CORA: Per-Slice Coherent Orthogonal Rotation for SVD-based Low-Rank Adaptation

Pengcheng Wang[1] Ziran Liu[2,3] Wei Wang[4] Wei Jiang[4]

[1]Purdue University [2]Shanghai Institute for Mathematics and Interdisciplinary Sciences
[3]Research Institute of Intelligent Complex Systems, Fudan University [4]Futurewei Technologies

NeurIPS 2026

Fine-tuning rotates both singular bases together

LoRA learns a low-rank update $\Delta W = BA$ on a frozen $W_0 = U_0 \Sigma_0 V_0^\top$. It leaves open which degrees of freedom of the weight the update should use: the left basis U_0, the right basis V_0 or the spectrum Σ_0.

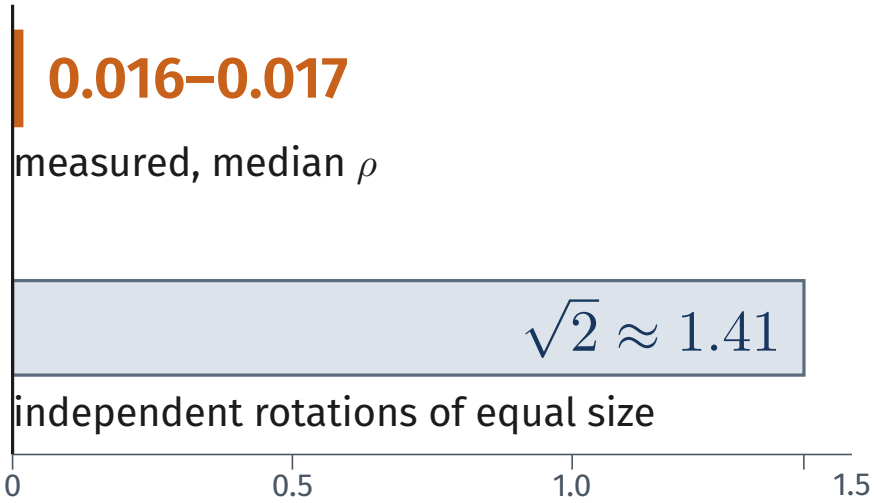
Our companion theory paper (Liu et al., 2026, arXiv:2605.02111) shows that a fully fine-tuned weight keeps the base singular frames up to a coherent rotation: nearly the same orthogonal Q on the left and on the right.

CORA is the adapter with exactly this form, one Q on both bases and a diagonal $\Delta\Sigma$ on the spectrum:

$$\widetilde{W} = U_0\, Q\, (\Sigma_0 + \Delta\Sigma)\, Q^\top V_0^\top$$

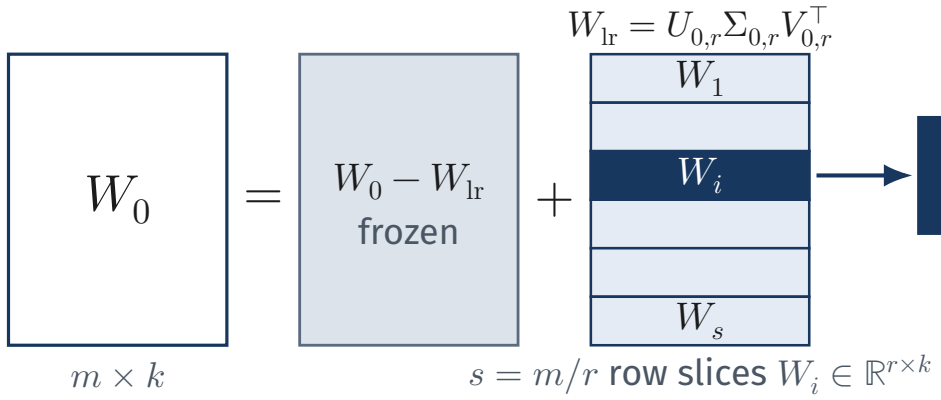
Measured on full fine-tuning

Mismatch between the left and right frame motions (normalized discrepancy ρ), 60,480 row slices of LLaMA-3.2-3B and five fully fine-tuned models

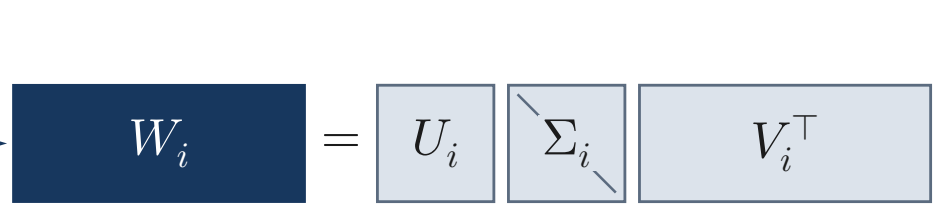


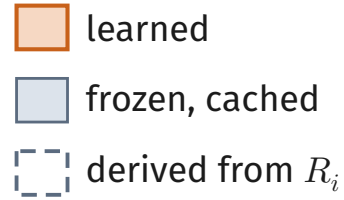
CORA: one coherent rotation per row slice

