KU LEUVEN

ECCV
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO
2024

Introducing Routing Functions to Vision-Language Parameter-Efficient Fine-Tuning with Low-Rank Bottlenecks

Tingyu Qu, Tinne Tuytelaars, Marie-Francine Moens

Background: PEFT with Low-Rank Bottlenecks

- colored modules → update
- down-project then up-project:

  → W_{down}: $d \rightarrow r$; W_{up}: $r \rightarrow d$ & $r \ll d$
- *For uni-modal tasks:*

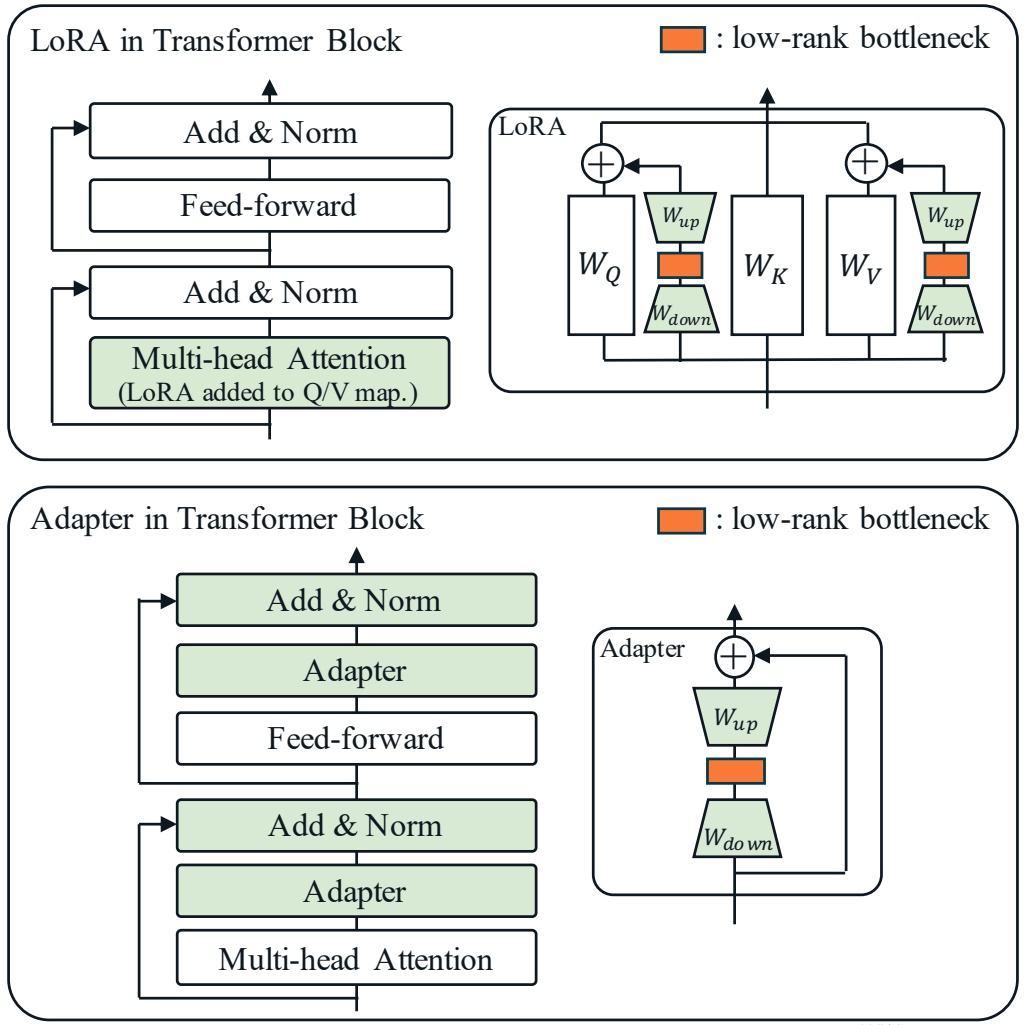
  compression from W_{down} suffices

  if $r >$ intrinsic dim. (minimum dim. required)
- *For multi-modal tasks:*

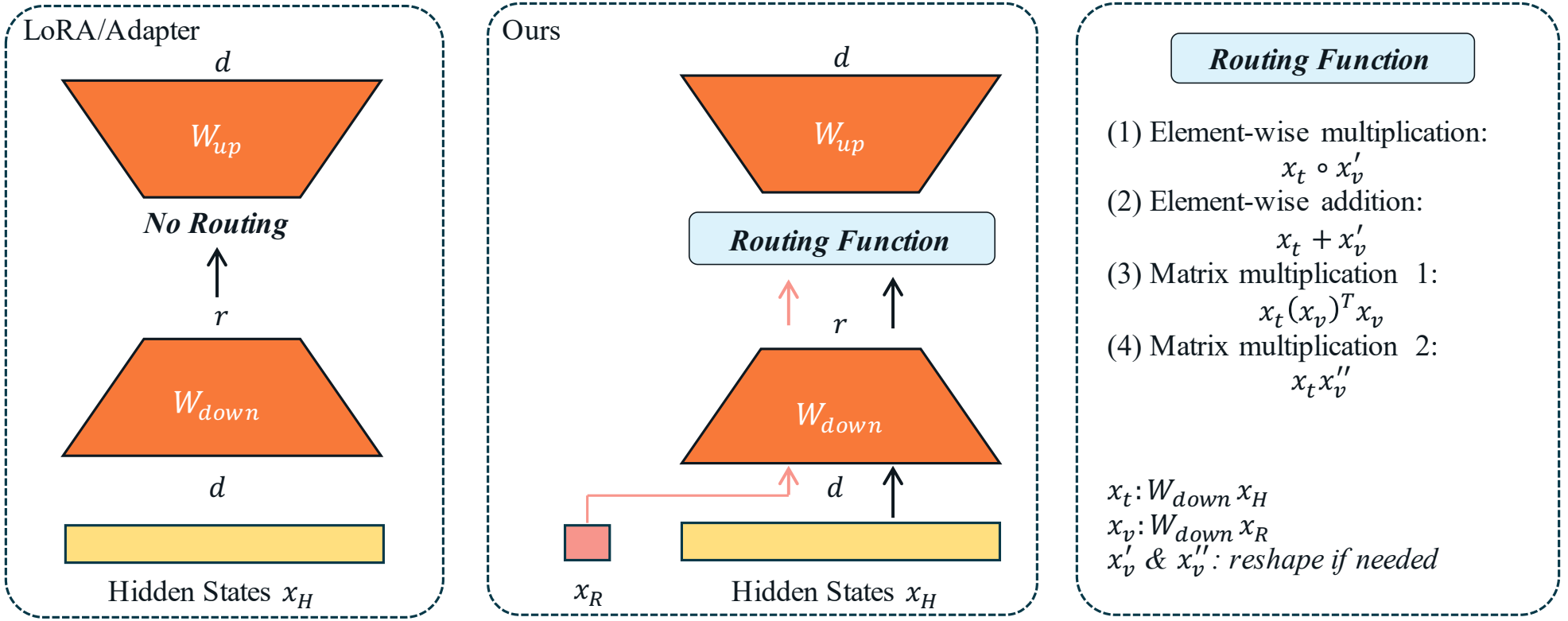
  How to balance modalities to enforce alignment

  with simple linear mapping (W_{down}) ?

