(0.70)}$	73.21$_{(0.49)}$	76.52$_{(0.39)}$
		`a photo of a {cls}.`	63.72$_{(1.09)}$	66.99$_{(0.52)}$	69.89$_{(1.14)}$	73.63$_{(0.55)}$	76.94$_{(0.37)}$
		Hand Engineered	63.64$_{(1.54)}$	67.35$_{(0.69)}$	70.50$_{(0.69)}$	**73.96**$_{(0.48)}$	77.05$_{(0.47)}$
		Template Mining (21 views)	**64.41**$_{(1.18)}$	**67.36**$_{(0.75)}$	**70.77**$_{(0.61)}$	73.94$_{(0.53)}$	**77.19**$_{(0.35)}$
	+RandomCrop (10 views)	Class name	62.18$_{(1.47)}$	66.01$_{(0.64)}$	69.47$_{(0.78)}$	73.27$_{(0.46)}$	76.60$_{(0.45)}$
		`a photo of a {cls}.`	64.00$_{(1.12)}$	67.08$_{(0.64)}$	70.22$_{(0.64)}$	73.70$_{(0.51)}$	76.96$_{(0.41)}$
		Hand Engineered	64.12$_{(1.38)}$	**67.63**$_{(0.64)}$	70.58$_{(0.59)}$	73.93$_{(0.39)}$	77.13$_{(0.38)}$
		Template Mining (21 views)	**64.57**$_{(1.00)}$	67.37$_{(0.62)}$	**70.86**$_{(0.54)}$	**74.02**$_{(0.41)}$	**77.27**$_{(0.38)}$

Table 11. **Ablation for augmentation under vision-language adaptation.** Salient conclusions: (1) Uni-modal adaptation is much worse than cross-modal adaptation even when doing aggressive image augmentation to increase the number of views, e.g. 10 random crops. (2) Doing both image augmentation and text augmentation can improve the results, but text augmentation has a more profound impact whereas image augmentation saturates with a few views. (3) Simple template mining can be as competitive as manually selected templates (cf. Table 18). Overall, we hope this preliminary investigation can encourage future work to explore more text augmentation strategies.

Method	Initialization	Number of shots				
		1	2	4	8	16
Linear Probing	Random	$36.58_{(1.47)}$	$48.85_{(1.43)}$	$58.87_{(0.82)}$	$66.46_{(0.74)}$	$71.63_{(0.50)}$
	Text	$58.32_{(0.71)}$	$61.39_{(0.74)}$	$65.25_{(0.61)}$	$68.54_{(0.58)}$	$71.90_{(0.33)}$
Cross-Modal Linear Probing	Random	$48.37_{(1.58)}$	$54.87_{(1.33)}$	$61.98_{(0.84)}$	$67.96_{(0.58)}$	$72.32_{(0.50)}$
	Text	$63.66_{(1.25)}$	$66.67_{(0.91)}$	$70.33_{(0.53)}$	$72.92_{(0.61)}$	$75.54_{(0.53)}$
Partial Finetuning	Random	$29.93_{(2.37)}$	$42.63_{(0.83)}$	$54.27_{(1.06)}$	$64.16_{(0.81)}$	$71.62_{(0.56)}$
	Text	$60.79_{(1.53)}$	$63.44_{(0.64)}$	$66.51_{(0.60)}$	$69.46_{(0.68)}$	$72.67_{(0.54)}$
Cross-Modal Partial Finetuning	Random	$42.03_{(1.91)}$	$50.85_{(1.20)}$	$59.74_{(0.89)}$	$66.98_{(0.90)}$	$72.92_{(0.42)}$
	Text	$64.27_{(0.96)}$	$67.14_{(0.58)}$	$70.26_{(0.55)}$	$73.53_{(0.51)}$	$76.53_{(0.48)}$

Table 12. **Ablation results for text-based vs random initialization for linear classifier weight.** We perform diligent analysis to confirm that initializing the linear classifier weights with text features is beneficial for the final performance. Still, cross-modal adaptation uniformly boosts the performance no matter the method or initialization. The text-based initialization is also more important for partial-finetuning than for linear probing, confirming the hypothesis [53] that a randomly initialized classifier will distort pre-trained features. Experiments in this table use center crop as image augmentation and Tip-Adapter's template as text augmentation for simplicity.

