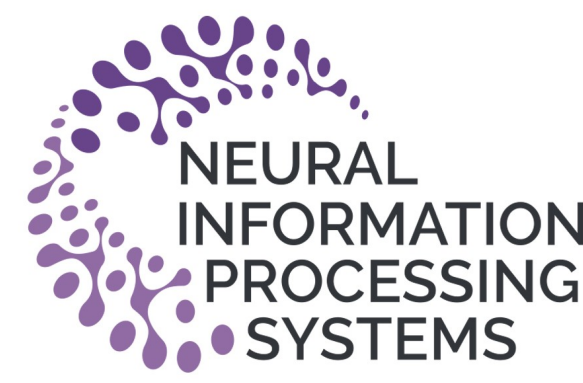


Improving Quantized Zeroth-Order Optimization through Reconstructed Low-Rank Structures

Fei Wang^{1,2*} Shuai Xie^{2*} Li Shen^{3,4} Chao Xue² Ye Liu¹ Changxing Ding^{1†}

¹ South China University of Technology ² JD.com ³ Shenzhen Campus of Sun Yat-sen University ⁴ Center for AI Theoretical Foundation and Systems, Shenzhen Loop Area Institute

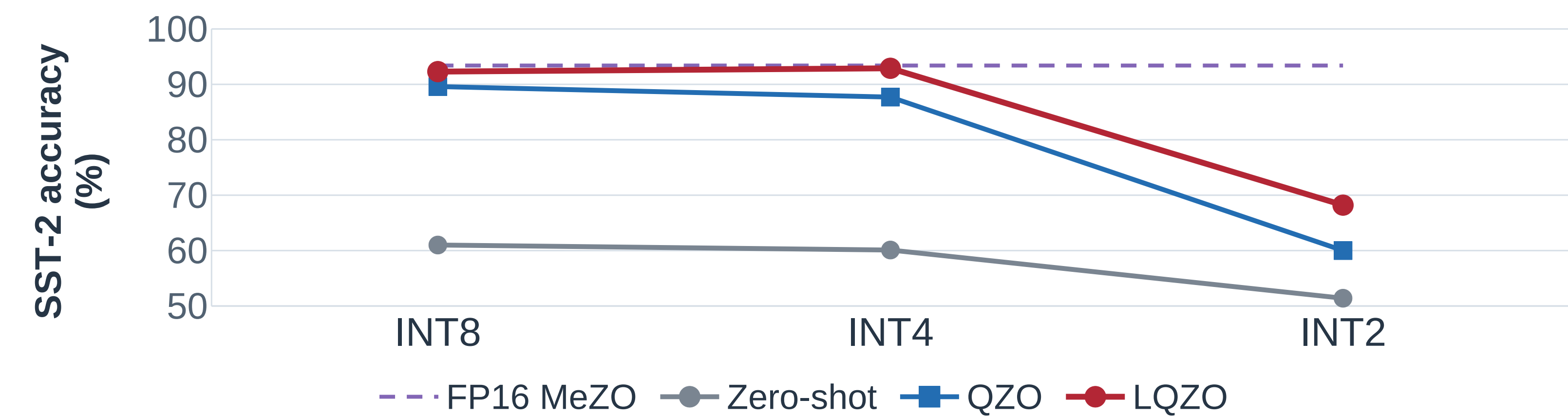


Abstract and main results

Quantized zeroth-order optimization fine-tunes large language models with low memory use, but performance deteriorates at low bit widths. Motivated by concentrated gradient spectra, LQZO combines low-rank perturbations, Binary-Aggregated Gaussian sampling, and scale reshaping through Isotropy-Enhanced Estimation. Across INT8, INT4, and INT2, it improves the 11-task average over QZO by 2.0% to 3.3%. LQZO improves INT2 trainability while retaining a gap to full precision. Experiments on quantized Vision Transformers show gains beyond language models.

