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,samueleyu,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 reading about dogs and listening 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?

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].

*Equal contribution.
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, reading about dogs and listening 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!

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-
all aforementioned methods including WiSE-FT optimize lightweight MLPs while freezing the encoder. Yet, by parameter-efficient finetuning methods particularly, CLIP-Adapter full-finetuning ample, CLIP recipes crafted for uni-modal foundation models. For example, trained models have excelled at classic vision tasks future directions.

most of the trending techniques are built upon successful efficient adaptation methods. However, we observe that 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.

**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 \(\phi(\cdot)\):

\[
\mathcal{L}_{uni-modal} = \sum_{i} \mathcal{H}(y_i, \phi(x_i))
\]

where \(\mathcal{H}\) is typically the softmax loss

\[
\mathcal{H}(y, f) = -\log \left( p(y|f) \right) = -\log \left( \frac{e^{wy_i,f}}{\sum_{y'} e^{wy',f}} \right).
\]

Our notation separates the feature extractor \(\phi\) 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 \(\phi\) 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:

\[
w_y, \phi(\cdot) \in R^N.
\]

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 straightforward: we assume each training example is accompanied by a discrete label \(m\) denoting its modality:

\[
(x_i, y_i) \rightarrow (x_i, y_i, m_i), \quad x_i \in X_{m_i}, \quad m_i \in M.
\]

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

\[
\mathcal{L}_{cross-modal} = \sum_{i} \mathcal{H}(y_i, \phi_m(x_i)),
\]

where we crucially assume access to modality-specific feature encoders \(\phi_m\) 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.

\[
w_y, \phi_m(\cdot) \in R^N.
\]

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 \(m \in M\):

\[
\hat{y} = \arg \max \limits_{y'} w_{y'} \cdot \phi_m(x).
\]

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:

\[
w_y = \sum_i \alpha_{iy} \phi_m(x_i) = \sum_{m \in M} w_y^m, \quad \text{where}
\]

\[
w_y^m = \sum_{\{i : m_i = m\}} \alpha_{iy} \phi_m(x_i).
\]

Linear classification via cross-modal adaptation solves for all weights \(\alpha_{iy} \text{ jointly} \), so as to minimize the empirical risk (or training loss). In contrast, prior art optimizes for image-specific \(\alpha_{iy}\)’s independently of the text-specific \(\alpha_{iy}\)’s, linearly combining them with a single global \(\alpha\) (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 \(\phi\), one can similarly finetune modality-specific encoders \(\phi_m\) 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,
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 \(\phi_{image}\) 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 \(i\) 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
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 adaptation 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. 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 Im-

ducible instructions can be found in our codebase.

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 \( \Delta \). 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).

