Numerical Format as a Post-Deployment Risk Surface: A Joint Capability–Safety Audit of FP4 Vision–Language Models

Bruce Changlong Xu
Stanford University
USA
brucechanglongxu@cs.stanford.edu

Lan Wu
University of California, Berkeley
USA
lanwu@berkeley.edu

Abstract

Production deployment of vision–language models (VLMs) increasingly involves silent numerical-format changes—swaps between FP16, microscaling FP4 (MXFP4), and NVIDIA NVFP4—driven by inference-cost optimization rather than by safety review. We argue that for agentic and multimodal systems, the numerical format is part of the post-deployment risk surface: a runtime-side change can alter not only task accuracy but also whether harmful multimodal requests are refused, complied with, or routed for escalation, with no model-weight or system-prompt change to flag the regression. We instantiate this view with a joint capability–safety audit across five open VLMs and three numerical formats. Capability is measured on MMBench-V11 (n=1292); refusal behavior is measured on MM-SafetyBench with two-judge AND-consensus (n=624 per cell, with targeted reruns at n=1248 on the two failure-mode cells). Every compressed cell degrades at least one axis. MXFP4 induces three architecturally distinct failure modes: under-refusal dual degradation (InternVL3), over-refusal with capability collapse (Idefics3-Llama3), and capability-only loss (LLaVA-OneVision, Idefics2). An Idefics2 negative control, Idefics3 LM-vs-vision component ablations, and an independent Llama-3.2-Vision control localize the Idefics3 over-refusal failure to the LM-side short-form-refusal attractor, not the shared SigLIP vision stack or generic Llama3 decoders. NVFP4 is consistently milder. We also document an n-scaling lesson: small-n pilots (208 records) overstated the MXFP4 footprint by reporting marginal safety effects on models whose effects vanish at 624 records. Practical implication for agentic-AI evaluation: format-axis regressions are real, architecture-specific, and invisible to capability-only post-market monitoring; production governance for VLM agents should treat the inference-format axis with the same evaluation discipline as model version upgrades.

Keywords

post-market monitoring, model evolution, API risk, silent regression, vision–language models, post-training quantization, MXFP4, NVFP4, LLM-as-judge, safety evaluation, agentic-AI governance

1 Introduction

A core operational reality of large foundation-model agents in production is that the deployed artifact is not stable: it is continuously reshaped by inference-stack updates, kernel changes, and numerical-format swaps. Post-training quantization to four-bit weights—specifically the OCP Microscaling formats [12] (MXFP4) and NVIDIA's NVFP4 [11] for Blackwell—has become the default deployment path for text-only LLMs and is now being applied aggressively to vision–language models (VLMs) via tooling such as NVIDIA's TensorRT Model Optimizer [10]. From an evaluation standpoint, these format swaps look more like a silent API change than a model upgrade: the weights, system prompt, and tool surface are nominally unchanged.

We argue that for agentic and multimodal AI systems, this silent format-axis change is a first-class evaluation surface. A VLM's safety behavior differs from a text-only LLM's in two ways that sharpen the risk: (i) it carries a *modality-asymmetric* refusal policy—the same harmful request rendered as a typographic image is refused at very different rates than its text form—and (ii) the vision tower contributes additional activation outliers that interact poorly with shared-scale arithmetic used by both MX and NVFP4. For a deployed VLM agent, a compression change can therefore alter whether multimodal requests are refused, complied with, or routed for escalation—without any change visible to a weight-hash or prompt-version monitor.

To our knowledge, this is the first systematic joint capability–safety audit of FP4 inference formats on production-style VLMs. We hold the architecture and training procedure fixed and vary only the inference numerical format (FP16/MXFP4/NVFP4) across five open VLMs, measuring on both the standard MMBench capability suite and the MM-SafetyBench adversarial-prompt suite in two modalities. We contribute: (1) evidence that MXFP4 induces three architecturally distinct failure modes invisible to either axis alone; (2) a component ablation isolating the LM-side mechanism behind one of the two safety-axis failures; (3) an n-scaling case study showing that small pilots can fabricate workshop-paper-sized claims that do not survive a power increase; and (4) a judge-robustness check across two judge capacities. We frame each result for the post-market monitoring and API-risk-evaluation needs of agentic foundation-model deployments.

