MP-ISMoE: Mixed-Precision Interactive Side Mixture-of-Experts for Efficient Transfer Learning

Yutong Zhang[1,2], **Zimeng Wu**[1,2], **Shengcai Liao**[3], **Shujiang Wu**[2], **Jiaxin Chen**[1,2,*]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]College of Information Technology, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

Abstract

Parameter-efficient transfer learning (PETL) has emerged as a pivotal paradigm for adapting pre-trained foundation models to downstream tasks, significantly reducing trainable parameters yet suffering from substantial memory overhead caused by gradient backpropagation during fine-tuning. While memory-efficient transfer learning (METL) circumvents this challenge by bypassing backbone gradient computation via lightweight small side networks, its stringent memory constraint severely limits learning capacity of side networks, thereby significantly compromising performance. To address these limitations, we propose a novel Mixed-Precision Interactive Side Mixture-of-Experts framework (MP-ISMoE). Specifically, we first propose a Gaussian Noise Perturbed Iterative Quantization (GNP-IQ) scheme to quantize weights into lower-bits while effectively decreasing quantization errors. By leveraging memory conserved from GNP-IQ, we subsequently employ Interactive Side Mixture-of-Experts (ISMoE) to scaling up side networks without sacrificing overall memory efficiency. Different from conventional mixture-of-experts, ISMoE learns to select optimal experts by interacting with salient features from frozen backbones, thus suppressing knowledge forgetting and boosting performance. Extensive experiments across diverse vision-language and language-only tasks demonstrate that MP-ISMoE remarkably promotes accuracy compared to state-of-the-art METL approaches, while maintaining comparable parameter and memory efficiency.

Code & Extended version —
https://github.com/Zhang-VKk/MP-ISMoE.git

Introduction

Large-scale foundation models, which are typically pre-trained on massive datasets, have demonstrated remarkable representation and generalization abilities across a wide range of domains, including computer vision (Fang et al. 2023), natural language processing (Touvron et al. 2023), and multi-modal tasks (Wu, Huang, and Wei 2024a). Transfer learning effectively unleashes the potential of these models, facilitating development of numerous task-specific models. However, with the ever-expanding model scale, the fully fine-tuning paradigm (Lv et al. 2023; Yin et al. 2025) becomes prohibitively resource-intensive.

Parameter-Efficient Transfer Learning (PETL) (Houlsby et al. 2019; Li and Liang 2021; Hu et al. 2022) has become a promising solution in balancing training cost and model capacity, by freezing most parameters of backbones and fine-tuning lightweight modules such as partial parameters (Zaken, Ravfogel, and Goldberg 2022), prompt vectors (Li and Liang 2021) or adapter networks (Chen et al. 2022). However, they still require backpropagation through large backbones, leading to excessive memory consumption that is disproportionate to the reduction in trainable parameters.

More recently, Memory-Efficient Transfer Learning (METL) (Zhang et al. 2020; Sung, Cho, and Bansal 2022; Liu, An, and Qiu 2024; Mercea et al. 2024) has emerged to achieve consistent reduction in both trainable parameters and memory overhead. Typically, METL introduces a lightweight trainable side network parallel to the frozen backbone, connecting paired features via ladder modules. It primarily turns gradient backpropagation to tiny side networks and ladders, thus improving training efficiency. However, existing METL methods predominantly focus on improving the efficiency and representational capacity of side networks, suffering from the following issues. (1) *Sub-optimal allocation of memory budget.* Given the dominant memory footprint of backbones, existing methods often allocate a small portion of memory to the side network, along with a stringent constraint on the amount of parameters. It inherently restrains the learning capacity of the side network, thus substantially limiting ultimate performance. (2) *Rigid and simplistic side network structure.* The side network is typically down-scaled proportionally from the backbone, inheriting its dense activation feature. Such fixed design leads to a sub-optimal trade-off between memory efficiency and model capacity, hampering effective transfer learning under tight memory budgets. (3) *Insufficient exploration of guidance from backbones.* Most existing METL methods leverage features from backbones to fine-tune side networks by directly combining it with those from side networks through weighted summation. They fail to elaborately explore the complementary knowledge from backbone to suppress over-fitting and knowledge forgetting, thus leaving much room for improvement.

To address above limitations, we propose a novel METL

*Corresponding Author

method, dubbed Mixed-Precision Interactive Side Mixture-of-Experts (**MP-ISMoE**). MP-ISMoE aims at enabling more effective memory allocation between backbone and side network, while facilitating efficient side network expansion during fine-tuning. To this end, we first introduce Gaussian Noise Perturbed Iterative Quantization (**GNP-IQ**), which applies iterative weight quantization with Gaussian noise perturbation to the backbone network, reducing massive memory footprint while maintaining model performance. Leveraging saved memory space by GNP-IQ, we design the Interactive Side Mixture-of-Experts (**ISMoE**) to expand the capacity of side network by employing a sparse MoE structure, without violating the overall memory budget. ISMoE further explores class tokens from backbone as guidance of general knowledge, and interactively apply them to adjust expert selection based on their correlations, thereby mitigating knowledge forgetting and over-fitting.

In summary, our main contributions lie in three-fold:

- We propose a novel mixed-precision fine-tuning framework with an MoE-based side network, dubbed as MP-ISMoE, for memory efficient transfer learning.
- We propose a Gaussian Noise Perturbed Iterative Quantization (GNP-IQ) process and an Interactive Side Mixture-of-Experts (ISMoE) structure to optimize memory-constrained resource allocation, expand the capacity of side network and mitigate knowledge forgetting, thereby enhancing transfer learning performance.
- We conduct extensive experiments on both vision-language and natural language processing tasks across multiple network architectures, demonstrating that MP-ISMoE remarkably outperforms the state-of-the-art METL methods, with comparable parameter and memory efficiency.

