Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models

Zhiqiu Lin* Samuel Yu* Zhiyi Kuang Deepak Pathak Deva Ramanan
Carnegie Mellon University
{zhiqiul,samuelyu,zkuang,dpathak,deva}@cs.cmu.edu

Abstract

The ability to quickly learn a new task with minimal instruction – known as few-shot learning – is a central aspect of intelligent agents. Classical few-shot benchmarks make use of few-shot samples from a single modality, but such samples may not be sufficient to characterize an entire concept class. In contrast, humans use cross-modal information to learn new concepts efficiently. In this work, we demonstrate that one can indeed build a better **visual** *dog classifier by* **read***ing about dogs and* **listen***ing to them bark. To do so, we exploit the fact that recent multimodal foundation models such as CLIP are inherently cross-modal, mapping different modalities to the same representation space. Specifically, we propose a simple* **cross-modal adaptation** *approach that learns from few-shot examples spanning different modalities. By repurposing class names as additional one-shot training samples, we achieve SOTA results with an embarrassingly simple linear classifier for vision-language adaptation. Furthermore, we show that our approach can benefit existing methods such as prefix tuning, adapters, and classifier ensembling. Finally, to explore other modalities beyond vision and language, we construct the first (to our knowledge) audiovisual few-shot benchmark and use cross-modal training to improve the performance of both image and audio classification. Project site at link.*

1. Introduction

Learning with minimal instruction is a hallmark of human intelligence [86, 91, 98], and is often studied under the guise of few-shot learning. In the context of few-shot visual classification [18, 20, 29, 46, 79, 82], a classifier is first pretrained on a set of base classes to learn a good feature representation and then adapted or finetuned on a small amount of novel class data. However, such few-shot setups often face an inherent ambiguity – if the training image contains a golden retriever wearing a hat, how does the learner know if the task is to find `dogs`, `golden retrievers`, or even `hats`? On the other hand, humans have little trouble understanding and even generalizing from as few as one example. How so?

*Equal contribution.

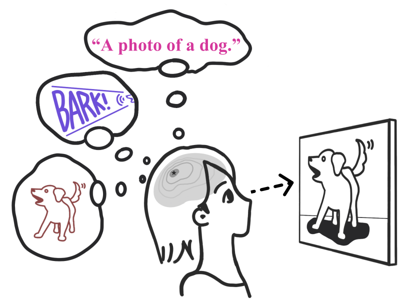

Figure 1. **Human perception is internally cross-modal.** When we perceive from one modality (such as vision), the same neurons will be triggered in our cerebral cortex as if we are perceiving the object from other modalities (such as language and audio) [24, 67, 70]. This phenomenon grants us a strong ability to learn from a few examples with cross-modal information [52, 67]. In this work, we propose to leverage cross-modality to adapt multimodal models (such as CLIP [81] and AudioCLIP [27]), that encode different modalities to the same representation space.

We argue that humans make use of multimodal signals and representations (Figure 1) when learning concepts. For example, verbal language has been shown to help toddlers better recognize visual objects given just a few examples [42, 90]. Indeed, there exists ample evidence from neuroscience suggesting that cognitive representations are inherently multimodal. For instance, visual images of a person evoke the same neurons as the textual strings of the person's name [80] and even audio clips of that person talking [70]. Even for infants as young as 1-5 months old, there is a strong correspondence between auditory-visual [52] as well as visual-tactile signals [67]. Such *cross-modal* or inter-modal representations are fundamental to the human perceptual-cognitive system, allowing us to understand new concepts even with few examples [24].

Cross-modal adaptation (our approach). In this paper, we demonstrate that cross-modal understanding of different modalities (such as image-text or image-audio) can improve the performance of individual modalities. That is, *read*ing about dogs and *listen*ing to them bark can help build a better *visual* classifier for them! To do so, we present a remarkably simple strategy for cross-modal few-shot adaptation: *we treat examples from different modalities as additional few-shot examples*. For example, given the "1-shot" task of learning a `dog` classifier, we treat *both* the textual `dog` label and the single visual image as training examples for learning a (visual) dog classifier. Learning is straightforward when using frozen textual and visual encoders, such as CLIP [81], that map different modalities to the same representational space. In essence, we have converted the "n-shot" problem to a "(n+1)-shot" problem (Figure 2)! We demonstrate that this basic strategy produces SOTA results across the board with a simple linear classifier, and can be applied to existing finetuning methods [100,111,113] or additional modalities (e.g. audio).

Why does it work? From one perspective, it may not be surprising that cross-modal adaptation improves accuracy, since it takes advantage of additional training examples that are "hidden" in the problem definition, e.g. a label name [104] or an annotation policy [68] for each class. However, our experiments demonstrate that multimodal cues are often complementary since they capture different aspects of the underlying concept; a `dog` label paired with a single visual example is often more performant than two images! For example, Figure 3 demonstrates a one-shot example where the target concept is ambiguous, but becomes clear once we add information from other modalities like language and sound.

