Sensing Which Modality Matters: Evidence-Gated Regularization for Robust VLA Policies

Yue Yang[1] **Diego Romeres**[2] **Chiori Hori**[2] **Gedas Bertasius**[1]
Daniel Szafir[1] **Siddarth Jain**[2]

[1]Department of Computer Science, University of North Carolina at Chapel Hill
[2]Mitsubishi Electric Research Laboratories (MERL)

yygx@cs.unc.edu | Project website

Abstract: Vision-Language-Action (VLA) policies fuse multimodal sensory inputs, but training on limited and homogeneous robot demonstrations encourages spurious inter-sensor correlations rather than task-relevant signal, a failure we term modality entanglement. Under real-world occlusions and distractors, this manifests as nuisance sensitivity to corruption of uninformative sensors and single-modality insufficiency when only one informative sensor remains intact. We propose **Evidence-Gated Regularization (EGR)**, a modality-agnostic training objective that introduces zero inference-time overhead. EGR derives a per-frame and per-sensor task-relevance signal to gate two state-conditional consistency objectives: invariance on low-evidence sensors, and single-sensor sufficiency on high-evidence ones. We introduce a benchmark based on BEHAVIOR-1K, comprising a fast inference-only diagnostic suite and 47 rollout-based skills targeting modality entanglement. We validate EGR on this benchmark and on two real-robot setups with fundamentally different embodiments: a bi-manual setup with two Kinova arms and three RGB cameras, and a single-arm MELFA ASSISTA setup combining vision and GelSight tactile sensors. EGR improves simulation success rates (SR) from **12.5% to 16.4%** under full modalities (+31%), from **9.4% to 16.5%** under uninformative-sensor corruption (+75%), and from **2.8% to 6.1%** under single-sensor fallback (+120%). Under physical-object distractors, EGR boosts SR from **30% to 85%** on the bi-manual setup (+183%) and from **55% to 70%** on the tactile setup (+27%).

Keywords: Multimodal Robustness, Vision-Language-Action Models, Evidence-Gated Regularization

1 Introduction

Robot learning has made substantial progress, enabling embodied agents to perform an increasingly broad range of complex tasks [1, 2, 3]. Vision-Language-Action (VLA) policies have recently emerged as a leading paradigm for general-purpose robot learning. These policies typically build on large-model backbones and rely on early multimodal fusion to integrate visual, linguistic, and action-relevant information [4, 5, 6]. However, early fusion can encourage shortcut learning, causing models to exploit spurious cross-modal correlations rather than task-relevant signals [7, 8]. VLA policies inherit this vulnerability, and the problem is amplified for robot demonstration data. Compared with the web-scale corpora used to train multimodal foundation models, robot demonstrations are typically far smaller and less diverse in operators, environments, and objects [9, 10]. We refer to the resulting failure mode as *modality entanglement*: a phenomenon in which VLA policies develop spurious dependencies on sensory modalities that are not task-relevant at the current state.

We illustrate modality entanglement effect through two representative, state-dependent failure modes (Fig. 1). As a first example, consider navigation, where the head camera provides the task-relevant information while the two downward-facing wrist cameras observe only the floor (Fig. 1(a)). At deployment, adding physical distractors to the wrist-camera views alone can stall the policy, even though the head-camera observation remains unchanged (Fig. 1(b)). This suggests that the policy

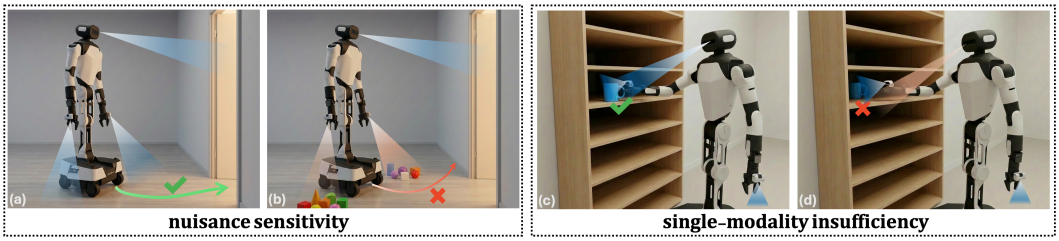

Figure 1: Two failure modes of modality entanglement. **(a-b) Nuisance sensitivity:** distractors in a task-irrelevant wrist camera break navigation despite an unchanged head camera. **(c-d) Single-modality insufficiency:** when a deeper-placed target makes the shelf board occlude the head camera, the policy stops despite an intact wrist view.

has spuriously associated the specific floor appearance in the wrist views with being on the correct path. We refer to this behavior as *nuisance sensitivity*: degradation caused by perturbations to sensors that, at the current state, contain no task-relevant signal. As a second example, consider reaching into a shelf, where both the head camera and the right wrist camera initially observe the target (Fig. 1(c)). At deployment, the target is placed deeper inside the shelf cell, causing the upper board to occlude the head-camera view while the right wrist camera continues to see the target. Nevertheless, the policy stops or behaves erratically (Fig. 1(d)). This suggests that the policy has not learned to treat the wrist camera alone as sufficient evidence for continuing the task. We refer to this behavior as *single-modality insufficiency*: when multiple sensors carry task-relevant signal but all are corrupted except one, the policy fails to fall back to that remaining sensor.

At its core, modality entanglement reflects a missing capability: the policy should be able to decide, at each frame, which sensors provide task-relevant evidence. Existing approaches differ in the signals they use, or fail to use, to guide this decision. Random modality dropout [11] and adversarial multi-sensor consistency [12] target generic robustness to sensor availability by treating sensors and timesteps uniformly. As a result, they improve tolerance to missing or perturbed inputs but do not explicitly perform state-dependent modality selection. Attention-based methods allow the policy to learn such selection end-to-end, but they rely on the same limited and homogeneous demonstration data that gives rise to modality entanglement in the first place. Consequently, the learned attention can inherit the same spurious co-occurrences as the underlying policy [13]. TacVLA [14] instead introduces an explicit architectural gate, activating tactile tokens only when contact is detected. However, this gate is specialized to a single modality and cue: tactile sensing and binary contact. Extending this idea to arbitrary sensor types would require redesigning the gating mechanism itself. What remains missing is a modality-selection method that is explicit, grounded in task structure rather than fragile demonstration statistics, and decoupled from any particular sensor modality.

We propose **Evidence-Gated Regularization (EGR)**, a training objective that satisfies these three desiderata. For each training frame, EGR assigns an evidence score that measures task-grounded signal from each sensor. Rather than learning this score end-to-end from demonstrations, we derive it from properties of the task. The score then gates two complementary consistency objectives: invariance to perturbations of low-evidence sensors, which encourages the policy to ignore uninformative inputs, and single-sensor sufficiency for high-evidence sensors, which encourages the policy to act on an informative signal. In this way, the evidence score serves as a soft inductive bias, allowing the policy to exploit additional useful structure. A key feature of EGR is its separation between modality-specific evidence construction and modality-agnostic regularization. The former defines task-grounded signal per sensor; the latter applies uniform regularization once that signal is computed. This design allows EGR to be applied consistently across heterogeneous sensor modalities while leaving the base VLA architecture unchanged. Because EGR is imposed only during training, it incurs no inference-time overhead. In summary, we contribute the following:

(1) We formalize modality entanglement in VLA policies and characterize it through two state-dependent failure modes—nuisance sensitivity and single-modality insufficiency.

(2) We introduce Evidence-Gated Regularization (EGR), a modality-agnostic training objective that leverages per-frame, per-sensor task-relevance evidence to gate two state-conditional consistency objectives, while adding zero inference-time overhead.

(3) We construct a BEHAVIOR-1K benchmark for modality entanglement, with a fast inference-only diagnostic suite and 47 rollout-based skills. On this benchmark, EGR improves success rates from **12.5% to 16.4%** under full modalities (+31%), from **9.4% to 16.5%** under uninformative-sensor corruption (+75%), and from **2.8% to 6.1%** under single-sensor fallback (+120%).

(4) We validate EGR on two real-robot setups with fundamentally different embodiments: a bi-manual platform with two Kinova arms and three RGB cameras, and a single-arm MELFA ASSISTA platform with one RGB camera and two GelSight tactile sensors. Under physical-object distractors, EGR boosts SR from **30% to 85%** on the bi-manual setup (+183%) and from **55% to 70%** on the tactile setup (+27%).

