

The Illusion of Expertise: Narrow Finetuning Induces Persistent Domain Priors in Vision-Language Models

Qinglin Zeng¹, Jusheng Zhang¹, Jiangqi Huang², Ningyuan Liu¹, Luozhang Xu¹,
Wenhao Wang^{3,*}, Keze Wang^{1,*}

¹School of Computer Science and Engineering, Sun Yat-sen University

²School of Artificial Intelligence, Sun Yat-sen University ³Vast Intelligence Lab

wangwenhao@vastilab.com, kezewang@gmail.com

Abstract

Narrow-domain finetuning makes vision-language models (VLMs) sound like domain experts while quietly making them less attentive to the image. We attribute this gap to a *persistent domain prior*: an input-agnostic shift in hidden state, installed by narrow finetuning, that biases the model toward domain-style outputs regardless of visual evidence. To test the claim we introduce ActDiff, a white-box diagnostic that isolates the finetuning trace by differencing base and finetuned hidden states under progressively degraded images. The trace is compact and persists under blank and Gaussian-noise inputs; the finetuned model is correspondingly *less*, not more, disrupted when the image is destroyed (gap 0.323 ± 0.036 nats on BIOMED, generic eval). Subtracting the trace at inference recovers 54% of the MMStar accuracy collapse on BIOMED, with a peak-and-collapse dose response that random and wrong-domain controls do not produce. A single first-moment regulariser added at training time shrinks the trace by $20\text{--}64\times$ across three AdaptLLM domains and improves generic NLL. The prior thus has a clean diagnostic, a causal handle at inference, and a one-line mitigation at training.

1 Introduction

Vision-language models are routinely adapted to specialised settings—biomedicine, document understanding, remote sensing—through narrow domain finetuning. The narrow task scores improve, but the adapted model becomes less reliable on inputs outside the finetuning distribution: it produces fluent domain prose that is only loosely anchored to the actual image. Existing accounts attribute this drift to weight-level mechanisms—erased pre-trained knowledge, representation collapse, modality bias in the language backbone—and accord-

ingly treat mitigation as a data-mixing or distillation problem.

We argue that an important part of the failure is not in the weights but in the residual stream. Narrow finetuning, in our reading, installs an *persistent domain prior*: an input-agnostic shift in hidden state that the unembedding reads out as a domain-style default, regardless of what the image carries. Under this view the finetuned model behaves like a confident *blind expert*—fluent in the target domain, but selectively less sensitive to visual evidence than the base model it started from. We call this the *illusion of expertise*.

We turn the claim into a falsifiable diagnostic. ActDiff (Section 3) differences base and finetuned hidden states under progressively degraded image inputs—shuffled, text-only, blank, and pure Gaussian noise. If the finetuning trace reflected genuine visual processing it should weaken as visual content is removed; instead, it is compact, late-layer, early-position, and just as strong on a blank image as on the original (Section 4.2). On a corruption-sensitivity probe the finetuned model is correspondingly *less* disturbed by destroyed images than the base—a 0.323 ± 0.036 nat gap on BIOMED under generic evaluation (Section 4.3). Subtracting the extracted direction at inference, without retraining, recovers 54% of the BIOMED MMStar collapse with a peak-and-collapse dose response that random-direction and wrong-domain controls do not produce (Section 4.4); the same intervention works on a second VLM family that shows the same behavioural signature, and fails on a third that does not. Finally, the prior is preventable at training: a single first-moment regulariser shrinks the diagnostic trace by $20\text{--}64\times$ across three AdaptLLM domains and improves generic NLL relative to data-mixing baselines, while a conditional KL alternative provably cannot (Section 4.5).

The contribution is a complete loop—diagnostic, causal test, formation account, and a one-line

*Corresponding authors. Code is available at <https://github.com/zqinglin/actdiff-vlm>.

mitigation—around the same input-agnostic mechanism, complementing rather than replacing the weight-level picture of finetuning degradation.

2 Related Work

Finetuning degradation and shortcut learning in VLMs. Adapting pretrained VLMs to specialized domains trades plasticity against stability (Goodfellow et al., 2014; Jung et al., 2023). Narrow supervised finetuning is known to induce representation drift and to erode zero-shot out-of-distribution behavior (Lundstrom-Imanov, 2026; Kachaev et al., 2025), an effect tied to *shortcut learning* and modality bias, where the language backbone overrides visual percepts in favor of strong linguistic priors (Chi et al., 2025; Nooralahzadeh et al., 2026; Zheng and Zhang, 2025; Lin et al., 2024; Deng et al., 2025). Most prior accounts localize this failure to erased weights or late-stage arbitration inside the LLM. We instead characterize it as an input-agnostic activation-space prior introduced by narrow finetuning, and test its causal role directly.

Mechanistic interpretability and activation steering. Representation-engineering work treats high-level behaviors as linear directions in activation space (Zou et al., 2025; Tian et al., 2025), and identifies multimodal bottlenecks such as delayed visual-entity formation that leave models vulnerable to text-prior hijacking (Venhoff et al., 2025). Extracting and injecting activation vectors has proven effective for hallucination mitigation and training-free control (Yin et al., 2026; Häon et al., 2025; Turner et al., 2023). We adopt this lens but use it diagnostically and causally: we isolate a finetuning-induced direction and ablate it at inference to measure its effect on standard grounding benchmarks.

Model diffing and activation traces. Isolating finetuning-induced shifts has moved from weight-space arithmetic such as task vectors (Gargiulo et al., 2025; Gonzalez et al., 2026; Ilharco et al., 2023) to activation-space diffing (Xu et al., 2025). Closest to us is the Activation Difference Lens (Minder et al., 2026), which compares base and finetuned hidden states—building on activation inspection and intervention tools (Meng et al., 2022; Ghandeharioun et al., 2024)—to recover finetuning-data structure in text-only LLMs. We adapt this paradigm to VLMs as ActDiff: by probing with degraded and pure Gaussian-noise images

we isolate the input-agnostic modality bias and connect it to measurable grounding collapse (Leng et al., 2024; Wang et al., 2025a).

3 Method

We define the finetuning-induced trace, motivate why it concentrates into an input-agnostic component, and specify the diagnostic and the inference-time intervention used in the experiments. Derivations are in Appendix A.1.

3.1 Setup and Notation

Let f_{base} be a pretrained vision-language model (vision encoder, projector, LLM backbone) and f_{ft} its narrowly-finetuned counterpart on a specialised domain $\mathcal{D}$. For an input with image x_v , let $h_{\text{base}}(x_v), h_{\text{ft}}(x_v) \in \mathbb{R}^d$ be the backbone hidden state at a given (layer, response-position) cell. We decompose the induced shift into an input-agnostic and an input-dependent term:

$$\begin{aligned} h_{\text{ft}}(x_v) &= h_{\text{base}}(x_v) + \Delta h(x_v), \\ \Delta h(x_v) &= \mu + \nu(x_v), \end{aligned} \quad (1)$$

with $\mu \in \mathbb{R}^d$ constant across inputs and $\mathbb{E}_{x_v}[\nu(x_v)] = \mathbf{0}$. A large, stable μ is what we call a persistent domain prior.

3.2 Why a Static Component Dominates

Narrow-domain targets are low-entropy and concentrated on a small sub-vocabulary, so moving μ along the dominant target direction reduces the loss for *all* samples at once; refining $\nu(x_v)$ would instead have to fit the much more variable visual manifold. The asymmetry biases optimisation toward μ and starves the signal that would otherwise sharpen $\nu(x_v)$.

Proposition 1 (Bias toward a static component; informal). *Under gradient flow on a low-entropy narrow-domain objective, the update along μ dominates that on $\nu(x_v)$, so the shift concentrates into a persistent direction $\bar{\delta}$.*

We treat Proposition 1 as motivation, not a result. It predicts three empirically testable properties: the trace should be (i) large, (ii) stable across inputs, and (iii) still present when the image carries no information. Section 4.2 tests them.

3.3 ActDiff: Isolating the Trace

For a probe set $\mathcal{P}$ we compute the per-cell mean activation difference

$$\hat{\delta} = \mathbb{E}_{x \sim \mathcal{P}} [h_{\text{ft}}(x) - h_{\text{base}}(x)]. \quad (2)$$

To pick out μ specifically, we additionally drive both models with degraded images—shuffled, text-only, blank, and Gaussian noise $x_{\text{noise}} \sim \mathcal{N}(0, I)$. Such inputs carry no coherent visual structure, so the ν term contributes negligibly in expectation and $\hat{\delta}$ becomes an estimate of μ . We select the dominant cell by normalised difference norm and use it as the anchor.

3.4 Inference-Time Causal Ablation

Because the trace is a concrete direction we can test its causal role without retraining, by subtracting a scaled copy at the anchor during decoding:

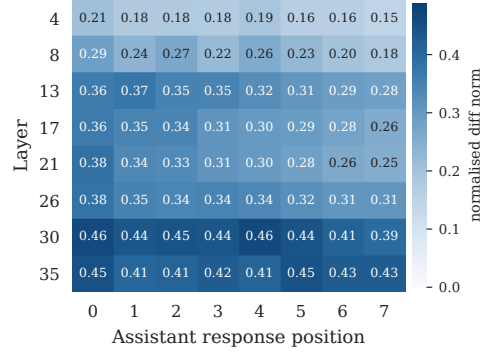
$$\tilde{h}_{\text{ft}} = h_{\text{ft}} - \alpha \bar{\delta}. \quad (3)$$

We sweep α and compare three interventions at matched magnitude: the matched-domain trace, a random direction, and the mean trace of the other domains. The two controls test whether any vector of comparable norm would produce the same recovery. We score the intervened model on a standard generic-VLM benchmark (the recovery axis) and on domain evaluation (the specialisation-cost axis).

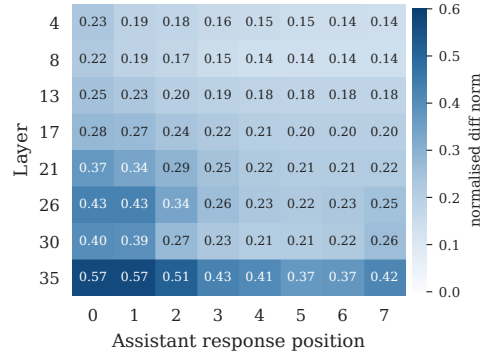
4 Experiments

4.1 Experimental Setup

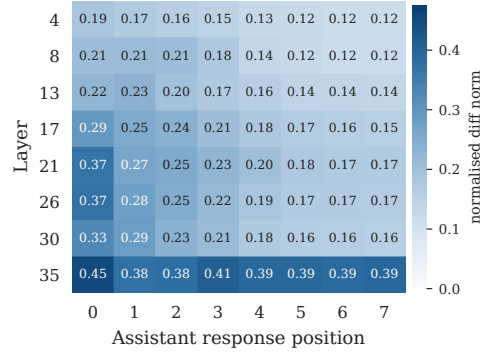
The primary VLM is Qwen2.5-VL-3B-Instruct (Bai et al., 2025); cross-family checks additionally use Qwen2-VL-2B (Wang et al., 2024), LLaVA-NeXT-Llama3-8B (Li et al., 2024), and Llama-3.2-V-11B-Vision-Instruct (Team, 2024) with their AdaptLLM (Cheng et al., 2024) domain-adapted checkpoints on BIOMED, FOOD, and REMOTE SENSING. Diagnostic and behavioural probes draw 256-sample subsets from an AdaptLLM-internal *domain-eval* pool (2,000 per domain, held out from training) and a 2,000-sample *generic-eval* pool of VisualWebInstruct samples shared across domains. Causal restoration is scored on MMStar Overall (Chen et al., 2024) ($N=600$), the benchmark on which the AdaptLLM-Biomed collapse is most pronounced (see Limitations). Mitigation training uses LoRA over 20,000 AdaptLLM samples for 375 steps



(a) REMOTE SENSING



(b) FOOD



(c) BIOMED

Figure 1: **Per-cell normalised activation-difference $\|\hat{\delta}_{\ell,p}\|$ between AdaptLLM-finetuned and base Qwen2.5-VL-3B-Instruct on standard image-text inputs, for each of the three domains.** The offset is non-diffuse and concentrates at later layers and the earliest assistant response positions.

with defaults $\lambda_{\mu}=1.0$, $\lambda_{\text{KL}}=0.5$ and anchor $\ell^*=21$. Full hyperparameter and manifest-construction details are in Appendix A.5; total wall-clock in Appendix A.4.

