Diagnosing and Repairing Visual Collapse in Compact Medical Multimodal LLMs

Jinjie Xie
Department of Computer Science
Loughborough University
J.Xie6@lboro.ac.uk

Baihua Li
Department of Computer Science
Loughborough University
B.Li@lboro.ac.uk

Qinggang Meng*
Department of Computer Science
Loughborough University
Q.Meng@lboro.ac.uk

Abstract

Compact medical multimodal models are the most realistic path to clinical deployment, yet their answers draw on the image less than their accuracy suggests, and how much they draw on it varies widely across benchmarks. We show that this weak visual use has a clear internal signature. In five compact medical backbones and across six medical VQA benchmarks, visual tokens lose their spatial diversity shortly after entering the language backbone, their residual updates are an order of magnitude smaller than text residuals, and attention on them freezes on content free regions of the image. Attention knockouts show that text positions draw most of their visual content within the first third of the backbone. We give a short analytical account that links the collapse of visual similarity to the residual dominance ratio, and we propose VGrip, a set of three lightweight losses that act on the three failure points without modifying the backbone architecture or the inference procedure. We measure visual use by comparing accuracy with the correct image, without the image and with a shuffled image, and summarise the comparison in a Visual Dependency Score. On Qwen2.5-VL-7B, VGrip lifts accuracy by 4.7 points and the Visual Dependency Score by 0.12 on average across the six benchmarks over a strong LoRA baseline, with consistent gains from two to eight billion parameters.

1 Introduction

Multimodal large language models have made rapid progress on medical image understanding, and there is growing interest in deploying them for clinical decision support [1, 2]. Real hospitals rarely allow large models to be used at full scale. On premise compute is limited, cloud inference is often forbidden by data protection rules, and patient images cannot be sent outside the institution [3, 4]. Compact models with two to seven billion parameters are therefore the most realistic path to clinical use [5]. These models are fluent and reach acceptable accuracies on standard benchmarks, but their answers depend on the image less than these accuracies suggest, and the extent of that dependence varies widely from one benchmark to another [6, 7]. For a system meant to support a radiologist this is a serious failure mode, since a confident wrong answer that looks correct is more dangerous than a refusal.

Most prior work treats this behaviour as a data or scale problem, and responds with more pretraining data, larger backbones, or specialised projectors [8, 9]. We find that scaling and extra data reduce

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

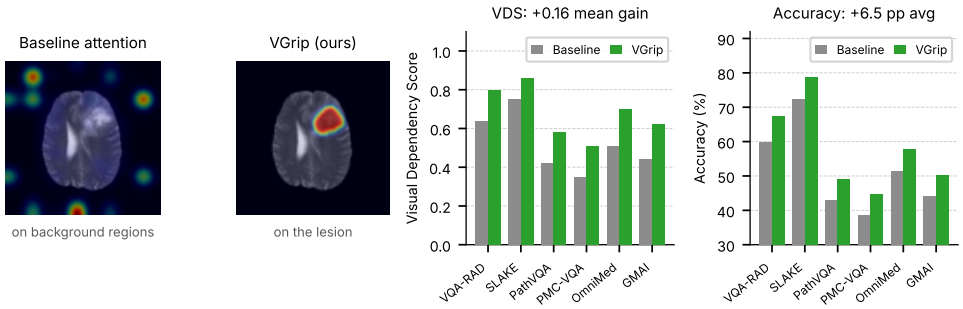

Figure 1: Compact medical multimodal models use the image, but weakly and unevenly across benchmarks. Left: The baseline spreads attention over background regions, while VGrip concentrates it on the lesion. Right: Visual Dependency Score and accuracy of the frozen Qwen2.5-VL-7B baseline and of VGrip on six medical VQA benchmarks.

the symptom but do not remove it. Something inside the model weakens the visual signal during inference, and that mechanism has not been pinned down. Instead of asking which dataset or which backbone gives the best score, we ask what happens to the visual representation as it travels through the language backbone. By measuring representation geometry, attention routing, and residual update strength layer by layer, we find a consistent picture across five compact models and six medical benchmarks. Visual tokens lose their spatial diversity shortly after they enter the backbone, their residual updates are roughly an order of magnitude smaller than text residuals, and attention on them ends up frozen on background regions of the image. Causal interventions show that text positions draw most of their visual content within the first third of the backbone, and that blocking this flow at later layers has little effect on the answer. The image enters the computation early and contributes less and less as depth increases.

A reasonable response would be that compact models simply lack the capacity for visual reasoning, and that the only fix is to make them larger. We show that this is not the case. Full fine-tuning raises accuracy but leaves the collapse signature largely in place, which points to the training dynamics as the binding constraint. Once the failure is understood as an internal dynamics problem, it can be repaired with three lightweight, mechanism guided losses that act on the same backbone without changing its architecture. The first loss preserves diversity in deep visual tokens through an effective rank penalty. The second rebalances residual updates between modalities using a small per layer scale. The third anchors shallow attention onto anatomically relevant regions. We call the combined approach VGrip. Figure 1 summarises the result. On Qwen2.5-VL-7B, VGrip lifts accuracy by 4.7 points on average across six benchmarks and raises the Visual Dependency Score, a normalised summary of how much the answer changes when image content is destroyed, from 0.56 to 0.68.

Contributions. We make three contributions. First, we provide a layer level account of weak visual use in compact medical multimodal models, showing that representation collapse, residual asymmetry, and frozen attention form a self reinforcing loop, and that this loop is consistent across five models and six benchmarks. Second, we introduce VGrip, a set of three lightweight losses that target the three failure points identified in the analysis, together with a short theoretical argument that links the diversity penalty to the residual dominance ratio. Third, we report visual use on six medical VQA benchmarks through accuracy with the correct image, without the image and with a shuffled image, and summarise it with the Visual Dependency Score. VGrip improves accuracy by 4.7 points and Visual Dependency Score by 0.12 on average over a strong LoRA baseline, with consistent gains from two to eight billion parameters.

2 Related Work

Medical multimodal models and modality bias. A growing line of work adapts general purpose vision language models to medical use through instruction tuning on radiology and pathology data [10, 11, 12], and recent systems push the backbone size down to two or seven billion parameters to fit hospital hardware [5, 13, 14]. These models reach competitive scores on VQA-RAD and SLAKE,

but several studies note that their answers are not always supported by the image [15, 16]. Asadi et al. [17] document the same effect at benchmark scale on frontier models, which keep 70 to 80 percent of their accuracy on standard multimodal benchmarks, medical VQA included, when the image is removed, and respond by removing text solvable questions from the evaluation. Their work fixes the measurement and ours fixes the model, by tracing why the failure arises inside compact open backbones and repairing it during training. The Visual Dependency Score we report belongs to the same family of delta metrics that compare accuracy with and without the image. The tendency to lean on language priors is well known in general visual question answering [18, 19]. Existing responses include causal reweighting [6, 7], data side fixes such as remixing [8, 9], and decoding time corrections that subtract a text only branch [19], contrast the output against a distorted image [20], or penalise over trust in a few summary tokens during beam search [21]. These methods act on the data, the loss weighting or the decoding distribution, and leave the internal dynamics of the forward pass untouched. We compare against the three decoding time corrections in Section 6.3. The compact medical setting we study turns out to be a strong stress test for these approaches, since neither extra data nor larger backbones fully remove the failure.

Mechanistic analysis and representation collapse. A separate body of work studies transformers through their internal geometry. Neural collapse and feature collapse have been analysed in classification and more recently in multimodal fusion [22, 23, 24]. Attention rollout [25] and residual stream analysis [26] let one trace how predictions are built up across layers. For multimodal models specifically, Yin [27] report that visual information becomes compressed into a small set of tokens after fusion, and Shukor [28] show that visual influence does not automatically grow with model scale. We build on these tools and apply them to compact medical multimodal models. The diversity preserving loss in our method is related to the effective rank regularisers used in self supervised learning [29, 30], and we apply it selectively to deep visual tokens inside the language backbone.

3 Mechanistic Diagnosis of Weak Visual Use

This section looks inside compact medical multimodal models to see what happens to the visual signal during inference. We trace the failure from behavioural evidence to internal dynamics and find a consistent picture across models and datasets.

3.1 Behavioural Evidence

We study Qwen2.5-VL-3B and Qwen2.5-VL-7B [13] as main backbones, and InternVL2.5-4B [14], LLaVA-Med-7B [10], and HuatuoGPT-Vision-7B [11] for cross model validation. All models are frozen during the diagnosis. We use VQA-RAD [31] and SLAKE [32], and evaluate each model under three input conditions. The **image plus text** condition is standard. In the **text only** condition the image and its tokens are removed from the input, so no visual positions enter the backbone. We do not substitute a blank image, since a blank image still produces visual tokens and would mix this condition with the shuffled one. In the **shuffled image** condition the visual tokens are randomly permuted so that their distribution matches the original but their content is destroyed.

Table 1 reports accuracy of the five models under these conditions. On VQA-RAD, Qwen2.5-VL-7B reaches 59.7 percent with image plus text, 39.1 percent with text only, and 43.8 percent with a shuffled image. Text only accuracy sits a few points above the random baseline on both benchmarks, so the image supplies most of what the model knows beyond guessing. A shuffled image recovers 18 to 30 percent of the image to text gap on VQA-RAD and 6 to 12 percent on SLAKE, so most of the gain depends on the content of the image. The medical specialists LLaVA-Med-7B and HuatuoGPT-Vision-7B reach the highest image plus text accuracy, but their lead over Qwen2.5-VL-7B is already present in the text only setting, where they score 5.4 and 7.1 points higher on VQA-RAD.

To make this observation quantitative we define the **Visual Dependency Score**, which we use throughout the paper. Let a_{it} be accuracy with the correct image, a_{sh} accuracy with a shuffled image, and a_{rand} the random baseline. We define

$$\text{VDS} = \frac{a_{\text{it}} - a_{\text{sh}}}{a_{\text{it}} - a_{\text{rand}}}. \tag{1}$$

A model that makes full use of the image has a VDS close to one, since shuffling should drop performance to the random baseline. A model that ignores the image has a VDS close to zero. The

Table 1: Behavioural evidence of visual use across compact medical MLLMs. We report accuracy (%) under three input conditions: standard **Image+Text** (I+T), **Text-only** (T), and **Shuffled-image** (I_{shuf}+T). The gap between I+T and I_{shuf}+T measures how much the prediction depends on the content of the image. The Visual Dependency Score (VDS, Equation 1) normalises this gap by the headroom above the random baseline, with $a_{rand} = 0.35$ on VQA-RAD and 0.40 on SLAKE.

