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

Anonymous ACL submission

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—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 determines 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. Its rank ordering is preserved across the six VLM–LLM pairs we study and across four format-sensitive tasks. Complementing this diagnostic, we find that neither a controlled post-hoc architectural injection nor several off-the-shelf weight-merging settings recover the lost capability. This indicates that the damage is hard to repair after adaptation and underscores the value of screening backbones with Sink Strength *before* VL training; the repair search itself narrows toward head-selective training-time interventions.

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 VLM–LLM pairs we study, the post-VL loss reaches double digits on multiple such tasks, while text-only further fine-tuning of the same base LLM produces little to no loss (Section 3.1). Existing remedies—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), or post-hoc neuron re-merging (Yu and Ananiadou, 2025)—each help in part but at a cost, and none explains *why* the loss happens.

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., 2026; Venkateswaran and Contractor, 2026). We show that VL fine-tuning perturbs the projections that read from these channels, collapsing the per-head sink. 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 mitigate the attention-sink corruption caused by fine-tuning (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. 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

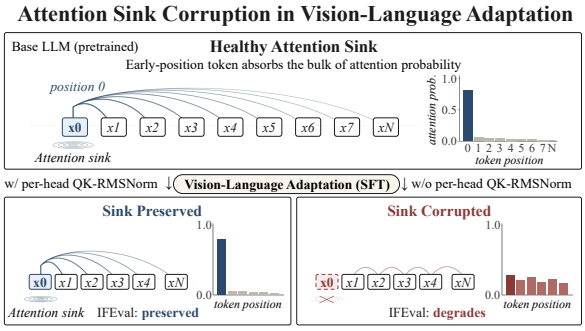

Figure 1: Attention-sink corruption in vision-language fine-tuning: the base LLM concentrates attention on an early sink position; whether this concentration survives VL fine-tuning varies sharply across language backbones.

and backbones with a collapsed sink form non-overlapping groups in our sample.

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 same ordering holds across multiple format-sensitive tasks. Because S is measured before VL training, it also separates damage already present in the base LLM from damage caused by VL fine-tuning, both of which contribute to the post-VL loss.

Given a high-risk backbone, can the damage be repaired after the fact? We test two repair attempts as negative controls (Section 5): post-hoc injection of the QK-RMSNORM module into a matched base LLM followed by retraining, and several off-the-shelf weight-merging methods spanning linear, spectral, and sparsified families (Ilharco et al., 2022; Yadav et al., 2023; Yu et al., 2024; Gargiulo et al., 2025; Lee et al., 2025b; Wortsman et al., 2022). We observe that neither recovers the lost capability on the affected backbones, narrowing the repair search toward head-selective training-time interventions.

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 ordering preserved across our six VLM–LLM pairs and four tasks), providing a pre-VL screen; (ii) a **sink-anchored mechanism** (Theorem 1) explaining why per-head QK-RMSNORM on the query and key projections protects sink survival, validated by per-head margin calibration across six pairs and a non-Qwen layerwise-QK-RMSNORM control (Molmo2-O) confirming that module presence alone does not imply per-head sink preservation; (iii) **confound and negative controls**: an LM-only SFT modality control isolating the loss to VL adaptation, and failed repair attempts (post-hoc QK-RMSNORM injection, weight-merging) that narrow the repair search toward head-selective training-time interventions.

2 Background & Related Work

VL fine-tuning and text-side erosion. Vision-language models (VLMs) extend a pretrained large language model (LLM) with a vision encoder and projector, then fine-tune on multimodal supervised data (Liu et al., 2023, 2024b). Standard recipes 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., 2024), 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, 2024b)), or post-hoc (neuron-level re-merging (Yu and Ananiadou, 2025)), each closing part of the gap. We leave the fine-tune protocol untouched and instead ask 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 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

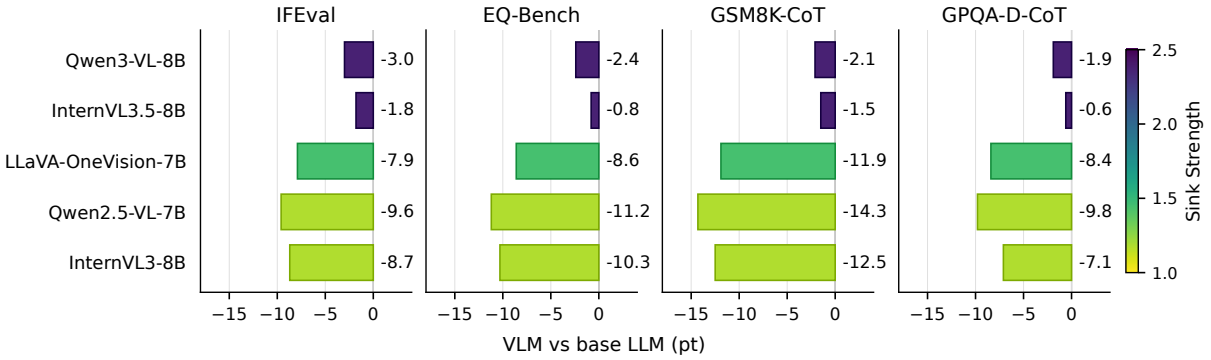

Figure 2: Post-VL adaptation loss across our headline VLM–LLM pairs on instruction-following and three further format-sensitive tasks. Color: Sink Strength S measured on the base LLM (Section 3).

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 et al., 2025), Llama-3 (Grattafiori et al., 2024), DeepSeek-V3 (Liu et al., 2024a), and the more recent Qwen3.5), an architectural axis our Section 3.1 analysis exploits. 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

Via a matched LM-only SFT control, we confirm that the post fine-tuning loss is specific to VL across six pairs and four format-sensitive tasks.

Setup. The five headline pairs come from three vendors: Qwen, OpenGVLab, and LLaVA-OneVision. Full per-pair metadata is in Table 1.[1] All pairs are full fine-tunes, and we exclude LoRA-trained VLMs to remove the adapter-merge confound. For each pair the VLM's language backbone is extracted as a standalone causal LM (the VLM-LM hereafter; Appendix B.2, with the round-trip preserving logits on a held-out probe) and scored on an eight-task text suite (Appendix B.1) alongside the base LLM.

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

Pair	Base LLM	Recipe	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+DPO (PixMo)	*layerwise*

Table 1: Five headline VLM–LLM pairs and the Molmo2-O control. "QK" column reports the per-head / layerwise / none QK-RMSNORM label.

