Pre-PEFT Probing: Weight Statistics and Perturbation Robustness for Layer Selection in VLM Vision Encoders

Qingtao Xia
Faculty of Computing
Harbin Institute of Technology
Harbin, China
25s103308@stu.hit.edu.cn

Jiahua Bao
Faculty of Computing
Harbin Institute of Technology
Harbin, China
bjh_samsara@163.com

Siyao Cheng*
Faculty of Computing
Harbin Institute of Technology
Harbin, China
csy@hit.edu.cn

Jie Liu
Faculty of Computing
Harbin Institute of Technology
Harbin, China
jieliu@hit.edu.cn

***Abstract*—We propose a pre-fine-tuning probing method for Parameter-Efficient Fine-Tuning (PEFT) layer selection, aiming to obtain more stable and higher gains with fewer trainable parameters when adapting large vision–language models (VLMs). Unlike the common practice of applying LoRA and other adapters to all layers at once—where layer selection often relies on heuristic rules—we focus on the vision encoder and directly evaluate the "adaptability" of each Transformer layer. Specifically, we characterize each layer from two perspectives: (i) the statistical properties of its Q/K/V projection weights (e.g., norms and condition numbers); (ii) robustness under controlled parameter perturbations. We then systematically compare these indicators with the downstream performance gains brought by applying PEFT to a single layer. Across experiments covering seven benchmarks and five PEFT variants, we observe a consistent correlation: layers (or matrices) with larger weight norms and higher condition numbers are usually more robust to perturbations and are more likely to yield larger fine-tuning gains. These results show that distribution-statistics analysis and perturbation tests before fine-tuning can provide practical signals for adaptation-layer selection, thereby maintaining or improving performance while reducing trainable parameters.**

***Index Terms*—Parameter-Efficient Fine-Tuning, Vision-Language Models, Layer Selection, LoRA, Perturbation Analysis, Vision Encoder**

I. Introduction

Vision–Language Models (VLMs) combine vision encoders and language decoders for multimodal understanding and generation, and models such as InternVL, LLaVA, and OpenFlamingo have shown strong image–text capabilities [1]–[3]. Most VLMs follow a modular design in which a pre-trained vision encoder is connected to an LLM through a lightweight connector [4]–[6].

For downstream adaptation, Parameter-Efficient Fine-Tuning (PEFT) is widely used because it updates only a small subset of parameters while often retaining competitive performance relative to full fine-tuning [7]–[9]. However, PEFT remains sensitive to where trainable modules are inserted, and layer selection is still largely heuristic.

* Corresponding author.

Even so, PEFT still faces a key bottleneck in practice: hyperparameters and insertion positions are highly sensitive, and there is a lack of explainable and predictable selection criteria. For example, which layers LoRA should be inserted into, whether all layers need to be covered, and how to choose layers when "only tuning part of the layers" can often significantly affect the final performance, while the trial-and-error cost is high. Moreover, many fine-tuning methods introduce additional trainable parameters (e.g., low-rank adapters or gating vectors). With standard random initialization and early-stage updates, these added degrees of freedom perturb the target layer computation in a structured manner. If, before fine-tuning starts, we can pre-judge which layers are more "worth fine-tuning" by analyzing in-layer weight statistics and controlled perturbation responses, then there is an opportunity to improve both the efficiency and the effectiveness of PEFT.

Figure 1 provides an overview of our pre-PEFT probing pipeline; perturbation details are given in Sec. III-B.

To reduce PEFT training parameters and improve model performance, it is crucial to identify which layers will have the most significant fine-tuning effects before starting the process. In both traditional Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), deeper layers are generally more effective at learning complex features [10]. Based on this, we quantify the parameter distribution for Query (Q), Key (K), and Value (V) across each ViT layer, as well as their robustness to perturbations. We then compare the fine-tuning effectiveness of each layer using five common PEFT methods. Our empirical results indicate a stable correspondence between these pre-tuning indicators and layer-wise PEFT gains across the vision encoders of InternVL2-1B (24 layers) and Qwen2.5-VL-3B (32 layers). For each experiment, we have conducted three trials and reported the median as the result; across both backbones, run-to-run variance is small (on the order of 0.3pp). Our contributions are as follows:

- We provide a unified layer-wise study linking perturbation robustness, Q/K/V weight statistics, and PEFT gains in VLM vision encoders.
- Across InternVL2-1B and Qwen2.5-VL-3B, more robust

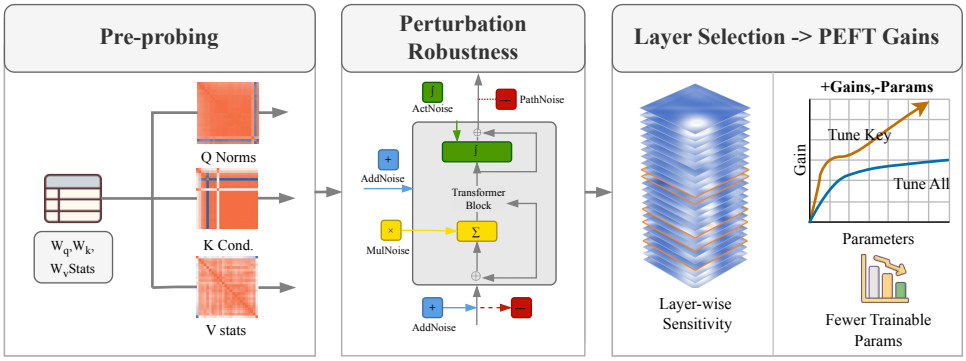

Fig. 1. Overview of our pre-PEFT probing pipeline for layer selection in the vision encoder. We quantify layer-wise weight statistics and perturbation robustness before fine-tuning, and use them as signals to guide PEFT layer selection.

layers often coincide with larger norms, higher condition numbers, and better layer-wise PEFT gains.

- We show that selective tuning of a small subset of layers can remain competitive with broader tuning, providing a practical pre-screening signal for PEFT layer selection.

The code and scripts used in this work are publicly available at https://github.com/atoz03/prepeft-probing.

II. Related Work

A. VLMs

VLMs align a vision encoder with a language decoder for multimodal understanding and generation. Representative systems such as LLaVA connect CLIP ViT-L/14 to an LLM through a lightweight projector [2], [11], while InternVL2 and Qwen2.5-VL adopt ViT-style vision encoders, lightweight connectors, and staged training recipes [1], [12]–[14].

B. PEFT with VLMs

To adapt VLMs under limited budgets, PEFT updates only a small subset of parameters, with common choices including LoRA, adapters, and prompt tuning [7], [8], [15]. These methods can be viewed as introducing structured perturbations to the original network during adaptation. Prior work has studied perturbation and robustness in CNN/Transformer models [16]–[18], but their connection to layer-wise PEFT effectiveness in VLM vision encoders remains underexplored. Our work studies this connection through weight statistics, perturbation robustness, and layer-wise PEFT gains.

III. Datasets and Perturbation Settings

We study pre-PEFT layer selection in the vision encoder. Our hypothesis is that robustness to controlled perturbations before fine-tuning correlates with subsequent PEFT effectiveness. Accordingly, we probe each layer without updating the base model, and later compare the probing signals with layer-wise PEFT results.

A. Datasets Preparation

We select 48 datasets that cover diverse VLM capabilities (image captioning, OCR, chart and table understanding, scientific image question answering, math question answering, etc.), including Screen2Words [19], OCR-VQA [20], Chart2Text [21], CLEVR-Math [22], and ScienceQA [23]. Since most downstream fine-tuning tasks have small data sizes and there is some overlap among datasets, the training distribution may be dominated by a few datasets. To more clearly observe the correlation between "perturbation scenarios–PEFT methods" and to better match real-world settings, we sample 27,000 pairs from 1.8 million $\langle image, text \rangle$ pairs to construct a training mixture for subsequent perturbation probing and PEFT training.

To evaluate model performance after PEFT, we choose seven benchmarks for comprehensive evaluation: MMStar [24] as the primary benchmark for overall comparison and consistency checking of basic capabilities, along with six additional benchmarks—ChartQA [25], OCR-VQA [20], HallusionBench [26], ScienceQA [23], RealWorldQA [27], and MMBench [28]—to evaluate diverse downstream tasks (English dialogue, form understanding, OCR, hallucination detection, science question answering, real-world question answering, etc.).

B. Perturbation Settings

Figure 2 summarizes these perturbation operators and injection points.

To align perturbation design with layer-wise controllability, we focus on five PEFT methods that can be naturally compared at the layer level: LoRA [7], VeRA [29], IA3 [30], LayerNorm-Tuning, and Partial Fine-Tuning.

We design five perturbation methods: Additive Noise (AddNoise), Multiplicative Noise (MulNoise), Linear Noise (LinearNoise), Activation Noise (ActNoise), and Path Noise (PathNoise). AddNoise and MulNoise sample from a normal distribution and apply element-wise noise scaled by a noise weight; LinearNoise and ActNoise introduce perturbations via a Kaiming-initialized linear layer [31] and ReLU [32]; PathNoise explicitly simulates the LoRA-style "added bypass" structure: we perform Singular Value Decomposition (SVD) on the original QKV generation linear layer to obtain U, S, and V, and then construct a new QKV generation matrix by down-sampling with $U \otimes S_u$ and up-sampling with $S_v \otimes V$ (where $S_u = S_v = \sqrt{S}$). The final output is a weighted sum of the original path and the new path.