Multimodal adaptation (prior art). In contrast to our cross-modal approach, most prior works simply follow the popular practice of finetuning uni-modal foundation models, such as large vision [12, 31, 32] or language models [8, 17, 62]. For example, CoOp [113] and other prompting methods [63,112,114] finetune CLIP via prefix tuning to replace hand-engineered prompts such as `"a photo of a {cls}"` with learned word tokens. Similarly, inspired by parameter-efficient tuning of language models [39], adapter-based methods [21,111] finetune CLIP by inserting lightweight multi-layer-perceptrons (MLPs). However, we aim to study the fundamental question of how to finetune *multi*-modal (as opposed to *uni*-modal) models. A crucial difference between prior art and ours is the use of textual information, as all existing methods [41,100,111,113] repurpose additional text features as *classifier weights* instead of *training samples*. We demonstrate in this paper that cross-modal adaptation is not only more performant but can also benefit prior uni-modal approaches.

Problem setup. We begin by replicating the existing evaluation protocol of other works [81, 111, 113] on few-shot adaptation of vision-language models, and report performance on 11 diverse downstream datasets. We produce state-of-the-art accuracy with an embarrassingly simple linear classifier that has access to additional "hidden" training examples in the form of textual labels, resulting in a system that is far more lightweight than prior art. Interestingly, we show that existing approaches [100,111,113], despite already repurposing text features as classifier weights, can still benefit from cross-modal learning. Finally, we extend our work to the audio domain by taking advantage of AudioCLIP [27] that maps audio to the same frozen CLIP representation space. We construct the first (to our knowledge) *cross-modal few-shot learning benchmark with audio* by intersecting ImageNet [15] and the ESC-50 audio classification dataset [77]. We show that cross-modal audiovisual learning helps for both downstream image and audio classification; in summary, one *can* train better dog image classifiers by listening to them bark!

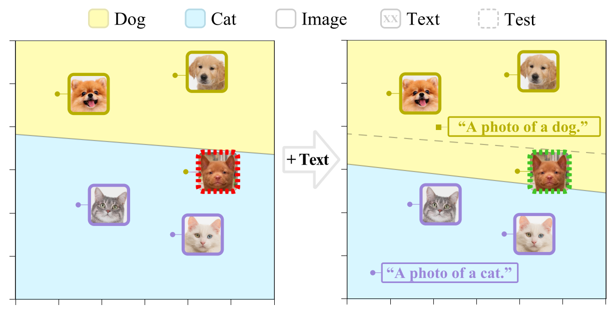

Figure 2. **Adding additional modalities helps few-shot learning**. Adding textual labels to a 2-shot cat-vs-dog classification task leads to better test performance (by turning the problem into a 3-shot cross-modal task!). We visualize cross-modal CLIP [21] features (projection to 2D with principal component analysis) and the resulting classifier learned from them, and observe a large shift in the decision boundary. See Figure 5 for more examples.

2. Related Works

Webly-supervised pre-training. Learning *foundation models* [5] from large-scale web data is becoming a predominant paradigm in AI. In NLP, models such as BERT [17] and GPT-3 [8] are pre-trained on a massive web text corpus with language-modeling objectives and can be transferred to a wide range of downstream tasks, even without explicit supervised finetuning [61, 94]. Self-supervision [11, 12, 32] is also a trending topic in the vision community, and recent methods [26, 31] demonstrate even stronger visual representations than fully-supervised pre-trained ones such as on ImageNet [15].

Multimodal foundation models. Recently, foundation models have shifted towards a multimodal supervi-

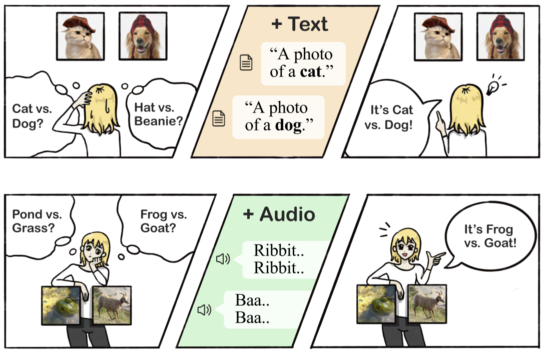

Figure 3. **Cross-modality reduces the ambiguity of few-shot learning.** Classic (uni-modal) few-shot learning is often *under*specified. Even for binary classification, when given only a single image per class (**left**), it is unclear whether the target class is the animal, the hat, or the background scene. Adding an extra modality, such as text or audio, helps clarify the problem setup (**right**). Notably, language usually comes “for free” in classification datasets in the form of a textual label per class.

sion paradigm. For visual representation learning, early works transform web image captions into structured outputs for supervised learning, such as multi-label targets [47] or visual n-grams [56]. More recently, CLIP [81] and ALIGN [43] propose a simple contrastive-based approach to embed images and captions into the same representation space, and demonstrate impressive “zero-shot” performance on downstream tasks. Follow-up works enhance multimodal pre-training by incorporating generative-based objectives [2, 57, 106], consistency regularization [60, 69], stronger visual priors [107], phrase-grounding tasks [58, 109], and audiovisual information through videos [27]. In this work, we focus on adapting CLIP [81] and AudioCLIP [27] for few-shot classification because contrastive-based multimodal models are stronger classifiers [2]. Adopting other multimodal models [2, 106] or adapting to tasks other than classification [92, 109] can be interesting future directions.