Image Encoder	Text Encoder	Number of shots				
		1	2	4	8	16
Frozen	Frozen	$63.66_{(1.25)}$	$66.67_{(0.91)}$	$70.33_{(0.53)}$	$72.92_{(0.61)}$	$75.54_{(0.53)}$
Finetune Attention Pooling Layer	Frozen	$64.13_{(1.29)}$	$67.23_{(0.51)}$	$70.44_{(0.55)}$	$73.64_{(0.47)}$	$76.65_{(0.44)}$
Frozen	Finetune Last Transformer Layer	$64.12_{(1.10)}$	$67.41_{(0.79)}$	$70.31_{(0.52)}$	$72.12_{(0.38)}$	$73.34_{(0.32)}$
Finetune Attention Pooling Layer	Finetune Last Transformer Layer	$64.09_{(1.28)}$	$67.06_{(0.76)}$	$70.38_{(0.57)}$	$73.64_{(0.48)}$	$76.68_{(0.39)}$

Table 13. **Ablation results for partial-finetuning.** Partial finetuning of the last layer of image encoder is much more effective than finetuning the last layer of text encoder, suggesting that one may simply freeze the text encoder for few-shot vision-language adaptation. Experiments in this table use center crop as image augmentation and Tip-Adapter's template as text augmentation for simplicity.

Backbone	Method	Number of shots				
		1	2	4	8	16
ResNet50	WiSE-FT	$59.09_{(0.69)}$	$61.80_{(0.77)}$	$65.29_{(0.71)}$	$68.43_{(0.59)}$	$71.64_{(0.30)}$
	Cross-Modal WiSE-FT	$63.76_{(1.08)}$	$66.40_{(0.84)}$	$68.95_{(0.60)}$	$71.74_{(1.21)}$	$74.09_{(0.83)}$
	Cross-Modal Prompting	$61.97_{(0.46)}$	$64.91_{(0.48)}$	$68.43_{(0.50)}$	$71.39_{(0.59)}$	$73.99_{0.54)}$
	Cross-Modal Adapter	$63.84_{(1.28)}$	$67.11_{(0.96)}$	$70.71_{(0.49)}$	$73.32_{(0.67)}$	$75.89_{(0.54)}$
	Linear Probing	$36.58_{(1.47)}$	$48.85_{(1.43)}$	$58.87_{(0.82)}$	$66.46_{(0.74)}$	$71.63_{(0.50)}$
	Cross-Modal Linear Probing	$63.66_{(1.25)}$	$66.67_{(0.91)}$	$70.33_{(0.53)}$	$72.92_{(0.61)}$	$75.54_{(0.53)}$
	Partial Finetuning	$29.93_{(2.37)}$	$42.63_{(0.83)}$	$54.27_{(1.06)}$	$64.16_{(0.81)}$	$71.62_{(0.56)}$
	Cross-Modal Partial Finetuning	$64.27_{(0.96)}$	$67.14_{(0.58)}$	$70.26_{(0.55)}$	$73.53_{(0.51)}$	$76.53_{(0.48)}$
ViT-B/16	WiSE-FT	$60.31_{(0.68)}$	$62.27_{(0.72)}$	$64.97_{(0.39)}$	$67.03_{(0.44)}$	$68.93_{(0.72)}$
	Cross-Modal WiSE-FT	$71.19_{(1.27)}$	$73.45_{(0.79)}$	$75.33_{(0.98)}$	$77.91_{(0.85)}$	$79.51_{(0.82)}$
	Linear Probing	$43.87_{(2.55)}$	$56.84_{(1.45)}$	$67.12_{(0.94)}$	$73.77_{(0.69)}$	$78.16_{(0.52)}$
	Cross-Modal Linear Probing	$71.21_{(1.13)}$	$73.70_{(1.03)}$	$76.78_{(0.48)}$	$78.89_{(0.37)}$	$81.07_{(0.30)}$
	Partial Finetuning	$35.44_{(3.49)}$	$52.04_{(1.52)}$	$65.50_{(0.99)}$	$74.05_{(0.94)}$	$79.58_{(0.53)}$
	Cross-Modal Partial Finetuning	$70.70_{(1.21)}$	$74.70_{(0.84)}$	$77.76_{(0.50)}$	$80.19_{(0.34)}$	$82.52_{(0.41)}$

