REWARD HACKING IN MULTIMODAL REINFORCEMENT LEARNING VIA LINGUISTIC PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has become a powerful post-training paradigm for multimodal large language models (MLLMs), which are increasingly deployed in safety-critical domains. However, we reveal a critical yet overlooked phenomenon: models optimized via RL may exploit linguistic priors encoded in their language backbone to predict answers without truly processing image content. We term this behavior ***Visually-bypassed Shortcut Learning (VSL)***. Unlike reward hacking in text-only models, VSL is invisible to existing monitoring, as both pathways yield superficially identical reasoning traces. We show that this shortcut is systematically favored during RL optimization due to its lower computational cost, reduced reward variance, and the rich statistical structures of the language backbone, causing RL training to actively degrade visual grounding. Using image-perturbation sensitivity metrics, we verify the pervasiveness of VSL across multiple MLLMs and task types in clean training environments, and find its severity strongly correlates with task-specific linguistic prior strength. These findings expose a silent trade-off: benchmark scores keep improving while visual grounding steadily erodes, yet no existing signal indicates when the degradation becomes unacceptable. We therefore propose ***Visual Information Evaluation (VIE)***, a low-cost detector requiring only forward passes and no training modification, which exploits the insight that shortcut-reliant models need less visual information to maintain high rewards. Experiments show that VIE reliably detects visual reward hacking under both in-distribution and out-of-distribution settings.

1 INTRODUCTION

The integration of visual modalities into large language models marks a major leap in artificial intelligence capabilities, enabling models to perceive, understand, and respond to the rich information contained in images. Multimodal large language models (MLLMs) Radford et al. (2021); Li et al. (2023); Liu et al. (2023; 2024) have demonstrated outstanding performance across a wide range of vision-language tasks, from visual question answering to complex visual reasoning. To further align these models with human preferences and improve task performance, reinforcement learning (RL) Schulman et al. (2017); Rafailov et al. (2023); Shao et al. (2024); Guo et al. (2025) has become a standard post-training paradigm. These post-trained models are increasingly deployed in safety-critical domains such as medical image analysis and autonomous driving, where accurate understanding of visual information is paramount and where confident yet ungrounded outputs can cause serious safety issues. However, a critical yet under-explored question arises: *when a multimodal model is optimized to maximize rewards, is it truly learning to understand images better, or is it learning to take cheating shortcuts?*

In language models, the problem of reward hacking Skalse et al. (2022); Bondarenko et al. (2025), in which models exploit vulnerabilities in proxy reward functions to gain high rewards without fulfilling the actual goal, has received systematic attention. Recent studies show that language reasoning models may develop implicit hacking strategies, and such shortcuts remain hidden outside their chain of thought, making monitoring difficult IT; Liu et al. (2026). Nevertheless, existing research on reward hacking is almost exclusively limited to the text modality, leaving a critical gap: the novel information risks introduced by the visual modality in multimodal reinforcement learning training still lack systematic investigation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We focus on the unique form of reward hacking introduced by new information modalities in multimodal reinforcement learning, which we term ***Visually-bypassed Shortcut Learning (VSL)***. Our core insight is that for a multimodal model, the correct answer to a visual question might be obtained through two paths. One is genuine visual perception, which processes and reasons over image content. The other is the use of linguistic priors, which relies on statistical patterns in training corpora to predict answers without genuinely utilizing visual inputs. Both paths may yield identical reward signals, but only the former reflects real visual understanding. More importantly, from the perspective of any text-based monitor or human reader, the reasoning process generated by the model appears identical in both cases, and visual shortcuts are completely invisible in the chain of thought.

The information asymmetry between different modalities in multimodal reinforcement learning creates conditions for reward hacking. The training optimization process does not care how rewards are obtained, and it naturally favors the path that provides more stable and easily optimized gradient signals. Our experimental validation shows that exploiting linguistic priors systematically meets this condition better than genuine visual perception. It incurs lower computational costs, produces reward signals with smaller variance Berchenko (2024), and benefits from the rich statistical structures already encoded in the language model backbone. Consequently, reinforcement learning training not only fails to improve visual grounding but also actively and gradually weakens it, driving the model to rely increasingly less on image content during training. This reveals a fundamental misalignment between proxy rewards and real objectives. The finding indicates that a model trained with visual reinforcement learning may perform well on standard benchmarks, yet its visual reasoning may merely be a post-hoc rationalization based on linguistically prior answers. When the model requires genuine visual perception, it fails silently and unpredictably. Furthermore, unlike explicit reward hacking in text models, VSL leaves no detectable traces in model outputs, making existing monitoring approaches extremely challenging. This raises a broader concern: multimodal RL alignment may be quietly optimizing for the wrong objective, and much of the progress on visual benchmarks could reflect increasingly linguistic shortcuts rather than genuine advances in visual understanding.

In this paper, we first verify the existence and pervasiveness of VSL in multimodal reinforcement learning training. Using carefully designed image-perturbation sensitivity metric, we quantify the model's true dependence on visual inputs for answering questions, and show that this dependence declines across multiple task types and models as reinforcement learning proceeds. Crucially, this degradation occurs in clean training environments without manually injected vulnerabilities, demonstrating that VSL is an inherent property of multimodal RL optimization rather than a product of adversarial design. ***Second,*** we find that the severity of VSL is highly correlated with the strength of linguistic priors in specific task types, that larger-scale models exhibit more severe VSL phenomena. ***Finally,*** these findings expose a pressing need that current training pipelines cannot satisfy: RL induces a silent trade-off in which benchmark scores keep improving while genuine visual grounding steadily erodes, and no existing signal tells practitioners *when* the decline in visual sensitivity becomes unacceptable, even though further training might still lift benchmark performance. Monitoring this trade-off during training calls for a detector that is low-cost, requires only forward passes, and operates without modifying the training process. To this end, we propose the ***Visual Information Evaluation (VIE)*** method. Our core insight is that a model relying on visual shortcuts requires less visual information to achieve high rewards than a model performing genuine visual reasoning. By gradually reducing image information and measuring the corresponding reward curve, VIE reliably detects the point at which the model's dependency on visual information has dropped to a negligible level, distinguishing shortcut-based answers from answers based on genuine visual grounding under both in-distribution and out-of-distribution settings. Our contributions are summarized as follows:

- We identify a unique reward hacking problem associated with the visual modality in multimodal models, named Visually-bypassed Shortcut Learning.
- Extensive experiments verify the existence and pervasiveness of VSL across multiple models and datasets, and reveal that VSL intensity is strongly correlated with the strength of task-specific linguistic priors.
- We propose the Visual Information Evaluation method to detect reward hacking in visual multimodal models, which achieves reliable detection under both in-distribution and out-of-distribution settings.

2 Related Work

Reinforcement Learning in Multimodal Models. The success of reinforcement learning Schulman et al. (2017); Rafailov et al. (2023); Shao et al. (2024); Guo et al. (2025); Zheng et al. (2025); Ahmadian et al. (2024) in enhancing the reasoning capabilities of LLMs has driven its rapid adoption in multimodal settings. Building on DeepSeek-R1 Guo et al. (2025), which shows that rule-based verifiable rewards combined with lightweight algorithms such as GRPO Shao et al. (2024) can elicit complex reasoning without heavy supervision, a growing body of work transfers this paradigm to visual question answering and spatial reasoning Pan et al. (2025); Meng et al. (2025); Zhou et al. (2025b), refines it with cold-start SFT or process-level rewards Huang et al. (2025); Yang et al. (2025b); Huang et al.; Fan et al. (2025), and extends it to segmentation, GUI agents, video, and audio Liu et al. (2025b); Lu et al. (2026); Feng et al. (2025); Li et al. (2025a). These pipelines largely inherit text-centric practices from LLMs: rewards are computed solely from the correctness of textual outputs, with no mechanism verifying whether visual inputs are genuinely utilized Zhou et al. (2025a). Recent studies have begun to notice the resulting risk that RL training weakens visual perception. PAPO Wang et al. (2026) observes that GRPO fails to adequately capture visual cues in perception-heavy tasks, and proposes perception-aware policy optimization with an implicit perception loss. Vision-SR1 Li et al. (2025b) finds that mainstream post-training pushes models toward language-dominant reasoning, and restores visual grounding via a three-stage self-reward strategy; modality-balancing preference optimization Liu et al. (2025a) similarly counteracts language dominance through adversarial negative mining. These observations are consistent with, and indirectly validate, our analysis. However, they tackle the symptom by designing new training algorithms, leaving open *why* RL erodes visual grounding, *how severely* it does so across training, and *at which step* the degradation becomes unacceptable even as benchmark scores keep rising. Our work answers these complementary questions: we formalize VSL, propose Visual Sensitivity (VS) to quantify the visual dependence of intermediate checkpoints, and introduce VIE as a low-cost, training-agnostic monitor that applies to both established and newly proposed algorithms.

Reward Hacking. Reward hacking refers to agents exploiting vulnerabilities in proxy rewards to gain high rewards without fulfilling the true objective Amodei et al. (2016); Bondarenko et al. (2025), classically interpreted as an instance of Goodhart's Law Goodhart (1984); Skalse et al. (2022). Existing countermeasures either repair the reward specification, e.g., Inverse Reward Design Hadfield-Menell et al. (2017) and reward-model ensembling Ramé et al. (2024), regularize policy deviation from a reference policy, e.g., KL regularization Wei et al. (2019) and ORPO Laidlaw et al. (2024), or involve humans through assistance games Fern et al. (2014); Shah et al. (2020). Nevertheless, this line of research is almost exclusively limited to the text modality, including recent studies on shortcuts hidden in the reasoning chains of language models IT; Liu et al. (2026), leaving the risks introduced by the visual modality in multimodal RL unexplored. We fill this gap by revealing VSL, a visually invisible form of reward hacking, and providing tools to quantify and detect it during training.

3 Methodology

3.1 Problem Formalization

Let $\mathcal{M}_\theta$ denote a multimodal large language model parameterized by θ, which takes as input an image $I \in \mathcal{I}$ and a language question $q \in \mathcal{Q}$, and produces a response from which a final answer $\hat{f}_\theta(I, q) \in \mathcal{A}$ is extracted. We consider a visual question answering task defined over a dataset $\mathcal{D} = \{(I_i, q_i, a_i^*)\}_{i=1}^N$, where a_i^* denotes the ground-truth answer. In the reinforcement learning post-training paradigm, the model is optimized to maximize a scalar reward $R(\hat{f}_\theta(I, q),\ a^*)$, which reflects answer correctness.

