This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Coeff-Tuning: A Graph Filter Subspace View for Tuning Attention-Based Large Models

Zichen Miao†, Wei Chen†, Qiang Qiu
Purdue University, IN, USA
{miaoz, chen2732, qqiu}@purdue.edu

Abstract

Transformer-based large pre-trained models have shown remarkable generalization ability, and various parameter-efficient fine-tuning (PEFT) methods have been proposed to customize these models on downstream tasks with minimal computational and memory budgets. Previous PEFT methods are primarily designed from a tensor-decomposition perspective that tries to effectively tune the linear transformation by finding the smallest subset of parameters to train. Our study adopts an orthogonal view by representing the attention operation as a graph convolution and formulating the multi-head attention maps as a convolutional filter subspace, with each attention map as a subspace element. In this paper, we propose to tune the large pre-trained transformers by learning a small set of combination coefficients that construct a more expressive filter subspace from the original multi-head attention maps. We show analytically and experimentally that the tuned filter subspace can effectively expand the feature space of the multi-head attention and further enhance the capacity of transformers. We further stabilize the fine-tuning with a residual parameterization of the tunable subspace coefficients, and enhance the generalization with a regularization design by directly applying dropout on the tunable coefficient during training. The tunable coefficients take a tiny number of parameters and can be combined with previous PEFT methods in a plug-and-play manner. Extensive experiments show that our approach achieves superior performances than PEFT baselines with neglectable additional parameters.[1]

1. Introduction

Large foundational models have demonstrated remarkable performance and generalization ability across a wide range of natural language processing and computer vision tasks [11, 28, 55, 68]. The success of these models can be attributed to their ability to learn rich, transferable representations by leveraging the power of the transformer architecture [68] and the extensive knowledge captured during

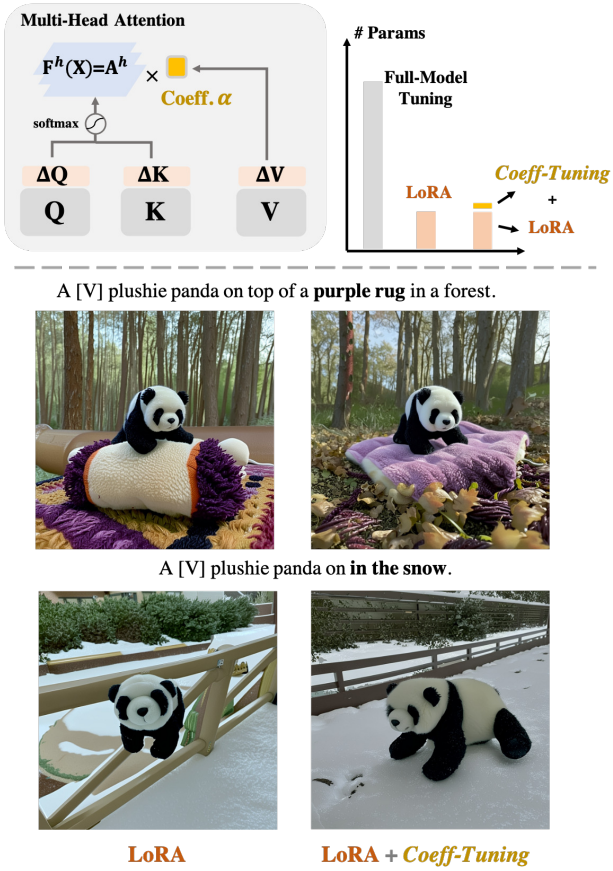


Figure 1. (Up) Under the graph convolution view of the attention, the proposed *Coeff-Tuning* enhances the expressiveness by combining the attention maps with a tiny tunable subspace coefficient α. It can be seamlessly integrated with other PEFT methods, e.g., LoRA [21], by introducing neglectable additional parameters. (Bottom) Combining *Coeff-Tuning* with LoRA achieves better results in the concept customization task with SDXL [51].

†Equal contribution.
[1]Code is available at: https://github.com/ZichenMiao/Coeff_Tuning

pre-training on large-scale datasets [1, 58]. However, fine-tuning these pre-trained transformers on downstream tasks with full-parameter updates can be challenging due to the high computational and memory costs.

To address this, various parameter-efficient fine-tuning (PEFT) methods have been developed to adapt pre-trained models on downstream tasks while minimizing the number of trainable parameters [21, 38, 53, 79]. A representative family of these approaches focuses on tuning the linear transformation layers in transformers, where they apply a lightweight parameterization, e.g., low-rank parameterization [21, 38], to adapt the weight matrices. However, none of these methods are designed from a holistic view of the core operation in transformers, the multi-head attention layer [68], which plays a key role in the expressiveness of transformers [10, 59].

In this paper, we propose a new PEFT method tailored to enhance the multi-head attention's expressiveness by tuning a minimal set of parameters. We first formulate the multi-head attention as a graph convolution with a filter subspace, with each graph filter as the attention map of each attention head. We then tune the filter subspace by simply learning a small set of subspace coefficients that linearly combine the pre-trained attention maps. In this way, we adapt for each attention head a new task-adaptive attention map that better captures the token-wise correlation.

We attribute the effectiveness of our approach to its ability to expand the feature space of the original multi-head attention layer. By applying the softmax function to the query-key correlation matrices, all elements in a row of the original attention map are non-negative and have a sum of one [68]. This limits the expressiveness of the multi-head attention as the output token embeddings lie within the convex hull of the value embeddings. Our method constructs a new filter subspace with non-constrained coefficients as shown in Figure 1, allowing negative elements learned in the combined attention map to capture more diverse token-wise relationships. This also enables the output tokens to obtain representation outside the convex hull of the value embeddings, which further enhances the capacity of the multi-head attention. We provide analysis and an illustrative toy example to better demonstrate this point.

We show that our approach can be seamlessly combined with previous PEFT methods in a plug-and-play manner. As the filter subspace in each attention layer is spanned by a tiny number of attention maps (usually 10$\sim$30), we only introduce neglectable additional tunable parameters and training costs over the base PEFT methods.

We validate our approach on a large variety of fine-tuning tasks and datasets, along with pre-trained transformers of different scales and modalities. Specifically, we first show the effectiveness of our method with ViT [11] on the few-shot image classification fine-tuning. We also show our method can achieve superior performance with vision-language transformers on image-text multi-modal fine-tuning [8]. Moreover, we also show our method can better customize the text-to-image diffusion models on personalized generation [56].

We summarize our contributions as follows,

- We provide a filter subspace view of the multi-head attention and propose to fine-tune a new filter subspace by learning a small set of subspace coefficients.
- We demonstrate that our method can expand the feature space of the original multi-head attention and enhance its expressiveness on downstream tasks.
- We validate our method on extensive fine-tuning tasks and various transformer architectures.

