A. Ablations

Architecture for Visual Adapter Layer. In the main paper, we use a linear layer for adapting visual features. To ablate the varying depth of $\mathbf{H}_{\boldsymbol{\theta}}$, we increase the number of layers (with ReLU activations in between). Table 4 shows the results on the ImageNet validation set using CLIP RN-50 on 16-shot classification.

Weighting Parameter α for different datasets. DAC improves the classification capability of both inter-modal and intra-modal classifiers. We use a scalar α to balance the contributions of each classifier towards the final accuracy. The value for α is selected based on the performance on the validation sets. A similar strategy was employed by Tip-Adapter [53]. However, in Tip-Adapter it is used to determine how much residual information should flow from intra-modal classifier to update the inter-modal predictions. To find the optimal value, we perform a grid search with a step size of 0.01, a search range in [0.1, 10], and the number of search steps being 10000. In this section, we present different values of α used to compute the final test performances of each dataset. In Fig. 8, we show how varying α influences the performance on 16-shot ImageNet classification. Table 5 lists our optimal values for α for all datasets (both DAC-V and DAC-VT). Since α is multiplied with the intra-modal logits, it can be seen that DAC-V consumes more information from the intra-modal classifier. Remember that in DAC-V, we only optimize the visual representations of CLIP without optimizing it for the upstream few-shot classification task. This further highlights the benefits of having better intra-modal representations in few-shot adaptation setting.

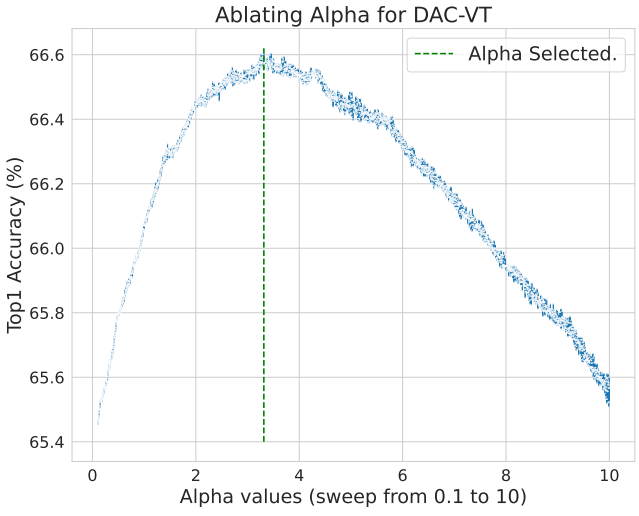


Figure 8. Ablating the α parameter for ImageNet using CLIP ResNet50.

B. Detailed Analysis on Error Inconsistencies

We analyze the error inconsistencies observed across various datasets in Fig. 9. This plot complements our analysis in Sec. 5 about the role of inter- and intra-modal classifiers in an ensembled setting, and further illustrates how DAC-VT reduces inconsistencies between intra and inter-modal classifiers. The consensus between the DAC-VT's sub-classifiers is higher for some datasets (*e.g.*, Flowers102, Caltech101), however, the inconsistencies for certain datasets (*e.g.*, FGVCAircraft) are still high.

C. A Case for Aligning Textual Representations in Target Domain

We further elaborate on why it is important to align textual features on each downstream task. Previous work [35] has shown that CLIP's zero-shot transfer is vulnerable to expansion of downstream vocabulary used for class labels. This becomes even more important when the visual concepts in the target domain get associated with different class labels, presented at different granularities. Figure 10 shows an example (taken from [40]) where multiple, different labels from ImageNet can be used to describe the same image. Such cases are particularly difficult for vision-language

Structure of $\mathbf{H}_{\boldsymbol{\theta}}$	DAC-V	DAC-VT
Linear Layer	64.89	66.61
2 Layer MLP	64.452	65.582
3 Layer MLP	64.08	65.274
4 Layer MLP	64.01	65.064

Table 4. Ablating structure of adapter layer $\mathbf{H}_{\boldsymbol{\theta}}$

Datasets	DAC-V α	DAC-VT α
UCF-101	3.78	1.16
Caltech101	2.40	1.33
ImageNet	8.32	3.31
SUN397	5.95	1.39
FGVCAircraft	8.2	6.91
StanfordCars	6.50	2.42
Flowers102	8.17	3.43
Food101	1.17	1.05
OxfordPets	1.07	0.73
DTD	3.05	1.11
EuroSAT	5.17	0.76

Table 5. Details of α used to weigh intra and inter-modal classifiers for different datasets in DAC-V and DAC-VT. In DAC-V the contribution from intra-modal features is weighted more which indicates that the adapted visual cache contains reliable information to update CLIP's inter-modal knowledge.

models to generalize to in zero-shot manner, unless more context is given by either prompts or some domain-specific training data.

