VLM-ROBUSTBENCH: A ROBUSTNESS BENCHMARK FOR VISION-LANGUAGE MODELS

Rohit Saxena[1] **Alessandro Suglia**[1] **Pasquale Minervini**[1,2]
[1]ILCC, School of Informatics, University of Edinburgh [2]Miniml.AI
{rohit.saxena, asuglia, p.minervini}@ed.ac.uk

ABSTRACT

Vision-language models (VLMs) achieve strong performance on standard, high-quality datasets, but we still do not fully understand how they perform under real-world image distortions. We present **VLM-RobustBench**, a benchmark spanning 49 augmentation types across noise, blur, weather, digital, and geometric perturbations, evaluated under graded severities (low/mid/high) and binary transforms, yielding 133 corrupted settings. We evaluate VLMs from four families (Qwen, InternVL, Molmo, Gemma) on two complementary benchmarks: MMBench (visually grounded) and MMMU-Pro (reasoning-oriented). Our results reveal that visual severity is a weak predictor of difficulty: low-severity spatial perturbations often degrade performance more than visually severe photometric corruptions. In particular, low-severity `glass_blur` reduces MMBench accuracy by about 8 pp on average across models, while the largest drops arise from resampling and geometric distortions (e.g., `upsample`, `elastic_transform`), reaching up to 34 pp. Overall, our findings suggest current VLMs are *semantically strong but spatially fragile*, motivating the definition of novel robustness evaluation protocols and training regimes that emphasize resampling and geometric invariances.

1 INTRODUCTION

The rapid evolution of Vision-Language Models (VLMs) has marked a transition from text-only specialists to multimodal generalist models that are capable of complex reasoning across a broad range of tasks. Recent VLMs (Bai et al., 2025; Clark et al., 2026; Wang et al., 2025a; Team et al., 2025) demonstrated remarkable proficiency on several benchmarks, exhibiting capabilities ranging from zero-shot generalisation to fine-grained image-text understanding Liu et al. (2024). These advancements catalysed the integration of VLMs into safety-critical pipelines, including autonomous driving perception stacks Zhou et al. (2024), medical diagnostic support systems (Sellergren et al., 2025), and automated document processing workflows (Wang et al., 2025b).

However, strong performance on curated benchmarks does not guarantee reliability under the distribution shifts encountered in deployment (Yu et al., 2024). Real-world visual inputs are rarely pristine: low-light sensor noise, adverse weather (rain, fog, snow), compression artefacts, and motion or defocus blur are common (Hendrycks & Dietterich, 2019). In addition, viewpoint changes induce geometric variations, such as scaling, rotation, and perspective distortion, that may be absent or simplified in training data (Barbu et al., 2019b; Zhou et al., 2022). For safety-critical use, we therefore need robustness evaluations that stress these everyday corruptions rather than measuring accuracy on a fixed dataset that includes only a few visual perturbations.

While the computer vision community has established rigorous benchmarks for robustness to common corruptions—most notably ImageNet-C (Hendrycks & Dietterich, 2019), the robustness landscape for modern VLMs remains less systematically characterised, especially across tasks and realistic corruption families (Usama et al., 2025; Ye et al., 2024). A central challenge is understanding whether language-side reasoning can compensate when visual perception is degraded, or whether certain perturbations induce sharp perceptual bottlenecks that dominate end performance (Liu et al., 2025; Fan et al., 2025; Zhou et al., 2025).

Moreover, corruption benchmarks often implicitly assume *severity monotonicity*: as visual distortion increases, inputs should become increasingly harder (Hendrycks & Dietterich, 2019). It remains

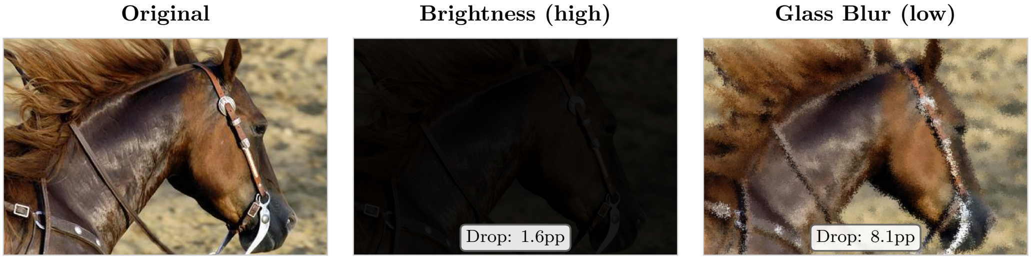

Figure 1: **Severity paradox (MMBench, mean over 9 models).** High-severity brightness drop: -1.6pp; low-severity glass blur: -8.1pp. Severity level does not always predict model difficulty.

unclear whether this assumption holds for VLMs, where perception and language reasoning are tightly coupled through cross-modal representations (Zhou et al., 2025). This motivates a dedicated benchmark that probes the interplay between visual corruptions and multimodal reasoning across a broad spectrum of perturbation types and severity levels.

In this work, we present **VLM-RobustBench**, a large-scale analysis of VLM robustness under visual corruption. We systematically evaluate 11 models spanning four major VLM families (Qwen, InternVL, Molmo, and Gemma) across 133 distinct augmentation configurations (42 corruptions at three severities plus 7 binary transforms) on two diverse datasets: MMBench (Liu et al., 2024) (more visually grounded) and MMMU-Pro (Yue et al., 2025) (more reasoning-oriented). Our results challenge prevailing assumptions and reveal that current VLMs are *semantically strong but spatially fragile*. We highlight three key contributions:

1. **The Spatial Fragility Finding:** VLMs are disproportionately sensitive to spatial and resampling artefacts. A resampling operation (`upsample`) or mild geometric distortion causes catastrophic failure (up to 34pp drop), whereas severe photometric degradations (e.g., noise, compression) are often handled robustly.
2. **Severity Mismatch:** We observe a decoupling of severity level and model difficulty (**??**). On MMBench, low-severity perturbations degrade performance more than high-severity perturbations of other types, complicating safety assurance (e.g., `glass_blur` and `solarize` at low severity result in an 8pp and 5.6pp drop, respectively).
3. **Family-Specific Vulnerabilities:** Robustness is not a function of parameter count. Distinct model families exhibit unique vulnerability "fingerprints," suggesting that architectural choices play a decisive role in determining failure modes.

2 Related Work

Vision–language model evaluation benchmarks. Standardized LVLM benchmarks probe complementary capabilities: multimodal perception/reasoning (MMBench) (Liu et al., 2024), expert knowledge and reasoning (MMMU / MMMU-Pro) (Yue et al., 2024; 2025), and perception–cognition decompositions (MME) (Fu et al., 2025). Some VLMs can exploit language priors with limited visual grounding (Tong et al., 2024), motivating vision-centric evaluations such as NaturalBench (Li et al., 2024). We build on this ecosystem but focus on robustness under visual degradation, using MMBench and MMMU-Pro as complementary anchors and introducing *visual gain* to quantify reliance on vision.

Robustness to natural corruptions in vision. Corruption robustness is formalized by ImageNet-C/ImageNet-P (Hendrycks & Dietterich, 2019), which define calibrated families of noise, blur, weather, digital, and geometric perturbations. Related benchmarks extend to broader shifts, including style/rendition changes (ImageNet-R) (Hendrycks et al., 2021) and viewpoint/background variation (ObjectNet) (Barbu et al., 2019a). VLM-RobustBench adapts this paradigm to LVLMs by expanding the augmentation taxonomy and separating graded severities from binary transforms.

Natural-corruption robustness in VLMs and VQA. LVLM robustness under visual corruption is less explored than in vision-only models. Recent studies evaluate VLMs on subsets of ImageNet-Cstyle corruptions and corrupted VQA, revealing task-dependent sensitivities (Usama et al., 2025). We extend this line with a broader suite—including resampling and geometry-focused stressors—across severity levels and both visually grounded and reasoning-oriented benchmarks, enabling analysis of when models rely on language priors versus visual grounding.

Adversarial robustness of vision–language models. Worst-case adversarial perturbations for vision–language models—including transferable and black-box attacks—have been studied (Zhao et al., 2023; Qi et al., 2024; Shayegani et al., 2024), and adversarial images can bypass safeguards and induce incorrect or unsafe outputs (Carlini et al., 2023). Defenses such as adversarially fine-tuned CLIP-style encoders improve robustness for downstream LVLMs with frozen vision backbones (Mao et al., 2023). Our focus is complementary: naturally occurring corruptions and operational artifacts rather than adaptive attacks.

Spatial robustness and resolution effects. Sensitivity to spatial and resampling distortions reflects known limitations of patch-based vision encoders, studied in Vision Transformer robustness analyses (Bhojanapalli et al., 2021; Paul & Chen, 2022). In LVLMs, preprocessing choices (e.g., resizing strategy and tokenization granularity) can strongly affect performance (McKinzie et al., 2024). Our benchmark provides a systematic testbed to quantify these effects and guide training strategies emphasizing geometric and resampling invariances.

3 METHOD

3.1 PROBLEM FORMULATION

Let $\mathcal{M}$ denote a vision-language model that takes an image $I \in \mathcal{I}$ from the space of images $\mathcal{I} = \mathbb{R}^{H \times W \times 3}$ and a text query Q as input, producing a textual response $\mathcal{M}(I, Q)$. We evaluate multiple-choice accuracy using an answer extractor $g(\cdot)$ that maps the response to a discrete option, yielding $\hat{y} = g(\mathcal{M}(I, Q))$.

We define a set of image augmentations $\mathcal{A} = \{A_1, \ldots, A_K\}$. Each severity-based augmentation A_k is a stochastic transformation parameterized by severity $s \in \mathcal{S}$ that maps an image $I \in \mathcal{I}$ to one of its augmented versions $\hat{I} \in \mathcal{I}$, with $\hat{I} \sim A_k(I, s)$. For reproducibility, we fix per-sample random seeds, yielding deterministic outputs for each (image, severity) pair. Binary augmentations are not parameterised by s and apply directly as $A_k(I)$.

Given a dataset $\mathcal{D} = \{(I_i, Q_i, y_i)\}_{i=1}^{N}$ with ground-truth answers y_i, we define clean accuracy and accuracy under augmentation A_k at severity s as:

$$\text{Acc}_{\text{clean}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[g(\mathcal{M}(I_i, Q_i)) = y_i] \,, \qquad \text{Acc}_{A_k,s} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[g(\mathcal{M}(A_k(I_i, s), Q_i)) = y_i] \,. \tag{1}$$

The robustness drop $\Delta_{A_k,s} = \text{Acc}_{\text{clean}} - \text{Acc}_{A_k,s}$ quantifies the performance degradation in percentage points. We additionally evaluate a no-image baseline $\text{Acc}_{\varnothing}$ where images are removed.

Visual Gain. To quantify reliance on visual information versus language priors, we define *Visual Gain* (VG) as $\text{VG} = \text{Acc}_{\text{clean}} - \text{Acc}_{\varnothing}$. A larger VG indicates stronger dependence on visual input, whereas a low VG suggests greater solvability from language priors alone.

Relative Corruption Error. To normalize corruption impact by a model's visual reliance, we define $\text{RCE}_{A_k,s} = \frac{\Delta_{A_k,s}}{\text{VG}} \times 100\%$. All model–dataset pairs in our experiments have $\text{VG} > 7$, so division is well-defined. RCE=100% means the corruption removes all visual benefit; RCE>100% means performance becomes worse than the no-image baseline.

3.2 AUGMENTATION TAXONOMY

We construct a suite of 49 image augmentations motivated by real-world degradations. The suite comprises 42 severity-based corruptions grouped into nine categories and 7 binary transforms (Table 1). Each severity-based corruption is evaluated at three levels $s \in \{\text{low}, \text{mid}, \text{high}\}$, whereas binary transforms are applied once (no severity parameter).

Category	N	Augmentations
Blur	5	Gaussian, motion, defocus, glass, zoom
Noise	4	Gaussian, shot, speckle, salt-pepper
Weather	5	Fog, frost, snow, rain, spatter
Digital	2	JPEG compression, pixelation
Geometric	5	Rotate, shear, affine, perspective_transform, elastic
Occlusion	3	Center, random, grid mask
Color/Tone	10	Brightness$^{\pm}$, contrast$^{\pm}$, saturation$^{\pm}$, gamma$^{\pm}$, hue shift, color jitter
Resolution	5	Downsample, upsample, sharpen, posterize, solarize
VLM-specific	3	Text overlay, watermark, border
Binary	7	Grayscale, invert, equalize, autocontrast, channel swap, flip (h/v)

Table 1: Augmentation suite: 42 severity corruptions (low/mid/high) + 7 binary (133 configs).

Family	Model	Type
Qwen3-VL	Qwen3-VL-4B	Instruct
	Qwen3-VL-8B	Instruct
	Qwen3-VL-30B-A3B	MoE, 3B
	Qwen3-VL-4B-Think	Thinking
	Qwen3-VL-8B-Think	Thinking
InternVL3.5	InternVL3.5-4B	Instruct
	InternVL3.5-8B	Instruct
	InternVL3.5-14B	Instruct
Molmo2	Molmo2-4B	Instruct
	Molmo2-8B	Instruct
Gemma 3	Gemma-3-12B-it	Instruct

Table 2: VLM families evaluated. Main: 9 instruct; Qwen `Think` separate.

