SDSA: Shallow-Deep Squeezing Adapter for Vision-Language Models

Hao Wang[1], Rui Zhu[2]*, Xiaoxu Li[1]*, Zhanyu Ma[3,4], and Jing-Hao Xue[5]

[1] Lanzhou University of Technology, Lanzhou 730050, China
{232081203025,lixiaoxu}@lut.edu.cn
[2] City St Geroge's, University of London, London EC1Y 8TZ, U.K.
rui.zhu@city.ac.uk
[3] Beijing University of Post and Telecommunication, Beijing 100876, China
[4] Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance
mazhanyu@bupt.edu.cn
[5] University College London, London WC1E 6BT, U.K.
jinghao.xue@ucl.ac.uk

Abstract. Recent adapters for vision-language models (VLMs) still suffer from dense cross-modal interactions and limited control over alignment capacity. To address this issue, we propose a two-stage Shallow-Deep Squeezing Adapter (SDSA) that explicitly regulates cross-modal interaction density and alignment capacity. In the shallow squeezing stage, SDSA leverages token-level masking to impose structured sparsity within a sparse bottleneck space. Then, in the deep squeezing stage, SDSA applies a shared low-rank transformation and a cross-modal attention module: the low-rank module consolidates alignment into a compact set of dominant shared directions, while cross-modal attention refines representations through selective interaction. Extensive experiments on 11 datasets show that SDSA delivers superior base-to-novel generalization and cross-dataset evaluation, and significantly improves the generalizability of VLMs under few-shot conditions. The code can be found at https://github.com/haowang-ac/SDSA.

Keywords: Vision-language models · Fine-tuning · Adapter · Few-shot classification

1 Introduction

Vision-language models (VLMs) are powerful approaches to learning aligned visual and textual representations [1, 4, 30]. By aligning semantic spaces between images and text, VLMs like Contrastive Language-Image Pre-training (CLIP) [26] entertain strong generalization and transfer capabilities across diverse downstream tasks, in particular on CLIP-based few-shot learning [11, 17, 18, 35, 39, 41, 42] by exploring prompt learning and lightweight adapters. For instance, MaPLe [17] pioneers the interaction between visual and textual features

* Corresponding authors.

for VLMs, achieving greater discriminability and generalizability than CoOp [42] and CoCoOp [41] do for few-shot classification. On this basis, by introducing adapters and leveraging prompt alignment and interaction, MMA [35] proposes a multi-modal prompt learning paradigm that shortens the training time.

Despite achieving promising performance in few-shot classification, existing methods such as MMA mainly enhance adaptation flexibility through multi-modal prompt interactions, but do not explicitly regulate cross-modal interaction density and alignment capacity. As a result, token-level interactions can remain relatively dense, and alignment may extend to minor or unstable directions, potentially limiting generalization under few-shot supervision.

To address this limitation, we propose the Shallow-Deep Squeezing Adapter (SDSA), a hierarchical two-stage mechanism designed to regulate cross-modal interaction density and alignment capacity. Specifically, shallow squeezing introduces token-level masking to impose structured sparsity and form a sparse bottleneck representation. Deep squeezing then applies a shared low-rank constraint followed by cross-modal attention, consolidating dominant semantic directions within a controlled shared subspace.

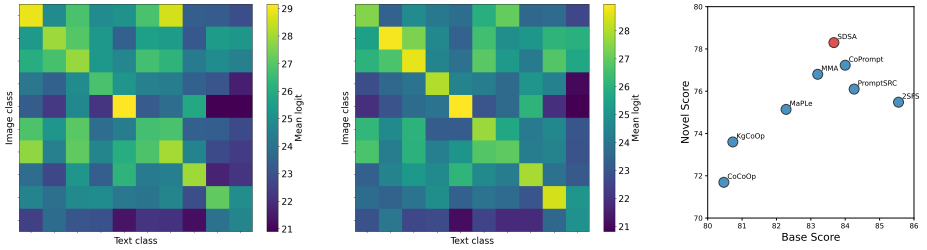
Shallow squeezing acts as a lightweight pre-filter that reduces token-level activations before deeper cross-modal interaction. We design a random token-level masking strategy to induce structured sparsity and limit unnecessary cross-modal coupling, effectively transforming dense token interactions into a sparse high-dimensional space. This process improves the interaction density of subsequent alignment while preserving salient visual and text responses. The sparse representation is subsequently projected through a fully connected layer into a bottleneck space.

Deep squeezing further regulates cross-modal alignment within the bottleneck by imposing a shared low-rank constraint across modalities. This transformation consolidates sparse features into a compact set of dominant shared alignment directions within a controlled low-dimensional subspace. By limiting alignment capacity, the model suppresses minor and unstable coupling directions that often capture spurious or dataset-specific correlations. After the low-rank transformation, cross-modal attention selectively exchanges complementary information between visual and textual tokens, enabling structured refinement while preserving principal semantic correspondences, thereby improving downstream generalization.

As shown in Fig. 1(a)-(b), our SDSA shows more accurate feature matching than MMA does. The logit distributions of SDSA are more concentrated along the diagonal, indicating stronger text–image alignment and reduced inter-class confusion. Furthermore, as illustrated in Fig. 1(c), although our SDSA is not the best on the base classes, its generalizability to novel classes is significantly better than other state-of-the-art methods.