Related Work

Parameter-Efficient Transfer Learning

Current PETL methods can be categorized into three paradigms: (1) *Partial Tuning* updates only a subset of original parameters, such bias (Zaken, Ravfogel, and Goldberg 2022), weights (Touvron et al. 2022), and task-specific ones (Sung, Nair, and Raffel 2021), while keeping the rest frozen. (2) *Prompt Tuning* embeds sparse manual-tuning (Zhang et al. 2023b) or dense randomly-initialized (Zhou et al. 2022) tokens into input or intermediate state of the backbone. (3) *Adapter Tuning* introduces lightweight learnable modules into frozen backbones, some methods (Hu et al. 2022; Liu et al. 2022; Mao et al. 2025) employ scaling/shifting factors, learnable vectors, or compact MLPs to refine feature projections, while others (Jie and Deng 2023; Zhang et al. 2023a) minimize trainable parameters via matrix decomposition and hyper-network prediction.

Memory-Efficient Transfer Learning

The METL endeavors to achieve the optimal performance-memory balance in resource-constrained scenarios. (1) *General Memory Optimization*. Mixed-precision training (Micikevicius et al. 2017) or quantization strategies (Wang et al. 2018b) introduce low bitwidth formats for weights, activations and gradients. Gradient checkpoint (Chen et al. 2016) selectively store critical intermediates, reconstructing (Gomez et al. 2017) discarded activations during backpropagation. (2) *Backpropagation Decoupling*. Orthogonally, another direction isolates gradient computation for extensive parameters of large models. Some manners (Raffel et al. 2020) update an extra projection layer, which follows the last backbone layer. While other methods (Zhang et al. 2020; Sung, Cho, and Bansal 2022; Diao et al. 2024b,a) introduce a parallel lightweight network to augment the static main network for new domains.

Weight-only Post-Training Quantization

The Weight-only Post-Training Quantization (Frantar et al. 2022; Kim et al. 2023; Lee et al. 2023; Lin et al. 2024) proves effective in accelerating the memory-bounded General Matrix-Vector Multiply (GEMV) operators while aims to convert weights from high-precision to low-precision with fewer bits, thus reducing the size of model and speeding up weight loading.

Mixture-of-Experts

Mixture-of-Experts (MoE) (Cai et al. 2025; Mu and Lin 2025) divides a model into specialized components (*i.e.*, Experts), each of which handles distinct tasks or data aspects, and combines the router (Liu et al. 2024; Harvey, Weale, and Yilmaz 2025) to selectively activate relevant experts, thereby leveraging a vast amount of expertise by increasing model capacity while maintaining computational efficiency. Typically, MoE can be categorized into two variants: *Dense MoE* (Pan et al. 2024; Wu, Huang, and Wei 2024b) activates all experts in each iteration, while *Sparse MoE* (Dai et al. 2024; Lieber et al. 2024; Wei et al. 2024) activates only some experts and thus generally has lower computational overhead.

Methodology

Framework Overview

Existing METL approaches based on side network typically impose strict constraints on the scale of trainable parameters, aiming to ensure low memory overhead during training. However, such limitation tend to compromise the representation capacity of model, thus leading to sub-optimal performance on downstream tasks. Therefore, we propose a novel framework, Mixed-Precision Interactive Side Mixture-of-Experts (MP-ISMoE), as depicted in Figure 1.

Specifically, Gaussian Noise Perturbed Iterative Quantization (GNP-IQ) module saves memory consumption of backbone weights by strategically reducing their numerical precision. To mitigate the increasing quantization error accumulated during fine-tuning, GNP-IQ employs an iterative re-quantization with injected Gaussian noise perturbation. Meanwhile, the Interactive Side Mixture-of-Experts (ISMoE) module improves the scalability of the side branch by introducing a memory-efficient MoE-based structure. To address the catastrophic knowledge forgetting issue, a cross-network representative token interaction mechanism is performed.

