MLAE: Masked LoRA Experts for Parameter-Efficient Fine-Tuning

Junjie Wang, Guangjing Yang, Wentao Chen, Huahui Yi, Xiaohu Wu, Zhouchen Lin, *Fellow, IEEE*, and Qicheng Lao

***Abstract*—In response to the challenges posed by the extensive parameter updates required for full fine-tuning of large-scale pre-trained models, parameter-efficient fine-tuning (PEFT) methods, exemplified by Low-Rank Adaptation (LoRA), have emerged. LoRA simplifies the fine-tuning process but may still struggle with a certain level of redundancy in low-rank matrices and limited effectiveness from merely increasing their rank. To address these issues, a natural idea is to enhance the independence and diversity of the learning process for the low-rank matrices. Therefore, we propose Masked LoRA Experts (MLAE), an innovative approach that applies the concept of masking to visual PEFT. Our method incorporates a cellular decomposition strategy that treats rank-1 components as experts defined under the chosen LoRA parameterization, thus enhancing diversity among update components. Additionally, we introduce a binary mask matrix that selectively activates these experts during training to promote more diverse and anisotropic learning, based on expert-level dropout strategies. Our investigations reveal that this selective activation not only enhances performance but also fosters a more diverse acquisition of knowledge with a marked decrease in parameter similarity among MLAE, significantly boosting the quality of the model. Remarkably, MLAE achieves new state-of-the-art (SOTA) performance with an average accuracy score of 78.8% on the VTAB-1k benchmark and 90.9% on the FGVC benchmark, surpassing the previous SOTA result by an average of 0.8% on both benchmarks. Moreover, MLAE shows strong generalization across diverse tasks, including LLM fine-tuning, semantic segmentation, and image/video–text understanding, underscoring its versatility and effectiveness in advancing PEFT.**

***Index Terms*—Parameter-efficient fine-tuning, pre-trained vision foundation models, masking, low-rank adaptation.**

Received 4 March 2025; revised 7 October 2025, 29 March 2026, and 22 May 2026; accepted 6 July 2026. Date of publication 20 July 2026; date of current version 28 July 2026. The work of Zhouchen Lin was supported in part by NSF China under Grant 62276004, in part by the Beijing Natural Science Foundation under Grant L257007, and in part by the Beijing Major Science and Technology Project under Grant Z251100008425006. The associate editor coordinating the review of this article and approving it for publication was Prof. Yipeng Liu. *(Junjie Wang and Guangjing Yang contributed equally to this work.) (Corresponding authors: Zhouchen Lin; Qicheng Lao.)*

Junjie Wang, Guangjing Yang, and Qicheng Lao are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: jjwang@bupt.edu.cn; ygj2018@bupt.edu.cn; qicheng.lao@gmail.com).

Wentao Chen is with Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wentaochen@sjtu.edu.cn).

Huahui Yi is with West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, Sichuan 610000, China (e-mail: huahui.yi1@gmail.com).

Xiaohu Wu is with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: xiaohu.wu@bupt.edu.cn).

Zhouchen Lin is with the State Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Beijing 100871, China (e-mail: zlin@pku.edu.cn).

Digital Object Identifier 10.1109/TIP.2026.3712351

I. Introduction

LARGE-SCALE pre-trained vision models [1], [2], [3], [4], [5] have manifested outstanding performance across a variety of downstream vision tasks [6], [7], [8], [9], [10]. However, full parameter fine-tuning poses significant challenges to computational resources and data storage. To address this issue, considerable efforts have been invested in developing parameter-efficient fine-tuning (PEFT) methods [11], [12], [13], [14], [15], [16], [17], [18], [19], which only fine-tunes a minimal number of parameters. Among these, Low-Rank Adaptation (LoRA) [15] stands out due to its simplicity and efficiency, which achieves high performance leveraging low-rank matrices. However, recent studies [20], [21], [22] have indicated that even within low-rank matrices, a certain level of redundancy persists, and merely increasing the rank of the update matrix does not necessarily enhance the quality of the model. Consequently, how to simultaneously reduce redundancy in low-rank matrices and enhance their quality represents a fundamental challenge.

Recent works exemplified by AdaLoRA [23] and IncreLoRA [24] have been proposed to dynamically adjust the matrix rank by adding or removing parameters in different layers of the model based on importance scores, addressing the issue of parameter redundancy in shallow networks. Inspired by these, we conceive the idea of temporarily removing certain parameters during the training phase, instead of permanently eliminating them without the chance of revitalization. This approach not only eliminates the need for calculating parameter scores, thereby reducing computational overhead, but also allows all parameters to participate during inference, enhancing the model's expressiveness and robustness. Therefore, a natural idea is employing masking to the parameters. However, it is impractical to apply masking directly to parameters of a low-rank matrix, as it may result in training instability and inefficiency, while individual parameters within such matrices do not carry any semantic information advocating diverse learning. We have observed that recent efforts [25], [26] have applied LoRA framework within the mixture of experts (MoE) [27] architecture, where each LoRA weight is treated as an expert, and a gating mechanism selects experts to achieve sparsity. This motivates us to explore applying masking to LoRA experts, rather than directly to the parameters. However, unlike these works, our goal is to promote diversity among fixed low-rank components through masking. Accordingly, we refer to each rank-1 component $b_j a_j^\top$ as an expert for brevity. These experts are defined under the chosen LoRA parameterization rather

than being intrinsically identifiable components of BA or learned routing experts, and our method does not constitute a MoE module: the backbone remains dense, no token-wise routing is employed, and all experts participate during inference.

In this work, we propose a novel method, MLAE (Masked LoRA Experts), which applies the concept of masking to the field of visual parameter-efficient fine-tuning (PEFT) by enhancing the independence and diversity of learning to address the issue of parameter redundancy. Specifically, MLAE addresses redundancy and enhances the quality of LoRA from two perspectives: First, it employs a cellular decomposition approach, where the low-rank matrix with rank r is further decomposed into r rank-1 experts under the chosen LoRA parameterization; Second, building on cellular decomposition, it further introduces a mask matrix with adaptive coefficients, applied to the decomposed expert matrix to implement MLAE. Through in-depth exploration of fixed, stochastic, and mixed masking with different configurations, we have discovered that the use of expert-level dropout to generate stochastic masks yields the best performance. Further analysis has revealed that the parameter similarity among these experts significantly decreases and they acquire a more diverse range of knowledge.

In summary, our contributions are mainly as follows:

- We tackle an interesting and relevant issue in the field of fine-tuning vision pre-trained models, specifically the problem of parameter diversity and redundancy in LoRA, which is a significant contribution to the domain of parameter-efficient fine-tuning.
- We have pioneered the introduction of masking strategies into the domain of visual PEFT, exploring various masking techniques including fixed, stochastic, and mixed masks. This innovation paves the way for future work in the field.
- We introduce MLAE, a method that is both simple and flexible. MLAE has attained state-of-the-art (SOTA) results on the VTAB-1k and FGVC benchmarks, achieving an average accuracy improvement of 0.8% over previous SOTA results on both benchmarks. In addition, MLAE demonstrates strong performance across diverse tasks, including LLM fine-tuning, semantic segmentation, and image/video–text understanding, further highlighting its effectiveness and generalization ability. This demonstrates its strong potential in advancing the field of PEFT.

II. Related Works

A. Parameter-Efficient Fine-Tuning

As the number of model parameters has continued to grow, the significant computational and storage costs with the traditional full fine-tuning have made Parameter-Efficient Fine-Tuning (PEFT) increasingly attractive. The PEFT methods for large-scale pre-trained models first emerged in the NLP field [28], where they achieve comparable performance to full fine-tuning while only requiring the fine-tuning of a few lightweight modules. Inspired by this success in NLP, researchers have begun applying PEFT to Pretrained Vision Models (PVMs). Existing visual PEFT methods can be categorized into addition-based, unified-based, and partial-based methods. Addition-based methods [11], [12], [13], [29], [30], [31], exemplified by VPT [12], AdaptFormer [11], Fact [32] and Prefix-tuning [31], incorporate additional trainable modules or parameters into the original PVMs to learn task-specific information. Unified-based methods [14], [33], [34], [35], [36], represented by NOAH [14], SPT [37] and GLoRA [34], integrate various fine-tuning approaches into a single architecture. Partial-based methods [15], [28], [38], [39], [40], such as LoRA [15], Adapter [28], SSF [39], and RepAdapter [40], update only a small portion of the inherent parameters while keeping most of the model's parameters unchanged during the adaptation process. In this paper, we explore more efficient and high-quality strategies for visual PEFT based on partial-based methods.

B. Masking Strategies

In recent years, masking strategies have been extensively employed in deep learning, primarily focusing on self-supervised pre-training through mask modeling. Representative works [41], [42], [43], [44], [45], [46] include language model pre-training spearheaded by BERT [47] and visual model pre-training exemplified by MAE [2] and SimMIM [48], both of which have achieved significant success. Typically, these works apply masking strategies to the input sequences, retaining a portion of the input and training models to predict the masked content. Meanwhile, a smaller body of work [49], [50], [51], [52], [53], [54], [55], [56], [57] has applied masking strategies to the model architecture, achieving preliminary results in continual learning and multi-task learning. Diverging from these approaches, we innovatively apply the masking strategy to the lightweight modules of the model based on the PEFT methods, conducting a comprehensive analysis that has yielded outstanding results.

III. Methods

Unlike existing approaches that adjust the incremental matrix rank based on importance scores or select the optimal experts through gating mechanisms, our methodology aims to alleviate parameter redundancy by enforcing diversity in the learning process. The key idea of our methodology is to enhance the capabilities of the model within a given budget by promoting varied learning rather than merely eliminating underutilized parameters. Therefore, we focus on achieving independence and diversity in parameter learning through two strategic approaches: by designing innovative model architectures and introducing masking strategy for training. The complete process is shown in Fig. 1.

A. Preliminary

1) Low-Rank Adaptation: LoRA [15] has been demonstrated to be a popular parameter-efficient fine-tuning approach to adapt pre-trained models to specific tasks. It leverages low-rank matrix decomposition of pre-trained weight matrices to significantly reduce the number of training parameters. Formally, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA creates two low-rank trainable matrices A and B with rank r, where $B \in \mathbb{R}^{d_{\text{out}} \times r}$, $A \in \mathbb{R}^{r \times d_{\text{in}}}$, and $r \ll \min\{d_{\text{out}}, d_{\text{in}}\}$.