In summary, our main contributions are:

- We introduce shallow squeezing, a token-level sparsity mechanism that suppresses unstable cross-modal interactions and enforces structured interaction



(a) MMA's text–image alignment heatmap on EuroSAT.

(b) SDSA's text–image alignment heatmap on EuroSAT.

(c) Performance on base and novel classes.

Fig. 1: (a)–(b) Text–image alignment heatmaps on the EuroSAT dataset, where each entry represents the mean logits averaged over all images in the corresponding class, for (a) MMA and (b) our SDSA method. (c) Performance comparison in terms of the accuracies on the base classes and those on the novel classes. Our SDSA performs the best in the generalizability on the novel classes.

density, yielding a compact and structured bottleneck representation for both visual and textual modalities.

- We propose deep squeezing, a shared low-rank transformation constraint that consolidates cross-modal alignment into a compact set of dominant shared directions, and cross-attention further facilitates structured token-level interaction. By restricting alignment capacity, it preserves principal semantic correspondence while suppressing minor and unstable coupling directions.
- Extensive experiments on 11 datasets show that SDSA delivers superior base-to-novel generalization and cross-dataset evaluation, and significantly improves the discriminability and generalizability of VLMs under few shots.

2 Related Work

2.1 Vision-Language Models

CLIP [26] and Align [16] have pioneered VLMs and drawn considerable attention to adapting these pre-trained VLMs to challenging downstream tasks such as those under data scarcity and domain shift. CLIP uses contrastive learning to monitor the distance at the same image-text embedding, and performed very well in zero-shot tasks. Align was proposed in the same year and trained on 1 billion image–text pairs. Although these pre-trained VLMs have learned generalized representations and significantly improved the performance of zero-shot tasks, effectively adapting them to specific downstream tasks remains a challenging direction. Many studies have shown that by using customized methods to fine-tune VLMs for few-shot image recognition [11, 17, 35, 40], object detection [10, 12, 31], and segmentation [21, 33], they can achieve better performance in downstream tasks.

2.2 Fine Tuning for CLIP

CoOp [42] modifies the manual prompts in its language branch to a continuous prompt vector to improve the performance of CLIP. CoCoOp [41] further expands CoOp by setting up image-specific prompt conditions. CoPrompt [28] applies two types of perturbation to the input, feeds one into an adapter-augmented model and the other into frozen CLIP, and then aligns their features via consistency constraints to reduce prompt overfitting. CLIP-Adapter [11] and Tip-Adapter [39] add a bottleneck layer after the image/text encoder of CLIP, and perform residual-style feature blending with the original pretrained features. PromptKD [20] adopts an unsupervised prompt distillation strategy, addressing both model lightweighting and performance improvement in few-shot and unlabeled scenarios. LoRA [38] inserts trainable low-rank matrices within linear layers (e.g., attention layers), enabling model adaptation to new tasks with an extremely small number of parameters. MaPLe [17] is a pioneer in enabling the joint interaction between visual and textual features for VLMs, achieving stronger discriminability and generalization than CoOp and CoCoOp in few-shot classification tasks. Subsequently, PromptSRC [18] enhances MaPLe with self-regularization to mitigate overfitting. On this basis, by introducing adapters and leveraging prompt alignment and interaction, MMA [35] proposes a multi-modal adapter paradigm that shortens the training time.

Unlike LoRA, which adapts only a single modality via low-rank weight updates, and MMA, which introduces multi-modal adapter learning without explicitly control token-level interaction density, SDSA adopts a hierarchical approach. In the shallow squeezing stage, SDSA imposes token-level masking to control cross-modal interaction density, producing a sparse bottleneck representation. In the deep squeezing stage, a shared low-rank adapter constrains alignment capacity, consolidating dominant semantic directions across modalities. This two-stage design enables both efficient adaptation of text and visual encoders and more stable, well-aligned multimodal adaper learning.

3 Methodology

Our proposed method SDSA involves adding a fast adapter module to the pretrained CLIP [26], to enhance its generalization ability in downstream tasks. Fig. 2 shows the overall architecture of SDSA, which shares weights between the image and text representations to learn shared features from different modalities. With this design, our SDSA enables feature-wise interactions between each image-text pair.

3.1 Contrastive Language-Image Pretraining

Here we briefly introduce the preliminary knowledge about CLIP [26], which takes input from image and text modalities. Image encoder $\mathcal{V}$ with K transformer layers $\{\mathcal{V}_i\}_{i=1}^{K}$ provides image patch embeddings $E_i = [x_i^1, x_i^2, \cdots, x_i^M] \in \mathbb{R}^{M \times d_v}$,

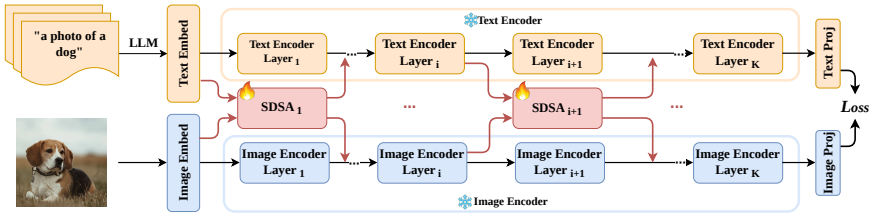


Fig. 2: The overall architecture of using SDSA in CLIP. SDSA operates on both the image and text encoders, with the entire pre-trained CLIP model frozen. SDSA shares weights between the image and text representations to learn shared features from different branches. With this design, our SDSA eliminates feature-wise interactions between each image-text pair, while keeping the computational cost minimal.