Table 14. **Complete results for all methods reported.** Experiments in this table use center crop as image augmentation and Tip-Adapter's template as text augmentation. Furthermore, we include ViT-B/16 results for completeness.

Dataset	Method	Number of Image Shots			
		0	1	2	4
ImageNet-ESC-19	Image-Only Linear Probing	-	$68.00_{(4.17)}$	$75.67_{(4.62)}$	$83.05_{(2.52)}$
	Image-Audio Linear Probing	-	$69.33_{(3.97)}$	$76.66_{(4.32)}$	$83.22_{(3.77)}$
	Image-Text Linear Probing	-	$85.69_{(5.36)}$	$86.94_{(2.41)}$	$89.21_{(3.04)}$
	Image-Audio-Text Linear Probing	-	$82.34_{(2.66)}$	$84.08_{(1.95)}$	$87.33_{(1.68)}$
	Audio-initialized Classifier	$36.74_{(9.36)}$	-	-	-
	Text-initialized Classifier	$84.95_{(0.00)}$	-	-	-
ImageNet-ESC-27	Image-Only Linear Probing	-	$60.13_{(3.97)}$	$71.81_{(2.96)}$	$79.01_{(2.50)}$
	Image-Audio Linear Probing	-	$60.87_{(4.41)}$	$73.32_{(2.46)}$	$78.94_{(2.66)}$
	Image-Text Linear Probing	-	$84.15_{(3.10)}$	$85.17_{(2.48)}$	$88.35_{(0.80)}$
	Image-Audio-Text Linear Probing	-	$75.96_{(2.77)}$	$79.81_{(1.95)}$	$83.41_{(1.19)}$
	Audio-initialized Classifier	$30.37_{(7.13)}$	-	-	-
	Text-initialized Classifier	$82.96_{(0.00)}$	-	-	-

Table 15. **ImageNet-ESC image-classification results.**

Dataset	Method	Number of Audio Shots			
		0	1	2	4
ImageNet-ESC-19	Audio-Only Linear Probing	-	$31.21_{(5.45)}$	$41.11_{(5.12)}$	$48.51_{(3.79)}$
	Audio-Image Linear Probing	-	$35.74_{(4.85)}$	$45.94_{(4.99)}$	$51.59_{(3.40)}$
	Audio-Text Linear Probing	-	$38.74_{(5.51)}$	$50.09_{(3.45)}$	$53.90_{(1.96)}$
	Audio-Image-Text Linear Probing	-	$42.33_{(4.06)}$	$49.32_{(4.67)}$	$53.61_{(2.44)}$
	Image-initialized Classifier	$34.21_{(1.17)}$	-	-	-
	Text-initialized Classifier	$38.16_{(0.00)}$	-	-	-
ImageNet-ESC-27	Audio-Only Linear Probing	-	$28.20_{(3.26)}$	$39.00_{(3.42)}$	$47.13_{(2.71)}$
	Audio-Image Linear Probing	-	$35.01_{(4.06)}$	$43.51_{(3.47)}$	$48.46_{(3.37)}$
	Audio-Text Linear Probing	-	$36.76_{(5.54)}$	$45.69_{(4.04)}$	$50.56_{(2.19)}$
	Audio-Image-Text Linear Probing	-	$36.06_{(5.36)}$	$46.19_{(2.96)}$	$50.79_{(2.49)}$
	Image-initialized Classifier	$29.00_{(0.84)}$	-	-	-
	Text-initialized Classifier	$31.02_{(0.00)}$	-	-	-