2. Related Work

Parameter-Efficient Fine-Tuning. Large attention-based transformers [11, 54, 55, 68] gain a significant amount of generalizable knowledge by pre-training on large-scale datasets [57, 58]. A common way to utilize those general skills is to fine-tune large transformers on domain-specific tasks and data. Recent advancements in parameter-efficient fine-tuning (PEFT) methods aim to reduce the computational and memory costs of adapting large models by fine-tuning only a subset of parameters. LoRA [21] and its variants [4, 24, 29, 36, 78, 82] fine-tune lower-rank matrices at each layer to represent weight updates. DoRA [38] enhances LoRA through weight decomposition into magnitude and direction, thus improving training stability and learning capacity without additional inference costs. The adapter [20] approach inserts small modules between layers and reduces parameters by only tuning these adapters [2, 24, 34, 50, 79]. Visual prompt tuning (VPT) [23, 60] has introduced a limited number of learnable parameters for optimization while keeping the backbone frozen. SSF [35] proposes scaling and shifting features of each layer in a pre-trained large model. OFT [53] tackles fine-tuning in text-to-image diffusion models by preserving the hyperspherical energy of neuron relationships. However, these methods primarily focus on efficient tuning of the parameter weight matrices, lacking a holistic view of the attention layer. In this paper, we take an orthogonal approach by tuning a tiny combination coefficient given a novel graph filter subspace view of the multi-head attention.

The Expressiveness of Attention. Transformer-based models have achieved success across vision and language tasks [8, 11, 14, 22, 31, 37] due to the effectiveness of the attention mechanism [68]. Despite their success, various studies have examined the limitations in the expressiveness of transformers [10, 59, 71, 80]. Shi et al. [59] and Zhou et al. [84] identify the attention collapse in transformers as

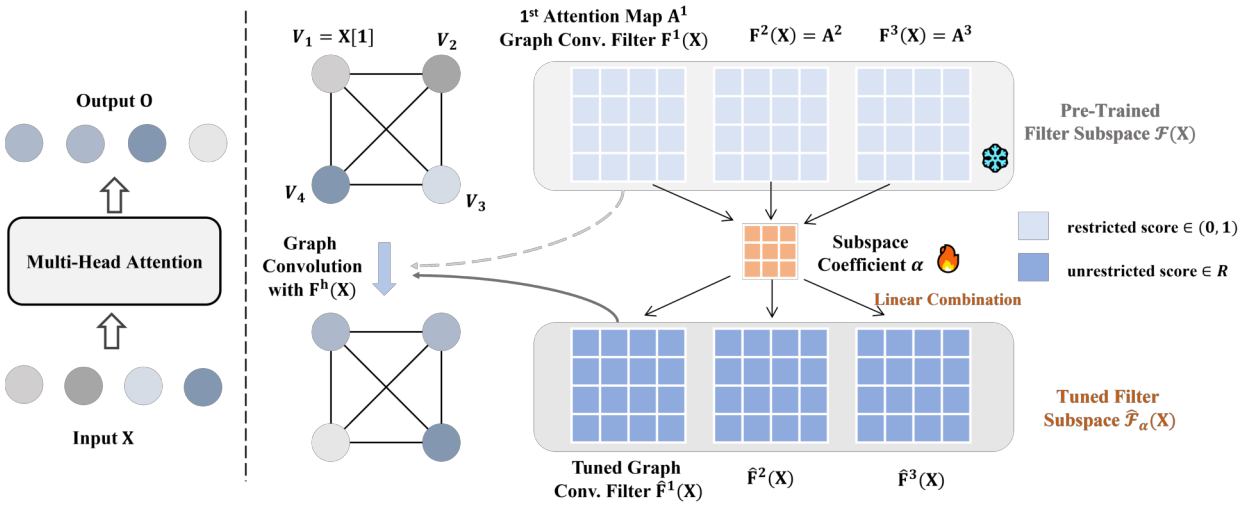


Figure 2. Illustration of the proposed *Coeff-Tuning*. We first view each token $\mathbf{X}[i]$ as a node V_i in an all-connected graph $\mathcal{G} = (V, E)$. The attention operation then turns into a graph convolution with each attention map $\mathbf{A}^h$ as a graph convolutional filter $\mathbf{F}^h(\mathbf{X})$. In this way, the attention maps in the multi-head attention layer form a filter subspace $\mathcal{F}(\mathbf{X})$, and each element in the filter is constrained in $(0, 1)$ after softmax. The proposed *Coeff-Tuning* then tunes a subspace coefficient $\boldsymbol{\alpha} \in \mathbb{R}^{H\times H}$ with a tiny number of parameters, that expands the range of attention scores by linearly combining the original filter subspace and enhances the expressiveness of the multi-head attention.

a result of self-attention's convergence to a low-rank space, which leads to unstable training as attention scores concentrate [10, 45]. To counteract this, techniques such as hierarchical fusion [59] and stabilizing attention entropy [80] have been proposed to prevent performance stagnation in vision transformers. DeepViT [84] proposes to re-generate diverse attention maps to avoid attention collapse in training vision transformers from scratch. In contrast, our work studies the expressiveness of the attention layer in the context of the large model fine-tuning. We also provide a novel interpretation of multi-head attention maps as a graph filter subspace, and propose the residual parameterization and regularization to stabilize the tuning process.

3. Method

3.1. A Filter Subspace View on Attention

Attention. The attention layer transforms the input sequence $\mathbf{X} \in \mathbb{R}^{N\times C_i}$ to the output sequence $\mathbf{O} \in \mathbb{R}^{N\times C_o}$, where N denotes the sequence length, C_i and C_o are the dimension of input and output features. The attention layer projects the input as $\mathbf{Q} = \mathbf{X}\mathbf{W}_q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_k$, $\mathbf{V} = \mathbf{X}\mathbf{W}_v$ using the corresponding projection matrices $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{C_i\times C_o}$, and calculates the attention map,

$$\begin{aligned}\mathbf{A} &= \text{softmax}(\mathbf{Q}\mathbf{K}^\intercal) \\ &= \text{softmax}(\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\intercal\mathbf{X}^\intercal).\end{aligned} \tag{1}$$

The attention map $\mathbf{A} \in \mathbb{R}^{N\times N}$ captures the token-wise relationship by doing inner-product in a space transformed by $\mathbf{W}_q, \mathbf{W}_k$. Then the output of the attention layer is,

$$\begin{aligned}\mathbf{O} &= \text{attn}(\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v; \mathbf{X}) \\ &= \text{softmax}(\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\intercal\mathbf{X}^\intercal)\mathbf{X}\mathbf{W}_v \\ &= \mathbf{A}\mathbf{X}\mathbf{W}_v.\end{aligned} \tag{2}$$