Adaptation of foundation models. As multimodal pre-trained models have excelled at classic vision tasks [81, 109], there has been surging interest in developing more efficient adaptation methods. However, we observe that most of the trending techniques are built upon successful recipes crafted for uni-modal foundation models. For example, CLIP [81] adopts linear probing [12,31,32,109] and full-finetuning [25, 31, 48, 99, 101, 109] when transferring to downstream tasks. Prompt adaptation of CLIP [63, 81, 105, 112, 114] is motivated by the success of prefix-tuning for language models [16,22,30,45,61,78,84,85,89]. Similarly, CLIP-Adapter [21] and Tip-Adapter [111] are inspired by parameter-efficient finetuning methods [39,44,110] that optimize lightweight MLPs while freezing the encoder. Yet, all aforementioned methods including WiSE-FT [100] use the other modality, e.g. textual labels, as *classifier weights* and still calculate a *uni-modal* softmax loss on the few-shot images. We instead show that incorporating other modalities as *training samples* is far more effective.

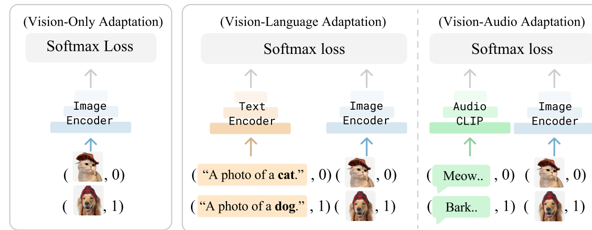

Figure 4. **Uni-modal (left) vs. cross-modal adaptation (right).** Prior work [21, 100, 111, 113] performs uni-modal adaptation by calculating the loss over a single modality. Cross-modal adaptation makes use of additional training samples from other modalities, exploiting pre-trained encoders that map different modalities to the same representation space. We show that cross-modal learning can also improve prior art and even extends to audio modalities with AudioCLIP [27].

Few-shot classification. Prior successful few-shot learning methods leverage meta learning [20, 82], metric learning [4, 91, 95], transfer learning [29, 79], and transductive learning [18, 46]. These classic algorithms usually assume a large meta-training set for pre-training the network, and then evaluate on multiple episodes of few-shot train (support) and test (query) sets. In this work, we instead follow the new evaluation protocol implemented by recent works on few-shot adaptation with CLIP [81, 111, 113]: (1) the meta-training phase is replaced with pre-trained CLIP models, and (2) the test sets are the official test splits of each dataset (thus not few-shot). Notably, none of the prior works [111, 113] we compare to in this paper perform optimization with test set samples, and we follow this practice to ensure a fair comparison. We leave semi-supervised [97] or transductive finetuning [18, 40] techniques as future work.

Cross-modal machine learning. Inspired by cross-modal human cognition [9, 49, 70], cross-modal learning [68, 104] is a subfield of multimodal machine learning [1,3,10,38,54,59,64,73,74,88,108] that aims to use data from additional modalities to improve a uni-modal task. Cross-modal learning does not require instance-wise alignment; for example, existing algorithms [68, 104] can benefit from class-level descriptions as opposed to image-level captions. In this work, we propose a lightweight cross-modal learning method by treating data from other modalities as additional training samples. Furthermore, we encourage future works to embrace cross-modal few-shot learning as opposed to the underspecified uni-modal setup (Figure 3).

3. Cross-Modal Adaptation

In this section, we mathematically formalize our approach to cross-modal few-shot learning.

Uni-modal learning. We begin by reviewing standard uni-modal few-shot classification, which learns a classifier from a small dataset of (x_i, y_i) pairs and pre-trained feature encoder $\phi(\cdot)$:

$$\mathcal{L}_{uni-modal} = \sum_i \mathcal{H}(y_i, \phi(x_i)) \tag{1}$$

where $\mathcal{H}$ is typically the softmax loss

$$\mathcal{H}(y, f) = -\log\Big(p(y|f)\Big) = -\log\Big(\frac{e^{w_y \cdot f}}{\sum_{y'} e^{w_{y'} \cdot f}}\Big). \tag{2}$$

Our notation separates the feature extractor ϕ from the final class weights w_y, since the former is typically pre-trained on a massive source dataset and the latter is trained on the few-shot target dataset. However, sometimes the representation ϕ can also be finetuned on the few-shot dataset (as we explore in our experiments). Importantly, both the class weights and feature extractor must live in the same N-dimensional space in order to compute their inner product:

$$w_y, \phi(\cdot) \in R^N. \tag{3}$$

Though we focus on classification, class models could be learned via other losses (such as centroid prototypes [91]).

