MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Breaking Modality Entanglement in Vision-Language-Action Policies

Yang, Yue; Romeres, Diego; Hori, Chiori; Bertasius, Gedas; Szafir, Daniel; Jain, Siddarth

TR2026-170 November 10, 2026

Abstract

Vision-Language-Action (VLA) policies integrate multiple sensors to support robust robotic task execution, yet limited and homogeneous demonstrations can encourage reliance on spurious inter-sensor correlations rather than task-relevant evidence. We call this failure modality entanglement. It takes two forms: nuisance sensitivity, where corrupting an uninformative sensor still degrades the policy, and single-modality insufficiency, where the policy cannot act on the one informative sensor that remains intact. We study both failures on a BEHAVIOR-1K benchmark of 47 skills and on two real robots, and address them with Evidence-Gated Regularization (EGR), a training objective that leverages simple task knowledge to guide the policy toward the most relevant sensor at each frame. Our study shows that with limited demonstrations more sensors can make a policy more fragile, that clean-input success hides both failures, and that gating the training loss by task-relevant evidence reduces them without new data or architectural changes. EGR more than doubles success under single-sensor fallback in simulation and raises success under physical distractors from 30% to 85% on a bi-manual robot.

Conference on Robot Learning (CoRL) Workshop on Everything Beneath the Policy 2026

© 2026 MERL. This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Breaking Modality Entanglement in Vision-Language-Action Policies

Yue Yang*†, Diego Romeres†, Chiori Hori†, Gedas Bertasius*, Daniel Szafir*, Siddarth Jain†
*University of North Carolina at Chapel Hill †Mitsubishi Electric Research Laboratories (MERL)

***Abstract*—Vision-Language-Action (VLA) policies integrate multiple sensors to support robust robotic task execution, yet limited and homogeneous demonstrations can encourage reliance on spurious inter-sensor correlations rather than task-relevant evidence. We call this failure modality entanglement. It takes two forms: nuisance sensitivity, where corrupting an uninformative sensor still degrades the policy, and single-modality insufficiency, where the policy cannot act on the one informative sensor that remains intact. We study both failures on a BEHAVIOR-1K benchmark of 47 skills and on two real robots, and address them with Evidence-Gated Regularization (EGR), a training objective that leverages simple task knowledge to guide the policy toward the most relevant sensor at each frame. Our study shows that with limited demonstrations more sensors can make a policy more fragile, that clean-input success hides both failures, and that gating the training loss by task-relevant evidence reduces them without new data or architectural changes. EGR more than doubles success under single-sensor fallback in simulation and raises success under physical distractors from 30% to 85% on a bi-manual robot.**

I. Introduction

Robot policies increasingly fuse several sensors, on the common assumption that more sensing makes a robot more robust. Vision-Language-Action (VLA) policies, a leading approach to general-purpose robot learning, build on large pretrained backbones and use early multimodal fusion to combine visual, linguistic, and action-relevant inputs [1]–[3]. Early fusion, however, invites shortcut learning, in which a model exploits spurious cross-modal correlations instead of task-relevant signals [4], [5]. VLA policies inherit this weakness, and robot data make it worse. Compared with the web-scale corpora behind multimodal foundation models, robot demonstrations are far smaller and less diverse in operators, environments, and objects [6], [7]. We call the resulting failure *modality entanglement*: the policy depends on sensors that carry no task-relevant signal at the current state.

Fig. 1 shows two state-dependent failure modes that this entanglement causes. In navigation, the head camera carries the task-relevant information, while the two downward-facing wrist cameras see only the floor (Fig. 1(a)). Placing physical distractors in the wrist-camera views alone can stall the policy, although the head-camera view is unchanged (Fig. 1(b)). We call this *nuisance sensitivity*: degradation caused by perturbing sensors that hold no task-relevant signal at the current state. In shelf reaching, the head camera and the right wrist camera both see the target at first (Fig. 1(c)). When the target sits deeper in the shelf cell, the upper board blocks the head camera, while the right wrist camera still sees the target. The policy nevertheless stops or acts erratically (Fig. 1(d)). We call this *single-modality insufficiency*: when several sensors carry task-relevant signal but all except one are corrupted, the policy fails to fall back on the remaining sensor.

Modality entanglement points to a missing capability: at each frame, the policy should identify which sensors provide task-relevant evidence. Random modality dropout [8] and adversarial multi-sensor consistency [9] target generic robustness to sensor availability by treating sensors and timesteps uniformly. 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 contribute to modality entanglement [10]. TacVLA [11] instead introduces an explicit architectural gate, activating tactile tokens only when contact is detected. What remains missing is a relevance signal that is explicit, grounded in task structure rather than fragile demonstration statistics, and decoupled from any particular sensor modality. Appendix A discusses further related work.

