Life of Research · Published 1 November 2026 · DOI 10.5281/zenodo.23858880

A Gap to a Reference Is Not a Loss: What Text-Capability Comparisons Between Vision-Language Models and Their Language Backbones Measure, Why Independent Audits and Vendor Reports Disagree in Sign, and the Controlled Trajectory That Would Assign the Cause

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A vision-language model is usually built by attaching an image encoder to a pretrained language model and training the combination on multimodal data, and a recurring claim is that this damages the language model's own text abilities. A 2026 study gives the strongest form of the claim: across six released model pairs, the language backbones of vision-language models score up to 18.7 points below a text-only reference on instruction following while knowledge benchmarks barely move, and a scalar that measures how strongly the reference model concentrates attention on an early "sink" position ranks the damage with a Spearman correlation of 0.97. Vendor reports for some of the same models say the opposite. InternVL3's report puts its backbone ahead of the same text-only reference on most language benchmarks, and Qwen3-VL's report puts its 8B model slightly ahead on instruction following. In this analysis the author reviews and synthesises published results to show that a gap between a vision-language backbone and a text reference is not, in general, a loss caused by vision-language training. The gap carries four indices that headlines leave implicit: lineage (whether the reference is the checkpoint the model was trained from), protocol (harness, template and answer extraction), mixture (how much text, and in what output format, the multimodal recipe contained) and construct (whether the benchmark probes knowledge, procedure or output format). For one pair, an independent audit and the vendor report agree on MMLU to within a tenth of a point but disagree in sign on GSM8K and on MBPP. With lineage and protocol held fixed, a loss does survive: in a matched 3B run, vision fine-tuning cost about 23 instruction-following points in the first thousand steps, against about 7 for text-only tuning. In the same run, however, the loss arrived before the attention sink weakened, and part of it came back while the sink kept falling. The paper consolidates the measured values in a table, sets out what four candidate mechanisms predict, and gives an attribution procedure that separates pipeline differences, protocol artefacts, recoverable losses of access and losses of capacity.

Keywords: vision-language models; catastrophic forgetting; instruction following; attention sinks; QK normalization; evaluation methodology; multimodal training data; model merging

1. Introduction

Most open vision-language models (VLMs) start life as a text-only large language model (LLM). A vision encoder and a small projector are attached, and the whole system is trained on images paired with text, usually in stages: an alignment stage that trains the projector, then one or more instruction-tuning stages that also update the language model (Liu et al., 2023; Liu et al., 2024). The language model is the part of the system that writes every answer, so a question follows naturally from the recipe. Does training it on images and image-conditioned answers make it worse at the things it did with text alone?

The literature answers in two voices. The first reports damage. LLaVA-1.5-13B scores 5.99 on MT-Bench against 6.57 for the Vicuna-1.5-13B model it was built on (Li et al., 2024a). A LLaVA-style model built on Qwen1.5 lost ground on 12 of 16 text benchmarks, by up to 9.70 points on MMLU and 13.33 on RACE-High (Zhang et al., 2024). Pretraining a Llama-2 7B model on image-caption pairs alone cut its MMLU accuracy from 46.0 to 28.8 percent (Lin et al., 2024). NVIDIA's comparison table shows three open 40B-to-76B VLMs averaging 6.3 to 6.9 points below their text backbones across MMLU, GSM8K, MATH and HumanEval (Dai et al., 2024). And a 2026 study by Choi, Kim and Kim, the anchor of this paper, extracts the language backbone from released VLMs and finds instruction-following gaps of 1.8 to 18.7 points against text-only references on its six-pair diagnostic set (up to 22.7 on its 17-pair panel, where one pair, Qwen3-VL-30B-A3B, scores 1.7 points above its reference), concentrated on tasks graded by strict parsers, with knowledge tasks nearly untouched (Choi et al., 2026).

The second voice reports preservation or gain. NVLM-D 72B averages 84.1 on the same four benchmarks against 79.8 for its Qwen2-72B-Instruct backbone, with MATH rising from 59.7 to 73.1 (Dai et al., 2024). InternVL3-8B scores above Qwen2.5-7B Chat on most of the 17 language benchmarks in its report, with an overall score of 72.9 against 69.4 (Zhu et al., 2025). Qwen2.5-VL-72B scores 86.3 on IFEval against 84.1 for Qwen2.5-72B (Bai

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

et al., 2025a). Qwen3-VL-8B-Instruct scores 83.7 on IFEval against 83.0 for Qwen3-8B (Bai et al., 2025b). VILA's authors found that switching to interleaved pretraining shrank the MMLU loss to about 5 points, and that blending text-only instruction data into fine-tuning closed the rest (51.4 percent against 51.2 for Llama-2 fine-tuned on the same text data), writing that the text capabilities were "temporarily hidden, but not forgotten" (Lin et al., 2024).

Two of these voices speak about the same models. Choi et al. (2026) report that InternVL3-8B's backbone trails Qwen2.5-7B-Instruct by 8.9 points on IFEval and 12.5 on GSM8K under their protocol; the InternVL3 report puts the same model three points ahead of the same reference on GSM8K under its own (Zhu et al., 2025). Choi et al. report Qwen3-VL-8B 5.4 points below Qwen3-8B on IFEval; the Qwen3-VL report puts it 0.7 points above. The question this paper asks is what a published gap between a VLM's language backbone and a text-only reference measures, and under what conditions it supports the claim that vision-language training removed something.

In this analysis the author reviews and synthesises published results to show that a gap to a reference is not, in general, a loss, and that the published record contains four distinct quantities under the one name. A reported text-capability gap carries four indices. The first is **lineage**: whether the reference model is the checkpoint the VLM's language component was trained from, a sibling trained from the same base by a different pipeline, or a model from a different generation. The second is **protocol**: the evaluation harness, chat template, few-shot count and answer-extraction rule under which both models were scored. The third is **mixture**: how much text-only data the multimodal recipe contained, and what output formats its supervision rewarded. The fourth is **construct**: whether a benchmark measures stored knowledge, a reasoning procedure, or compliance with an output format.

Fixing the four indices dissolves part of the disagreement and sharpens the rest. The paper argues three things. First, lineage and protocol alone can reverse the sign of a published gap. The anchor's own reference models are, in its authors' words, "a comparison point rather than an assumed immediate predecessor" of multimodal training (Choi et al., 2026), and for InternVL3 the vendor states that the language component was initialised from the pretrained base, not from the instruct model used as the reference (Zhu et al., 2025). Second, when lineage and protocol are fixed, a real loss attributable to vision-language training remains. In the anchor's matched 3B experiment, vision fine-tuning cost far more instruction-following accuracy than text-only fine-tuning at matched optimisation settings (Choi et al., 2026). Third, the evidence on mechanism is weaker than the evidence that the loss exists. In the same matched run, the loss came first and the attention-sink signature came later (Figure 1c). The published recovery results, from re-blended text data (Lin et al., 2024), preference tuning on images (Li et al., 2024a) and lineage-matched merging (Ratzlaff et al., 2024; Yu and Ananiadou, 2025), fit at least as well with a shift in output format and task inference as with corrupted attention.

The paper makes six contributions. It stipulates the terms and the four indices (Section 2). It reads the anchor result at the resolution of its own tables (Section 3) and sets out the opposing record (Section 4). It works through lineage, protocol, mixture and construct with the published results that isolate each (Sections 5 to 8). It states four candidate mechanisms and what each predicts about timing, recovery and architecture, and checks the predictions against the published trajectories (Sections 9 to 11). It consolidates the measured values in Table 1 (Section 12) and gives an attribution procedure, Algorithm 1, that a claim of "vision-language training damaged text capability" should pass (Section 13). It closes with the strongest objections, the studies that would settle what remains, and a flat statement of what is not known (Sections 14 to 16).

The argument is not that the anchor's finding is wrong. Its central observation, that strict-format instruction following is where released VLM backbones most often fall short of a text-only model, is supported by its own controls and by independent work (Li et al., 2024a; Srivastava et al., 2024). Nor is the argument that vendor reports are the truth and the audit an artefact; a vendor chooses its protocol and its comparison as freely as an auditor does. The argument concerns the inference from "the backbone scores below the reference" to "vision-language training damaged the backbone, and the attention sink is how". That inference needs lineage fixed, protocol fixed and the loss timed, and few published comparisons do all three.

2. Scope and terms

The paper concerns decoder-only VLMs in which visual features are projected into the input sequence of a pretrained language model and the language model's weights are updated during multimodal training. Cross-attention architectures with frozen language models are discussed only as a contrast, since freezing removes the phenomenon by construction (Dai et al., 2024). The paper concerns text-only behaviour of the language component: what the model does when no image is present. It does not cover the converse question of how well a VLM uses its image input, nor safety behaviour, except where safety results bear on recovery (Lee et al., 2024).

Reference. The text-only model against which a VLM's backbone is scored. It may be the exact checkpoint the VLM was initialised from, a sibling built from the same pretrained base, or a model of another generation chosen for convenience.