Table 16. **ImageNet-ESC audio-classification results.**

Method	Template	Number of shots				
		1	2	4	8	16
ProDA [63]	36 Learned Templates	**65.19**	**68.59**	**71.49**	74.21	76.78
Linear	Class name	$62.34_{(0.88)}$	$65.75_{(1.31)}$	$69.95_{(0.53)}$	$73.29_{(0.72)}$	$76.66_{(0.30)}$
	`a photo of a {cls}.`	$63.87_{(0.88)}$	$66.59_{(1.40)}$	$70.71_{(0.61)}$	$73.75_{(0.62)}$	$76.85_{(0.38)}$
	HandEngineered [111]	$64.52_{(1.43)}$	$67.31_{(1.26)}$	$70.97_{(0.51)}$	$73.77_{(0.84)}$	$77.21_{(0.41)}$
	Template Mining (21 views)	$64.37_{(1.38)}$	$67.62_{(1.03)}$	$71.00_{(0.70)}$	$74.17_{(0.61)}$	$77.15_{(0.47)}$
Partial	Class name	$62.58_{(1.87)}$	$66.46_{(0.81)}$	$70.29_{(0.61)}$	$74.22_{(0.51)}$	$77.73_{(0.57)}$
	`a photo of a {cls}.`	$64.38_{(1.14)}$	$67.48_{(0.67)}$	$70.59_{(1.38)}$	$74.68_{(0.45)}$	$78.34_{(0.45)}$
	HandEngineered [111]	$65.01_{(1.17)}$	$68.05_{(0.64)}$	$71.10_{(0.67)}$	$74.83_{(0.50)}$	$\mathbf{78.60}_{(0.40)}$
	Template Mining (21 views)	$64.89_{(1.16)}$	$68.03_{(0.74)}$	$71.04_{(0.97)}$	$\mathbf{74.90}_{(0.43)}$	$78.37_{(0.40)}$

Table 17. **Comparison to ProDA.** Since ProDA uses their own separate test split without releasing the code, it is not directly comparable to numbers reported in Table 1. Therefore, we reported results here with our best attempt to replicate their dataset split by using the official test splits of the datasets when available (Food101 [6] and DTD [14]). Note that ProDA reported results using 36 learned prompts, whereas our template mining only uses 21 templates searched on few-shot validation set without any learning. Since we do not know whether ProDA uses augmentation, we report center crop results in this table. Still, our approach is generally more performant than ProDA and we do not require deep finetuning which takes 100x training time.