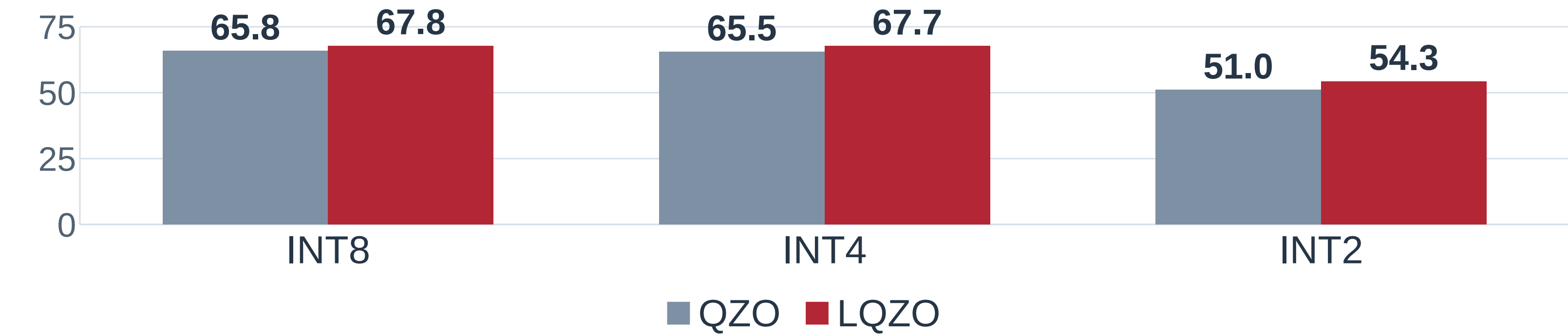
Low-rank motivation

Gradient spectra motivate structured exploration, while concentration varies across modules. Each perturbation has low rank, but refreshed factors let accumulated updates explore beyond one subspace.



SST-2 gains over QZO are 2.7%, 5.2%, and 8.2% at INT8, INT4, and INT2.