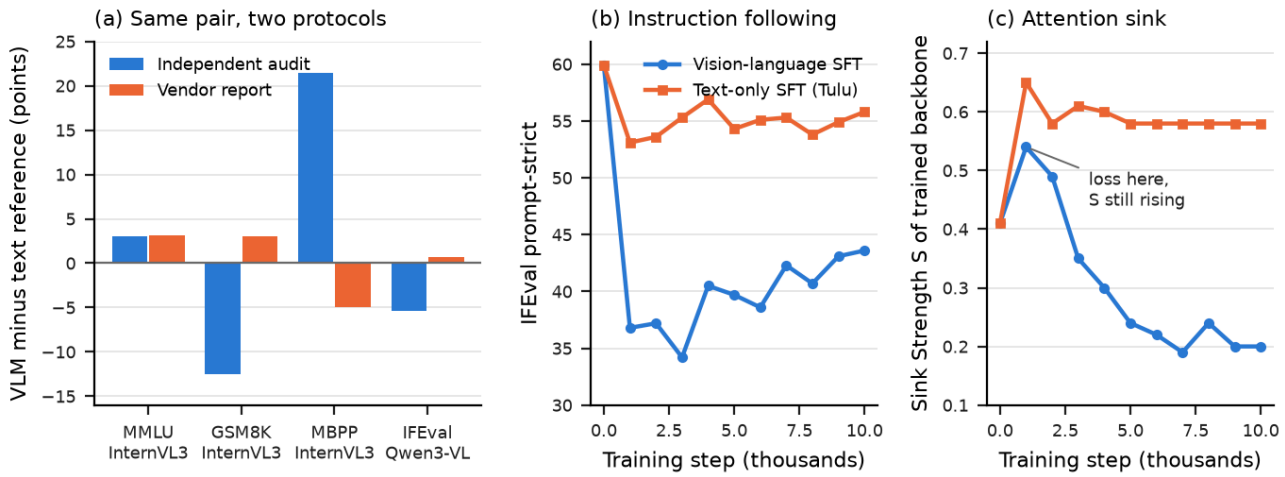


Figure 1. (a) Text-capability gap (VLM minus text reference, in points) for the same model pairs under an independent audit and under the vendor's own report. InternVL3-8B minus Qwen2.5-7B-Instruct on MMLU, GSM8K and MBPP (audit: Choi et al. 2026, Tables 6 and 8, lm-evaluation-harness, zero-shot, chat template; vendor: Zhu et al. 2025, Table 11, OpenCompass), and Qwen3-VL-8B minus Qwen3-8B on IFEval (audit: Choi et al. 2026; vendor: Bai et al. 2025b, Table 6). (b) IFEval prompt-strict accuracy and (c) Sink Strength of the trained backbone over training, for matched vision-language and text-only (Tulu) fine-tuning runs from Qwen2.5-3B-Instruct, as reported by Choi et al. 2026, Table 14.

Endpoint. The language component of a trained VLM, scored on text-only inputs. Choi et al. (2026) extract it as a standalone causal language model; vendor reports usually score the full VLM with no image.

Gap. The score of the endpoint minus the score of the reference on a named benchmark under a named protocol. A gap is a measurement of two models.

Loss. The change in a model's score caused by a training stage, measured by scoring the checkpoint before and after that stage under one protocol. A loss is a measurement of one model at two times. A gap equals a loss only when the reference is the endpoint's own pre-stage checkpoint and the protocol is shared.

Lineage-matched. A pair is lineage-matched when the reference is the endpoint's own initialisation. Studies that compare a VLM with the LLM it was trained from have lineage-matched pairs (Ratzlaff et al., 2024; Srivastava et al., 2024; Zhang et al., 2024). Comparisons of released models usually do not.

Protocol. The full scoring procedure: harness and version, whether a chat template and system prompt are applied, number of in-context examples, decoding settings, and the rule that turns generated text into a graded answer.

Format-sensitive task. Choi et al. (2026) apply the term to IFEval, EQ-Bench, GSM8K with chain-of-thought and GPQA-Diamond with chain-of-thought, on the ground that each is graded by parsing the output against a strict rule. IFEval itself consists of roughly 500 prompts carrying "verifiable instructions" of 25 types, such as minimum word counts or required keywords (Zhou et al., 2023a).

Access and capacity. The paper separates two ways a capability can disappear from a score. A loss of access means the model still holds what the task needs, but no longer produces it under the evaluated prompt: a small intervention such as a different prompt, a few training steps or a short data blend restores it. A loss of capacity means it does not come back that cheaply. The distinction follows the literature on spurious forgetting, where drops in continual fine-tuning often reflect lost task alignment rather than lost knowledge (Zheng et al., 2025), and on implicit task inference, where fine-tuning shifts which task a model thinks it is being asked to do (Kotha et al., 2024).

Attention sink. A position, usually the first token, that receives a large share of attention probability across heads and layers regardless of its content (Xiao et al., 2024). **Sink Strength S** is the anchor's summary statistic: the median, over late layers, heads and query positions, of the log-odds of attention to each head's most-attended key position (Choi et al., 2026).

3. The anchor result, read at the resolution of its tables

Choi et al. (2026) study five headline pairs of released VLMs and text references, plus a control. Qwen3-VL-8B and InternVL3.5-8B are paired with Qwen3-8B; Qwen2.5-VL-7B and InternVL3-8B with Qwen2.5-7B-Instruct; LLaVA-OneVision-7B with Qwen2-7B-Instruct; and Molmo2-O-7B, the control, with Olmo-3-7B-Instruct. Each VLM's language backbone is extracted and scored with lm-evaluation-harness under a zero-shot protocol with the chat template applied.

The gaps. On IFEval prompt-level strict accuracy the gaps are -1.8 (InternVL3.5), -5.4 (Qwen3-VL), -7.9

(LLaVA-OV), -8.9 (InternVL3), -9.6 (Qwen2.5-VL) and -18.7 (Molmo2-O). The per-prompt breakdown shows that the losses are spread over 43 to 126 distinct prompts per pair, so the gaps do not come from a few outliers. The failing prompts break rules a parser checks: word-count limits, required exact phrases, single-language constraints, and markers such as section headers. On Qwen2.5-VL, 28 of the 91 regressed prompts violate a punctuation rule, mostly the instruction not to use commas.

The contrast with knowledge tasks. Across the six pairs, MMLU and BoolQ never fall by more than 2.3 points, and their means are $+0.5$ and $+2.6$. The four format-sensitive tasks have means of -8.7 (IFEval), -16.3 (EQ-Bench), -14.2 (GSM8K) and -7.9 (GPQA). The authors rule out the reading that letter-choice tasks survive merely because they are scored by likelihood: GPQA-Diamond is multiple choice but drops by 7.1 to 9.8 points on the three pairs without QK normalisation, and GSM8K is open generation but stays within 2.1 points on the two pairs with it.

The predictor. Sink Strength, computed on the reference model from 15 prompts in about eight seconds, separates the pairs. The two Qwen3-based pairs share $S = +2.36$, the three older pairs fall between $+1.18$ and $+1.44$, and the Molmo2-O reference sits at $+0.24$. The rank correlation with the IFEval gap is 0.97 over six pairs. Under leave-one-pair-out fitting the mean prediction error is 2.54 points, against 4.40 for predicting the mean. The authors call this a sanity check rather than prospective evidence, which is the right description of six points. On a wider panel of 17 pairs the rank correlation is 0.88, but, as the authors note, pairs that share a reference share S, so the 17 pairs supply 12 distinct values of S, and 13 of the 17 are Qwen-family.

The architectural correlate. The S clusters line up with how the reference normalises queries and keys. Qwen3 applies RMS normalisation to each head's queries and keys (per-head QK-RMSNorm); Qwen2 and Qwen2.5 do not; Olmo-3 applies it across the whole layer. The authors give a perturbation bound under which per-head normalisation removes the dependence of the sink's logit margin on raw input magnitude, and argue that the layerwise variant does not. Within each vendor's line the newer, normalised backbone loses less, and the authors state plainly that this is not a controlled comparison because successive releases differ in much else.

Modality controls. Two controls address whether the gap is a property of vision-language training or of further training in general. In the first, a text-only Qwen2.5 endpoint (the long-context Qwen2.5-7B-Instruct-1M) shows no IFEval change against Qwen2.5-7B-Instruct where Qwen2.5-VL shows -9.6, and the VL endpoint's query and key projections have moved further from the reference, by median factors of 4.44 and 3.65 in operator norm. In the second, the authors run a matched experiment from Qwen2.5-3B-Instruct: one run fine-tunes on image-text instructions, the other on the Tulu text instruction set (Lambert et al., 2025), with batch size, learning rate and step count held equal. The vision run's IFEval falls from 59.9 to 36.8 by step 1,000 and ends at 43.6 after 10,000 steps. The text run falls to 53.1 at step 1,000 and ends at 55.8.

Negative controls. Adding per-head QK-RMSNorm, with unit scales, to Qwen2.5-3B-Instruct before vision training does not protect it: the modified model scores 11 on IFEval at its first evaluated checkpoint against 59.9 for the unmodified base. Merging the VLM backbone back toward its reference with seven off-the-shelf settings does not recover IFEval on Qwen2.5-VL; only the two mildest mixes avoid substantial further damage, and full-strength task arithmetic, task-singular-vector and spectral merges cost the unprotected pair tens of points.

Two results the paper reports but does not lean on. The first is a random-direction control. Gaussian noise added to Qwen2.5-7B-Instruct, scaled so that each sub-module moves as far in relative Frobenius norm as Qwen2.5-VL's backbone moved, drops IFEval to 10.72, a gap of -61.37, against the observed -9.6. The second is the step-by-step alignment of S with IFEval in the matched 3B run, which the authors describe candidly: most of the IFEval loss occurs by step 1,000, before S falls, and IFEval then partly recovers while S continues to decline. The paper returns to both in Section 9, because they bear more heavily on mechanism than the cross-pair correlation does.

Read at this resolution, the anchor makes two separable claims. One is descriptive: released VLM backbones, scored under one protocol against chosen references, fall short mostly on strictly graded tasks, and a cheap statistic of the reference ranks how far. The other is causal: vision-language fine-tuning corrupts the attention sink, and sink corruption is why format-sensitive tasks fail. The descriptive claim rests on six to seventeen pairs and a stated protocol. The causal claim rests mainly on the matched 3B run, a perturbation bound, and the architectural correlation. The rest of this paper asks what each claim needs.

