CALIP: Zero-Shot Enhancement of CLIP with Parameter-Free Attention

Ziyu Guo[1,2*], **Renrui Zhang**[2,3*†], **Longtian Qiu**[4*], **Xianzheng Ma**[3], **Xupeng Miao**[5], **Xuming He**[4], **Bin Cui**[1†]

[1] School of CS and Key Lab of HCST, Peking University
[2] The Chinese University of Hong Kong
[3] Shanghai AI Laboratory
[4] ShanghaiTech University
[5] Carnegie Mellon University
{guo.ziyu, bin.cui}@pku.edu.cn, zhangrenrui@pjlab.org.cn, {qiult, hexm}@shanghaitech.edu.cn

Abstract

Contrastive Language-Image Pre-training (CLIP) has been shown to learn visual representations with promising zero-shot performance. To further improve its downstream accuracy, existing works propose additional learnable modules upon CLIP and fine-tune them by few-shot training sets. However, the resulting extra training cost and data requirement severely hinder the efficiency for model deployment and knowledge transfer. In this paper, we introduce a free-lunch enhancement method, ***CALIP***, to boost ***CLIP***'s zero-shot performance via a parameter-free ***A***ttention module. Specifically, we guide visual and textual representations to interact with each other and explore cross-modal informative features via attention. As the pre-training has largely reduced the embedding distances between two modalities, we discard all learnable parameters in the attention and bidirectionally update the multi-modal features, enabling the whole process to be parameter-free and training-free. In this way, the images are blended with textual-aware signals and the text representations become visual-guided for better adaptive zero-shot alignment. We evaluate CALIP on various benchmarks of 14 datasets for both 2D image and 3D point cloud few-shot classification, showing consistent zero-shot performance improvement over CLIP. Based on that, we further insert a small number of linear layers in CALIP's attention module and verify our robustness under the few-shot settings, which also achieves leading performance compared to existing methods. Those extensive experiments demonstrate the superiority of our approach for efficient enhancement of CLIP. Code is available at https://github.com/ZiyuGuo99/CALIP.

Introduction

With the advance of learning theories and network architectures, supervised methods under a close-set assumption have achieved extraordinary results over a wide range of vision tasks, such as image classification (He et al. 2016; Krizhevsky, Sutskever, and Hinton 2012; Parmar et al. 2018; Mao et al. 2021), object detection (Ren et al. 2015; Carion et al. 2020; Zheng et al. 2020; Chen et al. 2017), and point cloud understanding (Qi et al. 2017a,b). Despite their success in those specific scenarios, they often lack the ability to attain general visual representations, which harms their transferability to open-set applications. Alternatively, based on exploiting the wide coverage of languages, Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021) proposes to conduct visual learning contrastively with descriptive natural language data. Pre-trained by large-scale image-text pairs, CLIP extracts both features of input images and texts by independent encoders, and aligns the paired ones within the same embedding space. On downstream tasks, given a new dataset with images of "unseen" classes, CLIP constructs the textual inputs by the category names and converts the original classification task into a image-text matching problem. As such, CLIP is able to achieve zero-shot recognition in open-vocabulary settings and obtains promising performance on various benchmarks.

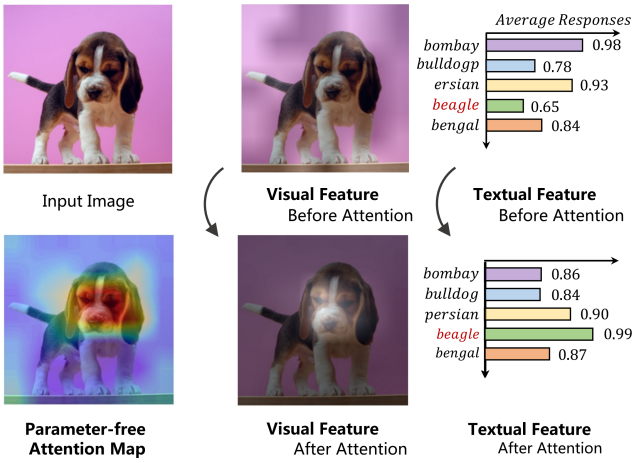

Figure 1: Visualization of Parameter-free Attention and the Interacted Features. Without any parameters, CALIP's cross-modal attention map (Left-Bottom) shows favorable weight distributions over the main objects, which well updates both visual and textual features: pixels within objects of ground-truth labels are enhanced and the corresponding category features in red are strengthened.

*These authors contributed equally.
†Corresponding author.

To further improve the downstream performance of CLIP, existing works introduce different fine-tuning methods for the few-shot classification. Inspired by prompt tuning (Li and Liang 2021) and adapters (Houlsby et al. 2019) in natural language processing, Context Optimization (CoOp) (Zhou et al. 2021), CLIP-Adapter (Gao et al. 2021) and Tip-Adapter (Zhang et al. 2021a) freeze CLIP's pre-trained weights and adopt learnable prompts or lightweight adapters to tune the textual and visual features. Despite the performance improvement, all existing methods with task-specific designs contain learnable parameters and rely on additional training phase with few-shot labeled data. This leads to extra resource cost and largely hinders CLIP's inherent advantage for efficient zero-shot knowledge transfer. As an example, existing methods are required to fine-tune CLIP separately for different downstream tasks, and deploy multiple model copies for different applications. Therefore, we ask the question: **Can we adapt CLIP by a more efficient and general method without additional few-shot data or training?**

