Text Capability Loss in Vision-Language Adaptation: An Attention-Sink Diagnosis

Minsik Choi[1,*] **Geewook Kim**[2,3,*] **Young Geun Kim**[1,†]
[1]Korea University [2]NAVER Cloud AI [3]KAIST AI
[*]Equal contribution. [†]Corresponding author.

Abstract

Fine-tuning a pretrained LLM into a vision-language model (VLM) can erode the backbone's text capability, with the damage concentrated on tasks that require following exact output rules, such as instruction following, chain-of-thought reasoning graded on a strictly parsed final answer, and similar evaluations with strict graders. We trace this gap to *attention-sink corruption*: VL fine-tuning perturbs the early sink position that anchors a large fraction of attention probability, and how well the base LLM preserves its sink tracks how much of the affected capability survives adaptation. Building on this view, we introduce *Sink Strength*, a single scalar computed on the base LLM in a few seconds on a single GPU that predicts post-VL degradation without any VL training. It consistently tracks relative degradation across the six VLM–LLM pairs and multiple format-sensitive tasks. Complementing this diagnostic, we find that post-pretraining QK-RMSNORM injection fails to reproduce the protection of native QK-RMSNORM, while several off-the-shelf weight-merging settings fail to recover the lost capability after VL training. These negative results underscore the value of screening backbones with Sink Strength before VL training and narrow the intervention space toward head-selective training-time protection.

1 Introduction

Vision-language models (VLMs) are commonly built by attaching a vision encoder and projector to a pretrained large language model (LLM) and jointly fine-tuning on multimodal data. We find that this adaptation can erode the LLM's text capability, with the damage concentrated on tasks we call *format-sensitive*: those that require following exact output rules, such as instruction following or chain-of-thought reasoning whose final answer is parsed under a strict rule. Across the released VLM–LLM pairs we study, the text-capability gap reaches double digits on multiple such tasks. A text-only further-training counterpart shows substantially smaller degradation, while a separately matched VL-versus-text training trajectory isolates the modality difference (Sections 3.1 and 5). Existing remedies each help in part but at a cost, and none explains *why* the loss happens. These include re-blending text-only data into the multimodal supervised fine-tuning (SFT) mix (Lin et al., 2024; Tong et al., 2024; Tu et al., 2025), freezing the LLM (Zhu et al., 2024), low-rank adapters (LoRA) (Liu et al., 2023; Hu et al., 2022), and post-hoc neuron re-merging (Yu and Ananiadou, 2025).

We trace this loss to *attention-sink corruption*. A modern LLM concentrates a large fraction of its attention onto a small number of early positions, anchored by a few channels with large normalization scales (Xiao et al., 2024; Barbero et al., 2025; Sun et al., 2024). This sink absorbs diffuse no-op attention mass that would otherwise spread across non-informative positions (Xiao et al., 2024; Barbero et al., 2025), stabilizing per-position attention to the surface tokens that format-sensitive evaluators grade (Zhang et al., 2025; Venkateswaran and Contractor, 2026). We show that VL fine-tuning perturbs the projections that read from these channels, collapsing the per-head sink (Figure 1). How well the base LLM's sink survives this perturbation predicts how much post-VL damage we see on format-sensitive tasks. Base LLMs whose query and key projections are less sensitive to input magnitude exhibit greater sink preservation under VL adaptation (Section 4). We identify per-head QK-RMSNORM as the clearest example of this in our sample. However, we also observe that base LLMs can already inherit attention-sink corruption from pre-training, depending on architectural choices. Molmo2-O, despite having the QK-RMSNORM module, applies it layerwise, leaving the per-head sink already weak in the base.

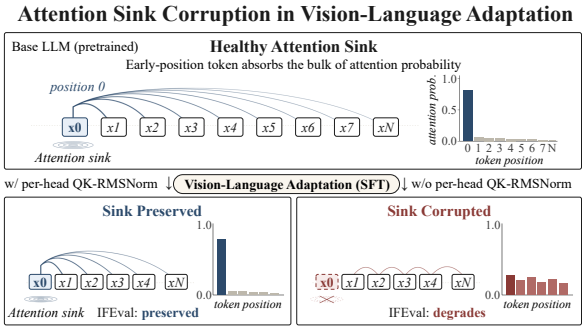

Figure 1: Attention-sink corruption in VL fine-tuning: the base LLM concentrates attention on an early sink position; whether this concentration survives VL fine-tuning varies sharply across language backbones.

This observation implies that predicting post-VL damage requires accounting for both base-LLM-inherited and VL-induced corruption. Consistent with this account, backbones with a preserved sink and backbones with a collapsed sink form non-overlapping groups on our headline task.

Building on this account, we introduce *Sink Strength* S (Section 3), a single scalar computed entirely on the base LLM using a small number of inference-only forward passes and no VL training. S quantifies the head-level sink concentration, ranks backbones by their post-VL damage (Spearman $\rho = 0.97$), and the ranking remains strongly correlated across multiple format-sensitive tasks. Because S is measured before VL training, it exposes base-side sink weakness independently of the subsequent VL update; Section 4 then combines this quantity with the post-VL gap perturbation to separate base-side vulnerability from update-induced corruption.

Given a high-risk backbone, can simple interventions prevent or recover the damage? We test two complementary negative controls (Section 5): post-pretraining injection of the QK-RMSNORM module into a matched base LLM followed by VL training, and several post-VL weight-merging methods spanning linear, spectral, and sparsified families (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024; Gargiulo et al., 2025; Lee et al., 2025b; Wortsman et al., 2022). The former fails to reproduce the protection associated with native QK-RMSNORM, while the latter fails to recover the lost capability under the settings we test, narrowing the intervention space toward head-selective training-time protection.

Our contributions are: (i) **Sink Strength** S, a base-LLM scalar predicting post-VL damage on format-sensitive tasks from inference-only forward passes (rank correlations remain strong across our six VLM–LLM pairs and four tasks, $\rho \geq 0.88$), providing a pre-VL screen; (ii) a **sink-anchored mechanism** (Theorem 1) showing how per-head QK-RMSNORM on the query and key projections removes explicit raw-input-magnitude sensitivity from the sink-gap perturbation bound, supported by margin calibration across the five headline pairs and two layerwise-QK-RMSNORM controls; (iii) **confound and negative controls**: text-only versus VL controls supporting modality-specific degradation, together with two complementary negative controls—post-pretraining QK-RMSNORM injection and post-VL weight-space merging.

2 Background & Related Work

VL fine-tuning and text-side erosion. Standard VLM recipes (Liu et al., 2023, 2024) update the LLM decoder weights during adaptation, typically over multiple stages of alignment and instruction tuning. A recurrent observation across this paradigm is that the adapted decoder loses some of the underlying LLM's text-side capability (Zhang et al., 2024; Zhai et al., 2024; Lee et al., 2025a; Ratzlaff et al., 2025; Srivastava et al., 2026), with instruction-following and other format-sensitive behaviors among the hardest hit. Existing remedies are data-side (text-only re-blending (Lin et al., 2024; Tong et al., 2024; Tu et al., 2025)), training-side (frozen LLM (Zhu et al., 2024) or low-rank adapter (Hu et al., 2022; Liu et al., 2023, 2024)), or post-hoc (neuron-level re-merging (Yu and Ananiadou, 2025)), each closing part of the gap. Our primary diagnostic analysis leaves the released VLM training recipes untouched and instead asks which base-LLM property predicts how much text capability survives adaptation.

Attention sinks and mechanistic position. The attention-sink phenomenon (Xiao et al., 2024) has been traced to massive activations (Sun et al., 2024), given an over-mixing rationale (Barbero et al., 2025), and its emergence conditions characterized (Gu et al., 2025), with the vision analogue motivating register tokens (Darcet et al., 2024). These analyze the sink at inference time on a trained network; we instead study what happens to it *during* fine-tuning. VL fine-tuning is known to shift attention patterns within the adapted decoder (Zhang et al., 2024), and we operationalize this shift as corruption of the early-position sink that anchors

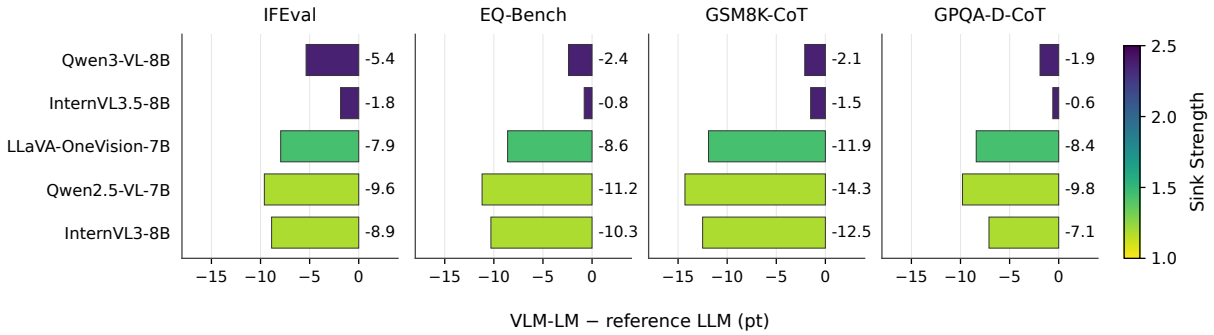

Figure 2: Text-capability gap across our headline VLM–LLM pairs on instruction-following and three further format-sensitive tasks. Color: Sink Strength S measured on the reference LLM (Section 3).

text-side stability. We introduce Sink Strength (Section 3) as a quantitative pre-VL diagnostic of the post-fine-tune fate of the sink, enabling ranking and screening of candidate base LLMs by their predicted post-VL damage before any VL training is committed. QK-RMSNORM extends an earlier query-key normalization idea (Henry et al., 2020) (originally an L2-norm variant) with RMSNorm in place of L2-norm, motivated by entropy-collapse stability arguments (Zhai et al., 2023). QK-RMSNORM adoption varies across recent base LLMs (present in Qwen3 (Yang et al., 2025), absent in Qwen2.5 (Qwen Team, 2024), Llama-3 (Grattafiori et al., 2024), and DeepSeek-V3 (DeepSeek-AI, 2024)), an architectural axis examined throughout our analysis. Our bound and saturation lemma are in the spirit of small-circuit attention analyses (Olsson et al., 2022; Geva et al., 2021) and recent fine-tuning circuit-stability work (Wang et al., 2025b).

3 Sink Strength: A Pre-VL Diagnostic

3.1 Problem Setup

An LM-only further-training counterpart shows substantially smaller text-capability degradation than the corresponding VL endpoint across the four format-sensitive tasks; Section 5 provides a separately matched VL-versus-text training trajectory from a shared base LLM.

Setup. The five headline pairs come from three model families: Qwen, InternVL, and LLaVA-OneVision. Full per-pair metadata is in Table 1.[1] All released VLMs in the headline panel update the language backbone rather than relying on LoRA-only adaptation, removing the adapter-merge confound. For each released VLM, we extract its language backbone as a standalone causal LM (the VLM-LM hereafter; Appendix B.2, with the round-trip preserving logits on a held-out probe) and score it on a nine-task text suite (Appendix B.1) against the corresponding text-only reference LLM, which serves as a comparison point rather than being assumed to immediately precede every stage of multimodal training.

Pair	Reference LLM	Post-training	QK
Qwen3-VL	Qwen3-8B	SFT+RL	**per-head**
InternVL3.5	Qwen3-8B	SFT+RL (CascadeRL)	**per-head**
Qwen2.5-VL	Qwen2.5-7B-Instruct	SFT+DPO	none
InternVL3	Qwen2.5-7B-Instruct	SFT+MPO	none
LLaVA-OV	Qwen2-7B-Instruct	SFT	none
Non-headline control (Appendix D.2)			
Molmo2-O	Olmo-3-7B-Instruct	SFT (PixMo)	*layerwise*

Table 1: Five headline VLM–LLM pairs and the Molmo2-O control. "QK" reports the per-head / layerwise / none QK-RMSNORM configuration. Per-head QK-RMSNORM normalizes each attention head by its own RMS, while the layerwise variant applies a single RMS across all concatenated heads.

Format-sensitive tasks. We call IFEval, EQ-Bench, GSM8K-CoT, and GPQA-Diamond-CoT *format-sensitive* because each grades a strict, parseable surface gate: per-prompt format compliance for IFEval (Zhou et al., 2023) (our headline metric), score-block parsing for EQ-Bench, and CoT answer-token extraction for GSM8K-CoT and GPQA-Diamond-CoT. Because these gates are parsed rather than scored semantically, small attention-pattern drift can disproportionately flip the outcome. Concretely, the regressing prompts we inspect often fail parseable rules: length constraints (a 1200-word essay truncated), exact-phrase requirements ("My answer is maybe"

[1] Hugging Face URLs and paper references for every model and evaluation dataset are collected in Appendix B.4.

Pair	QK	Both ✓	Lost	Gained	Both ×	Net (%)
Qwen3-VL	per-head	394	56	27	64	**5.4**
InternVL3.5	per-head	407	43	33	58	**1.8**
Qwen2.5-VL	none	299	**91**	39	112	9.6
InternVL3	none	294	**96**	48	103	8.9
LLaVA-OV	none	195	**83**	40	223	7.9
Molmo2-O	layerwise	317	**126**	25	73	**18.7**

Table 2: Per-prompt regression breakdown on the 541-prompt IFEval set, partitioning prompts by reference-LLM and VLM-LM pass/fail. Net = (Lost − Gained)/541.

reduced to "My answer"), single-language constraints (an Italian-only haiku gaining an English gloss), and format markers ("SECTION 1" rendered as "SECTION"). A per-category breakdown across pairs is in Appendix C.2.

Per-pair text-capability gap. Figure 2 (IFEval row) shows a clear VLM–LLM text-capability gap across our five headline pairs: two pairs show gaps of at most 5.4 pt (InternVL3.5 at −1.8, Qwen3-VL at −5.4), while three show gaps of at least 7.9 pt (LLaVA-OV −7.9, InternVL3 −8.9, Qwen2.5-VL −9.6). The two groups do not overlap. The non-headline pair (Molmo2-O on Olmo-3) shows a −18.7-pt gap and is analyzed separately as a control (Section 4).