2 Related Work

2.1 Shortcut Learning in Multimodal LLMs

Current Vision-Language-Action (VLA) policies and broader multimodal large language models (MLLMs) [15, 5, 6, 14, 16] combine large pretrained backbones with multimodal sensory input. Their generalization to unseen conditions depends critically on whether the model learns task-relevant signals or spurious cross-modal correlations. Geirhos et al. [7] characterized this as shortcut learning, where models rely on decision rules that work on the training distribution but fail under distribution shift. Dancette et al. [8] introduced VQA-CE to expose multimodal shortcuts involving question-image co-occurrence. Recent work characterizes similar effects at MLLM scale [17, 18, 19] and in VLA models [20, 21]. The phenomenon we characterize, where a VLA binds its policy to sensors that are uninformative at the current state, is a distinct form of inter-sensor shortcut that this prior line has not isolated.

2.2 Modality Robustness in Multi-Sensor Policies

Beyond robotics, multimodal learning has long observed that jointly training on multiple modalities can leave some under-utilized or cause over-reliance on the dominant one [22, 23, 24], motivating balancing and missing-modality-robust techniques [25, 26]. In robot manipulation, recent work weights modalities per stage or state via learned attention [27]. In VLA models, recent methods improve robustness through run-time distractor inpainting [28], attention-pathway restructuring [29], supervised attention specialization [30], and training on diverse visual clutters [31]. These methods rely on architectural changes, run-time intervention, or end-to-end learned attention, none grounding modality selection in per-frame task structure.

3 Methodology

EGR decouples per-frame sensor relevance from its effect on the policy. It first estimates how task-relevant each sensor is at each frame through a per-sensor evidence score (Sec. 3.2), then uses those scores uniformly across the regularization objectives that shape policy learning (Sec. 3.3). This separation allows the same framework to handle qualitatively different sensors while keeping inference cost identical to that of the base policy.

3.1 Problem Setup

A VLA policy $\pi_\theta(a_t \mid o_t)$ maps a multi-sensor observation $o_t = \{o_t^m\}_{m \in \mathcal{M}}$ to an action distribution, where $\mathcal{M}$ is the set of sensor modalities. The policy is trained via imitation learning on demonstration data $\mathcal{D}$. We use o_t^m to denote the observation in which only modality m is retained and all other modalities are corrupted. We say that π_θ exhibits **modality entanglement** at (t, m) when its behavioral dependence on modality m is misaligned with the modality's actual importance for task completion. This misalignment can arise in two complementary ways. First, under **nuisance sensitivity**, modality m at frame t is task-irrelevant, perturbing it does not change the task success rate, yet perturbing m changes the policy's action distribution. Second, under **single-modality insufficiency**, modality m at frame t is sufficient on its own but other task-relevant modalities coexist, executing π_θ using only m preserves task success rate, yet the policy's action distribution differs when only m is preserved compared to when all sensors are available.

3.2 Evidence as a Task-Relevance Indicator

The evidence score measures how much task-relevant signal a given sensor carries at each frame. For cameras, evidence reflects the visibility of task-relevant geometry; for tactile sensors, it reflects task-relevant contact patterns. More generally, evidence may be computed from any task-relevant signal available during training, whether obtained from task structure or learned representations.

Formally, for each frame t and sensor $m \in \mathcal{M}$, we assign an evidence score $E_{t,m} \in [0, 1]$. Scores are normalized across modalities at each frame, so they capture relative rather than absolute importance. We treat $E_{t,m}$ as a heuristic measure of task relevance, rather than ground truth. It indicates which failure mode (Sec. 3.1) the current frame is most susceptible to, thereby enabling selective regularization. The following two subsubsections instantiate this score for vision and tactile sensing.

3.2.1 Vision Instantiation

Vision evidence measures the amount of task-relevant geometry visible to each camera. Intuitively, a camera provides stronger evidence when focal or interaction objects occupy a larger visible area in its view. We assume access to per-frame instance segmentation, together with task annotations that identify focal objects, i.e., the objects being manipulated, and interaction objects, i.e., the receptacles or targets with which the focal objects interact as specified by the task. For example, in a "put coke can into trash bin" task, the coke can is the focal object and the trash bin is the interaction object. For camera m at frame t, let $A^{\text{focal}}_{t,m,k} \in [0, 1]$ denote the pixel-area ratio of focal object k and $A^{\text{inter}}_{t,m,k} \in [0, 1]$ the area ratio of interaction object k. We aggregate these object-level quantities into per-frame focal and interaction visibility scores, $a^{\text{focal}}_{t,m} = \sum_k A^{\text{focal}}_{t,m,k}$ and $a^{\text{inter}}_{t,m} = \sum_k A^{\text{inter}}_{t,m,k}$.

Using aggregated area alone can be misleading: large interaction objects, such as tables, may dominate the evidence in cameras that do not observe the focal object, thereby falsely inflating their scores. To prevent this, we include interaction objects only when the focal object is visible in the same camera, using an interaction gate $g_{t,m} = \mathbb{1}[a^{\text{focal}}_{t,m} > \epsilon_{\text{focal}}]$. The per-camera interaction score combines the focal contribution with the gated interaction contribution, $s_{t,m} = a^{\text{focal}}_{t,m} + \alpha \cdot g_{t,m} \cdot a^{\text{inter}}_{t,m}$, where the hyperparameter α controls the influence of interaction objects.

Cross-camera normalization yields the final evidence $E_{t,m} = s_{t,m}/(\max_{m'} s_{t,m'} + \varepsilon)$, and a global gate $G_t = \mathbb{1}[\max_m s_{t,m} > \epsilon_{\text{global}}]$ disables both regularization losses on search frames where no camera carries meaningful evidence, leaving the base imitation loss to drive policy behavior.

3.2.2 Tactile Instantiation

Tactile evidence is contact-based, with two regimes defined by the structure of the task. In the sparse-contact regime, which includes skills such as grasping, pressing, and insertion, the occurrence of contact is itself the task-relevant event. Evidence reduces to a binary contact indicator $E_{t,\text{tactile}} = \mathbb{1}[\text{contact at frame } t]$, similar in spirit to the contact-gating in prior work [14]. In the rich-contact regime, contact persists throughout the sub-task. Binary contact alone provides no frame-level discriminative signal. We therefore introduce a task-feature operator, ϕ_{task} that maps the current tactile observation to a similarity score against the target feature, yielding $E_{t,\text{tactile}} = \phi_{\text{task}}(o^{\text{tactile}}_t) \in [0, 1]$. The specific form of ϕ_{task} is task-dependent. In our evaluation, the target feature is a localized tactile pattern, such as a crack on a surface, and ϕ_{task} compares against reference signatures from demonstrations.

3.3 Evidence-Gated Regularization Objectives

The evidence score $E_{t,m}$ links the two failure modes formalized in Sec. 3.1 to the training objective. When evidence is low, it activates an invariance objective on modality m, addressing nuisance sensitivity. When evidence is high, it activates a sufficiency objective on m, addressing single-modality insufficiency. For intermediate evidence levels, training defaults to the base imitation loss.

