SG-LoRA: Semantic-guided LoRA Parameters Generation

Miaoge Li[1] , Yang Chen[1] , Zhijie Rao[1] , Can Jiang[2] , Kang Wei[3] , Jingcai Guo[1*]
[1]Department of COMP/LSGI, The Hong Kong Polytechnic University, Hong Kong SAR
[2]College of Artificial Intelligence, Xidian University, China
[3]School of Cyber Science and Engineer, Southeast University, China
jc-jingcai.guo@polyu.edu.hk

Abstract

Generating new Low-Rank Adaptation (LoRA) weights from pre-trained LoRAs has demonstrated strong generalization capabilities across various tasks, enabling the efficient transfer of AI models, particularly on resource-constrained edges. However, previous studies either merge base LoRAs via weighting coefficients or train a generative model under the closed-world assumption, limiting their efficiency and flexibility in complex edge user cases. This challenge may further increase when there are significant domain shifts between training and deployment. To this end, we propose ***Semantic-guided LoRA Parameter Generation (`SG-LoRA`)****, a tuning-free generative framework to efficiently produce task-specific parameters for unseen tasks in a semantic-to-LoRA pipeline. Concretely, `SG-LoRA` uses task descriptions as the semantic bridge, measuring their proximity to a set of known expert tasks in a shared embedding space. Based on this semantic guidance, it models the target task's LoRA parameter distribution to generate high-performing parameters for novel tasks. `SG-LoRA` enables the real-time construction of LoRA models aligned with individual intents by distilling knowledge from prominent LoRA experts and, while also offering a privacy-preserving solution for personalized model adaptation in a novel zero-shot open-world setting proposed in this work. Extensive experiments on multiple challenging tasks confirm the superior performance and remarkable adaptability of `SG-LoRA`* [1].

1. Introduction

In recent years, deep learning has seen remarkable progress, largely driven by the advent of large-scale pre-trained models (LPMs) [4, 18, 23, 34, 40]. Trained on massive and diverse datasets, these models demonstrate exceptional performance across a wide range of downstream tasks [1, 2, 19, 25, 26, 36, 40–42]. However, as both model and data scales continue to grow, retraining the entire model becomes increasingly computationally expensive and often infeasible in practice. To mitigate this challenge, parameter-efficient fine-tuning (PEFT) methods have drawn considerable attention [5, 20, 24, 33, 51, 52, 56]. Among them, Low-Rank Adaptation (LoRA) [8] has emerged as a prominent approach. LoRA adapts pre-trained models by introducing a small number of trainable low-rank matrices into existing layers, achieving strong task-specific performance while leaving the original model weights unchanged [39].

While an increasing number of pre-trained LoRA modules are becoming publicly available, effectively leveraging them in real-world scenarios remains a significant challenge. As shown in Figure. 1(a), we propose the Zero-Shot Open-world Adaption (ZSOA) in this paper, which aims to generate LoRA weights for unseen tasks based on a set of pre-trained LoRAs. ZSOA emphasizes two key aspects: *(1) No raw data is available for the unseen task, highlighting the need for rapid adaptability to evolving user intents* ; and *(2) Open-world task coverage, defined by a broad and unconstrained task space in which the unseen tasks may not be directly related to the seen tasks*. Compared to traditional LoRAs that need to be fine-tuned on downstream tasks, ZSOA shows data and computation-friendly strength, resulting in more flexibility in practice, particularly in edge environments where data privacy constraints and limited computational resources make large-scale retraining infeasible.

To enable broader applications of LoRA, prior research has explored two main directions, as illustrated in Figure. 1 (b-c), each partially addressing the challenges of ZSOA. The first line of work focuses on merging-based methods, which aim to rapidly construct task-specific models by directly fusing existing LoRA modules at hand [36, 45, 48]. Although these methods support open-world generation, the generated weights are obtained through deterministic fusion of existing LoRAs, resulting in limited diversity and constraining the model's ability to adapt to flexible or evolving requirements [36, 47]. Moreover, the merging process must be

*Corresponding author: *Jingcai Guo*.
[1]Code is available at: https://github.com/keepgoingjkg/SG-LoRA