For corruptions overlapping ImageNet-C Hendrycks & Dietterich (2019), we reuse the same corruption types but calibrate severity levels independently for VLM evaluation. For VLM-specific corruptions, we define monotonic, visually ordered severity schedules (full parameter schedules in Section F.5). Low severity corresponds to mild perturbations, while high severity corresponds to strongly degraded inputs. This yields 126 severity-based configurations (42 × 3 levels) plus 7 binary transforms, resulting in *133 augmentation configurations* per model–dataset pair.

4 Experimental Setup

4.1 Models

We evaluate **11 VLMs** spanning four model families of open-weights models: Qwen3-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025a), Molmo2 (Clark et al., 2026), and Gemma3 (Team et al., 2025) (Table 2). Our primary robustness comparisons focus on *9 instruction-following VLMs* evaluated under a consistent direct-answer prompting protocol. We additionally include **2 Qwen3-VL** `Think` models (4B and 8B) as a *test-time compute ablation*. To isolate the role of reasoning at inference time, we compare chain-of-thought prompting against the `Think` variants [1].

4.2 Datasets

We evaluate on two challenging multimodal benchmarks (with seed 42 for reproducibility):

MMMU-Pro A multimodal understanding benchmark covering subjects from STEM to humanities. We use the standard 10-option multiple choice variant, evaluating on a stratified 20% sample (532 samples) across 30 subject categories.

MMBench A comprehensive benchmark for multimodal perception and reasoning. We evaluate on the English development split using stratified 20% sampling (869 samples) to ensure category balance across all question types.

4.3 Evaluation Protocol

For each model-dataset pair, we evaluate on clean images, a no-image baseline (image removed), and all corrupted settings: 126 severity-based configurations (42 corruptions × 3 severity levels) plus 7 binary transforms, totaling *133 + 2 (baseline) evaluations* per pair. Corruptions are applied to images only; text prompts and answer formats are held fixed across conditions. We use stratified 20% subsampling above to keep the full corruption sweep tractable (135 settings per model-dataset pair) while preserving category balance.

[1] These results are reported in Appendix D.4 and are not included in the main aggregate robustness comparisons.

MMBench (Direct)

Model	Baseline↑	Worst-Case↓	Severe-Fail↓	Worst@Low↓	Benign@Low↑	VG↑	mRCE↓
Qwen3-VL-4B	88.4	26.3	3.8	7.0	88.1	48.2	3.9
Qwen3-VL-8B	90.2	30.2	5.3	8.3	73.8	48.2	5.2
Qwen3-VL-30B	90.7	29.4	3.8	6.1	88.1	47.7	**3.5**
InternVL3.5-4B	86.3	30.5	9.8	10.8	64.3	44.9	7.7
InternVL3.5-8B	89.1	31.7	8.3	9.6	71.4	50.8	6.5
InternVL3.5-14B	86.6	29.4	9.0	9.5	88.1	44.5	6.0
Molmo2-4B	88.5	33.1	4.5	7.4	78.6	43.6	5.5
Molmo2-8B	88.4	33.9	4.5	6.3	90.5	48.4	4.4
Gemma-3-12B	85.3	32.1	8.3	10.7	69.0	44.1	6.4

MMMU-Pro (Direct)

Model	Baseline↑	Worst-Case↓	Severe-Fail↓	Worst@Low↓	Benign@Low↑	VG↑	mRCE↓
Qwen3-VL-4B	31.5	7.4	3.8	2.8	95.2	7.1	$-$**6.8**
Qwen3-VL-8B	35.2	9.0	7.5	5.2	85.7	11.4	4.7
Qwen3-VL-30B	40.7	14.5	12.8	7.4	59.5	17.0	12.1
InternVL3.5-4B	37.3	11.1	14.3	5.6	76.2	12.3	11.6
InternVL3.5-8B	41.0	12.3	16.5	5.6	45.2	14.2	16.5
InternVL3.5-14B	42.0	9.6	9.0	5.9	85.7	15.1	5.2
Molmo2-4B	31.8	5.6	3.8	4.3	92.9	12.0	4.9
Molmo2-8B	31.2	5.6	5.3	4.0	90.5	7.4	**1.0**
Gemma-3-12B	33.0	9.9	27.1	9.0	28.6	10.8	24.2

Table 3: Robustness summary (9 direct-mode models). Worst-Case/Severe-Fail over 133 configs; Worst@Low/Benign@Low over 42 low-severity corruptions. VG and (m)RCE in Sec. 3.1.

Unless stated otherwise, we report results in direct mode (standard prompting) and aggregate over the 9 instruction-following checkpoints. Chain-of-thought (CoT) and thinking results are reported separately in Appendix D.4.

Metrics. We report clean accuracy $\mathrm{Acc}_{\mathrm{clean}}$ as defined in Section 3.1. Since mean accuracy aggregates across augmentation types and severities, it can mask severity-specific and tail failures; we therefore focus on drop-based metrics that highlight failure modes. For each configuration (A_k, s), we define the accuracy drop (in percentage points) as $\Delta_{A_k,s} = \mathrm{Acc}_{\mathrm{clean}} - \mathrm{Acc}_{A_k,s}$ (binary transforms omit s).

We additionally report: (i) *Worst-Case Drop* $\max_{k,s} \Delta_{A_k,s}$, the maximum drop over all 133 configurations (126 severity-based + 7 binary); (ii) *Severe-Failure Rate*, the fraction of the same 133 configurations for which performance drops by more than a relative threshold, $\Delta_{A_k,s} > 0.1\,\mathrm{Acc}_{\mathrm{clean}}$. (iii) *Worst@Low* $\max_k \Delta_{A_k,\mathrm{low}}$, the maximum drop over 42 severity-based corruptions at low severity (binary excluded); and (iv) *Benign@Low*, the fraction of these 42 low-severity corruptions with $\Delta \leq 1$. Additional implementation details are in Section F.

5 Results

5.1 Tiered Robustness Overview

Table 3 summarizes direct-mode robustness. We report clean accuracy $\mathrm{Acc}_{\mathrm{clean}}$ (*Baseline*) and drop-based tail-risk metrics over the full corruption suite: *Worst-Case Drop*, *Severe-Failure Rate*, *Worst@Low*, and *Benign@Low*. These are deployment-relevant because a small number of failure-inducing transforms (e.g., resampling in preprocessing) can dominate risk even when most corruptions are benign. We also report *VG* (visual reliance) and *mRCE* (mean relative corruption error) to normalize corruption impact by each model's visual benefit (Section 3.1).

Two patterns emerge. *(i) Large failures are sparse but consequential:* on MMBench, 65–90% of low-severity corruptions are benign ($\Delta \leq 1$ across the 42 low-severity settings, model-dependent), yet a small subset causes sharp drops. This is captured by *Severe-Failure Rate* (fraction of the 133 configurations with $\Delta > 0.1\,\mathrm{Acc}_{\mathrm{clean}}$); e.g., InternVL3.5-4B reaches 9.8% on MMBench (13/133 settings). *(ii) Visually mild perturbations can still be high-risk: Worst@Low* shows low-severity drops up to 10.8 pp on MMBench. MMMU-Pro shows a wider spread in *Benign@Low* (roughly 30–95% of the 42 low-severity settings), consistent with differing visual reliance across models and tasks.

Preempting the "destroyed image" intuition. Despite the visible degradation of some high-severity corruptions, perceived visual severity is a weak predictor of difficulty: subtle spatial perturbations (e.g., low-severity `glass_blur` and resampling artifacts) can match or exceed the harm from strong photometric distortions (e.g., JPEG compression).

5.2 Binary Augmentations: Trivial Transforms, Large Failures

Beyond the 126 severity-based configurations, our 133-config benchmark includes 7 binary (on/off) augmentations. Table 5 reveals that two trivial transformations, vertical flip and color inversion, are *catastrophic* on MMBench despite requiring no learned parameters.

Key insight. Vertical flip is more harmful than 39 of 42 high-severity corruptions on MMBench, exceeded only by `upsample`, `elastic_transform`, and `zoom_blur`, suggesting VLMs encode strong orientation priors. Color inversion causes catastrophic drops on MMBench (10.1pp) but only mild harm on MMMU-Pro (1.4pp), indicating perception tasks depend on absolute color relationships while reasoning does not.

Perception vs. Reasoning via Visual Gain. We analyze Visual Gain (VG $= \text{Acc}_{\text{clean}} - \text{Acc}_{\varnothing}$) as a proxy for reliance on visual input versus language priors. VG is computed per model and then averaged across direct-mode models. MMBench has substantially larger VG (46.7 points) than MMMU-Pro (11.9 points), indicating that MMBench decisions depend more strongly on visual grounding, whereas MMMU-Pro permits greater fallback to language priors. This aligns with MMBench exhibiting larger worst-case drops and higher severe-failure rates (Table 3).

5.3 Which Corruptions Drive Risk?

Figure 2 in Appendix B shows the most harmful corruptions at each severity, averaged over all direct-mode models. Per-model breakdowns are in Appendix D.8. Two consistent patterns emerge. First, *severity is an unreliable proxy for difficulty*: at *low* severity, `glass_blur` (8.1 drop on MMBench, 5.5 on MMMU-Pro) is among the top failures, exceeding most high-severity corruptions. Second, *catastrophic risk concentrates in a small set of resampling and geometric corruptions*: `upsample` and `elastic_transform` dominate mid/high severities on MMBench, while `zoom_blur` becomes more prominent on MMMU-Pro.

Glass Blur Anomaly. Low-severity `glass_blur` outperforms many high-severity photometric corruptions (e.g., JPEG compression), providing a concrete example of the severity–difficulty mismatch. Interestingly, `glass_blur` exhibits non-monotonic behavior, with low severity sometimes inducing larger drops than higher severities (Appendix D.9), further illustrating the decoupling of visual and model difficulty.

5.4 Quantifying Severity Mismatch

Severity mismatch can be quantified by checking whether performance degrades monotonically with increasing visual severity. For each model and severity-based corruption A_k, we compute: (i) A *monotonicity violation* indicator, set to 1 if $\Delta_{A_k,\text{low}} > \Delta_{A_k,\text{mid}}$ or $\Delta_{A_k,\text{mid}} > \Delta_{A_k,\text{high}}$ (strict inequality; ties do not count as violations), and 0 otherwise. (ii) The Spearman rank correlation (ρ) between severity level $\{\text{low}, \text{mid}, \text{high}\}$ and the robustness drop. All corruptions are deterministic given fixed parameters and per-sample seeds; we do not average over multiple stochastic draws, so observed violations reflect true model behavior rather than sampling variance.

We report these metrics in Appendix E.1. Overall, we observe substantial violation rates and weak-to-moderate correlations, indicating that visually ordered severity is an unreliable proxy for model difficulty, particularly on reasoning-oriented MMMU-Pro.

Dataset	Aug.	Sev.	RCE
MMBench	upsample	high	65.6
	elastic_transform	high	53.8
	upsample	mid	33.5
	zoom_blur	high	24.7
	solarize	high	22.6
	flip_v	binary	22.1
MMMU-Pro	zoom_blur	high	77.6
	elastic_transform	high	73.3
	upsample	high	72.9
	zoom_blur	mid	63.9
	upsample	mid	58.1
	glass_blur	low	44.6

Table 4: Top corruptions by Relative Corruption.

	MMBench		MMMU-Pro	
Augmentation	**Drop**	**Tier**	**Drop**	**Tier**
Autocontrast	0.0	Benign	−0.2	Benign
Grayscale	3.2	Moderate	0.4	Benign
Channel Swap	3.2	Moderate	0.0	Benign
Equalize	3.5	Moderate	0.3	Benign
Horizontal Flip	6.9	Moderate	3.6	Moderate
Invert	**10.1**	**Catastrophic**	1.4	Mild
Vertical Flip	**10.3**	**Catastrophic**	4.2	Moderate

Table 5: Binary transform drops (pp), mean over 9 direct-mode models. `flip_v` and `invert` are catastrophic on MMBench ($\Delta > 10$).

5.5 Catastrophic vs. Mild Distribution

Let $\Delta_{A_k,s} = \text{Acc}_{\text{clean}} - \text{Acc}_{A_k,s}$ denote the accuracy drop (percentage points). We define tiers: *benign* ($\Delta \leq 1$), *mild* ($1 < \Delta \leq 3$), *moderate* ($3 < \Delta \leq 10$), *catastrophic* ($\Delta > 10$), and *positive* ($\Delta < 0$, i.e., corruption improves accuracy). Table 18 reports tier distributions by severity level (direct mode, 9 models × 42 severity-based augmentations per level, plus 9 × 7 binary). A key finding: *binary augmentations on MMBench have the same catastrophic count (9) as mid-severity*, despite being trivial transforms—further evidence that spatial manipulations (flips) and color inversions are disproportionately harmful.

5.6 RCE Analysis: Severity Trends and Adversarial Regimes

RCE (defined in Section 3.1) normalizes corruption impact by visual reliance, enabling comparison across models with different VG. We report the *mean RCE over all 133 configurations*: $\text{mRCE} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \text{RCE}_c$, and additionally report severity-sliced means (Low/Mid/High over 42 configs each) and Binary (7 configs) in Table 20.