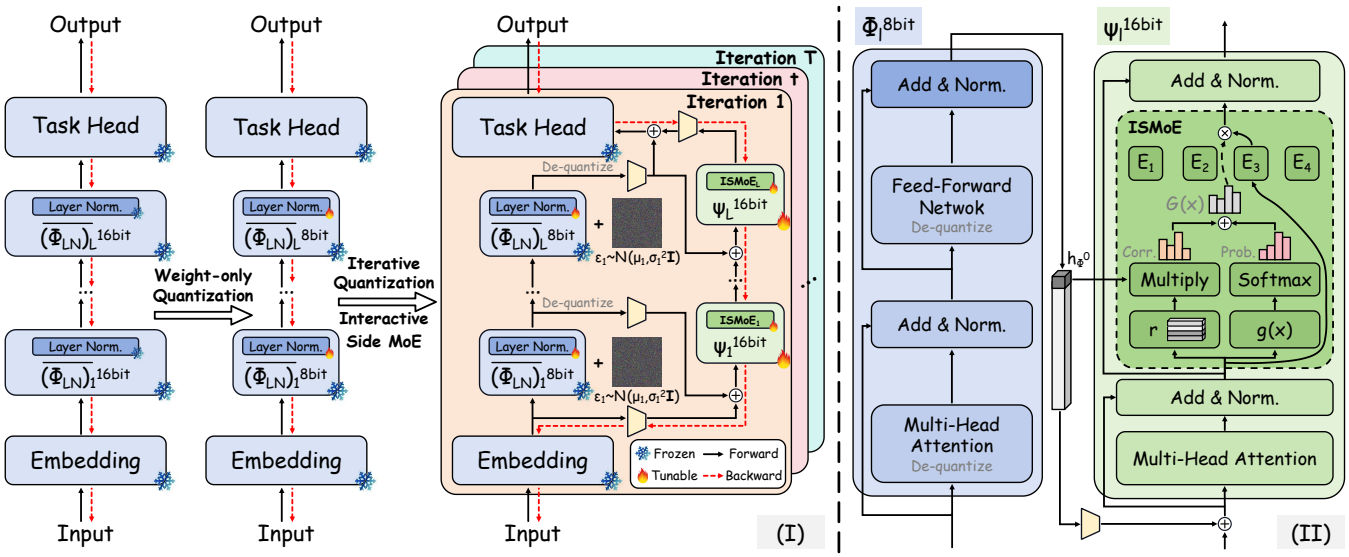

Figure 1: Diagram on the proposed framework. **Part (I)** presents Mixed-Precision Interactive Side Mixture-of-Experts for universal METL based on side networks. We first quantize most of weights in pre-trained backbone before fine-tuning, while preserving trainable Layer Normalization with full-precision. Subsequently, we introduce a learnable parallel Interactive Side Mixture-of-Experts (ISMoE) structure, connected to backbone via downsampling module. During fine-tuning phase, Gaussian Noise Perturbed Iterative Quantization (GNP-IQ) strategy introduces Gaussian noise perturbations into weights to mitigate quantization error. **Part (II)** illustrates detailed structure of l^{th} layer of MP-ISMoE. We adopt MoE to scale up side network, design representative feature for each expert to measure correlation with backbone, and select Topk experts based on the routing Probability and Correlation Score.

As for the overall fine-tuning process, the majority of the backbone parameters are frozen in 8-bit precision, while the Layer Normalization parameters remain trainable in full precision, and the side network is trained in 16-bit. In summary, our MP-ISMoE framework enables efficient mixed-precision fine-tuning, thereby achieving a superior trade-off between memory efficiency and downstream performance.

Gaussian Noise Perturbed Iterative Quantization

Quantization of Backbone. In context of a Transformer-based backbone network, we denote the set of parameters belonging to all Layer Normalization layers as Φ_{LN}, and the rest of the parameters as $\overline{\Phi_{\text{LN}}}$. Typically for memory efficient fine-tuning, Φ_{LN} and $\overline{\Phi_{\text{LN}}}$ are kept frozen. To further reduce the memory footprint of the backbone network, we perform weight-only quantization on $\overline{\Phi_{\text{LN}}}$.

Specifically, prior to fine-tuning, an initial asymmetric post-training quantization is applied, converting the full-precision weights into compact low-bit ones, formally as:

$$w_q = \text{clamp}(\lfloor \frac{w_f}{s} \rfloor + z; 0; 2^n - 1), \tag{1}$$

where $w_f \in \overline{\Phi_{\text{LN}}}$ denotes the original floating-point weight, and w_q is its quantized fixed-point counterpart. Here n is the bitwidth of the quantized value, and $\text{clamp}(\cdot; a; b)$ denotes truncating the value to the interval $[a, b]$. The scale factor s and zero-point z are calculated as:

$$\begin{aligned} s &= \frac{r_{\max} - r_{\min}}{q_{\max} - q_{\min}} = \frac{r_{\max} - r_{\min}}{2^n - 1}, \\ z &= \text{clamp}(\lfloor q_{\max} - \frac{r_{\max}}{s} \rfloor; 0; 2^n - 1), \end{aligned} \tag{2}$$

where $r_{\min/\max}$ and $q_{\min/\max}$ represent the numerical range of w_f and w_q, respectively. While w_q is stored in low-bit format, it is dequantized back to full-precision during forward pass computation by reversing Eq. (2) as:

$$w_d = s \cdot (w_q - z). \tag{3}$$

Iterative Quantization Strategy. Although most of the backbone parameters are frozen during fine-tuning, the aforementioned initially assigned quantization coefficients cannot remain fixed throughout the entire process. Noting that a small portion of quantized weights are updated during fine-tuning, the originally determined s and z may no longer be optimal for the evolving weight distribution. This mismatch leads to an increased quantization error Error_q, which can be expressed as:

$$\text{Error}_q = \frac{1}{U} \sum_{u=1}^{U} \left(w_f^{(u)} - w_d^{(u)}\right)^2, \tag{4}$$

where $U = |\overline{\Phi_{\text{LN}}}|$ denotes the number of quantized weights.

To mitigate this error, we introduce an iterative re-quantization mechanism for the backbone. To be precise, each time we randomly sample a small fraction $p\%$ of

weights from $\overline{\Phi_{\text{LN}}}$ and re-compute their scale and zero-point coefficients (*i.e.* s and z), thereby refining the quantization to better fit the evolving state of the model. To save the additional cost for quantization, this step is performed iteratively at a fixed interval of M epochs. Overall, this procedure is conducted $T = N_{\text{epoch}}/M$ times during fine-tuning, where N_{epoch} is the total number of fine-tuning epochs.

Gaussian Noise Perturbation. Furthermore, to bridge the gap between infrequent quantization steps and the continue shift in model dynamics, we inject a Gaussian noise perturbation into weights prior to each re-quantization, formally as Eq. (5).

$$w'_f = w_f + \epsilon_t, \text{ when } n_{\text{epoch}} = t \times M. \tag{5}$$

Here, w'_f denotes the perturbed weight for further quantization, and n_{epoch} is the index of current epoch. $\epsilon_t \sim \mathcal{N}(\mu_t, \sigma_t^2\mathbf{I}), t \in 1, 2, \cdots, T$ is a learnable perturbation draw from a Gaussian distribution. By optimizing the mean μ_t and standard variance σ_t, the perturbation serves to simulate the cumulative effect of latent parameter updates between re-quantizations, allowing the backbone to better anticipate its optimal quantized state during fine-tuning.

Interactive Side Mixture-of-Experts

Sparse MoE Based Side Network. With GNP-IQ scheme enabling a memory efficient backbone, we allocate the saved resources to expand the side network, thus enhancing its representation capability, while maintaining the overall memory budget. However, a direct dense expansion on the width of network falls short of achieving an optimal performance-efficiency balance, owing to the substantial increase in trainable parameters and memory consumption. Therefore, we adopt a sparse MoE in side network, enabling substantial capacity expansion under limiter memory constraints.

