Does Visual Degradation Amplify Instruction Sensitivity in Vision-Language-Action Models? An Empirical Study with OpenVLA

Jihwan Woo
Amazon Web Services
jihwanw@amazon.com

***Abstract*—Vision-Language-Action (VLA) models deployed in harsh environments face compounded challenges: degraded visual inputs and natural variation in operator instructions. This paper investigates whether visual degradation amplifies instruction sensitivity, using OpenVLA-7B as a case study. Three manipulation tasks are evaluated under four degradation conditions (low lighting, Gaussian noise, fog, partial occlusion) at three intensity levels. The effect is degradation-type-dependent. Under occlusion, the model's action outputs become highly stochastic (noise floor $= 0.250$), and instruction content has diminishing influence—a pattern consistent with model collapse rather than amplified sensitivity. Low lighting suppresses sensitivity ($F = 15.11$, $p < 0.001$), while noise and fog show no significant effect. The signal-to-noise ratio decreases from $3.0\times$ (clean) to $0.7\times$ (severe), crossing the SNR $= 1$ threshold. We propose this as a deployment criterion: if the ratio of instruction-induced variation to stochastic noise falls below 1, the model should not be trusted to follow instructions in that environment. Instruction canonicalisation reduces sensitivity by 60–90% for two of three tasks under clean conditions. These findings, while specific to OpenVLA-7B, provide a methodology for evaluating compounded vulnerabilities and indicate that harsh-environment deployment requires degradation-type-specific evaluation.**

I. Introduction

Vision-Language-Action (VLA) models generate robot actions from visual observations and natural language instructions [1]–[3]. As these models are considered for deployment in harsh environments—underwater manipulation, space operations, disaster response—two robustness challenges converge: (1) degraded visual inputs from poor lighting, sensor noise, or atmospheric interference, and (2) natural variation in how human operators phrase instructions.

Prior work has studied these challenges in isolation. Visual robustness has been evaluated through perturbation benchmarks [6], [7], and instruction sensitivity has been characterised across tasks [9], [15]. However, the *interaction* between visual degradation and instruction sensitivity—whether degraded perception amplifies linguistic sensitivity—has not been investigated.

This distinction matters for harsh-environment deployment. If visual degradation amplifies instruction sensitivity, then a VLA model that appears instruction-robust in clean laboratory conditions may become unpredictably sensitive when deployed underwater or in dusty planetary environments. Single-factor robustness evaluation would miss this compounded vulnerability.

This paper presents three contributions: (1) a systematic evaluation of VLA instruction sensitivity under four visual degradation conditions at three intensity levels; (2) evidence that the effect is degradation-type-dependent, with occlusion leading to model collapse rather than amplified sensitivity; (3) practical implications for harsh-environment deployment, including degradation-type-specific mitigation strategies.

II. Related Work

Vision-Language-Action models. VLA models unify perception, language understanding, and action generation in a single end-to-end architecture. RT-2 [1] showed that web-scale vision-language pretraining transfers to robotic control, achieving zero-shot generalisation to unseen objects. OpenVLA [2] open-sourced a 7B-parameter VLA trained on Open X-Embodiment [4], enabling reproducible research on instruction-conditioned manipulation. π_0 [3] introduced flow matching for continuous action generation. Octo [10] proposed a transformer-based generalist policy trainable on diverse robot datasets with a diffusion action head. A common feature of these models is that they accept *free-form* natural language instructions—the same task can be described in many ways. This flexibility is useful for human-robot interaction but introduces a risk: the model's output may change depending on *how* the instruction is worded, even when the intended task is identical. How this interacts with perceptual conditions is not known.