RCE by Severity. Table 20 reports mean RCE across models. On MMBench, RCE escalates from 1.6% (low) to 9.7% (high) to 11.5% (binary). MMMU-Pro shows higher RCE despite lower absolute drops because its smaller Visual Gain (11.9 vs. 46.7 points) amplifies relative impact. Notably, two configurations on MMMU-Pro exceed 100% RCE (`upsample:high` and `elastic_transform:high` for Qwen3-VL-4B), indicating truly adversarial corruptions.

Model-Specific RCE. Table 3 includes per-model RCE and Visual Gain. On MMBench, InternVL3.5-4B has the highest RCE (7.7%), losing nearly 8% of its visual contribution on average, while Qwen3-VL-30B is most resilient (3.5%). On MMMU-Pro, Gemma-3-12B suffers 24.2% RCE—corruptions destroy nearly a quarter of its visual benefit. Most strikingly, *Qwen3-VL-4B achieves negative RCE (−6.8%)* on MMMU-Pro, meaning corruptions *improve* performance relative to clean images, confirming its minimal visual reliance on reasoning tasks.

Worst-Case RCE by Augmentation. Table 4 lists the top corruptions by RCE. On MMBench, `upsample:high` destroys 65.6% of visual contribution—over half the benefit of having an image. On MMMU-Pro, `zoom_blur:high` reaches 77.6% RCE, and two configurations exceed 100% (adversarial). Binary transforms (e.g., `flip_v`) achieve 22% RCE on MMBench despite being trivial operations, confirming their outsized impact relative to visual reliance.

5.7 Mean Corruption Error (mCE)

Following ImageNet-C, we also compute mean Corruption Error (mCE) as a reference-normalized robustness ranking. We report mCE results and details in Appendix C.

5.8 Qualitative Analysis of Failure Modes

Figures 3–14 (Appendix G) visualize all 49 augmentations on a representative image. The dominant pattern is that corruptions that alter *spatial structure* are most harmful: even low-severity spatial/resampling effects (`glass_blur`, `upsample`, `elastic_transform`) often exceed the damage from visually severe photometric distortions (noise, compression, color shifts). This suggests VLMs rely strongly on spatial consistency and alignment; resampling artifacts that perturb boundaries or relative geometry are especially disruptive, whereas structure-preserving photometric degradations are comparatively well tolerated.

Why spatial fragility? We hypothesize this stems from patch-based ViT encoders: local rearrangements/distortions (glass blur, elastic transforms) and interpolation artifacts from resampling (`upsample`/`downsample`) can shift patch statistics and misalign pretrained features, while photometric changes largely preserve spatial relationships. This matches Figure 2, where resampling/geometry corruptions repeatedly rank among the most harmful, including at low severity.

Flip-rate evidence. Answer-flip analysis (correct on clean → incorrect under corruption) shows substantially higher flip rates for spatial/resampling corruptions than photometric ones on MMBench (Appendix D.12).

Systemic vs. unique failures. Catastrophic pairs are largely shared across models: on MMBench, the most consistent top failures are `upsample:high`, `elastic_transform:high`, and `upsample:mid`. On MMMU-Pro, catastrophic pairs are rarer; when they occur, `zoom_blur` (mid/high) and `elastic_transform:high` dominate (see Section D.1).

Resampling dominates tail risk. Across datasets, worst-case and catastrophic outcomes concentrate in resampling/geometry operations (e.g., `upsample`, `elastic_transform`, `zoom_blur`), implicating interpolation artifacts as a primary driver; we quantify their share among catastrophic cases ($\Delta > 10$) in Appendix E.2.

Family-specific vulnerabilities. Severe-failure rates vary by family and do not track parameter count: on MMBench, they range from 3.8% (Qwen3-VL-4B/30B) to 9.8% (InternVL3.5-4B). Gaps are pronounced for specific corruptions: `shot_noise:high` drops Gemma by 12.9 points vs. Qwen by 5.12; `pixelate:high` drops InternVL by 12.97 vs. Qwen by 6.06; and `downsample:high` drops InternVL by 9.73 vs. Qwen by 4.92. A compact family-by-augmentation matrix appears in Appendix D.7, with detailed breakdowns in Appendix D.9.

6 Conclusion

We introduced VLM-RobustBench, a benchmark showing that current VLMs are *semantically strong but spatially fragile*. Evaluating 11 open-weights LVLMs (4B–30B) across 49 augmentation types reveals four key findings: (1) visual severity is a weak proxy for difficulty—low-severity `glass_blur` (8–11pp) can exceed many high-severity corruptions; (2) trivial binary transforms can be catastrophic—vertical flip (10.3pp) and inversion (10.1pp) on MMBench; (3) resampling/geometric artifacts (`upsample`, `elastic_transform`) drive the largest failures (up to 34pp); and (4) visual reliance differs sharply by benchmark, with MMBench more visually grounded and MMMU-Pro more reliant on language priors.

Recommendations for VLM Development.

1. **Geometric Data Augmentation:** Go beyond color jitter/mixup to include resampling, elastic deformations, flips, and blur during pretraining.
2. **Robustness-Aware Evaluation:** Report spatial-corruption splits (e.g., “clean vs. flipped vs. resampled”) to penalize brittleness to simple geometry.
3. **Visual reliance:** Provide results on truly visually grounded inputs to better characterize visual reasoning capability.
4. **Family-Specific Curricula:** Target family-specific vulnerability fingerprints (e.g., flip sensitivity) rather than generic noise-only augmentation.

Acknowledgements

Rohit Saxena was supported by the Engineering and Physical Sciences Research Council (EPSRC) through the AI Hub in Generative Models (grant number EP/Y028805/1). Pasquale Minervini was partially funded by ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence), EPSRC (grant no. EP/W002876/1), and a donation from Accenture LLP.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10211–10221, 2021. doi: 10.1109/ICCV48922.2021.01007.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding, 2026. URL https://arxiv.org/abs/2601.10611.

Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R. Fung. Unveiling the lack of LVLM robustness to fundamental visual variations: Why and path forward. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20222–20242, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1037. URL https://aclanthology.org/2025.findings-acl.1037/.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme:

A comprehensive evaluation benchmark for multimodal large language models, 2025. URL https://arxiv.org/abs/2306.13394.

Ankit Goyal, Hugo Hadfield, Xuning Yang, Valts Blukis, and Fabio Ramos. Vla-0: Building state-of-the-art vlas with zero modification. *arXiv preprint arXiv:2510.13054*, 2025.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. URL https://openreview.net/pdf?id=HJz6tiCqYm.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, October 2021.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: evaluating vision-language models on natural adversarial samples. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. On the perception bottleneck of VLMs for chart understanding. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10829–10841, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.573. URL https://aclanthology.org/2025.findings-emnlp.573/.

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. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, pp. 216–233, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72657-6. doi: 10.1007/978-3-031-72658-3_13. URL https://link.springer.com/chapter/10.1007/978-3-031-72658-3_13.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models, 2023. URL https://arxiv.org/abs/2212.07016.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. MM1: methods, analysis and insights from multimodal LLM pre-training. In *ECCV (29)*, volume 15087 of *Lecture Notes in Computer Science*, pp. 304–323. Springer, 2024.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 2071–2081. AAAI Press, 2022. doi: 10.1609/AAAI.V36I2.20103. URL https://doi.org/10.1609/aaai.v36i2.20103.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i19.30150. URL https://doi.org/10.1609/aaai.v38i19.30150.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Erfan Shayegani, Yue Dong, and Nael B. Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=plmBsXHxgR.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris

Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9568–9578. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00914. URL https://doi.org/10.1109/CVPR52733.2024.00914.

Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. Analysing the robustness of vision-language-models to common corruptions, 2025. URL https://arxiv.org/abs/2504.13690.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025a. URL https://arxiv.org/abs/2508.18265.

Zhaowei Wang, Wenhao Yu, Xiyu REN, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. MMLongbench: Benchmarking long-context vision-language models effectively and thoroughly. In *ICML 2025 Workshop on Long-Context Foundation Models*, 2025b. URL https://openreview.net/forum?id=zsdJSkeS9S.

Junjie Ye, Yilong Wu, Songyang Gao, Caishuang Huang, Sixian Li, Guanyu Li, Xiaoran Fan, Qi Zhang, Tao Gui, and Xuanjing Huang. RoTBench: A multi-level benchmark for evaluating the robustness of large language models in tool learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 313–333, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.19. URL https://aclanthology.org/2024.emnlp-main.19/.

Han Yu, Jiashuo Liu, Xingxuan Zhang, Jiayun Wu, and Peng Cui. A survey on evaluation of out-of-distribution generalization, 2024. URL https://arxiv.org/abs/2403.01874.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.org/10.1109/CVPR52733.2024.00913.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN

979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Chenyue Zhou, Mingxuan Wang, Yanbiao Ma, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, Fei Luo, Xiaohua Chen, Xiaoshuai Hao, Hehan Li, Andi Zhang, Wenxuan Wang, Kaiyan Zhang, Guoli Jia, Lingling Li, Zhiwu Lu, Yang Lu, and Yike Guo. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models, 2025. URL https://arxiv.org/abs/2509.25373.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27378–27394. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhou22m.html.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, pp. 1–20, 2024. doi: 10.1109/TIV.2024.3402136.

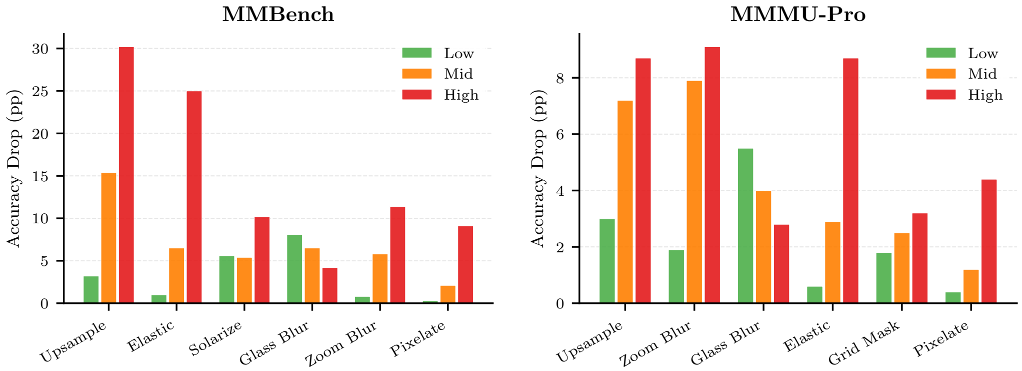

Figure 2: Top corruptions by severity (mean drop, 9 models). Resampling corruptions (`upsample`, `elastic_transform`) dominate at mid/high severity, while `glass_blur` shows an inverted pattern (Low > Mid > High) on both datasets.

A Impact Statement

This work aims to improve robustness evaluation for vision-language models. We do not anticipate direct negative societal impacts from the benchmark itself; however, insights from robustness gaps can inform safer deployment and failure mitigation in real-world applications. We consider this research particularly useful for the development of foundation models for robotics that directly leverage VLMs as their backbones (e.g., (Intelligence et al., 2025; Goyal et al., 2025), inter alia). Because these embodied systems are heavily reliant on VLMs for high-level reasoning and perception, they inherently inherit the foundational weaknesses of their backbones. These vulnerabilities are often exacerbated in physical settings, where robots are routinely exposed to diverse visual perturbations and environmental corruptions—ranging from lighting shifts to sensor noise—that can compromise safety and operational reliability.

B Additional Figures

C Mean Corruption Error (mCE)

Following ImageNet-C Hendrycks & Dietterich (2019), we compute *mean Corruption Error (mCE)* to compare model robustness against a reference baseline. For each corruption type c, we define:

$$\mathrm{CE}_c = \frac{\sum_s E_{c,s}^{\text{model}}}{\sum_s E_{c,s}^{\text{ref}}} \tag{2}$$

where $E = 1 - \mathrm{Acc}$ is the error rate. For the 42 severity-based corruptions, errors are summed over $s \in \{\text{low}, \text{mid}, \text{high}\}$; for the 7 binary corruptions, a single error term is used. The reference model is the one with the lowest baseline accuracy (analogous to AlexNet): *Gemma-3-12B* for MMBench (85.3%) and *Molmo2-8B* for MMMU-Pro (31.2%). Mean CE aggregates across all 49 corruption types: $\mathrm{mCE} = \frac{1}{49}\sum_c \mathrm{CE}_c$.

Table 6 reveals that *Qwen3-VL-30B is the most robust model on MMBench* with only 62.9% of the reference error rate, while *InternVL3.5-14B leads on MMMU-Pro at 85.0%*. Notably, mCE rankings differ from RCE rankings because mCE compares absolute error rates across models, while RCE measures each model's degradation relative to its own visual contribution. This complementary view shows that models with high baseline accuracy (Qwen family) tend to have lower mCE, while models with high visual gain but moderate baselines (InternVL3.5) show better RCE but higher mCE. In summary, we use tail-risk metrics (worst-case drop, severe-failure rate) for deployment risk assessment, mCE for cross-model robustness ranking, and RCE to factor out language-prior reliance.