We realize this gating through hinge weights anchored at two thresholds $\tau_{\text{low}} < \tau_{\text{high}}$:

$$w_{\text{inv}}(E_{t,m}) = \max\left(0,\ \frac{\tau_{\text{low}} - E_{t,m}}{\tau_{\text{low}}}\right), \qquad w_{\text{suff}}(E_{t,m}) = \max\left(0,\ \frac{E_{t,m} - \tau_{\text{high}}}{1 - \tau_{\text{high}}}\right). \quad (1)$$

Between τ_{low} and τ_{high} lies an ambiguous zone in which neither loss applies. To probe each loss, we introduce a corruption operator $\mathcal{C}_m(\cdot)$ that acts on modality m. This yields two corrupted observa-

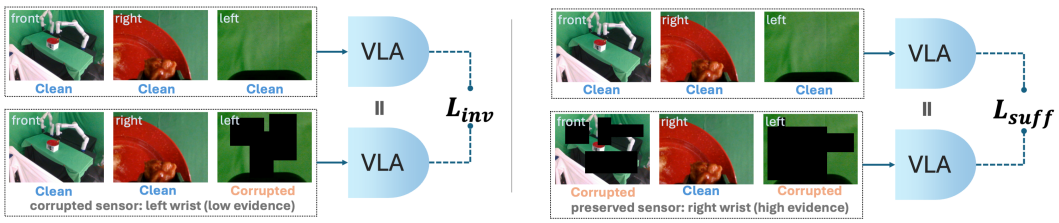

Figure 2: The two evidence-gated loss terms. **Left:** $\mathcal{L}_{\text{inv}}$ enforces invariance when a low-evidence sensor is corrupted. **Right:** $\mathcal{L}_{\text{suff}}$ enforces sufficiency when only a high-evidence sensor is preserved.

tions: $\tilde{o}_t^{(m)}$, in which modality m alone is corrupted by $\mathcal{C}_m$ and the other modalities are kept intact, and o_t^m, defined as in Sec. 3.1, in which all modalities other than m are corrupted. We instantiate $\mathcal{C}_m$ using random erasing on all visual inputs, including both camera RGB images and the RGB images produced by GelSight tactile sensors. However, the framework is agnostic to the specific choice of corruption.

The invariance loss targets the nuisance-sensitivity failure mode formalized in Sec. 3.1. When modality m provides little evidence, perturbations to m should not alter the policy's action (Fig. 2, left). Let v_θ denote the flow-matching velocity prediction of the underlying $\pi_{0.5}$ policy [6]. The global gate restricts the sum to informative frames, and the hinge weight w_{inv} activates only on low-evidence sensors; the resulting invariance loss is defined in Eq. (2).

$$\mathcal{L}_{\text{inv}} = \sum_{t,\, m\in\mathcal{M}} G_t \cdot w_{\text{inv}}(E_{t,m}) \cdot \left\| v_\theta(o_t) - v_\theta(\tilde{o}_t^{(m)}) \right\|_2^2. \tag{2}$$

The single-sensor sufficiency loss directly addresses the single-modality insufficiency failure mode formalized in Sec. 3.1. When evidence on modality m is high, the policy should produce the same action when only m is preserved (i.e., the observation o_t^m defined in Sec. 3.1) as when all sensors are available (Fig. 2, right). The hinge weight w_{suff} activates only on high-evidence sensors; the resulting sufficiency loss is defined in Eq. (3).

$$\mathcal{L}_{\text{suff}} = \sum_{t,\, m\in\mathcal{M}} G_t \cdot w_{\text{suff}}(E_{t,m}) \cdot \left\| v_\theta(o_t) - v_\theta(o_t^m) \right\|_2^2. \tag{3}$$

The final training objective combines the underlying flow-matching imitation loss $\mathcal{L}_{\text{flow}}$ with the two evidence-gated objectives:

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}} + \lambda_{\text{suff}}\mathcal{L}_{\text{suff}}. \tag{4}$$

Here $\lambda_{\text{inv}}, \lambda_{\text{suff}} \geq 0$ balance the two regularizers.

3.4 Efficient Training via Importance Sampling

Computing $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{suff}}$ as written in Eqs. (2) and (3) sums over all $|\mathcal{M}|$ modalities, requiring $2|\mathcal{M}|$ additional corrupted forward passes through π_θ per training step. We reduce this overhead with an unbiased importance-sampled estimator. For each sample b in the batch, let $W_b^{\text{inv}} = \sum_{m\in\mathcal{M}} w_{\text{inv}}(E_{b,m})$ and define the categorical distribution $p_{b,m}^{\text{inv}} = w_{\text{inv}}(E_{b,m})/W_b^{\text{inv}}$. We sample a single modality $m_b^\star \sim p_b^{\text{inv}}$, evaluate the per-modality loss term only on $m_b^\star$, and rescale by W_b^{inv}, yielding the estimator in Eq. (5).

$$\hat{\mathcal{L}}_{\text{inv}} = \sum_b G_t \cdot W_b^{\text{inv}} \cdot \left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m_b^\star)}) \right\|_2^2. \tag{5}$$

The same construct yields $\hat{\mathcal{L}}_{\text{suff}}$ using w_{suff} and W_b^{suff}. Both estimators are unbiased: $\mathbb{E}[\hat{\mathcal{L}}_{\text{inv}}] = \mathcal{L}_{\text{inv}}$ and $\mathbb{E}[\hat{\mathcal{L}}_{\text{suff}}] = \mathcal{L}_{\text{suff}}$ (see Appendix). Sampling proportional to hinge weights concentrates compute on modalities that actually contribute to the loss for each sample. The training overhead drops from $2|\mathcal{M}|$ to two extra forward passes per step regardless of the number of modalities.

4 Experiments and Results

Our evaluation is structured around three questions: **(Q1)** Does modality entanglement, as formalized in Sec. 3.1, emerge in trained VLA policies? **(Q2)** Does EGR mitigate this entanglement? and **(Q3)** Do the benefits of EGR transfer to real-robot platforms with different sensor configurations?

4.1 Implementation Details

We instantiate EGR on $\pi_{0.5}$ [6], augmenting its flow-matching imitation loss with $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{suff}}$ (Sec. 3). Vision evidence is computed from BEHAVIOR-1K's per-frame instance segmentation in simulation, and from Grounding DINO v2 [32] with manually-prompted bounding boxes followed by SAM2 [33] offline for real-robot training data; segmentation is used at training time only. The corruption operator $\mathcal{C}_m$ is instantiated as random erasing on all visual inputs (including GelSight RGB) during training, center erasing for sim deployment as a controlled diagnostic that approximates a large physical occlusion, and physical distractors for real-robot deployment.

4.2 Benchmark

Our benchmark provides 47 curated skill segments (11 navigation "NAV" + 36 manipulation "MAN") derived from BEHAVIOR-1K, together with two evaluation suites and the metrics for diagnosing modality entanglement at both the action level and the task-completion level.

4.2.1 Evaluation

We evaluate using two suites over the same set of benchmark segments. Suite 1 is an inference-only diagnostic that measures per-frame action deviation under modality perturbations, without rollout-based execution, making it inexpensive to compute. Suite 2 is a rollout-based evaluation that measures task success through real executions, providing a task-completion-level assessment that complements Suite 1's action-level diagnostics.

We evaluate four corruption regimes across both suites, using consistent terminology throughout the paper. **Clean** leaves all sensors intact. **NoUseless** corrupts the useless modality to test nuisance sensitivity (Sec. 3.1). **UsefulOnly** corrupts all modalities except the single useful one to test single-modality insufficiency (Sec. 3.1). **SingleUseful**, used only in setups with exactly one useful sensor, corrupts all remaining sensors and thus combines NoUseless and UsefulOnly into a single regime. For setups with multiple useful sensors, we evaluate Clean, NoUseless, and UsefulOnly separately to isolate the two failure modes. For setups with exactly one useful sensor, we evaluate Clean and SingleUseful. In Suite 2, each segment is rolled out 20 times to estimate task success rate.

4.2.2 Baselines and Metrics

Baselines. We compare EGR against two baselines. The first is vanilla $\pi_{0.5}$. The second, denoted **ModDrop**, augments $\pi_{0.5}$ training with random modality token replacement, implementing the modality dropout strategy of [11] on the $\pi_{0.5}$ backbone. Both baselines use the same architecture and training data as EGR.

Metrics. For Suite 1, we report Δ, defined as the increase in per-frame action MSE under corrupted inputs relative to full inputs, evaluated only on task-relevant action dimensions $\mathcal{D}$: the robot base for NAV, and the torso and both arms for MAN. A formal definition is provided in the Appendix. For Suite 2, we report task success rate (SR) for both NAV and MAN segments. For MAN, we also report end-effector contact rate (EC), the fraction of trials in which the active arm contacts the focal object, and the minimum end-effector distance to the target (Dist). These auxiliary metrics capture partial progress even when SR is zero. Standard errors, confidence intervals, and significance tests for all reported numbers are provided in the Appendix.