Cross-modal learning. Our extension to multiple modalities is staightforward; we assume each training example is accompanied by a discrete label m denoting its modality:

$$(x_i, y_i) \rightarrow (x_i, y_i, m_i), \quad x_i \in X_{m_i}, \quad m_i \in M. \tag{4}$$

For example, one may define the set of modalities to be $M = \{\text{visual}, \text{language}\}$ or $\{\text{visual}, \text{audio}\}$ (Figure 4). We can then define an associated loss:

$$\mathcal{L}_{cross-modal} = \sum_i \mathcal{H}(y_i, \phi_{m_i}(x_i)), \tag{5}$$

where we crucially assume access to modality-specific feature encoders ϕ_m for $m \in M$. While the individual datapoints x_i may come from different modalities with different dimensions, our formulation requires that the encoders map all modalities to the same fixed-dimensional space.

$$w_y, \phi_m(\cdot) \in R^N. \tag{6}$$

Note that this requirement is satisfied by many multimodal foundation models such as CLIP [81] and ALIGN [43] since they map different modalities into the same N-dimensional embedding. We provide training pseudocode for vision-language adaptation (section 3) in algorithm 1 for clarity.

Inference: The learned classifier can produce a label prediction for a test example x from *any* modality $m \in M$:

$$\hat{y} = \arg\max_{y'} w_{y'} \cdot \phi_m(x). \tag{7}$$

This means we can use the same classifier to classify different test modalities (e.g. images and audio clips). In this paper, we mainly evaluate on a single modality (like images) to emphasize that *multimodality helps unimodality*.

Cross-modal ensembles. We now show that cross-modal learning produces classifiers that are ensembles of modality-specific classifiers, exposing a connection to related approaches for ensembling (such as WiSE-FT [100]). We begin by appealing to the well-known *Representer Theorem* [87], which shows that optimally-trained classifiers can be represented as linear combinations of their training samples. In the case of a cross-modal linear probe, weights for class y must be a weighted combination of all i training features, across all modalities:

$$w_y = \sum_i \alpha_{iy}\phi_{m_i}(x_i) = \sum_{m\in M} w_y^m, \quad \text{where}$$
$$w_y^m = \sum_{\{i:m_i=m\}} \alpha_{iy}\phi_m(x_i). \tag{8}$$

Linear classification via cross-modal adaptation solves for all weights α_{iy} *jointly*, so as to minimize the empirical risk (or training loss). In contrast, prior art optimizes for image-specific α_{iy}'s *independently* of the text-specific α_{iy}'s, linearly combining them with a single global α (as in WiSE-FT [100]) or via text-based classifier initialization [21,111]. Our analysis suggests that the joint optimization enabled by cross-modal learning may help other adaptation methods, as our experiments do in fact show.

Extensions. Although we focus on uni-modal inference tasks (e.g. image classification), the above formulation allows the learned classifier to be applied to *multimodal* test sets, such as classifying videos by training on image and audio, and then ensembling predictions across the two modalities with Equation 7. Or, one can extend image classification by providing additional data such as captions and/or attributes. We leave these scenarios as future work. Finally, just as one can optimize uni-modal losses (1) by finetuning the encoder ϕ, one can similarly finetune modality-specific encoders ϕ_m in the cross-modal setting (5). We explore this finetuning method in the next section.

4. Vision-Language Adaptation

We now explore our cross-modal formulation for a particular multimodal setting. Many prior works [68, 104, 111, 113] explore the intersection of vision and language, and thus that is our initial focus. Interestingly, the influential "zero-shot" and "few-shot" evaluation protocols introduced by prior work [81, 102] can be mapped to our cross-modal setting, with one crucial difference; the textual label of each class can be treated as an explicit training sample (x_i, y_i, m_i). From this perspective, "zero-shot" learning may be more naturally thought of as one-shot cross-modal learning that learns a few-shot model on *text* and then infers with it on *images*.

Few-shot evaluation protocol. To ensure a fair comparison, we strictly follow the protocol of CoOp [113] by reporting test performance on 11 public image datasets (Table 5), with ResNet50 [33] as the image encoder backbone. For maximal reproducibility, we use CoOp's dataset splits [113] and the three-fold few-shot train sets sampled with the same random seeds. We adopt the given test split of each dataset as the test set. Some prior works [63, 111] apparently use the large-scale test set to tune hyperparameters for few-shot learning; we instead exercise due diligence by tuning hyperparameters (such as the learning rate, weight decay, and early stopping) on the given few-shot validation set with $min(n, 4)$ examples, where n is the number of training shots. We include PyTorch-style pseudocode (algorithm 1) and hyperparameter details (section 8).

Cross-modal adaptation outperforms SOTA. Table 1 shows the effectiveness of our proposal: we surpass all prior art with an embarrassingly simple linear classifier that requires significantly less training time than other carefully-crafted algorithms. In addition, partial finetuning of the last attentional pooling layer from ϕ_{image} sets the new SOTA. To ensure a fair comparison, we augment the class names into sentences using hand-engineered templates selected by Tip-Adapter [111] (Table 5) and follow their practice to initialize the linear layer with text features. Furthermore, we perform minimal image augmentation with a center crop plus a flipped view instead of random crops as in prior art [111, 113]. As such, we can pre-extract features before training the classifier, leading to significantly less training time as shown in Table 8. We also show that our method can benefit from both image and text augmentation in Table 6. In the appendix, we provide more ablations on classifier initialization (Table 12), partial finetuning (Table 13), and ViT-based backbone (Table 14). Per-dataset results are also in appendix Table 10.