D Additional Results

D.1 Worst-Case Augmentations

D.2 Dataset Category Sensitivity

D.3 Scaling Trends

D.4 Prompting-Mode Performance (Qwen)

D.5 Prompting-Mode Tier Distributions (Qwen)

D.6 Positive Augmentations

A small number of augmentations yield negative Δ (i.e., higher accuracy than baseline). On MMBench, `brightness:low/mid`, `gamma_up:low/mid`, and `gaussian_noise:low` show marginal gains (-0.1 to -0.2 pp); on MMMU-Pro, `hue_shift:low`, `gaussian_blur:low`, and `speckle_noise:low/mid` exhibit similar small effects. Given the magnitude (<0.5 pp), these may reflect noise, mild regularization, or dataset-specific priors rather than robust improvements; we note them for completeness but do not draw strong conclusions.

D.7 Family-Level Vulnerability Matrix

D.8 Per-Model Top-5 Corruptions

Tables 13 and 14 provide per-model top-5 most harmful corruptions at each severity. The aggregate patterns from Figure 2 hold consistently across models: `glass_blur` dominates at low severity, while `upsample` and `elastic_transform` dominate at mid/high severities.

Model	MMBench mCE	MMMU-Pro mCE
Qwen3-VL-30B	**62.9**	89.0
Qwen3-VL-8B	70.8	95.0
Qwen3-VL-4B	77.5	98.9
Molmo2-8B	78.2	100.0 (ref)
Molmo2-4B	79.2	100.0
InternVL3.5-8B	81.2	89.1
InternVL3.5-14B	92.0	**85.0**
InternVL3.5-4B	98.3	93.1
Gemma-3-12B	100.0 (ref)	101.1

Table 6: Mean Corruption Error (mCE) following ImageNet-C methodology. Lower is better; 100% matches the reference model.

Model	MMBench worst (aug, level, drop)	MMMU-Pro worst (aug, level, drop)
Qwen3-VL-4B	upsample (high, 26.3)	elastic_transform (high, 7.4)
Qwen3-VL-8B	upsample (high, 30.2)	elastic_transform (high, 8.9)
Qwen3-VL-30B	upsample (high, 29.4)	elastic_transform (high, 13.9)
InternVL3.5-4B	upsample (high, 30.6)	zoom_blur (high, 11.1)
InternVL3.5-8B	upsample (high, 31.6)	zoom_blur (high, 12.3)
InternVL3.5-14B	upsample (high, 29.4)	zoom_blur (high, 9.6)
Molmo2-4B	upsample (high, 33.1)	upsample (high, 5.6)
Molmo2-8B	upsample (high, 33.9)	upsample (high, 5.6)
Gemma-3-12B	upsample (high, 32.1)	upsample (high, 9.9)
Qwen3-VL-4B-Thinking	upsample (high, 29.5)	upsample (mid, 19.1)
Qwen3-VL-8B-Thinking	upsample (high, 30.8)	upsample (high, 23.1)

Table 7: Worst-case augmentation per model. Drop is baseline minus accuracy under that augmentation (percentage points); "bin" denotes binary augmentations.

MMBench		MMMU-Pro	
Category	Drop	Subject	Drop
image_style	5.30	Art	4.75
attribute_comparison	5.26	Music	3.90
structuralized_imagetext_understanding	4.81	Economics	3.69
social_relation	4.57	Art_Theory	3.46
nature_relation	4.51	Pharmacy	3.35

Table 8: Top-5 dataset categories with the largest average drops (percentage points), aggregated across models in direct mode.

Family	MMMU-Pro slope (R^2, n)	MMBench slope (R^2, n)
Qwen3-VL	+2.95 (1.00, n=3)	-0.38 (0.17, n=3)
InternVL3.5	-0.94 (0.12, n=3)	-1.44 (0.89, n=3)
Molmo2	-1.66 (1.00, n=2)	-1.00 (1.00, n=2)

Table 9: Scaling of robustness drop with model size within families (direct mode). Slope is the change in drop per log10(parameters); negative values indicate improved robustness with scale.

Model	MMMU-Pro		MMBench	
	Baseline	Drop	Baseline	Drop
Qwen3-VL-4B-COT	32.1	+1.3	86.9	+2.5
Qwen3-VL-8B-COT	42.0	+3.1	89.6	+3.2
Qwen3-VL-4B-Thinking	43.5	+2.5	89.3	+1.8
Qwen3-VL-8B-Thinking	50.0	+3.8	91.1	+3.0

Table 10: Baseline accuracy and mean drop for Qwen models under COT and thinking modes. Drop is averaged over the 133 corrupted configurations. Direct-mode results are in Table 3.

Mode	Mild	Moderate	Catastrophic	Positive
MMBench (Qwen models)				
Direct	23.6	17.0	3.5	21.6
COT	28.9	23.3	5.6	16.9
Thinking	27.1	21.1	4.1	16.2
MMMU-Pro (Qwen models)				
Direct	19.8	10.3	1.3	42.1
COT	19.5	21.4	6.4	26.7
Thinking	18.8	19.9	13.2	27.8

Table 11: Tier shares (%) for Qwen models by prompting mode.

Aug-Sev	Gemma	Qwen	InternVL	Molmo
shot_noise:high	12.90	5.12	12.39	5.62
pixelate:high	11.37	6.06	12.97	8.27
downsample:high	8.79	4.92	9.73	6.74
solarize:high	10.08	9.03	12.93	8.50

Table 12: Family-level mean drops (MMBench, direct) for selected aug-level pairs.

Table 13: Per-model top-5 most harmful corruptions at each severity (MMBench, direct mode). Values are accuracy drops in percentage points.

Model	Low	Mid	High
Gemma3-12B	glass_blur (10.7), solarize (3.6), shot_noise (2.5), upsample (2.5), text_overlay (2.3)	upsample (15.2), shot_noise (8.4), glass_blur (8.2), elastic_transform (6.7), zoom_blur (6.0)	upsample (32.1), elastic_transform (23.4), zoom_blur (13.2), shot_noise (12.9), pixelate (11.4)
InternVL3.5-4B	glass_blur (10.7), solarize (7.3), upsample (3.9), shot_noise (3.2), zoom_blur (2.3)	upsample (17.6), glass_blur (9.6), elastic_transform (8.6), solarize (7.3), zoom_blur (7.3)	upsample (30.6), elastic_transform (25.3), solarize (14.9), pixelate (13.0), shot_noise (12.7)
InternVL3.5-8B	glass_blur (9.3), solarize (8.1), upsample (4.5), shot_noise (3.4), grid_mask (2.8)	upsample (17.1), elastic_transform (8.7), zoom_blur (8.0), glass_blur (7.4), solarize (7.3)	upsample (31.6), elastic_transform (28.8), zoom_blur (15.9), shot_noise (13.7), pixelate (12.7)
InternVL3.5-14B	glass_blur (9.5), solarize (5.3), shot_noise (3.3), upsample (3.2), grid_mask (1.1)	upsample (17.9), glass_blur (7.8), elastic_transform (6.6), zoom_blur (6.5), solarize (5.5)	upsample (29.4), elastic_transform (24.7), pixelate (13.2), zoom_blur (12.5), solarize (11.4)
Molmo2-4B	glass_blur (7.4), solarize (5.5), upsample (4.1), shot_noise (2.1), grid_mask (1.9)	upsample (15.9), zoom_blur (6.3), glass_blur (5.7), elastic_transform (4.9), solarize (4.3)	upsample (33.1), elastic_transform (23.9), zoom_blur (10.0), center_occlusion (9.0), pixelate (7.8)
Molmo2-8B	glass_blur (6.3), solarize (4.6), upsample (3.5), grid_mask (1.1), zoom_blur (0.6)	upsample (18.9), glass_blur (6.7), elastic_transform (6.6), zoom_blur (4.5), solarize (4.3)	upsample (33.9), elastic_transform (25.6), zoom_blur (9.4), solarize (9.1), pixelate (8.7)
Qwen3-VL-4B	glass_blur (7.0), solarize (4.5), upsample (2.6), random_occlusion (1.2), elastic_transform (1.1)	upsample (13.1), elastic_transform (5.7), glass_blur (5.4), zoom_blur (4.5), solarize (4.1)	upsample (26.3), elastic_transform (25.8), zoom_blur (9.6), solarize (8.3), center_occlusion (6.0)
Qwen3-VL-8B	glass_blur (8.2), solarize (5.9), grid_mask (2.3), shot_noise (1.8), upsample (1.8)	upsample (14.2), glass_blur (7.0), elastic_transform (6.1), solarize (6.0), zoom_blur (5.2)	upsample (30.2), elastic_transform (27.0), zoom_blur (12.3), solarize (10.3), pixelate (7.8)
Qwen3-VL-30B	solarize (6.1), glass_blur (5.6), upsample (2.6), random_occlusion (1.2), watermark (1.1)	upsample (11.8), elastic_transform (6.2), solarize (4.9), glass_blur (4.3), zoom_blur (4.3)	upsample (29.4), elastic_transform (23.2), zoom_blur (10.9), solarize (8.4), center_occlusion (5.9)

Table 14: Per-model top-5 most harmful corruptions at each severity (MMMU-Pro, direct mode). Values are accuracy drops in percentage points.

Model	**Low**	**Mid**	**High**
Gemma3-12B	glass_blur (8.9), solarize (4.9), brightness (3.7), grid_mask (3.7), defocus_blur (3.4)	upsample (9.6), zoom_blur (8.9), elastic_transform (5.2), glass_blur (4.6), motion_blur (4.6)	upsample (9.9), zoom_blur (9.9), elastic_transform (9.3), downsample (5.9), center_occlusion (5.6)
InternVL3.5-4B	glass_blur (5.6), upsample (5.6), grid_mask (3.7), solarize (3.1), watermark (2.5)	zoom_blur (9.6), upsample (6.5), glass_blur (5.2), grid_mask (4.9), elastic_transform (4.6)	zoom_blur (11.1), elastic_transform (9.3), upsample (8.9), motion_blur (5.9), downsample (5.6)
InternVL3.5-8B	glass_blur (5.6), zoom_blur (4.6), grid_mask (4.3), motion_blur (3.7), rotate (2.8)	zoom_blur (11.7), upsample (8.0), random_occlusion (4.9), glass_blur (4.6), motion_blur (4.6)	zoom_blur (12.3), elastic_transform (10.5), upsample (9.6), pixelate (7.7), downsample (6.5)
InternVL3.5-14B	glass_blur (5.9), upsample (3.4), snow (1.5), zoom_blur (1.5), grid_mask (1.2)	zoom_blur (8.3), upsample (7.4), glass_blur (5.2), rotate (2.8), grid_mask (2.5)	zoom_blur (9.6), elastic_transform (9.3), upsample (9.3), pixelate (5.6), downsample (4.6)
Molmo2-4B	glass_blur (4.3), rotate (1.5), shot_noise (1.2), add_border (0.9), brightness_up (0.9)	upsample (4.3), glass_blur (3.1), zoom_blur (3.1), brightness_up (1.9), saturation (1.9)	upsample (5.6), zoom_blur (4.6), elastic_transform (4.0), downsample (2.5), brightness (1.9)
Molmo2-8B	glass_blur (4.0), zoom_blur (1.9), upsample (1.5), center_occlusion (1.2), gamma_up (0.6)	upsample (3.7), zoom_blur (3.7), elastic_transform (2.5), rotate (2.5), glass_blur (2.2)	upsample (5.6), elastic_transform (4.9), zoom_blur (4.6), rotate (3.4), gamma_up (1.9)
Qwen3-VL-4B	glass_blur (2.8), upsample (1.5), gamma (0.6), saturation (0.3), sharpen (0.0)	upsample (5.2), zoom_blur (5.2), glass_blur (2.2), solarize (1.5), downsample (0.9)	elastic_transform (7.4), upsample (7.4), zoom_blur (6.5), downsample (2.2), salt_pepper (2.2)
Qwen3-VL-8B	glass_blur (5.2), watermark (2.8), upsample (2.2), salt_pepper (1.2), snow (1.2)	zoom_blur (8.0), upsample (7.4), watermark (4.0), glass_blur (3.7), elastic_transform (2.2)	elastic_transform (8.9), upsample (8.9), zoom_blur (8.6), pixelate (5.2), brightness_up (2.5)
Qwen3-VL-30B	glass_blur (6.8), upsample (6.2), grid_mask (3.7), affine (2.8), watermark (2.8)	upsample (12.0), zoom_blur (12.0), elastic_transform (6.2), glass_blur (5.2), grid_mask (4.9)	elastic_transform (13.9), zoom_blur (13.9), upsample (12.7), grid_mask (7.4), pixelate (6.5)

D.9 Detailed Robustness Results by Family

We provide the complete breakdown of accuracy drops for key augmentation types across the four model families. Values represent the family-averaged drop (percentage points) at Low, Mid, and High severity on MMBench.

Table 15: **Qwen3-VL Family:** Mean accuracy drops on MMBench. Note the resilience to noise (e.g., Gaussian Noise) vs. fragility to resampling (Upsample).

Augmentation	**Low**	**Mid**	**High**
Upsample	2.31	13.05	28.65
Elastic Transform	0.78	6.02	25.32
Zoom Blur	0.55	4.65	10.94
Solarize	5.47	5.00	9.03
Glass Blur	6.96	5.59	4.10
Pixelate	0.16	0.35	6.06
Shot Noise	0.55	2.19	5.12
Brightness	0.23	0.23	3.99
JPEG Compression	0.20	0.27	0.31

Table 16: **InternVL3.5 Family:** Mean accuracy drops on MMBench. This family shows higher sensitivity to pixelation and noise compared to Qwen.