4.2.3 Benchmark Construction

Our benchmark is built on BEHAVIOR-1K [34], using its official per-frame instance segmentation and skill annotations. We focus on the 12 tasks proposed by the second-place team in the BEHAVIOR-1K Challenge [35], of which 11 yield usable data (one excluded for truncated annotation). For each candidate skill segment, we apply three vision-evidence-based filters that align segments with the evaluation protocol of Sec. 4.2.1. NAV segments naturally rely on only the head camera; we keep NAV segments where the head is useful and at least one wrist is useless, evaluating them under Clean and SingleUseful. MAN segments are more complex and may a priori have any

Table 1: Suite 1 (inference-only): action deviation under per-modality perturbation. Ratio is the percentage increase relative to Full; lower is better. Bold marks the lowest ratio per column.

	NAV			MAN				
Method	Full	SingleUseful	Ratio	Full	NoUseless	Ratio	UsefulOnly	Ratio
vanilla $\pi_{0.5}$	0.0616	0.0809	$+31.4\%$	0.0020	0.0022	$+12.7\%$	0.0052	$+158.5\%$
ModDrop	0.0853	0.1172	$+37.4\%$	0.0024	0.0026	$+10.9\%$	0.0056	$+137.7\%$
EGR (ours)	0.0780	0.0903	$\mathbf{+15.8\%}$	0.0023	0.0024	$\mathbf{+5.1\%}$	0.0051	$\mathbf{+121.8\%}$

Table 2: Suite 2 (rollout): task success rate (SR), end-effector contact rate (EC), and end-effector minimum distance to target (Dist). MAN columns report SR / EC / Dist. Bold marks the best per metric per regime.

	NAV ($n = 11$)		MAN ($n = 36$), SR↑ / EC↑ / Dist↓		
Method	Full	SingleUseful	Full	NoUseless	UsefulOnly
vanilla $\pi_{0.5}$	**42.27**	15.45	12.50 / **52.78** / **0.25**	9.44 / 47.22 / 0.26	2.78 / 22.78 / 0.45
ModDrop	40.00	15.45	7.50 / 40.97 / 0.29	6.81 / 38.89 / 0.29	2.78 / 22.36 / 0.47
EGR (ours)	40.91	**37.27**	**16.39** / 50.00 / 0.26	**16.53** / **53.89** / **0.25**	**6.11** / **30.14** / **0.37**

subset of cameras useful; for a uniform NoUseless / UsefulOnly evaluation across segments, we keep MAN segments where the head camera is useful and exactly one wrist is useless. After filtering, we re-decompose long-horizon demonstrations into segments and replay each demonstration to extract per-segment initial sim states. This yields 11 NAV + 36 MAN = 47 rollout-based segments. Filter statistics, threshold values, and replay procedure details are in the Appendix.

4.3 Diagnosing Modality Entanglement (Q1)

The vanilla $\pi_{0.5}$ rows of Table 1, in both the NAV and MAN panels, directly evidence both failure modes. On NAV, the SingleUseful condition corrupts both wrist cameras, which our filter has labeled as useless, while leaving the head camera (the only useful sensor) intact. Action deviation nonetheless increases by $+31.4\%$. Because NAV's SingleUseful condition simultaneously realizes NoUseless and UsefulOnly, it is a joint diagnosis of both failure modes. On MAN, corrupting the useless modality alone produces a $+12.7\%$ ratio, and keeping only the useful modality produces a $+158.5\%$ ratio. The first pair of numbers matches the nuisance sensitivity condition, and the third number matches the single-modality insufficiency condition. ModDrop shows comparable ratio patterns and higher absolute action deviation, indicating that random modality dropout does not eliminate this entanglement.

The rollout-based numbers in Table 2 corroborate the diagnosis at the task-completion level. Vanilla $\pi_{0.5}$ on MAN drops from 12.5% (Full) to 9.4% (NoUseless) to 2.8% (UsefulOnly); on NAV, from 42.3% (Full) to 15.5% (SingleUseful). Both suites converge: modality entanglement manifests at the action level (Table 1) and task-completion level (Table 2).

4.4 EGR Reduces Modality Entanglement (Q2)

The main result is the EGR row of Table 2. On MAN, EGR improves Full from 12.5 to 16.4 ($+31\%$ relative), NoUseless from 9.4 to 16.5 ($+75\%$), and UsefulOnly from 2.8 to 6.1 ($+120\%$). On NAV, Full is essentially preserved ($42.3 \rightarrow 40.9$) while SingleUseful rises from 15.5 to 37.3 ($+141\%$). Crucially, robustness gains do not come at the cost of clean-input performance, and MAN Full SR even improves substantially (12.5% to 16.4%, $+31\%$ relative), suggesting that evidence-gated regularization additionally helps the policy learn more task-relevant features even on clean inputs. The pattern of improvement matches the construction of the two losses, which target NoUseless and UsefulOnly directly. Table 1 corroborates this: EGR's action-deviation ratios are consistently lower than vanilla and ModDrop (NAV SingleUseful: 31.4%→15.8%; MAN UsefulOnly: 158.5%→121.8%).

4.5 Real-Robot Validation (Q3)

We evaluate EGR on two real-robot platforms with substantially different embodiments and modality combinations (Fig. 3). The bi-manual platform consists of two Kinova arms and three RGB cameras, with the sensor layout shown in the inset of Fig. 3c. It is evaluated on three short-horizon

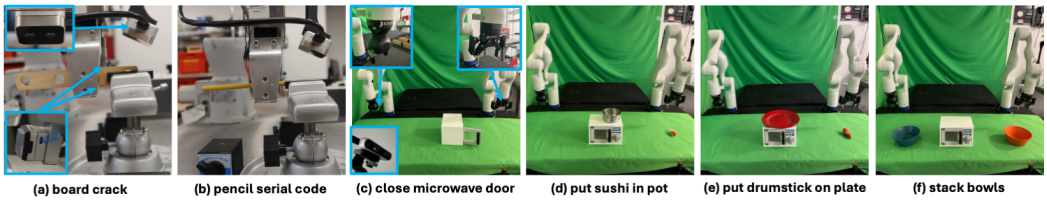

Figure 3: Real-robot platforms and tasks. **(a, b)** Vision-tactile platform. **(c-f)** Bi-manual platform. Sensor-detail zoom-ins are shown in (a) and (c).

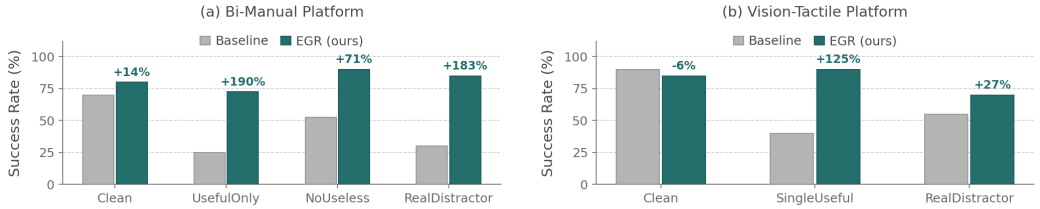

Figure 4: Real-robot results across two platforms (10 trials per condition). Each pair of bars compares the vanilla baseline (left, gray) and EGR (right, teal); relative gain is shown above each pair.

tasks (Fig. 3c-e) and one long-horizon bowl-stacking task (Fig. 3f). The vision–tactile platform consists of a single MELFA ASSISTA arm, one RGB camera, and two GelSight tactile sensors, with the sensor layout shown in the inset of Fig. 3a. It is evaluated on two tactile feature-localization tasks (Fig. 3a-b), which emulate industrial inspection settings where visual information may be limited and tactile feedback provides task-relevant signal. Following the evaluation regimes in Sec. 4.2.1, we test the bi-manual platform, which has two useful sensors, under Clean, NoUseless, and UsefulOnly conditions. We test the tactile platform, for which the right GelSight is the sole useful sensor, under Clean and SingleUseful conditions. For both platforms, we also include a **RealDistractor** condition in which a physical out-of-distribution object is placed in the camera view, providing a non-synthetic probe of robustness.

