Do Safety-Aligned Vision-Language Models Degrade Differently Under Common Image Corruptions?

Prince Mireku Kweku-Abeiku Attah-Anyen Annaliese Nartey
Nicole Nanka-Bruce Betty Blankson
Ashesi University

Abstract

Vision-language models are increasingly deployed in real-world settings where input images may be blurry, noisy, or otherwise corrupted, yet whether the safety-alignment procedures applied to these models help or harm their robustness to such corruptions remains poorly understood. Alignment is widely assumed to be net beneficial, but its interaction with perceptual robustness has not been systematically measured. We present the first controlled evaluation of this question, comparing four matched base/aligned model pairs that span three alignment paradigms; Mixed Preference Optimization (MPO), instruction tuning (DPO-style), and SFT+RLHF, across three model families and scales. We evaluate each pair on four benchmarks spanning spatial reasoning, general VQA, scene-text, and multi-discipline understanding under the ImageNet-C corruption suite, using Relative mean Corruption Error (Rel. mCE) to isolate alignment-induced changes in robustness while holding architecture and pre-training fixed within each pair. The effect is neither universally beneficial nor harmful: instruction tuning improves robustness for Qwen2-VL (Rel. mCE = 0.64) and Gemma-3-4B (0.90) but degrades it for Gemma-3-7B (1.19), while MPO yields mixed outcomes for InternVL2.5 (1.10). This is a preliminary study, and the divergent direction of effects across model families indicates that the alignment–robustness relationship cannot be assumed from any single pair; we release our evaluation pipeline and call for broader investigation across additional model families, scales, and alignment paradigms.

1. Introduction

Vision-language models (VLMs) have achieved strong performance on standard benchmarks [2, 3], but deployment in unconstrained environments exposes them to naturalistic image degradations such as sensor noise, atmospheric conditions, and compression artefacts. Separately, alignment training via RLHF, DPO, and Mixed Preference Optimisation (MPO) has become standard practice for producing helpful, harmless VLMs [4]. Yet the interaction between alignment and perceptual robustness has not been studied: does preference training alter how a model's task performance degrades when its visual inputs are corrupted?

Hendrycks and Dietterich [1] established the ImageNet-C benchmark as a standardised framework for evaluating corruption robustness, introducing the mean Corruption Error (mCE) as the primary metric. Recent work has extended this framework to VLMs [5, 6], but neither study isolates alignment as an independent variable. If the current over-specialization to clean distributions causes degraded performance under corruption, it would represent a systematic safety risk hidden beneath strong clean-data benchmark number.

We close this gap by evaluating four matched model pairs in which alignment training is the sole structural difference between base and aligned checkpoints. All pairs share vision encoder architecture, language backbone, and supervised fine-tuning (SFT) checkpoint (where applicable); see Section 3 for a discussion of the Gemma pre-training confound. Our contributions include:

- **C1** **First controlled base-vs-aligned evaluation under ImageNet-C**, spanning four model pairs and three alignment paradigms (MPO, instruction tuning/DPO, SFT+RLHF).
- **C2** **Identification and resolution of a verbosity-scoring mismatch** that artificially inflates the apparent capability gap between aligned and base models under standard VQA exact-match metrics, and a prompt-level fix that eliminates it.
- **C3** **Cross-method comparison** showing that alignment paradigm is a stronger predictor of robustness change than model scale, with instruction-tuned models consistently outperforming their base counterparts while MPO and RLHF-tuned models show dataset-dependent effects.
- **C4** **An ethical analysis** of the deployment implications arising from heterogeneous robustness profiles across alignment

methods.

2. Related Work

2.1. Corruption Robustness Benchmarks

Hendrycks and Dietterich [1] introduced ImageNet-C (15 corruption types, 5 severity levels) and the mCE metric as a standardized evaluation of model robustness to common corruptions. A key finding of their work is that raw accuracy improvements on clean data do not reliably transfer to robustness gains; models that are more accurate under clean conditions can still exhibit higher mCE, motivating the need for explicit corruption evaluation.

2.2. VLM Robustness

Usama *et al*. [5] conducted the first comprehensive VLM robustness analysis under the ImageNet-C framework, introducing TextVQA-C and GQA-C and finding task-specific vulnerability patterns: text recognition degrades most severely under blur and snow, while object reasoning is most sensitive to frost and impulse noise. VLM-RobustBench [6] extends evaluation to 11 open-weight models, including Qwen3-VL, InternVL3.5, Molmo2, and Gemma3, across 133 augmentation configurations, and raises the question of whether language-side reasoning can compensate for degraded visual inputs. Crucially, neither study compares base and aligned variants of the same model, leaving the effect of alignment training on corruption sensitivity unexplored.

2.3. Alignment and Adversarial Robustness

Carlini *et al*. [7] demonstrate that aligned VLMs remain vulnerable to gradient-based adversarial perturbations despite alignment training. Qi *et al*. [8] show that a single adversarially perturbed image can universally bypass safety filters across multiple aligned models. These results concern adversarial, worst-case perturbations; we study the complementary and more practically common setting of naturalistic, non-targeted corruptions from the ImageNet-C distribution. Liu *et al*. [9] introduce the term *safety alignment degradation*, showing that even adding a blank image or Gaussian noise to an input can break a VLM's refusal behavior, a direct empirical precursor to the task-performance axis studied here.

3. Methodology

3.1. Model Pairs

We evaluate four matched base/aligned model pairs. Within each pair, base and aligned models share identical vision encoder architecture, language backbone, and SFT checkpoint (where the base model has been SFT-tuned); alignment training is the sole varying factor. Table 1 summarizes all pairs.

Table 1. Model pairs evaluated. All pairs share vision encoder and backbone within each family. † Gemma-3-4B and Gemma-3-7B base models are raw pretrained models without SFT; see Section 4.5.

Pair	Base Model	Aligned Model	Method
P0	InternVL2.5-2B	InternVL2.5-2B-MPO	MPO
P1	Qwen2-VL-2B	Qwen2-VL-2B-Instruct	Instr. tuning
P2†	Gemma-3-4B-PT	Gemma-3-4B-IT	SFT+RLHF
P3†	Gemma-3-7B	Gemma-3-7B-IT	SFT+RLHF

3.2. Corruption Protocol

We apply four ImageNet-C corruption types [1]—Gaussian noise, defocus blur, brightness, and JPEG compression—at severity levels $\{1, 3, 5\}$ to all evaluation images (pilot shared suite). For Qwen2-VL (P1), we additionally evaluate an extended 11-corruption suite at all five severity levels; results are reported in Section 4.7. Corruptions are pre-generated at 224×224 resolution and saved as JPEGs before inference, ensuring deterministic and reproducible evaluation. The four pilot corruptions span the primary ImageNet-C families: noise (Gaussian), blur (defocus), weather (brightness), and digital (JPEG).