180 Templates (∗ indicates not in CoOp codebase)		
{cls}*	a tattoo of the {cls}.	a video of the person {cls}.
a photo of a {cls}.*	a photo of a person during {cls}.	a example of a person {cls}.
a picture of this {cls}.*	a photo of a clean {cls}.	a photo of a small {cls}.
a photo of my {cls}.*	a photo of a {cls} texture.	a photo of the small {cls}.
that is a {cls} photo.*	a bad photo of a {cls}.	the {cls} in a video game.
a picture of a {cls}.*	a video of the person during {cls}.	a demonstration of a person {cls}.
a {cls} photo.*	a drawing of the {cls}.	a photo of one {cls}.
this is a {cls} photo.*	a close-up photo of the {cls}.	a video of a person using {cls}.
a photo of these {cls}.*	a video of a person {cls}.	a blurry photo of a {cls}.
a picture of my {cls}.*	a good photo of a {cls}.	a photo of a person practicing {cls}.
a {cls} picture.*	a photo of a {cls} thing.	a photo of a {cls}, a type of flower.
that is a {cls} picture.*	a demonstration of the person practicing {cls}.	a painting of a {cls}.
a picture of those {cls}.*	itap of a {cls}.	a example of the person {cls}.
this is a {cls} picture.*	a photo of a {cls} pattern.	a example of the person performing {cls}.
that is a photo of a {cls}.*	itap of the {cls}.	a rendition of the {cls}.
a photo of your {cls}.*	a demonstration of a person using {cls}.	a cropped photo of a {cls}.
a picture of some {cls}.*	a cropped photo of the {cls}.	the origami {cls}.
a photo of those {cls}.*	a example of the person practicing {cls}.	a photo of the person {cls}.
a picture of these {cls}.*	a bright photo of a {cls}.	a example of the person doing {cls}.
{cls}, a picture.*	a photo of the hard to see {cls}.	a photo of the large {cls}.
a photo of an {cls}.*	a photo of a person using {cls}.	a example of a person doing {cls}.
a picture of the {cls}.*	a rendition of a {cls}.	a video of a person doing {cls}.
{cls}, a photo.*	a demonstration of a person during {cls}.	a sketch of the {cls}.
a photo of this {cls}.*	graffiti of the {cls}.	a photo of a nice {cls}.
a photo of the {cls}.*	a toy {cls}.	a good photo of the {cls}.
this is a photo of a {cls}.*	a jpeg corrupted photo of the {cls}.	a photo of a person performing {cls}.
a picture of your {cls}.*	a photo of the weird {cls}.	a pixelated photo of the {cls}.
a photo of a {cls}.*	a photo of a cool {cls}.	a photo of the dirty {cls}.
a picture of that {cls}.*	a video of the person practicing {cls}.	a photo of my new {cls}.
a photo of some {cls}.*	the plushie {cls}.	a sculpture of the {cls}.
a photo of my {cls}.*	a low resolution photo of a {cls}.	a photo of the person doing {cls}.
a photo of the {cls}.*	a photo of the person performing {cls}.	a photo of a {cls}, a type of pet.
a photo of that {cls}.*	the cartoon {cls}.	a centered satellite photo of the {cls}.
a picture of an {cls}.*	a video of a person practicing {cls}.	a photo of the {cls} texture.
a photo of the {cls}, a type of aircraft.	a photo of a {cls}, a type of aircraft.	a photo of a hard to see {cls}.
a bad photo of the {cls}.	a photo of the person using {cls}.	a black and white photo of a {cls}.
a photo of my dirty {cls}.	a centered satellite photo of {cls}.	itap of my {cls}.
a example of a person during {cls}.	a example of a person performing {cls}.	a video of the person doing {cls}.
a demonstration of the person doing {cls}.	a {cls} in a video game.	a demonstration of the person performing {cls}.
a demonstration of a person performing {cls}.	i love my {cls}!	art of a {cls}.
a photo of the person practicing {cls}.	a example of a person using {cls}.	a black and white photo of the {cls}.
a photo of a large {cls}.	a example of the person using {cls}.	a photo of the clean {cls}.
a photo of a weird {cls}.	a jpeg corrupted photo of a {cls}.	a photo of the nice {cls}.
a photo of a person {cls}.	a blurry photo of the {cls}.	a doodle of the {cls}.
a video of a person during {cls}.	a painting of the {cls}.	a close-up photo of a {cls}.
a photo of the {cls} thing.	a sculpture of a {cls}.	a low resolution photo of the {cls}.
the embroidered {cls}.	a demonstration of the person using {cls}.	a dark photo of a {cls}.
a photo of a {cls} object.	a sketch of a {cls}.	a video of the person performing {cls}.
a dark photo of the {cls}.	a drawing of a {cls}.	a photo of a dirty {cls}.
a photo of {cls}, a type of food.	a photo of the {cls} pattern.	a cartoon {cls}.
a example of the person during {cls}.	a photo of the cool {cls}.	the plastic {cls}.
a video of a person performing {cls}.	a photo of the {cls} object.	a photo of my clean {cls}.
a photo of many {cls}.	a video of the person using {cls}.	a photo of my old {cls}.
a photo of a person doing {cls}.	a demonstration of the person during {cls}.	a pixelated photo of a {cls}.
a plushie {cls}.	a centered satellite photo of a {cls}.	a demonstration of the person {cls}.
art of the {cls}.	a tattoo of a {cls}.	a doodle of a {cls}.
a photo of the person during {cls}.	graffiti of a {cls}.	the toy {cls}.
a bright photo of the {cls}.	a demonstration of a person practicing {cls}.	a plastic {cls}.
a rendering of a {cls}.	a embroidered {cls}.	a rendering of the {cls}.
a origami {cls}.	a example of a person practicing {cls}.	a demonstration of a person doing {cls}.