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{1}$$

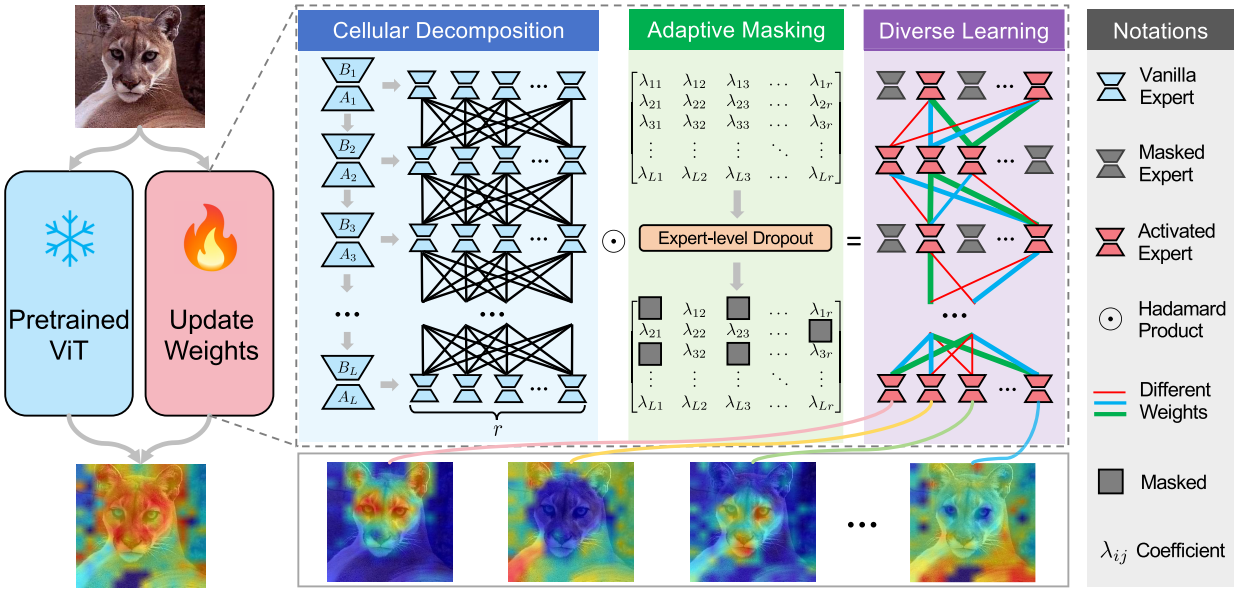


Fig. 1. Our proposed Masked LoRA Experts (MLAE) framework. MLAE decomposes a rank-r LoRA update into rank-1 experts and applies adaptive expert-level masking during training, encouraging diverse and less redundant learning.

During training, LoRA updates only the matrices A and B, while W_0 is frozen. The matrix A is initialized with a random Gaussian distribution and matrix B is initialized to zero.

B. Masked LoRA Experts

Vanilla LoRA significantly reduces the fine-tuning parameters of over-parametrized models by updating only low-rank matrices, nevertheless, without carefully imposed independence and diversity constraints, a certain degree of redundancy may still remain within these low-rank matrices.

1) Cellular Decomposition: To obtain fine-grained trainable units, we decompose the low-rank update into rank-1 components under the chosen LoRA parameterization. We refer to each column-row pair as an expert, while noting that these experts are defined under this parameterization rather than being naturally identifiable components of BA. Each expert can be separately initialized, weighted, and masked during training, which encourages diverse parameter updates without introducing additional routing or gating mechanisms. More specifically, the low-rank matrices B and A with rank r in Equation (1) can be written as $B = [b_1, b_2, \ldots, b_r]$ with $b_j \in \mathbb{R}^{d_{\text{out}} \times 1}$, and $A = [a_1, a_2, \ldots, a_r]^T$ with $a_j \in \mathbb{R}^{d_{\text{in}} \times 1}$. Accordingly, the update matrix in the i-th layer can be expressed as $\Delta W_i = B_i A_i = \sum_{j=1}^{r} b_{ij} a_{ij}^T$, where $\Delta W_i \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. For a pre-trained model with a total of L layers, each with r experts, the model features $L \times r$ experts in total, denoted as the matrix-valued array $\mathcal{E}$:

$$\begin{bmatrix} \Delta W_1 \\ \Delta W_2 \\ \vdots \\ \Delta W_L \end{bmatrix} = \begin{bmatrix} B_1 A_1 \\ B_2 A_2 \\ \vdots \\ B_L A_L \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{r} b_{1j} a_{1j}^T \\ \sum_{j=1}^{r} b_{2j} a_{2j}^T \\ \vdots \\ \sum_{j=1}^{r} b_{Lj} a_{Lj}^T \end{bmatrix}$$
$$\overset{\text{into array}}{\Longrightarrow} \mathcal{E} = \begin{bmatrix} b_{11}a_{11}^T & b_{12}a_{12}^T & \ldots & b_{1r}a_{1r}^T \\ b_{21}a_{21}^T & b_{22}a_{22}^T & \ldots & b_{2r}a_{2r}^T \\ \vdots & \vdots & \ddots & \vdots \\ b_{L1}a_{L1}^T & b_{L2}a_{L2}^T & \ldots & b_{Lr}a_{Lr}^T \end{bmatrix} \tag{2}$$

where $b_{ij}a_{ij}^T$ corresponds to the j-th expert in the i-th layer. Here, $\mathcal{E}$ should be interpreted as an $L \times r$ matrix-valued array rather than an ordinary block matrix: its (i, j)-th entry is the rank-1 matrix $b_{ij}a_{ij}^T \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Accordingly, for any scalar matrix $C = [c_{ij}] \in \mathbb{R}^{L \times r}$, the product $C \odot \mathcal{E}$ is defined blockwise by

$$(C \odot \mathcal{E})_{ij} = c_{ij} b_{ij} a_{ij}^T.$$

The multiplication by $\mathbf{1}_r$ then denotes summation over the r matrix-valued entries in each layer.

2) Masking Rank-1 LoRA Experts: On the basis of cellular decomposition, we introduce a mask matrix $M \in \mathbb{B}^{L \times r}$, where $\mathbb{B}$ denotes the binary set $\{0, 1\}$, and each element m_{ij} of M is a member of this set. The update matrices across L layers are:

$$\Delta W_i = \sum_{j=1}^{r} m_{ij}\, b_{ij} a_{ij}^T, \qquad i = 1, \ldots, L. \tag{3}$$

$$[\Delta W_1\ \Delta W_2\ \ldots\ \Delta W_L]^T = \big(M \odot \mathcal{E}\big)\, \mathbf{1}_r,\ \mathbf{1}_r := [\,1, \ldots, 1\,]^\top \tag{4}$$

where $\odot$ is the blockwise Hadamard product defined above. During the parameter initialization phase, each a_{ij}^T and b_{ij} is initialized separately to ensure that each expert starts the updating process in a distinct initial state, with a_{ij}^T being drawn from a random Gaussian distribution and b_{ij} set to zero.

C. Adaptive Masking Through Expert-Level Dropout

1) Adaptive Coefficients: We hypothesize that the contribution of each expert to the model's performance varies, thus necessitating the allocation of unique weights to each component. Specifically, $\Delta W_l = \sum_{j=1}^{r} \lambda_{lj} b_{lj} a_{lj}^T$. Thus, the total update matrices are:

$$[\Delta W_1\ \Delta W_2\ \ldots\ \Delta W_L]^T = (M \odot \Lambda \odot \mathcal{E})\, \mathbf{1}_r$$
$$= \left(M \odot \begin{bmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1r} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{L1} & \lambda_{L2} & \ldots & \lambda_{Lr} \end{bmatrix} \odot \mathcal{E} \right) \mathbf{1}_r \tag{5}$$

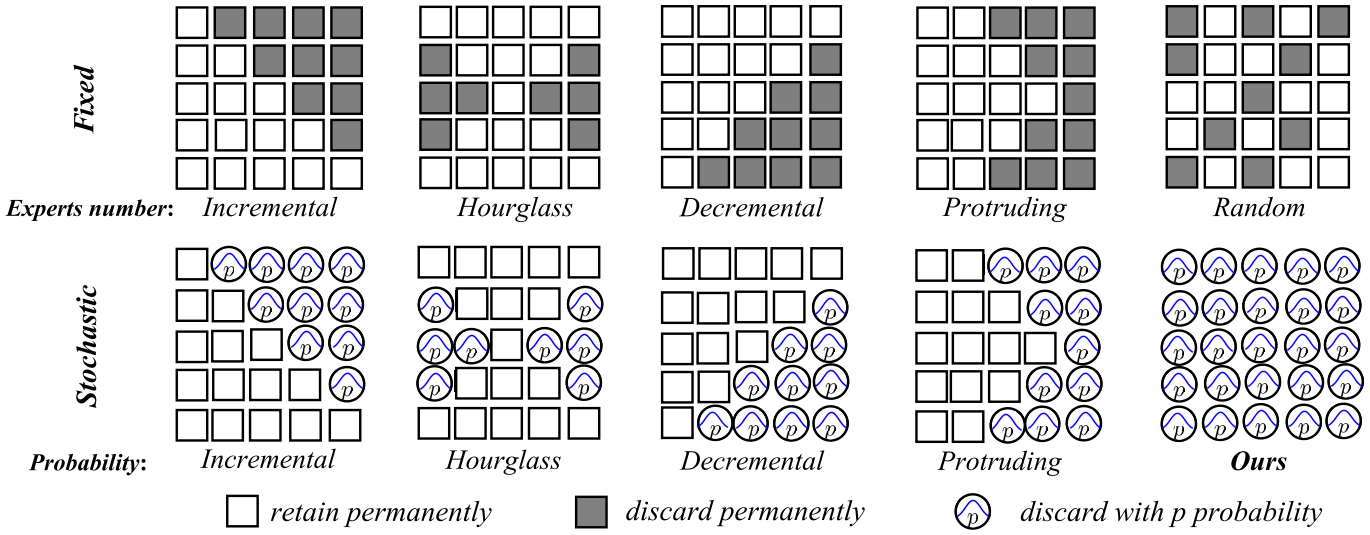


Fig. 2. Different masking strategies. The top row shows five representative fixed masking patterns, where selected experts are permanently discarded. The bottom row presents the corresponding stochastic masking variants, where different numbers of marked experts are used to illustrate different masking probability distributions across layers. For clarity, the unmarked experts in the bottom row do not indicate permanent retention, but are only shown to visualize the layer-wise probability patterns.

TABLE I

FULL RESULTS ON VTAB-1K BENCHMARK. "# PARAM" SPECIFIES THE TOTAL NUMBER OF TRAINABLE PARAMETERS IN BACKBONES. AVERAGE ACCURACY (%) AND # PARAM ARE AVERAGED OVER GROUP-WISE MEAN VALUES. THE BEST RESULT IS IN BOLD, AND THE SECOND-BEST RESULT IS UNDERLINED. GLORA† INDICATES THE CONFIGURATION WITH A PARAMETER COUNT COMPARABLE TO MLAE

