Supplementary Materials for *Efficient Fine-Tuning of Large Models via Nested Low-Rank Adaptation*

Lujun Li[1] Cheng Lin[2] Dezhi Li[3] You-Liang Huang[1] Wei Li[4] Tianyu Wu[1] Jie Zou[2]
Wei Xue[1] Sirui Han[1*] Yike Guo[1*]

[1] Hong Kong University of Science and Technology, [2] University of Electronic Science and Technology of China
[3] Southeast University, [4] University of Birmingham

[1]lilujunai@gmail.com, {siruihan,yikeguo}@ust.hk, [2]linchengtech@gmail.com, [3]lidezhi@seu.edu.cn

Our appendix provides supplementary information to the main paper, offering in-depth insights into our experimental procedures, extended discussions, and detailed setup configurations. It is organized into three main sections: (1) Extended Discussion, which elaborates on the differences between NoRA and existing work, acknowledges limitations, and considers potential societal impacts; (2) More Detailed Experiments, which presents additional results from our motivation experiments and extended NLP tasks; and (3) Experimental Setup and Hyperparameters, which outlines the specific configurations, hardware, software, and hyperparameters used in our studies. This comprehensive appendix aims to provide researchers with the necessary information to understand and potentially reproduce our results.

A. More Discussions

A.1. Ethics Statement

This research focuses exclusively on developing efficient techniques for Large Language Models (LLMs), utilizing publicly available datasets and models. The study does not directly address human ethics or privacy concerns. Instead, it aims to enhance the computational efficiency and adaptability of existing LLMs, which may indirectly contribute to their broader accessibility and application.

A.2. Reproducibility

The authors affirm the solid reproducibility of their results and provide specific code implementations in the appendix. The main experiments represent average outcomes from multiple repetitions, ensuring reliability and consistency. By presenting detailed results for different initial seeds, the researchers demonstrate the robustness and repeatability of their method across various conditions, further solidifying the reproducibility of their findings.

*Corresponding author. First three authors have equal contributions to experiments.

A.3. Summary of Innovations

(1) The study introduces NoRA, a novel nested parameter-efficient Low-Rank Adaptation (LoRA) design structure that optimizes the initialization and fine-tuning strategies of projection matrices. (2) The researchers propose an activation-aware Singular Value Decomposition (AwSVD) technique that adjusts weight matrices based on activation distributions, effectively managing outliers and accelerating model convergence. (3) The work constructs a unified design space for LoRA variants and develops comprehensive design guidelines, emphasizing the importance of specific design positions, serial structures, and the use of nested LoRA for enhanced performance and efficiency.

A.4. More Discussions for Related Work

The immense scale of modern Large Models presents significant challenges for their deployment and execution, necessitating the development of efficient compression and optimization techniques. Research in this area broadly spans model pruning, quantization, knowledge distillation, and automated methods for discovering optimal configurations. Model pruning aims to reduce model size by removing redundant parameters. Recent efforts have focused on automating the discovery of optimal pruning strategies. For instance, work has been done on discovering layer-wise sparsity allocations [13] and adapting layer sparsity based on activation correlation assessments [16]. Furthermore, evolutionary approaches have been proposed to generate symbolic pruning metrics from scratch, removing the need for manual design [3]. Quantization reduces the numerical precision of model weights, leading to smaller memory footprints and faster inference. Advances in this area include the development of structured binary LLMs that push beyond the 1-bit barrier [5]. To automate the complex process of mixed-precision quantization, methods have been developed to evolve training-free proxies that find efficient quantization strategies without fine-tuning [2]. For very large models,