4. The other side of the record

Set beside the anchor, the published record contains four kinds of evidence that point the other way or complicate the direction.

Vendor comparisons that show gains. NVLM reports that all three of its 72B variants beat their Qwen2-72B-Instruct backbone on the four-benchmark average, by 2.5 to 4.3 points, and attributes this to a curated text-only SFT set together with a large amount of multimodal mathematics data (Dai et al., 2024). The same table shows LLaVA-OneVision-72B, built on the same backbone, 6.3 points below it. InternVL3's language-

capability table compares six model sizes against the Qwen2.5 Chat models whose pretrained bases initialised them, and InternVL3 is ahead on the overall score at every size (Zhu et al., 2025). The authors attribute this to roughly 25 percent pure-language data in native multimodal pretraining, joint optimisation of all parameters, and high-quality text in post-training. Qwen3-VL's report states that multimodal models "are expected to match or surpass their text-only counterparts on language benchmarks" and presents its dense and mixture-of-experts models as matching or exceeding their text-only counterparts on most of them (Bai et al., 2025b).

Controlled studies in which the sign depends on the backbone. Ratzlaff et al. (2024) compare three LLaVA models with their own source LLMs on eight reasoning tasks. LLaVA-1.6-Mistral falls below Mistral on every task except CommonsenseQA, while LLaVA-1.5 and LLaVA-1.6, both built on Vicuna-1.5, perform similarly to or better than Vicuna on most tasks, with GSM8K the consistent exception. Srivastava et al. (2024) train LLaVA-style models on nine language models and report that instruction-tuned backbones show minimal or even negative forgetting, while base models forget more, although in their own Table 2 this holds only for Pythia-1.4B: at 7B the base Llama-2 shows more positive transfer (-1.84, where negative means a gain) than the instruction-tuned Vicuna (-0.36). They also report that forgetting falls with model scale and is much more severe on generation than on understanding tasks.

Recipes that restore text ability. VILA's MMLU loss under caption-only pretraining shrinks to 5.3 points with interleaved data and disappears with a text-blended fine-tuning stage: 51.4 percent against 51.2 for Llama-2 fine-tuned on the same text data (Lin et al., 2024). DeepSeek-VL reports that keeping at least 70 percent language data in joint pretraining is needed to preserve language ability, and that training on multimodal data alone produces severe forgetting (Lu et al., 2024). MM1 finds interleaved and text-only data important for text-only performance and settles on 45 percent captions, 45 percent interleaved and 10 percent text (McKinzie et al., 2024). MLAN reports that a text-heavy instruction mix performs on par with a vision-heavy one on both modalities while using as little as half the tokens (Tu et al., 2024). Li et al. (2024a) raise LLaVA-1.5-13B's MT-Bench score from 5.99 to 6.73, above Vicuna's 6.57, with direct preference optimisation on about 5,000 image-question pairs, without any added text-only data.

Frozen and partially frozen backbones. Llama 3-V freezes the language model and shows no text degradation by construction, and NVLM's authors report that freezing the language model in their cross-attention variant preserves text at a moderate cost in vision-language accuracy (Dai et al., 2024). Scale also matters: Srivastava et al. (2024), summarising Driess et al. (2023), note that a 12B PaLM-E model lost 87.3 percent of its language performance while the 562B model lost 3.9 percent. LoRA-trained VLM baselines forget less text than fully fine-tuned ones while doing worse on multimodal questions (Zhang et al., 2024), which matches the general finding that low-rank adaptation learns less and forgets less (Biderman et al., 2024a).

None of this contradicts the anchor's measurements. It shows that the sign and size of a text-capability gap vary with things other than the attention mechanism of the reference model: which backbone, which reference, which data, which protocol. The next four sections take those in turn.

5. Index one: lineage

A gap measures a loss only if the reference is where the endpoint started. For the released models in the anchor's panel, that condition is mostly unverified and in at least one case contradicted by the vendor.

The InternVL3 report states that its language components were "initialized solely from pre-trained base models, without employing instruction-tuned variants" (Zhu et al., 2025). The InternVL3 backbone and Qwen2.5-7B-Instruct are therefore siblings. Both descend from the Qwen2.5-7B base, one through Qwen's text post-training pipeline and the other through InternVL's native multimodal pretraining and multimodal post-training. A gap between them compares two post-training pipelines applied to one base. It is evidence about which pipeline produces better text behaviour under a given protocol, not evidence that multimodal training removed something the instruct model had, because the InternVL3 backbone never had what the instruct model learned in Qwen's text post-training.

The Qwen2.5-VL report says the model "is initialized with pre-trained weights from the Qwen2.5 LLM" without saying whether that means the base or the instruct checkpoint, and describes 4.1 trillion tokens of multimodal pretraining across three stages followed by SFT on about two million entries, half of them pure text, and DPO (Bai et al., 2025a). For small Qwen3-VL models, the report describes a "Strong-to-Weak Distillation" approach (Bai et al., 2025b). LLaVA-OneVision names the Qwen-2 family as its language model (Li et al., 2024b), and NVLM's comparison table lists Qwen2-72B-Instruct as the backbone of LLaVA-OneVision-72B (Dai et al., 2024), so the anchor's LLaVA-OneVision pair may be the closest thing to a lineage-matched pair in its headline panel. The others are long pipelines whose starting checkpoint is either stated to differ from the reference the anchor uses or not stated at all.

The anchor's authors are aware of this and phrase their claims accordingly. Their causal claims rest on the controls where lineage is fixed: the matched 3B trajectories and the injection experiment, both started from a stated checkpoint. The descriptive claim, that S

on the reference ranks the gap, does not depend on lineage, because S is computed on the reference and the gap is defined against it. What lineage changes is the reading. If an endpoint was never trained from its reference, then "how well the base LLM preserves its sink" (Choi et al., 2026) is not quite the quantity being correlated; the correlation is between a property of one model and the difference between that model and another model built from the same base by another team.

Lineage also bears on the comparison between pairs. The anchor's strongest architectural contrast sets Qwen3-based pairs ($S = +2.36$) against Qwen2.5-based pairs ($S = +1.18$). The Qwen3 and Qwen2.5 references differ in QK normalisation, and also in pretraining data, post-training (including Qwen3's hybrid thinking mode), and vendor tooling (Yang et al., 2025). Lineage-matched controlled studies avoid this confound by construction, since the same team trains the VLM from a stated model, and they show that the backbone matters. In Ratzlaff et al. (2024), Mistral-based LLaVA degrades on almost everything while Vicuna-based LLaVA does not. Neither is a per-head-normalised backbone of the Qwen3 kind, and the anchor's own extended panel lists Mistral-7B among the references without QK normalisation (Choi et al., 2026). Whatever separates the two, it is not the axis the anchor identifies, so backbone-dependent vulnerability exists without it. Prismatic VLMs add a related observation: a VLM trained from base Llama-2 and one trained from instruct-tuned Vicuna-1.5 show no statistically significant difference on vision-language benchmarks, although the instruct-derived models are more verbose and more prone to hallucination (Karamcheti et al., 2024). What the backbone was before multimodal training shapes how its outputs read afterwards.

A last lineage effect concerns the reference's own training. Srivastava et al. (2024) find instruction-tuned backbones more robust to linguistic forgetting than base models, though their two-model comparison shows this only at Pythia-1.4B and not at Llama-2 7B, and suggest the reason is distributional similarity between text instruction tuning and visual instruction tuning. Luo et al. (2023) report that general instruction tuning reduces forgetting in later fine-tuning. If an instruct reference is protected by its instruction tuning, then the distance between a base-initialised VLM and an instruct reference includes the instruction tuning the VLM never received from that pipeline.

6. Index two: protocol

The clearest evidence that protocol is a first-order variable comes from the pairs that both the anchor and a vendor report score against the same reference. The one with the most shared tasks is InternVL3-8B against Qwen2.5-7B-Instruct, which InternVL3's report calls Qwen2.5-7B Chat.

The two sources agree on MMLU. Choi et al. (2026) report $+3.0$ for the backbone; Zhu et al. (2025) report 77.3 against 74.2, a gap of $+3.1$. They disagree in sign on the two other tasks both report. On GSM8K the anchor reports -12.5 with chain-of-thought and flexible extraction, while the vendor reports 83.1 against 80.1, a gap of $+3.0$. On MBPP the anchor reports $+21.4$ for execution-graded pass@1, while the vendor reports 69.3 against 74.3, a gap of -5.0. The Qwen3-VL-8B pair shows the same pattern on IFEval: -5.4 in the anchor, $+0.7$ in the vendor's table (Bai et al., 2025b). Figure 1a sets these four comparisons side by side.

The other overlaps mostly point the same way. For InternVL3.5-8B against Qwen3-8B, the anchor reports $+3.0$ on MMLU and -1.5 on GSM8K, while the vendor's language table gives 78.2 against 76.9 ($+1.3$) and 90.9 against 89.8 ($+1.1$): agreement on MMLU, a sign flip on GSM8K (Wang et al., 2025). For InternVL3-2B against Qwen2.5-1.5B-Instruct, the anchor reports -18.0 on GSM8K and the vendor 72.5 against 61.6 ($+10.9$) (Zhu et al., 2025). On IFEval, the anchor reports -5.0 for Qwen3-VL-4B and -5.2 for Qwen3-VL-2B, and the vendor $+1.1$ for each. The exception is Qwen3-VL-30B-A3B on IFEval, where the anchor reports $+1.7$ and the vendor 85.8 against 84.7 ($+1.1$), so the two sources agree in sign (Bai et al., 2025b). Every sign flip in this set falls on a generative task, but not every generative comparison flips.

