Fine-Tuning via Linked Domains: A Closed-Form Dual Alignment Mechanism for Transferring Vision-Language Models

Peiyu Lu, Xiaoxu Li, *Member, IEEE*, Rui Zhu, Zhanyu Ma, *Senior Member, IEEE*, Jie Cao, Jing-Hao Xue, *Senior Member, IEEE*

Abstract*—Adapters and prompt learning have become two de facto strategies to fine-tune pre-trained vision-language models, mitigating the high computational cost of fine-tuning an entire model for downstream tasks. They can align the prediction from the fine-tuned model with that from the pre-trained model. However, the existing methods of these strategies primarily focus on aligning within a single modality, and the exploration of bidirectional interactions between modalities remains limited. To address this issue, we propose a closed-form dual alignment mechanism (DAM) that** *not only ensures the consistency in predictions within a single modality but also achieves the alignment of features across different modalities. In DAM, all alignments are achieved by closed-form solutions to ridge regression, without inducing a massive number of learnable parameters. Experimental results demonstrate that DAM outperforms the state-of-the-art methods on 11 benchmarks over various evaluation metrics. Our codes are available at https://github.com/Peiy-Lu/DAM.**

***Index Terms*—Vision-language model, Fine-tuning, Feature alignment.**

I. Introduction

THERE is a growing interest recently in efficiently fine-tuning pre-trained large-scale vision-language models (VLMs) [1] for downstream tasks, driven by an aim to mitigate the high computational cost associated with tuning an entire VLM.

Adapters [2]–[4] and prompt learning [5]–[7] are two de facto strategies for fine-tuning VLMs, both of which introduce only a small amount of parameters to different parts of the models while keeping pre-trained weights frozen, thereby reducing the computational cost. Adapters inserts learnable weights inside of the model, typically after the frozen pre-trained encoders, while prompt learning methods incorporate learnable vectors to the input space, replacing the hand-crafted prompt.

However, among the existing methods of these strategies, the exploration of bidirectional interactions between text and image modalities for fine-tuning VLMs remains limited. To learn the additional weights, some methods focus on ensuring the consistency between the prediction from the fine-tuned model with that from the pre-trained model [5], [8], [9], but primarily aligning within a single modality and neglecting the interaction between modalities. To resolve this problem, MaPLe [6] proposes to couple prompts between visual and textual branches, where the visual prompt is obtained by projecting the textual prompt through a linear layer; on the contrary, CoCoOp [5] and LFA [7] only consider the interaction from image to text, overlooking the interaction from text to image. On the other hand, CALIP [10] enhances bidirectional modality alignment by updating encoded features simply based on the similarity between intermediate visual and textual features in a parametric-free approach. MMA [11] proposes a multi-modal adapter which shares weights between image and text branches to learn shared representations of both modalities.

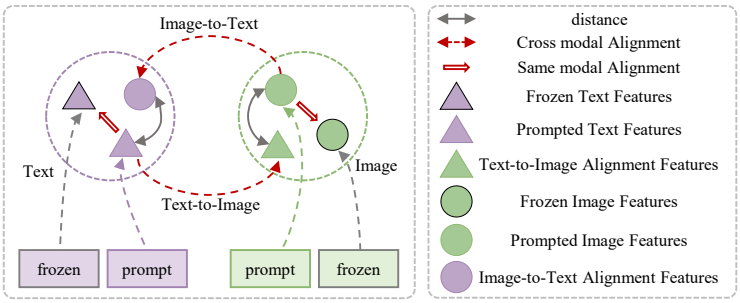


Fig. 1. The motivation of dual alignment mechanism (DAM). The frozen pre-trained features and the prompt features are aligned by same modal alignment (SMA) within same modalities. The prompt features of text and image modalities are then aligned through the bidirectional cross modal alignment (CMA).

In this paper, we propose a closed-form **D**ual **A**lignment **M**echanism (DAM) for transferring VLMs to downstream tasks, *featuring a bidirectional transformation between text and image modalities through ridge regression*. DAM integrates two key components: same-modal alignment (SMA) and cross-modal alignment (CMA). This design addresses the challenge of insufficient cross-modal information transfer while mitigating the risk of overfitting.

As illustrated in Figure 1, SMA aligns learnable prompts with pre-trained features within the same modality, using ridge regression with a closed-form solution. This approach requires only a very small number of learnable parameters to achieve

Corresponding author: Xiaoxu Li (lixiaoxu@lut.edu.cn), Rui Zhu (email: rui.zhu@city.ac.uk)

P. Lu, X. Li, J. Cao are with the School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China.

R. Zhu is with the Faculty of Actuarial Science and Insurance, Bayes Business School, City St George's, University of London EC1Y 8TZ, UK.

Z. Ma is with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China.

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, U.K.

performance comparable to adapter-based methods such as CoPrompt [12]. Ridge regression enables a balance between model complexity and feature alignment accuracy, ensuring stable performance across different tasks. Additionally, we introduce a same-modal alignment loss to further enhance consistency.

CMA, on the other hand, performs bidirectional alignments between text and image modalities, leveraging ridge regression for efficient computation. By aligning text-to-image and image-to-text features, CMA effectively captures intrinsic relationships between modalities, thereby enhancing generalization performance on downstream tasks. The residuals from the ridge regressions are used to compute the cross-alignment distance, from which the cross-entropy loss is calculated.

In short, our main contributions are as follows:

- We introduce a dual alignment mechanism (DAM) that guarantees both the same-modal prediction consistency and the cross-modal alignment.
- We develop a bidirectional cross-alignment scheme that seamlessly connects text and image modalities through both text-to-image and image-to-text projections.
- We implement all alignments using ridge regression with closed-form solutions, ensuring stable and efficient feature transformation.

II. Related Work

A. VLMs

VLMs have made significant strides in recent years, demonstrating their capabilities on zero-shot tasks [13]–[15]. CLIP [1] showcases remarkable zero-shot generalization capabilities by training on 400 million caption-image pairs collected from the internet. Researchers have since proposed various strategies to extend and optimize this model. SigLIP [16] achieves superior zero-shot performance with smaller batch sizes compared to CLIP. Llip [17] explores the phenomenon where images can be described in multiple ways and suggests using a cross-attention module to modulate image encoding based on target captions.

B. Transfer Learning for VLMs

Due to the enormous number of parameters in VLMs, transferring them to downstream tasks becomes extremely challenging. Fine-tuning techniques, which have proven effective for traditional deep neural networks [18], offer a promising approach for optimizing VLMs for specific applications. Previous studies attempted to fine-tune the entire VLM on downstream tasks, but this often led to a high computational cost and a decline in the model's generalization ability. To address this issue, researchers have proposed two fine-tuning paradigms that adjust downstream tasks without altering the pre-trained model parameters. First, prompt learning-based methods [5], [8] add a set of continuous learnable vectors as input prompt to achieve more accurate representations. CoOp [8] is the pioneer to learn the prompts. HPT [19] combines the structured and traditional language knowledge of large language models (LLMs) to enhance prompt effectiveness in a hierarchical manner. AAPL [20] introduces adversarial prompt embeddings by incorporating biases into the learnable prompts, distinguishing between low-level visual features and high-level class information, thus addressing the issue of traditional prompt learning being biased towards visible classes and negatively impacting unseen classes. Additionally, DECOOP [21] explores open-world prompt tuning via the decomposed prompt tuning (DePT) [22] framework and introduces new-class detectors to boost base-to-new discriminability, while KDPL [23] improves prompt generalization through unsupervised knowledge distillation from stronger models without requiring labeled data. Second, adapter-based methods [2] adapt downstream tasks by adding additional learnable parameters after the feature encoders. However, these methods only consider aligning the predictions from the fine-tuned model with that from the pre-trained model, neglecting the information flow between modalities.