	VQA-RAD				SLAKE			
Model	I+T	T	I_{shuf}+T	VDS	I+T	T	I_{shuf}+T	VDS
Qwen2.5-VL-3B	56.2	37.8	41.2	0.71	68.8	43.2	45.8	0.80
Qwen2.5-VL-7B	59.7	39.1	43.8	0.64	72.4	46.5	48.1	0.75
InternVL2.5-4B	55.1	36.2	40.5	0.72	67.1	41.8	44.2	0.85
LLaVA-Med-7B	63.4	44.5	50.1	0.47	76.5	52.8	55.6	0.57
HuatuoGPT-V-7B	64.8	46.2	51.5	0.45	77.8	54.1	57.0	0.55
Average	59.8	40.8	45.4	0.60	72.5	47.7	50.1	0.70

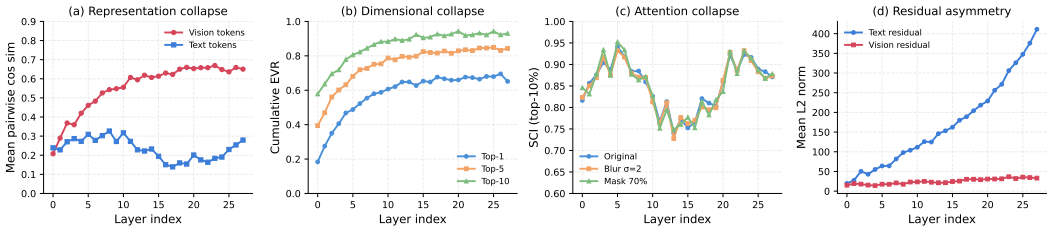

Figure 2: Layerwise diagnosis on Qwen2.5-VL-7B. (a) Vision tokens lose pairwise diversity while text tokens do not. (b) The top principal components explain most of the variance in deep layers. (c) Attention on visual tokens stays frozen under blur and masking. (d) Text residual updates dominate vision residuals by roughly an order of magnitude.

VDS columns of Table 1 range from 0.45 to 0.85, and every model scores higher on SLAKE than on VQA-RAD. The two medical specialists score lowest, at 0.45 to 0.57, because their shuffled image gap is smaller while their higher image plus text accuracy enlarges the denominator. Its main use is comparison across benchmarks with different headroom, so we report it as a summary next to the three raw accuracies. VDS measures how much the answer depends on image content and says nothing about whether that content is read correctly, a point we return to in Section 7.

3.2 Representation Collapse in the Language Backbone

To find where the visual signal weakens we follow the visual tokens through the language backbone and measure their geometry layer by layer. Figure 2 summarises four measurements on Qwen2.5-VL-7B.

Panel (a) shows the mean pairwise cosine similarity of visual tokens rising from 0.22 to above 0.6 within ten layers, while text tokens stay between 0.18 and 0.28. Panel (b) confirms this as dimensional compression, with the top principal component alone explaining more than sixty percent of deep visual variance. Panel (c) plots the Spatial Collapse Index, the share of attention mass in the top ten percent of visual positions, which stays between 0.85 and 0.97 and is nearly identical under strong blur or heavy masking. Panel (d) shows that text residual updates exceed vision residuals by an order of magnitude from layer five onwards. The last panel explains the first, since a shared update across all visual positions moves their mean without increasing their spread. High similarity alone could reflect benign abstraction, so we also probe what the deep visual tokens still encode. On the LoRA baseline, a linear probe on final layer visual tokens recovers patch row index at 32.4 percent top-1 against 3.1 percent chance, and organ identity at 44.8 percent against 8.3 percent chance, compared with 74.9 and 79.4 percent at shallow layers. The collapse removes spatial information that the shallow tokens still carry. These four effects reproduce on all five backbones and both datasets (Figure 3).

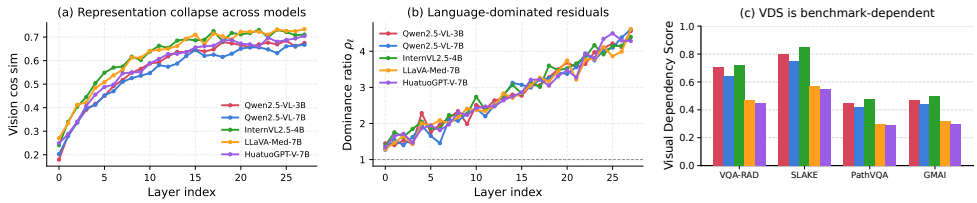

Figure 3: Cross model consistency of the diagnosis. (a) Vision token similarity across layers. (b) Residual dominance ratio. (c) Visual Dependency Score on four benchmarks. The collapse in (a) and the residual imbalance in (b) appear on all five compact backbones, while VDS in (c) depends strongly on the benchmark.

3.3 Causal Intervention

To test how much the visual pathway contributes to the prediction, we intervene on the forward pass with two knockouts. The residual knockout zeroes either the visual or the text residual update at layer ℓ. The attention knockout blocks all attention edges from text positions to visual positions from layer ℓ onward, which tests whether text positions still draw new content from the image at that depth.

Removing the visual residual update has almost no effect beyond the very first layer. Removing the text residual update has the opposite effect, with a drop of more than forty points at shallow layers and more than twenty points at late layers. The attention knockout explains this asymmetry. Blocking text to visual attention from layer 1 drops VQA-RAD accuracy from 59.7 to 41.8 percent, close to the text only baseline of 39.1. The drop shrinks with depth, to 51.3 at layer 5, 57.4 at layer 10 and 59.1 at layer 15, and is within noise beyond that. Most of the visual content reaches text positions within the first third of the backbone, and the residual knockout reads the tail of this flow. Visual information enters the computation early, and the deep visual tokens contribute little to the prediction once it has been copied into text positions.

Figure 3 confirms that this picture is not a quirk of one model. Across all five compact backbones the visual cosine similarity rises above 0.65 and the residual dominance ratio stays above one almost everywhere. The Visual Dependency Score varies substantially across benchmarks, with the highest values on VQA-RAD and SLAKE and lower values on PathVQA and GMAI-MMBench, and the two medical specialists score lowest on every benchmark.

3.4 Mechanistic Summary

The three effects form a self reinforcing loop. A shared residual update pushes the mean of the visual tokens without increasing their spread, so pairwise diversity collapses. Attention on the now similar tokens saturates onto a fixed set of positions, and text residuals continue to dominate because no new content flows back. The knockouts show the end result. Text positions read the image in the first layers, and the rest of the prediction is built mainly on the text pathway. The three entry points of the loop, namely representation diversity, residual balance, and attention anchoring, are the three targets of our method.

4 Theoretical Analysis

This section gives a short analytical account of the collapse observed in Figure 2 and uses it to motivate the first component of our method. It explains one specific observation, namely that a shared residual update across visual tokens drives their pairwise similarity toward one, and it predicts that the rate of this collapse is controlled by a quantity we can measure directly.

Let $v_1^{(\ell)}, \dots, v_{n_v}^{(\ell)} \in \mathbb{R}^d$ be the visual tokens at layer ℓ and let $\mu^{(\ell)}$ be their mean. Figure 2(d) shows that visual residual updates are small in norm, and they are also highly correlated across tokens. This motivates writing the update of each visual token as

$$\Delta v_i^{(\ell)} \;=\; a^{(\ell)} + \xi_i^{(\ell)}, \qquad \frac{1}{n_v}\sum_{i=1}^{n_v} \xi_i^{(\ell)} = 0, \tag{2}$$

where $a^{(\ell)}$ is the common component shared across all visual tokens and $\xi_i^{(\ell)}$ is a small token specific residual with $\|\xi_i^{(\ell)}\|_2 \ll \|a^{(\ell)}\|_2$. Empirically the norm of the common direction is an order of magnitude larger than the typical token specific residual across all models in our study.

Under this decomposition, the mean $\mu^{(\ell)}$ accumulates the shared updates $a^{(\ell)}$, while the centred tokens $\delta_i^{(\ell)} = v_i^{(\ell)} - \mu^{(\ell)}$ evolve only through the small residuals $\xi_i^{(\ell)}$. Let s_ℓ denote the mean pairwise cosine similarity of the visual tokens at layer ℓ,

$$s_\ell \;=\; \frac{1}{n_v(n_v-1)} \sum_{i \neq j} \frac{\langle v_i^{(\ell)}, v_j^{(\ell)} \rangle}{\|v_i^{(\ell)}\|_2 \, \|v_j^{(\ell)}\|_2}. \tag{3}$$

Expanding the cosine around the shared mean direction to second order gives

$$s_\ell \;=\; 1 - \frac{1}{2\,\|\mu^{(\ell)}\|_2^2} \cdot \frac{1}{n_v(n_v-1)} \sum_{i \neq j} \big\|\delta_i^{(\ell)} - \delta_j^{(\ell)}\big\|_2^2 \;+\; \mathcal{O}\big(\eta_\ell^4\big), \tag{4}$$

where $\eta_\ell = \max_i \|\delta_i^{(\ell)}\|_2 / \|\mu^{(\ell)}\|_2$. The full derivation is in Appendix A. The intuition is that $\|\mu^{(\ell)}\|_2^2$ grows as the sum of shared updates and pushes s_ℓ toward one, while the centred spread in the numerator acts as a brake. If the spread stays large, collapse is slow. If it shrinks, collapse is fast.

Equation 4 predicts that s_ℓ rises fastest at layers where the visual update adds little token specific spread relative to the drift of the mean. A convenient and measurable proxy for this regime is the residual dominance ratio

$$\rho_\ell \;=\; \frac{\|\Delta T^{(\ell)}\|_F}{\|\Delta V^{(\ell)}\|_F}, \tag{5}$$

where $\Delta T^{(\ell)}$ and $\Delta V^{(\ell)}$ are the residual updates on the text and vision substreams. A large ρ_ℓ marks layers where the visual update is small next to the text update and dominated by its shared component, so the centred spread barely grows and s_ℓ should rise quickly. We verify this prediction in Appendix A.4, where the layer to layer growth of s_ℓ correlates positively with ρ_ℓ on all three backbones we test, with Pearson correlations between 0.57 and 0.64.

Implication for the method. Since $\|\mu^{(\ell)}\|_2$ grows mechanically with layer index and is hard to control, the most direct way to slow the collapse is to keep the centred spread of the visual tokens large. A short calculation shows that this spread is lower bounded by a monotone increasing function of the effective rank of the centred visual token matrix. This points directly at a log determinant penalty on the visual gram matrix, which is the first component of our method described next.

5 VGrip: Repairing Visual Collapse

The diagnosis in Section 3 identified three failure points. Deep visual tokens collapse onto a few directions, residual updates are dominated by the text stream, and attention on visual tokens freezes on content free regions. VGrip adds three lightweight losses that target these three points. It leaves the backbone architecture and the inference procedure unchanged, and adds no inference time cost. Figure 4 gives an overview.

5.1 Deep Visual Diversity Regularisation

The first component targets the representation collapse in Figure 2(a,b). Let $\mathcal{D}$ be the set of deep layers on which the regulariser acts. In practice we use the last third of the backbone, where visual similarity has already saturated, and this band outperforms the first third, the middle third and all layers (Appendix C.1). For a layer $\ell \in \mathcal{D}$, let $V^{(\ell)} \in \mathbb{R}^{d \times n_v}$ be the matrix of centred visual tokens. Following Section 4, we penalise any loss of effective rank through a log determinant term on the gram matrix,

$$\mathcal{L}_{\text{DVDR}} \;=\; -\,\frac{1}{|\mathcal{D}|} \sum_{\ell \in \mathcal{D}} \log\det\Big(\tfrac{1}{n_v}\, V^{(\ell)\top} V^{(\ell)} + \varepsilon I_{n_v}\Big), \tag{6}$$

where $\varepsilon = 10^{-4}$ keeps the matrix positive definite. Minimising $\mathcal{L}_{\text{DVDR}}$ pushes the singular values of $V^{(\ell)}$ away from zero, which is exactly the condition the theory calls for. DVDR alone lifts accuracy from 62.4 to 64.1 and VDS from 0.68 to 0.72 on VQA-RAD.