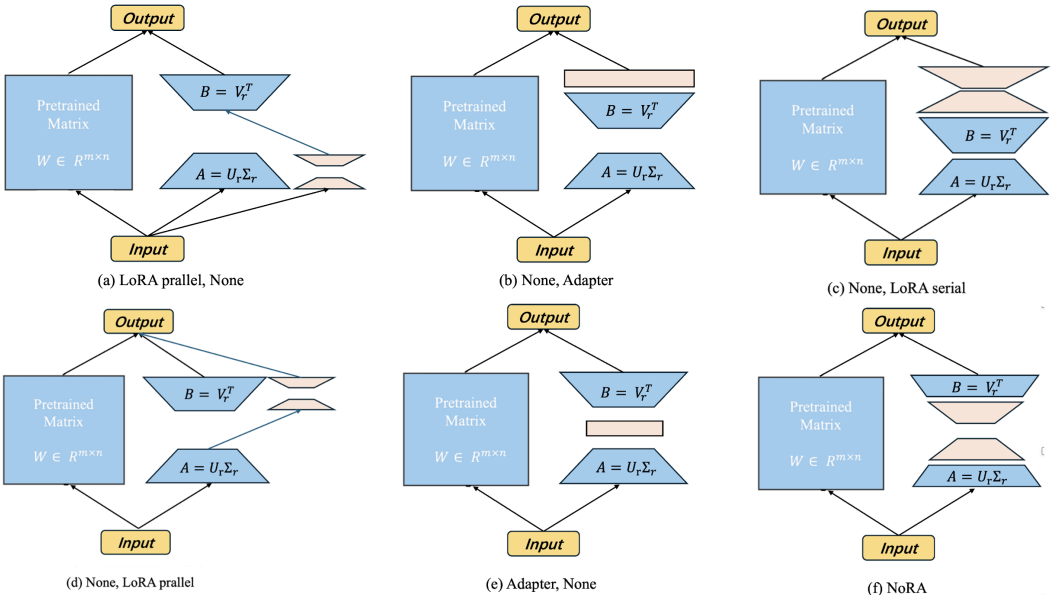

Figure 1. A subset of configurations within the unified design space (w_a, w_b).

such as Mixture-of-Experts (MoE), specialized compression techniques are crucial. These include structured compression via Singular Value Decomposition (SVD) [17] and the use of Delta Decompression to efficiently store and deploy MoE-based LLMs [6]. Knowledge Distillation (KD) is a popular compression paradigm where a smaller "student" model is trained to mimic the behavior of a larger "teacher" model. The field has explored various facets of this process. Novel frameworks like Shadow Knowledge Distillation have been introduced to bridge offline and online knowledge transfer [9]. Other approaches have focused on teacher-free distillation through self-regulated feature learning [8] or improving representation matching between teacher and student [22]. A significant trend is the automation of discovering efficient architectures and compression schemes, often using training-free, zero-cost proxies. This automation, a subset of Neural Architecture Search (NAS), avoids the prohibitive cost of repeatedly training large models. Researchers have developed parametric zero-cost proxies for efficient NAS [4] and methods for automatically discovering proxies for both generative architecture search [14] and distillation-aware search [19]. This principle has also been applied to find efficient ViT architectures [21] and to discover optimal attention patterns within them [15]. The automation extends to the KD process itself. Methodologies have been created to search for optimal student architectures [1] and to automate the entire KD process using techniques like Monte Carlo Tree Search [10]. Evolutionary algorithms have also been employed to create universal knowledge distillers that work for any teacher-student pair [12] and to search for optimal KD strategies for specific tasks like object detection [11].

A.5. Performance Gains

As the first nested LoRA method utilizing activation-aware SVD, NoRA demonstrates significant advantages in both performance and efficiency. (1) The performance gains compared to other LoRA variants are substantial, with NoRA achieving an average score of 84.4% on the LLaMA-3 8B model, surpassing LoRA's 82.8%. (2) In visual few-shot tasks, NoRA achieves the highest average accuracies of 80.9% (4 shots) and 86.1% (16 shots), outperforming existing methods. (3) The improvements in inference speed and memory optimization are notable strengths of NoRA, reducing the required parameters to as low as 4.1 million for the LLaMA-3 8B model while enhancing performance.

A.6. Comparison to Other Methods

(1) While other LoRA variants like AdaLoRA, LoRA-FA, VeRA, and LoRA-XS have made advancements in low-rank adaptation, NoRA distinguishes itself by addressing key limitations in existing approaches. The unified design space and nested structure of NoRA offer unique advantages in balancing parameter efficiency and task-specific adaptation. Unlike methods that focus solely on rank adjustment or activation memory reduction, NoRA's comprehensive approach to optimization, including its AwSVD technique and nested structure, provides a more holistic solution to the challenges of fine-tuning large language models.

