Dual-Pronged LoRA: Achieving Near-Zero Forgetting and High-Performance Adaptation for MLLMs

Jingqi Ye[1,2], Haonan He[1,2], Minglei Li[1,3], Fujun Han[1,4], Tao Chen[3,6], Peng Ye[1,5†]

[1] Shanghai Artificial Intelligence Laboratory, [2] University of Science and Technology of China, [3] Fudan University

[4] The Chinese University of Hong Kong, Shenzhen, [5] The Chinese University of Hong Kong, [6] Shanghai Innovation Institute

Motivation & Insight

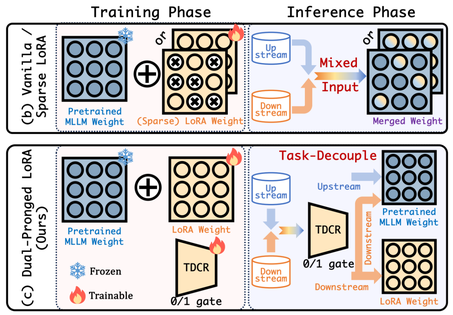
1. Catastrophic forgetting in MLLMs

MLLMs need fine-tuning for downstream tasks. However, downstream adaptation often degrade pretrained capabilities, causing **catastrophic forgetting**.

2. Pruning trades downstream adaptation for upstream retention

Full fine-tuning methods prune or mask weights and still incur high tuning costs.
LoRA methods reduces tuning cost, but prune also limits downstream fit.

3. From parameter entanglement to task decoupling

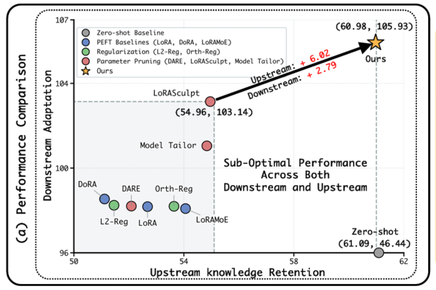


Parameter space entanglement forces a retention–adaptation trade-off. If we **decouple** the tasks, we can **retain upstream knowledge** and **maximize downstream performance**.

Q1. How can we learn to route inputs using only downstream data?

Q2. How can we allocate ranks to maximize downstream performance at a fixed budget?

4. A new retention–adaptation Pareto frontier



Rank-16-equivalent budget

Dual-Pronged LoRA achieves near-zero forgetting while attaining the best overall performance.

Mean Source
60.98 vs. 61.09 zero-shot+6.02 over LoRASculpt (54.96)

Mean Target
105.93 vs. 103.14 for LoRASculpt (+2.79)

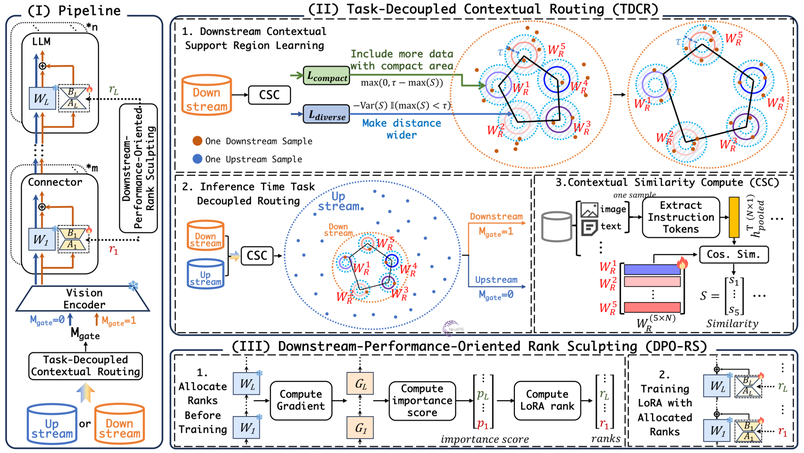
Wechat (Contact)

Paper (PDF)

Design & Methodology

Two components with complementary roles

TDCR decides when to activate LoRA. DPO-RS decides where to allocate its rank.



1. TDCR: Task-Decoupled Contextual Routing

- **One-class support learning.** Fit a compact region with diverse prototypes using only downstream data.
- **Inference-time decoupling.** Activate LoRA inside the region and use the frozen backbone outside.

$$S = \bar{h}_{\text{pooled}}{}^{\top} \bar{W}_R, \quad \mathcal{L}_{\text{compact}} = \max\big(0, \tau - \max(S)\big), \quad \mathcal{L}_{\text{diverse}} = -\text{Var}(S)\mathbb{I}[\max(S) < \tau]$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{compact}} + \lambda_2 \mathcal{L}_{\text{diverse}}$$

$$M_{\text{gate}} = \mathbb{I}[\max(S) > \tau - \delta], \quad y = W_0 x + \frac{\alpha}{r} M_{\text{gate}} BAx$$

2. DPO-RS: Downstream-Performance-Oriented Rank Sculpting

- **Rank Allocation Guided by Information Loss Minimization.** We formulate layer-wise LoRA rank allocation as minimizing the total information loss induced by low-rank updates under a fixed parameter budget.

$$\mathcal{E}_{\text{loss}}^{(l)} \approx \eta \sum_{i=2r_l+1}^{d^{(l)}} \left(\sigma_i^{(l)}\right)^2, \quad \min_{\{r_l\}} \sum_l \mathcal{E}_{\text{loss}}^{(l)}, \sum_l r_l c_l \le \mathcal{B}, \quad c_l = d_{\text{in}}^{(l)} + d_{\text{out}}^{(l)}, \quad r_l \in \mathbb{Z}_+$$

$$p_l = \frac{\|G_l\|_{\text{F}}}{\sum_k \|G_k\|_{\text{F}}}, \quad r_l = \text{clip}_{[r_{\min}, r_{\max}]} \left\lfloor \frac{\mathcal{B} p_l}{\sqrt{c_l}} \right\rfloor$$