To tackle this issue, we propose **CALIP**, which equips CLIP with a parameter-free attention module to conduct cross-modal interactions and avoid the need for extra downstream data or training, as shown in Figure 2. Before the CLIP outputting the final global feature of an image, we utilize its intermediate feature map, which preserves more fine-grained semantic information and contextual characteristics of the image. Then, we conduct a parameter-free cross-modal attention between the spatial visual feature and the textual feature, containing no learnable parameter. Different from traditional attention mechanism, our design consists of two key modifications, which are non-parametric and bidirectional. For the former, as the features of CLIP's two modalities have been well aligned during the contrastive pre-training, we are able to simply omit the linear layers within the attention, which were supposed to project the features into queries, keys and values. Therefore, their attention map can be directly calculated by matrix multiplication between features. For the latter, as there is no discrimination for queries, keys or values, we can simultaneously update both visual and textual features via the only attention map. With this attention mechanism, the visual feature is guided by category semantics from the texts, which converts their per-pixel features to be more distinctive for recognition. Correspondingly, the text counterpart adaptively explores features from informative regions on the image and becomes visual-aware and image-conditional, instead of remaining the same for the entire dataset. The visualization in Figure 1 demonstrates the effectiveness of our parametric-free attention. Finally, the zero-shot prediction of CALIP is obtained by matching between the visual and textual features after our proposed cross-modal interactions.

The whole process of CALIP is zero-shot, training-free and universal for various downstream tasks. We implement and evaluate CALIP on 14 datasets including zero-shot 2D image and 3D point cloud classification to illustrate its effectiveness. For some benchmarks, zero-shot CALIP without training even surpasses some prior methods after few-shot fine-tuning. On top of that, to fully unleash the power of cross-modal attention, we further add a small number of linear layers in the attention module and upgrade the parameter-free attention into a parametric version, named **CALIP-FS**. Under the few-shot fine-tuning, CALIP-FS achieves leading performance among all existing methods, which demonstrates the superiority of our proposed attention framework. The main contributions of CALIP are as follows:

- To our best knowledge, CALIP is the **first work** to conduct zero-shot enhancement over CLIP for downstream tasks without few-shot data or additional training.
- We design a parameter-free attention for cross-modal interactions upon CLIP to effectively exchange image-text informative features for better alignment.
- The parametric version, CALIP-FS with learnbale cross-modal attention modules, also achieves competitive performance among all existing few-shot methods.

Related Work

Downstream Adaption of CLIP. As a breakthrough in vision-language learning, CLIP (Radford et al. 2021) has shown great potential for obtaining generic visual representations by contrastive pre-training. Based on the superior transferable ability, the problem of effectively adapting CLIP to downstream tasks has been widely studied. Given few-shot training data, CoOp (Zhou et al. 2021) proposes the learnable prompts for textual inputs inspired by prompt learning (Li and Liang 2021), and VT-CLIP (Zhang et al. 2021c) introduces visual-guided texts for better vision-language alignment. Referring to adapters (Houlsby et al. 2019), CLIP-Adapter (Gao et al. 2021) appends a lightweight adapter module to produce adapted multi-modal features. Tip-Adapter (Zhang et al. 2021a) and CaFo (Zhang et al. 2022a, 2023) greatly reduce its training cost by constructing a key-value cache model. Besides 2D, PointCLIP (Zhang et al. 2021b; Zhu et al. 2022) extend CLIP into 3D data understanding by projecting point clouds into multi-view depth maps. Other works also apply CLIP for semantic segmentation (Rao et al. 2021), depth estimation (Zhang et al. 2022d), video analysis (Lin et al. 2022), and self-supervised pre-training (Zhang et al. 2022c; Gao et al. 2023). However, the existing downstream adaption of CLIP demands extra training data and the resources for fine-tuning, which weakens CLIP's core advantage of efficient zero-shot recognition. In this paper, we explore CALIP to enhance CLIP's downstream performance under zero-shot settings by interacting its two modalities with no parameter. In addition, our approach can be utilized for both 2D and 3D domains and is also well-performed when few-shot data are available, indicating great generalization ability.

Method

In this section, we first revisit CLIP for zero-shot recognition as the preliminary. Then we present the details of our zero-shot CALIP with parameter-free attention, followed by the parametric version, CALIP-FS.