Neither source is neutral. The vendor chose a toolkit and prompt versions, and notes that its Qwen2.5 scores "may differ from those officially reported" because of them (Zhu et al., 2025). The anchor chose lm-evaluation-harness, zero-shot prompting with the chat template, and flexible extraction, and checked that its reference-side numbers reproduce published Qwen values to within a point under matched configurations (Choi et al., 2026). The anchor also describes a second, leaderboard-style protocol with harness-default few-shot counts and no chat template, "whose disagreement on a single backbone is itself diagnostic". The version of the paper read here does not tabulate per-pair results under it.

The general literature on evaluation explains why the disagreement concentrates where it does. Sclar et al. (2024) find that meaning-preserving changes to prompt format move few-shot accuracy by up to 76 points for LLaMA-2-13B, and that which format suits a model correlates only weakly between models, which "puts into question" comparisons made under one fixed format. Biderman et al. (2024b) document the sensitivity of language-model scores to evaluation setup as a recurring practical problem. On the InternVL3 pair, the two tasks whose sign flips in Figure 1a are generative tasks graded by extracting an answer from free text: a final number for GSM8K, executable code for MBPP. The task on which the sources agree is graded by choosing a letter. The Qwen3-VL flip is on IFEval, which is generative and parser-graded too. The protocol-sensitive tasks are the

format-sensitive ones.

That coincidence admits two readings, and the evidence in hand does not choose between them. On the first, a VLM backbone's output habits really have drifted, so it is more fragile than the reference under any protocol that grades strictly. The protocol then decides how much of the fragility a benchmark exposes, but the fragility is real and is a model property. On the second, a VLM backbone formats its output differently, for instance more tersely, in another layout, or with a different chain-of-thought style, and whether a given extractor rewards or penalises that difference depends on the extractor. The gap is then a joint property of model and grader, and the reference and the endpoint can swap places under another grader. Figure 1a's MBPP pair, where the anchor's protocol favours the VLM backbone by 21 points and the vendor's favours the reference by 5, fits the second reading better than the first. A degraded ability to follow output rules should not, on its own, raise execution-graded coding accuracy by 21 points. The anchor itself reports that MBPP rises by a mean of 7.3 points across its five headline pairs.

The practical consequence is narrower than "vendor numbers are unreliable" or "audits are artefacts". A text-capability gap is a statement about two models under one grader, and a claim that it generalises across graders needs a second grader. The anchor's internal comparisons, with both models of a pair under one protocol, are sound as stated. The comparison across sources is not, and neither is a summary that treats either source's sign as a property of vision-language training.

7. Index three: mixture and supervision format

The published recipes differ by an order of magnitude in how much text-only data they contain, and the gaps track the difference in the expected direction. Qwen2.5-VL's SFT set is half pure text (Bai et al., 2025a). InternVL3 reports about 25 percent pure-language data in pretraining (Zhu et al., 2025), and InternVL3.5 a text-to-multimodal ratio of about 1:2.5 in pretraining and 1:3.5 in SFT (Wang et al., 2025). LLaVA-OneVision's single-image instruction set includes language-only data at about 14 percent (Li et al., 2024b). LLaVA-1.5 blends ShareGPT conversations into an instruction mix dominated by visual question answering (Liu et al., 2024). DeepSeek-VL keeps 70 percent or more language data in pretraining (Lu et al., 2024). The anchor's matched 3B vision run uses the LLaVA-1.5 instruction mix (Choi et al., 2026).

Two pieces of evidence suggest that the quantity of text is not the whole story, and that the format of the supervision matters as much as its modality.

The first concerns the very data the anchor's control used. LLaVA-1.5's authors diagnosed short-answer overfitting in InstructBLIP, which "can overfit to VQA training sets with short-answers, even on requests that require detailed responses", and traced it partly to ambiguous prompts that "can overfit an LLM behaviorally to short-form answers even for natural visual conversations" (Liu et al., 2024). To prevent it in their own mix, they appended format prompts such as "Answer the question using a single word or phrase" to academic VQA data, and report that LLaVA-1.5 then adjusts its output format to the user's instructions. The mix still contains a large share of single-word or single-phrase targets, and whether these leave a terse prior on text-only prompts is untested. The hypothesis is that a model fine-tuned on it may acquire such a prior, and IFEval's failing categories in the anchor include word-count limits and exact-phrase requirements cut short (Choi et al., 2026). The matched control's text arm uses the Tulu general instruction-tuning data, built for a general-purpose assistant (Lambert et al., 2025). The two arms therefore differ in output-length and output-format distribution as well as in modality. The anchor states that the runs differ "in training modality and data", and the control cannot separate the two.

The second is that changing the targets while keeping the modality reverses the loss. Li et al. (2024a) argue that VQA datasets lack the diversity and complexity of the text instruction data the backbone was trained on, and show that preference optimisation on about 5,000 image-question pairs with better-graded answers raises text MT-Bench above the source LLM's. NVLM's gains are attributed to the quality of its text-only SFT set and to multimodal mathematics data (Dai et al., 2024). In both cases the model kept training on images. What changed was the distribution of answers it was rewarded for.

The format-specialisation literature in text-only models predicts exactly this. Wang et al. (2024) show that fine-tuning a language model on a single task reduces its general in-context ability largely through "format specialization", overfitting to the fine-tuned task's output format, and that this happens at the very beginning of fine-tuning. The superficial-alignment results point the same way from the other side. LIMA concludes that almost all knowledge is learned in pretraining and that a small number of examples teaches response format (Zhou et al., 2023b). URIAL finds that aligned and base models differ mainly on stylistic tokens (Lin et al., 2023). If instruction-following behaviour sits in a thin, format-level layer, a large SFT stage with a skewed output distribution is the natural thing to disturb it, whatever the input modality.

This does not show that modality is irrelevant. The anchor's endpoint comparison finds the vision-language endpoint's attention projections displaced several times further than a text-only endpoint's, and the sink-to-other contrast in inputs about 1.12 times larger once vision tokens are present (Choi et al., 2026). Visual tokens enter the residual stream with statistics unlike text, and visual attention sinks form on a few image tokens with

massive activations (Kang et al., 2025), much as vision transformers repurpose high-norm background tokens for internal computation (Darcet et al., 2024). What the evidence does show is that the matched control, as run, identifies the effect of a data package, not the effect of modality, and that the published remedies that work change the data package.

8. Index four: construct

The anchor's finding that knowledge tasks survive while format-sensitive tasks degrade is one of its most useful results, and it disagrees with earlier work in a way that is informative.

Zhang et al. (2024) report MMLU falling by 9.70 points and CMMLU by 8.91 when a Qwen1.5 backbone is trained on LLaVA-style multimodal instructions under their own recipe. VILA's caption-only pretraining cut MMLU by 17.2 points (Lin et al., 2024). Choi et al. (2026) find MMLU changes between -2.3 and +3.0 across six released pairs. The difference is best explained by mixture and maturity of recipe rather than by construct. The studies that saw knowledge loss trained their own models with little text in the mix. The released models in the anchor's panel were trained by vendors who had observed this kind of loss and added large amounts of text to prevent it. The anchor's "knowledge preserved" result is therefore a finding about current release recipes, not a property of vision-language adaptation as such.

Within the anchor's panel, the construct split is real but less clean than the four-task label suggests. MBPP is execution-graded and so imposes a strict surface gate: code that fails to parse or run scores zero. Yet it rises by a mean of 7.3 points across the five headline pairs, and by 21.4 for InternVL3 (Choi et al., 2026). MMLU-Pro declines by a mean of only 2.7. The tasks that fall most share a stricter property than parseability: they reward following an instruction about the form of the answer that the model was not trained to expect, such as "no commas", "exactly three paragraphs", or "end with this phrase". This is closer to what IFEval was designed to measure (Zhou et al., 2023a) than to a general capacity for structured output.

Two other constructs are easy to conflate with these. One is the stated ability of a model to follow conversational instructions as judged by another model, which MT-Bench measures and on which LLaVA-1.5 trails Vicuna (Li et al., 2024a). The other is chain-of-thought reasoning graded on a parsed answer. Here GSM8K degrades in every model in Ratzlaff et al. (2024), including the Vicuna-based ones that otherwise improve, which suggests something more specific than format compliance, such as a shift in reasoning-trace style. A paper that reports "text capability" as one number across these constructs reports an average of quantities that move for different reasons.

9. Four mechanisms, and what each predicts

With the indices fixed as far as the evidence allows, a residual loss attributable to vision-language training remains: the anchor's matched 3B run shows one. Four candidate mechanisms are on offer in the literature. They are not mutually exclusive, but they make different predictions, and the predictions can be checked against published trajectories.

M1, sink corruption. Vision-language fine-tuning perturbs the query and key projections. Because the induced logit shift scales with input magnitude, and the sink position carries massive activations (Sun et al., 2024), the perturbation is largest exactly at the sink, which collapses the sink's share of attention. Losing the sink disrupts the per-position attention that strict-format tasks need (Choi et al., 2026). The broader sink literature supports the premise that sinks do work. They receive strong attention even when they carry no meaning, and keeping them in the cache is what rescues windowed attention (Xiao et al., 2024), they behave like key biases that hold excess attention (Gu et al., 2025), they keep deep transformers from over-mixing token representations (Barbero et al., 2025), and they serve as a "catch, tag, release" mechanism that marks groups of tokens with a shared direction (Zhang et al., 2025). M1 predicts that the loss follows the sink's decline in time, that restoring the sink restores the behaviour, and that architectures which pin per-head query and key magnitudes are protected.

M2, attention shift. Wings attributes text-only forgetting to a change in how attention is distributed between text before and after inserted image tokens. Across more than 100 trained VLMs, a larger shift goes with worse text-only performance (Zhang et al., 2024). M2 is a mechanism about mixed inputs that carries over to text-only inputs through changed weights. It predicts that architectural additions which absorb visual attention protect text, which Wings shows with low-rank residual attention learners.