Per-prompt regression. On the 541 IFEval validation prompts each backbone evaluates, we tag every prompt as *both_pass* (LLM and VLM-LM both score ≥ 0.5), *regression* (LLM pass, VLM-LM fail), *recovered* (LLM fail, VLM-LM pass), or *both_fail* (Table 2). The net regression rate reproduces the IFEval delta pair-by-pair. The regressions are spread across 43–126 distinct prompts per pair, so the behavioral split reflects broad instruction-following degradation rather than a few outlier prompts.

Modality attribution of the loss. A natural question is whether the observed gap reflects further optimization in general rather than VL training specifically. We first compare two Qwen2.5-family endpoints against the same Qwen2.5-7B-Instruct text-only reference: the text-only Qwen2.5-7B-Instruct-1M variant and Qwen2.5-VL-7B-Instruct. Under the same instruct protocol (0-shot with the chat template applied, Appendix B.1), the text-only variant matches the reference on IFEval prompt-strict while the VL endpoint shows a 9.6-pt gap; across the four format-sensitive tasks the text-only variant does not regress on IFEval or GSM8K and loses at most 0.42× the VL gap on EQ-Bench and GPQA (Table 3; details in Appendix F.1). The same comparison also indicates why the VL endpoint perturbs the sink more strongly. Across every layer of the late-layer window, its displacement from the reference is larger in operator norm than that of the text-only endpoint, with a median ratio of 4.44× on the query projection and 3.65× on the key projection. The sink-to-other input contrast is also ≈ 1.12× larger once vision tokens are present. Together, these observations are consistent with stronger sink perturbation on the VL side. This endpoint contrast supports modality as an important source of the gap rather than further optimization alone, with a separately matched training-time control provided in Section 5. This frames the predictor question: which candidate LLM backbones are most vulnerable to text-side degradation under VL adaptation?

Bench	Text-only Δ	VL Δ
IFEval prompt-strict	0.00	**−9.6**
GSM8K-CoT (flex)	+5.46	**−14.3**
GPQA-Diamond-CoT (flex)	−4.04	**−9.8**
EQ-Bench v2.1	−2.24	**−11.2**

Table 3: Endpoint modality comparison against the same Qwen2.5-7B-Instruct text-only reference: the text-only endpoint is Qwen2.5-7B-Instruct-1M, while the VL endpoint is Qwen2.5-VL-7B-Instruct.

3.2 Sink Strength

Sink Strength S, measured on each text-only reference LLM, captures the per-pair capability-gap split, whose clustering further aligns with per-head QK-RMSNORM in the reference LLM.

Definition. We capture per-head sink concentration on the base LLM as a stand-alone metric, the *Sink Strength*:

$$S \;:=\; \mathrm{median}_{(\ell,h,q)}\ \log \frac{a_{p_{\mathrm{sink}}}^{(\ell,h,q)}}{1 - a_{p_{\mathrm{sink}}}^{(\ell,h,q)}},$$

where $a_{p_{\mathrm{sink}}}^{(\ell,h,q)}$ is the attention probability at the per-head argmax key position p_{sink} at layer ℓ, head h, query position q, and the median is taken over the last ∼10 layers (design ablations in Appendix E.4). S is computed entirely on the base LLM (no VLM forward pass, no VL training) with 15 inference-only forward passes on calibration prompts disjoint from the IFEval test set. For a 7B backbone, this

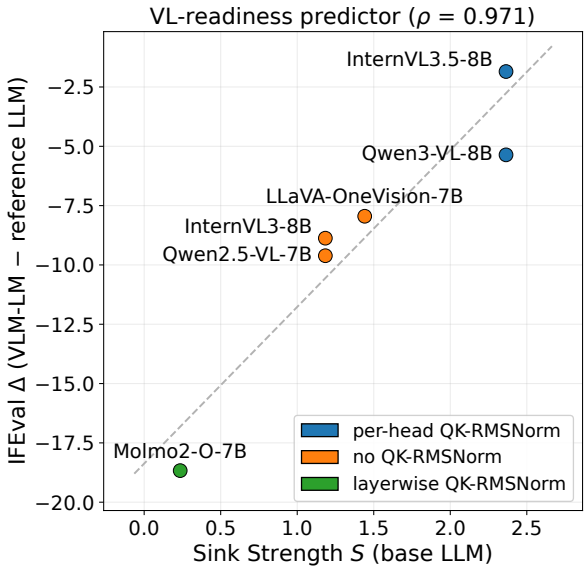

Figure 3: Sink Strength S on the reference LLM versus the VLM–LLM IFEval gap. Color denotes the QK-RMSNORM configuration. The dashed line is a global linear fit for visualization; reported prediction errors use leave-one-pair-out fitting.

takes ~ 8 seconds on a single A6000 GPU. Intuitively, S quantifies how concentrated the base LLM's attention is on its sink position: high S means a strong, well-formed sink, while $S \leq 0$ means the sink does not dominate the aggregate non-sink attention mass. Section 4 formalizes why this scalar captures vulnerability to VL-induced sink corruption via a sink-anchored perturbation bound.

Analysis. Color in Figure 2 indicates per-pair S measured on the reference LLM, and the same backbones that lose more on IFEval have smaller S. S separates the six pairs into non-overlapping ranges along both axes: the two lower-loss pairs share $S = +2.36$ (sharing the Qwen3-8B reference), the three higher-loss pairs fall in $S \in [+1.18, +1.44]$, and the Molmo2-O boundary case sits at $S = +0.24$ (Figure 3). The ordering induced by S strongly tracks the IFEval-gap ordering across the six pairs.

From sink collapse to format-sensitive loss. Low S in a candidate base LLM indicates a sink that fails to dominate non-informative positions: a sink collapse waiting to surface under VL fine-tuning. Format-sensitive tasks such as instruction following, score-block extraction, and CoT answer parsing grade strict, parseable rules over surface tokens that require stable per-position attention (Zhang et al., 2025; Venkateswaran and Contractor, 2026); the early-position sink anchors this stability (Xiao et al., 2024; Barbero et al., 2025), so sink collapse breaks surface-token tracking and flips format-checker outcomes. Knowledge tasks, which rely on facts encoded in MLP weights (Geva et al., 2021), do not depend on the same attention-pattern stability and are correspondingly preserved (Appendix C.1). The same holds on a broader capability panel spanning coding, long-context retrieval, and advanced reasoning, where the mean degradation remains substantially smaller than on the format-sensitive tasks (Appendix C.4).

Architectural correlate. The S clusters align with per-head QK-RMSNORM presence in the reference LLM, with finer resolution: per-head QK-RMSNORM pairs occupy the high-S range, no-QK-RMSNORM pairs the mid-S range, and Molmo2-O is a boundary case where QK-RMSNORM is present but applied layerwise. A coarse "has any QK-RMSNORM module" classifier would flag Molmo2-O as safe; S correctly identifies it as the highest-risk backbone among the six pairs, matching the largest measured drop in that set (-18.7 pt on IFEval). Because S is measured on the base LLM alone, this is identifiable *before* any VL training is committed, framing per-head sink preservation as a *pre-training architectural-design concern* (mechanism: Section 4.3). Within pairs sharing the same reference LLM, and hence the same S, the drop still varies by up to 3.6 pt (Qwen3-VL vs. InternVL3.5), which S cannot resolve and we attribute to VL recipe variation.

Empirical validation. We use S in two modes: rank ordering across backbones requires no outcome calibration, while magnitude prediction adds a single 1-D linear fit calibrated on known post-VL outcomes. Across the six pairs of Table 1, S strongly tracks the VLM–LLM IFEval gap (Spearman $\rho = 0.97$); under leave-one-pair-out validation, the average held-out prediction error is 2.54 pt on a 17-pt outcome range (best -1.8 pt to worst -18.7 pt), against 4.40 pt for a constant-mean baseline (per-pair detail in Appendix E.1). The validation holds out one pair, fits the linear predictor on the remaining five, predicts the held-out pair, and averages the error over all six choices. With six datapoints this is a sanity check on the linear relation rather than prospective evidence on unseen backbones. The one large headline residual

	IFEval	EQ-Bench	GSM8K	GPQA-D
Spearman ρ, n=6	0.971	0.971	0.971	0.883
Error, 6 pairs (pt)	2.54	19.09	7.91	1.96
Error, 5 headline pairs (pt)	1.63	0.95	1.03	1.69

Table 4: Sink Strength S as a cross-task predictor on the four format-sensitive tasks (held-out 1-D linear regression). The 5-headline row excludes Molmo2-O, the boundary case where the linear extrapolation breaks down (task-specific Molmo collapse, e.g., EQ-Bench -64.6); rank correlation holds on all six pairs.

is Qwen3-VL (4.5 pt held-out): it shares $S = 2.36$ with InternVL3.5 but the two differ by 3.6 pt in the observed gap, and no single-feature fit can separate backbones with identical S, which sets the resolution floor of the predictor. Below the $S \approx 1$ regime, the linear extrapolation underestimates the observed gap magnitude (-18.7 pt actual vs. -13.74 pt leave-one-out). While Theorem 1 controls the full theoretical margin $G_{\text{base}} - B_{\text{gap}}$ (both defined in Section 4), in our sample S tracks the IFEval gap more strongly than the full margin ($\rho = 0.97$ vs. 0.89); we therefore use S as a leading-order, low-cost base-side proxy for the theoretical margin.

3.3 Generalization and Practical Use

Generalization across format-sensitive tasks. The same S, measured once on the reference LLM, also ranks the backbones on each of the four format-sensitive tasks (Table 4). Across six pairs, Spearman $\rho \in [0.88, 0.97]$ on every task. On the five headline pairs (excluding Molmo2-O), the average held-out error stays under 2 pt on each task. S generalizes as a rank predictor across tasks without re-measurement; per-task magnitude prediction additionally requires a per-task linear fit but reuses the same reference-LLM S.

Predictor extension. We extend the predictor along two complementary axes. At fixed IFEval, a 17-pair panel spanning dense and MoE architectures, four reference-LLM families plus OLMoE, and 1.5B–32B scale achieves Spearman $\rho = 0.88$ (Appendix E.2). On a 9-pair subset evaluated across all four format-sensitive tasks, the rank correlation remains $\rho \geq 0.82$ (Appendix E.3).

Practical use. Given a candidate base LLM, S can be measured with 15 calibration prompts in seconds on a single GPU and used to rank expected post-VL IFEval damage. In our panels, every pair at $S \geq 2$ stays within 5.4 pt on IFEval, whereas pairs near $S \approx 1.2$ span -8.9 to -17.9 pt.

From diagnosis to mechanism. S is a VL-conditional, base-side diagnostic; for recipe-sensitive post-hoc analysis, the full margin $G_{\text{base}} - B_{\text{gap}}$ provides a refinement (Section 4). Since S is the median of G_{base}, it captures the base-side component of this margin, explaining why a pre-VL quantity carries information about post-VL outcomes.

4 Mechanism

A sink survives as long as its attention logit stays clearly ahead of the logits at the other key positions. This section makes that lead quantitative and bounds how much of it a fine-tuning update of a given size can close.

4.1 Setup and Notation

Setup. Fix a single attention head (hidden size d, head dim d_h; the layer has H heads), a single layer, and a single query token, with base-LLM hidden states $\{\xi_q, \xi_p\}_p$ of the surrounding tokens, where $\xi_q = \gamma_{\text{in}} \odot \text{RMSNorm}(x)$ for input-layernorm scale γ_{in} and RMSNorm denotes root-mean-square normalization. The attention logit at key position p is

$$l_p = \frac{Q \cdot K_p}{\sqrt{d_h}}, \qquad Q = W_q \xi_q, \qquad K_p = W_k \xi_p. \tag{1}$$

Under QK-RMSNORM each projection is divided by its own RMS before the scale is applied,

$$Q = \gamma_q \odot \frac{Q_{\text{pre}}}{\text{RMS}(Q_{\text{pre}})}, \tag{2}$$

where $Q_{\text{pre}} = W_q \xi_q$, and analogously for K_p. Here $\|\gamma_*\|_\infty := \max_d |\gamma_*[d]|$ denotes the channel-wise ℓ_∞ norm of a QK-RMSNORM scale vector. Let $l'_p = l_p + \delta_p$ be the post-perturbation logit, and let p_{sink} be the base LLM's per-head sink position (argmax key on the unperturbed forward pass). Two quantities anchored at p_{sink} drive the bound,

$$G_{\text{base}} := \log \frac{a_{p_{\text{sink}}}}{1 - a_{p_{\text{sink}}}}, \qquad B_{\text{gap}} := \sup_{p \neq p_{\text{sink}}} |\delta_p - \delta_{p_{\text{sink}}}|, \tag{3}$$

the aggregate sink gap on the unperturbed logits and the sink-anchored gap perturbation over the fixed key positions of the calibration prompt.

Perturbation model.

Assumption 1 (Perturbation model)**.** At the layer of interest, fine-tuning perturbs only the query and key projections, by ΔW_* with $\|\Delta W_*\|_{\mathrm{op}} \le \varepsilon$, and leaves the QK-RMSNORM scales γ_q, γ_k and the input-layernorm scale γ_{in} fixed. We hold the base hidden states $\{\xi_q, \xi_p\}$ at their LLM baseline values, so the bound isolates the effect of the projection drift at that layer.

Both parts have a plain reading. The operator-norm cap ε limits how far fine-tuning can move either projection. Holding the γ's fixed is a close approximation for the Qwen3-VL pair, where the relative-norm drift $\|\Delta\gamma\|/\|\gamma\|$ is at most 2.9% across the 36 q-norm and k-norm layers.

Notation summary. We use four sink-related quantities: G_{base} (per-head log-odds of sink attention, base LLM), B_{gap} (post-VL gap perturbation, Theorem 1), the margin $G_{\mathrm{base}} - B_{\mathrm{gap}}$ (controls sink survival, Lemma 1), and Sink Strength S (Section 3, median of G_{base} over (ℓ, h, q)).