We propose **Evidence-Gated Regularization (EGR)**, a training objective designed around these three properties. At each training frame, EGR gives every sensor an evidence score for the task-grounded signal it carries. We derive this score from the task structure instead of learning it end-to-end from demonstrations. 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. Because EGR is imposed only during training, it incurs no inference-time overhead. We test EGR on a BEHAVIOR-1K benchmark of 47 skill segments and on two real-robot platforms (Sec. III). EGR keeps the architecture and the demonstrations of the baseline fixed and changes only how the loss weighs each sensor. Sec. IV distills three lessons for multi-sensor policies.

II. Methodology

A. 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**,

This research was completed during Y. Yang's internship at MERL. yygx@cs.unc.edu

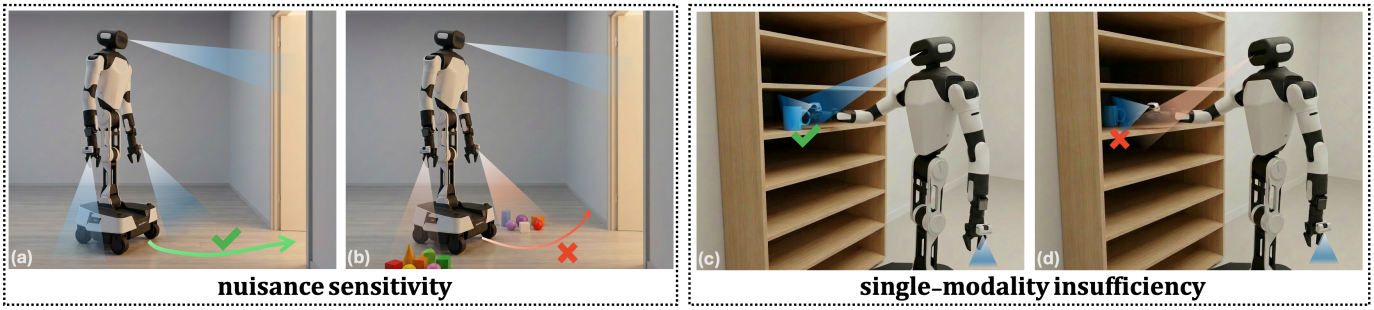


Fig. 1. Two failure modes of modality entanglement. **(a-b) Nuisance sensitivity:** distractors in a task-irrelevant wrist camera break navigation despite an unchanged head camera. **(c-d) Single-modality insufficiency:** when a deeper-placed target makes the shelf board occlude the head camera, the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

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.

B. 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. 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 as ground truth. It indicates which failure mode (Sec. II-A) the current frame is most susceptible to, thereby enabling selective regularization. For cameras, the score is the visible pixel area of the task's focal objects, plus that of interaction objects (e.g., receptacles) when a focal object is visible in the same camera, normalized across cameras. A global gate G_t turns off both regularizers on search frames where no camera sees these objects. Appendix B gives the full vision definition, and Appendix C the tactile one.

C. Evidence-Gated Regularization Objectives

The evidence score $E_{t,m}$ links the two failure modes formalized in Sec. II-A 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 $\tau_{\text{low}} < \tau_{\text{high}}$:

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

To probe each loss, we introduce a corruption operator $\mathcal{C}_m(\cdot)$ that acts on modality m. This yields two corrupted observations: $\tilde{o}_t^{(m)}$, 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. II-A, in which all modalities other than m are corrupted.

When modality m provides little evidence, perturbations to m should not alter the policy's action (Fig. 2 in Appendix D, left). Let v_θ denote the flow-matching velocity prediction of the underlying $\pi_{0.5}$ policy [3]. 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).

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

For sufficiency, 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. II-A) 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.

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 (Appendix D). The training overhead drops from $2|\mathcal{M}|$ to two extra forward passes per step regardless of the number of modalities.

III. Experiments and Results

Our evaluation is structured around three questions: **(Q1)** Does modality entanglement, as formalized in Sec. II-A,

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?

We instantiate EGR on $\pi_{0.5}$ [3], augmenting its flow-matching imitation loss with $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{suff}}$ (Sec. II). We fix the EGR hyperparameters at $\lambda_{\text{inv}} = 0.2$, $\lambda_{\text{suff}} = 0.1$, $\tau_{\text{low}} = 0.2$, $\tau_{\text{high}} = 0.5$ and $\alpha = 0.3$ for every run reported here. We did not tune the thresholds for performance.