M3, output-format and task-inference drift. Fine-tuning shifts the model's implicit inference about which task a prompt belongs to toward the fine-tuning distribution, and teaches that distribution's output format, without removing the underlying ability (Kotha et al., 2024; Wang et al., 2024). In vision-language SFT the shifted distribution is short-answer VQA and image description. M3 predicts that the loss is fast and early, that it is concentrated on tasks whose required format is far from the fine-tuning targets, that it is recovered cheaply by format-appropriate data, prompting or preference tuning, and that it depends on the output distribution of the training data more than on its input modality.

M4, diffuse parameter drift. Many small changes spread across neurons cumulatively degrade language ability, while a few large changes carry the new visual

skill (Yu and Ananiadou, 2025). Forgetting grows with the number of parameters updated and the number of steps (Kalajdzievski, 2024), and restricting the update rank reduces it (Biderman et al., 2024a). M4 predicts that loss grows with update size, that sparse or neuron-selective merging back toward the source recovers it, and that the architecture matters only through how far the weights move.

The published trajectories bear on these predictions in four places.

Timing. In the anchor's matched 3B vision run (Figure 1b and 1c), IFEval falls by 23.1 points between step 0 and step 1,000, while the trained backbone's Sink Strength rises from $+0.41$ to $+0.54$. From step 3,000 to step 10,000, S falls, with fluctuation, from $+0.35$ to $+0.20$ while IFEval climbs from 34.2 to 43.6 (Choi et al., 2026). Within the run, then, the sink signature and the behaviour move apart: the largest loss arrives while S rises, and the recovery happens while S falls. The authors read this as S tracking "the regime the run ends in rather than the step-level fluctuation". That reading is consistent with S as a predictor of vulnerability measured on the base. It does not fit M1 as a causal chain inside the run, where sink collapse should come before or alongside the behavioural loss. M3 predicts the observed shape: format specialisation happens at the very beginning of fine-tuning (Wang et al., 2024), and spurious forgetting shows early disruption of task alignment with partial recovery as training continues (Zheng et al., 2025; Palit et al., 2026). Palit et al. (2026) describe forgetting that "undoes itself": recall collapses, recovers while training continues on new material alone, and only later erodes. In their model, normalisation withdraws a shared displacement once the new material is learned. That account fits the IFEval curve in Figure 1b in shape, though it was derived for facts rather than formats, and nobody has tested it on vision-language training. A non-monotone course is not new to multimodal fine-tuning either: on the vision side, Zhai et al. (2023b) found that early fine-tuning of LLaVA on one image dataset improved performance on others, and that generalisation was lost only as fine-tuning continued.

Direction versus magnitude. The anchor's random-direction control is a strong result against the naive form of M4. A random perturbation as large as the real update destroys IFEval (-61.37), while the real update costs 9.6 points (Choi et al., 2026). Whatever vision-language training does to the backbone, it does not do it by moving weights a given distance. It moves them in directions that preserve almost everything. Neuron-level M4, in which small widespread changes accumulate, survives this, because it is a claim about which directions do the damage rather than how far the weights travel.

Recovery. M1 predicts that recovery requires restoring the sink, which the anchor argues cannot be done cheaply after training. The recovery literature shows text behaviour returning by routes that do not obviously touch the sink. Text-blended fine-tuning restores MMLU (Lin et al., 2024). Image-only preference tuning restores conversational instruction following (Li et al., 2024a). Lineage-matched merging restores reasoning-task accuracy at small merge weights (Ratzlaff et al., 2024) and through task arithmetic or TIES on Llama3 (Yu and Ananiadou, 2025). In the spurious-forgetting experiments, training on as few as ten safety examples, or on unrelated tasks, restored performance on older tasks (Zheng et al., 2025). Translating a prompt into another language restores abilities that fine-tuning had suppressed (Kotha et al., 2024). These results do not measure IFEval, so they do not refute M1 on the anchor's headline task. They do show that, on the constructs they measure, most of what vision-language and other fine-tuning removes is access rather than capacity.

Architecture. M1 predicts protection from per-head QK normalisation, and the anchor's cross-pair data fit that prediction. The other mechanisms predict no special protection from it. The evidence that would separate them is thin, and Section 11 takes it up.

A fair summary is that M1 is supported as a correlate across released backbones and contradicted in time within the one controlled run that measured both, that M3 predicts the controlled run's shape and the recovery literature but has not itself been measured on vision-language training, and that M2 and M4 each explain part of the remedy literature. None of the four has been tested with an intervention that holds the others fixed.

10. Recovery: why merging fails in one study and works in others

The anchor's negative merging result and the positive results in Ratzlaff et al. (2024), Yu and Ananiadou (2025) and Lee et al. (2024) look like a direct conflict. The four indices resolve most of it.

Lineage. Task arithmetic builds task vectors by subtracting a pretrained model's weights from the weights of the same model after fine-tuning (Ilharco et al., 2023). Weight averaging works because models fine-tuned from one pretrained checkpoint often lie in a single low-error basin (Wortsman et al., 2022). The positive merging studies satisfy this. Ratzlaff et al. (2024) merge LLaVA-1.6-Mistral with the Mistral model it was trained from, and Yu and Ananiadou (2025) merge LLaVA-NeXT models with the Mistral-7B and Llama3-8B models they were tuned from. The anchor merges released VLMs with references that, per Section 5, are comparison points rather than known parents. A task vector computed between siblings is not the vector a fine-tuning stage added, and arithmetic on it has no guarantee of landing anywhere useful.

Strength. The anchor's settings are the methods' standard strengths: task arithmetic and task singular vectors at a coefficient of 1, TIES at density 0.2, DARE at

a drop rate of 0.9 (Choi et al., 2026). The positive studies tune the strength. Ratzlaff et al. (2024) find that a weight proportion as small as 0.1 recovers most of the degraded language performance without hurting visual tasks, and in some cases improves them. That the anchor's mildest mixes, 95/5 and uniform averaging, were the only ones that did no further damage is consistent with this: the useful region is near the VLM end.

Method. Not every merge works even with lineage fixed. In Yu and Ananiadou (2025), on Llama3, the VLM backbone's language score is 57.14 against 62.91 for Llama3. Task arithmetic reaches 61.39 and TIES 62.43, while DARE (56.91) and DELLA (55.60) end below the unmerged VLM. The anchor likewise finds the sparsified merges most damaging. The agreement on DARE across two very different settings is a real finding: randomly dropping most of the delta parameters, which DARE was designed to do safely for homologous SFT models (Yu et al., 2024), does not carry over to multimodal deltas. TIES, which trims small changes and resolves sign conflicts (Yadav et al., 2023), behaves differently in the two settings, plausibly because lineage differs.

Construct. The positive merging studies measure multiple-choice and short-answer reasoning tasks. The anchor measures IFEval. A merge that restores knowledge and reasoning while leaving strict-format compliance broken would be consistent with all three studies. Nobody has published a lineage-matched merge evaluated on IFEval.

The question "can merging recover vision-language text loss?" therefore has no single answer in the literature. Under lineage-matched conditions, tuned strength and reasoning constructs, the published answer is mostly yes. Under sibling references, default strength and strict-format constructs, the anchor's answer is no. The two cells that would settle the matter, lineage-matched merging scored on IFEval and sibling merging scored on reasoning, are both empty.

Data-side recovery has fewer complications. Every study that adds text-only or format-diverse data to the vision-language stage reports less text loss or none (Lin et al., 2024; Lu et al., 2024; McKinzie et al., 2024; Tu et al., 2024; Dai et al., 2024). Cambrian-1 stresses the importance of balancing data sources and their ratios in instruction tuning (Tong et al., 2024). The anchor's own Tulu control shows text-only optimisation preserving the sink, and its authors frame data-side remedies as the mechanistically supported option whose cost is paid in data budget (Choi et al., 2026). Architectural remedies have positive published results as well: modality-specific low-rank learners (Zhang et al., 2024), freezing (Dai et al., 2024), and post-hoc sparse patches for multimodal fine-tuning that keep about 99 percent of original-task performance while replacing at most 10 percent of parameters (Zhu et al., 2024), though the last was evaluated on multimodal rather than text tasks. Continual-learning methods reduce linguistic forgetting by up to 15 percent over the standard LLaVA recipe (Srivastava et al., 2024).

11. What the QK-normalisation evidence licenses

The anchor's most quotable design implication is that per-head QK normalisation in the base model protects text ability under vision-language training, and that co-adaptation of the query and key projections during native pretraining, rather than the module's mere presence, is an important component of that protection (Choi et al., 2026). The evidence for the first half is a correlation across a small set of vendor lines. The evidence for the second half is one experiment whose design leaves a gap.

QK normalisation was introduced and adopted for training stability. Henry et al. (2020) proposed normalising queries and keys to keep the softmax from saturating, in low-resource translation. Dehghani et al. (2023) observed divergent training loss in ViT models of around 8B parameters, caused by very large attention logits producing near one-hot attention, and fixed it by normalising queries and keys. Wortsman et al. (2024) reproduced the attention-logit-growth instability in small models at high learning rates and found the same mitigation effective. Zhai et al. (2023a) linked low attention entropy to training instability. OLMo 2 adopts QK-norm to keep attention logits from growing too large (OLMo et al., 2024), and Qwen3 introduces it "to ensure stable training" (Yang et al., 2025). Nothing in this lineage concerns fine-tuning robustness, and the anchor's proposal that it also protects against vision-language damage is new.