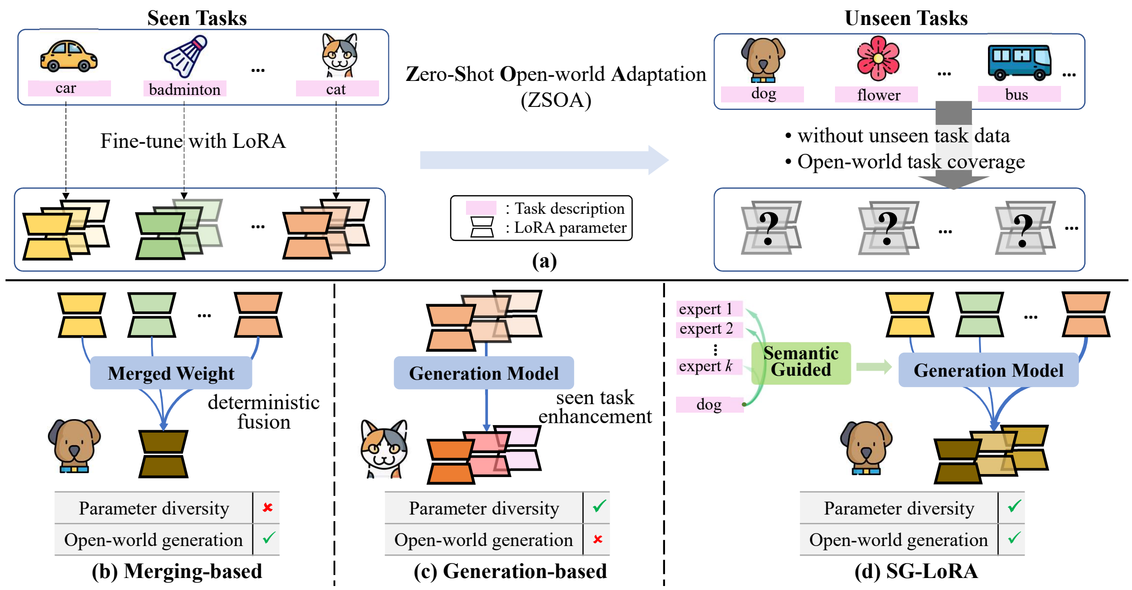

Figure 1. **Motivation of our `SG-LoRA`.** We consider a challenging scenario termed Zero-Shot Open-World Adaptation (ZSOA), where a model is provided rich LoRA resources for seen tasks but lacks access to data for unseen tasks during inference, with an unconstrained task space. Conventional LoRA adaptation methods are not suitable for ZSOA: merging-based approaches *struggle to explore the diversity of LoRA parameters*, while generation-based methods *primarily focus on LoRA enhancement for seen task*. Our `SG-LoRA` uses task descriptions as semantic guidance to enable conditional LoRA generation for unseen tasks in a data-free and open-world manner. (Each color family represents a set of LoRA parameters for the same task, for example, brown represents LoRAs for the *'Dog'* task, and yellow represents LoRAs for the *'Car'* task.)

carefully designed, as conflicts may arise when integrating LoRA modules trained on different tasks [53, 57]. In parallel, another direction explores generation-based methods, which leverage generative models, such as variational autoencoders (VAEs) [14] or diffusion models [7], to synthesize new LoRA parameters. By introducing stochasticity, these approaches enable greater diversity in parameter generation and bypass traditional fine-tuning pipelines. However, their success often relies on a closed-world assumption, where training and test tasks are drawn from similar distributions [38]. As a result, these methods are susceptible to task (or domain) shifts scenarios, and struggle to handle open-world tasks.

In this work, we draw inspiration from the human ability to intuitively infer semantic relationships between prior knowledge and new tasks [17], enabling effective generalization in unfamiliar situations. For example, after learning to recognize cat breeds such as *Birman* and *Egyptian Mau*, a person can identify *British Shorthair* based solely on its textual description by relating it to previously acquired concepts. Motivated by this analogy-driven reasoning process, we propose a **S**emantic-**G**uided **LoRA** Parameter generation framework (`SG-LoRA`) that adapt LPMs to the ZSOA setting. Specifically, given a set of expert (or base) LoRAs whose weights are trained on seen tasks, we aim to train a semantic-to-LoRA model that takes the semantics of the unseen tasks as inputs and outputs high-performing LoRA weights directly. Notably, the task semantics serve as a bridge between the seen and unseen tasks, guiding our `SG-LoRA` on how to leverage expert knowledge to generate task-specific LoRA weights. Motivated by prompt engineering [55], we adopt the task description to identify the task semantics. These descriptions, typically concise yet semantically rich, are processed by a frozen CLIP text encoder to capture task-level correlations without exposing user-specific data. The task semantics are then modeled as Gaussian distributions according to the relationship between the unseen and seen tasks.

Importantly, simply scaling the number of expert LoRAs does not guarantee performance gains, as they may provide contradictory or irrelevant task knowledge. To resolve this, we design a sparse aggregator that assembles the most semantically relevant expert to integrate rational prior knowledge for a target task. During inference, the system directly generates target LoRA modules aligned with user requirements using only textual task queries. Crucially, the stochastic nature of our trained generator converts deterministic LoRA construction into probabilistic parameter sampling, enhancing both parameter diversity and dynamic adaptability to evolving user intents. In summary, the main contributions of this work include :

- We introduce Semantic-Guided LoRA Generation, a versatile framework that harnesses semantic task relationships to enable zero-shot open-world adaptation. By condition-

ing on prior available task knowledge, `SG-LoRA` can synthesize high-performance LoRA parameters for arbitrary unseen tasks without retraining.
- By seamlessly integrating generated LoRA modules into off-the-shelf LPMs, our method enables fast personalization at inference time. It allows flexible configuration of the expert LoRA repository, supporting task-adaptive LoRAs both within and across datasets, thereby achieving scalable and adaptable model behavior.
- Comprehensive experiments on multiple image-text retrieval benchmarks demonstrate that the proposed method can rapidly generate LoRA parameters achieving performance comparable to traditional LoRA fine-tuning.

2. Related Work

2.1. Low-Rank Adaptation

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning method to adapt large models to novel tasks by approximating weight updates with low-rank matrices, drastically reducing trainable parameters while maintaining pre-trained knowledge [10, 50]. Specifically, given a pre-trained weight $\boldsymbol{W}_0$ with input $\boldsymbol{x}$ and hidden state $\boldsymbol{h}$, LoRA decomposes the weight update $\Delta \boldsymbol{W} \in \mathbb{R}^{a \times b}$ into two low-dimensional matrices $\boldsymbol{B} \in \mathbb{R}^{a \times r}$ and $\boldsymbol{A} \in \mathbb{R}^{r \times b}$:

$$\boldsymbol{h} = \boldsymbol{W}_0 \boldsymbol{x} + \Delta \boldsymbol{W} \boldsymbol{x} = \boldsymbol{W}_0 \boldsymbol{x} + \gamma \boldsymbol{B} \boldsymbol{A} \boldsymbol{x}, \tag{1}$$

where the rank $r \ll \min(a, b)$. γ is a constant scaling hyperparameter that controls the contribution of the LoRA update. Recently, a growing number of high-quality, pre-trained LoRA modules have become publicly available for various architectures , including transformers and vision–language models, offering rich resources to accelerate adaptation and deployment on specific downstream tasks.