For a multimodal model, the correct answer to a visual question can in principle be obtained through two fundamentally distinct pathways, as illustrated in Fig 1 (top). The first is *genuine visual perception*, in which the model processes and reasons over the visual content of I to produce $\hat{f}_\theta(I, q)$. The second is the exploitation of *linguistic priors*, in which the model leverages statistical patterns in its language backbone to predict a^* from q, without substantively utilizing the visual input. Both pathways can yield identical reward signals under standard evaluation, yet only the former reflects true visual understanding.

3.1.1 Formal Definition of Visually-bypassed Shortcut Learning

To quantify the degree to which a model relies on visual inputs, we introduce the **Visual Sensitivity** metric. Let $\mathcal{P}(I)$ denote a set of perturbed variants of image I, constructed such that the semantic

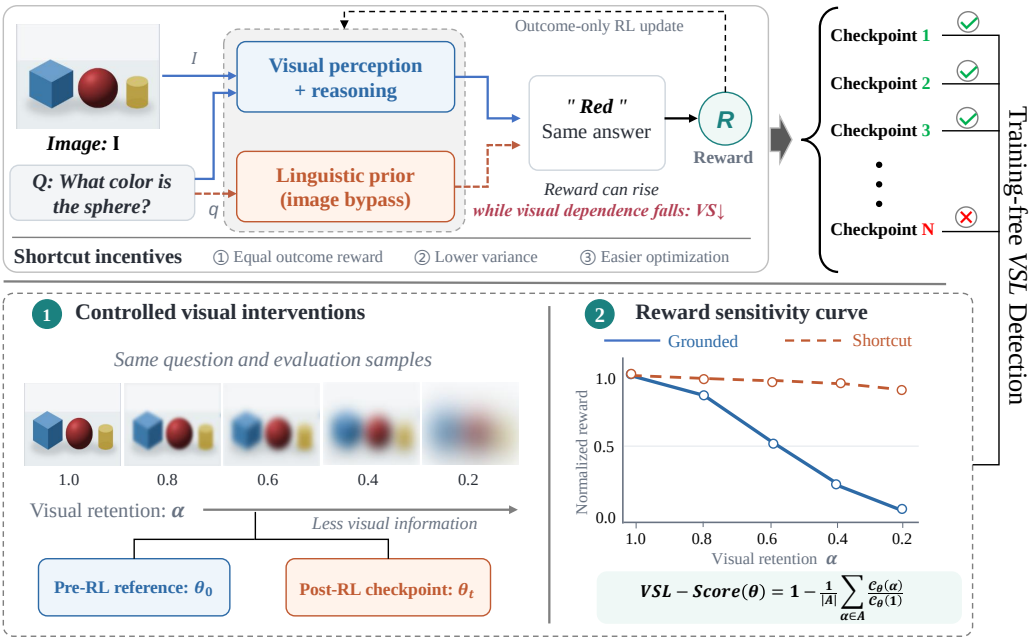

Figure 1: Overview of our framework. **Top**: the linguistic-prior pathway (image bypass) and genuine visual perception receive identical rewards, so reward rises while visual dependence falls (VS↓), driven by three shortcut incentives. **Bottom**: VIE degrades visual inputs, compares reward sensitivity curves across checkpoints, and yields a training-free VSL-Score.

content required to answer q is disrupted while low-level statistical features are partially preserved (detailed perturbation designs are described in Section 4.1). We define the visual sensitivity of a model $\mathcal{M}_\theta$ on a sample (I, q) as:

$$\mathrm{VS}_\theta(I, q) = 1 - \mathbb{E}_{\tilde{I}\sim\mathcal{P}(I)}\left[\mathbf{1}\left(\hat{f}_\theta(I, q) = \hat{f}_\theta(\tilde{I}, q)\right)\right] \tag{1}$$

where $\hat{f}_\theta(\cdot)$ denotes the extracted and normalized final answer. Intuitively, $\mathrm{VS}_\theta(I, q)$ measures the probability that perturbing the image causes the model to change its answer: a high value indicates strong dependence on visual content, whereas a low value indicates that the model's answer is robust to image perturbation and thus likely driven by linguistic priors rather than genuine visual perception. At the dataset level, we define the aggregate Visual Sensitivity as:

$$\mathrm{VS}_\theta = \mathbb{E}_{(I,q,a^*)\sim\mathcal{D}}\left[\mathrm{VS}_\theta(I, q)\right] \tag{2}$$

Answer Extraction. For models that produce chain-of-thought outputs (e.g., responses enclosed in `<think>...</think>` tags), we extract the final answer $\hat{f}_\theta$ from the content following the closing reasoning token, and apply string normalization prior to comparison. This ensures that variation in reasoning length or phrasing introduced by RL training does not confound the metric.

Definition of Visually-bypassed Shortcut Learning. We formally define *Visually-bypassed Shortcut Learning (VSL)* as the phenomenon in which RL training induces a systematic decline in VS_θ over the checkpoint sequence $\{\theta_t\}_{t=0}^{T}$ produced during training (Fig 1, top right): $\mathrm{VS}_{\theta_t} > \mathrm{VS}_{\theta_{t+\Delta}}$. Crucially, this degradation occurs in the absence of any deliberately injected vulnerabilities, and is not detectable from the model's textual outputs alone, as the chain-of-thought reasoning generated under both pathways appears superficially indistinguishable to any text-based observer.

3.1.2 Why RL Optimization Favors Visual Shortcuts

We argue that the linguistic shortcut pathway is *systematically favored* over genuine visual perception during RL optimization, due to three compounding factors, each empirically verified in Appendix F.

- ***Reward Signal Equivalence.*** Under a reward function $R(\hat{f}_\theta(I, q), a^*)$, both pathways yield identical reward when they produce the correct answer. Let p_{vis} and p_{lang} denote the probabilities of

obtaining a correct answer via the visual and linguistic pathways, respectively. The RL objective does not distinguish between these two sources, and thus provides no gradient signal to prefer the former:

$$\mathbb{E}\left[R \mid \text{visual path}\right] = \mathbb{E}\left[R \mid \text{linguistic path}\right] \quad \text{whenever } p_{\text{vis}} = p_{\text{lang}} \tag{3}$$

- ***Lower Reward Variance via Linguistic Priors.*** The variance of the reward signal plays a critical role in policy gradient optimization. For tasks with strong linguistic priors, the language pathway produces answers with higher consistency across rollouts, yielding a lower-variance reward signal. Formally, let Var_{lang} and Var_{vis} denote the reward variances under the two pathways. We have $\text{Var}_{\text{lang}}(R) \leq \text{Var}_{\text{vis}}(R)$, since linguistic predictions are more stable than visual interpretations which depend on fine-grained, high-dimensional perceptual features. Policy gradient methods such as GRPO (Shao et al., 2024) implicitly favor lower-variance reward paths, as they produce more consistent gradient estimates and enable more stable policy updates.

- ***Computational Asymmetry.*** The linguistic shortcut pathway bypasses the need for deep visual feature integration, relying predominantly on representations already encoded in the language model backbone. Consequently, the effective optimization path for linguistic shortcuts is shorter and encounters a smoother loss landscape than genuine visual reasoning. This asymmetry systematically biases the RL optimizer toward the linguistic pathway, even when the visual pathway would produce equally correct answers.

These three factors establish that RL training not only fails to penalize visual shortcut usage, but actively creates optimization pressure in its favor. The extent of this pressure is further modulated by the strength of task-specific linguistic priors: tasks for which p_{lang} is high are more susceptible to VSL, a prediction we empirically validate in Section 4.3.

3.2 Visual Information Evaluation Method

The invisibility of VSL in model outputs presents a fundamental challenge for detection: a model exploiting linguistic shortcuts produces reasoning traces that are superficially identical to those of a genuinely grounded model. We therefore seek a detection signal that operates not on the model's textual output, but on its behavioral response to controlled variations in visual input quality. Our key insight is the following: *A model relying on linguistic shortcuts requires less visual information to maintain high reward, whereas a model performing genuine visual reasoning exhibits a systematic reward degradation as visual information is progressively reduced.*

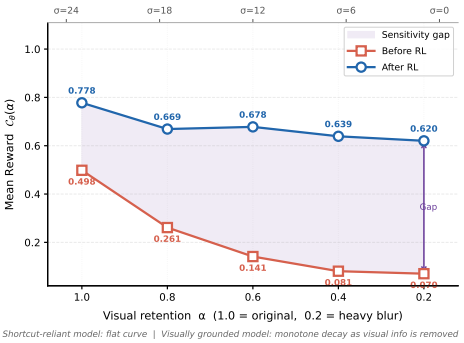

Figure 2: Reward sensitivity curves $\mathcal{C}_\theta(\alpha)$ for a visually-grounded model (Before RL) and a shortcut-reliant model (After RL) across five visual retention levels α.

We define a family of image degradation operators $\{\mathcal{F}_\alpha\}_{\alpha\in[0,1]}$, parameterized by a retention coefficient α, where $\mathcal{F}_1(I) = I$ (original image) and $\mathcal{F}_0(I) = I_\varnothing$ (fully uninformative image). For intermediate values of α, $\mathcal{F}_\alpha$ progressively removes visual information from I. In our implementation, we instantiate $\mathcal{F}_\alpha$ as a Gaussian blurring operator with kernel standard deviation $\sigma(\alpha) \propto (1-\alpha)$, applied at a discrete set of levels $\alpha \in \{1.0, 0.8, 0.6, 0.4, 0.2\}$, $\alpha = 0.0$ (uninformative) is excluded.

For a model $\mathcal{M}_\theta$ evaluated on dataset $\mathcal{D}$, we define the Reward Sensitivity Curve $\mathcal{C}_\theta$:

$$\mathcal{C}_\theta(\alpha) = \mathbb{E}_{(I,q,a^*)\sim\mathcal{D}}\left[R\Big(\hat{f}_\theta(\mathcal{F}_\alpha(I),\, q),\; a^*\Big)\right] \tag{4}$$

The shape of $\mathcal{C}_\theta(\alpha)$ encodes the degree to which the model's performance depends on visual input fidelity. As illustrated in Fig 1 (bottom) and Fig 2, two distinct behavioral profiles emerge:

Shortcut-reliant model: $\mathcal{C}_\theta(\alpha) \approx \mathcal{C}_\theta(1)$ for all α, indicating that reward is largely invariant to visual degradation, as the model's predictions are driven by linguistic priors independent of image content.

Visually grounded model: $\mathcal{C}_\theta(\alpha)$ exhibits a monotonic decay as $\alpha \to 0$, reflecting genuine dependence on visual information for answer generation.