		Natural							Specialized				Structured								
	# param (M)	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	Average
Traditional Fine-tuning																					
Full	85.8	68.9	87.7	64.3	97.2	86.9	87.4	38.8	79.7	95.7	84.2	73.9	56.3	58.6	41.7	65.5	57.5	46.7	25.7	29.1	68.9
Linear	0	64.4	85.0	63.2	97.0	86.3	36.6	51.0	78.5	87.5	68.5	74.0	34.3	30.6	33.2	55.4	12.5	20.0	9.6	19.2	57.6
PEFT methods																					
VPT-Deep [12]	0.53	**78.8**	90.8	65.8	98.0	88.3	78.1	49.6	81.8	96.1	83.4	68.4	68.5	60.0	46.5	72.8	73.6	47.9	32.9	37.8	72.0
Adapter [28]	0.16	69.2	90.1	68.0	98.8	89.9	82.8	54.3	84.0	94.9	81.9	75.5	80.9	65.3	48.6	78.3	74.8	48.5	29.9	41.6	73.9
AdaptFormer [11]	0.16	70.8	91.2	70.5	99.1	90.9	86.6	54.8	83.0	95.8	84.4	76.3	81.9	64.3	49.3	80.3	76.3	45.7	31.7	41.1	74.7
LoRA [15]	0.29	67.1	91.4	69.4	98.8	90.4	85.3	54.0	84.9	95.3	84.4	73.6	82.9	**69.2**	49.8	78.5	75.7	47.1	31.0	44.0	74.5
NOAH [14]	0.36	69.6	92.7	70.2	99.1	90.4	86.1	53.7	84.4	95.4	83.9	75.8	82.8	68.9	49.9	81.7	81.8	48.3	32.8	44.2	75.5
FacT [32]	0.07	70.6	90.6	70.8	99.1	90.7	88.6	54.1	84.8	96.2	84.5	75.7	82.6	68.2	49.8	80.7	80.8	47.4	33.2	43.0	75.6
SSF [39]	0.24	69.0	92.6	75.1	99.4	91.8	90.2	52.9	87.4	95.9	87.4	75.5	75.9	62.3	53.3	80.6	77.3	54.9	29.5	37.9	75.7
RepAdapter [40]	0.22	72.4	91.6	71.0	99.2	91.4	90.7	55.1	85.3	95.9	84.6	75.9	82.3	68.0	50.4	79.9	80.4	49.2	**38.6**	41.0	76.1
SPT-Adapter [37]	0.38	72.9	93.2	72.5	99.3	91.4	88.8	55.8	86.2	96.1	85.5	75.5	83.0	68.0	51.9	81.2	82.4	51.9	31.7	41.2	76.4
SPT-LoRA [37]	0.54	73.5	93.3	72.5	99.3	91.5	87.9	55.5	85.7	96.2	85.9	75.9	**84.4**	67.6	52.5	82.0	81.0	51.1	30.2	41.3	76.4
DA-VPT+ [64]	0.24	74.4	92.7	74.3	99.4	91.3	**91.5**	50.3	86.2	96.2	87.2	76.3	81.3	62.5	52.8	65.3	84.9	51.0	33.1	**48.7**	76.1
CaRA [65]	0.06	71.3	91.9	71.8	99.3	91.4	90.5	54.7	86.2	96.4	86.0	75.4	83.8	69.4	51.4	81.7	80.8	47.4	35.3	44.0	76.5
CDRA-SPT [66]	0.42	73.8	93.3	72.7	99.4	91.6	89.9	55.8	86.8	96.2	86.2	76.0	83.0	68.9	51.5	82.1	80.9	51.7	33.0	43.2	76.8
GLoRA† [34]	0.29	76.1	92.7	**75.3**	**99.6**	92.4	90.5	57.2	87.5	96.7	88.1	76.1	81.0	66.2	52.4	84.9	81.8	53.3	33.3	39.8	77.3
GLoRA [34]	0.86	76.4	92.9	74.6	**99.6**	92.5	90.5	57.8	87.3	**96.8**	88.0	76.0	83.1	67.3	54.5	**86.2**	83.8	52.9	37.0	41.4	78.0
MLAE	0.30	77.8	**94.7**	74.8	99.4	**92.6**	90.6	**58.4**	**88.4**	96.5	**88.5**	**76.7**	84.3	67.2	**55.7**	82.6	**87.8**	**57.1**	35.7	47.7	**78.8**

where Λ is the adaptive coefficient matrix, λ_{ij} denotes the weight assigned to the j-th expert in the i-th layer, and all the elements in Λ are initialized to 1 to ensure that all experts contribute equally during the initial phase. Notably, IncreLoRA [24] has also adopted a similar strategy of scaling coefficients. However, their primary goal is to enable the update matrix to be decomposed in a manner similar to SVD in AdaLoRA [23], facilitating subsequent incremental allocation. In their approach, the initial values of the scaling coefficients λ are set to zero.

2) Fixed Masking: Drawing inspiration from MoLA [20], which explores layer-wise expert allocation under the MoE framework, we specifically design fixed masks to investigate which layer (lower, middle, or upper) requires a more dense application of masks, as the five representative masks shown in Fig. 2. When fixed masking is applied, the masked experts are permanently discarded, meaning they do not undergo gradient updates and are not utilized during inference.

3) Stochastic Masking: Similar to the fixed patterns, we have implemented five variations of stochastic masking by

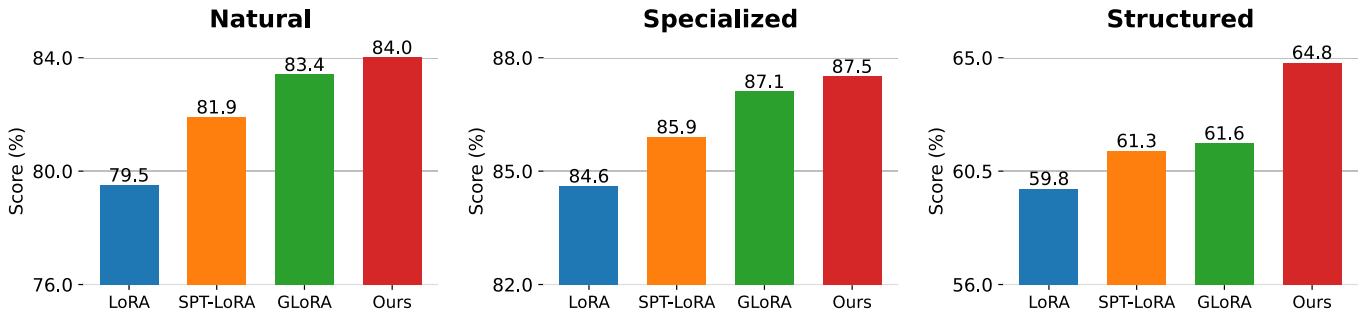


Fig. 3. Domain-wise average scores on VTAB-1k benchmark. Our MLAE performs best in all three domains, especially in Structured domain, surpassing GLoRA by 3.2% under a similar budget.

employing expert-level dropout to Λ and adjusting the dropout rate p across different layers, as illustrated in the bottom of Fig. 2. Accordingly, the total update matrix is given by:

$$[\Delta W_1\ \Delta W_2\ \ldots\ \Delta W_L]^T = \big(\text{Dropout}(\Lambda, p) \odot \mathcal{E}\big)\,\mathbf{1}_r \qquad (6)$$

where p represents the probability of an element being set to zero. During the training process, experts are randomly discarded, and the outputs of the remaining experts are scaled up to maintain the expected output magnitude. Consequently, these discarded experts do not undergo gradient updates during that iteration. During the inference, all experts are activated and used without any scaling.

4) Mixed Masking: Building on the aforementioned fixed masking and stochastic masking, we further propose mixed masking. In this strategy, certain experts are permanently masked, while stochastic masking is applied to the remaining experts. Accordingly, there are also five representative masks.

TABLE II
RESULTS ON FGVC BENCHMARK. "TUNED/TOTAL" DENOTES THE FRACTION OF TRAINABLE PARAMETERS

ViT-B/16	FGVC	
	Tuned/Total	Mean Acc (%).
Traditional Fine-tuning		
FULL	100%	88.5
PEFT methods		
Adapter	0.41%	85.7
LoRA-8	0.55%	86.0
SSF	0.39%	90.7
VPT-Deep	0.98%	89.1
SPT-LoRA	0.60%	90.1
MLAE (Ours)	0.56%	**90.9**

IV. EXPERIMENTS

A. Vision Pre-Trained Model Fine-Tuning

1) Image Classification: We now describe our experimental setup, including backbone, implementation details, and training strategies, to ensure reproducibility and fair comparison.

We evaluate our method on a total of 24 downstream tasks in two groups following VPT [12]: (1) VTAB-1k [58] is a large-scale transfer learning benchmark consisting of a collection of 19 vision datasets, which are clustered into three domains: Natural, Specialized, and Structured. The final model used for evaluation is trained using the full 1,000 examples in each dataset. Following [12], we use top-1 classification accuracy (%) averaged within each domain as the metric; (2) FGVC is a benchmark for fine-grained visual classification tasks including CUB-200-2011 [59], NABirds [60], Oxford Flowers [61], Stanford Dogs [62] and Stanford Cars [9]. We follow the validation splits in VPT [12]. We also report the top-1 accuracy averaged on the FGVC datasets. The average accuracy score on the test set with three runs is reported. For a fair comparison, we follow SPT [37] and use ViT-B/16 [1] pre-trained on supervised ImageNet-21K [6] as our default backbone. We only implement our proposed MLAE in the Q-K-V linear transformation layer across the 12 blocks of ViT-B, and this alone enables us to surpass GLoRA [34] which fine-tunes all linear layers in the model. However, MLAE can be implemented in any linear layer across various network architectures. All the experiments are trained using PyTorch on an NVIDIA GeForce RTX 3090 GPU. Following SPT [37], we employ the AdamW optimizer [63] with cosine learning rate decay in the training process. We set the batch size, learning rate, and weight decay as 64, 5e-4, 1e-4, respectively. Based on previous works [12], [14], [37], we also follow the same standard data augmentation pipeline, e.g., processing the image with random resized crop to 224 $\times$ 224 and a random horizontal flip for the FGVC datasets. Additionally, during data preprocessing, we use ImageNet inception mean and standard deviation for normalization. The number of training epochs is set to 500 for VTAB-1k and 300 for FGVC to achieve better convergence. Inspired by RepAdapter [40] for searching the hyperparameter scaling coefficient, we also explore the initialization values of the coefficient matrices and the dropout probability in our method. Dropout probability is searched from [0.0, 0.1, 0.3, 0.5, 0.7, 0.9] and the initialization value of expert coefficient is searched from [0.125, 0.25, 0.5, 1, 2, 4]. More details are provided in Table XII.

Our proposed approach achieves SOTA performance on both VTAB-1k and FGVC benchmarks. We first compare MLAE with the state-of-the-art PEFT methods on ViT, as reported in Table I. We initially note that all PEFT methods significantly surpass full fine-tuning in performance, whereas linear probing, which involves merely adjusting the classifier, yields substantially poorer results. Compared to these

TABLE III

INFERENCE EFFICIENCY COMPARISON OF MLAE WITH EXISTING METHODS DURING INFERENCE. ΔP AND ΔF DENOTE THE ADDITIONAL PARAMETERS AND FLOPS BY PEFT METHODS, RESPECTIVELY. THE INFERENCE THROUGHPUT IS DEFINED AS IMAGES PER SECOND (IMGS/SEC)

Method	ΔP (M) (↑)	ΔF (G) (↑)	**Throughput (imgs/sec)**		
			bs=1	**bs=4**	**bs=16**
Full tuning	0	0	91.5	375.7	539.5
VPT	0.55	5.6	86.1	283.5	381.5
Adapter	0.16	0.03	70.9	306.6	504.7
AdaptFormer	0.16	0.03	71.4	309.9	508.1
NOAH	0.12	0.02	72.1	312.7	492.9
LoRA	0	0	91.5	375.7	539.6
GLoRA			91.5	375.7	539.6
MLAE			91.5	375.2	538.5

TABLE IV

RESULTS ON VTAB-1K WITH CONVNEXT-B AS THE BACKBONE. ALL RESULTS EXCEPT THOSE OF MLAE ARE QUOTED FROM PRIOR WORK [12], [37], [40]