4.5.1 Bi-Manual Vision Platform

We run 4 tasks $\times$ 4 conditions $\times$ 10 trials per condition. Averaged across the four tasks, EGR improves Clean from 70% to 80% (+14%), NoUseless from 52.5% to 90% (+71%), UsefulOnly from 25% to 72.5% (+190%), and RealDistractor from 30% to 85% (+183%); see Fig. 4. The largest relative gain occurs on RealDistractor, where the distractor is a physical object the policy has never encountered, providing the strongest out-of-distribution signal among our conditions.

4.5.2 Vision-Tactile Platform

We run 2 tasks (pencil serial code, board crack) $\times$ 3 conditions $\times$ 10 trials per condition. Averaged across the two tasks, EGR preserves Clean within trial noise (90% to 85%, -6%), improves SingleUseful from 40% to 90% ($+125\%$), and improves RealDistractor from 55% to 70% ($+27\%$); see Fig. 4. EGR improves performance on the tactile platform under both synthetic SingleUseful corruption and physical RealDistractor corruption, confirming that the framework's gains transfer to a fundamentally different sensor combination.

5 Limitations

Our evidence instantiations derive from task structure: visible task-relevant geometry for vision and reference contact signatures for tactile. This covers settings in which the relevant signals can be specified a priori, but not cases where those signals must themselves be learned, such as in-hand object reorientation. Extending EGR to learn evidence from data, for example through information-theoretic objectives or self-supervised feature discovery, is a natural direction orthogonal to its regularization machinery. Our simulation benchmark also covers only 12 of the 50 BEHAVIOR-1K tasks so far; scaling to all 50 is straightforward future work using the same filtering and replay pipeline.

6 Conclusion

We presented Evidence-Gated Regularization (EGR), a training-time framework that mitigates modality entanglement in Vision-Language-Action policies. EGR uses a per-frame, per-sensor task-relevance signal to gate two state-conditional consistency objectives, requires no architectural change to the base policy, and adds zero inference cost. Across a BEHAVIOR-1K benchmark and two real-robot platforms with fundamentally different embodiments and sensors, EGR substantially improves success rates under both synthetic corruption and physical distractors.

References

[1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

[2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[3] K. Sivamayil, E. Rajasekar, B. Aljafari, S. Nikolovski, S. Vairavasundaram, and I. Vairavasundaram. A systematic study on reinforcement learning based applications. *Energies*, 16(3): 1512, 2023.

[4] R. Sapkota, Y. Cao, K. I. Roumeliotis, and M. Karkee. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025.

[5] M. J. Kim, F. Xia, R. Martin-Martin, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[6] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[7] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[8] C. Dancette, R. Cadene, D. Teney, and M. Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583, 2021.

[9] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[10] Y. Yang, B. Ikeda, G. Bertasius, and D. Szafir. Arcade: Scalable demonstration collection and generation via augmented reality for imitation learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2855–2861. IEEE, 2024.

[11] G.-H. Liu, A. Siravuru, S. Prabhakar, M. Veloso, and G. Kantor. Learning end-to-end multimodal sensor policies for autonomous navigation. In *Conference on robot learning*, pages 249–261. PMLR, 2017.

[12] J. Guo, Z. Wu, C. Tu, Y. Ma, X. Kong, Z. Liu, J. Ji, S. Zhang, Y. Chen, K. Chen, et al. On robustness of vision-language-action model against multi-modal perturbations. *arXiv preprint arXiv:2510.00037*, 2025.

[13] Y. Hao, R. Wang, Z. Cao, Z. Wang, Y. Cui, and D. Sadigh. Masked imitation learning: Discovering environment-invariant modalities in multimodal demonstrations. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–7. IEEE, 2023.

[14] K. Zhang, H. Zhang, Z. Xu, Z. Zhang, M. R. I. Prince, X. Li, X. Han, Y. Zhou, A. Ajoudani, and Y. She. Tacvla: Contact-aware tactile fusion for robust vision-language-action manipulation. *arXiv preprint arXiv:2603.12665*, 2026.

[15] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[16] J. Huang, S. Wang, F. Lin, Y. Hu, C. Wen, and Y. Gao. Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025.

[17] P. Hosseini, S. Nawathe, M. Moayeri, S. Balasubramanian, and S. Feizi. Seeing what's not there: Spurious correlation in multimodal llms. *arXiv e-prints*, pages arXiv–2503, 2025.

[18] R. Cai, B. Li, X. Wen, M. Chen, and Z. Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2025.

[19] Z. Wu, H.-Y. Huang, and Y. Wu. Beyond spurious signals: Debiasing multimodal large language models via counterfactual inference and adaptive expert routing. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3805–3825, 2025.

[20] S. Fei, S. Wang, J. Shi, Z. Dai, J. Cai, P. Qian, L. Ji, X. He, S. Zhang, Z. Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

[21] Y. Fang, Y. Feng, D. Jing, J. Liu, Y. Yang, Z. Wei, D. Szafir, and M. Ding. When vision overrides language: Evaluating and mitigating counterfactual failures in vlas. *arXiv preprint arXiv:2602.17659*, 2026.

[22] W. Wang, D. Tran, and M. Feiszli. What makes training multi-modal classification networks hard? *arXiv preprint arXiv:1905.12681*, 2019.

[23] X. Peng, Y. Wei, A. Deng, et al. Balanced multimodal learning via on-the-fly gradient modulation. *arXiv preprint arXiv:2203.15332*, 2022.

[24] C. Du, J. Teng, T. Li, et al. On uni-modal feature learning in supervised multi-modal learning. *arXiv preprint arXiv:2305.01233*, 2023.

[25] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

[26] M. K. Reza, A. Prater-Bennette, and M. S. Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 47(2):742–754, 2024.

[27] R. Feng, D. Hu, W. Ma, and X. Li. Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation. *arXiv preprint arXiv:2408.01366*, 2024.

[28] A. J. Hancock, A. Z. Ren, and A. Majumdar. Run-time observation interventions make vision-language-action models more visually robust. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9499–9506. IEEE, 2025.

[29] Y. Zhang, W. Yuan, Y. Zhang, X. Zhang, and J. Wan. Focusvla: Focused visual utilization for vision-language-action models. *arXiv preprint arXiv:2603.28740*, 2026.

[30] X. Jia, B. Yang, Z. Ge, X. Nie, Y. Zhou, C. Fan, Y. Li, Y. Chai, C. Jing, Z. Liang, et al. Guidedvla: Specifying task-relevant factors via plug-and-play action attention specialization. *arXiv preprint arXiv:2605.12369*, 2026.

[31] Y. Yang, L. Zhao, M. Ding, G. Bertasius, and D. Szafir. Boss: Benchmark for observation space shift in long-horizon task. *IEEE Robotics and Automation Letters*, 2025.

[32] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[33] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pages 28085–28128, 2025.

[34] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023.

[35] J. Bai, Y.-W. Chao, Q. Chen, J. Gu, M. J. Kim, Z. Li, X. Li, T.-Y. Lin, M.-Y. Liu, N. Ma, K. Mo, D. Qu, S. Sun, H. Xia, F. Wei, and X. Zeng. Openpi comet: Competition solution for 2025 behavior challenge. *arXiv preprint arXiv:2512.10071*, 2025. URL https://arxiv.org/abs/2512.10071.

This appendix supplements the main paper with implementation details, benchmark construction details, statistical analyses, and a theoretical derivation. Section A reports hyperparameters. Section B reports filter statistics, threshold values, and the sim state replay procedure. Section C formally defines the Suite 1 metric Δ and reports standard errors, confidence intervals, and significance tests for all numbers in the main paper. Section D proves that the importance-sampled estimators used during training are unbiased.

A Implementation Details

The values in this section specify the configuration used to train EGR on BEHAVIOR-1K in simulation and the synthetic corruption operator used to construct the Suite 1 and Suite 2 evaluations. The configuration is identical across all EGR simulation runs reported in the main text.

Table 3 lists the hyperparameters introduced by EGR itself: the two loss weights, the hinge thresholds, the evidence-preprocessing constants, and a flag governing how the per-camera mask channel is treated under corruption.