4.2 Perturbation Bound and Margin Saturation

We bound the post-VL logit-gap perturbation B_{gap} under projection drift, and convert the resulting margin $G_{\mathrm{base}} - B_{\mathrm{gap}}$ into an upper bound on post-VL non-sink mass.

The proof additionally requires a mild non-degeneracy condition on the projection-RMS denominators, which we state formally below.

Assumption 2 (Projection-RMS non-degeneracy)**.** Let $\mathcal{S}_q$ be the finite set of query-side inputs ξ_q over the calibration prompts at the layer and head of interest, and analogously $\mathcal{S}_k$ for key-side inputs ξ_p over all p. Assume $\|\xi\|_2 > 0$ and $\mathrm{RMS}(W_*\xi) > 0$ for every ξ in the relevant set, and define $m_q := \min_{\xi\in\mathcal{S}_q} \mathrm{RMS}(W_q\xi)/\|\xi\|_2$ and $m_k := \min_{\xi\in\mathcal{S}_k} \mathrm{RMS}(W_k\xi)/\|\xi\|_2$. Then $m_q, m_k > 0$ in the finite calibration set we measure.

Intuition. The sink margin is threatened when a small ΔW produces a large logit shift δ_p, and the three regimes differ in how the input ξ enters. Without QK-RMSNORM, $\delta_p \sim \Delta W \cdot \xi$ carries two input factors. The first is the query magnitude $\|\xi_q\|$, present at any query position whether or not the sink falls there. The second is the key-side magnitude term, which includes the sink input $\|\xi_{p_{\mathrm{sink}}}\|$ and can therefore be amplified when the sink carries a massive-activation channel (Sun et al., 2024). Per-head QK-RMSNORM divides Q and K_p by their own RMS, which absorbs $\|\xi\|$, so the bound depends only on the frozen $\|\gamma\|_\infty$ and the projection-RMS lower bounds, removing the explicit dependence on raw input magnitude. Layerwise QK-RMSNORM also absorbs $\|\xi\|$, through a single RMS over $Q_{\mathrm{cat}} = [Q_1; \cdots; Q_H]$, but the scale it leaves behind depends on each head's share of the layer's energy, so a weak head is compressed rather than normalized.

Theorem 1 (Logit-gap perturbation)**.** *Under Assumptions 1 and 2, in the regime where the first-order Taylor expansion of $\tilde{q}, \tilde{k}$ near W_q, W_k is valid (Appendix A.1), we have, with an absolute constant c and $O(\varepsilon^2)$ constants depending on the fixed weights W_q, W_k, scales $\gamma_q, \gamma_k, \gamma_{\mathrm{in}}$, calibration inputs $\{\xi_q, \xi_p\}$, and RMS lower bounds m_q, m_k,* No QK-RMSNORM:

$$B_{\mathrm{gap}} \le c\,\varepsilon\, d_h^{-1/2} \|\xi_q\| \,(\|W_q\|_{\mathrm{op}} + \|W_k\|_{\mathrm{op}}) \cdot \sup_{p \ne p_{\mathrm{sink}}} (\|\xi_{p_{\mathrm{sink}}}\| + \|\xi_p\|) + O(\varepsilon^2). \quad (4)$$

With QK-RMSNORM:

$$B_{\mathrm{gap}} \le c\,\varepsilon\, \|\gamma_q\|_\infty \|\gamma_k\|_\infty (m_q^{-1} + m_k^{-1}) + O(\varepsilon^2). \quad (5)$$

The two bounds make this precise, with explicit raw-input-magnitude factors surviving only in the no-QK-RMSNORM case.

Remark 1 (Layerwise QK-RMSNORM extension; semi-formal). Here the H heads share a single RMS denominator over the concatenated $H \cdot d_h$-dimensional Q_{cat}, with the layer's γ_q applied afterwards. Extracting head h from $\tilde{Q}_{\mathrm{cat}} = \gamma_q \odot Q_{\mathrm{cat}}/\mathrm{RMS}(Q_{\mathrm{cat}})$ gives

$$\|\tilde{Q}_h\| \;\le\; \|\gamma_q^{(h)}\|_\infty \|Q_h\| \frac{\sqrt{H d_h}}{\|Q_{\mathrm{cat}}\|}, \quad (6)$$

which replaces the head-energy-independent envelope $\|\gamma_q^{(h)}\|_\infty \sqrt{d_h}$ that per-head normalization provides; the same applies to keys. Writing $\rho_h := \|Q_h\|/\|Q_{\mathrm{cat}}\|$ for the head's share of layer energy, the two agree when $\rho_h = 1/\sqrt{H}$, while a head carrying less than its share is compressed and its attention logits, including the sink's lead, are attenuated with it. The margin can therefore fail from the G_{base} side rather than through a larger B_{gap}, and Molmo2-O is the empirical analogue with the weakest per-head G_{base} among the calibrated pairs

of Table 5 (full derivation in Appendix A.5; empirical control in Appendix D.2).

Lemma 1 (Saturation transfer)**.** *For sink position p_{sink} with pre-perturbation aggregate gap G_{base} and gap perturbation B_{gap},*

$$1 - a'_{p_{\text{sink}}} \;\le\; \sigma\big(B_{\text{gap}} - G_{\text{base}}\big), \qquad (7)$$

where $\sigma(x) = e^x/(1+e^x)$ is the logistic function.

The margin $G_{\text{base}} - B_{\text{gap}}$ controls an upper bound on the non-sink mass through $\sigma(B_{\text{gap}} - G_{\text{base}})$. Preserving the sink at level $1-\eta$ requires $G_{\text{base}} - B_{\text{gap}} \ge \log((1-\eta)/\eta)$, and negative margins are consistent with the observed sink collapse.

4.3 Empirical Sink Behavior

We calibrate the margin quantities on the seven pairs, measure the cross-pair sink-mass distribution, and use the layerwise controls to distinguish aggregate from per-head sink.

Calibration across the seven pairs. We compute $(G_{\text{base}}, B_{\text{gap}})$ across all five headline pairs and the two layerwise controls on a 15-prompt calibration set, aggregating over (layer, head, query position) in the last $\sim$10 layers where sink behavior is established beyond the first few local-context layers (Xiao et al., 2024). G_{base} is computed on the reference LLM at the per-head argmax key p_{sink}, and B_{gap} is reconstructed from the change in attention-probability ratios between the reference LLM and the VLM-LM at the same p_{sink}. The empirical B_{gap} is the realized reference-to-VLM-LM gap change and thus includes hidden-state and other parameter drift; it enters the margin of Lemma 1 but is not a direct numerical evaluation of Theorem 1's projection-only bound. For released pairs, the reference LLM is a comparison point rather than necessarily the immediate pre-VL checkpoint. Within each Qwen generation, the two pairs share the same reference LLM and therefore the same G_{base}; per-pair values are reported in Table 5. The marginal proxy $\text{median}(G) - \text{median}(B)$ is near zero on both per-head QK-RMSNORM pairs and strongly negative on all three no-QK-RMSNORM pairs. Both layerwise QK-RMSNORM pairs land at the negative end of the sample, with Molmo2-O driven by the weakest base sink in the set. The ordering tracks the behavioral split of Figure 2. Calibration sample-size sensitivity is reported in Appendix D.1.

Pair	QK	G_{base}	B_{gap}	$G-B$
Qwen3-VL	per-head	+2.36	+2.06	**+0.31**
InternVL3.5	per-head	+2.36	+2.47	**−0.11**
Qwen2.5-VL	none	+1.18	+5.01	**−3.83**
InternVL3	none	+1.18	+3.40	**−2.22**
LLaVA-OV	none	+1.44	+2.53	**−1.09**
Molmo2-O	layerwise	+0.24	+3.25	**−3.01**
MolmoE-1B	layerwise	+1.34	+4.19	**−2.84**

Table 5: Per-pair G_{base}, B_{gap}, and the marginal proxy $G - B := \text{median}(G_{\text{base}}) - \text{median}(B_{\text{gap}})$ of the Lemma 1 margin, on a 15-prompt calibration set.

Cross-pair sink and outcome bucketing. We partition each VLM-LM's IFEval prompts by per-prompt strict accuracy (pass at ≥ 0.5). For each bucket we measure position-0 attention, averaging over late-layer $\times$ head $\times$ trailing-query positions with bootstrap 95% CIs over 100 prompts per outcome bucket (Appendix C.3). The within-pair pass−fail gap is small (≤ 0.02 per-prompt mean), but the cross-pair span is large. The two per-head QK-RMSNORM pairs concentrate ≥ 0.70 of attention on position 0, while the no-QK-RMSNORM pairs span 5×10^{-5} to 0.50 and Molmo2-O (layerwise) sits at 0.25. Within the same OpenGVLab InternVL lineage, InternVL3 (no-QK-RMSNORM, 5×10^{-5}) and InternVL3.5 (per-head QK-RMSNORM, 0.74) span four orders of magnitude on this measure; the Qwen-VL family shows the same direction at smaller magnitude ($0.03 \to 0.70$ across Qwen2.5-VL $\to$ Qwen3-VL). Within-vendor framing reduces vendor-level confounding and strengthens the association with the architectural change, although recipe differences remain. The within-pair near-equality is consistent with the broken-sink state being a model-level head configuration rather than a prompt-level outcome marker. The cross-pair span is what tracks the behavioral split, so effective interventions should operate at the head level.

Per-head vs. aggregate sink. Molmo2-O (Olmo-3 base, layerwise QK-RMSNORM; Remark 1) separates *aggregate* from *per-head* sink: aggregate position-0 attention (0.25) is of the same order as LLaVA-OV's (0.50) yet IFEval damage is far worse, while per-head sink is the weakest among the calibrated pairs and already weak in the Olmo-3 reference LLM (Appendix D.2). The marginal proxy $G - B$ is strongly negative (-3.01, second only to the no-QK-RMSNORM Qwen2.5-VL at -3.83; Table 5), so Lemma 1's per-head margin

protection is not delivered head-by-head. MolmoE-1B, a second layerwise backbone on an MoE base, shows the same pattern, so both layerwise pairs carry strongly negative margins despite differing in base family and architecture. Together, the evidence supports per-head QK-RMSNORM as the protective configuration in our sample, with the layerwise variant insufficient.

4.4 Channel and Direction Ablations

Ablating the sink amplifier. For each of Qwen3-VL's three normalization-scale tensors ($\gamma_{\text{in}}, \gamma_q, \gamma_k$), replacing the top-10 channels per layer by the layer-wise mean disables the amplifier without touching the residual stream or projections (Appendix D.3): joint kill costs 44 pt on IFEval against 20 pt as the sum of singles, a 24-pt super-additive interaction that supports sink amplification as a redundant multi-norm system contributing to format-sensitive capability. For reference, a norm-matched isotropic random perturbation causes a $\sim 6\times$ larger IFEval drop than the observed Qwen2.5-VL gap (Appendix D.4). Thus, weight-space distance alone does not explain the degradation; perturbation direction matters.

5 Controls and Discussion

Within-vendor recipe control. Two within-vendor pairs (Qwen3-VL vs. Qwen2.5-VL; InternVL3.5 vs. InternVL3) each share a similar VL recipe and pretraining lineage while differing in per-head QK-RMSNORM, which reduces but does not eliminate the recipe/data confound. Co-varying changes such as the added RL stage and tokenizer updates may modulate the magnitude, but they never flip the sign across the within-class spread.

Modality control (VL vs. Tülu trajectory). The Section 3.1 comparison contrasts released 7B endpoints against a common text-only reference. We additionally run a matched 3B trajectory, comparing vision fine-tuning with text-only Tülu SFT from the same Qwen2.5-3B-Instruct base (Lambert et al., 2025). The vision run ends with a weaker sink ($S: +0.41 \rightarrow +0.20$) as SEEDBench rises $60.1 \rightarrow 64.8$, whereas the text-only run holds S at $+0.58$–$+0.65$ and loses only ~ 5 IFEval points from the shared base (Appendix F.2). Most VL-side IFEval loss occurs by step 1k before S declines, so we treat S as a regime-level diagnostic rather than a step-level mediator.

Post-pretraining QK-RMSNORM injection (negative). A natural question is whether the protection associated with native per-head QK-RMSNORM can be reproduced by injecting the module after LLM pretraining. We add per-head QK-RMSNORM with unit-scale initialization to a clean Qwen2.5-3B-Instruct and then run the matched two-stage VL recipe (Appendix F.3). The injected variant lags vanilla throughout training, scoring only 11 on IFEval at the first evaluated checkpoint against the 59.9 base. Combined with the Molmo2-O control (Section 4), this supports native-pretraining co-adaptation of the projections, rather than the module's mere presence, as an important component of the protective regime.

Weight-space merging (negative). Seven off-the-shelf merging settings spanning linear task-vector mixes (Ilharco et al., 2023), spectral decompositions (Gargiulo et al., 2025; Lee et al., 2025b), and sparsified merges (Yadav et al., 2023; Yu et al., 2024) on Qwen2.5-VL and Qwen3-VL fail to recover IFEval on the low-S side: only the two mildest mixes avoid substantial additional degradation (Appendix F.4).

What the negative results do and do not show. Existing remedies mitigate the loss at the cost of adaptation capacity or additional text-data compute (Section 2). Our matched Tülu trajectory provides a mechanistic rationale for the data-side remedy: text-only optimization preserves the sink whereas VL optimization disrupts it. The two negative controls argue against simpler shortcuts: post-pretraining QK-RMSNORM injection does not reproduce native protection, and post-VL merging does not recover the lost capability under our settings. This motivates pre-VL screening with S and head-selective training-time protection.

6 Conclusion

Across five headline VLM–LLM pairs from three model families, format-sensitive text-capability gaps track Sink Strength of the corresponding reference LLM, with per-head QK-RMSNORM as the strongest architectural correlate in our sample. The relation extends from the six-pair diagnostic set to a 17-pair panel spanning MoE and multiple LLM families. Negative injection and merging controls further motivate pre-VL screening and head-selective training-time protection.