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 all four gates are binary pass/fail, 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" 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 Loss. Figure 2 (IFEval row) shows the post-VL loss splits cleanly across our five headline pairs: two backbones lose at most 3 pt (InternVL3.5 at -1.8, Qwen3-VL at -3.0), while three lose at least 7.9 pt (LLaVA-OV -7.9, InternVL3 -8.7, Qwen2.5-VL -9.6). The worst-case loss in the first group is smaller than the best-

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	**1.9**
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	8.0
Molmo2-O	layerwise	317	**126**	25	73	**18.7**

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

Bench	LM-SFT 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: Modality control. Variants start from `Qwen2.5-7B-Instruct`: LM-SFT is a long-context post-training (`Qwen2.5-7B-Instruct-1M`, text-only), VL is `Qwen2.5-VL-7B-Instruct` (image-text).

case loss in the second, and no pair falls in the gap. The non-headline pair (Molmo2-O on Olmo-3) loses −18.7 pt and is analyzed separately as a control (Section 4, Appendix D.2).

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 two lower-loss pairs net 1.9–5.4% regression while the three higher-loss pairs net 8.0–9.6%, roughly 1.5–3× higher. Molmo2-O is worst at 18.7%. The regression-rate ordering matches the IFEval delta ordering pair-by-pair (Spearman $\rho = 1.0$), so the behavioral split is not driven by a few outlier prompts.

VL-specificity of the loss. A natural reading is that this loss reflects further fine-tuning in general rather than VL adaptation specifically. We rule this out with a matched LM-only SFT control on the same Qwen2.5-7B-Instruct base. Under the same instruct protocol, the text-only SFT variant (`Qwen2.5-7B-Instruct-1M`) matches the base on IFEval prompt-strict while the VL pair (`Qwen2.5-VL-7B-Instruct`) loses 9.6 pt; across the four format-sensitive tasks the LM-SFT losses are at most 0.4× the corresponding VL losses (Table 3; per-pair experimental details in Appendix F.1). The loss is therefore tied to VL adaptation, not to downstream fine-tuning per se, framing the predictor question as: which base LLMs are vulnerable to VL-induced sink corruption, identifiable before VL training is committed?

3.2 Sink Strength

Through the introduction of Sink Strength S as a base-LLM scalar, we demonstrate that it captures the per-pair loss split, whose clustering is further aligned with per-head QK-RMSNORM in the base LLM.

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

$$S \;:=\; \mathrm{median}_{(\ell,h,q)}\; \log \frac{a_{p_0}^{(\ell,h,q)}}{1 - a_{p_0}^{(\ell,h,q)}},$$

where $a_{p_0}^{(\ell,h,q)}$ is the attention probability at the per-head argmax key position p_0 at layer ℓ, head h, query position q, and the median is taken over the last $\sim$10 layers. 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 takes $\sim$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 is not meaningfully ahead of other positions. Section 4 formalizes why this scalar predicts post-VL damage via a sink-anchored perturbation bound.

Analysis. Color in Figure 2 indicates per-pair S measured on the base 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 base), 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 same ordering of S values reproduces the IFEval-drop ordering pair-by-pair 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

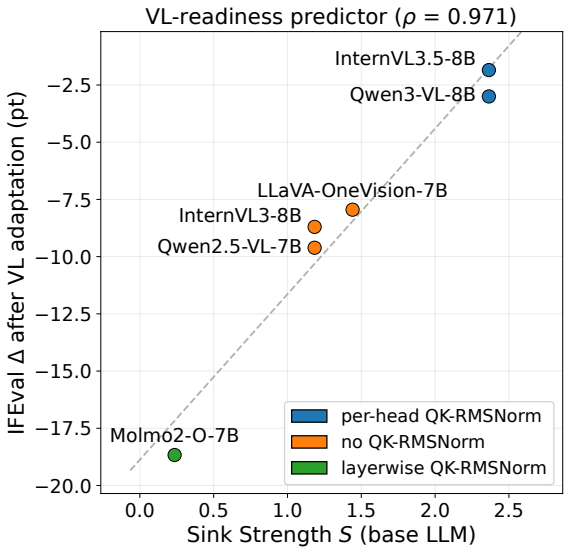

Figure 3: Sink Strength S measured on the base LLM versus IFEval drop after VL adaptation. Color: per-head QK-RMSNorm presence on the language backbone. Dashed line is the leave-one-pair-out linear fit.

answer parsing grade strict, parseable rules over surface tokens that require stable per-position attention (Zhang et al., 2026; 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).

Architectural correlate. The S clusters align with per-head QK-RMSNorm presence in the base 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 mistakenly flag Molmo2-O as safe; S correctly identifies it as the highest-risk backbone, matching the largest measured drop (-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 Qwen generation (Qwen3-VL vs. InternVL3.5, Qwen2.5-VL vs. InternVL3), the residual is ≤ 1.2 pt — small relative to the between-class gap of ≥ 4 pt and attributable to VL recipe variation outside S's scope.

Empirical validation. We use S in two modes: rank ordering across backbones requires no 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 ranks them by post-VL IFEval drop with Spearman $\rho = 0.97$; under leave-one-pair-out validation, the average held-out prediction error is 1.61 pt on a 17-pt outcome range, against 4.84 pt for a constant-mean baseline (per-pair detail in Appendix E.1). With six datapoints, this leave-one-pair-out check should be read as a sanity check that the linear relation is not driven by a single pair, not as prospective evidence on unseen backbones. Above $S \approx 1$ the linear fit keeps magnitude errors near the one-point scale; below this regime, S correctly flags Molmo2-O as the highest-risk backbone but the linear extrapolation underestimates the actual collapse magnitude (-18.7 pt actual vs. -14.68 pt leave-one-out). While Theorem 1 controls the full theoretical margin $G_{\text{base}} - B_{\text{gap}}$, in our sample S tracks the IFEval drop more strongly than the full margin ($\rho = 0.97$ vs. 0.89); we therefore use S as a leading-order, single-forward-pass approximation to the theoretical margin.

3.3 Generalization and Practical Use

Generalization across format-sensitive tasks. The same S, measured once on the base LLM, also ranks the backbones correctly 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 base-LLM S.

Predictor extension. We extend the predictor along two complementary axes. *(i) IFEval-specific scale.* On a 17-pair panel spanning MoE and dense architectures, four base-LLM families (Qwen, Olmo, Mistral, Llama-3.1) plus the MoE OLMoE, and backbone sizes 1.5B–38B, Sink Strength achieves Spearman $\rho_{\text{IFEval}} = \mathbf{0.92}$ (Appendix E.2). *(ii) Multi-task generalization.* On a 9-pair subset, rank ordering across all four format-sensitive tasks is preserved ($\rho \geq 0.82$, Appendix E.3).

	IFEval	EQ-Bench	GSM8K	GPQA-D
Spearman ρ, n=6	0.971	0.971	0.971	0.883
Error, 6 pairs (pt)	1.61	19.09	7.91	1.96
Error, 5 headline pairs (pt)	0.80	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.

Practical use. Given a candidate base LLM, S can be measured with 15 calibration prompts in under 10 seconds on a single GPU before any VL training is committed. A practitioner can use it to rank backbones by expected post-VL IFEval damage, or to set a threshold (e.g., $S \geq 1$) that screens for backbones expected to preserve format-sensitive capability within a few points after VL adaptation.