VLA robustness evaluation. Recent work has begun to characterise VLA failure modes, but almost exclusively along single axes. Liu et al. [5] benchmarked robustness to *semantic* perturbations (negation, distraction, spatial ambiguity) in LIBERO, finding that negation instructions cause up to 60% performance drops. Chen et al. [6] proposed VLATest for systematic *visual and linguistic* perturbation testing, but evaluated each modality independently—visual noise was tested with fixed instructions, and instruction variation was tested with clean images. Wen et al. [7] provided a multi-architecture benchmark but similarly did not cross visual and linguistic factors. Woo [15] showed that instruction phrasing sensitivity is task-dependent and introduced a noise floor methodology to separate instruction-induced variation from stochastic variation, but again under clean visual conditions only. This single-axis paradigm implicitly assumes that visual and linguistic

robustness are independent—an assumption that has not been tested. No prior work has evaluated the interaction between visual degradation and instruction sensitivity. Each modality has been tested in isolation, leaving open the question of whether their effects compound, cancel, or interact in more complex ways.

Perception in harsh environments. Underwater manipulation, space robotics, and disaster response operate under visual conditions far removed from the clean laboratory settings in which VLA models are trained and evaluated. Low lighting reduces contrast and colour information; particulate matter (dust, turbidity) introduces noise; fog and haze reduce visibility range; and physical obstacles cause partial occlusion [8]. These degradation types differ qualitatively: low lighting is a global intensity reduction, noise is a pixel-level perturbation, fog is a depth-dependent attenuation, and occlusion is a spatial information loss. If VLA instruction sensitivity depends on degradation *type*—not just severity—then deployment decisions require degradation-specific evaluation, not a single robustness score. This is particularly relevant for multi-phase missions where degradation conditions change (e.g., shallow-to-deep underwater transitions). This possibility has not been tested.

Instruction sensitivity in HRI. Sensitivity to instruction phrasing is a long-standing challenge in human-robot interaction. Tellex et al. [13] showed that grounding natural language to robot actions is sensitive to syntactic structure, and instruction granularity has been shown to affect execution in LLM-based planning systems [14]. Subsequent work on LLM-based robot systems, including Inner Monologue [11] and Code as Policies [12], showed that instruction format and granularity affect task execution quality. VLA models were expected to absorb this variation through large-scale pretraining, but recent evidence suggests they do not [15].

The compounding problem. The three lines of work above converge on an untested hypothesis: that visual degradation and instruction variation *interact*. A VLA model that appears instruction-robust in clean conditions may become unpredictably sensitive when deployed in a degraded environment. Conversely, a model that appears visually robust with fixed instructions may fail when operators use varied phrasing. Single-factor evaluation—the current standard—would miss both failure modes. This paper directly tests this interaction hypothesis across four degradation types, three intensity levels, and two instruction variation types, providing the first systematic characterisation of the compounded vulnerability.

III. Method

A. Model and Environment

OpenVLA-7B [2], trained on Open X-Embodiment [4], outputs 7-DoF actions $\mathbf{a} = [\Delta\mathbf{p}, \Delta\mathbf{r}, a_g] \in \mathbb{R}^7$ given image I and instruction l. Experiments use ManiSkill3 with SIMPLER Bridge scenes (WidowX arm) across three tasks: spoon-on-towel, carrot-on-plate, eggplant-in-basket. Table I summarises the experimental setup.

TABLE I
Experimental setup.

Component	Specification
VLA model	OpenVLA-7B (7B params, ViT + Llama 2)
Action space	7-DoF: $[\Delta x, \Delta y, \Delta z, \Delta r_x, \Delta r_y, \Delta r_z, g]$
Simulator	ManiSkill3 + SIMPLER Bridge
Robot	WidowX 250 (6-DoF arm + gripper)
Tasks	3 (spoon-on-towel, carrot-on-plate, eggplant-in-basket)
Instructions	5 phrasing + 5 specificity per task
Degradation	4 types $\times$ 3 levels + clean = 13
Samples	$n = 50$ pairs per cell
Total cells	3 tasks $\times$ 13 cond. $\times$ 3 types = 117

TABLE II
Experimental procedure for evaluating degradation $\times$ instruction sensitivity interaction.