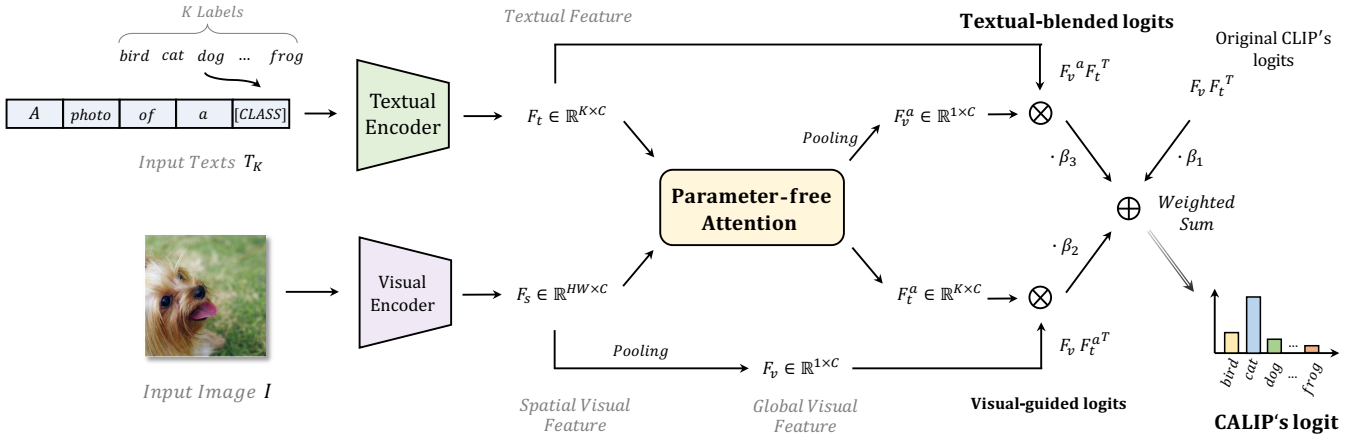

Figure 2: The Pipeline of CALIP. We introduce a parameter-free attention module for zero-shot enhancement of CLIP and require no extra data or training for downstream tasks. CALIP utilizes pre-trained encoders to extract spatial visual feature of the input image and K-category textual feature. Then, the proposed attention module updates their representations via cross-modal interactions and outputs the final zero-shot prediction by weighted summation of three classification logits.

Preliminary of CLIP

CLIP utilizes 400 million image-text pairs for contrastive pre-training in an unsupervised way, obtaining the ability to match "unseen" images with their corresponding categories. To extract features of both modalities, CLIP has two independent encoders: a ResNet (He et al. 2016) or vision transformer (ViT) (Dosovitskiy et al. 2021) for visual encoding, and a 12-layer transformer (Vaswani et al. 2017) for textual encoding, denoted as $\mathrm{VisEnc}(\cdot)$ and $\mathrm{TexEnc}(\cdot)$, respectively. For the downstream dataset with K categories, $\{C_1, C_2 \ldots, C_K\}$, CLIP places all category names into the [CLASS] token of a pre-defined textual template, e.g., "a photo of a [CLASS]", constructing K textual inputs $\mathrm{T_K}$. Then, their textual features are extracted as $F_t \in R^{K\times C}$, whose i-th row vector, $i = 1, \ldots, K$, represents the encoded knowledge of category C_i. For every input image I to be recognized, CLIP extracts its spatial feature map $F_s \in R^{H\times W\times C}$ and obtains the global visual representation $F_v \in R^{1\times C}$ by pooling operation. Finally, features from both encoders are matched via cosine similarities to produce the classification $logits \in R^{1\times K}$. The whole process is as

$$F_t = \mathrm{TexEnc}(\mathrm{T_K}), \tag{1}$$

$$F_v = \mathrm{Pooling}(F_s), \quad F_s = \mathrm{VisEnc}(\mathrm{I}), \tag{2}$$

$$logits = F_v F_t^T, \tag{3}$$

where we assume F_v and F_t^T are L2-normalized features and their matrix multiplication is equal to cosine similarities calculation. $logits$ denote the probabilities for all K categories and CLIP outputs the maximum one as the prediction.

CALIP with Parameter-free Attention

Motivation. While CLIP achieves promising results on zero-shot open-vocabulary recognition, which is concise and efficient, it still has room for improvement. We observe that the two modalities are totally isolated during encoding and there is no bridge for inter-modal information flow before the final matching. In addition, the spatial structures of images in F_s are largely left out by the pooling operation, which might harm the fine-grained visual understanding. More importantly, we aim to inherit the great strength of CLIP's zero-shot capacity for training-free transfer learning, which requires no downstream data. Therefore, we propose our parameter-free attention module (CALIP) to not only fulfill the cross-modal interactions, but also achieve the goal to conduct zero-shot enhancement over CLIP.

Design Details. After CLIP's encoding of two modalities, we utilize the intermediate spatial visual feature $F_s \in R^{H\times W\times C}$ and textual feature $F_t \in R^{K\times C}$ for interactions. We reshape F_s into a 1D vector sequence, $F_s \in R^{HW\times C}$ and obtain their attention weights directly by matrix multiplication without any projection,

$$A = F_s F_t^T \in R^{HW\times K}, \tag{4}$$

where A denotes the cross-modal attention map. Each element of A represents the attention weight, namely, the feature similarity between a category and one image pixel/site. Based on A, we bidirectionally update both textual and visual features as follows

$$F_s^a = \mathrm{SoftMax}(A/\alpha_t)F_t, \tag{5}$$

$$F_t^a = \mathrm{SoftMax}(A^T/\alpha_s)F_s, \tag{6}$$

where α_t and α_s modulate the attention magnitude for textual and visual modalities, respectively. Weighted by the attention scores representing similarity, two modalities both aggregate informative features from each other as visualized in Figure 1. For texts, as F_t encodes K-category knowledge, the signals of categories appearing on the image would be amplified and others would be restrained. Also, the textual features are now adaptive for different input images in a non-parametric manner, other than being fixed in all existing methods (Zhou et al. 2021; Gao et al. 2021; Zhang et al. 2021a). Likewise for the image, the pixel features within