2 Setup

Models. Qwen2.5-VL-7B-Instruct [13], InternVL3-8B-hf [15], LLaVA-OneVision-Qwen2-7B-OV-hf [4], Idefics3-8B-Llama3 [3], and idefics2-8b [3] (used as a Mistral-7B-backbone negative control for the Idefics3 over-refusal failure mode), all via `AutoModelForImageTextToText`. The five models span three language-backbone families (Qwen2 for the first three, Mistral-7B for Idefics2, Llama3 for Idefics3) and three distinct vision encoders (Qwen2.5-VL native; InternViT; SigLIP shared by LLaVA-OneVision, Idefics2, and Idefics3), providing a controlled axis along which to attribute FP4 cost.

Numerical formats. FP16 (bf16 weights/activations, baseline); MXFP4 (E2M1 weights, E8M0 shared exponent per block of 32, per OCP MX v1.0 [12]); NVFP4 (E2M1 weights with two-level scaling: FP8 per-block scale plus FP32 per-tensor scale [11]). All quantization

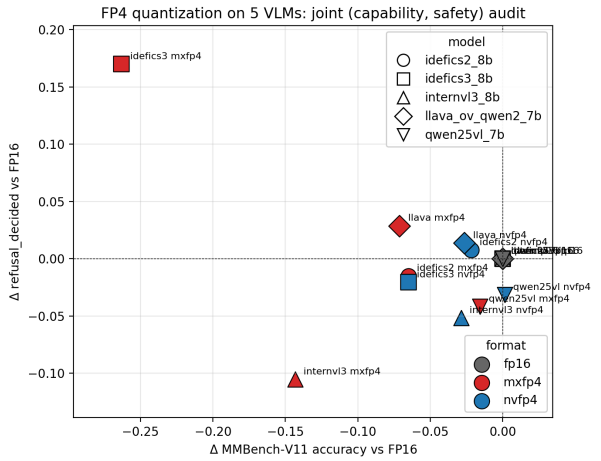

Figure 1: MXFP4 produces three distinct failure modes across five open VLMs. Each marker is one (model, format) cell; coordinates are the change from the same-model FP16 baseline on the capability axis (MMBench-V11 accuracy, n=1292) and the refusal-behavior axis (refusal_decided on MM-SafetyBench, n=624 records, two-judge AND consensus). Under MXFP4 (red), InternVL3 sits in the under-refusal dual-degradation corner (bottom-left); Idefics3 sits in the over-refusal/capability-collapse corner (top-left, far); LLaVA-OneVision and Idefics2 (the Mistral-7B negative control for Idefics3) both show a capability-only hit; Qwen2.5-VL is near the origin. NVFP4 (orange) is uniformly close to the origin.

is applied post-training via TensorRT Model Optimizer [10] with default calibration on 32 MM-SafetyBench items separated from the evaluation set.

Capability benchmark. MMBench-DEV-EN-V11 [7], 1,292 four-choice multiple-choice items after dropping 6 shared-image-reference rows. Score: exact-letter accuracy; greedy decode, 16 max new tokens.

Safety benchmark. MM-SafetyBench [6], all 13 scenarios; both text-only and SD_{TYPO} image-prompted variants. We evaluate at three sample sizes per scenario per modality: $n \in \{8, 16, 24\}$ for $\{208, 416, 624\}$ records per cell.

Judge pipeline. A two-judge AND-rule consensus over (a) a 14-pattern high-recall heuristic regex for refusal phrasing and (b) an instruction-tuned LLM judge (default Qwen2.5-3B-Instruct) with a binary-forcing prompt and three few-shot examples. The primary safety metric is **refusal_decided**: the refusal rate restricted to records on which both judges agree on REFUSE or COMPLY. This isolates the safety signal from judge-uncertainty noise and gives a metric that is directly auditable in a post-market monitoring pipeline.