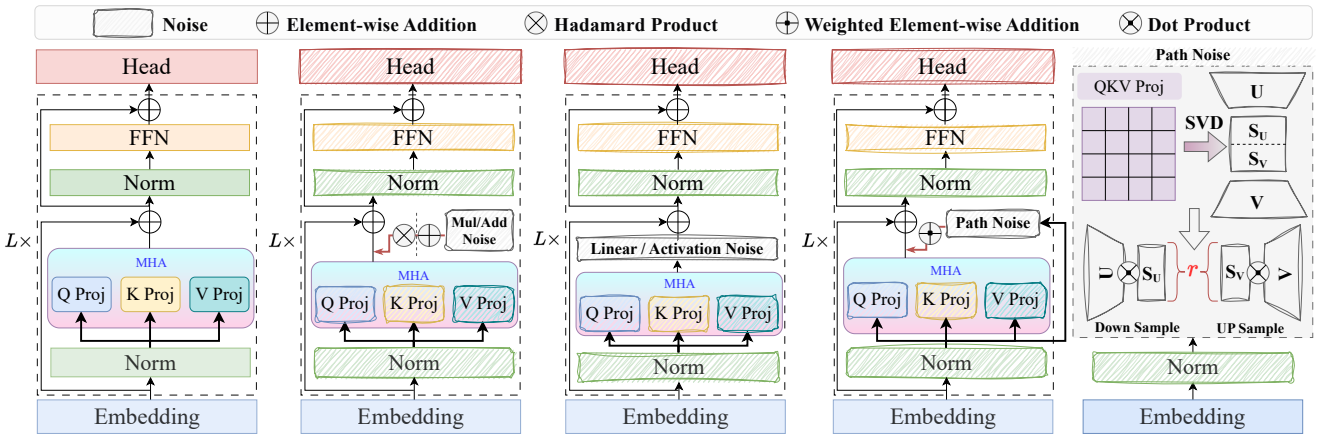

Fig. 2. Perturbation operators and injection points for pre-PEFT probing, including element-wise and activation perturbations.

In preliminary tests, we find that inserting a trainable linear layer (LinearNoise) into the vision encoder collapses the VLM (all benchmark scores drop to 0.0); thus LinearNoise is omitted from the reported tables.

a) Formalization: Let e denote the input embedding of a target layer and $qkv(\cdot)$ denote the original QKV projection. We instantiate perturbations by modifying the QKV output:

$$\begin{aligned} \textsc{AddNoise}: &\quad qkv(e) + \alpha\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \\ \textsc{MulNoise}: &\quad qkv(e) \odot (1 + \alpha\,\epsilon), \quad \epsilon \sim \mathcal{N}(0, I) \\ \textsc{ActNoise}: &\quad \phi(qkv(e))\,, \quad \phi(\cdot) = \mathrm{ReLU}(W(\cdot)) \\ \textsc{PathNoise}: &\quad (1-\alpha)\,qkv(e) + \alpha\,qkv'(e) \end{aligned} \tag{1}$$

where α is the perturbation weight and $qkv'(\cdot)$ is the SVD-constructed low-rank bypass path. We measure robustness by the change of downstream scores under the same evaluation protocol: layers with smaller drops are considered more robust.

b) Layer selection by probing: Given an adaptation budget of k layers, we rank candidate layers on a held-out split $\mathcal{D}$ by a probing score computed from ActNoise and PathNoise drops (smaller is better). Concretely, we define:

$$R_\ell = \frac{1}{2}\big(d_{\ell,\mathrm{act}} + d_{\ell,\mathrm{path}}\big), \quad d_{\ell,\mathrm{path}} = \mathbb{E}_{\alpha\in\mathcal{A}_{\mathrm{path}}}[s_0 - s_{\ell,\alpha}]\,, \tag{2}$$

where s_0 is the unperturbed score and $s_{\ell,\alpha}$ is the score after perturbing layer ℓ with strength α. We then select the top-k layers with the smallest R_ℓ for subsequent multi-layer joint PEFT.

c) Hyperparameter Settings (Two Backbones):

- **Backbones**: InternVL2-1B (24L ViT + MLP); Qwen2.5-VL-3B (32L ViT) [14]
- **Training**: lr=4e−5, wd=0.01, warmup=0.03, bf16
- **LoRA**: $r \in \{32, 128, 512\}$, α=256, dropout=0.05
- **Perturbation**: Add/Mul/Act: α=0.05; PathNoise: base α=0.5 (rank=32), sweep α=0.1–1.0
- **Inference**: temp=1.0, max_len=32768, sliding-window (Transformers 4.37.2)

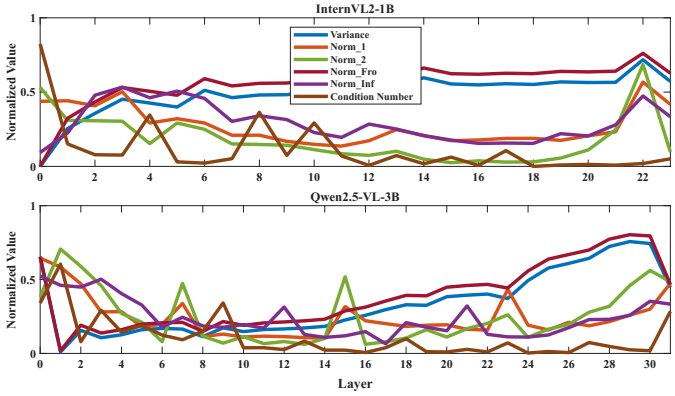

Fig. 3. Layer-wise statistics of the Q/K/V projection linear weights in the vision encoders of InternVL2-1B (top, 24 layers) and Qwen2.5-VL-3B (bottom, 32 layers). We report six metrics (variance, L_1, L_2, Frobenius, infinity norm, and condition number), each min–max normalized across layers within the same backbone.

IV. Experiments and Analysis

A. Distribution of Model Parameters

We analyze the Q/K/V projection matrices in the 24-layer InternVL2-1B and 32-layer Qwen2.5-VL-3B vision encoders. For coarse cross-backbone comparisons, we segment the encoders into contiguous 8-layer blocks and refer to higher-index blocks/layers as deeper.

To quantify differences in the weight distributions of the Q/K/V projection layers, we compute six metrics for each layer and each projection matrix: variance, L_1 norm, L_2 norm, Frobenius norm, infinity norm, and condition number. For each layer ℓ, we compute the statistics on the raw projection weights $W_q^{(\ell)}, W_k^{(\ell)}, W_v^{(\ell)}$. We define the condition number as $\kappa(W) = \sigma_{\max}(W)/\sigma_{\min}(W)$, where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and smallest singular values of W; $\kappa(\cdot)$ is computed on the raw weights without any attention-time scaling (e.g., $1/\sqrt{d_k}$) or additional normalization/reparameterization. For visualization and cross-layer comparison, we apply per-metric min–max scaling across layers within the same backbone to map each metric to $[0, 1]$.

TABLE I

BLOCK-LEVEL PERTURBATION RESULTS ON MMSTAR (LEFT) AND SIX BENCHMARKS (RIGHT) FOR INTERNVL2-1B. ♠/♣/♡/◇: COLUMN-BEST; ↑/↓: VS. UNPERTURBED.

(a) MMStar (InternVL2-1B)

Layers	Perturbation	Overall	CP	FGP	IR	LR	MA	S&T
0-7 (Block-1)	PathNoise	$37.61_{\downarrow 8.06}$	$49.33_{\downarrow 11.87}$	$28.22_{\downarrow 7.38}$	$40.53_{\downarrow 13.07}$	$27.96_{\downarrow 7.24}$	♠$44.13_{\downarrow 3.87}$	$35.47_{\downarrow 4.93}$
	ActNoise	$40.80_{\downarrow 4.87}$	$54.40_{\downarrow 6.80}$	$32.40_{\downarrow 3.20}$	♣$48.00_{\downarrow 5.60}$	$31.20_{\downarrow 4.00}$	$44.00_{\downarrow 4.00}$	♣$34.80_{\downarrow 5.60}$
	AddNoise	$45.07_{\downarrow 0.60}$	$61.20_{0.00}$	$33.60_{\downarrow 2.00}$	$53.20_{\downarrow 0.40}$	$34.80_{\downarrow 0.40}$	$47.60_{\downarrow 0.40}$	$40.00_{\downarrow 0.40}$
	MulNoise	$45.13_{\downarrow 0.53}$	$61.20_{0.00}$	$34.80_{\downarrow 0.80}$	$52.80_{\downarrow 0.80}$	$33.60_{\downarrow 1.60}$	◇$48.80_{\uparrow 0.80}$	$39.60_{\downarrow 0.80}$
8-15 (Block-2)	PathNoise	$36.12_{\downarrow 9.55}$	$44.22_{\downarrow 16.98}$	$28.98_{\downarrow 6.62}$	$39.07_{\downarrow 14.53}$	$27.69_{\downarrow 7.51}$	$43.78_{\downarrow 4.22}$	$32.98_{\downarrow 7.42}$
	ActNoise	$29.87_{\downarrow 15.80}$	$37.20_{\downarrow 24.00}$	$22.00_{\downarrow 13.60}$	$29.60_{\downarrow 24.00}$	$20.40_{\downarrow 14.80}$	$40.40_{\downarrow 7.60}$	$29.60_{\downarrow 10.80}$
	AddNoise	$45.33_{\downarrow 0.33}$	$61.20_{0.00}$	$34.00_{\downarrow 1.60}$	$53.20_{\downarrow 0.40}$	$35.20_{0.00}$	♡$48.40_{\uparrow 0.40}$	$40.00_{\downarrow 0.40}$
	MulNoise	$45.40_{\downarrow 0.27}$	$61.60_{\uparrow 0.40}$	◇$35.20_{\downarrow 0.40}$	$52.80_{\downarrow 0.80}$	$34.80_{\downarrow 0.40}$	$47.20_{\downarrow 0.80}$	◇$40.80_{\uparrow 0.40}$
16-23 (Block-3)	PathNoise	♠$39.88_{\downarrow 5.79}$	♠$54.13_{\downarrow 7.07}$	♠$30.44_{\downarrow 5.16}$	♠$45.07_{\downarrow 8.53}$	♠$30.13_{\downarrow 5.07}$	$43.20_{\downarrow 4.80}$	♠$36.31_{\downarrow 4.09}$
	ActNoise	♣$41.13_{\downarrow 4.53}$	♣$56.00_{\downarrow 5.20}$	♣$33.60_{\downarrow 2.00}$	$45.20_{\downarrow 8.40}$	♣$34.00_{\downarrow 1.20}$	♣$44.80_{\downarrow 3.20}$	$33.20_{\downarrow 7.20}$
	AddNoise	♡$45.60_{\downarrow 0.07}$	♡$61.60_{\uparrow 0.40}$	♡$34.80_{\downarrow 0.80}$	♡$53.60_{0.00}$	♡$36.00_{\uparrow 0.80}$	$47.20_{\downarrow 0.80}$	♡$40.40_{0.00}$
	MulNoise	◇$45.53_{\downarrow 0.13}$	◇$62.00_{\uparrow 0.80}$	$34.80_{\downarrow 0.80}$	◇$53.60_{0.00}$	◇$35.60_{\uparrow 0.40}$	$47.60_{\downarrow 0.40}$	$39.60_{\downarrow 0.80}$