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.
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 \(\Phi_{\text{audio}}\). 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 listening 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 approach is modality-agnostic.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Image Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-shot</td>
<td>2-shot</td>
</tr>
<tr>
<td>ImageNet-ESC-19</td>
<td>Image-Only Linear</td>
<td>68.0</td>
</tr>
<tr>
<td></td>
<td>Audio-Image Linear</td>
<td>69.3</td>
</tr>
<tr>
<td>ImageNet-ESC-27</td>
<td>Image-Only Linear</td>
<td>60.1</td>
</tr>
<tr>
<td></td>
<td>Audio-Image Linear</td>
<td>60.9</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Audio Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-shot</td>
<td>2-shot</td>
</tr>
<tr>
<td>ImageNet-ESC-19</td>
<td>Audio-Only Linear</td>
<td>31.2</td>
</tr>
<tr>
<td></td>
<td>Audio-Image Linear</td>
<td>35.7</td>
</tr>
<tr>
<td>ImageNet-ESC-27</td>
<td>Audio-Only Linear</td>
<td>28.2</td>
</tr>
<tr>
<td></td>
<td>Audio-Image Linear</td>
<td>35.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Class</th>
<th>Train</th>
<th>Val</th>
<th>Test</th>
<th>Hard-crafted Prompt [111]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUB-200</td>
<td>100</td>
<td>4,128</td>
<td>1,649</td>
<td>2,069</td>
<td>a photo of a [cls]</td>
</tr>
<tr>
<td>OxfordIIIT</td>
<td>10</td>
<td>2,064</td>
<td>736</td>
<td>528</td>
<td>a photo of a [cls], a type of pet.</td>
</tr>
<tr>
<td>StanfordCore50</td>
<td>196</td>
<td>6,609</td>
<td>6,609</td>
<td>8,041</td>
<td>a photo of a [cls]</td>
</tr>
<tr>
<td>Flowers102</td>
<td>102</td>
<td>4,091</td>
<td>1,634</td>
<td>2,463</td>
<td>a photo of a [cls], a type of flower.</td>
</tr>
<tr>
<td>Food101</td>
<td>101</td>
<td>50,500</td>
<td>20,200</td>
<td>30,300</td>
<td>a photo of [cls], a type of food.</td>
</tr>
<tr>
<td>FGVC-AirPac</td>
<td>100</td>
<td>3,334</td>
<td>1,353</td>
<td>5,333</td>
<td>a photo of a [cls], a type of aircraft.</td>
</tr>
<tr>
<td>SUN397</td>
<td>107</td>
<td>15,810</td>
<td>9,070</td>
<td>19,800</td>
<td>a photo of a [cls].</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>47</td>
<td>2,620</td>
<td>1,128</td>
<td>1,492</td>
<td>a centered satellite photo of [cls].</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>10</td>
<td>13,500</td>
<td>5,400</td>
<td>8,100</td>
<td>a photo of a person doing [cls].</td>
</tr>
</tbody>
</table>

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-
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 under specification 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 adaptation, 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.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-Shot CLIP</td>
<td>ImageNet</td>
<td>-V2</td>
</tr>
<tr>
<td>Linear Probing</td>
<td>58.2</td>
<td>51.3</td>
</tr>
<tr>
<td>CoOp (M=4)</td>
<td>55.9</td>
<td>46.0</td>
</tr>
<tr>
<td>CoOp (M=16)</td>
<td>63.0</td>
<td>55.1</td>
</tr>
<tr>
<td>WiSE-FT ($\alpha=0.5$)</td>
<td>63.3</td>
<td>55.4</td>
</tr>
<tr>
<td>Cross-Modal WiSE-FT ($\alpha=0.5$)</td>
<td>62.9</td>
<td>54.2</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Iteration</th>
<th>Time</th>
<th>Accuracy</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-shot CLIP [81]</td>
<td>0</td>
<td>0</td>
<td>60.33</td>
<td>0</td>
</tr>
<tr>
<td>Image-Only Linear</td>
<td>12k</td>
<td>15sec</td>
<td>56.44</td>
<td>-3.89</td>
</tr>
<tr>
<td>CoOp [113]</td>
<td>100k</td>
<td>4hr 40min</td>
<td>62.95</td>
<td>+2.62</td>
</tr>
<tr>
<td>ProGrad [113]</td>
<td>100k</td>
<td>17hr</td>
<td>63.45</td>
<td>+3.12</td>
</tr>
<tr>
<td>Tip-Adapter [111]</td>
<td>10k</td>
<td>5min</td>
<td>65.18</td>
<td>+5.18</td>
</tr>
<tr>
<td>Cross-Modal Linear</td>
<td>12k</td>
<td>15sec</td>
<td>64.51</td>
<td>+4.14</td>
</tr>
<tr>
<td>Cross-Modal Partial</td>
<td>12k</td>
<td>2.5min</td>
<td>65.95</td>
<td>+5.57</td>
</tr>
</tbody>
</table>

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


[27] Andrey Guzov, Federico Raue, Jörn Hees, and Andreas Dengel. Audiostream: Extending clip to image, text and audio, 2021. 1, 2, 3, 7
[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
[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
[43] Chao Jia, Yinfeng 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
[51] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization.


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