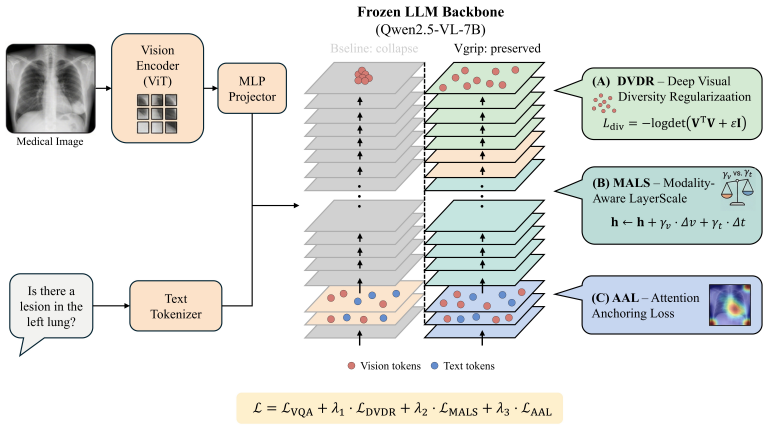

Figure 4: Overview of VGrip. Three lightweight losses are added to LoRA fine-tuning of a frozen backbone. DVDR preserves deep visual diversity, MALS rebalances residual updates between modalities, and AAL anchors shallow attention onto informative regions.

5.2 Modality Aware LayerScale

DVDR does not fix the underlying asymmetry in residual updates shown in Figure 2(d). If text residuals continue to grow an order of magnitude faster than vision residuals, DVDR has to work against a constant headwind. We introduce a per layer pair of learnable scalars $\gamma_v^{(\ell)}, \gamma_t^{(\ell)} \in \mathbb{R}_+$ and rescale the residual updates separately for each modality,

$$h_i^{(\ell+1)} \;=\; h_i^{(\ell)} + \gamma_{m(i)}^{(\ell)} \cdot F_\ell\Big(h^{(\ell)}\Big)_i, \qquad m(i) \in \{v, t\}, \tag{7}$$

where F_ℓ is the transformer block and $m(i)$ is the modality of token i. This is a modality aware variant of LayerScale [33]. We initialise both scalars at one and add a soft margin loss that discourages the text path from dominating,

$$\mathcal{L}_{\text{MALS}} \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \max\Big(0,\; \log \tilde{\rho}^{(\ell)} - \tau\Big)^2, \qquad \tilde{\rho}^{(\ell)} \;=\; \frac{\gamma_t^{(\ell)}\,\|\Delta T^{(\ell)}\|_F}{\gamma_v^{(\ell)}\,\|\Delta V^{(\ell)}\|_F}, \tag{8}$$

with $\tau = \log 2$. The loss is zero as long as the effective dominance ratio stays within a factor of two. MALS introduces just $2L$ parameters, and after training it amplifies the vision stream while leaving the text stream nearly unchanged (Appendix G).

5.3 Attention Anchoring Loss

The third component targets the frozen attention pattern in Figure 2(c). We want to nudge attention on visual tokens toward content rich regions without forcing a specific location. We use the norm of the visual token as a cheap and annotation free proxy for informativeness. In the radiology images that make up most of our benchmarks, the region of interest sits on a dark and uniform background, and background tokens have clearly lower norm than tokens on anatomy. Let $\bar{a}^{(\ell,h)} \in \mathbb{R}^{n_v}$ be the attention on visual tokens at layer ℓ and head h, averaged over queries, and define the anchor distribution

$$p_i^{(\ell)} \;=\; \frac{\|v_i^{(\ell)}\|_2}{\sum_{j=1}^{n_v} \|v_j^{(\ell)}\|_2}. \tag{9}$$

The attention anchoring loss is the KL divergence between the attention distribution and the anchor, averaged over the shallow layers $\mathcal{S}$ where attention is still malleable and over the heads,

$$\mathcal{L}_{\text{AAL}} \;=\; \frac{1}{|\mathcal{S}|\,H} \sum_{\ell \in \mathcal{S}} \sum_{h=1}^{H} \text{KL}\Big(\bar{a}^{(\ell,h)} \,\Big\|\, p^{(\ell)}\Big). \tag{10}$$

We set $\mathcal{S}$ to the first third of the backbone, where attention is still malleable. Anchoring deeper layers is less effective, since attention there has already frozen. AAL requires no segmentation masks and becomes inactive once attention already aligns with informative regions.

Table 2: Main results on six medical VQA benchmarks with Qwen2.5-VL-7B as the backbone. We report accuracy (%) and the Visual Dependency Score (VDS). **Baseline** is the frozen pretrained model; **LoRA-FT** applies standard LoRA fine-tuning; **+DVDR** adds the deep visual diversity regularizer; **+DVDR+MALS** additionally applies modality-aware LayerScale; **VGrip (full)** further adds the attention anchoring loss. OmniMed and GMAI denote OmniMedVQA and GMAI-MMBench. VDS uses a_{rand} of 0.35, 0.40, 0.38, 0.28, 0.30 and 0.32 on the six benchmarks in column order. Best per column in **bold**.

	VQA-RAD		SLAKE		PathVQA		PMC-VQA		OmniMed		GMAI		Average	
Method	Acc	VDS	Acc	VDS	Acc	VDS	Acc	VDS	Acc	VDS	Acc	VDS	Acc	VDS
Baseline (frozen)	59.7	0.64	72.4	0.75	42.8	0.42	38.5	0.35	51.3	0.51	44.1	0.44	51.5	0.52
LoRA-FT	62.4	0.68	73.8	0.77	44.6	0.46	40.1	0.39	53.0	0.55	45.8	0.48	53.3	0.56
+DVDR	64.1	0.72	75.3	0.80	46.0	0.50	41.5	0.42	54.7	0.60	47.2	0.53	54.8	0.60
+DVDR+MALS	65.9	0.76	77.1	0.83	47.8	0.54	43.2	0.46	56.5	0.65	48.9	0.57	56.6	0.64
VGrip (full)	**67.4**	**0.80**	**78.6**	**0.86**	**49.1**	**0.58**	**44.7**	**0.51**	**57.8**	**0.70**	**50.2**	**0.62**	**58.0**	**0.68**
Δ over LoRA	+5.0	+0.12	+4.8	+0.09	+4.5	+0.12	+4.6	+0.12	+4.8	+0.15	+4.4	+0.14	+4.7	+0.12

5.4 Full Objective

VGrip is trained on top of standard rank 16 LoRA adapters. Only the adapter parameters and the $2L$ MALS scalars are updated. The total objective is

$$\mathcal{L} \;=\; \mathcal{L}_{\text{VQA}} + \lambda_1 \, \mathcal{L}_{\text{DVDR}} + \lambda_2 \, \mathcal{L}_{\text{MALS}} + \lambda_3 \, \mathcal{L}_{\text{AAL}}, \tag{11}$$

where $\mathcal{L}_{\text{VQA}}$ is the standard next token cross entropy loss. We use $\lambda_1 = 0.05$, $\lambda_2 = 0.01$, and $\lambda_3 = 0.02$ throughout. VGrip adds no inference time cost, since the $\gamma^{(\ell)}$ scalars fold into the residual updates after training and the other two losses act only during training. On one A100 with Qwen2.5-VL-7B, VGrip increases training time per step by 11.5 percent and peak memory by 3.5 GB over plain LoRA. Full training details and a per loss breakdown are in Appendix B.

6 Experiments

We evaluate VGrip on six medical visual question answering benchmarks with five compact backbones. The goals are to show that VGrip improves accuracy and makes answers depend more on image content, that it outperforms decoding time corrections, that each of its three components contributes, that the improvement transfers across benchmarks and model sizes, and that the mechanistic findings of Section 3 are actually repaired.

6.1 Setup

We use VQA-RAD [31], SLAKE [32], PathVQA [34], PMC-VQA [35], OmniMedVQA [36], and GMAI-MMBench [37] under their official splits. Our main backbone is Qwen2.5-VL-7B [13], and for scaling and cross architecture comparison we also use Qwen2.5-VL-3B, InternVL2.5 at 2B/4B/8B [14], LLaVA-Med-7B [10], and HuatuoGPT-Vision-7B [11]. We compare VGrip against the frozen pretrained model, standard LoRA-FT, and two partial variants +DVDR and +DVDR+MALS, all using rank 16 LoRA with AdamW, peak learning rate 2×10^{-4}, batch size 32, three epochs, and three seeds.

6.2 Main Results

Table 2 reports accuracy and Visual Dependency Score for all five methods on all six benchmarks with Qwen2.5-VL-7B.

Three patterns stand out. First, VGrip lifts accuracy on every benchmark. The gain over the LoRA baseline averages 4.7 points, ranging from 4.4 on GMAI-MMBench to 5.0 on VQA-RAD. On SLAKE the full method reaches 78.6 percent, above the frozen medical specialists in Table 1, which were trained on much larger amounts of paired medical data. Second, VDS rises on every benchmark, from 0.56 to 0.68 on average, with the largest gains on OmniMedVQA and GMAI-MMBench and the smallest on SLAKE, where the LoRA baseline already scores 0.77. The accuracy gains therefore

Table 3: Comparison with decoding time corrections on Qwen2.5-VL-7B, accuracy (%). All corrections are applied to the LoRA-FT model under the protocol of Table 2.

Method	VQA-RAD	SLAKE
LoRA-FT	62.4	73.8
+ VCD [20]	63.6	75.1
+ OPERA [21]	63.9	75.4
+ CF-VQA style [19]	64.5	76.0
VGrip	67.4	78.6
VGrip + VCD	**68.0**	**79.1**

Table 4: Out of distribution accuracy (%) with Qwen2.5-VL-7B. VGrip improves on every setting, with the largest gain on cross hospital scans.

Method	In-dist. (VQA-RAD)	Cross-hospital	Corrupted scans	Unseen modality	Avg.
LoRA-FT	62.4	48.2	52.1	41.3	51.0
VGrip	**67.4**	**58.6**	**60.9**	**49.7**	**59.2**
Δ	+5.0	+10.4	+8.8	+8.4	+8.2

come with a stronger dependence on image content, which better language priors alone would not produce. Third, the three components contribute monotonically. Each row in Table 2 improves on the one above it, in both accuracy and VDS, without any regression on any benchmark.

6.3 Comparison with Decoding Time Corrections

Table 3 compares VGrip with three corrections that act at inference time on the LoRA-FT model: visual contrastive decoding (VCD) [20], OPERA [21], which we port to Qwen2.5-VL, and a CF-VQA style logit correction that subtracts the text only branch [19]. All three improve on LoRA-FT and all three fall short of VGrip. They act on the decoding distribution and leave the internal dynamics of the forward pass untouched, while VGrip changes those dynamics during training. Adding VCD on top of VGrip gives a further 0.6 and 0.5 points, so the two are complementary.

