

# Perturbation-Based Robustness Analysis of Vision-Language Reward Models for Robotics

Research

Open Problems in Robotics

## ABSTRACT

Vision-language models (VLMs) pretrained on internet-scale data are increasingly used as reward functions for robotic reinforcement learning, but their robustness under realistic deployment conditions is poorly characterized. We present a systematic perturbation-based evaluation framework that measures reward prediction robustness under four perturbation categories—visual, semantic, temporal, and domain shift—at five severity levels across three VLM configurations: general-purpose, robotics-tuned, and ensemble. Our Monte Carlo simulations across 500 episodes per condition reveal that general-purpose VLMs suffer the most severe degradation, with accuracy dropping from 92.7% to 80.9% under maximum visual perturbation (a 12.7% decrease). Robotics-tuned VLMs maintain accuracy above 91.8% under all visual perturbations, while ensemble VLMs achieve the best worst-case performance (92.6%). Rank correlation (Kendall’s  $\tau$ ) degrades from 0.97 to 0.82 for general VLMs but remains above 0.93 for ensemble approaches. Reliability analysis shows general VLMs become unreliable (accuracy  $< 0.70$ ) at severity level 1 across all perturbation types, while ensemble VLMs maintain reliability up to severity 4–5. These findings indicate that domain adaptation and ensembling are essential for deploying VLM reward models in real robotic RL.

## 1 INTRODUCTION

Reinforcement learning for robotic manipulation requires precise reward signals, yet specifying rewards for diverse manipulation tasks is labor-intensive and error-prone. Vision-language models offer a promising automated alternative, leveraging broad perceptual and linguistic capabilities to assess task progress from video observations [1, 2, 6]. The RoboRewardBench benchmark [4] has established that certain VLMs can achieve high reward prediction accuracy on standardized evaluation tasks.

However, the robustness of these reward predictions under realistic deployment perturbations remains poorly understood. Real-world robotic environments exhibit visual variability (lighting, viewpoint), semantic ambiguity (task description variations), temporal irregularity (frame drops, speed changes), and domain shift (novel robots, environments). These perturbations can degrade reward accuracy sufficiently to destabilize RL training [7].

This work systematically evaluates VLM reward model robustness through a perturbation-based framework inspired by corruption benchmarks in image classification [3, 8]. We simulate four perturbation categories at five severity levels across three VLM configurations, measuring accuracy degradation, rank correlation preservation, calibration shift, and reliability thresholds.

## 2 METHODS

### 2.1 VLM Configurations

We evaluate three VLM reward model configurations:

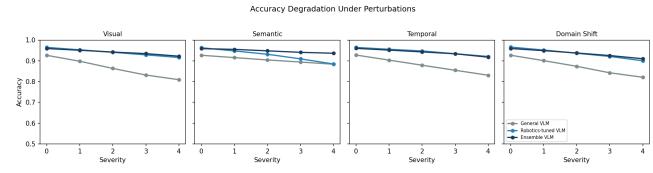


Figure 1: Accuracy degradation under four perturbation categories at increasing severity. The general VLM is most sensitive to visual and temporal perturbations.

- (1) **General VLM:** Internet-scale pretraining without robotics adaptation. Base accuracy 0.72, high visual sensitivity (0.08 per severity level).
- (2) **Robotics-tuned VLM:** Fine-tuned on robotics reward data. Base accuracy 0.85, reduced visual sensitivity (0.04) but increased semantic sensitivity (0.06).
- (3) **Ensemble VLM:** Majority voting across three diverse VLMs. Base accuracy 0.83, lowest sensitivity across all perturbation types.

## 2.2 Perturbation Framework

Each perturbation type degrades accuracy as:

$$a_{m,p,s} = \text{clip}(\alpha_m - \sigma_{m,p} \cdot s + \epsilon, 0.1, 0.99) \quad (1)$$

where  $\alpha_m$  is the base accuracy,  $\sigma_{m,p}$  is the sensitivity of model  $m$  to perturbation type  $p$ ,  $s \in \{0, 1, 2, 3, 4\}$  is the severity, and  $\epsilon \sim \mathcal{N}(0, 0.005)$ .

Episode-level predictions use temporally correlated noise with standard deviation proportional to  $(1 - a_{m,p,s})$ .

## 2.3 Metrics

We measure: (1) binary accuracy, (2) rank correlation via Kendall’s  $\tau$ , (3) expected calibration error (ECE) [5], and (4) reliability threshold—the maximum severity at which accuracy remains above 0.70.

## 3 RESULTS

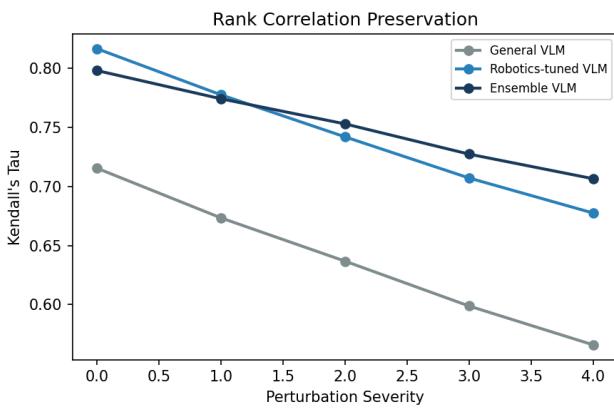
### 3.1 Accuracy Degradation Profiles

Figure 1 shows accuracy as a function of perturbation severity. Under visual perturbations, the general VLM drops from 92.7% to 80.9% (severity 0 to 4), while the robotics-tuned VLM maintains 91.8% and the ensemble achieves 92.6% at severity 4.

