VL-PET: Vision-and-Language Parameter-Efficient Tuning via Granularity Control

Zi-Yuan Hu, Yanyang Li, Michael R. Lyu, Liwei Wang

香港中文大學 The Chinese University of Hong Kong
Centre for Perceptual and Interactive Intelligence
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Project Page (Code Available)

ICCV23 PARIS

Background

- As the model size of **pre-trained language models (PLMs)** grows rapidly, full fine-tuning becomes prohibitively expensive for model training and storage.
- In **vision-and-language (VL), parameter-efficient tuning (PET)** techniques are proposed to integrate **modular modifications (e.g., Adapter and LoRA)** into encoder-decoder PLMs.

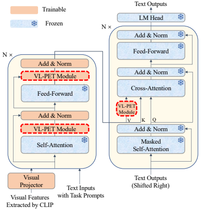
Critical Issues Neglected by Existing Methods

- Integrating **heavy and excessive** modular modifications into PLMs can greatly affect the intermediate output of the PLMs, leading to instability and performance degradation.
- For PLMs used in VL tasks, there exists **functionality gap between the encoders and decoders**.

Main Contribution

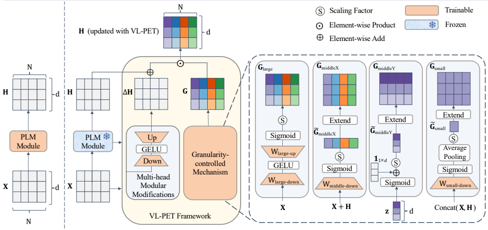
- We propose a **V**ision-and-**L**anguage **P**arameter-**E**fficient **T**uning **(VL-PET)** framework to impose effective control over modular modifications via a novel **granularity-controlled mechanism**.
- Considering different **granularity-controlled matrices** generated by this mechanism, a variety of **model-agnostic VL-PET modules** can be instantiated from our framework for better **efficiency and effectiveness trade-offs**.
- We further propose **lightweight PET module designs** to **enhance VL alignment and modeling for the encoders** and **maintain text generation for the decoders**.
- Extensive experiments conducted on four image-text tasks and four video-text tasks demonstrate the **efficiency, effectiveness and transferability** of our VL-PET framework.

Model Architecture



- Encoder-decoder generative PLMs (e.g., BART and T5)
- Model-agnostic VL-PET modules
- Lightweight PET module designs

VL-PET Framework



(a) Full Fine-tuning (b) Fine-tuning with VL-PET Framework (c) Examples of Granularity-controlled Matrices Generated by Granularity-controlled Mechanism

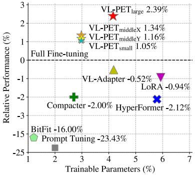
- **Unified formulation of PET**: $H \leftarrow H + \Delta H$
  H: intermediate hidden state from a PLM module
  ΔH: modular modifications (e.g., Adapter and LoRA)
- **Granularity-controlled mechanism**: $H \leftarrow G \odot (H + \Delta H)$
- **Multi-head modular modifications:**
  $\Delta H' = \phi(\text{Concat}(X'W^{(1)}_{\text{down}}, \cdots, X'W^{(N_h)}_{\text{down}}))W_{\text{up}}$

Level	Trainable Matrix	Granularity-controlled Matrix	Trainable Parameter Complexity
Large	$\mathbf{G}_{\text{large}} \in \mathbb{R}^{N\times d}$	$\mathbf{G}_{\text{large}} \in \mathbb{R}^{N\times d}$	$\mathcal{O}(dr)$
MiddleX	$\widetilde{\mathbf{G}}_{\text{middleX}} \in \mathbb{R}^{N\times 1}$	$\mathbf{G}_{\text{middleX}} \in \mathbb{R}^{N\times d}$	$\mathcal{O}(d)$
MiddleY	$\widetilde{\mathbf{G}}_{\text{middleY}} \in \mathbb{R}^{1\times d}$	$\mathbf{G}_{\text{middleY}} \in \mathbb{R}^{N\times d}$	$\mathcal{O}(d)$
Small	$\widetilde{\mathbf{G}}_{\text{small}} \in \mathbb{R}^{1\times 1}$	$\mathbf{G}_{\text{small}} \in \mathbb{R}^{N\times d}$	$\mathcal{O}(d)$