A. Benchmark

Our benchmark provides 47 curated skill segments (11 navigation "NAV" + 36 manipulation "MAN") derived from BEHAVIOR-1K.

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 in both suites. **Clean** leaves all sensors intact. **NoUseless** corrupts the useless modality to test nuisance sensitivity (Sec. II-A; Appendix F defines the useful and useless labels). **UsefulOnly** corrupts all modalities except the single useful one to test single-modality insufficiency (Sec. II-A). **SingleUseful**, used only in setups with exactly one useful sensor, corrupts all remaining sensors and thus combines NoUseless and UsefulOnly into a single regime. In Suite 2, each segment is rolled out 20 times to estimate task success rate, which totals 7,800 rollouts across the three methods.

2) Baselines and Metrics:

a) 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 [8] on the $\pi_{0.5}$ backbone. Both baselines use the same architecture and training data as EGR. The evidence score never enters the policy. It only sets the two hinge weights of Eq. (1), and segmentation runs offline at training time. All methods train on the same observations, and the only channel through which evidence acts is the weighting of the loss.

b) Metrics: 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.

3) Benchmark Construction: Our benchmark is built on BEHAVIOR-1K [12], using its official per-frame instance segmentation and skill annotations. We focus on 12 representative BEHAVIOR-1K Challenge tasks selected in [13], 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. III-A1. The three filters reduce 118 candidate segments to 103, then 99, then 47. Appendix F-A explains why this selection does not bias the comparison between methods.

B. Diagnosing Modality Entanglement (Q1)

The rollout-based numbers in Table I diagnose modality entanglement 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). The NoUseless drop is the clearest sign: the corrupted camera carries no task signal at these frames, yet success falls by a quarter. Sec. IV traces this to the demonstration data. Suite 1 (Appendix G) gives the same diagnosis at the action level: for vanilla $\pi_{0.5}$ on MAN, corrupting the useless modality raises action deviation by $+12.7\%$, and keeping only the useful modality raises it by $+158.5\%$. ModDrop shows comparable ratio patterns, indicating that random modality dropout does not eliminate this entanglement.

C. EGR Reduces Modality Entanglement (Q2)

The main results are summarized in the EGR row of Table I. 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 ($+141\%$). On average, clean-input performance therefore does not degrade under EGR. At the action level, EGR also lowers the deviation ratios (NAV SingleUseful: 31.4%→15.8%; MAN UsefulOnly: 158.5%→121.8%). The distribution of these gains across regimes indicates where they originate. The pattern tracks the construction of the two losses: NoUseless rises by 75% and UsefulOnly by 120%, both far above the 31% on Full, and the two regimes are exactly the ones $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{suff}}$ target. A regularizer that enforced consistency uniformly, without conditioning on state, would instead predict comparable gains across the three regimes. ModDrop is the closest control we have for that alternative, since it perturbs modalities uniformly and independently of state: it does not reproduce the gains and drops to 7.5% on clean MAN input. EGR does not win everywhere. It trails vanilla on 3 of 83 segment-regime cells under corruption but on 8 of 47 segments under clean inputs, and its largest regression is one NAV segment that vanilla always completes and EGR never does.

D. Real-Robot Validation (Q3)

We evaluate EGR on two real-robot platforms with different embodiments and modality combinations (Appendix I, Fig. 3). The bi-manual platform consists of two Kinova arms and three RGB cameras. The vision-tactile platform consists of a single MELFA ASSISTA arm, one RGB camera, and two GelSight tactile sensors. Under the regimes of Sec. III-A1, the bi-manual platform has two useful sensors, and the tactile platform has one, the right GelSight. 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

TABLE I

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. EVERY CELL AVERAGES 20 ROLLOUTS PER SEGMENT; SR CARRIES ITS CROSS-SEGMENT STANDARD ERROR.

	NAV ($n = 11$)		MAN ($n = 36$), SR↑ / EC↑ / Dist↓		
Method	Full	SingleUseful	Full	NoUseless	UsefulOnly
vanilla $\pi_{0.5}$	**42.27** $\pm$ 14.8	15.45 $\pm$ 9.3	12.50 $\pm$ 3.8 / **52.78** / **0.25**	9.44 $\pm$ 3.4 / 47.22 / 0.26	2.78 $\pm$ 2.8 / 22.78 / 0.45
ModDrop	40.00 $\pm$ 14.1	15.45 $\pm$ 9.2	7.50 $\pm$ 3.2 / 40.97 / 0.29	6.81 $\pm$ 3.1 / 38.89 / 0.29	2.78 $\pm$ 2.8 / 22.36 / 0.47
EGR (ours)	40.91 $\pm$ 14.5	**37.27** $\pm$ 13.8	**16.39** $\pm$ 4.1 / 50.00 / 0.26	**16.53** $\pm$ 4.2 / **53.89** / **0.25**	**6.11** $\pm$ 3.1 / **30.14** / **0.37**

non-synthetic probe of robustness. Fig. 4 in Appendix I plots these results.