Symbol	Value	Description
λ_{inv}	0.2	weight on $\mathcal{L}_{\text{inv}}$ in the total loss
λ_{suff}	0.1	weight on $\mathcal{L}_{\text{suff}}$ (denoted `lambda_solo` in the released code)
τ_{low}	0.2	hinge knee below which w_{inv} activates
τ_{high}	0.5	hinge knee above which w_{suff} activates
α	0.3	weight of the gated interaction-object contribution in the per-camera score $s_{t,m}$
ϵ_{focal}	10^{-3}	interaction-gate threshold on focal-object area $a^{\text{focal}}_{t,m}$
ϵ_{global}	10^{-3}	global-gate threshold on $\max_m s_{t,m}$
ε	10^{-8}	numerical stabilizer in cross-camera normalization
`set_mask`	False	corrupted cameras retain `image_mask`=True so patch tokens stay valid

Table 3: EGR-specific hyperparameters.

Two distinct corruption operators are used. The training-time operator $\mathcal{C}_m$ injects random rectangular erasures during the computation of $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{suff}}$ (Table 4). The deployment-time operator used in simulation evaluation applies a center-biased black erasure that approximates a large occlusion (Table 5). Both operators sample rectangles independently per camera.

Parameter	Setting
Number of rectangles	uniform integer in $\{2, 3, 4, 5\}$
Total erased-area fraction	$\text{Uniform}(0.30,\ 0.80)$
Per-rectangle aspect ratio	$\exp(\text{Uniform}(-1,\ 1)) \approx [0.37,\ 2.72]$
Rectangle placement	top-left position uniform over image
Fill value	per-pixel $\text{Uniform}(-1,\ 1)$ noise (images normalized to $[-1, 1]$)
Per-camera independence	yes

Table 4: Training-time corruption operator $\mathcal{C}_m$ used inside $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{suff}}$.

Parameter	Setting
Number of rectangles	uniform integer in $\{1,\ 2,\ 3\}$
Total erased-area fraction	$\text{Uniform}(0.30,\ 0.70)$
Per-rectangle aspect ratio	$\text{Uniform}(0.5,\ 2.0)$
Rectangle placement	Gaussian about image center with $\sigma = 0.10\,W$ and $0.10\,H$, clipped to fit
Fill value	0 (black)
Per-camera independence	yes

Table 5: Sim-deployment corruption operator used in Suite 1 and Suite 2 to instantiate the NoUseless, UsefulOnly, and SingleUseful regimes.

Table 6 lists the training-loop configuration. All EGR simulation runs warm-start from the $\pi_{0.5}$ checkpoint provided by [35], pre-trained on the 12 BEHAVIOR-1K Challenge tasks; only LoRA adapter weights are updated.

Setting	Value
Batch size	288
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$
Peak learning rate	2.5×10^{-5}
LR schedule	cosine decay, 1,000-step warmup, decay over 50,000 steps, end LR 0
Total training steps	50,000
Weight decay	10^{-10}
Gradient clipping	global-norm clip at 1.0
EMA	disabled
Numerical precision	bfloat16 model; float32 loss and parameter accumulators
Hardware	2× NVIDIA B200, FSDP across both devices
Dataloader workers	8
PaliGemma backbone	`gemma_2b_lora` (LoRA on the 2B base)
Action expert	`gemma_300m_lora` (LoRA on the 300M base)
Action horizon	32
Trainable parameters	LoRA adapters only; base weights frozen

Table 6: Training-loop configuration for EGR simulation runs.

B Benchmark Construction

B.1 Filter Pipeline Statistics

The benchmark curation pipeline starts from the 11 BEHAVIOR-1K Challenge tasks that yield usable annotations. Task 34 is excluded upstream because its annotation is truncated, removing 4 candidate segments before the evidence pipeline is run. The 11 retained tasks contribute 118 candidate segments, 39 navigation (NAV) and 79 manipulation (MAN). Three vision-evidence filters reduce this pool to the 47 segments evaluated in the main text. Filter 1 retains segments whose per-frame global gate $G_t = 1$ on at least 80% of frames, removing segments in which no camera observes the focal object on a typical frame. Filter 2 retains segments whose head camera is labeled useful, restricting evaluation to scenarios where the head view is informative. Filter 3 enforces a clean wrist-camera contrast: MAN segments are kept only when exactly one wrist is useless (so that the contrast between the useful wrist and the useless wrist is unambiguous), and NAV segments are kept when at least one wrist is useless (a looser criterion, since wrists are typically uninformative during locomotion). Table 7 reports the segment count after each filter, broken out by skill type.

Stage	All	NAV	MAN
Input (11 evaluation tasks)	118	39	79
After Filter 1 (segment-level global-gate pass)	103	25	78
After Filter 2 (head-camera useful)	99	25	74
After Filter 3 (wrist-camera ablation criterion)	47	11	36

Table 7: Segment counts through the three vision-evidence filters.

B.2 Threshold Values

Two thresholds operate at the per-frame level and drive the per-camera useful/useless classification: ϵ_{global} defines when a frame carries evidence at all, and `useless_threshold` defines when a frame counts as not seeing the focal object on a given camera. The remaining two thresholds aggregate frame-level outcomes to segment-level decisions: `g_threshold` on the fraction of frames passing the global gate, and `frac_threshold` on the fraction of frames below `useless_threshold`. A

camera m is labeled useless on a segment if at least 75% of its frames satisfy $E_{t,m} < 0.01$, and useful otherwise. Table 8 lists each threshold, its numeric value, and which filter consumes it.

Threshold	Value	What it gates	Consumed by
ϵ_{global}	10^{-3}	per-frame global gate, $G_t = 1$ iff $\max_m s_{t,m} > 10^{-3}$	Filter 1 (defines G_t)
`g_threshold`	0.80	minimum fraction of frames with $G_t = 1$ for a segment to pass	Filter 1
`useless_threshold`	0.01	per-frame, per-camera cutoff on $E_{t,m}$ below which the frame counts as not seeing the focal object	Filter 2 and Filter 3
`frac_threshold`	0.75	fraction of frames below `useless_threshold` required to label a camera useless on the segment	Filter 2 and Filter 3

Table 8: Threshold values used in the filter pipeline.

B.3 Sim State Replay Procedure

The official BEHAVIOR-1K HDF5 stores per-frame state as a slot-keyed serialized tensor; loading it into a Suite 2 evaluation scene with a different object set mis-aligns the slots, and a single-frame jump-in additionally loses hidden physics state such as joint velocities, contact constraints, and particle-system bookkeeping. We therefore replay each task's first episode from frame 0 in `VISUAL_ONLY` mode and, at each segment start frame, call `og.sim.dump_state(serialized=False)` to save a name-keyed snapshot covering all object poses, robot and gripper joint configurations, and particle-system state. At evaluation time, Suite 2 loads the snapshot directly into the evaluation scene and runs one zero-gravity settling step so PhysX re-establishes contact constraints before the policy begins acting.

C Metrics and Statistical Analysis

C.1 Definition of Δ for Suite 1

For each segment with relevant action dimensions $\mathcal{D}$ and frames $\mathcal{F}$,

$$\Delta = \underbrace{\frac{1}{|\mathcal{F}|}\sum_{t\in\mathcal{F}} \|\pi_\theta(o_t^{\text{corrupt}}) - a_t^\star\|_{\mathcal{D}}^2}_{\text{MSE}_{\text{corrupt}}} - \underbrace{\frac{1}{|\mathcal{F}|}\sum_{t\in\mathcal{F}} \|\pi_\theta(o_t^{\text{full}}) - a_t^\star\|_{\mathcal{D}}^2}_{\text{MSE}_{\text{full}}}, \tag{6}$$

where $a_t^\star$ is the ground-truth action chunk and $\mathcal{D}$ selects the robot base dimensions for NAV segments and the torso plus both arm dimensions for MAN segments.

C.2 Standard Errors and Confidence Intervals

This subsection reports per-cell uncertainty for every number used in Tables 1 and 2 and in Fig. 4 of the main paper. Aggregated means use $\text{SE} = \text{SD}/\sqrt{n}$ across the natural sampling unit: segments for the simulation suites and tasks for the real-robot platforms. Binomial rates additionally report the Wilson 95% interval over the pooled trial count. Suite 1 Ratios additionally report a 95% bootstrap interval from 10,000 paired resamples over segments. All computations use a fixed seed.