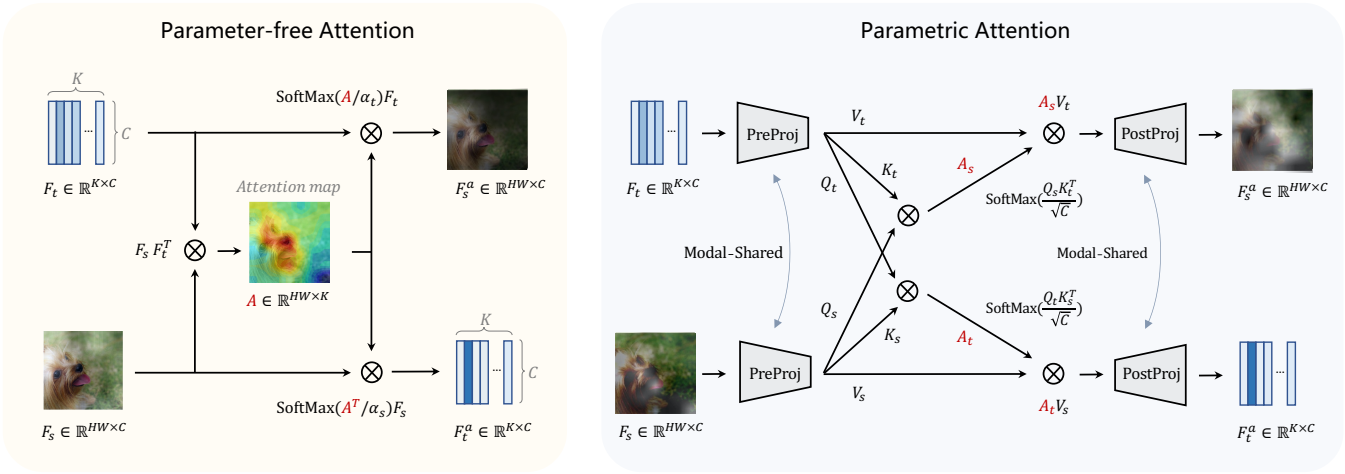

Figure 3: Structures of Parameter-free (Left) and Parametric Attention (Right). Parameter-free attention directly obtains the cross-modal attention map A by matrix multiplication and bidirectionally updates two features for zero-shot classification. Parametric attention is equipped with both pre-projection and post-projection layers for better few-shot performance.

foreground objects, which belong to the K categories, would become more notable. Meanwhile, the spatial feature map F_s provides pixel-level fine-grained information for the interaction, contributing to thorough cross-modal communication. Finally, we obtain the attention-interacted global visual feature by pooling and output the classification logits as

$$F_v^a = \text{Pooling}(F_s^a), \tag{7}$$

$$logits = \beta_1 \cdot F_v F_t^T + \beta_2 \cdot F_v F_t^{aT} + \beta_3 \cdot F_v^a F_t^T, \tag{8}$$

where $\beta_{1\sim3}$ denote the weights for three logits: the original CLIP's logits, visual-guided logits and textual-blended logits. By aggregation, CALIP achieves favorable zero-shot performance without few-shot fine-tuning or data.

Analysis. There are two differences between ours and the vanilla attention mechanism. The first is parametric-free: we involve no learnable parameters during the attention processing. The vanilla attention takes as input two terms and utilizes separate learnable linear layers to map them into the attention embedding space, where one as the query and the other as key and value. In contrast, our textual and visual features have already been pre-trained to be within the same space and can discard the linear layers for projection. The other difference is bidirectional. Traditional attention only updates one of the inputs, which is projected as the query, and maintains the other the same. Our design updates both of them for better interaction. As we have removed the difference for query, key and value, both input terms, visual and textual features are symmetric and play the same roles.

CALIP-FS with Parametric Attention

Motivation. Although the parameter-free attention enhances CLIP's zero-shot performance on a wide range of datasets, we expect to further unleash the power of cross-modal interactions under few-shot settings. Therefore, we construct CALIP-FS by inserting several learnable linear layers before and after the attention. We freeze the pre-trained encoders of CLIP and only fine-tune the inserted layers in the cross-modal attention for training efficiency.

Design Details. As shown in Figure 3, to save the parameters, we apply a modal-shared pre-projection layers to transform the textual feature F_t and spatial visual feature F_s into the C-dimensional query, key and value,

$$Q_t, K_t, V_t = \text{PreProject}(F_t), \tag{9}$$

$$Q_s, K_s, V_s = \text{PreProject}(F_s), \tag{10}$$

where $\text{PreProject}(\cdot)$ is composed of three linear layers respectively for query, key and value and shared for two modalities. Then, we calculate two attention maps,

$$A_t = \text{SoftMax}(\frac{Q_t K_s^T}{\sqrt{C}}) \in R^{K\times HW}, \tag{11}$$

$$A_s = \text{SoftMax}(\frac{Q_s K_t^T}{\sqrt{C}}) \in R^{HW\times K}, \tag{12}$$

where A_t and A_s are respectively for textual and visual features update. As the learnable projection layers are available, we could specify the attention maps to achieve modal-specific attention calculation. Afterwards, we obtain the updated features with shared post-projection layers,

$$F_t^a = \text{PostProject}(A_t V_s), \tag{13}$$

$$F_s^a = \text{PostProject}(A_s V_t), \tag{14}$$

where $\text{PostProject}(\cdot)$ only contains one linear layer. Then, we apply pooling to process F_t^a and acquire the final predicted logits by weighted summation of three terms, the same as the non-parametric version above. Equipped with such learnable projection layers, CALIP-FS significantly improves the performance over zero-shot CALIP and achieves competitive results among other state-of-the-art models by few-shot fine-tuning.