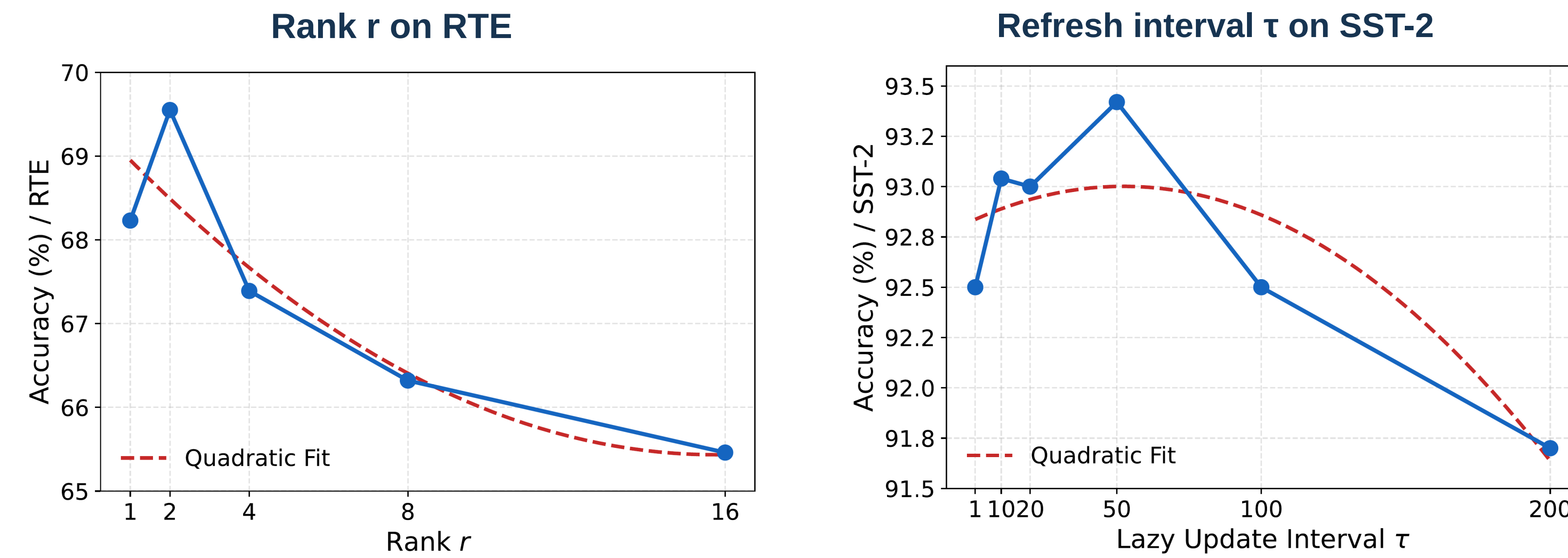
11-task average on OPT-6.7B



+2.0% to +3.3% over QZO

Means over three seeds. FP16 MeZO scores 68.5. INT2 zero-shot scores 37.6 versus 54.3 with LQZO. Scores use a 0–100 scale and gains are absolute differences.

Ablations on OPT-6.7B

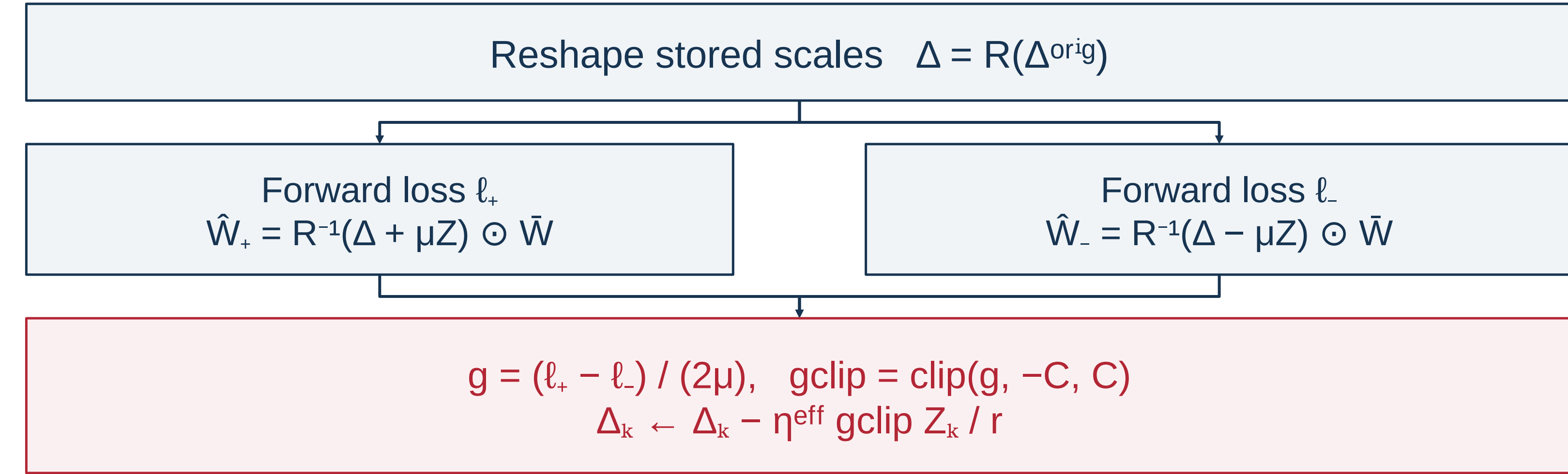


Practical defaults $r = 2, \tau = 50, \gamma = 8, \mu = 10^{-3}, C = 100$. Rank and interval ablations use RTE and SST-2, respectively.

Three designs in LQZO

Optimize continuous group scales with fixed integer coefficients. Both forward losses use the same mini-batch and perturbation Z .