where M is the number of patches and d_v is the dimension of images. and learnable class tokens (CLS), c_i:

$$[c_i, E_i] = \mathcal{V}_i([c_{i-1}, E_{i-1}]) \quad i = 1, 2, \cdots, K. \tag{1}$$

The text encoder $\mathcal{T}$ generates text features by tokenizing the words and projecting them to word embeddings $W_i = [w_i^1, w_i^2, \cdots, w_i^N] \in \mathbb{R}^{N \times d_l}$, where N is the number of tokens and d_l is the dimension of text features. At each stage, W_i is input to the $(i+1)^{th}$ transformer layer of text encoding branch$\mathcal{T}_{i+1}$:

$$[W_i] = \mathcal{T}_i(W_{i-1}), i = 1, 2, \cdots, K. \tag{2}$$

To obtain the final image representation x and text representation z, the class token c_K of last transformer layer ($\mathcal{V}_K$) and text embeddings w_k^N corresponding to the last token of the last transformer block ($\mathcal{T}_K$) are projected into a universal vision-language latent embedding space via ImageProj and TextProj:

$$x = \text{ImageProj}(c_K) \quad x \in \mathbb{R}^{d_{vl}}, \tag{3}$$

$$z = \text{TextProj}(w_K^N) \quad z \in \mathbb{R}^{d_{vl}}. \tag{4}$$

Then, CLIP makes a prediction for the image x:

$$P(y = c|x) = \frac{\exp(sim(x, z_y)/\tau)}{\sum_{c' \in C} \exp(sim(x, z_{c'})/\tau)}, \tag{5}$$

where $sim(\cdot)$ denotes the cosine similarity and τ denotes the temperature parameter.

3.2 Class Description Augmentation

We believe that one important reason for CLIP's limited performance is that its text tokens are significantly fewer than the image features. To obtain richer

linguistic knowledge, we leverage DeepSeek [22] to enhance textual descriptions as in [28]. We perform augmentation on the original text token *a class of a [CLS]* using DeepSeek with prompt key *the classname is description*, e.g., the original token *a photo of a [tench]* is augmented to *The [tench] is fish with olive-green skin and red eyes.* After passing through the LLM, W_i's are transformed into $T = \{t_i\}_{i=1}^{K}$.

3.3 Shallow-Deep Squeezing Adapter

We observe that in existing methods [17, 35], token-level interactions remain dense and cross-modal alignment may extend to unstable coupling directions. To mitigate this issue, we propose the **Shallow-Deep Squeezing Adapter (SDSA)** as shown in Fig. 3. The SDSA consists of a two-layer squeezing-expanding adapter framework. The masking stage randomly drops part of the activations to eliminate redundant and noisy information, subsequently the core low-dimensional compression stage is applied within the resulting sparse feature space enabling the extraction of the most critical semantic factors.

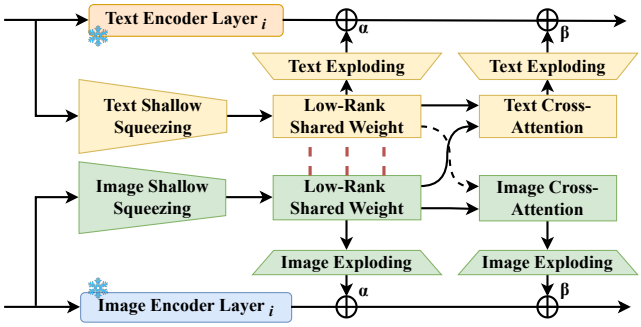


Fig. 3: SDSA structure. The unit contains separate shallow compression layers to compress text and image information, while a shared deep compression layer is equipped with a Low-Rank squeeze module and a cross-attention module for image and text, so as to establish a strong association between the vision branch and the language branch.

Shallow Squeezing. We observe that mask-based training, by randomly dropping a portion of token features, compels the model to reconstruct the full representation from the remaining information. This regularizes token interactions and prevents alignment from being dominated by a small subset of tokens, thereby improving robustness and generalization. In the shallow squeezing module, we first generate a Bernoulli-distributed random mask along the sequence dimension to perform subsampling and sparsification of the input during training. We predefine a $mask\_ratio \in (0, 1)$, and the retention probability:

$$\rho = 1 - mask\_ratio. \tag{6}$$

We then sample each position j of the image features and text features:

$$x_i^j \sim Bernoulli(\rho)\ x_i \in E_i,\ j \in d_v, \tag{7}$$

$$t_i^j \sim Bernoulli(\rho)\ t_i \in T_i,\ j \in d_l. \tag{8}$$

Deep Squeezing. We adjust the dimensionality after shallow squeezing using a fully connected layer, reducing it to dimension m, which then serves as the input to the shared low-rank squeezing module. The deep squeezing module is designed as a shared low-rank transformation that performs semantic reconstruction and alignment on multi-modal features within the bottleneck space. Following the low-rank layer, we obtain:

$$x_i = Dropout(\phi(x_i A)B),\ x_i \in E_i, \tag{9}$$

$$t_i = Dropout(\phi(t_i A)B),\ t_i \in T_i, \tag{10}$$

where $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times m}$, and ϕ denotes an element-wise non-linear activation(RELU in our implementation). Thu we enforce all deep transformations to be performed within a dimension-constrained subspace.