→ *Route features through the low-rank bottlenecks*



Method: VL PEFT with Routing Functions

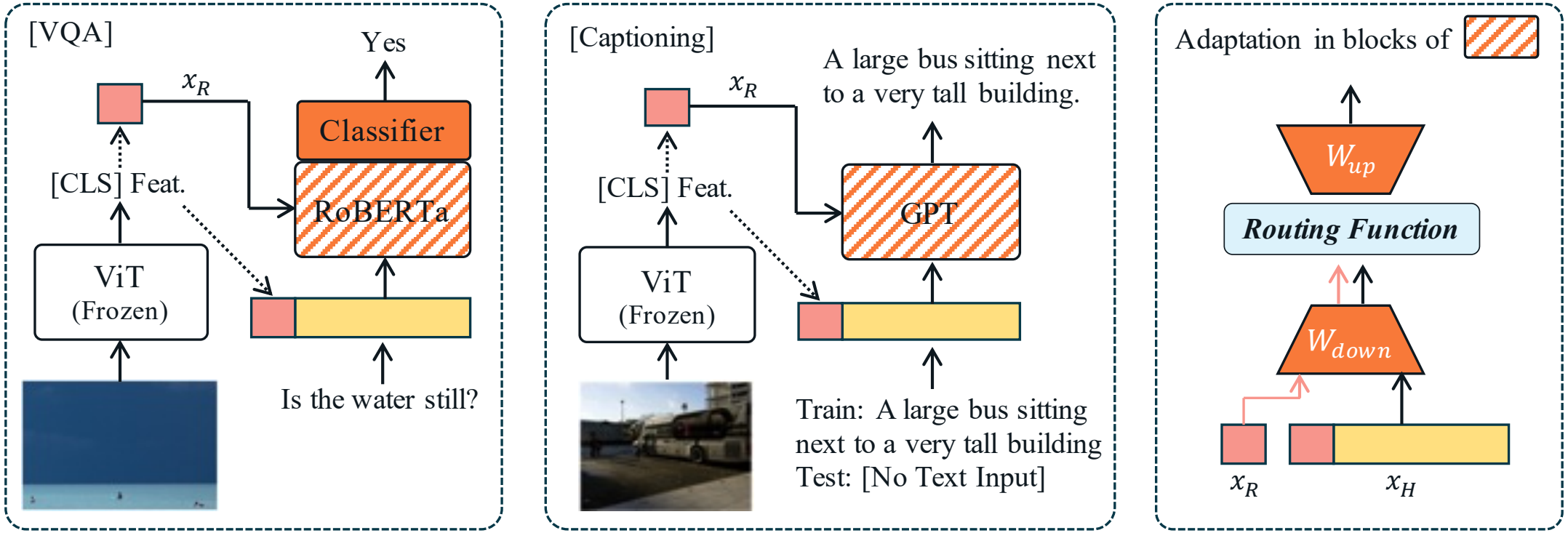


Given model hidden states (x_H) and features to be aligned to x_R (e.g. visual features):

Routing functions ▷ route $W_{down}x_H$ and $W_{down}x_R$ in the ***low-rank bottlenecks***

▷ use ***Linear operation with NO extra parameters***

Experiment with Encoder/Decoder-only Language Models



- Encoder-only LM: RoBERTa; decoder-only LM: GPT2
- Generative task: COCO Cap.; discriminative task: VQAv2
- Base PEFT module: Adapter/LoRA

Experiment with Encoder/Decoder-only Language Models

COCO Cap.:

Routing Functions	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
			LoRA, r=64		
None	18.3	17.4	36.8	55.8	11.8
$x_t \circ x'_v$	**26.1**$_{+7.8}$	**23.7**$_{+6.3}$	**48.6**$_{+11.8}$	**88.7**$_{+32.9}$	**17.3**$_{+5.5}$
$x_t + x'_v$	22.2$_{+3.9}$	20.3$_{+2.9}$	41.0$_{+4.2}$	73.5$_{+17.7}$	14.8$_{+3.0}$
$x_t(x_v)^\mathsf{T} x_v$	24.8$_{+6.5}$	22.6$_{+5.2}$	45.4$_{+8.6}$	84.9$_{+29.1}$	16.8$_{+5.0}$
$x_t x''_v$	23.9$_{+5.6}$	21.9$_{+4.5}$	43.9$_{+7.1}$	80.5$_{+14.7}$	16.2$_{+4.4}$
			Adapter, r=64		
None	15.9	18.5	37.0	61.6	14.1
$x_t \circ x'_v$	24.6$_{+8.7}$	23.1$_{+4.6}$	46.4$_{+9.4}$	84.5$_{+22.9}$	17.2$_{+3.1}$
$x_t + x'_v$	21.0$_{+5.1}$	21.5$_{+3.0}$	42.4$_{+5.4}$	75.0$_{+14.4}$	16.0$_{+1.9}$
$x_t(x_v)^\mathsf{T} x_v$	26.1$_{+10.2}$	**23.2**$_{+4.7}$	**46.9**$_{+9.9}$	85.4$_{+23.8}$	**17.3**$_{+3.2}$
$x_t x''_v$	**26.6**$_{+10.7}$	23.0$_{+4.5}$	46.8$_{+9.8}$	**85.8**$_{+24.2}$	17.2$_{+3.1}$

VQAv2:

Routing Functions	r = 64 LoRA	r = 64 Adapter	r = 128 LoRA	r = 128 Adapter
None	44.15	44.16	44.45	44.28
$x_t \circ x'_v$	53.51$_{+9.36}$	52.78$_{+8.62}$	**53.86**$_{+10.41}$	53.01$_{+8.73}$
$x_t + x'_v$	52.60$_{+8.45}$	53.94$_{+9.78}$	52.88$_{+8.43}$	53.95$_{+9.67}$
$x_t(x_v)^\mathsf{T} x_v$	53.88$_{+9.73}$	54.48$_{+10.32}$	53.09$_{+8.64}$	**55.06**$_{+10.78}$
$x_t x''_v$	**54.21**$_{+10.06}$	**54.96**$_{+10.80}$	51.88$_{+7.43}$	54.38$_{+10.10}$

Comparison to cross-attention:

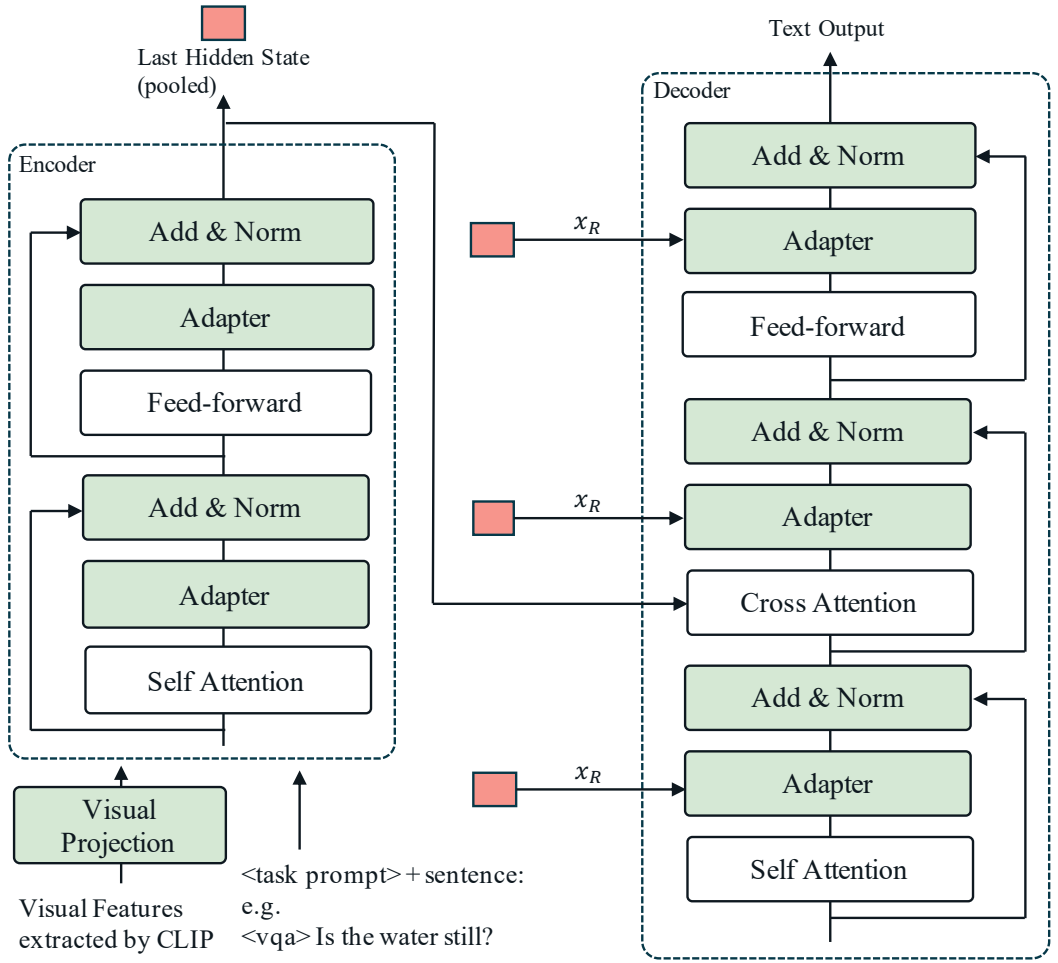
PEFT	Separate Map.	Alignment	Param.	BLEU-4	CIDEr
LoRA	✓	Cross-attn.	4.786M	28.7	92.2
LoRA	✓	$x_t(x_v)^\mathsf{T} x_v$	**3.932M**	30.7$_{+2.0}$	99.4$_{+7.2}$
LoRA	✗	$x_t(x_v)^\mathsf{T} x_v$	**4.746M**	30.0$_{+1.3}$	99.0$_{+6.8}$
Adapter†	✓	Cross-attn.	4.732M	30.7	99.8
Adapter†	✗	$x_t(x_v)^\mathsf{T} x_v$	**1.830M**	30.8$_{+0.1}$	98.8$_{-1.0}$