Method	**Avg.**	**Natural.**	**Specialized.**	**Structured.**
Full	74.0	78.0	83.7	60.4
Linear	63.6	74.5	81.5	34.8
VPT	68.7	78.5	83.0	44.6
LoRA	72.1	79.2	83.4	53.8
SPT-Adapter	78.4	83.7	86.2	65.3
SPT-LoRA	78.7	83.4	86.7	65.9
RepAdapter	79.0	83.5	86.7	66.8
MLAE	**81.6**	**86.1**	**89.4**	**69.2**

TABLE V

RESULTS ON SEMANTIC SEGMENTATION. COMPARISONS ON ADE20K [68] VAL WITH SETR [69] ON VIT-L BACKBONE. WE REPORT BOTH SINGLE-SCALE (S.S.) AND MULTI-SCALE (M.S.) MIOU RESULTS. THE BEST RESULT IS IN BOLD

Method	**Trainable Param.**	**mIoU-s.s. (%)**	**mIoU-m.s. (%)**
VPT	15.8M	44.0	45.6
LoRA	14.7M	43.9	45.9
Adapter	14.6M	44.4	46.6
SPT-LoRA	14.6M	45.4	47.5
SPT-Adapter	14.6M	45.2	47.2
MLAE	14.8M	**46.0**	**47.8**

TABLE VI

RESULTS ON DOMAIN GENERALIZATION. COMPARISON OF DIFFERENT PEFT METHODS ON FOUR DOMAIN-SHIFTED DATASETS (IMAGENET-V2, -SKETCH, -A, -R). MLAE ACHIEVES COMPETITIVE OR SUPERIOR PERFORMANCE OVERALL, SHOWING CLEAR IMPROVEMENTS ON MOST TARGET DOMAINS WHILE REMAINING COMPARABLE ON OTHERS

	Source	**Target**			
	ImageNet	-V2	-Sketch	-A	-R
Adapter	70.5	59.1	16.4	5.5	22.1
VPT	70.5	58.0	18.3	4.6	23.2
LoRA	70.8	59.3	20.0	6.9	23.3
NOAH	71.5	**66.1**	24.8	11.9	**28.5**
MLAE	**73.6**	63.6	**28.7**	**14.0**	28.1

methods, we can see that under a similar budget (e.g., 0.29M trainable parameters for GLoRA [34] vs. 0.30M for MLAE), MLAE performs better on VTAB-1k. In particular, by slightly adjusting the dropout rates across different datasets, MLAE can achieve SOTA performance with an average accuracy score of 78.8%, even better than GLoRA with almost tripled budget (78.0%). MLAE achieves the best performance in 9 out of 19 datasets in the VTAB-1k benchmark and secures the second-best results in 7 additional datasets. Moreover, intuitively from Fig. 3, we can see that compared to the previous LoRA-based PEFT methods, MLAE performs the best in all three domains, especially in the Structured domain, surpassing the previous SOTA method GLoRA by 3.2% under a comparable budget. Notably, SPT [37] reveals that, by measuring the maximum mean discrepancy between the source and target domains, Structured datasets have larger domain gaps from the pre-training source [6] compared to Natural and Specialized datasets. The significant improvement in the Structured domain indicates that MLAE may contribute to mitigating domain gaps by compelling each expert to extract more diverse information from images.

To validate the effectiveness of our method in fine-grained visual classification tasks, we conduct experiments on the FGVC benchmark using three random seeds. The experimental results indicate that MLAE achieves the best performance across all seeds, demonstrating stable performance and parameter efficiency. As shown in Table II, the traditional fine-tuning method FULL, despite using 100% of the parameters, only achieves an accuracy of 88.5%. In contrast, our method achieves an accuracy of 90.9% with only 0.56% trainable parameters. Compared to other PEFT methods, Adapter (85.7%), LoRA-8 (86.0%), SSF (90.7%), and SPT-LoRA (90.1%), MLAE achieves the highest accuracy with the fewest tuned parameters. These results highlight our method's capability to achieve optimal performance with minimal trainable parameters, proving the effectiveness and superiority of MLAE in fine-grained visual classification tasks. More detailed experimental results can be found in XV.

TABLE VII

RESULTS ON LLM FINE-TUNING. ACCURACY (%) COMPARISON OF LLAMA 7B [75], LLAMA2 7B [76], AND LLAMA3 8B [77] WITH VARIOUS PEFT METHODS ON EIGHT COMMONSENSE REASONING DATASETS. RESULTS FOR ALL THE BASELINE METHODS ARE TAKEN FROM [78], [79], WHILE RESULTS FOR MLAE ARE OBTAINED USING THE HYPERPARAMETERS DESCRIBED IN XVII. DORA†: THE ADJUSTED VERSION OF DORA WITH THE RANK HALVED AS IN [78]

Model	PEFT Method	# Params (%)	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Avg.
ChatGPT	-	-	73.1	85.4	68.5	78.5	66.1	89.8	79.9	74.8	77.0
LLaMA-7B	Prefix	0.11	64.3	76.8	73.9	42.1	72.1	72.9	54.0	60.6	64.6
	Series	0.99	63.0	79.2	76.3	67.9	75.7	74.5	57.1	72.4	70.8
	Parallel	3.54	67.9	76.4	78.8	69.8	78.9	73.7	57.3	75.2	72.2
	LoRA	0.83	68.9	80.7	77.4	78.1	78.8	77.8	61.3	74.8	74.7
	DoRA†	0.43	70.0	82.6	79.7	83.2	80.6	80.6	65.4	77.6	77.5
	DoRA	0.84	69.7	83.4	78.6	87.2	81.0	81.9	66.2	79.2	78.4
	MLAE	0.83	70.0	82.8	78.9	86.9	82.6	81.9	65.8	82.2	**78.9**
LLaMA2-7B	LoRA	0.83	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6
	DoRA†	0.43	72.0	83.1	79.9	89.1	83.0	84.5	71.0	81.2	80.5
	DoRA	0.84	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4	79.7
	MLAE	0.83	71.4	83.0	80.4	90.8	83.3	83.5	68.3	83.8	**80.6**
LLaMA3-8B	LoRA	0.70	70.8	85.2	79.9	91.7	84.3	84.2	71.2	79.0	80.8
	DoRA†	0.35	74.5	88.8	80.3	95.5	84.7	90.1	79.1	87.2	85.0
	DoRA	0.71	74.6	89.3	79.9	95.5	85.6	90.5	80.4	85.8	85.2
	MLAE	0.70	75.6	88.9	81.1	95.8	87.9	91.3	80.7	87.8	**86.1**

TABLE VIII

THE MULTI-TASK EVALUATION RESULTS ON VQA, GQA, NLVR^2 AND COCO CAPTION WITH THE VL-BART BACKBONE. THE REPORTED METRICS ARE ACCURACY (%) FOR VQA^{v2}, GQA, AND NLVR^2, AND CIDER SCORE FOR COCO CAPTION

Method	# Params (%)	VQA^{v2}	GQA	NLVR^2	COCO Cap	Avg.
FT	100	66.9	56.7	73.7	112.0	77.3
LoRA	5.93	65.2	53.6	71.9	115.3	76.5
DoRA	5.96	65.8	54.7	73.1	115.9	77.4
MLAE	5.98	66.1	55.4	73.5	115.7	**77.7**

TABLE IX

THE MULTI-TASK EVALUATION RESULTS ON TVQA, HOW2QA, TVC, AND YC2C WITH THE VL-BART BACKBONE. THE REPORTED METRICS ARE ACCURACY (%) FOR TVQA AND HOW2QA, AND CIDER-D SCORE FOR TVC AND YC2C

Method	# Params (%)	TVQA	How2QA	TVC	YC2C	Avg.
FT	100	76.3	73.9	45.7	154	87.5
LoRA	5.17	75.5	72.9	44.6	140.9	83.5
DoRA	5.19	76.3	74.1	45.8	145.4	85.4
MLAE	5.20	76.5	74.9	45.9	146.0	**85.8**

To verify the applicability of MLAE on other vision architectures, following the protocol of VPT [12], we fine-tune ConvNeXt [67] with MLAE and compare it against LoRA, SPT-LoRA, and RepAdapter on the VTAB-1k benchmark. As shown in Table IV, MLAE achieves the best overall performance, with an average accuracy of 81.6%, clearly surpassing RepAdapter (79.0%) and SPT-LoRA (78.7%). Moreover, MLAE consistently outperforms all baselines across the natural, specialized, and structured task groups, with particularly notable gains on structured tasks (+2.4 points over RepAdapter). These results demonstrate that MLAE effectively generalizes beyond transformer-based architectures.

2) Semantic Segmentation: To verify the feasibility of MLAE beyond visual classification, we further apply MLAE to the SETR-PUP [69] framework for semantic segmentation on the ADE20k [68] dataset, following the protocol established in VPT [12]. We report the *val* mIoU results in Table V. Notably, MLAE achieves the best performance among all compared methods, surpassing both LoRA and SPT-LoRA. In particular, it improves the multi-scale mIoU to 47.8%, demonstrating stronger generalization and robustness in dense prediction tasks. These results highlight the effectiveness of MLAE beyond visual classification, showing clear advantages when applied to semantic segmentation.

3) Domain Generalization: The ability to generalize across out-of-domain settings is crucial for large-scale neural networks [70]. Therefore, following NOAH [14], we fine-tune models using MLAE on ImageNet [6] with 16 shots per category, and then directly evaluate them on four ImageNet variants that involve distinct types of domain shifts. From Table VI, we observe that MLAE achieves the highest accuracy on the source domain (ImageNet) and consistently delivers competitive performance on all four domain-shifted datasets. Notably, MLAE surpasses NOAH on the ImageNet-Sketch [71] and ImageNet-A [72] variants by a clear margin, reaching 28.7% and 14.0%, respectively. While NOAH performs slightly better on ImageNet-V2 [73] and ImageNet-R [74], MLAE demonstrates a more balanced improvement across diverse shifts. These results suggest that MLAE generalizes well from the source domain to unseen domains.