To facilitate interaction text features to "attend to" image features and image features to "attend to" text features, this approach enables the explicit modeling of cross-modal dependencies, as opposed to each modality performing self-attention independently, we design a cross-attention module at the end of the deep squeezing stage. After passing through the low-rank layer, we obtain $\hat{E}_i = [\hat{x_i^1}, \hat{x_i^2}, \cdots, \hat{x_i^M}] \in \mathbb{R}^{M\times d_v}$ and $\hat{T}_i = [\hat{t_i^1}, \hat{t_i^2}, \cdots, \hat{t_i^N}] \in \mathbb{R}^{N\times d_l}$,

$$\hat{x_i} = CrossAttn(Q = x_i, K = t_i, V = t_i),\ x_i \in E_i,\ t_i \in T_i, \tag{11}$$

$$\hat{t_i} = CrossAttn(Q = t_i, K = x_i, V = x_i),\ x_i \in E_i,\ t_i \in T_i. \tag{12}$$

To connect the two deeply compressed components, we use residual connections to aggregate these two parts into the output of the original CLIP encoder:

$$[c_j, E_j] = \mathcal{V}_j([c_{j-1}, E_{j-1}]) + \alpha(\Phi[c_{j-1}, E_{j-1}]) + \beta(\Phi[c_{j-1}, \hat{E}_{j-1}]),\ j = k, k+1, ..., L, \tag{13}$$

$$[T_j] = \mathcal{T}_j([T_j]) + \alpha(\Phi[T_{j-1}]) + \beta(\Phi[\hat{T}_{j-1}]),\ j = k, k+1, ..., L, \tag{14}$$

where α and β denote weight coefficients, Φ denotes exploding.

After the final layer K of the CLIP encoder, we obtain $x \in \mathbb{R}^{d_v}$ and $z \in \mathbb{R}^{d_l}$ via Eq. (3) and Eq. (4) and then follow CLIP's classification rule Eq. (5).

4 Experiments

4.1 Experiments Setup

Tasks. *Base-to-Novel Generalization.* For this evaluation, we split the dataset classes evenly into base and novel classes. The model is trained on the base classes under a 16-shot setting, and its performance is evaluated on both base and novel classes, with the novel classes assessed in a zero-shot setting.

Cross-Dataset Evaluation. In this setting, we fine-tune the model on the ImageNet source domain with all 1000 classes and evaluate it on the other ten datasets. Following CoCoOp [41], our model is trained under a 16-shot setting.

Domain Generalization. To assess the out-of-distribution transferability of our method, we perform an evaluation following the cross-dataset setting, where the model is trained on ImageNet and evaluated on four ImageNet dataset variants.

Few-Shot Learning. To evaluate the model's performance in scenarios with limited labeled data, we train the model with 1, 2, 4, 8, and 16 shots per class, and then evaluate it on a test set that includes all classes.

Datasets. For base-to-novel generalization, we consider a suite of 11 datasets used in CoOp [42]. Specifically, these include ImageNet [6], Caltech101 (CAL) [9], Oxford Pets (PETS) [25], Stanford Cars (CARS) [19], Oxford Flowers 102 (FLWR) [24], Food-101 (FOOD) [2], FGVC Aircraft (AIR) [23], SUN39 [34], Describable Textures (DTD) [5], EuroSAT (ESAT) [13], and UCF-101 (UCF) [29], and we use the splits of [42]. For cross-dataset evaluation, we fine-tune the model on the ImageNet [6] dataset and evaluate it on the other ten datasets mentioned above as target datasets. For domain generalization, we adopt ImageNet as the source dataset and four variant datasets as the target datasets, including ImageNet-V2 [27], ImageNet-Sketch [32], ImageNet-A [15] and ImageNet-R [14].

Training Details. We use pre-trained ViT-B/16 CLIP model [7] as the backbone and set the $mask\_ratio = 0.10$, $m = 128$, $r = 16$, $k = 1$, $\alpha = 0.05$, $\beta = 0.01$. For base-to-novel evaluation, the training is conducted for 20 epochs with a batch size of 16 (10 epochs and a batch size of 32 are used for the ImageNet dataset) and an initial learning rate of 0.02 using the SGD optimizer. We report accuracies for the base and novel classes, along with their harmonic mean (HM), averaged over three standard random seeds (1, 2, and 3). For cross-dataset and domain generalization evaluation, the model is trained on 5 epochs with an initial learning rate of 0.005 using the SGD optimizer. For few-shot learning, we train the model for 30 epochs with a learning rate of 0.02 (10 epochs are used for the ImageNet dataset) using the SGD optimizer. All experiments are conducted on an RTX 5090 GPU.

4.2 Quantitative Performance on Various Tasks

Base-to-Novel Generalization. In Tab. 1, we conduct a comprehensive evaluation of our method against state-of-the-art (SOTA) approaches on the base-

Table 1: Comparison with state-of-the-art methods on base-to-novel generalization. "Base" and "Novel" are the recognition accuracies for the base class and novel class. "HM" is the harmonic mean of base and new accuracy, providing the trade-off between adaptation and generalization. SDSA achieves strong generalization over existing methods across 11 recognition datasets. The best results are labeled in bold.

