MED-DSLC: Multi-Expert-Domain Classification via Domain Supervision and Logit Calibration

Zheng Zeng*[1], Deepak Sridhar*[1], and Nuno Vasconcelos[1]

University of California, San Diego, USA
{zhz396,desridha}@ucsd.edu
*Equal contribution

Abstract. Vision-language models (VLMs) such as CLIP enable zero-shot classification by comparing image features with text prompts in a shared embedding space. A fundamental property underlying this capability is the global comparability of logits across arbitrary candidate classes. However, VLMs are often adapted to fine-grained domains using techniques such as LoRA. While this improves in-domain accuracy, out-of-domain accuracy degrades. This leads to a higly fragmented model ecosystem, with thousands of specialized models. *Multi-Expert-Domain* (`MED`) classification seeks to address this problem, by merging LoRAs trained independently on specialized domains. However, due to the independent training, the various domain experts no longer produce globally calibrated logits. As a result, when evaluating over the union of multiple domain-specific class sets, heterogeneous logit scales induce cross-domain interference and artificially high confidence for out-of-domain classes, inducing prediction errors. In this work, we identify domain supervision and cross-domain logit miscalibration as the key issue to scalable multi-domain zero-shot recognition. We propose a mixture-of-experts `MED` architecture, `MED-DSLC`, combining domain supervised training and domain-wise logit scaling, to explicitly restore global logit comparability. `MED-DSLC` is a lightweight solution for `MED` classification, which is shown to preserve within-domain discrimination while reducing cross-domain logit interference with minimal data. Extensive experiments across diverse fine-grained benchmarks demonstrate that it substantially improves mean accuracy (+15%), cross-domain robustness, and scalability in the size of `MED` classification problem. Our results show that restoring output-level calibration is essential under highly data imbalanced settings for achieving a truly zero-shot VLM under multi-domain specialization. Code is publicly available at MED-DSLC.

Keywords: Few-shot Open-set Recognition · Adapter Merging · Mixture of Experts · Generalization

1 Introduction

Vision-language models (VLMs), such as CLIP [16], have transformed visual recognition by enabling zero-shot classification with open class vocabularies. By

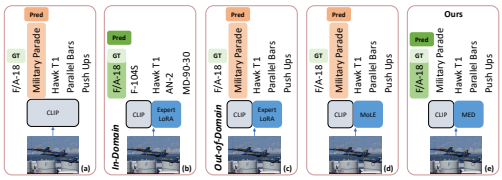

Fig. 1: Logits for classification of an image of the class 'FA-18'. Only largest logit value is shaded. (a) CLIP logits. (b) Logits of LoRA adapted domain expert trained on the image domain. (c) Logits of an expert trained on another domain. (d) Logit miscalibration across domains of an existing `MED` model (MoLE). (e) Logits of the proposed `MED-DSLC` classifier.

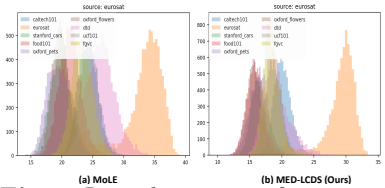

Fig. 2: Logit histograms for images of the EuroSAT dataset and classes of a `MED` classifier with 10 domain experts (identified by color). (a) Logits produced by MoLE exhibit overlap between domains (EuroSAT and DTD). (b) The overlap is significantly reduced by `MED-DSLC`.

aligning images and text in a shared embedding space, these models can classify images with respect to arbitrary class names without retraining. Given feature vectors for image and class labels, a similarity score (logit) is computed per label and a softmax function used to output class probabilities. A central property underlying the zero shot capability is that the logits of all classes are globally comparable, allowing softmax normalization over any set of labels. Despite this strong zero-shot generalization, performance often degrades on fine-grained or domain-specific datasets involving image classes not commonly found in the public domain. This is illustrated in Fig. 1(a), which visualizes the logits (only largest one shaded) of the CLIP model for the image shown and multiple-domain class labels. While CLIP maintains a single globally calibrated logit space, it performs poorly due to lack of domain knowledge. In this case, it fails to classify the image into the fine-grained class of 'F-18 jets', instead assigning it to the coarser and only loosely related class of 'Military Parade'.

To address this limitation, parameter-efficient finetuning (PEFT) methods, such as prompt tuning [26] and low-rank adaptation (LoRA) [5], are widely used to specialize VLMs to individual domains. In practice, however, this results in a collection of independently adapted domain experts, each achieving strong in-domain accuracy but typically exhibiting weak performance outside that domain. This is illustrated In Fig. 1(b-c), where domain-experts achieve strong in-domain performance but poor out-of-domain generalization. While in Figure 1 (b) a domain-expert LoRA, trained on an airplane dataset, assigns the image to the correct airplane class, in Figure 1 (c) an out-of domain expert (LoRA trained on another dataset) produces an even larger logit for the original CLIP prediction. This creates an extremely fragmented model ecosystem where, for each VLM, many specialized adapters must be maintained. Furthermore, for many applications, the appropriate domain is typically unknown at test time or there is a need to combine classes from many domains. In this case, it is not even clear which domain expert to use.

This problem has created interest in the multi-expert-domain (`MED`) task, where the goal is to merge adapter parameters or their outputs [3, 4, 21] to consolidate several domain-specific experts into a single unified model. While

merging alleviates model fragmentation, it is not trivial to guarantee that the MED model preserves the performance of the individual experts across domains. There are two major difficulties. The first is the routing problem, namely to decide which expert to use for each image at test time. A recent trend in the literature [12, 21], which we follow, is to rely on a mixture of experts (MoE) architecture [18], where a gating network selects the expert LoRA. However, these approaches follow a literal MoE implementation, where domain assignments are learned without supervision. Hence, they inherit the MoE problem of expert load balancing [2, 10], which is not trivial to solve. We note that, for MED classifiers, domain labels can be easily derived from class labels, which are always available for training. We thus propose an alternative MoE training scheme, based on supervised domain learning.