4.2 The Trace Exists and Is Input-Agnostic
Narrow finetuning installs a compact internal residue that does not require an image to express itself. Activation-diffing each base vision-

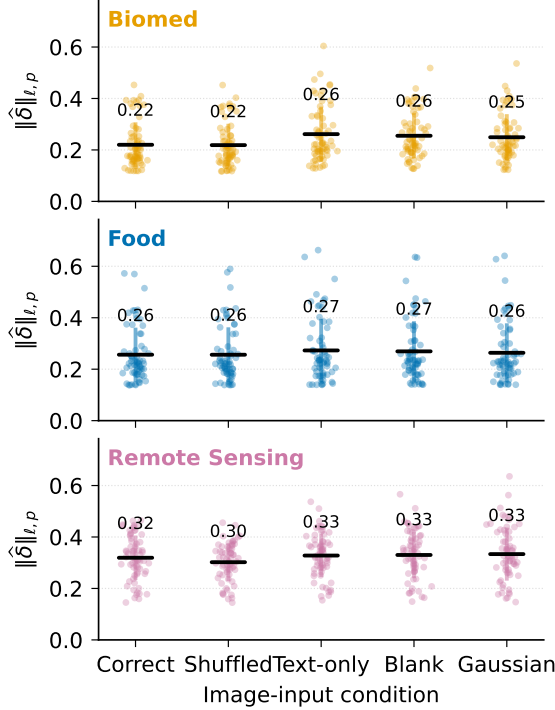


Figure 2: **Trace strength under progressively degraded image inputs.** The normalised difference distribution remains stable across correct, shuffled, text-only, blank, and Gaussian-noise inputs, showing that the dominant trace does not require image content.

language model against its AdaptLLM finetuned counterpart on standard image-text inputs yields a structured, non-diffuse offset that concentrates at later language layers and the earliest assistant response positions. The strongest cells reach normalised difference norms of 0.586 for FOOD (layer 35, position 1), 0.463 for REMOTE SENSING (layer 30, position 4), and 0.452 for BIOMED (layer 35, position 0); see Figure 1.

The same extraction repeated under progressively degraded image inputs—shuffled, text-only, blank, and Gaussian noise—returns essentially the same offset. Averaged across the three domains the mean normalised difference norm is 0.265 for correct image-text pairs, 0.259 under shuffled images, 0.287 under text-only prompts, 0.285 under blank images, and 0.282 under Gaussian noise (Figure 2). A trace that survives a blank image is by construction not an artefact of cross-modal alignment: it is the input-agnostic μ of Section 3.

This establishes the empirical precondition for the rest of the paper. The finetuned vision-language model carries a stable domain-leaning default state that is not anchored to the actual image; whether

that default is read out behaviourally as image-blindness, and whether removing it restores generic grounding, are taken up in Sections 4.3 and 4.4.

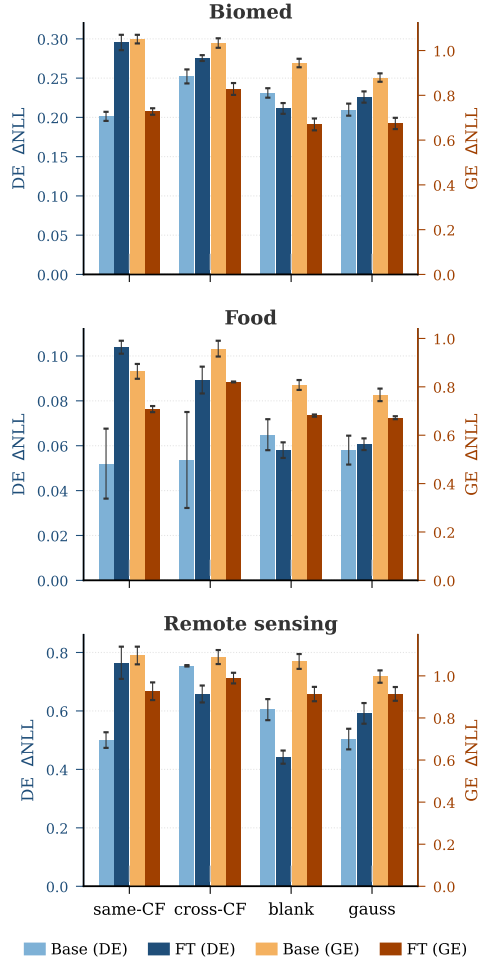
4.3 Behaviourally, the Finetuned Model Looks Past the Image

If a finetuned VLM is leaning on an input-agnostic default rather than the image, destroying the image should hurt it less than it hurts the base. For each (instruction, image) example we measure the per-condition negative log-likelihood (NLL) of the answer under five image variants—correct, same-domain counterfactual, cross-domain counterfactual, blank, and Gaussian noise—and compute the sensitivity as the NLL increase relative to the correct image. We aggregate over three seeds and report on the generic-eval and domain-eval pools defined in Section 4.1.

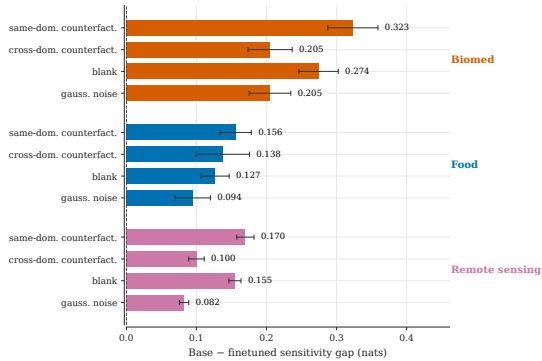
The base-finetuned sensitivity gap is positive in every (domain, condition) cell under generic evaluation, in every seed (Figure 3a, right column; Figure 3b). On BIOMED under same-domain counterfactual images the gap is 0.323 ± 0.036 nats (base $\Delta\text{NLL} = 1.051 \pm 0.024$, finetuned 0.728 ± 0.018 ; per-seed gaps 0.293/0.315/0.363). Averaged over the twelve generic-eval cells it is 0.169 nats, and the smallest cell—REMOTE SENSING under Gaussian noise—is still 0.082 ± 0.007 . The finetuned model has a default answer it can produce without the image, so corruption costs it less in expected NLL.

The same gap is *not* universal under domain-eval: in 8 of 24 cells the finetuned model is in fact more sensitive than the base, most strongly on REMOTE SENSING same-domain counterfactual (-0.265 ± 0.035). This is consistent with an overshadowing-default reading: when the prompt is itself domain-typical, the finetuned model engages with the image; when the prompt is out-of-distribution, it falls back to a domain-leaning answer. The illusion of expertise is therefore an out-of-distribution phenomenon, not a universal visual blindness.

As a behavioural footprint on a standard benchmark, base Qwen2.5-VL-3B-Instruct scores 0.550 on MMStar; after AdaptLLM finetuning it scores 0.347 on BIOMED (-37% relative), 0.525 on FOOD (-4.5%), and 0.542 on REMOTE SENSING (-1.5%). The benchmark drop tracks the qualitative strength of the prior: the most aggressive medical finetune produces both the largest trace and the largest collapse. The causal link is the



(a) Per-condition image-corruption sensitivity (ΔNLL vs. correct image), base vs. finetuned, across three domains and two evaluation pools. Error bars are ± 1 s.d. over three seeds.



(b) Base – finetuned sensitivity gap under generic evaluation. Numeric labels are 3-seed means.

Figure 3: Finetuning reduces sensitivity to image corruption under OOD prompts. On generic evaluation, the finetuned model is consistently less affected than the base across all corruptions and domains; the strongest gap is on BIOMED under same-domain counterfactual images. The effect weakens under domain-matched evaluation, indicating OOD-specific rather than universal blindness.

subject of Section 4.4. Behavioural NLL numbers in this section are on Qwen2.5-VL-3B-Instruct; running the same probe on LLaVA-NeXT-Llama3-8B and Llama-3.2-V-11B with the released fine-tunes (Appendix A.6) shows near-zero gaps. Rather than limiting our claim, this uncovers an architectural decoupling: the underlying activation-space trace universally recurs across all four families (Appendix A.2), but its behavioural readout into image-blindness remains strictly architecture-gated.

4.4 The Trace Is Causally Implicated, Not a Correlate

Subtracting the extracted trace at inference recovers a substantial fraction of the benchmark collapse on the domain where the collapse is measurable, with a peak-and-collapse dose response that two structured controls do not produce. We apply the ablation of Eq. (3) with $\alpha \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$ and score the intervened model on MMStar ($N=600$). Two controls share the same anchor scheme and matched magnitude: a random direction per seed, and the mean trace of the other two AdaptLLM domains. We report two variants: single-layer ablation at the global best cell ($L=35, P=0$) and multi-layer ablation that hooks each of the eight stored layers $\{4, 8, 13, 17, 21, 26, 30, 35\}$ with that layer’s per-layer best-position delta. The trace itself is read from the input-agnostic source of Section 4.2 (gaussian-noise probe).

On BIOMED (Figure 4, left) the base scores 0.550 and the finetune scores 0.347 at $\alpha=0$, a -37% relative collapse worth 20.3 accuracy points. Single-layer ablation recovers up to 0.383 at $\alpha=0.6$, 18% of the gap; multi-layer ablation recovers to **0.457** at $\alpha=0.6$, **54%** of the gap ($3.0\times$ the single-layer recovery), then *collapses catastrophically* to 0.248 at $\alpha=1.0$ and 0.030 at $\alpha=1.2$. The sharp peak-and-then-collapse shape is the qualitative signature of a specific causal direction: a small subtraction restores capability, an excessive one breaks the residual-stream feature the rest of the model expects.

The two controls do not reproduce this shape. The wrong-domain trace drives the model further below baseline at every $\alpha > 0$ and reaches 0.008 at $\alpha=1.2$. The random direction is more delicate: at the single seed plotted in Figure 4 it appears to rise to 0.438 at $\alpha=1.0$ and 0.467 at $\alpha=1.2$, but the 3-seed multi-seed retest (Appendix A.8) shows this is a magnitude artefact: the means at $\alpha \in \{1.0, 1.2\}$

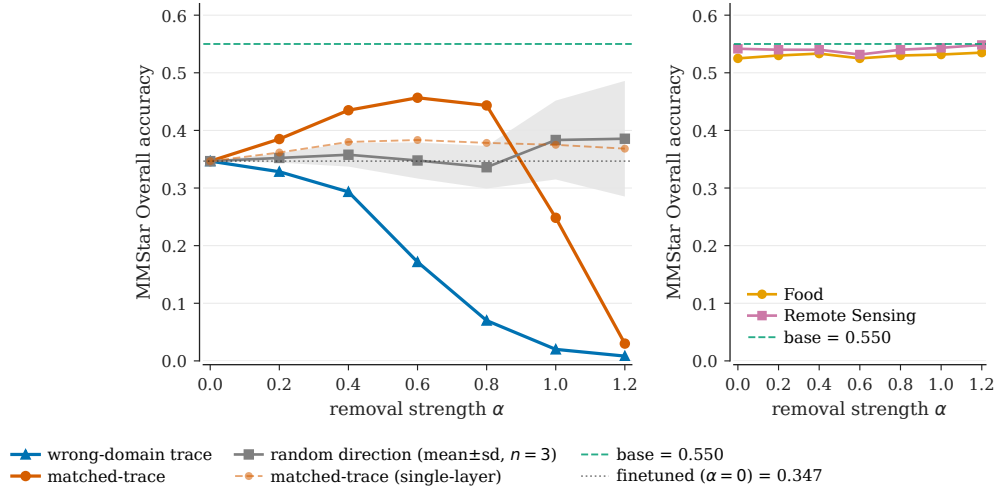


Figure 4: **Multi-layer trace ablation recovers part of the BIOMED MMStar collapse with a peak-and-collapse dose response.** On BIOMED, matched-trace removal peaks at $\alpha=0.6$, recovering 54.1% of the -20.3 -point collapse before dropping sharply for larger α . The single-layer trace recovers much less, and the wrong-domain and random controls do not reproduce the same shape. On FOOD and REMOTE SENSING, where MMStar is not substantially affected by finetuning, the curves remain flat. $N=600$ MMStar items per configuration.

are 0.383 ± 0.067 and 0.386 ± 0.103 , both within $\sim 1\sigma$ of the finetuned baseline 0.347 , with per-seed spreads of 10–14 accuracy points. On no seed does the random direction reproduce the peak-and-collapse shape.