Methods	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [26]	69.34	74.22	71.70	72.43	68.16	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp [42]	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp [41]	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
KgCoOp [36]	80.73	73.60	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
MaPLe [17]	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
PromptSRC [18]	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
CoPrompt [28]	84.00	77.23	80.48	77.67	**71.27**	**74.33**	98.27	94.90	96.55	95.67	**98.10**	**96.87**
MMA [35]	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	98.07	96.72
2SFS [8]	**85.55**	75.48	80.20	**77.71**	70.99	74.20	**98.71**	94.43	96.52	95.32	97.82	96.55
SDSA *[Ours]*	83.77	**78.34**	**80.96**	77.53	71.13	74.20	98.27	**95.80**	**97.02**	**95.77**	96.47	96.11

Methods	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [26]	63.37	74.89	68.65	72.08	**77.80**	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp [42]	78.12	60.40	68.13	97.60	59.67	76.06	88.33	82.26	85.19	40.44	22.30	28.75
CoC0Op [41]	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
KgCoOp [36]	71.76	**75.04**	73.36	95.00	74.73	83.65	90.50	91.70	91.09	36.21	33.55	34.83
MaPLe [17]	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	**91.38**	37.44	35.61	36.50
PromptSRC [18]	78.27	74.97	76.58	98.07	76.50	82.56	90.67	91.53	91.10	42.73	37.87	40.15
CoPrompt [28]	76.97	74.40	75.66	97.27	76.60	85.71	**90.73**	**92.07**	91.40	40.20	39.33	39.76
MMA [35]	78.50	73.10	75.70	97.75	77.53	85.48	90.13	91.30	90.71	40.57	36.33	38.33
2SFS [8]	**82.50**	74.80	**78.46**	**98.29**	76.17	85.83	89.11	91.34	90.21	**47.48**	35.51	40.63
SDSA *[Ours]*	79.43	74.97	77.14	97.80	77.33	**86.37**	90.50	91.30	90.90	42.97	**45.13**	**44.02**

Methods	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [26]	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp [42]	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp [41]	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
KgCoOp [36]	80.29	76.53	78.36	77.55	54.99	64.55	83.64	64.34	73.48	82.89	76.67	79.65
MaPLe [17]	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
PromptSRC [18]	**82.67**	78.47	80.52	**83.37**	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
CoPrompt [28]	82.63	**80.03**	**81.31**	83.13	64.73	72.79	94.60	78.57	85.84	86.90	79.57	83.07
MMA [35]	82.27	78.57	80.38	83.20	**65.63**	73.38	85.46	82.34	83.87	86.23	80.03	82.20
2SFS [8]	82.59	78.91	80.70	84.60	65.01	**73.52**	**96.91**	67.09	79.29	**87.85**	78.19	82.74
SDSA *[Ours]*	82.33	78.67	80.46	81.67	63.77	71.60	88.43	**86.33**	**87.37**	86.73	**80.87**	**83.67**

to-novel task, including CLIP [26], CoOp [42], CoCoOp [41], KgCoOp [36], MaPLe [17], PromptSRC [18], CoPrompt [28], MMA [35] and 2SFS [8]. The average metrics across all 11 datasets are presented in the first column of the Tab. 1. Our method achieves highly competitive performance on this task, achieving strong results in novel accuracy and harmonic mean (HM) accuracy across all evaluated datasets. Even when our method does not achieve optimal perfor-

Table 2: Performance comparison on cross-dataset evaluation.

Methods	Source	Target										
	Imagenet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF101	Avg.
CoOp [42]	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp [41]	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLE [17]	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
PromptSRC [18]	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
MMA [35]	71.00	93.80	90.30	66.13	72.07	86.12	25.33	**68.17**	46.57	**49.24**	68.32	66.61
SDSA*[Ours]*	70.77	93.43	**92.17**	**66.30**	**73.37**	**86.47**	**30.33**	67.43	**47.77**	45.47	**70.43**	**67.32**

Table 3: Performance comparison on domain generalization.

Methods	Source	Target			
	ImageNet	-V2	-Sketch	-A	-R
CoOP [42]	**71.51**	64.20	47.99	49.71	75.21
MaPLe [17]	70.72	64.07	49.15	50.90	76.98
CoPrompt [28]	70.80	64.25	**49.43**	50.50	77.51
MMA [35]	71.00	**64.33**	49.13	51.12	76.98
SDSA *[Ours]*	70.77	64.17	49.37	**51.33**	**77.57**

mance on individual datasets, the performance gap between it and the top-performing approach is minimal, underscoring its overall robustness. Most notably, our method excels on the novel class with an average score of 78.34%. This demonstrates that SDSA's integrated masking and feature extraction capabilities effectively mitigate noise associated with base classes during training, thereby significantly boosting the model's generalization performance on unseen novel classes and driving this superior novel-class accuracy.

Cross-Dataset Evaluation. Tab. 2 presents the cross-dataset evaluation results. SDSA attains the highest average accuracy of 67.32%, outperforming all competing approaches. Notably, it achieves the best performance on seven target datasets and, even when not ranked first, consistently remains among the top-performing methods. These results highlight the strong zero-shot transferability of SDSA.

Domain Generalization. In Tab. 3, we report the results of the domain generalization task. SDSA achieves the best performance on the -A and -R target domains, while delivering competitive results on -V2 and -Sketch. These findings demonstrate that SDSA offers domain generalization performance on par with SOTA methods.