A second problem is that independently adapted experts implicitly reshape both the geometry and magnitude of their output logits. When candidate labels from multiple domains are evaluated jointly, either through naïve merging or mixture-based aggregation, logits originating from different domain specialists directly compete under softmax normalization. Because these logits are not calibrated across domains, systematic scale discrepancies arise. Domains producing higher-magnitude logits can dominate predictions even for out-of-domain inputs, leading to artificial overconfidence and degraded cross-domain accuracy. This is shown in Fig. 1(d), for a model (MoLE) that combines logits from different experts, including those of Fig. 1(b)-(c), by mixture-based aggregation [21]. While each expert has independently reshaped its logit geometry and scale, the expert logits compete directly under a softmax that spans all classes. Since the expert logits are no longer comparable, this results in cross-domain interference: domains producing larger-magnitude logits dominate predictions, even for out-of-domain inputs. Fig. 2 (a) illustrates the problem by showing the histogram of logits produced for the various classes of a MED problem by the MoE classifier, for images of the EuroSAT dataset. Logit miscalibration produces significant overlap between logits of EuroSAT classes and those of other domains. Importantly, this phenomenon intensifies as the class imbalance or the number of specialized domains increases, fundamentally undermining scalability.

While domain supervision mitigates this problem, there can still be interference from out-of-domain experts that produce unusually high logits. To overcome this problem, we augment domain supervision with a simple and efficient domain-wise logit scaling. This consists of learning a temperature parameter per domain to normalize its logits before joint softmax evaluation. This domain-wise logit scaling preserves within-domain discrimination while aligning cross-domain magnitudes, preventing any single domain from dominating predictions. We denote the combination of domain supervised routing and logit calibration ***MED*** *by Domain Supervision and Logit Calibration* (MED-DSLC). As shown in Fig. 1(e), this combination restores cross-domain logit alignment, enabling the correct predictions. Fig. 2 (b) shows that MED-DSLC learns to better separate the logits of classes from different domains. Extensive experiments across diverse fine-grained recognition benchmarks demonstrate that this combination substantially im-

proves cross-domain robustness and stabilizes performance as the number of domains grows. Compared to existing MED strategies, our approach achieves higher mean accuracy over union label spaces and significantly reduces overconfidence on out-of-domain classes.

Overall this paper makes the following contributions :

1. We identify unsupervised domain routing and cross-domain logit miscalibration as primary barriers to the MED problem.
2. We propose MED-DSLC, a combination of domain-wise logit calibration and domain-aware expert routing to restore global logit comparability.
3. We demonstrate consistent improvements in cross-domain and union-class evaluation, enabling robust truly zero-shot recognition across specialized domains.

2 Related Work

Vision–Language Pretraining and Parameter-Efficient Finetuning. Contrastive trained VLMs, such as CLIP [16], have become a foundation for open-vocabulary recognition. These models learn aligned image-text embeddings from large-scale web data, enabling zero-shot classification across diverse concepts. However, they often underperform on fine-grained or domain-specific benchmarks, motivating parameter-efficient finetuning (PEFT) strategies such as prompt learning [8,25,26] and low-rank adaptation (LoRA) [5,22]. LoRA adds low-rank updates to frozen weights, enabling efficient specialization to fine-grained domains. However, these approaches typically produce one set of adapted parameters per domain, resulting in a collection of specialized experts rather than a unified model capable of handling multiple domains jointly. This is the starting point for the MED classification problem , which seeks to produce that model.

LoRA Composition and Mixture-of-Experts. Recent works explore combining multiple LoRA adapters through dynamic composition [6] or mixture-of-experts (MoE) [7, 9, 18] mechanisms. LoraHub [6] proposes dynamic LoRA composition to improve cross-task generalization. Model-based Clustering [13] proposes maintaining a library of LoRA modules that can be composed to extend model capabilities through dynamic adapter routing. Several works investigate merging independently trained LoRAs into a single model: while [15] extends model soups to LoRA via weight averaging, [23] studies rank-wise clustering to enhance modularity, and [17] demonstrates effective composition of subject and style LoRAs through structured merging. KnOTS [19] method uses singular value decomposition to map LoRA updates into a shared representation space, improving merging quality and generality without using any task data. Other works [1, 14, 20, 24] explore similar themes, such as data-free merging LoRA modules in a common low-rank core space that preserves efficient adaptation while improving merge accuracy. While these approaches highlight the flexibility of low-rank updates, they primarily focus on skill or style composition, they typically target language tasks and do not explicitly address cross-domain logit calibration, which is central to the MED problem.

MoE methods such as Mixture-of-LoRA-Experts (MoLE) [21] introduces load-balanced expert routing over LoRA modules while [12] studies routing strategies with pretrained LoRAs and gating vectors to improve unseen-domain transfer. These methods typically rely on unsupervised gating or learned routing without explicitly leveraging domain priors or availability of specialized fine-tuned LoRAs. They do not explicitly address domain routing or cross-domain interference when multiple domain-specialized experts are jointly evaluated. These problems are crucial to `MED` and the focus of our work.

3 Method

In multi-expert-domain (`MED`) classification, a VLM has been adapted to multiple expert domains, using PEFT techniques like LoRA, and must perform classification on a label set drawn from the union of all classes, including novel class compositions that span several domains, without domain knowledge at test-time. The `MED` classifier should leverage the VLM and a set of adaptation parameters (e.g. LoRAs) independently trained on the different domains. Ideally, both of these should be kept frozen during `MED` training. The independent domain adaptation induces cross-domain logit miscalibration, which compromises global zero-shot consistency. To address this problem, we propose `MED-DSLC`, a `MED` classifier with domain-supervised routing and domain-aware logit scaling.

3.1 Domain adapted foundation models

In the era of foundation models and open set classification, image classification is usually done by a VLM, such as CLIP [16], composed by a pair v, g of vision $v(x)$ and text $g(t)$ encoders, trained to map pairs (x, t) of image x and text t into a semantic embedding $(v(x), g(t))$ where the two modalities are aligned. Given an image x and a label set $\mathcal{Y}$, the VLM produces a similarity logit

$$z_c(x) = \langle f(x), g(t_c) \rangle. \tag{1}$$

for each class name $t_c \in \mathcal{Y}$. A posterior class distribution is then evaluated with a softmax

$$p(c \mid x) = \frac{\exp(z_c(x))}{\sum_{c' \in \mathcal{Y}} \exp(z_{c'}(x))}, \tag{2}$$

and the class of largest probability assigned to x.

While this allows zero-shot classification for any label set $\mathcal{Y}$, the VLM performance is not uniform over label sets. While classification accuracies tend to be high for coarser-grained problems involving relatively non-similar classes containing images popular on the internet, performance can degrade substantially when $\mathcal{Y}$ includes fine-gained classes, or classes from image domains for which data is not widely available in the public domain. Figure 1 (a) gives the example of fine-grained classification of airplanes. We refer to these as *expert domains*[1] in

[1] Here, "expert" refers to domains that challenge CLIP, not necessarily humans; e.g., fine-grained face recognition.