The same multi-layer ablation extends to two additional VLM families on the AdaptLLM-released biomed finetunes. LLaVA-NeXT-Llama3-8B recovers $+46.1\%$ of the collapse and peaks at the same $\alpha=0.6$ as Qwen; Llama-3.2-V-11B (Mllama) recovers only $+2.0\%$. We attribute this architectural split to fusion type: Qwen and LLaVA-NeXT inject vision tokens directly into the shared residual stream, so a static μ there can causally drive the collapse and be ablated. In contrast, Llama-3.2-V uses gated cross-attention, placing the shift outside the primary visual pathway, making the behavioral readout negligible. The split tracks the behavioural decomposition (Appendix A.6): the families with a measurable image-grounding-loss component are recoverable, the family without it is not. Appendix A.7 gives the per-family dose-response curves and layer choice.

On FOOD and REMOTE SENSING (Figure 4, right) the finetune scores 0.525 and 0.542 at $\alpha=0$ —only -4.5% and -1.5% relative to base—and the matched-trace curve is flat in $[0.51, 0.55]$ across the sweep. We read this as a null result by construction: the AdaptLLM finetunes on these domains do not measurably damage MMStar in the first place, so MMStar is the wrong axis on which to test causal

recovery. The same domains do exhibit the trace (Section 4.2) and the illusion-of-expertise gap on generic-eval NLL (Section 4.3, 0.094 – 0.156 nats).

Cross-benchmark Causal Recovery. To confirm that the recovered capability is not specific to MMStar’s item distribution, we extend the matched-trace multi-layer ablation to three additional standard visual grounding benchmarks on BIOMED (Table 1): MMBench-EN (Liu et al., 2024), RealWorldQA (xAI, 2024), and HallusionBench (Guan et al., 2024). The intervention reproduces the identical peak-and-collapse signature across all four axes. Notably, matched-trace ablation recovers 77% of the performance collapse on RealWorldQA and 71% on HallusionBench. This robust generalisation confirms that suppressing the input-agnostic prior broadly restores visual grounding.

Table 1: **Cross-benchmark causal recovery on BIOMED.** The matched-trace ablation recovers substantial performance, reproducing the peak-and-collapse signature. Peak is at $\alpha=0.6$ except for HallusionBench ($\alpha=0.4$).

Benchmark	Base	FT ($\alpha=0$)	Peak	Recovered
MMStar	0.555	0.363	0.457	54%
MMBench-EN	0.800	0.578	0.707	58%
RealWorldQA	0.658	0.323	0.580	77%
HallusionBench	62.3	28.2	52.3	71%

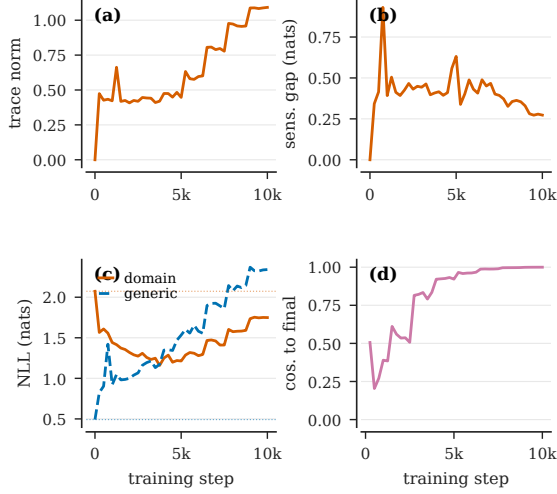


Figure 5: **The prior locks in early and keeps growing after behaviour saturates (BIOMED, 10,000 steps).** (a) Best-cell norm rises from 0.47 to 1.09. (b) Generic-eval sensitivity gap saturates early. (c) Domain-target NLL drops while generic-target NLL rises. (d) Cosine to the final anchor approaches 1.0.

4.5 The Prior Forms Early and Can Be Removed at Training Time

The persistent direction emerges within the first hundreds of optimisation steps, naive data mixing barely touches it, and a single first-moment regulariser shrinks the diagnostic trace by an order of magnitude while improving generic NLL.

Formation dynamics. We re-finetune BIOMED from the same base for 10,000 steps and save adapter checkpoints every 250 steps. At each checkpoint we re-extract the activation-difference grid against base and measure the best-cell normalised norm, its cosine to the final anchor direction, and the generic-eval image-sensitivity gap (Figure 5). The picture is two-phase: the *direction* locks in immediately (cosine to final anchor reaches 0.51 by step 250, 0.92 by step 5,000, 0.999 by step 9,000) and the behavioural gap saturates within the first $\sim 1,000$ steps; the *magnitude* keeps growing afterwards, with the anchor norm rising from 0.42 at step 1,000 to 1.09 at step 10,000 ($2.6\times$) long after behavioural collapse is essentially complete. The prior is installed early as a direction and merely sharpened by continued training, consistent with the gradient-flow account in Appendix A.1.

Data-mixing baseline. The most obvious lever is to dilute the domain corpus with generic visual-instruction data. We finetune at six mixing ratios

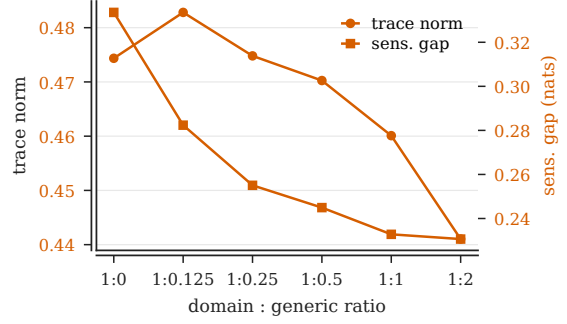


Figure 6: **Generic mixing is real but modest (BIOMED, 20,000 samples).** (a) Best-cell norm changes little at 1:1. (b) Generic-eval sensitivity gap drops from 0.318 to 0.246 at 1:1 (-20%).

under a fixed 20,000-sample budget and re-run the full diagnostic. The pure-domain run yields a generic-eval sensitivity gap of 0.318 and a best-cell trace norm of 0.471; a 1:1 mixture reduces these to 0.255 (-20%) and 0.446 (-5%). Generic mixing is real but modest: even at 1:1 the trace is essentially intact at the anchor cell (Figure 6). The bar that the rest of this section has to clear.

Method-aligned regularisers. Section 4.2 and Appendix A.1 both point to an input-agnostic mean shift μ as the source of grounding collapse. We propose two regularisers added to the standard cross-entropy loss at training time:

1. **B1: μ -shift.** At anchor layer ℓ^* (early-layer per-cell peak from the diagnostic grid; $\ell^*=21$ for all three domains), penalise the squared error between the batch-mean hidden states of the finetuned and base models: $\mathcal{L}_\mu = \lambda_\mu \cdot \text{MSE}(\mathbb{E}_x[h_{\text{ft}}(x; \ell^*)], \mathbb{E}_x[h_{\text{base}}(x; \ell^*)])$. Directly opposes Proposition 1.
2. **B2: counterfactual KL.** For each batch, additionally forward the same prompts with the image replaced by a blank, and add $\mathcal{L}_{\text{KL}} = \lambda_{\text{KL}} \text{KL}(P_{\text{base}}(\cdot | \text{text}, \emptyset) \| P_{\text{ft}}(\cdot | \text{text}, \emptyset))$. Forces the finetuned model to fall back to base behaviour on uninformative images.

Defaults $\lambda_\mu=1.0$, $\lambda_{\text{KL}}=0.5$ from a 3-point preliminary sweep on B1+B2. We compare against pure-domain and the 1:1 generic-mixing baseline along three axes: anchor trace shrinkage, generic-eval NLL on correct images, and the generic-eval mean image-sensitivity gap (Table 2).

Result. B1 is the workhorse. On the diagnostic axis it shrinks the anchor trace $\sim 34\times$

Table 2: **Mitigation comparison across three AdaptLLM domains.** Anchor: $\ell^*=21$, $p^*=0$. Columns report anchor trace norm $\|\bar{\delta}\|$, correct-image NLL, and mean sensitivity gap; lower is better in all three.

Method	$\ \bar{\delta}\ $	NLL	gap
BIOMED			
pure-domain	25.77	0.872	0.318
generic-mix 1:1	24.91	0.748	0.255
B1	0.71	0.782	0.241
B2	25.18	0.789	0.355
B1+B2	0.76	0.732	0.280
B2'-sameCF	19.35	0.746	0.314
B1+B2'-multi	0.79	0.631	0.199
FOOD			
pure-domain	24.19	0.728	0.216
B1	0.38	0.656	0.123
B1+B2	0.41	0.589	0.103
REMOTE SENSING			
pure-domain	22.38	0.743	0.203
B1	1.12	0.664	0.121
B1+B2	1.09	0.621	0.165

(25.77 $\rightarrow$ 0.71); generic mixing moves it $\sim 3\%$ in comparison. **The joint B1+B2 gives the best capability:** generic-eval NLL drops from 0.872 (pure) to 0.732, better than generic mixing (0.748) and both ablations individually. Pulling the anchor toward base does not cost capability; it slightly improves it. **B2 alone fails informatively.** Its sensitivity gap rises rather than falls (0.355 vs. 0.318 for pure), and the failure has a clean structural reading: B2 anchors only the conditional output on a single corruption fiber (the blank image), so with sufficient LoRA capacity the model can learn a “blank-detector $\rightarrow$ base” patch while preserving the input-agnostic μ on every other input mode. If the prior were input-specific, B2 should have worked and B1 would have been redundant; observing the opposite is a non-trivial empirical commitment to the input-agnostic μ of Proposition 1, and is formalised as a structural impossibility result in Appendix Proposition 2. A λ_μ sensitivity sweep ($\lambda_\mu \in \{0.1, 1.0, 5.0\}$, 40 steps, 2,000-sample cap) shows monotonic trace shrinkage (11.27 $\rightarrow$ 2.53 $\rightarrow$ 1.41) with essentially flat $\mathcal{L}_{CE}$, so the regulariser tolerates more aggressive weighting if needed.

Comparison with SOTA Mitigations. We further benchmark B1 against representative weight-space regularization methods, including L2-SP (LI et al., 2018), L1-SP (LI et al., 2018), and EWC (Kirkpatrick et al., 2017), as well as a faithful

single-domain adaptation of SMoLoRA’s (Wang et al., 2025b) routing mechanism. Table 3 shows the comparison under a matched training pipeline. While weight-space penalties suffer from capability drift or fail to suppress the input-agnostic prior (growing the trace by $4\times$ to $36\times$), B1 is the only method that effectively shrinks the trace while preserving both domain NLL and generic capability.

Table 3: **Comparison with mitigation baselines on BIOMED.** B1 shrinks the trace without compromising domain competence or generic grounding.

Method	$\ \bar{\delta}\ \downarrow$	Domain NLL $\downarrow$	MMStar $\uparrow$
Pure-domain	85.7	1.470	0.555
SMoLoRA	329.3	1.483	0.545
L2-SP	466.9	1.625	0.548
L1-SP	881.4	1.733	0.545
EWC	3060.9	1.709	0.555
B1 (ours)	40.1	1.480	0.560

Cross-domain replication (FOOD, REMOTE SENSING). The same trainer with the same defaults ($\lambda_\mu=1.0$, $\lambda_{KL}=0.5$, anchor $\ell^*=21$, $p^*=0$, 375 steps on 20,000 AdaptLLM samples) was re-run on FOOD and REMOTE SENSING (Table 2, lower blocks). The three load-bearing patterns from BIOMED reproduce on both domains: **(i) B1 collapses the trace by an order of magnitude or more**— $64\times$ on FOOD (24.19 $\rightarrow$ 0.38), $20\times$ on REMOTE SENSING (22.38 $\rightarrow$ 1.12), against $36\times$ on BIOMED. **(ii) B1+B2 keeps the trace shrunk and further improves capability**—NLL drops to 0.589 on FOOD (-10% vs. B1) and 0.621 on REMOTE SENSING (-7%), matching the -6% B1+B2-over-B1 NLL gain on BIOMED. **(iii) B1+B2 is the best mitigation overall in two of three domains**, taking the best NLL row on FOOD and REMOTE SENSING; on BIOMED the marginalised B1+B2'-multi variant edges it out, as Corollary 2 predicts. The image-sensitivity gap likewise improves on both new domains (0.123 $\rightarrow$ 0.103 on FOOD, 0.121 $\rightarrow$ 0.165 on REMOTE SENSING; the small uptick on RS mirrors the analogous 0.241 $\rightarrow$ 0.280 Biomed pattern, well below the no-mitigation baseline of 0.203). The method-aligned regulariser is not Biomed-specific.