Input: VLA model f_θ, tasks $\mathcal{K}$, instructions $\mathcal{L}^P, \mathcal{L}^S$, degradation functions $\{g_d\}$, levels $\{1, 2, 3\}$
Output: Sensitivity matrix $\bar{\Delta}^v_{k,d}$, ANOVA results

Render initial image I from simulator for task k
for each degradation $d \in \{\text{clean}\} \cup \{g_d \times \text{level}\}$
  Apply degradation: $I_d = g_d(I, \text{level})$
  for each variation type $v \in \{P, S\}$
    for each instruction pair $(l_i, l_j) \in \binom{\mathcal{L}^v}{2}$
      $\mathbf{a}_i = f_\theta(I_d, l_i)$; $\mathbf{a}_j = f_\theta(I_d, l_j)$ (same I_d)
      Record $\Delta_d(l_i, l_j) = \|\mathbf{a}_i - \mathbf{a}_j\|_2$
  Compute noise floor: $\Delta_d^{\text{NF}} = \|f_\theta(I_d, l) - f_\theta(I_d, l)\|_2$ (same instruction, different forward passes)
Compute $\bar{\Delta}^v_{k,d}$, $\text{AR}_{k,d}$, SNR_d
Run two-way ANOVA: level $\times$ variation type

B. Instruction Variation

Following [15], two balanced sets of 5 instruction variants per task:

- **Phrasing** ($\mathcal{L}^P$): 5 synonymous verb substitutions
- **Specificity** ($\mathcal{L}^S$): original, detailed, vague, imperative, conditional

C. Visual Degradation Conditions

Four degradation types are applied to the input image I before feeding to the model, each at three intensity levels (mild, moderate, severe):

1) **Low lighting**: Brightness scaling $I' = \alpha I$, where $\alpha \in \{0.7, 0.5, 0.3\}$
2) **Gaussian noise**: $I' = I + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma \in \{0.05, 0.10, 0.20\}$
3) **Fog/haze**: $I' = I \cdot t + A(1 - t)$, transmission $t \in \{0.7, 0.5, 0.3\}$, atmospheric light $A = 1$
4) **Partial occlusion**: Random rectangular patches covering $\{10\%, 20\%, 30\%\}$ of image area

D. Metrics

The experimental procedure is summarised in Table II.
Pairwise action sensitivity under degradation condition d:

$$\Delta_d(l, l') = \|f_\theta(I_d, l) - f_\theta(I_d, l')\|_2 \tag{1}$$

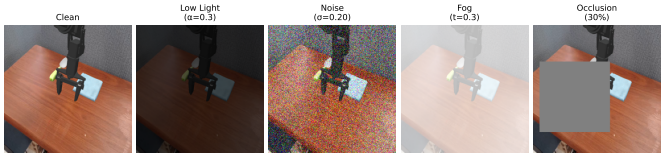

Fig. 1. Example input images under severe degradation. From left: clean, low lighting ($\alpha = 0.3$), Gaussian noise ($\sigma = 0.20$), fog ($t = 0.3$), and partial occlusion (30% area). Task: spoon-on-towel (WidowX arm, SIMPLER Bridge scene).

where I_d is the degraded image. The mean sensitivity per condition:

$$\bar{\Delta}^{v}_{k,d} = \frac{1}{|\mathcal{P}^v| \cdot |\mathcal{I}|} \sum_{(l,l') \in \binom{\mathcal{L}^v}{2}} \sum_{I \in \mathcal{I}} \Delta_d(l, l') \tag{2}$$

yielding $n = 50$ measurements per cell. The amplification ratio quantifies how much degradation increases sensitivity:

$$\text{AR}_{k,d} = \frac{\bar{\Delta}^{v}_{k,d}}{\bar{\Delta}^{v}_{k,\text{clean}}} \tag{3}$$

Note that the L_2 distance aggregates position deltas (metres) and rotation deltas (radians) into a single scalar; the two components have different physical units. We report the aggregate metric for consistency with prior work [15], but acknowledge that per-component analysis (position vs rotation vs gripper) would provide finer-grained insight.

A two-way ANOVA tests the interaction between degradation level and variation type:

$$\Delta = \mu + \alpha_{\text{deg}} + \beta_{\text{var}} + (\alpha\beta)_{\text{deg}\times\text{var}} + \epsilon \tag{4}$$

IV. Results

Fig. 1 shows example images under each degradation condition at severe intensity. The visual impact ranges from subtle (low lighting) to severe information loss (occlusion).