Suite 1. Each cell reports the Full and corrupt MSE means with standard errors and the Ratio with a bootstrap 95% interval; the EGR row has the lowest Ratio in every panel.

Panel	Method	n	Full MSE (mean $\pm$ SE)	Corrupt MSE (mean $\pm$ SE)	Ratio % (mean $\pm$ SE) [95% CI]
NAV / SingleUseful	vanilla	11	0.0616 ± 0.0105	0.0809 ± 0.0115	$+31.35 \pm 9.33$ $[+16.8, +53.4]$
NAV / SingleUseful	ModDrop	11	0.0853 ± 0.0130	0.1172 ± 0.0172	$+37.42 \pm 14.38$ $[+15.6, +72.3]$
NAV / SingleUseful	EGR	11	0.0780 ± 0.0115	0.0903 ± 0.0128	$+15.83 \pm 4.66$ $[+8.0, +26.3]$
MAN / NoUseless	vanilla	36	0.0020 ± 0.0002	0.0022 ± 0.0002	$+12.74 \pm 4.93$ $[+4.0, +23.3]$
MAN / NoUseless	ModDrop	36	0.0024 ± 0.0002	0.0026 ± 0.0002	$+10.90 \pm 2.65$ $[+6.0, +16.4]$
MAN / NoUseless	EGR	36	0.0023 ± 0.0003	0.0024 ± 0.0002	$+5.07 \pm 2.67$ $[+0.4, +10.9]$
MAN / UsefulOnly	vanilla	36	0.0020 ± 0.0002	0.0052 ± 0.0007	$+158.54 \pm 28.36$ $[+105.0, +216.4]$
MAN / UsefulOnly	ModDrop	36	0.0024 ± 0.0002	0.0056 ± 0.0008	$+137.73 \pm 19.82$ $[+99.4, +176.7]$
MAN / UsefulOnly	EGR	36	0.0023 ± 0.0003	0.0051 ± 0.0007	$+121.82 \pm 25.10$ $[+74.0, +172.4]$

Table 9: Suite 1 per-cell statistics. Standard errors are computed across segments; Ratio confidence intervals are from 10,000 paired bootstrap resamples over segments.

Suite 2 (NAV). Per-segment SR is a binomial proportion over $n = 20$ trials, and we report the cross-segment mean $\pm$ SE together with a bootstrap 95% CI over $n = 11$ segments.

Condition	Method	n_{seg}	SR (mean $\pm$ SE)	95% CI
Full	vanilla	11	42.27 ± 14.78	$[15.9, 70.0]$
Full	ModDrop	11	40.00 ± 14.13	$[15.5, 66.8]$
Full	EGR	11	40.91 ± 14.47	$[15.0, 68.2]$
SingleUseful	vanilla	11	15.45 ± 9.28	$[1.8, 34.5]$
SingleUseful	ModDrop	11	15.45 ± 9.18	$[1.8, 34.5]$
SingleUseful	EGR	11	37.27 ± 13.76	$[13.6, 63.6]$

Table 10: Suite 2 NAV success rate per cell. Means and SEs are computed across $n_{\text{seg}} = 11$ segments; CIs are from 10,000 bootstrap resamples over segments.

Suite 2 (MAN). For each cell, the mean is computed across $n_{\text{seg}} = 36$ MAN segments; each per-segment value summarizes 20 trials. We report SE and bootstrap 95% CI across segments for all three metrics.

		SR (%)		EC (%)		EE distance (m)	
Condition	Method	mean $\pm$ SE	95% CI	mean $\pm$ SE	95% CI	mean $\pm$ SE	95% CI
Full	vanilla	12.50 ± 3.77	$[6.0, 20.4]$	52.78 ± 6.15	$[40.8, 64.6]$	0.249 ± 0.049	$[0.16, 0.35]$
Full	ModDrop	7.50 ± 3.17	$[2.5, 14.6]$	40.97 ± 6.24	$[28.9, 52.9]$	0.287 ± 0.048	$[0.20, 0.39]$
Full	EGR	16.39 ± 4.10	$[8.9, 24.7]$	50.00 ± 5.95	$[38.5, 61.4]$	0.257 ± 0.049	$[0.17, 0.36]$
NoUseless	vanilla	9.44 ± 3.41	$[3.8, 16.8]$	47.22 ± 6.01	$[35.7, 58.6]$	0.264 ± 0.045	$[0.18, 0.36]$
NoUseless	ModDrop	6.81 ± 3.05	$[1.9, 13.3]$	38.89 ± 5.96	$[27.5, 50.4]$	0.292 ± 0.047	$[0.21, 0.39]$
NoUseless	EGR	16.53 ± 4.21	$[9.2, 25.1]$	53.89 ± 6.07	$[41.9, 65.6]$	0.246 ± 0.046	$[0.16, 0.34]$
UsefulOnly	vanilla	2.78 ± 2.78	$[0.0, 8.3]$	22.78 ± 6.29	$[11.2, 35.1]$	0.447 ± 0.059	$[0.34, 0.57]$
UsefulOnly	ModDrop	2.78 ± 2.78	$[0.0, 8.3]$	22.36 ± 6.34	$[10.8, 35.0]$	0.470 ± 0.064	$[0.35, 0.60]$
UsefulOnly	EGR	6.11 ± 3.05	$[1.4, 12.9]$	30.14 ± 6.01	$[18.9, 42.2]$	0.366 ± 0.052	$[0.27, 0.47]$

Table 11: Suite 2 MAN per-cell statistics for the three reported metrics. $n_{\text{seg}} = 36$ for every cell.

Real bi-manual platform. Each (task, condition, method) cell contains $n = 10$ trials, pooled across 4 tasks to $N = 40$ trials per (condition, method). We report the pooled rate with its Wilson 95% interval over the 40 trials, together with the cross-task SE over $n_{\text{task}} = 4$.

	Baseline			EGR		
Condition	Pooled n/N	Wilson 95% CI	Cross-task SE	Pooled n/N	Wilson 95% CI	Cross-task SE
Clean	28/40	$[0.55, 0.82]$	0.70 ± 0.07	32/40	$[0.65, 0.90]$	0.80 ± 0.07
NoUseless	21/40	$[0.37, 0.67]$	0.53 ± 0.05	36/40	$[0.77, 0.96]$	0.90 ± 0.07
UsefulOnly	10/40	$[0.14, 0.40]$	0.25 ± 0.10	29/40	$[0.57, 0.84]$	0.73 ± 0.10
RealDistractor	12/40	$[0.18, 0.45]$	0.30 ± 0.15	34/40	$[0.71, 0.93]$	0.85 ± 0.09

Table 12: Real bi-manual platform SR per condition. The pooled count n aggregates 4 tasks $\times$ 10 trials; the cross-task SE is over $n_{\text{task}} = 4$ task-level SRs.

Real tactile platform. SR per (condition, method) cell pools the 2 tasks $\times$ 10 trials into $N = 20$ trials.

	Baseline		EGR	
Condition	Pooled n/N	Wilson 95% CI	Pooled n/N	Wilson 95% CI
Clean	18/20	$[0.70, 0.97]$	17/20	$[0.64, 0.95]$
SingleUseful	8/20	$[0.22, 0.61]$	18/20	$[0.70, 0.97]$
RealDistractor	11/20	$[0.34, 0.74]$	14/20	$[0.48, 0.85]$

Table 13: Real tactile platform SR per condition, pooled across the 2 tasks.

C.3 Significance Tests

For the simulation suites we use the one-sided paired Wilcoxon signed-rank test on per-segment values, with EGR as the test arm and one of the two baselines as the reference. The alternative hypothesis follows the metric's direction: for Suite 1 Ratio we test EGR $\Delta <$ baseline Δ where $\Delta = \mathrm{MSE}_{\mathrm{corrupt}} - \mathrm{MSE}_{\mathrm{full}}$; for Suite 2 SR and EC we test EGR $>$ baseline; for Suite 2 EE distance we test EGR $<$ baseline. For the real-robot platforms we use Fisher's exact test (two-sided) on the pooled trial counts. Significance markers throughout: $^{*}\, p < 0.05$, $^{**}\, p < 0.01$, $^{***}\, p < 0.001$.