3.3. Datasets and Scoring

We evaluate on four VQA benchmarks: (1) **GQA** [10] for object and spatial reasoning, scored via case-insensitive word-boundary string containment to accommodate sentence-level model outputs; (2) **VQAv2** [11] for general-purpose visual QA, scored via soft match against annotator consensus answers; (3) **TextVQA** [12] for text-in-image VQA, scored via exact match against the ground-truth answer set; and (4) **MMBench** [13] for multiple-choice reasoning, scored via letter extraction (A/B/C/D) from free-form responses.

Verbosity fix. Instruction-tuned and aligned models generate verbose answers (*e.g.*, "There are two bike handles" vs. ground truth "2"), which standard exact-match scorers cannot match even when the answer is semantically correct. We append a concise-answer instruction: "Answer with a single word or short phrase only.", to all open-ended VQA prompts. Multiple-choice (MMBench) and containment-scored (GQA) datasets are unaffected. All reported results use corrected prompts. This fix constitutes Contribution C2 and is discussed further in Section 5.

3.4. Metrics

Following Hendrycks and Dietterich [1], for each corruption c at severities $s \in \{1, 3, 5\}$ we compute the Corruption Error:

$$\mathrm{CE}_c^f = \frac{\sum_s (1 - \mathrm{acc}_{s,c}^f)}{\sum_s (1 - \mathrm{acc}_{s,c}^{\mathrm{base}})}, \tag{1}$$

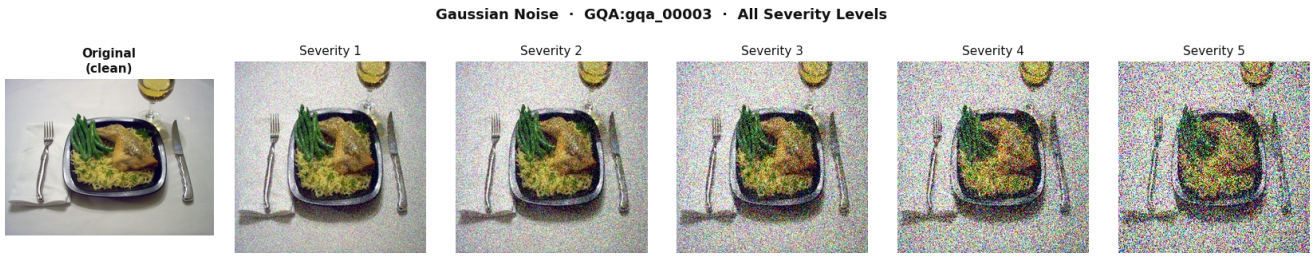

Figure 1. **Sample corruption (Gaussian noise) at the various severity levels.** The original clean image (left) is progressively degraded from mild grain at severity 1 to near-total loss of fine detail at severity 5, illustrating the perceptual range across which all models are evaluated.

where $\mathrm{acc}^f_{s,c}$ is model f's accuracy under corruption c at severity s, and the denominator uses the *base model's corrupted error* as a within-pair normaliser. This differs from the original ImageNet-C formulation (which uses AlexNet as a universal reference) and from the global-reference variant we compute in Section 4.2; it enables a cleaner intra-pair comparison independent of the reference model choice. Mean Corruption Error (mCE) averages CE over all corruption types. Our primary comparison statistic is:

$$\text{Rel. mCE} = \frac{\text{mCE(aligned)}}{\text{mCE(base)}}, \tag{2}$$

where values < 1 indicate that alignment *improves* robustness and values > 1 indicate that alignment *hurts* robustness.

3.5. Inference Setup

All experiments use greedy decoding (temperature $= 0$, `do_sample=False`, `max_new_tokens=64`) for reproducibility. InternVL2.5 (P0) experiments used an NVIDIA GTX 1650 (4 GB VRAM) with 4-bit NF4 quantisation via BitsAndBytes in `torch.float16`; $n = 50$ samples per dataset. Qwen2-VL (P1) and Gemma (P2, P3) experiments used Apple Silicon (MPS backend) in `bfloat16` with no quantisation and flash-attention disabled; $n = 10$ samples per dataset. Models are loaded once per (model, dataset) sweep to avoid redundant initialisation costs. A fixed random seed (`seed=42`) is used for all sample selection.

4. Experiments

4.1. Clean Accuracy

Table 2 reports clean (uncorrupted) accuracy for all eight models. Figure 2 shows these values as grouped bar charts. Results differ strikingly by pair. For **P0** (InternVL2.5/MPO), the base model outperforms the aligned model on all four datasets, with the largest gap on VQAv2 ($\Delta = -0.073$ pp). This is notable: MPO alignment does not improve task accuracy and slightly reduces it, suggesting that preference optimisation trades a small amount of task performance for other properties such as chain-of-thought reasoning quality [4].

Table 2. Clean accuracy. $n = 50$ for P0; $n = 10$ for P1–P3. Δ = aligned − base. † Base model is a raw pretrained checkpoint without SFT.

Model	GQA	VQAv2	TextVQA	MMBench
P0 — InternVL2.5-2B (MPO)				
Base (IVL-SFT)	0.58	0.72	0.66	0.58
Aligned (IVL-MPO)	0.56	0.65	0.64	0.52
Δ	−0.02	−0.07	−0.02	−0.06
P1 — Qwen2-VL-2B (Instruction tuning)				
Base (QW-base)	0.30	0.50	0.10	0.30
Aligned (QW-inst)	0.80	0.60	0.40	0.80
Δ	+0.50	+0.10	+0.30	+0.50
P2† — Gemma-3-4B (SFT+RLHF)				
Base (G4B-pt)	0.54	0.11	0.02	0.68
Aligned (G4B-it)	0.62	0.41	0.62	0.72
Δ	+0.08	+0.30	+0.60	+0.04
P3† — Gemma-3-7B (SFT+RLHF)				
Base (G7B-pt)	0.50	0.30	0.04	0.50
Aligned (G7B-it)	0.30	0.28	0.08	0.26
Δ	−0.20	−0.02	+0.04	−0.24

For **P1** (Qwen2-VL/instruction tuning), alignment produces large gains, particularly on GQA, TextVQA, and MMBench (+0.50 each), reflecting that the Qwen2-VL base model suffers from the verbosity confound on TextVQA (0.10 clean accuracy) even after our prompt fix, and from poor task-following on GQA and MMBench. For **P2** (Gemma-3-4B), the gap is most extreme on TextVQA (+0.60) and VQAv2 (+0.30), largely attributable to the raw pre-trained base model's inability to follow VQA task formats (see Section 4.5). For **P3** (Gemma-3-7B), the aligned model is *worse* on three of four datasets ($\Delta = -0.20$ on GQA, -0.24 on MMBench), which is unexpected and discussed in Section 4.8.