Average over 11 2D Datasets		
Model	Acc.	Shot Num.
CLIP	58.53	0-Shot
CALIP	**59.45**	0-Shot
-	-	-

ImageNet		
Model	Acc.	Shot Num.
CLIP	60.32	0-Shot
CALIP	**60.57**	0-Shot
CoOp	59.99	**4-Shot**

Caltech101		
Model	Acc.	Shot Num.
CLIP	83.94	0-Shot
CALIP	**87.71**	0-Shot
CoOp	87.53	**1-Shot**

SUN397		
Model	Acc.	Shot Num.
CLIP	58.53	0-Shot
CALIP	**58.59**	0-Shot
Linear.	54.49	**4-shot**

Food101		
Model	Acc.	Shot Num.
CLIP	77.32	0-Shot
CALIP	**77.42**	0-Shot
CLIP-A.	77.20	**2-Shot**

Flowers102		
Model	Acc.	Shot Num.
CLIP	66.10	0-Shot
CALIP	**66.38**	0-Shot
Linear.	58.07	**1-Shot**

StanfordCars		
Model	Acc.	Shot Num.
CLIP	55.71	0-Shot
CALIP	**56.27**	0-Shot
CoOp	55.59	**1-Shot**

FGVCAircraft		
Model	Acc.	Shot Num.
CLIP	17.10	0-Shot
CALIP	**17.76**	0-Shot
CoOp	9.64	**1-Shot**

OxfordPets		
Model	Acc.	Shot Num.
CLIP	85.83	0-Shot
CALIP	**86.21**	0-Shot
CoOp	85.32	**8-Shot**

DTD		
Model	Acc.	Shot Num.
CLIP	40.07	0-Shot
CALIP	**42.39**	0-Shot
Linear.	39.48	**2-Shot**

EuroSAT		
Model	Acc.	Shot Num.
CLIP	37.54	0-Shot
CALIP	**38.90**	0-Shot
-	-	-

UCF101		
Model	Acc.	Shot Num.
CLIP	61.33	0-Shot
CALIP	**61.72**	0-Shot
Linear.	53.55	**2-Shot**

Table 1: Zero-Shot Performance (%) of CALIP on Eleven 2D Datasets. Our zero-shot CALIP can consistently outperform CLIP and even surpass some methods with few-shot fine-tuning. "Linear." and "CLIP-A." denote Linear-probe CLIP and CLIP-Adapter, respectively.

Average over 3 3D Datasets		
Model	Acc.	Shot Num.
PointCLIP	21.90	0-Shot
CALIP	**23.60**	0-Shot

ModelNet10		
Model	Acc.	Shot Num.
PointCLIP	30.13	0-Shot
CALIP	**32.44**	0-Shot

ModelNet40		
Model	Acc.	Shot Num.
PointCLIP	20.18	0-Shot
CALIP	**21.47**	0-Shot

ScanObjectNN		
Model	Acc.	Shot Num.
PointCLIP	15.38	0-Shot
CALIP	**16.90**	0-Shot

Table 2: Zero-Shot Performance (%) of CALIP on Three 3D Datasets. We extend CALIP for 3D point cloud recognition based on PointCLIP under zero-shot settings, where CALIP shows stable performance enhancement.

Experiments

Zero-shot CALIP

Datasets To fully evaluate the zero-shot enhancement of CALIP, we experiment on a wide range of benchmarks including 11 image 2D datasets and 3 point cloud 3D datasets. 2D datasets contain a variety of visual concepts, e.g., real-world scenarios, satellite-captured landscapes and detailed textures, which are ImageNet (Jia et al. 2009), Caltech101 (Li, Fergus, and Perona 2004), OxfordPets (Vedaldi 2012), StanfordCars (Krause et al. 2014), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Gool 2014), FGVCAircraft (Maji et al. 2013), SUN397 (Xiao et al. 2010), DTD (Cimpoi et al. 2013), EuroSAT (Helber et al. 2017) and UCF101 (Soomro, Zamir, and Shah 2012). The 3D datasets include both synthetic and sensor-scanned point clouds: ModelNet10 (Wu et al. 2015), ModelNet40 (Wu et al. 2015) and ScanObjectNN (Uy et al. 2019). As CALIP requires no downstream data for training, we utilize no training sets of the datasets and directly evaluate on their full test sets.

Settings We adopt ResNet-50 (He et al. 2016) as the visual encoder and a 12-layer transformer as the textual encoder. Following CLIP's (Radford et al. 2021) pre-processing, we resize all test images into 224×224 resolutions and H, W, C of visual spatial feature F_s denote 7, 7, 1024. We set α_t and α_s for modulating textual and visual attention magnitude both as 2. For the pooling operation of F_s^a, we select the combination of maximum and average poolings for better features integration. We adopt varying $\beta_1, \beta_2, \beta_3$ for different datasets to adapt their specific domains. As for textual templates, we refer to CLIP adopting handcrafted ones. Regarding 3D point cloud recognition, CALIP follows PointCLIP (Zhang et al. 2021b) to project point clouds onto 6-view depth maps with the distance 1.2 and aggregate view-wise zero-shot predictions as the final output.