Few-Shot Learning. Tab. 4 presents the few-shot learning results under 1, 2, 4, 8, and 16-shot settings. The results show that our model achieves accuracy comparable to the current SOTA methods. From these observations, we can conclude that our method performs well even under limited sample conditions.

Table 4: Performance comparison on few-shot setting on the ViT-B/16 backbone.

Shots	Method	Imagenet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF101
1-shot	CoOp [42]	65.7	92.5	90.2	67.5	78.3	84.3	20.8	67.0	50.1	56.4	71.2
	MaPLe [17]	69.7	93.6	91.4	67.6	74.9	85.4	28.1	**69.3**	50.0	29.1	71.1
	PLOT [3]	66.5	**94.3**	**91.9**	68.0	80.5	86.2	28.6	66.8	**54.6**	**65.4**	**74.3**
	MMA [35]	69.2	92.9	91.2	67.9	**83.6**	83.0	28.7	64.0	52.3	55.0	74.2
	SDSA/*Ours*/	**69.9**	93.5	90.9	**69.9**	79.2	**86.0**	**32.6**	69.1	48.6	58.7	73.2
2-shot	CoOp	67.0	93.1	89.9	70.4	88.0	84.4	25.9	67.0	54.1	56.1	74.1
	MaPLe	70.0	**94.9**	91.8	68.5	79.8	86.5	29.5	70.7	50.6	59.4	74.0
	PLOT	68.3	94.7	**92.3**	**73.2**	89.8	86.3	31.1	68.1	**56.7**	**76.8**	76.8
	MMA	**70.4**	94.0	92.0	71.8	**90.3**	83.0	28.7	64.0	52.3	55.0	74.2
	SDSA/*Ours*/	**70.4**	94.4	91.3	72.8	87.4	**86.6**	**34.3**	**71.0**	56.1	58.7	**77.1**
4-shot	CoOp	68.8	94.5	92.5	74.4	92.2	84.5	30.9	69.7	59.5	69.7	77.6
	MaPLe	70.6	95.0	**93.3**	70.1	84.9	86.7	30.1	71.4	59.0	69.9	77.1
	PLOT	70.4	95.1	92.6	76.3	92.9	**86.5**	35.3	71.7	62.4	**83.2**	79.8
	MMA	71.0	94.3	92.2	**76.5**	**93.0**	82.1	37.6	70.0	**63.9**	79.4	80.1
	SDSA/*Ours*/	**71.2**	**95.2**	93.0	76.2	92.9	86.4	**38.0**	**72.8**	61.5	66.7	**80.3**
8-shot	CoOp	70.6	94.5	91.3	79.0	94.9	82.7	38.5	71.9	64.8	77.1	80.0
	MaPLe	71.3	95.1	93.1	71.3	90.5	**87.2**	33.8	73.2	63.0	82.8	79.5
	PLOT	71.3	**95.5**	93.0	81.3	95.4	86.6	41.4	73.9	66.5	**88.4**	82.8
	MMA	71.8	95.4	92.8	**81.4**	**96.0**	83.0	**44.8**	72.3	**68.0**	86.5	**83.4**
	SDSA/*Ours*/	**72.1**	**95.5**	**93.4**	80.2	95.4	86.5	43.7	**74.9**	66.5	81.1	83.2
16-shot	CoOp	**71.9**	95.8	92.0	82.9	96.8	84.2	43.2	74.9	69.7	85.0	83.1
	MaPLe	**71.9**	95.4	93.2	74.3	94.2	**87.4**	36.8	74.5	68.4	87.5	81.4
	PLOT	72.6	96.0	93.6	84.6	97.6	87.1	46.7	76.0	71.4	92.0	85.3
	MMA	73.1	**96.3**	93.2	**85.7**	**98.0**	84.6	**52.7**	74.6	**73.5**	**92.4**	**86.3**
	SDSA/*Ours*/	73.2	96.1	**94.0**	84.2	97.5	87.0	50.7	**76.7**	72.2	89.9	85.3

As the number of shots increases, SDSA extracts class features more accurately, thus amplifying the benefits of using learnable adapters.

4.3 Ablation Studies

Impact of SDSA and Scaling Factor α and β. We conduct ablation experiments on two key components of our SDSA: α denotes the ablation on the low-rank transformation, and β refers to the ablation of the cross-attention. The ablation results are presented in Tab. 5(a). Clearly, using both modules yields the best performance across all 11 datasets for Base, Novel, and HM, indicating that the combination of low-rank constraint and cross-modal refinement is essential for effective alignment. The scaling factor plays a critical role in balancing the importance of general and task-specific features. We systematically evaluated the impact of the scaling factor, and the results are presented in Tab. 5(b) and Tab. 5(c). For the scaling factor α, a larger value improves the performance in the base classes, while the performance of the novel-class peaks at $\alpha = 0.05$ and then decreases slightly. For the scaling factor β, a larger value improves the performance of the base-class but consistently degrades the performance of the novel classes; conversely, a smaller β favors the novel classes. To facilitate rapid adaptation to novel classes downstream tasks, we adopted $\alpha = 0.05$ and $\beta = 0.01$ as the hyperparameters for SDSA.

Table 5: Ablation on SDSA and scaling factors α and β ($\beta = 0.01$ for (b)α evaluation, $\alpha = 0.05$ for (c)β evaluation).