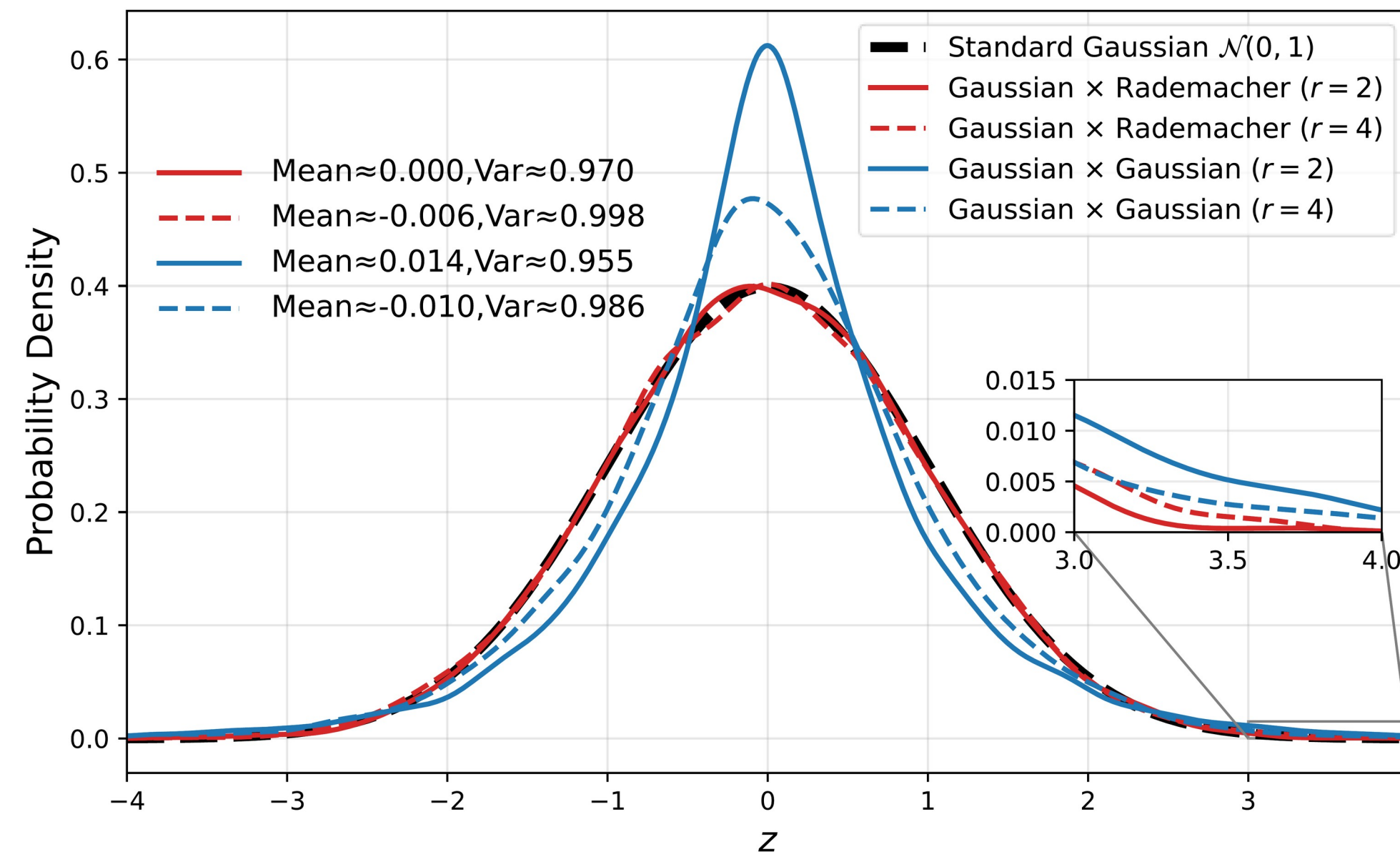
1 Low-rank perturbations $Z_k = U_k V_k^T$



Two forward passes and no backpropagation. Map scales back with R^{-1} . The applied learning rate is $\eta^{\text{eff}} = 2^{4-b}\eta$ for b -bit weights.

2 Binary-Aggregated Gaussian Perturbation

$Z_k = U_k V_k^T$. Draw standard Gaussian U_k each step and Rademacher $V_k \in \{-1, +1\}$ every τ steps. This reduces higher moments without extra loss evaluations.