Concretely, ISMoE constructs a set of N distinct experts $\{E_i\}_{i=1}^N$ by replicating the original FFN blocks from the side network. A sparse gating mechanism is then introduced to dynamically select the top-k most relevant experts for each input token. To be precise, a linear projection function $g(\cdot) \in \mathbb{R}^N$ first computes raw gating scores over experts. These scores are then sparsified using a masked top-k selection operator $\mathcal{M}_k(\cdot)$, which retains only the top-k values and set the rest to $-\infty$. The resulting sparse scores are normalized via a Softmax function to produce the final routing probability $G(\cdot)$. Let $\boldsymbol{x}^{in}$ and $\boldsymbol{x}^{out}$ denote the input and output feature of the ISMoE module, respectively, the final output is computed as the probability weighted sum of the selected expert outputs. The complete forward pass computation is formally defined as Eq. (6).

$$\begin{aligned} G(\boldsymbol{x}) &= \text{Softmax}(\mathcal{M}_k(g(\boldsymbol{x}))), \\ \mathcal{M}_k(\boldsymbol{x})_i &= \begin{cases} \boldsymbol{x}_i, & \text{if } \boldsymbol{x}_i \in \text{top-}k(\boldsymbol{x})). \\ -\infty, & \text{otherwise,} \end{cases} \\ \boldsymbol{x}^{out} &= \sum_{i=1}^{N} G_i(\boldsymbol{x}^{in}) \cdot E_i(\boldsymbol{x}^{in}). \end{aligned} \tag{6}$$

Cross-Network Interaction Guided Expert Selection. Although the MoE-based side network benefits from flexibly activated experts, unconstrained fully training of these parameters may entangle the roles of side experts and the backbone, which causes over-fitting to task-specific patterns and forgetting of general knowledge, ultimately leading to sub-optimal performance. To address this issue, we introduce a cross-network interaction mechanism that explicitly serializes the capability of both branches, thereby fostering the learning of complementary knowledge across different network branches. In particular, the general knowledge encoded in the backbone is directly utilized to guide expert selection in the side network, ensuring complementary collaboration between two branches.

Specifically, we extract a salient token ${\boldsymbol{h}_\Phi}^0 \in \mathbb{R}^D$, which typically refers to the [CLS] token in context of the Transformer-based architecture, from the backbone as a proxy for general-purpose representations. For experts in the side network, we initiate a learnable matrix of representative tokens $\boldsymbol{r} \in \mathbb{R}^{N \times D}$, where each row $\boldsymbol{r}_i$ corresponds to the affinity of expert E_i with the requirement from general knowledge. Then, a correlation score vector $\boldsymbol{c} \in \mathbb{R}^N$ is computed by measuring the similarity between the similarity between the salient token and the expert-wise representative tokens as Eq. (7), which serve as a general-knowledge-informed prior over the expert selection process.

$$\boldsymbol{c} = \text{Norm}({\boldsymbol{h}_\Phi}^0 \times \boldsymbol{r}), \tag{7}$$

where $\text{Norm}(\cdot)$ represents normalization operation. Finally, we integrate this prior into the expert selection operation by modulating the routing probability for top-k expert selection:

$$g'(\boldsymbol{x}) = \frac{\text{Softmax}(g(\boldsymbol{x})) + c}{2}, \tag{8}$$

where $g'(\cdot)$ denotes the refined routing probability, and the subsequent gating operations follow the same process as Eq. (6).

Experimental Results and Analysis

Experimental Settings

Datasets and Evaluation Metrics. We validate our proposed MP-ISMoE on both Vision-Language (VL) and Natural Language Processing (NLP) tasks. Specifically, for VL tasks, we conduct experiments on *image-text retrieval* (ITR: Flickr30K (Young et al. 2014), MSCOCO (Lin et al. 2014)), *video-text retrieval* (VTR: MSVD (Chen and Dolan 2011), MSR-VTT (Xu et al. 2016)), *visual and compositional question answering* (VQA: VQAv2 (Goyal et al. 2017), GQA: GQA (Hudson and Manning 2019)), and *visual grounding* (VG: RefCOCO, RefCOCO+ (Yu et al. 2016), RefCOCOg (Mao et al. 2016)). By following (Diao et al. 2024b), we report Recall@1 (R@1) and Rsum of R@1,5,10 on cross-modal retrieval tasks, overall Accuracy on QA tasks, and mean Average Precision (mAP) on VG tasks. In the case of NLP task, we adopt GLUE benchmark (Wang et al. 2018a) and present Accuracy Metric, F1 Score, Matthew's Correlation, Pearson-Spearman Correlation as the evaluation metrics for various datasets respectively. Detailed descriptions are provided in *Extended Version*.

Method	Params. (M)↓	Mem. (G)↓	Flickr30K I-T↑	Flickr30K T-I↑	Flickr30K Rsum↑	MSCOCO1K I-T↑	MSCOCO1K T-I↑	MSCOCO1K Rsum↑	MSCOCO5K I-T↑	MSCOCO5K T-I↑	MSCOCO5K Rsum↑	Params. (M)↓	Mem. (G)↓	MSR-VTT T-V↑	MSR-VTT V-T↑	MSR-VTT Rsum↑	MSVD T-V↑	MSVD V-T↑	MSVD Rsum↑
Fully-FT	201.2	176.8	85.6	73.3	546.6	83.1	71.7	542.7	64.2	51.2	468.9	151.3	48.8	42.8	42.1	389.2	45.2	57.1	425.5
LST	**9.7**	**24.4**	82.1	66.5	529.5	78.2	64.8	525.8	57.8	43.1	434.5	11.2	32.0	37.0	37.8	356.7	35.5	55.4	407.2
UniPT	12.4	**24.4**	84.8	69.1	537.4	80.6	67.5	532.9	61.1	45.9	445.3	**9.6**	**13.6**	38.9	39.3	361.3	40.9	59.7	432.1
SHERL	11.3	**24.4**	86.1	71.1	542.3	81.8	69.2	537.5	62.5	47.3	450.8	**9.6**	**13.6**	39.2	40.6	363.7	40.9	60.2	429.7
Ours†	12.9	25.5	86.5	71.3	543.1	81.9	69.1	538.7	62.2	47.1	449.1	10.1	14.5	39.9	41.1	365.5	42.0	60.4	435.6
Ours‡	11.8	25.4	**87.4**	**73.3**	**547.0**	**82.8**	**71.0**	**542.6**	**63.4**	**48.7**	**453.8**	10.1	14.6	**40.3**	**41.3**	**366.9**	**42.1**	**60.7**	**435.8**