4.2. Mean Corruption Error

Table 3 reports mCE computed with the global reference (InternVL2.5-2B SFT as the universal reference model,

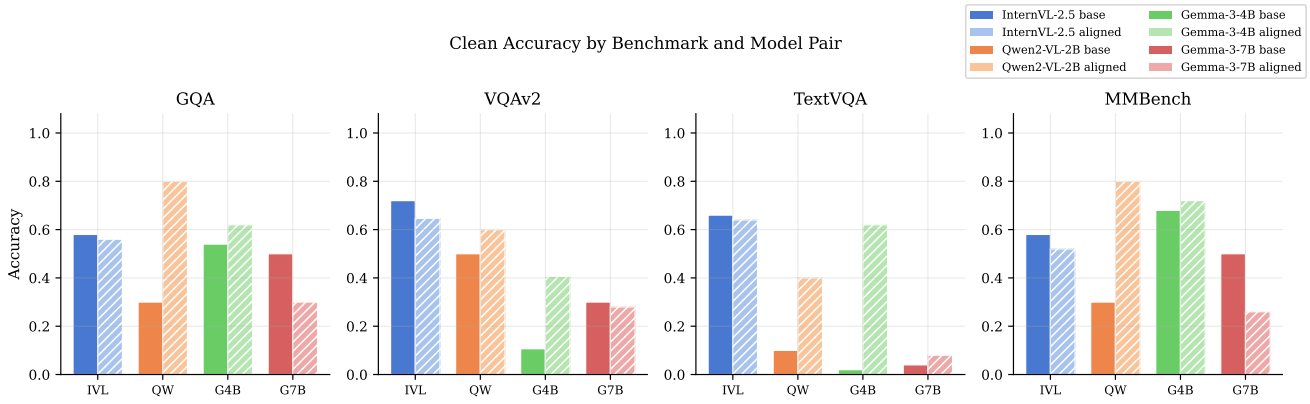

Figure 2. Clean accuracy by benchmark and model pair (base = solid, aligned = hatched). P0 is the only pair where alignment consistently reduces clean accuracy.

analogous to AlexNet in [1]). Figure 3 shows the per-benchmark comparison. Under this formulation, IVL-SFT has $\mathrm{mCE} = 1.0$ by construction across all datasets.

Table 3. mCE summary (global reference = InternVL2.5-2B SFT; lower = more robust). Macro-mCE is the unweighted mean over four datasets.

Model	GQA	VQAv2	TVA	MMB	Macro-MCE
IVL-SFT (ref)	1.000	1.000	1.000	1.000	1.000
QW-inst	0.846	1.202	1.072	1.207	1.082
IVL-MPO	0.830	1.150	1.005	1.415	1.100
G4B-it	0.941	1.925	0.994	0.978	1.209
G7B-pt	1.040	1.342	1.413	1.419	1.303
G4B-pt	1.236	2.059	1.402	0.824	1.380
G7B-it	1.213	1.879	1.130	1.987	1.552
QW-base	1.569	1.928	1.194	2.427	1.779

We made three main observations. First, both Qwen-Instruct and IVL-MPO have lower macro-mCE than their base counterparts in absolute terms, but the macro aggregation masks important per-dataset heterogeneity. Second, G4B-it achieves the best absolute mCE on TextVQA (0.994) and MMBench (0.978) among all non-reference models, while G4B-pt has the worst VQAv2 score (2.059), again partly reflecting the task-following confound. Third, and most striking, QW-base is the worst-performing model overall (macro-mCE $= 1.779$), while QW-inst is among the best (1.082); a reversal of 0.697 mCE points attributable entirely to instruction tuning.

4.3. Alignment Effect: Relative mCE

Table 4 reports the intra-pair Relative mCE (Eq. 2). Figure 6 in the appendix shows the per-benchmark breakdown.
The results contradict a simple "alignment hurts robustness" hypothesis. For P1 (instruction tuning), alignment *consistently improves* robustness across all four datasets (H1 supported 4/4), with particularly large gains on MMBench (0.497) and GQA (0.539). P2 (Gemma-4B SFT+RLHF) follows a similar pattern (3/4 datasets). In contrast, P0 (MPO) and P3 (Gemma-7B RLHF) show macro Rel. mCE above 1.0, indicating that alignment *increases* corruption sensitivity on balance. The divergence between P2 and P3 is notable: two models from the same family with the same alignment procedure at different scales produce qualitatively opposite results, suggesting that scale and capacity may mediate the alignment-robustness relationship in a non-monotonic way.

Table 4. Relative mCE (aligned/base) using the intra-pair formula (Eq. 2). Values < 1 (shown in blue) indicate alignment improves robustness. "H1 supp." counts datasets where alignment helps. $*$ QW TextVQA base clean accuracy $= 0.10$; mCE normalisation less reliable. † Gemma base models are raw pre-trained; see Section 4.5.

Pair	GQA	VQAv2	TextVQA	MMBench	Macro
P0 (MPO)	0.83	1.15	1.005	1.415	1.100
P1 (Instr.)$*$	0.539	0.624	0.898	0.497	0.640
P2† (RLHF)	0.761	0.935	0.709	1.186	0.898
P3† (RLHF)	1.167	1.400	0.800	1.400	1.192

4.4. Per-Corruption Breakdown

Table 5 reports CE by corruption type on GQA under the global reference (IVL-SFT).

Several patterns emerge. **Brightness** is the corruption most sensitive to alignment: P0 (MPO) achieves a CE of 0.60 (versus 1.0 for its base), and P1 (instruction tuning) achieves 0.831—both substantial improvements. This may reflect that RLHF and DPO data collection processes expose models to a wider variety of image conditions, implicitly improving brightness robustness. **Defocus blur** consistently produces

Table 5. CE per corruption type on GQA (global reference = IVL-SFT). Lower = more robust. Models shown as B = base, A = aligned within each pair.

	P0 (MPO)		P1 (Inst.)		P2 (G4B)		P3 (G7B)	
Corruption	B	A	B	A	B	A	B	A
Gauss. noise	1.000	0.889	1.556	0.879	1.167	1.056	1.000	1.167
Defocus blur	1.000	0.950	1.400	0.816	1.150	0.900	0.900	1.050
Brightness	1.000	0.600	1.733	0.831	1.333	0.867	1.200	1.400
JPEG comp.	1.000	0.882	1.588	0.860	1.294	0.941	1.059	1.235
mCE	1.000	0.830	1.569	0.846	1.236	0.941	1.040	1.213

the highest CE values and shows the smallest alignment benefit, suggesting that blur-type corruptions are not well addressed by preference training on text-image pairs with clean images. **P3 (G7B-IT)** is the only aligned model that is *worse* than its base counterpart on Gaussian noise and brightness; this reversal occurs across the board and is not corruption-specific, pointing to a structural degradation from alignment rather than a sensitivity to a particular corruption family.

4.5. Methodological Note: Gemma Pre-training Confound

The Gemma-3-4B (P2) and Gemma-3-7B (P3) base models (`gemma-3-4b-pt`, `gemma-3-7b`) are raw pre-trained language models without supervised fine-tuning. They cannot reliably follow VQA task prompts, producing near-zero accuracy on TextVQA (2% for G4B-pt, 4% for G7B-pt) and very low accuracy on VQAv2 (11% and 30%) while generating language continuations rather than answers. The aligned models (G4B-it, G7B-it) score 62% and 8% on the same TextVQA task, respectively. On most datasets, this gap measures instruction-following capability, not alignment-induced robustness change.