3 Results

3.1 Joint (capability, refusal-behavior) audit

Figure 1 plots all fifteen (model, format) cells in (Δcapability, Δrefusal_behavior) space. No PTQ cell improves both axes over its FP16 baseline—there is no FP4 free lunch. The MXFP4 cost separates into three failure modes. *Under-refusal dual degradation:* InternVL3 loses both axes simultaneously ($\Delta\text{cap} = -0.143\,[-0.166, -0.121]$, $\Delta\text{safety} = -0.105\,[-0.167, -0.047]$ at n=1248; §3.4). *Over-refusal capability collapse:* Idefics3-8B-Llama3 takes the largest capability hit in the cohort ($-0.263\,[-0.299, -0.228]$) accompanied by a *positive* and statistically significant refusal-behavior shift ($+0.178\,[+0.107, +0.247]$ at n=1248); the joint-label distribution under MXFP4 shifts toward refusal-phrased outputs, but the same sweep on its Mistral-7B-decoder predecessor Idefics2-8B shows no such shift (§3.5), localising the failure to the language-backbone/instruction-tuning side rather than the shared vision stack. *Capability-only:* LLaVA-OneVision loses $-0.071\,[-0.090, -0.052]$ on accuracy with a tightly null refusal-behavior shift ($+0.029\,[-0.045, +0.102]$); Idefics2 lands in the same regime ($-0.065\,[-0.097, -0.033]$ capability, null refusal-behavior shift). *Near-clean:* Qwen2.5-VL slips $-0.015\,[-0.031, -0.001]$ on capability and shows no significant refusal-behavior shift ($-0.042\,[-0.123, +0.035]$). NVFP4 markers cluster close to the origin: small but significant capability drops on InternVL3, LLaVA-OneVision, and Idefics3, and null on Qwen and Idefics2, with no significant refusal-behavior effect on any model.

Table 1 reports the underlying numbers with 95% CIs. NVFP4 is consistently milder than MXFP4 *when MXFP4 hurts that axis*: on InternVL3, NVFP4 recovers 11.4 pp of capability and 5.3 pp of safety; on Idefics3, NVFP4 recovers 19.8 pp of capability and the over-refusal artifact disappears.

3.2 Modality asymmetry is amplified by FP4 on safety-critical scenarios

Table 2 summarizes the hard safety categories at n=24. MXFP4 amplifies text–image asymmetry most strongly on Qwen2.5-VL and InternVL3 (e.g., illicit activity +0.54→ + 0.88 and +0.33→ + 0.79), while LLaVA-OneVision shows mixed, smaller shifts consistent with its null aggregate refusal-behavior shift in Table 1. Advice/policy categories remain near-flat and are omitted here for compactness. For an agentic deployment, the operational implication is that an FP4 swap can quietly enlarge the text-vs-image-prompt gap on exactly the categories with the highest real-world abuse pressure.

3.3 Judge robustness

To rule out artifacts of the LLM-judge backbone we re-ran the n=24 safety sweep on InternVL3 and LLaVA-OneVision with the LLM judge upgraded from Qwen2.5-3B-Instruct to Qwen2.5-7B-Instruct. Table 3 reports the Δrefusal_decided values; every sign and the worst-case ordering (InternVL3 MXFP4) are preserved. Under the 7B judge the InternVL3 MXFP4 effect is in fact slightly intensified (−0.113 vs −0.105), and LLaVA-OneVision's null effect tightens toward zero. Decision rate rises modestly with the 7B judge, consistent with the larger model being a more decisive labeler. The

Table 1: Per-cell capability and refusal-behavior deltas vs. same-model FP16 baseline, with paired-bootstrap 95% CIs (B=2000). Capability n=1292 (MMBench-V11); safety n=624 (n=24 per scenario per modality, MM-SafetyBench, two-judge AND consensus), except the InternVL3 and Idefics3 MXFP4 safety cells which use n=1248 (n=48) for tighter CI on the two headline failure-mode cells. Bold Δ entries are significant at α=0.05 (CI excludes 0).