A.7. Societal Impacts

The development of NoRA has potential societal implications: (1) Democratization of AI: By reducing computational requirements, NoRA could make fine-tuning large models

Table 1. Detailed results for 5 datasets with the ViT-B/16 as visual backbone. Top-1 accuracy averaged over 3 random seeds is reported. Highest value is highlighted in bold, and the second highest is underlined.

Shots 4						
(WA,WB)	Food	Pets	DTD	UCF	Cars	Average
Random, Random	85.94	93.24	64.07	79.25	73.61	79.22
$\mathbf{U\Sigma, V}$	**87.02**	**93.70**	63.77	79.12	73.39	79.40
$\mathbf{U, \Sigma V}$	86.69	93.59	**64.89**	79.75	**74.65**	**79.91**
$\mathbf{U\sqrt{\Sigma}, \sqrt{\Sigma}V}$	86.81	93.92	64.18	**79.28**	73.78	79.59

more accessible to researchers and organizations with limited resources. (2) Environmental Benefits: Increased efficiency in model adaptation could lead to reduced energy consumption and carbon footprint associated with AI research and deployment.

B. More Detailed Experiments

B.1. Motivation Experiment Results

Our motivation experiments focused on comparing different initialization strategies and architectural configurations. Key findings include:

- Figure 1 illustrates a subset of the structures within our unified design framework.
- SVD vs. Random Initialization: As shown in Table 1, SVD consistently outperformed random initialization across all tested datasets. For instance, in the Fine-tuning Vision-Language Models task, the maximum difference in average accuracy between SVD initialization and random initialization across the five datasets is 0.69 and 0.58 for 4-shot and 16-shot scenarios, respectively.
- AwSVD Performance: As shown in Figure 2, the Activation-aware SVD (AwSVD) method further improved upon standard SVD, showing about 10% reduction in output errors.
- Architectural Configurations: As shown in Table 6, the CLIP model with LoRA serial configuration outperforms the parallel configuration on diverse datasets. The average performance improvement is 2.5% and 2.55% for 4-shot and 16-shot, respectively. Additionally, compared to the adapter architecture, the LoRA serial configuration reduces the number of trainable parameters by 94%, leading to a more efficient parameter utilization.

C. Theoretical Analysis

C.1. Training Stability: Decomposition Error Bound (Theorem 1)

Theorem C.1 (Decomposition Error Bound)**.** *The spectral norm error of NoRA's approximation satisfies:*

$$\|\mathbf{W} - \mathbf{W}_{original}\|_2 \leq \sigma_{r+1} \cdot \kappa(\mathbf{S}),$$

Table 2. Detailed results for 5 datasets with the ViT-B/16 as visual backbone. Top-1 accuracy averaged over 3 random seeds is reported. Highest value is highlighted in bold, and the second highest is underlined.

Shots 16						
(WA,WB)	Food	Pets	DTD	UCF	Cars	Average
Random, Random	87.12	94.33	71.28	86.02	84.72	84.69
$\mathbf{U\Sigma, V}$	**87.60**	**94.49**	**72.70**	86.12	**85.46**	**85.27**
$\mathbf{U, \Sigma V}$	87.44	94.25	72.64	86.62	84.72	85.13
$\mathbf{U\sqrt{\Sigma}, \sqrt{\Sigma}V}$	87.56	94.17	72.40	**86.41**	85.01	85.11

Table 3. Detailed results for 5 datasets with the ViT-B/16 as visual backbone. Top-1 accuracy averaged over 3 random seeds is reported. Highest value is highlighted in bold, and the second highest is underlined. #Param represents the number of trainable parameters.

Shots 4							
w_a	#Param	Food	Pets	DTD	UCF	Cars	Average
LoRA Serial	0.59M	**87.02**	**93.65**	**66.61**	**79.73**	**74.10**	**80.22**
LoRA parallel	0.38M	85.44	93.38	62.35	74.86	72.57	77.72
Adapter Serial	10.62M	86.21	88.36	63.53	77.35	73.64	77.82

Table 4. Detailed results for 5 datasets with the ViT-B/16 as visual backbone. Top-1 accuracy averaged over 3 random seeds is reported. Highest value is highlighted in bold, and the second highest is underlined. #Param represents the number of trainable parameters.