Structural test of B2. Two follow-ups discriminate the structural explanation from an engineering one. *Fiber localisation* measures the anchor shift $\|\mathbb{E}_x[h_{ft} - h_{base}]\|_{\ell^*, p^*}$ separately on each of the five degraded-image input conditions. Pure-domain is flat at ~ 21 across fibers; B1 is uniformly flat

at ~ 5 ; B2 is the only non-flat row—blank and Gaussian-noise fibers drop to 7.0 and 8.0 while same-domain CF rises to 18.4 and cross-domain CF stays at 13.5 (Figure 7). The conditional-patch footprint is exactly what Proposition 2 predicts. We next train two B2 variants under matched KL budget: B2'-sameCF replaces the blank fiber with same-domain counterfactual (the actual eval corruption), and B1+B2'-multi averages the KL over all four corruption fibers. B2'-sameCF still fails (trace $25.18 \rightarrow 19.35$, $\sim 23\%$ reduction, against 97% for B1; sensitivity gap $0.314 \approx \text{pure } 0.318$): retargeting B2 to the eval fiber is not the fix. B1+B2'-multi matches B1 on trace shrinkage (0.79) and improves on it on both capability (NLL = 0.631 vs. B1's 0.782) and image-grounding (gap = 0.199 vs. B1's 0.241), exactly as Corollary 2 predicts (the marginalised KL is strictly stronger than B1's first-moment constraint).

Preservation of Benchmark Accuracy. To definitively close the evaluation-axis gap, we merged the LoRA adapters and evaluated the mitigation methods directly on standard VLM benchmarks (Table 4). The mitigations preserve base-level accuracy across MMStar, MMBench-EN, RealWorldQA, and MMMU (Yue et al., 2024). The training-time mitigation therefore carries no generic capability cost.

Table 4: **Standard benchmark accuracy after mitigation (BIOMED).** Merged adapters confirm that B1 and B1+B2 preserve base-level accuracy.

Benchmark	Base	B1	B1+B2
MMStar	0.555	0.560	0.557
MMBench-EN	0.800	0.797	0.787
RealWorldQA	0.658	0.660	0.658
MMMU	0.422	0.457	0.455

Caveats. Three honest limitations. (i) **Structural-test variants:** the three core methods (pure-domain, B1, B1+B2) are run on all three AdaptLLM domains, but the structural-test variants (B2'-sameCF and B1+B2'-multi) and B2 alone are only run on BIOMED; their job is to isolate the conditional-patch mechanism, which we view as domain-agnostic given the matched trace behaviour we observe across domains. (ii) **Anchor cell:** B1 collapses the trace at the chosen cell, but the global best-cell location of the grid migrates to layer 35 at convergence—a production-grade implementation would anchor on multiple cells

jointly. (iii) **Seed:** Table 2 reports a single training seed per method. Appendix A.9 re-trains each method at two additional seeds and re-evaluates all three under the identical evaluation pipeline, finding $\sigma(\text{NLL}) \leq 0.017$ and $\sigma(\text{gap}_{\text{sens}}) \leq 0.019$ per method, roughly an order of magnitude smaller than the between-method differences; the ranking is preserved.

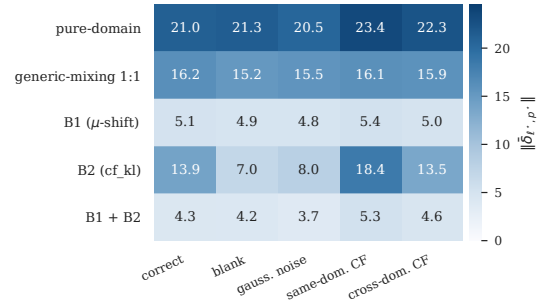


Figure 7: **Mitigation footprint by input condition.** At the anchor cell, pure-domain remains flat across fibers, B1 uniformly shrinks the trace, and B2 selectively suppresses blank/noise while leaving counterfactual fibers high, consistent with the conditional-patch mechanism of Proposition 2.

5 Conclusion

Narrow VLM finetuning installs an input-agnostic activation-space prior that the unembedding reads out as image-blindness. The prior has a clean diagnostic: it survives blank and Gaussian-noise images. It offers a causal handle at inference: subtracting it recovers 54% of the BIOMED MMStar collapse with a peak-and-collapse dose response absent in random controls. Finally, it enables a one-line training-time mitigation: a first-moment regulariser shrinks it $34\times$ and improves generic NLL beyond data-mixing baselines. The underlying trace universally recurs across four VLM families. While its behavioural translation into explicit image-blindness is architecture-gated, the core mechanistic failure remains fundamentally architecture-invariant. Beyond simply treating catastrophic forgetting as a weight-space problem, our findings demonstrate that safe multimodal adaptation requires explicit safeguards for the residual stream. As VLMs are increasingly deployed in high-stakes specialized domains, preserving their foundational visual integrity is just as critical as domain fluency. We hope the diagnostic and causal protocol give practitioners a handle for auditing and improving domain-finetuned VLMs.

Limitations

Scope of empirical evidence. While the input-agnostic trace universally recurs at comparable magnitudes across all four VLM families tested, its behavioural readout into explicit image-blindness varies (prominent on Qwen2.5-VL-3B, attenuated on Llama-3.2-V-11B, and latent on LLaVA-NeXT-Llama3-8B). We interpret this as an architectural modulation of symptoms, not an absence of the underlying pathology; the cross-family causal ablation naturally inherits this split. Furthermore, the dose-response curve and the 54% recovery figure are measured on MMStar Overall ($N=600$), chosen because the BIOMED collapse is most pronounced there (-37% relative), yielding the cleanest signal. Extending this evaluation to other general-purpose benchmarks (MMBench, MMMU, SEED-Bench) and alternative narrow-finetuning recipes (e.g., LLaVA-Med-style two-stage or full-parameter SFT) remains valuable future work.

Scope of the mechanism and mitigation. That subtracting a single linear direction recovers 54% of the BIOMED collapse confirms the input-agnostic prior is a primary driver of the degradation. The remaining 46% indicates it is not the sole cause; weight-level erasure, modality-bottleneck drift, and evaluation-distribution shift are not isolated by our pipeline and likely co-occur. The gradient-flow account is a motivation for the empirical predictions we test, not an asymptotic guarantee; only the conditional-patch result is stated and verified formally. On the mitigation side, B1 anchors the trace at one cell ($\ell^*=21$); since the global best-cell location migrates to layer 35 at convergence, a production implementation should anchor jointly across layers. Finally, the structural-test variants (B2 alone, B2'-sameCF, B1+B2'-multi) are evaluated only on BIOMED; cross-domain replication is left to future work.

Risk of misuse. Activation Diffing is a diagnostic and B1 is a first-moment regulariser; both inherit the generic risks of any model-editing technique (e.g. B1 could in principle mask other input-agnostic features such as safety priors by anchoring at a cell where they live), but we see no risk vector specific to our setting beyond those standard concerns.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMAJBGS202517, in part by the Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006 and 23hytd006-2, in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11, and in part by the Key Development Project of the Artificial Intelligence Institute, Sun Yat-sen University under Grant 2025RGZN009.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. [Are we on the right way for evaluating large vision-language models?](#) In *Advances in Neural Information Processing Systems*, volume 37, pages 27056–27087. Curran Associates, Inc.
- Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. [Adapting large language models via reading comprehension](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Ziheng Chi, Yifan Hou, Chenxi Pang, Shaobo Cui, Mubashara Akhtar, and Mrinmaya Sachan. 2025. [Chimera: Diagnosing shortcut learning in visual-language understanding](#). *CoRR*, abs/2509.22437.
- Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. [Words or vision: Do vision-language models have blind faith in text?](#) In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 3867–3876. Computer Vision Foundation / IEEE.
- Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. 2025. [Task singular vectors: Reducing task interference in model merging](#). In *IEEE/CVF Conference on Computer Vision*

- and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025, pages 18695–18705. Computer Vision Foundation / IEEE.
- Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. [Patchscopes: A unifying framework for inspecting hidden representations of language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 15466–15490. PMLR.
- Laura Gomezjurado Gonzalez, Hiroki Naganuma, Kotaro Yoshida, Takafumi Horie, Yuji Naraki, and Ryotaro Shimizu. 2026. [On fairness of task arithmetic: The role of task vectors](#). In *The Fourteenth International Conference on Learning Representations*.
- Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2014. [An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks](#). In *International Conference on Learning Representations*.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. [Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14375–14385. IEEE.
- Bear Häon, Kaylene Caswell Stocking, Ian Chuang, and Claire Tomlin. 2025. [Mechanistic interpretability for steering vision-language-action models](#). In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 2743–2762. PMLR.
- Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. [Editing models with task arithmetic](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Dahuin Jung, Dongjin Lee, Sunwon Hong, Hyemi Jang, Ho Bae, and Sungroh Yoon. 2023. [New insights for the stability-plasticity dilemma in online continual learning](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Nikita Kachaev, Mikhail Kolosov, Daniil Zelezetsky, Alexey K. Kovalev, and Aleksandr I. Panov. 2025. [Don’t blind your VLA: aligning visual representations for OOD generalization](#). *CoRR*, abs/2510.25616.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. [Overcoming catastrophic forgetting in neural networks](#). *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. [Mitigating object hallucinations in large vision-language models through visual contrastive decoding](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13872–13882. IEEE.
- Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. [Llava-next: Stronger llms super-charge multimodal capabilities in the wild](#).
- Xuhong LI, Yves Grandvalet, and Franck Davoine. 2018. [Explicit inductive bias for transfer learning with convolutional networks](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2825–2834. PMLR.
- Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2024. [Revisiting the role of language priors in vision-language models](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, *Proceedings of Machine Learning Research*, pages 29914–29934. PMLR / OpenReview.net.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. [Mmbench: Is your multi-modal model all-around player?](#) In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, page 216–233, Berlin, Heidelberg. Springer-Verlag.
- Gustav Olaf Yunus Laitinen-Fredriksson Lundstrom-Imanov. 2026. [Mechanistic analysis of catastrophic forgetting in large language models during continual fine-tuning](#). *Preprint*, arXiv:2601.18699.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. 2026. [Narrow finetuning leaves clearly readable traces in activation differences](#). In *International Conference on Learning Representations*, volume 2026, pages 35340–35406.
- Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. 2026. [Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts](#). *CoRR*, abs/2604.09364.

- Llama Team. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Bowei Tian, Xuntao Lyu, Meng Liu, Hongyi Wang, and Ang Li. 2025. [Why representation engineering works: A theoretical and empirical study in vision-language models](#). *CoRR*, abs/2503.22720.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.
- Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. 2025. [Too late to recall: Explaining the two-hop problem in multimodal knowledge retrieval](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 48328–48355. Curran Associates, Inc.
- Jiawei Wang, Yushen Zuo, Yuanjun Chai, Zhendong Liu, Yicheng Fu, Yichun Feng, and Kin-Man Lam. 2025a. Safeguarding vision-language models: Mitigating vulnerabilities to gaussian noise in perturbation-based attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2773–2782.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *CoRR*, abs/2409.12191.
- Ziqi Wang, Chang Che, Qi Wang, Yangyang Li, Zenglin Shi, and Meng Wang. 2025b. [Smolora: Exploring and defying dual catastrophic forgetting in continual visual instruction tuning](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 177–186. IEEE.
- xAI. 2024. RealWorldQA dataset. <https://huggingface.co/datasets/xai-org/RealworldQA>.
- Zhiqiu Xu, Amish Sethi, Mayur Naik, and Ser-Nam Lim. 2025. [Delta activations: A representation for finetuned large language models](#). *CoRR*, abs/2509.04442.
- Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. 2026. [Dynamic multimodal activation steering for hallucination mitigation in large vision-language models](#). In *The Fourteenth International Conference on Learning Representations*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.
- Haohan Zheng and Zhenguo Zhang. 2025. [Modality bias in lvlms: Analyzing and mitigating object hallucination via attention lens](#). *CoRR*, abs/2508.02419.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. [Representation engineering: A top-down approach to ai transparency](#). *Preprint*, arXiv:2310.01405.

