LEARNING RATE AS STRUCTURE IN MIXTURES OF LORA EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixtures of LoRA experts (MoE-LoRA) are reported to outperform LoRA in multi-task learning. However, they outperform LoRA at a shared learning rate, not at LoRA's finely tuned one. We show that MoE-LoRA structures mainly act as hidden learning-rate multipliers (*e.g.*, finely tuned learning rates remove 89% (vision) and 96% (language) of the accuracy difference between methods). We further show that tasks need different learning rates, which one shared multiplier cannot provide. For instance, on the same smallNORB images, the camera's azimuth needs a large learning rate and its elevation a small one. We explain these differences with a Bayesian update, where the learning rate sets the gain (*i.e.*, how far training moves the pretrained model toward the labels). Because a confident prior needs a small gain, we measure this confidence with frozen features and predict before training which tasks need a small learning rate. We confirm every prediction after training. We also show that tasks should share parameters only when they need the same learning rate. We therefore propose PrismLoRA, which splits the LoRA rank into blocks at different learning rates and lets each task use the blocks near its optimal learning rate. PrismLoRA sets each task's blocks automatically from frozen features and validation curves. PrismLoRA outperforms all eight tuned baselines with matched parameters. Specifically, PrismLoRA improves over the strongest baseline by 2.63 points on VTAB-6 and 1.35 on Structured-8, and ranks first on language tasks. How much to trust the prior should therefore be set per task by the model's structure. Code: `https://anonymous.4open.science/r/PrismLoRA-182E`

1 INTRODUCTION

Mixtures of LoRA experts (MoE-LoRA) fit one model to several input–output distributions in label-scarce domains, *e.g.*, 14 findings per chest radiograph (Wang et al., 2017; Irvin et al., 2019). MoE-LoRA outperforms LoRA (Hu et al., 2022) in published multi-task comparisons, but none of them tunes the learning rate of LoRA. For instance, MoLoRA (Zadouri et al., 2024), MOELoRA (Liu et al., 2024) and MTL-LoRA (Yang et al., 2025) report 0.81 to 1.23 points above LoRA with equal trainable parameters. What does MoE-LoRA add over LoRA at its own tuned learning rate, *i.e.*, the grid value with the best validation accuracy?

We find that each MoE-LoRA method trains mainly like LoRA at a multiplied learning rate, and that its remaining improvement comes from per-distribution parameters. We tune LoRA and seven MoE-LoRA methods on one learning-rate grid on two VTAB-1k suites, VTAB-6 and Structured-8 (Zhai et al., 2019). Tuning each method, instead of the standard 10^{-3} (Zhang et al., 2025b), removes 89% of the accuracy variance across methods on VTAB-6 (Fig. 2c). It removes 96% on nine language distributions with LLaMA-2-7B (Hu et al., 2023; Cobbe et al., 2021; Touvron et al., 2023). A known equivalence between a forward scale and the learning rate under Adam explains the multiplier (Kingma & Ba, 2015; Hayou et al., 2024). The remaining improvement over LoRA is at most 0.99 points on VTAB-6. On Structured-8, it is 2.75 to 6.42 points for the five methods with per-distribution parameters, and -0.10 and -0.91 for the two others.

Distributions need different gains, even on the same inputs, where the gain K is the fraction of the residual r that training adds to the pretrained weights μ_0,

$$\mu_1 = \mu_0 + K\, r, \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

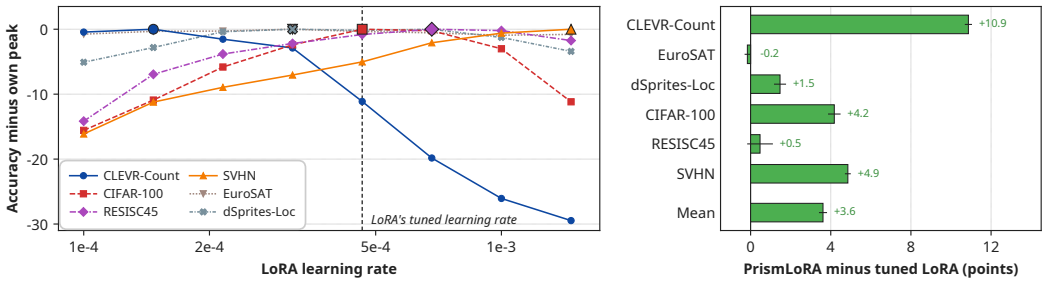

Figure 1: **At the tuned learning rate of LoRA, CLEVR-Count and SVHN fall 11.12 and 5.03 points below their best, and PrismLoRA improves them most.** (a) Test accuracy of LoRA on each VTAB-6 distribution minus its best over the learning-rate grid. Large markers show the best learning rate of each distribution, and the dashed line the tuned learning rate of LoRA. (b) Accuracy of PrismLoRA minus that of tuned LoRA, paired by seed (mean $+3.61$). ViT-B/16, three seeds, standard-error whiskers.

and r is the weight change that fits the training labels (§3). Under Adam, K grows with the learning rate, so distributions that share parameters and a learning rate share one K. One K cannot serve all, because the best learning rates of LoRA on VTAB-6 span a factor of 10 (Fig. 1a). For instance, on the same CLEVR images, learning rate 10^{-3} costs counting 25.62 points and distance 3.42 (Fig. 3b).

Frozen features predict, before training, which distributions need a K below the shared one. The Kalman gain explains this, as the K that minimizes the expected squared error of μ_1 (Ollivier, 2018),

$$K = \frac{P}{P + R}, \tag{2}$$

where P is the prior variance of μ_0 along r, and R is the variance that the finite training set adds to r. Where frozen features already predict the label, P is small, so K is small. For ordinal labels (*e.g.*, count, distance, angle), a ridge regression on frozen features makes this call. It matches the measured best learning rates on 8 of 8 Structured-8 distributions and 4 of 4 structured distributions trained with CIFAR-100 and SVHN.