A proper alignment ablation requires both models to share an SFT checkpoint and differ only in preference optimization, as P0 (InternVL2.5) achieves correctly, with IVL-SFT as the base checkpoint. Gemma results from Table 4 are therefore interpreted with caution: where base accuracy is near zero, the CE denominator becomes small and mCE values are less reliable (notably TextVQA for G4B-pt and G7B-pt). GQA and MMBench results are more interpretable for P2 (G4B-pt achieves 54% on GQA), but G7B-pt's clean MMBench accuracy (0.50) being dramatically higher than G7B-it (0.26) suggests a further confound in the 7B pair.

This finding has broader methodological relevance: many published VLM robustness studies compare instruction-tuned models against raw pre-trained models without acknowledging this confound, potentially attributing instruction-following gains to architectural or scaling factors.

4.6. Ablation Study: Corruption Severity

Table 6 reports GQA accuracy at each of the three evaluated severity levels for P0 (InternVL2.5, $n = 50$), the pair with the most reliable base/aligned comparison.

Table 6. GQA accuracy vs. severity for P0 (InternVL2.5/MPO). B = base (IVL-SFT), A = aligned (IVL-MPO). ↑ aligned is better; ↓ base is better.

	Sev. 1		Sev. 3		Sev. 5	
Corruption	B	A	B	A	B	A
Gaussian noise	0.50	0.60↑	0.30	0.40↑	0.40	0.40
Defocus blur	0.20	0.50↑	0.30	0.30	0.50	0.30↓
Brightness	0.50	0.60↑	0.50	0.80↑	0.50	0.70↑
JPEG comp.	0.50	0.50	0.50	0.50	0.30	0.50↑

Three severity-dependent patterns are apparent. First, **brightness** shows a monotone advantage for the aligned model at all severity levels, with the gap peaking at severity 3 ($+0.30$ pp). Second, **defocus blur** reveals a striking **severity-dependent reversal**: the aligned model is substantially better at low severity ($+0.30$ at sev. 1) but becomes *worse* at high severity (-0.20 at sev. 5). This inversion may reflect that the aligned model uses language priors more aggressively to "fill in" mildly degraded images (which works at sev. 1) but hallucinates under severe blur (sev. 5) where the visual signal is too degraded to anchor the prior. Third, **Gaussian noise** shows a consistent but diminishing advantage for the aligned model across severities, vanishing at sev. 5, consistent with the hypothesis that alignment benefits erode as corruptions approach perceptual limits. Figure 5 illustrates severity degradation curves for all VLM pairs across benchmarks.

4.7. Extended Analysis: Qwen2-VL (11 Corruptions)

Qwen2-VL was additionally evaluated on the full 11-corruption suite at all five severity levels. CE is computed with QW-base as the intra-pair reference. Table 7 summarizes mCE across datasets.

Table 7. Qwen2-VL 11-corruption mCE (intra-pair reference = QW-base). Rel. CE = CE(aligned)/CE(base). $*$ TextVQA base clean accuracy $= 0.10$; values less reliable.

Dataset	mCE (base)	mCE (aligned)	Rel. mCE
GQA	0.104	0.722	6.964
VQAv2	0.542	0.927	1.711
TextVQA$*$	0.141	0.265	1.874
MMBench	0.177	0.977	5.532

The extremely high Rel. CE values (6.96 on GQA, 5.53 on MMBench) reflect the near-zero performance of QW-base on

certain corruptions; this is when the denominator is close to zero, small differences in the numerator produce large ratios. Interpreting these values as evidence of alignment harm would be misleading; they are a consequence of the task-following confound in QW-base noted in Section 4.5. The intra-pair global reference (Table 3) is more interpretable for Qwen: QW-inst has mCE $= 0.846$ vs. IVL-SFT on GQA, better than every base model, confirming the robust instruction-tuning benefit identified in Section 4.3.

4.8. Discussion

Does alignment hurt robustness (H1)? Results are pair-dependent and paradigm-dependent. H1 is clearly *contradicted* for P1 (instruction tuning, Qwen2-VL): alignment consistently and substantially improves robustness across all datasets and corruption families (Rel. mCE $= 0.64$ macro). H1 receives *mixed support* for P0 (MPO, InternVL2.5): alignment helps on GQA (0.83) but hurts on MMBench (1.42) and VQAv2 (1.15), with a macro Rel. mCE of 1.10. H1 is *supported* for P3 (Gemma-3-7B/RLHF): alignment consistently hurts across three of four datasets (macro 1.19), though the pre-training confound limits interpretation.

Alignment method comparison. The two most interpretable pairs, P0 (MPO) and P1 (instruction tuning), show qualitatively opposite profiles: MPO yields mixed results while instruction tuning yields consistent improvements. Whether this difference is driven by the alignment method itself (MPO's quality-loss term versus DPO's preference loss) or by the base model family (InternVL vs. Qwen) cannot be determined from two pairs alone. Future work pairing the same backbone with different alignment procedures would allow cleaner attribution.

Which corruption family is most harmful? Across all pairs, defocus blur consistently produces the highest CE values and the smallest alignment benefit, consistent with prior findings that blur disproportionately degrades VLM text recognition [5]. Brightness shows the largest alignment benefit across P0, P1, and P2, suggesting that preference training may incidentally improve robustness to photometric shifts.

Limitations. Three limitations bound interpretation. First, sample sizes are small ($n = 10$–50), producing high-variance accuracy estimates; differences of ± 0.1 are within the noise band for binomial proportions at these sample sizes. Second, hardware differences between P0 (GTX 1650, 4-bit quantisation) and P1–P3 (Apple MPS, bfloat16) may introduce inference-level variation that is difficult to disentangle from alignment effects. Third, the Gemma pre-training confound (Section 4.5) limits the interpretability of P2 and P3 Relative mCE values on TextVQA and VQAv2. Larger-scale runs with matched hardware are needed for firm conclusions.

5. Ethical Analysis

5.1. Verbosity Penalty as Evaluation Bias

The four benchmarks used in this study (GQA, VQAv2, TextVQA, MMBench) are drawn predominantly from Western and English-language image sources. GQA and VQAv2 derive from Visual Genome and COCO, respectively, both of which over-represent North American and European visual contexts. TextVQA images are sourced from Open Images, which similarly skews toward Western infrastructure and signage (English stop signs, Latin-script text). This creates a systematic bias: robustness conclusions drawn from these benchmarks may not generalize to visual contexts common in Africa, South Asia, or East Asia, where different camera hardware, lighting conditions, and environmental factors produce distinctive corruption signatures.

For AI systems deployed in Ghanaian or broader African settings for example, a context directly relevant to Ashesi University, these bias warrants serious attention. We recommend that future work explicitly include African visual benchmarks (e.g., images from Accra marketplaces, Kumasi Road scenes, or agricultural settings) when evaluating robustness, to ensure that reliability claims are not geographically contingent.