Three pieces of evidence complicate the proposal. First, per-head QK normalisation does not remove sinks: Zhang et al. (2025) report that in recent models with query-key normalisation, sinks "remain just as prevalent". This is compatible with the anchor's account, in which normalisation makes the sink robust rather than absent, but it means S, not the module, is the operative variable, which the anchor's Molmo2-O case already shows. Second, a sink is not uniformly protective in continual learning. Bai et al. (2024) find that pretrained encoders that put high attention on ubiquitous sink tokens can over-smooth and interfere across sequential tasks, and that encouraging attention diversity away from sinks improves continual-learning performance. Their setting, BERT and RoBERTa classifiers, is far from VLM instruction tuning, so this is a caution against assuming that a stronger sink always means safer adaptation, not a contradiction. Third, attention designs that remove the sink are reported to improve stability. A head-specific sigmoid gate after attention mitigates the sink, in a design its authors call attention-sink-free, while improving training stability and long-context

extrapolation (Qiu et al., 2025), and replacing softmax with unnormalised sigmoid attention prevents sinks from emerging in models up to 1B parameters (Gu et al., 2025). If sink strength protected fine-tuned behaviour in general, sink-free designs would be expected to be fragile under fine-tuning. Nobody has tested whether they are.

The injection experiment has a design gap. The anchor adds per-head RMS normalisation, with unit scales, to Qwen2.5-3B-Instruct and runs the two-stage LLaVA recipe. In stage one only the projector is trained (Choi et al., 2026). Normalising queries and keys at unit scale changes every attention logit of a model that was trained without it, so the modified model computes a different function before any vision-language data arrives, and in stage one its new scales cannot adapt. The reported score of 11 on IFEval at the first evaluated checkpoint therefore mixes two effects: whatever protection the module fails to give, and whatever damage the surgery did by itself. Separating them needs the modified model's text scores before any vision-language training, or a short text-only re-adaptation of the injected scales before the vision stage. The paper as read reports neither. The claim that native co-adaptation matters may well be right, but this experiment cannot show it.

What the QK evidence licenses, then, is a screening heuristic rather than a design rule. On the backbones sampled, a reference with high S, which so far means one trained natively with per-head QK normalisation, has lost less strict-format instruction following under its vendors' VL recipes, measured under one protocol. Whether that is caused by the normalisation, by co-varying vendor choices, or by something S happens to index is open, as the anchor's limitations section says.

12. Consolidated evidence

Table 1 collects the measured values on which the argument rests. Each row gives one published result and its source. Signs follow the convention VLM minus reference, so negative values mean the VLM side scores lower.

Three things stand out in Table 1. The rows marked C, comparisons of released models against chosen references, span gaps from -18.7 to +21.4 points, and on two pairs the sign changes with the source. The rows marked M, where the reference is the starting point, show losses that are real but that respond strongly to the data recipe: the same Qwen2-72B-Instruct backbone ends 6.3 points down under one recipe and 4.3 points up under another. And the only rows that pair a lineage-matched vision-language arm with a matched text-only arm and log a mechanism statistic at every checkpoint are the anchor's two 3B trajectories. Other mechanism claims rest on correlations, across released C pairs (Sink Strength) or across self-trained lineage-matched models (Wings' attention shift; Zhang et al., 2024).

The Ovis2-8B row deserves separate mention. It shares its reference, and therefore its S, with Qwen2.5-VL and InternVL3, but its gap is -17.9 against their -9.6 and -8.9 (Choi et al., 2026). At fixed S, recipe variation spans about nine points, which is larger than the 4.2-point difference between the Qwen2.5-VL gap and the Qwen3-VL gap that the architectural account explains. The anchor's authors draw the same conclusion: S orders backbones by risk but does not predict the size of the loss a particular recipe will produce.

13. An attribution procedure

Algorithm 1 states the checks a claim of the form "vision-language training damaged text capability on task t, through mechanism m" would need to pass. The steps are ordered so that the cheap checks that dissolve a claim come first.

The procedure has three points. First, steps 1 to 8 decide whether a published gap is evidence about vision-language training at all. Most published gaps between released models stop at step 2, since the reference is not the initialisation. Some also stop at step 6: on the evidence in Figure 1a, InternVL3's GSM8K and MBPP gaps are protocol-bound. Neither label says the gap is unimportant. A user who swaps a text model for a VLM backbone, under the user's own protocol, experiences the gap whatever its cause. The labels say only that the gap cannot be cited as a loss.

Second, steps 9 to 12 replace the anchor's matched control with a three-arm design. The anchor's text arm used different data with a different output distribution. The middle arm, text-only data whose answers match the vision mix's length and format profile, separates modality from supervision format. If the format-matched text arm loses as much as the vision arm, the loss is a format effect; if it loses as little as the Tulu arm, the loss is a modality effect. The third arm measures how much of either a text blend prevents.

Third, steps 13 to 19 apply the access-capacity distinction and the timing test before any mechanism is credited. A mechanism earns credit when its signature moves before or with the behaviour, and when an intervention on the signature moves the behaviour. The anchor proposes such an intervention itself: freezing the layers below the first sink layer and projecting sink-head updates away from the sink direction, or disabling per-head normalisation during Qwen3's VL training to see whether IFEval collapses (Choi et al., 2026). Spurious-forgetting work supplies a matching test for M3: freezing bottom layers protects task alignment (Zheng et al., 2025), and removing one shared direction from each weight update restores old facts (Palit et al., 2026). The same interventions, applied to VL training and scored on IFEval, would test M1 and M3 against each other directly.

Table 1. Published text-capability gaps and losses for vision-language models, with the indices each result fixes. "Lineage" is M when the reference is the VLM's stated initialisation (lineage-matched) and C when it is a comparison point. All values are as reported by the cited source.

Source	Model pair or run	Lineage	Task	Reported value	Protocol
Choi et al. 2026	Qwen2.5-VL-7B vs Qwen2.5-7B-Instruct	C	IFEval strict	-9.6	lm-eval, 0-shot, template
Choi et al. 2026	Molmo2-O-7B vs Olmo-3-7B-Instruct	C	IFEval strict	-18.7	lm-eval, 0-shot, template
Choi et al. 2026	Ovis2-8B vs Qwen2.5-7B-Instruct	C	IFEval strict	-17.9 at $S = +1.18$	lm-eval, 0-shot, template
Choi et al. 2026	InternVL3-8B vs Qwen2.5-7B-Instruct	C	GSM8K / MBPP / MMLU	-12.5 / +21.4 / +3.0	lm-eval, 0-shot, template
Zhu et al. 2025	InternVL3-8B vs Qwen2.5-7B Chat	C (base-initialised)	GSM8K / MBPP / MMLU	83.1 vs 80.1 / 69.3 vs 74.3 / 77.3 vs 74.2	OpenCompass
Choi et al. 2026	Qwen3-VL-8B vs Qwen3-8B	C	IFEval strict	-5.4	lm-eval, 0-shot, template
Bai et al. 2025b	Qwen3-VL-8B-Instruct vs Qwen3-8B	C	IFEval	83.7 vs 83.0	vendor
Bai et al. 2025a	Qwen2.5-VL-72B vs Qwen2.5-72B	C	IFEval	86.3 vs 84.1	vendor
Dai et al. 2024	LLaVA-OV-72B / NVLM-D-72B vs Qwen2-72B-Instruct	M / M (as listed)	4-task mean	-6.3 / +4.3	vendor
Li et al. 2024a	LLaVA-1.5-13B vs Vicuna-1.5-13B	M	MT-Bench	5.99 vs 6.57; 6.73 after DPO	MT-Bench judge
Zhang et al. 2024	Qwen1.5 + SigLIP, LLaVA-style vs Qwen1.5	M	MMLU	-9.70	own suite
Lin et al. 2024	Llama-2-7B, caption-only pretraining	M	MMLU	46.0 to 28.8	own suite
Lin et al. 2024	Llama-2-7B, interleaved + text-blended SFT	M	MMLU	51.4 vs 51.2 (text SFT)	own suite
Yu and Ananiadou 2025	LLaVA-NeXT-Llama3 vs Llama3-8B	M	7-task language mean	57.14 vs 62.91; TIES 62.43	lm-eval (0-shot; GSM8K 8-shot CoT)
Choi et al. 2026	3B run, VL SFT, step 1k / 10k	M	IFEval strict	59.9 to 36.8 / 43.6	lm-eval, 0-shot, template
Choi et al. 2026	3B run, Tulu text SFT, step 1k / 10k	M	IFEval strict	59.9 to 53.1 / 55.8	lm-eval, 0-shot, template
Choi et al. 2026	Random update, Qwen2.5-VL-matched norm	M	IFEval strict	-61.37	lm-eval, 0-shot, template

Algorithm 1. Attributing a VLM text-capability gap

```
Input: VLM endpoint V, reference R, task set T, protocols P1, P2
1  if R is not V's stated initialisation:
2      label every gap "pipeline difference"; go to 9 for a loss
3  for each task t in T:
4      g1 <- score(V,t,P1) - score(R,t,P1)
5      g2 <- score(V,t,P2) - score(R,t,P2)
6      if sign(g1) != sign(g2): label t "protocol-bound"
7  report g1, g2 per task with protocol named; stop if all
8      tasks are protocol-bound
9  from R, run matched arms: VL mix; text mix with the VL mix's
10     output-length and format profile; VL mix + text blend
10b and a long-form text arm (e.g. Tulu) as the general-SFT baseline
11 L_vl, L_fmt, L_blend, L_tulu <- per-arm loss on t, same protocol
12 modality effect <- L_vl - L_fmt ; format effect <- L_fmt - L_tulu
13 for the VL arm, test access: (a) format exemplars in prompt,
14     (b) 100-step text SFT, (c) lineage-matched merge, swept weight
15 if (a), (b) or (c) restores t within tolerance: label "access"
16 else label "capacity"
17 mechanism test: log S, attention shift and format statistics
18     at every checkpoint; require the candidate to change before
19     or with the behaviour, and an intervention on it to move t
```

Applied to the anchor's six pairs as published, the procedure stops most of them at step 2. InternVL3 and InternVL3.5 are resolved against lineage by their vendors; the InternVL3.5 report compares its models with Qwen3 models "whose corresponding pre-trained base models are employed as the initialization of the language component" (Wang et al., 2025). The Qwen2.5-VL and Qwen3-VL pairs are unresolved, and so is Molmo2-O, which the anchor describes as built on Olmo-3 without naming the checkpoint (Choi et al., 2026). Only the LLaVA-OneVision pair has a third-party listing (Dai et al., 2024) suggesting that its reference is its starting point. That pair would proceed to steps 3 to 8, which need a second protocol the anchor has not tabulated. Applied to the anchor's 3B trajectories, it passes step 1 and the single-protocol half of steps 3 to 8 (steps 5 to 7 need a second protocol, which the run did not use), records a modality-plus-format effect of about

16 points at step 1,000 (a 23.1-point vision loss against a 6.8-point text loss), and stops at step 12 because the format-matched arm has not been run. That is a real finding about a data package. It is not yet a finding about modality or about the attention sink.