Augmentation	**Low**	**Mid**	**High**
Upsample	3.83	17.55	30.56
Elastic Transform	0.78	7.94	26.30
Zoom Blur	1.60	7.23	13.36
Pixelate	0.55	1.60	12.97
Solarize	6.88	6.68	12.93
Shot Noise	3.28	5.98	12.39
Glass Blur	9.81	8.28	7.11
Motion Blur	0.74	3.05	7.31
JPEG Compression	0.20	0.27	0.63

Table 17: **Gemma 3 & Molmo 2 Families:** Comparison of key failure modes (High Severity Drops).

Augmentation (High)	**Gemma 3 (12B)**	**Molmo 2 (Avg)**
Upsample	32.12	33.47
Elastic Transform	23.45	24.74
Zoom Blur	13.25	9.67
Shot Noise	12.90	5.62
Pixelate	11.37	8.26
Solarize	10.08	8.50
Glass Blur	5.51	6.10

D.10 Tier Distributions

Table 18 reports tier distributions by severity level (direct mode, 9 models × 42 severity-based augmentations per level, plus 9 × 7 binary).

Table 19 breaks down tier shares per model, highlighting that catastrophic and positive rates vary widely even within families.

D.11 RCE by Severity

Table 20 reports mean RCE across models by severity level. Higher RCE on MMMU-Pro reflects its smaller Visual Gain denominator. Two configurations on MMMU-Pro exceed 100% RCE (`upsample:high` and `elastic_transform:high` for Qwen3-VL-4B), indicating truly adversarial corruptions.

Dataset	Severity	Benign	Mild	Moderate	Catastrophic
MMBench	Low	304	48	24	2
	Mid	182	127	60	9
	High	94	113	133	38
	Binary	9	12	33	**9**
MMMU-Pro	Low	276	77	25	0
	Mid	230	93	52	3
	High	188	105	79	6
	Binary	37	15	11	0

Table 18: Tier distribution by severity level (counts out of 378 for severity-based, 63 for binary). Tiers use fixed thresholds: Catastrophic = $\Delta > 10$pp. Binary augmentations on MMBench produce 9 catastrophic cases—matching mid-severity—driven by vertical flip and color inversion.

MMBench (Direct)				
Model	**Mild %**	**Moderate %**	**Catastrophic Rate (%)**	**Positive %**
Qwen3-VL-4B	21.8	18.0	2.3	24.1
Qwen3-VL-8B	28.6	19.5	4.5	11.3
Qwen3-VL-30B	18.0	13.5	3.8	29.3
InternVL3.5-4B	29.3	30.1	8.3	3.8
InternVL3.5-8B	26.3	28.6	6.8	11.3
InternVL3.5-14B	18.0	21.1	6.8	14.3
Molmo2-4B	27.1	21.1	2.3	8.3
Molmo2-8B	21.1	17.3	2.3	17.3
Gemma-3-12B	35.3	18.8	6.8	6.0
MMMU-Pro (Direct)				
Model	**Mild %**	**Moderate %**	**Catastrophic Rate (%)**	**Positive %**
Qwen3-VL-4B	9.0	3.8	0.0	74.4
Qwen3-VL-8B	15.0	7.5	0.0	39.8
Qwen3-VL-30B	35.3	19.5	3.8	12.0
InternVL3.5-4B	24.8	16.5	0.8	26.3
InternVL3.5-8B	38.3	24.8	2.3	6.0
InternVL3.5-14B	17.3	10.5	0.0	37.6
Molmo2-4B	15.8	5.3	0.0	19.5
Molmo2-8B	12.8	5.3	0.0	51.1
Gemma-3-12B	49.6	32.3	0.0	3.0

Table 19: Per-model tier shares (%) across 133 augmentation configurations (direct mode). Tiers use fixed thresholds: Catastrophic = $\Delta > 10$pp (distinct from the relative Severe-Failure Rate in Table 3). Benign shares are omitted for brevity.

D.12 PER-EXAMPLE FLIP DECOMPOSITION

Table 21 shows answer-flip rates (fraction of questions correct on clean that become incorrect under corruption) for a representative model (Qwen3-VL-8B) on MMBench. Spatial/resampling corruptions cause substantially more flips than photometric ones, even when the latter are at high severity.

Accuracy drop Δ conflates two opposing effects: examples the model previously answered correctly but now fails (*harmful flips*, Flip$^+$), and examples it previously failed but now succeeds (*helpful flips*, Flip$^-$):

$$\text{Flip}^+ = \Pr(\text{correct}_{\text{clean}} \wedge \text{wrong}_{\text{corrupted}}) \tag{3}$$

$$\text{Flip}^- = \Pr(\text{wrong}_{\text{clean}} \wedge \text{correct}_{\text{corrupted}}) \tag{4}$$

with net accuracy drop $\Delta = \text{Flip}^+ - \text{Flip}^-$. Flip rates are computed per example then averaged across the dataset; when aggregating across models or corruptions, we report macro-averages. This decomposition reveals whether drops stem from genuine failures or are partially masked by compensating gains.

Dataset	Severity	Mean RCE (%)	Interpretation
MMBench	Low	1.6	Minimal visual loss
	Mid	4.0	Moderate impact
	High	9.7	∼10% visual loss
	Binary	11.5	Highest relative harm
MMMU-Pro	Low	2.7	Low but > MMBench
	Mid	6.3	Moderate
	High	13.0	Severe relative loss
	Binary	10.1	High relative harm

Table 20: Mean Relative Corruption Error by severity. RCE measures what fraction of visual contribution is destroyed. Higher RCE on MMMU-Pro reflects its smaller Visual Gain denominator.

Corruption	Severity	Flip rate
`glass_blur`	low	11.7%
`upsample`	high	36.3%
`elastic_transform`	high	32.9%
`brightness`	high	4.4%
`jpeg_compression`	high	1.6%

Table 21: Answer-flip rates for Qwen3-VL-8B on MMBench. Spatial/resampling corruptions (top) cause substantially more flips than photometric ones (bottom), even at high severity.

Flip Rates by Severity. Table 22 reports flip rates aggregated by severity level. On MMBench, harmful flips escalate sharply (1.8% at low → 6.3% at high), while helpful flips remain low (1.1–1.8%), confirming that accuracy drops reflect genuine failures rather than noise. Binary augmentations show the highest harmful flip rate (12.3%)—over 6× the low-severity rate—with minimal compensation (1.6% Flip^{-}). For severity-based corruptions, MMMU-Pro exhibits smaller $\text{Flip}^{+}/\text{Flip}^{-}$ ratios (1.2–1.6× vs. 1.7–3.6× on MMBench), consistent with its lower visual reliance.

Model-Specific Patterns. Table 23 reveals striking model differences. On MMBench, Molmo2 models exhibit the highest $\text{Flip}^{+}/\text{Flip}^{-}$ ratios (3.9–4.2), indicating "clean" degradation with minimal lucky compensations. Gemma-3-12B and InternVL3.5 models have the highest absolute Flip^{+} rates (5.1–5.5%), making them the most fragile. On MMMU-Pro, Qwen3-VL-4B achieves a ratio below 1.0 (0.85), meaning corruptions *help more than hurt*—strong evidence of minimal visual reliance on reasoning tasks.

Binary Augmentation Flips. Table 24 drills into per-augmentation flip rates for binary transforms. On MMBench, `flip_v` and `invert` consistently cause 10–15% harmful flips across models, with InternVL3.5-4B reaching 15.7% for vertical flip. On MMMU-Pro, the same augmentations show dramatically lower Flip^{+} (5–11%) and higher Flip^{-}, with some models (Molmo2-8B) showing near-zero or negative net flips for `flip_h`—confirming that spatial transforms harm perception far more than reasoning.

D.13 Answer-Flip Rates Across Models

Table 25 extends the flip-rate analysis to all direct-mode models for selected corruptions. Flip rate is defined as the fraction of questions answered correctly on clean images that become incorrect under corruption.

Dataset	Severity	Flip$^+$ (%)	Flip$^-$ (%)	Ratio
MMBench	Low	1.79	1.05	1.70
	Mid	3.37	1.49	2.26
	High	6.33	1.78	3.56
	Binary	12.27	1.63	7.53
MMMU-Pro	Low	2.16	1.74	1.24
	Mid	3.20	2.31	1.39
	High	4.47	2.82	1.59
	Binary	5.30	2.65	2.00

Table 22: Flip rates by severity level (averaged over models). Flip$^+$ = harmful (correct→wrong), Flip$^-$ = helpful (wrong→correct). Higher ratios indicate "purer" degradation with less compensating gains.

	MMBench		MMMU-Pro	
Model	Flip$^+$	Ratio	Flip$^+$	Ratio
Qwen3-VL-4B	3.69	2.55	2.37	**0.85**
Qwen3-VL-8B	4.46	2.76	3.64	1.21
Qwen3-VL-30B	3.72	2.20	4.16	2.08
InternVL3.5-4B	**5.13**	3.79	3.42	1.79
InternVL3.5-8B	**5.27**	3.27	3.93	2.74
InternVL3.5-14B	4.45	3.01	2.94	1.44
Molmo2-4B	3.63	3.94	2.75	1.32
Molmo2-8B	3.24	**4.19**	2.50	1.06
Gemma-3-12B	5.53	2.31	5.13	2.19

Table 23: Model flip rates (%) and Flip$^+$/Flip$^-$ ratios. Higher ratios indicate purer degradation. Bold: highest Flip$^+$ (most fragile) and extreme ratios.

	MMBench			MMMU-Pro		
Augmentation	Flip$^+$	Flip$^-$	Net	Flip$^+$	Flip$^-$	Net
flip_v	12.4	2.0	10.4	8.3	4.1	4.2
flip_h	9.0	1.9	7.1	7.2	3.6	3.6
invert	12.1	1.7	10.4	3.7	2.4	1.3
channel_swap	4.5	1.2	3.3	1.1	1.2	−0.1
equalize	5.1	1.6	3.5	2.7	2.6	0.1
grayscale	4.6	1.5	3.2	1.8	1.5	0.3
autocontrast	0.2	0.2	0.0	0.3	0.6	−0.2

Table 24: Binary augmentation flip rates (%) averaged over 9 direct-mode models. Vertical flip and invert dominate harmful flips on MMBench but show reduced impact on MMMU-Pro.

Model	`glass:low`	`ups:high`	`elast:high`	`bright:high`	`jpeg:high`
Qwen3-VL-4B	8.4	29.2	29.1	2.7	1.4
Qwen3-VL-8B	10.6	32.4	29.1	4.0	1.4
Qwen3-VL-30B	8.6	31.2	26.7	2.5	1.5
InternVL3.5-4B	12.8	33.2	27.9	4.8	2.3
InternVL3.5-8B	12.4	34.7	32.0	5.5	1.9
InternVL3.5-14B	12.2	32.4	27.2	3.9	1.4
Molmo2-4B	9.0	34.7	26.4	3.9	0.8
Molmo2-8B	7.9	35.4	27.1	3.8	1.1
Gemma3-12B	14.2	35.8	26.6	6.6	4.3

Table 25: Answer-flip rates (%) across 9 direct-mode models on MMBench. Spatial/resampling corruptions (columns 2–4) consistently cause higher flip rates than photometric ones (columns 5–6). Thinking models are excluded due to different output format.

E Quantitative Robustness Metrics

E.1 Severity Mismatch Metrics

Table 26 reports the consistency between visual severity levels and model performance drops. A high violation rate indicates that increasing visual severity does not reliably lead to larger performance drops.

Table 26: Severity mismatch metrics. **Violation Rate**: Fraction of augmentation trajectories where drop does not strictly increase with severity. **Mean Spearman** ρ: Rank correlation between severity and drop (averaged across models/augmentations).

Dataset	Violation Rate (%)	Mean Spearman ρ
MMBench	30.2	0.71
MMMU-Pro	56.1	0.34

E.2 Tail Risk Share from Spatial/Resampling Corruptions

Table 27 quantifies the contribution of spatial/resampling corruptions to catastrophic failures. We define "Spatial/Resampling" augmentations as: upsample, downsample, elastic_transform, zoom_blur, rotate, shear, affine, perspective_transform, and pixelate (included as a resolution/resampling artifact that disrupts spatial structure).

Table 27: Fraction of catastrophic cases ($\Delta > 10$) attributable to spatial/resampling corruptions.

Dataset	Share from Spatial/Resampling (%)	Top Contributors
MMBench	65.5	upsample, elastic_transform, zoom_blur
MMMU-Pro	100.0	zoom_blur, elastic_transform, upsample

E.3 Mean Corruption Error by Category

Following ImageNet-C methodology, we compute mean Corruption Error (mCE) to compare model robustness against a reference baseline. For each corruption type c, $\text{CE}_c = \frac{\sum_s E^{\text{model}}_{c,s}}{\sum_s E^{\text{ref}}_{c,s}}$ where $E = 1 - \text{Acc}$ is the error rate. The reference model is the one with lowest baseline accuracy: **Gemma-3-12B** (85.3%) for MMBench and **Molmo2-8B** (31.2%) for MMMU-Pro. Values below 100% indicate better robustness than the reference; values above 100% indicate worse robustness.

Table 28 reveals category-specific robustness patterns. On MMBench, **Qwen3-VL-30B** achieves the lowest mCE across all categories (56–71%), demonstrating consistent robustness. Noise corruptions show the best relative robustness (mean 76.9% across models), while Binary transforms show the worst (mean 86.2%). Notably, InternVL3.5-4B approaches or exceeds 100% mCE in most categories (Blur 101.2%, Binary 100.8%), indicating it is less robust than the reference Gemma model despite having similar baseline accuracy.