To use *Visual Information Evaluation Method* for quantitative detection, we summarize $\mathcal{C}_\theta(\alpha)$ into a scalar *VSL Score*:

$$\text{VSL-Score}(\theta) = 1 - \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \frac{\mathcal{C}_\theta(\alpha)}{\mathcal{C}_\theta(1)} \tag{5}$$

where $\mathcal{A} = \{0.2, 0.4, 0.6, 0.8\}$ is the set of degradation levels excluding the original image, and the ratio $\mathcal{C}_\theta(\alpha)/\mathcal{C}_\theta(1)$ normalizes performance relative to the full-image baseline. A low VSL-Score (close to 0) indicates that performance is well-maintained under degradation, characteristic of shortcut reliance. A high VSL-Score (close to 1) indicates strong sensitivity to visual information, characteristic of genuine visual grounding.

4 Experiments

4.1 Experiment Setting

Models and Datasets: To verify the existence and pervasiveness of VSL, we conduct experiments on five representative open-source MLLMs spanning different model families and parameter scales. Specifically, we select three models from the Qwen3-VL series (Bai et al., 2025): Qwen3-VL-2B, Qwen3-VL-4B, and Qwen3-VL-8B, which are built upon the Qwen3 backbone (Yang et al., 2025a). We additionally include LLaVA-OneVision-4B (An et al., 2025), a strong open-source model that demonstrates competitive performance across single-image, multi-image, and video scenarios, and InternVL3.5-4B (Wang et al., 2025). For RL fine-tuning in our experiments, all models are trained from their respective base or instruct checkpoints using GRPO (Shao et al., 2024). **"Before"** and **"After"** refer to the model performance immediately after SFT and after RL training, respectively. Further experimental details can be found in Appendix A.

We employ three benchmark datasets that provide a controlled gradient of linguistic prior strength, enabling systematic analysis of how prior intensity modulates VSL severity. VQAv2 (Goyal et al., 2017) is a large-scale real-image VQA dataset designed to reduce language bias by pairing each question with two complementary images that elicit different answers. OK-VQA (Marino et al., 2019) is a knowledge-based VQA dataset whose questions require external world knowledge beyond image content to answer correctly CLEVR (Johnson et al., 2017) is a synthetically rendered diagnostic dataset designed explicitly to test compositional visual reasoning over geometric properties, spatial relationships, and object attributes.

Evaluation Metrics: We adopt two complementary metrics to assess the degree of visually-bypassed shortcut learning: **(1) Visual Sensitivity (VS); (2) Task Accuracy**.

Our primary metric is the Visual Sensitivity score defined in Eq. equation 2, which measures the probability that a model changes its answer when the input image is perturbed. To compute VS, we apply three categories of image perturbation to each evaluation sample, as detailed below.

(a) Patch Shuffling. The image is divided into a 4×4 grid whose patches are randomly permuted. This destroys object identities, spatial relationships, and scene layout while preserving local textures and low-level statistics, testing whether the model depends on holistic semantic structure.

(b) Gaussian Noise Substitution. The image is replaced entirely with white noise from $\mathcal{N}(0, \sigma^2)$, with σ matching the original pixel distribution. This **extreme perturbation** removes all visual information and serves as an upper-bound reference: a model invariant to it almost certainly exploits linguistic priors exclusively.

(c) Gaussian Blurring. A Gaussian kernel with $\sigma = 10$ (moderate, α_{low}) and $\sigma = 30$ (severe, α_{high}) progressively removes high-frequency detail while preserving coarse layout and color. As a **graded perturbation** retaining partial visual structure, it bridges the VS metric and the VIE framework.

4.2 RQ1: Does VSL Exist in Multimodal RL Training?

We begin by investigating whether Visually-bypassed Shortcut Learning emerges as an inherent consequence of RL training in multimodal models. We conduct RL fine-tuning on Qwen3-VL-4B using GRPO on the OK-VQA training set, and evaluate the model's task accuracy and Visual Sensitivity (VS) before and after training under all three perturbation types. Experimental analysis regarding different reinforcement learning methods and reward settings is provided in Appendix A.

Table 1 presents the core results. On unperturbed images, RL training yields a clear accuracy gain, confirming that GRPO successfully optimizes the reward signal and improves benchmark performance. However, when we examine the VS metric across all perturbation conditions, a strikingly different picture emerges: the average VS score *declines* from 0.4730 to 0.3063, indicating that the model's

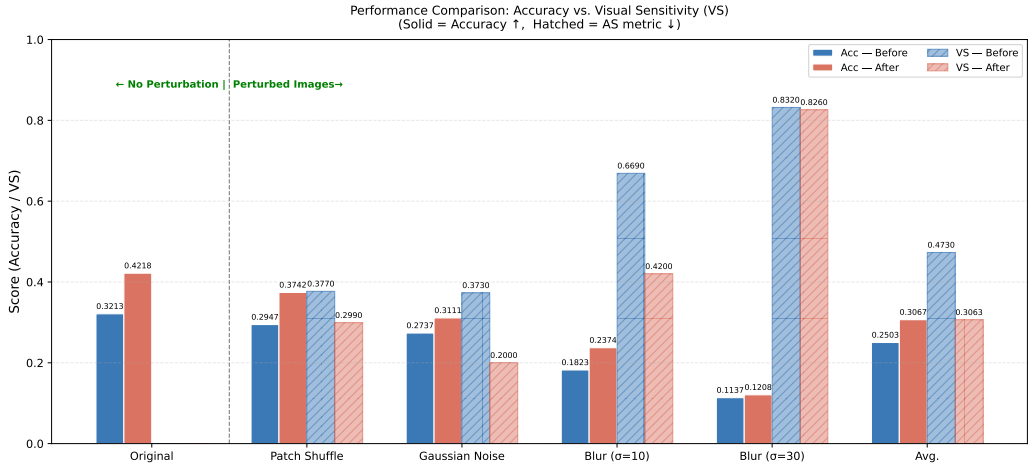

Figure 3: Visualization of accuracy and Visual Sensitivity (VS) before and after Reinforcement Learning across different visual perturbations.

Table 1: Accuracy and Visual Sensitivity (VS) before and after Reinforcement Learning. The results show that accuracy increases while VS values decrease across all perturbations, indicating a reduced reliance on visual information after RL. The "**average**" column represents the mean VS across *Patch Shuffle, Gaussian Noise, and Gaussian Blur* ($\sigma = 10$).

Visual transformation	Accuracy ↑		VS ↓	
	Before	After	Before	After
Original	0.3213	0.4218$_{\uparrow 0.1005}$	-	-
Patch Shuffle	0.2947	0.3742$_{\uparrow 0.0795}$	0.3770	0.2990$_{\downarrow 0.0780}$
Gaussian Noise	0.2737	0.3111$_{\uparrow 0.0374}$	0.3730	0.2000$_{\downarrow 0.1730}$
Blur ($\sigma = 10$)	0.1823	0.2374$_{\uparrow 0.0551}$	0.6690	0.4200$_{\downarrow 0.2490}$
Blur ($\sigma = 30$)	0.1137	0.1208$_{\uparrow 0.0071}$	0.8320	0.8260$_{\downarrow 0.0060}$
Average	0.2503	0.3067$_{\uparrow 0.0564}$	0.4730	0.3063$_{\downarrow 0.1667}$

outputs have become substantially less sensitive to image content after training. This divergence constitutes the defining empirical signature of VSL, and directly validates our theoretical prediction in Section 3.1.2 that RL optimization simultaneously maximizes proxy rewards and erodes genuine visual grounding. The corresponding visualization results are shown in Figure 3.

The degradation in VS is not an artifact of any single perturbation design but holds robustly across all three perturbation categories, as shown in Figure 4. As perturbation strength increases from left (Patch Shuffle) to right (Blur $\sigma = 30$), accuracy decreases monotonically for both before- and after-RL models. However, the VS curves show the opposite trend — VS *increases* with perturbation strength because stronger perturbations are more likely to force an answer change even in a shortcut-reliant model. Notably, Gaussian blur perturbations at severe blur ($\sigma = 30$), VS remaining relatively stable across training. This suggests that the post-RL model still responds to the near-total removal of visual information, it retains awareness that a completely uninformative image differs from the original. This result implies that RL training does not eliminate all visual processing, but selectively weakens the fine-grained, content-specific perception that is most necessary for genuine visual reasoning.

4.3 RQ2: What Factors Drive the Severity of VSL?

In this section, we investigate which factors modulate VSL severity. Table 2 reports VS scores before and after RL training across five representative MLLMs and three datasets. Figure 5 provides a visual of the two key factors analyzed below. Detailed analysis for all models can be found in Appendix B.

As shown in Table 2, VS declines after RL training universally: every model-dataset combination exhibits a drop in VS under all three types of perturbations. This includes models from three independent architectures — Qwen3-VL, LLaVA-OneVision, and InternVL3.5 — evaluated on various datasets. The universality of this finding across architectures and task types confirms that VSL

Table 2: Visual Sensitivity (VS) scores before and after RL training across five MLLMs and three datasets under three perturbation types. **Bold** values indicate post-RL results. A consistent decline in VS is observed across all model-dataset combinations, confirming the pervasiveness of VSL.

Models and Datasets		Patch Shuffle		Gaussian Noise		Blur ($\sigma = 10$)		Average	
		Before	**After**	Before	**After**	Before	**After**	Before	**After**
Qwen3-VL-2B	OKVQA	0.3280	**0.3104**	0.3100	**0.2420**	0.6170	**0.4840**	0.4183	**0.3454**
	VQAv2	0.4950	**0.2180**	0.3050	**0.1130**	0.5610	**0.3020**	0.4537	**0.2110**
	CLEVR	0.6420	**0.4870**	0.6090	**0.5350**	0.7590	**0.5630**	0.6700	**0.5283**
Qwen3-VL-4B	OKVQA	0.3770	**0.2990**	0.3730	**0.2000**	0.6690	**0.4200**	0.4730	**0.3063**
	VQAv2	0.3750	**0.3190**	0.2110	**0.2070**	0.4140	**0.3230**	0.3333	**0.2830**
	CLEVR	0.6860	**0.5550**	0.6890	**0.5160**	0.7510	**0.5620**	0.7086	**0.5443**
Qwen3-VL-8B	OKVQA	0.3877	**0.2149**	0.3612	**0.2174**	0.5871	**0.3751**	0.4453	**0.2691**
	VQAv2	0.3985	**0.3144**	0.4029	**0.2715**	0.6104	**0.4023**	0.4706	**0.3294**
	CLEVR	0.5030	**0.4100**	0.7380	**0.5140**	0.8010	**0.7330**	0.6807	**0.5523**
LLaVA-OneVision-4B	OKVQA	0.3420	**0.2540**	0.3070	**0.2430**	0.4820	**0.2790**	0.3770	**0.2586**
	VQAv2	0.3620	**0.2250**	0.2390	**0.0900**	0.4460	**0.2640**	0.3490	**0.1930**
	CLEVR	0.4630	**0.3940**	0.4700	**0.4400**	0.5770	**0.4760**	0.5033	**0.4367**
InternVL3.5-4B	OKVQA	0.3530	**0.2060**	0.3740	**0.2310**	0.6630	**0.4160**	0.4633	**0.2843**
	VQAv2	0.2950	**0.2420**	0.1930	**0.1310**	0.3860	**0.3560**	0.2913	**0.2430**
	CLEVR	0.4080	**0.2870**	0.5080	**0.4490**	0.5420	**0.5170**	0.4860	**0.4176**

is not a model-specific artifact but an inherent property of the RL optimization process, consistent with our theoretical analysis in Section 3.1.2.