(b) Benchmarks (InternVL2-1B)

Layers	Perturbation	CQA	HB	MMB	OCRVQA	RWQA	SQA
0-7 (Block-1)	PathNoise	$28.79_{\downarrow 29.09}$	$29.59_{\downarrow 5.62}$	$47.69_{\downarrow 19.49}$	$31.98_{\downarrow 11.53}$	♠$42.99_{\downarrow 5.77}$	$78.83_{\downarrow 8.58}$
	ActNoise	$27.84_{\downarrow 30.04}$	♣$30.82_{\downarrow 4.39}$	♣$60.91_{\downarrow 6.27}$	♣$35.37_{\downarrow 8.13}$	♣$44.31_{\downarrow 4.44}$	$82.30_{\downarrow 5.11}$
	AddNoise	♡$58.16_{\uparrow 0.28}$	$35.49_{\uparrow 0.29}$	♡$67.35_{\uparrow 0.17}$	♡$43.54_{\uparrow 0.03}$	♡$48.76_{0.00}$	♡$87.36_{\downarrow 0.05}$
	MulNoise	$58.04_{\uparrow 0.16}$	$31.27_{\downarrow 3.93}$	$67.18_{0.00}$	$43.51_{0.00}$	◇$48.89_{\uparrow 0.13}$	$87.26_{\downarrow 0.15}$
8-15 (Block-2)	PathNoise	$24.57_{\downarrow 33.31}$	♠$30.52_{\downarrow 4.68}$	$42.50_{\downarrow 24.68}$	$27.96_{\downarrow 15.55}$	$42.67_{\downarrow 6.09}$	$78.09_{\downarrow 9.32}$
	ActNoise	$12.88_{\downarrow 45.00}$	$26.22_{\downarrow 8.98}$	$28.69_{\downarrow 38.49}$	$20.64_{\downarrow 22.87}$	$37.52_{\downarrow 11.24}$	$73.18_{\downarrow 14.23}$
	AddNoise	$57.96_{\uparrow 0.08}$	♡$36.14_{\uparrow 0.94}$	$67.18_{0.00}$	$43.53_{\uparrow 0.03}$	$48.24_{\downarrow 0.52}$	$87.31_{\downarrow 0.10}$
	MulNoise	◇$58.08_{\uparrow 0.20}$	$35.10_{\downarrow 0.10}$	◇$67.70_{\uparrow 0.52}$	◇$43.54_{\uparrow 0.04}$	$47.84_{\downarrow 0.92}$	◇$87.46_{\uparrow 0.05}$
16-23 (Block-3)	PathNoise	♠$36.39_{\downarrow 21.49}$	$29.77_{\downarrow 5.44}$	♠$53.00_{\downarrow 14.18}$	♠$33.74_{\downarrow 9.77}$	$42.95_{\downarrow 5.81}$	♠$82.23_{\downarrow 5.17}$
	ActNoise	♣$49.52_{\downarrow 8.36}$	$30.41_{\downarrow 4.79}$	$51.46_{\downarrow 15.72}$	$34.76_{\downarrow 8.74}$	♣$44.31_{\downarrow 4.44}$	♣$82.90_{\downarrow 4.51}$
	AddNoise	♡$58.16_{\uparrow 0.28}$	$31.56_{\downarrow 3.64}$	$67.27_{\uparrow 0.09}$	$43.50_{\downarrow 0.01}$	$48.63_{\downarrow 0.13}$	♡$87.36_{\downarrow 0.05}$
	MulNoise	$57.76_{\downarrow 0.12}$	◇$35.68_{\uparrow 0.47}$	$67.27_{\uparrow 0.09}$	$43.50_{0.00}$	$48.50_{\downarrow 0.26}$	$87.41_{0.00}$

TABLE II

BLOCK-LEVEL PERTURBATION RESULTS ON MMSTAR (LEFT) AND SIX BENCHMARKS (RIGHT) FOR QWEN2.5-VL-3B. ♠/♣/♡/◇: COLUMN-BEST; ↑/↓: VS. UNPERTURBED.

(a) MMStar (Qwen2.5-VL-3B)

Layers	Perturbation	Overall	CP	FGP	IR	LR	MA	S&T
0-7 (Block-1)	PathNoise	♠$54.64_{\downarrow 0.26}$	♠$66.40_{\downarrow 1.20}$	$46.40_{\uparrow 0.40}$	$60.80_{\uparrow 0.40}$	♠$57.43_{\uparrow 2.41}$	♠$60.80_{\downarrow 1.60}$	$36.00_{\downarrow 2.00}$
	ActNoise	♣$55.57_{\uparrow 0.67}$	♣$68.00_{\uparrow 0.40}$	$46.00_{0.00}$	$60.40_{0.00}$	♣$57.43_{\uparrow 2.41}$	♣$63.20_{\uparrow 0.80}$	$38.40_{\uparrow 0.40}$
	AddNoise	$55.30_{\uparrow 0.40}$	$67.60_{0.00}$	$44.80_{\downarrow 1.20}$	$60.00_{\downarrow 0.40}$	♡$57.43_{\uparrow 2.41}$	♡$63.60_{\uparrow 1.20}$	$38.40_{\uparrow 0.40}$
	MulNoise	$55.10_{\uparrow 0.20}$	◇$67.60_{0.00}$	◇$47.60_{\uparrow 1.60}$	$60.40_{0.00}$	$55.42_{\uparrow 0.40}$	$61.60_{\downarrow 0.80}$	$38.00_{0.00}$
8-15 (Block-2)	PathNoise	$53.57_{\downarrow 1.33}$	$65.20_{\downarrow 2.40}$	$46.00_{0.00}$	$58.80_{\downarrow 1.60}$	$54.62_{\downarrow 0.40}$	$60.80_{\downarrow 1.60}$	$36.00_{\downarrow 2.00}$
	ActNoise	$55.44_{\uparrow 0.54}$	$68.00_{\uparrow 0.40}$	♣$46.40_{\uparrow 0.40}$	$60.80_{\uparrow 0.40}$	$57.03_{\uparrow 2.01}$	$62.00_{\downarrow 0.40}$	$38.40_{\uparrow 0.40}$
	AddNoise	♡$55.97_{\uparrow 1.07}$	♡$68.00_{\uparrow 0.40}$	♡$47.60_{\uparrow 1.60}$	$61.20_{\uparrow 0.80}$	$56.22_{\uparrow 1.20}$	$62.80_{\uparrow 0.40}$	♡$40.00_{\uparrow 2.00}$
	MulNoise	$55.24_{\uparrow 0.34}$	$65.60_{\downarrow 2.00}$	$46.00_{0.00}$	$60.80_{\uparrow 0.40}$	◇$57.83_{\uparrow 2.81}$	◇$62.00_{\downarrow 0.40}$	$39.20_{\uparrow 1.20}$
16-23 (Block-3)	PathNoise	$52.30_{\downarrow 2.60}$	$65.20_{\downarrow 2.40}$	$44.80_{\downarrow 1.20}$	$60.40_{0.00}$	$52.61_{\downarrow 2.41}$	$58.40_{\downarrow 4.00}$	$32.40_{\downarrow 5.60}$
	ActNoise	$55.10_{\uparrow 0.20}$	$67.20_{\downarrow 0.40}$	$45.20_{\downarrow 0.80}$	♣$61.60_{\uparrow 1.20}$	$55.02_{0.00}$	$62.80_{\uparrow 0.40}$	$38.80_{\uparrow 0.80}$
	AddNoise	$55.37_{\uparrow 0.47}$	$68.00_{\uparrow 0.40}$	$45.20_{\downarrow 0.80}$	♡$61.60_{\uparrow 1.20}$	$55.82_{\uparrow 0.80}$	$61.60_{\downarrow 0.80}$	$40.00_{\uparrow 2.00}$
	MulNoise	$54.57_{\downarrow 0.33}$	$66.80_{\downarrow 0.80}$	$46.40_{\uparrow 0.40}$	$60.00_{\downarrow 0.40}$	$55.02_{0.00}$	$61.20_{\downarrow 1.20}$	$38.00_{0.00}$
24-31 (Block-4)	PathNoise	$53.77_{\downarrow 1.13}$	$64.80_{\downarrow 2.80}$	♠$46.80_{\uparrow 0.80}$	♠$62.00_{\uparrow 1.60}$	$52.61_{\downarrow 2.41}$	$59.60_{\downarrow 2.80}$	♠$36.80_{\downarrow 1.20}$
	ActNoise	$55.17_{\uparrow 0.27}$	$67.60_{0.00}$	$46.00_{0.00}$	$61.20_{\uparrow 0.80}$	$54.62_{\downarrow 0.40}$	$61.60_{\downarrow 0.80}$	♣$40.00_{\uparrow 2.00}$
	AddNoise	$55.17_{\uparrow 0.27}$	$67.20_{\downarrow 0.40}$	$45.20_{\downarrow 0.80}$	$61.60_{\uparrow 1.20}$	$54.62_{\downarrow 0.40}$	$62.80_{\uparrow 0.40}$	$39.60_{\uparrow 1.60}$
	MulNoise	◇$55.37_{\uparrow 0.47}$	$66.80_{\downarrow 0.80}$	$46.80_{\uparrow 0.80}$	◇$61.20_{\uparrow 0.80}$	$55.82_{\uparrow 0.80}$	$62.00_{\downarrow 0.40}$	◇$39.60_{\uparrow 1.60}$