On MMMU-Pro, all models cluster near 100% mCE (range 85–102%), reflecting the harder dataset where even the reference model struggles. **InternVL3.5-14B** leads with the lowest overall mCE (85.0%), showing particular strength in Color/Tone (83.8%) and Weather (83.9%) categories. Conversely, Gemma-3-12B exceeds 100% mCE in 9/10 categories (up to 102.3% on Blur), indicating worse robustness than the Molmo2-8B reference. The tight clustering suggests that on challenging reasoning tasks, relative robustness differences between models diminish.

F Experiment Details

This section provides implementation details for reproducibility.

Table 28: Mean Corruption Error (mCE, %) by model and corruption category. Lower is better; 100% matches the reference model. Reference models: Gemma-3-12B for MMBench, Molmo2-8B for MMMU-Pro. Categories match Table 1.

MMBench (Reference: Gemma-3-12B, Baseline 85.3%)

Model	Base	Blur	Noise	Weath	Digi	Geom	Occl	Color	Resol	VLM	Bin	All
Qwen3-VL-4B	88.4	76.3	69.1	79.1	73.4	78.8	80.5	77.7	76.8	76.6	81.9	77.5
Qwen3-VL-8B	90.2	70.6	64.4	69.1	65.6	72.9	74.0	69.4	71.4	69.1	76.5	70.8
Qwen3-VL-30B	90.7	**58.7**	**56.0**	**64.0**	**56.7**	**63.1**	**64.5**	**62.0**	**64.8**	**61.0**	**70.6**	**62.9**
InternVL3.5-4B	86.3	101.2	93.6	99.0	96.0	98.8	99.2	96.7	99.7	94.9	100.8	98.3
InternVL3.5-8B	89.1	81.1	79.7	77.9	77.8	85.0	82.6	77.7	82.7	75.8	88.4	81.2
InternVL3.5-14B	86.6	93.2	86.0	91.1	89.3	91.1	94.2	91.7	92.2	86.4	98.2	92.0
Molmo2-4B	88.5	80.0	72.9	80.3	75.0	79.9	85.6	78.5	79.8	78.8	80.2	79.2
Molmo2-8B	88.4	79.5	70.1	76.9	77.4	80.5	81.0	77.8	81.9	75.1	79.6	78.2
Gemma-3-12B (ref)	85.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 29: Mean Corruption Error (mCE, %) by model and corruption category. Lower is better; 100% matches the reference model. Reference models: Gemma-3-12B for MMBench, Molmo2-8B for MMMU-Pro. Categories match Table 1.

MMMU-Pro (Reference: Molmo2-8B, Baseline 31.2%)

Model	Base	Blur	Noise	Weath	Digi	Geom	Occl	Color	Resol	VLM	Bin	All
Qwen3-VL-4B	31.5	99.2	99.8	98.7	99.0	96.4	97.6	98.5	100.5	98.4	100.3	98.9
Qwen3-VL-8B	35.2	95.2	95.2	94.5	95.6	94.4	93.8	94.3	95.1	96.0	96.2	95.0
Qwen3-VL-30B	40.7	89.9	88.1	88.8	89.7	89.2	91.2	87.5	89.9	88.7	89.7	89.0
InternVL3.5-4B	37.3	94.8	92.9	91.7	92.5	92.4	94.7	91.6	93.9	93.7	94.6	93.2
InternVL3.5-8B	41.0	91.6	88.8	88.0	90.0	88.7	91.2	87.6	89.0	87.0	90.6	89.1
InternVL3.5-14B	42.0	86.6	84.8	**83.9**	85.5	**84.2**	85.8	**83.8**	86.0	**84.4**	86.5	**85.0**
Molmo2-4B	31.8	99.2	100.7	100.4	99.6	97.6	99.1	100.4	99.7	100.7	101.5	100.0
Molmo2-8B (ref)	31.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Gemma-3-12B	33.0	102.3	101.8	101.1	100.2	99.5	101.9	100.1	101.6	101.0	101.9	101.1

F.1 Random Seeds

We use fixed random seeds throughout all experiments to ensure reproducibility:

- **Sampling seed**: 42 — used for stratified dataset sampling to select the 20% evaluation subset.
- **Augmentation seed**: 1234 — base seed for deterministic per-sample augmentation. Each sample i receives seed $(1234 \times 1000003 + i) \mod 2^{32}$ to ensure reproducible yet varied augmentations.
- **Generation seed**: 42 — used for thinking-mode models that require sampling-based decoding.

F.2 Dataset Sampling

To reduce computational costs while maintaining statistical validity, we evaluate on a 20% stratified subset of each benchmark:

- **MMBench**: 869 samples from 4,329 total (stratified by `category` field)
- **MMMU-Pro**: 532 samples from 2,658 total (stratified by `subject` field)

Stratified sampling ensures proportional representation of all question categories/subjects in the evaluation subset.

Table 30: Augmentation parameters by severity level. Values at low, mid, and high severity are used in experiments.

Category	Augmentation	Param	Low	Mid	High	Note
Blur	gaussian_blur	radius	0.5	1.5	2.5	pixels
	motion_blur	ksize	5	9	15	kernel size
	defocus_blur	radius	1.0	3.0	5.0	pixels
	zoom_blur	factor	0.02	0.06	0.10	zoom amount
	glass_blur	sigma	0.5	0.9	1.3	blur sigma
Noise	gaussian_noise	std	0.02	0.06	0.10	normalized
	shot_noise	scale	25	10	5	lower=more
	speckle_noise	std	0.05	0.15	0.25	normalized
	salt_pepper	amount	0.01	0.04	0.08	pixel fraction
Weather	fog	intensity	0.2	0.6	1.0	opacity
	frost	intensity	0.2	0.6	1.0	opacity
	snow	intensity	0.1	0.3	0.5	density
	rain	intensity	0.1	0.3	0.5	density
	spatter	intensity	0.1	0.3	0.5	coverage
Digital	jpeg_compression	quality	80	50	20	lower=worse
	pixelate	scale	0.9	0.5	0.2	lower=coarser
Geometric	rotate	degrees	5	15	30	rotation
	shear	degrees	5	15	25	shear angle
	affine	degrees	5	15	30	rotation+scale
	perspective_transform	magnitude	0.05	0.15	0.25	distortion
	elastic_transform	alpha	30	80	180	deformation
Color/Tone	brightness	factor	0.7	0.3	0.1	lower=darker
	brightness_up	factor	1.3	1.7	2.5	higher=brighter
	contrast	factor	0.7	0.3	0.1	lower=flatter
	contrast_up	factor	1.3	1.8	3.0	higher=sharper
	saturation	factor	0.5	0.1	0.0	lower=grayer
	saturation_up	factor	1.5	2.5	4.0	higher=vivid
	gamma	factor	0.7	0.4	0.2	lower=brighter
	gamma_up	factor	1.3	2.0	3.0	higher=darker
	hue_shift	degrees	10	40	90	color rotation
	color_jitter	range	0.1	0.3	0.5	random B/C/S
Occlusion	random_occlusion	ratio	0.05	0.15	0.25	area blocked
	grid_mask	ratio	0.1	0.2	0.3	grid density
	center_occlusion	ratio	0.1	0.3	0.5	center blocked
Resolution	downsample	scale	0.75	0.35	0.15	lower=smaller
	upsample	scale	1.5	3.0	6.0	interpolation
	sharpen	factor	1.5	3.0	6.0	edge enhance
	posterize	bits	6	4	2	lower=fewer
	solarize	threshold	200	128	64	lower=more
VLM-specific	text_overlay	fontsize	24	48	72	pixels
	watermark	fontsize	24	48	72	pixels
	add_border	width	10	30	60	pixels

F.3 Prompting Templates

We use two prompting modes with standardized templates:

Direct Mode. Designed for short, single-letter responses:

```
Please select the correct answer from the options above.
Respond with only the letter of the correct option.  Do not
explain.  Answer:
```

Chain-of-Thought (CoT) Mode. Designed for reasoning-based responses:

```
Answer the preceding multiple choice question.  The last
line of your response should be of the following format:
'Answer:  $LETTER' (without quotes) where LETTER is one of
options.  Think step by step before answering.
```

F.4 Generation Parameters

All models use deterministic decoding with `max_new_tokens=2048`. Thinking models (Qwen3-VL-Thinking) require sampling-based decoding and use: `max_new_tokens=8192`, `temperature=0.6`, `top_p=0.95`, `top_k=20`.

F.5 Augmentation Parameters

Table 30 lists the parameter values for each severity level across all severity-based augmentations. We evaluate at low, mid, and high severities. Binary augmentations have no severity variation.

Binary Augmentations. The following 7 augmentations have no severity levels: `flip_h`, `flip_v`, `grayscale`, `invert`, `channel_swap`, `equalize`, `autocontrast`.

F.6 Augmentation Application

Augmentations are applied deterministically based on sample index:

1. For each sample index i, compute per-sample seed: $s_i = (1234 \times 1000003 + i) \mod 2^{32}$
2. Initialize augmentation RNG with s_i
3. Apply augmentation to all images in the sample

This ensures: (1) identical augmentation across model runs for fair comparison, (2) different random variations per sample for stochastic augmentations (noise, blur, occlusion positions).

F.7 Evaluation Protocol

Correctness. A response is correct if the extracted letter matches the ground truth answer field.

Metrics. All metrics (accuracy, flip rates, RCE, mCE) are computed on the same 20% stratified subset across all models and augmentations, enabling direct comparison.

G Augmentation Visualization

Figures 3–14 visualize all 49 augmentations applied to a representative MMBench image at low, mid, and high severity levels. Binary augmentations have no severity variation.

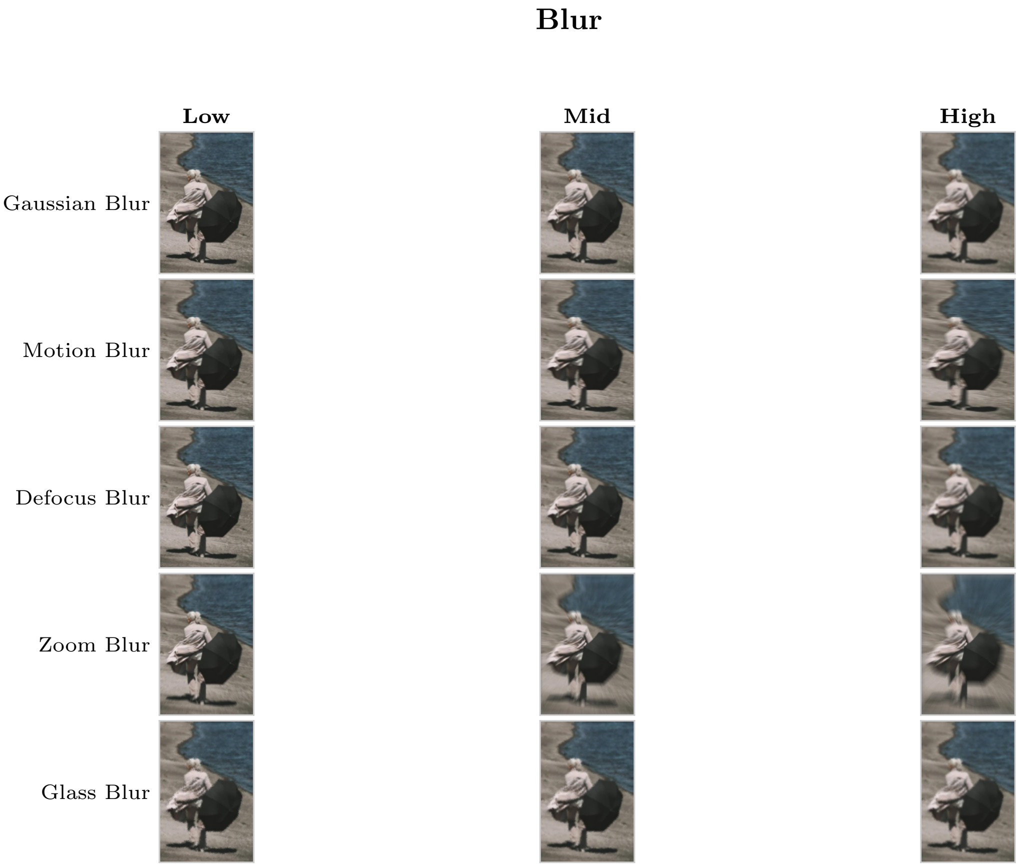

Figure 3: Augmentation Visualization: Blur augmentations at low, mid, and high severity.