4) Efficiency Analysis: Following RepAdapter [40], we present the final inference throughput of various PEFT methods in Table III, all computed on an NVIDIA RTX 3090 GPU. In the table, we employ ViT-B/16 as the vision model. Our method, which builds upon LoRA, maintains the same count

TABLE X
RESULTS OF THREE MASKING STRATEGIES UNDER DIFFERENT CONFIGURATIONS ILLUSTRATED IN FIG. 2 (ALL WITH THE SAME BUDGET OF TOTAL RANK = 96)

Masking strategy	Incremental	Hourglass	Decremental	Protruding	Random	Ours
fixed	76.8	75.5	75.7	76.1	76.5	
stochastic	77.7	76.3	76.0	77.6	—	**78.8**
mixed	78.2	77.5	77.7	78.0	78.1	

TABLE XI
ABLATIONS ON KEY COMPONENTS IN MLAE

Decomposition	Mask	Adaptive	VTAB-1k			
			Natural	Specialized	Structured	Mean Acc.
✓	✗	✗	80.4	86.1	63.4	76.6
✓	✗	✓	80.5	86.0	63.2	76.6
✓	✓	✗	83.9	87.4	64.6	78.6
✓	✓	✓	**84.0**	**87.5**	**64.8**	**78.8**

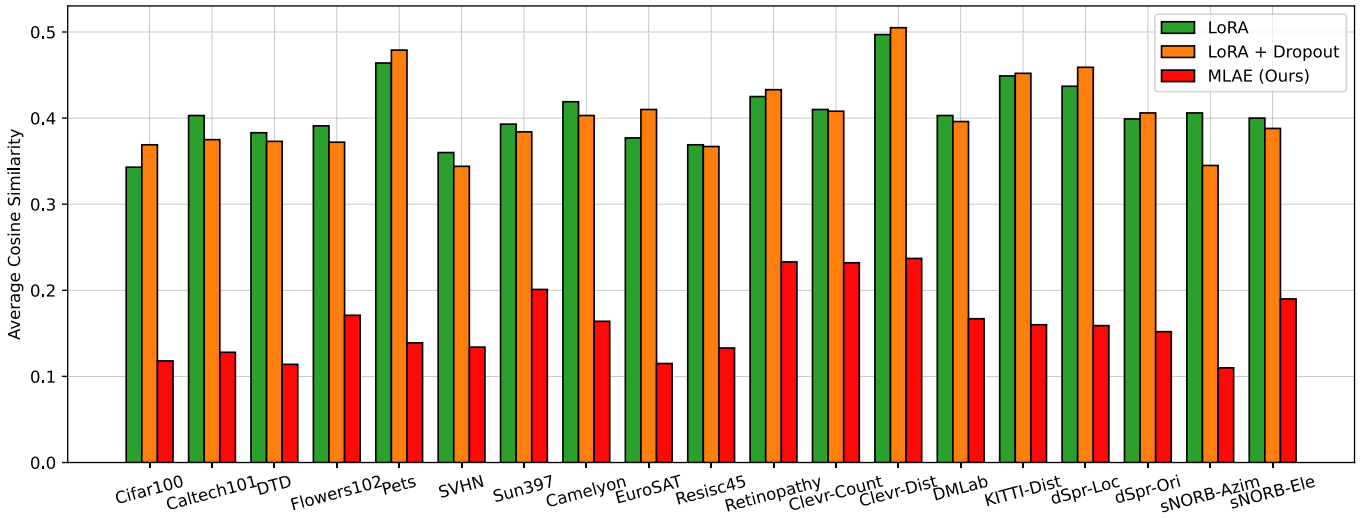


Fig. 4. Comparisons of average parameter similarity between MLAE and LoRA baselines. It is worth noting that LoRA+dropout refers to applying dropout directly to the entire output of LoRA.

TABLE XII
SUBMATRIX RANK

Submatrix	Mean Acc.
$\underbrace{4\ 4}_{2\times4}$	78.2
$\underbrace{2\ 2\ 2\ 2}_{4\times2}$	78.3
$\underbrace{1\ 1\ 1\ 1\ 1\ 1\ 1\ 1}_{8\times1}$	**78.8**

of additional parameters, FLOPs, and throughput as LoRA (i.e., ours: 538.5 imgs/sec vs. 539.6 imgs/sec for LoRA). We also trained both the LoRA and MLAE methods on the VTAB-1k benchmark for 100 epochs per dataset. We recorded the training time for each dataset and repeated the process 5 times to calculate the average training time. Notably, the speed of MLAE is comparable to that of LoRA, with MLAE even demonstrating slight acceleration in certain cases. Specifically, the average training time for MLAE was 1133 seconds, while LoRA took 1149 seconds. The GPU memory consumption for MLAE is 9.8 GB, compared to 9 GB for LoRA.

B. LLM Fine-Tuning

To assess whether the benefits of MLAE generalize from vision backbones to large language models (LLMs), we conduct a comprehensive evaluation in the LLM fine-tuning regime. Specifically, following DoRA [78], we evaluate MLAE against LoRA, DoRA, and several baseline parameter-efficient fine-tuning methods [28], [31], [80], on

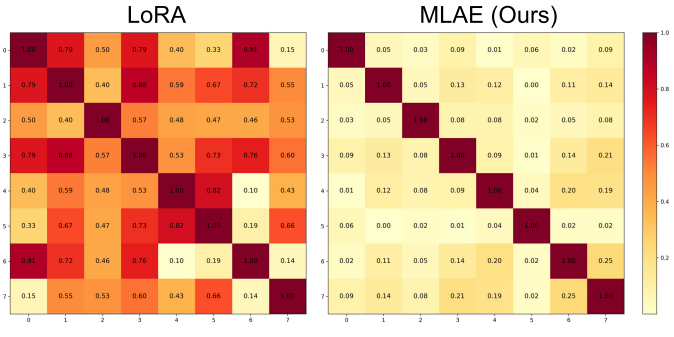


Fig. 5. Confusion matrix comparison of overall parameter similarity on CIFAR-100 ($r = 8$). Compared with LoRA, MLAE exhibits markedly lower similarity, indicating reduced redundancy among experts.

the LLaMA-7B [75] model for commonsense reasoning tasks. This benchmark consists of eight subtasks, each with its own designated training and testing sets. In addition, we further compare MLAE with LoRA and DoRA on the LLaMA2-7B [76] and LLaMA3-8B [77] models. MLAE consistently outperforms existing PEFT methods across all three backbone models. On LLaMA-7B, MLAE achieves the best average accuracy (78.9%), slightly higher than DoRA (78.4%) and notably above LoRA (74.7%). On LLaMA2-7B, MLAE reaches 80.6%, again surpassing both DoRA (79.7%) and LoRA (77.6%). The advantage becomes clearer on LLaMA3-8B, where MLAE obtains 86.1%, leading DoRA (85.2%) and LoRA (80.8%). Overall, MLAE demonstrates robust improvements in commonsense reasoning, with consistent gains as model size increases.

C. Vision-Language Model Fine-Tuning

Following DoRA [78], we evaluate the effectiveness of MLAE for multi-modal model fine-tuning. Specifically, we conduct experiments on multi-task image/video–text understanding using VL-BART.

1) Image/Video-Text Understanding: We compare MLAE with LoRA, DoRA, and full fine-tuning on VL-BART, which comprises a vision encoder (CLIP-ResNet101 [81]) and an encoder–decoder language model ($\text{BART}_{\text{Base}}$ [82]). We evaluate performance on four image–text tasks—VQA^{v2} [83] and GQA [84] for visual question answering, NLVR^2 [85] for visual reasoning, and MSCOCO [86] for image captioning, as well as four video–text tasks from the VALUE benchmark [87]: TVQA [88] and How2QA [89] for video question answering, and TVC [90] and YC2C [91] for video captioning. We follow the same framework as [92] and fine-tune VL-BART using the same setup as LoRA outlined in [92] when applying MLAE. See Table XVIII for the complete hyperparameters. The results of FT, LoRA, and DoRA for both tasks are directly quoted from DoRA [78].

As shown in Tables VIII and IX, MLAE achieves strong results on both image–text and video–text tasks. With only about 6% of trainable parameters, MLAE matches or slightly surpasses full fine-tuning performance, demonstrating high parameter efficiency. On image–text benchmarks, MLAE improves the average score by 0.3% over DoRA, mainly due to better visual reasoning and cross-modal alignment. On video–text tasks, it further gains 0.4%, showing enhanced temporal modeling ability. Overall, MLAE delivers systematic improvements with minimal parameter overhead, confirming its effectiveness for efficient multimodal fine-tuning.

TABLE XIII
TOTAL BUDGET

#Experts	Mean Acc.
$\underbrace{1\ 1}_{2\times1}$	77.3
$\underbrace{1\ 1\ 1\ 1}_{4\times1}$	78.2
$\underbrace{1\ 1\ 1\ 1\ 1\ 1\ 1\ 1}_{8\times1}$	**78.8**

D. Analysis and Visualizations

To gain deeper insights into the underlying mechanisms of MLAE, we first provide a theoretical intuition explaining why expert-level dropout leads to diversified experts and how reducing parameter similarity contributes to performance gains. Second, we conduct analyses and visualizations from the perspectives of parameter similarity and feature attention maps to substantiate our findings. The results indicate that MLAE significantly reduces parameter similarity compared to vanilla LoRA, allowing each expert to extract more diverse feature information, ultimately enhancing model performance.

1) Theoretical Intuition: Previous work [93], [94] has theoretically demonstrated that Dropout can be interpreted as an approximation to a Bayesian posterior over network weights. This Bayesian perspective suggests that Dropout facilitates averaging predictions across a distribution of various network architectures. By randomly omitting neurons during training, Dropout effectively emulates sampling from a distribution of different network subsets. This introduces uncertainty into the model, enabling it to train across multiple sub-networks [95]. Specifically, we employ different random initializations for various experts; due to unique initial conditions and stochasticity in the training process, each expert can converge to a different basin of the posterior, thereby exhibiting diversity. Additionally, [96], [97] enhanced the performance of Bayesian Model Averaging (BMA) [93] by increasing the diversity among ensemble members. Given that Dropout is a mathematical equivalent form of BMA, reducing parameter similarity (under rank-1 decomposition) consequently implies increased diversity among experts which can finally lead to performance gains.

2) Parameter Cosine Similarity Analysis: We explore the differences in cosine similarity among parameters within the incremental matrices learned by LoRA and our proposed MLAE. First, we introduce how to compute the parameter cosine similarity between MLAE and LoRA methods. Since MLAE decomposes the incremental matrix ΔW into r parameter matrices of size $d_{\text{out}} \times d_{\text{in}}$, we can directly flatten each expert's parameter matrix into a one-dimensional tensor, then calculate the cosine similarity between each pair of tensors, and finally average the similarity values within all 12 blocks of the model to represent the overall model's cosine similarity. When calculating the parameter similarity for ΔW learned

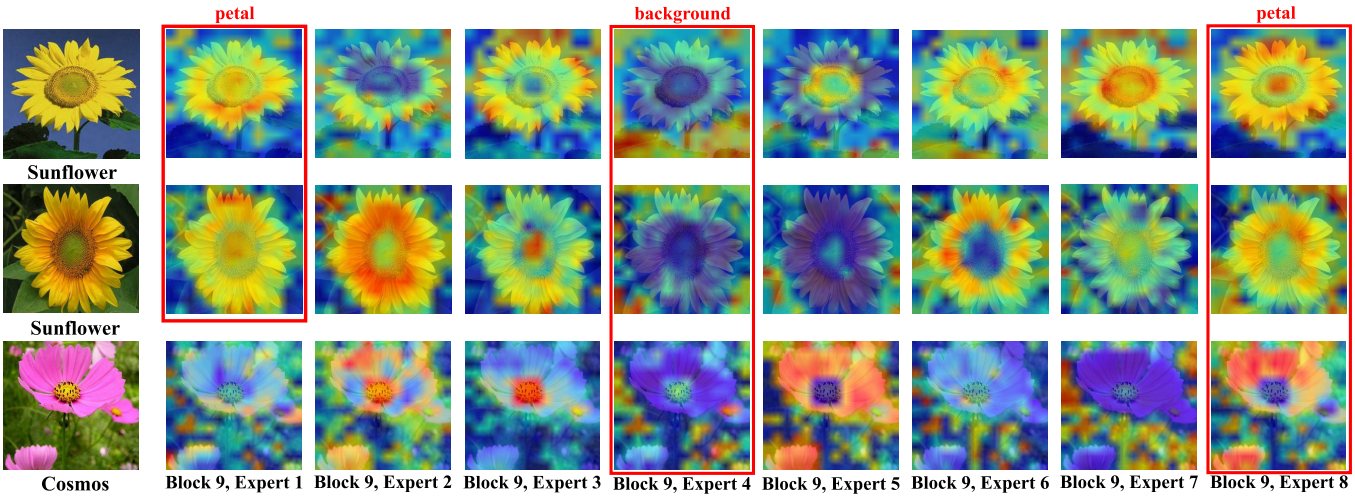


Fig. 6. Visualization of attention patterns within Block 9 of Flowers102 model across different samples. Horizontally, different experts within the same block attend to distinct regions of the same image, such as petals, flower centers, or background, revealing clear expert specialization. Vertically, the same expert often exhibits consistent attention patterns across samples from the same class, while still adapting its focus when the class changes (e.g., from Sunflower to Cosmos). These patterns highlight the diversity and complementarity of expert representations in MLAE.