14. The strongest case against this reading

"**Users do not care about lineage. A VLM backbone that scores below the text model they would otherwise deploy is worse for them, whatever the cause.**" This is correct, and the paper does not dispute it. The anchor's practical recommendation, to screen candidate backbones before committing to vision-language training, is useful whatever the mechanism. A screening statistic needs only to predict. The objection does not reach the causal claims, and it does not make a sibling comparison into evidence about what training did.

"**The anchor's matched trajectory fixes lineage and protocol, and it still shows a large loss with a sink collapse. That is the controlled evidence.**" It shows a large loss, with lineage and protocol fixed. It does not show the sink collapse driving the loss. In the vision run, the largest single drop (step 0 to 1k) ends at the checkpoint where S is highest, and the recovery takes place while S falls (Figure 1b and 1c). The authors themselves say S tracks the end regime rather than the step-level behaviour. A predictor of vulnerability that is not a timed cause within the run is still useful, but it is a different claim from "VL fine-tuning perturbs the sink, and that is why format tasks fail".

"**Vendor reports are marketing, and their protocols are chosen to flatter.**" Possibly, and Section 6 treats them as one protocol among several rather than as ground truth. But the argument does not need vendor numbers to be right. It needs only the observation that on the same pair, two careful protocols disagree in sign on generative, extraction-graded tasks and agree on a letter-choice task. That pattern is a property of the evaluation situation whichever source is closer to the truth, and it matches what the prompt-format literature predicts (Sclar et al., 2024).

"**The format-drift account is a just-so story. Nobody has shown it in vision-language training.**" That is true, and this paper says so in Section 9: M3 predicts the published curves but has not been tested directly. The asymmetry is that M3's predictions about timing, recovery and data dependence are each matched by published results in text-only fine-tuning (Wang et al., 2024; Kotha et al., 2024; Zheng et al., 2025) and are each consistent with the vision-language data in hand. M1's timing prediction is contradicted by the one trajectory that logged both quantities. That makes M3 a stronger default hypothesis, not an established one.

"**Knowledge preservation in the anchor's panel shows that the loss is specific to format, which is what the sink account explains and the forgetting accounts do not.**" The format-drift account explains it too, and arguably better, since format specialisation is by definition a loss of format compliance with knowledge intact (Wang et al., 2024). The diffuse-drift and attention-shift accounts are the ones that struggle to explain a construct split, and earlier lineage-matched studies with thin text mixes did see knowledge loss (Zhang et al., 2024; Lin et al., 2024).

"**The random-direction control shows that the update is structured, which supports a specific mechanism such as sink corruption.**" It shows that the update is far from random. Every learned update is. It rules out the account in which any perturbation of a given size causes the observed damage. It does not discriminate among M1, M2, M3 and neuron-level M4, all of which are claims about structured updates.

15. Studies that would settle the open part

Five studies, all feasible with open models and public data, would close most of what remains.

A three-arm matched trajectory. From one instruct checkpoint, run vision-language SFT on the LLaVA-1.5 mix, text-only SFT on a set whose answers match that mix's length and format distribution, and text-only SFT on a long-form set such as Tulu. Log IFEval, MMLU, GSM8K and S at every 250 steps. This separates modality from supervision format (Algorithm 1, steps 9 to 12). If the format-matched arm loses as much as the vision arm, M3 is the main account. If it loses little, modality matters, and M1 or M2 gain support.

A two-protocol audit. Score every pair in the anchor's 17-pair panel under the anchor's instruct protocol, its community-default protocol and one vendor-style toolkit, and report the per-pair, per-task sign agreement. The anchor's authors have already built the pipeline. The per-pair results under the second protocol, which they describe as diagnostic, would show how much of the descriptive result is protocol-bound.

Lineage-matched merging on IFEval. Take models whose initialisation is stated and public, such as the LLaVA-1.6 and LLaVA-NeXT families, sweep merge weight and method, and score IFEval. If merging restores IFEval when the parent is the reference, the anchor's negative merging result is a lineage effect. If it does not, strict-format loss is harder to reverse than the reasoning losses that merging is known to recover.

A sink intervention with timing. Run the anchor's proposed head-selective protection, freezing layers below the first sink layer and projecting sink-head updates away from the sink direction, against the spurious-forgetting baseline of freezing the same number of bottom layers without any sink targeting (Zheng et al., 2025). If the sink-targeted version protects IFEval and the untargeted one does not, M1 gains causal support

that the current correlation cannot provide.

A clean injection. Repeat the QK-normalisation injection with a short text-only re-adaptation of the new scales before the vision stage, and report text scores at the point of injection. This separates "the module protects only when co-adapted in pretraining" from "the surgery damaged the model".

16. What is not known

It is not known whether vision-language training removes strict-format instruction-following capacity, or only access to it. No published study has run an access test on IFEval after vision-language training.

It is not known how much of any released VLM's text gap is due to vision-language training. For the released models in the anchor's panel, the starting checkpoint is in two cases stated to differ from the reference and in others not stated at all.

It is not known whether per-head QK normalisation protects fine-tuned behaviour. The evidence is a correlation across two vendor lines that differ in many other ways, and one injection experiment that cannot separate protection from surgical damage.

It is not known whether the attention sink causes the format loss. In the only run that logged both, the sink and the behaviour moved in opposite directions for most of training.

It is not known whether the result that knowledge tasks survive generalises beyond current release recipes. Earlier recipes with less text lost knowledge, and current ones may simply have added enough text to stop it.

17. Limitations

This analysis reads published papers and reports, and inherits their choices. Several sources are vendor technical reports whose evaluation details are partly undisclosed. The anchor is a recent preprint, read in its first arXiv version, and later versions may report the second-protocol results or the pre-training injection scores whose absence Sections 6 and 11 note. The lineage judgements in Section 5 rest on what each report says about initialisation. Where a report is ambiguous, as Qwen2.5-VL's is about base versus instruct initialisation, the paper treats lineage as unresolved rather than guessing. The four mechanisms are not an exhaustive list, and a combination of them is likely. Finally, the attribution procedure in Section 13 is a proposal; its thresholds (tolerance for recovery, step counts for access tests) would need to be fixed in advance by whoever runs it.

18. Conclusion

The published claim that vision-language training erodes a language model's text ability is a family of claims with four hidden parameters. Fixing lineage turns many comparisons of released models into comparisons between pipelines. Fixing protocol shows that, for at least one pair, the sign of a reasoning or coding gap depends on the grader. Fixing mixture shows that the size of the effect tracks how much text, and what kind of answers, the recipe contained. Fixing construct shows that, in current releases, the loss is concentrated on following instructions about answer form. When all four are fixed, a real loss remains. It is large, fast and early, and in the one run that timed it, it did not follow the attention sink.

The anchor's Sink Strength statistic is a useful, cheap screen for backbones under the protocol and recipes it was validated on, and its authors are careful to present it as a ranking tool rather than a predictor of magnitude. The causal story the statistic is offered with, in which sink corruption drives format loss, is not yet established. On the evidence in hand, a shift in the model's inferred task and output format explains the timing and the recovery at least as well. The experiment that would decide between them, a format-matched text arm run alongside the vision arm with both quantities logged at every checkpoint, is cheap by the standards of the field.

References

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., et al. (2025a). Qwen2.5-VL Technical Report. arXiv:2502.13923. doi:10.48550/arXiv.2502.13923.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., et al. (2025b). Qwen3-VL Technical Report. arXiv:2511.21631. doi:10.48550/arXiv.2511.21631.

Bai, X., Sun, Y., & Balasubramanian, N. (2024). Does RoBERTa Perform Better than BERT in Continual Learning: An Attention Sink Perspective. COLM 2024. arXiv:2410.05648. doi:10.48550/arXiv.2410.05648.

Barbero, F., Arroyo, A., Gu, X., Perivolaropoulos, C., Bronstein, M., Velickovic, P., & Pascanu, R. (2025). Why do LLMs attend to the first token? arXiv:2504.02732. doi:10.48550/arXiv.2504.02732.

Biderman, D., Portes, J., Gonzalez Ortiz, J. J., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., et al. (2024a). LoRA Learns Less and Forgets Less. Transactions on Machine Learning Research. arXiv:2405.09673. doi:10.48550/arXiv.2405.09673.

Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A. F., Ammanamanchi, P. S., et al. (2024b). Lessons from the Trenches on Reproducible Evaluation of Language Models. arXiv:2405.14782. doi:10.48550/arXiv.2405.14782.

Choi, M., Kim, G., & Kim, Y. G. (2026). Text Capability Loss in Vision-Language Adaptation: An Attention-Sink Diagnosis. EMNLP 2026. arXiv:2609.00746. doi:10.48550/arXiv.2609.00746.