Multi-head Attention. Multi-head attention extends this by allowing multiple attention mechanisms, *i.e.*, heads, to work in parallel, with each head independently learning attention patterns. For each attention head, we have,

$$\mathbf{O}^h = \text{attn}(\mathbf{W}_q^h, \mathbf{W}_k^h, \mathbf{W}_v^h; \mathbf{X}), \tag{3}$$

where $h = 1, \cdots, H$ and H is the number of heads. The project matrices for each head are $\mathbf{W}_q^h, \mathbf{W}_k^h, \mathbf{W}_v^h \in \mathbb{R}^{C_i\times \frac{C_o}{H}}$. The multi-head attention represents as,

$$\begin{aligned}\mathbf{Z} &= \text{Concat}\left[\mathbf{O}^1, \cdots, \mathbf{O}^H\right]\mathbf{W}_o \\ &= \text{Concat}\left[\mathbf{A}^1\mathbf{X}\mathbf{W}_v^1, \cdots, \mathbf{A}^H\mathbf{X}\mathbf{W}_v^H\right]\mathbf{W}_o \\ &= \sum_{h=1}^{H}\mathbf{A}^h\mathbf{X}\mathbf{W}_{vo}^h,\end{aligned} \tag{4}$$

where $\mathbf{W}_o \in \mathbb{R}^{C_o\times C_o}$ and $\mathbf{W}_{vo}^h = \mathbf{W}_v^h(\mathbf{W}_o^h)^\intercal$, $\mathbf{W}_o^h \in \mathbb{R}^{C_o\times\frac{C_o}{H}}$, $\mathbf{W}_{vo}^h \in \mathbb{R}^{C_i\times C_o}$.

A Filter Subspace View on Multi-head Attention. We now present our filter subspace view on the multi-head attention. By viewing each token embedding $\mathbf{X}[n] \in \mathbb{R}^{C_i}$ as a node V_n, which are connected with all the other nodes with edges $E[m, n] = 1, \forall m \in [N]$, we obtain a graph representation of all the tokens, $\mathcal{G} = (V, E)$.

We then reformulate the attention operation as the message passing as in Graph Convolution Nerworks [27]. Specifically, inspired by [6], we write the single-head attention as a graph convolution with the filter $\mathbf{F}^h(\mathbf{X})$,

$$\mathbf{O}^h[n,:] = \sum_{\substack{n'\in[N]\\ c'\in[C_i]}} \mathbf{F}^h(\mathbf{X})[n,n']\,\mathbf{X}[n',c']\mathbf{W}^h[c',:], \tag{5}$$

where $\mathbf{W}^h = \mathbf{W}^h_{vo} \in \mathbb{R}^{C_i\times C_0}$ denotes the weight for linear projection on each node, and the n-th row of the filter, i.e., $\mathbf{F}^h(\mathbf{X})[n,:]$, represents the local graph filter around the n-th node (token). Comparing with (1) and (2), we have,

$$\mathbf{F}^h(\mathbf{X}) = \text{softmax}(\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\intercal\mathbf{X}^\intercal). \tag{6}$$

The multi-head attention individually learns H filters $\{\mathbf{F}^1(X),...,\mathbf{F}^H(X)\}$ to aggregate complementary information across nodes. These H filters naturally form a graph convolutional filter subspace $\mathcal{F}(\mathbf{X})$ as,

$$\mathcal{F}(\mathbf{X}) = \{\mathbf{F}^h(\mathbf{X})\}_{h=1}^H \subseteq (0,1)^{N\times N}, \tag{7}$$

and we denote each $\mathbf{F}^h(\mathbf{X})$ as a filter atom. The multi-head attention can then be represented in the way of decomposed graph convolution [6] as,

$$\mathbf{O}[n,:] = \sum_{h=1}^H \sum_{\substack{n'\in[N]\\ c'\in[C_i]}} \mathbf{F}^h(\mathbf{X})[n,n']\mathbf{X}[n',c']\mathbf{W}^h[c',:]. \tag{8}$$

It can be seen from (8) that the parameters $\mathbf{F}^h(\mathbf{X}), \mathbf{W}^h$ plays two key yet independent roles in the graph convolution. The $\mathbf{W}^h$ updates the knowledge for each node by projecting its representation into a new space, while the graph filter $\mathbf{F}^h(\mathbf{X})$ guides the information flow across nodes. Intuitively, with a more expressive graph convolutional filter subspace $\mathcal{F}(\mathbf{X})$, each node can aggregate more informative features from other nodes.

3.2. Subspace Coefficient Tuning

Given the re-formulation of multi-head attention maps as a graph convolutional filter subspace, we present an efficient fine-tuning method specifically designed for transformers to enhance attention's expressiveness. Our motivation is that after the softmax function, each graph filter $\mathbf{F}^h(\mathbf{X})$ has values in the range $(0,1)$ with row-sum as one. The normalized attention maps restrict the expressiveness in the way that each output embedding lies in the convex hull of the value embeddings.

To expand the feature space of the output embeddings, we propose *Coeff-Tuning*, to tune the filter subspace on the target dataset with a small set of combination coefficients, i.e., **subspace coefficient** $\boldsymbol{\alpha} \in \mathbb{R}^{H\times H}$. As shown in Figure 2, the tunable coefficient $\boldsymbol{\alpha}$ constructs a more expressive filter subspace $\hat{\mathcal{F}}_\alpha(\mathbf{X})$ by linearly combining the pre-trained multi-head attention maps $\mathbf{F}(\mathbf{X})$. Each element $\hat{\mathbf{F}}^h(\mathbf{X})$ in the learned filter subspace can be written as,

$$\hat{\mathbf{F}}^h(\mathbf{X}) = \boldsymbol{\alpha}[h,:]^\intercal\mathcal{F}(\mathbf{X}) = \sum_{i=1}^H \boldsymbol{\alpha}[h,i]\mathbf{F}^i(\mathbf{X}). \tag{9}$$

The multi-head attention then represents as,

$$\mathbf{O} = \sum_{h=1}^H \hat{\mathbf{F}}^h(\mathbf{X})\mathbf{X}\mathbf{W}^h = \sum_{h=1}^H\sum_{i=1}^H \boldsymbol{\alpha}[h,i]\mathbf{F}^i(\mathbf{X})\mathbf{X}\mathbf{W}^h. \tag{10}$$

Note that, mathematically, the subspace coefficient $\boldsymbol{\alpha}$ is unconstrained, allowing it to combine attention maps with a wider range of values from the original filter subspace $\mathbf{F}(\mathbf{X})$. With attention scores exceeding the range $(0,1)$, we effectively expand the space for $\mathbf{O}$ by allowing a more flexible combination of value matrices.