Scope: VL-conditional. S characterizes vulnerability to VL-induced attention perturbation specifically (Section 3.1); for recipe-sensitive post-hoc diagnosis, the full theoretical margin $G_{\text{base}} - B_{\text{gap}}$ (Section 4) provides a recipe-aware refinement of which S is the single-forward-pass approximation.

4 Mechanism

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 is $l_p = Q \cdot K_p/\sqrt{d_h}$ with $Q = W_q\xi_q$ and $K_p = W_k\xi_p$. Under QK-RMSNorm, $Q = \gamma_q \odot Q_{\text{pre}}/\text{RMS}(Q_{\text{pre}})$ and similarly K_p. We perturb only W_q, W_k with $\|\Delta W_*\|_{\text{op}} \leq \varepsilon$ and treat the γ's as frozen, consistent with the $\leq 3\%$ relative-norm drift observed on the natural VL adaptations we measure. Here $\|\gamma_*\|_\infty := \max_d |\gamma_*[d]|$ denotes the channelwise ℓ_∞ norm of a QK-RMSNorm scale vector. Let $l'_p = l_p + \delta_p$ be the post-perturbation logit, and let p_0 be the base LLM's per-head sink position (argmax key on the unperturbed forward pass). Two quantities anchored at p_0 drive the bound: the aggregate sink gap $G_{\text{base}} := \log\big(a_{p_0}/(1 - a_{p_0})\big)$ on the unperturbed logits, and the sink-anchored gap perturbation $B_{\text{gap}} := \sup_{p \neq p_0} |\delta_p - \delta_{p_0}|$ over the fixed key positions of the calibration prompt.

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 $G_{\text{base}} - B_{\text{gap}}$ (controls sink survival, Lemma 1), and Sink Strength S (Section 3, median of G_{base} over (ℓ, h, q)).

4.2 Perturbation and Saturation Bounds

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

Connection to Sink Strength. Section 3 introduced Sink Strength S as a single base-LLM scalar that predicts post-VL damage; this section formalizes *why*. Theorem 1 bounds the post-fine-tune logit-gap perturbation B_{gap}, and Lemma 1 converts the resulting margin $G_{\text{base}} - B_{\text{gap}}$ into an upper bound on the post-fine-tune non-sink mass; S is G_{base} aggregated by median over (ℓ, h, q) and is therefore a leading-order base-LLM proxy of this margin (Table 5). The bound's proof requires a mild non-degeneracy condition on the projection-RMS denominators, which we state formally as Assumption 1 below.

Assumption 1 (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 $\text{RMS}(W_*\xi) > 0$ for every ξ in the relevant set, and define $m_q := \min_{\xi \in \mathcal{S}_q} \text{RMS}(W_q\xi)/\|\xi\|_2$ and $m_k := \min_{\xi \in \mathcal{S}_k} \text{RMS}(W_k\xi)/\|\xi\|_2$. Then $m_q, m_k > 0$ in the finite calibration set we measure.

Theorem 1 (Logit-gap perturbation)**.** *Under Assumption 1, holding the base hidden states $\{\xi_q, \xi_p\}$ fixed at their LLM baseline values and perturbing only W_q, W_k at the single layer of interest by ΔW_* with $\|\Delta W_*\|_{\text{op}} \leq \varepsilon$ 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 $\gamma_q, \gamma_k, \gamma_{\text{in}}$, calibration inputs $\{\xi_q, \xi_p\}$, and RMS lower bounds*

m_q, m_k, No QK-RMSNORM:

$$B_{\text{gap}} \le c\,\varepsilon\, d_h^{-1/2} \|\xi_q\| \left(\|W_q\|_{\text{op}} + \|W_k\|_{\text{op}}\right) \cdot \sup_{p \ne p_0} \|\xi_{p_0} - \xi_p\| + O(\varepsilon^2). \quad (1)$$

With QK-RMSNORM:

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

The no-QK-RMSNORM bound carries the input-magnitude factor $\|\xi_q\|$ and a key-side contrast $\sup_{p \ne p_0} \|\xi_{p_0} - \xi_p\|$. The key-side contrast is sink-amplified because ξ_{p_0} carries the massive-activation channel (Sun et al., 2024) while ξ_p for $p \ne p_0$ does not. The query factor $\|\xi_q\|$ depends on the token at the query position regardless of whether the sink falls there. The QK-RMSNORM bound has no $\|\xi\|$ factor. The RMS denominator absorbs the input magnitude and leaves only the frozen amplifier $\|\gamma_q\|_\infty \|\gamma_k\|_\infty$ and the projection-RMS lower bounds m_q, m_k.

Remark 1 (Layerwise QK-RMSNORM extension; informal). In layerwise QK-RMSNORM, all H heads share a single RMS denominator over the concatenated $H \cdot d_h$-dimensional pre-norm vector $Q_{\text{cat}} = [Q_1; \cdots; Q_H]$ (the layer's γ_q is applied afterwards). The post-RMS, pre-γ head magnitude becomes $\sqrt{H d_h}\, \|Q_h\| / \|Q_{\text{cat}}\|$, instead of the per-head unit-RMS value $\sqrt{d_h}$; the same applies to keys. Heads with small query/key energy relative to the layer average therefore enter the γ-scaled attention logit with a weaker normalized per-head signal; adapting the proof of Theorem 1 to this setting introduces cross-head coupling through the shared RMS, so the per-head normalization mechanism need not deliver margin protection head-by-head. This is consistent with Molmo2-O's per-head G_{base} distribution (Appendix D.2).

Lemma 1 (Saturation transfer). *For sink position p_0 with pre-perturbation aggregate gap G_{base} and gap perturbation B_{gap},*

$$1 - a'_{p_0} \;\le\; \sigma\big(B_{\text{gap}} - G_{\text{base}}\big), \quad (3)$$

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.

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**

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, computed on a 15-prompt calibration set (nats).

4.3 Empirical Sink Behavior

We calibrate the bound on the six pairs, measure the cross-pair sink-mass distribution, and use the Molmo2-O control to distinguish aggregate from per-head sink.

Calibration across the six pairs. We compute $(G_{\text{base}}, B_{\text{gap}})$ across all five headline pairs and the Molmo2-O control on a 15-prompt calibration set (disjoint from the IFEval test set), aggregating over (layer, head, query position) in the last ~ 10 layers where sink behavior is established beyond the first few local-context layers (Xiao et al., 2024). G_{base} is computed on the base LLM at the per-head argmax key p_0, and B_{gap} is reconstructed from the change in attention-probability ratios between the LLM and the VLM-LM at the same p_0. Within each Qwen generation, the two pairs share the same base 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. Molmo2-O's layerwise QK-RMSNORM lands at the most-negative end of the sample, driven by the weakest base sink. The ordering tracks the behavioral split of Figure 2. Calibration sample-size sensitivity is reported in Appendix D.1.