Dai, W., Lee, N., Wang, B., Yang, Z., Liu, Z., Barker, J., Rintamaki, T., Shoeybi, M., et al. (2024). NVLM: Open Frontier-Class Multimodal LLMs. arXiv:2409.11402. doi:10.48550/arXiv.2409.11402.

Darcet, T., Oquab, M., Mairal, J., & Bojanowski, P. (2024). Vision Transformers Need Registers. ICLR 2024. arXiv:2309.16588. doi:10.48550/arXiv.2309.16588.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A., Caron, M., et al. (2023). Scaling Vision Transformers to 22 Billion Parameters. arXiv:2302.05442. doi:10.48550/arXiv.2302.05442.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., et al. (2023). PaLM-E: An

Embodied Multimodal Language Model. arXiv:2303.03378. doi:10.48550/arXiv.2303.03378.

Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., & Lin, M. (2025). When Attention Sink Emerges in Language Models: An Empirical View. ICLR 2025. arXiv:2410.10781. doi:10.48550/arXiv.2410.10781.

Henry, A., Dachapally, P. R., Pawar, S., & Chen, Y. (2020). Query-Key Normalization for Transformers. Findings of EMNLP 2020. arXiv:2010.04245. doi:10.48550/arXiv.2010.04245.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., & Farhadi, A. (2023). Editing Models with Task Arithmetic. ICLR 2023. arXiv:2212.04089. doi:10.48550/arXiv.2212.04089.

Kalajdzievski, D. (2024). Scaling Laws for Forgetting When Fine-Tuning Large Language Models. arXiv:2401.05605. doi:10.48550/arXiv.2401.05605.

Kang, S., Kim, J., Kim, J., & Hwang, S. J. (2025). See What You Are Told: Visual Attention Sink in Large Multimodal Models. arXiv:2503.03321. doi:10.48550/arXiv.2503.03321.

Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., & Sadigh, D. (2024). Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models. ICML 2024. arXiv:2402.07865. doi:10.48550/arXiv.2402.07865.

Kotha, S., Springer, J. M., & Raghunathan, A. (2024). Understanding Catastrophic Forgetting in Language Models via Implicit Inference. ICLR 2024. arXiv:2309.10105. doi:10.48550/arXiv.2309.10105.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., et al. (2025). Tulu 3: Pushing Frontiers in Open Language Model Post-Training. arXiv:2411.15124. doi:10.48550/arXiv.2411.15124.

Lee, S., Kim, G., Kim, J., Lee, H., Chang, H., Park, S. H., & Seo, M. (2024). How Does Vision-Language Adaptation Impact the Safety of Vision Language Models? arXiv:2410.07571. doi:10.48550/arXiv.2410.07571.

Li, S., Lin, R., & Pei, S. (2024a). Multi-modal Preference Alignment Remedies Degradation of Visual Instruction Tuning on Language Models. Proceedings of ACL 2024 (Volume 1: Long Papers), 14188-14200. arXiv:2402.10884. doi:10.48550/arXiv.2402.10884.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., et al. (2024b). LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326. doi:10.48550/arXiv.2408.03326.

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., & Choi, Y. (2023). The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. arXiv:2312.01552. doi:10.48550/arXiv.2312.01552.

Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., et al. (2024). VILA: On Pre-training for Visual Language Models. CVPR 2024. arXiv:2312.07533. doi:10.48550/arXiv.2312.07533.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning. NeurIPS 2023. arXiv:2304.08485. doi:10.48550/arXiv.2304.08485.

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved Baselines with Visual Instruction Tuning. CVPR 2024. arXiv:2310.03744. doi:10.48550/arXiv.2310.03744.

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., et al. (2024). DeepSeek-VL: Towards Real-World Vision-Language Understanding. arXiv:2403.05525. doi:10.48550/arXiv.2403.05525.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., & Zhang, Y. (2023). An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. arXiv:2308.08747. doi:10.48550/arXiv.2308.08747.

McKinzie, B., Gan, Z., Fauconnier, J.-P., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., et al. (2024). MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training. arXiv:2403.09611. doi:10.48550/arXiv.2403.09611.

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., et al. (2024). 2 OLMo 2 Furious. arXiv:2501.00656. doi:10.48550/arXiv.2501.00656.

Palit, V., Draye, F., Zucchet, N., Jin, Z., & Scholkopf, B. (2026). When Forgetting is not Catastrophic: On the Mechanics of Spurious Forgetting. arXiv:2610.08718. doi:10.48550/arXiv.2610.08718.

Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., et al. (2025). Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free. arXiv:2505.06708. doi:10.48550/arXiv.2505.06708.

Ratzlaff, N., Luo, M., Su, X., Lal, V., & Howard, P. (2024). Training-Free Mitigation of Language Reasoning Degradation After Multimodal Instruction Tuning. arXiv:2412.03467. doi:10.48550/arXiv.2412.03467.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. ICLR 2024. arXiv:2310.11324. doi:10.48550/arXiv.2310.11324.

Srivastava, S., Harun, M. Y., Shrestha, R., & Kanan, C. (2024). Improving Multimodal Large Language Models Using Continual Learning. CoLLAs 2025. arXiv:2410.19925. doi:10.48550/arXiv.2410.19925.

Sun, M., Chen, X., Kolter, J. Z., & Liu, Z. (2024). Massive Activations in Large Language Models. COLM 2024. arXiv:2402.17762. doi:10.48550/arXiv.2402.17762.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., et al. (2024). Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. NeurIPS 2024. arXiv:2406.16860. doi:10.48550/arXiv.2406.16860.

Tu, J., Ni, Z., Crispino, N., Yu, Z., Bendersky, M., Gunel, B., Jia, R., Liu, X., et al. (2024). MLAN: Language-Based Instruction Tuning Preserves and Transfers Knowledge in Multimodal Language Models. arXiv:2411.10557. doi:10.48550/arXiv.2411.10557.

Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., et al. (2025). InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. arXiv:2508.18265. doi:10.48550/arXiv.2508.18265.

Wang, Y., Si, S., Li, D., Lukasik, M., Yu, F., Hsieh, C.-J., Dhillon, I. S., & Kumar, S. (2024). Two-stage LLM Fine-tuning with Less Specialization and More Generalization. ICLR 2024. arXiv:2211.00635. doi:10.48550/arXiv.2211.00635.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., et al. (2022). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. ICML 2022. arXiv:2203.05482. doi:10.48550/arXiv.2203.05482.

Wortsman, M., Liu, P. J., Xiao, L., Everett, K., Alemi, A., Adlam, B., Co-Reyes, J. D., Gur, I., et al. (2024). Small-scale proxies for large-scale Transformer training instabilities. ICLR 2024. arXiv:2309.14322. doi:10.48550/arXiv.2309.14322.

Xiao, G., Tian, Y., Chen, B., Han, S., & Lewis, M. (2024). Efficient Streaming Language Models with Attention Sinks. ICLR 2024. arXiv:2309.17453. doi:10.48550/arXiv.2309.17453.

Yadav, P., Tam, D., Choshen, L., Raffel, C., & Bansal, M. (2023). TIES-Merging: Resolving Interference When Merging Models. NeurIPS 2023. arXiv:2306.01708. doi:10.48550/arXiv.2306.01708.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., et al. (2025). Qwen3 Technical Report. arXiv:2505.09388. doi:10.48550/arXiv.2505.09388.

Yu, L., Yu, B., Yu, H., Huang, F., & Li, Y. (2024). Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. ICML 2024. arXiv:2311.03099. doi:10.48550/arXiv.2311.03099.

Yu, Z., & Ananiadou, S. (2025). Locate-then-Merge: Neuron-Level Parameter Fusion for Mitigating Catastrophic Forgetting in Multimodal LLMs. Findings of EMNLP 2025. arXiv:2505.16703. doi:10.48550/arXiv.2505.16703.

Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., & Susskind, J. (2023a). Stabilizing Transformer Training by Preventing Attention Entropy Collapse. ICML 2023. arXiv:2303.06296. doi:10.48550/arXiv.2303.06296.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., & Ma, Y. (2023b). Investigating the Catastrophic Forgetting in Multimodal Large Language Models. arXiv:2309.10313. doi:10.48550/arXiv.2309.10313.

Zhang, S., Khan, M., & Papyan, V. (2025). Attention Sinks: A 'Catch, Tag, Release' Mechanism for Embeddings. arXiv:2502.00919. doi:10.48550/arXiv.2502.00919.

Zhang, Y.-K., Lu, S., Li, Y., Ma, Y., Chen, Q.-G., Xu, Z., Luo, W., Zhang, K., et al. (2024). Wings: Learning Multimodal LLMs without Text-only Forgetting. NeurIPS 2024. arXiv:2406.03496. doi:10.48550/arXiv.2406.03496.

Zheng, J., Cai, X., Qiu, S., & Ma, Q. (2025). Spurious Forgetting in Continual Learning of Language Models. ICLR 2025. arXiv:2501.13453. doi:10.48550/arXiv.2501.13453.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., et al. (2023b). LIMA: Less Is More for Alignment. arXiv:2305.11206. doi:10.48550/arXiv.2305.11206.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023a). Instruction-Following Evaluation for Large Language Models. arXiv:2311.07911. doi:10.48550/arXiv.2311.07911.

Zhu, D., Sun, Z., Li, Z., Shen, T., Yan, K., Ding, S., Kuang, K., & Wu, C. (2024). Model Tailor: Mitigating Catastrophic Forgetting in Multi-modal Large Language Models. arXiv:2402.12048. doi:10.48550/arXiv.2402.12048.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., et al. (2025). InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479. doi:10.48550/arXiv.2504.10479.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23858880