Residual Parameterization. To facilitate the tuning and preserve the most knowledge in the pre-trained transformer, we propose a residual parameterization of the subspace coefficient,

$$\begin{aligned} \boldsymbol{\alpha}' &= \boldsymbol{\alpha} + \mathbf{I}, \\ \hat{\mathbf{F}}^h(\mathbf{X}) &= \mathbf{F}^h(\mathbf{X}) + \sum_{i=1}^H \boldsymbol{\alpha}[h,i]\mathbf{F}^i(\mathbf{X}). \end{aligned} \tag{11}$$

In this way, we maintain the original filter subspace with the identity matrix, while parameterizing the subspace coefficient $\boldsymbol{\alpha}'$ to learn the residual combination of attention maps on fine-tuning tasks.

Reguralization on Subspace Coefficient. The proposed *Coeff-Tuning* substantially enhances the expressiveness of multi-head attention with tunable coefficient $\boldsymbol{\alpha}$. However, it may also lead to overfitting and preventing channel projection parameters $\mathbf{W}_v, \mathbf{W}_o$ from learning generalizable features. To alleviate this, we propose a simple regularization design that performs dropout [63] directly on the tuned subspace coefficient $\boldsymbol{\alpha}$ during training,

$$\boldsymbol{\alpha}' = \text{Dropout}(\boldsymbol{\alpha}; p) + \mathbf{I}. \tag{12}$$

In this way, we randomly remove an element from $\boldsymbol{\alpha}$ with a probability of p to avoid overfitting.

4. Analysis

In this section, we analyze that adjusting the filter subspace using the proposed subspace coefficient $\boldsymbol{\alpha}$ can effectively enhance the expressiveness of the multi-head attention.

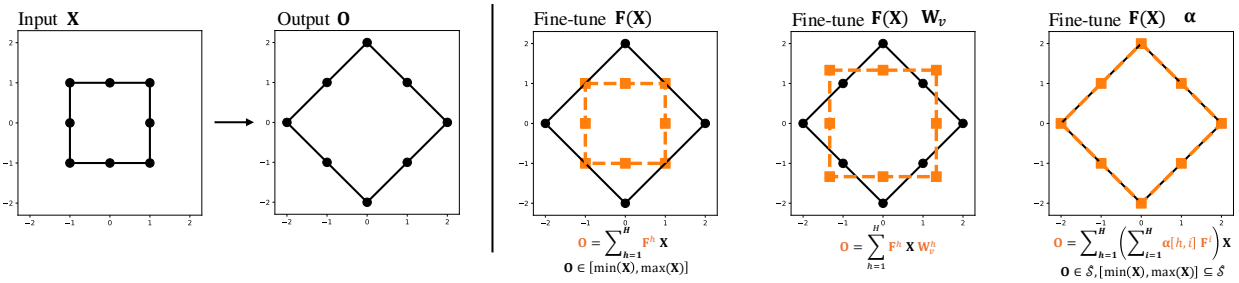


Figure 3. The effectiveness of tuning subspace coefficient $\boldsymbol{\alpha}$. The goal is to convert the input graph $\mathbf{X}$ to the target graph $\mathbf{O}$ using a single attention layer, $\mathbf{O} = \sum_{h=1}^{H} \mathbf{F}^h(\mathbf{X})\mathbf{X}\mathbf{W}_v^h$, where $\mathbf{F}(\mathbf{X}) = \text{softmax}(\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\intercal\mathbf{X}^\intercal)$. We have the following observations: (i) When only tuning $\mathbf{W}_q$ and $\mathbf{W}_k$, *i.e.*, updating $\mathbf{F}(\mathbf{X})$, the positions of output nodes are constrained within the range of the input nodes $\mathbf{X}$, since the values of $\mathbf{F}(\mathbf{X})$ are in the range of $(0, 1)$. (ii) In addition to tuning $\mathbf{F}(\mathbf{X})$, updating $\mathbf{W}_v$ projects the input as $\mathbf{X}\mathbf{W}_v$, but the output graph is still not aligned with the target graph. (iii) In addition to tuning $\mathbf{F}(\mathbf{X})$, adapting $\boldsymbol{\alpha}$ successfully transforms the input graph into the target graph. Compared with (i), $\boldsymbol{\alpha}$ expand the positions of output nodes $\sum_{i=1}^{H} \boldsymbol{\alpha}[:, i]\mathbf{F}^i(\mathbf{X})\mathbf{X}$ beyond the constraints of the input nodes, enhancing the expressiveness of the attention effectively.

For the filter $\mathbf{F}(\mathbf{X})$ in the filter subspace $\mathcal{F}$, since $\mathbf{F}(\mathbf{X}) = \text{softmax}(\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\intercal\mathbf{X}^\intercal)$, their elements $\mathbf{F}(\mathbf{X})[n, n']$ satisfy the following properties,

$$0 < \mathbf{F}(\mathbf{X})[n, n'] < 1, \quad \sum_{n' \in [N]} \mathbf{F}(\mathbf{X})[n, n'] = 1. \tag{13}$$

Definition 4.1. *Given N nodes $\{\mathbf{x}_i\}_{i=1}^N$, the set $\mathcal{S}(\mathbf{x}) = \{\sum_{i=1}^N \lambda_i \mathbf{x}_i \mid \lambda_i > 0, \sum_i \lambda_i = 1\}$ is a bounded convex set.*

According to the Definition 4.1, when the projected nodes $\mathbf{X}\mathbf{W}^h \in \mathbb{R}^{N \times C_o}$ of each head h are given, the output $\mathbf{O}^h[n, :] = \sum_{n' \in [N]} \mathbf{F}^h(\mathbf{X})[n, n']\mathbf{X}[n']\mathbf{W}^h$ lies within $\mathcal{S}(\mathbf{X}\mathbf{W}^h)$, which means the outputs $\mathbf{O}^h[n, :]$ are bounded. This property also applies to the multi-head attention mechanism.

Proposition 4.2. *If $\mathcal{S}^h$ is a bounded convex set, $\forall h = 1, \cdots, H$, the set $\mathcal{S} = \mathcal{S}^1 + \cdots + \mathcal{S}^H = \{\mathbf{x}^1 + \cdots + \mathbf{x}^H \mid \forall \mathbf{x}^h \in \mathcal{S}^h\}$ is also a bounded convex set.*

According to Proposition 4.2, when the projected nodes $\mathbf{X}\mathbf{W}^h$ are given, the outputs of the multi-head attention $\mathbf{O}[n, :] = \sum_{h=1}^H \sum_{n' \in [N]} \mathbf{F}^h(\mathbf{X})[n, n']\mathbf{X}[n']\mathbf{W}^h$ are also bounded.