Experiments & Analysis

1. Main results under matched budgets

Methods	IconQA							COCO-Caption						
	OKVQA	OCRVQA	GQA	TextVQA	*Source*	*Target*	*Avg*	OKVQA	OCRVQA	GQA	TextVQA	*Source*	*Target*	*Avg*
Zero-shot	57.99	66.20	61.93	58.23	61.09	23.18	42.13	57.99	66.20	61.93	58.23	61.09	40.40	50.74
Budget matched to LoRA Rank=16														
LoRA	51.10	54.90	55.56	49.48	52.76	84.14	68.45	47.34	60.55	56.91	45.61	52.60	112.22	82.41
DoRA	49.94	54.40	55.39	47.26	51.75	84.48	68.11	45.38	59.75	55.40	41.38	50.48	112.60	81.54
AdaLoRA	53.02	54.05	58.20	51.57	54.21	84.56	69.39	25.57	54.00	38.92	27.13	36.41	121.53	78.97
GoRA	42.54	58.00	50.63	22.11	43.32	83.09	63.21	46.25	63.05	57.02	42.15	52.12	117.56	84.84
L2-Reg	49.96	51.40	55.15	47.47	51.00	84.66	67.83	46.05	61.50	56.60	43.54	51.92	111.83	81.88
Orth-Reg	53.12	56.10	57.43	51.00	54.41	84.52	69.47	48.11	59.85	56.83	46.69	52.87	111.87	82.37
Model Tailor	52.68	56.55	56.26	51.41	54.23	83.26	68.74	51.83	59.75	58.73	51.43	55.44	118.84	87.14
DARE	51.10	54.30	55.55	49.20	52.54	84.20	68.37	46.64	59.45	56.01	44.46	51.64	112.21	81.93
LoRAMoE	53.58	55.15	57.66	51.72	54.53	84.90	69.71	48.81	59.55	57.46	48.54	53.59	111.77	82.68
LoRASculpt	54.07	57.00	57.66	52.24	55.24	85.02	70.13	50.86	59.10	57.64	51.10	54.68	121.26	87.97
Dual-Pronged LoRA (Ours)	57.92	66.15	61.75	57.60	**60.86**	**85.72**	**73.29**↑3.16	57.97	66.25	61.93	58.21	**61.09**	**126.13**	**93.61**↑5.64
Budget matched to LoRA Rank=32														
LoRA	45.06	48.40	49.22	36.47	44.79	82.52	63.65	44.92	59.90	54.88	39.75	49.86	110.27	80.07
DoRA	46.35	48.50	40.80	33.01	42.17	84.82	63.49	44.10	60.05	54.69	40.17	49.75	109.25	79.50
AdaLoRA	53.19	51.60	57.68	51.55	53.51	**85.89**	69.70	28.71	55.20	43.32	31.26	39.62	121.32	80.47
GoRA	56.53	63.45	61.03	56.69	59.43	85.47	72.45	25.59	24.50	26.36	18.95	23.85	92.26	58.06
L2-Reg	30.16	26.65	16.62	17.16	22.65	82.36	52.50	43.60	60.00	54.09	36.53	48.56	109.39	78.97
Orth-Reg	41.51	40.45	53.58	30.92	41.62	83.14	62.38	43.77	60.70	53.74	38.61	49.21	110.44	79.82
Model Tailor	50.07	53.85	53.04	46.92	50.97	81.76	66.37	50.99	60.60	58.54	49.65	54.95	117.64	86.29
DARE	44.39	46.85	48.75	34.84	43.71	82.28	62.99	42.98	58.65	53.92	38.61	48.54	108.57	78.56
LoRAMoE	51.51	55.05	56.65	49.83	53.26	84.77	69.02	48.64	59.75	57.05	47.59	53.26	112.86	83.06
LoRASculpt	53.52	59.50	57.63	53.76	56.10	85.26	70.68	49.99	58.65	57.63	50.73	54.25	120.35	87.30
Dual-Pronged LoRA (Ours)	57.99	66.25	61.93	56.12	**60.57**	85.85	**73.21**↑2.76	57.98	66.25	61.93	58.21	**61.09**	**124.07**	**92.58**↑5.28

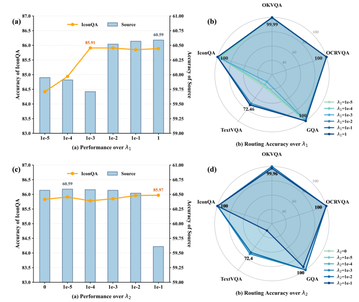
2. Component ablation

Method	IconQA			COCO-Caption		
	Source	Target	Avg	Source	Target	Avg
Zero-shot	61.09	23.18	42.13	61.09	40.40	50.74
LoRA	54.32	83.90	69.11	52.29	112.17	82.23
+ TDCR	60.48	83.90	72.19	**61.10**	112.17	86.63
+ DPO-RS	54.58	**85.97**	70.24	53.45	123.74	88.59
Both	**60.57**	85.85	**73.21**	61.09	**124.07**	**92.58**

3. Domain generalization

Method	ArxivQA			Flickr30k		
	Source	Target	Avg	Source	Target	Avg
Zero-shot	61.09	53.80	57.45	61.09	58.44	59.77
LoRA	38.05	89.77	63.91	54.23	80.78	67.51
LoRASculpt	36.10	91.43	63.76	55.95	85.42	70.68
Ours	**60.78**	**92.17**	**76.47**	**61.09**	**90.13**	**75.61**

4. TDCR sensitivity



5. TextVQA case study



Please refer to our paper for more details!