1. **Significant improvements** on both generative (COCO Cap.) and discriminative (VQAv2) tasks.
2. Comparable to cross-attention, with **fewer parameters & linear operations**

Experiment with Encoder-Decoder Language Models

- Multi-task learning of four VL tasks (VQAv2, GQA, NLVR^2 and COCO Cap.)
- Single Adapter/LoRA for all tasks:

Routing Functions	Single LoRA VQA	GQA	NLVR^2	Avg.	COCO Cap.
None†	65.15	53.66	72.58	63.80	115.01
$x_t \circ x'_v$	**65.68**	**53.96**	73.42	**64.35**	113.94
$x_t + x'_v$	65.14	53.73	73.51	64.13	114.96
$x_t (x_v)^\mathsf{T} x_v$	64.94	53.56	**73.60**	64.03	117.80
$x_t x''_v$	64.84	53.13	72.98	63.65	**119.26**
Routing Functions	**Single Adapter VQA**	**GQA**	**NLVR^2**	**Avg.**	**COCO Cap.**
None†	65.76	54.16	73.19	64.37	114.61
$x_t \circ x'_v$	**65.92**	**54.34**	74.23	**64.83**	114.38
$x_t + x'_v$	65.89	54.18	73.90	64.66	114.39
$x_t (x_v)^\mathsf{T} x_v$	65.84	53.65	**74.31**	64.03	117.65
$x_t x''_v$	65.83	53.61	73.27	64.24	**118.86**



1. **Consistent improvements** especially for COCO Cap.
2. See paper: **Combining multiple adapters w/ routing functions**

Takeaways

Routing functions

→ **help guide the feature learning in low-rank bottlenecks for PEFT**.

→ work for **various types of vision-language models**.

→ can **potentially be beneficial to more tasks** when feature routing is needed.

- Please refer to our paper for more results and detailed analyses.

Thank you for your attention

KU LEUVEN