Proposition 4.3. *After applying the subspace coefficients $\boldsymbol{\alpha}$, the outputs of the reconstructed filter subspace $\hat{\mathcal{F}}_\alpha$ form a set $\hat{\mathcal{S}} = \{\sum_{h=1}^H \sum_{i=1}^H \boldsymbol{\alpha}[h, i]\mathbf{F}^i(\mathbf{X})[n, n']\mathbf{X}\mathbf{W}^h \mid \mathbf{F}^i(\mathbf{X})[n, n'] > 0, \sum_j \mathbf{F}^i(\mathbf{X})[n, n'] = 1, \boldsymbol{\alpha}[h, i] \in \mathbb{R}\}$, which satisfies $\mathcal{S} \subseteq \hat{\mathcal{S}}$.*

According to Proposition 4.3, when applying filter subspace coefficients $\boldsymbol{\alpha}$, we can reconstruct the filter subspace $\mathcal{F}$ and expand the set $\mathcal{S}$. Specifically, given $\mathbf{X}\mathbf{W}^h$, the output $\mathbf{O}[n, :] = \sum_{h=1}^H \sum_{i=1}^H \boldsymbol{\alpha}[h, i]\mathbf{F}^i(\mathbf{X})[n, n']\mathbf{X}\mathbf{W}^h$ can be in the set $\hat{\mathcal{S}} \setminus \mathcal{S}$. Our method obtains an expanded set, enhancing the expressiveness of the model. We provide the support of our analyses in Appendix A.

4.1. Validation Experiment

We assess the effectiveness of the subspace coefficients using a single attention layer. Given an input graph $\mathbf{X}$, the output of the attention layer is calculated as $\mathbf{O} = \mathbf{F}(\mathbf{X})\mathbf{X}\mathbf{W}_v$, where $\mathbf{F}(\mathbf{X}) = \text{softmax}(\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\intercal\mathbf{X}^\intercal)$.

Our goal is to optimize $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ to transform the input graph $\mathbf{X}$ into the target graph $\mathbf{O}$ as shown in Figure 3. The input graph contains 8 nodes $\mathbf{X} = [[-1, 1], [0, 1], [1, 1], [1, 0], [1, -1], [0, -1], [-1, -1], [-1, 0]]$ positioned in a range of $[-1, 1]$ and the target graph $\mathbf{O} = \big[[-1, 1], [0, 2], [1, 1], [2, 0], [1, -1], [0, -2], [-1, -1], [-2, 0]\big]$, with node coordinates in the range of $[-2, 2]$.

As shown in Figure 3, by only tuning the $\mathbf{W}_q$ and $\mathbf{W}_k$, *i.e.*, $\mathbf{F}(\mathbf{X}) = \text{softmax}(\mathbf{X}\mathbf{W}_q\mathbf{W}_k^\intercal\mathbf{X}^\intercal)$, the output is constrained inside the range of the input, which is aligned with Proposition 4.1. When we tune $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$, the output exceeds the input range but still cannot fully match the target graph. However, when we tune the subspace $\mathbf{F}(\mathbf{X})$ and subspace coefficient $\boldsymbol{\alpha}$, the attention layer can successfully transform the input graph into the target graph as in Figure 3. This observation is aligned with Proposition 4.3, and indicates that the proposed *Coeff-Tuning* can effectively expand the feature space of the multi-head attention layer.

4.2. Complexity Analysis

We now formally analyze the complexity of the proposed *Coeff-Tuning*, and compare it with other baseline PEFT methods. The number of parameters required for our method is $H \times H$, where H is the number of heads at each attention layer. Considering $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{C_i \times C_o}$ and $\mathbf{W}_o \in \mathbb{R}^{C_o \times C_o}$ of the attention layer, our method introduces relative $\frac{H^2}{3C_iC_o + C_o^2}$ parameters. In modern large transformer models, we typically have $C_i = 2048$ and $C_o = 1280$. With $H = 10$, our method requires approximately 0.001% additional parameters. Compared with other PEFT methods such as LoRA, our method introduces

Table 1. Fine-tuning Results on VTAB-1k with ViT

	Natural							Specialized				Structured										
	C100	Cal101	DTD	F102	Pets	SVHN	S397	Came	E_SAT	Res45	Retino	Cle_cnt	Cle_dis	DMLab	KITTI	dS_loc	dS_ori	S_azi	S_ele	**Mean**	**Param.**	**GFLOPs**
Full fine-tuning	68.9	87.7	64.3	97.2	86.9	87.4	38.8	79.7	95.7	84.2	73.9	56.3	58.6	41.7	65.5	57.5	46.7	25.7	29.1	65.57	85.8	45.15
Adapter	74.1	86.1	63.2	97.7	87.0	34.6	50.8	76.3	88.0	73.1	70.5	45.7	37.4	31.2	53.2	30.3	25.4	13.8	22.1	55.82	0.234	16.44
BitFit	72.8	87.0	59.2	97.5	85.3	59.9	51.4	78.7	91.6	72.9	69.8	61.5	55.6	32.4	55.9	66.6	40.0	15.7	25.1	62.05	0.105	15.08
VPT	77.7	86.9	62.6	97.5	87.3	74.5	51.2	78.2	92.0	75.6	72.9	50.5	58.6	40.5	67.1	68.7	36.1	20.2	34.1	64.85	0.06	16.68
LoRA	67.1	91.4	69.4	98.8	90.4	85.3	54.0	84.9	95.3	84.4	73.6	82.9	69.2	49.8	81.7	81.8	48.3	32.8	44.2	72.91	0.305	15.74
SPT-Adapter	72.9	93.2	72.5	99.3	91.4	84.6	55.2	85.3	96.0	83.9	75.8	82.2	68.0	49.3	80.0	82.4	51.9	31.7	41.2	73.52	0.266	16.44
Linear Probing	63.4	85.0	63.2	97.0	86.3	36.6	51.0	78.5	87.5	68.6	74.0	34.3	30.6	33.2	55.4	12.5	20.0	9.6	19.2	52.94	0	15.05
Coeff. α Only (Ours)	69.5	90.8	74.5	99.1	90.9	83.0	51.1	83.6	95.3	84.6	75.8	57.4	59.1	46.5	78.6	75.0	51.4	25.6	34.2	**69.78**	0.002	15.06
FACT-TT≤16	71.3	89.6	70.7	98.9	91.0	87.8	54.6	85.2	95.5	83.4	75.7	82.0	69.0	49.8	80.0	79.2	48.4	34.2	41.4	73.04	0.040	15.14
w/ Coeff-Tuning (Ours)	**72.1**	**89.7**	**71.8**	**99.1**	**91.2**	**88.9**	**54.8**	**85.3**	**95.7**	**84.2**	75.3	**83.0**	68.8	**51.0**	**81.3**	**80.0**	**49.3**	**35.0**	**42.7**	**73.64**	0.042	15.15
SNELL-8	73.7	92.7	72.4	99.2	91.4	89.2	55.4	84.9	96.1	86.4	75.2	84.0	68.5	53.5	81.0	82.7	49.9	33.9	39.2	74.17	0.268	15.47
w/ Coeff-Tuning (Ours)	**74.2**	**92.9**	**73.1**	99.2	**91.6**	**90.3**	**55.6**	**85.2**	**96.2**	86.3	**75.8**	**84.1**	**68.8**	**53.7**	**81.3**	**84.0**	**52.1**	**35.0**	**39.9**	**74.70**	0.270	15.48