2.2. Model Merging

Model merging integrates parameter-level knowledge from multiple independently trained networks into a unified model, achieving enhanced capabilities [11, 12, 21, 46]. The pioneer work Model Soups [45] establishes weight averaging as a foundational paradigm, showing that averaging fine-tuned models from identical pre-trained bases with varied hyperparameters consistently outperforms individual models. AdapterSoup [3] generalizes the Model Soups paradigm to cross-domain adaptation by dynamically averaging domain-specific adapters at test time. This approach preserves the base model's integrity while enhancing out-of-distribution generalization through selective weight-space interpolation of relevant domain knowledge. Recent advances have extended this paradigm to LoRA-based module fusion. For instance, LoraHub [10] dynamically composes pre-trained LoRA modules by optimizing their weights through few-shot examples from new tasks, leveraging black-box optimization techniques (e.g., CMA-ES) to achieve efficient adaptation without backpropagation. Meanwhile, SemLA [32] introduces a training-free approach by directly comparing test images' visual features with known domain prototypes, using the resulting similarity to efficiently guide adapter retrieval and fusion. However, they either require unknown-task data or involve loading and unloading multiple LoRA adapters for each input, which can be computationally impractical. Moreover, they rely on inflexible, deterministic fusion for unseen tasks.

2.3. Neural Network Parameters Generation

Although generative modeling has advanced considerably, the direct generation of network weights for pre-trained models remains an emerging area of study [15, 16, 43, 54]. Approaches such as generative hyper-representation learning [6], neural network diffusion [9, 13, 30], and kernel density estimation–based methods [38] have shown promise but remain fundamentally limited to small architectures and unconditional weight generation within fixed distributions. Consequently, these methods struggle to generalize to unseen tasks, constraining their broader applicability. While meta-learning frameworks [27, 49] have enabled powerful joint model generation for visual recognition and few-shot learning, they often neglect the personalization and parameter diversity, restricting the generator's output to classifier heads rather than more flexible and expressive parameter sets, such as LoRA. ICM-LoRA[35] innovatively explores the parameter relationships among tasks through task vectors, but focuses on closed-world, task-specific enhancements of LoRA parameters. As a result, the question of whether one can rapidly generate efficient, user-intent–focused LoRA parameters in open-world settings remains unexplored.

3. The Proposed Model

In this section, we begin with an overview of essential concepts for understanding semantic-guided LoRA parameter generation, followed by detailed descriptions of our proposed methods.

3.1. Problem Defination and Preliminary

We define Zero-Shot Open-world Adaptation (ZSOA) as a novel and challenging task setting that requires models to generalize across semantically diverse tasks in open-world scenarios. Unlike conventional zero-shot learning or task transfer—often confined to fixed label spaces or narrow domains—*ZSOA focuses on tasks that share a common structural format (e.g., image-text retrieval) but differ substantially in domain, content, or distribution*. It emphasizes that queried tasks during inference time are drawn from an undefined and unbounded set, requiring rapid adaptation to novel tasks without access to raw data. This setting reflects realistic deployment scenarios, where task-level generalization must rely solely on prior experience.

In this work, we instantiate ZSOA in the context of fine-grained image-text retrieval, with each task formulated as a retrieval problem over a specific semantic category (e.g., animal species, flower type). Formally, given a set of fine-tuned LoRA module $\mathcal{W}$ trained on known tasks $\mathcal{T}$, our goal is to adapt the model to unseen task $\mathcal{T}^*$ without accessing any labeled image-text pairs. Let $f(\mathcal{T})$ denote the textual description of task $\mathcal{T}$, we learn a generator G that predicts LoRA parameters for $\mathcal{T}^*$ based on semantic descriptions and $\mathcal{W}$:

$$\mathcal{W}^* = G\left(f(\mathcal{T}^*), \mathcal{W}, f(\mathcal{T})\right), \tag{2}$$

The synthesized LoRA parameter $\mathcal{W}^*$ is then used to modulate a frozen vision-language backbone, enabling it to perform the image-text retrieval task defined by $\mathcal{T}^*$. ZSOA thus extends traditional zero-shot learning by enabling parameter-level generalization, accommodating diverse user-instructed tasks with open-world queries.

3.2. LoRA Parameter Dataset Construction

3.2.1. Task-Specific LoRA Training

The first stage involves constructing a dataset of LoRA parameters. Consider a collection of N distinct tasks $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_N\}$, where each $\mathcal{T}_n$ corresponds to a specific retrieval task (e.g., image-text retrieval for *Cat* category). In this context, the task is dataset-agnostic—that is, it may originate from the same dataset or from a different one. Given a pre-trained vision-language model (VLM), we train a task-specific LoRA for each $\mathcal{T}_n$ using corresponding image-text pairs, applying LoRA modules at consistent positions within the VLM. To ensure comparability and reduce variance, all LoRA modules are trained using identical network configurations by default. After training stabilizes, we extract and store the LoRA from the final M epochs, yielding task-specific parameter data:

$$\Delta \boldsymbol{W}_n = \{\Delta \boldsymbol{W}_n^m\}_{m=1}^M, \quad \mathbf{d}_n = f(\mathcal{T}_n), \tag{3}$$

where $\Delta \mathbf{W}_n^m = [\boldsymbol{B}_n^m, \boldsymbol{A}_n^m]$ denotes the concatenation of LoRA parameters in layer-wise order for task n from the m-th saved epoch. $\mathbf{d}_n$ denotes the textual description associated with $\mathcal{T}_n$, generated using the template *'a photo of a <class name>'*, and encoded by the frozen CLIP text encoder $f(\cdot)$ to serve as a global semantic representation. Collectively, these form the LoRA parameter dataset:

$$\mathcal{W} = \{\Delta \boldsymbol{W}_1, \dots, \Delta \boldsymbol{W}_N\}, \tag{4}$$

where each element is optimized for image-text alignment within its respective task.