Different gains need separate parameters, because a shared weight moves by the summed updates of all distributions that use it. On VTAB-6, per-distribution learning rates and Adam states on one shared LoRA, as in AdaTask (Yang et al., 2023), reach 80.26 mean accuracy. Two owned blocks, one for CLEVR-Count and one for the rest, reach 82.51 (§4.3). So distributions that need the same gain should share one block.

PrismLoRA splits the rank of LoRA into blocks and trains block z at learning rate $\eta_0 \rho^z$ (base η_0, ratio $\rho > 1$), so each block has its own gain (Fig. 4). PrismLoRA weights the blocks of each distribution by a Gaussian over z, centered at a learning rate estimated on validation data from the sweep every baseline already runs (§5.2). So PrismLoRA shares fewer blocks between distributions whose gains differ more.

PrismLoRA outperforms all eight tuned baselines on VTAB-6 and Structured-8 (Table 1). On VTAB-6, PrismLoRA reaches 84.42 mean accuracy (MOELoRA 81.79, LoRA 80.80), with the largest improvements on CLEVR-Count and SVHN (+10.87 and +4.85, Fig. 1). On Structured-8, PrismLoRA reaches 57.93 (MTL-LoRA 56.58, LoRA 50.16), and on the language distributions 70.14 (MTL-LoRA 69.92, LoRA 69.30).

We make three contributions.

- **Explanation.** MoE-LoRA trains mainly like LoRA at a multiplied learning rate, plus per-distribution parameters.
- **Principle.** Distributions need different gains, predictable from frozen features, and different gains need separate parameters.
- **Design.** PrismLoRA beats the strongest tuned baseline by 2.63 (VTAB-6) and 1.35 points (Structured-8).

2 Related Work

Parameter-efficient fine-tuning. Parameter-efficient fine-tuning (PEFT) outperforms full fine-tuning when labels are scarce (Hu et al., 2022; Chen et al., 2022; Jie & Deng, 2023; Dutt et al., 2024; Mai et al., 2025). PEFT tunes one learning rate per task on validation data, one model per task as in VTAB-1k (Jia et al., 2022; Lian et al., 2022; Xin et al., 2024; Le et al., 2026; Zhai et al., 2019). Tuned PEFT variants reach similar accuracy (Mai et al., 2025; Lee et al., 2026; He et al., 2026). PrismLoRA brings per-distribution learning rates into one jointly trained model.

Multi-task PEFT and mixtures of LoRA experts. Multi-task LoRA and MoE-LoRA share one model across tasks through task-specific parameters or routed experts (Zadouri et al., 2024; Liu et al., 2024; Tian et al., 2024; Yang et al., 2025; Ponti et al., 2023; Caccia et al., 2023; Dou et al., 2024; Wu et al., 2024; Wang et al., 2024; Agiza et al., 2024; Gao et al., 2025; Tian et al., 2026). They credit their improvement to routing and expert specialization (Chen et al., 2024; Shen et al., 2024; Liao et al., 2025; Wang et al., 2025). Several read the LoRA scale as a learning rate (Hu et al., 2022; Dou et al., 2024; Kalajdzievski, 2023), or correct expert scales or gradients (Fan et al., 2025; Sun et al., 2025). Yet they use one learning rate, even for task-specific parameters (Yang et al., 2025; Tian et al., 2024; Agiza et al., 2024) and rank-wise experts (Zhang et al., 2025a; Zhao et al., 2026). PrismLoRA reads mixture structure as a gain and sets one per distribution on owned blocks.

Learning rates and optimizers. Optimizers set learning rates per parameter group, parameter or model. Parameter groups are layers (Howard & Ruder, 2018; You et al., 2017; Clark et al., 2020; Bao et al., 2022), LoRA matrices (Hayou et al., 2024) or ranks (Wang et al., 2026). Adaptive optimizers set a per-parameter gain from gradient statistics (Kingma & Ba, 2015; Balles & Hennig, 2018), as in Kalman and Bayesian filtering (Singhal & Wu, 1988; de Freitas et al., 2000; Ollivier, 2018; Aitchison, 2020). Learning-rate-free optimizers set one learning rate per model (Baydin et al., 2018; Defazio & Mishchenko, 2023; Ivgi et al., 2023; Mishchenko & Defazio, 2024) (two are baselines in Table 3). Meta-learning learns per-task learning rates for separate per-task parameter copies (Antoniou et al., 2019; Li et al., 2017; Baik et al., 2020; Lee et al., 2020).

Multi-task learning and task grouping. Multi-task optimization reweights task losses or edits task gradients (Chen et al., 2018; Kendall et al., 2018; Yu et al., 2020; Liu et al., 2021; Navon et al., 2022; Liu et al., 2023), and tuned scalarization often matches it (Kurin et al., 2022; Xin et al., 2022; Royer et al., 2023). Others keep per-task optimizer states for shared parameters (Malkiel & Wolf, 2021; Yang et al., 2023). Task grouping partitions tasks by transfer affinity (Standley et al., 2020; Fifty et al., 2021). PrismLoRA groups distributions by gain and gives each group owned blocks.

3 Preliminaries

Problem. We fit one model to n input–output distributions with 1,000 labeled examples each, and we know the distribution of every training and test example. We define distribution d by its joint density $p_d(x, y)$, so the same inputs with a new label set form a new distribution (*e.g.*, CLEVR counting and distance (Zhai et al., 2019)). We minimize the mean loss $\frac{1}{n}\sum_{d=1}^{n}\mathbb{E}_{p_d}\,\ell\big(f_\theta(x;d),y\big)$ (Caruana, 1997) and report the mean accuracy over distributions.