MaPLe [6] projects the textual prompt through a linear layer to obtain the visual prompt, only considering text-to-image projection. Methods such as CoPrompt [12] follows the same approach as MaPLe to align modalities, however it involves an implicit consistency regularization in the loss function to achieve the alignment. CoCoOp [5] and LFA [7] only consider the interaction from image to text, overlooking the interaction from text to image. CALIP [10] updates the encoded features bidirectionally via a simple similarity between intermediate visual and textual features. Ma et al. [24] study the cause of the overfitting issue in CoOp and propose a Novel Feature Learner (NFL) to resolve this problem of the learned prompts. MMA [11] proposes a multi-modal adapter which shares weights between image and text branches to learn shared representations of both modalities. HIPL [25] proposes to explore hierarchical prompt learning method and cross-modal prompt interaction to ensure mutual synergy between visual features and text features. MAP [26] jointly explores textual and visual prompting at the attribute level.

Different from the above methods, we introduce DAM that explicitly considers both same-modal alignment and cross-modal alignment to enhance generalization performance. Ridge regression with closed-form solutions are adopted for all alignments to reduce the number of learnable parameters in the network and boost generalization.

III. Method

The overall workflow of DAM is illustrated in Figure 2. To mitigate overfitting to downstream tasks, we adopt input augmentation strategies similar to those used in HPT [19] and CoPrompt [12]. The detailed text descriptions and the original images are processed by learnable text and image encoders, while the individual tokens and the augmented images are processed through fixed, pre-trained encoders. We then apply same-modal alignment (SMA) to both branches to ensure consistent predictions between the pre-trained and learnable encoders within each modality. Finally, the bidirectional cross-modal alignment (CMA) is applied to the outputs from SMA.

A. Input Augmentation

For text augmentation, following HPT, we first generate descriptions with structured relationships from language in-

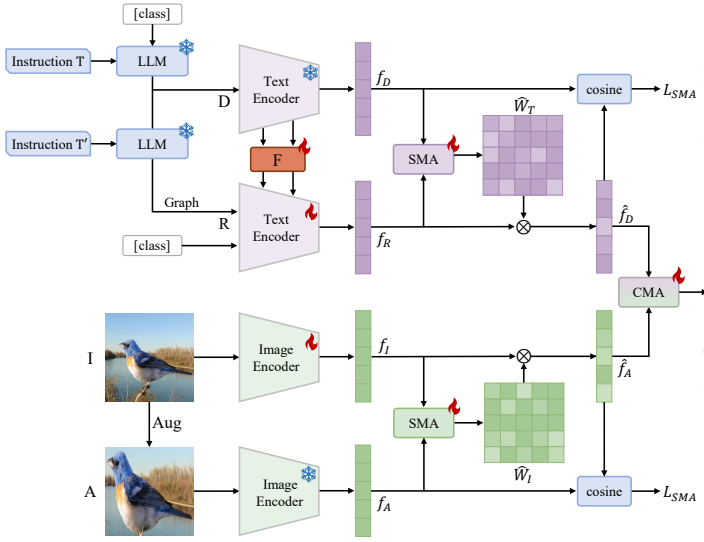


Fig. 2. The overall structure of DAM. To reduce overfitting, text and image inputs are augmented through a hierarchical approach in HPT and random resized cropping, respectively. Next, the inputs go through text and image encoders to obtain pre-trained text feature f_D, learnable text feature f_R, pre-trained image feature f_A and learnable image feature f_I. These features are then aligned through SMA within single modalities through alignment matrices $\hat{W}_T$ and $\hat{W}_I$. Finally, the outputs from SMA, text feature $\hat{f}_D$ and SMA image feature $\hat{f}_A$, are aligned through bidirectional CMA. The whole process is supervised by a linear combination of the SMA loss, L_{SMA}, and the cross-entropy (CE) loss, L_{CE}, calculated from the cross-alignment distance.

structions T containing [CLASS], resulting in $D = LLM(T)$. For the descriptions D, HPT designs an additional instruction T', using LLM for generating structured knowledge $R = LLM([T', LLM(T)])$. In Figure 2, the linear projection layer F between the two text encoders captures the association between individual tokens and the complete description. Then, we input D and R into the pre-trained text encoder and the learnable prompt text encoder, respectively, and obtain text features $f_D \in \mathbb{R}^{c\times 1}$ and $f_R \in \mathbb{R}^{c\times 1}$, where c is the dimension of features.

For image augmentation, we apply random resized cropping to generate the augmented image $A = \text{Aug}(I)$ from the original image I. We then feed the augmented image A and the original image I into the pre-trained image encoder and the learnable prompt image encoder, respectively, to obtain the image features $f_A \in \mathbb{R}^{c\times 1}$ and $f_I \in \mathbb{R}^{c\times 1}$.

B. Same-Modal Alignment (SMA)

After obtaining the text and image features, to prevent significant deviations of the learnable encoders from the pre-trained encoders, we perform SMA through closed-form ridge regression.

Our goal in SMA is to find alignment matrices $W_T \in \mathbb{R}^{c\times c}$ and $W_I \in \mathbb{R}^{c\times c}$ to align the features f_R or f_I to the features f_D or f_A, respectively, i.e., $W_T f_R = f_D$ and $W_I f_I = f_A$. Note that W_T and W_I are specific for each text description and image, respectively. For simplicity, we slightly abuse notation and omit the indices of text and image. We propose to obtain the alignment matrices via ridge regression, which includes an L_2 regularization to balance feature alignment accuracy and model complexity, thereby promoting generalization. The strength of the regularization is controlled by the learnable hyperparameter λ. The closed-form solutions can provide the global optimum directly, eliminating the risk of convergence to local minimum associated with iterative gradient descent.

Since the calculations are the same for both text and image alignment, we present them for text alignment below as an example:

$$W_T^* = \underset{W_T}{\text{argmin}} \|W_T f_R - f_D\|_2^2 + \lambda \|W_T\|_2^2. \tag{1}$$

The optimal alignment matrix W_T^* can be obtained through the following closed-form solution:

$$W_T^* = f_D f_R^T (f_R f_R^T + \lambda I)^{-1}, \tag{2}$$

where $I \in \mathbb{R}^{c\times c}$ is an identity matrix. Note that the above optimization problem is solved for each pair of pre-trained and learnable features; that is, we obtain N different W_T^*'s, where N is the number of classes.

Additionally, to prevent from excessively fitting to the learnable features, we perform an additional smoother operation as follows:

$$\hat{W}_T = \rho I + (1-\rho) W_T^*, \tag{3}$$

where alignment conversion ratio ρ is a learnable parameter. When ρ is small, the aligned features remain similar to the pre-trained features. Conversely, when ρ is large, the aligned features tend to learn more knowledge from fine-tuning.