(b) Benchmarks (Qwen2.5-VL-3B)

Layers	Perturbation	CQA	HB	MMB	OCRVQA	RWQA	SQA
0-7 (Block-1)	PathNoise	♠$84.28_{\uparrow 0.20}$	♠$61.09_{\uparrow 0.73}$	♠$82.27_{\downarrow 0.16}$	♠$76.80_{\downarrow 0.50}$	♠$65.23_{\downarrow 1.18}$	$74.81_{\downarrow 0.75}$
	ActNoise	$84.12_{\uparrow 0.04}$	$60.15_{\downarrow 0.21}$	♣$82.53_{\uparrow 0.10}$	$77.30_{0.00}$	♣$65.75_{\downarrow 0.66}$	♣$75.46_{\downarrow 0.10}$
	AddNoise	♡$84.16_{\uparrow 0.08}$	♡$61.20_{\uparrow 0.84}$	♡$82.59_{\uparrow 0.16}$	$77.00_{\downarrow 0.30}$	$65.49_{\downarrow 0.92}$	$75.06_{\downarrow 0.50}$
	MulNoise	$83.80_{\downarrow 0.28}$	◇$60.78_{\uparrow 0.42}$	$82.53_{\uparrow 0.10}$	$76.90_{\downarrow 0.40}$	$65.23_{\downarrow 1.18}$	$75.16_{\downarrow 0.40}$
8-15 (Block-2)	PathNoise	$83.48_{\downarrow 0.60}$	$60.46_{\uparrow 0.10}$	$80.89_{\downarrow 1.54}$	$76.60_{\downarrow 0.70}$	$63.53_{\downarrow 2.88}$	$74.71_{\downarrow 0.85}$
	ActNoise	$84.08_{0.00}$	$59.73_{\downarrow 0.63}$	$82.20_{\downarrow 0.23}$	$77.20_{\downarrow 0.10}$	$65.23_{\downarrow 1.18}$	$75.01_{\downarrow 0.55}$
	AddNoise	$84.04_{\downarrow 0.04}$	$60.78_{\uparrow 0.42}$	$82.41_{\downarrow 0.02}$	$76.70_{\downarrow 0.60}$	♡$65.88_{\downarrow 0.53}$	$75.11_{\downarrow 0.45}$
	MulNoise	$83.84_{\downarrow 0.24}$	$60.67_{\uparrow 0.31}$	$82.74_{\uparrow 0.31}$	$77.10_{\downarrow 0.20}$	$65.10_{\downarrow 1.31}$	$74.91_{\downarrow 0.65}$
16-23 (Block-3)	PathNoise	$82.44_{\downarrow 1.64}$	$60.15_{\downarrow 0.21}$	$81.63_{\downarrow 0.80}$	$75.40_{\downarrow 1.90}$	$63.40_{\downarrow 3.01}$	$74.47_{\downarrow 1.09}$
	ActNoise	$84.12_{\uparrow 0.04}$	♣$60.57_{\uparrow 0.21}$	$82.47_{\uparrow 0.04}$	$76.60_{\downarrow 0.70}$	$65.62_{\downarrow 0.79}$	$75.26_{\downarrow 0.30}$
	AddNoise	$83.96_{\downarrow 0.12}$	$60.78_{\uparrow 0.42}$	$82.33_{\downarrow 0.10}$	$76.90_{\downarrow 0.40}$	$65.62_{\downarrow 0.79}$	♡$75.21_{\downarrow 0.35}$
	MulNoise	◇$84.00_{\downarrow 0.08}$	$60.15_{\downarrow 0.21}$	◇$82.78_{\uparrow 0.35}$	$77.10_{\downarrow 0.20}$	◇$66.41_{0.00}$	$75.16_{\downarrow 0.40}$
24-31 (Block-4)	PathNoise	$82.16_{\downarrow 1.92}$	$59.10_{\downarrow 1.26}$	$81.96_{\downarrow 0.47}$	$76.80_{\downarrow 0.50}$	$62.48_{\downarrow 3.93}$	♠$74.91_{\downarrow 0.65}$
	ActNoise	♣$84.16_{\uparrow 0.08}$	$60.57_{\uparrow 0.21}$	$82.49_{\uparrow 0.06}$	♣$77.70_{\uparrow 0.40}$	$65.62_{\downarrow 0.79}$	$75.01_{\downarrow 0.55}$
	AddNoise	$83.96_{\downarrow 0.12}$	$60.78_{\uparrow 0.42}$	$82.35_{\downarrow 0.08}$	♡$77.10_{\downarrow 0.20}$	$65.88_{\downarrow 0.53}$	$75.01_{\downarrow 0.55}$
	MulNoise	$84.00_{\downarrow 0.08}$	$60.46_{\uparrow 0.10}$	$82.61_{\uparrow 0.18}$	◇$77.20_{\downarrow 0.10}$	$65.36_{\downarrow 1.05}$	◇$75.31_{\downarrow 0.25}$

In Figure 3, deeper layers generally exhibit larger variance and norm-related metrics, while condition-number fluctuations are more concentrated in earlier layers.

B. Segmented Analysis of Vision Encoder Layers

To estimate promising PEFT layers before tuning, we inject perturbations into the predefined 8-layer blocks and evaluate MMStar plus six benchmarks; PathNoise is used as the LoRA-like perturbation. Tables I(a) and II(a) report the block-level MMStar results for InternVL2-1B and Qwen2.5-VL-3B, respectively.

Across both backbones, ActNoise and PathNoise produce the largest degradation, whereas AddNoise and MulNoise are much milder; moreover, block-level robustness is not monotonic with depth (Tables I and II). This motivates the finer layer-wise analysis below.

C. Layer-Wise Analysis of Vision Encoder

Table III shows substantial within-block variation on InternVL2-1B, with several middle layers consistently more fragile under both ActNoise and PathNoise, suggesting layer-level PEFT selection rather than contiguous block selection.

D. PEFT Analysis of Vision Encoder

We evaluate layer-wise PEFT on InternVL2-1B with LoRA ($r = 32/128/512$), VeRA, IA3, LayerNorm-Tuning, and Partial-Tuning, and on Qwen2.5-VL-3B with LoRA ($r = 32$) plus the same baselines.

Figure 4 and Table IV show that the best PEFT gains cluster in early and late layers, broadly overlapping with perturbation-robust, high-statistics layers.

a) Strong same-budget baselines.: Under the same adaptation budget ($k = 5$), budget-matched heuristics yield only marginal and unstable gains. On InternVL2-1B, first-k LoRA reaches 43.70% versus the 43.63% base, while random subsets and single-component tuning are weaker; on Qwen2.5-VL-3B, middle-k LoRA reaches 55.24% versus the 54.90% base, whereas attention-only and MLP-only tuning drop to 54.70% and 54.50%.

E. Paired Significance Testing on Aggregated Benchmarks

We conduct significance testing on the pooled six-benchmark set (ChartQA, HallusionBench, MMBench, OCRVQA, RealWorldQA, and ScienceQA) using a two-sided McNemar test and a paired percentile bootstrap 95% CI (27k resamples) for the accuracy difference Δ.

VeRA (layer 1) improves pooled accuracy by $\Delta = +2.14$pp (95% CI $[1.63, 2.64]$pp; $p = 1.2 \times 10^{-16}$), and LoRA (layer 2) by $\Delta = +2.08$pp (95% CI $[1.57, 2.58]$pp; $p = 8.236 \times 10^{-16}$), mainly driven by HallusionBench and RealWorldQA; OCRVQA slightly decreases for VeRA (-1.23pp; $p = 0.1038$). Partial-Tuning (layer 13) also yields a small HallusionBench gain ($\Delta = +1.58$pp; $p = 0.02007$). In summary, probing-selected layers provide statistically significant gains on the pooled benchmarks, with improvements concentrated on HallusionBench and RealWorldQA.

Across both backbones and multiple PEFT families, perturbation-robust layers are distributed in both shallow and deep blocks and usually deliver stronger PEFT gains, supporting pre-PEFT probing as a practical layer-selection signal.

TABLE III

LAYER-WISE PERTURBATION ON MMSTAR (A) AND SIX BENCHMARKS (B) FOR INTERNVL2-1B. BEST IN BOLD; WORST UNDERLINED.