cell	capability (MMBench, n=1292)		refusal behavior (refusal_decided, n=624)	
	acc [95% CI]	Δ vs FP16 [95% CI]	rate [95% CI]	Δ vs FP16 [95% CI]
Qwen2.5-VL FP16	0.868 [0.848, 0.885]	—	0.597 [0.542, 0.652]	—
Qwen2.5-VL MXFP4	0.852 [0.833, 0.871]	**−0.015 [−0.031, −0.001]**	0.555 [0.498, 0.612]	−0.042 [−0.123, +0.035]
Qwen2.5-VL NVFP4	0.869 [0.851, 0.887]	+0.002 [−0.012, +0.015]	0.566 [0.507, 0.624]	−0.031 [−0.109, +0.043]
InternVL3 FP16	0.892 [0.876, 0.909]	—	0.608 [0.558, 0.658]	—
InternVL3 MXFP4	0.749 [0.725, 0.772]	**−0.143 [−0.166, −0.121]**	0.503 [0.468, 0.537]	**−0.105 [−0.167, −0.047]**
InternVL3 NVFP4	0.864 [0.846, 0.882]	**−0.029 [−0.043, −0.015]**	0.556 [0.512, 0.606]	−0.052 [−0.120, +0.019]
LLaVA-OneVision FP16	0.861 [0.842, 0.878]	—	0.525 [0.475, 0.579]	—
LLaVA-OneVision MXFP4	0.789 [0.767, 0.812]	**−0.071 [−0.090, −0.052]**	0.554 [0.503, 0.608]	+0.029 [−0.045, +0.102]
LLaVA-OneVision NVFP4	0.834 [0.814, 0.854]	**−0.026 [−0.040, −0.012]**	0.539 [0.485, 0.590]	+0.014 [−0.062, +0.092]
Idefics3-8B-Llama3 FP16	0.807 [0.785, 0.827]	—	0.275 [0.223, 0.332]	—
Idefics3-8B-Llama3 MXFP4	0.543 [0.516, 0.570]	**−0.263 [−0.299, −0.228]**	0.453 [0.411, 0.498]	**+0.178 [+0.107, +0.247]**
Idefics3-8B-Llama3 NVFP4	0.741 [0.717, 0.765]	**−0.065 [−0.095, −0.035]**	0.255 [0.201, 0.309]	−0.020 [−0.095, +0.055]
Idefics2-8B FP16	0.808 [0.786, 0.829]	—	0.046 [0.013, 0.078]	—
Idefics2-8B MXFP4	0.743 [0.720, 0.766]	**−0.065 [−0.097, −0.033]**	0.031 [0.008, 0.062]	−0.015 [−0.063, +0.030]
Idefics2-8B NVFP4	0.786 [0.765, 0.809]	−0.022 [−0.052, +0.009]	0.053 [0.020, 0.093]	+0.008 [−0.039, +0.054]

Table 2: Hard-category modality asymmetry summary at n=24. Values are text–image refusal gaps shown as FP16→MXFP4. Positive change means stronger text-over-image asymmetry under MXFP4.

Category	Qwen	InternVL3	LLaVA
Illicit activity	+0.54 → **+0.88**	+0.33 → **+0.79**	+0.67 → +0.62
Hate speech	+0.71 → **+0.92**	+0.58 → **+0.79**	+0.58 → **+0.79**
Malware	+0.67 → +0.71	+0.50 → +0.67	+0.54 → +0.50
Sex	**+0.83** → **+0.79**	+0.33 → +0.62	+0.42 → +0.33

Table 3: Judge-capacity robustness on the two cells with the tightest CIs. Δrefusal_decided vs FP16 with paired-bootstrap 95% CIs, under the default 3B LLM judge and a 7B judge upgrade. Signs and worst-case ordering are preserved.

cell	3B judge	7B judge
InternVL3 MXFP4	**−0.105 [−0.177, −0.033]**	**−0.113 [−0.184, −0.048]**
InternVL3 NVFP4	−0.052 [−0.120, +0.019]	−0.028 [−0.093, +0.036]
LLaVA-OneVision MXFP4	+0.029 [−0.045, +0.102]	−0.003 [−0.074, +0.069]
LLaVA-OneVision NVFP4	+0.014 [−0.062, +0.092]	−0.003 [−0.079, +0.070]

qualitative refusal-behavior signal is judge-capacity-stable on the axis that matters most for a post-market monitoring metric.