Analysis As shown in Table 1, we compare zero-shot CALIP with CLIP and some few-shot models for 2D image classification. Our CALIP with parameter-free attention consistently outperforms CLIP on all downstream benchmarks by +0.92% average accuracy. We largely surpass CLIP by +3.77% on Caltech101 and +1.36% on EuroSAT. CALIP without training even beats existing learnable methods under few-shot fine-tuning, e.g., **surpassing 1-shot Linear-probe CLIP by +8.89% on Flowers102, and 8-shot CoOp by +0.89% on OxfordPets.** As for 3D point cloud classification in Table 2, CALIP also enhances PointCLIP on 3 datasets by +1.70% average accuracy without parameters.

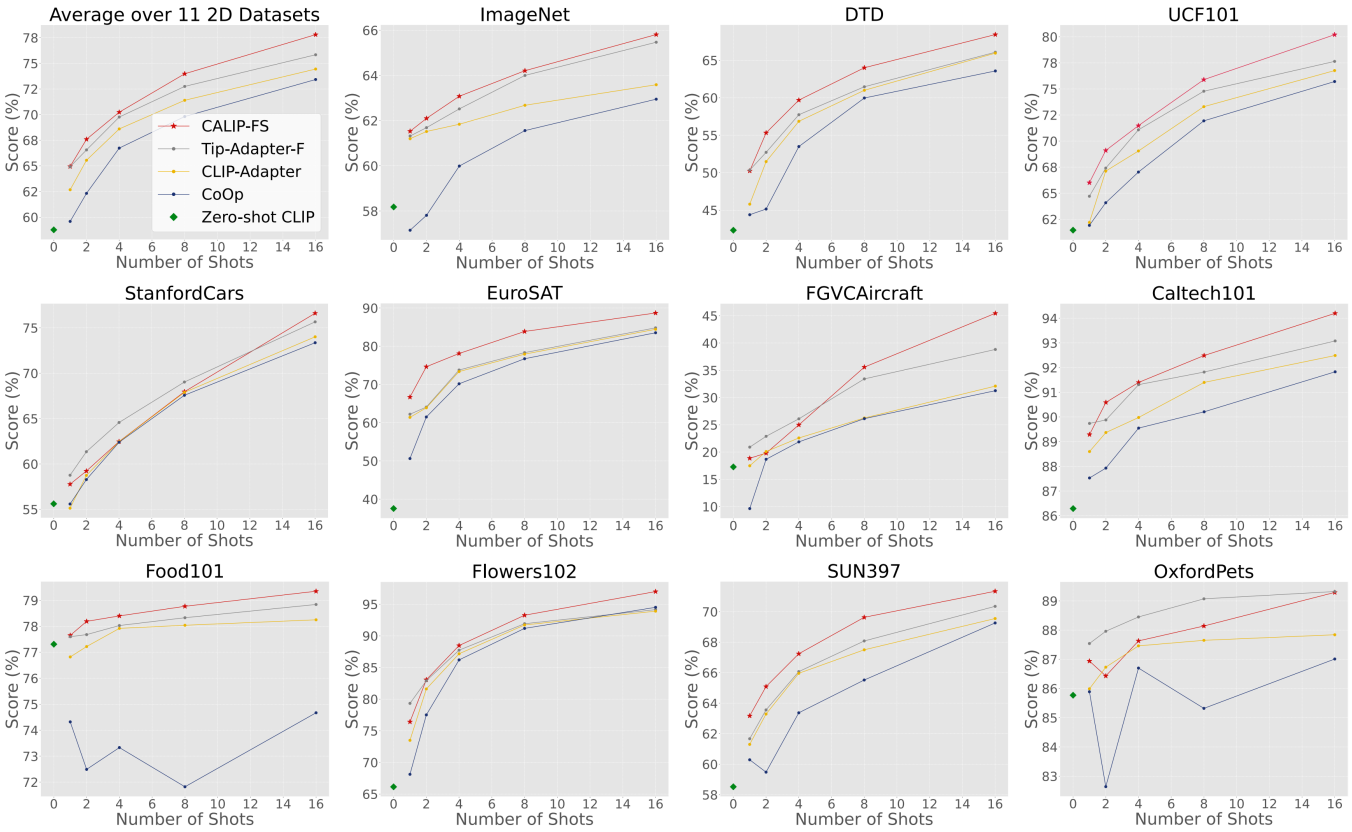

Figure 4: Few-shot Performance pf CALIP-FS on Eleven 2D Datasets. CALIP-FS shows the overall best performance over previous baselines for few-shot recognition of a wide range of visual concepts.

Datasets	Source	Target			
	ImageNet	-V2	-A	-R	-Sketch
CLIP	60.32	53.27	23.61	60.42	35.44
CALIP	**60.57**	**53.70**	**23.96**	**60.81**	**35.61**
Linear-probe	56.13	45.61	12.71	34.86	19.13
CoOp	62.95	54.58	23.06	54.96	31.04
CALIP-FS	**65.81**	**55.98**	**23.42**	**56.74**	**35.37**

Table 3: Performance (%) on Distribution Shift.

Few-shot CALIP-FS

Datasets We evaluate CALIP-FS for few-shot classification on 11 2D datasets mentioned above and compare ours with the state-of-the-art methods: zero-shot CLIP (Radford et al. 2021), CoOp (Zhou et al. 2021), CLIP-Adapter (Gao et al. 2021) and Tip-Adapter-F (Zhang et al. 2021a). We follow the widely-adopted few-shot protocols, which randomly sample 1, 2, 4, 8 and 16 shots of each category for training and test models on the full test set.