6.4 Component Analysis, Sensitivity, and Scaling

We isolate the contribution of each VGrip component through a 2^3 ablation on VQA-RAD, with full results in Appendix C. DVDR is the single most effective component on its own, raising accuracy from 62.4 to 64.1 and VDS from 0.68 to 0.72, which is consistent with the prediction in Section 4 that preserving deep visual diversity slows the collapse of s_ℓ. MALS adds 1.2 points and AAL adds 0.7 points on their own. Any pair of components recovers more than the sum of the individual gains because the three failure modes they target are coupled. The full triple recovers a 5.0 point gain in accuracy and a 0.12 gain in VDS. Sweeping the DVDR weight λ_1 gives a broad plateau between $\lambda_1 = 0.025$ and $\lambda_1 = 0.1$ within which accuracy stays within 1.2 points and VDS within 0.05 of their best values.

We also test whether the gains disappear at larger scale. On five backbones from 2B to 8B, the accuracy gain of VGrip over LoRA stays between 4.7 and 5.8 points and narrows slightly with size, and the VDS gain stays between 0.11 and 0.15. Larger backbones reduce the gap without closing it, and VGrip on Qwen2.5-VL-3B already outperforms the LoRA baseline on Qwen2.5-VL-7B (64.5 against 62.4), which is consistent with weak visual use being a dynamics problem more than a capacity problem.

Replacing LoRA with full fine-tuning raises accuracy to 64.8 but leaves the collapse largely in place, with final layer visual similarity at 0.63 against 0.67, and adding VGrip on top reaches 68.9 with a similarity of 0.32.

6.5 Robustness and Generalisation

A model that depends on the image should respond to changes in the image. Under increasing Gaussian blur the LoRA baseline loses 4.9 points, most of it at the mildest level, and then stays nearly flat, while VGrip drops by 15.3 points and falls below the baseline beyond $\sigma = 2$ (Appendix D.1). On clean images VGrip is 5.0 points ahead. Table 4 reports three distribution shift settings, cross hospital scans, corrupted scans and unseen imaging modalities. VGrip improves accuracy on every setting, and the gain is largest on cross hospital scans, where the question distribution is unchanged and only the images shift.

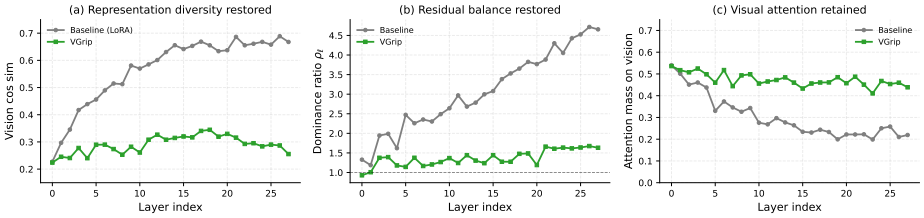

Figure 5: Mechanistic verification of the repair on Qwen2.5-VL-7B. (a) Vision token cosine similarity across layers. VGrip keeps the final layer value near 0.30, while the LoRA baseline collapses to 0.67. (b) Residual dominance ratio ρ_ℓ. VGrip holds the ratio near one across all layers while the baseline drifts toward three. (c) Fraction of attention mass on visual tokens. VGrip preserves visual attention above 0.45 while the baseline lets it decay below 0.2.

Text solvable questions. We apply the B-Clean procedure of Asadi et al. [17] to VQA-RAD, removing every question that any candidate model answers correctly without the image, which leaves 238 questions. On this subset LoRA-FT scores 58.9 and VGrip 65.7, so the gap widens from 5.0 to 6.8 points. Removing text solvable questions penalises the model that relies on them more heavily.

6.6 Mechanistic Verification

If VGrip works for the reasons we claim, the three failure modes of Section 3 should be measurably repaired. Figure 5 compares the LoRA baseline and VGrip on Qwen2.5-VL-7B. VGrip keeps the final layer visual cosine similarity at 0.30 against 0.67, holds the residual dominance ratio near one instead of letting it drift toward three, and keeps the attention mass on visual tokens above 0.45 while the baseline lets it decay below 0.2. The repaired tokens also carry content. Under VGrip a linear probe on final layer visual tokens recovers patch row index at 71.6 percent and organ identity at 76.2 percent, against 32.4 and 44.8 percent for the LoRA baseline and close to the shallow layer values of 74.9 and 79.4 percent.

7 Discussion and Conclusion

Compact medical multimodal models use the image, but less than their accuracy suggests and unevenly across benchmarks. Inside the language backbone, visual tokens lose spatial diversity shortly after fusion, their residual updates are an order of magnitude smaller than text residuals, and their attention freezes on content free regions. The collapse persists under full fine-tuning and across model sizes, which points to the training dynamics of the backbone more than to its capacity. VGrip targets each failure point with a lightweight loss and improves accuracy by 4.7 points and VDS by 0.12 over a strong LoRA baseline on six medical VQA benchmarks.

The Visual Dependency Score is a necessary condition for grounding. It measures how much the answer changes when image content is destroyed, so a model that misreads the image can score as high as one that reads it correctly, and it belongs next to the raw accuracies, accuracy under distribution shift and expert review.

Limitations. Our analysis covers models from two to eight billion parameters, and the gain of VGrip narrows slightly with size. VGrip gives the model no explicit incentive to abstain when the image is absent or unreadable. Our benchmarks are in English and carry known demographic and equipment biases, and we make no claim of clinical readiness. Further limitations are in Appendix H.

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge the financial support of the UK Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/X035433/1].

References

[1] Xiang Li, Like Li, Yuchen Jiang, Hao Wang, Xinyu Qiao, Ting Feng, Hao Luo, and Yong Zhao. Vision-language models in medical image analysis: From simple fusion to general large models. *Information Fusion*, 118:102995, 2025.

[2] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

[3] Béria Chingnabé Kalpélbé, Angel Gabriel Adaambiik, and Wei Peng. Vision language models in medicine. *arXiv preprint arXiv:2503.01863*, 2025.

[4] Mirja Mittermaier, Marium Raza, and Joseph C Kvedar. Collaborative strategies for deploying ai-based physician decision support systems: challenges and deployment approaches. *NPJ Digital Medicine*, 6(1):137, 2023.

[5] Jinwei Ye, Yaokang Wang, Weibin Kong, Gengshen Wu, and Wenjian Liu. Medlite: A lightweight medical multimodal model based on knowledge distillation and inference optimization. In *International Conference on Intelligent Computing*, pages 435–445. Springer, 2025.

[6] Ali Vosoughi, Shijian Deng, Songyang Zhang, Yapeng Tian, Chenliang Xu, and Jiebo Luo. Cross modality bias in visual question answering: A causal view with possible worlds vqa. *IEEE Transactions on Multimedia*, 26:8609–8624, 2024.

[7] Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, 2024.

[8] Jie Ma, Pinghui Wang, Dechen Kong, Zewei Wang, Jun Liu, Hongbin Pei, and Junzhou Zhao. Robust visual question answering: Datasets, methods, and future challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5575–5594, 2024.

[9] Xiaoyu Ma, Hao Chen, and Yongjian Deng. Improving multimodal learning balance and sufficiency through data remixing. *arXiv preprint arXiv:2506.11550*, 2025.

[10] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[11] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370, 2024.

[12] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner (2023). *URL: https://arxiv. org/abs/2307.15189*, 2023.

[13] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[14] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei 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*, pages 24185–24198, 2024.

[15] Robert Kaczmarczyk, Theresa Isabelle Wilhelm, Ron Martin, and Jonas Roos. Evaluating multimodal ai in medical diagnostics. *NPJ Digital Medicine*, 7(1):205, 2024.

[16] Shohreh Ghorbani, Chenyu Zhang, Minsol Kim, and Jingyao Wu. Beyond accuracy: A diagnostic protocol for fairly evaluating multimodal reasoning. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

[17] Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[19] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710, 2021.

[20] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882. IEEE, 2024.

[21] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427. IEEE, 2024.

[22] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[23] Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse. *arXiv preprint arXiv:2505.22483*, 2025.

[24] Xin An, Peng Zhao, Guiqin Wang, Cong Zhao, and Shusen Yang. Transformer feature collapse of temporal action detection via multi-granularity semantic enhancement. *Neurocomputing*, 626:129543, 2025.

[25] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4190–4197, 2020.

[26] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

[27] Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in mllms: unlocking pathways to faster inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9382–9391, 2025.

[28] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–23, 2025.

[29] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[30] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[31] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

[32] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[33] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021.

[34] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[35] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[36] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

[37] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

[38] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[40] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Appendix

A. Derivations for Section 4 15
A.1. Empirical Support for the Shared Update Model 15
A.2. Sketch of Equation 4 15
A.3. Why an Effective Rank Penalty 15
A.4. Empirical Verification of the Correlation Prediction 16
B. Implementation Details 16
B.5. Training Cost 17
C. Component Ablation, Sensitivity, and Scaling 17
C.1. Component Ablation 17
C.2. Sensitivity to the DVDR Weight 18
C.3. Scaling Across Backbone Sizes 19
C.4. Full Fine-Tuning 19
D. Robustness and Mechanistic Repair 19
D.1. Robustness to Image Perturbations and OOD Generalisation 20
D.2. Mechanistic Repair 20
E. Cross Model Diagnosis 21
E.5. Knockout Interventions 23
F. Behavioural Details 23
F.1. Full Behavioural Evaluation 23
F.2. Per Question Type Analysis 24
F.3. Closer Look at the Shuffled Image Baseline 24
F.4. VDS Sensitivity to the Random Baseline 25
G. MALS Dynamics and Language Preservation 25
H. Additional Limitations 27

A Derivations for Section 4

This appendix expands the argument of Section 4. We verify empirically that the shared update model is a reasonable description of what we observe, sketch the second order expansion that leads to Equation 4, and explain why a log determinant penalty is a natural surrogate for keeping the visual effective rank large.

A.1 Empirical Support for the Shared Update Model

The derivation assumes that the visual residual update decomposes as $\Delta v_i^{(\ell)} = a^{(\ell)} + \xi_i^{(\ell)}$, with $\|\xi_i^{(\ell)}\|_2 \ll \|a^{(\ell)}\|_2$. We measure the ratio $\|a^{(\ell)}\|_2 / \mathrm{mean}_i \|\xi_i^{(\ell)}\|_2$ on Qwen2.5-VL-7B, InternVL2.5-4B, and LLaVA-Med-7B. Across the three models and across layers the ratio stays between 8 and 23, with a median near 12. The common component is consistently an order of magnitude larger than the typical token specific residual, which is what the derivation below needs.