LoRA and MoE-LoRA. We call a block one LoRA pair (B, A) of rank q on a frozen weight W_0 (Hu et al., 2022). We initialize $B = 0$, so training starts at pretrained weights μ_0. We write the M-block update for input x of distribution d as

$$\Delta W(x, d) = s \sum_{m=1}^{M} g_m(x, d)\, B_m A_m, \tag{3}$$

with forward scale s and gate $g_m(x, d) \geq 0$. We recover LoRA with $M = 1$, $g_1 \equiv 1$. We thus specify a method by which distributions read each block m and at which scale $s\, g_m$ (Section 4.1). We train every method with AdamW (Loshchilov & Hutter, 2019) under one recipe, with a learning rate η per block and head.

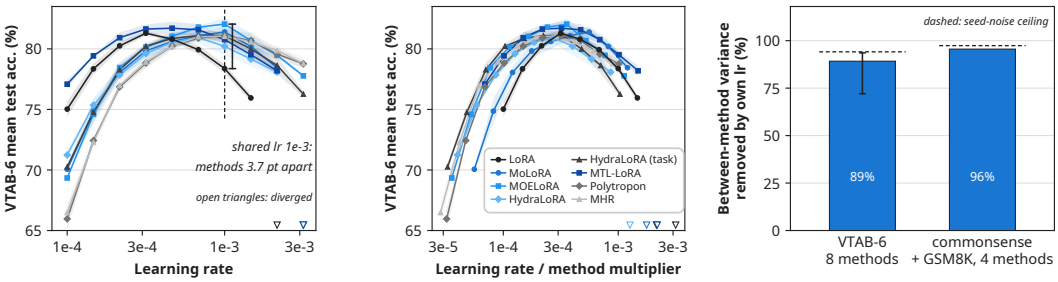

Figure 2: **MoE-LoRA methods differ at one shared learning rate and nearly coincide once each learning rate is divided by the hidden multiplier of its method.** (a) VTAB-6 mean test accuracy against the learning rate. (b) The same curves against the learning rate divided by the multiplier of each method, fitted on separate runs. (c) Share of the between-method variance at one shared learning rate that the own learning rate of each method removes, with seed-noise ceilings dashed. ViT-B/16 (a–c) and LLaMA-2-7B (c), three seeds.

The learning rate sets the gain. Under Adam, the learning rate η sets the gain K_d of Eq. (1) for two reasons. First, each coordinate moves by about η per iteration, whatever the gradient scale (Kingma & Ba, 2015; Balles & Hennig, 2018). Second, T iterations of gradient descent at η on a quadratic loss match ridge regression toward μ_0 with precision $1/(\eta T)$ (Sjöberg & Ljung, 1995; Ali et al., 2019) (Appendix A). So a larger η at fixed T weakens the prior and raises K_d (Ollivier, 2018; Aitchison, 2020).

4 Motivation Analysis: Structure as a Hidden Learning Rate

4.1 A Common Hidden Gain

We implement LoRA and seven MoE-LoRA methods (§6.1) as special cases of Eq. (3) (Appendix C).

Each structural element of MoE-LoRA acts mainly as a learning-rate multiplier. Under Adam, a constant c on a block output (*e.g.*, a gate weight) multiplies its learning rate and initialization by $\sqrt{c}$ (Appendix A). A uniform gate therefore turns MoE-LoRA into LoRA at another learning rate. The rank acts the same way, because the best learning rate scales as rank$^{-0.46}$ (predicted rank$^{-1/2}$; Fig. 11). So each method trains like LoRA at its learning rate divided by a hidden multiplier, from $\times 1.45$ (MTL-LoRA) to $\times 3.52$ (MHR) (Fig. 12).

This multiplier explains most of the accuracy gap between methods (Fig. 2). At 10^{-3}, VTAB-6 accuracy ranges from 78.37 (LoRA) to 82.05 (MOELoRA). Dividing each learning rate by its multiplier reduces the root-mean-square gap between the curves from 1.74 to 0.82 points (seed noise 0.22; Fig. 2b). So at one learning rate, a MoE-LoRA method changes each distribution as LoRA at a smaller learning rate does (Pearson 0.94; Fig. 10).

Routing is not the main source of improvement, because it does not split distributions. One trained routing decision carries at most 5.52% of the entropy of the distribution label, even when the gate reads the distribution identity (Fig. 14). The exception is MTL-LoRA, whose gate separates distributions already at initialization. So routing leaves one gain common to all distributions.

4.2 One Gain per Distribution

The best learning rate differs across distributions by up to a factor of 10. For LoRA on VTAB-6, it ranges from 1.47×10^{-4} (CLEVR-Count) to 1.47×10^{-3} (SVHN) (Fig. 1a). Per-distribution best learning rates beat the best shared one by 2.02 points on VTAB-6 (24 of 24 method–seed pairs).

By Eq. (2), a distribution needs a small gain when frozen features already predict its labels, because the prior variance P along r is then small. Because the labels define r, the gain depends on the label set and not only on the inputs.

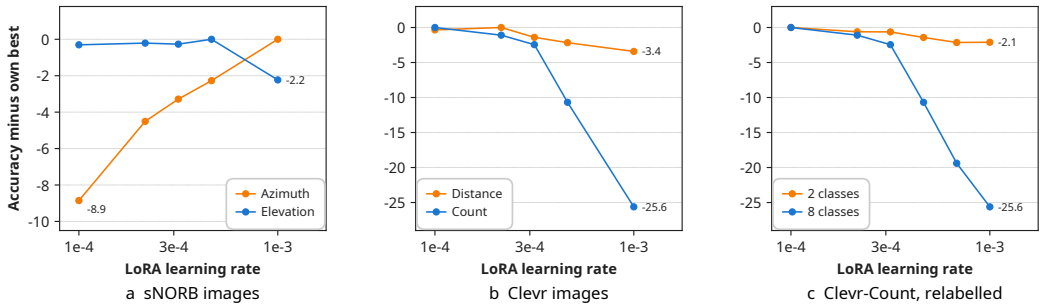

Figure 3: **On the same images, the label set decides how much a large learning rate costs.** LoRA test accuracy minus its best on the plotted grid, against the learning rate. (a) smallNORB azimuth and elevation. (b) CLEVR counting and distance. (c) CLEVR counting with 8 classes and with 2 (at most 6 against at least 7 objects). ViT-B/16, 20 epochs, three seeds for 8-class counting and two for the other curves.