3.2.2. LoRA Expert Repository Formation

To construct a reliable expert LoRA space that simulates diverse LoRA resources, we first curate a representative subset of tasks from the full corpus $\mathcal{W}$, where each task is associated with its corresponding LoRA parameters and semantic embedding. For each selected task, we compute the mean LoRA parameters $\boldsymbol{\mu}_e$ over all available adaptations, yielding a distilled representation of its task-specific adaptation pattern. These averaged parameters, along with their associated semantic embeddings, constitute our expert repository:

$$\mathcal{W}_{\text{expert}} = \{(\boldsymbol{\mu}_e, \mathbf{d}_e) \mid e \in \mathcal{E}\}, \quad \boldsymbol{\mu}_e = \frac{1}{M} \Delta \boldsymbol{W}_e, \tag{5}$$

where $\mathcal{E}$ represents the selected expert index and $\mathcal{W}_{\text{expert}}$ serves as a compact yet expressive basis for capturing the essential characteristics of each knowledge domain. The remaining LoRA data are then split into training and evaluation sets for subsequent model development and evaluation.

3.3. Semantic-guided LoRA Parameter Generation

Given the collected LoRA repository $\mathcal{W}_{\text{expert}}$, our `SG-LoRA` framework generates conditional LoRA parameters under semantic guidance. We first define the task semantics by selecting and combining the most relevant expert LoRAs from the repository using a sparse aggregator. A conditional variational autoencoder (CVAE)[37] is then trained to generate target LoRA parameters aligned with the corresponding task semantics. Once trained, the model can generalize to unseen tasks in an open-world query setting without further training.

3.3.1. Construction of task semantics

Intuitively, not all experts contribute equally to an unseen task. Therefore, it is essential to identify and prioritize the most beneficial experts. Fortunately, the CLIP textual encoder is well-suited for this purpose, as it effectively captures rich semantic relationships across tasks. As shown in Eq. 3, we use each task's textual embedding as a global semantic descriptor for its LoRA parameters. For an unseen task $\mathcal{T}^*$ with textual embedding $\mathbf{d}^*$, we compute cosine similarities with all expert embeddings $\{\boldsymbol{d}_e | e \in \mathcal{E}\}$ and select the top-K experts with the highest similarity scores to form a semantically tailored expert set, indexed by $\mathcal{I}_{\text{top-}K}$. Then, we normalize their similarity scores using the softmax function to obtain the fusion coefficients:

$$\alpha_k = \frac{\exp\left(\text{sim}(\mathbf{d}^*, \mathbf{d}_k)/\tau\right)}{\sum_{k' \in \mathcal{I}_{\text{top-}k}} \exp\left(\text{sim}(\mathbf{d}^*, \mathbf{d}_{k'})/\tau\right)}, \quad k \in \mathcal{I}_{\text{top-}K}, \tag{6}$$

where $\tau > 0$ is a temperature parameter. The semantic vector for task $\mathcal{T}^*$ is computed as a weighted sum:

$$\boldsymbol{\mu}^* = \sum_{k \in \mathcal{I}_{\text{top-}K}} \alpha_k \cdot \boldsymbol{\mu}_k, \tag{7}$$

The attention strategy in Eq. 7 guides our model to understand the unseen task with the expert LoRAs from the textual perspective, resulting in high-quality semantic representation.

To capture the semantic diversity of the task, we also consider estimating the element-wise variance for task $\mathcal{T}^*$ under the Law of Total Variance theory:

$$\boldsymbol{\sigma}^{*2} = \sum_{k=1}^{K} \alpha_k \boldsymbol{\sigma}_k^2 + \sum_{k=1}^{K} \alpha_k (\boldsymbol{\mu}_k - \boldsymbol{\mu}^*) \odot (\boldsymbol{\mu}_k - \boldsymbol{\mu}^*), \quad (8)$$

where $\odot$ denotes element-wise multiplication and $\boldsymbol{\sigma}_k$ denotes the variance of the k-th expert. This flexible formulation enables the model to better reflect the statistical properties of new tasks, which is crucial for generative modeling. The estimated mean and variance guide the generative process, resulting in more accurate and context-aware outputs, thereby improving generalization and robustness. We leave the derivation of Eq. 8 in Appendix B. To simplify, we will use c to represent the task semantics of $\mathcal{T}^*$ in the following descriptions, *e.g.*, $c = \{\boldsymbol{\mu}^*, \boldsymbol{\sigma}^{*2}\}$.

3.3.2. Conditional LoRA Parameter Generation

We adopt a conditional variational autoencoder framework to generate target LoRA parameters based on the task semantics calculated above. Given a batch of training LoRA tensor $\boldsymbol{X}$, the encoder approximates the posterior distribution $q(\boldsymbol{z}|\boldsymbol{X}, c)$ using a multi-layer perceptron (MLP) that takes the $\boldsymbol{X}$ to be reconstructed and the task semantics c as input. A latent code $z \sim q(\boldsymbol{z}|\boldsymbol{X}, c)$ is sampled and passed to the decoder along with c as condition to reconstruct the original input $\boldsymbol{X}$. Unlike traditional VAEs that adopt $p(\boldsymbol{z}) = \mathcal{N}(0, \mathbf{I})$ as the prior distribution, we here develop a semantic-aware prior for each task $p(\boldsymbol{z}|c)$. $p(\boldsymbol{z}|c)$ is parameterized with stacked MLPs, allowing the model to flexibly represent a task-specific prior distribution based on domain-level statistics.

The model is trained by minimizing the negative evidence lower bound (ELBO), which consists of two terms: the reconstruction and the regularization term :

$$\mathcal{L}_{\text{CVAE}} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{X},c)} \left[\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 \right] + \lambda \cdot KL(q(\boldsymbol{z}|\boldsymbol{X}, c) \| p(\boldsymbol{z}|c)). \quad (9)$$

where $\hat{\boldsymbol{X}}$ denotes the reconstructed LoRA parameters, $KL(\cdot \parallel \cdot)$ is the Kullback-Leibler divergence, and λ controls the relative weight of the *KL* term. The first term encourages the decoder to reconstruct accurate LoRA parameters, while the second term regularizes the latent space to align the task-specific prior. During inference, a sample z is drawn from the prior distribution $p(\boldsymbol{z}|c)$, and the decoder generates the corresponding custom LoRA parameter.