A.2 Sketch of Equation 4

Write each visual token as $v_i = \mu + \delta_i$ where μ is the visual mean and δ_i is its centred part with $\sum_i \delta_i = 0$. Let $\eta = \max_i \|\delta_i\|_2 / \|\mu\|_2$, which is small because the common update pushes the mean out faster than the centred spread grows. Expanding the inner product and the norms gives

$$\langle v_i, v_j \rangle \;=\; \|\mu\|_2^2 + \langle \mu, \delta_i + \delta_j \rangle + \langle \delta_i, \delta_j \rangle, \tag{12}$$

$$\|v_i\|_2 \|v_j\|_2 \;=\; \|\mu\|_2^2 \left(1 + \frac{\langle \mu, \delta_i + \delta_j \rangle}{\|\mu\|_2^2} + \mathcal{O}(\eta^2) \right). \tag{13}$$

Dividing Equation 12 by Equation 13 and keeping terms up to order η^2, the linear terms in $\delta_i + \delta_j$ cancel, and the remaining quadratic term combines into $-\frac{1}{2}\|\delta_i - \delta_j\|_2^2 / \|\mu\|_2^2$. This gives

$$\cos(v_i, v_j) \;=\; 1 - \frac{1}{2} \frac{\|\delta_i - \delta_j\|_2^2}{\|\mu\|_2^2} + \mathcal{O}(\eta^4). \tag{14}$$

Averaging over all pairs $i \neq j$ gives Equation 4. Two features of this expansion matter for the rest of the argument. The leading correction only involves the centred spread $\|\delta_i - \delta_j\|_2^2$ in the numerator and the accumulated shared updates $\|\mu\|_2^2$ in the denominator, so the cosine similarity can only fall short of one if the spread is large relative to the mean. And the approximation becomes more accurate as η shrinks, so it is most informative in the deep layers where the collapse is most severe.

A.3 Why an Effective Rank Penalty

Equation 4 tells us that slowing the collapse requires keeping the pairwise spread large. Using $\sum_i \delta_i = 0$, the pairwise spread can be written in terms of the singular values of the centred visual token matrix V,

$$\frac{1}{n_v(n_v - 1)} \sum_{i \neq j} \|\delta_i - \delta_j\|_2^2 \;=\; \frac{2}{n_v - 1} \sum_k \sigma_k^2. \tag{15}$$

The right hand side is the total spectral energy of V. LayerNorm inside the backbone keeps this total roughly constant, so the only way to shrink the spread is to let the energy concentrate on a few singular directions. The effective rank, defined as $\mathrm{erank}(V) = \exp(-\sum_k p_k \log p_k)$ with $p_k = \sigma_k^2 / \sum_j \sigma_j^2$, measures exactly this concentration. A large effective rank forces the energy to spread across many directions and therefore keeps the pairwise spread large.

Maximising the effective rank directly is awkward because it depends non smoothly on the spectrum. The log determinant $\log\det(V^\top V + \varepsilon I) = \sum_k \log(\sigma_k^2 + \varepsilon)$ is a convenient smooth surrogate. Maximising it under a fixed energy budget $\sum_k \sigma_k^2 = E$ gives the stationary condition

$$\frac{1}{\sigma_k^2 + \varepsilon} = \nu \quad \text{for all } k, \tag{16}$$

which forces all singular values to be equal. This is also where the effective rank reaches its maximum. Penalising the negative log determinant therefore pushes the visual spectrum toward a flat profile, which is what the DVDR loss does.

A.4 Empirical Verification of the Correlation Prediction

Section 4 also predicts that the layer to layer growth of visual similarity $g_\ell = s_{\ell+1} - s_\ell$ should correlate positively with the residual dominance ratio ρ_ℓ, since a large ρ_ℓ marks layers where the visual update adds little token specific spread. We measure both quantities across all layers of Qwen2.5-VL-7B. The Pearson correlation is 0.61 with p value below 10^{-3}, and the fitted slope is positive and significant. Results on InternVL2.5-4B and LLaVA-Med-7B give Pearson correlations of 0.57 and 0.64. The scatter is noisy because real transformer layers do more than the shared update model captures, but the sign of the effect is stable across models. The scatter and fitted line for Qwen2.5-VL-7B are shown in Appendix D.1, Figure 8(c).

B Implementation Details

This appendix gives the implementation details that did not fit in the main text. All code and configuration files will be released upon publication.

B.1 Backbone and LoRA Configuration

We use the publicly released checkpoints of Qwen2.5-VL-3B and Qwen2.5-VL-7B [13], InternVL2.5 at 2B, 4B, and 8B [14], LLaVA-Med-v1.5-7B [10], and HuatuoGPT-Vision-7B [11]. The vision encoder and the projection module are frozen. LoRA adapters [38] are inserted into the attention query, key, value, and output projections, and into the MLP gate, up, and down projections of every transformer block in the language backbone. We use rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. This gives roughly 35 million trainable parameters for the 7B backbones and 18 million for the 3B backbone, which is below 0.5 percent of the total parameter count in every case.

B.2 Training Hyperparameters

All fine tuning runs use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01, and peak learning rate 2×10^{-4}. We use a cosine schedule with a linear warm up over the first three percent of steps. Batch size is 32 and we train for three epochs on VQA-RAD, SLAKE, and PathVQA, and for two epochs on PMC-VQA, OmniMedVQA, and GMAI-MMBench because their training splits are larger. Gradient clipping is applied at global norm 1.0. Training uses mixed precision with bfloat16 for activations and float32 for optimizer states. Images are resized so that the longer side is 448 pixels, which is the default resolution of the Qwen2.5-VL family. For InternVL2.5 backbones we follow the native dynamic resolution scheme of the model. Questions and answers are tokenised with the backbone tokeniser and padded to the longest length in the batch.

B.3 VGrip Specific Settings

The DVDR loss is applied to the last third of the transformer layers. For a 28 layer backbone this is layers 19 through 28. We set $\varepsilon = 10^{-4}$ in the log determinant and centre the visual tokens by subtracting the per layer mean before computing the gram matrix. The DVDR loss is evaluated on a random subset of 64 visual tokens per image to keep the computation affordable, which matches earlier work on effective rank regularisation [29].

The MALS scalars $\gamma_v^{(\ell)}$ and $\gamma_t^{(\ell)}$ are initialised at one so that the model starts from the plain LoRA baseline. The soft margin τ is set to $\log 2$, which allows the text residual to be up to twice the norm of the vision residual without penalty. Gradients on the MALS scalars are scaled by an extra factor of 0.1 during the first 500 steps to prevent early instability.

The AAL loss is applied to the first third of the layers, which is layers 1 through 9 for a 28 layer backbone. We average attention weights over queries and then over heads before computing the KL divergence to the norm based anchor distribution. The norm in the anchor is computed on a detached copy of the visual tokens so that gradients from AAL do not flow back into the token norms. AAL is warmed up over the first epoch with a linear ramp from zero to λ_3 in order to avoid interfering with early task learning.

Table 5: Training cost of each VGrip component on one A100 with Qwen2.5-VL-7B, batch size 32, averaged over 200 steps.

Setting	s / step	Overhead	Peak memory (GB)
LoRA-FT	1.82	reference	47.3
+ DVDR	1.88	$+3.3\%$	48.1
+ MALS	1.83	$+0.5\%$	47.4
+ AAL	1.94	$+6.6\%$	49.6
VGrip (full)	2.03	$+11.5\%$	50.8

The three loss weights are $\lambda_1 = 0.05$, $\lambda_2 = 0.01$, and $\lambda_3 = 0.02$ for all experiments unless stated otherwise. These values were chosen on a held out portion of VQA-RAD and then fixed across all other benchmarks and backbones.

B.4 Evaluation Protocol

For VQA-RAD, SLAKE, and PathVQA we follow the open and closed question protocol of the original benchmarks and report overall accuracy. For PMC-VQA and OmniMedVQA we use the official multiple choice evaluation and report the percentage of questions answered correctly. For GMAI-MMBench we use the official evaluation toolkit and report the aggregate score averaged across its 285 tasks.

The Visual Dependency Score in Equation 1 requires a shuffled image baseline. We generate shuffled images by dividing each visual token sequence into a grid that matches the native tokenisation of the backbone and permuting the tokens uniformly at random, while keeping positional embeddings fixed. This destroys spatial content but preserves the global visual statistics that the model might rely on as a generic prior. For the random baseline a_{rand} we use the empirical accuracy of a model that predicts the most frequent answer in the training set. This gives 0.35 on VQA-RAD, 0.40 on SLAKE, 0.38 on PathVQA, 0.28 on PMC-VQA, 0.30 on OmniMedVQA and 0.32 on GMAI-MMBench.

All numbers in the main text are averaged over three seeds, namely 41, 42, and 43. Standard deviations across seeds are below 0.6 points of accuracy and below 0.03 points of VDS on every benchmark, and full error bars are available in the code release.

B.5 Training Cost

Table 5 reports training time and peak memory on one A100 with Qwen2.5-VL-7B, batch size 32 and VQA-RAD, averaged over 200 steps. DVDR is cheap because it acts only on the last third of the layers and on 64 subsampled visual tokens per image, so each layer needs a 64×64 log determinant instead of one over all n_v tokens. AAL is the most expensive of the three losses because it needs per head attention over visual positions. Inference cost is unchanged in every setting, since the MALS scalars fold into the residual updates after training.

C Component Ablation, Sensitivity, and Scaling

This appendix gives the full results of the component analysis, hyperparameter sensitivity, and scaling studies summarised in Section 6.4. Figure 6 plots the main results of Table 2, and Figure 7 visualises the three studies below.

C.1 Component Ablation

We isolate the contribution of each VGrip component by running all seven non trivial combinations of DVDR, MALS, and AAL on VQA-RAD with Qwen2.5-VL-7B. All other settings match the main VGrip configuration. Results are in Table 6.

Three patterns emerge. First, every single component is individually useful. Turning on any one of the three losses improves both accuracy and VDS over the LoRA baseline, with DVDR being the strongest and AAL the weakest when applied on its own. This ordering matches the hierarchy suggested by the diagnosis in Section 3. Representation collapse is the earliest failure point and

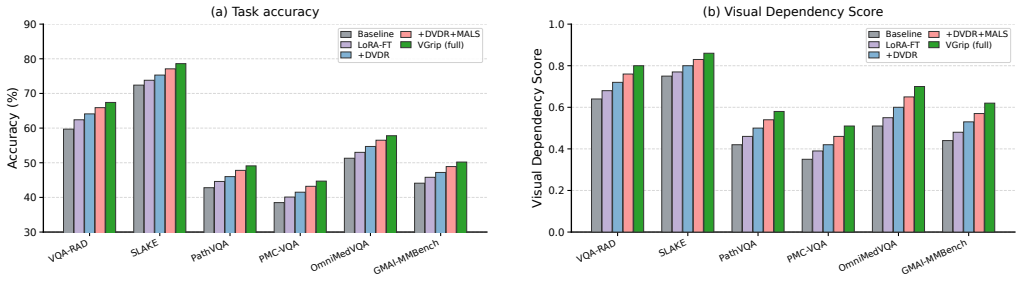

Figure 6: Main results on six medical VQA benchmarks with Qwen2.5-VL-7B. (a) Accuracy. (b) Visual Dependency Score.

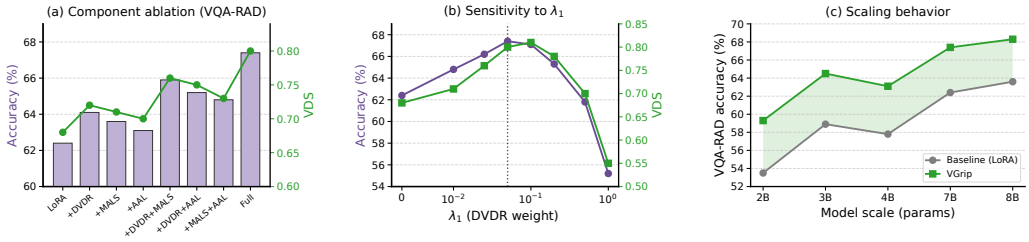

Figure 7: Component analysis and sensitivity. (a) Accuracy and Visual Dependency Score on VQA-RAD as each of the three losses is switched on. (b) Sensitivity of VGrip to λ_1 while λ_2 and λ_3 stay at their defaults. (c) Accuracy on VQA-RAD across five backbone sizes from 2B to 8B.

preventing it has the broadest effect. Residual imbalance and attention freezing are downstream consequences, so addressing them alone gives a smaller gain.