Method	Params. (M)↓	Mem. (G)↓	VQAv2 Test_D↑	VQAv2 Test_S↑	GQA Test_D↑	GQA Test_S↑	Params. (M)↓	Mem. (G)↓	RefCOCO Val↑	RefCOCO TestA↑	RefCOCO TestB↑	RefCOCO+ Val↑	RefCOCO+ TestA↑	RefCOCO+ TestB↑	RefCOCOg Val↑	RefCOCOg Test↑
Fully-FT	236.8	82.0	76.71	76.86	60.25	61.44	185.2	39.6	86.51	89.13	81.22	79.54	84.54	70.63	80.92	80.95
LST	13.4	25.6	75.29	75.44	59.93	60.75	0.9	12.6	81.63	85.19	76.03	71.32	78.20	62.06	72.53	73.67
UniPT	**10.3**	**11.6**	75.33	75.53	60.10	60.72	**0.7**	**6.8**	82.71	86.25	78.16	72.94	79.18	64.49	77.04	77.33
SHERL	13.0	14.0	75.53	75.82	60.16	60.82	**0.7**	**6.8**	83.02	86.39	78.41	73.29	80.11	64.59	77.80	77.33
Ours†	10.9	12.6	75.82	76.87	60.78	61.41	0.8	7.3	83.32	87.09	**79.20**	73.59	79.68	64.88	77.86	77.95
Ours‡	13.6	14.9	**76.21**	**76.91**	**60.91**	**61.44**	0.8	7.4	**83.49**	**87.26**	79.19	**73.92**	**80.51**	**65.02**	**78.39**	**78.08**

Table 1: Comparison results (%) with METL approaches across various architectures and distinct VL tasks, in terms of amount of learnable parameters, memory usage, and other task-specific metrics. The best results are highlighted in **bold**, and the second best results are underlined.

Counterparts. We compare MP-ISMoE with full fine-tuning and two representative efficient adaptation paradigms: (1) *Memory-Efficient* approaches exemplified by LST (Sung, Cho, and Bansal 2022), UniPT (Diao et al. 2024b), and SHERL (Diao et al. 2024a); (2) *Parameter-Efficient* methods including Partial Tuning (BitFit (Zaken, Ravfogel, and Goldberg 2022)), Prompt Tuning (Prompt (Li and Liang 2021)), and Adapter Tuning (Adapter (Houlsby et al. 2019), LoRA (Hu et al. 2022)).

Implementation Details. To ensure rigorous and fair comparisons, we maintain consistent experimental configurations with UniPT (Diao et al. 2024b) and SHERL (Diao et al. 2024a), including the optimizer, warm-up scheduler, batch size, training epochs, *etc*. Besides, our method is implemented based on UniPT/SHERL, denoted as Ours$^{\dagger/\ddagger}$, respectively. Additional training details are depicted in *Extended Version*.

Main Results

Baselines. Similar to (Diao et al. 2024a), in order to conduct a more exhaustive and challenging evaluation, we present a comparison on diverse VL and NLP tasks with various pre-trained architectures, including:

- ITR task: *VSE*∞ (Chen et al. 2021) leverages BERT-base as text and Instagram (WSL) pre-trained ResNeXt-101(32×8d) as vision backbones.
- VTR task: *CLIP4Clip* (Luo et al. 2021) adapts pre-trained CLIP's dual-Transformer framework (ViT-B/32 + Text Transformer) through temporal domain adaptation from image-text to video-text spaces.
- QA task: *CLIP-ViL* (Shen et al. 2021) utilizes frozen CLIP image encoder with text embeddings, followed by a cross-modal fusion Transformer.
- VG task: *MDETR* (Kamath et al. 2021) combines ResNet-101 and RoBERTa-B for image and text encoding, with a query-attended encoder-decoder Transformer.
- NLP task: *T5-series* (Raffel et al. 2020) imports text encoder and autogressive decoder, with balanced layer reduction (6/24 total layers of side network, equally split for encoder and decoder for *base/large*).

MP-ISMoE outweights METL methods in memory-constrained scenarios. We compare MP-ISMoE with state-of-the-art METL methods on five VL tasks. As shown in Table 1, our MP-ISMoE achieves superior performance, with the minimal discrepancy from the fully fine-tuned model. Concretely, it demonstrates the following advantages: (1) Remarkable performance improvement. MP-ISMoE outperforms LST across a range of tasks and backbones. When integrated with UniPT/SHERL (*i.e.*, Ours$^{\dagger/\ddagger}$), it yields an average improvement of 1.4/1.2% in R@1 and 4.6/4.4% in Rsum for cross-modal retrieval, 0.80/0.79% improvements for question answering, and 5.47/4.92% for visual grounding. These results in challenging pattern matching and limited data-driven scenarios strongly underscore the efficacy of MP-ISMoE. (2) Comparable training memory consumption. In most cases, MP-ISMoE reduces training memory usage by approximately 50% compared to LST. Despite a slight memory increase (*e.g.* 1GB for the retrieval task) over the baselines, we consider it acceptable in light of the gains in performance. (3) Negligible inference cost. With the sparse MoE-based side network where only a fixed number of experts are activated, MP-ISMoE introduces no