TABLE XIV
INITIALIZATION

Coeff.	Mean Acc.
0.125	78.0
0.25	78.2
0.5	78.2
1	78.3
2	78.1
4	77.9

TABLE XV
PROBABILITY

p	Mean Acc.
0	76.6
0.1	77.1
0.3	77.3
0.5	77.5
0.7	77.3
0.9	74.4

by LoRA, we perform a special operation by extracting a column from matrix B ($B \in \mathbb{R}^{d_{\text{out}} \times r}$) and a row from matrix A ($A \in \mathbb{R}^{r \times d_{\text{in}}}$) for matrix multiplication, resulting in a parameter submatrix of size $d_{\text{out}} \times d_{\text{in}}$. We can thus obtain r parameter submatrices, and the subsequent processing is the same as for MLAE. Fig. 4 displays the average cosine similarity among model parameters for 19 visual tasks on the VTAB-1k benchmark across three different settings. After performing the computation, Fig. 5 shows the overall parameter similarity confusion matrices on CIFAR-100, comparing LoRA and MLAE (Ours), with MLAE exhibiting significantly lower parameter similarity. We can observe from the Fig. 4: (i) Applying dropout directly on ΔW, i.e., LoRA+Dropout, does not always effectively reduce parameter similarity. This interesting phenomenon suggests that directly applying dropout to the incremental matrix may impact the learning of model parameters due to its randomness, resulting in more uniform parameter updates on certain datasets and thus increasing similarity; (ii) Our method MLAE exhibits lower average cosine similarity across all datasets. Unlike LoRA+Dropout, MLAE performs expert-level dropout by stochastically masking certain experts during the training process. Results from Table I and Fig. 4 demonstrate that MLAE can effectively and significantly reduce parameter similarity, thereby achieving higher model performance.

3) Feature Attention-map Visualization: To visualize the feature representations learned by each expert in MLAE, focusing on a single block, we show a few representative feature attention maps of the same expert across different samples in Fig. 6 and Fig. 7. For Fig. 6, from a horizontal perspective, there are notable variations among experts within the same block. For instance, with the first sunflower, some experts mainly focus on the petals (*Expert 2, 3*), while others concentrate more on the center of the flower (*Expert 5*). Vertically, we notice consistency among experts when focusing on the same class. For example, *Expert 1, 8* focus on the sunflower's petals, while *Expert 4* consistently focuses on the background, etc. When the image switches to another flower (e.g., Cosmos), the attention distribution shifts. While some experts still focus on the petals and background, others shift their attention to the center (*Expert 3*) or the background (*Expert 7*). For Fig. 7, the top two rows show the attention patterns of the same experts for two different images of elephants, demonstrating consistent specialization in focusing on the body, nose, and background across similar samples. The bottom row shows the feature maps for a kangaroo image, showing that the attention patterns of the same experts may vary when applied to a different class of samples. Both figures indicate that the experts in MLAE exhibit a distinct degree of specialization while simultaneously capturing a

TABLE XVI
OPTIMAL VALUES FOR INITIALIZATION AND PROBABILITY ON THE VTAB-1K BENCHMARK

Dataset	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele
Dropout probability	0.7	0.1	0.7	0.7	0.7	0.5	0.7	0.7	0.5	0.5	0.7	0	0	0.5	0.5	0.5	0.7	0	0
Coeff Initialization	2	0.5	0.25	1	2	4	4	1	4	2	0.25	0.125	0.125	2	1	1	4	1	0.25

TABLE XVII
HYPERPARAMETERS OF MLAE FOR FINE-TUNING VL-BART ON IMAGE/VIDEO-TEXT TASKS

Model	LLaMA-7B	LLaMA2-7B	LLaMA3-8B
Rank r		32	
α		64	
Expert dropout		0.5	
Lora dropout		0.05	
LR	3e-4	3e-4	1e-4
LR Scheduler		Linear	
Batch size		16	
Warmup Steps		100	
Epochs		3	
Where		Q, K, V, Up, Down	

TABLE XVIII
HYPERPARAMETERS OF MLAE FOR FINE-TUNING VL-BART ON IMAGE/VIDEO-TEXT TASKS

Hyperparameter (MLAE)	image-text	video-text
Rank r	128	
α	128	
Expert dropout	0.5	0.7
Lora dropout	0.0	
LR	1e-3	3e-4
LR Scheduler	Linear	
Optimizer	AdamW	
Batch size	300	40
Warmup Ratio	0.1	
Epochs	20	7
Where	Q, K	

broad spectrum of feature representations, highlighting the diversity and complementarity of the expert modules in feature extraction and representation. We speculate that such differentiated focus areas and consistency help the model obtain more comprehensive and diverse feature representations when dealing with complex tasks, thereby enhancing the overall performance and generalization ability of the model.

To further quantify the degree of specialization observed in the attention maps, we conduct a statistical analysis across multiple samples. Specifically, we randomly select 200 images from each dataset and feed them into the model. For each sample, we compute the mean Intersection over Union (IoU) between the output feature maps of a given expert and those of the other experts within the same block. This metric provides a quantitative measure of the diversity among expert outputs, where lower IoU values indicate more diverse and specialized representations. For instance, in Block 9 of Flowers102, the mean IoU between expert outputs is 0.26 ± 0.04, while in Block 10 of Caltech101 the value is 0.29 ± 0.06. An IoU of 0.26 implies that only about 26% of the pixels in the feature maps overlap across different experts, which is consistent with our qualitative observations and further supports the claim that experts within the same block capture complementary and diverse feature patterns from the dataset.

E. Masking Strategy Comparisons

We conduct an in-depth investigation of masking strategies with different configurations for MLAE, as illustrated in Fig. 2.

Configurations:

1) Fixed Masking: Recent work such as MoLA [20] manually assigns different numbers of LoRA experts to different layers. Following this idea, we study a similar strategy within MLAE, termed *fixed masking*, where certain layers are assigned fewer experts than the uniform-allocation baseline, so the missing experts can be regarded as permanently masked in a fixed manner. For ViT-B, the layer-wise expert allocation follows five representative patterns across the 12 blocks. The ***incremental*** pattern gradually increases the number of experts, i.e., [2, 2, 2, 6, 6, 6, 10, 10, 10, 14, 14, 14], whereas the ***decremental*** pattern adopts the opposite trend [14, 14, 14, 10, 10, 10, 6, 6, 6, 2, 2, 2]. In contrast, the ***hourglass*** pattern concentrates more experts at the beginning and end, i.e., [14, 14, 14, 2, 2, 2, 2, 2, 2, 14, 14, 14], while the ***protruding*** pattern allocates more experts to the middle blocks [2, 2, 2, 14, 14, 14, 14, 14, 14, 2, 2, 2]. For the ***random*** pattern, each layer is assigned an integer number of experts between 1 and 14, with the average across layers constrained to 8. All configurations are evaluated under the same trainable parameter budget.

2) Stochastic Masking: We further explore *stochastic masking*, where each layer is assigned 8 experts for fair comparison and expert-level masking is introduced by applying dropout to the expert coefficients so that zeroed coefficients deactivate the corresponding experts during training. In this way, layer-wise variation is controlled by dropout probability rather than expert allocation, and the pattern names therefore refer to the expected number of active experts rather than the dropout rates themselves. For ViT-B, we define five representative dropout profiles over the 12 blocks. The ***incremental*** profile gradually decreases the dropout probability toward deeper blocks, thereby retaining more active experts in later blocks, while the ***decremental*** profile follows the opposite trend. The ***hourglass*** profile uses higher dropout probabilities in the middle blocks, whereas the ***protruding*** profile uses lower ones. Their corresponding schedules are [0.8, 0.8, 0.7, 0.7, 0.6, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0], [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6,

TABLE XIX
PERFORMANCE COMPARISONS ON FIVE FGVC DATASETS WITH VIT-B/16 MODELS PRE-TRAINED ON IMAGENET-21K

Method \ Dataset	CUB-200-2011	NABirds	Oxford Flowers	Stanford Dogs	Stanford Cars	Mean	Params. (M)
Full fine-tuning	87.3	82.7	98.8	89.4	84.5	88.54	85.98
Linear probing	85.3	75.9	97.9	86.2	51.3	79.32	0.18
Adapter	87.1	84.3	98.5	89.8	68.6	85.67	0.41
Bias	88.4	84.2	98.8	91.2	79.4	88.41	0.28
VPT-Shallow	86.7	78.8	98.4	90.7	68.7	84.62	0.25
VPT-Deep	88.5	84.2	99.0	90.2	83.6	89.11	0.85
SPT-LoRA	88.6	83.4	99.5	91.4	87.4	90.06	0.69
SSF	89.5	**85.7**	**99.6**	89.6	**89.2**	90.72	0.39
MLAE (seed 0)	89.6	84.5	99.2	91.6	89.0	90.78	0.29
MLAE (seed 1)	89.6	85.1	99.2	91.7	89.1	90.94	0.29
MLAE (seed 42)	89.6	85.3	99.3	91.5	88.5	90.84	0.29
MLAE (Avg.)	**89.6**	85	99.2	**91.6**	88.9	**90.86**	0.29

TABLE XX
DETAILED RESULTS OF DIFFERENT RANK SUBMATRIX ON VTAB-1K BENCHMARK VARIANTS

Sub-matrix Rank	Natural							Specialized				Structured								Average
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	
$\underbrace{4\ 4}_{2\times4}$	77.0	94.4	74.5	99.3	92.5	89.8	58.1	87.3	96.6	88.1	76.5	84.4	67.0	55.5	82.1	86.2	56.2	33.7	46.1	78.2
$\underbrace{2\ 2\ 2\ 2}_{4\times2}$	77.7	94.6	74.5	99.4	92.5	88.9	58.2	88.4	96.3	88.4	76.5	82.6	65.3	54.4	82.6	87.8	55.8	35.7	47.3	78.3
$\underbrace{1\ 1\ 1\ 1\ 1\ 1\ 1\ 1}_{8\times1}$	77.8	94.7	74.8	99.4	92.6	90.6	58.4	88.4	96.5	88.5	76.7	84.3	67.5	55.7	82.6	87.8	57.1	35.7	47.7	78.8

TABLE XXI
DETAILED RESULTS OF DIFFERENT BUDGETS ON VTAB-1K BENCHMARK

# Experts	Natural							Specialized				Structured								Average
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	
$\underbrace{1\ 1}_{2\times1}$	75.5	92.3	73.7	99.3	92.1	87.9	58.2	86.8	95.9	87.7	75.8	83.4	66.0	51.4	80.2	82.5	56.7	34.7	47.1	77.3
$\underbrace{1\ 1\ 1\ 1}_{4\times1}$	76.7	94.2	74.4	99.3	92.4	89.6	58.1	87.4	96.1	88.5	76.1	83.8	68.3	53.4	80.7	86.9	56.5	35.4	47.4	78.2
$\underbrace{1\ 1\ 1\ 1\ 1\ 1\ 1\ 1}_{8\times1}$	77.8	94.7	74.8	99.4	92.6	90.6	58.4	88.4	96.5	88.5	76.7	84.3	67.5	55.7	82.6	87.8	57.1	35.7	47.7	78.8

0.6, 0.7, 0.7, 0.8, 0.8], [0.0, 0.1, 0.3, 0.5, 0.6, 0.8, 0.8, 0.6, 0.5, 0.3, 0.1, 0.0], and [0.8, 0.6, 0.5, 0.3, 0.1, 0.0, 0.0, 0.1, 0.3, 0.5, 0.6, 0.8], respectively. We also consider a uniform dropout rate across all layers (**Ours**).