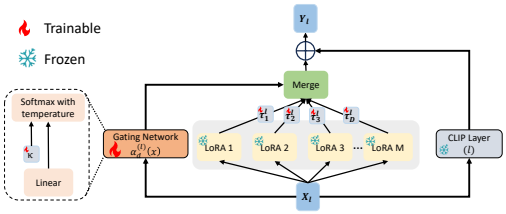

Fig. 3: The `MED-DSLC` network combines a VLM (such as CLIP) and a set of LoRAs learned independently for different image domains, which are kept frozen. `MED-DSLC` is an MoE architecture that relies on a combination of 1) domain routing by a gating network $\alpha^{(l)}(x)$ trained with domain supervision and 2) logit rescaling using a set of learned domain-specific temperature parameters $\tau_d^{(l)}$ in the last layer before softmax.

this work. To overcome this problem, several PEFT techniques [5,8,25,26] have been developed. These augment the VLM with a small set of *domain-specific* parameters W_d, and train them with expert domain data $\mathcal{D}_d$, while keeping f and g frozen, usually in the low-shot regime. Given expert domains datasets $\{\mathcal{D}_d\}_{d=1}^D$ and associated label sets $\{\mathcal{Y}_d\}_{d=1}^D$, the VLM is adapted independently to each domain, producing D specialized encoders $\{f_d = f \oplus W_d\}_{d=1}^D$. While these techniques are very effective, the adapted models tend to loose the open set robustness of the VLM. In general, f_d achieves substantially improved performance for labels in $\mathcal{Y}^d$ but degrades classification outside of this set, as illustrated in Figures 1 (b)-(c). This results in a highly fragmented model ecosystem, where practitioners develop thousands of individual models for specific domains.

3.2 Multi-expert-domain classification

In this work, we consider the problem of how to leverage these domain-specific experts to implement a *multi-expert-domain* (`MED`) classifier. This is a classifier of images x over a label set $\mathcal{Y} = \{\mathcal{Y}_1, \dots \mathcal{Y}_D\}$ composed of several expert domain label sets. PEFT parameters W_d are available for each of the domains, and the task is to combine domain parameters W_d and VLM encoder f into an encoder

$$f^{\texttt{MED}} = f \oplus_{d=1}^{D} W_d \tag{3}$$

of good performance for any label set $\mathcal{C} \subseteq \mathcal{Y}$, while having no access to domain labels at inference time. While `MED` classification can be developed for any PEFT technique, we focus on LoRA [5], due to is large popularity and demonstrated success for computer vision problems.

LoRA-based PEFT consists of adding to a parameter matrix $W \in \mathbb{R}^{p \times q}$ of the VLM encoder $f(x)$ an adaptation matrix $\Delta W = BA$, where $B \in \mathbb{R}^{p \times k}$ and $A \in \mathbb{R}^{k \times q}$ are two low-rank matrices, of shared intermediate dimensionality k, typically much smaller than p or q. The residual parameters ΔW are learned with a cross-entropy loss, while leaving W fixed. This allows adaptation with a small number of parameters and few training examples. To implement `MED`, we adopt the Mixture-of-Experts (MoE) [18] architecture of Figure 3. At each

adapted layer $\ell \in \{1, \ldots, L\}$, the LoRA residual from expert d is denoted as $\Delta h_d^{(\ell)}(x) = B_d^{(\ell)} A_d^{(\ell)} x$, where $B_d^{(\ell)}, A_d^{(\ell)}$ are the LoRA matrices available for domain d. Given a set of input-dependent domain mixing weights $\alpha_d^{(\ell)}(x) \geq 0$ such that $\sum_{d=1}^{D} \alpha_d^{(\ell)}(x) = 1$, the `MED`-LoRA residual is computed with

$$\Delta h^{(\ell)}(x) = \sum_{d=1}^{D} \alpha_d^{(\ell)}(x)\, \Delta h_d^{(\ell)}(x). \tag{4}$$

As usual for MoEs [18], the weights $\alpha_d^{(\ell)}(x)$ are computed by a gating network $\alpha(x)$ learned from a dataset of examples from all domains. This network learns to route the features derived from image x to the associated domain expert. This produces a unified image encoder $f^{\texttt{MED}}(\cdot)$ that dynamically combines domain-specific expertise without requiring domain labels at inference. In our implementation, the gating network combines linear layer W_g and non-linearity ρ,

$$g(x) = \rho(W_g x). \tag{5}$$

This generalizes two architectures previously proposed for LoRA mixing: MoLE [21] if ρ is the softmax function

$$\rho_d(x) = \frac{\exp(W_d x)}{\sum_{c=1}^{D} \exp(W_c x)}, \tag{6}$$

and Phatgoose [12] if ρ consists of a vector of sigmoid functions

$$\rho_d(x) = \frac{1}{1 + \exp(-W_d x)}. \tag{7}$$

However, there are significant differences to these works that we discuss next.

3.3 Training

Like most MoE methods, MoLE and Phatgoose assume that domain labels are unavailable during training. A significant challenge, in this case, is to guarantee that the training produces well-balance experts, e.g. avoiding the scenario where one expert is used for most of the data and the remaining only infrequently. This balancing is usually encouraged by introducing a balancing loss during training. Several such losses have been proposed in the literature [11]. MoLE relies on

$$\mathcal{L}_b = -\log(\prod_{d=1}^{D} \alpha_d), \quad \alpha_d = \frac{1}{L} \sum_{l=1}^{l} \alpha_d^{(l)}, \tag{8}$$

However, in the `MED` problem, the domain of each example x can be trivially derived from its class label y. It suffices to search for the label in each of the domain label sets $\mathcal{Y}_d$, i.e.

$$d^* = \arg\max_d \mathbb{1}(y \in \mathcal{Y}_d) \tag{9}$$

where $\mathbb{1}(.)$ is equal to one when the argument holds and 0 otherwise. This motivates us to use a training objective consisting of two losses. The first is the image *classification loss*, a cross-entropy loss over the complete label set $\mathcal{Y}$,

$$\mathcal{L}_{\text{cls}} = -\log p(y \mid x), \tag{10}$$

where y is the ground-truth class of x. The second is the *domain classification loss* that supervises the gating function g. Let $\alpha_d(x)$ be as in (8). Given ground-truth domain d^* for image x,

$$\mathcal{L}_{\text{domain}} = -\log \alpha_{d^*}(x). \tag{11}$$

This loss promotes domain-discriminative expert selection, reducing cross-domain interference. Its domain supervised nature guarantees much stronger discrimination between domains than (8), therefore inducing much more accurate domain selections by the gating network α. Furthermore, if the number of training examples per domain is balanced, i.e. the same for each domain, the experts are automatically balanced. The overall training loss is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{domain}}, \tag{12}$$

where λ balances classification and routing supervision.