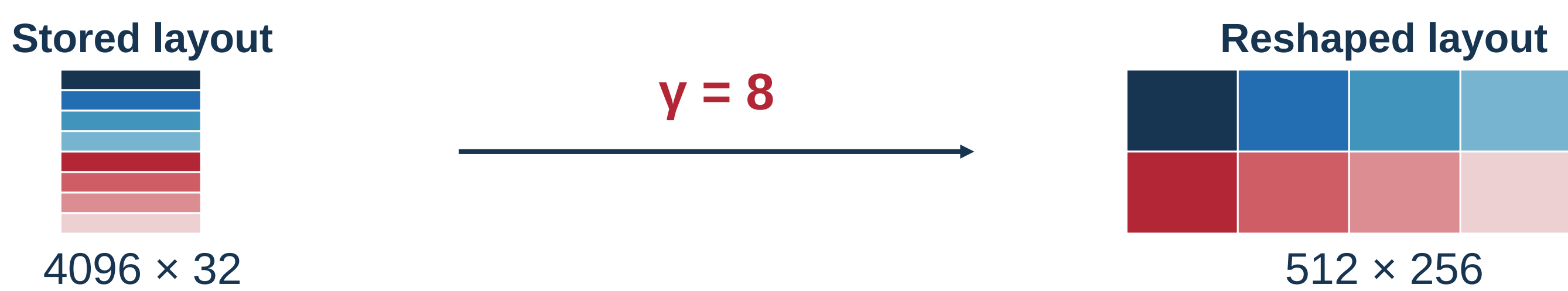
10,000 perturbation samples. Binary factors reduce tail mass relative to two Gaussian factors.

QZO	LQZO Gaussian V	LQZO Rademacher V
67.4	70.5	71.3

Six-task average on OPT-6.7B. Rademacher V adds 0.8% over Gaussian V, improving five of six tasks.

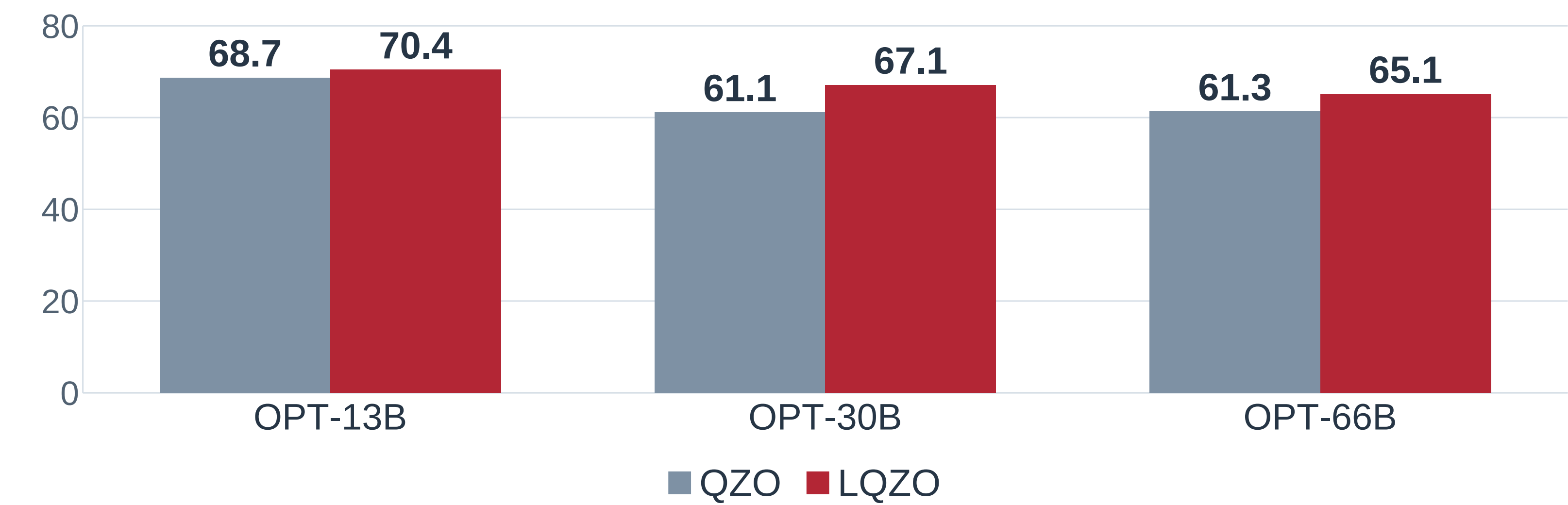
3 Isotropy-Enhanced Estimation

In this example, all 131,072 scales remain trainable. In the 512×256 layout, a rank-2 factor pair contains 1,536 entries, compared with 131,072 for a dense perturbation.