We test this by changing only the labels of fixed images (Fig. 3). On smallNORB, azimuth is best at 10^{-3} and elevation at 4.64×10^{-4}. On CLEVR, 10^{-3} costs counting 25.62 points and distance 3.42 against their best. Merging the 8 count classes into 2 cuts this cost from 25.62 to 2.11 points.

We predict a small gain before training when a ridge regression on frozen features explains over 70% of the variance of an ordinal label. On Structured-8, this rule names only CLEVR-Count, and all eight methods confirm it. From its validation-selected learning rate to 10^{-3}, LoRA loses 21.4 points on CLEVR-Count and under 10 elsewhere.

Best learning rates separate only once training fits the training data. On GSM8K and commonsense with 15k examples and 2 epochs, training does not fit, and their best learning rates coincide (Fig. 18). With 1,000 examples per dataset, LoRA at 1.7×10^{-4}, the best learning rate of the nine-dataset mean, fits 93.4% of GSM8K training problems. GSM8K then prefers 5.5×10^{-5}, raising its test accuracy from 20.92 to 24.03 (Fig. 19). Standard budgets fit label-scarce data, *e.g.*, 100 epochs on 1,000 VTAB-1k images (Jia et al., 2022; Zhang et al., 2025b).

4.3 The Gain Lands on Owned Parameters

A block gives one gain to all distributions that read it, because under Adam each weight entry moves by about one learning rate per iteration, summed over all readers. A block is owned when only one distribution or group reads it.

We test this on VTAB-6 with per-distribution learning rates and Adam states on one shared LoRA, as in AdaTask (Yang et al., 2023). It falls 2.24 points below two owned blocks (3 of 3 seeds) and 6.80 on CLEVR-Count. Keeping the learning rate of CLEVR-Count fixed, we raise only those of the other five by $\times 3.15$. CLEVR-Count still loses 2.83 points, so the other readers of the block set its gain.

Owned blocks realise per-distribution gains, and distributions with equal gains can share one block (§6.2). They also separate distributions that share images, lifting dSprites location on Structured-8 from 58.31 to 83.50 (Table 7).

Because the other structural elements act as learning-rate multipliers (§4.1), a configuration has only three coordinates. They are sharing (which distributions read which block), the learning rate of each block up to its cap (the largest stable learning rate), and the head learning rate. The head learning rate differs by distribution (Appendix E) but does not replace block gains (§6.3).

5 Our PrismLoRA: Learning Rate as Structure

PrismLoRA makes the learning rate a coordinate of the adapter, so each distribution trains at its own gain on blocks that its group owns (Section 4).

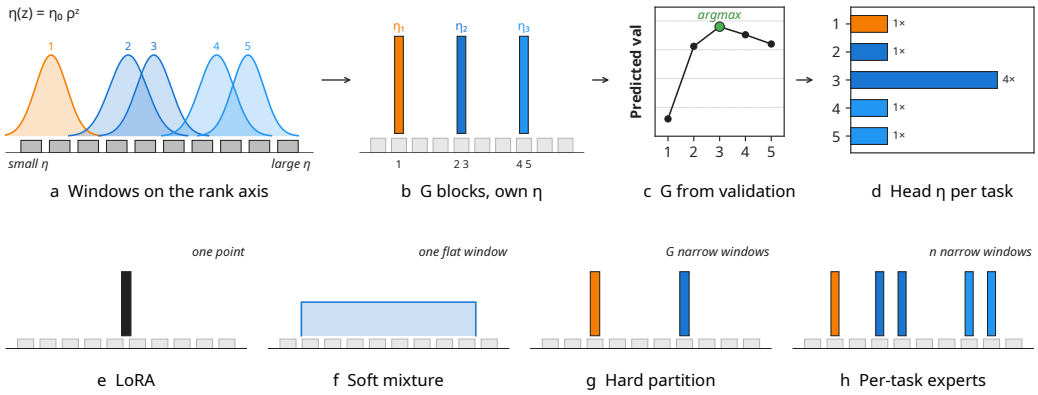

Figure 4: **PrismLoRA trains the rank blocks of LoRA at learning rates along one axis, and LoRA and MoE-LoRA structures are special windows on this axis.** (a) Each distribution reads the blocks through a Gaussian window; block m trains at $\eta_0 \rho^{z_m}$. (b) Close windows merge into G groups, each one block at its own learning rate. (c) G maximizes the validation accuracy predicted from the sweep. (d) Each head trains at 1 or 4 times the base learning rate. (e) LoRA is one block. (f) A uniformly gated MoE-LoRA is one flat window. (g) Hard grouping is G zero-width windows. (h) One expert per distribution is n zero-width windows. Positions and curves are illustrative.

5.1 The Learning-Rate Axis

PrismLoRA splits the rank of LoRA into M blocks (B_m, A_m) of rank q, places block m at position z_m on a learning-rate axis (Fig. 4a), and trains it with AdamW (Loshchilov & Hutter, 2019) at learning rate

$$\eta_m = \eta_0 \, \rho^{z_m}, \tag{4}$$

with base learning rate η_0 and grid step $\rho = 10^{1/6}$. So each block has its own gain K (Eq. (1)).

The gate of Eq. (3) becomes a fixed Gaussian window w_d with centre ζ_d and width σ_d,

$$\Delta W(x, d) = s \sum_{m=1}^{M} w_{d,m} \, B_m A_m, \qquad w_{d,m} \propto \exp\!\big(-(z_m - \zeta_d)^2 / 2\sigma_d^2\big), \quad \textstyle\sum_m w_{d,m} = 1. \tag{5}$$

The centre ζ_d sets the gain of d, and the width σ_d sets its sharing (Appendix B.1).