3.4 Domain-sensitive logit calibration

The main difficulty of `MED` classification is to combine LoRAs trained on non-overlaping datasets $\mathcal{D}_d$ from different domains d. This induces two challenges. The first is to guarantee that each example x is routed to the LoRA $h_{d^*}(x)$ of the corresponding domain d^*. The supervised training of the previous section is a strong solution to this problem. The second is to guarantee that the logits $z_c(x)$ can be effectively used for classification with (2). This is not guaranteed when the logits are produced by independently trained networks, as in `MED`, because the probabilities $p(c|x)$ are invariant to many rigid logit transformations. For example, due to the monotonicity of the exponential function

$$\begin{aligned} \arg\max\left(p(a|x), p(b|x)\right) &= \arg\max(e^{z_a(x)}, e^{z_b(x)}) \\ &= \arg\max(e^{\gamma z_a(x)}, e^{\gamma z_b(x)}), \forall \gamma > 0 \end{aligned} \tag{13}$$

and the relative ordering of the probabilities remains constant under the scaling of the logits by a common non-negative constant γ. This is commonly exploited to control the smoothness of output probability distributions of machine learning models using a temperature parameter. Since logit rescaling only affects the relative magnitudes of the probabilities, not their ranking, the logits of the optimal classifier are only defined up to a scaling by γ. While this is not a problem when all network parameters are learned simultaneously, it can derail a `MED` classifier, where the logits of each domain are produced by an independently trained LoRA. In this case, logits of domain d are likely to be scaled by a value $\gamma_d \in \mathbb{R}^{|\mathcal{Y}_d|}$

different from the scaling γ_m of domain m. It follows that, if class a belongs to domain d and class b to domain m, the guarantee of (13) no longer holds, i.e

$$\exists \gamma_a \neq \gamma_b, \text{such that} \arg\max\left(p(a|x), p(b|x)\right) \neq \arg\max(e^{\gamma_a z_a(x)}, e^{\gamma_b z_b(x)}).$$

Since, when evaluating over the entire label set $\mathcal{Y}$, logits from heterogeneous domains $\mathcal{Y}_d$ compete directly under softmax normalization, a domain that systematically produces larger-magnitude logits will induce the classifier to predict its classes even for out-of-domain examples. For example, in Figure 1, the logit produced for the groundtruth-class of the image ('F-18 jet') by the expert of its domain (Figure 1 (b)) is smaller than the logit produced for another class ('Military Parade') by an out-of-domain expert (Figure 1 (c)). This breaks the fundamental assumption underlying VLM-based zero-shot inference: that similarity scores are globally calibrated across candidate classes. Hence, the `MED-DSLC` residual of (4) can be dominated by the logit of the incorrect class, making it the class predicted by the `MED`-classifier, as shown in Figure 1 (d). As the number of domains increases, the probability of spurious high-confidence predictions grows, leading to degraded cross-domain performance.

To restore global logit comparability, we introduce a learnable temperature parameter $\tau_d > 0$ for each domain. For any class $c \in \mathcal{Y}_d$, logits are rescaled with

$$\tilde{z}_c(x) = \frac{z_c(x)}{\tau_d}. \tag{14}$$

This is equivalent to applying domain-specific temperature scaling prior to softmax evaluation over $\mathcal{Y}$, $p(c \mid x) = \frac{\exp(\tilde{z}_c(x))}{\sum_{c' \in \mathcal{Y}} \exp(\tilde{z}_{c'}(x))}$. This normalization preserves within-domain ranking while aligning cross-domain magnitudes. Learning the temperature parameters $\{\tau_d\}_{d=1}^{D}$ compensates for domain-specific amplification effects introduced by independent LoRA adaptation. Importantly, this mechanism introduces only D scalar parameters and does not modify the VLM.

By combining domain-supervised routing with domain-sensitive logit scaling, `MED-DSLC` restores the key property required for zero-shot inference: logits across arbitrary candidate classes are globally comparable. This enables robust classification over any union of domain-specific classes without requiring domain labels at test time, thereby recovering scalable, truly zero-shot CLIP behavior under multi-domain specialization. Figure 2 illustrates the benefits of domain supervision and domain-sensitive logit scaling for `MED`. The figure shows histograms of the logits for classes of the EuroSAT dataset with (`MED-DSLC`) and without (MoLE) these two components. While the MoLE histogram exhibits a significant overlap between the logits of DTD and EuroSAT classes, this overlap is drastically reduced for `MED-DSLC`.

4 Experimental Results

We evaluate `MED-DSLC` on a diverse suite of nine fine-grained recognition benchmarks spanning object categories, textures, scenes, and actions. Following prior