4. Experiments

4.1. Experimental Settings

Datasets. The proposed `SG-LoRA` is evaluated on three benchmark datasets. Specifically, we use the widely adopted MS-COCO dataset [22], a standard benchmark for image-text retrieval, known for its diverse scenes and rich linguistic annotations. To evaluate the model's generalization ability, we further include the OxfordPets dataset [29] and the Flowers102 dataset [28], both of which are originally designed for fine-grained image classification. Moreover, given the inherent ambiguity and limited informativeness of MS-COCO captions and the absence of captions in the other two datasets, we construct synthetic textual descriptions using Qwen2-VL [44]. For MS-COCO, we use the original training split but regenerate captions for each image, effectively creating a new image–caption dataset, and then divide it into training, validation, and test sets.

Metrics. The evaluation metric used is Recall@K (R@K), which quantifies the proportion of correct matches appearing in the top-K retrieved candidates. We report R@1, R@5, and R@10 for both image-to-text and text-to-image retrieval scenarios.

Implementation Details. We adopt CLIP ViT-B/16 as our backbone, injecting rank-2 LoRA adapters into the W_q, W_k, and W_v projection matrices of every Transformer block in the visual encoder. Training is carried out with the Adam optimizer. The CVAE's encoder and prior network each consist of two-layer MLPs with ReLU activations, whereas the decoder is realized as a three-layer MLP with ReLU activations. We set the default values of M, K, and λ in the model to 100, 4, and 1, respectively. All experiments were performed on a single NVIDIA A6000 GPU.

4.2. Comparative Methods

To evaluate the effectiveness of the proposed method, we compare it with the following methods: (1) **Zero-Shot CLIP**: The original CLIP model without any adaptation of LoRA modules. (2) **Model Soups**: Consistent with [45] all LoRA experts in $\mathcal{W}_{\text{expert}}$ are uniformly averaged without considering their relevance to the target task. (3) **AdapterSoup (Top-*k* LoRA Merging)**: The top-K experts from the expert repository are selected based on the semantic similarity vector and with equal weight, assigning each a coefficient of $1/K$. This can be seen as an variant of [3] revised to our setting. (4) **Top-*K* LoRA Weighted**: The top-K experts are selected based on the semantic similarity vector, and their weights are computed by applying a softmax function over the similarity scores for adaptive merging. (5) **`SG-LoRA`**: Our proposed method, which generates task-specific LoRAs based on semantic proximity. (6) **Oracle:** For each task, LoRA parameters are trained individually on the specific dataset where we evaluate.

4.3. Main Results and Discussion

4.3.1. In-Dataset Image-Text Retrieval

We first conducted in-dataset evaluations on MS-COCO and OxfordPets dataset separately, with results shown in Table.1.

Table 1. Model Performance Comparison on MS-COCO and OxfordPets Datasets. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	MS-COCO						OxfordPets					
	I2T			T2I			I2T			T2I		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-Shot CLIP	66.43	84.31	89.14	41.66	64.63	73.01	40.45	66.27	77.53	26.03	50.66	62.98
Oracle	72.45	88.91	93.41	53.10	76.47	83.97	55.84	81.84	89.13	40.99	70.41	80.39
Model Soups	69.37	85.96	90.95	47.38	69.54	77.97	52.54	77.80	85.59	33.51	61.77	72.93
AdapterSoup	70.70	86.57	91.09	48.64	70.51	78.79	52.59	78.52	86.09	34.05	62.70	73.93
Top-K LoRA Weighted	71.55	87.54	91.69	49.85	71.79	79.66	53.96	79.41	86.53	35.42	64.08	74.99
SG-LoRA	**74.31**	**88.78**	**92.50**	**54.42**	**75.45**	**82.18**	**57.15**	**80.40**	**88.04**	**37.62**	**67.16**	**77.44**

Table 2. Cross-Dataset Generalization Performance: Bidirectional Evaluation between MS-COCO and OxfordPets. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	MS-COCO → OxfordPets						OxfordPets → MS-COCO					
	I2T			T2I			I2T			T2I		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-Shot CLIP	40.45	66.27	77.53	26.03	50.66	62.98	66.43	84.31	89.14	41.66	64.63	73.01
Oracle	55.84	81.84	89.13	40.99	70.41	80.39	72.45	88.91	93.41	53.10	76.47	83.97
Model Soups	44.67	70.91	80.78	30.45	56.77	68.52	68.58	85.67	90.62	44.09	66.55	75.08
AdapterSoup	45.96	71.83	81.42	30.88	57.32	69.08	68.74	85.83	90.63	44.19	66.58	75.31
Top-K LoRA Weighted	48.13	73.43	82.73	33.34	59.53	70.89	68.75	85.77	90.67	44.60	66.76	75.25
SG-LoRA	**55.41**	**80.73**	**87.33**	**38.84**	**66.77**	**76.69**	**70.81**	**86.83**	**91.41**	**46.50**	**68.73**	**77.19**

Several key observations are as follows: 1). Compared to the Zero-Shot CLIP baseline and consistent with previous findings [32], directly merging all experts in the expert repository leads to performance improvements. 2). Selecting semantically relevant experts, like those most related to the current query task from the repository, can further enhance performance. However, naively treating all selected experts equally may result in degraded performance. For instance, in Table 1, AdapterSoup underperforms compared to Top-K LoRA Weighted. This may be because different experts contribute unevenly to the target task. Assigning equal weight ignores these differences and may amplify noise from less relevant experts. By contrast, incorporating semantic weighting coefficients allows the fusion process to account for varying degrees of relevance, leading to more effective integration of expert knowledge and improved retrieval performance. 3). While merging-based approaches still fall short of the performance achieved by directly fine-tuning LoRA on the unseen task, SG-LoRA fully recovers the performance of oracle adapters.