1 Slice the rank-r part

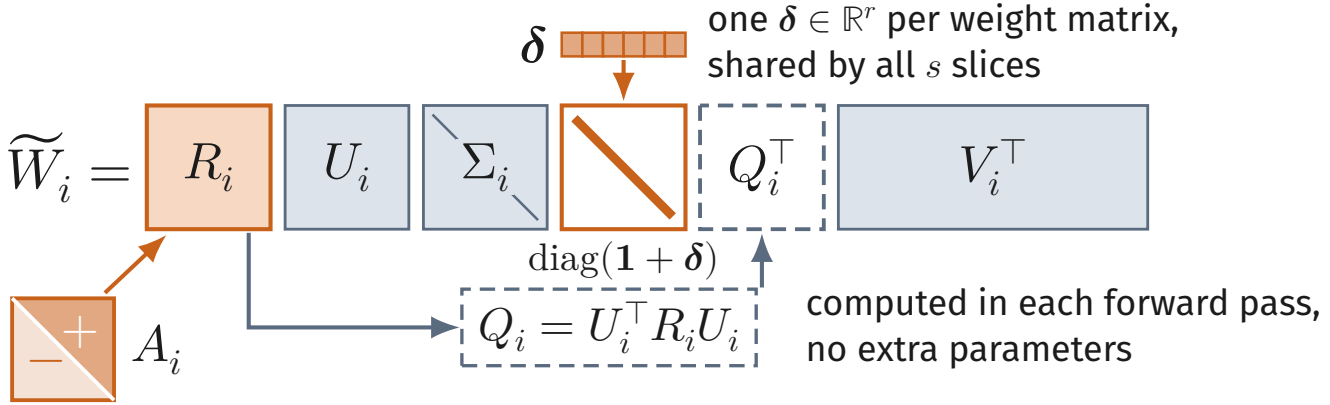


2 Cache the SVD of each slice (offline)



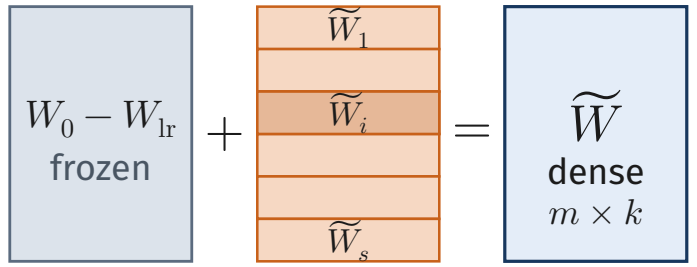


3 Learn R_i on the left, derive Q_i for the right



A_i skew-symmetric: $r(r-1)/2$ learned entries
Cayley map: $R_i = (I - A_i)(I + A_i)^{-1}$

4 Merge into one dense weight



$$\widetilde{W} = (W_0 - W_{\text{lr}}) + [\,\widetilde{W}_1^\top \cdots \widetilde{W}_s^\top\,]^\top$$

$R_i = I,\ \boldsymbol{\delta} = \mathbf{0} \Rightarrow \widetilde{W} = W_0$

$m(r-1)/2 + r$ trainable parameters per weight matrix
no adapter computation at inference

Coherent by construction, small to train, free at inference

Adapted row slice

$$\widetilde{W}_i = \underbrace{U_i Q_i}_{\text{left frame}} \; \Sigma_i \; \underbrace{\operatorname{diag}(\mathbf{1}+\boldsymbol{\delta})}_{\text{shared scale}} \; \underbrace{Q_i^\top V_i^\top}_{\text{right frame}}$$

Since $R_i U_i = U_i Q_i$, every trained slice $R_i U_i \Sigma_i \operatorname{diag}(\mathbf{1}+\boldsymbol{\delta}) Q_i^\top V_i^\top$ has this form: the same $Q_i \in \mathrm{O}(r)$ rotates both singular bases.

- The Cayley map $R_i = (I - A_i)(I + A_i)^{-1}$ keeps every R_i orthogonal, with no penalty term.
- At $A_i = 0$ and $\boldsymbol{\delta} = \mathbf{0}$ the adapted weight is exactly W_0.
- After training the slices merge into one dense $m \times k$ weight: no adapter computation remains at inference.

Trainable parameters per $m \times k$ weight

$$|\theta^{(W)}_{\mathrm{CORA}}| = \tfrac{1}{2}\, m\,(r-1) + r$$

$$\frac{|\theta^{(W)}_{\mathrm{CORA}}|}{|\theta^{(W)}_{\mathrm{LoRA}}|} = \frac{m(r-1)}{2r(m+k)} + \frac{1}{m+k}$$

The count does not depend on k. On square projections it is about one quarter of LoRA's at the same r.