Second, the components are complementary. Each pairwise combination recovers more than the sum of the two individual gains. DVDR alone gives $+1.7$ accuracy and MALS alone gives $+1.2$, while the pair recovers $+3.5$. The three failure modes are coupled, and breaking one leaves the others partially in place. MALS without DVDR rebalances the residual magnitudes but does not prevent the visual tokens from collapsing, so the text side still drives inference.

Third, the full triple recovers 5.0 accuracy points and 0.12 of VDS, against $+3.5$ and $+0.08$ for the best pairwise combination. The added cost is small, since AAL adds no trainable parameters and MALS adds only $2L$ scalars.

Layer bands. DVDR acts on the last third of the layers and AAL on the first third. Table 7 moves each loss to other bands while keeping the rest of VGrip fixed. Applying DVDR to the first third gives little over LoRA-FT, since the tokens have not yet collapsed at those depths. The middle third recovers most of the effect, and applying it to every layer is slightly worse than the deep band because the shallow penalty competes with fusion. AAL shows the mirror pattern. Anchoring the first third works best because routing is still malleable there and the effect propagates downstream, while at deeper layers attention has already frozen onto content free regions and the KL target is hard to reach.

C.2 Sensitivity to the DVDR Weight

We sweep the DVDR weight λ_1 over eight values from zero to one while keeping $\lambda_2 = 0.01$ and $\lambda_3 = 0.02$ fixed at their defaults. Results on VQA-RAD with Qwen2.5-VL-7B are in Table 8.

Both metrics have a broad plateau between $\lambda_1 = 0.025$ and $\lambda_1 = 0.1$, where accuracy stays within 1.2 points and VDS within 0.05 of their best values. Performance degrades at large λ_1 because the diversity term starts to dominate the task loss and the model loses language fluency. At $\lambda_1 = 1.0$ accuracy falls below the LoRA baseline and VDS falls to 0.55, so the model spreads its visual representation so aggressively that it can no longer combine it with the question. The default

Table 6: Component ablation of VGrip on VQA-RAD with Qwen2.5-VL-7B. Each row turns one or more of the three losses on. Best values in bold.

Configuration	DVDR	MALS	AAL	**Acc (%)**	**VDS**
LoRA-FT (no VGrip)				62.4	0.68
+ DVDR only	✓			64.1	0.72
+ MALS only		✓		63.6	0.71
+ AAL only			✓	63.1	0.70
+ DVDR + MALS	✓	✓		65.9	0.76
+ DVDR + AAL	✓		✓	65.2	0.75
+ MALS + AAL		✓	✓	64.8	0.73
VGrip (full)	✓	✓	✓	**67.4**	**0.80**

Table 7: VQA-RAD accuracy (%) of VGrip on Qwen2.5-VL-7B when DVDR or AAL is applied to a different band of layers. Defaults in bold.

Band	**DVDR moved**	**AAL moved**
First third	63.8	**67.4**
Middle third	65.6	65.9
Last third	**67.4**	64.2
All layers	66.5	66.3

$\lambda_1 = 0.05$ sits near the top of the plateau. We did not sweep λ_2 and λ_3 in the same way, since their smaller defaults already keep MALS and AAL in a soft correction regime.

C.3 Scaling Across Backbone Sizes

We repeat the main experiment on five backbones from 2B to 8B, using the same LoRA configuration and the same VGrip hyperparameters as in the main results. Results on VQA-RAD are in Table 9.

The accuracy gain of VGrip over the LoRA baseline stays between 4.7 and 5.8 points across all five backbones, and the gain in VDS stays between 0.11 and 0.15. Both gains narrow slightly as the backbone grows, but larger models do not close the gap on their own. The plain LoRA baseline improves by about 10 accuracy points between the 2B and 8B backbones, which is the kind of improvement one would expect from more capacity. Within each model family, VGrip on the smaller backbone reaches or exceeds the plain LoRA baseline on a backbone about twice its size, with 64.5 on Qwen2.5-VL-3B against 62.4 on Qwen2.5-VL-7B, 59.3 on InternVL2.5-2B against 57.8 on InternVL2.5-4B, and 63.1 on InternVL2.5-4B against 63.6 on InternVL2.5-8B. A targeted intervention at a small scale therefore matches the effect of roughly doubling the backbone.

C.4 Full Fine-Tuning

To test whether the collapse is specific to low rank adaptation, we replace LoRA with full fine-tuning of the language backbone on Qwen2.5-VL-7B and VQA-RAD, keeping all other settings fixed. Table 10 reports accuracy together with the final layer visual cosine similarity and the mean residual dominance ratio $\bar{\rho}$.

Full fine-tuning raises accuracy while leaving the collapse signature largely in place, since nothing in the standard training signal rewards visual diversity. Adding VGrip closes the mechanistic gap in the same way it does on LoRA. Its 1.5 point gain over LoRA + VGrip is smaller than the 2.4 point gain of full fine-tuning over LoRA-FT, which points to the collapse, more than adapter capacity, as the binding constraint at this scale.

D Robustness and Mechanistic Repair

This appendix expands the robustness and verification studies of Sections 6.5 and 6.6.

Table 8: Sensitivity of VGrip to the DVDR weight λ_1 on VQA-RAD with Qwen2.5-VL-7B. Default value in bold.

λ_1	0	0.01	0.025	**0.05**	0.1	0.2	0.5	1.0
Acc (%)	62.4	64.8	66.2	**67.4**	67.1	65.3	61.8	55.2
VDS	0.68	0.71	0.76	**0.80**	0.81	0.78	0.70	0.55

Table 9: Scaling of VGrip across five backbones on VQA-RAD. Accuracy (%) and Visual Dependency Score are shown for the plain LoRA baseline and for full VGrip.

		LoRA-FT		**VGrip**			
Backbone	**Params**	Acc	VDS	Acc	VDS	ΔAcc	ΔVDS
InternVL2.5-2B	2.0B	53.5	0.68	59.3	0.83	+5.8	+0.15
Qwen2.5-VL-3B	3.8B	58.9	0.72	64.5	0.85	+5.6	+0.13
InternVL2.5-4B	4.2B	57.8	0.73	63.1	0.85	+5.3	+0.12
Qwen2.5-VL-7B	7.6B	62.4	0.68	67.4	0.80	+5.0	+0.12
InternVL2.5-8B	8.1B	63.6	0.69	68.3	0.80	+4.7	+0.11

D.1 Robustness to Image Perturbations and Out of Distribution Generalisation

Sensitivity to Gaussian blur. A model that uses the image should degrade as the image is degraded. We apply Gaussian blur with standard deviation σ from 0 to 3 to each input image and measure accuracy (Table 11).

The LoRA baseline loses 3.5 points at the mildest blur and only 1.4 more between $\sigma = 0.5$ and $\sigma = 3$. Once fine detail is gone, further degradation barely matters to it. VGrip drops by 15.3 points over the full range and keeps losing accuracy as the image degrades, which is what one would expect from a model that relies on visual content. At $\sigma = 1.5$ the gap has shrunk to 1.7 points, and from $\sigma = 2$ onwards VGrip falls below LoRA, since it has learned to rely on an image that no longer carries information. On clean inputs VGrip is 5.0 points ahead.

Out of distribution evaluation. All three settings use models trained on VQA-RAD and SLAKE only. **Cross hospital** scans come from the MIMIC-CXR-VQA test split. Its images come from a different institution than the MedPix contributors behind VQA-RAD and the curated SLAKE pool, with no site overlap, giving 1,285 images and 2,478 questions. **Corrupted** scans apply five perturbations to the VQA-RAD test set, namely motion blur, beam hardening on CT cases, Rician noise on MRI cases, low dose Poisson noise and JPEG compression at quality 20, each at three severities following the medical adaptation of the ImageNet-C protocol. This gives 6,765 evaluations over 451 base questions. **Unseen modality** uses the held out OmniMedVQA subset with ultrasound and dermoscopy, both absent from the training splits, giving 1,842 images and 3,976 questions. Results are in Table 4 and Figure 8(b).

VGrip improves accuracy on every setting, with an average gain of 8.2 points. The largest gain of 10.4 points is on cross hospital scans, where language priors help the least. The questions follow the same templates across hospitals while the images shift, so a model that answers from question templates gains nothing from the new images. The gain of 8.4 points on unseen modalities fits the same reading. The further the image distribution moves from training, the more a model that reads the image gains over one that relies on question templates.

D.2 Mechanistic Repair

This section gives details for the three panels of Figure 5 and for the linear probe of Section 6.6.

Representation collapse is prevented. The baseline reaches a final layer visual similarity of about 0.67, the same value we measured in Section 3.2. VGrip keeps it at 0.30, close to the similarity of text tokens on the same backbone. The effect is visible from the first third of the backbone onwards, which fits DVDR acting on deep layers together with the earlier routing that AAL establishes.

Table 10: Full fine-tuning against LoRA on VQA-RAD with Qwen2.5-VL-7B. s_L is the final layer visual cosine similarity and $\bar{\rho}$ the mean residual dominance ratio across layers.

Setting	Acc (%)	s_L	$\bar{\rho}$
LoRA-FT	62.4	0.67	2.20
LoRA + VGrip	67.4	0.30	1.05
Full-FT	64.8	0.63	2.05
Full-FT + VGrip	**68.9**	0.32	1.08

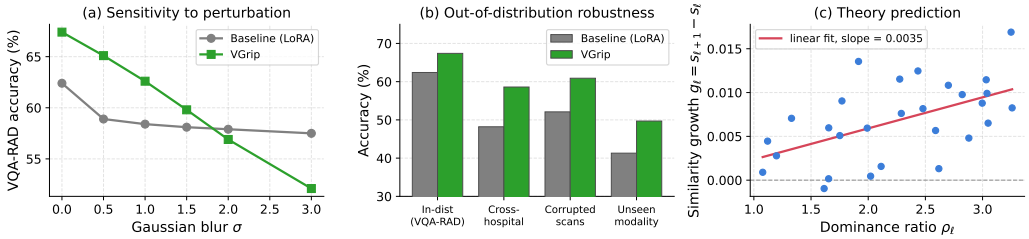

Figure 8: Robustness and theory validation on Qwen2.5-VL-7B. (a) Accuracy on VQA-RAD under Gaussian blur of increasing σ. The LoRA baseline changes little beyond the mildest blur, while VGrip degrades as the image degrades. (b) Accuracy on the in-distribution test set and three distribution shift settings. (c) Layer to layer growth of visual similarity $g_\ell = s_{\ell+1} - s_\ell$ plotted against the residual dominance ratio ρ_ℓ for the baseline model. The fitted line has a positive slope, as Section 4 predicts.

Residual balance is restored. The baseline dominance ratio drifts upward from about 1.2 at shallow layers to above 3.0 near the output, reproducing the asymmetry in Figure 2(d). VGrip holds ρ_ℓ near one throughout, with a median of 1.05 and a maximum of 1.4. MALS achieves this by modestly amplifying the visual residuals while leaving the text residuals nearly unchanged (Appendix G).