5.2. Misuse Risk of Robustness Findings

Our finding that instruction-tuned models are more robust to common corruptions could be misread to justify deploying VLMs in safety-critical contexts (medical imaging, autonomous driving, surveillance) without adequate corruption-specific testing. This would be a dangerous over-extrapolation: the corruptions in ImageNet-C are stylised, algorithmically generated approximations of real-world degradation, and robustness on these benchmarks does not guarantee robustness to the specific corruption distributions encountered in clinical imaging noise, automotive sensor artefacts, or compressed security footage. Any deployment in safety-critical applications should be accompanied by domain-specific corruption testing, uncertainty quantification, and human oversight.

5.3. Model-Level Fairness and Alignment Robustness

Instruction tuning datasets are known to over-represent certain demographic groups, cultural contexts, and languages. If aligned models are more robust to corruption in part because they leverage language priors from skewed instruction-following datasets, this robustness benefit may not extend equally to queries about underrepresented contexts. A model that compensates for corrupted visual input by defaulting to language priors could produce systematically incorrect or culturally inappropriate outputs for images from low-resource settings. Future work should disaggregate robust-

ness evaluation by demographic and cultural subgroups to determine whether the robustness gains we observe are evenly distributed.

6. Conclusion

We presented a preliminary controlled evaluation of base versus alignment-tuned VLMs under the ImageNet-C corruption benchmark, comparing four model pairs across three alignment paradigms (MPO, instruction tuning, SFT+RLHF) and four VQA datasets. While the scope is intentionally limited, the comparison is, to our knowledge, the first to isolate alignment-induced changes in perceptual robustness while holding architecture and pre-training fixed within each pair.

Our central observation is that alignment training does not uniformly increase or decrease corruption robustness: MPO-aligned InternVL2.5 shows mixed per-dataset effects (macro Rel. mCE $= 1.10$), instruction-tuned Qwen2-VL shows consistent and substantial robustness improvements (0.64), Gemma-3-4B instruction tuning mostly improves robustness (0.90), while Gemma-3-7B instruction tuning degrades it (1.19). We also observed a severity-dependent reversal for defocus blur, in which the aligned model is more robust at low severity but less robust at high severity—a pattern with direct deployment implications that, given our sample size, warrants confirmation at larger scale. Methodologically, we identified a verbosity-scoring mismatch that systematically biases standard VQA metrics against aligned models, proposed a prompt-level fix, and documented an evaluation design confound that inflates apparent alignment gaps in prior work when base models lack SFT fine-tuning.

We emphasize that these results are preliminary. The divergent direction of effects across model families is itself the strongest signal in our data: it indicates that the alignment–robustness relationship cannot be inferred from any single pair, and motivates broader, better-powered evaluation before any general claim about alignment and perceptual robustness can be made.

Future work. Three extensions are most pressing. First, scaling to larger sample sizes ($n \geq 500$ per condition) with matched hardware would provide the statistical power needed to move from suggestive to firm conclusions, particularly for severity-dependent effects such as the defocus-blur reversal. Second, completing evaluation across all 15 ImageNet-C corruption types and all 5 severity levels would enable full mCE computation comparable to the original benchmark and broaden coverage beyond the subset examined here. Third, and most consequential, evaluating whether corruption erodes the *safety posture* of aligned models remain entirely open; given that safety is the stated purpose of alignment, this is the highest-value next step for the research agenda this paper opens.

Declaration of Ai Usage. We used Claude (Anthropic) to assist with code debugging, LaTeX formatting, and experimental pipeline design. All experimental results, analysis, and writing are original work by the team. GitHub Copilot was used for code auto-completion during pipeline development.

References

[1] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1903.12261.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[3] Z. Chen *et al*. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[4] W. Wang, Z. Chen *et al*. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

[5] M. Usama *et al*. Analysing the robustness of vision-language models to common corruptions. *arXiv preprint arXiv:2504.13690*, 2025.

[6] VLM-RobustBench Authors. VLM-RobustBench: A comprehensive benchmark for robustness of vision-language models. *arXiv preprint arXiv:2603.06148*, 2025.

[7] N. Carlini *et al*. Are aligned neural networks adversarially aligned? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[8] X. Qi *et al*. Visual adversarial examples jailbreak aligned large language models. In *AAAI Conference on Artificial Intelligence*, 2024.

[9] Q. Liu *et al*. Unraveling and mitigating safety alignment degradation of vision-language models. *arXiv preprint arXiv:2410.09047*, 2024. To appear in ACL Findings, 2025.

[10] D. A. Hudson and C. D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE/CVF CVPR*, 2017.

[12] A. Singh *et al*. Towards VQA models that can read. In *IEEE/CVF CVPR*, 2019.

[13] Y. Liu, H. Duan *et al*. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, 2024.

[14] J. Bai *et al*. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[15] Meta AI. Llama 3.2: Multimodal large language models. Technical report, Meta AI, 2024.

7. Appendix

Table 8. Benchmark performance comparison across various Vision Language Models. Higher scores indicate better performance; best results are highlighted in **bold**.

Model	GQA	VQAv2	TextVQA	MMBench
IVL-SFT (base)	0.580	**0.720**	**0.660**	0.580
IVL-MPO (aligned)	0.560	0.650	0.640	0.520
Qwen2-VL-2B	0.300	0.500	0.100	0.300
Qwen2-VL-2B-Inst	**0.800**	0.600	0.400	**0.800**
Gemma-3-4B-PT	0.540	0.110	0.020	0.680
Gemma-3-4B-IT	0.620	0.410	0.620	0.720
Gemma-3-7B	0.500	0.300	0.040	0.500
Gemma-3-7B-IT	0.300	0.280	0.080	0.260

Table 9. Robustness summary on the mCE metric. All values are normalized against the IVL-SFT (base) reference. **Lower values indicate better robustness**; best results per benchmark are highlighted in bold.

Model	GQA ↓	VQAv2 ↓	TextVQA ↓	MMBench ↓
IVL-SFT (base)	1.000	**1.000**	1.000	1.000
IVL-MPO (aligned)	**0.830**	1.150	1.005	1.415
Qwen2-VL-2B	1.569	1.928	1.194	2.427
Qwen2-VL-2B-Inst	0.846	1.202	1.072	1.207
Gemma-3-4B-PT	1.236	2.059	1.402	**0.824**
Gemma-3-4B-IT	0.941	1.925	**0.994**	0.978
Gemma-3-7B	1.040	1.342	1.413	1.419
Gemma-3-7B-IT	1.213	1.879	1.130	1.987

Table 10. Comprehensive mCE and mCR summary macro-averaged over four datasets. Values are normalized against the IVL-SFT (base) reference. **Lower mCE** (↓) and **higher mCR** (↑) indicate superior robustness; best results are in bold.