A. Main Results

Table III presents the mean pairwise action distance aggregated across three tasks. The effect of visual degradation on instruction sensitivity is degradation-type-dependent. Occlusion consistently increases action distance at all intensity levels, with phrasing sensitivity reaching 0.126 at severe level (2.8× the clean baseline). Low lighting, counterintuitively, *decreases* sensitivity at severe levels, possibly because reduced brightness compresses the model's output distribution. Noise and fog produce inconsistent effects across levels.

B. Per-Degradation-Type ANOVA

Table IV reports two-way ANOVA results (degradation level × variation type) computed separately for each degradation type. The omnibus ANOVA across all conditions is not significant ($p = 0.16$), but disaggregation reveals that the effect is concentrated in specific degradation types.

Occlusion shows a significant main effect of degradation level ($F = 4.69$, $p = 0.003$), confirming that partial occlusion amplifies instruction sensitivity. Low lighting also shows a significant level effect ($F = 15.11$, $p < 0.001$) with a significant interaction ($p = 0.02$), though the direction is reversed: sensitivity *decreases* with darker images. Noise and fog show no significant effects.

TABLE III
MEAN PAIRWISE ACTION DISTANCE (L_2) AGGREGATED ACROSS THREE TASKS. P = PHRASING, S = SPECIFICITY. AR = AMPLIFICATION RATIO VS CLEAN.

Degradation	Level	$\bar{\Delta}^P$	$\bar{\Delta}^S$	NF	AR^P
Clean	—	.046	.072	.019	1.0×
Low light	Mild	.057	.059	.008	1.2×
	Mod.	.037	.038	.018	0.8×
	Severe	.031	.023	.035	0.7×
Noise	Mild	.063	.072	.049	1.4×
	Mod.	.063	.063	.065	1.4×
	Severe	.062	.058	.031	1.3×
Fog	Mild	.047	.057	.004	1.0×
	Mod.	.052	.052	.011	1.1×
	Severe	.055	.059	.039	1.2×
Occlusion	Mild	.104	.119	.180	2.3×
	Mod.	.111	.087	.016	2.4×
	Severe	.127	.111	.250	2.8×

NF = noise floor (stochastic variation from identical instruction). $n = 150$ per cell (50 per task × 3 tasks).

TABLE IV
TWO-WAY ANOVA (DEGRADATION LEVEL × VARIATION TYPE) BY DEGRADATION TYPE. EACH ANALYSIS INCLUDES CLEAN AS LEVEL 0.

Degradation	Level effect		Interaction	
type	F	p	F	p
Low light	15.11	<0.001***	3.29	0.020*
Noise	0.78	0.507	2.16	0.091
Fog	0.71	0.547	1.79	0.147
Occlusion	4.69	0.003**	0.94	0.420

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. $\text{df}_{\text{level}} = 3$, $\text{df}_{\text{interaction}} = 3$.

C. Per-Task Analysis

Table V reports per-task ANOVA results pooling all degradation types. All three tasks show significant degradation level effects, but the interaction between degradation and variation type is significant only for eggplant-in-basket ($F = 3.21$, $p = 0.02$). This task involves the most complex geometry (inserting into a container), suggesting that task difficulty modulates the degradation–sensitivity coupling.

D. Signal-to-Noise Ratio

Table VI shows the signal-to-noise ratio (instruction-induced sensitivity divided by stochastic noise floor) at each degradation severity. Under clean conditions, instruction variation produces 3.0× more action variation than stochastic noise. At severe degradation, this ratio drops to 0.7×—instruction sensitivity becomes indistinguishable from random noise. This

TABLE V
TWO-WAY ANOVA BY TASK (ALL DEGRADATION TYPES POOLED).

Task	Level effect		Interaction	
	F	p	F	p
Spoon-on-towel	2.68	0.046*	0.24	0.870
Carrot-on-plate	9.30	<0.001***	2.51	0.057
Eggplant-in-basket	4.72	0.003**	3.21	0.022*

TABLE VI
SIGNAL-TO-NOISE RATIO BY DEGRADATION SEVERITY (AGGREGATED ACROSS ALL DEGRADATION TYPES AND TASKS).