Visual attention is retained. The fraction of attention mass on visual tokens, averaged over heads and queries, decays in the baseline from about 0.55 at layer zero to below 0.2 at the final layer. VGrip keeps it above 0.45 throughout, so queries can still pull content from the visual tokens in deep layers. AAL is only active in the first third of the layers, and its effect carries into deeper layers because the routing it sets up early is kept once the downstream dynamics are balanced.

Linear probe. We train linear probes on frozen visual tokens to predict the row index of each patch among 32 rows and the organ shown in the image among 12 categories drawn from VQA-RAD. Chance is 3.1 and 8.3 percent. On shallow layers of the frozen backbone the probes reach 74.9 and 79.4 percent. On final layer tokens they reach 32.4 and 44.8 percent for the LoRA baseline and 71.6 and 76.2 percent for VGrip. The baseline keeps part of the organ information and loses most of the spatial layout, and VGrip restores both close to their shallow layer values.

Theory prediction. Section 4 predicts that the layer to layer growth of visual similarity g_ℓ correlates positively with the residual dominance ratio ρ_ℓ. Figure 8(c) plots g_ℓ against ρ_ℓ for each layer transition of the baseline Qwen2.5-VL-7B model, and the fitted line has a positive slope. Appendix A.4 gives the correlations on all three backbones.

E Cross Model Diagnosis

Section 3.2 reported the layerwise diagnosis on Qwen2.5-VL-7B. This appendix reports the same measurements on all five backbones used in the diagnosis, namely Qwen2.5-VL-3B, Qwen2.5-VL-7B, InternVL2.5-4B, LLaVA-Med-7B, and HuatuoGPT-Vision-7B, together with the knockout curves of Section 3.3.

Table 11: Accuracy (%) on VQA-RAD with Qwen2.5-VL-7B under Gaussian blur of increasing σ.

Method	σ=0	σ=0.5	σ=1.0	σ=1.5	σ=2.0	σ=3.0
LoRA-FT	62.4	58.9	58.4	58.1	57.9	57.5
VGrip	**67.4**	65.1	62.6	59.8	56.9	52.1
Δ (clean $\rightarrow$ σ=3)			LoRA: -4.9	VGrip: -15.3		

Table 12: Representation collapse across backbones. s_L is the mean pairwise cosine similarity of visual tokens at the final layer, and $\ell_{0.5}$ is the first layer at which s_ℓ exceeds 0.5, reported as a fraction of total depth.

	VQA-RAD		**SLAKE**	
Backbone	s_L	$\ell_{0.5}$	s_L	$\ell_{0.5}$
Qwen2.5-VL-3B	0.68	0.36	0.71	0.33
Qwen2.5-VL-7B	0.67	0.38	0.69	0.35
InternVL2.5-4B	0.72	0.30	0.73	0.28
LLaVA-Med-7B	0.74	0.26	0.76	0.24
HuatuoGPT-V-7B	0.70	0.32	0.72	0.30
Mean	0.70	0.32	0.72	0.30

E.1 Setup

All backbones are evaluated frozen on VQA-RAD and SLAKE. For each model we compute the mean pairwise cosine similarity of visual tokens, the residual dominance ratio, and the Visual Dependency Score. Because the five backbones have different numbers of layers, we report layerwise statistics on a normalised depth axis from zero to one where available, and report end of backbone values in the tables. VDS is also reported on PathVQA, PMC-VQA, OmniMedVQA, and GMAI-MMBench.

E.2 Representation Collapse

The mean pairwise cosine similarity of visual tokens rises sharply in the shallow layers of every model we tested. Table 12 reports the final layer value s_L for each model on both datasets, along with the depth at which s_ℓ first exceeds 0.5, denoted $\ell_{0.5}$.

Every model has a final layer similarity above 0.65, and all of them cross 0.5 within the first 40 percent of their depth. The two medical specialists, LLaVA-Med-7B and HuatuoGPT-Vision-7B, reach the highest terminal values despite pretraining on large amounts of paired medical data, so specialisation does not prevent the collapse.

E.3 Residual Dominance

The residual dominance ratio ρ_ℓ stays above one across almost all layers of every model (Table 13).

On average the text residual norm is 2.2 times the vision residual norm, and the text stream dominates at more than 85 percent of the layers in every model. The two medical specialists show slightly stronger asymmetry. A model that knows more about medicine in language form produces larger text updates, and the visual stream loses further ground.

E.4 Visual Dependency Score Across Six Benchmarks

Table 14 reports the Visual Dependency Score of each backbone on all six benchmarks. Figure 3(c) shows four of these columns.

Across the 30 model and benchmark combinations VDS ranges from 0.23 to 0.85, with a mean of 0.48. The benchmark matters more than the model. Every backbone scores highest on SLAKE and VQA-RAD and lowest on PMC-VQA, where the mean is 0.31. PMC-VQA draws from figure panels that often carry textual labels, so part of the answer can be read from text in the image or inferred

Table 13: Distribution of the residual dominance ratio ρ_ℓ across layers on VQA-RAD. $\bar{\rho}$ is the mean across layers, $\rho_{\max}$ is the maximum, and **frac** is the fraction of layers at which $\rho_\ell > 1$.

Backbone	$\bar{\rho}$	$\rho_{\max}$	**frac** > 1
Qwen2.5-VL-3B	2.1	3.4	0.89
Qwen2.5-VL-7B	2.2	3.3	0.93
InternVL2.5-4B	1.9	3.0	0.86
LLaVA-Med-7B	2.4	3.8	0.93
HuatuoGPT-V-7B	2.3	3.6	0.90
Mean	2.2	3.4	0.90

Table 14: Visual Dependency Score of five frozen compact medical multimodal models on six benchmarks.

Backbone	**VQA-RAD**	**SLAKE**	**PathVQA**	**PMC-VQA**	**OmniMed**	**GMAI**	**Mean**
Qwen2.5-VL-3B	0.71	0.80	0.45	0.36	0.55	0.47	0.56
Qwen2.5-VL-7B	0.64	0.75	0.42	0.35	0.51	0.44	0.52
InternVL2.5-4B	0.72	0.85	0.48	0.38	0.58	0.50	0.59
LLaVA-Med-7B	0.47	0.57	0.30	0.24	0.38	0.32	0.38
HuatuoGPT-V-7B	0.45	0.55	0.29	0.23	0.36	0.30	0.36
Mean	0.60	0.70	0.39	0.31	0.48	0.41	0.48

from the question. The two medical specialists score lowest on every benchmark, because their higher image plus text accuracy enlarges the denominator while their shuffled image gap is smaller.

E.5 Knockout Interventions

Figure 9 shows the residual knockout of Section 3.3 at every layer, and Table 15 gives the attention knockout on VQA-RAD. Blocking text to visual attention from the first layer brings accuracy close to the text only condition, and the effect fades as the block starts deeper in the backbone.

E.6 Discussion

The tables above show representation collapse and residual asymmetry in every compact medical multimodal model we tested, at similar magnitude, while the dependence on image content varies mostly with the benchmark. The medical specialists are not exempt. Their specialisation makes the text side stronger, which widens the dominance ratio and lowers their VDS. More pretraining or a narrower domain focus therefore does not remove the failure by itself, since it sits in the dynamics of the language backbone more than in what the backbone knows about medicine.

F Behavioural Details

This appendix expands the behavioural diagnosis of Section 3.1. We give the full bar plot for all five backbones, break accuracy down by question type, look more closely at the shuffled image baseline, and report how VDS depends on the random baseline.

F.1 Full Behavioural Evaluation

Figure 10 plots accuracy under the three input conditions for all five backbones on VQA-RAD and SLAKE, using the numbers of Table 1.

Three observations stand out. First, the image matters. The average gap between image plus text and text only is 19.0 points on VQA-RAD and 24.8 points on SLAKE. A shuffled image recovers 18 to 30 percent of that gap on VQA-RAD and 6 to 12 percent on SLAKE, so most of the gain depends on the content of the image and a smaller part on generic visual statistics.

Second, the two medical specialists have higher absolute accuracy in every condition but do not depend more on the image. LLaVA-Med-7B improves on Qwen2.5-VL-7B by 3.7 points with image

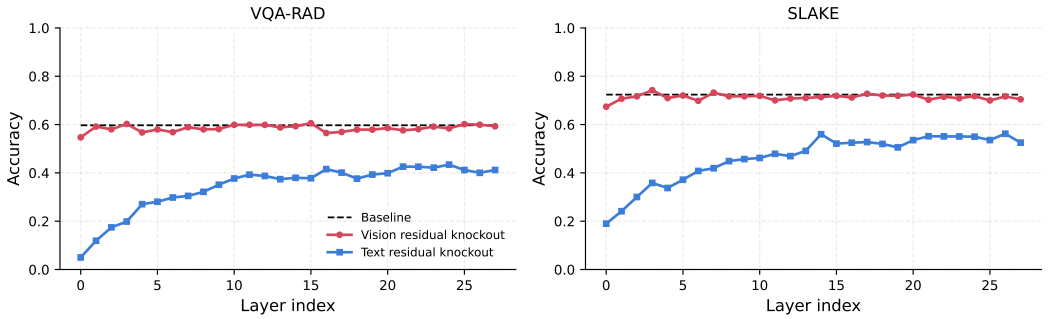

Figure 9: Residual knockout on Qwen2.5-VL-7B. Zeroing the vision residual update changes accuracy little beyond the first layer, consistent with text positions having read most of the visual content by then (Section 3.3). Zeroing the text residual update causes large drops at every depth, largest at shallow layers.

Table 15: Attention knockout on Qwen2.5-VL-7B and VQA-RAD. All attention edges from text positions to visual positions are blocked from the given layer onward. Accuracy (%).

Block from layer	1	5	10	15	none	text only
Accuracy	41.8	51.3	57.4	59.1	59.7	39.1

plus text and by 5.4 points with text only. Specialist training has made the model more knowledgeable about medicine, and the extra knowledge sits on the text side.

Third, the pattern is stable. Standard deviations across three seeds are below 0.6 points in every cell of Table 1, and the shuffled image accuracy stays within 0.4 points of its mean across permutation seeds.

F.2 Per Question Type Analysis

Aggregate accuracy can hide category specific patterns. A model that scores well on closed yes or no questions through linguistic priors can make up for poor performance on localisation questions that need visual evidence. Figure 11 breaks accuracy down by six common medical VQA question categories.

The six categories span the range of reasoning in medical VQA. **Yes or no** questions have a binary answer and can often be solved from the question alone. **Modality** asks what imaging type is shown, and can be solved either visually or from a modality hint in the question. **Organ** asks which organ is depicted. **View** asks about imaging orientation, such as axial or coronal. **Location** asks where a finding is, and **laterality** asks left or right. The last three categories need visual evidence.

The yes or no category has the smallest gap between text only and image plus text on every model, since linguistic priors are often enough for binary questions about common anatomy. Modality and organ sit in the middle. Location and laterality show the largest text only deficit and the largest VGrip gain, which is where spatial evidence matters most and where the collapse of spatial information in deep visual tokens hurts most. The ordering of categories is the same for all three models, with LLaVA-Med-7B the strongest baseline in every category.