Method	Params. (%) ↓	Memory (G) ↓ Train	Test	CoLA	SST-2	MRPC	QQP	MNLI	QNLI	RTE	STS-B	Avg.
Fully-FT	100	17.6	0.86	62.8	93.9	91.9	89.9	86.2	92.5	74.1	90.3	85.2
Adapter	1.63	13.0	0.87	64.4	94.2	88.9	88.9	86.4	93.1	75.1	91.1	85.3
LoRA	1.71	12.6	0.86	63.3	94.3	90.1	89.0	86.3	93.2	75.5	90.9	85.3
BitFit	0.13	10.7	0.86	61.8	94.3	91.0	88.7	85.6	93.1	67.6	90.8	84.1
Prompt	0.03	22.2	0.87	0	90.3	74.6	88.5	82.5	92.5	59.5	90.1	72.2
LST	1.74	5.5	0.88	58.1	94.1	90.4	88.8	85.6	93.3	71.9	90.7	84.1
UniPT	1.36	**2.9**	**0.86**	62.2	94.2	90.8	88.9	85.5	93.3	69.8	89.7	84.3
SHERL	**0.85**	**2.9**	0.87	61.1	93.7	89.4	88.8	85.3	93.3	71.9	90.9	84.3
Ours†	2.41	3.2	**0.86**	**63.4**	**94.6**	**91.6**	89.1	85.7	93.4	70.9	90.0	84.8
Ours‡	1.48	3.3	0.87	62.4	94.3	89.6	**89.3**	**85.9**	**93.2**	**72.8**	**91.5**	**84.9**
LST (T5-large)	1.23	12.2	2.88	65.3	95.7	91.6	89.7	88.6	94.1	79.9	92.4	87.1
UniPT (T5-large)	0.92	9.1	2.82	65.7	95.8	92.0	89.7	88.2	94.2	79.6	92.0	87.2
SHERL (T5-large)	**0.64**	**7.1**	**2.80**	65.6	95.8	92.9	89.6	88.6	94.2	80.8	92.1	87.5
Ours† **(T5-large)**	1.75	9.9	2.82	**66.7**	**96.5**	93.1	90.0	88.7	94.7	79.8	92.4	87.7
Ours‡ **(T5-large)**	0.81	7.6	**2.80**	66.4	**96.5**	**93.4**	**90.3**	**88.9**	**94.9**	**81.6**	**92.7**	**88.1**

Table 2: Comparison results with PETL **(Top)** and METL **(Bottom)** methods on GLUE benchmark, with *T5-base/large*. We report the number of learnable parameters and memory usage as efficiency metrics, and accuracy, F1 score, Matthew's Correlation, and Pearson-Spearman Correlation as performance indicators. The best results are highlighted in **bold**, and the second best results are underlined.

extra inference cost, which ensures its practical applicability for real-world deployment.

To further assess the generalization ability of our approach, we conduct additional evaluations on NLP tasks. As shown in Table 2, MP-ISMoE improves the overall performance of baseline UniPT and SHERL by 0.6%, with comparable memory consumption. More significantly, it outperforms LST comparable trainable parameters while reducing training memory usage by over 30%.

In summary, MP-ISMoE enables a larger trainable parameter space and scales up model capacity without notably increasing memory overhead, thereby achieving a more favorable trade-off between memory efficiency and performance.

MP-ISMoE outperforms PETL methods with similar training memory consumption. We also evaluate the proposed method with state-of-the-art PETL methods on the GLUE benchmark for NLP tasks. As shown in Table 2, with the *T5-base* model as backbone, MP-ISMoE significantly reduces the training memory overhead from 17.6GB to 3.2GB, up to 81.8% of the fully fine-tuning, while in context of the same scale of trainable parameters, the prevailing Adapter and LoRA methods only gain a reduction ratio of 25.6%. Besides, MP-ISMoE surpasses BitFit and Prompt by 0.8% and 12.7% on average, respectively, while requiring only 29.9% and 14.4% of their training memory overhead. These results demonstrate that MP-ISMoE achieves significantly higher memory efficiency than both Full-FT and other PETL methods. To further exploit the memory efficiency and validate the scalability on larger backbones, we continue our comparison on the *T5-large* backbone. Remarkably, MP-ISMoE achieves a 15.7% performance gain over Prompt under similar or even lower training memory consumption, and consistently outperforms other PETL baselines without incurring additional inference memory overhead.

Ablation Study

On Main Components. We evaluate the effect of the main components, including Gaussian Noise Perturbed Iterative Quantization (GNP-IQ) and Interactive Side Mix-of-Experts (ISMoE), on the *VSE*∞ for ITR task. Here, we use UniPT as baseline. As summarized in Table 3, GNP-IQ significantly reduces training memory consumption from 24.4GB to 18.4GB (a reduction of 24.6%) by quantizing the pre-trained backbone into lower-bit precision weights. Although this inevitably leads to a slight performance degradation, it frees up substantial memory for scaling up the side network. Conversely, the single introduction of ISMoE significantly boosts performance, by improving R@1 and Rsum by 2.2% and 6.7%, respectively, at a cost of increased training memory usage. These results highlight the inherent strengths of the two modules, *i.e.* the memory efficiency of GNP-IQ and the accuracy advantage of ISMoE. When further combined, GNP-IQ and ISMoE complement each other by reallocating part of the memory budget from the backbone to the expanded side network, ultimately yielding average improvements of 1.5% in R@1 and 5.1% in RSum, respectively, with only a negligible increase in memory overhead.

On Effect of GNP-IQ. We further evaluate the individual effect of designs in GNP-IQ on the *VSE*∞ (Chen et al. 2021) for ITR task. As shown in the Table 4, on the basis of the baseline, we first fine-tune with the frozen pre-trained backbone using different weight precision formats. Specifically, the introductions of low unified precision and mixed-precision reduce training memory by 36.9% and 39.3%,

GNP-IQ	ISMoE	Params. (M)↓	Memory (G)↓	Flickr30K			MSCOCO1K			MSCOCO5K		
				I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑
		12.4	<u>24.4</u>	84.8	69.1	537.4	80.6	67.5	532.9	61.1	45.9	445.3
✓		<u>12.5</u>	**18.4**	83.7	68.1	534.8	78.9	66.3	530.4	59.9	44.7	442.7
	✓	12.7	32.4	**86.9**	**71.7**	**544.3**	**82.5**	**69.8**	**540.4**	**63.4**	**47.7**	**450.9**
✓	✓	12.9	25.5	<u>86.5</u>	<u>71.3</u>	<u>543.1</u>	<u>81.9</u>	<u>69.1</u>	<u>538.7</u>	<u>62.2</u>	<u>47.1</u>	<u>449.1</u>

Table 3: Ablation results (%) of the main components using *VSE*∞ on ITR tasks. The best results are highlighted in **bold**, and the second best results are <u>underlined</u>.