Model	GQA ↓	VQAv2 ↓	TextVQA ↓	MMBench ↓	Macro-mCE ↓	Macro-mCR ↑
IVL-SFT (base)	1.000	**1.000**	1.000	1.000	**1.000**	**0.000**
Qwen2-VL-2B-Inst	0.846	1.202	1.072	1.207	1.082	-0.082
IVL-MPO (aligned)	**0.830**	1.150	1.005	1.415	1.100	-0.100
Gemma-3-4B-IT	0.941	1.925	**0.994**	0.978	1.209	-0.209
Gemma-3-7B	1.040	1.342	1.413	1.419	1.303	-0.303
Gemma-3-4B-PT	1.236	2.059	1.402	**0.824**	1.380	-0.380
Gemma-3-7B-IT	1.213	1.879	1.130	1.987	1.552	-0.552
Qwen2-VL-2B	1.569	1.928	1.194	2.427	1.779	-0.779

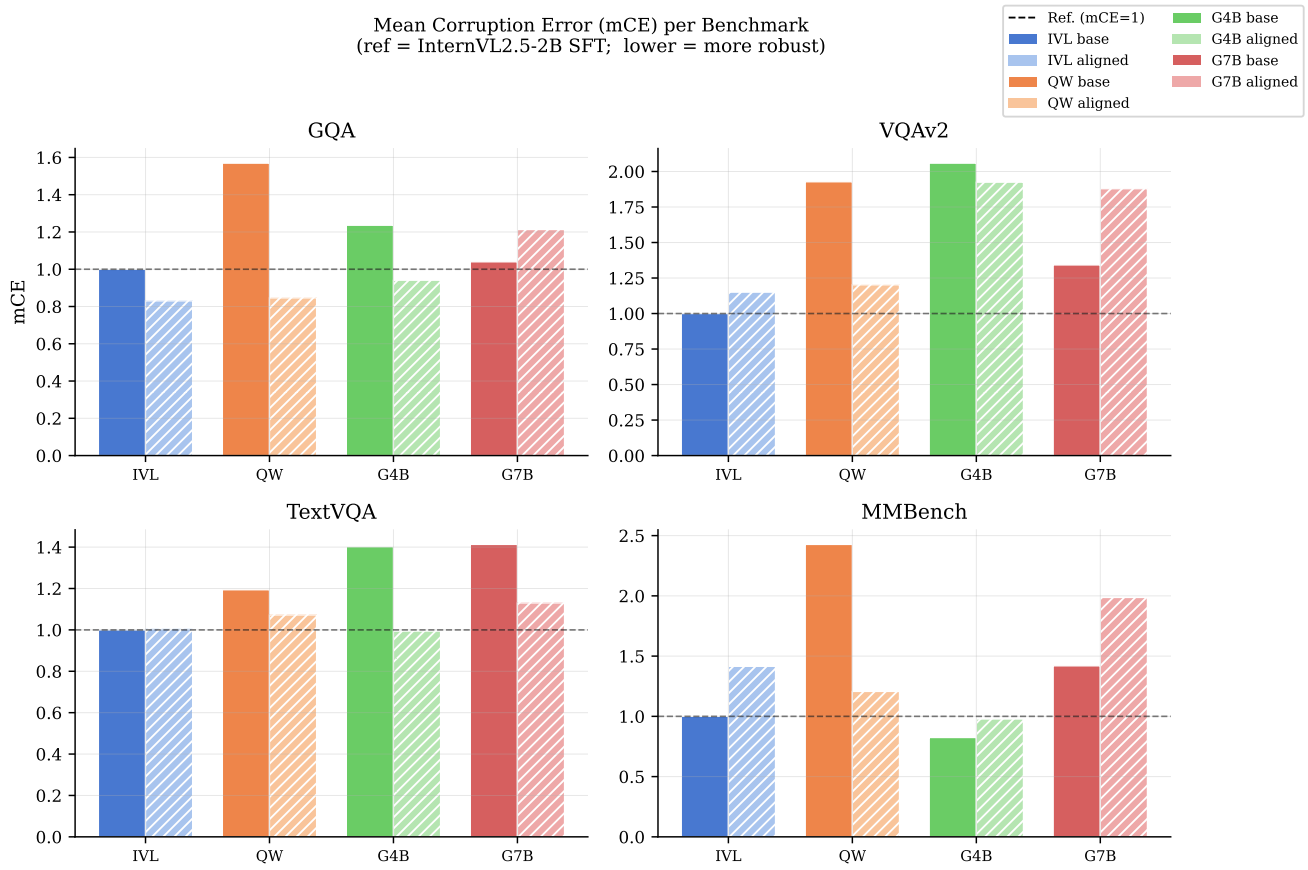

Figure 3. **mCE per benchmark for each model pair** (global reference = InternVL2.5-2B SFT; lower = more robust). Solid bars = base; hatched bars = aligned. The dashed line marks the reference level (mCE = 1.0). Aligned models in the IVL and QW pairs sit at or below the reference on GQA, while all models exceed it on VQAv2 and MMBench, with Gemma-3-4B-PT and Qwen2-VL-2B base showing the largest deviations.

Table 11. Detailed robustness evaluation across four benchmarks. We report Corruption Error (CE) for base and aligned models, alongside Relative CE (Rel.CE). All values are macro-averaged across corruption levels. **Lower is better** for all metrics.

	GQA			**VQAv2**			**TextVQA**			**MMBench**		
Corruption	CE_b	CE_a	Rel.	CE_b	CE_a	Rel.	CE_b	CE_a	Rel.	CE_b	CE_a	Rel.
Brightness	0.198	0.849	4.29	0.680	1.063	1.56	0.200	0.469	2.34	0.286	0.935	3.27
Contrast	0.086	0.728	8.49	0.440	0.760	1.73	0.133	0.422	3.17	0.200	1.000	5.00
Defocus Blur	0.143	0.644	4.51	0.560	0.921	1.64	0.111	0.171	1.54	0.086	0.857	9.99
Elastic	0.029	0.729	25.5	0.640	0.960	1.50	0.133	0.311	2.33	0.143	1.000	7.00
Fog	0.114	0.768	6.72	0.400	1.000	2.50	0.178	0.378	2.13	0.286	1.143	4.00
Gauss. Noise	0.057	0.687	12.0	0.560	0.922	1.65	0.156	0.191	1.23	0.143	0.924	6.47
Impulse Noise	0.029	0.689	24.1	0.560	0.880	1.57	0.133	0.156	1.17	0.143	1.086	7.60
JPEG Comp.	0.086	0.742	8.66	0.440	1.054	2.40	0.156	0.240	1.54	0.286	0.975	3.41
Pixelate	0.343	0.778	2.27	0.640	0.920	1.44	0.089	0.178	2.00	0.200	1.143	5.71
Shot Noise	0.057	0.689	12.1	0.560	0.920	1.64	0.111	0.178	1.60	0.086	1.000	11.7
Zoom Blur	0.000	0.640	—	0.480	0.800	1.67	0.156	0.222	1.43	0.086	0.686	8.00
mCE	**0.104**	**0.722**	**6.96**	**0.542**	**0.927**	**1.71**	**0.141**	**0.265**	**1.87**	**0.177**	**0.977**	**5.53**