(a) MMStar (InternVL2-1B)

L	ActNoise							PathNoise						
	Overall	CP	FGP	IR	LR	MA	S&T	Overall	CP	FGP	IR	LR	MA	S&T
0	**45.6$_{\downarrow0.1}$**	60.0$_{\downarrow1.2}$	36.4$_{\uparrow0.8}$	**53.6$_{0.0}$**	**37.2$_{\uparrow2.0}$**	47.6$_{\downarrow0.4}$	38.8$_{\downarrow1.6}$	44.5$_{\downarrow1.1}$	61.6$_{\uparrow0.4}$	34.0$_{\downarrow1.6}$	51.6$_{\downarrow2.0}$	33.6$_{\downarrow1.6}$	46.4$_{\downarrow1.6}$	40.0$_{\downarrow0.4}$
1	45.0$_{\downarrow0.7}$	60.0$_{\downarrow1.2}$	36.0$_{\uparrow0.4}$	**53.6$_{0.0}$**	34.0$_{\downarrow1.2}$	45.6$_{\downarrow2.4}$	40.8$_{\uparrow0.4}$	44.2$_{\downarrow1.5}$	57.6$_{\downarrow3.6}$	36.0$_{\uparrow0.4}$	51.6$_{\downarrow2.0}$	35.2$_{0.0}$	46.4$_{\downarrow1.6}$	38.4$_{\downarrow2.0}$
2	44.9$_{\downarrow0.7}$	60.0$_{\downarrow1.2}$	33.6$_{\downarrow2.0}$	53.2$_{\downarrow0.4}$	34.8$_{\downarrow0.4}$	47.6$_{\downarrow0.4}$	40.4$_{0.0}$	**45.3$_{\downarrow0.4}$**	61.2$_{0.0}$	36.4$_{\uparrow0.8}$	51.6$_{\downarrow2.0}$	36.8$_{\uparrow1.6}$	47.2$_{\downarrow0.8}$	38.4$_{\downarrow2.0}$
3	45.5$_{\downarrow0.1}$	62.0$_{\uparrow0.8}$	35.2$_{\downarrow0.4}$	51.6$_{\downarrow2.0}$	34.0$_{\downarrow1.2}$	**49.6$_{\uparrow1.6}$**	40.8$_{\uparrow0.4}$	43.8$_{\downarrow1.9}$	58.8$_{\downarrow2.4}$	33.2$_{\downarrow2.4}$	51.2$_{\downarrow2.4}$	36.0$_{\uparrow0.8}$	44.4$_{\downarrow3.6}$	39.2$_{\downarrow1.2}$
4	45.1$_{\downarrow0.5}$	60.4$_{\downarrow0.8}$	36.0$_{\uparrow0.4}$	53.2$_{\downarrow0.4}$	33.2$_{\downarrow2.0}$	49.2$_{\uparrow1.2}$	38.8$_{\downarrow1.6}$	44.9$_{\downarrow0.8}$	59.6$_{\downarrow1.6}$	36.4$_{\uparrow0.8}$	51.2$_{\downarrow2.4}$	36.0$_{\uparrow0.8}$	47.2$_{\downarrow0.8}$	38.8$_{\downarrow1.6}$
5	45.4$_{\downarrow0.3}$	60.0$_{\downarrow1.2}$	36.4$_{\uparrow0.8}$	52.4$_{\downarrow1.2}$	35.6$_{\uparrow0.4}$	47.2$_{\downarrow0.8}$	40.8$_{\uparrow0.4}$	44.5$_{\downarrow1.2}$	61.2$_{0.0}$	35.6$_{0.0}$	51.2$_{\downarrow2.4}$	**37.2$_{\uparrow2.0}$**	45.6$_{\downarrow2.4}$	36.0$_{\downarrow4.4}$
6	44.6$_{\downarrow1.1}$	60.4$_{\downarrow0.8}$	36.0$_{\uparrow0.4}$	51.2$_{\downarrow2.4}$	36.0$_{\uparrow0.8}$	45.6$_{\downarrow2.4}$	38.4$_{\downarrow2.0}$	42.7$_{\downarrow3.0}$	57.2$_{\downarrow4.0}$	33.2$_{\downarrow2.4}$	51.2$_{\downarrow2.4}$	33.2$_{\downarrow2.0}$	46.8$_{\downarrow1.2}$	34.4$_{\downarrow6.0}$
7	44.7$_{\downarrow0.9}$	60.0$_{\downarrow1.2}$	33.6$_{\downarrow2.0}$	53.2$_{\downarrow0.4}$	34.8$_{\downarrow0.4}$	47.2$_{\downarrow0.8}$	39.6$_{\downarrow0.8}$	**45.3$_{\downarrow0.4}$**	60.0$_{\downarrow1.2}$	35.6$_{0.0}$	51.2$_{\downarrow2.4}$	36.4$_{\uparrow1.2}$	46.4$_{\downarrow1.6}$	**42.0$_{\uparrow1.6}$**
8	44.8$_{\downarrow0.9}$	60.4$_{\downarrow0.8}$	36.0$_{\uparrow0.4}$	52.0$_{\downarrow1.6}$	34.4$_{\downarrow0.8}$	46.4$_{\downarrow1.6}$	39.6$_{\downarrow0.8}$	44.6$_{\downarrow1.1}$	58.0$_{\downarrow3.2}$	**37.2$_{\uparrow1.6}$**	53.6$_{0.0}$	34.8$_{\downarrow0.4}$	46.8$_{\downarrow1.2}$	37.2$_{\downarrow3.2}$
9	44.5$_{\downarrow1.1}$	59.6$_{\downarrow1.6}$	35.2$_{\downarrow0.4}$	52.0$_{\downarrow1.6}$	36.0$_{\uparrow0.8}$	45.2$_{\downarrow2.8}$	39.2$_{\downarrow1.2}$	44.6$_{\downarrow1.1}$	58.0$_{\downarrow3.2}$	**37.2$_{\uparrow1.6}$**	53.6$_{0.0}$	34.8$_{\downarrow0.4}$	46.8$_{\downarrow1.2}$	37.2$_{\downarrow3.2}$
10	44.7$_{\downarrow1.0}$	60.4$_{\downarrow0.8}$	33.2$_{\downarrow2.4}$	52.8$_{\downarrow0.8}$	35.6$_{\uparrow0.4}$	46.0$_{\downarrow2.0}$	40.0$_{\downarrow0.4}$	44.5$_{\downarrow1.1}$	59.6$_{\downarrow1.6}$	35.6$_{0.0}$	52.0$_{\downarrow1.6}$	34.8$_{\downarrow0.4}$	45.2$_{\downarrow2.8}$	40.0$_{\downarrow0.4}$
11	45.1$_{\downarrow0.6}$	**62.4$_{\uparrow1.2}$**	36.8$_{\uparrow1.2}$	52.4$_{\downarrow1.2}$	34.8$_{\downarrow0.4}$	45.2$_{\downarrow2.8}$	38.8$_{\downarrow1.6}$	44.5$_{\downarrow1.1}$	59.6$_{\downarrow1.6}$	35.6$_{0.0}$	52.0$_{\downarrow1.6}$	34.8$_{\downarrow0.4}$	45.2$_{\downarrow2.8}$	40.0$_{\downarrow0.4}$
12	43.9$_{\downarrow1.8}$	59.2$_{\downarrow2.0}$	32.4$_{\downarrow3.2}$	50.4$_{\downarrow3.2}$	34.8$_{\downarrow0.4}$	47.2$_{\downarrow0.8}$	39.2$_{\downarrow1.2}$	44.1$_{\downarrow1.5}$	60.4$_{\downarrow0.8}$	35.6$_{0.0}$	51.2$_{\downarrow2.4}$	35.6$_{\uparrow0.4}$	47.2$_{\downarrow0.8}$	34.8$_{\downarrow5.6}$
13	44.0$_{\downarrow1.7}$	59.6$_{\downarrow1.6}$	32.4$_{\downarrow3.2}$	52.8$_{\downarrow0.8}$	34.4$_{\downarrow0.8}$	47.6$_{\downarrow0.4}$	37.2$_{\downarrow3.2}$	42.7$_{\downarrow3.0}$	56.4$_{\downarrow4.8}$	32.8$_{\downarrow2.8}$	49.6$_{\downarrow4.0}$	34.0$_{\downarrow1.2}$	46.0$_{\downarrow2.0}$	37.2$_{\downarrow3.2}$
14	44.0$_{\downarrow1.7}$	60.4$_{\downarrow0.8}$	32.4$_{\downarrow3.2}$	50.8$_{\downarrow2.8}$	34.4$_{\downarrow0.8}$	48.4$_{\uparrow0.4}$	37.6$_{\downarrow2.8}$	43.9$_{\downarrow1.7}$	59.6$_{\downarrow1.6}$	31.2$_{\downarrow4.4}$	53.2$_{\downarrow0.4}$	34.8$_{\downarrow0.4}$	47.2$_{\downarrow0.8}$	37.6$_{\downarrow2.8}$
15	44.4$_{\downarrow1.3}$	57.2$_{\downarrow4.0}$	33.6$_{\downarrow2.0}$	**53.6$_{0.0}$**	36.8$_{\uparrow1.6}$	47.2$_{\downarrow0.8}$	38.0$_{\downarrow2.4}$	44.5$_{\downarrow1.1}$	59.2$_{\downarrow2.0}$	34.4$_{\downarrow1.2}$	52.0$_{\downarrow1.6}$	35.2$_{0.0}$	47.2$_{\downarrow0.8}$	39.2$_{\downarrow1.2}$
16	44.5$_{\downarrow1.2}$	60.4$_{\downarrow0.8}$	34.4$_{\downarrow1.2}$	50.4$_{\downarrow3.2}$	35.6$_{\uparrow0.4}$	46.8$_{\downarrow1.2}$	39.2$_{\downarrow1.2}$	44.4$_{\downarrow1.3}$	58.8$_{\downarrow2.4}$	31.6$_{\downarrow4.0}$	52.8$_{\downarrow0.8}$	36.4$_{\uparrow1.2}$	**48.0$_{0.0}$**	38.8$_{\downarrow1.6}$
17	44.1$_{\downarrow1.6}$	59.6$_{\downarrow1.6}$	32.0$_{\downarrow3.6}$	52.0$_{\downarrow1.6}$	36.4$_{\uparrow1.2}$	45.2$_{\downarrow2.8}$	39.2$_{\downarrow1.2}$	44.1$_{\downarrow1.5}$	56.8$_{\downarrow4.4}$	34.4$_{\downarrow1.2}$	51.2$_{\downarrow2.4}$	35.6$_{\uparrow0.4}$	46.0$_{\downarrow2.0}$	40.8$_{\uparrow0.4}$
18	45.3$_{\downarrow0.3}$	61.6$_{\uparrow0.4}$	34.8$_{\downarrow0.8}$	52.4$_{\downarrow1.2}$	36.4$_{\uparrow1.2}$	46.8$_{\downarrow1.2}$	40.0$_{\downarrow0.4}$	44.7$_{\downarrow1.0}$	60.0$_{\downarrow1.2}$	33.2$_{\downarrow2.4}$	52.8$_{\downarrow0.8}$	35.2$_{0.0}$	46.0$_{\downarrow2.0}$	40.8$_{\uparrow0.4}$
19	45.0$_{\downarrow0.7}$	61.2$_{0.0}$	35.2$_{\downarrow0.4}$	**53.6$_{0.0}$**	34.0$_{\downarrow1.2}$	46.0$_{\downarrow2.0}$	40.0$_{\downarrow0.4}$	45.0$_{\downarrow0.7}$	60.0$_{\downarrow1.2}$	35.6$_{0.0}$	52.4$_{\downarrow1.2}$	35.6$_{\uparrow0.4}$	45.6$_{\downarrow2.4}$	40.8$_{\uparrow0.4}$
20	44.8$_{\downarrow0.9}$	62.0$_{\uparrow0.8}$	35.2$_{\downarrow0.4}$	52.0$_{\downarrow1.6}$	34.0$_{\downarrow1.2}$	47.6$_{\downarrow0.4}$	38.0$_{\downarrow2.4}$	44.9$_{\downarrow0.8}$	60.4$_{\downarrow0.8}$	36.0$_{\uparrow0.4}$	51.6$_{\downarrow2.0}$	36.0$_{\uparrow0.8}$	44.8$_{\downarrow3.2}$	40.4$_{0.0}$
21	44.6$_{\downarrow1.1}$	60.0$_{\downarrow1.2}$	33.2$_{\downarrow2.4}$	51.6$_{\downarrow2.0}$	35.6$_{\uparrow0.4}$	48.4$_{\uparrow0.4}$	38.8$_{\downarrow1.6}$	44.1$_{\downarrow1.6}$	59.6$_{\downarrow1.6}$	33.2$_{\downarrow2.4}$	52.8$_{\downarrow0.8}$	35.6$_{\uparrow0.4}$	44.8$_{\downarrow3.2}$	38.4$_{\downarrow2.0}$
22	44.9$_{\downarrow0.7}$	60.8$_{\downarrow0.4}$	36.0$_{\uparrow0.4}$	51.2$_{\downarrow2.4}$	32.8$_{\downarrow2.4}$	47.6$_{\downarrow0.4}$	**41.2$_{\uparrow0.8}$**	44.5$_{\downarrow1.1}$	**62.0$_{\uparrow0.8}$**	35.2$_{\downarrow0.4}$	**54.0$_{\uparrow0.4}$**	36.4$_{\uparrow1.2}$	42.0$_{\downarrow6.0}$	37.6$_{\downarrow2.8}$
23	44.7$_{\downarrow1.0}$	60.4$_{\downarrow0.8}$	**37.6$_{\uparrow2.0}$**	51.2$_{\downarrow2.4}$	35.6$_{\uparrow0.4}$	46.0$_{\downarrow2.0}$	37.2$_{\downarrow3.2}$	44.1$_{\downarrow1.5}$	59.6$_{\downarrow1.6}$	33.6$_{\downarrow2.0}$	51.2$_{\downarrow2.4}$	33.2$_{\downarrow2.0}$	46.8$_{\downarrow1.2}$	40.4$_{0.0}$