Figure 5(a) plots the mean VS drop (Before − After RL) averaged across all five models for each dataset. A clear descending trend is observed as linguistic prior strength decreases: the mean VS drop is largest on OK-VQA, which has the strongest linguistic priors due to its reliance on external world knowledge that can be partially retrieved from the language backbone alone, and progressively diminishes through VQAv2 to CLEVR, where questions require genuine compositional visual reasoning and linguistic priors offer minimal predictive power. This monotonic relationship directly validates the core prediction of our framework: tasks that offer a richer linguistic shortcut provide a more stable and lower-variance reward signal, intensifying the RL optimization pressure toward the shortcut pathway.

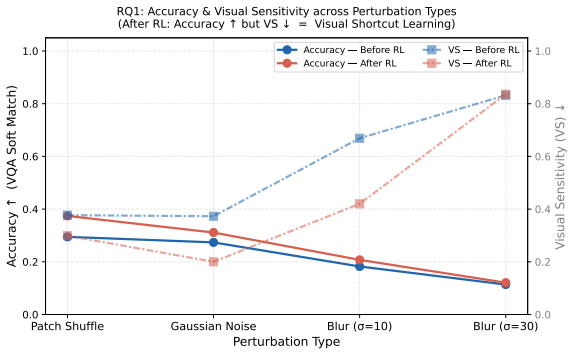

Figure 4: Trends in accuracy and Visual Sensitivity (VS) before and after Reinforcement Learning under various visual perturbations.

Figure 5(b) isolates the effect of model scale by comparing VS Before and After RL training across the Qwen3-VL-2B, 4B, and 8B variants. On high-prior OK-VQA, the VS drop increases monotonically with scale, suggesting that larger models, with richer and more expressive language backbones, provide more exploitable shortcuts precisely where linguistic priors are most predictive. This trend does not strictly hold on mid- and low-prior datasets: the 2B model exhibits an anomalous drop on VQAv2 (analyzed as an outlier in Appendix B), and cross-scale differences on CLEVR are small. We characterize model scale as an interacting factor rather than a universal amplifier: larger models intensify VSL most strongly on high-prior tasks, where the shortcut is most exploitable, whereas on low-prior tasks the achievable shortcut severity is inherently bounded by the limited predictive power of linguistic priors. Scaling up model parameters, typically associated with improved visual capabilities, thus does not mitigate VSL and may amplify it on shortcut-prone tasks.

4.4 RQ3: How Reliably Can We Detect VSL?

Having established the existence and driving factors of VSL, we now evaluate whether our proposed Visual Information Evaluation (VIE) method can reliably detect it.

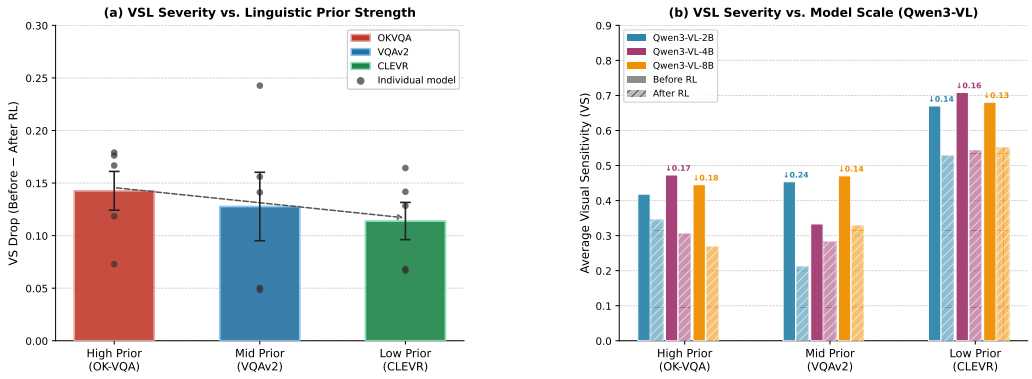

Figure 5: **(a)** Mean VS drop (Before−After RL) across all five models as a function of linguistic prior strength. Dots indicate individual model values; error bars denote standard error. The dashed arrow highlights the monotonic decreasing trend. **(b)** Average VS Before (solid) and After (hatched) RL training for the Qwen3-VL model family across three prior levels, illustrating that the scale-induced amplification of VSL is most pronounced on high-prior tasks.

Figure 2 presents the reward sensitivity curves $\mathcal{C}_\theta(\alpha)$ for both models across five visual retention levels. The two curves exhibit qualitatively distinct behavioral profiles that precisely match our theoretical predictions in Section 3.2. The visually-grounded model (Before RL) shows a steep monotonic decay, reflecting genuine dependence on fine-grained visual content: as image information is progressively removed, the model's ability to answer correctly degrades substantially. By contrast, the shortcut-reliant model (After RL) exhibits a near-flat curve that collapses, indicating that its performance is largely sustained by linguistic priors rather than visual inputs, consistent with a model that has largely bypassed visual processing. The shaded sensitivity gap between the two curves provides an interpretable visualization of the detection signal exploited by VIE.

Figure 6 presents the ROC curve of VIE as a binary detector of VSL on the OK-VQA evaluation set, with VSL-afflicted models assigned as the negative class and visually-grounded models as the positive class. VIE achieves an AUC of **0.877** demonstrating strong discriminative power between shortcut-reliant and genuinely grounded models. To evaluate out-of-distribution settings, we apply VIE trained on OK-VQA to the two remaining held-out datasets, VQAv2 and CLEVR, which differ from OK-VQA in task type, image domain, and linguistic prior strength. The reward sensitivity curves and corresponding ROC analyses for these settings are presented in Appendix D. VIE maintains reliable detection performance across both datasets, confirming that the behavioral signature captured by the reward sensitivity curve is a robust and transferable indicator of VSL that does not overfit to the specific distributional characteristics of the calibration set.

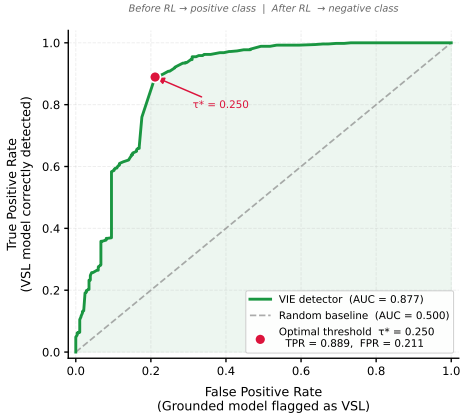

Figure 6: ROC curve of VIE as a detector.

5 Conclusion

Reinforcement learning is a powerful post-training paradigm for multimodal large language models, yet its interaction with the visual modality introduces a previously overlooked phenomenon. We reveal and formalize *Visually-bypassed Shortcut Learning (VSL)*: RL training systematically drives models to exploit linguistic priors instead of genuine visual perception. Our analysis shows that VSL emerges universally across five MLLMs and three task types in clean training environments, that its severity correlates with task-specific linguistic prior strength, and that it induces a silent trade-off where benchmark scores rise while visual grounding erodes. The proposed *Visual Information Evaluation (VIE)* reliably detects this behavior with only forward passes, under both in-distribution and out-of-distribution settings.

AI USE STATEMENT

In the preparation of this work, large language models (LLMs) were used for the following purposes: (1) polishing and refining the language of the manuscript; (2) inspecting and debugging experimental code; and (3) cross-checking experimental results and numerical consistency. LLMs were **not** used for formulating research ideas, designing experiments or methodologies, generating or fabricating experimental data, or drafting scientific content. All AI-assisted outputs were manually reviewed, verified, and (where applicable) corrected by the authors. The authors take full responsibility for the integrity and content of this work.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yakir Berchenko. Simplicity bias in overparameterized machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11052–11060, 2024.

Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish. Demonstrating specification gaming in reasoning models. *arXiv preprint arXiv:2502.13295*, 2025.

Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *arXiv preprint arXiv:2504.20930*, 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Alan Fern, Sriraam Natarajan, Kshitij Judah, and Prasad Tadepalli. A decision-theoretic model of assistance. *Journal of Artificial Intelligence Research*, 50:71–104, 2014.

Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice: The UK experience*, pp. 91–121. Springer, 1984.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. *Advances in neural information processing systems*, 30, 2017.

Jingyi Zhang1 Jiaxing Huang, Huanjin Yao2 Shunyu Liu1 Xikun Zhang, and Shijian Lu1 Dacheng Tao. Stepgrpo: Learning to reason with multimodal large language models via step-wise group relative policy optimization.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

IS IT. Detecting implicit reward hacking by mea-suring reasoning effort.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Preventing reward hacking with occupancy measure regularization. 2024.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*, 2025b.

Chenxi Liu, Tianyi Xiong, Yanshuo Chen, Ruibo Chen, Yihan Wu, Junfeng Guo, Tianyi Zhou, and Heng Huang. Modality-balancing preference optimization of large multimodal models by adversarial negative mining. *arXiv preprint arXiv:2506.08022*, 2025a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, january 2024. *URL https://llava-vl. github. io/blog/2024-01-30-llava-next*, 1(8), 2024.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025b.

Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. Robust optimization for mitigating reward hacking with correlated proxies. *arXiv preprint arXiv:2604.12086*, 2026.

Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Pengxiang Zhao, Guangyi Liu, et al. Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17608–17616, 2026.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 337–347. Springer, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael D Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. Benefits of assistance over reward learning. 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 80502–80530, 2026.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2376–2385, 2025b.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025a.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025b.