Notably, SG-LoRA even outperforms Oracle in certain cases—for example, R@1 for bidirectional retrieval on MS-COCO and R@1 for image-to-text retrieval on OxfordPets. This improvement gains from the efficient compression of expert LoRAs: our trained CVAE integrates the target LoRA using compact yet semantically rich task representations, enabling the generation of target-aligned LoRA parameters by modeling their distribution in the parameter space. Moreover, the Oracle LoRA fine-tuned on unseen tasks sometimes suffers from overfitting, especially when trained on a small set of image-caption pairs. Our SG-LoRA helps mitigate this performance drop, likely due to its ability to generalize without relying on target-task data.

4.3.2. Cross-Dataset Image-Text Retrieval

Given the flexibility of SG-LoRA, we conducted a more challenging cross-dataset evaluation. As shown in Table.2, SG-LoRA consistently outperforms merging-based approaches in these settings. Interestingly, we also observed that models trained on MS-COCO were able to generate LoRA parameters that, in some cases, outperformed those trained directly on the OxfordPets (e.g., a relative improvement of 1.22% in T2I R@1). This may be attributed to the richer expert knowledge available in MS-COCO, which, due to its broader data diversity, enables more extensive exploration of the parameter space during generation, a capability not achievable when generating within the narrower

Table 3. Ablation on expert repository strategy for cross-dataset evaluation. We evaluate how the MS-COCO *Cat* expert affects retrieval on two unseen OxfordPets cat tasks (marked with gray text).

Expert strategy	*Egyptian Mau* **I2T**			*Egyptian Mau* **T2I**			Expert strategy	*Persian* **I2T**			*Persian* **T2I**		
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10
w/o *Cat* expert	36.08	62.89	71.13	15.21	31.70	44.07	w/o *Cat* expert	44.00	80.00	87.00	34.00	62.25	72.75
w/ *Cat* expert	37.11	63.92	72.16	15.21	35.05	46.91	w/ *Cat* expert	47.00	79.65	86.00	36.75	64.25	73.75

scope of OxfordPets. Another possible reason is that the uniform LoRA training configuration used across datasets (as described in Section 3.2.1) may not be optimal for OxfordPets. Conversely, we find that when the generation model is trained on OxfordPets and applied to MS-COCO, its performance is generally worse than that of models trained on MS-COCO. More comparisons are reported in Appendix C.

4.4. Evaluation on Standard Image-Text Retrieval

Considering another complex scenario where the test task may contain image-text pairs from multiple categories, we evaluated retrieval performance on the Flickr30K test set [31]. Since this dataset has no clear category distinction during retrieval, we randomly selected one caption per image, fed it into the CLIP textual encoder to obtain the textual embedding, and then calculated the mean value as the task description for retrieval. The experimental results are shown in Figure.2, where `SG-LoRA` outperforms Zero-Shot CLIP. Additionally, `SG-LoRA` trained on MS-COCO achieves better results than that trained on OxfordPets. This is because MS-COCO provides more comprehensive expert knowledge, covering a wider range of categories, while OxfordPets primarily focuses on fine-grained distinctions within just two broad categories—cats and dogs. This also indicates that when semantic guidance is more powerful and comprehensive, or more relevant to the downstream task, the generated LoRA parameters are also superior.

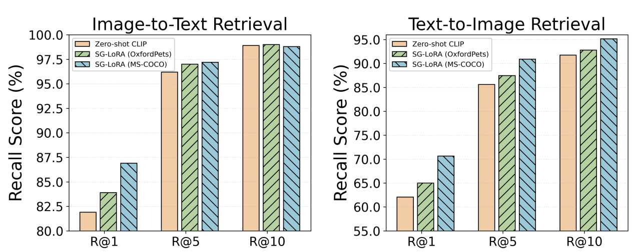

Figure 2. Simulation of task-agnostic evaluation on the general image-text retrieval dataset. We train `SG-LoRA` on MS-COCO and OxfordPets, respectively, and evaluate them on the Flickr30K test set (zoom in for a better view).

4.5. Ablation Stduy

Impact of expert repository configuration. In the experiment of Table. 2, we observed that the *Cat* expert from the MS-COCO dataset was frequently selected as a semantic condition. To further evaluate the impact of semantically salient experts, we assessed how the inclusion of the *Cat* expert from MS-COCO influences retrieval performance on two unseen cat classes on OxfordPets. As shown in Table.3, incorporating the *Cat* expert into the expert repository improves performance in most cases, particularly for text-to-image retrieval. This highlights the effectiveness of semantically guided expert selection. This finding also demonstrates the flexibility of our method in constructing task-adaptive expert repositories—particularly when a richer pool of LoRA resources is available.

Given that our model supports open-world expert repository construction, we further conducted a case study where experts from OxfordPets, MS-COCO, and Flowers102 were combined into a mixed-source repository. We also combined training data from both datasets to train the `SG-LoRA` model accordingly. Figure.3 presents a comparison between single-source and mixed-source expert configurations. Additionally, we present the top-4 experts selected by `SG-LoRA` under the mixed-source setup for the unseen *Yorkshire Terrier* task, along with their corresponding weights in Table.4. As shown, the mixed-source experts yield better performance than the single-source experts. The performance even surpasses that of the oracle LoRA model in text-to-image retrieval. These demonstrate the potential of our method in more realistic, real-time application scenarios, where expert repositories are constructed dynamically from heterogeneous data sources.

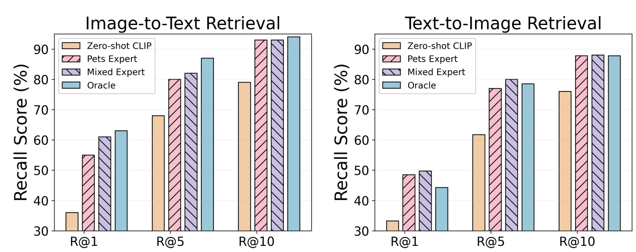

Figure 3. Comparison on expert repository configuration: single-source experts vs. mixed-source experts.