Why does cross-modal learning help? As stated earlier, one reason that cross-modal learning helps is that it turns the original n-shot problem to an $(n + 1)$-shot one. However, Table 1 shows that 1-shot cross-modal linear probing outperforms the 2-shot results of most prior methods. This suggests that training samples from other modalities tend to contain complementary cues [68, 100, 104]. One can loosely observe this in Figure 2 and Figure 5,

Algorithm 1: An example of PyTorch-style pseudocode for cross-modal (vision-language) adaptation. Notably, the image and text samples do not need to be paired and one may sample different numbers of them per batch. For simplicity, we omit linear classifier initialization and early stopping with validation performance. One can also disable the corresponding `grad` field of the encoders for partial finetuning, or pre-extract intermediate features to speed up training.

```
# w:  linear layer initialized with text features
# T: temperature scaling (default is 100)
for _ in iteration:
    # Randomly sample images and texts
    im, im_labels = image_loader.next()
    tx, tx_labels = text_loader.next()

    # Extract image and text features
    im_f = image_encoder(im)
    tx_f = text_encoder(tx)

    # Concatenate then L2 normalize
    features = cat((im_f, tx_f))
    features = normalize(features)
    labels = cat((im_labels, tx_labels))

    # Compute softmax (cross entropy) loss
    logits = w(features)
    loss = cross_entropy_loss(logits / T, labels)
    loss.backward()

    # Update linear layer
    update(w.params)
    # [optional] Update (partial or full) encoders
    update(image_encoder.params)
    update(text_encoder.params)
```

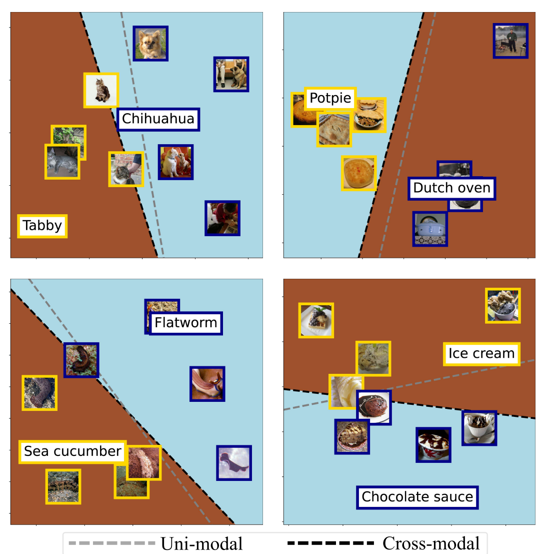

Figure 5. **Additional PCA projection plots for random pairs of classes in ImageNet [15].** Adding one-shot text as training samples can oftentimes aggressively shift the decision boundary.

whereby visual and text examples lie in slightly different parts of the embedding space (indicating the potential to aggressively shape the final decision boundary). In fact, WiSE-FT [100] is inspired by similar reasons to ensemble

Method	Number of shots					Train speed
	1	2	4	8	16	
Zero-Shot CLIP (58.8)	-	-	-	-	-	-
Linear Probing	36.7	47.6	57.2	65.0	71.1	<1min
WiSE-FT [100]	59.1	61.8	65.3	68.4	71.6	<1min
CoOp [113]	59.6	62.3	66.8	69.9	73.4	14hr
ProGrad [114]	62.6	64.9	68.5	71.4	74.0	17hr
Tip-Adapter [111]	64.5	66.7	69.7	72.5	75.8	5min
Tip-Adapter† [111]	63.3	65.9	69.0	72.2	75.1	5min
Cross-Modal Linear Probing	64.1	67.0	70.3	73.0	76.0	<1min
Cross-Modal Partial Finetuning	**64.7**	**67.2**	**70.5**	**73.6**	**77.1**	<3min

Table 1. **Comparison to SOTA using the CoOp [113] protocol**, which reports top-1 accuracy across 11 test sets in Table 5. We include per-dataset results and standard deviation in section 9. For a fair comparison, we reuse the same few-shot visual samples and hand-engineered text prompts used by Tip-Adapter [111]. The original Tip-Adapter searches over hyperparameters (e.g. early stopping) on the large-scale test set, which may not be realistic for few-shot scenarios. Instead, we rerun their codebase and early-stop on a few-shot validation set (as we do), denoted by †. We reproduce WiSE-FT in our codebase since the original work does not provide few-shot results. In summary, by incorporating one-shot text samples into our training set, a simple cross-modal linear probe already outperforms *all* prior methods across *all* shots. Additionally, partial finetuning further improves performance, especially for 8 and 16 shots. Finally, our methods are faster to train than prior work, sometimes significantly (full report in Table 8).