LLaMA-2-7B: CORA at $r = 16$ trains 6.6 M parameters, LoRA at $r = 32$ trains 56 M. Parameter, gradient and optimizer-state storage shrink by about 8.5×.

Commonsense reasoning: higher accuracy, fewer parameters

Method	LLaMA-2-7B		LLaMA-3-8B	
	Params	Average	Params	Average
LoRA, $r = 32$	56 M	77.6	63 M	80.8
DoRA, $r = 32$	57 M	79.7	63 M	85.2
PiSSA, $r = 32$	56 M	73.8	63 M	75.4
MiLoRA, $r = 32$	56 M	79.2	63 M	81.9
CORA, $r = 128$ (full W_0)	55.7 M	**82.0 ± 0.7**	87.4 M	**86.9 ± 0.2**
CORA, $r = 64$	27.6 M	81.6 ± 1.4	43.3 M	86.5 ± 0.3
CORA, $r = 32$	13.6 M	81.1 ± 1.7	21.3 M	86.7 ± 0.2
CORA, $r = 16$	6.6 M	81.8 ± 0.6	10.3 M	86.5 ± 0.3
CORA, $r = 8$	–	–	4.8 M	85.2 ± 0.7

Eight-task average on commonsense_170k; best average in bold. CORA: mean ± sample s.d. over three runs; baselines from the DoRA and MiLoRA papers. $r = 128$ (full W_0) slices the full W_0 with height 128; LLaMA-2-7B at $r = 8$ was run once.

CORA at $r = 16$

6.6 M parameters on LLaMA-2-7B: 81.8

10.3 M parameters on LLaMA-3-8B: 86.5

Above LoRA, DoRA, PiSSA and MiLoRA on both backbones, which train 56–63 M parameters.

On LLaMA-2-13B, CORA reaches 84.4 with 21.3 M parameters ($r = 32$) and 84.0 with 10.3 M ($r = 16$).

Code generation and mathematical reasoning

HumanEval Pass@1, CodeAlpaca training

	LLaMA-3-8B		Mistral-7B	
Config	Params	Pass@1	Params	Pass@1
LoRA r=32 (ref)	84 M	34.7	84 M	33.8
DoRA r=32 (ref)	85 M	33.1	85 M	33.7
LoRI-D r=32 (ref)	44 M	43.2	44 M	33.8
DiaBlo N=128 (ref)	61 M	39.4	61 M	34.0
DiaBlo N=64 (ref)	121 M	43.2	122 M	34.4
CORA, $r = 64$	43.3 M	47.0	43.3 M	**40.9**
CORA, $r = 32$	21.3 M	45.1	21.3 M	**40.9**
CORA, $r = 16$	10.3 M	**48.2**	10.3 M	40.2
CORA, $r = 8$	4.8 M	45.1	4.8 M	40.2

(ref): published results under their own implementations; LoRA, DoRA, LoRI-D from LoRI; DiaBlo with N diagonal blocks. CORA single runs. Best in bold.

GSM8K, MetaMathQA training

Backbone	Method	Params	GSM8K
LLaMA-2-7B	LoRA, $r = 64$	113 M	60.6
	CORA, $r = 128$ (full W_0)	55.7 M	**62.4**
	CORA, $r = 64$	27.6 M	62.0
Mistral-7B	LoRA (Gaussian)		69.50 ± 0.42
	CORA, $r = 128$	87.4 M	**70.0**
Gemma-7B	LoRA (Gaussian)		**75.11 ± 0.64**
	CORA, $r = 128$	59.6 M	**75.1**

Best per backbone in bold; on Gemma-7B the two agree to one decimal. LLaMA-2-7B baseline as in the MiLoRA and DiaBlo papers; Mistral-7B and Gemma-7B: Gaussian-initialized LoRA from the PiSSA paper (parameter count not reported), three-run mean ± s.d.; CORA two-run means there, single runs on LLaMA-2-7B.

Takeaways

- Fully fine-tuned weights keep their singular frames up to one coherent rotation. CORA builds the adapter on this property: one learned rotation per row slice, its right-side partner computed from it, and one shared scale vector per weight matrix.
- It trains $m(r-1)/2 + r$ parameters per weight matrix, about one quarter of LoRA's count on square projections, and merges into a dense weight for inference.
- Across repeated commonsense runs it is above LoRA, DoRA, PiSSA and MiLoRA with roughly one sixth to one eighth of their parameters. On HumanEval it is above every listed baseline on LLaMA-3-8B and Mistral-7B, and on GSM8K it matches or exceeds LoRA on all three backbones.
- Rank 16 or 32 is a good start for commonsense and code; math gains from a higher rank at a fixed budget.

Paper 31441. Theory: Liu et al., 2026, arXiv:2605.02111.