Shots 16							
w_a	#Param	Food	Pets	DTD	UCF	Cars	Average
LoRA Serial	0.59M	**87.74**	94.33	**72.40**	**86.70**	**87.25**	**85.68**
LoRA parallel	0.38M	86.30	**94.36**	70.57	85.09	79.31	83.13
Adapter Serial	10.62M	86.80	94.06	70.80	85.70	83.24	84.27

Table 5. Different calibration datasets with the ViT-B/16 as visual backbone. Top-1 accuracy averaged over 3 random seeds is reported.

Calibration dataset	Test dataset	Test Accuracy
dtd	ucf101	78.64
dtd	stanford_cars	73.27
dtd	dtd	64.83
dtd	oxford_pets	93.73
dtd	food101	86.42
food101	food101	86.42
food101	dtd	64.60
food101	stanford_cars	73.52
food101	oxford_pets	93.76
food101	ucf101	78.98
oxford_pets	stanford_cars	73.50
oxford_pets	dtd	64.83
oxford_pets	food101	86.43
oxford_pets	oxford_pets	93.87
oxford_pets	ucf101	79.04

Table 6. Different calibration dataset sizes with the ViT-B/16 as visual backbone. Top-1 accuracy averaged over 3 random seeds is reported.

cal_batch_size	oxford_pets	food101	stanford_cars	dtd	ucf101
256	93.87	86.42	73.37	64.66	78.77
128	93.98	86.42	73.49	64.72	78.88
64	94.03	86.42	73.54	64.72	78.91

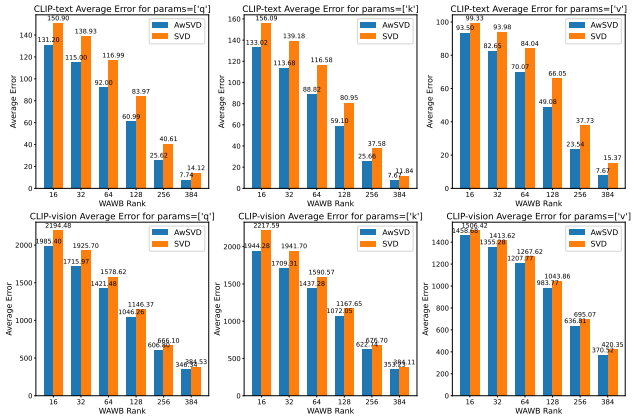

Figure 2. Comparison of SVD decomposition errors in CLIP text-encoder and vision-encoder across query projection, key projection, and value projection.

where σ_{r+1} is the $(r+1)$-th singular value of $\mathbf{W}_{aw}$, and $\kappa(\mathbf{S}) = \|\mathbf{S}\|_2\|\mathbf{S}^{-1}\|_2$ is the condition number of the activation scaling matrix $\mathbf{S}$.

Proof. Let $\mathbf{W}_{aw} = \mathbf{U}\Sigma\mathbf{V}^\top$ be the activation-weighted SVD. NoRA approximates $\mathbf{W}_{\text{original}} = \mathbf{W}_{aw}\mathbf{S}^{-1}$ using rank-r components:

$$\mathbf{W} = \mathbf{U}_r\Sigma_r\mathbf{V}_r^\top\mathbf{S}^{-1}.$$

The approximation error is:

$$\|\mathbf{W} - \mathbf{W}_{\text{original}}\|_2 = \|(\mathbf{U}_r\Sigma_r\mathbf{V}_r^\top - \mathbf{U}\Sigma\mathbf{V}^\top)\mathbf{S}^{-1}\|_2 \quad (1)$$

$$= \|(\mathbf{U}_\perp\Sigma_\perp\mathbf{V}_\perp^\top)\mathbf{S}^{-1}\|_2 \quad \text{(truncated components)} \quad (2)$$

$$\leq \|\mathbf{U}_\perp\Sigma_\perp\mathbf{V}_\perp^\top\|_2 \cdot \|\mathbf{S}^{-1}\|_2 \quad \text{(submultiplicativity)} \quad (3)$$

$$= \sigma_{r+1} \cdot \|\mathbf{S}^{-1}\|_2. \quad (4)$$

Using $\kappa(\mathbf{S}) = \|\mathbf{S}\|_2\|\mathbf{S}^{-1}\|_2$, we obtain:

$$\|\mathbf{W} - \mathbf{W}_{\text{original}}\|_2 \leq \sigma_{r+1} \cdot \kappa(\mathbf{S}).$$