Method	Number of shots				
	1	2	4	8	16
Linear Probing	36.7	47.6	57.2	65.0	71.1
Cross-Modal Linear Probing	64.1	67.0	70.3	73.0	76.0
Δ	27.4	19.4	13.1	8.0	4.9
WiSE-FT [100]	59.1	61.8	65.3	68.4	71.6
Cross-Modal WiSE-FT	63.8	66.4	69.0	71.7	74.1
Δ	4.7	4.6	3.7	3.3	2.5
CoOp [113]	59.6	62.3	66.8	69.9	73.4
Cross-Modal Prompting	62.0	64.9	68.6	71.4	74.0
Δ	2.4	2.6	1.8	1.5	0.6
Tip-Adapter† [111]	63.3	65.9	69.0	72.2	75.1
Cross-Modal Adapter	64.4	67.6	70.8	73.4	75.9
Δ	1.1	1.7	1.8	1.2	0.8

Table 2. **Cross-modal adaptation improves existing methods.** We follow the same protocol as Table 1, reporting the delta accuracy between uni-modal and cross-modal variants of various state-of-the-art methods. The consistent boost suggests that cross-modal training is orthogonal to techniques for uni-modal adaptation, such as prompting [113], adapter [39], and robust finetuning [100].

the uni-modal visual classifier with a "zero-shot" (one-shot-text) classifier (in the linear probing case). However, Equation 8 shows that cross-modal adaptation can also be seen as jointly learning an ensemble, while WiSE-FT [100] learns the visual classifier independently of the text classifier. This suggests that other adaptation methods may benefit from cross-modal learning, as we show next.

Cross-modal adaptation helps prior art (Table 2). This includes prompting (CoOp [113]), adapters (Tip-Adapter [111]), and robust-finetuning (WiSE-FT [100]). We see a large improvement in the low-data regime (1 and 2 shots). Notably, we do not need to tune any methods, and simply reuse the reported hyperparameters. For prompting, we follow CoOp [113] to optimize 16 continuous tokens with the same training setting. For the Adapter model, we follow the same 2-layer MLP architecture of CLIP-Adapter [21] with the given residual ratio of 0.2; we outperform Tip-Adapter without relying on their training-free initialization of MLP. For WiSE-FT, we adopt the given ratio (0.5) to post-hoc ensemble the learned and the zero-shot classifiers. Overall, our experiments suggest that cross-modal adaptation is consistently effective, and should likely be a baseline moving forward given its ease-of-implementation (algorithm 1). For example, instead of separately benchmarking on "zero-shot" (one-shot-text) and few-shot-vision, a cross-modal linear prob would suffice to evaluate representations of a multimodal model.

5. Vision-Audio Adaptation

We now explore cross-modal adaption for other modalities such as audio. We pose the following question: can one learn a better dog *visual* classifier by *listening* to a dog barking? To examine this question, we curate the first audiovisual benchmark that supports few-shot classification of both image and audio.

Our ImageNet-ESC benchmark.[1] We construct our audiovisual benchmark by intersecting two of the most popular image and audio datasets: ImageNet [15] with 1000 types of objects and ESC-50 [77] with 50 types of environmental sounds (including animal, nature, human activity, domestic, and urban noises). We use the class names of the two datasets for class matching. For each class in ESC-50, we check whether there is a corresponding ImageNet class that may produce this type of sound. In this process, we observe that the audio-to-object matching can sometimes be one-to-many. For example, the `clock-alarm` class in ESC-50 can be mapped to either `digital clock` or `analog clock` in ImageNet; the `dog` (barking) class in ESC-50 can be matched to any of the 120 dog species. In such scenarios, we randomly match the classes, e.g. `clock alarm` to `digital clock` and `dog` to `otterhound`. Also, we find that some audio classes loosely match with some visual objects, such as `drinking-sipping` to `water bottle` and `pouring-water` to `water jug`. As such, we create two versions of the dataset: (1) **ImageNet-ESC-27**, which represents the *maximal* intersection consisting of all loose matches, and (2) **ImageNet-ESC-19**, a subset of the for-

[1]Download instructions can be found in our codebase.

mer version consisting of more accurate matches. The final matches are shown in appendix Table 9.

Few-shot evaluation protocol. We use five-fold few-shot splits sampled from ImageNet, with each split divided into half for training and validation. Test performance is recorded on the official ImageNet validation set of the corresponding classes. We adopt the predefined five folds of ESC-50, where each fold contains 8 samples per class. We construct 5 splits from ESC-50 by selecting one fold for training and validation, and record test performance on the other 4 folds. We report averaged performance over 25 runs (since we have 5 random splits for each modality). To keep consistent with our vision-language experiments, we adopt a uni-modal validation and test set and leave cross-modal testing for future work.

Audio encoding. We use AudioCLIP [27] with an ES-ResNeXT backbone [28] as the audio encoder ϕ_{audio}. Because AudioCLIP is trained on a large-scale video dataset (AudioSet [23]) while freezing the pre-trained CLIP text and image encoder, it produces audio embeddings in the same representation space. While AudioCLIP is pretrained on a sizable amount of data, we note that it does not come close to matching the scale of CLIP pretraining [27, 81]. Thus, it does not perform favorably compared to the SOTA for downstream "zero-shot" audio (i.e. one-shot text) classification tasks [27]. However, scaling up audio pretraining is orthogonal to our investigation.