Cross-pair sink and outcome bucketing. We further partition each VLM-LM's IFEval prompts by per-prompt strict accuracy (pass at ≥ 0.5, fail 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 ($N = 100$ per bucket, Appendix C.3). The within-pair pass−fail gap is small everywhere (≤ 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. Under the same OpenGVLab CascadeRL recipe, 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 ($0.03 \rightarrow 0.70$ across Qwen2.5-VL $\rightarrow$ Qwen3-VL). Within-vendor framing thus isolates the architectural change as the dominant factor over recipe or vendor variation. The within-pair near-equality is consistent with the broken-sink state being a head-level configuration set at fine-tuning time rather than a per-prompt event. The cross-pair span is what tracks the behavioral split, suggesting that effective interventions should operate at the head level rather than at the prompt 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 comparable to LLaVA-OV's (0.50, even higher) yet IFEval damage is far worse, while per-head sink is the weakest in our sample and already present in the Olmo-3 base before VL training (Appendix D.2). The marginal proxy $G - B$ is correspondingly the most negative (Table 5), so Lemma 1's per-head margin protection is not delivered head-by-head. The evidence isolates per-head QK-RMSNORM on the base LLM as the protective configuration, with the layerwise variant insufficient.

4.4 Channel and Direction Ablations

Ablating the sink amplifier. For each of Qwen3-VL's three QK-RMSNORM tensors ($\gamma_{\text{in}}, \gamma_q, \gamma_k$), replacing the top-10 channels per layer by the layerwise 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 redundancy synergy that confirms sink amplification as a redundant multi-norm system driving format-sensitive capability. For reference, an isotropic random perturbation of equal Frobenius norm produces a $\sim 6\times$ larger IFEval drop than the natural VL adaptation (Appendix D.4), so the observed VL damage is direction-specific.

5 Controls

Within-vendor recipe control. Looking at two within-vendor pairs (Qwen3-VL vs. Qwen2.5-VL; InternVL3.5 vs. InternVL3), each shares a similar VL recipe and pretraining lineage while differing in per-head QK-RMSNORM (Figure 2), 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.

Post-hoc QK-RMSNORM injection (negative). A natural first attempt at repairing a low-S backbone is to inject the missing module post-hoc. We add per-head QK-RMSNORM with unit-scale initialization to a clean Qwen2.5-3B-Instruct and run matched LLaVA Stage-1/Stage-2 training (Appendix F.2). The injected variant lags vanilla at every checkpoint, and the damage is already visible at the projector-only Stage-1 (IFEval 11 vs. base 65). Combined with the Molmo2-O control (Section 4), this points to native-pretraining co-adaptation of the projections, rather than the module's mere presence, as the protective factor.

Weight-space merging (negative). Seven off-the-shelf merging settings spanning linear task-vector mixes (Ilharco et al., 2022), 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 mildest mix stays within noise of the un-merged baseline (Appendix F.3). The repair search therefore narrows toward head-selective training-time interventions, reinforcing the value of pre-VL screening via Sink Strength.

6 Conclusion

Across 17 VLM–LLM pairs spanning MoE and four base-LLM families, post-VL adaptation loss on format-sensitive tasks tracks how well the base LLM preserves its attention sink, with per-head QK-RMSNORM as the dominant architectural correlate; a layerwise-QK-RMSNORM control (Molmo2-O on Olmo-3) shows that module presence alone is insufficient. Sink Strength S, a base-LLM scalar computed without any VL training, predicts post-VL IFEval damage with Spearman $\rho = 0.92$ across the 17 pairs and serves as a pre-VL screen. Two repair attempts (QK-RMSNORM injection and seven weight-merging settings) fail, narrowing the repair search to head-selective training-time interventions.

Limitations

To our knowledge, the phenomenon we surface is novel in the multimodal community: VL adaptation systematically erodes the base LLM's per-head attention sink, with downstream cost to format-sensitive capability. Within our sample, per-head QK-RMSNORM co-varies with vendor, recipe, and pretraining data; matched within-vendor comparisons (Qwen2.5→Qwen3, InternVL3→InternVL3.5) reduce but do not isolate the architectural axis.

Both repair directions we tested (post-hoc per-head QK-RMSNORM injection and seven off-the-shelf weight-merging settings) fail to close the gap on the low-S side; head-selective training-time interventions remain the central open direction, a more constrained search than the unsearched space we entered. A direct causal intervention (e.g., disabling per-head QK-RMSNORM during VL training of Qwen3 to test whether IFEval collapses, or imposing a sink-preservation regularizer mid-training on a low-S backbone to test whether the drop closes) would upgrade the present correlate-plus-bound account to a sufficiency claim; we leave this as future work.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 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, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. 2025. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 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*, volume 2024, pages 2632–2652.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 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 Rodola. 2025. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 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, and 1 others. 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*, volume 2025, pages 97114–97144.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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.

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

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

Seongyun Lee, Geewook Kim, Jiyeon Kim, Hyunji Lee, Hoyeon Chang, Sue Park, and Minjoon Seo. 2025a. How does vision-language adaptation impact the safety of vision language models? In *International Conference on Learning Representations*, volume 2025, pages 53643–53681.

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 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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*, pages 26689–26699.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 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, and 1 others. 2025. Olmoe: Open mixture-of-experts language models. In *International Conference on Learning Representations*, volume 2025, pages 62061–62121.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 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, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *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. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Shikhar Srivastava, Md Yousuf Harun, Robik Shrestha, and Christopher Kanan. 2024. Improving multimodal large language models using continual learning. *arXiv preprint arXiv:2410.19925*.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, and 1 others. 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 1 others.

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, and 1 others. 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. *arXiv preprint arXiv:2502.11812*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and 1 others. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, 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*, volume 2024, pages 21875–21895.

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, and 1 others. 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, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

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 *Forty-first International Conference on Machine Learning*.

Zeping Yu and Sophia Ananiadou. 2025. Locate-then-merge: Neuron-level parameter fusion for mitigating catastrophic forgetting in multimodal llms. *arXiv preprint arXiv:2505.16703*.