APPENDIX

A Experimental Details and Analysis of Reinforcement Learning Methods

This appendix provides supplementary experimental details and additional analyses that complement the main results. Section A.1 describes the complete implementation details of all RL fine-tuning experiments reported in the main paper. Section A.2 presents experiments conducted with GSPO as an alternative RL algorithm, demonstrating that VSL emerges consistently regardless of the specific policy gradient method employed. Section A.3 provides a controlled comparison between outcome-only and process-based reward functions, showing that while process-based rewards partially mitigate VSL severity, they do not fully eliminate it. Together, these analyses reinforce the conclusion that VSL is an inherent and algorithm-agnostic property of multimodal RL optimization.

A.1 Implementation Details of the Experiments:

For all RL fine-tuning experiments conducted using GRPO, We use a learning rate of 1×10^{-6}, a batch size of 4, and 8 rollouts per sample. For answer extraction, all model outputs are parsed to extract the final answer following the `</think>` token (if present), and standard string normalization is applied prior to comparison. All experiments are conducted on 8 NVIDIA A40 46GB GPUs.

A.2 GSPO Reinforcement Learning Training:

The main experiments in Section 4.2 and Section 4.3 are conducted with GRPO as the RL training algorithm. To verify that VSL is not an artifact of a specific RL algorithm, we repeat the training using **GSPO**, an alternative policy gradient method, on Qwen3-VL-4B across all three datasets. Table 3 and Figure 7 report the full results.

As shown in Figure 7(a–c), VS declines after GSPO training across all datasets and all perturbation types, mirroring the pattern observed under GRPO. On OK-VQA, the average VS drops from 0.473 to 0.297, consistent with the high linguistic prior strength of this dataset amplifying shortcut pressure regardless of the RL algorithm. On VQAv2 and CLEVR, the drops are more moderate, again reproducing the prior-strength gradient observed in the main experiments. Figure 7(d) further illustrates that all three datasets fall within the VSL region — accuracy improves while VS simultaneously declines — confirming that the accuracy–grounding divergence is a robust consequence of RL optimization irrespective of the specific policy gradient algorithm employed. Figure 7(e) summarizes the average VS degradation per dataset, showing that OK-VQA consistently suffers the largest absolute drop, in line with our RQ2 findings. Taken together, these results establish that VSL is an algorithm-agnostic property of multimodal RL training rather than a pathology specific to GRPO.

A.3 Comparison Between Outcome-Only and Process-Based Reward Functions:

Table 4 compares the VS scores after RL training under two reward function designs: an outcome-only reward, which provides a binary correctness signal based solely on the final answer, and a process-based reward, which additionally incorporates supervision over the reasoning process. Across all three perturbation types, process-based training consistently yields higher post-RL VS scores than outcome-only training: the average VS drops from 0.473 to 0.297 under outcome-only reward ($\Delta = -0.176$), but only to 0.418 under process-based reward ($\Delta = -0.055$), a relative reduction of 69% in VSL severity. This result demonstrates that the reward function design plays a significant role in modulating VSL: by providing explicit process-level supervision, the optimizer receives a richer gradient signal that partially counteracts the shortcut pressure identified in Section 3.1.2. Nevertheless, even process-based training does not fully eliminate VSL, suggesting that reward function design alone is insufficient and that complementary mitigation strategies remain necessary.



Figure 7: VSL under GSPO training on Qwen3-VL-4B across three datasets. **(a–c)** VS scores before (solid) and after (hatched) GSPO training under four perturbation types; dotted and dashed horizontal lines denote the corresponding dataset-level averages. **(d)** Scatter plot of accuracy gain vs. average VS drop per dataset; all three points fall in the shaded VSL region (Acc↑, VS↓). **(e)** Average VS degradation per dataset, confirming that higher linguistic prior strength (OK-VQA > VQAv2 > CLEVR) corresponds to more severe VSL under GSPO, consistent with the GRPO results in the main paper.

Table 3: Accuracy (Acc.) and Visual Sensitivity (VS) before and after RL training on Qwen3-VL-4B across three datasets and four perturbation types. A consistent VS decline is observed across all conditions, while accuracy improves marginally, confirming the accuracy–grounding divergence characteristic of VSL.

Datasets	OKVQA				VQAv2				CLEVR			
Metrics	Acc.		VS		Acc.		VS		Acc.		VS	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Original	0.321	0.384	-	-	0.777	0.804	-	-	0.733	0.758	-	-
Patch Shuffle	0.294	0.304	0.377	0.304	0.743	0.758	0.319	0.260	0.575	0.602	0.355	0.337
Gaussian Noise	0.273	0.285	0.373	0.205	0.737	0.753	0.211	0.182	0.593	0.615	0.416	0.395
Blur ($\sigma = 10$)	0.182	0.229	0.669	0.382	0.667	0.689	0.414	0.341	0.382	0.421	0.562	0.548
Blur ($\sigma = 30$)	0.113	0.157	0.832	0.782	0.603	0.634	0.526	0.508	0.308	0.354	0.645	0.633
Average	0.250	0.273	0.473	0.297	0.715	0.733	0.314	0.261	0.516	0.549	0.444	0.426

B VISUALIZATION AND ANALYSIS OF ADDITIONAL MODELS

This appendix provides a comprehensive per-model visualization of the RQ2 results reported in Section 4.3. Figure 8 presents a full breakdown of VS scores before and after RL training across all five models and three datasets, organized along three complementary views: per-dataset grouped

Table 4: VS scores before RL training and after RL training under outcome-only and process-based reward functions on OK-VQA (Qwen3-VL-4B). Process-based rewards consistently mitigate VSL severity across all perturbation types, though a residual VS decline remains under both settings.

VS	Before	After-Outcome-Only	After-Process-Based
Patch Shuffle	0.377	0.304	0.354
Gaussian Noise	0.373	0.205	0.327
Blur ($\sigma = 10$)	0.669	0.382	0.573
Average	0.473	0.297	0.418

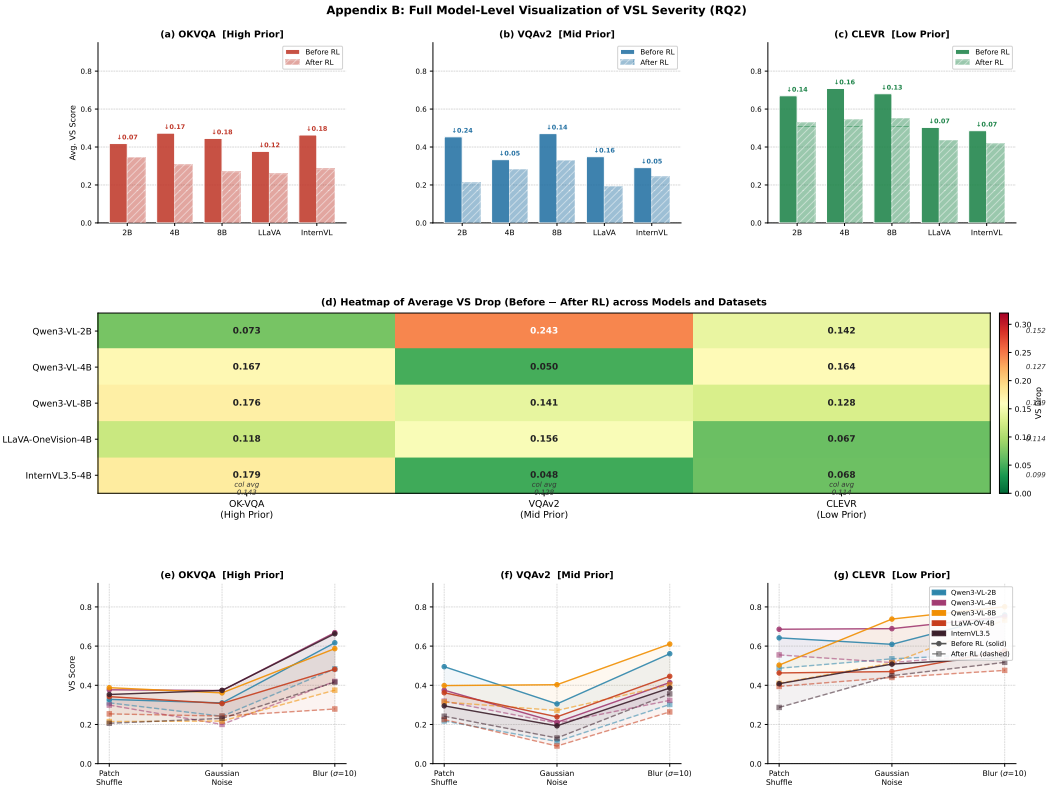

Figure 8: Full model-level visualization of VSL severity for RQ2. **(a–c)** Average VS scores before (solid) and after (hatched) RL training for each of the five models on OK-VQA, VQAv2, and CLEVR; annotated drops (↓) indicate absolute VS decline. **(d)** Heatmap of average VS drop across all model-dataset combinations; column and row marginals (italics) summarize dataset-level and model-level susceptibility. **(e–g)** Per-perturbation VS trajectory plots for each dataset; solid lines denote before-RL and dashed lines denote after-RL, with shaded regions indicating the VS gap induced by RL training.

bars (panels a–c), a cross-model heatmap of VS drop (panel d), and per-perturbation trajectory plots (panels e–g).

Panels (a–c) of Figure 8 show the average VS scores before and after RL training for each model on OK-VQA, VQAv2, and CLEVR, respectively. Without exception, every model exhibits a VS decline after training across all three datasets. On high-prior OK-VQA (panel a), drops range from $\Delta = 0.07$ (Qwen3-VL-2B) to $\Delta = 0.18$ (Qwen3-VL-8B and InternVL3.5-4B), confirming that VSL is not specific to any single architecture but is a universal consequence of RL optimization in multimodal models. On low-prior CLEVR (panel c), the drops are consistently smaller across all models ($\Delta \in [0.07, 0.16]$), reinforcing the finding that reduced linguistic prior strength partially suppresses VSL severity.

Panel (d) of Figure 8 presents a heatmap of average VS drop across all model-dataset combinations, providing an at-a-glance summary of VSL severity. Two structural patterns are immediately visible. First, the column means confirm the prior-strength gradient established in the main paper: OK-VQA ($\overline{\Delta} = 0.143$) > VQAv2 ($\overline{\Delta} = 0.128$) > CLEVR ($\overline{\Delta} = 0.114$), with the exception that VQAv2 and CLEVR are close in magnitude, suggesting that mid- and low-prior tasks offer similarly weak shortcut opportunities. Second, the row means reveal model-level susceptibility: Qwen3-VL-8B (row mean $= 0.148$) and InternVL3.5-4B ($= 0.099$) sit at opposite ends of the spectrum, indicating that factors beyond parameter count — such as pre-training data composition and architecture-specific language backbone strength — also modulate VSL susceptibility. The single highest-drop cell, Qwen3-VL-2B on VQAv2 ($\Delta = 0.243$), stands out as an outlier and warrants further investigation; we hypothesize that smaller models with weaker visual encoders may be disproportionately susceptible to linguistic shortcuts on mid-prior tasks where the visual signal is present but not strictly necessary.