Proposition 1 (Special cases)**.** *Eqs. (4) and (5) include (i) LoRA, with $M = 1$; (ii) an MoE-LoRA with uniform gates, with all z_m equal and $\sigma_d \to \infty$; (iii) hard grouping into G groups, with $\sigma_d \to 0$ at G distinct positions; (iv) one expert per distribution, with $\sigma_d \to 0$ at n distinct positions.*

Appendix B gives the proofs.

5.2 Solving the Configuration from Data

PrismLoRA solves every quantity on validation data, using the learning-rate sweep that every baseline runs (Algorithm 1).

1. **Positions.** PrismLoRA fuses two estimates of each position by the Kalman weight (Eq. (2)) (Efron & Morris, 1973; Ignatiadis & Wager, 2019),

$$\zeta_d = \bar{\zeta}_d + \frac{P}{P+R}\,(\hat{\zeta}_d - \bar{\zeta}_d) - \kappa\sqrt{\frac{PR}{P+R}}. \tag{6}$$

   The prior $\bar{\zeta}_d$ (variance P) comes from frozen features (Section 4.2), and the observation $\hat{\zeta}_d$ (variance R) is the smoothed arg max of LoRA's validation curve on d (Appendix B.2). The term with $\kappa = 0.3$ lowers ζ_d, because an overly large learning rate costs more accuracy than an overly small one (Arrow et al., 1951).

2. **Groups.** PrismLoRA cuts the sorted positions at the $G-1$ largest gaps, because optimal partitions are intervals (Fisher, 1958; Topkis, 1978). Each group trains one block at the mean position of its members, *i.e.*, $\sigma_d \to 0$. Distributions with the same inputs own separate copies of their group's block (Section 4.3).
3. **Number of groups.** PrismLoRA keeps the G with the highest mean validation accuracy, read from the curve of LoRA on each distribution at the learning rate of its group. $G = 1$, *i.e.*, tuned LoRA, is always a candidate.
4. **Cap.** PrismLoRA caps a block of rank q at $\eta_{\mathrm{L}}\sqrt{q/q_{\mathrm{L}}}$, where LoRA of rank q_{L} diverges at η_{L}.
5. **Global and head learning rates.** PrismLoRA shifts the axis by -1, 0 or $+1$ grid steps and trains each head at 1 or 4 times the tuned learning rate of LoRA, both chosen on validation.
6. **Update scale.** At evaluation, PrismLoRA scales $\Delta W(x, d)$ by γ_d, chosen on validation, so the gain becomes $\gamma_d K$ (Appendix B.5).

5.3 Justification

Appendix B justifies each step. PrismLoRA selects on validation because training loss does not identify the gain (Table 3). PrismLoRA has 9.73M parameters, fewer than LoRA (9.88M; Appendix B.8).

6 Experiments

6.1 Setup

Benchmarks. We report equal-weight mean test accuracy on three label-scarce benchmarks. VTAB-6 holds six VTAB-1k distributions with 800 training and 200 validation images each (Zhai et al., 2019; Zhang et al., 2025b). Structured-8 holds the eight structured VTAB-1k distributions, whose CLEVR, dSprites and smallNORB images each carry two label sets. The language benchmark holds eight commonsense datasets and GSM8K, 1,000 training examples each (Clark et al., 2019; Bisk et al., 2020; Sap et al., 2019; Zellers et al., 2019; Sakaguchi et al., 2020; Clark et al., 2018; Mihaylov et al., 2018; Cobbe et al., 2021).

Backbones, baselines and protocol. We adapt ViT-B/16 (ImageNet-21k) for 20 epochs with the recipe of NOAH, and LLaMA-2-7B for 5 epochs with that of LLM-Adapters (Dosovitskiy et al., 2021; Ridnik et al., 2021; Zhang et al., 2025b; Loshchilov & Hutter, 2019; Touvron et al., 2023; Hu et al., 2023; Yang et al., 2025). We compare with LoRA (Hu et al., 2022) and seven MoE-LoRA methods, *i.e.*, MoLoRA (Zadouri et al., 2024), HydraLoRA (learned or task gate) (Tian et al., 2024), MOELoRA (Liu et al., 2024), MTL-LoRA (Yang et al., 2025), Polytropon (Ponti et al., 2023) and MHR (Caccia et al., 2023). All methods match the trainable parameters of MoLoRA. We select every learning rate on validation from one grid per benchmark, and report three seeds on vision and seed 42 on language (App. F).

Solved configurations. On VTAB-6, PrismLoRA selects $G = 3$ rank-22 groups, {CLEVR-Count}, {CIFAR-100, EuroSAT, dSprites-Loc} and {RESISC45, SVHN}. On Structured-8, it selects $G = 4$ and eight rank-8 blocks, one per distribution. On both, it selects global learning rate 2.15×10^{-4} and $\gamma_d = 1$, and 4 times the tuned LoRA learning rate (4.64×10^{-4}) only for the CIFAR-100, dSprites-Loc and dSprites-Ori heads. On language, it selects $G = 1$ and $\gamma_d = 0.75$ only on GSM8K, HellaSwag and ARC-e.

6.2 Results

PrismLoRA has the highest accuracy on all three benchmarks (Table 1). On VTAB-6, its worst seed (84.25) exceeds the best seed of MOELoRA (81.95), the strongest baseline. On Structured-8, it exceeds MTL-LoRA at every seed ($+1.26$, $+2.15$, $+0.62$), and on language by 0.22 points.

The improvement lands on the distributions that one shared learning rate trades away (Fig. 5). On VTAB-6, it is largest on CLEVR-Count and SVHN, whose best learning rates lie a factor of 10 apart (Fig. 1a). On Structured-8, it is largest on distributions that share images (dSprites-Loc $+33.79$,

Table 1: **At matched parameters, PrismLoRA outperforms LoRA and seven MoE-LoRA methods, each at its validation-selected learning rate.** Params: adapter parameters on VTAB-6, heads excluded.