Following the same procedure, we can find $\hat{W}_I$'s to align pairs of pre-trained and learnable image features:

$$\begin{aligned} \hat{W}_I &= \rho I + (1-\rho) W_I^* \\ &= \rho I + (1-\rho) f_A f_I^T \left(f_I f_I^T + \lambda I\right)^{-1}. \end{aligned} \tag{4}$$

Finally, we obtain the aligned features:

$$\begin{aligned} \hat{f}_D &= \hat{W}_T f_R, \\ \hat{f}_A &= \hat{W}_I f_I. \end{aligned} \tag{5}$$

To further enhance consistency, we introduce the following same-modal alignment loss:

$$\mathcal{L}_{\text{SMA}} = 2 - \text{sim}(\hat{f}_D, f_D) - \text{sim}(\hat{f}_A, f_A), \tag{6}$$

where $\text{sim}(\cdot)$ is the cosine similarity.

C. Cross-Modal Alignment (CMA)

To fine-tune VLMs, existing research primarily focuses on learning within a single modality or on unidirectional modality projections, often overlooking the bidirectional relationship between modalities. In our work, we propose CMA that aligns different modalities from both directions, i.e. image-to-text and text-to-image. Through this bidirectional alignment, we can more effectively capture the relationships between text and image features, thereby enhancing generalization.

Let's take the image-to-text alignment as an example. We aim to learn a cross alignment matrix $W_{IT} \in \mathbb{R}^{c\times c}$ to align the SMA image feature $\hat{f}_A$ with the SMA text features $\hat{f}_D$,

Algorithm 1 Dual Alignment Mechanism (DAM)

Input: Frozen text features $F_D \in \mathbb{R}^{N\times c}$,
Text prompt features $F_R \in \mathbb{R}^{N\times c}$,
Frozen image features $F_A \in \mathbb{R}^{B\times c}$,
Image prompt features $F_I \in \mathbb{R}^{B\times c}$
Output: Distance between image features and category text features $d \in \mathbb{R}^{B\times N}$

Initialize:
N: number of classes
B: number of training images
c: feature dimensionality
Same-modal alignment matrix:
$F_D \leftarrow F_D$.unsqeeze(2)
$F_A \leftarrow F_A$.unsqeeze(2)
$\hat{W}_T \leftarrow$ Text_SMA(F_R, F_D)
$\hat{W}_I \leftarrow$ Image_SMA(F_I, F_A)
Same-modal features alignment:
$\hat{F}_D \leftarrow F_R \times \hat{W}_T$
$\hat{F}_A \leftarrow F_I \times \hat{W}_I$
Cross-modal features alignment:
$d_{CMA} \leftarrow$ CMA$(\hat{F}_A, \hat{F}_D)$
Combining CMA and cosine distances:
$d \leftarrow \beta \text{sim}(\hat{F}_A, \hat{F}_D) + (1-\beta) d_{CMA}$
return d

such that $W_{IT}\hat{f}_A = \hat{f}_D$. This can be achieved by solving the following ridge regression:

$$W^*_{IT} = \underset{W_{IT}}{\text{argmin}} \|W_{IT}\hat{f}_A - \hat{f}_D\|_2^2 + \lambda \|W_{IT}\|_2^2. \tag{7}$$

The closed-form solution of W^*_{IT} can be obtained as follows:

$$W^*_{IT} = \hat{f}_D \hat{f}_A^T \left(\hat{f}_A \hat{f}_A^T + \lambda I\right)^{-1}. \tag{8}$$

Similarly, by swapping $\hat{f}_D$ and $\hat{f}_A$ in the above equation, we can derive the alignment matrix W^*_{TI} from text features $\hat{f}_D$ to image features $\hat{f}_A$:

$$W^*_{TI} = \hat{f}_A \hat{f}_D^T \left(\hat{f}_D \hat{f}_D^T + \lambda I\right)^{-1}. \tag{9}$$

As with SMA, W_{IT} and W_{TI} are specific to each text and image pair.

D. Final Loss

Based on the cross-alignment matrices, we compute the cross-alignment similarity:

$$\begin{aligned} d_{\text{CMA}} = \delta(\omega[-\alpha\|\hat{f}_D - \mu_{IT} W^*_{IT}\hat{f}_A\|_2^2 - \\ (1-\alpha)\|\hat{f}_A - \mu_{TI} W^*_{TI}\hat{f}_D\|_2^2]), \end{aligned} \tag{10}$$

where $\delta(\cdot)$ is the log softmax operation, μ and ω are learnable hyperparameters and $\alpha \in [0,1]$ is a learnable hyperparameter used to balance the two distances.

To obtain the final similarity, we propose to add d_{CMA} to the original cosine similarity, enriching the similarity by explicitly modelling the bidirectional relationship between modalities. The final similarity is calculated as follows:

$$d = \beta \text{sim}(\hat{f}_A, \hat{f}_D) + (1-\beta) d_{\text{CMA}}, \tag{11}$$

where $\beta \in [0,1]$ is a learnable hyperparameter used to balance the two measures.

Next, for the k-th input image I_k, the classification loss is obtained by computing its cross-entropy loss over N text categories:

$$\mathcal{L}_{\text{CE}} = -\log \frac{e^{d_{k,y_k}/\tau}}{\sum_{n=1}^{N} e^{d_{k,n}/\tau}}, \tag{12}$$

where y_k is the true label of I_k and τ is the temperature parameter.

Finally, the total loss is obtained by weighting the sum of the cross-entropy loss and the same-modal alignment loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \gamma \mathcal{L}_{\text{SMA}}, \tag{13}$$

where γ is a hyperparameter to balance the two losses.

The overall algorithm of DAM involving both SAM and CAM is summarized in Algorithm 1.

IV. Experiments

Following CoPrompt, we evaluate our method under three generalization settings: base-to-new generalization, cross-dataset evaluation and domain generalization. We compare the performance of DAM with the baseline zero-shot CLIP and seven state-of-the-art fine-tuning methods, including Co-CoOp, LFA, MaPle, PromptSRC, CoPrompt, HPT DeCoOp [21], 2SFS [27] and FCPrompt-G [28]. Comprehensive ablation studies are also conducted to demonstrate the effectiveness of each component in DAM.

A. Datasets

For base-to-new generalization and cross-dataset evaluation, we report results on 11 datasets used in CoOp [8]: ImageNet [29], Caltech [30], OxfordPets [31], StanfordCars [32], Flowers [33], Food101 [34], FGVCAircraft [35], EuroSAT [36], UCF101 [37], DTD [38], and SUN397 [39].

For domain generalization, we follow previous works [5], [8] and report classification results on four ImageNet variants: ImageNet-V2 [40], ImageNet-Sketch [41], ImageNet-A [42], and ImageNet-R [43].

B. Implementation Details

We use ViT-B/16 as the vision backbone and perform prompt tuning on the pre-trained CLIP [1] model. The optimization process employs the SGD optimiser with an initial learning rate of 0.0025 for base-to-new generalization tasks and 0.001 for other tasks. All text and image encoders are the same as those used in HPT. For the text branch, we follow HPT to generate descriptions and structured knowledge with LLM. For the γ parameter in Equation (13), we set it to 4.0 for FGVCAircraft, 6.0 for SUN397 and StanfordCars, 7.0 for OxfordPets, EuroSAT and UCF101, 8.0 for Caltech101, OxfordFlowers, Food101 and DTD, 9.0 for ImageNet.