Audio improves image classification. Table 3 shows that adding a random one-shot-audio improves upon naive image-only linear probing, especially in an extremely low-shot setting. This reaffirms Figure 3's hypothesis that cross-modality can reduce the ambiguity of the uni-modal few-shot setup; in other words, one can learn a better *image* classifier by *listen*ing to object sounds. One exception is the 4-shot performance on ImageNet-ESC-27, where adding audio does not help. We posit that (1) loosely-matched classes can result in noisier training data, and (2) the audio representations are not as robust due to smaller-scale pretraining. This suggests that cross-modal adaptation is less effective when representations are not aligned well or insufficiently trained. Nevertheless, under most scenarios, cross-modal adaptation helps. Table 15 shows that adding the language modality (i.e. label names) can significantly boost the performance, which is expected because our benchmark is curated with textual information. For all experiments, we follow an identical procedure to vision-language experiments in section 3 and provide details in appendix section 8.

Vision improves audio classification. We additionally evaluate the *reverse* task – whether adding a random one-shot *image* sample for downstream audio classification can improve upon audio-only training. Table 4 shows the results, where we see the same favorable trend. This success concludes that our approachis modality-agnostic.

Dataset	Method	Image Classification		
		1-shot	2-shot	4-shot
ImageNet-ESC-19	Image-Only Linear	68.0	75.7	83.1
	Image-Audio Linear	**69.3**	**76.7**	**83.2**
ImageNet-ESC-27	Image-Only Linear	60.1	71.8	**79.0**
	Image-Audio Linear	**60.9**	**73.3**	78.9

Table 3. **Image classification results on ImageNet-ESC benchmark.** Adding one audio shot can improve image classification under most few-shot scenarios, even when the audio and vision modalities are only loosely aligned.

Dataset	Method	Audio Classification		
		1-shot	2-shot	4-shot
ImageNet-ESC-19	Audio-Only Linear	31.2	41.1	48.5
	Audio-Image Linear	**35.7**	**45.9**	**51.6**
ImageNet-ESC-27	Audio-Only Linear	28.2	39.0	47.1
	Audio-Image Linear	**35.0**	**43.5**	**48.5**

Table 4. **Audio classification results on ImageNet-ESC benchmark.** Similar to Table 3, adding one image shot improves few-shot audio classification.

Dataset	Classes	Train	Val	Test	Hand-crafted Prompt [111]
Caltech101 [19]	100	4,128	1,649	2,465	a photo of a {cls}.
OxfordPets [75]	37	2,944	736	3,669	a photo of a {cls}, a type of pet.
StanfordCars [50]	196	6,509	1,635	8,041	a photo of a {cls}.
Flowers102 [71]	102	4,093	1,633	2,463	a photo of a {cls}, a type of flower.
Food101 [6]	101	50,500	20,200	30,300	a photo of {cls}, a type of food.
FGVCAircraft [66]	100	3,334	3,333	3,333	a photo of a {cls}, a type of aircraft.
SUN397 [103]	397	15,880	3,970	19,850	a photo of a {cls}.
DTD [14]	47	2,820	1,128	1,692	{cls} texture.
EuroSAT [35]	10	13,500	5,400	8,100	a centered satellite photo of {cls}.
UCF101 [93]	101	7,639	1,898	3,783	a photo of a person doing {cls}.
ImageNet [15]	1000	1.28M	N/A	50,000	itap of a {cls}. a bad photo of the {cls}. a origami {cls}. a photo of the large {cls}. a {cls} in a video game. art of the {cls}. a photo of the small {cls}.

Table 5. **Detailed statistics of the 11 datasets.** We adopt the hand-engineered templates selected by Tip-Adapter [111] unless otherwise stated. Note that this set of templates is identical to the ones selected by CLIP [81] and CoOp [113], except for ImageNet.

6. Ablation Studies

We present a few selected ablation studies in this section. For comprehensive results, please refer to section 9.

Data augmentation of text samples. Like most prior works [81, 113], we also find that data augmentation can improve downstream performance during vision-language adaptation (cf. Table 1). Notably, since the class names are included as training samples, one can explore augmentation techniques for text (just as random cropping for images). Besides the fixed template `a photo of a {cls}` and hand-crafted templates (Table 5), we also try a **template mining** strategy that does not rely on the selected dataset-specific templates. To automatically mine for the templates, we search among a pool of 180 templates for 21 templates with the best zero-shot performance on the few-shot vali-

Finetuning	ImageAugment	TextAugment	Number of shots: 1	2	4	8	16
Linear	CenterCrop	Classname	61.8	65.3	69.0	72.0	74.9
		`a photo of a {cls}.`	63.2	66.2	69.7	72.5	75.3
		Template Mining	63.5	67.2	70.3	73.1	75.7
		Hand Engineered [111]	63.7	66.7	70.3	72.9	75.5
	+Flipped View	Hand Engineered [111]	**64.1**	**67.0**	**70.3**	**73.0**	**76.0**
Partial	CenterCrop	Classname	62.5	65.7	69.3	72.9	76.2
		`a photo of a {cls}.`	63.8	66.8	69.8	73.4	76.7
		Template Mining	64.3	67.1	70.3	73.5	76.5
		Hand Engineered [111]	64.6	67.2	70.2	73.7	76.9
	+Flipped View	Hand Engineered [111]	**64.7**	**67.7**	**70.6**	**73.8**	**77.2**

Table 6. **Augmentation for cross-modal adaptation.** We evaluate the impact of selected augmentation techniques following the same CoOp protocol as in Table 1.

dation set of each dataset. We discuss how we collect the 180 templates in appendix section 8. For image augmentation, we perform standard flipping and random cropping. We show a subset of results in Table 6, and find that all text augmentation techniques provide a sizable boost in performance. We also report comprehensive ablations in appendix Table 11 and compare it to the SOTA prompting method ProDA [63]. The salient conclusions include (1) the performance gain from image augmentation is saturated after more than two views, and (2) template mining can be as competitive as a large number of 36 carefully-tuned prompts. In fact, prompting [61, 63, 113] can be viewed as another *text augmentation* technique under cross-modal adaptation, and we leave this exploration to future work.