Table 18. **Templates used during template mining.** Most of the templates we use come from the original CoOp codebase [113]. In addition, we add 31 random templates by paraphrasing [45] the standard template `a photo of a {cls}`. We encourage future work to try out more sophisticated techniques to generate templates, e.g. through automated prompting [113] or with the help of language models [45].

References

[1] Mohamed Afham, Salman Khan, Muhammad Haris Khan, Muzammal Naseer, and Fahad Shahbaz Khan. Rich semantics improve few-shot learning. *arXiv preprint arXiv:2104.12709*, 2021. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 3

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020. 3

[4] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14493–14502, 2020. 3

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 7, 15, 21

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 17

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 2

[9] Gemma Calvert, Edward Bullmore, M.J. Brammer, Ruth Campbell, Steven Williams, Philip Mcguire, Peter Woodruff, S.D. Iversen, and Anthony David. Activation of auditory cortex during silent lipreading. science, 276(5312), 593-596. *Science (New York, N.Y.)*, 276:593–6, 05 1997. 3

[10] Cătălina Cangea, Petar Veličković, and Pietro Lio. Xflow: Cross-modal deep neural networks for audiovisual classification. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3711–3720, 2019. 3

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3

[13] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 17

[14] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 7, 15, 21

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5, 6, 7, 14, 17

[16] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*, 2022. 3

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[18] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019. 1, 3

[19] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 7

[20] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 1, 3

[21] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 2, 3, 4, 6, 14

[22] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 3

[23] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 7

[24] Eleanor J Gibson. Principles of perceptual learning and development. 1969. 1

[25] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. *Advances in neural information processing systems*, 30, 2017. 3

[26] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. 2

[27] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. 1, 2, 3, 7

[28] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresne(x)t-fbsp: Learning robust time-frequency transformation of audio, 2021. 7

[29] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE international conference on computer vision*, pages 3018–3027, 2017. 1, 3

[30] Adi Haviv, Jonathan Berant, and Amir Globerson. Bertese: Learning to speak to bert. *arXiv preprint arXiv:2103.05327*, 2021. 3

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3

[32] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[34] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2017. 17

[35] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 7

[36] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 8

[37] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 8

[38] Danfeng Hong, Naoto Yokoya, Gui-Song Xia, Jocelyn Chanussot, and Xiao Xiang Zhu. X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:12–23, 2020. 3

[39] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2, 3, 6

[40] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 3

[41] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *arXiv preprint arXiv:2208.05592*, 2022. 2

[42] Ray Jackendoff. On beyond zebra: The relation of linguistic and visual information. *Cognition*, 26(2):89–114, 1987. 1

[43] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3, 4

[44] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 3

[45] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. 3, 22

[46] Thorsten Joachims et al. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pages 200–209, 1999. 1, 3

[47] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. 3

[48] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[49] Stephen M. Kosslyn, Giorgio Ganis, and William L. Thompson. 3Multimodal images in the brain. In *The neurophysiological foundations of mental and motor imagery*. Oxford University Press, 01 2010. 3

[50] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 7

[51] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization.