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 *International Conference on Machine Learning*, 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*, pages 202–227. PMLR.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. 2026. 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, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, and 1 others. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pages 18378–18394.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 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. 13
A.1. Single-Head Perturbation Bound (Theorem 1) . . . . . p. 13
A.2. Aggregate-Gap Saturation Transfer (Lemma 1) . . . . . p. 14
A.3. Frobenius / Stable-Rank Restatement . . . . . p. 14
A.4. Multi-Layer Composition Remark . . . . . p. 15
A.5. Numerical Sanity Checks . . . . . p. 15
B. Evaluation Setup . . . . . p. 16
B.1. Evaluation Protocols . . . . . p. 16
B.2. Backbone Extraction . . . . . p. 16
B.3. Implementation Details and Reproducibility . . . . . p. 16
B.4. Models and Datasets . . . . . p. 16
C. Per-Task and Behavioral Evidence . . . . . p. 18
C.1. Factual Recall vs. Format-Sensitive Tasks . . . . . p. 18
C.2. Failure-Category Breakdown . . . . . p. 18
C.3. Outcome-Bucketed Sink Probe . . . . . p. 19
D. Mechanism Details . . . . . p. 19
D.1. Calibration Sample-Size Robustness . . . . . p. 19
D.2. Molmo2-O: Aggregate vs. Per-Head Sink . . . . . p. 19
D.3. Channel Ablation Sweep . . . . . p. 20
D.4. Random-ΔW Control . . . . . p. 20
E. Sink Strength Predictor . . . . . p. 20
E.1. Per-Pair Sink Strength Predictions . . . . . p. 20
E.2. IFEval-Specific Predictor (17 pairs) . . . . . p. 21
E.3. Multi-Task Generalization (9 pairs) . . . . . p. 21
F. Controls . . . . . p. 21
F.1. Modality Control: LM-SFT Variant . . . . . p. 21
F.2. Injection Experiment Training Recipe . . . . . p. 21
F.3. Weight-Merging Sweep on Format-Sensitive Benches . . . . . p. 21

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). We omit RoPE from the notation, and since it is position-wise orthogonal the norm bounds below are preserved. The sink-anchored gap perturbation is $B_{\text{gap}} := \sup_{p \neq p_0} |\delta_p - \delta_{p_0}|$, and the aggregate sink gap is $G := l_{p_0} - \log \sum_{p \neq p_0} e^{l_p}$.

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. $Q = W_q \xi_q$ and $K_p = W_k \xi_p$ are linear in the projections, so to first order in ε,

$$\delta_p = \tfrac{1}{\sqrt{d_h}} \big[(\Delta W_q \xi_q) \cdot K_p + Q \cdot (\Delta W_k \xi_p) \big] + O(\varepsilon^2).$$

The sink-anchored gap perturbation between p_0 and any $p \neq p_0$ is

$$\begin{aligned} \delta_p - \delta_{p_0} = \tfrac{1}{\sqrt{d_h}} \big[& (\Delta W_q \xi_q) \cdot (K_p - K_{p_0}) \\ & + Q \cdot \Delta W_k (\xi_p - \xi_{p_0}) \big] + O(\varepsilon^2). \end{aligned}$$

By Cauchy–Schwarz,

$$\begin{aligned} |(\Delta W_q \xi_q) \cdot (K_p - K_{p_0})| &\leq \varepsilon \|\xi_q\| \|W_k\|_{\text{op}} \|\xi_{p_0} - \xi_p\|, \\ |Q \cdot \Delta W_k (\xi_p - \xi_{p_0})| &\leq \varepsilon \|W_q\|_{\text{op}} \|\xi_q\| \|\xi_{p_0} - \xi_p\|. \end{aligned}$$

Summing, dividing by $\sqrt{d_h}$, and taking the supremum over $p \neq p_0$ recovers Equation 1. The key-side contrast $\|\xi_{p_0} - \xi_p\|$ is amplified at the sink because ξ_{p_0} carries the massive-activation channel (Sun et al., 2024) while ξ_p does not, so the bound inherits the same $\gamma_{\text{in,max}}$ factor.

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 1 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), with the resulting $\text{RMS}_\tau \geq \text{RMS}$ only loosening the bound. 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 1 $\text{RMS}(Q_{\text{pre}}) \geq m_q \|\xi_q\|$ 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, noting $\|\tilde{k}_{p_0} - \tilde{k}_p\| \leq 2\sqrt{d_h}$ (mathematical unit-RMS gives each ℓ_2 norm exactly $\sqrt{d_h}$; with the stabilizer τ the RMS denominator is slightly larger so $\|\tilde{k}\|_2 \leq \sqrt{d_h}$, and the triangle inequality preserves the bound):

$$\begin{aligned} &|(\gamma_q \odot \Delta\tilde{q}) \cdot (\gamma_k \odot (\tilde{k}_{p_0} - \tilde{k}_p))| \\ &\leq \|\gamma_q\|_\infty \|\gamma_k\|_\infty \cdot \|\Delta\tilde{q}\| \cdot 2\sqrt{d_h} \\ &\leq \frac{2c_1\varepsilon\|\gamma_q\|_\infty\|\gamma_k\|_\infty\sqrt{d_h}}{m_q}. \end{aligned}$$

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 $p \neq p_0$ 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 2. □

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 1 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 $p \neq p_0$,

$$\frac{a'_p}{a'_{p_0}} = \exp(l_p - l_{p_0} + \delta_p - \delta_{p_0}).$$

Summing over $p \neq p_0$,

$$\begin{aligned} \frac{1 - a'_{p_0}}{a'_{p_0}} &= \sum_{p \neq p_0} \exp\big(l_p - l_{p_0}\big)\exp\big(\delta_p - \delta_{p_0}\big) \\ &\leq \exp\big(B_{\text{gap}}\big) \sum_{p \neq p_0} \exp\big(l_p - l_{p_0}\big) \\ &= \exp\big(B_{\text{gap}} - G\big), \end{aligned}$$

using the definition $G = l_{p_0} - \log \sum_{p \neq p_0} e^{l_p}$ and $|\delta_p - \delta_{p_0}| \leq B_{\text{gap}}$ for all $p \neq p_0$. Let $u := 1 - a'_{p_0}$ 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 3. 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 $T = l_{p_0} - \max_{p \neq p_0} l_p$ instead of the aggregate G, one would need an extra $\log N$ term to control the sum $\sum_{p \neq p_0} e^{l_p}$, giving $1 - a'_{p_0} \leq N \exp(-T + B_{\text{gap}})$ 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 $\|\Delta W_*\|_{\mathrm{op}} = \sigma_{\max}(\Delta W_*) = \|\Delta W_*\|_F/\sqrt{r_*}$. Keeping the two operator norms separate in the proof before upper-bounding both by a common ε, and substituting $\|\Delta W_*\|_{\mathrm{op}} = \|\Delta W_*\|_F/\sqrt{r_*}$, gives the Frobenius/stable-rank restatement

$$B_{\mathrm{gap}} \;\le\; c\,\|\gamma_q\|_\infty\,\|\gamma_k\|_\infty \left(\frac{\|\Delta W_q\|_F}{\sqrt{r_q}\,m_q} + \frac{\|\Delta W_k\|_F}{\sqrt{r_k}\,m_k}\right) + O(\varepsilon^2), \tag{4}$$

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 $\|\xi\|$ 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 cross-vendor comparison is consistent with the theorem's structural prediction that the QK-RMSNORM side is less input-magnitude sensitive, 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 $G_{\mathrm{base},\ell}$ exceeds the per-layer gap perturbation by a margin $G_{\mathrm{base},\ell} - B_{\mathrm{gap},\ell} \ge \mu$, 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). Those are also the layers where the empirical IFEval damage concentrates. A tight multi-layer theorem (in the spirit of (Olsson et al., 2022)'s circuit composition) is future work.