	Clean	Mild	Moderate	Severe
Signal (mean $\bar{\Delta}$)	.059	.072	.063	.066
Noise floor	.019	.060	.028	.089
SNR	3.0×	1.2×	2.3×	0.7×

collapse is driven primarily by the noise floor increasing under occlusion (from 0.019 to 0.250).

E. Per-Task Amplification Under Severe Degradation

Table VII presents the amplification ratio for each task–degradation combination at severe intensity, with Mann-Whitney U test p-values. Three patterns emerge. First, occlusion amplifies phrasing sensitivity across all tasks (AR $= 1.8$–$5.9\times$), with the strongest effect on eggplant-in-basket ($5.9\times$, $p = 0.025$). Second, the amplification is asymmetric: phrasing sensitivity increases under occlusion, but specificity sensitivity does not consistently increase. Third, low lighting suppresses sensitivity in all tasks (AR < 1).

F. Baseline Sensitivity Predicts Amplification

Tasks with lower baseline sensitivity exhibit higher amplification under degradation. The Spearman correlation between clean-condition phrasing sensitivity and amplification ratio is $r = -0.55$ ($p < 0.001$). Eggplant-in-basket, which has the lowest clean sensitivity ($\bar{\Delta}^P = 0.021$), shows the highest amplification under occlusion ($5.9\times$). A possible explanation is that tightly clustered outputs reflect the model's high confidence in a narrow action region; when degradation disrupts the visual features anchoring this confidence, the outputs scatter more widely than those of tasks where the model already exhibits broader action distributions.

G. Visualisation

Fig. 2 shows the amplification ratio for each task–degradation combination at severe intensity. Occlusion dominates across all tasks, with eggplant-in-basket reaching $5.9\times$. Fig. 3 illustrates the SNR collapse from $3.0\times$ (clean) to $0.7\times$ (severe), crossing the SNR $= 1$ threshold below which instruction sensitivity is indistinguishable from stochastic noise.

V. DISCUSSION

Effect sizes and confidence intervals. Table VIII reports Cohen's d and bootstrap 95% confidence intervals for

TABLE VII
AMPLIFICATION RATIO (AR) UNDER SEVERE DEGRADATION BY TASK. BOLD = SIGNIFICANT ($p < 0.05$, MANN-WHITNEY U).

Task	Degrad.	AR^P	p^P	AR^S	p^S
Spoon	Low lt.	0.5	.041	0.4	.020
	Noise	1.0	.383	1.2	.102
	Fog	0.5	.931	0.6	.515
	Occl.	**1.8**	**.020**	2.6	.199
Carrot	Low lt.	0.7	.084	0.2	.062
	Noise	**1.8**	**<.001**	0.6	.672
	Fog	**1.4**	**<.001**	0.6	.245
	Occl.	**2.9**	**<.001**	1.3	.287
Eggpl.	Low lt.	1.3	.035	0.5	.002
	Noise	**1.6**	**.011**	0.7	.467
	Fog	**3.0**	**<.001**	**1.6**	**<.001**
	Occl.	**5.9**	**.025**	0.7	.565

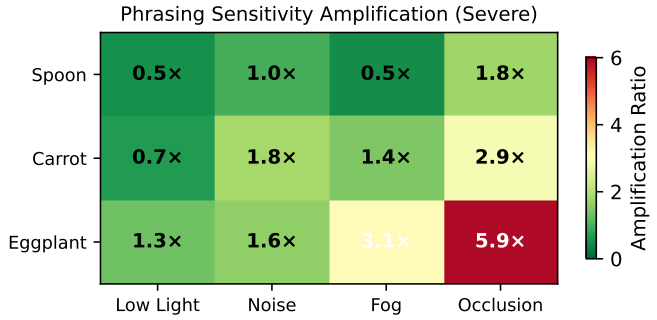

Fig. 2. Phrasing sensitivity amplification ratio under severe degradation. Occlusion produces the strongest amplification across all tasks, with eggplant-in-basket reaching $5.9\times$.

key comparisons. Occlusion produces a medium effect size ($d = 0.41$) on phrasing sensitivity, with the eggplant task showing the largest individual effect ($d = 0.53$, AR 95% CI $[2.65, 10.73]$). The SNR collapse from clean to severe is robust: the 95% CI for severe SNR is $[0.64, 0.86]$, entirely below 1.0, confirming that instruction sensitivity is indistinguishable from noise under severe degradation.