		ViT-B/16		Commonsense + GSM8K
Method	Params (M)	VTAB-6	Structured-8	1k per task
LoRA	9.88	80.80±0.17	50.16±0.85	69.30
MoLoRA	9.95	81.08±0.30	49.25±0.23	69.46
HydraLoRA	9.98	80.92±0.47	50.06±0.61	69.26
HydraLoRA (task gate)	10.20	81.10±0.37	53.67±0.36	68.99
MOELoRA	9.44	81.79±0.13	52.91±0.19	69.03
MTL-LoRA	9.99	81.71±0.53	56.58±0.43	69.92†
Polytropon	9.44	81.09±0.56	53.97±0.74	68.28
MHR	9.58	81.05±0.36	54.05±0.42	68.28
PrismLoRA (Ours)	9.73	**84.42**±0.25	**57.93**±0.37	**70.14**

Vision columns: mean ± standard deviation over three seeds.

Language column: seed 42 for every method, each at its own validation-selected learning rate.

†Learning rate selected at the lower edge of the grid (1.7×10^{-4}); the grid was not extended.

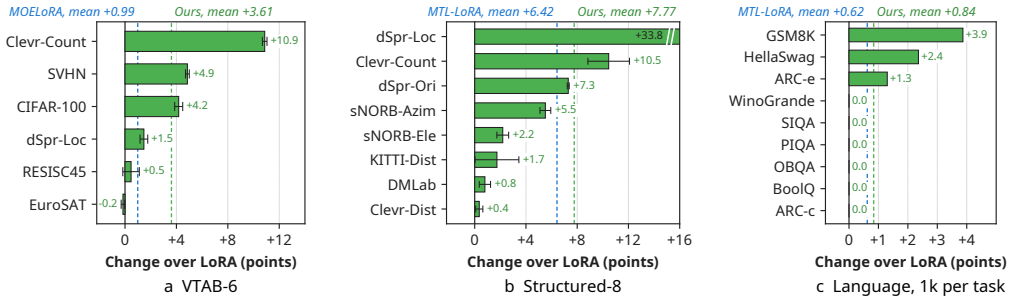

Figure 5: **PrismLoRA improves most on the distributions that one shared learning rate trades away.** Accuracy of PrismLoRA minus tuned LoRA per distribution of (a) VTAB-6, (b) Structured-8 and (c) the language benchmark. Dashed lines mark the mean change of PrismLoRA (green) and of the strongest tuned baseline (blue). (a, b) Three seeds, paired standard-error whiskers; the dSprites-Loc bar (+33.79) is cut at +16. (c) Seed 42.

CLEVR-Count +10.46, dSprites-Ori +7.30) and smaller on KITTI-Dist and DMLab (+1.73, +0.79). On language, it lands where $\gamma_d = 0.75$ (GSM8K +3.87, HellaSwag +2.36, ARC-e +1.30).

One block per distribution (zero-width windows) falls below grouping at matched parameters (72.02 against 83.89 on VTAB-6; Table 4), because its rank-11 blocks have a lower cap (Section 5.2) and SVHN collapses to 19.27.

PrismLoRA reduces to tuned LoRA when one gain suffices (Table 9). On Commonsense-15k and GSM8K after two epochs, even per-dataset learning rates chosen on test add at most 0.26 points. There, PrismLoRA selects $G = 1$ on validation and equals tuned LoRA on test (73.77; App. F.2).

6.3 Sources of the Improvement

We add the elements of PrismLoRA to tuned LoRA one at a time (Table 2). Grouping at one re-selected learning rate adds 1.89 points on VTAB-6 and 6.74 on Structured-8. Per-group learning rates add another 1.20 and 0.96 points. At the LoRA learning rate, the same VTAB-6 groups add only 0.27 points (Section 4.3).

LoRA with learning-rate-free optimizers falls short, as it keeps one learning rate for all distributions (Table 3). With Prodigy (Mishchenko & Defazio, 2024), it stays below tuned LoRA on all three benchmarks (78.22 against 80.80 on VTAB-6). With D-Adaptation (Defazio & Mishchenko, 2023), it collapses at 1 of 3 seeds on VTAB-6 and 2 of 3 on Structured-8.

Table 2: **Grouping and per-group learning rates give most of the improvement of PrismLoRA.** Second columns: change over the row above, ± paired standard error (three seeds).

Configuration	VTAB-6		Structured-8	
	Accuracy	Change	Accuracy	Change
LoRA at its own learning rate	80.80	—	50.16	—
+ grouping, one learning rate[†]	82.69	+1.89±0.06	56.90	+6.74±0.57
+ per-group learning rates	83.89	+1.20±0.15	57.86	+0.96±0.23
+ per-distribution head learning rate (PrismLoRA)	**84.42**	+0.52±0.12	**57.93**	+0.07±0.17
Grouping at LoRA's learning rate	81.07	+0.27±0.27	—	—

[†]On Structured-8 this row includes owned parameters for the three same-image task pairs.

Table 3: **An automatic global learning rate does not reach the improvement of PrismLoRA.** Prodigy and D-Adaptation: published learning rate 1.0. Default: 10^{-3} on vision, 3×10^{-4} on language. Language: Commonsense-15k and GSM8K, two epochs. D-Adaptation means include its collapsed seeds.

Method	VTAB-6	Structured-8	CS + GSM8K
LoRA, learning rate tuned on validation	80.80	50.16	**73.77**
LoRA, default learning rate	78.33	49.94	71.29
LoRA with Prodigy	78.22	49.76	72.94
LoRA with D-Adaptation	55.28	27.80	56.63
PrismLoRA (Ours)	**84.42**	**57.93**	**73.77**[†]

[†]= LoRA, $G = 1$ (the rule reduces PrismLoRA to LoRA here).

The per-distribution head learning rate adds only 0.52 points on VTAB-6 (all on CIFAR-100) and 0.07 on Structured-8 (paired standard error 0.17), and LoRA with it stays 2.73 and 6.85 points below PrismLoRA (Table 8).