F.3 Closer Look at the Shuffled Image Baseline

We consider three variants of the shuffled image. **Token shuffle** randomly permutes the visual tokens after the vision encoder, which is the variant used in the main text. **Patch shuffle** permutes image patches before the vision encoder. **Pixel shuffle** permutes pixels at full resolution, which is the most destructive variant.

Table 16 reports the three variants. All three remove most of the 20.6 point gain that the image provides, and the more destructive the shuffle, the closer the model gets to the text only baseline.

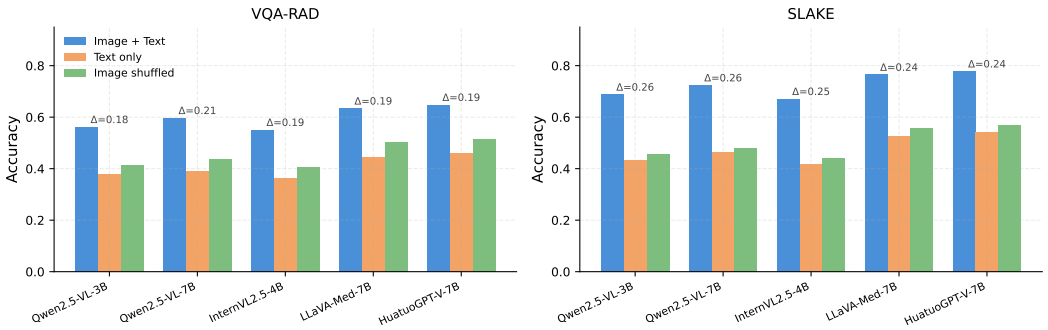

Figure 10: Behavioural evaluation of all five backbones on VQA-RAD and SLAKE. The image plus text, text only, and shuffled image conditions are shown side by side. The gap Δ between image plus text and text only is displayed above each group.

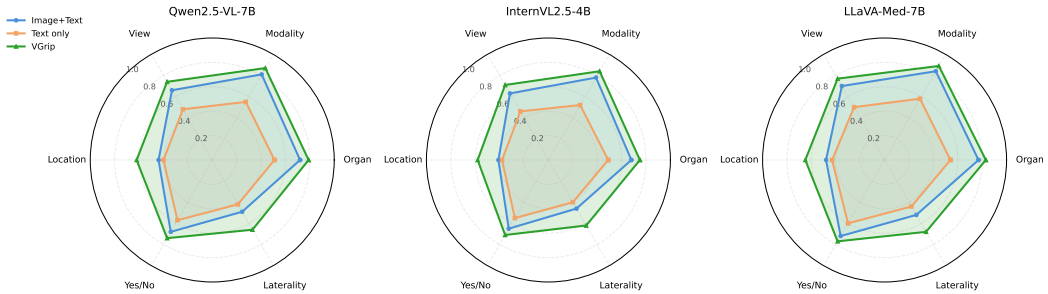

Figure 11: Per question type accuracy for three models. Each radar plot compares image plus text, text only, and full VGrip. Categories that require visual grounding, such as location and laterality, show the largest gap between text only and VGrip.

The few points that survive even the pixel shuffle come from coarse global statistics. Chest X-rays look different from brain MRI at the coarsest level, and this difference survives any permutation of tokens or pixels. We use token shuffle in the VDS because it keeps the forward pass identical up to the vision encoder output and isolates what happens inside the language backbone.

F.4 VDS Sensitivity to the Random Baseline

The Visual Dependency Score in Equation 1 normalises by $a_{\text{it}} - a_{\text{rand}}$, so its absolute scale depends on how a_{rand} is defined. Varying a_{rand} from 0.28 to 0.41, a range that covers all six benchmarks, moves the frozen Qwen2.5-VL-7B from 0.50 to 0.85 on VQA-RAD and from 0.55 to 0.77 on SLAKE, and InternVL2.5-4B reaches 1.04 on VQA-RAD at the upper end because its shuffled accuracy falls below a_{rand}. The absolute value of VDS should therefore be read with the a_{rand} of its benchmark, which is why we fix one value per benchmark and report the raw accuracies next to it. Comparisons between methods on the same benchmark use the same a_{rand}, and the ordering of models within a benchmark is the same at both ends of the range.

G MALS Dynamics and Language Preservation

This appendix takes a closer look at the modality aware LayerScale component. We verify the claim in Section 5.2 that MALS amplifies the visual stream while leaving the text stream nearly unchanged, and we check that this intervention does not hurt the language ability of the backbone.

G.1 Learned Values of the Scalars

Figure 12 plots the learned values of $\gamma_v^{(\ell)}$ and $\gamma_t^{(\ell)}$ across layers at the end of training on VQA-RAD with Qwen2.5-VL-7B. Both scalars are initialised at 1.0. After training, $\gamma_v^{(\ell)}$ drifts upward to a mean

Table 16: Accuracy (%) of Qwen2.5-VL-7B on VQA-RAD under three shuffling schemes.

Condition	Accuracy (%)
Image plus text	59.7
Text only	39.1
Token shuffle (ours)	43.8
Patch shuffle	42.2
Pixel shuffle	40.5

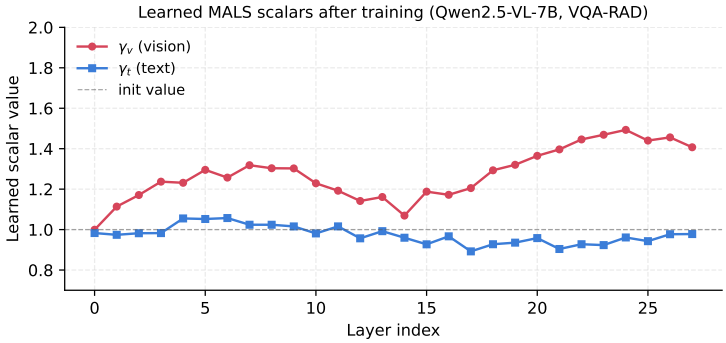

Figure 12: Learned MALS scalars after training on VQA-RAD. The vision scalar γ_v drifts upward, while the text scalar γ_t stays near one. MALS lifts the visual residual path and leaves the text path close to its initial scale.

of 1.41 and a maximum of 1.83, while $\gamma_t^{(\ell)}$ stays close to its initial value with a mean of 0.97 and a range between 0.88 and 1.04.

The asymmetry matters for two reasons. First, the residual dominance ratio reported in Section 6.6 is restored primarily by pulling the visual side up to meet the text side, not by pushing the text side down. A training signal that suppressed text residuals would risk degrading the language capabilities that the backbone relies on for fluent answers. Second, a scalar near one for the text stream means that MALS can be added to a pretrained model without disturbing the pretrained text dynamics more than a LoRA adapter already does.

G.2 Language Preservation

A concern with any intervention on residual magnitudes is that it might degrade the language capabilities of the backbone. We test this by measuring perplexity on a held out slice of C4 text [39] that is unrelated to the medical training distribution.

Table 17 shows that perplexity under VGrip is 8.49, within 0.1 of the plain LoRA baseline at 8.51 and the frozen pretrained model at 8.42. The small degradation from 8.42 to 8.49 comes from task specific fine tuning, since plain LoRA shows a similar degradation. VGrip does not make the language capabilities of the backbone measurably worse.

G.3 Ablation of Initialisation and Learning Rate

The MALS scalars have two extra hyperparameters. The initialisation value, which we set to 1.0, and the learning rate scaling during the first 500 steps, which we set to 0.1. Table 18 tests whether VGrip is sensitive to these choices on VQA-RAD with Qwen2.5-VL-7B.

Initialising both scalars at 1.0 is important because the model starts from exactly the LoRA baseline and can drift away from it gradually. A cold start at 0.5 or an aggressive start at 2.0 both hurt performance because the model begins training in a regime that the backbone was not pretrained in. The gradient scaling during the first 500 steps matters for a similar reason. Without the warm up, the scalars move too quickly during the first epoch when the task gradients are still noisy, which can cause the vision and text pathways to become unbalanced in a way that is then hard to recover from.

Table 17: Perplexity on a held out slice of C4 for Qwen2.5-VL-7B under different training setups. VGrip is within measurement noise of plain LoRA, which in turn is within noise of the frozen pretrained model.

Training setup	**Perplexity**
Frozen pretrained	8.42
LoRA-FT on VQA-RAD	8.51
VGrip on VQA-RAD	8.49

Table 18: Sensitivity of VGrip to the MALS hyperparameters on VQA-RAD. The default initialisation is $\gamma = 1.0$ and the default gradient scaling during the first 500 steps is 0.1.

Setting	**Acc (%)**
γ_v, γ_t init $= 0.5, 0.5$	65.1
γ_v, γ_t init $= 1.0, 1.0$ (default)	**67.4**
γ_v, γ_t init $= 1.5, 1.0$	66.9
γ_v, γ_t init $= 2.0, 2.0$	63.2
No gradient warm up	65.8
Gradient scale $= 0.01$	66.5
Gradient scale $= 0.1$ (default)	**67.4**
Gradient scale $= 1.0$	64.3

Within a reasonable range the method is robust to both hyperparameters, and unit initialisation with a gentle warm up is the simplest default.

H Additional Limitations

This appendix expands on the limitations in Section 7.

H.1 AAL Relies on Norm Based Informativeness

The attention anchoring loss uses the norm of a visual token as a proxy for how informative that token is. On radiology images this heuristic works well, because the anatomical region of interest usually sits on a dark and uniform background, which produces a clear contrast in token norms. The proxy has known failure cases. Large vision transformers can produce high norm artifact tokens in uninformative background regions [40], and an anchor built on such tokens would pull attention toward the background.

The heuristic also works less well when the foreground does not have higher norm than the background. Histopathology is the clearest example. Tissue fills most of the field of view and the clinically important regions differ in fine grained texture more than in bulk contrast. AAL alone contributes 0.4 accuracy points on PathVQA against 0.7 on VQA-RAD, so the 4.5 point VGrip gain on PathVQA comes mainly from DVDR and MALS. A segmentation head, saliency over backbone features or a learned foreground predictor would give a more reliable anchor for pathology, at the cost of extra training requirements.

H.2 No Continued Pretraining

All of our runs start from publicly released checkpoints and adapt them on the training split of the target benchmark. We did not apply VGrip during continued pretraining on medical corpora, which is the usual route to a specialist medical multimodal model. The diagnosis in Section 3 suggests that the collapse is already present after pretraining, since it appears in every frozen backbone. Applying VGrip from the start could keep the backbone out of the collapsed regime and might give larger gains than the ones we report. Testing this requires roughly an order of magnitude more compute than all experiments in this paper.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are supported by experiments in Sections 3-6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In both the main text and the appendix.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are already stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided experiment descriptions in the main paper and appendix. Furthermore, we will release our GitHub implementation if the paper is accepted

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release the code upon acceptance, or provide an anonymous link if requested by reviewers.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided details for experiments in the paper

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experiments are performed on diverse datasets following the evaluation protocols of previous studies for fair comparison.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This information is included in our results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discusses both clinical deployment benefits and risks of visual blindness.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks

    Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly acknowledged all papers and resources used in our experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: This work does not involve crowdsourcing or any research involving human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work does not involve crowdsourcing or any research involving human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a component of the method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.