Degradation-type-dependent effects. Visual degradation does not uniformly amplify instruction sensitivity. The effects are qualitatively different across degradation types. Low lighting suppresses sensitivity, likely by compressing the output distribution as visual contrast decreases. Noise and fog produce no significant effect, suggesting that the model's visual encoder is partially robust to pixel-level and depth-dependent perturbations.

Occlusion: amplification or collapse? Occlusion requires careful interpretation. While the amplification ratio reaches $5.9\times$ for eggplant-in-basket, the noise floor under severe occlusion (0.250) exceeds the phrasing sensitivity (0.127). This means the model's outputs are more variable across repeated identical inputs than across different instructions. Rather than "amplified sensitivity to instructions," this pattern is more consistent with *model collapse*: the model produces near-

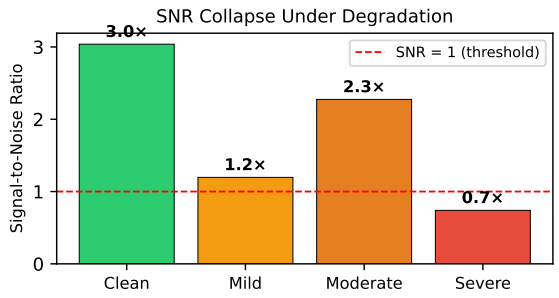

Fig. 3. Signal-to-noise ratio by degradation severity. The dashed line marks SNR $= 1$, below which instruction-induced variation is indistinguishable from stochastic noise.

TABLE VIII
EFFECT SIZES (COHEN'S d) AND BOOTSTRAP 95% CI FOR SEVERE DEGRADATION VS CLEAN (PHRASING SENSITIVITY, AGGREGATED ACROSS TASKS).

Degradation	Cohen's d	Size	95% CI (AR)
Low light	-0.23	Small	[0.42, 0.98]
Noise	0.23	Small	[0.87, 1.92]
Fog	0.13	Negligible	[0.79, 1.63]
Occlusion	0.41	Medium	[1.38, 4.72]
SNR: Clean $= 3.1\times$ [2.39, 3.92]; Severe $= 0.7\times$ [0.64, 0.86]			

stochastic outputs because 30% of the visual input is missing, and instruction content has minimal influence on the output. The practical implication is the same—the model is unreliable under severe occlusion—but the mechanism differs from true instruction sensitivity amplification. We recommend that future work distinguish between these two failure modes by comparing the noise floor against the instruction sensitivity signal. Additionally, occlusion location matters: a patch covering the target object likely has a different effect than one covering the robot arm or background. Spatially structured degradation is a natural extension of this work. Real underwater environments also exhibit colour distortion (blue/green shift) that is absent from the current degradation set; this domain-specific perturbation may interact with instruction sensitivity in ways not captured by the four generic degradation types tested here.

SNR collapse. The decrease in SNR from $3.0\times$ to $0.7\times$ has a practical interpretation: under severe degradation, the model's response to different instructions is no larger than its response to running the same instruction twice. At this point, instruction variation is effectively meaningless—the model's behaviour is dominated by stochastic noise rather than linguistic content. This threshold could serve as a deployment criterion: if SNR < 1, the model should not be trusted to follow instructions reliably. In practice, this can be operationalised as a pre-deployment calibration protocol: run n identical instructions and n varied instructions in the target environment, compute SNR, and issue a go/no-go decision. With $n = 10$ per condition, the test takes under five minutes and provides a quantitative basis for deployment decisions in new environments.