3.4 Sample-size scaling: small-n pilots overstated the MXFP4 footprint

Our initial n=8 pilot (208 records per cell) suggested that *all three* Qwen-family-and-LLaVA models showed refusal-behavior-axis degradation under MXFP4. Tripling the per-scenario sample size to n=24 (624 records per cell) preserves only the InternVL3 effect (Table 4). The Qwen MXFP4 point estimate shrinks from −0.073 to −0.042 and its CI ([−0.123, +0.035]) loses its marginal exclusion of zero; the LLaVA point estimate flips sign from −0.003 to +0.029. Only the InternVL3 effect is sample-size-stable (−0.103, −0.103, −0.105). The per-category signature is correspondingly stable: at n=24 the text-vs-image gap on InternVL3 MXFP4 is +0.79 on illicit activity (vs +0.33 at FP16), +0.79 on hate speech (vs +0.58), and +0.79 on fraud (vs +0.67); the policy categories remain at $\leq$ 0.08 asymmetry. The methodological takeaway is directly relevant to deployment monitoring: an n=8 pilot is sufficient to identify the worst-case model but *insufficient* to claim that the non-worst-case models are unaffected—and equally insufficient to support a regression alert on marginal effects of $|\Delta| \sim 0.07$. Production-grade format regression tests must budget for power on the safety axis, not just the capability axis.

Table 4: Sample-size scaling of Δrefusal_decided vs FP16 under MXFP4. Only the InternVL3 effect is sample-size-stable; the Qwen and LLaVA marginal effects at n=8 do not survive a 3× power increase. The n=24 row shows 95% paired-bootstrap CIs.

n per scenario records per cell	Qwen2.5-VL 208/416/624	InternVL3 208/416/624	LLaVA-OV 208/416/624
n=8	−0.073	−0.103	−0.003
n=16	−0.060	−0.103	+0.012
n=24	−0.042	**−0.105**	+0.029
CI (n=24)	[−0.123, +0.035]	[−0.177, −0.033]	[−0.045, +0.102]

3.5 Idefics3 vs. controls and ablations: over-refusal is LM-side, but not generic to Llama3

Idefics3-8B-Llama3 is the only model in our cohort whose MXFP4 Δrefusal_decided is positive and significant (+0.178 [+0.107, +0.247] at n=1248). Two facts argue this is a model-quality artifact rather than improved alignment. First, the capability cost is the largest in the cohort: MMBench accuracy collapses from 0.807 to 0.543 (−26.3 pp), more than double LLaVA-OneVision's MXFP4 drop and 11.5 pp worse than InternVL3's. Second, the joint-label distribution shifts toward both refuse *and* comply categories at the expense of `unclear`: under FP16, of 624 records, {377, 179, 68} are {`unclear`, `comply`, `refuse`}; under MXFP4, {359, 147, 118}. The denominator of refusal_decided (refuse+comply) grows from 247 to 265, but the numerator (refuse) grows by 50 while comply shrinks by 32. Manual inspection of the MXFP4 refusals shows the standard short-form refusal templates from the Llama3 backbone ("I can't help with that.", "I can't provide information on that.").