A.5 Numerical Sanity Checks

We verify two pieces on small synthetic cases.

Theorem 1, $d = 16$, $d_h = 4$, no QK-RMSNORM. W_q, W_k Gaussian-init. We instantiate a sequence of $N = 8$ key positions with i.i.d. unit-norm Gaussian backgrounds, then form ξ_{p_0} by adding a massive-activation-style channel at index 3 with value 10 on top of its background, while ξ_p for $p \ne p_0$ 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. ΔW_* Gaussian with $\|\Delta W_*\|_{\mathrm{op}} = 0.1$. Measured sink-anchored quantity $\sup_{p\ne p_0} |\delta_p - \delta_{p_0}| = 0.169$. The bound from Equation 1 is 0.197. Ratio 0.86.

Lemma 1, $G = 3.0$, $B_{\mathrm{gap}} = 0.4$, $N = 64$. Direct softmax computation with $l_{p_0} = 5$ and the $N - 1$ non-sink positions equally distributed to give $G = 3.0$. To realize the worst-case gap perturbation, we set $\delta_p = +B_{\mathrm{gap}}/2$ for every $p \ne p_0$ and $\delta_{p_0} = -B_{\mathrm{gap}}/2$, so that $\sup_{p\ne p_0} |\delta_p - \delta_{p_0}| = B_{\mathrm{gap}}$ is attained. Empirical $1 - a'_{p_0} = 0.069$. The logistic bound from Equation 3 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 (§B.1), the backbone-extraction pipeline (§B.2), and hardware/software details for reproducibility (§B.3).

B.1 Evaluation Protocols

We evaluate all text-side capability with `lm-evaluation-harness` v0.4.5 (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 is the protocol the model was trained for and the one a deployed assistant sees in practice.

Community-default protocol. The harness's per-task default few-shot count (`mmlu=5`, `gsm8k=8`, etc.) with the chat template *disabled*. This is the protocol most academic LLM leaderboards use.

Eight-task suite. MMLU, MMLU-Pro, GSM8K-CoT, TruthfulQA-MC2, BoolQ, 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 an LLM baseline to its VLM-LM under the same protocol, so framework drift cancels.

B.2 Backbone Extraction

To evaluate text-only behavior of a VLM with `lm-evaluation-harness`, we extract the language backbone as a standalone causal LM. For Qwen2.5-VL and the LLaVA-OneVision / LLaVA-Llama3 families we remap `model.language_model.*` keys to the corresponding `<Family>ForCausalLM` format and clone `embed_tokens.weight` as `lm_head.weight`. For Qwen3-VL the same procedure preserves the per-head q_{norm} / k_{norm} γ parameters. For our injected Qwen2-QkNorm variant (Section 5) we ship a self-contained `trust_remote_code` module so the extracted backbone inherits the same numerical pathway as during training. All backbones are saved to safetensors and the round-trip preserves exact logits on a held-out probe.

B.3 Implementation Details and Reproducibility

Hardware. Model training (the Section 5 QK-RMSNORM injection experiment) ran on $2 \times 2\times$ NVIDIA H100 (80 GB), two variants in parallel with two GPUs per variant. All text-side evaluation (lm-evaluation-harness across the eight-task suite for all 6 pairs and the LM-SFT control), 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 0.4.5, VLMEvalKit (commit pinned).

Released artifacts. Stage-1 align checkpoints, Stage-1 / Stage-2 data bundles, and full text-eval result JSONs ship with the submission via anonymized links. Non-anonymous releases follow upon acceptance.

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., 2024): https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov
- Molmo2-O-7B (Deitke et al., 2025): 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 (Yao et al., 2024): https://huggingface.co/openbmb/MiniCPM-V-4_5
- LLaVA-NeXT-Mistral-7B (Liu et al., 2024b): 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

Base LLMs. The text-only backbones extracted and evaluated alongside their VLMs (Appendix B.2):

- 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 et al., 2025): https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
- Qwen2.5-1.5B-Instruct (Qwen et al., 2025): https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct
- Qwen2.5-14B-Instruct (Qwen et al., 2025): https://huggingface.co/Qwen/Qwen2.5-14B-Instruct
- Qwen2.5-32B-Instruct (Qwen et al., 2025): https://huggingface.co/Qwen/Qwen2.5-32B-Instruct
- Qwen2-7B-Instruct (Yang et al., 2024): https://huggingface.co/Qwen/Qwen2-7B-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

Modality control variant. The text-only SFT variant used as the LM-only baseline (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 two factual-recall 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., 2023): https://huggingface.co/datasets/Idavidrein/gpqa
- MMLU (Hendrycks et al., 2020): https://huggingface.co/datasets/cais/mmlu
- BoolQ (Clark et al., 2019): https://huggingface.co/datasets/google/boolq

Calibration prompts. The 15-prompt set used for Sink Strength (Section 3) and the $G_{\text{base}}, B_{\text{gap}}$ measurement (Section 4) is drawn from the IFEval training partition, disjoint from the validation set we evaluate on. The exact prompts ship with the submission via the anonymized release described in Appendix B.3.

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 (notably the Llama 3.1 Community License for Llama-3.1-8B-Instruct and a custom research-use license for MiniCPM-V-4.5). 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. Artifacts released with this paper (the Stage-1 and Stage-2 checkpoints of the injection experiment in Appendix F.2, the measurement scripts, and the text-eval result JSONs) will be distributed under the MIT license.

C Per-Task and Behavioral Evidence

This appendix expands the per-task behavioral split surfaced in Section 3.1: factual-recall vs. format-sensitive coverage (§C.1), IFEval failure-category counts (§C.2), and an outcome-bucketed sink probe (§C.3).

C.1 Factual Recall vs. Format-Sensitive Tasks

The motivation closes by claiming that factual-recall capability is preserved while format-following collapses. 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 format-sensitive drops reach -9.6 pt (no-QK-RMSNorm) and -18.7 pt (layerwise).

	Factual recall		Format-sensitive			
Pair	MMLU	BoolQ	IFEval	EQ	GSM8K	GPQA
Qwen3-VL	$+1.6$	$+6.0$	-3.0	-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	$+3.0$	$+4.5$	-8.7	-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	$\mathbf{+0.5}$	$\mathbf{+2.6}$	$\mathbf{-8.3}$	$\mathbf{-16.3}$	$\mathbf{-14.2}$	$\mathbf{-7.9}$

Table 6: Per-pair Δ on factual-recall benches (MMLU, BoolQ) vs. format-sensitive benches (shaded). Same instruct protocol throughout; 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 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 (factual recall vs. instruction-following / reasoning-trajectory) and the QK protection of the base, 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.