Analysis The main results are presented in Figure 4. The average accuracy over 11 datasets on the top-left corner indicates CALIP-FS's superior few-shot performance over all other baselines. Based on zero-shot CLIP, CALIP-FS achieves significant performance improvements, especially on DTD and EuroSAT, ranging from +20% to +50%. Compared to other few-shot methods, we only lag behind Tip-Adapter-F on OxfordPets, and largely outperform others on DTD, EuroSAT and SUN397. More importantly, rather than Tip-Adapter-F's complicated two-step fine-tuning by storing all training samples, CALIP-FS is more efficient and simple with the one-step training.

Out-of-distribution Performance

Robustness to distribution shift is a common benchmark to evaluate the generalization ability of deep-learning models. We evaluate the out-of-distribution performance of CALIP and CALIP-FS by training on ImageNet and testing on ImageNetV2 (Recht et al. 2019), ImageNet-Sketch (Wang et al. 2019), ImageNet-A (Hendrycks et al. 2021b) and ImageNet-R (Hendrycks et al. 2021a). These test datasets contain compatible categories with ImageNet but within different visual domains. In Table 3, we compare ours with the published results of zero-shot CLIP, Linear-probe CLIP and CoOp. As shown, CALIP acquires better generalization ability than CLIP without training. By 16-shot fine-tuning, CALIP-FS also surpasses CoOp on four out-of-distribution datasets.

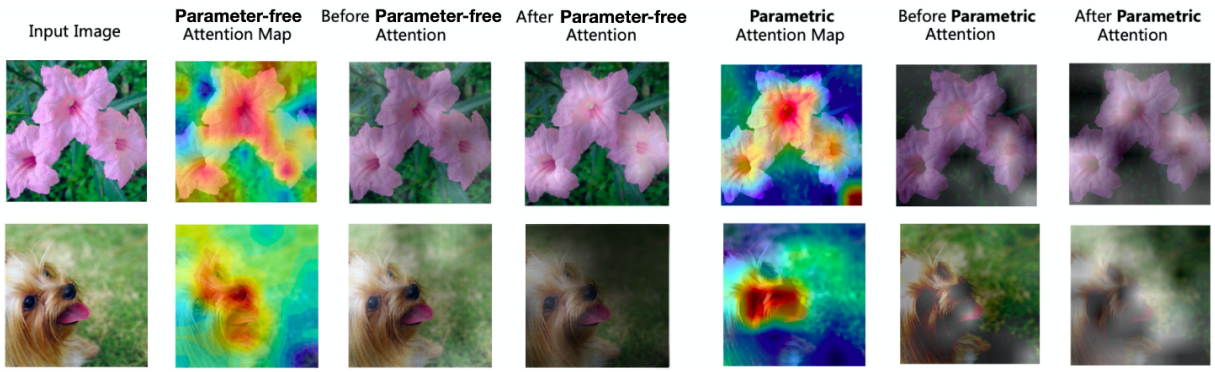

Figure 5: Visualization of Attention Maps and Spatial Visual Features in CALIP and CALIP-FS.

Combination of Logits					
$F_v F_t^T$	$F_v F_t^{aT}$	$F_v^a F_t^T$	$F_v^a F_t^{aT}$	CALIP	CALIP-FS
✓	-	-	-	83.94%	83.94%
✓	✓	-	-	84.02%	94.42%
✓	-	✓	-	85.10%	93.39%
✓	-	-	✓	81.96%	94.60%
✓	✓	✓	-	**85.66%**	**94.75%**
✓	✓	✓	✓	85.34%	94.66%

Table 4: Ablation Study of Logits Combination.

Visual Projection		Textual Projection		Accuracy
Pre-Proj.	Post-Proj.	Pre-Proj.	Post-Proj.	
-	-	-	-	87.71%
✓	-	✓	-	89.75%
✓	-	✓	✓	90.36%
✓	✓	✓	-	93.94%
✓	✓	✓	✓	**94.75%**

Table 5: Ablation Study of Pre/Post-Projection Designs.

Ablation Study

To further demonstrate the theory of our approach, we conduct ablation studies on Caltech101 dataset with zero-shot CALIP and 16-shot CALIP-FS. We report our results on the official validation set for tuning hyperparameters and network structures.

Cross-modal Attention The attention aggregates three terms of logits for the final output: $F_v F_t^T$, $F_v F_t^{aT}$ and $F_v^a F_t^T$, where the first term is the CLIP's original prediction and the other two respectively contains the attention-interacted F_t^a and F_v^a. There actually exists the fourth term: $F_v^a F_t^{aT}$, that is, the logits predicted by the updated features of both modalities. In Table 4, we explore their best combination form and observe that, for both CALIP and CALIP-FS, the fourth term $F_v^a F_t^{aT}$ would adversely influence the predicted logits, since its too much cross-modal interaction might harm the already well-aligned knowledge from pre-trained CLIP. In contrast, the combination of logits that only interact one modality via the attention performs better. It not only preserves the effective pre-trained CLIP's knowledge, but also fuses newly-interacted cross-modal knowledge.

Pre/Post-Projection Layers We explore where to insert learnable linear layers in CALIP's parameter-free attention to construct CALIP-FS. As shown in Table 5, equipping both pre/post-projection layers for two modalities achieves the best performance. This design decouples the embedding space of attention calculation from the previous one by the former projecting-in and the latter projecting-out layers, which produces better attention map for interactions.