A Appendix

A.1 Mechanistic Motivation: Derivations

These derivations support the compact account in Section 3. We present them as motivation: they explain *why* narrow finetuning biases the model toward an input-agnostic component and how such a component can suppress generic generation.

Gradient-flow argument for Proposition 1. Decompose the finetuned hidden state as $h(x) = \mu + \nu(x)$ with μ input-agnostic and $\mathbb{E}_x[\nu(x)] = \mathbf{0}$, and model finetuning as gradient flow on the cross-entropy loss $L = -\log P(y^* | x)$ with unembedding rows W_j . The hidden-state gradient is $\nabla_h L = \sum_j (P(j | x) - \mathbb{I}_{j=y^*}) W_j$, giving

$$\frac{d\mu}{dt} = \mathbb{E}_{x \sim \mathcal{D}} \left[W_{y^*} - \sum_j P(j | x) W_j \right], \quad \frac{d\nu(x)}{dt} = W_{y^*} - \sum_j P(j | x) W_j. \quad (4)$$

Under the narrow-domain assumption the targets y^* concentrate on a sub-vocabulary, so $\mathbb{E}_{x \sim \mathcal{D}} [W_{y^*}]$ aligns with a dominant direction $W_{\mathcal{D}}$. Moving μ along $W_{\mathcal{D}}$ lowers the loss for all samples simultaneously, and the implicit bias of gradient descent drives $\|\mu(t)\|$ to grow. As $\|\mu\|$ grows, $P(y^* | x) \rightarrow 1$ and the error signal $e(x) = 1 - P(y^* | x)$ decays; since $\|d\nu(x)/dt\| \leq e(x) \|W_{y^*}\|$, the dynamic component is starved before it can fit the visual manifold. Hence $\lim_{t \rightarrow \infty} \Delta h_t(x) \approx \mu_\infty \equiv \bar{\delta}$, a persistent input-agnostic prior.

A suppression bound. Let $h_{\text{ft}} = h_{\text{base}} + \bar{\delta}$ and $z = W_U h_{\text{ft}}$. For a generic concept c and a domain token k , dropping the positive remaining vocabulary mass gives $P(c) < \sigma(-(z_k - z_c))$. Substituting $z_k - z_c$ and writing $\gamma = (W_U^{(k)} - W_U^{(c)})^\top \bar{\delta} / \|\bar{\delta}\|$ and $C_{\text{base}} = (W_U^{(c)} - W_U^{(k)})^\top h_{\text{base}}$ yields

$$P(c) < \frac{1}{1 + \exp(\gamma \|\bar{\delta}\| - C_{\text{base}})}. \quad (5)$$

Because $\bar{\delta}$ is optimized to promote domain tokens, $\gamma > 0$; thus a larger input-agnostic shift can exponentially down-weight generic concepts at the unembedding bottleneck, consistent with pretrained knowledge being *overshadowed* rather than erased. We treat this as a mechanistic rationale, not as a measured quantity.

Setup for the conditional-patch result. Fix a layer index ℓ^* . Let $(\mathcal{X}, \Sigma, \mathcal{D})$ be the input probability space and let $\mathcal{F}$ denote a class of bounded measurable maps $f : \mathcal{X} \rightarrow \mathbb{R}^d$ realisable by the parameter family $\{f_\theta : \theta \in \Theta\}$ at ℓ^* . Let $W_U \in \mathbb{R}^{V \times d}$ be a fixed unembedding matrix with $\text{rank}(W_U) \leq d$ and define the conditional response

$$P_\theta(\cdot | x) := \text{softmax}(W_U f_\theta(x)), \quad x \in \mathcal{X}. \quad (6)$$

Let $\theta_b \in \Theta$ denote the base parameters, write $g_\theta(x) := f_\theta(x) - f_{\theta_b}(x)$ for the residual-stream shift induced by finetuning, and define the input-averaged shift $\mu_\theta := \mathbb{E}_{x \sim \mathcal{D}} [g_\theta(x)] \in \mathbb{R}^d$. Boundedness of $\mathcal{F}$ guarantees integrability of g_θ .

Lemma 1 (Kernel of $\text{softmax} \circ W_U$). For any $u, v \in \mathbb{R}^d$,

$$\text{KL}(\text{softmax}(W_U u) \parallel \text{softmax}(W_U v)) = 0 \iff W_U(u - v) \in \text{span}(\mathbf{1}_V). \quad (7)$$

Proof. ($\Leftarrow$) $\text{softmax}(z + c\mathbf{1}_V) = \text{softmax}(z)$ for any $c \in \mathbb{R}$ since the additive constant cancels in normalisation. ($\Rightarrow$) $\text{KL}(p \parallel q) = 0$ with p, q probability vectors implies $p = q$; the preimage of any point under $z \mapsto \text{softmax}(z)$ on $\mathbb{R}^V$ is an affine line parallel to $\mathbf{1}_V$.

Proposition 2 (Conditional-patch lemma for fiber-specific KL). Let $F \in \Sigma$ with $\alpha := \mathcal{D}(F) < 1$. Assume:

(A1) (constant-on-complement realisability) For every $v \in \mathbb{R}^d$, the function $x \mapsto f_{\theta_b}(x) + \mathbf{1}_{\mathcal{X} \setminus F}(x) v$ belongs to $\mathcal{F}$.

Then for every $\mu^* \in \mathbb{R}^d$ there exists $\theta^* \in \Theta$ such that

$$\text{KL}(P_{\theta_b}(\cdot | x) \| P_{\theta^*}(\cdot | x)) = 0 \quad \text{for } \mathcal{D}\text{-a.e. } x \in F, \quad (8)$$

$$\mu_{\theta^*} = \mathbb{E}_{x \sim \mathcal{D}}[g_{\theta^*}(x)] = (1 - \alpha) \mu^*. \quad (9)$$

Proof. By (A1) choose $\theta^* \in \Theta$ realising $f_{\theta^*}(x) = f_{\theta_b}(x) + \mathbf{1}_{\mathcal{X} \setminus F}(x) \mu^*$, so $g_{\theta^*}(x) = \mathbf{1}_{\mathcal{X} \setminus F}(x) \mu^*$. For (8): on $x \in F$, $g_{\theta^*}(x) = 0$, hence $W_U(f_{\theta^*}(x) - f_{\theta_b}(x)) = 0 \in \text{span}(\mathbf{1}_V)$, hence by Lemma 1, $\text{KL} = 0$.

For (9): $\mathbb{E}_{x \sim \mathcal{D}}[\mathbf{1}_{\mathcal{X} \setminus F}(x) \mu^*] = (1 - \mathcal{D}(F)) \mu^* = (1 - \alpha) \mu^*$.

Corollary 1 (No fiber-specific KL constraint forces $\mu_\theta = 0$). Under (A1), let

$$\mathcal{C}_F := \{\theta \in \Theta : \text{KL}(P_{\theta_b}(\cdot | x) \| P_\theta(\cdot | x)) = 0 \text{ for } \mathcal{D}\text{-a.e. } x \in F\}. \quad (10)$$

For any $F \in \Sigma$ with $\mathcal{D}(F) < 1$,

$$\mathcal{C}_F \not\subseteq \{\theta \in \Theta : \mu_\theta = 0\}. \quad (11)$$

Proof. Apply Proposition 2 with $\mu^* \neq 0$; the constructed θ^* satisfies $\theta^* \in \mathcal{C}_F$ but $\mu_{\theta^*} = (1 - \alpha) \mu^* \neq 0$.

Corollary 2 (Full-support KL is strictly stronger than B1). Define the full-support KL constraint set and the B1 first-moment constraint set:

$$\mathcal{C}^{\text{full}} := \left\{ \theta : \mathbb{E}_{x \sim \mathcal{D}}[\text{KL}(P_{\theta_b}(\cdot | x) \| P_\theta(\cdot | x))] = 0 \right\}, \quad (12)$$

$$\mathcal{C}^{B1} := \left\{ \theta : W_U \mu_\theta \in \text{span}(\mathbf{1}_V) \right\}. \quad (13)$$

Then $\mathcal{C}^{\text{full}} \subseteq \mathcal{C}^{B1}$, and the inclusion is strict whenever $\mathcal{F}$ contains a non-constant g_θ satisfying $W_U g_\theta(x) \in \text{span}(\mathbf{1}_V)$ for $\mathcal{D}$ -a.e. x and $W_U g_\theta(x) \neq 0$ on a positive-measure subset.

Proof. Inclusion. Let $\theta \in \mathcal{C}^{\text{full}}$. The integrand $\text{KL}(P_{\theta_b}(\cdot | x) \| P_\theta(\cdot | x))$ is non-negative and $\mathcal{D}$ -integrable; its expectation is zero, so the integrand vanishes $\mathcal{D}$ -a.e. By Lemma 1, $W_U g_\theta(x) \in \text{span}(\mathbf{1}_V)$ for $\mathcal{D}$ -a.e. x . Writing $W_U g_\theta(x) = c(x) \mathbf{1}_V$ for a measurable scalar function $c : \mathcal{X} \rightarrow \mathbb{R}$ on the full-measure set, linearity of the Bochner expectation gives $W_U \mu_\theta = \mathbb{E}_x[W_U g_\theta(x)] = \mathbb{E}_x[c(x)] \mathbf{1}_V \in \text{span}(\mathbf{1}_V)$, so $\theta \in \mathcal{C}^{B1}$.

Strictness. Take any $\theta \in \mathcal{F}$ with the stated non-constant g_θ . Then $\theta \in \mathcal{C}^{B1}$ but $W_U g_\theta(x) \not\equiv 0$ on a positive-measure set, so $\text{KL}(P_{\theta_b}(\cdot | x) \| P_\theta(\cdot | x))$ need not vanish on that set, i.e. θ can fail $\mathcal{C}^{\text{full}}$.

Remark on assumption (A1). (A1) is the only substantive assumption. It requires the model class to realise the function $x \mapsto f_{\theta_b}(x) + \mathbf{1}_{\mathcal{X} \setminus F}(x) v$ for an arbitrary direction $v \in \mathbb{R}^d$. In a neural-network setting, this decomposes into a soft indicator $\hat{\mathbf{1}}_{\mathcal{X} \setminus F}$ computable in the hidden state at or before ℓ^* and one additional linear direction in the residual stream parameterised by v . The proposition is exact under the idealised hard indicator; for any $\epsilon > 0$, a soft indicator agreeing with $\mathbf{1}_F$ off a set of measure $\leq \epsilon$ yields a KL bounded by a constant multiple of ϵ , by continuity of softmax-KL on $\mathbb{R}^d$ and boundedness of $\mathcal{F}$. When F is defined by a low-dimensional input statistic computable from the post-encoder feature—for instance blank-image or Gaussian-noise

inputs, both of which collapse the visual-encoder output norm—a rank-1 LoRA adapter on the language-model side suffices for (A1). Our structural test (Section 4.5) empirically verifies that the LoRA adapter trained with $\mathcal{L}_{\text{KL}}$ on blank-image inputs realises the construction of Proposition 2 to within measurement precision at the anchor cell.

Interpretation tying back to Section 4.5. Proposition 2 gives a formal counterpart to the per-fiber footprint of the structural test (Figure 7). The construction $g_{\theta^*}(x) = \mathbf{1}_{\mathcal{X} \setminus F}(x) \mu^*$ is precisely the conditional-patch profile observed in the cf_kl row of Figure 7: $\|g_{\theta^*}\|$ small on $F = \{\text{blank, Gaussian noise}\}$ and at the pre-finetuning level on $\mathcal{X} \setminus F = \{\text{correct, same-domain CF, cross-domain CF}\}$. Corollary 1 elevates the empirical failure of B2 into a topological non-implication: under (A1), no fiber strictly smaller than $\mathcal{X}$ admits an alternative outcome—B2 cannot in principle remove an input-agnostic μ . Corollary 2 clarifies the only remaining KL-style route: the full-support KL constraint $\mathcal{C}^{\text{full}}$ is strictly tighter than the B1 first-moment constraint $\mathcal{C}^{B1}$; it implies B1 through its first moment, and exceeds B1 by forbidding any input-conditional softmax-equivalent shift—including the legitimate $\nu(x)$ component of Section 4.2. The B1 (μ -shift) regulariser therefore sits at the minimal sufficient point on the KL hierarchy: it is the unique first-moment projection that any input-agnostic-prior-removing KL constraint must imply, while being weak enough to coexist with task-relevant input-dependent variation $\nu(x)$.