(b) Benchmarks (InternVL2-1B)

Layer	ActNoise						PathNoise					
	CQA	HB	MMB	OCRVQA	RWQA	SQA	CQA	HB	MMB	OCRVQA	RWQA	SQA
0	55.8$_{\downarrow2.0}$	31.5$_{\downarrow3.7}$	67.4$_{\uparrow0.3}$	43.3$_{\downarrow0.2}$	48.4$_{\downarrow0.4}$	87.5$_{0.0}$	53.7$_{\downarrow4.2}$	32.3$_{\downarrow2.9}$	67.0$_{\downarrow0.2}$	42.8$_{\downarrow0.7}$	48.1$_{\downarrow0.7}$	**87.8$_{\uparrow0.3}$**
1	57.4$_{\downarrow0.5}$	31.8$_{\downarrow3.4}$	67.7$_{\uparrow0.5}$	**43.5$_{0.0}$**	48.5$_{\downarrow0.3}$	87.6$_{\uparrow0.1}$	54.0$_{\downarrow3.9}$	33.2$_{\downarrow2.0}$	65.6$_{\downarrow1.5}$	42.7$_{\downarrow0.8}$	47.2$_{\downarrow1.6}$	86.9$_{\downarrow0.5}$
2	58.2$_{\uparrow0.3}$	32.4$_{\downarrow2.8}$	67.8$_{\uparrow0.6}$	**43.5$_{0.0}$**	48.4$_{\downarrow0.4}$	87.6$_{\uparrow0.1}$	54.3$_{\downarrow3.6}$	32.9$_{\downarrow2.3}$	67.0$_{\downarrow0.2}$	42.8$_{\downarrow0.7}$	46.0$_{\downarrow2.7}$	86.7$_{\downarrow0.7}$
3	55.5$_{\downarrow2.4}$	35.6$_{\uparrow0.4}$	67.7$_{\uparrow0.5}$	43.3$_{\downarrow0.2}$	47.3$_{\downarrow1.4}$	87.6$_{\uparrow0.1}$	47.5$_{\downarrow10.4}$	34.2$_{\downarrow1.0}$	65.4$_{\downarrow1.8}$	42.3$_{\downarrow1.2}$	45.2$_{\downarrow3.5}$	87.0$_{\downarrow0.4}$
4	57.1$_{\downarrow0.8}$	31.9$_{\downarrow3.3}$	66.7$_{\downarrow0.5}$	43.4$_{\downarrow0.1}$	48.2$_{\downarrow0.5}$	87.5$_{0.0}$	54.9$_{\downarrow3.0}$	31.8$_{\downarrow3.4}$	67.3$_{\uparrow0.1}$	43.0$_{\downarrow0.5}$	48.4$_{\downarrow0.4}$	87.1$_{\downarrow0.3}$
5	56.6$_{\downarrow1.3}$	**35.7$_{\uparrow0.5}$**	67.3$_{\uparrow0.1}$	43.3$_{\downarrow0.2}$	47.5$_{\downarrow1.3}$	87.5$_{\uparrow0.1}$	55.0$_{\downarrow2.8}$	34.3$_{\downarrow1.0}$	**67.6$_{\uparrow0.4}$**	43.1$_{\downarrow0.4}$	45.9$_{\downarrow2.9}$	87.5$_{\uparrow0.1}$
6	52.8$_{\downarrow5.0}$	33.4$_{\downarrow1.8}$	67.6$_{\uparrow0.4}$	43.0$_{\downarrow0.5}$	48.1$_{\downarrow0.7}$	87.5$_{\uparrow0.1}$	50.2$_{\downarrow7.6}$	32.0$_{\downarrow3.2}$	65.0$_{\downarrow2.1}$	41.8$_{\downarrow1.7}$	45.4$_{\downarrow3.4}$	85.9$_{\downarrow1.5}$
7	57.8$_{\downarrow0.1}$	35.3$_{\uparrow0.1}$	67.2$_{0.0}$	43.4$_{\downarrow0.2}$	47.3$_{\downarrow1.4}$	87.7$_{\uparrow0.2}$	56.8$_{\downarrow1.1}$	31.5$_{\downarrow3.7}$	67.4$_{\uparrow0.3}$	43.3$_{\downarrow0.2}$	46.1$_{\downarrow2.6}$	87.3$_{\downarrow0.1}$
8	56.9$_{\downarrow1.0}$	32.5$_{\downarrow2.7}$	66.7$_{\downarrow0.5}$	43.2$_{\downarrow0.3}$	45.2$_{\downarrow3.5}$	**88.1$_{\uparrow0.6}$**	55.7$_{\downarrow2.2}$	31.2$_{\downarrow4.0}$	67.2$_{0.0}$	43.4$_{\downarrow0.1}$	43.0$_{\downarrow7.2}$	87.5$_{0.0}$
9	56.6$_{\downarrow1.3}$	35.1$_{\downarrow0.1}$	67.0$_{\downarrow0.2}$	43.1$_{\downarrow0.4}$	47.8$_{\downarrow0.9}$	87.9$_{\uparrow0.4}$	55.3$_{\downarrow2.7}$	32.7$_{\downarrow2.5}$	67.3$_{\uparrow0.1}$	43.2$_{\downarrow0.3}$	46.5$_{\downarrow2.2}$	87.8$_{\uparrow0.3}$
10	56.5$_{\downarrow1.4}$	31.6$_{\downarrow3.7}$	67.5$_{\uparrow0.3}$	43.3$_{\downarrow0.2}$	45.6$_{\downarrow3.1}$	87.9$_{\uparrow0.4}$	55.1$_{\downarrow2.8}$	34.7$_{\downarrow0.5}$	67.4$_{\uparrow0.3}$	43.3$_{\downarrow0.2}$	47.0$_{\downarrow1.7}$	87.7$_{\uparrow0.2}$
11	56.9$_{\downarrow1.0}$	32.4$_{\downarrow2.8}$	67.7$_{\uparrow0.5}$	43.2$_{\downarrow0.3}$	45.2$_{\downarrow3.6}$	87.8$_{\uparrow0.3}$	55.2$_{\downarrow2.7}$	33.2$_{\downarrow2.0}$	67.6$_{\uparrow0.4}$	43.0$_{\downarrow0.5}$	45.9$_{\downarrow2.8}$	87.6$_{\uparrow0.1}$