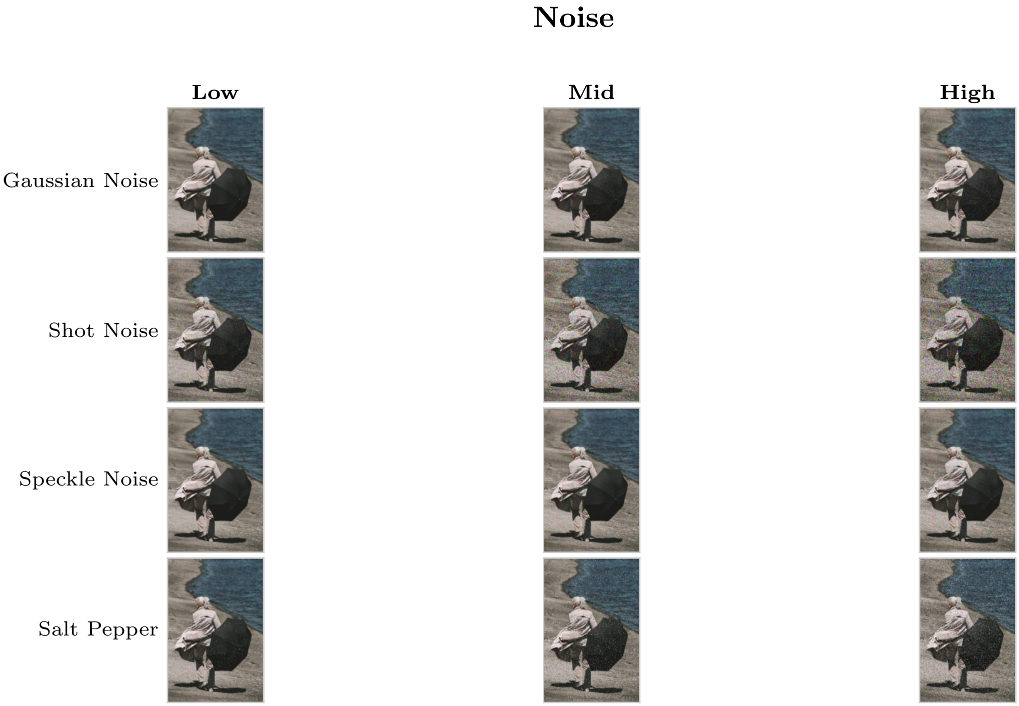

Figure 4: Augmentation Visualization: Noise augmentations at low, mid, and high severity.

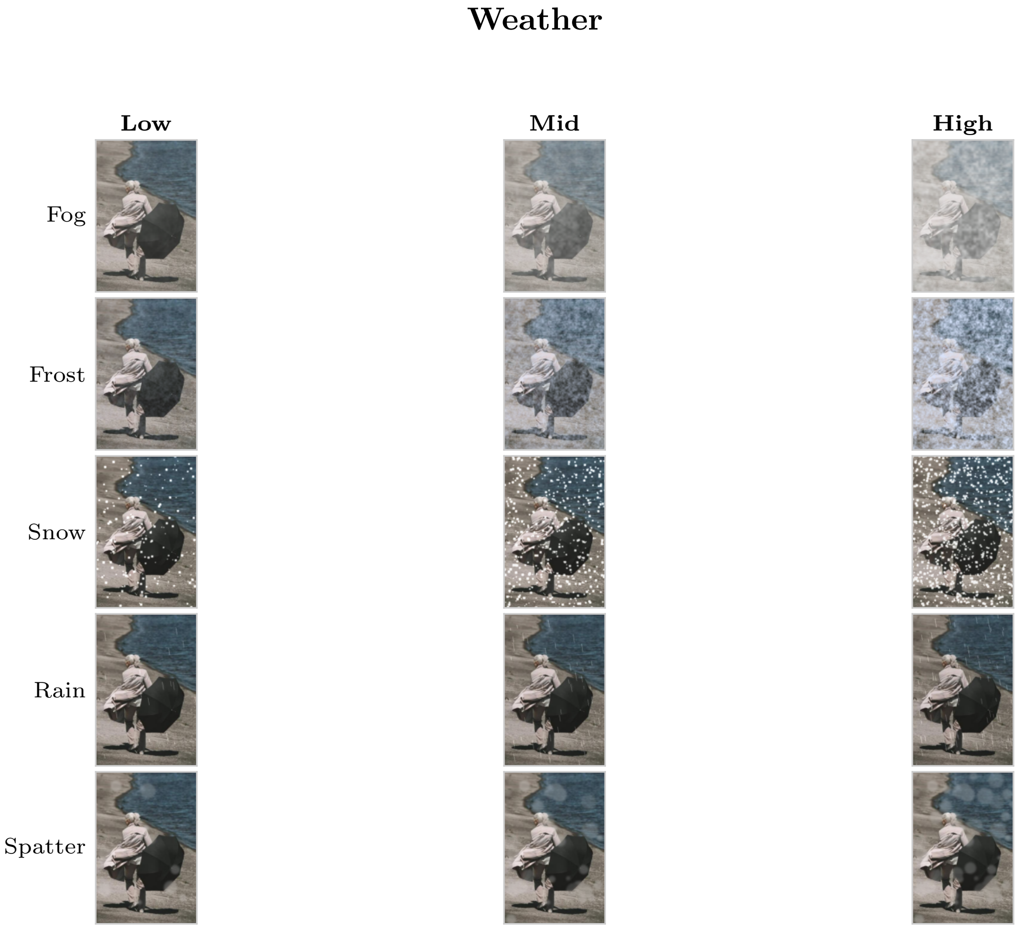

Figure 5: Augmentation Visualization: Weather augmentations at low, mid, and high severity.

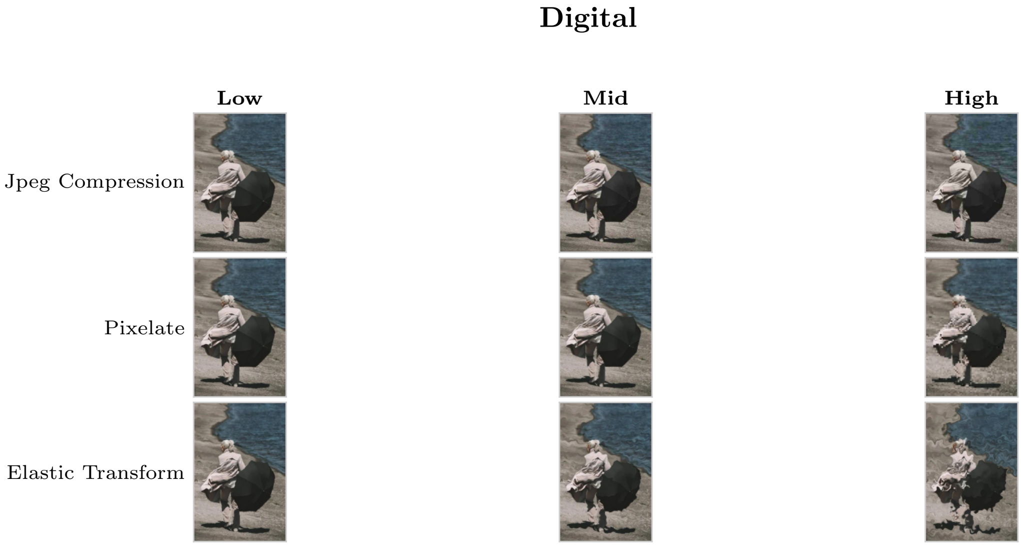

Figure 6: Augmentation Visualization: Digital augmentations at low, mid, and high severity.

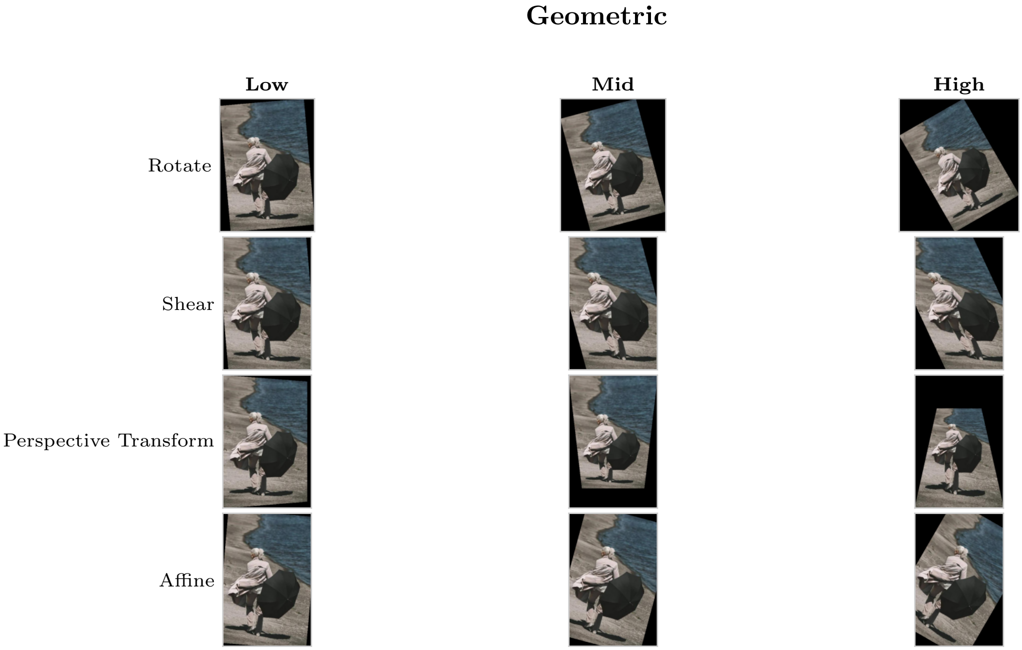

Figure 7: Augmentation Visualization: Geometric augmentations at low, mid, and high severity.

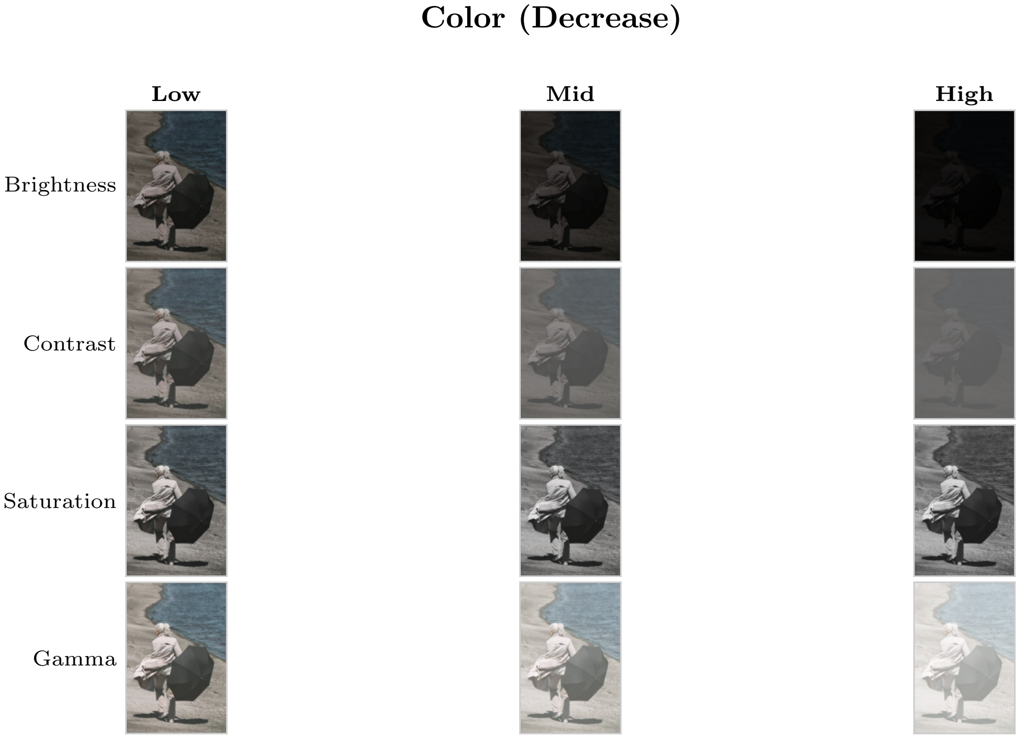

Figure 8: Augmentation Visualization: Color/Tone decrease augmentations at low, mid, and high severity.

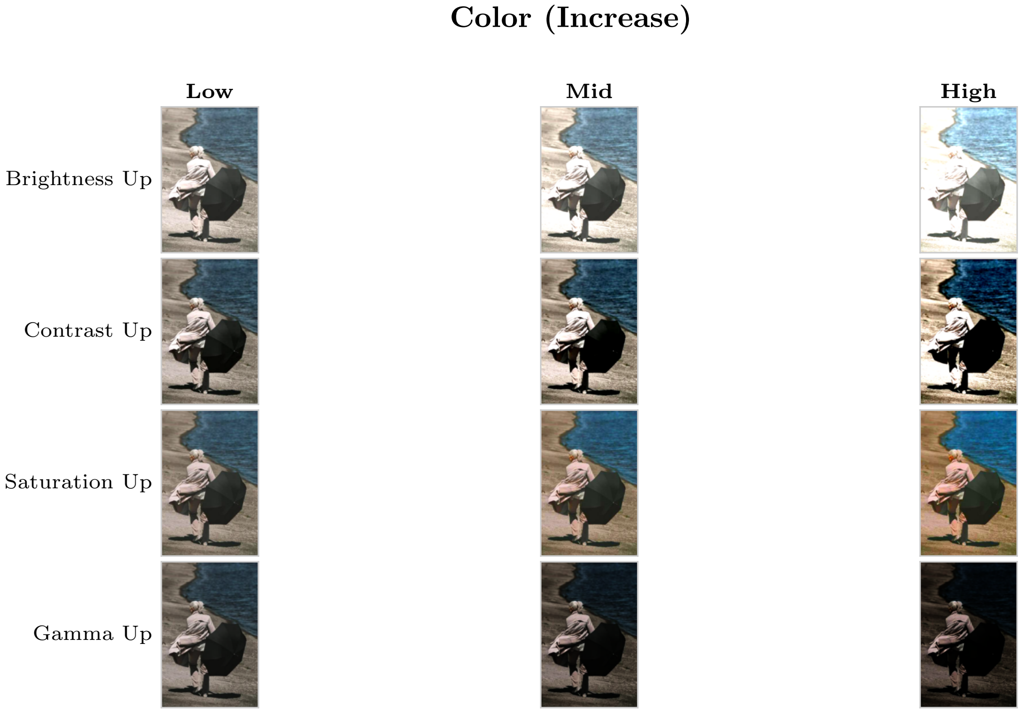

Figure 9: Augmentation Visualization: Color/Tone increase augmentations at low, mid, and high severity.