In cross-dataset and domain generalization tasks, all categories from ImageNet are used during the training phase, leading to a tensor of $1000 \times c \times c$ containing W_T's, resulting in expensive computational cost. To address this, we calculate a single $W_T \in \mathbb{R}^{c\times c}$ across all text features during training and testing, thereby reducing the computational overhead.

TABLE I
BENCHMARK RESULTS ON DIFFERENT DATASETS. THE "BASE", "NEW", AND "HM" (HARMONIC MEAN) COLUMNS REPRESENT ACCURACIES ON BASE CLASSES, NEW CLASSES AND THEIR HARMONIC MEANS, RESPECTIVELY. DAM OUTPERFORMS ON 9 OUT OF 11 DATASETS IN HARMONIC MEAN, RESULTING IN 0.72% AND 0.97% AVERAGE IMPROVEMENT OVER TWO MAIN COMPETITORS, COPROMPT AND HPT, RESPECTIVELY.

(a) Average

	Base	Novel	HM
CLIP	69.34	74.22	71.70
Co-CoOp	80.47	71.69	75.83
LFA	83.62	74.56	78.83
MaPle	82.28	75.14	78.55
PromptSRC	84.26	76.10	79.97
CoPrompt	84.00	77.23	80.48
HPT	84.32	76.86	80.42
DECoOp	-	-	76.13
2SFS	**85.55**	75.48	80.20
FCPrompt-G	85.28	75.72	80.22
DAM	84.53	**78.12**	**81.20**

(b) ImageNet

	Base	Novel	HM
CLIP	72.43	68.14	70.22
Co-CoOp	75.98	70.43	73.10
LFA	76.89	69.36	72.93
MaPle	76.66	70.54	73.47
PromptSRC	77.60	70.73	74.01
CoPrompt	77.67	71.27	74.33
HPT	77.95	70.74	74.17
DECoOp	-	-	72.98
2SFS	77.71	70.99	74.20
FCPrompt-G	77.33	71.11	74.09
DAM	**78.00**	**71.53**	**74.63**

(c) Caltech101

	Base	Novel	HM
CLIP	96.84	94.00	95.40
Co-CoOp	97.96	93.81	95.84
LFA	98.41	93.93	96.13
MaPle	97.74	94.36	96.02
PromptSRC	98.10	94.03	96.02
CoPrompt	98.27	94.90	96.55
HPT	98.37	94.98	96.65
DECoOp	-	-	96.52
2SFS	**98.71**	94.43	96.52
FCPrompt-G	98.41	94.07	96.19
DAM	98.20	**96.13**	**97.15**

(d) OxfordPets

	Base	Novel	HM
CLIP	91.17	97.26	94.12
Co-CoOp	95.20	97.69	96.43
LFA	95.13	96.23	95.68
MaPle	95.43	97.76	96.58
PromptSRC	95.33	97.30	96.30
CoPrompt	95.67	98.10	96.87
HPT	95.78	97.65	96.71
DECoOp	-	-	95.27
2SFS	95.32	97.82	96.55
FCPrompt-G	**96.40**	97.89	**97.14**
DAM	95.97	**98.17**	97.06

(e) StanfordCars

	Base	Novel	HM
CLIP	63.37	74.89	68.65
Co-CoOp	70.49	73.59	72.01
LFA	76.32	74.88	75.59
MaPle	72.94	74.00	73.47
PromptSRC	78.27	74.97	76.58
CoPrompt	76.97	74.40	75.66
HPT	76.95	74.23	75.57
DECoOp	-	-	73.24
2SFS	**82.50**	74.80	**78.46**
FCPrompt-G	80.87	73.03	76.75
DAM	79.97	**75.30**	77.57

(f) Flowers102

	Base	Novel	HM
CLIP	72.08	77.80	74.83
Co-CoOp	94.87	71.75	81.71
LFA	97.34	75.44	85.00
MaPle	95.92	72.46	82.56
PromptSRC	98.07	76.50	85.95
CoPrompt	97.27	76.60	85.71
HPT	98.17	78.37	87.16
DECoOp	-	-	84.16
2SFS	98.29	76.17	85.83
FCPrompt-G	**98.32**	74.82	84.98
DAM	97.47	**80.30**	**88.06**

(g) Food101

	Base	Novel	HM
CLIP	90.10	91.22	90.66
Co-CoOp	90.70	91.29	90.99
LFA	90.52	91.48	91.00
MaPle	90.71	92.05	91.38
PromptSRC	90.67	91.53	91.10
CoPrompt	90.73	92.07	91.40
HPT	90.46	91.57	91.01
DECoOp	-	-	90.68
2SFS	89.11	91.34	90.21
FCPrompt-G	90.76	**92.19**	**91.47**
DAM	**90.97**	91.53	91.25

(h) FGVCAircraft

	Base	Novel	HM
CLIP	27.19	36.29	31.09
Co-CoOp	33.41	23.71	27.74
LFA	41.48	32.29	36.31
MaPle	37.44	35.61	36.50
PromptSRC	42.73	37.87	40.15
CoPrompt	40.20	**39.33**	39.76
HPT	42.68	38.13	**40.28**
DECoOp	-	-	31.44
2SFS	**47.48**	35.51	**40.63**
FCPrompt-G	43.68	36.97	40.05
DAM	41.93	38.40	40.09

(i) SUN397

	Base	Novel	HM
CLIP	69.36	75.35	72.23
Co-CoOp	79.74	76.86	78.27
LFA	82.13	77.20	79.59
MaPle	80.82	78.70	79.75
PromptSRC	82.67	78.47	80.52
CoPrompt	82.63	80.03	81.31
HPT	82.57	79.26	80.88
DECoOp	-	-	78.11
2SFS	82.59	78.91	80.70
FCPrompt-G	**82.93**	79.06	80.95
DAM	82.87	**80.03**	**81.43**

(j) DTD

	Base	Novel	HM
CLIP	53.24	59.90	56.37
Co-CoOp	77.01	56.00	64.85
LFA	81.29	60.63	69.46
MaPle	80.36	59.18	68.16
PromptSRC	83.37	62.97	71.75
CoPrompt	83.13	64.73	72.79
HPT	83.84	63.33	72.16
DECoOp	-	-	62.72
2SFS	84.60	65.01	73.52
FCPrompt-G	**84.61**	60.31	70.42
DAM	84.13	**64.07**	**73.32**

(k) EuroSAT

	Base	Novel	HM
CLIP	56.48	64.05	60.03
Co-CoOp	87.49	60.04	71.21
LFA	93.40	71.24	80.83
MaPle	94.07	73.23	82.35
PromptSRC	92.90	73.90	82.32
CoPrompt	94.60	78.57	85.84
HPT	94.24	77.12	84.82
DECoOp	-	-	74.61
2SFS	**96.91**	67.09	79.29
FCPrompt-G	96.80	75.21	84.65
DAM	93.10	**82.07**	**87.24**

(l) UCF101

	Base	Novel	HM
CLIP	70.53	77.50	73.85
Co-CoOp	82.33	73.45	77.64
LFA	86.97	77.48	81.95
MaPle	83.00	78.66	80.77
PromptSRC	87.10	78.80	82.74
CoPrompt	86.90	79.57	83.07
HPT	86.52	80.06	83.16
DECoOp	-	-	77.67
2SFS	87.85	78.19	82.74
FCPrompt-G	**87.99**	78.28	82.85
DAM	87.20	**80.90**	**83.93**

C. Base-to-New Generalization

Similar to existing CoOp-based methods, an important evaluation of fine-tuning involves a base-to-new generalization setting composed of two disjoint subsets: base classes and novel classes. The trainable parameters are obtained by using labeled base classes and evaluated on both base classes and the unseen novel classes.