Impact of the number of experts. We conduct ablation studies on the number of experts K used in task semantic construction. As shown in Figure.4, using too few experts leads to insufficient knowledge for generalizing to unseen tasks, while increasing K incorporates more semantic information but may also introduce irrelevant context. Overall,

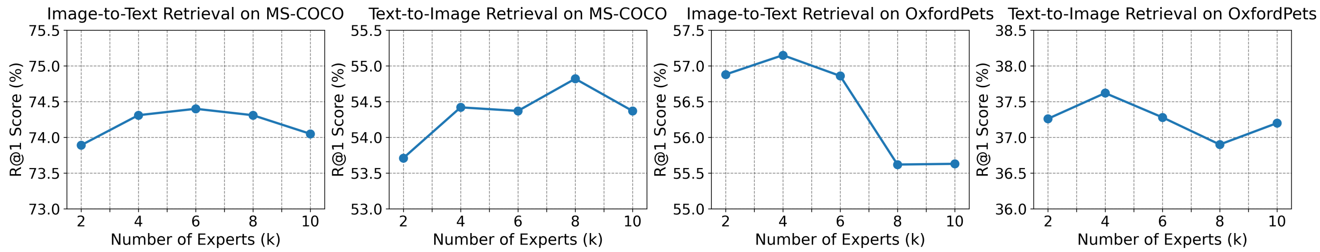

Figure 4. Ablaton study on the number of experts.

$K = 4$ shows a good balance between expert diversity and relevance, delivering good performance on both datasets.

Table 4. Top-4 expert for *Yorkshire Terrier* image-text retrieval under mixed-source experts configurations.

Expert Task	Source Dataset	Contribution
Scottish Terrier	OxfordPets	0.9221
Dog	MS-COCO	0.0692
Cat	MS-COCO	0.0082
American Bulldog	OxfordPets	0.0005

Ablation study on modalities of semantic prior condition. The construction of semantic priors serves as the foundation for our SG−LoRA. In Table.5, we compare the performance of semantic conditions across different modalities, where the conditional task description from each modality influences the selection of experts for unseen tasks by affecting α_k in Eq.6. The visual condition is obtained by averaging the visual embeddings of training set images within each task dataset using a frozen CLIP visual encoder. Experimental results show that the textual condition better captures the semantic relationships between tasks. This could be attributed to two factors. Firstly, the high degree of condensation in textual semantics might play a role. Secondly, discrepancies between training and test images, as well as the presence of noisy samples, may introduce inaccuracies into the visual prior condition.

Table 5. Ablation study on modalities of semantic prior condition.

Condition	Metrics		Dataset
	I2T R@1	T2I R@1	
Visual	73.16	52.70	MS-COCO
Textual	74.31	54.42	
Visual	86.30	70.12	Flickr30K
Textual	86.90	70.66	

5. Qualitative Analysis

To further investigate the parameter diversity of SG−LoRA, we conducted evaluations on the unseen '*Zebra*' task from the MS-COCO dataset and visualized the generated LoRAs using t-SNE. As shown in Figure 5, we observe that the distribution of LoRAs generated by SG−LoRA (in green) exhibits diversity rather than extensively overlapping with the Oracle. This indicates that, by injected stochasticity, our method effectively explores the high-performance LoRAs in the parameter space.

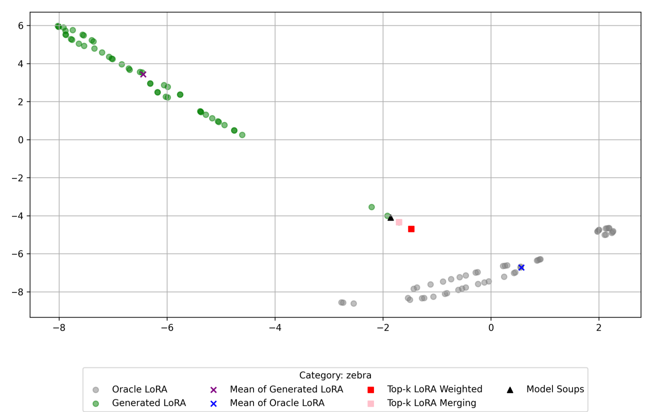

Figure 5. t-SNE visualization of LoRA parameters on the unseen '*Zebra*' retrieval task. For SG−LoRA and Oracle LoRA, we randomly sampled 50 samples each. (zoom-in for more details).

6. Conclusion

In this work, we introduce a novel and challenging task setting termed Zero-Shot Open-world Adaptation (ZSOA), which requires models to generalize across semantically diverse tasks in open-world scenarios. To achieve this, we propose a flexible and efficient approach that dynamically generates task-specific LoRA parameters guided by available LoRA resources. By identifying the most relevant expert knowledge based on semantic similarity and leveraging task semantics in a conditional generative framework, our SG−LoRA models the distribution of unseen parameters in a tuning-free manner. The inherent stochasticity of our generation process further introduces diversity, enhancing adaptability to previously unseen tasks. SG−LoRA is scalable and naturally privacy-preserving, making it well-suited for deployment in sensitive and dynamic real-world environments.

Acknowledgments

This research was supported by the Hong Kong RGC General Research Fund (Grant Nos. 15221123, 15216424, and 15211525) and the Hong Kong PolyU Internal Research Fund (Grant Nos. P0058468 and P0056171).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Yang Chen, Jingcai Guo, Miaoge Li, Zhijie Rao, and Song Guo. Star++: Region-aware conditional semantics via interpretable side information for zero-shot skeleton action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[3] Alexandra Chronopoulou, Matthew E Peters, Alexander Fraser, and Jesse Dodge. Adaptersoup: Weight averaging to improve generalization of pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2054–2063, 2023.

[4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[5] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235, 2023.

[6] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1 (2):3, 2022.

[9] Wei Hu, QiHao Zhao, Yangyu Huang, and Fan Zhang. P-diff: Learning classifier with noisy labels based on probability difference distributions. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1882–1889. IEEE, 2021.

[10] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023.

[11] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277, 2022.

[12] Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model stock: All we need is just a few fine-tuned models. In *European Conference on Computer Vision*, pages 207–223. Springer, 2024.