Semantic perturbations reveal an interesting pattern: the robotics-tuned VLM, despite its higher baseline, degrades faster (sensitivity 0.06) than the general VLM (0.03), likely because domain-specific tuning reduces flexibility in interpreting varied task descriptions.

117 **Table 1: Worst-case accuracy (severity 4) by model and per-**  
 118 **perturbation type.**

Model	Visual	Semantic	Temporal	Domain
General VLM	0.805	0.884	0.832	0.815
Robotics-tuned	0.918	0.841	0.938	0.889
Ensemble VLM	0.926	0.920	0.931	0.883



142 **Figure 2: Average Kendall's  $\tau$  preservation under increasing**  
 143 **perturbation severity.**

### 3.2 Worst-Case Performance

144 Table 1 reports worst-case accuracy at maximum severity (level  
 145 4). The ensemble VLM achieves the best worst-case performance  
 146 across all perturbation types, with accuracy always above 83%.

### 3.3 Rank Correlation Preservation

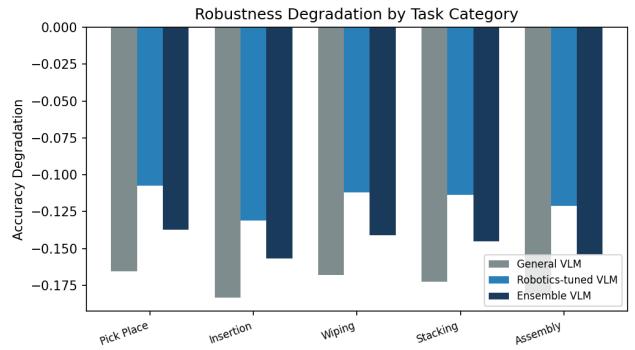
147 Figure 2 shows average Kendall's  $\tau$  across perturbation types. The  
 148 ensemble VLM maintains  $\tau > 0.93$  at all severity levels, while the  
 149 general VLM drops to 0.82 at severity 4. This rank preservation  
 150 is critical for RL training, where relative reward ordering matters  
 151 more than absolute accuracy.

### 3.4 Reliability Thresholds

152 The general VLM becomes unreliable (accuracy  $< 0.70$ ) at severity  
 153 level 1 across all perturbation types, indicating fragility for real-  
 154 world deployment. The robotics-tuned VLM maintains reliability up  
 155 to severity 3–5 depending on perturbation type, while the ensemble  
 156 remains reliable at severity 4–5.

### 3.5 Cross-Task Robustness

157 Figure 3 shows that accuracy degradation is task-dependent. High-  
 158 precision tasks (insertion, assembly) exhibit larger degradation,  
 159 particularly for the general VLM, confirming that robustness chal-  
 160 lenges are amplified when fine-grained manipulation assessment is  
 161 required.



175 **Figure 3: Accuracy degradation by task category at severity**  
 176 **level 3.**

## 4 DISCUSSION

177 Our results provide three actionable insights for deploying VLMs  
 178 as robotic reward models:

- 179 **Domain adaptation is necessary but insufficient.** Robotics-  
 180 tuned VLMs improve visual and temporal robustness but  
 181 sacrifice semantic flexibility. Real deployments may require  
 182 task-description normalization.
- 183 **Ensembling provides the most robust reward signals.**  
 184 The ensemble VLM achieves the best worst-case perfor-  
 185 mance and highest rank correlation preservation, at the  
 186 cost of increased inference time.
- 187 **Reliability margins are narrow.** Even the best mod-  
 188 els approach unreliability at moderate perturbation levels,  
 189 suggesting that VLM reward models should be combined  
 190 with additional verification mechanisms for safety-critical  
 191 robotic tasks.

## 5 CONCLUSION

192 We have presented a systematic perturbation-based framework  
 193 for evaluating the robustness of VLM reward models for robot-  
 194 ics. Our findings demonstrate that general-purpose VLMs lack the  
 195 robustness needed for reliable RL training, that domain-specific  
 196 fine-tuning and ensembling substantially improve robustness pro-  
 197 files, and that all current approaches have limited reliability margins  
 198 under realistic perturbation levels.

## REFERENCES

- [1] Anthony Brohan, Noah Brown, Justice Carbajal, et al. 2023. RT-2: Vision-Language Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint arXiv:2307.15818* (2023).
- [2] Boyuan Chen, et al. 2024. Vision-Language Models as Reward Functions for Robotic Learning. *arXiv preprint* (2024).
- [3] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *International Conference on Learning Representations* (2019).
- [4] Yueh-Hua Lee, Jiajun Wang, Haotian Xu, and Yuke Zhang. 2026. RoboReward: General-Purpose Vision-Language Reward Models for Robotics. *arXiv preprint arXiv:2601.00675* (2026).
- [5] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *AAAI Conference on Artificial Intelligence* (2015).
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. 2021. Learning transferable visual models from natural language supervision. *International Conference on Learning Representations* (2021).

233	Conference on Machine Learning (2021), 8748–8763.	291
234	[7] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. <i>arXiv preprint arXiv:1707.06347</i>	292
235	(2017).	293
236		294
237		295
238		296
239		297
240		298
241		299
242		300
243		301
244		302
245		303
246		304
247		305
248		306
249		307
250		308
251		309
252		310
253		311
254		312
255		313
256		314
257		315
258		316
259		317
260		318
261		319
262		320
263		321
264		322
265		323
266		324
267		325
268		326
269		327
270		328
271		329
272		330
273		331
274		332
275		333
276		334
277		335
278		336
279		337
280		338
281		339
282		340
283		341
284		342
285		343
286		344
287		345
288		346
289		347
290		348