Table I presents the performance of DAM on 11 benchmark datasets under the base-to-new generalization setting. The accuracies on the base set and the novel set, as well as their harmonic means (HM) are reported. The average accuracies over all datasets are summarised in Table I(a). The best results are highlighted in bold. Obviously, our method outperforms on 7 out of 11 datasets in harmonic mean. Compared to our maim competitors, Coprompt and HPT, DAM improves the harmonic mean by 0.72% and 0.97% on average, respectively. Notably, the harmonic mean improvement on the Eurosat dataset, which has a large domain gap, is 2.42% higher than HPT, indicating that our method maintains strong robustness even with significant domain gaps. This demonstrates that our

TABLE II
PERFORMANCE OF DAM ON CROSS-DATASET EVALUATION, I.E. THE MODELS ARE TRAINED ON IMAGENET AND EVALUATED ON OTHER 10 DATASETS WITH POTENTIALLY DIFFERENT DISTRIBUTIONS. DAM IS THE BEST IN 6 OUT OF 10 CASES AND ACHIEVES THE HIGHEST AVERAGE, DEMONSTRATING ITS SUPERIOR CROSS-DATA GENERALIZATION PERFORMANCE.

	Source	**Target**										
	ImageNet	**Caltech**	**Pets**	**Cars**	**Flowers**	**Food**	**Aircraft**	**SUN397**	**DTD**	**EuroSAT**	**UCF101**	**Avg.**
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	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	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
Bayesian Prompt	70.93	93.67	90.63	65.00	70.90	86.30	24.93	67.47	46.10	45.87	68.67	65.95
PromptSRC	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
CoPrompt	70.80	94.50	90.73	65.67	72.30	**86.43**	24.00	67.57	47.07	**51.90**	69.73	67.00
HPT	71.72	94.20	92.63	**66.33**	74.84	86.21	25.68	68.75	**50.87**	47.36	70.50	67.74
DAM	**72.13**	**94.60**	**92.87**	66.07	**75.53**	86.07	**26.10**	**69.00**	50.27	47.57	**71.30**	**67.94**

TABLE III
PERFORMANCES ON DOMAIN GENERALIZATION. DAM OUTPERFORMS SOTAS ACROSS ALL DOMAINS SHIFTS.

	Source	**Target**				
	ImageNet	**-V2**	**-S**	**-A**	**-R**	**Avg.**
CLIP	66.73	60.83	46.15	47.77	73.96	57.17
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
Co-CoOp	71.02	64.07	48.75	50.63	76.18	59.90
LFA	72.65	64.72	48.01	51.50	76.09	60.08
MaPLe	70.72	64.07	49.15	50.90	76.98	60.26
Bayesian Prompt	70.93	64.23	49.20	51.33	77.00	60.44
PromptSRC	71.27	64.35	49.55	50.90	77.80	60.65
CoPrompt	70.80	64.25	49.43	50.50	77.51	60.42
HPT	71.72	65.25	49.36	50.85	77.38	60.71
DAM	**72.13**	**66.00**	**50.30**	**51.62**	**78.07**	**61.50**

DAM excels not only in same-modal feature alignment but also in effectively handling cross-modal feature alignment, thus enhancing the model's generalization capability on new classes.

D. Cross-Dataset Evaluation

Unlike traditional generalization methods that assume similar data distributions of the training and test datasets, cross-dataset generalization trains on a base dataset (e.g., ImageNet) and evaluates on 10 other datasets, potentially with different distributions. This evaluation method aims to test the model's zero-shot capability in a cross-dataset setting.

The cross-data accuracies are reported in Table II with the average across 10 target datasets presented in the last column. DAM achieves the highest average of 67.94%, showcasing its superior generalization performance on datasets with different distributions.

E. Domain Generalization

To evaluate our method's robustness on domain shifts, we treat ImageNet as the source domain and its variants as target domains. We fine-tune the ImageNet-based model and evaluate it on four ImageNet variants that exhibit different domain shifts. The comparison with existing methods is summarized in Table III. DAM outperforms all SOTAs on all target domains and achieves the best average performance of 61.50%, demonstrating its generalization capability across different domain shifts.

F. Plug-in Performance of CMA

CMA can also be applied as a plug-in module on top of existing zero-shot methods. To evaluate the plug-in performance of CMA, we fix all learnable parameters: setting λ to 0.001, μ to 1, α to 0.5, β to 0.65, and ω to 1. We use ResNet-50 as the visual encoder for CLIP and apply CMA after CLIP's visual and text encoders. As shown in Table IV, our method improves the average performance across these 11 datasets by 0.51% compared to CLIP.

G. Performance on Long-tailed Data

Following Candle [44], we evaluate the proposed approach on generalization tasks from base classes to novel classes on long-tailed data. Specifically, in the imbalanced setting, all training data are obtained by downsampling the base class data with varying imbalance ratios (IR), i.e. the ratio between the maximum and minimum number of samples across all classes. During dataset construction, the maximum number of samples per class is set to 100 if this condition is satisfied in the original data; otherwise, it is set according to the maximum per-class number of samples available in the original dataset.

We report the HMs in Table V under three IRs: 10, 20, and 50. On average, DAM consistently outperforms SOTAs across the three IRs. Under the most challenging setting with IR=50, DAM achieves a 78.82% average accuracy, surpassing Candle by 0.62%. This advantage increases to 1.72% and 1.32% as IR decreases to 10 and 20, respectively. DAM is particularly effective in fine-grained and texture recognition tasks: on EuroSAT, DAM is 4.45% higher than Candle, while on DTD, it is 2.23% above Candle. Although DAM is not specifically designed for long-tailed data, it remains effective because the per-image alignment matrices with closed-form solutions are unaffected by class imbalance.

H. Learnable Hyperparameter Values

The values of learnable hyperparameters are reported in Table VI. The ridge regression coefficient λ varies substantially across datasets, with ImageNet requiring a large value of 0.9767 to manage high intra-class variation and avoid overfitting, while EuroSAT prefers a small value of 0.0094. α adjusts bidirectional weights in d_{CMA} and is close to 0.5 across datasets but higher on ImageNet, indicating that both

TABLE IV
USING CMA AS A PLUG-IN MODULE TO IMPROVE THE PERFORMANCE OF CLIP.

Model	ImageNet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF101	Avg.
CLIP	58.20	85.20	82.07	55.17	66.75	74.94	16.95	57.48	40.07	36.07	58.64	57.41
CLIP+CMA	**58.66**	**85.87**	**82.34**	**55.84**	**67.36**	**74.98**	**17.25**	**58.11**	**40.43**	**36.85**	**59.43**	**57.92**

TABLE V
THE HARMONIC MEANS OF BASE AND NOVEL ACCURACIES UNDER DIFFERENT IMBALANCE RATIOS.