In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 17

[52] Patricia K Kuhl and Andrew N Meltzoff. The intermodal representation of speech in infants. *Infant behavior and development*, 7(3):361–381, 1984. 1

[53] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 19

[54] Jet-Tsyn Lee, Danushka Bollegala, and Shan Luo. "touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4276–4282. IEEE, 2019. 3

[55] Li, Andreeto, Ranzato, and Perona. Caltech 101, Apr 2022. 17

[56] Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017. 3

[57] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 3

[58] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 3

[59] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 3

[60] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 3

[61] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 2, 3, 8

[62] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv:2103.10385*, 2021. 2

[63] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 2, 3, 5, 8, 15, 21

[64] Shan Luo, Wenzhen Yuan, Edward Adelson, Anthony G Cohn, and Raul Fuentes. Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2722–2727. IEEE, 2018. 3

[65] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 17

[66] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 7

[67] Andrew N Meltzoff and Richard W Borton. Intermodal matching by human neonates. *Nature*, 282(5737):403–404, 1979. 1

[68] Jesse Mu, Percy Liang, and Noah Goodman. Shaping visual representations with language for few-shot classification. *arXiv preprint arXiv:1911.02683*, 2019. 2, 3, 4, 5

[69] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022. 3

[70] Bence Nanay. Multimodal mental imagery. *Cortex*, 105:125–136, 2018. 1, 3

[71] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 7

[72] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 17

[73] Frederik Pahde, Main Nabi, Tassila Klein, and Patrick Jahnichen. Discriminative hallucination for multi-modal few-shot learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 156–160. IEEE, 2018. 3

[74] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2644–2653, 2021. 3

[75] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 7

[76] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 17

[77] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 2, 6, 14

[78] Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*, 2022. 3

[79] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018. 1, 3

[80] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005. 1

[81] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 1, 2, 3, 4, 5, 7, 8, 14

[82] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 1, 3

[83] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 8

[84] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few-shot text classification and natural language inference. *Computing Research Repository*, arXiv:2001.07676, 2020. 3

[85] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *Computing Research Repository*, arXiv:2009.07118, 2020. 3

[86] Lauren A Schmidt. *Meaning and compositionality as statistical induction of categories and constraints*. PhD thesis, Massachusetts Institute of Technology, 2009. 1

[87] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001. 4

[88] Eli Schwartz, Leonid Karlinsky, Rogerio Feris, Raja Giryes, and Alex Bronstein. Baby steps towards few-shot learning with multiple semantics. *Pattern Recognition Letters*, 160:142–147, 2022. 3

[89] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 3

[90] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005. 1

[91] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 1, 3, 4

[92] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022. 3

[93] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7, 17

[94] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2

[95] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 3

[96] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 8

[97] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14657, 2022. 3

[98] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7278–7286, 2018. 1

[99] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Growing a brain: Fine-tuning by increasing model capacity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2471–2480, 2017. 3

[100] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 2, 3, 4, 5, 6, 8, 14

[101] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Transferring textual knowledge for visual recognition. *arXiv preprint arXiv:2207.01297*, 2022. 3

[102] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 5

[103] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 7, 17

[104] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 4, 5

[105] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, and Yanning Zhang. Class-aware visual prompt tuning for vision-language pre-trained model. *arXiv preprint arXiv:2208.08340*, 2022. 3

[106] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3

[107] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer.

Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 3

[108] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 3

[109] Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 3

[110] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *European Conference on Computer Vision*, pages 698–714. Springer, 2020. 3

[111] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2, 3, 4, 5, 6, 7, 8, 14, 16, 21

[112] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 2, 3, 8

[113] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 2, 3, 4, 5, 6, 7, 8, 14, 15, 22

[114] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022. 2, 3, 6