Test-time distribution shifts. We examine how robust our approach is against test-time distribution shifts in Table 7. Specifically, we follow the CoOp [113] protocol to report the test performance of a classifier trained on the source dataset (16-shot ImageNet) to 4 distribution-shifted target test sets, including ImageNet-V2 [83], ImageNet-Sketch [96], ImageNet-A [37], and ImageNet-R [36]. As shown in Table 7, cross-modal adaptation can significantly boost the robustness of image-only linear probing and is competitive against baselines designed to address robustness such as CoCoOp [112] and WiSE-FT [100]. Cross-Modal adaptation also improves upon WiSE-FT [100] and sets the new SOTA. We can conclude that language modality plays an important role in robustness, similar to how humans rely on textual cues for recognition [37].

Efficiency. As shown in Table 8, our approaches are much more lightweight because we do not rely on deep finetuning [112, 113] or heavy image augmentations. This allows us to speed up training by pre-extracting features, resulting in rather fast training speeds.

7. Discussion and Limitations

We show that cross-modal training is a lightweight and effective approach for adapting pre-trained multimodal models for downstream uni-modal tasks. One reason for its effectiveness is that it naturally addresses the underspecification of few-shot learning. In the context of vision-language adaptation, one can achieve SOTA results by using existing text labels as free training samples. In the context of vision-audio adapation, one can learn better visual object classifiers by listening to object sounds (and better audio classifiers by looking at objects!). One attractive aspect of cross-modal learning is that the learned models naturally apply to multimodal test data, such as the classification of videos that contain both visual and audio signals. However, cross-modal learning is less effective when model representations are not well-aligned or insufficiently trained. Nevertheless, due to its simplicity and effectiveness, we hope cross-modal learning becomes a tool for future research on multi-modal adaptation.

Method	Source: ImageNet	Target: -V2	-Sketch	-A	-R
ResNet50					
Zero-Shot CLIP	58.2	51.3	33.3	21.7	56.0
Linear Probing	55.9	46.0	19.1	12.7	34.9
CoOp (M=4)	63.0	55.1	32.7	22.1	55.0
CoOp (M=16)	63.3	55.4	34.7	**23.1**	56.6
WiSE-FT (α=0.5)	62.9	54.2	33.3	20.3	57.4
Cross-Modal WiSE-FT (α=0.5)	**65.2**	**56.6**	**35.6**	22.6	**59.5**
Cross-Modal Linear Probing	64.5	55.3	33.1	20.0	56.4
ViT-B/16					
Zero-Shot CLIP	66.7	60.8	46.2	47.8	74.0
Linear Probing	65.9	56.3	34.8	35.7	58.4
CoOp (M=4)	71.9	64.2	46.7	48.4	74.3
CoOp (M=16)	71.7	64.6	47.9	49.9	75.1
CoCoOp	71.0	64.1	48.8	**50.6**	76.2
WiSE-FT (α=0.5)	73.0	65.2	49.1	49.8	77.6
Cross-Modal WiSE-FT (α=0.5)	72.9	**65.4**	**49.2**	50.5	**77.8**
Cross-Modal Linear Probing	**73.2**	64.8	47.9	48.3	76.4

Table 7. **Robustness under test-time distribution shifts.** We follow CoOp [113]'s protocol for evaluating the test-time performance on variants of ImageNet. We report results with two image encoders (ResNet50 and ViT-B/16), and mark the **best** and second best results. Salient conclusions: (a) Cross-modal linear probing is much more robust than its uni-modal counterpart while being competitive to previous SOTA methods such as WiseFT and CoOp, and (b) it can be further augmented with post-hoc modification through WiseFT to achieve new the SOTA.

Method	Iteration	Time	Accuracy	Gain
Zero-shot CLIP [81]	0	0	60.33	0
Image-Only Linear	12k	**15sec**	56.44	-3.89
CoOp [113]	100k	14h 40min	62.95	+2.62
ProGrad [113]	100k	17hr	63.45	+3.12
Tip-Adapter [111]	10k	5min	65.18	+5.18
Cross-Modal Linear	12k	**15sec**	64.51	+4.14
Cross-Modal Partial	12k	2.5min	**65.95**	**+5.57**

Table 8. **Efficiency and accuracy for different methods on ImageNet-16-shot.** All experiments are tested with batch size 32 on a single NVIDIA GeForce RTX 3090 GPU. Our approaches take less time and achieve SOTA performance.

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