7 Conclusion

We showed that each MoE-LoRA structure acts mainly as a hidden multiplier on the learning rate of LoRA (§4.1), and that distributions need different gains K on owned parameters (§4.2, §4.3). On these findings we built PrismLoRA, which outperforms the strongest tuned baseline on VTAB-6 and Structured-8. Practitioners should therefore compare MoE-LoRA methods at their own tuned learning rates.

MoE-LoRA serves label-scarce domains, and labels will stay scarce, because a model that cannot verify its outputs cannot make its own training data (Song et al., 2025; Huang et al., 2024). The gain with which each distribution's labels move the model will therefore keep mattering.

AI Use Statement

The authors used large language models (Claude, by Anthropic) to debug experiment code, summarize experimental results and revise the text of the paper. The authors led the research direction, checked all results and take full responsibility for the content.

Reproducibility Statement

The code is available at `https://anonymous.4open.science/r/PrismLoRA-182E`. Section 6.1 and Appendix F give the datasets, backbones, learning-rate grids and seeds. Appendix B gives the configuration procedure of PrismLoRA, and Appendix A gives the derivations.

REFERENCES

Ahmed Agiza, Marina Neseem, and Sherief Reda. MTLoRA: Low-rank adaptation approach for efficient multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16196–16205, 2024.

Laurence Aitchison. Bayesian filtering unifies adaptive and non-adaptive neural network optimization methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378, 2019.

Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. In *International Conference on Learning Representations (ICLR)*, 2019.

Kenneth J. Arrow, Theodore Harris, and Jacob Marschak. Optimal inventory policy. *Econometrica*, 19(3):250–272, 1951. doi: 10.2307/1906813.

Sungyong Baik, Myungsub Choi, Janghoon Choi, Heewon Kim, and Kyoung Mu Lee. Meta-learning with adaptive hyperparameters. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 20755–20765, 2020.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413, 2018.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Atılım Güneş Baydin, Robert Cornish, David Martínez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. In *International Conference on Learning Representations (ICLR)*, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, New York, second edition, 1985. doi: 10.1007/978-1-4757-4286-2.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569, 2018.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239.

Lucas Caccia, Edoardo Maria Ponti, Zhan Su, Matheus Pereira, Nicolas Le Roux, and Alessandro Sordoni. Multi-head adapter routing for cross-task generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. doi: 10.1023/A:1007379606734.

Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. Revisiting parameter-efficient tuning: Are we really there yet? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2612–2626, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.168.

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, and Jing Shao. Octavius: Mitigating task interference in MLLMs via LoRA-MoE. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 2924–2936, 2019. doi: 10.18653/v1/N19-1300.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*, 2020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2021.

João F. G. de Freitas, Mahesan Niranjan, and Andrew H. Gee. Hierarchical Bayesian models for regularization in sequential learning. *Neural Computation*, 12(4):933–953, 2000. doi: 10.1162/089976600300015655.

Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by D-adaptation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7449–7479, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 1932–1945, 2024. doi: 10.18653/v1/2024.acl-long.106.

Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A. Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. In *Proceedings of the International Conference on Medical Imaging with Deep Learning (MIDL)*, volume 250 of *Proceedings of Machine Learning Research*, pp. 406–425. PMLR, 2024.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973. doi: 10.1080/01621459.1973.10481350.

Chenghao Fan, Zhenyi Lu, Sichen Liu, Chengfeng Gu, Xiaoye Qu, Wei Wei, and Yu Cheng. Make LoRA great again: Boosting LoRA with adaptive singular values and mixture-of-experts optimization alignment. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15804–15832, 2025.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 27503–27516, 2021.

Walter D. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789–798, 1958. doi: 10.1080/01621459.1958.10501479.

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Ruibo Liu, Baochen Sun, Yawen Zhang, Daiyi Peng, Xiaoyuan Guo, and V. S. Subrahmanian. MoLA: MoE LoRA with layer-wise expert allocation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5112–5127, 2025. doi: 10.18653/v1/2025.findings-naacl.284.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016. Section 7.8, pp. 246–249, eqs. (7.44)–(7.45).

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806, 2024.

Haonan He, Jingqi Ye, Minglei Li, Zhengbo Wang, Tao Chen, Lei Bai, and Peng Ye. A unified study of LoRA variants: Taxonomy, review, codebase, and empirical evaluation. *arXiv preprint arXiv:2601.22708*, 2026.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 328–339, 2018. doi: 10.18653/v1/P18-1031.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. LLM-Adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5254–5276, 2023.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024.

Nikolaos Ignatiadis and Stefan Wager. Covariate-powered empirical Bayes estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 9617–9629, 2019.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pp. 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Maor Ivgi, Oliver Hinder, and Yair Carmon. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14465–14499, 2023.

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 (ECCV)*, pp. 709–727, 2022.

Shibo Jie and Zhi-Hong Deng. FacT: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 1060–1068, 2023. doi: 10.1609/aaai.v37i1.25187.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2018. doi: 10.1109/CVPR.2018.00781.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Frederik Kunstner, Jacques Chen, J. Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *International Conference on Learning Representations (ICLR)*, 2023.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 12169–12183, 2022.

Minh Le, Anh Nguyen, Huy Nguyen, Chau Nguyen, Anh Tuan Tran, and Nhat Ho. Revisit visual prompt tuning: The expressiveness of prompt experts. In *International Conference on Learning Representations (ICLR)*, 2026.

Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. In *International Conference on Learning Representations (ICLR)*, 2020.

Yu-Ang Lee, Ching-Yun Ko, Pin-Yu Chen, and Mi-Yen Yeh. Learning rate matters: Vanilla LoRA may suffice for LLM fine-tuning. *arXiv preprint arXiv:2602.04998*, 2026.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-SGD: Learning to learn quickly for few-shot learning. arXiv preprint arXiv:1707.09835, 2017.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 109–123, 2022.