Method	Mean (%)	Δ (%)	Caltech	EuroSAT	Cars	Food	Pets	Flowers	DTD	UCF	FGVC
Base Split											
Expert-LoRA	85.48	+15.36	96.67	94.81	81.71	88.56	95.96	**97.91**	**81.02**	**83.61**	**49.10**
MED-DSLC	**85.51**	**+15.39**	**97.40**	**94.90**	**81.73**	88.70	**96.17**	**97.91**	80.67	83.20	48.92
KnOTS-DARE-TIES [19]	78.17	+8.05	93.76	78.81	74.74	89.62	94.21	88.51	67.94	77.51	38.42
KnOTS-TIES [19]	74.91	+4.79	93.51	71.76	70.84	89.86	94.15	81.67	61.81	74.82	35.77
PHATGOOSE [12]	72.97	+2.85	92.62	66.12	70.01	89.68	93.67	79.20	57.18	74.04	34.21
MoLE	71.80	+1.69	93.11	72.90	63.47	89.16	92.40	72.74	60.07	72.08	30.31
LoRA-Mean	70.12	+0.00	89.94	60.29	66.42	**89.70**	93.04	75.97	53.59	71.15	30.97
Single-LoRA	42.35	-27.77	83.54	44.17	26.39	60.56	53.32	32.10	33.68	45.40	1.98
CLIP	66.22	-3.90	84.43	52.69	63.07	89.46	89.79	72.17	49.07	67.27	28.03
All Split											
Expert-LoRA	71.60	+6.63	92.56	67.58	70.10	83.29	91.03	**78.44**	**56.50**	**72.56**	**32.34**
MED-DSLC	**71.80**	**+6.83**	**92.95**	**68.43**	**70.33**	83.73	**91.55**	78.32	56.26	72.43	32.19
KnOTS-DARE-TIES [19]	68.66	+3.69	90.64	58.28	**70.51**	84.79	90.57	76.00	49.88	69.73	27.51
KnOTS-TIES [19]	67.70	+2.73	90.69	54.19	69.22	85.72	90.81	74.95	46.10	69.18	28.47
PHATGOOSE [12]	59.77	-5.20	82.85	23.10	65.86	78.33	88.33	67.48	47.10	61.49	23.37
MoLE	65.85	+0.88	90.19	56.25	65.03	85.02	88.66	70.85	44.86	66.56	25.20
LoRA-Mean	64.97	+0.00	87.68	45.81	66.98	**85.82**	89.78	73.12	42.67	66.72	26.13
Single-LoRA	36.28	-28.69	73.63	28.10	27.27	57.44	54.18	22.25	25.65	36.98	0.99
CLIP	62.52	-2.45	85.26	40.21	65.13	85.04	87.38	70.60	40.54	63.94	24.57

Table 1: Cross-domain results. Rows grouped by base/all splits. Best per column in bold, second-best underlined. Δ is improvement over LoRA-Mean within each split.

work [22], we train domain-specific LoRA experts and evaluate merged models under both **in-domain** and **cross-domain** protocols. The in-domain setting evaluates each dataset independently, while the cross-domain protocol forms a unified label space over the union of all classes, requiring models to discriminate across domains without domain labels at test time. This setting explicitly measures robustness to cross-domain interference. For all experiments, we split each dataset into two halves, *base* (seen during training) and *remaining* (unseen during training) classes, and report results on *base* and *all* (seen+unseen) splits. The *all* split is the most comprehensive and challenging evaluation, as it measures joint generalization across both seen and unseen classes within a unified label space.

Datasets and training. We use a suite of fine-grained recognition datasets commonly used to evaluate few-shot learning [26]. This includes Caltech101, EuroSAT, FGVC Aircraft, Cars, Pets, Flowers, DTD, UCF101 and Food101. We note that certain domain classes conflict with other domains. For example, "water_lilly" in Caltech101 overlaps with classes in Oxford Flowers. To avoid this duplication, we exclude the following classes from Caltech101: water_lilly, lotus, pagoda, pizza, sunflower, airplane, barrel, ice_cream, car_side, lobster, helicopter, soccer_ball. In total, we obtain 783 classes across all datasets after filtering. For training, we follow the standard setup of prompt learning [26]. We train rank-2 LoRA experts with 16-shot images per class of the Base split. The AdamW optimizer is used with a batch size of 16 and a learning rate of 1e-4. For more details, refer to supplementary section 1.

Baselines. The proposed MED-DSLC approach was compared to several baselines. MED-DSLC is implemented with domain supervision during training, logit rescaling, and the softmax gating function of (6). CLIP is the original CLIP model, without adaptation. Single-LoRA is the standard LoRA adaptation over

Method	Mean (%)	Δ (%)	Caltech	EuroSAT	Cars	Food	Pets	Flowers	DTD	UCF	FGVC
Base Split											
Expert-LoRA	**85.96**	**+14.10**	97.89	95.02	81.71	88.69	**96.17**	**97.91**	**83.45**	**83.66**	**49.10**
MED-DSLC	85.89	+14.04	**98.05**	**95.05**	**81.73**	88.78	**96.17**	**97.91**	82.99	83.40	48.92
KnOTS-DARE-TIES [19]	79.29	+7.44	97.65	81.62	74.74	89.75	94.84	88.51	70.37	77.77	38.42
KnOTS-TIES [19]	76.26	+4.41	97.40	74.60	70.84	89.99	94.63	81.67	66.32	75.13	35.77
PHATGOOSE [12]	74.78	+2.93	92.05	55.95	76.44	84.65	94.42	89.08	70.72	74.20	35.53
MoLE	72.90	+1.05	96.92	73.95	63.47	89.27	92.72	72.74	64.47	72.29	30.31
LoRA-Mean	71.85	+0.00	96.43	64.00	66.42	**89.80**	93.46	75.97	58.22	71.41	30.97
Single-LoRA	43.40	-28.45	85.16	44.33	26.61	61.21	54.70	32.67	35.42	48.55	1.98
CLIP (Base)	67.99	-3.87	89.86	57.10	63.07	89.57	90.11	72.17	54.40	67.58	28.03
All Split											
Expert-LoRA	72.27	+5.77	**94.43**	67.89	70.19	83.79	91.31	**78.48**	**59.16**	**72.80**	**32.37**
MED-DSLC	**72.33**	**+5.82**	94.38	**68.63**	**70.35**	83.90	**91.63**	78.32	58.69	**72.80**	32.22
KnOTS-DARE-TIES [19]	69.78	+3.27	93.69	60.78	**70.53**	85.18	91.06	76.00	53.31	69.92	27.51
KnOTS-TIES [19]	68.92	+2.42	93.79	56.98	69.22	86.02	91.25	74.95	50.18	69.44	28.47
PHATGOOSE [12]	62.66	-3.84	87.48	38.44	65.86	78.72	88.85	67.56	50.30	63.34	23.40
MoLE	66.97	+0.46	93.54	57.79	65.03	85.33	88.91	70.85	49.11	66.93	25.20
LoRA-Mean	66.50	+0.00	92.46	48.57	66.98	**86.05**	90.16	73.12	47.87	67.17	26.13
Single-LoRA	37.47	-29.03	76.34	28.58	27.50	58.07	55.14	22.53	27.60	40.52	0.99
CLIP (Base)	63.87	-2.63	87.88	43.00	65.13	85.29	87.65	70.60	46.22	64.47	24.57

Table 2: In-domain results. Rows grouped by base/all splits. Best per column in bold, second-best underlined. Δ is improvement over LoRA-Mean within each split.

the entire set of 10 datasets, using a LoRA with 10x the number of parameters of the individual expert LoRAs of MED-DSLC. LoRA-mean is a representative of methods that combine LoRA parameters through arithmetic operations, in this case the average of the 10 domain LoRAs. Expert-LoRA is the performance achieved with a domain oracle, i.e. when the domain of the test image is known and the corresponding LoRA always chosen. Intuitively, this is an upper bound for performance, since the "right"expert is always selected. However, we will see that a combination of LoRAs can usually slightly outperform this result. MoLE is identical to MED-DSLC without domain supervision and logit scaling.