(a) Imbalance Ratio = 10

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF101	Avg.
CoOp + Logit Adjusted Loss	91.78	94.10	69.23	71.86	89.25	32.12	72.15	54.88	54.42	64.74	70.66
CoCoOp + Logit Adjusted Loss	95.09	**96.69**	71.91	77.61	**91.20**	33.71	77.99	65.11	60.28	76.78	75.67
Linear Feature Alignment	95.69	94.09	72.72	84.38	90.44	34.27	78.39	67.43	69.56	82.71	76.97
Candle	95.89	95.99	74.30	85.03	90.80	37.78	79.26	68.13	80.51	83.17	79.34
DAM	**96.97**	96.61	**75.71**	**87.81**	90.78	**39.02**	**80.42**	**72.80**	**83.69**	**84.36**	**81.06**

(b) Imbalance Ratio = 20

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF101	Avg.
CoOp + Logit Adjusted Loss	92.65	94.15	67.39	73.72	86.38	29.33	68.93	55.18	62.64	60.12	70.08
CoCoOp + Logit Adjusted Loss	95.25	**96.64**	71.38	80.31	**91.20**	32.78	77.29	61.31	58.82	71.70	74.48
Candle	95.84	95.89	73.49	84.92	90.75	**38.02**	78.53	67.32	80.96	82.59	79.08
DAM	**96.87**	96.19	**73.83**	**87.93**	90.33	36.55	**79.43**	**72.06**	**85.73**	**83.39**	**80.40**

(c) Imbalance Ratio = 50

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF101	Avg.
CoOp + Logit Adjusted Loss	93.30	93.34	67.18	75.45	87.65	29.20	65.91	51.42	57.35	61.62	69.92
CoCoOp + Logit Adjusted Loss	94.90	95.44	69.97	76.84	**91.10**	31.45	76.18	59.37	64.99	77.53	74.32
Candle	94.95	**95.83**	**71.78**	84.62	90.70	**36.68**	78.05	65.69	80.17	**81.72**	78.20
DAM	**96.59**	95.04	71.06	**86.83**	89.73	35.05	**78.45**	**67.92**	**84.62**	81.64	**78.82**

TABLE VI
THE VALUES OF LEARNABLE HYPERPARAMETERS.

hyperparameters	ImageNet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF101
λ	0.9767	0.0979	0.0358	0.1917	0.0991	0.0979	0.0975	0.3862	0.0443	0.0094	0.0976
α	0.7900	0.4996	0.4985	0.5075	0.4999	0.4986	0.4998	0.5401	0.4990	0.4996	0.4995
β	0.9749	0.8326	0.7476	0.9288	0.8843	0.8470	0.8720	0.9477	0.8376	0.6578	0.8822
ω	0.8723	0.9901	0.9964	0.9805	0.9898	0.9895	0.9901	0.9614	0.9956	0.9993	0.9894
μ_{IT}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
μ_{TI}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

directions are important. β controls d_{CMA}'s contribution to the final similarity d. All β values greater than 0.5 suggests that the original similarity dominates, though non-trivial $1-\beta$, notably on Oxford Pets and EuroSAT, demonstrates the relevance of d_{CMA}. The scaling factor ω is close to 1 for all datasets with a slightly lower value for ImageNet, suggesting a preference for the standard log softmax function. Finally, μ_{IT} and μ_{TI}, which adjust the ridge regression solutions, are exactly 1.0000, confirming that the closed-form solution already provides effective alignment.

I. Ablation Experiments

a) Influence of DAM Components: We conduct an ablation study on various combinations of three components of DAM, i.e. SMA for image branch (SMA_I), SMA for text branch (SMA_T) and CMA, for base-to-new generalization. As shown in Table VII, experimental results demonstrate that SMA benefits both text and image modalities and so does CMA. Moreover, with all components enabled, DMA clearly achieves the highest accuracies. These results reveal that the superior generalization performance of DMA is achieved by enforcing effective same-modal and cross-modal alignments.

TABLE VII
THE IMPACT OF SMA FOR IMAGE BRANCH (SMA_I), SMA FOR TEXT BRANCH (SMA_T) AND CMA.

SMA_I	SMA_T	CMA	Base	New	HM
			84.32	76.86	80.23
✓			84.23	77.43	80.69
	✓		83.95	77.25	80.46
		✓	84.15	77.61	80.75
✓	✓		84.26	77.82	80.91
✓		✓	84.22	78.03	81.01
	✓	✓	84.19	77.92	80.93
✓	✓	✓	**84.53**	**78.12**	**81.20**

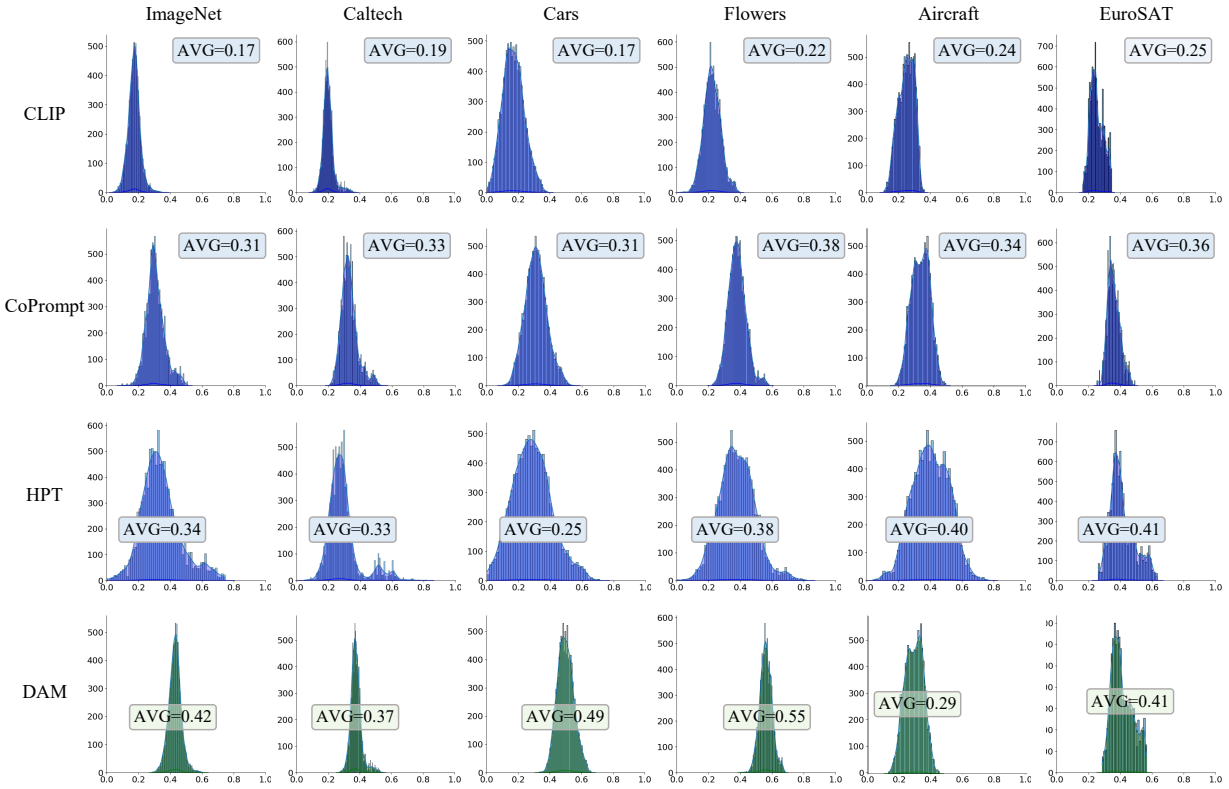


Fig. 3. The histograms of similarity scores between image and text feature pairs. DAM has higher similarity scores over CLIP, CoPrompt and HPT across all datasets, except for Aircraft.

