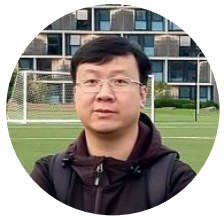
Mixture-of-Control: State-Aware Fine-Tuning for Transformer-based Models

Duc-Anh Nguyen[1,2,*], Tien Ngoc Luu[1,*], Tung Pham[1], Toan Tran[1]

[1] Qualcomm AI Research, [2] University of Virginia, * Equal Contribution



NeurIPS 2026

Fine-tuning today: memory-efficient OR cross-layer communication

Method	Memory-Efficient	Cross-Layer Communication
LoRA/ DoRA Weight-based FT	✗	✗
Control State-based FT Zhang et al., ICML 2025	✓	✗
MoLEx Weight-based FT, comm. Teo & Nguyen, ICLR 2025	✗	✓
MoC (Ours) State-based FT & comm.	✓	✓

Why memory?

−50%

DoRA rank → half the parameters

−0.26 GB

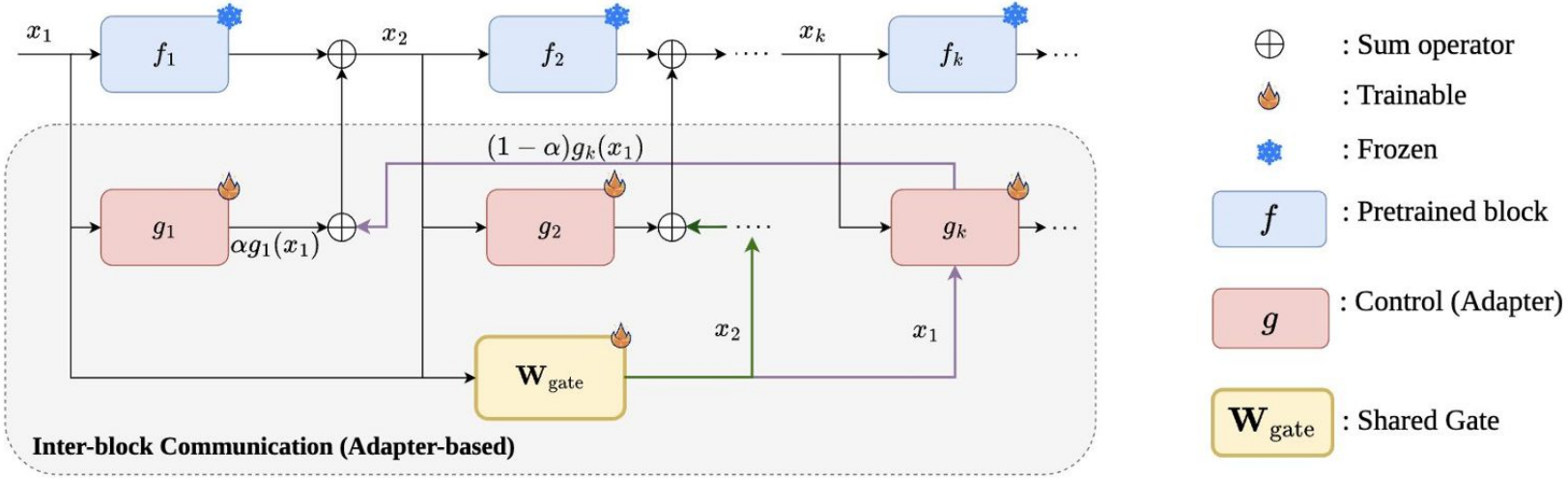
GPU memory saved, out of ~69 GB (8B model)

Activations, not parameters, dominate memory.

Key Messages

- Control tunes blocks, not weights, so it saves memory, but each block's adapter only sees its own layer.
- Current communication-based methods are computational expensive.

Methodology



Mixture-of-Control (MoC) Forward Pass

$$x_{t+1} = x_t + f_t(W_t x_t) + \alpha g_t(x_t) + (1-\alpha)\sum_{k=1}^{L} \pi_k(x_t) g_k(x_t),$$

where

$$\pi_k(x_t) = \text{softmax}(\text{TopK}(\mathbf{W}_{\text{gate}} x_t))_k$$

Experimental Results

Table 1: Comparison of algorithm performance across commonsense reasoning benchmarks. * marks previously published numbers. Best results are in **bold**.