Limitations

To our knowledge, the multimodal literature has not previously characterized the phenomenon we surface, in which VL adaptation erodes the base LLM's per-head attention sink and carries a downstream cost to format-sensitive capability. Within our sample, per-head QK-RMSNORM covaries with vendor, recipe, and pretraining data. Within-vendor comparisons (Qwen2.5→Qwen3, InternVL3→InternVL3.5) reduce but do not isolate the architectural axis. We validate S against VL adaptation outcomes, so whether it also predicts instruction-following drift under adaptation regimes we did not sample, such as continual pretraining, remains open and is a natural next test of the diagnostic's reach.

The post-VL merging methods we test do not recover the lost capability, while post-pretraining QK-RMSNORM injection fails to reproduce the native protective regime, leaving head-selective training-time intervention as an open direction (Section 5). A direct causal intervention would test sufficiency, for instance disabling per-head QK-RMSNORM during VL training of Qwen3 to see whether IFEval collapses. We leave this to future work.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. 2025. Why do LLMs attend to the first token? In *Second Conference on Language Modeling*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, et al. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28652–28668.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision transformers need registers. In *International Conference on Learning Representations*.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *Preprint*, arXiv:2412.19437.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2025. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18695–18705.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. 2020. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, et al. 2025. Tülu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: Insights and future directions. In *NeurIPS 2024 Workshops: RBFM*.

Seongyun Lee, Geewook Kim, Jiyeon Kim, Hyunji Lee, Hoyeon Chang, Sue Hyun Park, and Minjoon Seo. 2025a. How does vision-language adaptation impact the safety of vision language models? In *International Conference on Learning Representations*.

Yu-Ang Lee, Ching-Yun Ko, Tejaswini Pedapati, I-Hsin Chung, Mi-Yen Yeh, and Pin-Yu Chen. 2025b. STAR: Spectral truncation and rescale for model merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 496–505.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. 2025. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Samuel J. Paech. 2023. EQ-Bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Neale Ratzlaff, Man Luo, Xin Su, Vasudev Lal, and Phillip Howard. 2025. Training-free mitigation of language reasoning degradation after multimodal instruction tuning. In *Proceedings of the AAAI Symposium Series*, volume 5, pages 384–388.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*.

Shikhar Srivastava, Md Yousuf Harun, Robik Singh Shrestha, and Christopher Kanan. 2026. Improving multimodal large language models using continual learning. In *Proceedings of The 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pages 736–755. PMLR.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. In *First Conference on Language Modeling*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Jianhong Tu, Zhuohao Ni, Nicholas Crispino, Zihao Yu, Michael Bendersky, Beliz Gunel, Ruoxi Jia, Xin Liu, Lingjuan Lyu, Dawn Song, and Chenguang Wang. 2025. MLAN: Language-based instruction tuning preserves and transfers knowledge in multimodal language models. In *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 59–74.

Praveen Venkateswaran and Danish Contractor. 2026. Spotlight your instructions: Instruction-following with dynamic attention steering. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3752–3770.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025a. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. 2025b. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 63088–63112. PMLR.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57755–57775. PMLR.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. 2025. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Zeping Yu and Sophia Ananiadou. 2025. Locate-then-merge: Neuron-level parameter fusion for mitigating catastrophic forgetting in multimodal LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7065–7078.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. 2023. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40770–40803. PMLR.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and*

Learning, volume 234 of *Proceedings of Machine Learning Research*, pages 202–227. PMLR.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. 2025. Attention sinks: A 'catch, tag, release' mechanism for embeddings. *Advances in Neural Information Processing Systems*, 38:83140–83181.

Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2024. Wings: Learning multimodal LLMs without text-only forgetting. *Advances in Neural Information Processing Systems*, 37:31828–31853.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Table of Contents

A. Theory and Proofs ..... p. 15
A.1. Single-Head Perturbation Bound (Theorem 1) ..... p. 15
A.2. Aggregate-Gap Saturation Transfer (Lemma 1) ..... p. 16
A.3. Frobenius / Stable-Rank Restatement ..... p. 17
A.4. Multi-Layer Composition Remark ..... p. 17
A.5. Why Layerwise QK-RMSNorm Does Not Provide the Same Per-Head Guarantee ..... p. 18
A.6. Numerical Sanity Checks ..... p. 18
B. Evaluation Setup ..... p. 20
B.1. Evaluation Protocols ..... p. 20
B.2. Backbone Extraction ..... p. 20
B.3. Implementation Details and Reproducibility ..... p. 20
B.4. Models and Datasets ..... p. 20
C. Per-Task and Behavioral Evidence ..... p. 22
C.1. Capability Preservation vs. Format-Sensitive Degradation ..... p. 22
C.2. Failure-Category Breakdown ..... p. 23
C.3. Outcome-Bucketed Sink Probe ..... p. 23
C.4. Extended Capability Panel ..... p. 23
D. Mechanism Details ..... p. 24
D.1. Calibration Sample-Size Robustness ..... p. 24
D.2. Layerwise Backbones: Aggregate vs. Per-Head Sink ..... p. 25
D.3. Channel Ablation Sweep ..... p. 26
D.4. Random-ΔW Control ..... p. 26
E. Sink Strength Predictor ..... p. 26
E.1. Per-Pair Sink Strength Predictions ..... p. 26
E.2. IFEval-Specific Predictor (17 pairs) ..... p. 26
E.3. Multi-Task Generalization (9 pairs) ..... p. 26
E.4. Sink Strength Design Ablations ..... p. 26
F. Controls ..... p. 27
F.1. Modality Comparison: Text-Only Endpoint ..... p. 27
F.2. Modality Control: VL vs. Text-Only Trajectory ..... p. 27
F.3. Injection Experiment Training Recipe ..... p. 28
F.4. Weight-Merging Sweep on Format-Sensitive Benches ..... p. 29

A Theory and Proofs

Notation as in Section 4. We work with a single attention head, hidden size d, head dim d_h. The pre-norm inputs are $\xi_q, \xi_p = \gamma_{\text{in}} \odot \text{RMSNorm}(x_q), \gamma_{\text{in}} \odot \text{RMSNorm}(x_p) \in \mathbb{R}^d$ at the query and key-position-p tokens, and the pre-QK-RMSNORM projections are $Q_{\text{pre}} = W_q \xi_q$, $K_{\text{pre},p} = W_k \xi_p$ with $W_q, W_k \in \mathbb{R}^{d_h \times d}$. Under QK-RMSNORM we set $\tilde{q} = Q_{\text{pre}}/\text{RMS}(Q_{\text{pre}})$ and $Q = \gamma_q \odot \tilde{q}$ (similarly K_p). For clarity, the main text omits RoPE from the notation. Here, let R_q and R_p denote the orthogonal RoPE transformations at the query and key positions. Orthogonality preserves the norm of each rotated vector, while differences across positions are bounded by the triangle inequality; this yields the sum-of-norms term in Equation 4. The sink-anchored gap perturbation is $B_{\text{gap}} := \sup_{p \neq p_{\text{sink}}} |\delta_p - \delta_{p_{\text{sink}}}|$, and the aggregate sink gap is $G := l_{p_{\text{sink}}} - \log \sum_{p \neq p_{\text{sink}}} e^{l_p}$.

A.1 Single-Head Perturbation Bound (Theorem 1)

Proof of Theorem 1. Let $\Delta W_q, \Delta W_k$ be the projection perturbations with $\|\Delta W_*\|_{\text{op}} \leq \varepsilon$. The two cases are treated separately because the dependence of Q, K_p on the projections is linear without QK-RMSNORM and nonlinear (through the RMS denominator) with QK-RMSNORM: the no-QK-RMSNORM case is exactly linear up to the explicit second-order cross term, whereas the QK-RMSNORM case requires a local Taylor expansion through the RMS denominator.

No QK-RMSNORM. Let R_q and R_p denote the orthogonal RoPE transformations at the query position and key position p, respectively. The attention logit is

$$\ell_p = \frac{(R_q W_q \xi_q)^\top (R_p W_k \xi_p)}{\sqrt{d_h}}.$$

Since the projections are linear, the first-order perturbation is

$$\delta_p = \frac{1}{\sqrt{d_h}} \Big[(R_q \Delta W_q \xi_q)^\top (R_p W_k \xi_p) + (R_q W_q \xi_q)^\top (R_p \Delta W_k \xi_p) \Big] + O(\varepsilon^2).$$

Let $s = p_{\text{sink}}$. The sink-anchored gap perturbation between s and any $p \neq s$ is

$$\begin{aligned} \delta_p - \delta_s = \frac{1}{\sqrt{d_h}} \Big[& (R_q \Delta W_q \xi_q)^\top \big(R_p W_k \xi_p - R_s W_k \xi_s \big) \\ & + (R_q W_q \xi_q)^\top \big(R_p \Delta W_k \xi_p - R_s \Delta W_k \xi_s \big) \Big] + O(\varepsilon^2). \end{aligned} \tag{8}$$

Because each R_p is orthogonal,

$$\|R_p W_k \xi_p - R_s W_k \xi_s\| \leq \|W_k\|_{\text{op}} \big(\|\xi_p\| + \|\xi_s\| \big), \tag{9}$$

$$\|R_p \Delta W_k \xi_p - R_s \Delta W_k \xi_s\| \leq \varepsilon \big(\|\xi_p\| + \|\xi_s\| \big). \tag{10}$$

Also,

$$\|R_q \Delta W_q \xi_q\| \leq \varepsilon \|\xi_q\|, \qquad \|R_q W_q \xi_q\| \leq \|W_q\|_{\text{op}} \|\xi_q\|.$$

Applying Cauchy–Schwarz therefore gives

$$|\delta_p - \delta_s| \leq \frac{\varepsilon}{\sqrt{d_h}} \|\xi_q\| \big(\|W_q\|_{\text{op}} + \|W_k\|_{\text{op}} \big) \big(\|\xi_s\| + \|\xi_p\| \big) + O(\varepsilon^2).$$

Taking the supremum over $p \neq s$ and absorbing fixed numerical factors into the absolute constant c yields Equation 4.

With QK-RMSNORM. The logit is $l_p = (\gamma_q \odot \tilde{q}) \cdot (\gamma_k \odot \tilde{k}_p)/\sqrt{d_h}$ where $\tilde{q} = Q_{\text{pre}}/\text{RMS}(Q_{\text{pre}})$ and analogously $\tilde{k}_p$. The map from W_q, W_k to $\tilde{q}, \tilde{k}_p$ is differentiable on the open set where $\text{RMS}(Q_{\text{pre}}), \text{RMS}(K_{\text{pre},p}) > 0$, which Assumption 2 ensures. We work with the mathematical RMS here, and production implementations add a small stabilizer $\tau > 0$ inside the square root (distinct from the perturbation magnitude ε). Since $\text{RMS}_\tau \geq \text{RMS}$, the unstabilized bound used below is conservative. To first order in ε on a small neighborhood the logit perturbation decomposes into a Q-side and K-side term:

$$\delta_p = \tfrac{1}{\sqrt{d_h}}\big[(\gamma_q \odot \Delta\tilde{q}) \cdot (\gamma_k \odot \tilde{k}_p) + (\gamma_q \odot \tilde{q}) \cdot (\gamma_k \odot \Delta\tilde{k}_p)\big] + O(\varepsilon^2).$$

We bound each term separately. For the Q-side perturbation,

$$\Delta\tilde{q} = \frac{\Delta W_q \xi_q}{\text{RMS}(Q_{\text{pre}})} - \frac{Q_{\text{pre}}\,\Delta\text{RMS}}{\text{RMS}^2(Q_{\text{pre}})},$$

with $|\Delta\text{RMS}| \leq \|\Delta W_q\|_{\text{op}}\|\xi_q\|/\sqrt{d_h}$. Both terms are $O(\varepsilon\|\xi_q\|/\text{RMS}(Q_{\text{pre}}))$, and by Assumption 2 $\text{RMS}(Q_{\text{pre}}) \geq m_q\|\xi_q\|$ on the calibration distribution, so

$$\|\Delta\tilde{q}\| \leq \frac{c_1\varepsilon}{m_q},$$

with c_1 an absolute constant. The crucial point is that $\|\xi_q\|$ has cancelled, since numerator and denominator both scale linearly with $\|\xi_q\|$.

For the sink-anchored gap perturbation we then bound the Q-side contribution after RoPE. Since each R_p is orthogonal and mathematical unit-RMS gives $\|\tilde{k}_p\|_2 = \sqrt{d_h}$ (with the stabilizer τ, $\|\tilde{k}_p\|_2 \leq \sqrt{d_h}$), we have

$$\big\|R_{p_{\text{sink}}}(\gamma_k \odot \tilde{k}_{p_{\text{sink}}}) - R_p(\gamma_k \odot \tilde{k}_p)\big\| \leq 2\|\gamma_k\|_\infty\sqrt{d_h}.$$

Therefore,

$$\begin{aligned}
&\Big|R_q(\gamma_q \odot \Delta\tilde{q}) \cdot \Big[R_{p_{\text{sink}}}(\gamma_k \odot \tilde{k}_{p_{\text{sink}}}) - R_p(\gamma_k \odot \tilde{k}_p)\Big]\Big| \\
&\leq \|\gamma_q\|_\infty\|\Delta\tilde{q}\| \cdot 2\|\gamma_k\|_\infty\sqrt{d_h} \\
&\leq \frac{2c_1\varepsilon\|\gamma_q\|_\infty\|\gamma_k\|_\infty\sqrt{d_h}}{m_q}.
\end{aligned}$$

A symmetric argument for the K-side gives the $1/m_k$ term. Summing the two sides, dividing by the outer $\sqrt{d_h}$ in the logit definition, and taking the supremum over $p \neq p_{\text{sink}}$ yields

$$B_{\text{gap}} \leq c\,\varepsilon\,\|\gamma_q\|_\infty\,\|\gamma_k\|_\infty\big(\tfrac{1}{m_q} + \tfrac{1}{m_k}\big) + O(\varepsilon^2),$$

which is Equation 5. □

Input-magnitude cancellation. The $\|\xi_q\|$ that the no-QK-RMSNORM bound carries is cancelled in the QK-RMSNORM pathway by the RMS denominator. What replaces it is the projection-RMS lower bound m_q, and if a calibration prompt puts ξ_q in or near the null space of W_q, m_q degrades and the bound loosens. Assumption 2 excludes exact degeneracy; near-null cases appear as small m_q or m_k and therefore loosen the bound, consistent with the empirical $m_q, m_k > 0$ we measure on the calibration set.