Generalization and efficiency

Gains extend to 66B parameters



INT4 six-task means over three seeds on SST-2, RTE, BoolQ, WiC, WSC, and SQuAD. Gains over QZO are 1.7%, 6.0%, and 3.8%.

Beyond OPT

INT4 evaluation	QZO	LQZO	Gain
LLaMA-2-7B 11-task average	69.8	71.4	+1.6%
LLaMA-3.1-8B 11-task average	70.3	74.1	+3.8%
Qwen3-4B-Base Five-task average	83.44	84.59	+1.15%
ViT-base CIFAR-100	37.8	51.2	+13.4%

Qwen3 improves on four of five tasks, with SST-2 close to QZO. On RTN-quantized ViT, the three-task average rises from 65.9 to 73.7. GPTQ groups 32, 64, and 128 give SST-2 gains of 1.30% to 5.24% on OPT-6.7B.

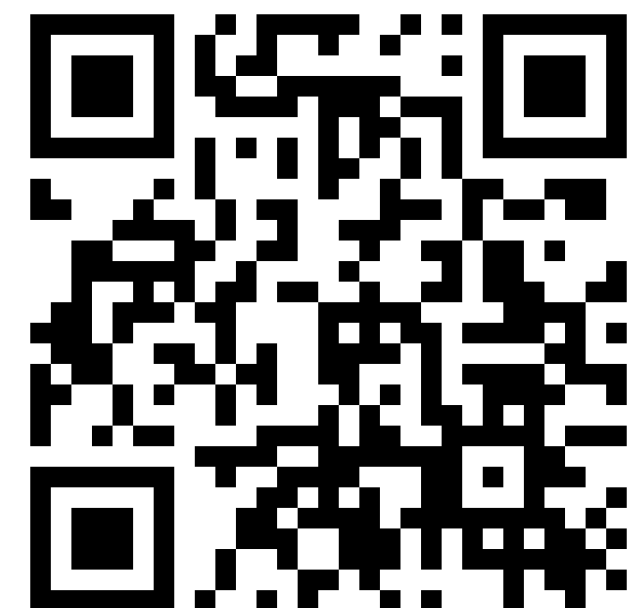
Small memory and time overhead

About 1 MB extra peak memory vs QZO

GPU time vs QZO +5.1% (6.7B), +5.0% (13B)

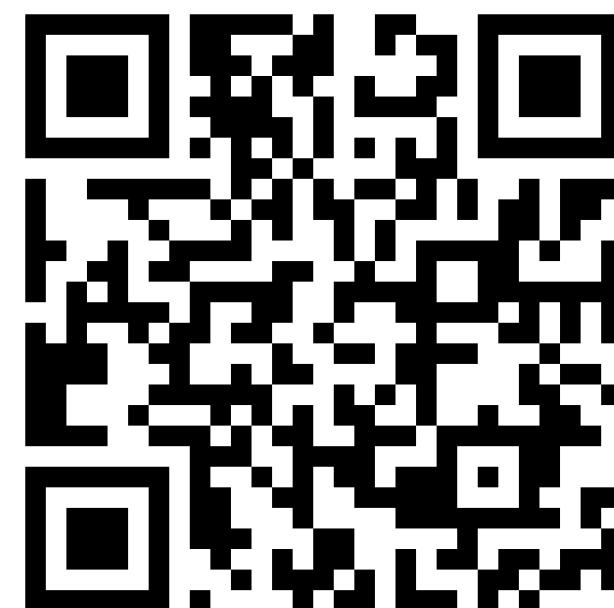
INT4 on one A100-40GB, sequence length 128, batch size 16. Mean GPU time over 100 steps after 10 warmup steps. Excludes data loading and host preprocessing.

Paper



[OpenReview](#)

Code



[GitHub / UnifiedZO](#)

* Equal contribution. This work was done while Fei Wang was at JD Explore Academy.

† Corresponding author. wang_fei2019@163.com chxding@scut.edu.cn

The convergence analysis includes finite-difference bias for the independently refreshed, unclipped base estimator. Lazy updates and scalar clipping are evaluated empirically.