5 Results

Cross-Domain Evaluation. Table 1 reports performance under the cross-domain protocol, where class labels from multiple datasets are unified and logits compete under a single softmax. The table is organized in three sections. The top section shows the expected upper bound performance of Expert-LoRA. The middle section compares different MED approaches. Finally, the bottom section includes Single-Lora approach and the original CLIP model.

Base Split. On the base split, MED-DSLC achieves the best mean accuracy of **85.51%**, improving over LoRA-Mean (70.12%) by **+15.39** points. Notably, MED-DSLC slightly surpasses the oracle Expert-LoRA (85.48%). This can be explained by the LoRA combination of (4), which increases robustness to light overfitting by the expert LoRAs. It also achieves a large gain (+12.99) over MoLE, demonstrating the advantages of domain supervision and logit scaling. All models involving expert LoRAs improve dramatically on CLIP and Single-LoRA. The latter has significantly lower performance than even CLIP, perhaps due to

overfitting of the large number of parameters being learned. Compared to existing parameter merging and model-editing approaches with LoRAs, `MED-DSLC` outperforms KnOTS-DARE-TIES by more than **7** points and the gap is even larger relative to PHATGOOSE, which trails `MED-DSLC` by almost **12** points.

All Split. The **all split** has qualitatively similar results. All methods have weaker performance, due to the need to generalize to unseen classes, but the relative performances are similar. Again `MED-DSLC` slightly outperforms Expert-Lora and achieves large gains (close to 6 points) over the other `MED` methods. The gains of this split are particularly meaningful, as it maximizes cross-domain interference: logits from independently trained experts must remain comparable across both domains and class partitions, and generalize to classes unseen at training.

In-Domain Evaluation Table 2 presents in-domain results, where each dataset is evaluated independently and no cross-domain competition occurs. The model includes all LoRAs, and must ideally route images through the dataset expert.

Base Split. `MED-DSLC` achieves 85.89%, closely matching Expert-LoRA (85.96%). The gains over the other `MED` methods are again significant (7-13 points). The significant gain over MoLE and PHATGOOSE illustrates the advantage of domain supervision. Because the weaker gating function of MoLE routes images through LoRAs of mismatched domains, performance degrades. On the other hand, the average LoRA of LoRA-mean is not very effective for most domains, illustrating the limitations of adaptation by parameter arithmetic.

All Split. Again, the results are qualitatively similar to those of the Base split.

6 Ablations

`MED-DSLC` components. Table 3 compares various implementations of `MED-DSLC` in the cross-domain setting, varying in terms of the inclusion, or not, of domain scaling and supervised learning, and the use of the softmax of (6) or sigmoid vector of (7) as non-linearity of the gating network. The results are qualitatively the same for the Base and All splits. The weakest performance is achieved by the model without logit scaling and domain supervision. For the same non-linearity, the addition of domain supervision during training results in a large performance gain. Further adding logit scaling results in a smaller additional gain. We note that the gains of logit scaling are likely to increase for larger numbers of domains. Comparing the gating non-linearities, the softmax function clearly outperforms the sigmoid array, for comparable network configurations. Please refer to supplementary for ablation studies for the in-domain setting.

Sensitivity to class imbalance. We conduct a target-domain imbalance sweep experiment to better understand the role of LS in Figure 4. Here, one target domain is restricted to k classes while the remaining domains retain their original class sets. Figure 4 (left) shows that LS is more useful in the highly imbalanced settings. In particular, at $k = 2$, `MED-DSLC` improves target-domain accuracy over MED-DS by +2.2 points; at $k = 4$, the gains are smaller, at +1.3 points.

Log Scal	Dom Sup	Soft	Sig	Mean (%)	Δ (%)	Caltech	EuroSAT	Cars	Food	Pets	Flowers	DTD	UCF	FGVC
Base Split														
✓	✓	✓		**85.51**	**+15.39**	**97.40**	94.90	**81.73**	88.70	96.17	**97.91**	80.67	83.20	48.92
	✓	✓		85.43	+15.31	96.59	**94.95**	81.41	88.48	**96.28**	**97.91**	**80.90**	**83.25**	**49.10**
✓	✓		✓	79.21	+9.09	96.27	89.55	73.46	89.90	94.74	87.75	68.98	76.73	35.53
	✓		✓	78.55	+8.43	95.46	89.38	73.69	**89.95**	94.26	84.14	66.44	76.63	36.97
		✓		71.80	+1.69	93.11	72.90	63.47	89.16	92.40	72.74	60.07	72.08	30.31
All Split														
✓	✓	✓		**71.80**	**+6.83**	**92.95**	68.43	**70.33**	83.73	**91.55**	78.32	**56.26**	72.43	32.19
	✓	✓		71.61	+6.64	91.77	67.47	70.25	83.53	91.20	**78.60**	**56.26**	**73.09**	**32.34**
✓	✓		✓	69.97	+5.00	92.02	**69.04**	69.39	**85.94**	91.17	76.13	49.76	69.44	26.85
	✓		✓	69.67	+4.71	91.87	68.94	69.08	85.79	90.71	74.75	49.35	69.71	26.88
		✓		65.85	+0.88	90.19	56.25	65.03	85.02	88.66	70.85	44.86	66.56	25.20

Table 3: Cross-domain ablations. Rows grouped by base/all splits. Best per column in bold, second-best underlined. Δ is improvement over LoRA-Mean within each split.

Table 4: Ablation on LoRA ranks, SigLIP LoRA, and data efficiency in the cross-domain experiment under all split.