1) Bi-Manual Vision Platform: We run 4 tasks $\times$ 4 conditions $\times$ 10 trials, giving $N = 40$ pooled trials per condition. EGR improves Clean from 70% to 80% (28/40 to 32/40), NoUseless from 52.5% to 90% (21/40 to 36/40, $p = 0.0004$, Fisher exact), UsefulOnly from 25% to 72.5% (10/40 to 29/40, $p < 0.0001$), and RealDistractor from 30% to 85% (12/40 to 34/40, $p < 0.0001$).

2) Vision-Tactile Platform: We run 2 tasks (pencil serial code, board crack) $\times$ 3 conditions $\times$ 10 trials, giving $N = 20$ pooled trials per condition. EGR preserves Clean within trial noise (90% to 85%), improves SingleUseful from 40% to 90% (8/20 to 18/20, $p = 0.0022$, Fisher exact), and improves RealDistractor from 55% to 70% (11/20 to 14/20).

IV. DISCUSSION AND LIMITATIONS

Our results indicate that, in low-demonstration settings, the training data largely determine which sensors a policy learns to rely on. From these findings, we draw three key lessons.

More sensors can make a policy more fragile. With limited demonstrations, all sensors are observed together under a narrow range of conditions, encouraging vanilla $\pi_{0.5}$ to overfit to their joint patterns rather than identify task-relevant sensors. Adding sensors introduces more opportunities for distribution shifts at test time, even when those sensors are unnecessary for the task. Consequently, policy performance degrades both when irrelevant sensors are corrupted (NoUseless: $12.5\% \rightarrow 9.4\%$ on MAN) and when only the useful sensor remains (UsefulOnly: $12.5\% \rightarrow 2.8\%$).

Clean-input success hides both failures. On NAV, vanilla $\pi_{0.5}$ and EGR achieve comparable clean success rates (42.3% and 40.9%, respectively), yet differ substantially under SingleUseful (15.5% versus 37.3%). Thus, clean-input evaluations can mask sensitivity to irrelevant sensors and insufficient reliance on task-relevant sensors. Evaluating sensor contributions through targeted corruption and single-sensor retention reveals vulnerabilities that standard clean evaluations overlook.

Training guidance with task-relevant evidence. Generic augmentation alone is insufficient: ModDrop perturbs all sensors indiscriminately and fails to address either failure mode. Simple task knowledge, identifying the relevant object and whether a sensor observes it, proved sufficient to selectively gate the training losses. This improves robustness without requiring additional demonstrations or architectural changes.

Limitations: Our evidence instantiations rely on task structure: visible task-relevant geometry for vision and reference contact signatures for tactile sensing. This approach applies to settings where task-relevant signals can be specified a priori. For the real-robot experiments, evidence construction required 423 manually annotated bounding boxes (Appendix E). Extending EGR to learn evidence directly from data, through information-theoretic objectives or self-supervised feature discovery, is an important direction for future work. However, saliency derived from the trained policy may inherit its modality entanglement, motivating the use of an independent evidence source. Another open question is whether the observed gains arise primarily from evidence gating, the consistency objectives, or their combination. A controlled study that varies the evidence signal while keeping the objectives fixed could help disentangle these effects. Comparing EGR against attention-based sensor-selection methods would also provide further insight. Although EGR uses continuous, per-frame evidence scores and add no regularization on ambiguous frames, the benchmark currently relies on binary, segment-level labels. Evaluating time-varying sensor relevance therefore remains future work. Finally, our simulation benchmark covers 12 of the 50 BEHAVIOR-1K tasks, leaving broader task coverage for future investigation.

V. CONCLUSION

We investigated how multi-sensor VLA policies learn to rely on different sensor inputs under limited demonstrations and how this reliance affects their robustness to sensor perturbations. We introduced Evidence-Gated Regularization (EGR), a training-time framework that reduces modality entanglement by encouraging policies to rely on task-relevant sensor information. Evaluations across BEHAVIOR-1K tasks and two real-robot platforms demonstrate improved robustness to synthetic sensor corruption and physical distractors without additional demonstrations or architectural modifications. Our findings highlight that clean-input performance alone is insufficient to assess multi-sensor robustness and emphasize the importance of evaluating whether policies rely on the sensors that are truly relevant to the task.