(a)Impact of SDSA

α	β	Base	Novel	HM
		68.85	75.15	71.86
✓		83.25	77.83	80.45
	✓	72.82	76.31	74.52
✓	✓	**83.77**	**78.34**	**80.96**

(b) Scaling Factor α

α	Base	Novel	HM
0.01	76.63	77.57	77.10
0.05	83.77	**78.34**	80.96
0.10	**84.82**	77.77	**81.14**

(c) Scaling Factor β

β	Base	Novel	HM
0.01	83.77	**78.34**	**80.96**
0.05	84.00	77.74	80.75
0.10	**84.06**	76.54	80.12

Impact of Using SDSA in Different Layers. Tab. 6 shows the performance impact of using SDSA in different layers. Each row in the table corresponds to a configuration where the number of layers applying the module is gradually incorporated, starting from layers 1, 3, 6, and 9, up to the final layer. Classification performance consistently improves as more layers apply the module. The model achieves the highest accuracy when the SDSA module is applied to all layers, as shown in the last row (1→12) of Tab. 6, corresponding to the SDSA configuration.

Table 6: Impact of SDSA in different layers.

Layers	Base	Novel	HM	Trainable
9→12	78.65	76.34	77.48	1.60M
6→12	82.11	77.40	79.69	2.80M
3→12	83.22	78.00	80.53	4.00M
1→12	**83.77**	**78.34**	**80.96**	4.79M

Table 7: Impact of rank in low-rank approximation.

r	Base	Novel	HM	Trainable
8	82.94	78.22	80.51	4.77M
16	83.77	**78.34**	80.96	4.79M
32	84.31	78.14	**81.11**	4.84M
64	**84.69**	77.68	81.03	4.95M

Impact of Rank in Low-Rank Approximation. Tab. 7 illustrates how the rank r in the low-rank approximation affects performance. A smaller rank reduces the number of parameters but limits the model's capacity, negatively impacting both base and novel accuracy. In contrast, a larger rank enhances base performance but may hinder generalization to novel classes. An intermediate rank achieves the strongest novel-class performance while maintaining competitive base accuracy and model compactness. This behavior suggests that cross-modal alignment requires a controlled shared subspace. Excessively high rank expands the alignment space to unstable minor directions that often encode dataset-specific co-occurrences, whereas overly low rank over-compresses dominant semantic axes and restricts expressiveness. We choose rank 16 as the default choice, as it exhibits stronger generalization ability.

Table 8: Impact of mask ratio in random masking.

ratio	Base	Novel	HM
0.00	**83.83**	78.18	80.91
0.05	83.78	78.28	80.94
0.10	83.77	**78.34**	**80.96**
0.15	83.67	78.30	80.90
0.30	83.22	78.20	80.63

Table 9: Comparison of model total and learnable parameters.

Methods	Adapter	Total	Trainable
CoOp [42]	None	124.33M	0.04M
TCP [37]	Shared	149.92M	0.33M
MMA [35]	Shared	125.00M	0.67M
SDSA*[Ours]*	Shared	129.12M	4.79M

Impact of Mask Ratio in Random Masking. The mask ratio determines the sparsity level imposed during shallow squeezing. As shown in Tab. 8, performance varies with different masking strengths. Without masking, interactions remain fully dense. Introducing a moderate masking ratio ($mask\_ratio = 0.10$) yields the best Novel and HM results, indicating that controlled sparsification improves generalization by regulating interaction density while retaining principal semantic signals.

4.4 Computational Complexity

We compare the total and learnable parameters of several methods using the ViT-B/16 backbone. Tab. 9 demonstrates that our SDSA shared adapter has only a minor difference in total and trainable parameters compared with existing shared adapter methods. Removing SDSA from certain layers (as shown in Tab. 6) can further reduce the computational cost, albeit with a slight drop in performance.

4.5 Qualitative Performance

Fig. 1(a)-(b) show the text–image alignment heatmaps on the EuroSAT dataset. It can be observed that SDSA achieves significantly better text–image matching performance than MMA.In comparison with the baseline MMA, our approach yields logit distributions more sharply concentrated along the diagonal, with higher diagonal values and lower off-diagonal values.This pattern quantitatively demonstrates stronger text–image alignment and lower inter-class confusion, indicating that images from each category are more reliably matched to their corresponding textual labels.

Fig. 4 shows the t-SNE visualization of image features for both our method and MMA, with MMA selected due to its similarity in type to ours. On the EuroSAT dataset, our method achieves strong intra-class compactness and clear inter-class separability for the base and novel classes, while also exhibiting fewer outliers. On the Caltech101 dataset, even with an increasing number of categories, our method maintains superior inter-class separation compared to MMA. While there are a few outliers in the plots, they are more concentrated within

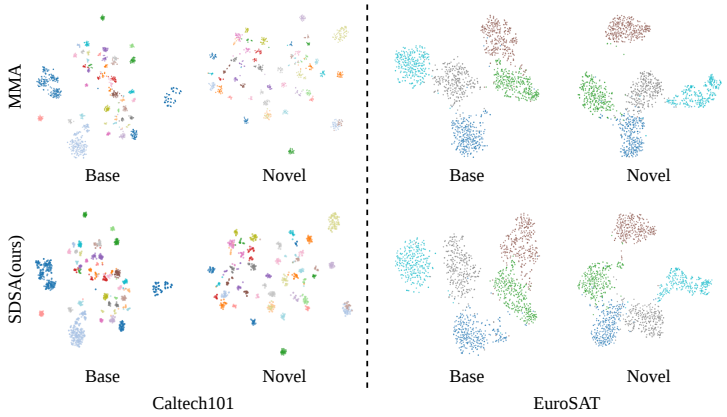


Fig. 4: t-SNE plots of image embeddings generated by MMA and our method (SDSA) on Caltech101 and EuroSAT datasets. SDSA shows better separation in both base and novel classes.