Model	Method	% Params	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Avg (↑)	Time (↓) (h)	GPU Mem. (↓) (GB)
LLaMA2-7B	LoRA*	0.83	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6	6.1	48.0
	DoRA*	0.85	71.8↑2.0	**83.7**↑3.8	76.0↓3.5	89.1↑5.5	82.6↑0.0	**83.7**↑3.9	68.2↑3.5	**82.4**↑1.4	79.7↑2.1	10.5	80.2
	MoLEx	0.85	69.7↓0.1	81.9↑2.0	78.4↓1.1	91.3↑7.7	81.5↓1.1	81.5↑1.7	65.8↑1.1	80.8↓0.2	78.9↑1.3	9.8	77.4
	FacT_TK	0.51	69.8↑0.0	80.6↑0.7	77.8↓1.7	90.8↑7.2	81.8↓0.8	79.7↓0.1	65.4↑0.7	78.6↓2.4	78.1↑0.5	6.0	48.5
	StructLoRA	0.83	70.9↑1.1	82.0↑2.1	78.6↓0.9	91.6↑8.0	82.1↓0.5	81.6↑1.8	66.6↑1.9	80.6↓0.4	79.3↑1.7	8.5	71.6
	FlexLoRA	0.83	69.5↓0.3	81.2↑1.3	78.1↓1.4	92.9↑9.3	78.7↓3.9	82.8↑3.0	66.5↑1.8	77.0↓4.0	78.3↑0.7	7.0	54.8
	Control*	0.50	**72.3**↑2.5	82.5↑2.6	79.2↓0.3	89.1↑5.5	83.1↑0.5	83.0↑3.2	68.5↑3.8	79.0↓2.0	79.6↑2.0	4.2	42.6
	MoC	0.51	71.8↑2.0	83.2↑3.3	**79.6**↑0.1	**93.3**↑9.7	**84.2**↑1.6	83.6↑3.8	**68.6**↑3.9	80.4↓0.6	**80.5**↑2.9	5.5	47.6
LLaMA3-8B	LoRA*	0.71	70.8	85.2	79.9	91.7	84.3	84.2	71.2	79.0	80.8	7.5	59.1
	DoRA*	0.71	74.6↑3.8	89.3↑4.1	79.9↑0.0	95.5↑3.8	85.6↑1.3	90.5↑6.3	80.4↑9.2	85.8↑6.8	85.2↑4.4	10.5	69.1
	MoLEx	0.72	70.1↓0.7	86.2↑1.0	80.4↑0.5	93.2↑1.5	84.1↓0.2	84.8↑0.6	71.3↑0.1	82.2↑3.2	81.5↑0.7	11.0	94.9
	FacT_TK	0.43	70.7↓0.1	86.0↑0.8	79.5↓0.4	94.2↑2.5	84.2↓0.1	86.5↑2.3	73.3↑2.1	81.8↑2.8	82.0↑1.2	7.7	57.6
	StructLoRA	0.71	70.9↑0.1	86.6↑1.4	79.5↓0.4	93.9↑2.2	84.0↓0.3	87.0↑2.8	73.9↑2.7	82.6↑3.6	82.3↑1.5	9.8	75.4
	FlexLoRA	0.71	73.4↑2.6	88.5↑3.3	78.6↓1.3	95.3↑3.6	84.4↑0.1	90.2↑6.0	79.2↑8.0	85.6↑6.6	84.4↑3.6	8.6	62.4
	Control*	0.42	74.1↑3.3	87.8↑2.6	80.7↑0.8	95.5↑3.8	86.0↑1.7	90.8↑6.6	80.0↑8.8	87.8↑8.8	85.3↑4.5	4.4	54.0
	MoC	0.43	**75.5**↑4.7	**89.5**↑4.3	**82.0**↑2.1	**96.6**↑4.9	**88.8**↑4.5	**91.5**↑7.3	**81.6**↑10.4	**88.0**↑9.0	**86.7**↑5.9	6.5	56.7
Mistralv3-7B	LoRA	0.78	68.7	83.5	77.2	90.3	82.5	85.7	73.4	80.5	80.2	6.5	58.4
	DoRA	0.79	70.3↑1.5	84.5↑1.0	77.8↑0.6	89.9↓0.4	83.1↑0.5	87.1↑1.4	72.2↓1.2	82.4↑1.9	80.9↑0.7	10.1	70.4
	MoLEx	0.80	71.4↑2.7	85.4↑2.0	78.2↑1.0	90.3↓0.0	83.5↑1.0	84.1↓1.6	75.0↑1.5	82.7↑2.2	81.3↑1.1	10.5	82.3
	FacT_TK	0.48	70.5↑1.8	86.4↑2.9	80.3↑3.1	94.1↑3.8	84.2↑1.7	83.9↓1.8	71.0↓2.4	84.4↑3.9	81.9↑1.7	6.3	56.0