Experiment Results on Image-Text Tasks

Method	Trainable Params (%)	VQA Acc. (%)	GQA Acc. (%)	NLVR² Acc. (%)	COCO Cap. (CIDEr)	Avg.
Backbone: BART-base						
Full Fine-tuning♣	100	$66.88_{0.17}$	$56.79_{0.41}$	$73.66_{0.21}$	$112.01_{0.93}$	$77.33_{0.39}$
BitFit♣ [54]	1.21	$52.94_{0.52}$	$43.15_{0.94}$	$52.29_{0.76}$	$111.44_{0.41}$	$64.96_{0.17}$
Prompt Tuning♣ [26]	2.00	$44.12_{0.45}$	$36.37_{0.35}$	$51.34_{0.83}$	$105.02_{0.24}$	$59.21_{0.21}$
Compacter♣ [21]	2.70	$64.63_{0.09}$	$52.70_{0.24}$	$71.11_{0.35}$	$114.69_{0.42}$	$75.78_{0.21}$
HyperFormer♣ [38]	5.79	$64.62_{0.67}$	$52.55_{0.64}$	$70.74_{1.45}$	$114.84_{0.38}$	$75.69_{0.74}$
LoRA♣ [16]	5.93	$65.15_{0.16}$	$53.66_{0.84}$	$72.58_{0.73}$	$115.01_{0.26}$	$76.60_{0.32}$
VL-Adapter♣ [49]	4.18	$65.76_{0.28}$	$54.16_{0.44}$	$73.19_{0.71}$	$114.61_{0.26}$	$76.93_{0.25}$
$\text{VL-PET}_{\text{small}}$	2.98	$65.43_{0.06}$	$54.03_{0.14}$	$72.43_{0.22}$	$120.68_{0.35}$	$78.14_{0.11}$
$\text{VL-PET}_{\text{middleX}}$	2.98	$65.54_{0.09}$	$54.53_{0.15}$	$72.66_{0.17}$	$120.72_{0.51}$	$78.37_{0.14}$
$\text{VL-PET}_{\text{middleY}}$	2.98	$65.36_{0.15}$	$53.83_{0.39}$	$\mathbf{73.43}_{0.78}$	$120.31_{0.09}$	$78.23_{0.19}$
$\text{VL-PET}_{\text{large}}$	4.16	$\mathbf{66.17}_{0.27}$	$\mathbf{55.11}_{0.17}$	$\mathbf{73.43}_{0.35}$	$\mathbf{122.03}_{0.46}$	$\mathbf{79.18}_{0.14}$
Backbone: T5-base						
Full Fine-tuning♠	100	$67.10_{0.10}$	$56.30_{0.30}$	$74.30_{0.40}$	$112.20_{0.30}$	$77.50_{0.30}$
BitFit♠ [54]	0.83	$55.10_{0.20}$	$45.50_{0.20}$	$51.70_{1.10}$	$101.20_{0.20}$	$63.40_{0.10}$
Prompt Tuning♠ [26]	1.26	$47.40_{0.70}$	$40.60_{0.40}$	$51.00_{0.40}$	$96.10_{0.90}$	$58.80_{0.60}$
LoRA♠ [16]	7.54	$63.70_{0.20}$	$53.30_{0.10}$	$70.00_{0.30}$	$110.30_{0.40}$	$74.30_{0.10}$
VL-Adapter♠ [49]	7.98	$\mathbf{67.10}_{0.10}$	$56.00_{0.40}$	$72.70_{0.30}$	$111.80_{0.10}$	$76.90_{0.20}$
LST♠ [48]	7.46	$66.50_{0.10}$	$55.90_{0.10}$	$71.60_{0.30}$	$113.50_{0.30}$	$76.90_{0.10}$
$\text{VL-PET}_{\text{small}}$	4.51	$65.88_{0.31}$	$54.96_{1.01}$	$72.64_{0.09}$	$120.05_{0.41}$	$78.38_{0.37}$
$\text{VL-PET}_{\text{middleX}}$	4.50	$66.63_{0.14}$	$55.87_{0.25}$	$\mathbf{74.11}_{0.37}$	$120.41_{0.31}$	$79.26_{0.26}$
$\text{VL-PET}_{\text{middleY}}$	4.50	$66.62_{0.20}$	$55.87_{0.13}$	$73.91_{0.45}$	$120.26_{0.40}$	$79.17_{0.08}$
$\text{VL-PET}_{\text{large}}$	7.31	$66.95_{0.21}$	$\mathbf{56.06}_{0.21}$	$73.42_{0.46}$	$\mathbf{121.66}_{0.06}$	$\mathbf{79.52}_{0.21}$

Relatively Improvement



- $\text{VL-PET}_{\text{large}}$ with lightweight PET module designs significantly outperforms VL-Adapter by 2.92% (3.41%) and LoRA by 3.37% (7.03%) with BART-base (T5-base) on image-text tasks