Activation-weighted SVD prioritizes components with higher activation magnitudes, accelerating σ_{r+1} decay and reducing $\kappa(\mathbf{S})$, thereby stabilizing training. □ □

C.2. Convergence Acceleration: Gradient Norm Preservation (Theorem 2)

Theorem C.2 (Gradient Norm Preservation)**.** *For NoRA's outer LoRA layer initialized with AwSVD, the gradient norm satisfies:*

$$\|\tilde{g}\| \propto s \cdot \left(\|U_rU_r^\top g\| + \|gV_rV_r^\top\|\right),$$

where U_r and V_r are top-r singular vectors from activation-weighted SVD.

Proof. The equivalent gradient $\tilde{g}_t$ under LoRA decomposition $\tilde{W}_t = W_{\text{init}} + sB_tA_t$ is:

$$\tilde{g}_t = s^2\left(B_tB_t^\top g_t + g_tA_t^\top A_t\right).$$

1. Initialization with AwSVD:

$$B_t = U_r \quad \text{(orthonormal basis for column space)}, \quad (5)$$

$$A_t = \Sigma_rV_r^\top\mathbf{S}^{-1} \quad \text{(scaled row space basis)}. \quad (6)$$

2. Projection Analysis:

$$B_tB_t^\top = U_rU_r^\top \quad \text{(projection onto column space)}, \quad (7)$$

$$A_t^\top A_t = \mathbf{S}^{-\top}V_r\Sigma_r^2V_r^\top\mathbf{S}^{-1} \quad \text{(row space scaling)}. \quad (8)$$

3. Gradient Norm Decomposition:

$$\|\tilde{g}_t\| = s^2\|B_tB_t^\top g_t + g_tA_t^\top A_t\| \quad (9)$$

$$\leq s^2\left(\|U_rU_r^\top g_t\| + \|g_t\mathbf{S}^{-\top}V_r\Sigma_r^2V_r^\top\mathbf{S}^{-1}\|\right) \quad (10)$$

$$\leq s^2\left(\|g_{\|U_r}\| + \sigma_{\max}^2(\Sigma_r)\cdot\|g_t\mathbf{S}^{-1}\|\right). \quad (11)$$

4. Simplification via AwSVD Properties: Since $\mathbf{S}$ is diagonal and Σ_r contains dominant singular values:

$$\|g_t\mathbf{S}^{-1}\| \approx \|g_tV_rV_r^\top\|,$$

leading to:

$$\|\tilde{g}_t\| \propto s\cdot\left(\|U_rU_r^\top g_t\| + \|g_tV_rV_r^\top\|\right).$$

The scaling factor s amplifies gradients along critical subspaces, accelerating convergence. □ □

C.3. Adaptation Flexibility: Dual Modulation (Proposition 3)

Proposition C.3 (Weight Decomposition Dynamics)**.** *Starting from the weight decomposition $W = \|W\|_c \cdot \frac{W}{\|W\|_c}$, NoRA's update satisfies:*

$$\Delta\mathbf{W} = (\mathbf{U}_r\Sigma_r)(\mathbf{B}'\mathbf{A}')(\mathbf{V}_r^\top\mathbf{S}^{-1}),$$

where the inner matrices $\mathbf{B}'\mathbf{A}'$ modulate both magnitude (ΔM) and direction (ΔD) of updates:

$$\Delta M = \|\mathbf{B}'\mathbf{A}'\|_F\cdot\|\Sigma_r\|_F, \quad \Delta D = 1 - \cos(\mathbf{B}'\mathbf{A}', \mathbf{I}_{r\times r}).$$