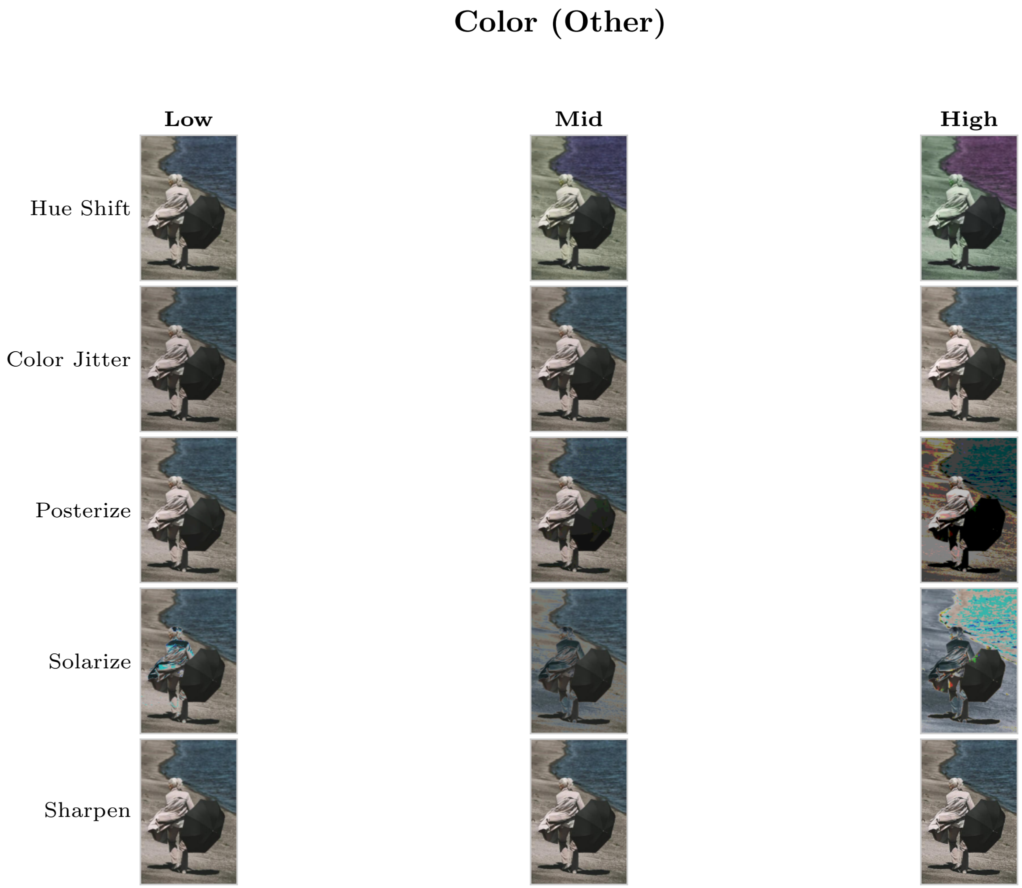

Figure 10: Augmentation Visualization: Other Color/Tone augmentations at low, mid, and high severity.

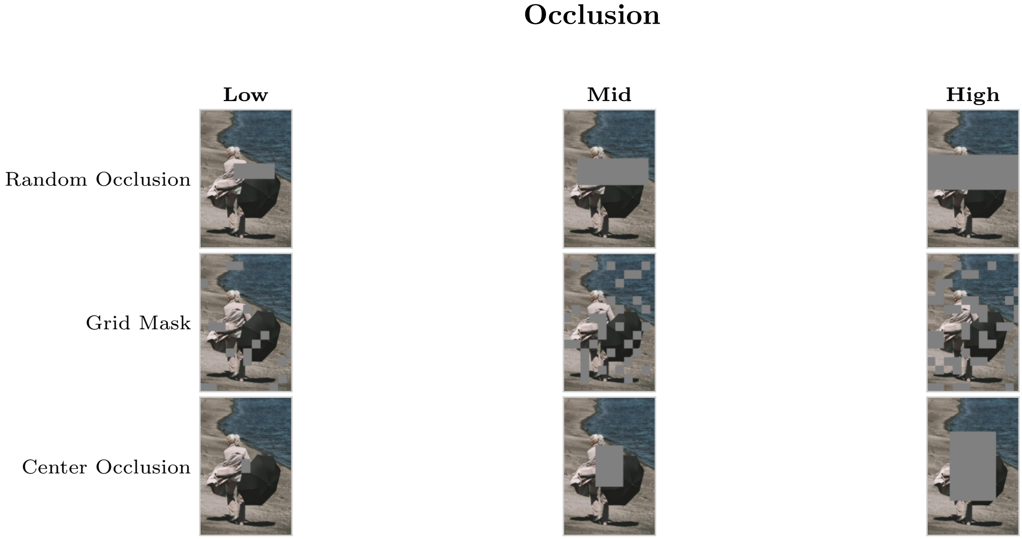

Figure 11: Augmentation Visualization: Occlusion augmentations at low, mid, and high severity.

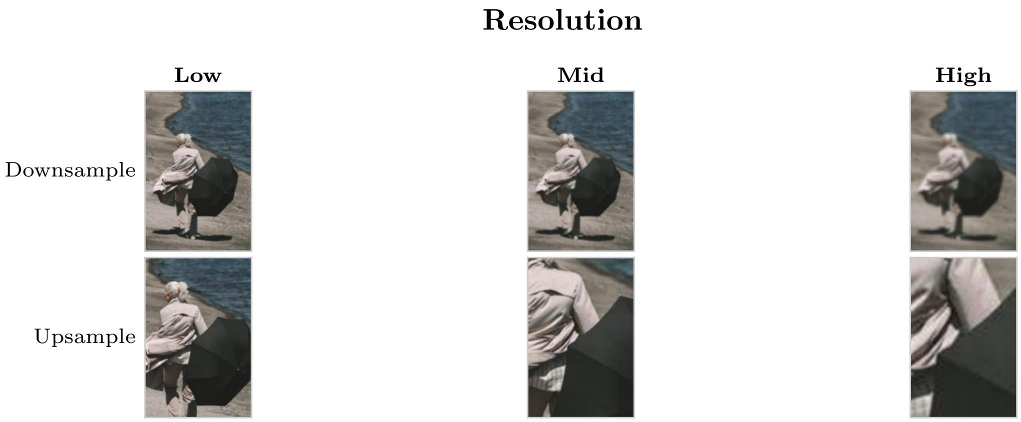

Figure 12: Augmentation Visualization: Resolution augmentations at low, mid, and high severity.

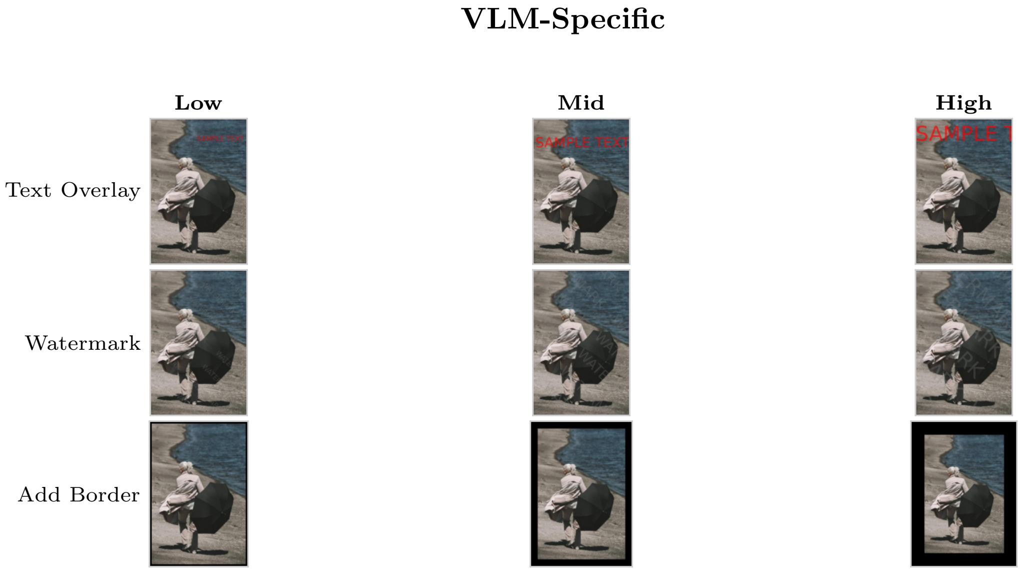

Figure 13: Augmentation Visualization: VLM-specific augmentations at low, mid, and high severity.

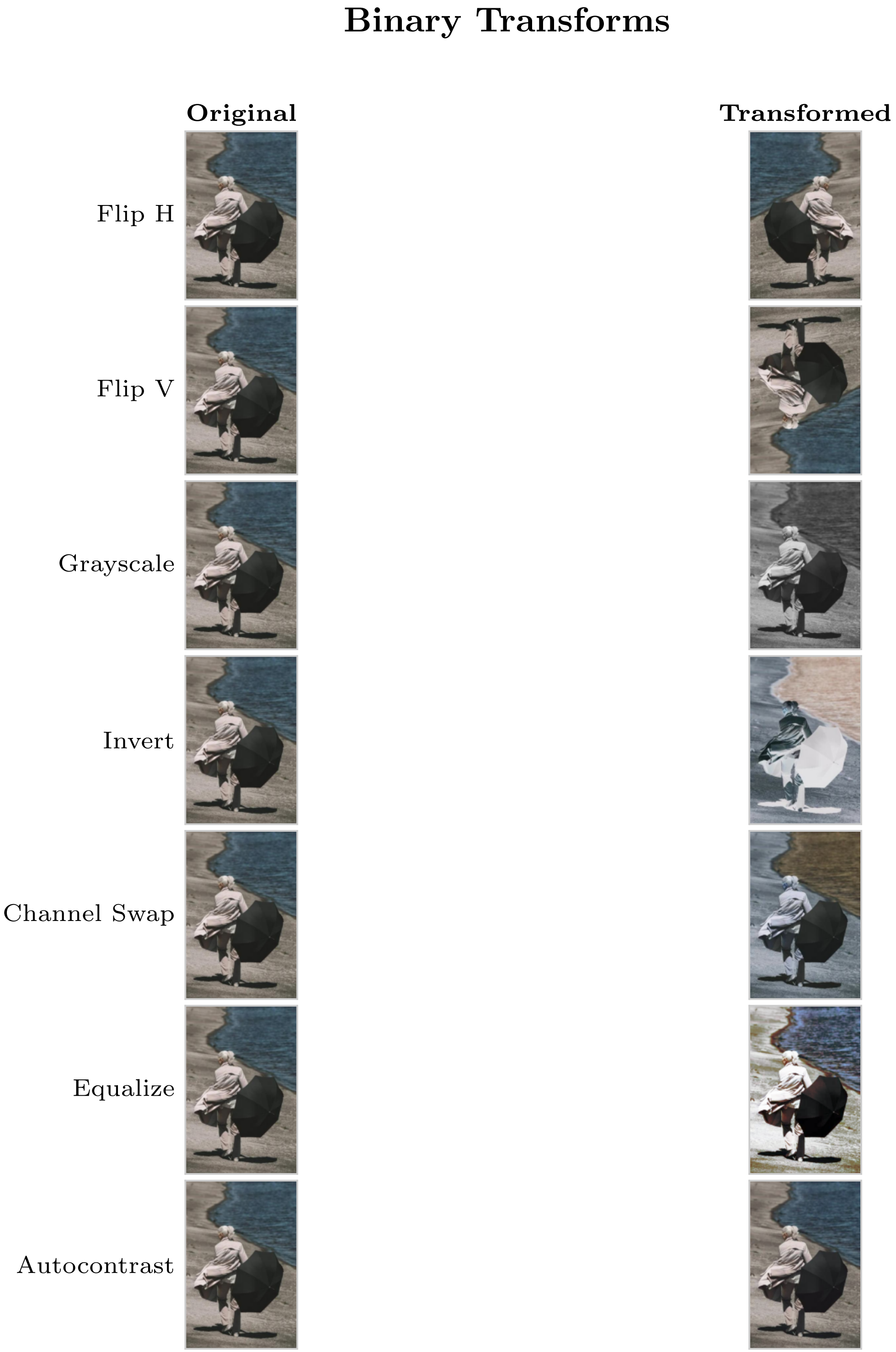

Figure 14: Augmentation Visualization: Binary transforms (no severity variation).