Method	LoRA Rank			SigLIP LoRA	Data Efficiency			
	r2	r4	r8	r2	1	4	8	16
LoRA-Mean	64.97	64.84	65.07	51.64	64.97	64.97	64.97	64.97
MoLE	65.85	65.82	65.51	50.81	65.26	65.63	65.88	65.85
PHATGOOSE	59.77	61.36	61.98	49.67	61.51	59.42	57.75	59.77
KnOTS-TIES	67.70	67.29	67.84	53.23	67.70	67.70	67.70	67.70
`MED-DSLC`	**71.80**	**72.20**	**72.48**	**62.27**	**71.70**	**71.50**	**71.53**	**71.80**

This indicates that LS compensates for domain-level score inflation rather than only fitting balanced priors.

Comparisons to MED baselines with SigLIP and different LoRA ranks Table 4 shows the backbone architecture and LoRA rank ablation, for all-split mean accuracies. The gains of `MED-DSLC` are almost constant across LoRA ranks, and even more significant for SigLIP than CLIP. It also shows that the method is model agnostic and does not rely on the biases of an architecture.

Domain mixing. In the `MED` problem, the classifier should maintain good performance over any label subset $\mathcal{C} \subseteq \mathcal{Y}$ combining classes from any of the domains. To test the robustness of to the size of the label set $\mathcal{C}$, we performed an experiment with increasing number of classes k per domain. For a given k, the difficulty of the classification also depends on the class composition. To further analyze robustness to variations in label-space complexity, we consider two approaches for class selection: (1) **First-k**, which uses the first k classes from each domain, and (2) **Random-k** (Top-k), where k classes are randomly sampled per domain. The rationale is that datasets are usually organized by class similarity. So, while First-k should create a more homogeneous classification problem, Random-k has more heterogeneous coverage of the domain classes. We consider $k \in \{5, 10, 20, 40\}$ and report mean accuracy across domains for *All Split*.

Tables 5 and 6 show that `MED` performance is indeed sensitive to the class composition. Classification performance degrades as the number of classes increases and is higher for Random-k than for First-k. This is expected, since problems with more classes and greater class similarity tend to be more chal-

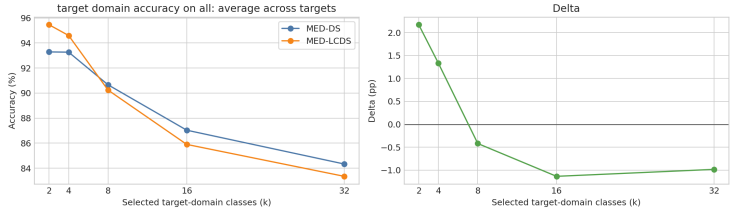

Fig. 4: Effect of logit scaling on performance for a target-domain imbalance sweep. Left: accuracy vs number of target-domain classes (k) for `MED-DSLC` and MED-DS. Right: difference between the two approaches.

Method	$k=5$	$k=10$	$k=20$	$k=40$
Expert-LoRA	91.14	85.37	83.58	79.15
LoRA-Mean	81.34	74.88	72.43	66.63
MoLE	82.94	76.80	73.58	66.99
`MED-DSLC`	**91.37**	**85.34**	**83.52**	**79.39**
Δ	8.43	8.54	9.94	12.4

Table 5: First-k class selection (All split). Δ is the difference between `MED-DSLC` and MoLE. Mean accuracy (%) across domains.

Method	$k=5$	$k=10$	$k=20$	$k=40$
Expert-LoRA	92.33	90.07	83.49	78.00
LoRA-Mean	85.48	86.39	79.57	71.50
MoLE	87.47	86.92	79.72	72.25
`MED-DSLC`	**92.30**	**90.21**	**83.53**	**77.97**
Δ	4.83	3.29	3.81	5.72

Table 6: Random-k (Top-k) class selection (All split). Δ is the difference between `MED-DSLC` and MoLE. Mean accuracy (%) across domains.

lenging. However, in both cases, `MED-DSLC` is superior to LoRA-Mean and MoLE, and tracks the performance of Expert-LoRA, achieving gains similar to those of Table 1 where k is the size of the entire label set $\mathcal{Y}$. The gains (Δ) of `MED-DSLC` over MoLE are also larger for the more difficult problem and grow with the total number of classes. This shows that `MED-DSLC` is quite robust to the class composition of the `MED` problem.

Data Efficiency. To evaluate the data efficiency of `MED-DSLC`, we consider a **one-shot training** setting where the expert LoRAs remain fixed and only the gating network and temperature scaling are trained using a single labeled example per class. This experiment measures how much supervision is required to learn effective routing between pretrained experts. Table 7 shows that `MED-DSLC` achieves a mean accuracy of **71.71%** on the *All Split*, outperforming LoRA-Mean (64.97%) by **+6.75** points and MoLE (65.26%) by **+6.46** points. Table 4 additionally shows the results for 4, 8 and 16 shots. The gains are consistent across most domains as compared to the 16-shot setting in Table 1. This demonstrates that `MED-DSLC` is highly data-efficient, maintaining strong cross-domain performance even when the routing mechanism is trained with minimal data.

Overall, these ablations confirm that domain-aware routing with logit scaling enables efficient, scalable expert merging that stays accurate and interference-free as class count grows.

7 Qualitative Analysis

Figure 5 compares the predictions of `MED-DSLC` without (orange) and with (green) logit scaling. When combining domain-specialized experts without logit scaling,

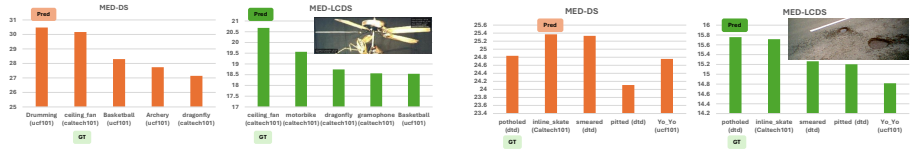

Fig. 5: MED logits with (green) and without (orange) logit scaling. Class (domain) shown in the horizontal axis, and ground-truth (predicted) class shown at the bottom (top). In the absence of logit scaling, logits are poorly calibrated across domains, and the predicted class can be due to a spurious logit outside the image domain.