Proof. **1. Update Decomposition:**

$$\Delta\mathbf{W} = \underbrace{(\mathbf{U}_r\Sigma_r)}_{\text{frozen outer}}\cdot\underbrace{(\mathbf{B}'\mathbf{A}')}_{\text{trainable inner}}\cdot\underbrace{(\mathbf{V}_r^\top\mathbf{S}^{-1})}_{\text{frozen outer}}.$$

2. Magnitude Modulation:

$$\|\Delta\mathbf{W}\|_F = \|\mathbf{B}'\mathbf{A}'\|_F\cdot\|\Sigma_r\|_F,$$

where $\|\mathbf{B}'\mathbf{A}'\|_F$ controls the update magnitude.

3. Direction Modulation:

$$\Delta D = 1 - \frac{\langle\mathbf{B}'\mathbf{A}', \mathbf{I}_{r\times r}\rangle}{\|\mathbf{B}'\mathbf{A}'\|_F\cdot\|\mathbf{I}_{r\times r}\|_F},$$

measuring alignment with identity matrix. This allows independent control of update direction while preserving pre-trained structure. □ □

Table 7. GLUE Benchmark.

Method	Trainable Parameters	QNLI
Full FT	355M	94.7
LoRA	800K	94.8
NoRA	70K	94.6

C.4. Additional Experiment Results

Extended results for natural language processing tasks:

- Based on the data in the table, we compared the performance of LoRA and NoRA methods on commonsense reasoning tasks using the LlaMA 7B model. Notably, NoRA demonstrated strong performance across multiple tasks, achieving an average score of 75.8%, which is slightly higher than LoRA's scores of 74.4% (r=16) and 75.3% (r=32).
- Question Natural Language Inference: QNLI (Question Natural Language Inference) is a task from the GLUE (General Language Understanding Evaluation) benchmark. Using the QNLI dataset, NoRA achieved an accuracy of 94.6%, compared to 94.8% for LoRA and 94.7% for full fine-tuning, while reducing trainable parameters by 91% compared to LoRA and by 99.8% compared to full fine-tuning (see Table 7).

In addition, Figure 3 presents more visual comparisons to show that our method can outperform the effects of LoRA on the generation task.

D. Experimental Setup and Hyperparameters

D.1. Model Configurations

- CLIP ViT-B/16 vision encoder: 86.19 Million parameters, 12 layers, 768 hidden size
- CLIP ViT-B/16 text encoder: 63.43 Million parameters, 12 layers, 512 hidden size
- Mistral-7B: 7 billion parameters, 32 layers, 4096 hidden size

D.2. Hardware and Software

- GPUs: 8 x NVIDIA V100S (32GB)
- Framework: PyTorch 1.10.0
- CUDA Version: 11.3

D.3. Hyperparameters

Instruction Tuning: We perform the instruction tuning experiments on Mistral-7B-v0.1 [7] , Gemma-7B [20] and LlaMA-3 8B models. We use a batch size of 128 and train for 2 epochs on 100k samples of the MetaMathQA dataset. Models are evaluated on the GSM8K and MATH datasets. The learning rate is set to 7E-3 with the AdamW optimizer [18]. The warmup ratio is 0.02, and a cosine learning rate scheduler is used. The parameter α for NoRA modules is always equal to the rank. In NoRA (0.92M), the Outer and Inner LoRA ranks are 64 and 32, respectively. We used 8 $\times$ V100S 32GB GPUs for the finetuning

Fine-tuning of Vision-Language Models: Table 9 details our hyperparameter settings for CLIP ViT-B/16, which remain consistent across all 5 datasets.

Common hyperparameters across experiments:

- Batch size: 32
- Learning rate: 1e-4 (AdamW optimizer)
- Weight decay: 0.01
- Warmup steps: 500
- Max steps: 20,000

Task-specific adjustments:

- GSM8K and Math: Increased max steps to 30,000
- Few-shot CLIP: Reduced batch size to 16, max steps to 5,000

D.4. Evaluation Metrics

- NLP tasks: Accuracy, F1 score
- Math reasoning: Pass@1 score
- Few-shot image classification: Top-1 accuracy

References

[1] Peijie Dong, Lujun Li, and Zimian Wei. Diswot: Student architecture search for distillation without training. In *CVPR*, 2023. 2