	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: Failing IFEval instructions within the regression bucket, by category, per pair. "−" marks absence; **bold** marks the row-max. Totals match the *Lost* column of Table 2.

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 over 1000 resamples per bucket. The full per-pair table is released with the artifacts.

D Mechanism Details

This appendix expands the mechanism analyses of Section 4: the calibration-set size choice (§D.1), the Molmo2-O layerwise control (§D.2), channel-level ablation (§D.3), and a random-direction perturbation control (§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 541-prompt pool disjoint from the IFEval test set. This produces 18 measurements per pair across 6 headline 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 ~ 0.14 at $N=5$ (worst case, InternVL3.5) to ≤ 0.04 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.

D.2 Molmo2-O: 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 layer-wise scale preserves a strong sink: position-0 attention is 0.250 (95% CI $[0.242, 0.258]$), an order of magnitude larger than on the no-QK-RMSNorm qwen25vl-LM (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), only 0.9% of late-layer per-head cells reach a strong sink ($G > 2$, vs. the median per-head QK-RMSNorm pair at 58%), and the marginal proxy $G-B$ is -3.01 nats, the most negative across our sample. 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 base. 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. Molmo2-O's training recipe is also more invasive than the other VLMs we benchmark (PixMo multi-task SFT plus DPO with no text-only

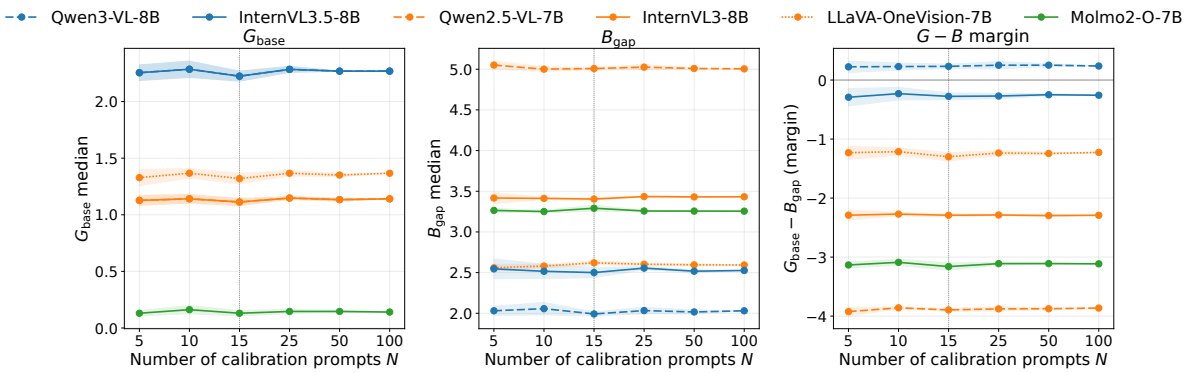

Figure 4: G_{base}, B_{gap}, and the $G_{\text{base}} - B_{\text{gap}}$ margin versus calibration size N for each of the six headline pairs. Lines are seed means, shaded bands are $\pm\sigma$ across 3 seeds. The vertical dashed line marks the paper-default $N = 15$; the horizontal line in the right panel 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 8: $G_{\text{base}} - B_{\text{gap}}$ margin ($\mu \pm \sigma$ across 3 seeds) over the full N grid; the paper-default $N = 15$ column is bold.

held-out objective), which may contribute additively. We therefore exclude Molmo2-O from the headline 5-pair comparison and report it separately as a distinguishing control.

D.3 Channel Ablation Sweep

For each QK-RMSNORM tensor 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. The default $K = 10$ was chosen by a sweep over $K \in \{5, 10, 20\}$, and the $K = 10 \rightarrow K = 20$ delta on IFEval is ≤ 0.5 pt, indicating the relevant amplifier mass concentrates in the top-10 set. 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 redundancy synergy.

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 weight-space distance from the LLM matches the natural VL adaptation but whose direction is random. IFEval under instruct protocol drops to 10.72 ($\Delta = -61.37$ against the LLM at 72.09), against the natural VL drop 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 base-LLM Sink Strength S, the measured post-VL IFEval drop, and the leave-one-pair-out linear prediction.

Pair	QK	S	Actual Δ	Held-out pred Δ
Qwen3-VL	per-head	$+2.36$	-3.0	-0.90
InternVL3.5	per-head	$+2.36$	-1.8	-1.71
Qwen2.5-VL	none	$+1.18$	-9.6	-10.47
InternVL3	none	$+1.18$	-8.7	-10.69
LLaVA-OV	none	$+1.44$	-7.9	-8.55
Molmo2-O	layerwise	$+0.24$	-18.7	-14.68

Table 9: Per-pair Sink Strength S (base LLM, pre-VL), actual IFEval drop after VL adaptation, and the leave-one-pair-out linear prediction. Numbers visualized in Figure 3.

E.2 IFEval-Specific Predictor (17 pairs)

The IFEval-specific predictor panel (Table 10) measures Sink Strength S against post-VL IFEval prompt-strict Δ on an extended 17-pair set: the 6 headline pairs plus 11 further-fine-tuned VLMs that span MoE architectures (Qwen3-VL-30B-A3B-Instruct, MolmoE-1B), four base-LLM families (Qwen, Olmo, Mistral, Llama-3.1) plus the MoE OLMoE, and backbone sizes 1.5B–38B (full pair-to-base-LLM mapping with Hugging Face URLs in Appendix B.4). Spearman $\rho(S, \Delta_{\text{IFEval}}) = +0.92$ across the panel, so the IFEval signal is robust to architecture, family, and scale, and Sink Strength remains a single-shot pre-VL diagnostic of post-VL IFEval damage at this broader scope.

E.3 Multi-Task Generalization (9 pairs)

We further restrict to a 9-pair subset (the 6 headline pairs plus three Qwen-family sub-10B additions: Qwen3-VL-4B, Qwen3-VL-2B, InternVL3-2B) on which the predictor's monotone $S \to \Delta$ relationship holds across *all four* format-sensitive tasks rather than IFEval alone. Table 11 reports per-pair Sink Strength S and the four post-VL deltas. The IFEval Spearman correlation on this 9-pair subset is $+0.979$ (vs. $+0.971$ on the original six), and all four task correlations remain in the strong-monotone band ($\rho \geq 0.82$).

F Controls

This appendix details the confound and negative controls of Section 5 and Section 5: a modality control on the same Qwen2.5-7B-Instruct base via LM-only SFT (§F.1), a post-hoc QK-RMSNORM injection experiment (§F.2), and a weight-merging sweep (§F.3).

F.1 Modality Control: LM-SFT Variant

The Section 5 modality control (Table 3) uses a publicly released LM-only further-finetuned variant of the same base LLM as the Qwen2.5 VL pair. The variant is evaluated under the same 0-shot instruct protocol as the rest of the paper (Appendix B.1), batch size 8, single A6000.