Transfer VL-PET Modules to Video-Text Tasks

Method	Trainable Params (%)	TVQA Acc. (%)	How2QA Acc. (%)	TVC Cap. (CIDEr)	YC2C Cap. (CIDEr)	Avg.
Full Fine-tuning	100	77.69	74.79	50.56	151.71	88.69
BitFit	0.38	66.05	65.42	31.16	115.23	69.47
Prompt Tuning	1.18	24.51	27.76	30.22	108.04	47.63
Compacter	1.89	73.78	72.14	41.39	140.52	81.96
LoRA	5.17	75.51	72.69	44.17	142.72	83.77
VL-Adapter	3.39	77.06	74.73	46.72	153.28	87.95
$\text{VL-PET}_{\text{small}}$	2.18	77.69	74.89	47.92	150.24	87.69
$\text{VL-PET}_{\text{middleY}}$	2.17	**77.76**	75.40	**48.30**	150.25	87.93
$\text{VL-PET}_{\text{middleY}}$	2.17	77.58	75.15	47.93	151.13	87.95
$\text{VL-PET}_{\text{large}}$	3.37	76.97	**75.60**	47.53	**154.41**	**88.63**

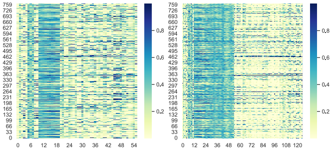
Selected Ablation Studies

Method	Params (%)	VQA (%)	GQA (%)	NLVR² (%)	COCO (CIDEr)	Avg.
VL-PET w/o **G**	2.97	$65.22_{0.14}$	$53.35_{0.39}$	$72.65_{0.44}$	$120.19_{0.68}$	$77.85_{0.34}$
$\text{VL-PET}_{\text{small}}$	2.98	$65.43_{0.06}$	$54.03_{0.14}$	$72.43_{0.22}$	$120.68_{0.35}$	$78.14_{0.11}$
$\text{VL-PET}_{\text{middleX}}$	2.98	$65.54_{0.09}$	$54.53_{0.15}$	$72.66_{0.17}$	$120.72_{0.51}$	$78.37_{0.14}$
$\text{VL-PET}_{\text{middleY}}$	2.98	$65.36_{0.15}$	$53.83_{0.39}$	$\mathbf{73.43}_{0.78}$	$120.31_{0.09}$	$78.23_{0.19}$
$\text{VL-PET}_{\text{large}}$	4.16	$\mathbf{66.17}_{0.27}$	$\mathbf{55.11}_{0.17}$	$\mathbf{73.43}_{0.35}$	$\mathbf{122.03}_{0.46}$	$\mathbf{79.18}_{0.14}$

Applying VL-PET Designs to Existing Methods

Method	Params (%)	VQA (%)	GQA (%)	NLVR² (%)	COCO (CIDEr)	Avg.
Compacter	2.70	$64.63_{0.09}$	$52.70_{0.24}$	$71.11_{0.35}$	$114.69_{0.42}$	$75.78_{0.21}$
+ $\mathbf{G}_{\text{small}}$ + LW	2.08	$64.14_{0.05}$	$52.84_{0.45}$	$71.04_{0.54}$	$118.77_{0.31}$	$76.70_{0.08}$
+ $\mathbf{G}_{\text{middleX}}$ + LW	2.07	$64.35_{0.10}$	$53.10_{0.73}$	$70.57_{0.44}$	$119.02_{0.37}$	$76.76_{0.17}$
+ $\mathbf{G}_{\text{middleY}}$ + LW	2.07	$64.00_{0.12}$	$52.49_{0.60}$	$70.81_{0.68}$	$117.48_{0.11}$	$76.20_{0.33}$
+ $\mathbf{G}_{\text{large}}$ + LW	3.28	$\mathbf{65.60}_{0.15}$	$\mathbf{54.05}_{0.30}$	$\mathbf{71.66}_{0.17}$	$\mathbf{119.55}_{0.23}$	$\mathbf{77.72}_{0.15}$
VL-Adapter	4.18	$65.76_{0.28}$	$54.16_{0.44}$	$73.19_{0.71}$	$114.61_{0.26}$	$76.93_{0.25}$
+ $\mathbf{G}_{\text{small}}$ + LW	2.98	$65.56_{0.17}$	$54.34_{0.32}$	$71.95_{0.23}$	$\mathbf{119.10}_{0.25}$	$77.74_{0.11}$
+ $\mathbf{G}_{\text{middleX}}$ + LW	2.98	$65.73_{0.12}$	$54.90_{0.10}$	$73.04_{0.50}$	$118.34_{0.91}$	$78.00_{0.31}$
+ $\mathbf{G}_{\text{middleY}}$ + LW	2.98	$65.67_{0.17}$	$54.11_{0.29}$	$73.18_{0.34}$	$117.38_{0.52}$	$77.58_{0.08}$
+ $\mathbf{G}_{\text{large}}$ + LW	4.16	$\mathbf{66.31}_{0.23}$	$\mathbf{55.09}_{0.37}$	$\mathbf{73.46}_{0.37}$	$119.05_{0.83}$	$\mathbf{78.47}_{0.11}$

Qualitative Analysis of Granularity Control



(a) VQA input: "vqa: <question><visual>" with $G \in \mathbb{R}^{56\times 768}$ (b) NLVR² input: "nlvr: <question><visual>" with $G \in \mathbb{R}^{126\times 768}$