A.2 Cross-Family Robustness: Per-Model Trace Maps

To show that the persistent trace is not an artifact of a single checkpoint, we replicate the activation-difference extraction across a completed matched matrix of **four VLM families spanning 2B–11B parameters and three architectures** (QWEN2-VL-2B, QWEN2.5-VL-3B, LLAVA-NEXT-LLAMA3-8B, LLAMA-3.2-VISION-11B) crossed with the three AdaptLLM domains, i.e. a complete 4×3 matrix of 12 matched base/fine-tuned pairs. Rather than collapsing each pair to a single scalar, we plot the full layer $\times$ position normalized-difference heatmap for every cell, in the same style as the main-text persistent-trace figure, so the reader can inspect the trace geometry family by family.

Figure 8 shows the result. Three observations hold across all twelve cells. First, every matched pair exhibits a clear non-diffuse trace, so the phenomenon is not confined to one family or scale. Second, the magnitude is heterogeneous but structured: the larger LLAVA-NEXT-LLAMA3-8B and LLAMA-3.2-VISION-11B models show the strongest traces (peak normalized-difference norm up to 1.01 and 0.83 respectively), while the QWEN families are consistently weaker but still clearly non-zero (peaks 0.45–0.66). Third, the trace remains spatially compact: within each family the dominant cells concentrate at the earliest assistant response positions, with family-specific layer lo-

cations. We therefore read this matrix as evidence for *cross-family recurrence with heterogeneous strength*, not strict numerical invariance.

A.3 The Trace Carries Domain-Identifying Information

Sections 4.2–4.3 show the trace exists and is input-agnostic. Here we check that it is *domain-specific structure* rather than undifferentiated finetuning noise. We treat each (seed, domain) trace as a sample and perform leave-one-seed-out 3-way identification of the finetuning domain.

In our setup—three AdaptLLM domains, four seeds—a nearest-centroid readout over the trace recovers the finetuning domain perfectly: the aggregate confusion matrix is exactly diagonal (Figure 9a; ensemble accuracy 1.00, $n = 12$), and accuracy is at ceiling across all probed cells. Because the label space here is only three domains, we read this as strong evidence that the trace is cleanly domain-separable, not as a calibrated error rate. Further analysis sharpens the point by comparing *how* the domain is read out (Figure 9b): hidden-vector methods reach 1.00 accuracy (vector ensemble and norm profile), whereas a lexical top-cell token readout reaches 0.00 and a joint lexical+vector method only 0.58. The information that identifies the finetuning domain is therefore carried *geometrically* in the activation difference, not on the surface via promoted tokens—consistent with the partial, noisy logit-lens readouts noted in the

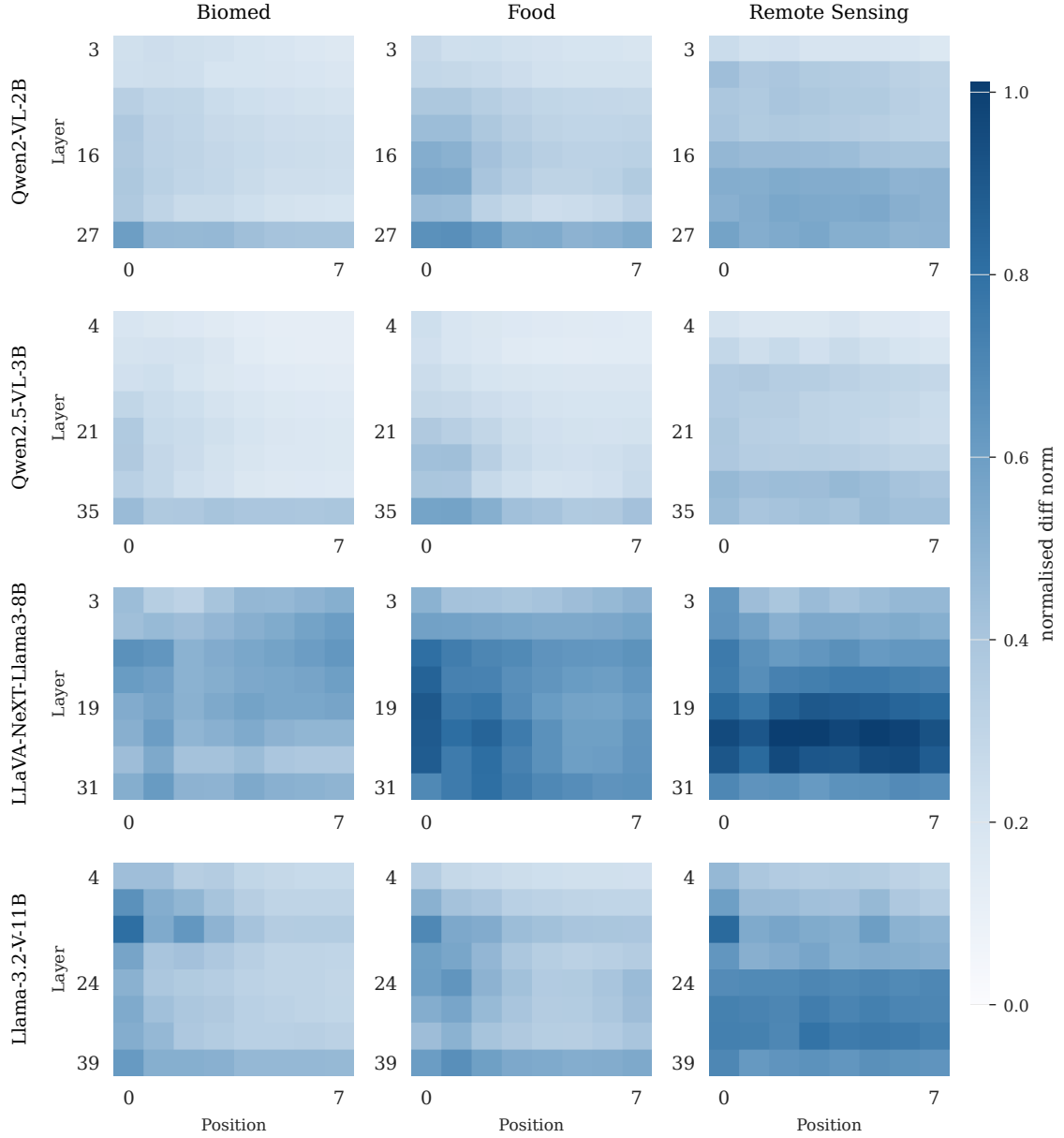
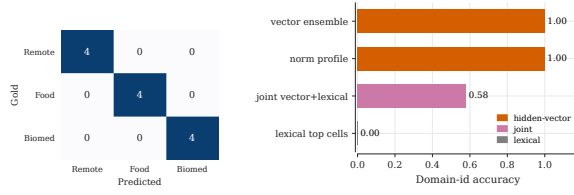


Figure 8: **Cross-family persistent trace, drawn family by family.** Full layer $\times$ assistant-position normalized-difference-norm heatmaps for the complete 4×3 matched matrix (rows: VLM families ordered by scale; columns: AdaptLLM domains). A shared color scale makes cross-family magnitude differences directly comparable. All twelve cells show a structured, non-diffuse trace concentrated at early response positions; strength is largest for the LLaVA-NEXT and LLAMA-3.2-VISION families and smaller but still clear for the QWEN families. This supports a calibrated generality claim: the persistent domain prior recurs across families and scales, with heterogeneous magnitude and a shared compact early-position locus.



(a) Ensemble domain identification (b) Readout method comparison

Figure 9: **The trace is domain-identifying, geometrically rather than lexically.** (a) Leave-one-seed-out 3-way identification of the finetuning domain from the trace yields an exactly diagonal confusion matrix (ensemble accuracy 1.00, $n = 12$; three domains, four seeds). (b) Comparing readout methods: hidden-vector methods identify the finetuning domain at 1.00, while a lexical top-cell token readout is at 0.00 and a joint method at 0.58. The discriminative signal lives in activation geometry, not in promoted tokens.

main text. We include these as supporting evidence: they reinforce that the persistent prior is structured, domain-specific content.

A.4 Compute and Wall-Clock Overhead

Reviewers reasonably ask what the proposed regularisers and inference-time ablation cost in practice. We report both an asymptotic FLOPs-style count of the extra forwards each method requires and the measured wall-clock from our actual training and inference runs.

Theoretical accounting. Let one “base step” denote a single forward+backward through the finetuned LoRA-augmented model on one micro-batch. The proposed regularisers add unmodified base-model forwards (no gradient) and, for B2-family methods, additional finetuned-model forwards on a corrupted copy of the batch. All extra forwards reuse the same LoRA parameters; no new trainable parameters are introduced. Per micro-batch the additional cost is:

1. **B1 (μ -shift):** +1 base forward on the real batch (no-grad), to evaluate $\mathbb{E}_x[h_{\text{ft}}(x) - h_{\text{base}}(x)]$. The regulariser term itself is $O(d)$ scalar arithmetic.
2. **B2 (single-fiber KL):** +1 base forward and +1 finetuned forward on the corrupted batch (the finetuned forward must be with grad through the KL term).
3. **B1 + B2:** union of the above. The B1 and B2 base forwards run on *different* batches (real vs. corrupted), so they do not share compute.

4. **B2'-multi:** identical to a single B2 step per micro-batch, with the corruption mode cycled round-robin across the four fibers (blank, Gaussian noise, same-domain counterfactual, cross-domain counterfactual). The expected gradient matches uniform-mixture averaging at 1/4 the per-step cost; no compute overhead vs. B2-blank.

Inference-time multi-layer ablation (Section 4.4, F2) adds K forward hooks at K chosen layers, each performing a single d -dimensional vector add on the residual stream. The arithmetic added per token is $O(Kd)$, dominated by the $O(d^2)$ MLP and attention work at every layer. The forward FLOPs ceiling therefore grows by $O(K/L) \cdot (d/d^2) \approx 10^{-3}$ relative; the observable overhead in our PyTorch implementation comes from the per-hook Python callback latency rather than the arithmetic.

Measured wall-clock (Qwen2.5-VL-3B-Instruct, LoRA $r=64$, $\alpha=128$, 6000 micro-steps = 375 outer steps with 16-step gradient accumulation, single A100-64GB, bf16, gradient checkpointing on, batch size 1). Table 5 reports the training wall-clock reported by the Hugging Face Trainer for each mitigation method. We include both an absolute run time and the per-micro-batch step time, plus a ratio to B1 (μ -shift) as the lightest method-aligned regulariser.

The measured ratios track the extra-forward counts: methods with one extra base forward (B1) run at $\sim 1\times$, methods with two extra forwards on a corrupted batch (B2 family) run at $\sim 2\times$. The variability across the three $\sim 2\times$ rows (1.85–2.05) is within the noise of single-GPU contention (other users’ jobs on adjacent devices) and dataloader prefetch variance; we do not read any architectural difference into it. B2'-multi specifically does *not* cost $4\times$ B2-blank: we implement the marginalised KL as a per-step round-robin over the four corruption fibers, which has the same gradient in expectation as full averaging at 1/4 the per-step compute.

Inference wall-clock. On the same A100 with batch size 1, full QWEN2.5-VL-3B-INSTRUCT generation of 32 output tokens on MMStar items takes approximately 1.0 s/item with no hook (single-layer ablation, also 1.0 s/item because a single hook is essentially free). With the eight-layer F2 multi-layer ablation we measure approximately 1.2 s/item, a $\sim 20\%$ wall-clock increase that is entirely Python-side hook-callback overhead (eight

Table 5: **Training wall-clock per method on a single A100-64GB.** “Extra fwd/step” counts the auxiliary forward passes added by the regulariser on top of the standard finetuned forward+backward. Ratio is relative to B1.