Note that the adaptation of textual representations introduced in *cf*. 4.2 aims to caters for such confusing examples as it modulates the overall textual embedding (including the class name). Such an optimization allows the textual cache to adapt the class description according to the visual concepts defined by a few observed images.

Few-shots	1	2	4	8	16
Linear-probe CLIP	22.17	31.98	41.20	49.52	56.13
CoOp	57.15	57.81	59.99	61.56	62.95
CLIP-Adapter	61.20	61.52	61.84	62.68	63.59
Tip-Adapter	60.70	60.92	60.95	61.48	62.00
Tip-Adapter-F	61.19	61.75	62.48	63.84	65.47
DAC-V	60.71	61.48	61.87	63.38	64.89
DAC-VT	**61.32**	**62.39**	**63.11**	**64.78**	**66.61**

Table 6. Top1 accuracy of different methods on ImageNet at different shots.

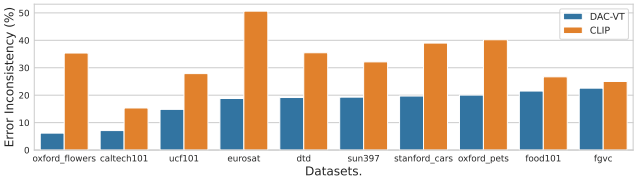
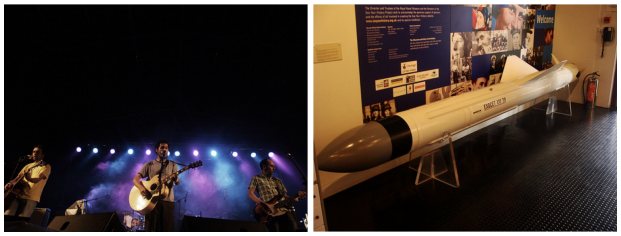


Figure 9. **Comparative analysis of error inconsistencies between intra-modal and inter-modal classifiers of CLIP and DAC-VT on 10 different datasets** (sorted by DAC-VT's performance). We observe that DAC-VT significantly reduces the error inconsistencies, however, the performance gap reduces on certain datasets such Food101 and FGVCAircrafts.



(a) stage (b) missile

Figure 10. **Examples of confusing labels in ImageNet [40]**. The labels above appear to correctly describe the visual concepts, however, ImageNet assigns *acoustic_guitar* and *projectile* labels to the images, respectively.

D. Understanding Inter-Modal and Intra-Modal Representations Alignment

In this section, we delve into understanding how DAC-VT modulates the interactions between inter-modal and intra-modal representations. We look at them from the perspective of cone effects occurrences in representations distances that's been extensively studied in [24]. In Fig. 11, we showcase the range of cosine similarities scores obtained by computing similarities between inter-modal and intra-modal representations. It can be seen that even after updating textual representations, DAC-VT maintains the same range of inter-modal similarity between images and text as in CLIP. The bigger shift is observed in intra-modal alignment where the visual representations tuned with DAC have a different support in comparison to TIP and CLIP based intra-modal alignments. We conjecture that this shift happens because the supervised contrastive objective used to tune visual representations introduce a different learning inductive bias than what was used to aligning image-text representations.

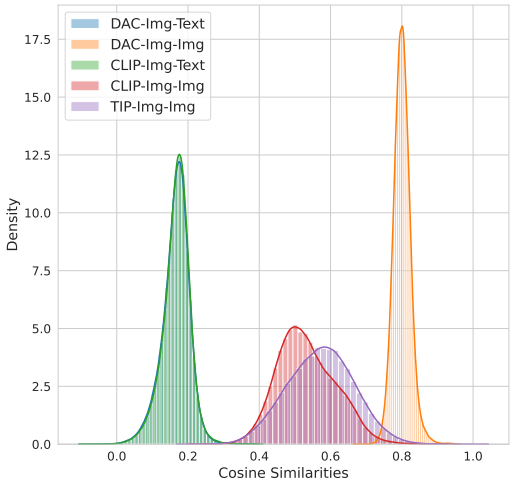


Figure 11. **Pictorial depiction of modality gaps between intra-modal and inter-modal representations of different methods** (illustrated by cosine similarities). It can be seen that the DAC-VT's and CLIP image-text similarities remain within the same range.