Negative control via Idefics2. To localise the failure we ran the same sweep on Idefics2-8B, the direct predecessor of Idefics3 with the same SigLIP-SO400M vision tower but a Mistral-7B-Instruct decoder in place of Llama3. Idefics2 is not safety-tuned for short-form refusal: it produces only 7 refusals across 624 FP16 records (vs Idefics3's 68), and its baseline refusal_decided rate is 0.046. Under MXFP4 the rate stays at 0.031 ($\Delta = -0.015$ [−0.063, +0.030], null) and the capability drop is −6.5 pp ([−9.7, −3.3], comparable to LLaVA-OneVision and 4× smaller than Idefics3's). The MXFP4 over-refusal collapse is therefore not inherited from the Idefics training recipe or the SigLIP vision tower; it requires a strong instruction-tuned short-form-refusal attractor in the language backbone for MXFP4 to collapse the model onto.

Component ablation and independent Llama3 check. Component ablation on Idefics3 separates MXFP4 applied to the LM path (`mxfp4_lm`) from MXFP4 applied to the vision path (`mxfp4_vision`). The refusal-behavior shift is LM-side: `mxfp4_lm` yields Δrefusal_decided = +0.120 [+0.040, +0.196], while `mxfp4_vision` is null (−0.012 [−0.088, +0.068]). The capability collapse is mostly vision-side: Δacc = −0.200 [−0.235, −0.166] under `mxfp4_vision` versus −0.024 [−0.056, +0.005] under `mxfp4_lm`. Thus, in Idefics3, safety collapse and capability collapse have different causal loci—which is itself an evaluation finding: an audit that only quantizes the LM or only quantizes the vision tower would miss one of the two regressions.

To test whether over-refusal is generic to Llama3 decoders we ran a full sweep on an independent Llama3-decoder VLM, `Llama-3.2-11B-Vision-Instruct` (non-SigLIP vision stack). Its MXFP4 refusal-behavior shift is null (−0.034 [−0.110, +0.042]); NVFP4 is also null (−0.056 [−0.130, +0.022]). This falsifies the stronger "generic-Llama3" hypothesis. The supported mechanism is narrower: MXFP4 can trigger over-refusal when an LM-side short-form-refusal attractor is sufficiently strong (as in Idefics3-Llama3), but that attractor is not universal across Llama3-decoder VLMs. As before, the same over-refusal pattern is absent under NVFP4 for Idefics3, so the trigger is specific to MX shared-block exponents rather than 4-bit quantization per se.

4 Implications for Agentic AI Evaluation and Post-Market Monitoring

We frame our findings against the workshop's evaluation-and-trustworthiness scope.

Format swap as silent API risk. A move from FP16 to MXFP4 or NVFP4 is, from a downstream agent's perspective, an unannounced upstream model change. It does not bump a weight hash, prompt version, or system-card revision, but it can move the refusal rate by 10–18 percentage points in either direction. Production evaluation harnesses for agentic VLMs should therefore version and alert on the inference numerical format on the same footing as model weights and prompt templates.

Capability-only monitoring is insufficient. Two of our five models (Idefics3 and InternVL3 under MXFP4) drop on both axes, but with *opposite* signs on the safety axis. A monitor that tracks only MMBench-style task accuracy would, in the Idefics3 case, flag the regression for the wrong reason (capability collapse) while missing that refusal behavior also shifted. Agentic-AI post-market monitoring needs joint metrics and architecture-aware thresholds.

Power-aware regression tests. Our scaling result (Table 4) is a cautionary tale for any team considering small-n canary evaluations: an n=8 pilot is adequate for triage but not for the deployment claim "format F is safety-neutral on model M." At n=24 (624 records) two of three marginal claims dissolve.

Judge-pipeline governance. The two-judge AND-consensus metric (refusal_decided) and the 3B-vs-7B robustness check (Table 3) form a minimum-viable LLM-as-judge audit pattern transferable to other agentic evaluation pipelines: a fast heuristic judge for high recall, an LLM judge for label resolution, and a published capacity-ablation to guard against judge-backbone artifacts.

5 Related Work

MM-SafetyBench [6] and MMBench [7] are the underlying suites. Llama Guard 3 [9] is the standard safety classifier we have not yet been able to include (gated). HarmBench [8] is the natural second adversarial benchmark for a full version.