[2] Peijie Dong, Lujun Li, Zimian Wei, Xin Niu, Zhiliang Tian, and Hengyue Pan. Emq: Evolving training-free proxies for automated mixed precision quantization. *arXiv preprint arXiv:2307.10554*, 2023. 1

[3] Peijie Dong, Lujun Li, Zhenheng Tang, Xiang Liu, Xinglin Pan, Qiang Wang, and Xiaowen Chu. Pruner-zero: Evolving symbolic pruning metric from scratch for large language models. In *ICML*, 2024. 1

[4] Peijie Dong, Lujun Li, Zhenheng Tang, Xiang Liu, Zimian Wei, Qiang Wang, and Xiaowen Chu. Parzc: Parametric zero-cost proxies for efficient nas. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16327–16335, 2025. 2

[5] Peijie Dong, Lujun Li, Yuedong Zhong, Dayou Du, Ruibo Fan, Yuhan Chen, Zhenheng Tang, Qiang Wang, Wei Xue, Yike Guo, et al. Stbllm: Breaking the 1-bit barrier with structured binary llms. In *ICLR*, 2025. 1

[6] Hao Gu, Wei Li, Lujun Li, Zhu Qiyuan, Mark Lee, Shengjie Sun, Wei Xue, and Yike Guo. Delta decompression for moe-based llms compression. *arXiv preprint arXiv:2502.17298*, 2025. 2

[7] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 5

[8] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In *ECCV*, 2022. 2

Table 8. Commonsense reasoning on LlaMA 7B

Model	Method	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Avg
LlaMA 7B	LoRA$_{r=16}$	68.9	80.7	77.4	78.1	78.8	77.8	61.3	74.8	74.4
	LoRA$_{r=32}$	68.5	81.0	77.4	77.1	79.0	77.8	63.3	77.9	75.3
	NoRA	68.1	80.3	76.8	80.6	79.6	80.5	62.6	77.8	75.8