Suite 1. EGR's per-segment Δ is significantly smaller than vanilla in all three panels and smaller than ModDrop in two of the three; the MAN NoUseless comparison against ModDrop is marginal ($p = 0.061$).

Panel	Comparison	n	W	p
NAV / SingleUseful	EGR vs vanilla	11	14.0	0.0261^{*}
NAV / SingleUseful	EGR vs ModDrop	11	10.0	0.0105^{*}
MAN / NoUseless	EGR vs vanilla	36	235.0	0.0400^{*}
MAN / NoUseless	EGR vs ModDrop	36	248.0	0.0606
MAN / UsefulOnly	EGR vs vanilla	36	213.0	0.0181^{*}
MAN / UsefulOnly	EGR vs ModDrop	36	200.0	0.0107^{*}

Table 14: Suite 1 paired Wilcoxon signed-rank tests on per-segment Δ. One-sided, EGR $\Delta <$ baseline Δ.

Suite 2 (NAV). With only 11 NAV segments the paired Wilcoxon attains its floor of $p = 2^{-4} = 0.0625$ under SingleUseful; all four segments on which the methods differ favor EGR, so the limitation is statistical power rather than effect direction or consistency.

Condition	Comparison	n_{seg}	nonzero diffs	W	p
Full	EGR vs vanilla	11	3	1.0	0.8750
Full	EGR vs ModDrop	11	3	4.0	0.3750
SingleUseful	EGR vs vanilla	11	4	10.0	0.0625
SingleUseful	EGR vs ModDrop	11	4	10.0	0.0625

Table 15: Suite 2 NAV paired Wilcoxon. With only 11 segments and few nonzero differences, the smallest attainable p is $2^{-(\text{nonzero})}$. Under SingleUseful all 4 nonzero per-segment differences favor EGR, attaining the floor $p = 2^{-4} = 0.0625$ for both baseline comparisons.

Suite 2 (MAN). The Full-condition SR gain is significant against ModDrop ($p < 0.001$) and marginal against vanilla ($p = 0.065$); it is a secondary effect, since EGR primarily targets the corruption conditions, where NoUseless and UsefulOnly are significant against both baselines.

Metric	Condition	Comparison	p
SR	Full	EGR vs vanilla	0.0649
SR	Full	EGR vs ModDrop	0.0003^{***}
SR	NoUseless	EGR vs vanilla	0.0055^{**}
SR	NoUseless	EGR vs ModDrop	0.0002^{***}
SR	UsefulOnly	EGR vs vanilla	0.0033^{**}
SR	UsefulOnly	EGR vs ModDrop	0.0033^{**}
EC	Full	EGR vs vanilla	0.8079
EC	Full	EGR vs ModDrop	0.0029^{**}
EC	NoUseless	EGR vs vanilla	0.0186^{*}
EC	NoUseless	EGR vs ModDrop	0.0001^{***}
EC	UsefulOnly	EGR vs vanilla	0.0057^{**}
EC	UsefulOnly	EGR vs ModDrop	0.0073^{**}
EE distance	Full	EGR vs vanilla	0.7540
EE distance	Full	EGR vs ModDrop	0.0051^{**}
EE distance	NoUseless	EGR vs vanilla	0.1990
EE distance	NoUseless	EGR vs ModDrop	0.0025^{**}
EE distance	UsefulOnly	EGR vs vanilla	0.0000^{***}
EE distance	UsefulOnly	EGR vs ModDrop	0.0000^{***}

Table 16: Suite 2 MAN paired Wilcoxon signed-rank tests across $n_{\text{seg}} = 36$ segments. SR/EC use H_1 : EGR $>$ baseline; EE distance uses H_1 : EGR $<$ baseline.

Real bi-manual. All three corruption conditions (NoUseless, UsefulOnly, RealDistractor) are significant at $p < 0.001$ against the baseline, while Clean shows no significant difference, consistent with EGR not targeting clean performance.

Condition	Baseline n/N	EGR n/N	Odds ratio	p
Clean	28/40	32/40	1.71	0.4391
NoUseless	21/40	36/40	8.14	0.0004^{***}
UsefulOnly	10/40	29/40	7.91	0.0000^{***}
RealDistractor	12/40	34/40	13.22	0.0000^{***}

Table 17: Real bi-manual Fisher's exact test (two-sided) on pooled trial counts across the 4 tasks.

Real tactile. On RealDistractor the tactile platform pools only $N = 20$ trials and the EGR advantage (14/20 vs 11/20) does not reach significance ($p = 0.51$); the same RealDistractor condition on the bi-manual platform, with $N = 40$ trials, is strongly significant ($p < 0.001$, Table 17), indicating that the tactile result is underpowered rather than absent.

Condition	Baseline n/N	EGR n/N	Odds ratio	p
Clean	18/20	17/20	0.63	1.0000
SingleUseful	8/20	18/20	13.50	0.0022^{**}
RealDistractor	11/20	14/20	1.91	0.5145

Table 18: Real tactile platform Fisher's exact test (two-sided) on pooled SR. RealDistractor shows a positive point estimate that does not reach significance at $N = 20$ pooled trials.

D Unbiased Importance-Sampled Estimator for $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{suff}}$

The per-sample contribution of $\mathcal{L}_{\text{inv}}$ in Eq. (2) of the main paper is

$$\ell_b^{\text{inv}} = G_t \sum_{m \in \mathcal{M}} w_{\text{inv}}(E_{b,m}) \cdot \left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m)}) \right\|_2^2. \tag{7}$$

Let $W_b^{\text{inv}} = \sum_{m \in \mathcal{M}} w_{\text{inv}}(E_{b,m})$. When $W_b^{\text{inv}} > 0$, define the categorical distribution

$$p_{b,m}^{\text{inv}} = \frac{w_{\text{inv}}(E_{b,m})}{W_b^{\text{inv}}}, \qquad \sum_{m \in \mathcal{M}} p_{b,m}^{\text{inv}} = 1. \tag{8}$$

Rewriting the per-sample loss using this distribution,

$$\ell_b^{\text{inv}} = G_t \cdot W_b^{\text{inv}} \cdot \sum_{m \in \mathcal{M}} p_{b,m}^{\text{inv}} \cdot \left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m)}) \right\|_2^2 = G_t \cdot W_b^{\text{inv}} \cdot \mathbb{E}_{m \sim p_b^{\text{inv}}} \left[\left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m)}) \right\|_2^2 \right]. \tag{9}$$

Our estimator draws a single sample $m_b^\star \sim p_b^{\text{inv}}$ and forms

$$\hat{\ell}_b^{\text{inv}} = G_t \cdot W_b^{\text{inv}} \cdot \left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m_b^\star)}) \right\|_2^2. \tag{10}$$

Taking expectation over $m_b^\star$,

$$\mathbb{E}_{m_b^\star \sim p_b^{\text{inv}}} \left[\hat{\ell}_b^{\text{inv}} \right] = G_t \cdot W_b^{\text{inv}} \cdot \mathbb{E}_{m \sim p_b^{\text{inv}}} \left[\left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m)}) \right\|_2^2 \right] = \ell_b^{\text{inv}}. \tag{11}$$

The W_b^{inv} multiplier is the importance-sampling correction; without it the estimator would be biased downward by a factor of W_b^{inv}. When $W_b^{\text{inv}} = 0$, no modality on sample b triggers the invariance hinge; we then fall back to a uniform $1/|\mathcal{M}|$ prior over modalities for sampling, and the loss contribution remains zero through the W_b^{inv} multiplier. Summing $\hat{\ell}_b^{\text{inv}}$ over the batch gives the estimator $\hat{\mathcal{L}}_{\text{inv}}$ of Eq. (5) of the main paper, so $\mathbb{E}[\hat{\mathcal{L}}_{\text{inv}}] = \mathcal{L}_{\text{inv}}$. The derivation for $\hat{\mathcal{L}}_{\text{suff}}$ is identical, with w_{suff} and W_b^{suff} replacing their invariance counterparts. This estimator preserves the expected gradient of the full-sum loss while reducing the per-step cost from $2|\mathcal{M}| + 1$ forward passes to three total (one clean, one for $\hat{\mathcal{L}}_{\text{inv}}$, one for $\hat{\mathcal{L}}_{\text{suff}}$), independent of the number of modalities.