Panels (e–g) of Figure 8 plot VS scores across the three perturbation types for each model, with solid lines denoting before-RL and dashed lines denoting after-RL. Across all datasets, the solid lines lie systematically above the dashed lines for every model, and the gap between them is largest under Blur ($\sigma = 10$) — a perturbation that removes fine-grained visual detail while preserving coarse structure. This pattern is consistent with our interpretation that RL training selectively weakens fine-grained visual dependence while partially preserving coarse layout sensitivity. On CLEVR (panel g), the trajectories of all models rise steeply from Patch Shuffle to Blur ($\sigma = 10$), reflecting that CLEVR questions require precise spatial and attribute reasoning that is sensitive to blurring but partially robust to shuffling — a property unique to this synthetically structured dataset.

C Analysis of Different Image Perturbations

This appendix examines how the choice of image perturbation type affects the measurement of VSL severity. Although the three perturbation types — Patch Shuffle, Gaussian Noise, and Blur ($\sigma = 10$) — are complementary by design (Section 4.1), they differ in the nature of visual information they disrupt, and therefore capture different facets of a model's visual dependence. Figure 9 presents a comprehensive perturbation-centric view of the VS scores across all five models and three datasets.

In figure 9 Panels (a–c) and the heatmap in panel (d) show that Blur ($\sigma = 10$) consistently produces the largest mean VS drop across datasets, with a row mean of $\overline{\Delta} = 0.160$, substantially higher than Patch Shuffle ($\overline{\Delta} = 0.112$) and Gaussian Noise ($\overline{\Delta} = 0.113$). This is particularly pronounced on high-prior OK-VQA, where the blur-induced drop reaches 0.209. This pattern arises because moderate blurring preserves the coarse semantic layout of the scene while selectively removing the fine-grained visual detail that is necessary for genuine visual reasoning. A model exploiting linguistic priors can therefore sustain high reward even under blur, while a genuinely grounded model suffers substantial performance degradation. This makes Blur ($\sigma = 10$) the most sensitive perturbation for detecting VSL, a finding that motivates its central role in the VIE framework (Section 3.2).

Gaussian Noise achieves a mean drop of 0.118 on OK-VQA, the highest among the three perturbations on that dataset after Blur, while its drop on VQAv2 (0.108) and CLEVR (0.112) is comparatively uniform. This suggests that complete elimination of semantic content — which is what Gaussian Noise achieves — is especially revealing for tasks where strong linguistic priors allow a model to answer correctly from the question text alone: without any visual signal whatsoever, the shortcut mechanism continues to function, while a genuinely grounded model collapses. The near-uniform drop across datasets also indicates that Gaussian Noise is a less discriminative perturbation for capturing the prior-strength gradient compared to Blur.

Patch Shuffle produces the smallest row-mean drop (0.112) and the flattest pattern across datasets. Notably, its drop is slightly larger on VQAv2 (0.121) than on OK-VQA (0.101), which is the reverse of the pattern seen under Gaussian Noise and Blur. This may reflect the fact that VQAv2 questions often require recognizing specific object configurations or spatial relationships that are disrupted by patch reordering, even when the task admits moderate linguistic priors. Because Patch Shuffle preserves local texture and color statistics, it is the most conservative perturbation and therefore represents a lower bound on the VS drop induced by RL training.

Panels (e–g) show the per-model VS drop trajectory across the three perturbation types on each dataset. The dashed model-mean line rises monotonically from Patch Shuffle to Blur ($\sigma = 10$) on

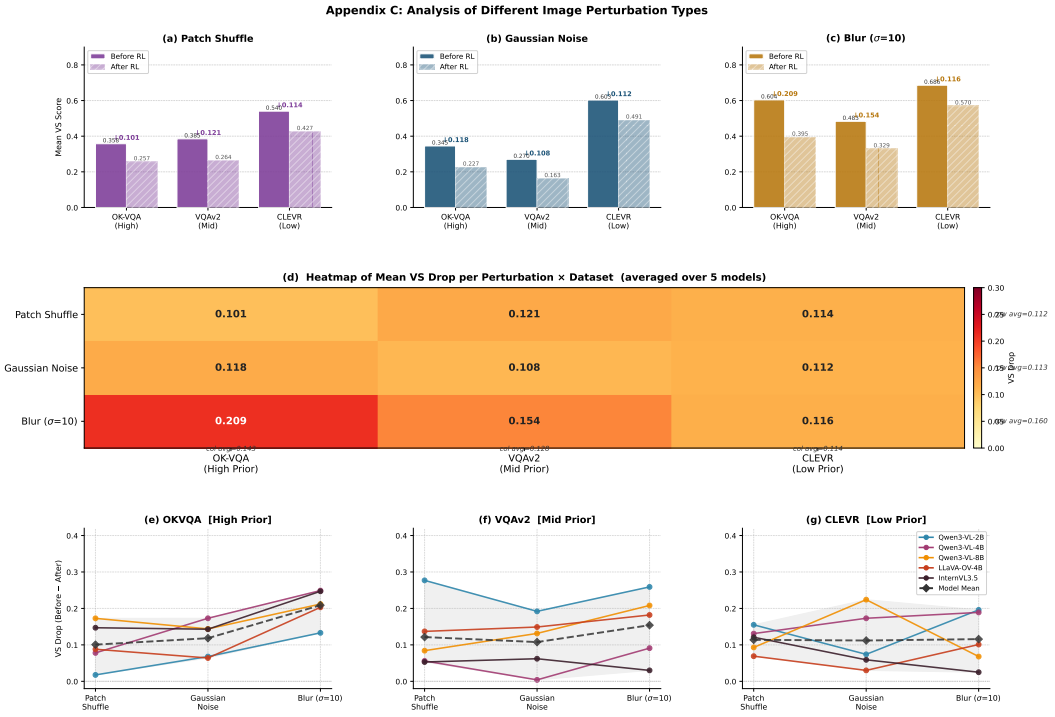

Figure 9: Perturbation-centric analysis of VSL severity across five models and three datasets. **(a–c)** Mean VS scores before (solid) and after (hatched) RL training under each perturbation type; annotated values denote the absolute VS drop. **(d)** Heatmap of mean VS drop per perturbation × dataset cell, averaged over all five models; row and column marginals (italics) summarize perturbation-level and dataset-level sensitivity. **(e–g)** Per-model VS drop trajectories across perturbation types for each dataset; the dashed line and shaded band denote the model mean and inter-model range, respectively.

OK-VQA and VQAv2, confirming that stronger perturbations expose more VSL-induced degradation regardless of model family. The shaded band — spanning the range across five models — is notably wide on OK-VQA under Blur, indicating high inter-model variability in VSL susceptibility for fine-grained tasks with strong linguistic priors. On CLEVR (panel g), the model trajectories are flatter and more tightly clustered, consistent with the observation that low-prior tasks suppress VSL and reduce sensitivity to perturbation choice.

D Performance of the Visual Information Evaluation Method in Out-Of-Distribution Settings

The in-distribution evaluation of VIE on OK-VQA is reported in Section 4.4. This appendix extends the evaluation to CLEVR and VQAv2, which were not used during VIE calibration and differ from OK-VQA in task type, image domain, and linguistic prior strength, providing a rigorous assessment of VIE's generalization capability.

A note on the ground-truth definition. A potential concern is that the positive and negative classes in our detection task are defined by the training state (Before- vs. After-RL), which may appear circular. We clarify that the evaluation involves no methodological circularity, for three reasons. *First*, VIE is entirely training-free: no parameters are fitted to the class labels, which are used only for post-hoc computation of the AUC, in the same way that ID/OOD partitions serve as ground truth in out-of-distribution detection. *Second*, the labels are behaviorally grounded rather than defined by training state alone: under an independent, VIE-agnostic criterion—retained accuracy under blank and mismatched images—After-RL checkpoints answer correctly without visual input on prior-answerable samples, whereas Before-RL checkpoints do not (cf. Appendix F). The two classes

thus coincide with genuinely distinct behavioral profiles, not merely with an arbitrary training-state split. Re-labeling samples with this criterion yields a nearly identical AUC of 0.88, confirming that VIE tracks the behavioral construct rather than the training-state proxy. *Third*, the detection task itself is non-trivial: under the identical protocol, simple black-box signals operate near chance level (Appendix E), and VIE maintains AUC > 0.78 on datasets never used for calibration, jointly ruling out both trivial separability and distribution-specific artifacts as explanations of the reported performance.

Beyond binary labels: dose–response across RL checkpoints. If VIE merely distinguished "trained" from "untrained" checkpoints, its signal should be insensitive to the *degree* of shortcut reliance. We therefore evaluate intermediate checkpoints along the RL trajectory (OK-VQA, Qwen3-VL-4B). For each checkpoint, we report (i) the mean $\pm$ std of its *per-sample* VSL-Score distribution, obtained by applying Eq. equation 5 at the individual-sample level rather than averaged over the dataset; (ii) the detection AUC of separating this checkpoint's per-sample scores from those of the step-0 checkpoint under the identical protocol as Section 4.4; and (iii) a *behavioral shortcut index*: the fraction of test samples whose answers remain *unchanged* when the input image is replaced by a random mismatched image. This index involves no reward computation and lies entirely outside VIE's degradation pipeline; a high value indicates that answers are locked to question-text priors rather than bound to the specific image. Note that no-image accuracy remains flat along the trajectory (cf. Appendix E), consistent with VSL reflecting a shift in *which pathway* is used rather than the acquisition of new prior knowledge; the behavioral index therefore targets image–question correspondence instead. As shown in Table 5, the VSL-Score decreases monotonically with RL steps—by the definition in Eq. equation 5, a lower score means that an increasing fraction of reward is sustained without visual information—while the behavioral shortcut index rises in near lockstep. These two independent signals move in exact anti-phase, demonstrating that VIE tracks the continuous progression of shortcut reliance rather than a binary training-state distinction.