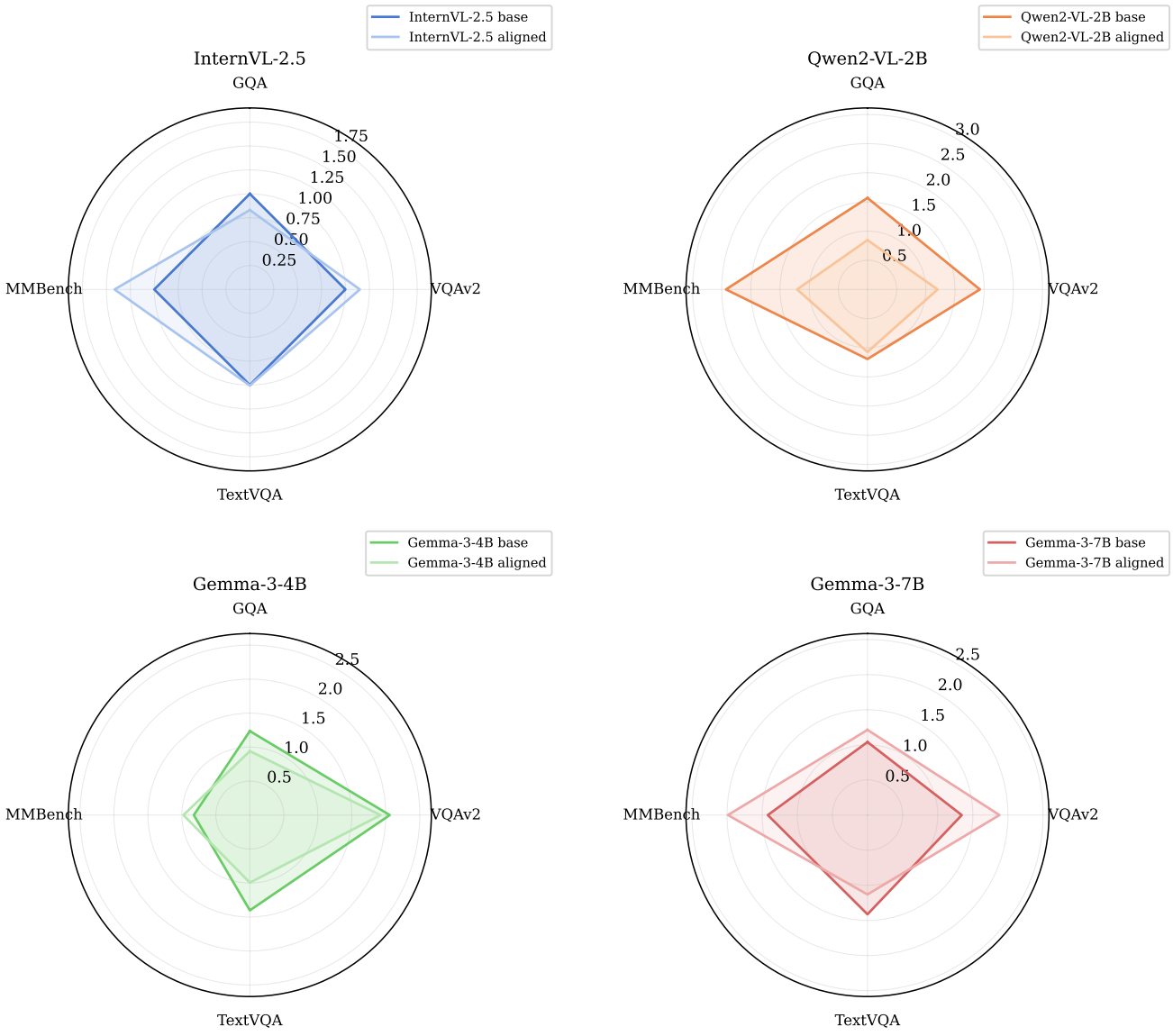

Figure 4. **Robustness radar charts: mCE across benchmarks per model pair.** Each axis represents one benchmark (GQA, VQAv2, TextVQA, MMBench); shorter spokes indicate lower mCE and therefore greater robustness. Solid lines = base model; dashed lines = aligned model. InternVL-2.5 (blue, top-left) shows a GQA-only benefit with a notably longer MMBench spoke for the aligned model. Qwen2-VL-2B (orange, top-right) has a uniformly smaller aligned polygon across all axes, confirming consistent robustness gains from instruction tuning. Gemma-3-4B (green, bottom-left) shows a similar but weaker benefit, with VQAv2 nearly identical between base and aligned. Gemma-3-7B (red, bottom-right) is the only pair where the aligned polygon is *larger* on GQA, VQAv2, and MMBench, visually confirming the degradation reported in Table 4.

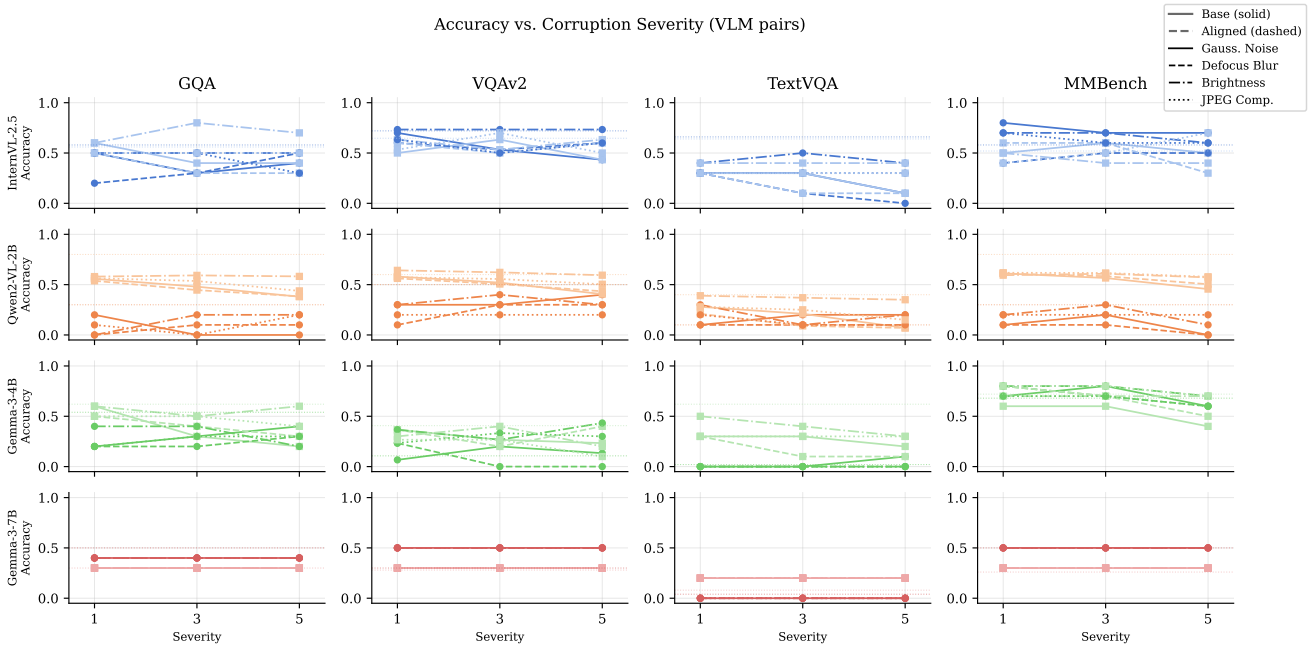

Figure 5. Accuracy vs. corruption severity for all three VLM pairs (base = solid lines, aligned = dashed lines). Dotted horizontal lines indicate clean baseline accuracy. The aligned-better-at-low-severity pattern is most pronounced for Qwen2-VL across all datasets and corruption types.