Mengqi Liao, Wei Chen, Junfeng Shen, Shengnan Guo, and Huaiyu Wan. HMoRA: Making LLMs more effective with hierarchical mixture of LoRA experts. In *International Conference on Learning Representations (ICLR)*, 2025.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18878–18890, 2021.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. FAMO: Fast adaptive multitask optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 57226–57243, 2023.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. When MOE meets LLMs: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1104–1114, 2024. doi: 10.1145/3626772.3657722.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Zheda Mai, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Quang-Huy Nguyen, Li Zhang, and Wei-Lun Chao. Lessons and insights from a unifying study of parameter-efficient fine-tuning (PEFT) in visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14845–14857, 2025.

Itzik Malkiel and Lior Wolf. MTAdam: Automatic balancing of multiple training loss terms. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10713–10729, 2021. doi: 10.18653/v1/2021.emnlp-main.837.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 7697–7711, 2022.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2381–2391, 2018. doi: 10.18653/v1/D18-1260.

Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35779–35804, 2024.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations (ICLR)*, 2021.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446, 2022.

Yann Ollivier. Online natural gradient as a Kalman filter. *Electronic Journal of Statistics*, 12(2), 2018. doi: 10.1214/18-EJS1468.

Edoardo Maria Ponti, Alessandro Sordoni, Yoshua Bengio, and Siva Reddy. Combining parameter-efficient modules for task-level generalisation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 687–702, 2023.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. ImageNet-21K pretraining for the masses. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Amelie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. Scalarization for multi-task and multi-domain learning at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 16917–16941, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. Multimodal instruction tuning with conditional mixture of LoRA. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 637–648, 2024. doi: 10.18653/v1/2024.acl-long.38.

Sharad Singhal and Lance Wu. Training multilayer perceptrons with the extended Kalman algorithm. In D. Touretzky (ed.), *Advances in Neural Information Processing Systems (NIPS)*, volume 1. Morgan-Kaufmann, 1988.

Jonas Sjöberg and Lennart Ljung. Overtraining, regularization and searching for a minimum, with application to neural networks. *International Journal of Control*, 62(6):1391–1407, 1995. doi: 10.1080/00207179508921605.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham M. Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132, 2020.

Mengyang Sun, Yihao Wang, Tao Feng, Dan Zhang, Yifan Zhu, and Jie Tang. A stronger mixture of low-rank experts for fine-tuning foundation models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57712–57727, 2025.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. HydraLoRA: An asymmetric LoRA architecture for efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 9565–9584, 2024. doi: 10.52202/079017-0304.

Zichen Tian, Antoine Ledent, and Qianru Sun. Scalable multi-task low-rank model adaptation. In *International Conference on Learning Representations (ICLR)*, 2026.

Donald M. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26(2): 305–321, 1978. doi: 10.1287/opre.26.2.305.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Haowen Wang, Tao Sun, Congyun Jin, Yingbo Wang, Yibo Fan, Yunqi Xu, Yuliang Du, and Cong Fan. Customizable combination of parameter-efficient modules for multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Huiyi Wang, Daijiao Liu, Lina Yao, and Dong Gong. One rate is not enough: Adaptive anisotropic learning rates for LoRA fine-tuning. *arXiv preprint arXiv:2609.05885*, 2026.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, 2017. doi: 10.1109/CVPR.2017.369.

Xujia Wang, Haiyan Zhao, Shuo Wang, Hanqing Wang, and Zhiyuan Liu. MALoRA: Mixture of asymmetric low-rank adaptation for enhanced multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5624–5641, 2025. doi: 10.18653/v1/2025.findings-naacl.312.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *International Conference on Learning Representations (ICLR)*, 2024.

Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54488–54510, 2024.

Derrick Xin, Behrooz Ghorbani, Justin Gilmer, Ankush Garg, and Orhan Firat. Do current multi-task optimization methods in deep learning even help? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 13597–13609, 2022.

Yi Xin, Siqi Luo, Xuyang Liu, Yuntao Du, Haodi Zhou, Xinyu Cheng, Christina Lee, Junlong Du, Haozhe Wang, Mingcai Chen, Ting Liu, Guimin Hu, Zhongwei Wan, Rongchao Zhang, Aoxue Li, Mingyang Yi, and Xiaohong Liu. V-PETL bench: A unified visual parameter-efficient transfer learning benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 80522–80535, 2024.

Enneng Yang, Junwei Pan, Ximei Wang, Haibin Yu, Li Shen, Xihua Chen, Lei Xiao, Jie Jiang, and Guibing Guo. AdaTask: A task-aware adaptive learning rate approach to multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10745–10753, 2023. doi: 10.1609/aaai.v37i9.26275.

Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, Weizhu Chen, and Yunhai Tong. MTL-LoRA: Low-rank adaptation for multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22010–22018, 2025. doi: 10.1609/aaai.v39i20.35509.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient MoE for instruction tuning. In *International Conference on Learning Representations (ICLR)*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4791–4800, 2019. doi: 10.18653/v1/P19-1472.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Dacao Zhang, Kun Zhang, Shimao Chu, Le Wu, Xin Li, and Si Wei. MoRE: A mixture of low-rank experts for adaptive multi-task learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1311–1324, 2025a. doi: 10.18653/v1/2025.findings-acl.68.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:5268–5280, 2025b. doi: 10.1109/TPAMI.2024.3435939.

Ziyu Zhao, Yixiao Zhou, Xin Yu, Zhi Zhang, Didi Zhu, Tao Shen, Zexi Li, Jinluan Yang, Xuwu Wang, Jing Su, Kun Kuang, Zhongyu Wei, Fei Wu, and Yu Cheng. Each rank could be an expert: Single-ranked mixture of experts LoRA for multi-task learning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Volume 1*, pp. 1998–2009, 2026. doi: 10.1145/3770854.3780222.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 55006–55021, 2023.