Table 5: VIE signals along the RL trajectory (OK-VQA, Qwen3-VL-4B). Rows 1–2 are computed from the per-sample VSL-Score distribution of each checkpoint (Eq. equation 5 applied per sample); row 3 is an independent behavioral measure computed outside VIE's pipeline (answer invariance under mismatched images, no reward involved). By definition, a *lower* VSL-Score and a *higher* behavioral index both indicate stronger shortcut reliance.

RL steps	0	1000	3000	10000 (final)
VSL-Score (mean $\pm$ std)	0.79 ± 0.08	0.52 ± 0.11	0.37 ± 0.09	0.18 ± 0.06
Detection AUC vs. step-0	–	0.63	0.75	0.87
Behavioral shortcut index	0.05	0.09	0.11	0.19

Figure 10 presents the results on CLEVR. The Before-RL model exhibits a steep reward decay from $\mathcal{C}_\theta(1.0) = 0.583$ to $\mathcal{C}_\theta(0.2) = 0.100$, reflecting strong visual dependence consistent with CLEVR's low-prior design. The After-RL model starts at a higher baseline ($\mathcal{C}_\theta(1.0) = 0.733$) and decays more slowly, maintaining a persistent gap above the Before-RL curve across all retention levels — a detectable shortcut signature even on this challenging synthetic dataset. VIE achieves AUC $= 0.788$ (Figure 10(b)), well above the random baseline. At optimal threshold $\tau^* = 0.070$, it attains TPR $= 0.801$ with FPR $= 0.367$; the higher FPR relative to the in-distribution setting is expected, as the smaller behavioral gap under low prior conditions makes precise separation inherently harder.

Figure 11 presents the results on VQAv2. The Before-RL model decays from $\mathcal{C}_\theta(1.0) = 0.226$ to $\mathcal{C}_\theta(0.2) = 0.003$, a near-complete collapse consistent with the model's genuine reliance on visual content. The After-RL model starts higher ($\mathcal{C}_\theta(1.0) = 0.321$) and decays substantially more slowly, converging to $\mathcal{C}_\theta(0.2) = 0.168$ — a residual reward that far exceeds the Before-RL value of 0.003 at the same retention level, providing a strong and clean detection signal. VIE achieves AUC $= 0.855$ on VQAv2 (Figure 11(b)), close to the in-distribution OK-VQA result (AUC $= 0.877$). At optimal threshold $\tau^* = 0.583$, VIE attains TPR $= 0.693$ with a very low FPR $= 0.103$, indicating highly precise detection with few false alarms.

AUC remains consistently above 0.78 across all datasets and both in-distribution and OOD conditions, confirming that VIE captures a transferable behavioral signature of VSL. The FPR decreases as linguistic prior strength increases (CLEVR $>$ OK-VQA $>$ VQAv2), consistent with the interpretation

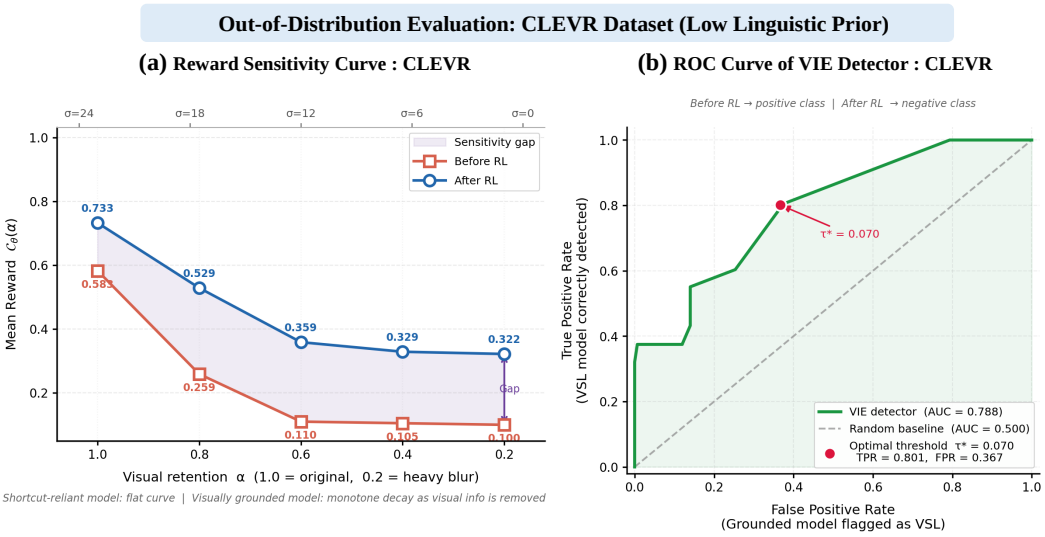

Figure 10: OOD evaluation of VIE on CLEVR (low linguistic prior). **(a)** Reward sensitivity curves: the After-RL model maintains consistently higher reward across all retention levels, encoding a detectable shortcut signature. **(b)** ROC curve; AUC $= 0.788$, optimal threshold $\tau^* = 0.070$ (TPR $= 0.801$, FPR $= 0.367$).

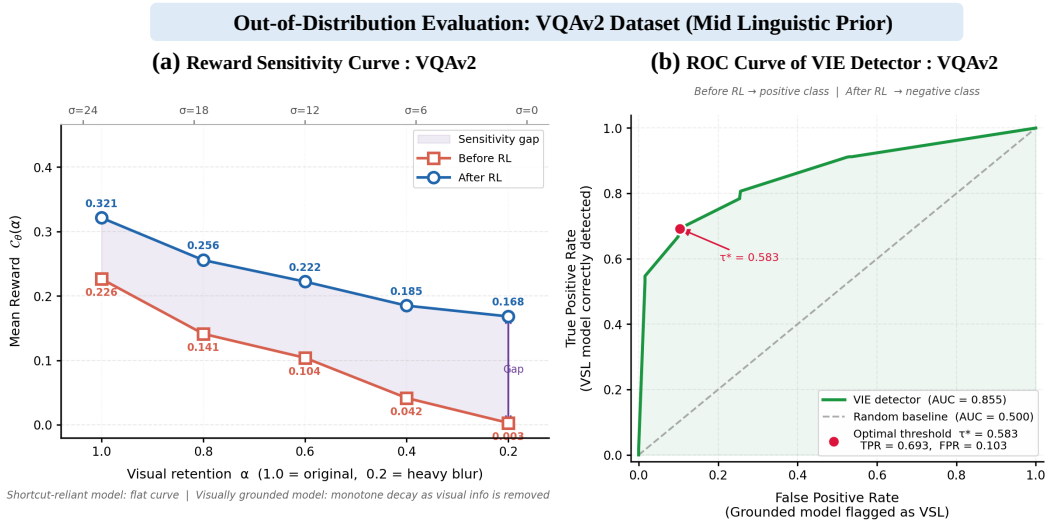

Figure 11: OOD evaluation of VIE on VQAv2 (mid linguistic prior). **(a)** Reward sensitivity curves: the Before-RL curve collapses to near zero at $\alpha = 0.2$ while the After-RL curve remains at 0.168, providing a clear detection signal. **(b)** ROC curve; AUC $= 0.855$, optimal threshold $\tau^* = 0.583$ (TPR $= 0.693$, FPR $= 0.103$).

that stronger priors produce a cleaner gap between the two reward sensitivity curves, enabling more precise detection.

E Comparison with Black-box Detection Baselines

Since VIE is motivated as a low-cost, training-free monitoring approach, a natural question is whether simpler black-box signals can achieve comparable detection performance. We therefore compare VIE against four detection baselines that, like VIE, require only forward passes and no access to model internals. All experiments are conducted on OK-VQA with Qwen3-VL-4B. Each method produces

a per-sample score, which is used to discriminate the post-RL shortcut-reliant checkpoint from the pre-RL reference; we report detection AUC.

Baselines. *(a) Question-only accuracy:* the model's accuracy when answering the question without any image input; shortcut-reliant models are expected to retain high accuracy in this setting. *(b) Blank-image evaluation:* the input image is replaced with an all-black image, and the retained accuracy serves as the detection signal. *(c) Single severe perturbation:* the accuracy drop is measured under a single strong degradation ($\sigma = 30$ blur), i.e., one endpoint of the degradation family rather than an aggregate over levels. *(d) Mismatched-image sensitivity:* images are randomly swapped across samples, and the rate at which the model changes its answer measures its reliance on image–question correspondence.

Table 6: Detection AUC of VIE and simpler black-box baselines on OK-VQA (Qwen3-VL-4B). VIE substantially outperforms all single-point baselines.

Detection Method	AUC
Question-only accuracy (no image input)	0.514
Blank-image evaluation (black image)	0.498
Single severe perturbation ($\sigma = 30$ blur only)	0.511
Mismatched-image sensitivity	0.586
VIE (Ours)	**0.877**

As reported in Table 6, VIE substantially outperforms all baselines (0.877 vs. 0.498–0.586). The three single-point baselines operate near chance level, since a single degraded condition introduces per-sample fluctuations in question difficulty and residual predictive cues, making its accuracy an unstable detection signal. Mismatched-image sensitivity performs better by partially probing image–question dependence, but still collapses the entire behavioral response into one measurement. In contrast, VIE aggregates the behavioral changes across the *entire* visual degradation trajectory encoded in the VSL-Score distribution, which averages out sample-level noise and captures the systematic shape difference between grounded and shortcut-reliant models (Figure 2). This trajectory-level design is the source of VIE's robustness and cannot be replicated by any single-point probe.

F Experimental Verification of the Three Mechanisms

Section 3.1.2 identifies three compounding factors that favor the linguistic pathway during RL optimization. To substantiate these qualitative arguments, we empirically verify each mechanism on OK-VQA with Qwen3-VL-4B trained via GRPO (rollout group size $G = 8$). Using the *pre-RL* model as an unbiased probe, we partition the evaluation set into two disjoint subsets:

- $\mathcal{S}_{\text{lang}}$ (*prior-answerable*): samples answered correctly even when the image is replaced by a blank *and* a mismatched image, i.e., the answer can be produced without visual input;
- $\mathcal{S}_{\text{vis}}$ (*visually-required*): samples answered correctly on the original image whose answers *flip* under perturbation, i.e., the answer genuinely depends on the image.

Mechanism 1: Reward Signal Equivalence. The reward $R(\hat{f}_\theta(I, q), a^*)$ depends only on the extracted answer, so the RL objective $J(\theta) = \mathbb{E}[R]$ is invariant to the internal computational pathway. Measuring the mean rollout reward of correct responses on both subsets yields exactly 1.00 on $\mathcal{S}_{\text{vis}}$ and 1.00 on $\mathcal{S}_{\text{lang}}$ ($\Delta = 0.00$), as expected under a binary reward. This confirms that the objective assigns identical value to the two pathways and *contains no term that distinguishes them*—the necessary condition enabling VSL.