[13] Xiaolong Jin, Kai Wang, Dongwen Tang, Wangbo Zhao, Yukun Zhou, Junshu Tang, and Yang You. Conditional lora parameter generation. *arXiv preprint arXiv:2408.01415*, 2024.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Boris Knyazev, Michal Drozdzal, Graham W Taylor, and Adriana Romero Soriano. Parameter prediction for unseen deep architectures. *Advances in Neural Information Processing Systems*, 34:29433–29448, 2021.

[16] Boris Knyazev, Doha Hwang, and Simon Lacoste-Julien. Can we scale transformers to predict parameters of diverse imagenet models? In *International Conference on Machine Learning*, pages 17243–17259. PMLR, 2023.

[17] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[19] Miaoge Li, Dongsheng Wang, Xinyang Liu, Zequn Zeng, Ruiying Lu, Bo Chen, and Mingyuan Zhou. Patchct: Aligning patch set and label set with conditional transport for multi-label image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15348–15358, 2023.

[20] Miaoge Li, Jingcai Guo, Richard Yi Da Xu, Dongsheng Wang, Xiaofeng Cao, Zhijie Rao, and Song Guo. Tsca: On the semantic consistency alignment via conditional transport for compositional zero-shot learning. *arXiv preprint arXiv:2408.08703*, 2024.

[21] Weishi Li, Yong Peng, Miao Zhang, Liang Ding, Han Hu, and Li Shen. Deep model fusion: A survey. *arXiv preprint arXiv:2309.15698*, 2023.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[24] Xinyang Liu, Dongsheng Wang, Bowei Fang, Miaoge Li, Zhibin Duan, Yishi Xu, Bo Chen, and Mingyuan Zhou. Patch-prompt aligned bayesian prompt tuning for vision-language models. *arXiv preprint arXiv:2303.09100*, 2023.

[25] Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18733–18741, 2024.

[26] Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. Adsqa: Towards advertisement video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23396–23407, 2025.

[27] Elvis Nava, Seijin Kobayashi, Yifei Yin, Robert K Katzschmann, and Benjamin F Grewe. Meta-learning via classifier (-free) diffusion guidance. *arXiv preprint arXiv:2210.08942*, 2022.

[28] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[29] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[30] William Peebles, Ilija Radosavovic, Tim Brooks, Alexei A Efros, and Jitendra Malik. Learning to learn with generative models of neural network checkpoints. *arXiv preprint arXiv:2209.12892*, 2022.

[31] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[32] Reza Qorbani, Gianluca Villani, Theodoros Panagiotakopoulos, Marc Botet Colomer, Linus Härenstam-Nielsen, Mattia Segu, Pier Luigi Dovesi, Jussi Karlgren, Daniel Cremers, Federico Tombari, et al. Semantic library adaptation: Lora retrieval and fusion for open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9804–9815, 2025.

[33] Zhijie Rao and Jingcai Guo. Balancing cross-modal attention for generalized zero-shot learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3360–3369, 2025.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[35] Yihua Shao, Minxi Yan, Yang Liu, Siyu Chen, Wenjie Chen, Xinwei Long, Ziyang Yan, Lei Li, Chenyu Zhang, Nicu Sebe, et al. In-context meta lora generation. *arXiv preprint arXiv:2501.17635*, 2025.

[36] Donald Shenaj, Ondrej Bohdal, Mete Ozay, Pietro Zanuttigh, and Umberto Michieli. Lora. rar: Learning to merge loras via hypernetworks for subject-style conditioned image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16132–16142, 2025.

[37] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[38] Bedionita Soro, Bruno Andreis, Hayeon Lee, Wonyong Jeong, Song Chong, Frank Hutter, and Sung Ju Hwang. Diffusion-based neural network weights generation. *arXiv preprint arXiv:2402.18153*, 2024.

[39] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237, 2022.

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[41] Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities. *Advances in Neural Information Processing Systems*, 36:52792–52810, 2023.

[42] Dongsheng Wang, Jiequan Cui, Miaoge Li, Wang Lin, Bo Chen, and Hanwang Zhang. Instruction tuning-free visual token complement for multimodal llms. In *European Conference on Computer Vision*, pages 446–462. Springer, 2024.

[43] Kai Wang, Dongwen Tang, Boya Zeng, Yida Yin, Zhaopan Xu, Yukun Zhou, Zelin Zang, Trevor Darrell, Zhuang Liu, and Yang You. Neural network diffusion. *arXiv preprint arXiv:2402.13144*, 2024.

[44] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[45] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

[46] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.

[47] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024.

[48] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

[49] Baoquan Zhang, Chuyao Luo, Demin Yu, Xutao Li, Huiwei Lin, Yunming Ye, and Bowen Zhang. Metadiff: Meta-learning with conditional diffusion for few-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 16687–16695, 2024.

[50] Jinghan Zhang, Junteng Liu, Junxian He, et al. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36:12589–12610, 2023.

[51] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and

Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

[52] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[53] Ziyu Zhao, Tao Shen, Didi Zhu, Zexi Li, Jing Su, Xuwu Wang, Kun Kuang, and Fei Wu. Merging loras like playing lego: Pushing the modularity of lora to extremes through rank-wise clustering. *arXiv preprint arXiv:2409.16167*, 2024.

[54] Andrey Zhmoginov, Mark Sandler, and Maksym Vladymyrov. Hypertransformer: Model generation for supervised and semi-supervised few-shot learning. In *International Conference on Machine Learning*, pages 27075–27098. PMLR, 2022.

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[56] Xingyu Zhu, Shuo Wang, Beier Zhu, Miaoge Li, Yunfan Li, Junfeng Fang, Zhicai Wang, Dongsheng Wang, and Hanwang Zhang. Dynamic multimodal prototype learning in vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2501–2511, 2025.

[57] Xiandong Zou, Mingzhu Shen, Christos-Savvas Bouganis, and Yiren Zhao. Cached multi-lora composition for multi-concept image generation. *arXiv preprint arXiv:2502.04923*, 2025.