their respective categories. Notably, our method demonstrates significantly improved separation for novel categories. We attribute this improvement to the Shallow-Deep Squeezing Adapter: within the bottleneck space of the adapter, our model more effectively extracts discriminative features while regulating cross-modal interactions and alignment capacity, thereby enhancing intra-class compactness and inter-class separability.

5 Conclusion

In this paper, we propose the Shallow-Deep Squeezing Adapter (SDSA), a hierarchical alignment regulation framework for vision-language models. Unlike conventional squeezing-expanding adapters that primarily enhance adaptation flexibility, SDSA explicitly controls cross-modal interaction density and alignment capacity through a shallow–deep design. Specifically, we insert lightweight adapters into both the text and visual encoders of CLIP, and regulate cross-modal alignment via token-level masking, shared low-rank transformation, and cross-attention module. This design restricts alignment to a compact set of dominant shared directions while preventing unstable coupling, leading to more robust generalization. Extensive experiments on 11 widely used image classification datasets demonstrate that SDSA achieves superior base-to-novel generalization and strong cross-dataset transfer performance. Nevertheless, performance remains sensitive in extremely few-shot scenarios, highlighting the need for more robust adaptation strategies for vision-language models under severely limited supervision.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant 62225601, U23B2052), the Scientific Research Innovation Capability Support Project for Young Faculty under Grant SRICSPYF-BS2025009, the Beijing Natural Science Foundation Project No. L242025, the Gansu Province Top Leading Talents Foundation, the Gansu Provincial Key Talent Project under Grant 2025RCXM002, the Guizhou Province Science and Technology Plan Project (No. QianKeHeZhongDa [2025]031), the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance, and the Royal Society under International Exchanges Award IEC/NSFC/201071.

References

1. Bordes, F., Pang, R.Y., Ajay, A., Li, A.C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al.: An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247 (2024) 1
2. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014) 8
3. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: Prompt learning with optimal transport for vision-language models. In: International Conference on Learning Representations. pp. 4372–4392 (2023) 11
4. Chen, J., Zhou, Z., Tong, Y., Chang, D., Luo, Y., Ma, Z.: Seeing as experts do: A knowledge-augmented agent for open-set fine-grained visual understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41446–41455 (2026) 1
5. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014) 8
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 8
7. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 8
8. Farina, M., Mancini, M., Iacca, G., Ricci, E.: Rethinking few-shot adaptation of vision-language models in two stages. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29989–29998 (2025) 9
9. Fei-Fei, L., Fergus, R., Perona, P.: 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. pp. 178–178. IEEE (2004) 8
10. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: European conference on computer vision. pp. 701–717. Springer (2022) 3
11. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024) 1, 3, 4

12. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021) 3
13. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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) 8
14. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021) 8
15. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15262–15271 (2021) 8
16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021) 3
17. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19113–19122 (2023) 1, 3, 4, 6, 9, 10, 11
18. Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15190–15200 (2023) 1, 4, 9, 10
19. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013) 8
20. Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., Yang, J.: Promptkd: Unsupervised prompt distillation for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26617–26626 (2024) 4
21. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7061–7070 (2023) 3
22. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024) 6
23. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013) 8
24. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008) 8
25. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012) 8
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 1, 3, 4, 9

27. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019) 8
28. Roy, S., Etemad, A.: Consistency-guided prompt learning for vision-language models. arXiv preprint arXiv:2306.01195 (2023) 4, 6, 9, 10
29. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 8
30. Tong, Y., Chang, D., Yin, Z., Liu, X., Fang, Y., Ma, Z.: Reversing the flow: Generation-to-understanding synergy in large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings. pp. 6976–6986 (June 2026) 1
31. Vidit, V., Engilberge, M., Salzmann, M.: Clip the gap: A single domain generalization approach for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3219–3229 (2023) 3
32. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in neural information processing systems **32** (2019) 8
33. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022) 3
34. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 3485–3492. IEEE (2010) 8
35. Yang, L., Zhang, R.Y., Wang, Y., Xie, X.: Mma: Multi-modal adapter for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23826–23837 (2024) 1, 2, 3, 4, 6, 9, 10, 11, 13
36. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6757–6767 (2023) 9
37. Yao, H., Zhang, R., Xu, C.: Tcp: Textual-based class-aware prompt tuning for visual-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23438–23448 (2024) 13
38. Zanella, M., Ben Ayed, I.: Low-rank few-shot adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1593–1603 (2024) 4
39. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021) 1, 4
40. Zhang, Z., Yang, X., Li, X., Ma, Z., Xue, J.H.: Class-aware cross-attention helps prompt learning with limited samples. IEEE Transactions on Multimedia pp. 1–12 (2026) 3
41. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022) 1, 2, 4, 8, 9, 10
42. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) 1, 2, 4, 8, 9, 10, 11, 13