relative $\frac{H^2}{3rC_i+5rC_o}$ parameters, where r is the LoRA rank. Using the same values for C_i, C_o, H, and setting $r = 4$, our approach requires less than 0.2% additional parameters compared to LoRA and its variants. This also indicates that our *Coeff-Tuning* can maintain almost the same quantity of parameters when combined with other PEFT methods.

5. Experiments

In this section, we validate the proposed *Coeff-Tuning* with various tasks and datasets, including tuning the vision transformer [11] on few-shot image classification, tuning text-to-image diffusion model [51] for concept customization, and tuning vision-language BART [8] on image-to-text understanding. Combining with other weight-based PEFT methods, the proposed *Coeff-Tuning* consistently achieves superior results with neglectable parameters added.

5.1. Few-shot Image Classification

we first validate the proposed method by fine-tuning Vision Transformers on a series of downstream datasets for few-shot image classification. Our method achieves superior results while maintaining a high parameter efficiency compared with baseline methods.

Experimental Settings. We utilize the VTAB-1k [81] dataset which comprises 19 diverse computer vision tasks across three domains: standard natural imagery, specialized capture, and synthetic/structured environments. The benchmark evaluates model performance on various task objectives. With only 1,000 training samples per task, it presents a significant few-shot learning challenge. We also benchmark our method on five fine-grained datasets in few-shot settings, including CUB-200-2011 [69], NABirds [67], Oxford Flowers [44], Stanford Dogs [25] and Stanford Cars [13]. A detailed description is provided in Appendix B.2.

Table 2. Fine-tuning Results on the FGVC with ViT.

	CUB-200	NABirds	Flowers	S-Dogs	S-Cars	Mean
Full fine-tuning	87.3	82.7	98.8	89.4	84.5	88.5
Linear Probing	85.3	75.9	97.9	86.2	51.3	79.3
Adapter	87.3	84.3	98.4	88.8	68.4	85.5
BitFit	88.4	84.2	98.8	91.2	79.4	88.4
VPT	86.7	78.8	98.4	90.7	68.7	84.6
LoRA	84.9	79.0	98.1	88.1	79.8	86.0
SPT-LoRA	88.6	83.4	99.5	91.4	87.3	90.1
SNELL-8	89.6	86.8	99.3	92.1	89.9	91.5
w/ Coeff-Tuning (Ours)	**89.8**	**87.1**	**99.5**	**92.2**	**90.7**	**91.9**

As for the vision transformer, we follow VPT [23], SSF [35], and select ViT-B/16 [11] pre-trained on ImageNet-21K [9] as the base model for fine-tuning. We benchmark the proposed *Coeff-Tuning* with a series of PEFT methods for few-shot learning including SSF [35], VPT [23], Adapter [20], LoRA [21], and BitFit [79].

We present two settings of *Coeff-Tuning* in this task. We first explore the representation ability of **tuning subspace coefficient α only**, which tunes the transformer bottleneck with only tuning H^2 (**144** for ViT [11]) parameters for each transformer block. We also show the performance of combining *Coeff-Tuning* and SSF [35] to demonstrate that our method can plug and play with other PEFT methods. We set the coefficient dropout rates as $p = 0$ and $p = 0.1$ for these two settings respectively, and provide other hyperparameters in Appendix B.2.

Results. As shown in Table 1, the proposed *Coeff-Tuning* shows superior performances in both settings. Compared with the linear probing baseline which achieves a mean accuracy of 61.69, *Coeff-Tuning* obtains a significantly higher result of **73.49** with tuning only $(\mathbf{12} * \mathbf{12}) * \mathbf{12} = \mathbf{1,728}$ additional subspace coefficients $\boldsymbol{\alpha}$ in ViT [11] (ViT [11] has 12 attention layers, each attention has 12 heads). Thus, we achieve a large performance increase in few-shot image classification (**11.8**↑) with neglectable additional parameter costs. Note that this result has already surpassed almost all baselines. We further achieve the best performance by combining *Coeff-Tuning* with SSF while maintaining almost the same parameter cost.



Figure 4. Results on Concept Customization. Compared to the baseline method (middle row), *Coeff-Tuning* (bottom row) leads to higher concept fidelity and text alignment of the generated images. For instance, our method accurately generates images of a "$\langle V \rangle$ plushie sloth in the snow," while the baseline fails to align with the text prompt. Additionally, our method generates a "$\langle V \rangle$ wooden pot" with higher fidelity to the input image.

Table 3. Concept Customization Results with SDXL.

Method	# Params (%)	Text Align.	Concept Align.	Diversity
DoRA	0.0371	0.33	0.73	10.9
Ours + DoRA	0.0374	**0.34**	**0.75**	**11.37**
LoRA	0.0306	0.33	0.74	10.08
Ours + LoRA	0.0309	**0.34**	**0.75**	**10.45**

5.2. Personalized Text-to-Image Generation

We now show the effectiveness of *Coeff-Tuning* with the concept customization with Stable Diffusion (SDXL) [51, 55]. We validate the effectiveness of *Coeff-Tuning* by combining it with baseline PEFT methods, LoRA [21] and DoRA [38], and compare the performance with baselines without tuning $\boldsymbol{\alpha}$. Specifically, we fine-tune SDXL [51] to associate the special token with a concept image from Dreambooth [56] and CustomConcept101 [30] datasets. We choose 10 different concepts, and fine-tune the model on $4 \sim 9$ images. We utilize 25 different text prompts following [53] and generate 4 images for each prompt. We evaluate the fidelity, diversity, and text-to-image alignment of the generated images. Other training details are described in Appendix B.3.

Evaluation metrics. We assess the concept alignment with the average cosine similarity between CLIP embeddings [54] of the generated images and the concept images. The diversity of generated images is measured by the Vendi score [12] calculated with the DINOv2 embeddings [46]. The text alignment of the generated images is measured via average cosine similarity in the CLIP feature space [54].