Weight Precision	Gaussian Noise	Params. (M)↓	Memory (G)↓	Flickr30K			MSCOCO1K			MSCOCO5K		
				I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑
Full-Pre		**12.4**	24.4	**84.8**	**69.1**	**537.4**	**80.6**	**67.5**	**532.9**	**61.1**	**45.9**	**445.3**
Low-Pre		**12.4**	**14.8**	81.6	66.4	529.3	77.5	64.6	524.7	58.4	43.0	433.8
Mixed-Pre		**12.4**	<u>15.4</u>	82.1	66.7	531.4	77.8	65.0	526.9	58.8	43.5	437.2
Mixed-Pre	✓	<u>12.5</u>	18.4	<u>83.7</u>	<u>68.1</u>	<u>534.8</u>	<u>78.9</u>	<u>66.3</u>	<u>530.4</u>	<u>59.9</u>	<u>44.7</u>	<u>442.7</u>
MoE Structure	Network Correlation	Params. (M)↓	Memory (G)↓	Flickr30K			MSCOCO1K			MSCOCO1K		
				I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑	I-T ↑	T-I ↑	Rsum ↑
		12.4	**24.4**	84.8	69.1	537.4	80.6	67.5	532.9	61.1	45.9	445.3
✓		<u>12.7</u>	<u>31.6</u>	<u>86.3</u>	<u>71.3</u>	<u>543.1</u>	<u>82.0</u>	<u>69.2</u>	<u>538.8</u>	<u>62.9</u>	<u>47.1</u>	<u>449.4</u>
✓	✓	<u>12.7</u>	32.4	**86.9**	**71.7**	**544.3**	**82.5**	**69.8**	**540.4**	**63.4**	**47.7**	**450.9**

Table 4: (**Top**) Ablation results (%) of GNP-IQ with various backbone weights precision (Full-, Mixed-, and Low-Precision), w/ or w/o Gaussian noise perturbation. (**Bottom**) Ablation results (%) of ISMoE w/ or w/o MoE structure and measuring correlation between networks. All experiments are conducted on ITR tasks with *VSE*∞. The best results are highlighted in **bold**, and the second best results are <u>underlined</u>.

respectively. However, the retrieval accuracy incurs severe degradation, although the drop under mixed-precision fine-tuning is relatively moderate. Furthermore, under mixed-precision, the introduction of Gaussian noise perturbation enables recovery of retrieval accuracy while maintaining a relatively low memory consumption. This is attributed to the noise-induced perturbations effectively simulating long-term weight updates, thereby mitigating the accumulated quantization error that would otherwise impair fine-tuning.

On Effect of ISMoE. We also evaluate the effect of detailed designs in ISMoE on the *VSE*∞ (Chen et al. 2021) for ITR task. As previously discussed, the memory overhead introduced by this module can be compensated by the reduction achieved from the GNP-IQ module, therefore, memory consumption is not the focus of this section. As shown in Table 4, introducing the sparse MoE structure significantly improves the R@1 and Rsum by 1.6% and 5.2%, respectively. This result underscores the effectiveness of MoE-based scaling up in enhancing transfer learning. Upon further incorporating expert selection based on salient token from the backbone, these metrics are continuously increased by 2.2% and 6.7%, respectively, while keeping the amount of learnable parameters and memory consumption basically constant. This improvement can be attributed to the more effective utilization of general knowledge from the backbone in guiding expert selection, which in turn alleviates overfitting and mitigates the forgetting of general knowledge.

We also extensively study the **influence of the ratio p of re-quantized weights** in each iteration, **impact of the number of experts N** in Eq. (6) in MoE structures. Due to space limitation, we summarize the detailed results in *Extended Version*.

Conclusion

In this paper, we propose a novel METL method dubbed Mixed-Precision Interactive Side Mixture-of-Experts (MP-ISMoE), which effectively addresses the inherent limitations of existing methods regarding the scalability and representational capabilities of side networks. We develop the Gaussian Noise Perturbed Iterative Quantization (GNP-IQ) process that enables mixed-precision training, effectively compressing the memory footprint of the backbone while preserving more performance. Furthermore, the Interactive Side Mixture-of-Experts (ISMoE) structure is introduced, scaling up the side network by reallocating the previously saved memory, and mitigating knowledge forgetting by leveraging salient token-guided expert selection. Experimental results on multiple vision-language and natural language processing tasks demonstrate that our method achieves superior balance between trainable parameters, memory efficiency and transfer learning performance, by surpassing existing state-of-the-art METL methods in accuracy with comparable memory overhead.

Acknowledgements

This work was partly by the National Natural Science Foundation of China (No. 62202034), the Beijing Natural Science Foundation (No. 4242044), the Aeronautical Science Foundation of China (No. 2023Z071051002), CCF Baidu Open Fund, the Graduate Education and Development Research Special Fund of Beihang University, and the Fundamental Research Funds for the Central Universities.

References

Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; and Huang, J. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.

Chen, D.; and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics*, 190–200.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15789–15798.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Proceedings of the Advances in Neural Information Processing Systems*, 16664–16678.

Chen, T.; Xu, B.; Zhang, C.; and Guestrin, C. 2016. Training Deep Nets with Sublinear Memory Cost. *arXiv preprint arXiv:1604.06174*.

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Diao, H.; Wan, B.; Jia, X.; Zhuge, Y.; Zhang, Y.; Lu, H.; and Chen, L. 2024a. SHERL: Synthesizing High Accuracy and Efficient Memory for Resource-Limited Transfer Learning. In *Proceedings of the European Conference on Computer Vision*, 75–95.