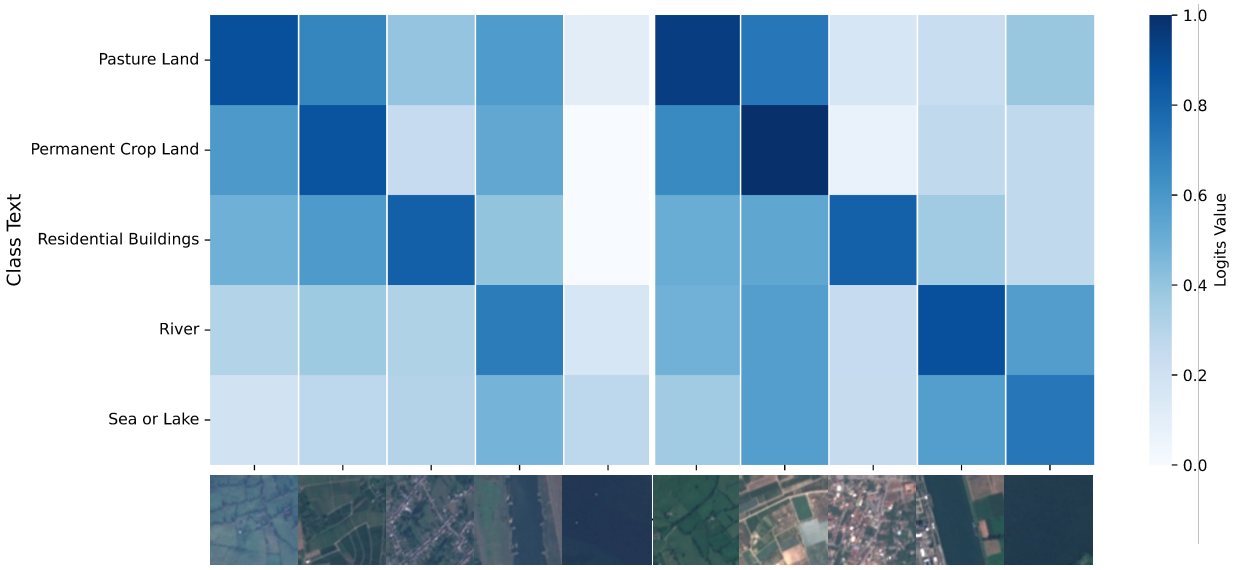


Fig. 4. The visualization of image-text matching performance on the EuroSAT dataset. The vertical axis is the text category, the horizontal axis is the satellite image, the color indicates the model's prediction confidence (dark blue denotes higher confidence), and the diagonal corresponds to the true category of the image.

b) Influence of CMA Components: We further perform ablation study on CMA, evaluating the effects of using only text-to-image alignment (TtoI), only image-to-text alignment (ItoT), and using both of them. The baseline results of this ablation study use only SMA for consistency alignment within modality and cosine similarity as the distance metric. As shown in Table VIII, the experimental results indicate that incorporating a single cross-alignment improves accuracy, while utilizing both cross-alignment strategies yields the best performance. This finding highlights that the bidirectional alignment strategies in CMA provide complementary information to each other, and are essential for improving the model's

TABLE VIII
THE IMPACT OF TWO CMA COMPONENTS, TEXT-TO-IMAGE ALIGNMENT (TTOI) AND IMAGE-TO-TEXT ALIGNMENT (ITOT).

ItoT	TtoI	Base	New	HM
		84.26	77.82	80.91
✓		84.39	78.07	81.11
	✓	84.43	77.95	81.06
✓	✓	**84.53**	**78.12**	**81.20**

TABLE IX
THE IMPACT OF CLOSED-FORM SOLUTIONS OF W'S. ✓INDICATES THAT W'S ARE LEARNABLE.

SMA	CMA	Base	New	HM
✓	✓	84.12	76.76	80.27
✓		83.81	77.31	80.43
	✓	83.89	77.64	80.64
		84.53	**78.12**	**81.20**

generalization capabilities.

c) Influence of Closed-form Solutions of W's: We replace the closed-form solutions with learnable matrices, where ✓indicates that the W's in the corresponding modules are replaced with learnable ones. In Table IX, we report the average accuracies of base and novel classes, as well as their HMs, across 11 datasets. It is clear that using learnable matrices in both SMA and CMA provides the worst performance on new classes, compared with using learnable matrices in only one of the modules. In contrast, using closed-form solutions in both modules achieves the best performance on both base and new classes. It is worth noting that using learnable matrices in both modules yield barely comparable results with baseline HPT in Table I, showing that the performance gain of DAM is due to effective alignment through ridge regression rather than the increase in model capacity.

J. Computational Cost

Table X compares the number of learnable parameters and training time of CLIP, HPT, CoPrompt, and DAM. The number of learnable parameters in DAM is slightly higher than that in HPT, with 11 additional parameters, but is more than 16 times lower than that in CoPrompt owing to the use of closed-form solutions from ridge regression. This suggests that DAM can achieve the best generalization performance with minimal additional learnable parameters. Furthermore, we conduct training time evaluations on the EuroSAT dataset. Our method incurs an additional training time of 0.07 seconds per epoch compared to the baseline HPT, which is acceptable considering the performance gain. Notably, the majority of this extra computational overhead is attributed to matrix inversion operations.

K. Visualization

In Figure 3, we visualize the similarity scores between the image features and the text features in their corresponding classes. The similarity score is measured by cosine similarity in CLIP, CoPrompt and HPT and d in Equation (11) for DAM.

TABLE X
COMPARISON OF THE NUMBERS OF TOTAL PARAMETERS, LEARNABLE PARAMETERS, AND TRAINING TIME PER EPOCH.

Model	Total	Learnable	Training time (s)
CLIP (ViT-B/16)	149620737	-	-
HPT	149916697	295960	0.11
CoPrompt	154622978	4740096	0.10
DAM (Ours)	149916708	295971	0.18

All similarity scores are normalized to $[0, 1]$. We observe that image and text feature pairs are more reliably aligned using DAM, resulting in higher similarity scores compared to traditional methods.

Furthermore, in Figure 4, we visualize the performance of image-text matching on the EuroSAT dataset, illustrated with four example categories: pasture land, permanent crop land, residential buildings, and river. It is clear that the model can accurately match the land scenes with typical textures, such as buildings, land boundaries, and rivers. However, its performance is limited in the sea or lake category, where images lack clear and discriminative textures.

V. CONCLUSION

We propose DAM for fine-tuning VLMs for downstream tasks. DAM integrates SMA for consistent prediction within the same modalities and CMA for bidirectional alignment of different modalities. The alignments are achieved by ridge regression with closed-form solutions, enabling a balance between model complexity and feature alignment accuracy and the minimal addition to the number of learnable parameters. Extensive experimental results and ablation studies showcase the superior generalization performance of DAM.

ACKNOWLEDGEMENT

This work was partly supported by the National Nature Science Foundation of China (Grants 62176110, 62225601, U23B2052), Key Talent Program of Gansu Province, China under Grant 2025RCXM002, Science and Technology Program of Hebei, China under Grant SZX2020034, Hong-Liu Distinguished Young Talents Foundation of Lanzhou University of Technology, Beijing Natural Science Foundation Project No. L242025, and the Royal Society under International Exchanges Award IEC\NSFC\201071.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[2] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. J. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *ArXiv*, vol. abs/2110.04544, 2021.