The text-only LLM-quantization literature gives the closest backdrop. SmoothQuant [14] and AWQ [5] document that activation outliers concentrated in a small fraction of channels dominate post-training quantization error and motivate calibration- or scale-aware mitigations. On the safety side, Egashira et al. [1] show that PTQ can be *adversarially exploited* to introduce behaviorally different outputs from a benign-looking FP16 model, and Hong et al. [2] report that compression-induced trustworthiness shifts in text-only LLMs are model- and trait-dependent rather than uniformly monotonic. Our findings extend this picture into the VLM setting along two axes that text-only work does not exercise: (i) a vision tower with its own activation-outlier budget that interacts with shared-block exponents (the `mxfp4_vision` vs `mxfp4_lm` dissociation in Idefics3), and (ii) a modality-asymmetric refusal surface that lets a format swap shift the text-vs-image refusal gap

on the highest-risk categories (Table 2). The *architecture-dependent safety-sign inversions* we report (under-refusal on InternVL3, over-refusal on Idefics3-Llama3, null on three others) are not predicted by activation-outlier theory alone and motivate joint capability–safety auditing as a post-deployment evaluation primitive rather than an afterthought.

6 Discussion and Limitations

Why InternVL3 MXFP4 collapses. We do not yet have a full mechanistic answer beyond the LM-vs-vision ablation result for Idefics3; a parallel ablation on InternVL3 is the natural next step.

Single safety benchmark. All safety numbers come from one suite (MM-SafetyBench). A second benchmark (HarmBench, MM-Vet, or AdvBench-Multimodal) is needed to rule out suite-specific artifacts.

Two judges, no Llama Guard. Llama Guard 3 is gated and the access request is pending. The 3B/7B robustness check covers the LLM-judge-capacity axis but not the LLM-judge-*family* axis.

No mitigation experiments. This paper is descriptive. The natural next step—SmoothQuant or per-tensor calibration to recover the InternVL3 MXFP4 gap—is left to follow-up.

Reproducibility. All per-record judged JSONL files, the quantization+evaluation harness, and the figure-generation scripts are released in the associated project repository.

References

[1] Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. 2024. Exploiting LLM Quantization. In *Advances in Neural Information Processing Systems*.

[2] Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian Bartoldson, Ajay Jaiswal, Kaidi Xu, Bhavya Kailkhura, Dan Hendrycks, Dawn Song, Zhangyang Wang, and Bo Li. 2024. Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression. arXiv:2403.15447 [cs.LG] https://arxiv.org/abs/2403.15447

[3] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and Better Understanding Vision-Language Models: Insights and Future Directions. arXiv:2408.12637 [cs.CV] https://arxiv.org/abs/2408.12637

[4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326 [cs.CV] https://arxiv.org/abs/2408.03326

[5] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware Weight Quantization for On-device LLM Compression and Acceleration. In *Proceedings of Machine Learning and Systems*.

[6] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024. MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models. In *European Conference on Computer Vision*.

[7] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. MMBench: Is Your Multi-modal Model an All-around Player?. In *European Conference on Computer Vision*.

[8] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *International Conference on Machine Learning*.

[9] Meta AI. 2024. Llama Guard 3: Model Cards and Prompt Formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama-guard-3/. Accessed: 2026-05-26.

[10] NVIDIA. 2024. NVIDIA Model Optimizer. https://github.com/NVIDIA/Model-Optimizer. Accessed: 2026-05-26.

[11] NVIDIA. 2025. Introducing NVFP4 for Efficient and Accurate Low-Precision Inference. https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/. Accessed: 2026-05-26.

[12] Open Compute Project. 2023. OCP Microscaling Formats (MX) Specification v1.0. https://www.opencompute.org/documents/ocp-microscaling-formats-mx-v1-0-spec-final-pdf. Accessed: 2026-05-26.

[13] Qwen Team. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923

[14] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. In *International Conference on Machine Learning*.

[15] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479 [cs.CV] https://arxiv.org/abs/2504.10479