A.2 Aggregate-Gap Saturation Transfer (Lemma 1)

Proof of Lemma 1. Write the perturbed softmax probability at position p as $a'_p = e^{l'_p}/Z'$ with $l'_p = l_p + \delta_p$ and $Z' = \sum_q e^{l'_q}$. Then for $p \neq p_{\text{sink}}$,

$$\frac{a'_p}{a'_{p_{\text{sink}}}} = \exp(l_p - l_{p_{\text{sink}}} + \delta_p - \delta_{p_{\text{sink}}}).$$

Summing over $p \neq p_{\text{sink}}$,

$$\begin{aligned}\frac{1 - a'_{p_{\text{sink}}}}{a'_{p_{\text{sink}}}} &= \sum_{p \neq p_{\text{sink}}} \exp\big(l_p - l_{p_{\text{sink}}}\big) \exp\big(\delta_p - \delta_{p_{\text{sink}}}\big) \\ &\leq \exp\big(B_{\text{gap}}\big) \sum_{p \neq p_{\text{sink}}} \exp\big(l_p - l_{p_{\text{sink}}}\big) \\ &= \exp\big(B_{\text{gap}} - G\big),\end{aligned}$$

using the definition $G = l_{p_{\text{sink}}} - \log \sum_{p \neq p_{\text{sink}}} e^{l_p}$ and $|\delta_p - \delta_{p_{\text{sink}}}| \leq B_{\text{gap}}$ for all $p \neq p_{\text{sink}}$. Let $u := 1 - a'_{p_{\text{sink}}}$ and write the inequality as $u/(1-u) \leq e^{B_{\text{gap}} - G}$. Solving for u gives $u \leq e^{B_{\text{gap}} - G}/(1 + e^{B_{\text{gap}} - G}) = \sigma(B_{\text{gap}} - G)$, which is the logistic form of Equation 7. The weaker exponential bound $u \leq e^{B_{\text{gap}} - G}$ follows from $\sigma(x) \leq e^x$. □

Comparison to top-2 gap arguments. If one tried to state the same lemma using the top-2 gap $T = l_{p_{\text{sink}}} - \max_{p \neq p_{\text{sink}}} l_p$ instead of the aggregate G, one would need an extra $\log N$ term to control the sum $\sum_{p \neq p_{\text{sink}}} e^{l_p}$, giving $1 - a'_{p_{\text{sink}}} \leq N \exp(-T + B_{\text{gap}})$ or equivalently $\exp(-T + B_{\text{gap}} + \log N)$, which is vacuous unless $T > \log N + B_{\text{gap}}$. The aggregate-gap form sidesteps this and matches the measured quantity directly.

A.3 Frobenius / Stable-Rank Restatement of Theorem 1

The bound is stated in terms of $\|\Delta W_q\|_{\text{op}}$ and $\|\Delta W_k\|_{\text{op}}$. Practitioners often report the Frobenius norms $\|\Delta W_q\|_F, \|\Delta W_k\|_F$ instead, and the conversion is via per-projection stable ranks.

The restatement. Let $r_q := \|\Delta W_q\|_F^2 / \sigma_{\max}^2(\Delta W_q)$ and analogously r_k. By definition $\|\Delta W_*\|_{\text{op}} = \sigma_{\max}(\Delta W_*) = \|\Delta W_*\|_F / \sqrt{r_*}$. Keeping the two operator norms separate in the proof before upper-bounding both by a common ε, and substituting $\|\Delta W_*\|_{\text{op}} = \|\Delta W_*\|_F / \sqrt{r_*}$, gives the Frobenius/stable-rank restatement

$$B_{\text{gap}} \;\leq\; c\, \|\gamma_q\|_\infty \, \|\gamma_k\|_\infty \left(\frac{\|\Delta W_q\|_F}{\sqrt{r_q}\, m_q} + \frac{\|\Delta W_k\|_F}{\sqrt{r_k}\, m_k} \right) + O(\varepsilon^2), \tag{11}$$

with empirical inputs $\{r_q, r_k, \|\Delta W_q\|_F, \|\Delta W_k\|_F, m_q, m_k, \|\gamma_q\|_\infty, \|\gamma_k\|_\infty\}$. This is a rewriting using stable ranks, not a tightening. If one perturbation is zero, its corresponding term is taken to be zero; otherwise the expression applies with the stable rank defined above.

Role of the restatement. The Frobenius form makes the structural observation explicit: the QK-RMSNORM side has no explicit raw-input-magnitude factor, only the stable ranks r_q, r_k and the γ, m constants. The number that enters the main text is the directly measured B_{gap} on the calibration set (Section 4), and the measured value on Qwen3 (2.06 nats) is about $2.4\times$ smaller than on Qwen2.5 (5.01 nats) on the same 15-prompt calibration set (Table 5). The Qwen3-versus-Qwen2.5 comparison is consistent with the theorem's structural prediction that the QK-RMSNORM side is less sensitive to raw input magnitude, but we do not claim it as a causal isolation of QK-RMSNORM alone, since Qwen3 and Qwen2.5 differ in pre-training data, RL stage, and tokenizer in addition to QK-RMSNORM.

A.4 Multi-Layer Composition Remark

Theorem 1 and Lemma 1 are single-layer statements. A formal multi-layer composition requires additional control of the value vectors, residual stream, MLP, and post-attention LayerNorm, which we do not provide here. We instead state an informal chaining intuition. If at every layer ℓ the per-layer base aggregate gap $G_{\text{base},\ell}$ exceeds the per-layer gap perturbation by a margin $G_{\text{base},\ell} - B_{\text{gap},\ell} \geq \mu$, then by Lemma 1 the perturbed non-sink mass at that layer is at most $e^{-\mu}$, and the next layer *is approximately fed* a hidden state close to the LLM baseline under additional Lipschitz assumptions on the intervening operators. We do not formalize these assumptions and report this remark only as motivation for the per-layer diagnostic.

Caveat. In our calibration measurements, the marginal proxy is negative in the top late layers of Qwen2.5 (Section 4). A tight multi-layer theorem (in the spirit of (Olsson et al., 2022)'s circuit composition) is future work.

A.5 Why Layerwise QK-RMSNorm Does Not Provide the Same Per-Head Guarantee

This subsection derives the claim of Remark 1, that a shared RMS denominator does not provide the same per-head margin guarantee as per-head QK-RMSNORM.

Per-head normalization. Under per-head QK-RMSNORM, head h is divided by its own RMS, $\mathrm{RMS}(Q_h) = \|Q_h\|/\sqrt{d_h}$, so

$$\|\tilde{Q}_h\| = \left\|\gamma_q^{(h)} \odot \frac{Q_h}{\mathrm{RMS}(Q_h)}\right\| \le \|\gamma_q^{(h)}\|_\infty \sqrt{d_h}. \tag{12}$$

The input magnitude $\|Q_h\|$ cancels, which is exactly why the QK-RMSNORM branch of Theorem 1 carries no $\|\xi\|$ factor.

Layerwise normalization. Under layerwise QK-RMSNORM, all H heads share one denominator over the concatenated vector $Q_{\mathrm{cat}} = [Q_1; \cdots; Q_H]$, so $\mathrm{RMS}(Q_{\mathrm{cat}}) = \|Q_{\mathrm{cat}}\|/\sqrt{Hd_h}$. Extracting the h-th block of $\tilde{Q}_{\mathrm{cat}} = \gamma_q \odot Q_{\mathrm{cat}}/\mathrm{RMS}(Q_{\mathrm{cat}})$ gives

$$\|\tilde{Q}_h\| \le \|\gamma_q^{(h)}\|_\infty \|Q_h\| \frac{\sqrt{Hd_h}}{\|Q_{\mathrm{cat}}\|}, \tag{13}$$

which replaces the per-head envelope of Equation 12. The same derivation applies to keys.

What the shared denominator costs. Write $\rho_h := \|Q_h\|/\|Q_{\mathrm{cat}}\|$ for the share of layer energy carried by head h, so that Equation 13 reads $\|\gamma_q^{(h)}\|_\infty \rho_h \sqrt{Hd_h}$. Both normalizations remove the input magnitude, since ρ_h is unchanged when ξ_q is rescaled. What per-head normalization additionally provides is a scale independent of the head energy, namely $\sqrt{d_h}$ up to the learned factor $\|\gamma_q^{(h)}\|_\infty$. The layerwise scale instead additionally tracks ρ_h. Under equal per-head energy $\rho_h = 1/\sqrt{H}$ and Equation 13 collapses to Equation 12, but a head carrying less than its share has $\rho_h < 1/\sqrt{H}$ and is compressed relative to its neighbours.

Consequence for the margin. Compression attenuates that head's attention logits, and with them the sink's logit lead, so G_{base} can be small at heads that carry little of the layer's energy. The margin $G_{\mathrm{base}} - B_{\mathrm{gap}}$ can then fail from the base side rather than through a larger perturbation, which is the opposite of the no-QK-RMSNORM failure mode where $\|\xi\|$ enters the bound on B_{gap} directly. Aggregate sink mass is a sum over heads and can stay high while individual compressed heads carry almost no lead, so preservation of the aggregate does not imply preservation of the per-head margin. Appendix D.2 reports the empirical counterpart, where Molmo2-O has the lowest per-head G_{base} among the calibrated pairs at an unremarkable B_{gap}, while the no-QK-RMSNORM Qwen2.5-VL shows the reverse profile.

A.6 Numerical Sanity Checks

We verify two pieces on small synthetic cases.

Theorem 1, $d = 16$, $d_h = 4$, no QK-RMSNORM. We consider a RoPE-free synthetic case with Gaussian-initialized W_q, W_k. We instantiate a sequence of $N = 8$ key positions with i.i.d. unit-norm Gaussian backgrounds, then form $\xi_{p_{\mathrm{sink}}}$ by adding a massive-activation-style channel at index 3 with value 10 on top of its background, while ξ_p for $p \neq p_{\mathrm{sink}}$ retain only the unit-norm background. The sink hidden state thus has a sink/non-sink magnitude gap dominated by the size-10 channel rather than being constrained to unit norm. We draw Gaussian ΔW_* with $\|\Delta W_*\|_{\mathrm{op}} = 0.1$. The measured sink-anchored quantity is $\sup_{p \neq p_{\mathrm{sink}}} |\delta_p - \delta_{p_{\mathrm{sink}}}| = 0.169$. Because this synthetic case omits RoPE, the proof admits the sharper specialization with $\|\xi_{p_{\mathrm{sink}}} - \xi_p\|$ before the triangle-inequality relaxation used in Equation 4. This specialized bound is 0.197, giving a ratio of 0.86. Since $\|\xi_{p_{\mathrm{sink}}} - \xi_p\| \le \|\xi_{p_{\mathrm{sink}}}\| + \|\xi_p\|$, the more general RoPE-aware bound in Equation 4 is also satisfied.

Lemma 1, $G = 3.0$, $B_{\text{gap}} = 0.4$, $N = 64$. Direct softmax computation with $l_{p_{\text{sink}}} = 5$ and the $N - 1$ non-sink positions equally distributed to give $G = 3.0$. To realize the worst-case gap perturbation, we set $\delta_p = +B_{\text{gap}}/2$ for every $p \neq p_{\text{sink}}$ and $\delta_{p_{\text{sink}}} = -B_{\text{gap}}/2$, so that $\sup_{p \neq p_{\text{sink}}} |\delta_p - \delta_{p_{\text{sink}}}| = B_{\text{gap}}$ is attained. Empirical $1 - a'_{p_{\text{sink}}} = 0.069$. The logistic bound from Equation 7 is $\sigma(0.4 - 3.0) = \sigma(-2.6) \approx 0.069$. Ratio 1.00 (the bound is tight under the worst-case perturbation).

B Evaluation Setup

This appendix collects evaluation infrastructure used across the paper: the harness protocols (Appendix B.1), the backbone-extraction pipeline (Appendix B.2), and hardware/software details for reproducibility (Appendix B.3).

B.1 Evaluation Protocols

We evaluate all text-side capability with `lm-evaluation-harness` v0.4.12 (Gao et al., 2024) under two protocols whose disagreement on a single backbone is itself diagnostic.

Instruct protocol. `–apply_chat_template`, `–num_fewshot 0`, `–batch_size 8`, bf16. This matches the intended chat-style usage of instruction-tuned models.

Community-default protocol. The harness's per-task default few-shot count (`mmlu`=5, `gsm8k`=8, etc.) with the chat template *disabled*. This provides a common leaderboard-style comparison.

Nine-task suite. MMLU, MMLU-Pro, BoolQ, MBPP (code, pass@1), RULER (long-context), GSM8K-CoT, IFEval (`prompt_level_strict_acc`), GPQA-Diamond-CoT, and EQ-Bench (raw points). IFEval is the headline metric for sink corruption because its checker grades surface format directly (Zhou et al., 2023).

Pipeline sanity. Our LLM-side numbers reproduce the Qwen blog values on Qwen2.5-7B-Instruct GSM8K-CoT to within 0.93 pt (90.67 vs. 91.6) and MMLU-Pro to within 0.46 pt (55.84 vs. 56.30) when run under matched OpenCompass configurations. The deltas reported throughout this paper compare a reference LLM to its VLM-LM under the same protocol, reducing sensitivity to framework-specific evaluation differences.

B.2 Backbone Extraction

To evaluate text-only behavior of a VLM with `lm-evaluation-harness`, we extract its language backbone as a standalone causal LM. For Qwen2.5-VL and the LLaVA-OneVision / LLaVA-Llama3 families, we remap the language-backbone parameters to the corresponding `<Family>ForCausalLM` namespace while preserving the VLM's original input embeddings and output language-model head, including the original `lm_head.weight`. For Qwen3-VL, the same procedure additionally preserves the `q_norm` / `k_norm` scale parameters used in per-head normalization. For our QK-RMSNorm-injected Qwen2.5-3B variant (Section 5), we use a self-contained `trust_remote_code` module so that the extracted backbone follows the same numerical pathway as during training. All extracted backbones are saved to `safetensors`, and the round-trip reproduces the original language module's logits on a held-out text-only probe.