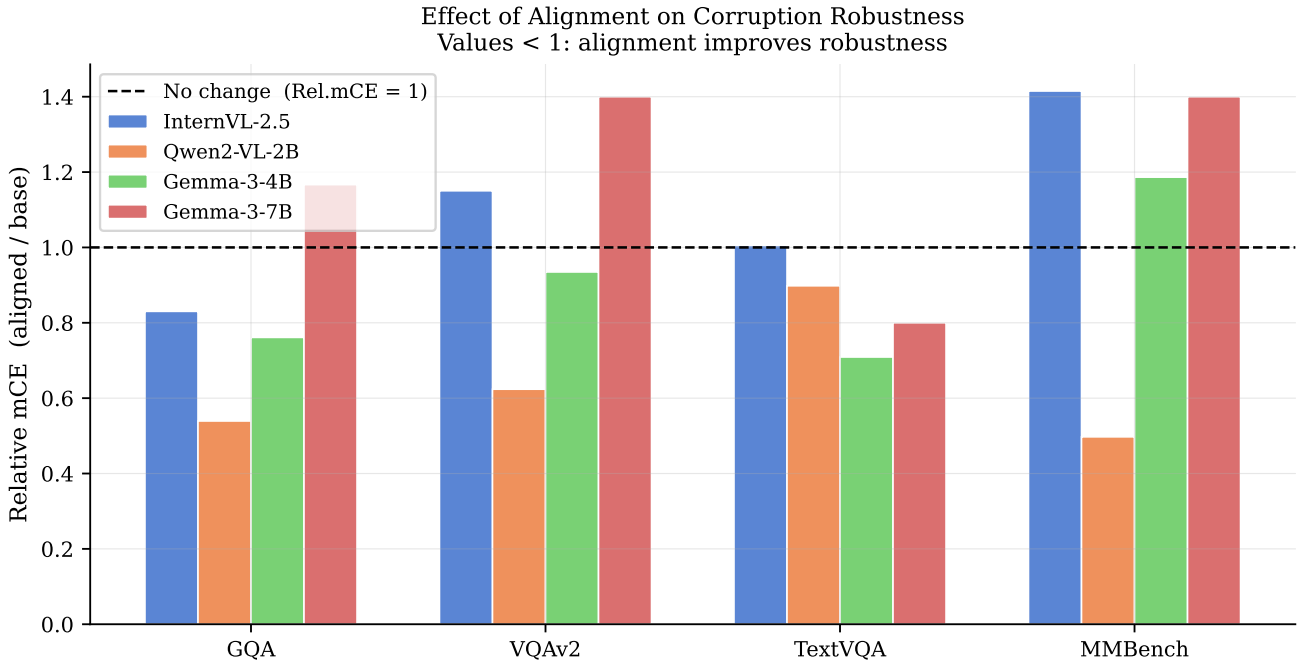

Figure 6. **Effect of alignment on corruption robustness across benchmarks and model pairs.** Each bar shows Relative mCE ($mCE_{aligned}$ / mCE_{base}); values below the dashed line (Rel. mCE = 1) indicate the aligned model is *more* robust, values above indicate it is *more* brittle. Qwen2-VL-2B (orange) falls below the line on all four benchmarks (macro = 0.64); Gemma-3-4B (green) improves on three of four (macro = 0.90); InternVL-2.5 (blue) shows mixed dataset-dependent effects (macro = 1.10); and Gemma-3-7B (red) exceeds the line on three benchmarks, indicating overall degradation (macro = 1.19). No single alignment paradigm uniformly improves or degrades robustness— the direction and magnitude of the effect depend on both alignment method and model family. Gemma base models are raw pre-trained checkpoints; their values are interpreted with caution (Section 4.5).

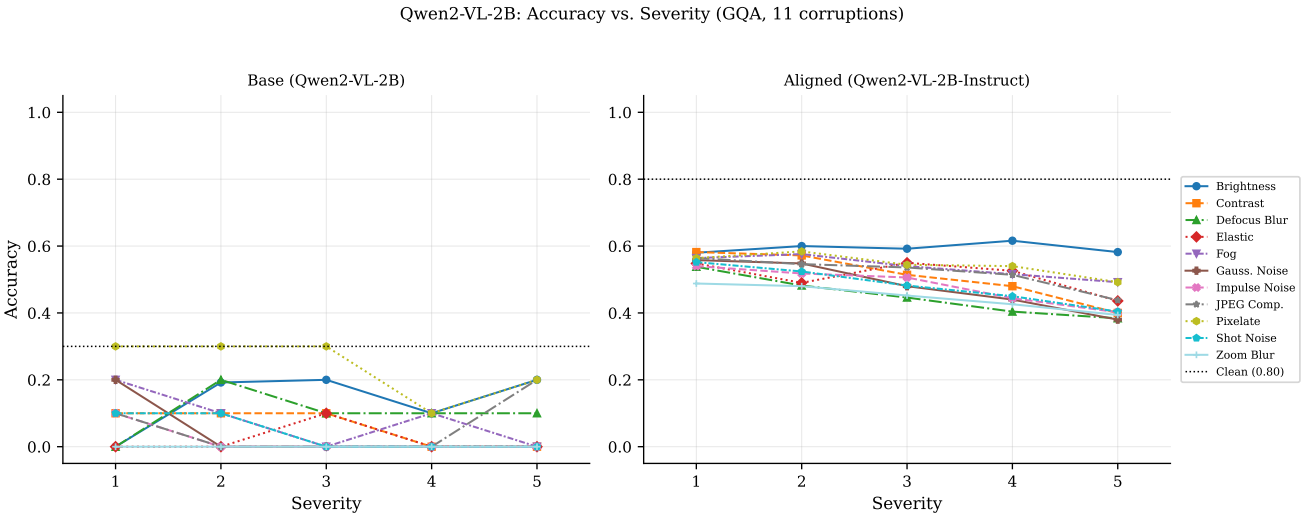

Figure 7. **Qwen2-VL-2B accuracy vs. corruption severity on GQA (11 corruptions, severities 1–5).** The base model (left) collapses to near-zero on most corruptions by severity 3, while the aligned model (right) degrades smoothly from ≈0.6 to ≈0.4, maintaining meaningful accuracy across all corruption types and severity levels.