Visualization

In Figure 5, we visualize attention maps, spatial visual features before and after the CALIP's parameter-free attention and CALIP-FS's parametric attention, respectively. As shown, for both variants, the attention maps concentrate well around the object pixels, and the visual features become more distinctive guided by category texts as expected. Also, after few-shot fine-tuning, the distributions of attention maps and visual features all get more intensive, which indicates the improvements resulted from learnable parameters.

Conclusion

We propose CALIP, the first work to conduct zero-shot enhancement over CLIP via a parameter-free attention module. CALIP interacts visual and textual features without any parameters or training and achieves favorable performance over a wide range of 2D and 3D benchmarks. Then, we introduce the parametric version CALIP-FS to further boost its classification accuracy under few-shot fine-tuning and acquire competitive results among existing state-of-the-art methods. We hope our work could inspire future researches for zero-shot enhancement of pre-trained large-scale multi-modal models. Concerning limitations, we will further extend our parameter-free methods for wider vision tasks, or even develop purely non-parametric networks like Point-NN (Zhang et al. 2022b).

Acknowledgements

This work is supported by NSFC (No. 61832001 and U22B2037).

References

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101 – Mining Discriminative Components with Random Forests. In *Springer International Publishing*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229. Springer.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1907–1915.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2013. Describing Textures in the Wild. *IEEE*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *arXiv preprint arXiv:2110.04544*.

Gao, P.; Zhang, R.; Fang, R.; Lin, Z.; Li, H.; Li, H.; and Yu, Q. 2023. Mimic before Reconstruct: Enhancing Masked Autoencoders with Feature Mimicking. *arXiv preprint arXiv:2303.05475*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2017. 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*.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Jia, D.; Wei, D.; Socher, R.; Li, L. J.; Kai, L.; and Li, F. F. 2009. ImageNet: A large-scale hierarchical image database. 248–255.

Krause, J.; Stark, M.; Deng, J.; and Li, F. F. 2014. 3D Object Representations for Fine-Grained Categorization. In *IEEE International Conference on Computer Vision Workshops*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

Li, F.; Fergus, R.; and Perona, P. 2004. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *IEEE*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Lin, Z.; Geng, S.; Zhang, R.; Gao, P.; de Melo, G.; Wang, X.; Dai, J.; Qiao, Y.; and Li, H. 2022. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, 388–404. Springer.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-Grained Visual Classification of Aircraft. *HAL - INRIA*.

Mao, M.; Zhang, R.; Zheng, H.; Gao, P.; Ma, T.; Peng, Y.; Ding, E.; Zhang, B.; and Han, S. 2021. Dual-stream network for visual recognition. *arXiv preprint arXiv:2105.14734*.

Nilsback, M. E.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *Sixth Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*.

Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; and Tran, D. 2018. Image transformer. In *International Conference on Machine Learning*, 4055–4064. PMLR.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2021. DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting. *arXiv preprint arXiv:2112.01518*.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 5389–5400. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *Computer Science*.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; and Yeung, S.-K. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1588–1597.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vedaldi, A. 2012. Cats and dogs. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3498–3505.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition*.

Zhang, R.; Deng, H.; Li, B.; Zhang, W.; Dong, H.; Li, H.; Gao, P.; and Qiao, Y. 2022a. Collaboration of Pre-trained Models Makes Better Few-shot Learner. *arXiv preprint arXiv:2209.12255*.

Zhang, R.; Fang, R.; Gao, P.; Zhang, W.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2021a. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. *arXiv preprint arXiv:2111.03930*.

Zhang, R.; Guo, Z.; Gao, P.; Fang, R.; Zhao, B.; Wang, D.; Qiao, Y.; and Li, H. 2022b. Point-M2AE: Multi-scale Masked Autoencoders for Hierarchical Point Cloud Pre-training. *arXiv preprint arXiv:2205.14401*.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2021b. PointCLIP: Point Cloud Understanding by CLIP. *arXiv preprint arXiv:2112.02413*.

Zhang, R.; Hu, X.; Li, B.; Huang, S.; Deng, H.; Li, H.; Qiao, Y.; and Gao, P. 2023. Prompt, Generate, then Cache: Cascade of Foundation Models makes Strong Few-shot Learners. *arXiv preprint arXiv:2303.02151*.

Zhang, R.; Qiu, L.; Zhang, W.; and Zeng, Z. 2021c. VT-CLIP: Enhancing Vision-Language Models with Visual-guided Texts. *arXiv preprint arXiv:2112.02399*.

Zhang, R.; Wang, L.; Qiao, Y.; Gao, P.; and Li, H. 2022c. Learning 3D Representations from 2D Pre-trained Models via Image-to-Point Masked Autoencoders. *arXiv preprint arXiv:2212.06785*.

Zhang, R.; Zeng, Z.; Guo, Z.; and Li, Y. 2022d. Can Language Understand Depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, 6868–6874.

Zheng, M.; Gao, P.; Wang, X.; Li, H.; and Dong, H. 2020. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2021. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*.

Zhu, X.; Zhang, R.; He, B.; Zeng, Z.; Zhang, S.; and Gao, P. 2022. PointCLIP V2: Adapting CLIP for Powerful 3D Open-world Learning. *arXiv preprint arXiv:2211.11682*.