B.3 Implementation Details and Reproducibility

Hardware. Model training (the Section 5 QK-RMSNorm injection experiment) ran on four NVIDIA H100 GPUs (80 GB), with two GPUs allocated to each variant. All text-side evaluation (lm-evaluation-harness across the nine-task suite for all 6 pairs and the text-only endpoint control, with MMLU-Pro and RULER measured on the five headline pairs and Molmo2-O excluded for those two), the Sink Strength calibration forward passes (Section 3, $\sim$ 8 s per 7B backbone), and the $G_{\text{base}}, B_{\text{gap}}$ measurement on the 15-prompt calibration set ($\sim$ 52 s/pair, full pipeline including VLM-LM forward and m_q/m_k compute) ran on a single NVIDIA A6000 (48 GB) in bf16.

Software. PyTorch 2.10+cu128, transformers 4.57.6, DeepSpeed 0.19.0, lm-evaluation-harness 0.4.12, VLMEvalKit.

B.4 Models and Datasets

This subsection collects Hugging Face URLs and paper references for every model and dataset used in the paper.

Vision-language models (headline). The five headline VLMs and the Molmo2-O control (Section 3.1):

- Qwen3-VL-8B-Instruct (Bai et al., 2025a): `https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct`
- InternVL3.5-8B (Wang et al., 2025a): `https://huggingface.co/OpenGVLab/InternVL3_5-8B`
- Qwen2.5-VL-7B-Instruct (Bai et al., 2025b): `https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct`
- InternVL3-8B (Zhu et al., 2025): `https://huggingface.co/OpenGVLab/InternVL3-8B`

- LLaVA-OneVision-Qwen2-7B-OV (Li et al., 2025): https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov
- Molmo2-O-7B (Clark et al., 2026): https://huggingface.co/allenai/Molmo2-O-7B

Vision-language models (extended 17-pair panel). The eleven additional VLMs in the IFEval-specific 17-pair panel (Appendix E.2):

- Qwen3-VL-4B-Instruct (Bai et al., 2025a): https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct
- Qwen3-VL-2B-Instruct (Bai et al., 2025a): https://huggingface.co/Qwen/Qwen3-VL-2B-Instruct
- Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025a): https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct
- Ovis2-8B (Lu et al., 2024): https://huggingface.co/AIDC-AI/Ovis2-8B
- Ovis2-34B (Lu et al., 2024): https://huggingface.co/AIDC-AI/Ovis2-34B
- MolmoE-1B-0924 (Deitke et al., 2025): https://huggingface.co/allenai/MolmoE-1B-0924
- InternVL3-2B (Zhu et al., 2025): https://huggingface.co/OpenGVLab/InternVL3-2B
- InternVL3-14B (Zhu et al., 2025): https://huggingface.co/OpenGVLab/InternVL3-14B
- MiniCPM-V-4.5 (Yu et al., 2025): https://huggingface.co/openbmb/MiniCPM-V-4_5
- LLaVA-NeXT-Mistral-7B (Liu et al., 2024): https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf
- Idefics3-8B-Llama3 (Laurençon et al., 2024): https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3

Text-only reference LLMs. The text-only reference models evaluated alongside the extracted VLM-LMs:

- Qwen3-8B (Yang et al., 2025): https://huggingface.co/Qwen/Qwen3-8B
- Qwen3-4B (Yang et al., 2025): https://huggingface.co/Qwen/Qwen3-4B
- Qwen3-1.7B (Yang et al., 2025): https://huggingface.co/Qwen/Qwen3-1.7B
- Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025): https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507
- Qwen2.5-7B-Instruct (Qwen Team, 2024): https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
- Qwen2.5-1.5B-Instruct (Qwen Team, 2024): https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct
- Qwen2.5-14B-Instruct (Qwen Team, 2024): https://huggingface.co/Qwen/Qwen2.5-14B-Instruct
- Qwen2.5-32B-Instruct (Qwen Team, 2024): https://huggingface.co/Qwen/Qwen2.5-32B-Instruct
- Qwen2-7B-Instruct (Yang et al., 2024): https://huggingface.co/Qwen/Qwen2-7B-Instruct
- Qwen2.5-3B-Instruct (Qwen Team, 2024): https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
- Olmo-3-7B-Instruct: https://huggingface.co/allenai/Olmo-3-7B-Instruct
- OLMoE-1B-7B-0924-Instruct (Muennighoff et al., 2025): https://huggingface.co/allenai/OLMoE-1B-7B-0924-Instruct
- Mistral-7B-Instruct-v0.2 (Jiang et al., 2023): https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
- Llama-3.1-8B-Instruct (Grattafiori et al., 2024): https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Vision tower and projector donor. The injection experiment of Appendix F.3 takes its vision tower and projector from Qwen2.5-VL-3B-Instruct (Bai et al., 2025b), https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct.

Modality-control variant. The text-only further-training endpoint used in the 7B modality comparison (Appendix F.1): Qwen2.5-7B-Instruct-1M, https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-1M.

Text evaluation datasets. Four format-sensitive tasks and five capability tasks reported in the paper, all run via `lm-evaluation-harness` (Gao et al., 2024) (Appendix B.1):

- IFEval (Zhou et al., 2023): https://huggingface.co/datasets/google/IFEval
- EQ-Bench v2.1 (Paech, 2023): https://github.com/EQ-bench/EQ-Bench
- GSM8K (Cobbe et al., 2021): https://huggingface.co/datasets/openai/gsm8k
- GPQA-Diamond (Rein et al., 2024): https://huggingface.co/datasets/Idavidrein/gpqa
- MMLU (Hendrycks et al., 2021): https://huggingface.co/datasets/cais/mmlu
- BoolQ (Clark et al., 2019): https://huggingface.co/datasets/google/boolq
- MMLU-Pro (Wang et al., 2024): https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro
- MBPP (Austin et al., 2021): https://huggingface.co/datasets/google-research-datasets/mbpp
- RULER (Hsieh et al., 2024): https://github.com/NVIDIA/RULER

Calibration prompts. The 15-prompt set used for Sink Strength (Section 3) and the $G_{\text{base}}, B_{\text{gap}}$ measurement (Section 4) consists of short instruction-following prompts we wrote in the style of IFEval. No IFEval instance is used, so the calibration set is disjoint from the data we evaluate on by construction.

Licenses and terms of use. All Hugging Face checkpoints and evaluation datasets listed above are used under their published licenses, predominantly Apache 2.0 and MIT, with a small number under model-specific licenses (the Llama 3.1 Community License for Llama-3.1-8B-Instruct, and the research-only Qwen Research License for Qwen2.5-3B-Instruct and Qwen2.5-VL-3B-Instruct, which we use for research purposes only). We download each artifact under its stated terms and do not redistribute the underlying weights or data. The evaluation harness `lm-evaluation-harness` (Gao et al., 2024) is MIT-licensed.

C Per-Task and Behavioral Evidence

This appendix expands the per-task behavioral split surfaced in Section 3.1: capability vs. format-sensitive coverage (Appendix C.1), IFEval failure-category counts (Appendix C.2), an outcome-bucketed sink probe (Appendix C.3), and an extended capability panel on coding, reasoning, and long-context retrieval (Appendix C.4).

C.1 Capability Preservation vs. Format-Sensitive Degradation

Section 3.1 shows that several non-format-sensitive capabilities are comparatively preserved while format-following degrades. We verify this here on the knowledge- and comprehension-oriented benchmarks surfaced in the main text, and Appendix C.4 extends the check to coding, long-context retrieval, and advanced reasoning.

Knowledge and comprehension. Table 6 reports the per-pair Δ on MMLU and BoolQ alongside the four format-sensitive anchors under the same instruct protocol. MMLU and BoolQ drops never exceed 2.3 pt across the six pairs, and on three pairs they even improve. On the same backbones the headline IFEval gaps reach -9.6 pt (no-QK-RMSNorm) and -18.7 pt (layerwise).

	Knowledge / comprehension		Format-sensitive			
Pair	MMLU	BoolQ	IFEval	EQ	GSM8K	GPQA
Qwen3-VL	+1.6	+6.0	−5.4	−2.4	−2.1	−1.9
InternVL3.5	+3.0	+7.6	−1.8	−0.8	−1.5	−0.6
Qwen2.5-VL	−2.1	−1.2	−9.6	−11.2	−14.3	−9.8
InternVL3	+3.0	+4.5	−8.9	−10.3	−12.5	−7.1
LLaVA-OV	−2.3	+0.1	−7.9	−8.6	−11.9	−8.4
Molmo2-O	−0.0	−1.5	−18.7	−64.6	−42.7	−19.7
Mean	**+0.5**	**+2.6**	**−8.7**	**−16.3**	**−14.2**	**−7.9**

Table 6: Per-pair Δ on knowledge- and comprehension-oriented benchmarks (MMLU, BoolQ) vs. format-sensitive benchmarks (shaded). Same instruct protocol throughout, except that the Molmo2-O GPQA baseline is the published Olmo-3-7B-Instruct score rather than our own matched run, which would give -10.6. Pair-level QK category in Table 1. EQ is EQ-Bench v2.1 (raw points); GSM8K is GSM8K-CoT flexible-extract; GPQA is GPQA-Diamond-CoT flexible-extract.

Ruling out a format confound. A sceptical reader could attribute the contrast to MCQA-vs-open-generation format rather than capability: MMLU and BoolQ preserve because they are loglikelihood-graded letter-picks, and the format-sensitive tasks degrade because they are open generation. Two facts in our data rule out this simple reading. First, GPQA-Diamond-CoT is MCQA in format (the answer is one of A/B/C/D,

scored by flexible-extract on a free-text rationale), yet it drops by -7.1 to -9.8 pt on the three no-QK-RMSNORM pairs, almost the full IFEval-scale magnitude. Second, GSM8K-CoT is open generation with a multi-step reasoning trajectory, yet it stays within ± 2.1 pt on the two per-head QK-RMSNORM pairs. Preservation therefore tracks the capability being probed (knowledge/comprehension vs. instruction-following / reasoning-trajectory) and the QK protection of the reference LLM, not the task's MCQA-vs-generation format.

C.2 Failure-Category Breakdown

For each VLM-LM we partition the IFEval set into the four buckets reported in Table 2 (*both_pass*, *regression*, *recovered*, *both_fail*). The *regression* bucket (prompts the LLM passes but the VLM-LM fails) is the locus of the per-pair degradation. Table 7 reports the dominant failing IFEval instruction categories per pair (counts of failed instructions within the regression bucket). "Punctuation" is largely `no_comma`, "combination" is largely `repeat_prompt` or `two_responses`, and "length_constraints" is largely `number_words` or `number_paragraphs`. "Language" is the `response_language` constraint, "startend" covers `quotation` and `end_checker`, and "detectable_format" covers `multiple_sections` and `number_bullet_lists`. "Change_case" covers `english_capital` and `english_lowercase`, and "keywords" covers `existence` and `letter_frequency`.

C.3 Outcome-Bucketed Sink Probe

For each VLM-LM we partition the IFEval prompts by `prompt_level_strict_acc` into a pass bucket (≥ 0.5) and a fail bucket (< 0.5), then run a single forward pass per prompt with `output_attentions=True` and aggregate position-0 attention probability over (late layer $\times$ head $\times$ last-8 query positions). We report the per-prompt mean (so the bootstrap is over prompts, not flattened head/layer observations) and a 95% percentile CI from bootstrap over 100 prompts sampled per outcome bucket (200 per pair).

C.4 Extended Capability Panel

Beyond the knowledge/comprehension benchmarks and the four format-sensitive tasks, the nine-task suite of Appendix B.1 covers coding (MBPP, execution pass@1), advanced multi-domain reasoning (MMLU-Pro), and long-context retrieval (RULER). Table 8 reports these three on the five headline pairs; Molmo2-O, the layerwise control, is not included in this panel.

The picture across the three is mixed rather than uniform. Coding rises (mean $+7.3$), long-context retrieval is roughly flat (mean -1.1), and advanced reasoning declines mildly (mean -2.7, worst -7.4 on Qwen2.5-VL). What separates them from the format-sensitive cluster is the comparison on the same five pairs, where the four format-sensitive means run -5.6 to -8.5 pt. None of these three benchmarks separates by per-head QK-RMSNORM, unlike the format-sensitive four. The loss therefore tracks the demand for controlled output formatting rather than the task's domain or difficulty alone, which is what the mechanism of Section 4 predicts.

	Qwen3-VL	InternVL3.5	Qwen2.5-VL	InternVL3	LLaVA-OV	Molmo2-O
length_constraints	9	8	17	16	7	**20**
detectable_format	8	5	10	8	20	**30**
keywords	9	6	13	18	4	**19**
change_case	5	8	11	**17**	16	8
combination	14	-	6	13	7	**22**
startend	7	4	4	7	12	**18**
language	5	-	7	5	8	**10**
punctuation	3	1	**28**	21	8	3
detectable_content	3	3	3	5	1	**11**
Total	**56**	**43**	**91**	**96**	**83**	**126**

Table 7: Regressed IFEval prompts, counted per violated instruction category, per pair. A prompt that violates several categories is counted in each, and a prompt whose violations fall outside the listed categories is counted in none, so the columns need not sum to the Total row. The Total row is the regressed-prompt count and matches the *Lost* column of Table 2. "$-$" marks absence; **bold** marks the row-max.