Results. The quantitative results are shown in Table 3. By combining with LoRA [21] or DoRA [38], the proposed *Coeff-Tuning* generates images with greater diversity and better alignment with the both text prompt and input concept images. The qualitative results of our method compared with LoRA [21] are shown in Figure 4. The generated images with *Coeff-Tuning* have higher concept fidelity, preserving more characteristics from the input images.

As an example, consider the concept "$\langle V \rangle$ cat" as shown in the 4th column in Figure 4, the image generated using our method retains the cat's color from the input image, while the one generated by the baseline method alter the color. Additionally, the backgrounds in images generated with our method show more detailed features compared to those from the baseline method. *Coeff-Tuning* achieves the above improvements by introducing a limited number of additional parameters. We provide more generated images in Appendix B.3.

5.3. Image-Text Understanding

As we have shown the effectiveness of *Coeff-Tuning* in fine-tuning models on both few-shot image classification and personalized image generation, we now validate it on multi-task image-text understanding, which is a much larger-scale multi-modal fine-tuning task that requires a higher model capacity. By combining with baseline PEFT meth-

Table 4. Multi-Model Fine-tuning Results with VL-Bart.

Method	# Params (%)	VQA^{v2}	GQA	NVLR^2	COCO Cap	Avg.
Full Model	100	66.9	56.7	73.7	112.0	77.3
LoRA	5.94	65.2	53.6	71.9	**115.3**	76.5
Ours + LoRA	5.94	**66.2**	**54.1**	**73.0**	115.2	**77.1**
DoRA	5.96	65.8	54.7	**73.1**	115.9	77.4
Ours + DoRA	5.96	**66.4**	**54.9**	73.0	**116.4**	**77.7**

ods, our method again achieves superior performance with neglectable parameter increase.

Experimental Settings. We evaluate our method on VL-BART [8] multi-modal transformer, which utilizes CLIP-ResNet101 [54] for vision encoding and sends the visual features to BART-Base [31] for image-text joint modeling. Following [65], we evaluate our method with the multi-task fine-tuning of VL-Bart [8] on four large-scale image-text understanding tasks simultaneously: VQAv2 [14] and GQA [22] for visual question answering, NLVR2 [64] for visual reasoning, and MSCOCO [5] for image captioning. Following the plug-and-play practice, we combine the proposed *Coeff-Tuning* with LoRA [21] and DoRA [38] respectively, and set the subspace coefficient dropout rate to $p = 0.2$. Detailed descriptions of other settings and hyper-parameters are provided in Appendix B.4.

Results. As shown in Table 4, our method surpasses baseline LoRA [21] and DoRA [38] respectively in almost all image-text datasets consistently, and achieves higher average scores. The *Coeff-Tuning* combined with LoRA [21] obtains comparable results with full-model fine-tuning, and it achieves better results than full-model tuning when it is combined with DoRA [38]. Notably, these improvements are achieved while maintaining the same parameter budgets as we only tune a hundred parameters for each attention.

5.4. Ablation Study

In this section, we examine the effectiveness of each component of our design with ablation experiments.

Table 5. Ablation studies on (a) residual parameterization and (b) coefficient dropout showing their effects on model performance.

Parametrization	Avg.
$\boldsymbol{\alpha}$ w/ rand. init.	44.28
$\boldsymbol{\alpha}$ w/ identity init.	61.57
$I + \boldsymbol{\alpha}$, zero init.	**73.49**

(a)

Dropout p	Avg. Score
0.0	75.1
0.1	77.3
0.2	**77.7**
0.4	72.3

(b)

Residual Parameterization. We first demonstrate the importance of residual parameterization in Eq. 11 in *Coeff-Tuning*. We compare our residual parameterization with the direct parameterization where we directly tune $\boldsymbol{\alpha}$ without adding it to the identity matrix, and initialize it with either a random matrix or an identity matrix. We adopt the few-shot tuning task as in Sec. 5.1 and only tune $\boldsymbol{\alpha}$. Results in Table 5a show that models with residual parameterization achieve superior performance than direct parameterization with two different initializations. This validates our design that maintaining the original filter subspace while learning the residual combination helps preserve pre-trained knowledge and leads to better fine-tuning performance.

Subspace Coefficient Dropout. The dropout regularization on the subspace coefficient as in Eq. 12 plays a crucial role in preventing overfitting. We adopt the experiment setting in multi-modal fine-tuning as in Sec. 5.3. We experiment with different dropout rates $p \in \{0, 0.1, 0.2, 0.4\}$ during training. The results in Table 5b demonstrate that a moderate dropout rate ($p = 0.2$) yields the best performance while either a lower or higher dropout rate leads to sub-optimal results. This suggests that applying dropout directly on the tuned subspace coefficient $\boldsymbol{\alpha}$ can effectively prevent overfitting and lead to better generalization.

6. Conclusion

In this paper, we presented *Coeff-Tuning*, a new parameter-efficient fine-tuning approach that enhances multi-head attention's expressiveness by learning a set of subspace coefficients. By reformulating attention as a graph convolution with a filter subspace, we demonstrated both theoretically and empirically that our method expands the feature space of multi-head attention beyond the original convex hull constraints. The proposed *Coeff-Tuning* can be seamlessly integrated with existing PEFT methods while introducing negligible additional parameters. Through extensive experiments across diverse tasks and modalities, we consistently demonstrated superior performance compared to competitive PEFT baselines. Our method achieved significant improvements in various metrics while maintaining high parameter efficiency, suggesting a promising direction for enhancing the adaptability of large pre-trained transformers.

References

[1] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[2] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 2022. 2

[3] Wei Chen, Zichen Miao, and Qiang Qiu. Inner product-based neural network similarity. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[4] Wei Chen, Zichen Miao, and Qiang Qiu. Large convolutional model tuning via filter subspace. *arXiv preprint arXiv:2403.00269*, 2024. 2, 1

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 8

[6] Xiuyuan Cheng, Zichen Miao, and Qiang Qiu. Graph convolution with low-rank learnable local filters. *arXiv preprint arXiv:2008.01818*, 2020. 4

[7] Xiuyuan Cheng, Zichen Miao, and Qiang Qiu. Graph convolution with low-rank learnable local filters. In *International Conference on Learning Representations*, 2021. 1

[8] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 2, 6, 8

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6, 1

[10] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021. 2, 3

[11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1, 2, 6

[12] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022. 7

[13] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 6

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2, 8

[15] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023. 1

[16] Liu He and Daniel Aliaga. Globalmapper: Arbitrary-shaped urban layout generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 454–464, 2023. 1

[17] Liu He and Daniel Aliaga. Coho: Context-sensitive city-scale hierarchical urban layout generation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[18] Liu He, Yijuan Lu, John Corring, Dinei Florencio, and Cha Zhang. Diffusion-based document layout generation. In *International Conference on Document Analysis and Recognition*, pages 361–378. Springer, 2023.