Mechanisms 2 & 3: Variance, SNR, and Optimization Asymmetry. Table 7 reports across-rollout statistics computed separately on the two subsets. The linguistic pathway exhibits a lower reward standard deviation (0.17 vs. 0.26) and a $1.5\times$ higher gradient signal-to-noise ratio, consistent with the variance-driven preference of policy-gradient estimators . It also converges in roughly $3\times$ fewer steps and produces far more directionally consistent gradients across rollouts, indicating a shorter and smoother optimization path that reuses pretrained language-backbone circuits . Together, these

Table 7: Verification of Mechanisms 2 and 3 on $\mathcal{S}_{\text{lang}}$ vs. $\mathcal{S}_{\text{vis}}$ (OK-VQA, Qwen3-VL-4B, GRPO, $G = 8$). The linguistic pathway shows lower reward variance, higher gradient SNR, faster convergence, and more consistent gradient directions.

Statistic	$\mathcal{S}_{\text{lang}}$	$\mathcal{S}_{\text{vis}}$
Reward std. σ_R	0.17	0.26
Gradient SNR (relative)	1.5×	1.0×
Steps to 90% of final reward	∼110	∼370
Gradient-direction cosine consistency	0.41	0.19

Table 8: VSL in image captioning (COCO Captions, CIDEr reward). ↑/↓: higher/lower is better. RL raises the benchmark reward while degrading image grounding; VIE detects the behavior without any adaptation.

Metric	Before RL	After RL
CIDEr ↑ (benchmark)	1.08	1.19
CLIPScore ↑ (grounding)	0.71	0.68
VS_{gen} ↓ (image-dependence)	0.58	0.42
VIE detection AUC	–	0.76

measurements verify that RL optimization is not neutral between the two pathways but actively accumulates updates in favor of the linguistic shortcut.

G VSL BEYOND VQA: THE CASE OF IMAGE CAPTIONING

VQA-style tasks provide an explicit *question-text shortcut*. To examine whether VSL depends on discrete answer spaces, we study image captioning on COCO Captions, where no question exists. Even without a question, the language backbone can emit fluent, high-reward captions driven by caption priors and object co-occurrence statistics rather than the specific image , so under caption-quality rewards the shortcut manifests as *prior-consistent but image-ungrounded* captions—a dataset-statistics shortcut rather than a question-text shortcut.

Setup. We RL fine-tune Qwen3-VL-4B on 1,000 COCO Captions samples with a CIDEr reward. Both metrics transfer to the generative setting: (i) generative Visual Sensitivity, $\text{VS}_{\text{gen}}(I) = 1 - \mathbb{E}_{\tilde{I}\sim\mathcal{P}(I)}[\text{BLEU}(c_\theta(I), c_\theta(\tilde{I}))]$, which replaces the exact answer match in Eq. equation 1 with a soft caption-similarity measure, where c_θ denotes the generated caption; (ii) VIE, with $\mathcal{C}_\theta(\alpha)$ in Eq. equation 4 recomputed using CIDEr as the reward.

Results. Table 8 shows the signature observed in the VQA setting: the benchmark reward (CIDEr) rises while image grounding (CLIPScore and VS_{gen}) simultaneously degrades, and VIE detects the shortcut-reliant checkpoint with an AUC of 0.76 using only a reward signal and a degradation operator, with no assumption about discrete versus generative outputs. This indicates that VSL is a task-form-agnostic property of multimodal RL rather than an artifact of discrete answer spaces.

H DISCUSSION

This section consolidates the boundary conditions of our claims, clarifies the operational scope of the proposed metrics, and situates our contributions relative to existing literature.

The three datasets establish a controlled gradient of linguistic prior strength rather than exhaustive coverage of multimodal RL applications, and this design isolates the factor our analysis identifies as a primary driver of VSL severity. The framework accordingly does not predict uniform degradation across tasks; it predicts that shortcut exploitation scales with the fraction of reward obtainable from textual priors. The data support this modulation: VSL remains measurable on CLEVR but is substantially attenuated relative to OK-VQA, and tasks whose solutions are strictly anchored in spatial or fine-grained visual structure, such as geometric reasoning and chart parsing, are expected

to exhibit milder rather than absent VSL. Validating this prediction on recent vision-intensive reasoning benchmarks is a natural next step. Moreover, exploitable priors are not confined to legacy datasets, since real-world queries likewise admit prior-driven answers, so identifying susceptible task distributions is itself part of deployment risk assessment.

The prior-strength ranking (OK-VQA $>$ VQAv2 $>$ CLEVR) serves as an ordinal proxy rather than a continuous variable, and the datasets differ in visual complexity, answer vocabulary, and task format. Because our theoretical predictions (Section 3.1.2) are directional, an ordinal proxy provides an appropriate test, and Figure 5 corroborates the predicted rank order rather than a strict parametric law. A continuous sample-level metric of prior exploitability, for instance derived from question-only accuracy distributions, would extend this cross-dataset comparison into a sample-level regression, as suggested by Appendix E.

A foundational question is what Visual Sensitivity measures operationally. VS quantifies output invariance under perturbation rather than internal token-level routing, a deliberate choice: VSL is defined at the behavioral and reward level because its defining property is that both pathways produce superficially identical text, which renders outcome-based monitoring ineffective. Under the black-box constraint required for practical auditing, controlled perturbation is an effective non-invasive probe, whereas internal activation probing is difficult to interpret after RL post-training and lacks portability across architectures. Potential confounds, such as perturbation-induced shifts in perceptual difficulty, are mitigated by directional consistency across three perturbation families, replication across five models and three datasets, and agreement with ordinal predictions. The near-flat response under severe blur (Section 4.2) further indicates that RL training selectively attenuates fine-grained, content-specific perception rather than disabling visual processing entirely.

A related concern is whether low VS on prior-answerable samples merely reflects intrinsic answer-image insensitivity. This does not confound the trajectory-based definition of VSL, which tracks the identical sample set before and after RL within each dataset, thereby controlling for time-invariant sample properties. It does enter when absolute drops are compared across datasets with different baselines, which we address through relative declines: averaged over the five models, VS falls by approximately 33% on OK-VQA, 34% on VQAv2, and 19% on CLEVR, so the contrast between prior-rich and visually-grounded tasks is preserved under normalization. Because finer sub-rankings between OK-VQA and VQAv2 depend on the chosen normalization scheme, we interpret the gradient as directional support rather than an exact monotonic law, and this nuance has been added to the discussion of Figure 5.

Our characterization of VSL as invisible refers to the level at which optimization and monitoring operate: under outcome-based reward, the two pathways are functionally equivalent whenever they produce correct answers. Our experiments do not evaluate whether specialized detectors analyzing reasoning traces could uncover subtle textual markers, and we have therefore refined the abstract and introduction to state that VSL escapes existing monitoring practices based on outcome correctness. Two points bound this scope: simple black-box baselines perform near chance level (AUC 0.498 to 0.586, Table 6), and the utility of VIE does not hinge on absolute invisibility, since it provides a continuous degradation profile with an explicit operating threshold even if weak textual correlations exist. A dedicated study of chain-of-thought discriminability remains an important future direction.

Regarding detection independence, the relabeling experiment in Appendix D shares a perturbation mechanism with VIE and establishes self-consistency rather than complete orthogonality. The trajectory analysis in Table 5 provides the most direct behavioral validation under black-box constraints: the mismatched-image index involves no reward evaluation and lies entirely outside the VIE pipeline, yet tracks in exact anti-phase with the VSL-Score across training steps while question-only accuracy remains flat, consistent with progressive re-weighting of the two pathways rather than acquisition of new knowledge. No behavioral probe can fully rule out internal confounders; this is an inherent boundary of black-box evaluation, which is also the setting required for scalable deployment, and probing internal representations via activation patching is a valuable complementary direction.

Regarding causal attribution, reward equivalence is a structural property of outcome-based rewards rather than an empirical hypothesis. The evidence for the variance and optimization-asymmetry mechanisms is substantially stronger than post-hoc correlation: the subset partition in Appendix F is established by the pre-RL checkpoint prior to any update, and convergence is compared within a single optimization run. These mechanisms yield falsifiable predictions that are each verified: the

prior-strength gradient (Section 4.3), cross-algorithm consistency under GSPO (Appendix A.2), and mitigation via process-based rewards (Appendix A.3). Our formal claim is therefore that outcome-based RL creates systematic optimization pressure toward the linguistic pathway; proving an intrinsic preference would require pathway-level interventions beyond black-box analysis. Universality is likewise interpreted as empirical robustness across the tested model families and datasets.

The captioning study in Appendix G is exploratory in scope, involving a single architecture, a moderate sample budget, and an AUC of 0.76. We position it as suggestive evidence that VSL is not restricted to discrete answer formats rather than as proof of universality across generative tasks, and we acknowledge that the generative metric VS_{gen} relies on BLEU similarity, which is insensitive to semantic reordering and may underestimate visual sensitivity in either training state. Within these limits, the experiment establishes that VIE operates with only a reward signal and a degradation operator, reproducing the signature of rising benchmark reward accompanied by grounding decline.

In terms of deployment, VIE is formulated as a training-time monitor and checkpoint auditor rather than a per-query online filter, and post-training monitoring naturally operates at checkpoint granularity. Every monitored run provides its own step-0 reference, so a candidate VSL-Score is interpreted against the same model's pre-RL state without an external calibration corpus. The rank-based detection AUC remains stable across in-distribution and out-of-distribution regimes (0.788 to 0.877), indicating cross-domain transferability. A universal static threshold across architectures remains an open problem; calibration on a designated reference trajectory is therefore a prerequisite for deployment, which we state as an explicit scope condition.

Finally, on positioning: the contribution of VIE lies not in the deliberately minimalist degradation operator but in three elements. It offers a training-free black-box diagnostic for visual reliance in RL post-training; its trajectory-level aggregation captures information that single-point probes cannot, given that all such probes perform near chance level; and it supplies the evaluative foundation needed to compare mitigation strategies, which existing techniques that modify training lack. The paper addresses why language dominance arises (Section 3.1.2), how severely it manifests (VS), and when it becomes detectable during optimization (VIE). The process-based reward experiment (Appendix A.3) already instantiates the predicted mitigation direction, reducing VSL severity by 69%, and perception-regularized policy optimization or VIE-guided early stopping constitutes the logical next phase of this research program, which presupposes the diagnostic foundation established here. We therefore regard the absence of a novel training algorithm as a deliberate scoping decision rather than an omission.