Low-light suppression. The finding that low lighting *suppresses* instruction sensitivity ($F = 15.11$, $p < 0.001$) is arguably the cleanest result in this study, as it is not contaminated by noise floor inflation. Two mechanisms may explain this. First, the visual encoder may map dark images to a narrow region of feature space, compressing the representation and reducing the influence of all other inputs, including language. Second, severely darkened images may fall outside the training distribution, causing the model to default to a fixed action regardless of instruction content. In either case, the practical implication is that low-light environments may mask instruction sensitivity problems: a model that appears robust to instruction variation under dim lighting may become sensitive again when lighting improves. This has direct relevance for underwater and nighttime operations where lighting conditions fluctuate. For example, an underwater robot descending from surface to depth experiences continuous light reduction; a VLA tested at depth may appear instruction-robust, but become sensitive again when the robot ascends or when artificial lighting is activated.

Mitigation strategies. We consider two approaches: (1) instruction canonicalisation, mapping diverse operator instructions to a fixed template before feeding to the VLA model, which is most effective when SNR > 1; (2) degradation-aware fine-tuning, augmenting training data with degraded images, particularly occlusion, to build robustness at the source.

To validate the first strategy, we conducted a canonicalisation experiment: instead of feeding diverse instructions, all instructions were mapped to a single canonical form (e.g., "put the spoon on the towel") before querying the model. Table IX reports the results. Under clean conditions, canonicalisation reduces phrasing sensitivity by 60–90% for spoon and carrot tasks. Under occlusion, the reduction is more variable (16–99%), but the canonical action distance is consistently lower than the diverse-instruction distance. In practice, canonicalisation can be implemented as a lightweight preprocessing module in the VLA pipeline: *operator instruction* $\rightarrow$ *LLM normaliser* $\rightarrow$ *canonical instruction* $\rightarrow$ *VLA model* $\rightarrow$ *action*. A more sophisticated variant could learn the instruction that minimises action variance from a calibration phase, rather than mapping to a fixed template. We also tested LLM-based canonicalisation (Claude Haiku), where each instruction is normalised before being passed to the VLA. For spoon, LLM canonicalisation achieves comparable reduction (75%) to the fixed template (83%). However, for eggplant, the LLM *increases* sensitivity (-523%) because it rewrites single-step instructions as multi-step commands (e.g., "pick the eggplant, put it inside the basket"), introducing new variation. This highlights that normalisation quality matters: a poorly designed normaliser can be worse than no normalisation. This echoes findings in LLM-based planning [11], where over-specified instructions degrade execution. In this pilot experiment ($n = 5$ per condition), the simple fixed-template approach is more reliable than LLM-based normalisation and is effective for spoon and carrot under clean conditions (83–90% reduction).

TABLE IX
CANONICALISATION EXPERIMENT: PHRASING SENSITIVITY ($\bar{\Delta}^P$). FIXED = SINGLE CANONICAL TEMPLATE; LLM = CLAUDE HAIKU NORMALISATION. RED.^{F} = FIXED-TEMPLATE REDUCTION.

Task	Condition	Diverse	Fixed	LLM	Red.^{F}
Spoon	Clean	.033	.006	.032	83%
	Occlusion	.009	.008	—	16%
Carrot	Clean	.117	.012	.015	90%
	Occlusion	.024	.012	—	51%
Eggplant	Clean	.089	.083	.081	6%
	Occlusion	.134	.002	—	99%

The eggplant-clean anomaly (6% reduction) likely reflects that the canonical instruction "put the eggplant in the basket" already produces high action variance due to the geometric complexity of inserting into a container—canonicalisation cannot reduce sensitivity that originates from task difficulty rather than instruction wording. Under occlusion, the eggplant result inverts (99% reduction), consistent with the model collapsing to a near-deterministic default action when visual information is severely degraded.