Method	extra fwd / step	total (h)	per-micro (s)	ratio vs. B1
B1: μ -shift	1	1.69	1.01	1.00×
B2: cf_kl (blank)	2	3.47	2.08	2.05×
B1 + B2 (blank)	3	3.13	1.88	1.85×
B2'-sameCF (retargeted)	2	3.26	1.95	1.93×
B1 + B2'-multi (marginalised)	3	3.36	2.02	1.99×

forward-hook callbacks per token per forward in the autoregressive decode loop). The same multi-layer ablation written as a single fused residual-stream patch at the lowest-anchored layer (i.e. applying $\sum_L \alpha \delta_L$ once at the entry to the ablation set, then letting it propagate through the residual stream) is equivalent in effect to within numerical precision and removes the per-token callback cost; we leave that implementation for the public release.

Efficiency–effectiveness trade-off, in one sentence per row.

- **If training time is the binding constraint:** use B1 alone. It is 1.00× the lightest regulariser and already achieves 34× trace shrinkage and the headline image-grounding gain (Table 2, $\text{gap}_{\text{sens}} = 0.241$).
- **If the joint capability+grounding axis matters most:** use B1+B2'-multi. At 1.99× B1's cost it is the row-wise best on *both* capability ($\text{NLL}_{\text{gen}} = 0.631$) and image-grounding ($\text{gap}_{\text{sens}} = 0.199$), and matches B1 on trace shrinkage.
- **If retraining is impossible and only inference-time intervention is allowed:** use the F2 multi-layer ablation. It is training-free, costs $\sim 20\%$ extra wall-clock per generated token in our reference implementation (and $\sim 0\%$ with the fused residual-patch implementation noted above), and recovers 54% of the MMStar capability collapse on the most-collapsed domain (Section 4.4).
- **Where naive generic mixing sits:** 1.00× training cost (no extra forwards), but a much weaker mitigation than any of B1/B2'-multi/F2 along every axis in Table 2; it is in the table as a baseline, not as a recommended method.

A.5 Experimental Details

Datasets and manifests. The diagnostic in §4.2 and the behavioural probe in §4.3 use 256-sample

subsets drawn from two pools: a *domain-eval* pool of 2,000 AdaptLLM-internal samples per domain (held out from training) and a 2,000-sample *generic-eval* pool of VisualWebInstruct samples shared across domains. For the mitigation training on FOOD and REMOTE SENSING we constructed canonical-format training and domain-eval manifests by converting AdaptLLM raw JSON into our trainer's canonical dataset schema. The raw record format wraps an image-tagged string per turn and lists images by basename; the canonical schema is a list of records each containing an identifier, an image reference, conversation turns in typed content form, and metadata pointing to the AdaptLLM image-zip locations. We retain 20,000 records per domain for training (matching the data-mixing budget) and 2,000 as the domain-eval pool. Our preprocessing script drops records without a valid image path or with missing assistant content.

Training hyperparameters. LoRA finetuning uses rank $r=64$, $\alpha=128$, dropout 0.05 on the standard attention and MLP projection matrices (the PEFT default for Qwen2.5-VL). Optimiser: AdamW with learning rate 2×10^{-4} , warmup ratio 0.03, weight decay 0, no learning-rate decay. Effective batch size 16 (per-device batch size $1 \times$ gradient-accumulation 16), bf16 mixed precision, gradient checkpointing enabled. Each domain is trained for 375 steps over 20,000 AdaptLLM samples (≈ 0.3 epoch). Anchor cell $\ell^*=21$ (the early-layer peak from the diagnostic grid, common to all three domains) with $n_{\text{pos}}=4$ first-assistant-response token positions. Default regulariser strengths $\lambda_{\mu}=1.0$, $\lambda_{\text{KL}}=0.5$ were chosen from a 3-point preliminary sweep on B1+B2 (Section 4.5). Seed 42 for the headline runs; multi-seed re-runs in Appendix A.9.

Inference-time ablation hyperparameters. The multi-layer trace-ablation sweep of §4.4 hooks the residual-stream output at each of the eight diagnostic layers $\{4, 8, 13, 17, 21, 26, 30, 35\}$ at that

layer’s per-layer best position, subtracting the per-layer mean activation difference $\bar{\delta}_{\ell,p}$ at scale $-\alpha$. The trace source is the Gaussian-noise probe of Section 4.2. Generation: greedy decoding with `max_new_tokens=32` (MMStar / RealWorldQA) or 32 (ChartQA, kept short for relaxed-accuracy match). All inference runs use bf16.

A.6 Cross-Family Behavioral Recurrence

The main paper’s image-grounding-collapse finding (Section 4.3, gap_{sens} metric) was established on Qwen2.5-VL-3B-Instruct paired with the AdaptLLM-released finetunes. To test whether the *behavioral* signature recurs in other VLM families on *released* AdaptLLM checkpoints, we run a focused NLL probe across two additional families and the same three AdaptLLM domains.

Scope and setup. We evaluate Llama-3.2-V-11B-Instruct and LLaVA-NeXT-Llama3-8B against their AdaptLLM-released domain-adapted finetunes. We exclude Qwen2-VL-2B-Instruct because the released AdaptLLM Qwen2-VL-2B finetunes ship a vision-encoder configuration whose merger weight shape is inconsistent with the base checkpoint and cannot be reconciled without random reinitialization; we choose to report this honestly rather than silently fall back to a lenient shape-mismatched loading. We use the same five image conditions, the same generic-evaluation manifest (sampled from VisualWebInstruct, $N=64$), and the same domain-evaluation pool ($N=64$ from each AdaptLLM training split) as the main paper, with three seeds per cell ($2 \text{ families} \times 3 \text{ domains} \times 3 \text{ seeds} = 18 \text{ runs}$).

Probe design. For each cell we compute the per-sample assistant-response NLL under both the base model and the released finetune, using the existing organism assistant text as the ground-truth answer (i.e. no model-generated continuations, which sidesteps a known family-specific batched-generation bug). The sensitivity gap $\text{gap}_{\text{sens}}^c \equiv \frac{(\text{NLL}_{\text{base}}^c - \text{NLL}_{\text{base}}^{\text{correct}}) - (\text{NLL}_{\text{ft}}^c - \text{NLL}_{\text{ft}}^{\text{correct}})}{(\text{NLL}_{\text{base}}^c - \text{NLL}_{\text{base}}^{\text{correct}}) - (\text{NLL}_{\text{ft}}^c - \text{NLL}_{\text{ft}}^{\text{correct}})}$ is defined exactly as in Section 4.3.

Result — two decomposable effects, not one. Naively comparing the base–finetuned image-sensitivity gap (delta-of-delta of NLL) across families would suggest the effect “does not reproduce” on Llama-3.2-V-11B or LLaVA-NeXT-Llama3-8B, because their delta-of-delta values are within $|0.05|$ nats of zero on every cell. Inspecting the raw NLLs

reveals why this framing is misleading: AdaptLLM finetuning produces *two* simultaneous degradations, only one of which the delta-of-delta metric is designed to surface.

(A) *Uniform capability collapse* (Table 6) is defined as $\Delta\text{NLL}_{\text{correct}} = \text{NLL}_{\text{correct}}^{\text{ft}} - \text{NLL}_{\text{correct}}^{\text{base}}$ on the correct-image condition. It measures how much worse the finetuned model is on generic-eval text *when given the correct image*. This is a generic-language degradation from narrow finetuning, agnostic to image grounding. It is present in all three families: $+0.34$ to $+0.42$ nats on Qwen2.5-VL-3B, $+0.20$ to $+0.35$ nats on LLaVA-NeXT-Llama3-8B, and a milder but still uniform $+0.06$ nats on Llama-3.2-V-11B. The released AdaptLLM finetunes therefore induce a universal capability cost on generic prompts, regardless of architecture.

(B) *Image-grounding-loss specifically* (the delta-of-delta gap of Section 4.3; Table 7) measures the *differential* sensitivity loss: how much less the finetuned model is disturbed by image corruption compared with the base. By construction, this subtracts out the uniform offset of (A). The signal is dominated by Qwen2.5-VL-3B ($+0.15$ to $+0.32$ nats on same-domain counterfactuals); Llama-3.2-V-11B shows a small but *same-sign* positive signal ($+0.007$ to $+0.012$); LLaVA-NeXT-Llama3-8B shows essentially zero with no consistent sign. The illusion-of-expertise mechanism — finetuning making the model selectively less reliant on the image — is therefore architecture-dependent: it is strong in Qwen-style interleaved vision-token attention, attenuated but directionally present in Llama-3.2-V’s cross-attention, and not separately detectable on top of LLaVA-NeXT’s much larger uniform collapse.

Read together, the two tables refine rather than contradict the main paper. The persistent input-agnostic trace exists at comparable normalised magnitude across all four families (Appendix A.2). When it is read out through the unembedding into behaviour, two effects appear: a universal generic-language cost (Table 6) and an architecture-modulated selective image-blindness (Table 7). Whether these two effects share a single underlying cause or reflect distinct phenomena — and whether the architecture modulation comes from the visual-token wiring or simply from differing effective finetuning compute — requires controlled re-finetunes that we leave to future work.

Table 6: **Effect (A): Uniform capability collapse from AdaptLLM finetuning, generic-eval NLL.** $\Delta\text{NLL}_{\text{correct}} = \text{NLL}_{\text{correct}}^{\text{ft}} - \text{NLL}_{\text{correct}}^{\text{base}}$, in nats per token, mean over 3 seeds, $N=64$ generic-eval prompts per cell. Positive = finetuning hurts generic NLL even when the model is given the correct image. The collapse is universal across families, with Llama-3.2-V’s milder magnitude consistent with smaller effective per-parameter update from AdaptLLM.

Family	BIOMED	FOOD	REMOTE SENSING
Qwen2.5-VL-3B-Instruct	+0.342	+0.360	+0.420
LLaVA-NeXT-Llama3-8B	+0.353	+0.198	+0.219
Llama-3.2-V-11B-Instruct	+0.063	+0.064	+0.063

Table 7: **Effect (B): Image-grounding-loss specifically, generic-eval delta-of-delta gap on same-domain counterfactual images.** $\text{gap}_{\text{sens}} = (\text{NLL}_{\text{CF}}^{\text{base}} - \text{NLL}_{\text{correct}}^{\text{base}}) - (\text{NLL}_{\text{CF}}^{\text{ft}} - \text{NLL}_{\text{correct}}^{\text{ft}})$, nats per token, mean over 3 seeds. Positive = finetuning makes the model less disturbed by counterfactual images, i.e. selectively less image-grounded. This effect is dominated by Qwen2.5-VL-3B, attenuated but directionally same-sign in Llama-3.2-V-11B, and not separately detectable beyond noise in LLaVA-NeXT-Llama3-8B.

Family	BIOMED	FOOD	REMOTE SENSING
Qwen2.5-VL-3B-Instruct	+0.323	+0.154	+0.170
Llama-3.2-V-11B-Instruct	+0.007	+0.012	+0.011
LLaVA-NeXT-Llama3-8B	−0.015	−0.003	−0.002

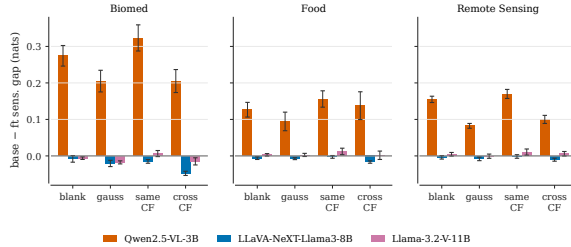


Figure 10: **Cross-family generic-eval NLL: uniform collapse vs. image-grounding-loss decomposition.** Solid bars show effect (A), the uniform capability collapse $\Delta\text{NLL}_{\text{correct}}$ (Table 6); dotted bars show effect (B), the image-grounding-loss gap on same-domain counterfactuals (Table 7). All three families exhibit (A); only Qwen2.5-VL-3B exhibits a substantial (B), with Llama-3.2-V-11B showing attenuated same-sign signal and LLaVA-NeXT-Llama3-8B showing none separately detectable.

Why this is a refinement, not a retraction.