12	57.4$_{\downarrow0.5}$	33.3$_{\downarrow1.9}$	67.2$_{0.0}$	43.3$_{\downarrow0.2}$	47.8$_{\downarrow0.9}$	87.6$_{\uparrow0.1}$	54.9$_{\downarrow3.1}$	35.0$_{\downarrow0.3}$	67.2$_{0.0}$	43.3$_{\downarrow0.2}$	47.6$_{\downarrow1.2}$	87.5$_{0.0}$
13	51.9$_{\downarrow6.0}$	35.5$_{\uparrow0.3}$	67.6$_{\uparrow0.4}$	42.5$_{\downarrow1.0}$	46.9$_{\downarrow1.8}$	87.0$_{\downarrow0.4}$	43.6$_{\downarrow14.2}$	30.6$_{\downarrow4.6}$	65.5$_{\downarrow1.6}$	39.9$_{\downarrow3.6}$	47.2$_{\downarrow1.6}$	87.4$_{0.0}$
14	52.5$_{\downarrow5.4}$	31.4$_{\downarrow3.9}$	67.3$_{\uparrow0.1}$	42.6$_{\downarrow0.9}$	44.3$_{\downarrow4.4}$	87.4$_{0.0}$	55.3$_{\downarrow2.6}$	34.4$_{\downarrow0.8}$	66.5$_{\downarrow0.7}$	43.2$_{\downarrow0.3}$	47.2$_{\downarrow1.6}$	87.5$_{\uparrow0.1}$
15	57.2$_{\downarrow0.7}$	31.2$_{\downarrow4.0}$	67.4$_{\uparrow0.2}$	43.0$_{\downarrow0.6}$	46.9$_{\downarrow1.8}$	87.5$_{0.0}$	56.0$_{\downarrow1.9}$	36.0$_{\uparrow0.8}$	67.4$_{\uparrow0.3}$	42.8$_{\downarrow0.7}$	**49.5$_{\uparrow0.8}$**	87.4$_{0.0}$
16	56.9$_{\downarrow1.0}$	35.3$_{\uparrow0.1}$	**67.9$_{\uparrow0.7}$**	42.9$_{\downarrow0.6}$	48.5$_{\downarrow0.3}$	87.5$_{\uparrow0.1}$	**58.0$_{\uparrow0.1}$**	31.4$_{\downarrow3.8}$	66.8$_{\downarrow0.4}$	43.1$_{\downarrow0.4}$	47.7$_{\downarrow1.0}$	87.1$_{\downarrow0.3}$
17	58.5$_{\uparrow0.6}$	35.2$_{0.0}$	67.0$_{\downarrow0.2}$	43.3$_{\downarrow0.2}$	48.5$_{\downarrow0.3}$	87.0$_{\downarrow0.4}$	57.9$_{0.0}$	**36.2$_{\uparrow1.0}$**	66.7$_{\downarrow0.5}$	43.3$_{\downarrow0.2}$	47.1$_{\downarrow1.7}$	87.6$_{\uparrow0.2}$
18	57.5$_{\downarrow0.4}$	34.1$_{\downarrow1.1}$	67.4$_{\uparrow0.2}$	43.3$_{\downarrow0.2}$	47.3$_{\downarrow1.4}$	87.3$_{\downarrow0.1}$	56.9$_{\downarrow1.0}$	31.5$_{\downarrow3.7}$	66.1$_{\downarrow1.1}$	43.2$_{\downarrow0.3}$	47.3$_{\downarrow1.4}$	87.3$_{\downarrow0.1}$
19	57.5$_{\downarrow0.4}$	34.9$_{\downarrow0.3}$	67.0$_{\downarrow0.2}$	43.3$_{\downarrow0.2}$	47.3$_{\downarrow1.4}$	87.6$_{\uparrow0.1}$	55.6$_{\downarrow2.3}$	31.2$_{\downarrow4.0}$	67.2$_{0.0}$	43.4$_{\downarrow0.1}$	46.8$_{\downarrow2.0}$	87.2$_{\downarrow0.2}$
20	**58.7$_{\uparrow0.8}$**	31.7$_{\downarrow3.6}$	67.1$_{\downarrow0.1}$	43.3$_{\downarrow0.2}$	47.8$_{\downarrow0.9}$	87.4$_{0.0}$	56.9$_{\downarrow1.0}$	32.2$_{\downarrow3.0}$	67.1$_{\downarrow0.1}$	43.3$_{\downarrow0.2}$	48.1$_{\downarrow0.7}$	87.1$_{\downarrow0.3}$
21	57.7$_{\downarrow0.2}$	31.7$_{\downarrow3.5}$	66.8$_{\downarrow0.4}$	43.4$_{\downarrow0.1}$	**49.2$_{\uparrow0.4}$**	87.0$_{\downarrow0.4}$	56.4$_{\downarrow1.5}$	32.9$_{\downarrow2.3}$	65.9$_{\downarrow1.3}$	43.5$_{0.0}$	47.5$_{\downarrow1.3}$	87.3$_{\downarrow0.1}$
22	58.0$_{\uparrow0.1}$	30.3$_{\downarrow4.9}$	67.1$_{\downarrow0.1}$	43.3$_{\downarrow0.2}$	48.5$_{\downarrow0.3}$	87.8$_{\uparrow0.3}$	55.1$_{\downarrow2.8}$	30.1$_{\downarrow5.1}$	64.9$_{\downarrow2.3}$	42.9$_{\downarrow0.6}$	47.2$_{\downarrow1.6}$	86.9$_{\downarrow0.5}$
23	**58.7$_{\uparrow0.8}$**	30.2$_{\downarrow5.0}$	67.6$_{\uparrow0.4}$	**43.5$_{0.0}$**	47.5$_{\downarrow1.3}$	87.6$_{\uparrow0.2}$	56.4$_{\downarrow1.4}$	30.7$_{\downarrow4.5}$	67.5$_{\uparrow0.3}$	**43.6$_{\uparrow0.1}$**	48.8$_{0.0}$	87.6$_{\uparrow0.1}$

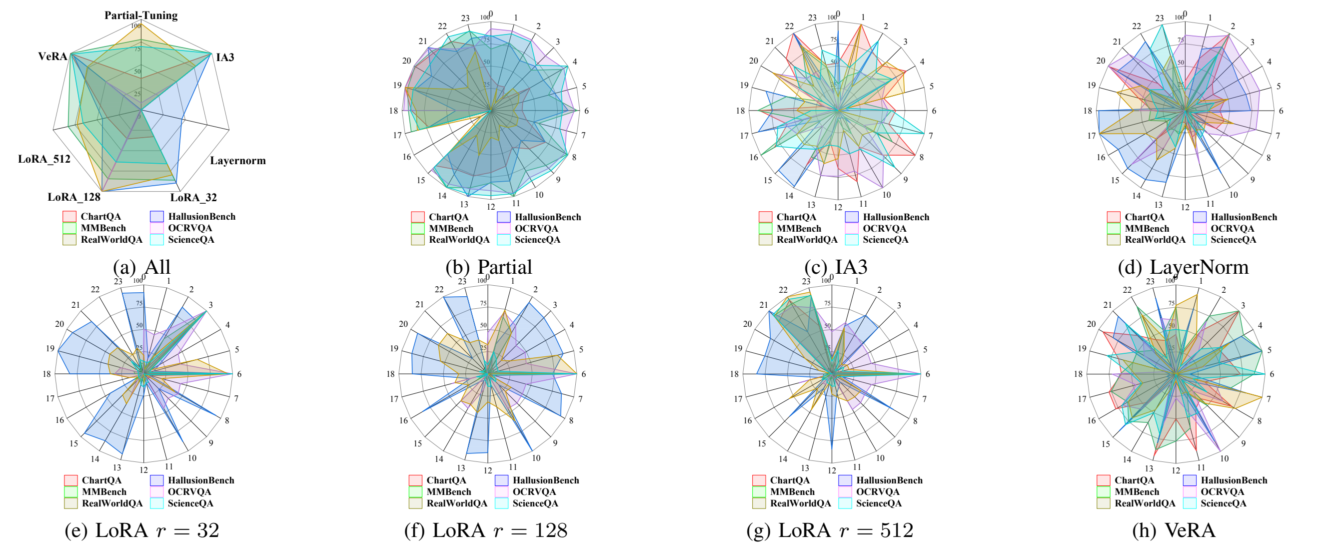

Fig. 4. Benchmark scores for all-layer tuning (a) and per-layer PEFT (b–h): Partial-Tuning, IA3, LayerNorm-Tuning, LoRA ($r = 32/128/512$), and VeRA.