[3] T. Yu, Z. Lu, X. Jin, Z. Chen, and X. Wang, "Task residual for tuning vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 899–10 909.

[4] X. Li, D. Lian, Z. Lu, J. Bai, Z. Chen, and X. Wang, "Graphadapter: Tuning vision-language models with dual knowledge graph," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 795–16 804, 2022.

[6] M. U. Khattak, H. A. Rasheed, M. Maaz, S. H. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19 113–19 122, 2022.

[7] Y. Ouali, A. Bulat, B. Martínez, and G. Tzimiropoulos, "Black box few-shot adaptation for vision-language models," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15 488–15 500, 2023.

[8] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, pp. 2337 – 2348, 2021.

[9] A. Bulat and G. Tzimiropoulos, "Language-aware soft prompting for vision & language foundation models," *ArXiv*, vol. abs/2210.01115, 2022.

[10] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, and B. Cui, "Calip: Zero-shot enhancement of clip with parameter-free attention," in *AAAI Conference on Artificial Intelligence*, 2022.

[11] L. Yang, R.-Y. Zhang, Y. Wang, and X. Xie, "Mma: Multi-modal adapter for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 23 826–23 837.

[12] S. Roy and A. Etemad, "Consistency-guided prompt learning for vision-language models," *ArXiv*, vol. abs/2306.01195, 2023.

[13] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2400–2413, 2020.

[14] Y. Liu, X. Gao, J. Han, and L. Shao, "A discriminative cross-aligned variational autoencoder for zero-shot learning," *IEEE Transactions on Cybernetics*, vol. 53, no. 6, pp. 3794–3805, 2023.

[15] K. Wei, C. Deng, X. Yang, and D. Tao, "Incremental zero-shot learning," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13 788–13 799, 2022.

[16] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11 941–11 952, 2023.

[17] S. Lavoie, P. Kirichenko, M. Ibrahim, M. Assran, A. G. Wilson, A. Courville, and N. Ballas, "Modeling caption diversity in contrastive vision-language pretraining," *ArXiv*, vol. abs/2405.00740, 2024.

[18] Q. Chang, K.-W. Hung, and J. Jiang, "Fine tuning of deep contexts toward improved perceptual quality of in-paintings," *IEEE Transactions on Cybernetics*, vol. 52, no. 6, pp. 4850–4854, 2022.

[19] Y. Wang, X. Jiang, D. Cheng, D. Li, and C. Zhao, "Learning hierarchical prompt with structured linguistic knowledge for vision-language models," in *AAAI Conference on Artificial Intelligence*, 2023.

[20] G. Kim, S. Kim, and S. Lee, "Aapl: Adding attributes to prompt learning for vision-language models," *ArXiv*, vol. abs/2404.16804, 2024.

[21] Z. Zhou, M. Yang, J.-X. Shi, L.-Z. Guo, and Y.-F. Li, "Decoop: Robust prompt tuning with out-of-distribution detection," *ArXiv*, vol. abs/2406.00345, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270219133

[22] Z. Shi and A. Lipani, "Dept: Decomposed prompt tuning for parameter-efficient fine-tuning," *ArXiv*, vol. abs/2309.05173, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261682057

[23] M. Mistretta, A. Baldrati, M. Bertini, and A. D. Bagdanov, "Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation," in *European Conference on Computer Vision*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270923864

[24] C. Ma, Y. Liu, J. Deng, L. Xie, W. Dong, and C. Xu, "Understanding and mitigating overfitting in prompt tuning for vision-language models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4616–4629, 2023.

[25] X. Yin, J. Wu, W. Yang, X. Zhou, S. Zhang, and T. Zhang, "Hierarchy-aware interactive prompt learning for few-shot classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, pp. 12 221–12 232, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271414720

[26] X. Liu, J. Wu, W. Yang, X. Zhou, and T. Zhang, "Multi-modal attribute prompting for vision-language models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[27] M. Farina, M. Mancini, G. Iacca, and E. Ricci, "Rethinking few-shot adaptation of vision-language models in two stages," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29 989–29 998, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277043339

[28] L. Liu, N. Wang, C. Chen, D. Liu, X. Yang, X. Gao, and T. Liu, "Frequency-based comprehensive prompt learning for vision-language models." *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:280690395

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[30] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*. IEEE, 2004, pp. 178–178.

[31] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.

[32] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[33] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008, pp. 722–729.

[34] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*. Springer, 2014, pp. 446–461.

[35] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[36] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.

[37] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[38] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.

[39] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492.

[40] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.

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

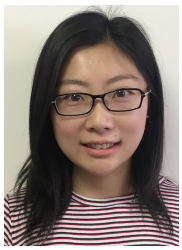
[42] H. Gao, H. Zhang, X. Yang, W. Li, F. Gao, and Q. Wen, "Generating natural adversarial examples with universal perturbations for text classification," *Neurocomputing*, vol. 471, pp. 175–182, 2022.

[43] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *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*, 2021, pp. 8340–8349.

[44] J.-X. Shi, C. Zhang, T. Wei, and Y.-F. Li, "Efficient and long-tailed generalization for pre-trained vision-language model," *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270562385

Peiyu Lu He is currently working toward the M.E. degree with Lanzhou University of Technology. His research interests include computer vision and few-shot learning.

Xiaoxu Li received the Ph.D. degree from Beijing University of Posts and Telecommunications in 2012. She is currently a Professor with the School of Computer and Communication, Lanzhou University of Technology. Her research interests include machine learning fundamentals with a focus on applications in image and video understanding. She is also a member of the China Computer Federation.



Rui Zhu received the Ph.D. degree in statistics from University College London in 2017. She is a Senior Lecturer in Statistics in the Faculty of Actuarial Science and Insurance, City St George's, University of London. Her research interests include machine learning, computer vision and interdisciplinary applications in actuarial science. She serves as the Associate Editor for IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Circuits and Systems for Video Technology and Neurocomputing.

Zhanyu Ma is currently a Professor at Beijing University of Posts and Telecommunications, Beijing, China, since 2019. He received the Ph.D. degree in electrical engineering from KTH Royal Institute of Technology, Sweden, in 2011. From 2012 to 2013, he was a Postdoctoral Research Fellow with the School of Electrical Engineering, KTH. He has been an Associate Professor with the Beijing University of Posts and Telecommunications, Beijing, China, from 2014 to 2019. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in computer vision, multimedia signal processing. He is a Senior Member of IEEE.

Jie Cao received the M.E. degree from Xi'an Jiaotong University, China, in 1994. She is currently a Professor and a Vice President of the Lanzhou University of Technology. Her research interests include machine learning, pattern recognition, speech and speaker recognition, information fusion, and computer vision.

Jing-Hao Xue received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is a Professor of Statistical Pattern Recognition in the Department of Statistical Science, University College London. His research interests include statistical pattern recognition, machine learning, and computer vision. He received the Best Associate Editor Award of 2021 from the IEEE Transactions on Circuits and Systems for Video Technology, and the Outstanding Associate Editor Award of 2022 from the IEEE Transactions on Neural Networks and Learning Systems.