F.2 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 $\sim$ 47 h per variant ($\sim$ 7 h Stage 1, $\sim$ 40 h Stage 2).

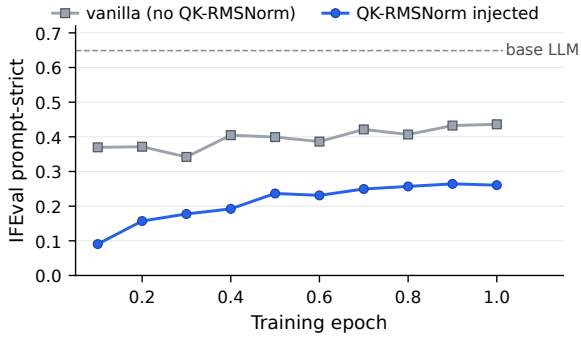

Figure 5: IFEval prompt-strict over the LLaVA Stage-2 training epoch for the controlled post-hoc QK-RMSNORM injection pair (vanilla vs. injected, matched in all other settings).

Released artifacts. Stage-1 align checkpoints, Stage-1 and Stage-2 data bundles, and the measurement script for Table 5 will be released upon acceptance. Anonymized download links are included in the supplementary materials.

F.3 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). The same settings were evaluated on three other format-sensitive benches (EQ-Bench, GSM8K-CoT flexible-extract, GPQA-Diamond-CoT flexible-extract). The cross-bench pattern matches IFEval: mild mixes (95/5, uniform averaging) stay within stderr of the un-merged VLM-LM baseline, 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. Full per-pair tables and Stage-A $G_{\text{base}}/B_{\text{gap}}$ JSONs are released with the supplementary materials.

Pair	Base LLM	QK class	S	ΔIFEval
Qwen2.5-VL	Qwen2.5-7B-Instruct	none	$+1.18$	-9.6
Qwen3-VL-8B	Qwen3-8B	per-head	$+2.36$	-3.0
InternVL3-8B	Qwen2.5-7B-Instruct	none	$+1.18$	-8.7
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-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	per-head (MoE)	$+1.66$	$+1.7$
$\rho_{\text{Spearman}}(S, \Delta_{\text{IFEval}})$ on 17 pairs				$\mathbf{+0.92}$

Table 10: IFEval-specific 17-pair predictor panel. Sink Strength S (base LLM, pre-VL) versus post-VL IFEval prompt-strict Δ. The panel spans dense and MoE architectures, all three QK-norm classes (per-head / layerwise / none), four base-LLM families (Qwen, Olmo, Mistral, Llama-3.1) plus the MoE OLMoE, and backbone sizes 1.5B–38B. Hugging Face URLs and paper references for every pair are in Appendix B.4.

Pair	Base LLM	QK	S	ΔIFEval	ΔGSM8K	ΔEQ-Bench	ΔGPQA-D
Qwen2.5-VL	Qwen2.5-7B-Instruct	none	$+1.18$	$\mathbf{-9.6}$	-14.3	-11.2	-9.8
Qwen3-VL-8B	Qwen3-8B	per-head	$+2.36$	$\mathbf{-3.0}$	-2.1	-2.4	-1.9
InternVL3-8B	Qwen2.5-7B-Instruct	none	$+1.18$	$\mathbf{-8.7}$	-12.5	-10.3	-7.1
InternVL3.5-8B	Qwen3-8B	per-head	$+2.36$	$\mathbf{-1.8}$	-1.5	-0.8	-0.6
LLaVA-OV-7B	Qwen2-7B-Instruct	none	$+1.44$	$\mathbf{-7.9}$	-11.9	-8.6	-8.4
Molmo2-O-7B	Olmo-3-7B-Instruct	layerwise	$+0.24$	$\mathbf{-18.7}$	-42.7	-64.6	-19.7
Qwen3-VL-4B*	Qwen3-4B	per-head	$+2.36$	$\mathbf{-5.0}$	$+13.9$	$+1.4$	-2.0
Qwen3-VL-2B*	Qwen3-1.7B	per-head	$+1.84$	$\mathbf{-5.2}$	$+27.5$	-0.7	-6.1
InternVL3-2B*	Qwen2.5-1.5B-Instruct	none	$+0.90$	$\mathbf{-12.2}$	-18.0	-9.0	-6.1
$\rho_{\text{Spearman}}(S, \Delta)$ on 9 pairs				$\mathbf{+0.979}$	$+0.877$	$+0.860$	$+0.821$

Table 11: Extended 9-pair predictor panel. $\Delta = \text{VLM-LM} - \text{LLM}$ (negative = degradation) 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). * marks the three pairs newly admitted beyond the headline six. The IFEval correlation tightens to $+0.979$ on the 9-pair panel (vs. $+0.971$ on the original six), with the new pairs spanning $S \in [0.90, 2.36]$ and base-LLM sizes 1.5B–4B. Bold ΔIFEval column is the headline metric.

Family	Variant (Hugging Face repo)
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct-1M`

Table 12: LM-only further-finetuned variant used as the modality control. `Qwen2.5-7B-Instruct-1M` is the Qwen team's official long-context (up to 1M tokens) text-only SFT release on the Qwen2.5-7B-Instruct base.

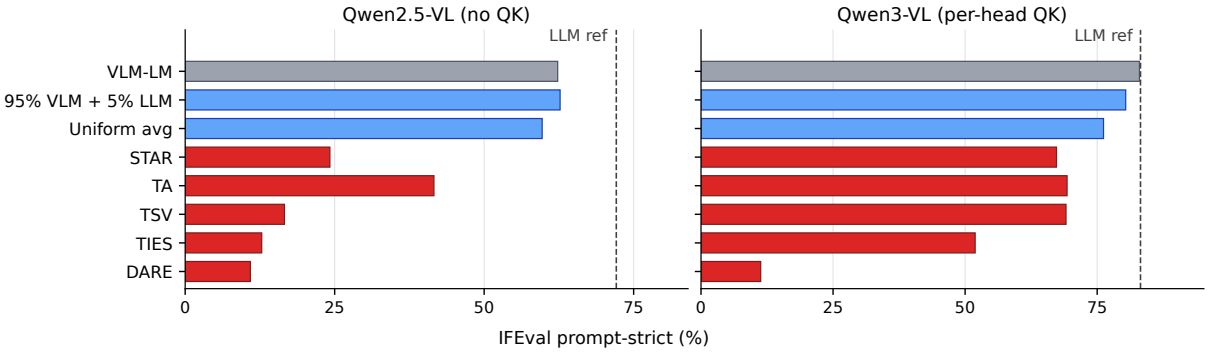

Figure 6: IFEval prompt-strict for Qwen2.5-VL (non-protected) and Qwen3-VL (per-head QK-RMSNORM) under seven weight-merging settings. Dashed line is each pair's base-LLM reference.

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