TABLE IV

MMSTAR RESULTS OF LAYERNORM-TUNING, LORA ($r = 32$), AND PARTIAL-TUNING FOR EACH LAYER AND ALL-LAYER TUNING (ALL).

Layer	Layernorm-Tuning							LoRA ($r = 32$)							Partial-Tuning						
	OR	CP	FGP	IR	LR	MA	S&T	OR	CP	FGP	IR	LR	MA	S&T	OR	CP	FGP	IR	LR	MA	S&T
0	43.2	**64.0**	32.4	50.8	32.9	**42.4**	36.4	43.4	62.8	32.8	52.0	34.5	40.8	37.6	43.7	**64.0**	33.2	52.0	34.5	40.8	37.6
1	**43.8**	63.2	32.4	53.2	34.5	**42.4**	36.8	43.1	62.8	32.8	51.6	34.5	41.2	35.6	43.3	62.8	32.4	52.8	32.9	41.2	37.6
2	43.6	63.2	**34.4**	52.0	34.1	40.8	37.2	43.7	62.4	32.0	53.2	**34.9**	41.6	38.0	43.3	63.6	32.4	52.8	34.5	41.2	35.2
3	43.6	63.2	33.6	51.6	34.5	42.0	36.8	43.4	63.6	32.8	51.6	34.1	42.0	36.4	43.6	**64.0**	31.2	**53.6**	34.5	41.2	37.2
4	43.6	63.2	33.2	52.8	34.1	41.2	36.8	43.7	62.4	**35.2**	52.8	34.1	40.4	37.2	42.8	62.0	31.6	52.8	33.7	40.4	36.4
5	**43.8**	62.0	32.4	52.8	34.5	42.0	**38.8**	43.9	63.6	34.0	**53.6**	33.7	41.2	37.2	**43.8**	**64.0**	32.8	52.8	34.5	41.6	37.2
6	43.2	63.2	32.4	52.4	34.5	40.8	35.6	43.6	63.6	34.4	52.0	34.1	41.2	36.4	43.7	63.2	**34.0**	52.8	33.7	40.4	38.0
7	43.4	63.6	32.0	52.0	34.5	41.2	36.8	43.8	62.8	33.2	52.0	34.5	42.0	38.0	43.4	63.6	33.6	51.2	34.1	40.0	38.0
8	43.4	63.6	32.8	53.2	33.7	41.6	35.6	43.5	63.2	33.6	52.0	34.1	41.2	36.8	43.1	63.2	32.8	51.6	34.5	40.0	36.4
9	43.3	62.8	32.8	52.0	34.5	41.6	36.0	**44.1**	**64.0**	34.4	52.4	34.1	41.6	38.0	43.7	62.8	33.2	52.4	34.5	40.8	**38.4**
10	43.3	63.2	32.4	51.6	34.1	42.0	36.4	43.4	62.8	33.2	52.0	33.7	40.4	38.0	43.4	63.2	32.0	52.4	33.7	40.8	38.0
11	43.7	63.2	32.8	**53.6**	34.5	41.2	36.8	43.3	63.2	33.6	51.6	34.1	41.6	35.6	43.5	63.2	32.8	52.8	35.3	39.6	37.2
12	43.4	63.6	32.8	52.4	34.1	41.2	36.4	43.4	62.8	33.2	52.4	34.1	40.4	37.6	43.6	63.6	**34.0**	52.0	34.1	40.4	37.6
13	43.2	62.0	33.2	51.6	33.7	41.2	37.6	43.8	63.2	33.6	**53.6**	34.1	41.2	36.8	43.4	63.2	32.4	52.8	33.3	41.6	36.8
14	43.3	63.6	32.4	51.6	33.7	40.8	37.6	43.4	62.8	33.2	52.4	34.5	41.6	36.0	43.4	63.2	33.2	52.8	33.7	40.4	36.8
15	43.3	63.2	33.2	51.6	33.7	40.8	37.2	43.9	63.6	33.2	52.8	33.7	**42.4**	37.6	**43.8**	62.8	33.2	52.4	34.9	41.6	37.6
16	43.3	63.2	32.4	52.4	32.9	42.0	36.8	43.6	63.6	33.6	52.4	34.1	40.8	37.2	43.5	63.6	33.2	52.0	34.1	41.6	36.4
17	42.9	63.2	32.0	52.0	32.9	40.8	36.4	43.6	63.6	33.6	51.6	34.1	41.6	37.2	42.9	62.8	32.0	51.6	33.3	40.8	36.8
18	43.7	63.2	32.8	52.4	**34.9**	41.2	37.6	43.8	63.6	32.8	52.8	34.5	40.8	38.0	42.9	62.8	32.0	52.0	34.1	41.2	35.2
19	43.6	**64.0**	32.0	52.4	34.1	41.2	37.6	43.5	62.4	34.0	51.2	34.1	42.0	37.2	43.6	63.6	32.4	53.2	34.1	**42.0**	36.0
20	43.4	63.2	31.6	52.8	34.5	40.4	38.0	43.4	63.2	32.8	51.2	**34.9**	41.6	36.8	43.7	63.2	33.2	53.2	33.7	40.8	38.0
21	43.5	62.8	32.8	52.4	33.7	42.0	37.2	43.6	62.8	33.6	51.2	34.1	41.2	**38.8**	43.4	62.8	33.2	52.0	34.9	40.8	36.4
22	43.4	63.2	32.8	53.2	33.7	40.8	36.8	43.6	**64.0**	33.2	51.6	34.5	41.2	36.8	43.5	63.6	32.4	52.0	33.7	41.6	37.6
23	43.6	62.0	34.0	52.0	34.5	41.6	37.2	43.4	62.0	32.8	52.0	34.1	42.0	37.2	43.4	63.6	30.8	52.0	34.1	41.2	**38.4**
ALL	43.2	63.6	32.8	52.4	32.9	40.8	36.4	43.6	62.8	33.2	**53.6**	34.1	41.6	36.0	43.6	62.4	33.6	51.6	**35.7**	41.6	36.8

V. Conclusion

We study pre-PEFT layer selection for VLM vision encoders through Q/K/V weight statistics and perturbation probing. Across two backbones, layers that are more robust to perturbations tend to yield better PEFT outcomes, often together with larger weight norms and higher condition numbers in our evaluated settings. These results suggest that robustness-guided selection of a small subset of layers can reduce trainable parameters while remaining competitive with broader tuning in limited-data regimes.

Acknowledgment

This work is partly supported by the Project granted by the Ministry of Agriculture and Rural Affairs, the Key Research and Development Program of Heilongjiang Province under Grant Nos. 2022ZX01A22, 2021ZXJ05A03, the National Natural Science Foundation of China under Grant Nos. 62350710797, 61972114, 62106061, the National Science and Technology Major Project of China under Grant Nos. 2021ZD0110901, the collaborative innovation and promotion system of the modern agricultural industry technology for watermelon and melon in Heilongjiang Province.

References

[1] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[3] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa *et al.*, "OpenFlamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[4] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

[5] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[6] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[8] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[9] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "LLaMA-adapter: Efficient fine-tuning of language models with zero-init attention," *arXiv preprint arXiv:2303.16199*, 2023.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[12] OpenGVLab Team, "InternVL2," Online, 2024. [Online]. Available: https://internvl.github.io/blog/2024-07-02-InternVL-2.0/

[13] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[14] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-VL technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[15] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008. [Online]. Available: https://papers.neurips.cc/paper/7181-attention-is-all-you-need

[17] D. Coppola and H. K. Lee, "Investigating and unmasking feature-level vulnerabilities of cnns to adversarial perturbations," *arXiv preprint arXiv:2405.20672*, 2024.

[18] F. N. Nokabadi, J.-F. Lalonde, and C. Gagné, "Reproducibility study on adversarial attacks against robust transformer trackers," *arXiv preprint arXiv:2406.01765*, 2024.

[19] B. Wang, G. Li, X. Zhou, Z. Chen, T. Grossman, and Y. Li, "Screen2words: Automatic mobile ui summarization with multimodal learning," in *The 34th Annual ACM Symposium on User Interface Software and Technology*, 2021, pp. 498–510.

[20] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, "OCR-VQA: Visual question answering by reading text in images," in *2019 international conference on document analysis and recognition (ICDAR)*. IEEE, 2019, pp. 947–952.

[21] S. Kantharaj, R. T. K. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty, "Chart-to-text: A large-scale benchmark for chart summarization," *arXiv preprint arXiv:2203.06486*, 2022.

[22] A. D. Lindström and S. S. Abraham, "CLEVR-Math: A dataset for compositional language, visual and mathematical reasoning," *arXiv preprint arXiv:2208.05358*, 2022.

[23] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," in *Advances in Neural Information Processing Systems*, vol. 35, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html

[24] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin *et al.*, "Are we on the right way for evaluating large vision-language models?" *arXiv preprint arXiv:2403.20330*, 2024.

[25] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," *arXiv preprint arXiv:2203.10244*, 2022.

[26] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob *et al.*, "HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 375–14 385.

[27] xAI, "RealWorldQA," Hugging Face, 2024. [Online]. Available: https://huggingface.co/datasets/xai-org/RealworldQA

[28] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "MMBench: Is your multi-modal model an all-around player?" *arXiv preprint arXiv:2307.06281*, 2023.

[29] D. J. Kopiczko, T. Blankevoort, and Y. M. Asano, "VeRA: Vector-based random matrix adaptation," *arXiv preprint arXiv:2310.11454*, 2023.

[30] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1950–1965, 2022.

[31] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4918–4927.

[32] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.