Limitations. This study evaluates a single model (OpenVLA-7B); the findings may not generalise to architecturally different VLAs such as RT-2, π_0, or Octo. All experiments use simulation with synthetic degradation—real harsh-environment imagery exhibits spatially non-uniform degradation (e.g., depth-dependent fog, object-specific occlusion) that uniform synthetic perturbations do not capture. The analysis is single-step: action vectors are compared at one time step, and trajectory-level effects are not evaluated. OpenVLA operates closed-loop across steps (new image each step), so initial action perturbations may be corrected or amplified over a trajectory; which outcome occurs is an open question. We attempted full-episode rollouts to measure task success rate, but OpenVLA-7B achieved 0% success across all conditions including clean—consistent with the known sim-to-real gap for this model in ManiSkill3/SIMPLER [8]. Consequently, success rate comparison across degradation conditions was not feasible, and the analysis relies on action distance as a proxy for behavioural change. Action distance remains informative because it captures the model's *intended* behaviour change in response to instruction and degradation variation, independent of whether the downstream controller successfully executes the action. Episodes were capped at 40 steps; Bridge manipulation tasks may require 50–100 steps for completion, and the cap may have contributed to the 0% success rate. The canonicalisation experiment uses 5 samples per condition, limiting statistical power.

VI. CONCLUSION

This paper makes two contributions. First, it identifies two distinct failure modes under compounded visual-linguistic perturbation: *sensitivity amplification* (the model responds more to instruction differences) and *model collapse* (the model ignores instructions and produces stochastic output). Second, it presents a methodology for evaluating the interaction between visual degradation and instruction sensitivity, applied to OpenVLA-7B. The interaction is degradation-type-dependent: low lighting suppresses sensitivity, noise and fog have no significant effect, and occlusion leads to model collapse where instruction content has diminishing influence. The signal-to-noise ratio collapses from $3.0\times$ to $0.7\times$ under severe conditions, providing a quantitative deployment criterion. Instruction canonicalisation reduces sensitivity by 60–90% for two of three tasks under clean conditions. While these findings are specific to one model and synthetic degradation, the experimental framework—crossing visual and linguistic perturbation factors with noise floor analysis—is applicable to any VLA architecture and can inform deployment decisions for harsh-environment robotics. Future work should evaluate whether these findings generalise across VLA architectures, whether action distance differences translate to task success rate changes over full trajectories, how spatially non-uniform degradation affects the interaction, and whether internal model representations (e.g., cross-attention weights between vision and language tokens) reveal the mechanistic basis for the observed modality interaction patterns.

REFERENCES

[1] A. Brohan et al., "RT-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv:2307.15818*, 2023.

[2] M. J. Kim et al., "OpenVLA: An open-source vision-language-action model," *arXiv:2406.09246*, 2024.

[3] K. Black et al., "π_0: A vision-language-action flow model for general robot control," *arXiv:2410.24164*, 2024.

[4] Open X-Embodiment Collaboration, "Open X-Embodiment: Robotic learning datasets and RT-X models," *arXiv:2310.08864*, 2024.

[5] Y. Liu et al., "In-depth robustness analysis of vision-language-action models," *arXiv:2510.13626*, 2025.

[6] S. Chen et al., "VLATest: Testing and evaluating vision-language-action models," *arXiv:2409.12894*, 2024.

[7] B. Wen et al., "Benchmarking vision, language, & action models on robotic learning tasks," *arXiv:2411.05821*, 2024.

[8] X. Li et al., "Evaluating real-world robot manipulation policies in simulation," in *Proc. CoRL*, 2024.

[9] J. Woo, "Does how you ask matter? Instruction phrasing affects VLA action variance," in *ICRA 2026 Workshop*, 2026.

[10] Octo Model Team, "Octo: An open-source generalist robot policy," in *Proc. RSS*, 2024.

[11] W. Huang et al., "Inner monologue: Embodied reasoning through planning with language models," in *Proc. CoRL*, 2023.

[12] J. Liang et al., "Code as policies: Language model programs for embodied control," in *Proc. ICRA*, 2023.

[13] S. Tellex et al., "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proc. AAAI*, 2011.

[14] M. Ahn et al., "Do as I can, not as I say: Grounding language in robotic affordances," in *Proc. CoRL*, 2022.

[15] J. Woo, "Task-dependent sensitivity of VLA models to instruction wording," *Electronics Letters*, under review, 2026.