	StructLoRA	0.78	70.0↑1.3	83.2↓0.3	77.4↑0.2	89.9↓0.4	81.7↓0.8	87.2↑1.5	74.8↑1.4	81.4↑0.9	80.7↑0.5	9.2	70.1
	FlexLoRA	0.78	73.9↑5.2	88.4↑4.9	79.2↑2.0	**95.9**↑5.6	86.0↑3.5	**90.2**↑4.5	78.8↑5.4	86.8↑6.3	84.9↑4.7	7.5	61.3
	Control	0.46	70.6↑1.8	86.9↑3.5	79.0↑1.8	94.2↑3.9	84.8↑2.2	87.9↑2.2	76.5↑3.0	84.8↑4.3	83.1↑2.9	4.3	53.2
	MoC	0.48	**74.2**↑5.5	**88.9**↑5.4	**81.0**↑3.8	95.6↑5.3	**87.9**↑5.4	88.9↑3.2	**79.5**↑6.1	**87.8**↑7.3	**85.5**↑5.3	5.5	55.6
Qwen2.5-14B	LoRA	0.63	73.4	90.4	82.2	96.0	88.6	94.7	86.1	90.6	87.7	9.4	62.1
	DoRA	0.64	75.1↑1.7	90.1↓0.3	82.0↓0.2	96.3↑0.3	89.4↑0.8	95.0↑0.3	88.2↑2.1	93.0↑2.4	88.6↑0.9	14.3	94.7
	MoLEx	0.64	73.5↑0.1	90.5↑0.1	81.9↓0.3	95.5↓0.5	88.2↓0.4	94.4↓0.3	85.8↓0.3	92.2↑1.6	87.8↑0.1	20.3	124.8
	FacT_TK	0.43	74.7↑1.3	90.1↓0.3	82.3↑0.1	96.2↑0.2	89.6↑1.0	94.7↑0.0	88.1↑2.0	92.4↑1.8	88.5↑0.8	8.0	61.5
	StructLoRA	0.63	74.8↑1.4	89.5↓0.9	81.0↓1.2	95.8↓0.2	89.3↑0.7	94.7↑0.0	87.4↑1.3	92.4↑1.8	88.1↑0.4	13.5	77.3
	FlexLoRA	0.63	74.2↑0.8	91.4↑1.0	82.0↓0.2	96.4↑0.4	87.7↓0.9	**96.7**↑2.0	90.0↑3.9	95.4↑4.8	89.2↑1.5	10.5	67.4
	Control	0.43	72.8↓0.6	88.8↓1.6	81.3↓0.9	94.6↓1.4	87.5↓1.1	94.4↓0.3	86.9↑0.8	91.2↑0.6	87.2↓0.5	6.5	58.7
	MoC	0.43	**75.5**↑2.1	**91.8**↑1.4	**82.4**↑0.2	**96.8**↑0.8	**91.0**↑2.4	96.6↑1.9	**90.2**↑4.1	**96.4**↑5.8	**90.1**↑2.4	8.0	61.3

MoC may cut training memory up to **50%** while outperforming state-based methods.

Conclusion & Future Directions

- We identify the **key limitations** of prior adaptation works: state-based FT (no communication) & communication overhead in weight-based FT.
- We propose **Mixture-of-Control (MoC)**, an efficient cross-layer communication at control-level memory efficiency, improving adaptation quality with minimal additional cost.
- Shows improvement on diverse models & benchmarks, supported by both theoretically & empirically (10 pretrained models + 33 datasets).

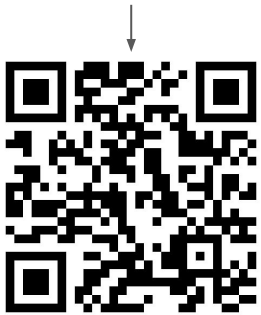
Future Directions

- Inference efficiency bottleneck.
- Adaptive Learning Schemes for the Hyperparameters.

October 2026

THANK YOU FOR YOUR ATTENTION!

Enjoy our paper in



Questions?

Duc-Anh Nguyen (https://duc-anh-2002.github.io/)

University of Virginia