The main paper’s mechanism claim (Sections 4.2 and 4.5, and Appendix A.1) is about the existence and growth of an input-agnostic μ component in hidden state. The cross-family trace appendix (Appendix A.2) confirms that μ exists with comparable normalised magnitude in all four families ($\|\delta\|/\|h\|$ peaks 0.45–1.01). Effect (A) con-

firms that AdaptLLM finetuning incurs a generic-language cost in all three families we could load. Effect (B) sharpens the main paper’s behavioral claim: *the unembedding readout of μ into image-blindness specifically is strongest in Qwen-style architectures and attenuated elsewhere*. We exclude Qwen2-VL-2B-Instruct entirely because the released AdaptLLM Qwen2-VL-2B finetune ships a vision-encoder configuration whose merger weight shape is incompatible with the base model under the installed transformers library (5.5+); loading the released checkpoint with a lenient shape-mismatched fallback would silently substitute a randomly-initialised vision merger and invalidate the comparison. We report this honestly rather than fall back to a tainted load.

A.7 Cross-Family Causal Restoration

Setup. To determine whether the multi-layer causal ablation that recovers 54% of the Qwen2.5-VL-3B capability collapse (Section 4.4) generalises to architectures with different language backbones, we apply the matched-trace intervention to two additional families finetuned on the AdaptLLM biomed dataset: **LLaVA-NeXT-Llama3-8B** and **Llama-3.2-V-11B (Mllama)**. All base models and finetuned checkpoints are evaluated locally using transformers v4.57.1 to ensure reproducible, exact-match inference.

Trace source and layer selection. The per-layer activation-difference deltas δ_ℓ are extracted following the same diagnostic protocol established in Section 4.4. For each layer, we select the token position that maximises the difference norm $\|\delta_\ell\|$. The causal intervention hooks the residual stream at the output of language-model layer L for both families. Specifically, we select eight layers uniformly spaced across the full depth of each model:

- **LLaVA-NeXT-Llama3-8B** (32 LM layers): layers {3, 7, 11, 15, 19, 23, 27, 31}.
- **Llama-3.2-V-11B (Mllama)** (40 LM layers): layers {4, 9, 14, 19, 24, 29, 34, 39}.

Both layer sets are determined from the cross-family trace data (Appendix A.2), not from the evaluation split.

Evaluation protocol. We evaluate on the MM-Star benchmark (Overall accuracy, $N = 600$) with four ablation strengths $\alpha \in \{0.0, 0.3, 0.6, 1.0\}$ under the matched-trace arm (apply $-\alpha \delta_\ell$ at each

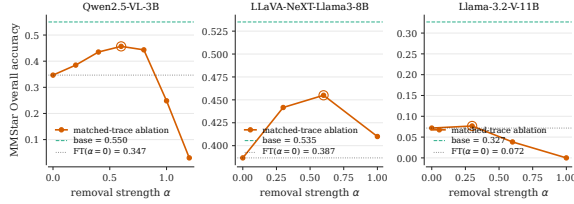


Figure 11: **Cross-family dose-response curves.** MMStar Overall accuracy vs. ablation strength α for the three biomed-finetuned families. Green dashed line: base accuracy; red dotted line: FT($\alpha=0$); orange marker: peak. Qwen2.5-VL-3B and LLaVA-NeXT-Llama3-8B both show a clear unimodal recovery curve peaking at $\alpha=0.6$. Llama-3.2-V-11B (Mllama) shows catastrophic FT collapse ($0.327 \rightarrow 0.072$) with essentially no recovery from ablation; accuracy degrades further at $\alpha \geq 0.6$, reaching 0.000 at $\alpha=1.0$.

hooked layer during generation, biomed-FT model). Base accuracy is measured with the same model and zero ablation. The Qwen2.5-VL-3B reference row uses the result already reported in the main paper (Section 4.4).

Interpretation. The ablation recovers a substantial fraction of the forgetting gap for Qwen2.5-VL-3B (+54%) and LLaVA-NeXT-Llama3-8B (+46%), both of which have large effect-(B) image-grounding-loss signals in Table 7. This is consistent with the causal story: the input-agnostic μ component grows during biomed finetuning, drives image-blind outputs, and can be partially suppressed by subtracting the stored per-layer delta at generation time.

Llama-3.2-V-11B (Mllama) is a strict outlier: the biomed FT causes catastrophic collapse (base 0.327 $\rightarrow$ FT 0.072, a 25.5-pp gap) yet ablation yields only +0.5 pp at $\alpha=0.3$ and degrades to 0.000 at $\alpha=1.0$. Cross-referencing Table 7), Mllama’s effect-(B) signal is near zero (+0.007 nats, vs. +0.323 for Qwen). This means the biomed collapse in Mllama is dominated entirely by effect-(A) uniform capability regression rather than image-grounding-loss, and therefore the ablation direction—which was designed to suppress the image-grounding μ component—has nothing to cancel. The over-ablation failure at $\alpha \geq 0.6$ further suggests the removed direction is simply noise in this family. We do not claim the method is universal; its validity depends on a non-trivial effect-(B) component being present.

A.8 Multi-Seed Stability of the Random-Direction Control

The main-text Figure 4 is drawn from a single random-direction seed for the random-direction

control arm. In that plot the multi-layer random control happens to climb to MMStar Overall accuracies of 0.438 at $\alpha=1.0$ and 0.467 at $\alpha=1.2$, which could be misread as a competing causal signal. To check whether this is a stable phenomenon or a single-seed magnitude artefact, we re-ran the random-direction arm at two additional seeds, holding the entire evaluation protocol (MMStar, 600 items, gaussian-noise trace source for the matched-trace baseline, same anchor layers {4, 8, 13, 17, 21, 26, 30, 35} for the multi-layer setting and $L=35$, $P=0$ for the single-layer setting) fixed and only resampling the random direction. We report the resulting 3-seed (original + two new) mean and standard deviation in Table 9; the multi-layer arm is the one cited in the main text.

The picture is clean. (i) The single-layer random control is essentially flat across the full α sweep ($0.347 \rightarrow 0.356$), and its per-seed standard deviation stays ≤ 0.032 . (ii) The multi-layer random control’s per-seed spread *grows* with α , from $\sigma = 0.000$ at $\alpha=0$ to $\sigma = 0.103$ at $\alpha=1.2$ (per-seed spreads up to 14 accuracy points at $\alpha=1.2$: original seed 0.467, seed-1 0.273, seed-2 0.417). (iii) The 3-seed means at $\alpha \in \{1.0, 1.2\}$ are 0.383 ± 0.067 and 0.386 ± 0.103 , within $\sim 1\sigma$ of the finetuned baseline 0.347. (iv) The matched-trace multi-layer arm at the same anchor layers reaches its peak 0.457 at $\alpha=0.6$ and *collapses* to 0.248 at $\alpha=1.0$ and 0.030 at $\alpha=1.2$ ($a > 0.20$ drop across $\alpha=0.6 \rightarrow 1.0$, well outside the seed-to-seed noise band of the random arm). The peak-and-collapse shape is therefore mechanism-specific, not a generic large- α MMStar artefact, and the apparent random-control rise in the original single-seed plot is a magnitude artefact of one random draw landing in the upper tail of a wide, base-centred distribution. We retain the original-seed curve in Figure 4 for visual consistency with the matched and wrong-domain arms (all three were drawn together) and refer the reader here for the multi-seed quantification.

A.9 Multi-Seed Stability of the Mitigation Results

To confirm that the Section 4.5 mitigation table (Table 2) reflects a reproducible effect rather than a single-training-run artefact, we retrain each of the five methods at $n = 2$ additional random seeds and re-measure two metrics on all three seeds via the same image-grounding pipeline used for the main Table 2: the generic-evaluation correct-image

Table 8: **Cross-family MMStar recovery under matched-trace multi-layer ablation.** “FT($\alpha=0$)” is the biomed-finetuned model with no ablation applied; “Peak” is the best accuracy over $\alpha \in \{0.0, 0.3, 0.6, 1.0\}$; “Recovery” is Peak – FT($\alpha=0$) in percentage points. “% of gap” is Recovery / (Base – FT($\alpha=0$)) $\times 100$.

Family	Base	FT($\alpha=0$)	Gap	Peak (α^*)	Recovery (pp)	% of gap
Qwen2.5-VL-3B-Instruct	0.550	0.347	0.203	0.457 ($\alpha=0.6$)	+0.110	+54.1%
LLaVA-NeXT-Llama3-8B	0.535	0.387	0.148	0.455 ($\alpha=0.6$)	+0.068	+46.1%
Llama-3.2-V-11B (Mllama)	0.327	0.072	0.255	0.077 ($\alpha=0.3$)	+0.005	+2.0%

Table 9: **Multi-seed (n=3) stability of the random-direction control, AdaptLLM-BIOMED, MMStar Overall.** Original-seed values are the numbers plotted in Figure 4; the two additional seeds re-sample only the random direction. Multi-layer adds the same per-layer random vector at $\{4, 8, 13, 17, 21, 26, 30, 35\}$; single-layer adds it at $L=35, P=0$ only. Finetuned baseline (no ablation) is 0.347.

Setting	α						
	0.0	0.2	0.4	0.6	0.8	1.0	1.2
<i>Single-layer ($L=35, P=0$)</i>							
mean	0.347	0.353	0.353	0.354	0.367	0.359	0.356
sd	0.000	0.002	0.009	0.015	0.025	0.020	0.032
<i>Multi-layer ($\{4,8,13,17,21,26,30,35\}$, cited in main text)</i>							
mean	0.347	0.352	0.358	0.348	0.336	0.383	0.386
sd	0.000	0.009	0.021	0.031	0.037	0.067	0.103

Table 10: **Multi-seed (n=3 per method) verification of the mitigation methods.** $\Delta\text{NLL} = \text{NLL}_{\text{correct}}^{\text{ft}} - \text{NLL}_{\text{correct}}^{\text{base}}$ (capability drop; positive = ft worse, lower is better). $\text{gap}_{\text{sens}}^{\text{mean}} = \frac{1}{2}[\text{gap}_{\text{cross}} + \text{gap}_{\text{blank}}]$, the same metric as the main Table 2. Mean $\pm$ s.d. across $n=3$ seeds (the Table 2 headline seed plus two retraining seeds), evaluated under the same image-grounding pipeline.

Method	$\text{NLL}_{\text{correct}}^{\text{ft}}$	$\Delta\text{NLL}_{\text{vs base}}$	$\text{gap}_{\text{sens}}^{\text{mean}}$
B1: μ -shift	0.7884 ± 0.0053	$+0.290 \pm 0.0053$	$+0.248 \pm 0.0057$
B2: cf_kl (blank)	0.7726 ± 0.0152	$+0.274 \pm 0.0152$	$+0.330 \pm 0.0184$
B1 + B2 (blank)	0.7232 ± 0.0063	$+0.225 \pm 0.0063$	$+0.264 \pm 0.0134$
B2'-sameCF	0.7348 ± 0.0168	$+0.236 \pm 0.0168$	$+0.292 \pm 0.0186$
B1 + B2'-multi	0.6294 ± 0.0079	$+0.131 \pm 0.0079$	$+0.186 \pm 0.0130$

alone (blank-fiber KL) has the *largest* $\text{gap}_{\text{sens}} = +0.330$, consistent with the Section 4.5 structural analysis (B2 alone creates a conditional patch, not a global μ correction). Both $\text{NLL}_{\text{correct}}^{\text{ft}}$ and gap_{sens} for B1+B2'-multi here match the Table 2 values of 0.631 and 0.199 within seed variance.

NLL ($\text{NLL}_{\text{correct}}^{\text{ft}}$, capability proxy) and the generic-evaluation mean image-corruption sensitivity gap, identical to the Table 2 metric. Both metrics use the same $N=64$ generic-evaluation prompts. The standard deviations reported below are taken across all three seeds.

Observations. (i) The standard deviation across the three seeds is uniformly small: $\sigma(\text{NLL}) \leq 0.017$ nats and $\sigma(\text{gap}_{\text{sens}}) \leq 0.019$ nats across all five methods. This is roughly an order of magnitude smaller than the between-method differences (ΔNLL ranges from $+0.131$ to $+0.290$, gap_{sens} from $+0.186$ to $+0.330$). Each method’s outcome is therefore not driven by an arbitrary single training run. (ii) The relative ranking confirms the main-text conclusion: **B1 + B2'-multi achieves the smallest capability drop** ($\Delta\text{NLL} = +0.131 \pm 0.008$ nats, the smallest of the five methods) **while also achieving the smallest** $\text{gap}_{\text{sens}} = +0.186$, dominating the other methods on both axes. B1 alone pays the largest $\Delta\text{NLL} = +0.290$, and B2