3) Mixed Masking: Combining the two masking schemes above, we further investigate a third strategy, termed *mixed masking*, which preserves the layer-wise expert allocation of fixed masking while applying a uniform dropout rate across all layers. We use a shared dropout rate because stochastic masking shows that this setting yields the best performance. This experiment is designed to examine whether expert-level stochastic masking can further improve fixed masking, and more broadly, to validate the general effectiveness of our masking design.

Results. All experiments with the above settings are conducted under an identical budget. From Table X, we have the following findings: (i) The configuration with a uniform dropout rate applied across all layers through stochastic masking (Ours) outperforms the other configurations, achieving a score of 78.8%. (ii) Both stochastic and mixed masking consistently outperform fixed masking, indicating that dynamic expert-level masking is preferable. These results support the

TABLE XXII
DETAILED RESULTS OF DIFFERENT INITIALIZATIONS ON VTAB-1K BENCHMARK

Initialization	Natural							Specialized				Structured								Average
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	
0.125	76.9	94.7	74.5	99.3	92.3	87.6	58.1	87.7	96.2	87.8	76.3	84.3	67.2	52.9	81.0	87.4	55.8	34.0	47.7	78.0
0.25	77.2	94.4	74.8	99.4	92.4	87.8	58.1	88.1	96.2	88.1	76.7	84.0	65.5	53.6	81.3	87.1	55.8	35.2	47.7	78.2
0.5	77.5	94.7	74.5	99.4	92.6	88.6	58.1	87.9	96.2	88.1	75.7	83.6	66.3	54.6	82.0	87.7	55.7	35.3	47.4	78.2
1	77.7	94.6	74.5	99.4	92.5	88.9	58.2	88.4	96.3	88.4	76.5	82.6	65.3	54.4	82.7	87.8	55.8	35.7	47.3	78.3
2	77.8	94.3	74.3	99.4	92.6	89.1	58.2	88.1	96.3	88.5	75.7	81.4	64.1	55.7	82.0	87.7	56.8	35.3	46.0	78.1
4	77.0	93.6	73.8	99.4	92.5	90.6	58.4	87.2	96.5	88.5	75.3	80.7	64.2	55.5	81.9	87.1	57.1	34.8	44.7	77.9

TABLE XXIII
DETAILED RESULTS OF PROBABILITY ON VTAB-1K BENCHMARK (WITHOUT THE BEST INITIALIZATION)

p	Natural							Specialized				Structured								Average
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	
0	70.7	94.9	70.2	99.1	90.1	88.0	53.0	88.0	95.9	86.3	73.6	82.3	64.7	52.9	81.6	84.7	53.0	36.1	48.1	76.6
0.1	72.4	94.9	72.4	99.3	91.3	88.2	54.3	88.4	96.3	87.1	75.6	81.6	64.3	53.0	81.7	86.9	53.6	35.9	44.5	77.1
0.3	74.9	94.8	73.0	99.4	91.3	88.4	56.3	88.3	96.2	87.7	76.1	77.7	62.8	54.4	81.0	87.4	54.8	35.4	43.9	77.3
0.5	76.5	94.6	74.1	99.3	91.9	88.5	57.4	88.1	96.6	88.4	76.1	74.8	62.5	54.5	82.8	87.8	55.6	35.2	43.5	77.5
0.7	77.6	93.8	74.2	99.3	92.5	88.7	58.3	88.6	96.4	88.3	76.3	71.3	62.3	54.5	80.7	87.3	56.1	33.3	41.8	77.3
0.9	76.2	91.5	73.2	99.3	92.2	85.0	58.1	86.6	96.1	87.4	76.8	61.6	57.5	51.5	79.6	65.4	54.2	26.8	36.7	74.4

effectiveness of stochastic masking by promoting more diverse and complementary expert representations.

F. Ablation Study

Key Components. To thoroughly evaluate the proposed MLAE, we conduct extensive ablation studies. In Table XI, we evaluate the impact of three key components on model performance. Results show that when using cellular decomposition alone, the model achieves an average accuracy of only 76.6%. However, the performance significantly improves by 2.0% when both cellular decomposition and expert-level masking techniques are applied. Employing all three techniques together elevates the model's performance to a new SOTA accuracy of 78.8%. This indicates that while cellular decomposition serves as a solid foundation, expert-level masking can significantly enhance performance, and the adaptive coefficients technique further optimizes model behavior.

Submatrix Rank and Budget. While fine-tuning the matrix ΔW under the same budget of trainable parameters, we assess the effect of different expert rank configurations. The results in Table XII indicate that the rank-1 submatrix configuration achieves the highest performance. Next, we explore the impact of tuning ΔW using different numbers of rank-1 experts in Table XIII. The results indicate that increasing the number of rank-1 experts can significantly improve model performance, possibly because more experts can process diverse features more finely, thus enhancing the overall performance. However, it is worth noting that indiscriminately increasing the number of experts may not continuously yield performance gains and instead can increase computational demands and resource consumption. Therefore, choosing an appropriate number of experts to balance performance improvement and resource consumption is crucial.

Hyperparameters of Initialization and Probability. Finally, inspired by RepAdapter [40] for searching the hyperparameter scaling coefficient, we also explore the initialization values of the coefficient matrices and the dropout probability in our method in Table XIV and Table XV. The first observation is that an initialization value of 1 for the coefficient matrices yields slightly better performance while a probability of 0.5 achieves the best performance on the VTAB-1k benchmark. However, it is noteworthy that the optimal values for the initialization and probability may vary across different datasets to achieve peak performance.

V. CONCLUSION

In this paper, we propose MLAE, a novel visual PEFT method that introduces masking into LoRA to reduce redundancy and improve parameter quality. By decomposing a low-rank update into multiple rank-1 experts and selectively activating them during training, MLAE encourages more independent and diverse learning. Through extensive exploration of fixed, stochastic, and mixed masking strategies, we find that stochastic expert-level masking via dropout achieves the best overall performance. Further analysis shows that MLAE reduces parameter similarity among experts and helps the model acquire more diverse knowledge.

Extensive experiments across the VTAB-1k and FGVC benchmarks demonstrate the effectiveness and superiority

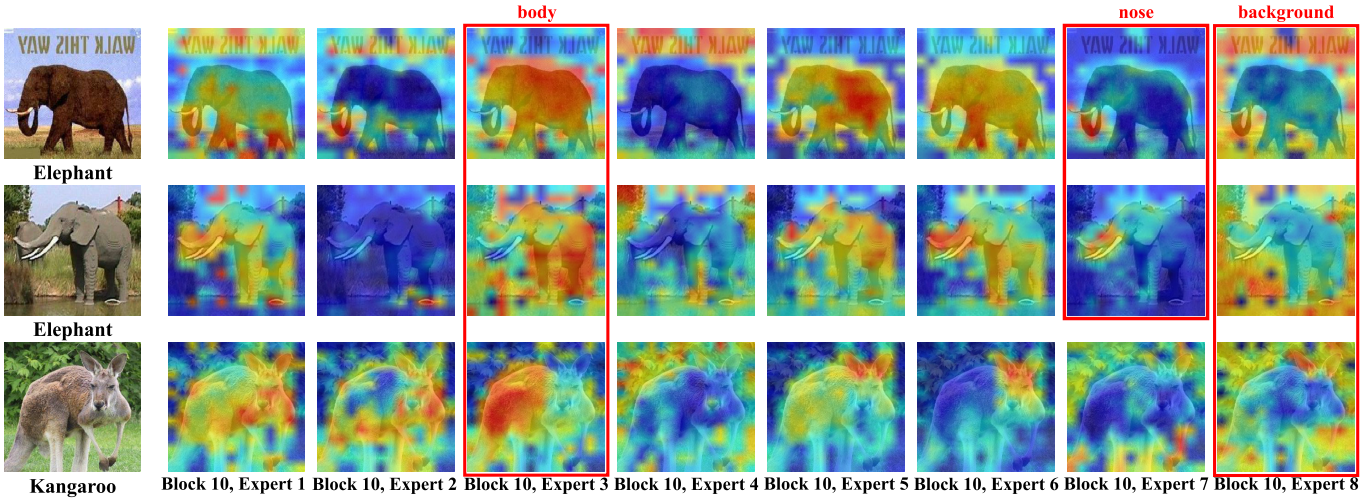


Fig. 7. Visualization of feature maps in Block 10 of Caltech101 model across different samples. The same experts show consistent focus on semantic regions such as the body, nose, and background across elephant images, while shifting their attention on a kangaroo image, revealing both specialization and adaptability to different semantic categories.

of MLAE. Moreover, MLAE also shows strong generalization ability on LLM fine-tuning, semantic segmentation, and image/video-text understanding, highlighting its flexibility and potential as a simple yet effective direction for PEFT.

APPENDIX
DATASET DETAILS

See Tables XVI–XXIII and Fig. 7.

REFERENCES

[1] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[2] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[3] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[4] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12104–12113.

[5] S. Zhao, R. Quan, L. Zhu, and Y. Yang, "CLIP4STR: A simple baseline for scene text recognition with pre-trained vision-language model," *IEEE Trans. Image Process.*, vol. 33, pp. 6893–6904, 2024.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[7] R. Goyal et al., "The 'something something' video database for learning and evaluating visual common sense," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5842–5850.

[8] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[9] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. 1st Workshop Fine-Grained Vis. Categorization*, Jun. 2011, pp. 1–2.

[10] B. Zhou et al., "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019.

[11] S. Chen et al., "AdaptFormer: Adapting vision transformers for scalable visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 16664–16678.

[12] M. Jia et al., "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 709–727.

[13] S. Jie and Z.-H. Deng, "Convolutional bypasses are better vision transformer adapters," 2022, *arXiv:2207.07039*.

[14] Y. Zhang, K. Zhou, and Z. Liu, "Neural prompt search," 2022, *arXiv:2206.04673*.

[15] E. J. Hu et al., "Lora: Low-rank adaptation of large language models," in *Proc. 10th Int. Conf. Learn. Represent.*, Apr. 2022, pp. 1–26.

[16] J. Liao, X. Xu, M. C. Nguyen, A. Goodge, and C. S. Foo, "COFT-AD: COntrastive fine-tuning for few-shot anomaly detection," *IEEE Trans. Image Process.*, vol. 33, pp. 2090–2103, 2024.

[17] J. Dong, Y. Yao, W. Jin, H. Zhou, Y. Gao, and Z. Fang, "Enhancing few-shot out-of-distribution detection with pre-trained model features," *IEEE Trans. Image Process.*, vol. 33, pp. 6309–6323, 2024.