Diao, H.; Wan, B.; Zhang, Y.; Jia, X.; Lu, H.; and Chen, L. 2024b. UniPT: Universal Parallel Tuning for Transfer Learning with Efficient Parameter and Memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 28729–28740.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 19358–19369.

Frantar, E.; Ashkboos, S.; Hoefler, T.; and Alistarh, D. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Gomez, A. N.; Ren, M.; Urtasun, R.; and Grosse, R. B. 2017. The Reversible Residual Network: Backpropagation Without Storing Activations. In *Proceedings of the Advances in Neural Information Processing Systems*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. 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*, 6904–6913.

Harvey, D. F.; Weale, G.; and Yilmaz, B. 2025. Optimizing MoE Routers: Design, Implementation, and Evaluation in Transformer Models. *arXiv preprint arXiv:2506.16419*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; Laroussilhe, Q. D.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the International Conference on Machine Learning*, 2790–2799.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *Proceedings of the International Conference on Learning Representations*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6700–6709.

Jie, S.; and Deng, Z.-H. 2023. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1060–1068.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 1780–1790.

Kim, S.; Hooper, C.; Gholami, A.; Dong, Z.; Li, X.; Shen, S.; Mahoney, M. W.; and Keutzer, K. 2023. Squeezellm: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*.

Lee, C.; Jin, J.; Kim, T.; Kim, H.; and Park, E. 2023. Owq: Lessons learned from activation outliers for weight quantization in large language models. *CoRR*.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Annual Meeting of the Association for Computational Linguistics*, 4582–4597.

Lieber, O.; Lenz, B.; Bata, H.; Cohen, G.; Osin, J.; Dalmedigos, I.; Safahi, E.; Meirom, S.; Belinkov, Y.; Shalev-Shwartz, S.; et al. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*.

Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.-M.; Wang, W.-C.; Xiao, G.; Dang, X.; Gan, C.; and Han, S. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6: 87–100.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Liu, H.; Tam, D.; Muqeeth, M.; Mohta, J.; Huang, T.; Bansal, M.; and Raffel, C. A. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35: 1950–1965.

Liu, T.; Blondel, M.; Riquelme, C.; and Puigcerver, J. 2024. Routers in vision mixture of experts: An empirical study. *arXiv preprint arXiv:2401.15969*.

Liu, Y.; An, C.; and Qiu, X. 2024. Y-tuning: An efficient tuning paradigm for large-scale pre-trained models via label representation learning. *Frontiers of Computer Science*, 18(4): 184320.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Lv, K.; Yang, Y.; Liu, T.; Gao, Q.; Guo, Q.; and Qiu, X. 2023. Full parameter fine-tuning for large language models with limited resources. *arXiv preprint arXiv:2306.09782*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11–20.

Mao, Y.; Ge, Y.; Fan, Y.; Xu, W.; Mi, Y.; Hu, Z.; and Gao, Y. 2025. A survey on lora of large language models. *Frontiers of Computer Science*, 19(7): 197605.

Mercea, O.-B.; Gritsenko, A.; Schmid, C.; and Arnab, A. 2024. Time-Memory-and Parameter-Efficient Visual Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5536–5545.

Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Mu, S.; and Lin, S. 2025. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*.

Pan, B.; Shen, Y.; Liu, H.; Mishra, M.; Zhang, G.; Oliva, A.; Raffel, C.; and Panda, R. 2024. Dense training, sparse inference: Rethinking training of mixture-of-experts language models. *arXiv preprint arXiv:2404.05567*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; and Zhou, Y. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Shen, S.; Li, L. H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; and Keutzer, K. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Sung, Y.-L.; Cho, J.; and Bansal, M. 2022. LST: Ladder Side-Tuning for Parameter and Memory Efficient Transfer Learning. In *Proceedings of the Advances in Neural Information Processing Systems*, 12991–13005.

Sung, Y.-L.; Nair, V.; and Raffel, C. A. 2021. Training neural networks with fixed sparse masks. In *Proceedings of the Advances in Neural Information Processing Systems*, 24193–24205.

Touvron, H.; Cord, M.; El-Nouby, A.; Verbeek, J.; and Jégou, H. 2022. Three Things Everyone Should Know about Vision Transformers. In *Proceedings of the European Conference on Computer Vision*, 497–515.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, N.; Choi, J.; Brand, D.; Chen, C.-Y.; and Gopalakrishnan, K. 2018b. Training deep neural networks with 8-bit floating point numbers. *Advances in neural information processing systems*, 31.

Wei, T.; Zhu, B.; Zhao, L.; Cheng, C.; Li, B.; Lü, W.; Cheng, P.; Zhang, J.; Zhang, X.; Zeng, L.; et al. 2024. Skywork-moe: A deep dive into training techniques for mixture-of-experts language models. *arXiv preprint arXiv:2406.06563*.

Wu, X.; Huang, S.; and Wei, F. 2024a. Mixture of LoRA Experts. *arXiv preprint arXiv:.13628*.

Wu, X.; Huang, S.; and Wei, F. 2024b. Mixture of LoRA Experts. In *The Twelfth International Conference on Learning Representations*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5288–5296.

Yin, D.; Hu, L.; Li, B.; Zhang, Y.; and Yang, X. 2025. 5%¿ 100%: Breaking performance shackles of full fine-tuning on visual recognition tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20071–20081.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*, 69–85.

Zaken, E. B.; Ravfogel, S.; and Goldberg, Y. 2022. BitFit: Simple Parameter-Efficient Fine-Tuning for Transformer-Based Masked Language-Models. In *Annual Meeting of the Association for Computational Linguistics*, 1–9.

Zhang, J. O.; Sax, A.; Zamir, A.; Guibas, L.; and Malik, J. 2020. Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks. In *Proceedings of the European Conference on Computer Vision*, 698–714.

Zhang, Q.; Chen, M.; Bukharin, A.; Karampatziakis, N.; He, P.; Cheng, Y.; Chen, W.; and Zhao, T. 2023a. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16816–16825.