[19] Liu He, Yizhi Song, Hejun Huang, Daniel Aliaga, and Xin Zhou. Kubrick: Multimodal agent collaborations for synthetic video generation. *arXiv preprint arXiv:2408.10453*, 2024. 1

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 2, 6

[21] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 1, 2, 6, 7, 8

[22] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2, 8

[23] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2, 6, 1

[24] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 2021. 2, 1

[25] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, 2011. 6

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 1

[27] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[29] Dawid J Kopiczko, Tijmen Blankevoort, and Yuki M Asano. Vera: Vector-based random matrix adaptation. *ICLR*, 2024. 2

[30] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 7, 1

[31] M Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 2, 8

[32] Junbo Li, Zichen Miao, Qiang Qiu, and Ruqi Zhang. Training bayesian neural networks with sparse subspace variational inference. *arXiv preprint arXiv:2402.11025*, 2024. 1

[33] Tianchun Li, Chengxiang Wu, Pengyi Shi, and Xiaoqian Wang. Cumulative difference learning vae for time-series with temporally correlated inflow-outflow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13619–13627, 2024. 1

[34] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the Association for Computational Linguistics*, 2021. 2

[35] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35:109–123, 2022. 2, 6, 1

[36] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022. 2

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[38] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 2, 7, 8, 1

[39] Tianci Liu, Ruirui Li, Yunzhe Qi, Hui Liu, Xianfeng Tang, Tianqi Zheng, Qingyu Yin, Monica Xiao Cheng, Jun Huan, Haoyu Wang, et al. Unlocking efficient, scalable, and continual knowledge editing with basis-level representation fine-tuning. *arXiv preprint arXiv:2503.00306*, 2025. 1

[40] Zichen Miao, Ze Wang, Wei Chen, and Qiang Qiu. Continual learning with filter atom swapping. In *International Conference on Learning Representations*, 2021. 1

[41] Zichen Miao, Ze Wang, Xiuyuan Cheng, and Qiang Qiu. Spatiotemporal joint filter decomposition in 3d convolutional neural networks. *Advances in Neural Information Processing Systems*, 34:3376–3388, 2021. 1

[42] Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024. 1

[43] Zichen Miao, Zhengyuan Yang, Kevin Lin, Ze Wang, Zicheng Liu, Lijuan Wang, and Qiang Qiu. Tuning timestep-distilled diffusion model using pairwise sample optimization. *arXiv preprint arXiv:2410.03190*, 2024. 1

[44] 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. 6

[45] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022. 3

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[47] Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. *arXiv preprint arXiv:2503.06339*, 2025. 1

[48] Gaurav Patel, Konda Reddy Mopuri, and Qiang Qiu. Learning to retain while acquiring: Combating distribution-shift in adversarial data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7786–7794, 2023.

[49] Gaurav Patel, Christopher Sandino, Behrooz Mahasseni, Ellen L Zippi, Erdrin Azemi, Ali Moin, and Juri Minxha. Efficient source-free time-series adaptation via parameter subspace disentanglement. *arXiv preprint arXiv:2410.02147*, 2024. 1

[50] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020. 2

[51] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ICLR*, 2024. 1, 6, 7

[52] Qiang Qiu, Xiuyuan Cheng, Guillermo Sapiro, et al. Dcfnet: Deep neural network with decomposed convolutional filters. In *International Conference on Machine Learning*, pages 4198–4207. PMLR, 2018. 1

[53] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023. 2, 7

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 2, 7, 8

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 7

[56] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 7, 1

[57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015. 2

[58] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[59] Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen Lee, and James T Kwok. Revisiting over-smoothing in bert from the perspective of graph. *arXiv preprint arXiv:2202.08625*, 2022. 2, 3

[60] Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania, Han Zhang, Yuan Hao, Irfan Essa, and Lu Jiang. Visual prompt tuning for generative transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[61] Yizhi Song, Liu He, Zhifei Zhang, Soo Ye Kim, He Zhang, Wei Xiong, Zhe Lin, Brian Price, Scott Cohen, Jianming Zhang, et al. Refine-by-align: Reference-guided artifacts refinement through semantic alignment. *arXiv preprint arXiv:2412.00306*, 2024. 1

[62] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8048–8058, 2024. 1

[63] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 4

[64] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 8

[65] 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. 8

[66] Gemma Canet Tarrés, Zhe Lin, Zhifei Zhang, Jianming Zhang, Yizhi Song, Dan Ruta, Andrew Gilbert, John Collomosse, and Soo Ye Kim. Thinking outside the bbox: Unconstrained generative object compositing. *arXiv preprint arXiv:2409.04559*, 2024. 1

[67] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015. 6

[68] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1, 2

[69] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[70] Haoyu Wang, Tianci Liu, Ruirui Li, Monica Cheng, Tuo Zhao, and Jing Gao. Roselora: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. *arXiv preprint arXiv:2406.10777*, 2024. 1

[71] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*, 2022. 2

[72] Ze Wang, Xiuyuan Cheng, Guillermo Sapiro, and Qiang Qiu. Stochastic conditional generative networks with basis decomposition. *arXiv preprint arXiv:1909.11286*, 2019. 1

[73] Ze Wang, Seunghyun Hwang, Zichen Miao, and Qiang Qiu. Image generation using continuous filter atoms. *Advances in Neural Information Processing Systems*, 34:17826–17838, 2021.

[74] Ze Wang, Zichen Miao, Jun Hu, and Qiang Qiu. Adaptive convolutions with per-pixel dynamic filter atom. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2021. 1

[75] Ze Wang, Zichen Miao, Xiantong Zhen, and Qiang Qiu. Learning to learn dense gaussian processes for few-shot learning. *Advances in Neural Information Processing Systems*, 34:13230–13241, 2021. 1

[76] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239, 2023. 1

[77] Zhexiao Xiong, Wei Xiong, Jing Shi, He Zhang, Yizhi Song, and Nathan Jacobs. Groundingbooth: Grounding text-to-image customization. *arXiv preprint arXiv:2409.08520*, 2024. 1

[78] SHIH-YING YEH, Yu-Guan Hsieh, Zhidong Gao, Bernard BW Yang, Giyeong Oh, and Yanmin Gong. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In *International Conference on Learning Representations*, 2023. 2

[79] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. 2, 6

[80] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023. 2, 3

[81] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning

with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 6, 1

[82] 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. *ICLR*, 2022. 2

[83] Yibo Zhong, Haoxiang Jiang, Lincan Li, Ryumei Nakada, Tianci Liu, Linjun Zhang, Huaxiu Yao, and Haoyu Wang. Neat: Nonlinear parameter-efficient adaptation of pre-trained models. *arXiv preprint arXiv:2410.01870*, 2024. 1

[84] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021. 2, 3