Pair	QK	MBPP	MMLU-Pro	RULER
Qwen3-VL	per-head	+7.6	−3.4	−0.2
InternVL3.5	per-head	−2.4	+0.6	−0.4
Qwen2.5-VL	none	+14.2	−7.4	+0.3
InternVL3	none	+21.4	−0.2	+0.6
LLaVA-OV	none	−4.4	−3.2	−6.0
Mean		**+7.3**	**−2.7**	**−1.1**

Table 8: Extended capability panel: per-pair Δ (VLM-LM − reference LLM, pp) on coding (MBPP, execution pass@1), advanced multi-domain reasoning (MMLU-Pro), and long-context retrieval (RULER), across the five headline pairs; Molmo2-O excluded. Same instruct protocol; per-pair QK category as in Table 1.

D Mechanism Details

This appendix expands the mechanism analyses of Section 4: the calibration-set size choice (Appendix D.1), the two layerwise controls (Appendix D.2), channel-level ablation (Appendix D.3), and a random-direction perturbation control (Appendix D.4).

D.1 Calibration Sample-Size Robustness

The $(G_{\text{base}}, B_{\text{gap}})$ measurement of Section 4 and the Sink Strength S of Section 3 both use a default of $N = 15$ calibration prompts. We verify this choice is converged by sweeping $N \in \{5, 10, 15, 25, 50, 100\}$ with 3 random seeds per cell, drawing each subset uniformly from a 100-prompt calibration pool disjoint from the IFEval test set. At $N = 100$ every seed draws the entire pool, so the seed scatter vanishes there by construction. This produces 18 measurements per pair across the 6 pairs, 108 filled cells in total. The hyperparameter is measurement-time only: it never trains a model and never sees benchmark labels, so it controls only the variance of the estimator.

Observations. (i) *Plateau from $N \approx 10$.* Seed-mean shifts are ≤ 0.05 in G_{base}, ≤ 0.05 in B_{gap}, and ≤ 0.06 in $G-B$ between $N=10$ and $N=100$ for every pair; by $N = 15$ the curves are indistinguishable from the $N = 100$ asymptote within seed scatter. (ii) *Seed scatter contracts as expected.* $\sigma(G - B)$ drops from ~ 0.15 at $N=5$ (worst case, InternVL3.5) to ≤ 0.07 by $N = 15$ and ≤ 0.02 by $N = 50$. (iii) *Pair ordering invariant.* Across all 108 (pair, N, seed) cells, the sign of $G - B$ is preserved (positive only for Qwen3-VL, negative for the other five), and the rank ordering of pairs on $G-B$ does not flip at any cell. The behavioral split and the magnitude ordering used in Section 4 therefore do not depend on the calibration set choice.

Sink Strength S. The sweep above tracks the margin $G_{\text{base}} - B_{\text{gap}}$. The predictor of Section 3 uses S alone, which is less stable under small N, so we report it separately. Moving from $N = 15$ to $N = 5$ shifts S on the six pairs by at most 0.11 (mean 0.09), against the ≤ 0.06 margin shift above. The predictor is nevertheless unchanged, with $\rho(S, \Delta_{\text{IFEval}}) = 0.971$ at both sizes and no reordering of the six pairs. Since $N=5$ also runs about $3\times$ faster (~ 3 s versus ~ 8 s per 7B backbone on an A6000), it reproduces the headline predictor at a third of the cost. We keep $N=15$ as the default because the extra prompts are inexpensive and buy a margin against seed variance on broader panels, where the smaller set is not a safe drop-in replacement.

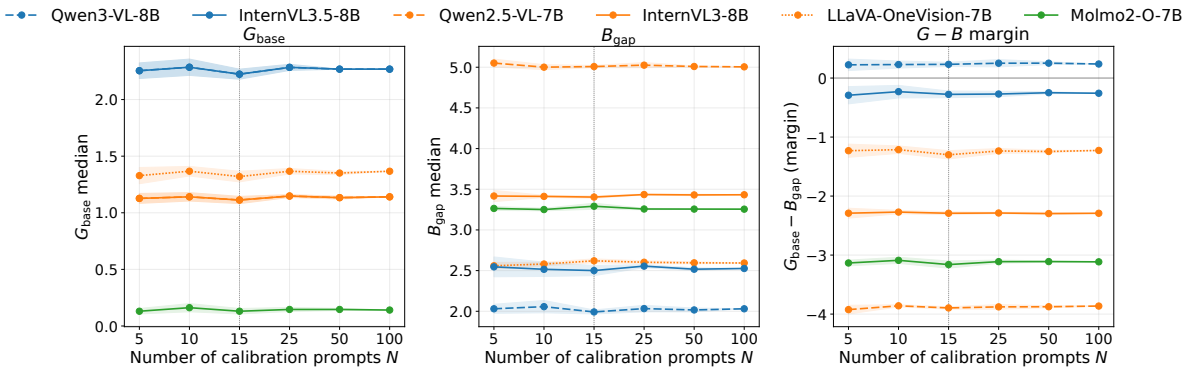

Figure 4: G_{base}, B_{gap}, and the $G_{\text{base}} - B_{\text{gap}}$ margin versus calibration size N for each of the six pairs. Lines are seed means, shaded bands are $\pm\sigma$ across 3 seeds. The vertical dashed line marks the paper-default $N = 15$; the horizontal line in the right panel marks the $G - B = 0$ sign-flip boundary.

Pair	$N=5$	$N=10$	$N=\mathbf{15}$	$N=25$	$N=50$	$N=100$
Qwen2.5-VL-7B	-3.92 ± 0.07	-3.86 ± 0.02	$\mathbf{-3.90 \pm 0.02}$	-3.88 ± 0.04	-3.88 ± 0.02	-3.86 ± 0.00
InternVL3-8B	-2.29 ± 0.08	-2.27 ± 0.03	$\mathbf{-2.29 \pm 0.03}$	-2.29 ± 0.01	-2.30 ± 0.02	-2.29 ± 0.00
LLaVA-OneVision-7B	-1.23 ± 0.11	-1.21 ± 0.06	$\mathbf{-1.30 \pm 0.07}$	-1.24 ± 0.04	-1.25 ± 0.02	-1.23 ± 0.00
Qwen3-VL-8B	$+0.22 \pm 0.10$	$+0.23 \pm 0.06$	$\mathbf{+0.23 \pm 0.04}$	$+0.25 \pm 0.06$	$+0.25 \pm 0.02$	$+0.24 \pm 0.00$
InternVL3.5-8B	-0.29 ± 0.15	-0.23 ± 0.11	$\mathbf{-0.28 \pm 0.06}$	-0.27 ± 0.05	-0.25 ± 0.02	-0.26 ± 0.00
Molmo2-O-7B	-3.13 ± 0.05	-3.09 ± 0.05	$\mathbf{-3.16 \pm 0.06}$	-3.11 ± 0.03	-3.11 ± 0.01	-3.11 ± 0.00

Table 9: $G_{\text{base}} - B_{\text{gap}}$ margin ($\mu \pm \sigma$ across 3 seeds) over the full N grid; the paper-default $N = 15$ column is bold. Each cell averages 3 random N-prompt subsets drawn from the 100-prompt sweep pool, which is separate from the fixed 15-prompt calibration set used in Table 5 and throughout the paper. This is why the $N = 15$ column differs slightly from that table. The pair ordering is identical in both.

D.2 Layerwise Backbones: Aggregate vs. Per-Head Sink

Molmo2-O-7B is built on Olmo-3, which applies QK-RMSNORM *layerwise*: at each layer, RMS is computed across the full $H \cdot d_h$-dimensional concatenated Q (and K) vector before γ_q, γ_k are applied, rather than across each head's d_h-dimensional slice independently (as in Qwen3 / InternVL3.5). At the model-mean level the layerwise scale preserves a strong sink: position-0 attention is 0.250 (95% CI $[0.242, 0.258]$), over $8\times$ larger than on the no-QK-RMSNORM Qwen2.5-VL backbone (0.030). At the per-head level the distribution is different. On the same calibration set (15-prompt, top-$\sim$10 late layers), median G_{base} across (layer, head, query position) is $+0.24$ nats (the lowest among the six pairs), strong-sink prevalence ($G_{\text{base}} > 2$, equivalently $a_{\text{sink}} > 0.881$) over non-degenerate late-layer (prompt, layer, head, query) observations is 8.5%, against 58% on the per-head QK-RMSNORM backbone (Qwen3-8B). The per-head margin protection of Lemma 1 is therefore not delivered head-by-head, and IFEval drops by 18.7 pt, GSM8K-CoT by 42.7 pt, EQ-Bench by 64.6 pt, GPQA-Diamond-CoT by 19.7 pt against the Olmo-3-7B-Instruct reference (the GPQA figure is measured against the published Olmo-3 score; our own matched run gives 10.6 pt). The reading is that aggregate model-level sink mass is not sufficient and that the evidence points to per-head sink strength as the relevant quantity. We therefore exclude Molmo2-O from the headline 5-pair comparison and report it separately as a distinguishing control.

A second layerwise pair. MolmoE-1B, built on OLMoE-1B-7B-0924 and scored here against the released OLMoE-1B-7B-0924-Instruct, applies QK-RMSNORM across all heads at once on an MoE backbone and is a second natural layerwise case. It reaches $S = +1.34$ with $G_{\text{base}} - B_{\text{gap}} = -2.84$ and an IFEval drop of 9.4 pt, so it follows the same pattern as Molmo2-O ($S = +0.24$, $G_{\text{base}} - B_{\text{gap}} = -3.01$, 18.7 pt) with the lower-S pair taking the larger drop. Both layerwise pairs sit well below the per-head pairs, which share $S = +2.36$, and both carry strongly negative margins while differing in base family and in dense versus MoE architecture. The two are not indepen-

dent of recipe, since both come from the Molmo and PixMo line, so the recipe-independent derivation of Appendix A.5 is what carries the argument and these two pairs stand as convergent empirical support rather than as a controlled comparison. Neither is a headline pair, and both are reported here as layerwise controls.

D.3 Channel Ablation Sweep

For each of the three normalization-scale tensors in Qwen3-VL-8B's language backbone (`input_layernorm` γ_{in}, `q_norm` γ_q, `k_norm` γ_k), we identify the top-K channels by $|\gamma|$ at each layer and replace those entries with the layer-wise channel mean. This disables the amplifier while leaving the residual stream and projections untouched. We fix $K = 10$ for all three tensors, so the ablation targets the largest-magnitude amplifier channels at each layer. Joint kill (all three norms simultaneously) costs 44.36 pt on IFEval, against a sum of individual costs of 19.96 pt, a 24.40-pt super-additive interaction.

D.4 Random-ΔW Control

We add Gaussian noise $G_W \sim \mathcal{N}(0, \sigma_W^2 I)$ to each attention / MLP projection W of Qwen2.5-7B-Instruct with σ_W chosen so that the per-sub-module relative Frobenius norm $\|G_W\|_F / \|W\|_F$ exactly matches the corresponding ratio observed between Qwen2.5-VL-7B's extracted LM and Qwen2.5-7B-Instruct. The γ parameters of every RMSNorm are left untouched. The result is a model whose per-submodule relative Frobenius distances from the reference LLM match those observed for Qwen2.5-VL, but whose perturbation direction is random. IFEval under the instruct protocol drops to 10.72 ($\Delta = -61.37$ against the reference LLM at 72.09), compared with the observed Qwen2.5-VL gap of -9.6 pt.

E Sink Strength Predictor

E.1 Per-Pair Sink Strength Predictions

Per-pair numerical detail behind Figure 3 (Section 3): the reference-LLM Sink Strength S, the observed VLM–reference IFEval gap, and the leave-one-pair-out linear prediction.

Pair	QK	S	Actual Δ	Held-out pred Δ
Qwen3-VL	per-head	$+2.36$	-5.4	-0.92
InternVL3.5	per-head	$+2.36$	-1.8	-3.41
Qwen2.5-VL	none	$+1.18$	-9.6	-10.78
InternVL3	none	$+1.18$	-8.9	-10.95
LLaVA-OV	none	$+1.44$	-7.9	-9.04
Molmo2-O	layerwise	$+0.24$	-18.7	-13.74

Table 10: Per-pair Sink Strength S measured on the reference LLM, the observed IFEval gap between the VLM-LM and reference LLM, and the leave-one-pair-out linear prediction.

E.2 IFEval-Specific Predictor (17 pairs)

The IFEval-specific predictor panel (Table 11) measures Sink Strength S against the VLM–reference IFEval prompt-strict Δ on an extended 17-pair set: the 6 pairs plus 11 additional VLMs that span MoE architectures (Qwen3-VL-30B-A3B-Instruct, MolmoE-1B), four reference-LLM families (Qwen, Olmo, Mistral, Llama-3.1) plus the MoE OLMoE, and backbone sizes 1.5B–32B (full pair-to-reference-LLM mapping with Hugging Face URLs in Appendix B.4). The panel therefore covers MoE and dense architectures, four reference-LLM families, and 1.5B–32B scale. Over it, Spearman $\rho(S, \Delta_{\text{IFEval}}) = +0.88$. Two caveats bound what a single rank correlation can carry here. Pairs that share a reference LLM share S by construction, so the 17 pairs supply 12 distinct S values, and 13 of the 17 are Qwen-family, so the four non-Qwen backbones are single points rather than a family-level sample.

E.3 Multi-Task Generalization (9 pairs)

Where Appendix E.2 widens the model set at a fixed task, this panel widens the task set instead. It adds three sub-10B Qwen-family pairs to the six (Qwen3-VL-4B, Qwen3-VL-2B, InternVL3-2B), which extends the panel down to 1.5B reference LLMs while keeping all four format-sensitive tasks on the unified protocol. Table 12 reports per-pair Sink Strength S and the four VLM–reference deltas. The IFEval correlation is $+0.945$ against $+0.971$ on the six, and the other three tasks fall between $+0.82$ and $+0.88$.

E.4 Sink Strength Design Ablations

Section 3 defines S as a median taken over the last ~ 10 decoder layers. This subsection isolates the three choices that definition makes.