Method	Mean (%)	Δ (%)	Caltech	EuroSAT	Cars	Food	Pets	Flowers	DTD	UCF	FGVC
					All Split						
MED-DSLC	**71.71**	**+6.75**	**92.07**	**69.42**	**70.15**	83.46	**91.06**	**78.20**	**56.09**	**72.85**	**32.13**
MoLE	65.26	+0.29	90.00	50.00	66.32	85.22	89.32	71.05	42.02	66.16	27.27
LoRA-Mean	64.97	+0.00	87.68	45.81	66.98	**85.82**	89.78	73.12	42.67	66.72	26.13

Table 7: Cross-domain results for all split with one-shot training. Best per column in bold, second-best underlined. Δ is improvement over LoRA-Mean within each split.

the model exhibits more cross-domain interference: logits from unrelated domains attain larger magnitudes and dominate the joint softmax, leading to incorrect predictions. The use of domain-wise logit scaling (right) aligns logit magnitude across experts while preserving within-domain discrimination. This restores global comparability in the joint label space and reduces artificial dominance by out-of-domain classes. As a result, the correct class receives the highest calibrated logit, demonstrating that logit scaling effectively mitigates cross-domain interference and stabilizes multi-domain predictions. For additional qualitative results, refer to supplementary material.

8 Conclusion

In this work, we studied the MED classification problem, which seeks to merge domain-specific LoRA adapters for CLIP under realistic open-set settings, where base (seen) and new (unseen) classes from multiple datasets must be evaluated jointly. Our analysis revealed *cross-domain interference* as the key limitation of existing approaches: independently trained experts produce miscalibrated logits that compete improperly under a shared softmax, leading to degradation in unified label spaces, especially in the challenging *all* class split. We proposed MED-DSLC, a domain-aware data-efficient mixture-of-experts framework that combines supervised domain routing with domain-wise logit scaling. The routing enforces structured expert selection, while learned temperature parameters restore cross-domain logit comparability without affecting within-domain discrimination. Across extensive experiments, MED-DSLC consistently outperformed previous approaches to the MED classification problem. It substantially outperforms naive merging and matches or surpasses oracle expert selection, while preserving in-domain accuracy. Ablations with varying label-space sizes further confirm that MED scales robustly as cross-domain competition increases. In summary, we show that restoring global logit alignment is essential for building unified, interference-resistant multi-domain vision–language systems.

References

1. Alipour, M., Mohammadi Amiri, M.: Towards reversible model merging for low-rank weights (2025), https://arxiv.org/abs/2510.14163, arXiv preprint 4
2. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research (JMLR) **23**, 1–39 (2022) 3
3. Feng, W., Hao, C., Zhang, Y., Han, Y., Wang, H.: Mixture-of-loras: An efficient multitask tuning method for large language models. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 11371–11380 (2024) 2
4. Gou, Y., Liu, Z., Chen, K., Hong, L., Xu, H., Li, A., Yeung, D.Y., Kwok, J.T., Zhang, Y.: Mixture of cluster-conditional lora experts for vision-language instruction tuning. arXiv preprint arXiv:2312.12379 (2023) 2
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022) 2, 4, 6
6. Huang, C., Liu, Q., Lin, B.Y., Pang, T., Du, C., Lin, M.: Lorahub: Efficient cross-task generalization via dynamic lora composition. arXiv preprint arXiv:2307.13269 (2023) 4
7. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation **3**(1), 79–87 (1991). https://doi.org/10.1162/neco.1991.3.1.79 4
8. Jia, M., Tang, L., Chen, B.C., et al.: Visual prompt tuning. In: European Conference on Computer Vision (ECCV). pp. 709–726 (2022) 4, 6
9. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the em algorithm. Neural Computation **6**(2), 181–214 (1994) 4
10. Lepikhin, D., Lee, H., Xu, Y., et al.: Gshard: Scaling giant models with conditional computation and automatic sharding. In: International Conference on Learning Representations (ICLR) (2021) 3
11. Mu, S., Lin, S.: A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. arXiv preprint arXiv:2503.07137 (2025) 7
12. Muqeeth, M., Liu, H., Liu, Y., Raffel, C.: Learning to route among specialized experts for zero-shot generalization. arXiv preprint arXiv:2402.05859 (2024) 3, 5, 7, 10, 11
13. Ostapenko, O., Su, Z., Ponti, E., Charlin, L., Le Roux, N., Caccia, L., Sordoni, A.: Towards modular LLMs by building and reusing a library of LoRAs. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 38885–38904. PMLR (2024) 4
14. Panariello, A., Marczak, D., Magistri, S., Porrello, A., Twardowski, B., Bagdanov, A.D., Calderara, S., van de Weijer, J.: Accurate and efficient low-rank model merging in core space (2025), https://arxiv.org/abs/2509.17786, arXiv preprint 4
15. Prabhakar, A., Li, Y., Narasimhan, K., Kakade, S., Malach, E., Jelassi, S.: Lora soups: Merging loras for practical skill composition tasks. In: Proceedings of the 31st International Conference on Computational Linguistics (Industry Track). pp. 644–655. Association for Computational Linguistics (2025) 4
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, T., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International Conference on Machine Learning (ICML). PMLR (2021) 1, 4, 5
17. Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., Jampani, V.: Ziplora: Any subject in any style by effectively merging loras. European Conference on Computer Vision (ECCV) 2024, LNCS **15059**, 422–438 (2025) 4
18. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (ICLR) (2017) 3, 4, 6, 7
19. Stoica, G., Ramesh, P., Ecsedi, B., Choshen, L., Hoffman, J.: Model merging with svd to tie the knots. In: International Conference on Learning Representations (ICLR) (2025) 4, 10, 11
20. Wei, Y., Tang, A., Shen, L., Hu, Z., Yuan, C., Cao, X.: Modeling multi-task model merging as adaptive projective gradient descent. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=EqoKRSR5Pa 4
21. Wu, X., Huang, S., Wei, F.: Mixture of lora experts. In: International Conference on Learning Representations (2024), https://openreview.net/forum?id=uWvKBCYh4S 2, 3, 5, 7
22. Zanella, M., Ben Ayed, I.: Low-rank few-shot adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1593–1603 (2024) 4, 10
23. Zhao, Z., Shen, T., Zhu, D., Li, Z., Su, J., Wang, X., Wu, F.: Merging LoRAs like playing lego: Pushing the modularity of lora to extremes through rank-wise clustering. In: International Conference on Learning Representations (ICLR) 2025 (2025) 4
24. Zheng, S., Wang, H., Huang, C., Wang, X., Chen, T., Fan, J., Hu, S., Ye, P.: Decouple and orthogonalize: A data-free framework for lora merging (2025), https://arxiv.org/abs/2505.15875, arXiv preprint 4
25. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16837–16847 (2022) 4, 6
26. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. In: International Journal of Computer Vision (IJCV). Springer (2022) 2, 4, 6, 10