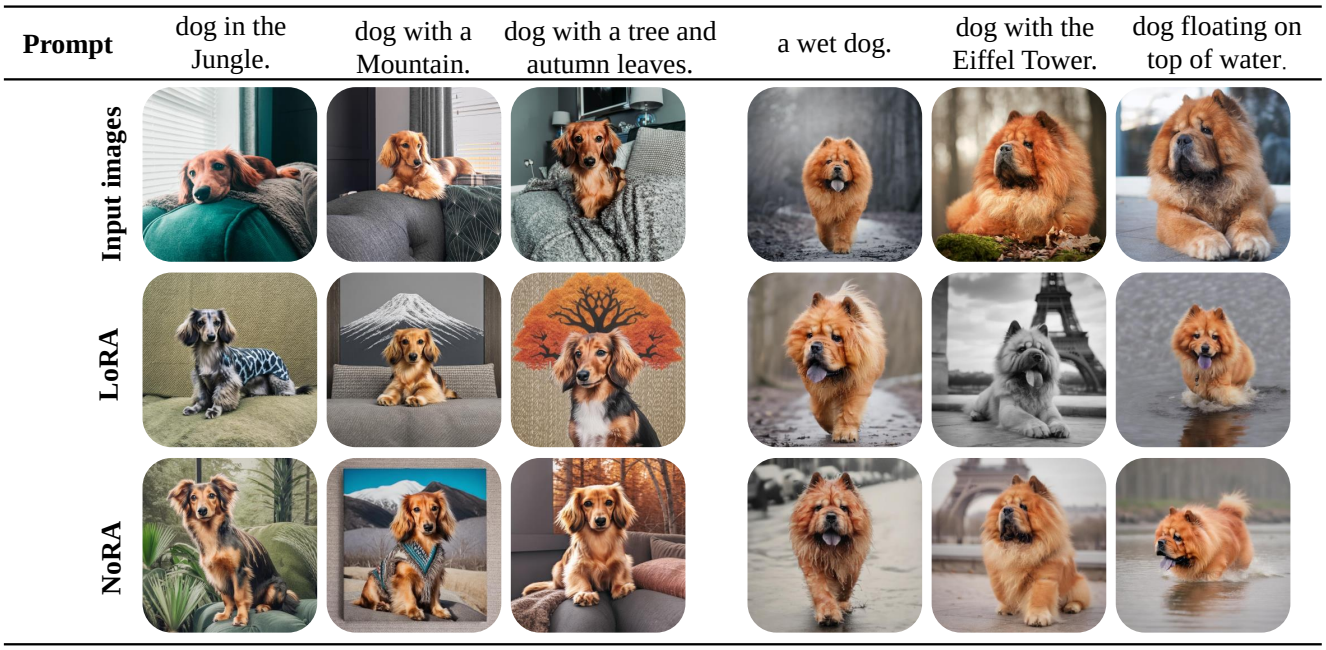

Figure 3. More visualization of LoRA and NoRA performance on subject-driven image generation task. The illustration demonstrates the benefit of NoRA for models that adapt input images based on diverse prompts (e.g., "cat in the jungle" or "dog on the beach"), emphasizing the maintenance of thematic consistency and the accurate representation of diverse environments.

Table 9. Our hyperparameter configuration on fine-tuning of Vision-Language model experiments.

Hyperparameters	**LoRA Serial**
Batch size	64
Learning rate	5e-4
Scheduler	CosineAnnealingLR
Optimizer	AdamW
Weight decay	0.01
Dropout rate	0.25
Placement	query, key, value
n_iters	400
(W_B, W_A) Init.	$(\mathbf{U\Sigma}, \mathbf{VS^{-1}})$
Outer LoRA rank	256
Inner LoRA rank	16

[9] Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. In *NeuIPS*, 2022. 2

[10] Lujun Li, Peijie Dong, Zimian Wei, and Ya Yang. Automated knowledge distillation via monte carlo tree search. In *ICCV*, 2023. 2

[11] Lujun Li, Yufan Bao, Peijie Dong, Chuanguang Yang, Anggeng Li, Wenhan Luo, Qifeng Liu, Wei Xue, and Yike Guo. Detkds: Knowledge distillation search for object detectors. In *ICML*, 2024. 2

[12] Lujun Li, Peijie Dong, Anggeng Li, Zimian Wei, and Ya Yang. Kd-zero: Evolving knowledge distiller for any teacher-student pairs. *NeuIPS*, 2024. 2

[13] Lujun Li, Peijie, Zhenheng Tang, Xiang Liu, Qiang Wang, Wenhan Luo, Wei Xue, Qifeng Liu, Xiaowen Chu, and Yike Guo. Discovering sparsity allocation for layer-wise pruning of large language models. In *NeuIPS*, 2024. 1

[14] Lujun Li, Haosen Sun, Shiwen Li, Peijie Dong, Wenhan Luo, Wei Xue, Qifeng Liu, and Yike Guo. Auto-gas: Automated proxy discovery for training-free generative architecture search. ECCV, 2024. 2

[15] Lujun Li, Zimian Wei, Peijie Dong, Wenhan Luo, Wei Xue, Qifeng Liu, and Yike Guo. Attnzero: efficient attention discovery for vision transformers. In *ECCV*, 2024. 2

[16] Wei Li, Lujun Li, Mark Lee, and Shengjie Sun. Als: Adaptive layer sparsity for large language models via activation correlation assessment. In *NeuIPS*, 2024. 1

[17] Wei Li, Lujun Li, You-Liang Huang, Mark G. Lee, Shengjie Sun, Wei Xue, and Yike Guo. Structured mixture-of-experts

LLMs compression via singular value decomposition. In *ICML*, 2025. 2

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[19] Haosen Sun, Lujun Li, Peijie Dong, Zimian Wei, and Shitong Shao. Auto-das: Automated proxy discovery for training-free distillation-aware architecture search. ECCV, 2024. 2

[20] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 5

[21] Zimian Wei, Peijie Dong, Zheng Hui, Anggeng Li, Lujun Li, Menglong Lu, Hengyue Pan, and Dongsheng Li. Auto-prox: Training-free vision transformer architecture search via automatic proxy discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 2

[22] Liu Xiaolong, Li Lujun, Li Chao, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching. In *ICLR*, 2023. 2