Why the late layers. We split each reference LLM's decoder into equal early, middle, and late thirds and recompute S within each band (Table 13). Sink concentration and its correlation with the outcome rise together with depth. The early

Pair	Reference LLM	QK class	S	ΔIFEval
Qwen2.5-VL	Qwen2.5-7B-Instruct	none	$+1.18$	-9.6
Qwen3-VL-8B	Qwen3-8B	per-head	$+2.36$	-5.4
InternVL3-8B	Qwen2.5-7B-Instruct	none	$+1.18$	-8.9
InternVL3.5-8B	Qwen3-8B	per-head	$+2.36$	-1.8
LLaVA-OV-7B	Qwen2-7B-Instruct	none	$+1.44$	-7.9
Molmo2-O-7B	Olmo-3-7B-Instruct	layerwise	$+0.24$	-18.7
Qwen3-VL-4B	Qwen3-4B	per-head	$+2.36$	-5.0
Qwen3-VL-2B	Qwen3-1.7B	per-head	$+1.84$	-5.2
Ovis2-8B	Qwen2.5-7B-Instruct	none	$+1.18$	-17.9
Ovis2-34B	Qwen2.5-32B-Instruct	none	$+1.39$	-6.7
MolmoE-1B	OLMoE-1B-7B-0924-Instruct	layerwise (MoE)	$+1.34$	-9.4
InternVL3-2B	Qwen2.5-1.5B-Instruct	none	$+0.90$	-12.2
InternVL3-14B	Qwen2.5-14B-Instruct	none	$+1.30$	-5.2
MiniCPM-V-4.5	Qwen3-8B	per-head	$+2.36$	-4.4
LLaVA-NeXT-Mistral-7B	Mistral-7B-Instruct-v0.2	none	-0.19	-10.9
Idefics3-8B	Llama-3.1-8B-Instruct	none	-0.21	-22.7
Qwen3-VL-30B-A3B	Qwen3-30B-A3B-Instruct-2507	per-head (MoE)	$+1.66$	$+1.7$
$\rho_{\text{Spearman}}(S, \Delta_{\text{IFEval}})$ on 17 pairs				$+0.88$

Table 11: IFEval-specific 17-pair predictor panel. Sink Strength S measured on the reference LLM versus the VLM-LM–reference-LLM IFEval prompt-strict Δ. The panel spans dense and MoE architectures, all three QK-norm classes (per-head / layerwise / none), four reference-LLM families (Qwen, Olmo, Mistral, Llama-3.1) plus the MoE OLMoE, and backbone sizes 1.5B–32B. Hugging Face URLs and paper references for every pair are in Appendix B.4.

third carries almost no signal, while the late third alone reproduces the full predictor at $\rho = 0.97$. The late third overlaps the last-$\sim$10-layer default for every headline pair, so the emphasis on late layers reflects where the sink lives rather than a tuned window.

Sensitivity to the window size. Recomputing S over the last K layers for $K \in \{3, 5, 10, 15, 20\}$ leaves the rank correlation unchanged at $\rho = 0.971$ for every K. The magnitudes of S shift with K, but no pairwise ordering reverses, so the last-$\sim$10-layer window is a robust choice rather than a tuned one.

Why the median. The statistic should report whether the *typical* head forms a sink, which is what a median over (ℓ, h, q) measures, whereas a mean is inflated by a few strong heads. Substituting the mean preserves both the sign and the pair ordering, so this choice changes none of the conclusions drawn from S. Appendix D.2 is the case that motivates it, where a strong model-level aggregate coexists with a weak per-head distribution that only the per-head statistic exposes.

F Controls

This appendix details the modality comparison of Section 3.1 and the controls of Section 5: a text-only Qwen2.5-family endpoint (Appendix F.1), a step-wise VL versus text-only trajectory from a shared 3B base (Appendix F.2), a post-pretraining QK-RMSNORM injection experiment (Appendix F.3), and a weight-merging sweep (Appendix F.4).

F.1 Modality Comparison: Text-Only Endpoint

The Section 3.1 modality comparison (Table 3) uses Qwen2.5-7B-Instruct-1M, the Qwen team's official long-context text-only release, as a Qwen2.5-family further-training endpoint (Hugging Face URL in Appendix B.4).

Both this model and Qwen2.5-VL-7B-Instruct are evaluated against the same Qwen2.5-7B-Instruct text-only reference. The text-only variant is evaluated under the same 0-shot instruct protocol as the rest of the paper (Appendix B.1), with batch size 8 on a single A6000.

F.2 Modality Control: VL vs. Text-Only Trajectory

The Appendix F.1 comparison contrasts released 7B endpoints against a common text-only reference. This subsection instead tracks the two modalities step by step on the 3B base of the injection experiment, so the sink can be observed while it changes. Starting from `Qwen2.5-3B-Instruct`, we run vision fine-tuning on image-text instruction data (the Stage-2 recipe of Appendix F.3) and a matched text-

Pair	Reference LLM	QK	S	ΔIFEval	ΔGSM8K	ΔEQ-Bench	ΔGPQA-D
Qwen2.5-VL	Qwen2.5-7B-Instruct	none	+1.18	**−9.6**	−14.3	−11.2	−9.8
Qwen3-VL-8B	Qwen3-8B	per-head	+2.36	**−5.4**	−2.1	−2.4	−1.9
InternVL3-8B	Qwen2.5-7B-Instruct	none	+1.18	**−8.9**	−12.5	−10.3	−7.1
InternVL3.5-8B	Qwen3-8B	per-head	+2.36	**−1.8**	−1.5	−0.8	−0.6
LLaVA-OV-7B	Qwen2-7B-Instruct	none	+1.44	**−7.9**	−11.9	−8.6	−8.4
Molmo2-O-7B	Olmo-3-7B-Instruct	layerwise	+0.24	**−18.7**	−42.7	−64.6	−19.7
Qwen3-VL-4B⋆	Qwen3-4B	per-head	+2.36	**−5.0**	+13.9	+1.4	−2.0
Qwen3-VL-2B⋆	Qwen3-1.7B	per-head	+1.84	**−5.2**	+27.5	−0.7	−6.1
InternVL3-2B⋆	Qwen2.5-1.5B-Instruct	none	+0.90	**−12.2**	−18.0	−9.0	−6.1
$\rho_{\text{Spearman}}(S, \Delta)$ on 9 pairs				+0.945	+0.877	+0.860	+0.821

Table 12: Extended 9-pair predictor panel. Δ denotes VLM-LM minus the corresponding reference LLM (negative = lower performance relative to the reference) on IFEval prompt-strict, GSM8K-CoT flexible-extract, EQ-Bench v2.1, and GPQA-Diamond-CoT flexible-extract; all evaluated under the unified protocol (Appendix B.1), except that the Molmo2-O GPQA baseline is the published Olmo-3-7B-Instruct score rather than our own matched run. ⋆ marks the three pairs added beyond the six, which span $S \in [0.90, 2.36]$ and reference-LLM sizes 1.5B–4B. Bold ΔIFEval column is the headline metric.

Third	S mean	$\rho(S, \Delta_{\text{IFEval}})$
Early	+0.20	+0.09
Middle	+0.91	+0.79
Late	+1.56	+0.97

Table 13: Sink Strength recomputed within equal thirds of the decoder, on the six pairs. Both the magnitude of S and its rank correlation with the VLM–reference IFEval gap increase with depth.

only run on Tülu general instruction SFT (Lambert et al., 2025), at the same effective batch size, learning rate, and number of steps, matching the optimization settings while varying the training modality and data. At each checkpoint we measure Sink Strength S on the trained backbone, vision understanding on SEEDBench-IMG (Li et al., 2023), and IFEval prompt-strict.

Table 14 reports the trajectories. The vision run gains vision understanding while its sink collapses, with SEEDBench-IMG rising from 60.1 to 64.8 and S falling from +0.41 at the shared base to +0.20 at step 10k, and IFEval sitting in the 34–44 band throughout. The text-only run moves the sink in the opposite direction, holding S between +0.58 and +0.65, and gives up about 5 pt of IFEval. The step-0 IFEval of 59.9 reproduces the published `Qwen2.5-3B-Instruct` score of 58.2 (Qwen Team, 2024) to within 1.7 pt, so the drops start from a correctly measured baseline. Within the vision run the two curves are not step-by-step aligned. IFEval takes most of its loss by step 1k, before S has fallen, and then recovers slightly while S continues to decline, so S tracks the regime the run ends in rather than the step-level fluctuation.

	VL adaptation			Tülu (text-only)	
Step	S	SEED	IFEval	S	IFEval
0	+0.41	–	59.9	+0.41	59.9
1k	+0.54	60.1	36.8	+0.65	53.1
2k	+0.49	60.9	37.2	+0.58	53.6
3k	+0.35	62.1	34.2	+0.61	55.3
4k	+0.30	63.0	40.5	+0.60	56.9
5k	+0.24	63.9	39.7	+0.58	54.3
6k	+0.22	63.9	38.6	+0.58	55.1
7k	+0.19	64.3	42.3	+0.58	55.3
8k	+0.24	64.7	40.7	+0.58	53.8
9k	+0.20	65.0	43.1	+0.58	54.9
10k	+0.20	64.8	43.6	+0.58	55.8

Table 14: Step-wise modality control on `Qwen2.5-3B-Instruct`. S is Sink Strength on the trained backbone, SEED is SEEDBench-IMG overall accuracy, and IFEval is prompt-strict. Both runs start from the same base and use the same batch size, learning rate, and step count, while differing in training modality and data.

F.3 Injection Experiment Training Recipe

The injection experiment trains two 3B VLMs from the same Qwen2.5-3B-Instruct base under the same LLaVA Stage-1 / Stage-2 recipe. The two variants are *vanilla* (no QK-RMSNORM) and *QK-RMSNorm injected* with unit-scale initialization $\gamma_q = \gamma_k = 1$. Vision tower and projector come from Qwen2.5-VL-3B-Instruct. Both runs proceed on dedicated 2× NVIDIA H100 nodes with DeepSpeed ZeRO-2. Total walltime is ∼ 47 h per variant (∼ 7 h Stage 1, ∼ 40 h Stage 2).

Hyperparameter	Stage 1 (align)	Stage 2 (instruct)
Trainable params	projector only	full LM + projector
Dataset	LLaVA-Pretrain (558K)	LLaVA-Mix (665K)
Image resolution	336^2 (112,896 px)	672^2 (451,584 px)
Per-device batch	32	4
Grad accumulation	2 (NGPU= 2)	16
Effective batch	128	128
Optimizer	AdamW	AdamW
Learning rate	1×10^{-4}	2×10^{-5}
LR schedule	const + warmup	cosine
Warmup ratio	0.03	0.03
Weight decay	0.0	0.0
Max grad norm	1.0	1.0
Epochs	1	1
Steps	4361	$\sim$ 5195
Precision	bf16	bf16
DeepSpeed	ZeRO-2	ZeRO-2
Save every	end of stage	500 steps
Save total limit	—	2

Table 15: Training recipe for the post-pretraining QK-RMSNORM injection experiment (vanilla and injected variants are matched on all rows; the QK-RMSNORM module is the only difference).

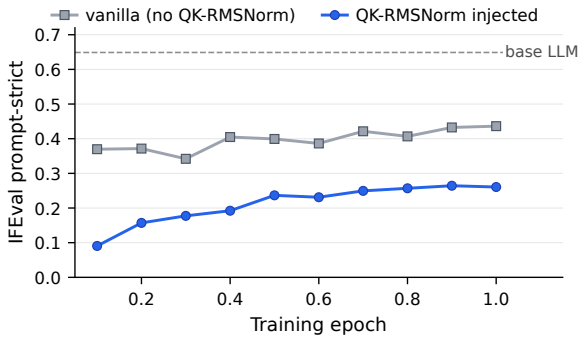

Figure 5: IFEval prompt-strict over the LLaVA Stage-2 training epoch for the controlled post-pretraining QK-RMSNORM injection pair (vanilla vs. injected, matched in all other settings).

F.4 Weight-Merging Sweep on Format-Sensitive Benches

The body reports IFEval on the two pairs (Qwen2.5-VL, Qwen3-VL) for seven weight-merging settings: the asymmetric 95% VLM + 5% LLM mix, uniform averaging, full Task Arithmetic (λ=1), Task Singular Vectors (α=1), STAR (η=40%), TIES (density 0.2, λ=1), and DARE (mask rate 0.9, λ=1). We also ran the same settings on EQ-Bench, GSM8K-CoT flexible-extract, and GPQA-Diamond-CoT flexible-extract, and summarize the pattern here rather than reproducing three further figures. It matches IFEval, in that the mild mixes (95/5, uniform averaging) avoid substantial additional degradation, full-strength operators (TA, TSV, STAR) drop the non-protected pair by tens of points, and sparsified merges (TIES, DARE) are the most damaging across both pairs.

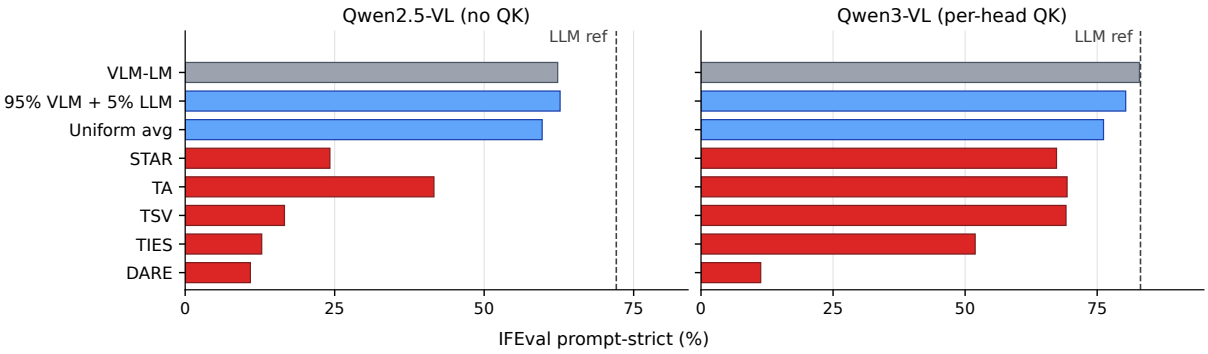

Figure 6: IFEval prompt-strict for Qwen2.5-VL (non-protected) and Qwen3-VL (per-head QK-RMSNORM) under seven weight-merging settings. Dashed line is each pair's reference LLM.