[18] S. Wang, C. Li, Z. Yan, W. Liang, Y. Yuan, and G. Wang, "HAda: Hyper-adaptive parameter-efficient learning for multi-view ConvNets," *IEEE Trans. Image Process.*, vol. 34, pp. 85–99, 2025.

[19] Z. Yu, D. Shen, Z. Jin, J. Huang, D. Cai, and X.-S. Hua, "Progressive transfer learning," *IEEE Trans. Image Process.*, vol. 31, pp. 1340–1348, 2022.

[20] C. Gao et al., "Higher layers need more LoRA experts," 2024, *arXiv:2402.08562*.

[21] N. Ding et al., "Sparse low-rank adaptation of pre-trained language models," 2023, *arXiv:2311.11696*.

[22] J. Cheng et al., "Revisiting LoRA through the lens of parameter redundancy: Spectral encoding helps," 2025, *arXiv:2506.16787*.

[23] Q. Zhang et al., "Adaptive budget allocation for parameter-efficient fine-tuning," in *Proc. 11th Int. Conf. Learn. Represent.*, Kigali, Rwanda, May 2023. [Online]. Available: https://openreview.net/forum?id=lq62uWRJjiY

[24] F. Zhang, L. Li, J. Chen, Z. Jiang, B. Wang, and Y. Qian, "IncreLoRA: Incremental parameter allocation method for parameter-efficient fine-tuning," 2023, *arXiv:2308.12043*.

[25] Y. Zhu et al., "SiRA: Sparse mixture of low rank adaptation," 2023, *arXiv:2311.09179*.

[26] X. Wu, S. Huang, and F. Wei, "Mole: Mixture of LoRA experts," in *Proc. 12th Int. Conf. Learn. Represent.*, Vienna, Austria, May 2024, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=uWvKBCYh4S

[27] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," 2017, *arXiv:1701.06538*.

[28] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.

[29] Y.-C. Liu, C.-Y. Ma, J. Tian, Z. He, and Z. Kira, "Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36889–36901.

[30] B. Dong, P. Zhou, S. Yan, and W. Zuo, "LPT: Long-tailed prompt tuning for image classification," in *Proc. 11th Int. Conf. Learn. Represent.*, Kigali, Rwanda, May 2023, pp. 1–20.

[31] X. Lisa Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," 2021, *arXiv:2101.00190*.

[32] S. Jie and Z.-H. Deng, "Fact: Factor-tuning for lightweight adaptation on vision transformer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 1060–1068.

[33] B. X. B. Yu, J. Chang, L. Liu, Q. Tian, and C. Wen Chen, "Towards a unified view on visual parameter-efficient transfer learning," 2022, *arXiv:2210.00788*.

[34] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-all: Generalized LoRA for parameter-efficient fine-tuning," 2023, *arXiv:2306.07967*.

[35] Z. Jiang, C. Mao, Z. Huang, Y. Lv, D. Zhao, and J. Zhou, "Rethinking efficient tuning methods from a unified perspective," 2023, *arXiv:2303.00690*.

[36] Q. Gao et al., "A unified continual learning framework with general parameter-efficient tuning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11449–11459.

[37] H. He, J. Cai, J. Zhang, D. Tao, and B. Zhuang, "Sensitivity-aware visual parameter-efficient fine-tuning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11791–11801.

[38] C.-L. Fu, Z.-C. Chen, Y.-R. Lee, and H.-y. Lee, "AdapterBias: Parameter-efficient token-dependent representation shift for adapters in NLP tasks," 2022, *arXiv:2205.00305*.

[39] D. Lian, D. Zhou, J. Feng, and X. Wang, "Scaling & shifting your features: A new baseline for efficient model tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 109–123.

[40] G. Luo et al., "Towards efficient visual adaption via structural re-parameterization," 2023, *arXiv:2302.08106*.

[41] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[42] J. Zhou et al., "IBOT: Image BERT pre-training with online tokenizer," 2021, *arXiv:2111.07832*.

[43] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 162, 2022, pp. 1298–1312.

[44] X. Geng, H. Liu, L. Lee, D. Schuurmans, S. Levine, and P. Abbeel, "Multimodal masked autoencoders learn transferable representations," 2022, *arXiv:2205.14204*.

[45] T. Arici et al., "MLIM: Vision-and-language model pre-training with masked language and image modeling," 2021, *arXiv:2109.12178*.

[46] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto, "Masked vision and language modeling for multi-modal representation learning," 2022, *arXiv:2208.02131*.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[48] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9653–9663.

[49] K. Książek and P. Spurek, "HyperMask: Adaptive hypernetwork-based masks for continual learning," 2023, *arXiv:2310.00113*.

[50] A. Mallya and S. Lazebnik, "PackNet: Adding multiple tasks to a single network by iterative pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7765–7773.

[51] S. Golkar, M. Kagan, and K. Cho, "Continual learning via neural pruning," 2019, *arXiv:1903.04476*.

[52] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4548–4557.

[53] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang, "Long live the lottery: The existence of winning tickets in lifelong learning," in *Proc. 9th Int. Conf. Learn. Represent.*, May 2021, pp. 9533–9551. [Online]. Available: https://openreview.net/forum?id=LXMSvPmsm0g

[54] H. Kang et al., "Forget-free continual learning with winning subnetworks," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 10734–10750.

[55] J. Wu, W. Lin, G. Shi, X. Wang, and F. Li, "Pattern masking estimation in image with structural uncertainty," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4892–4904, Dec. 2013.

[56] H. Dadkhahi and M. F. Duarte, "Masking strategies for image manifolds," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4314–4328, Sep. 2016.

[57] J. Zhu, H. Ma, J. Chen, and J. Yuan, "High-quality and diverse few-shot image generation via masked discrimination," *IEEE Trans. Image Process.*, vol. 33, pp. 2950–2965, 2024.

[58] X. Zhai et al., "A large-scale study of representation learning with the visual task adaptation benchmark," 2019, *arXiv:1910.04867*.

[59] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200–2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, Jul. 2011.

[60] G. Van Horn et al., "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 595–604.

[61] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, India, Dec. 2008, pp. 722–729.

[62] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and L. Fei-Fei, "Fine-grained car detection for visual census estimation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017, pp. 4502–4508.

[63] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in Adam," 2018, *arXiv:1711.05101*.

[64] L. Ren, C. Chen, L. Wang, and K. Hua, "DA-VPT: Semantic-guided visual prompt tuning for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 4353–4363.

[65] L. Veeramacheneni, M. Wolter, H. Kuehne, and J. Gall, "Canonical rank adaptation: An efficient fine-tuning strategy for vision transformers," in *Proc. 42nd Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 267, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., Jul. 2025, pp. 61108–61125. [Online]. Available: https://proceedings.mlr.press/v267/veeramacheneni25a.html

[66] T. Chen et al., "Sensitivity-aware efficient fine-tuning via compact dynamic-rank adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 9655–9664.

[67] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986.

[68] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 633–641.

[69] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[70] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023.

[71] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 10506–10518.

[72] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15257–15266.

[73] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5389–5400.

[74] D. Hendrycks et al., "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8320–8329.

[75] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[76] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[77] A. Grattafiori et al., "The llama 3 herd of models," 2024, *arXiv:2407.21783*.

[78] S.-Y. Liu et al., "DoRA: Weight-decomposed low-rank adaptation," 2024, *arXiv:2402.09353*.

[79] Z. Hu et al., "LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models," 2023, *arXiv:2304.01933*.

[80] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," 2021, *arXiv:2110.04366*.

[81] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[82] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[83] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6904–6913.

[84] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6700–6709.

[85] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," 2018, *arXiv:1811.00491*.

[86] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[87] L. Li et al., "VALUE: A multi-task benchmark for video-and-language understanding evaluation," 2021, *arXiv:2106.04632*.

[88] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "TVQA: Localized, compositional video question answering," 2018, *arXiv:1809.01696*.

[89] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "HERO: Hierarchical encoder for Video+Language omni-representation pre-training," 2020, *arXiv:2005.00200*.
[90] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVR: A large-scale dataset for video-subtitle moment retrieval," in *Proc. 16th Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2020, pp. 447–463.
[91] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, pp. 7590–7598.
[92] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 5227–5237.
[93] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
[94] S.-I. Maeda, "A Bayesian encourages dropout," 2014, *arXiv:1412.7003*.
[95] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 6402–6413.
[96] Y. Grewal and T. D. Bui, "Diversity is all you need to improve Bayesian model averaging," in *Proc. Bayesian Deep Learn. Workshop 35th Conf. Neural Inf. Process. Syst.*, Dec. 2021, pp. 1–4.
[97] G. Nam, J. Yoon, Y. Lee, and J. Lee, "Diversity matters when learning from ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8367–8377.

Junjie Wang received the B.S. degree in artificial intelligence from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2025. He is currently pursuing the Ph.D. degree with Peking University, Beijing. His research interests include machine learning and computer vision.

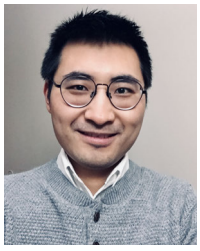
Guangjing Yang received the B.S. degree in information engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2022, where he is currently pursuing the Ph.D. degree in artificial intelligence with the School of Artificial Intelligence. His current research interests include parameter-efficient fine-tuning of large models and large vision-language models.

Wentao Chen received the B.S. degree in information engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2023. He is currently pursuing the joint Ph.D. degree with Shanghai Jiao Tong University (SJTU) and the University of Michigan—Shanghai Jiao Tong University Joint Institute.

Huahui Yi received the B.Eng. degree in computer science from the Civil Aviation University of China in 2022. He is currently a Research Assistant at West China Biomedical Big Data Center, West China Hospital, Sichuan University. His primary research interests are multimodality learning, continual learning, and representation learning.

Xiaohu Wu received the Ph.D. degree from Télécom Paris of Institut Polytechnique de Paris, France, in 2016. He is currently a Faculty Member at Beijing University of Posts and Telecommunications. He has held research or visiting positions at Nanyang Technological University, Singapore, Fondazione Bruno Kessler, Trento, Italy, Aalto University, Espoo, Finland, the University of California, Berkeley, and Singapore University of Technology and Design. His research interests include machine learning, modeling and performance analysis, optimization, and network economics.

Zhouchen Lin (Fellow, IEEE) received the Ph.D. degree in applied mathematics from Peking University in 2000. He is currently a Boya Special Professor with the State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University. He has published over 340 technical articles, collecting more than 38 000 Google Scholar citations. His research interests include machine learning and numerical optimization. He is a fellow of IAPR, AAIA, and CSIG. He has been the Program Co-Chair of ICPR 2022, the Area Chair of ACML, ACCV, CVPR, ICCV, NIPS/NeurIPS, AAAI, IJCAI, ICLR, and ICML, and the Senior Area Chair of ICML, NeurIPS, CVPR, and ICLR for many times. He was an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and currently is an Associate Editor of *International Journal of Computer Vision* and *Optimization Methods and Software*.



Qicheng Lao received the B.S. degree in medicine from Fudan University, China, the M.Sc. degree in experimental medicine from McGill University, Canada, and the Ph.D. degree in computer science from Concordia University, Canada. He was a Post-Doctoral Fellow with the Montreal Institute for Learning Algorithms (MILA), University of Montreal. He is currently a Professor with Beijing University of Posts and Telecommunications (BUPT). His research interests include multimodal representation learning and machine learning methods applied to healthcare.