REFERENCES

[1] 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.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," in *Conference on Robot Learning (CoRL)*. PMLR, 2025, pp. 2679–2713.

[3] 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," in *Conference on Robot Learning (CoRL)*. PMLR, 2025, pp. 17–40.

[4] 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*, vol. 2, no. 11, pp. 665–673, 2020.

[5] 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*, 2021, pp. 1574–1583.

[6] 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)*. IEEE, 2024, pp. 6892–6903.

[7] 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)*. IEEE, 2024, pp. 2855–2861.

[8] 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*. PMLR, 2017, pp. 249–261.

[9] 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," in *International Conference on Learning Representations (ICLR)*, 2026.

[10] 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)*. IEEE, 2023, pp. 1–7.

[11] 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.

[12] 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*. PMLR, 2023, pp. 80–93.

[13] 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.

[14] 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*. PMLR, 2023, pp. 2165–2183.

[15] 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.

[16] P. Hosseini, S. Nawathe, M. Moayeri, S. Balasubramanian, and S. Feizi, "Spurlens: Automatic detection of spurious cues in multimodal llms," *arXiv preprint arXiv:2503.08884*, 2025.

[17] 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.

[18] 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*, 2025, pp. 3805–3825.

[19] S. Fei, S. Wang, J. Shi, Z. Dai, J. Cai, P. Qian, L. Ji, X. He, S. Zhang, Z. Fei, *et al.*, "Libero-plus: A progressive robustness benchmark for visual-language-action models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026, pp. 38 574–38 583.

[20] 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.

[21] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 695–12 705.

[22] X. Peng, Y. Wei, A. Deng, *et al.*, "Balanced multimodal learning via on-the-fly gradient modulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8238–8247.

[23] C. Du, J. Teng, T. Li, *et al.*, "On uni-modal feature learning in supervised multi-modal learning," in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 8632–8656.

[24] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, "Deep multimodal learning with missing modality: A survey," *Transactions on Machine Learning Research*, 2026.

[25] 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*, vol. 47, no. 2, pp. 742–754, 2024.

[26] R. Feng, D. Hu, W. Ma, and X. Li, "Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation," in *Conference on Robot Learning (CoRL)*. PMLR, 2025, pp. 340–363.

[27] 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)*. IEEE, 2025, pp. 9499–9506.

[28] 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.

[29] 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," in *Robotics: Science and Systems (RSS)*, 2026.

[30] 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.

[31] Y. Yang, S. Cheng, Y. Fang, H. Bharadhwaj, M. Ding, G. Bertasius, and D. Szafir, "Lilo-vla: Compositional long-horizon manipulation via linked object-centric policies," *arXiv preprint arXiv:2602.21531*, 2026.

[32] Z. Cao, D. Tian, R. Guan, Y. Mu, X. Sun, S. Liang, D. Liu, T. Huang, Y. Yue, H. Ding, *et al.*, "Tactile-based multimodal fusion in embodied intelligence: A survey of vision, language, and contact-driven paradigms," *arXiv preprint arXiv:2605.17336*, 2026.

[33] L. Chen, Z. Lyu, and W. J. Beksi, "Reconvla: An uncertainty-guided and failure-aware vision-language-action framework for robotic control," *arXiv preprint arXiv:2604.16677*, 2026.

[34] 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*. Springer, 2024, pp. 38–55.

[35] 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 (ICLR)*, 2025.

APPENDIX A
RELATED WORK

A. Shortcut Learning in Multimodal LLMs

Current Vision-Language-Action (VLA) policies and broader multimodal large language models (MLLMs) [2], [3], [11], [14], [15] 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. [4] 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. [5] introduced VQA-CE to expose multimodal shortcuts involving question-image co-occurrence. Recent work characterizes similar effects at MLLM scale [16]–[18] and in VLA models [19], [20]. 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.

B. 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 [21]–[23], motivating balancing and missing-modality-robust techniques [24], [25]. In robot manipulation, recent work weights modalities per stage or state via learned attention [26]. In VLA models, recent methods improve robustness through run-time distractor inpainting [27], attention-pathway restructuring [28], supervised attention specialization [29], and training on diverse visual clutters [30], [31]. Tactile sensing raises the same question for a sensor whose reliability varies with contact rather than with view, and a recent survey maps how vision, language, and touch are fused in embodied policies [32]. Closer to our formulation, ReconVLA [33] attaches a calibrated per-step confidence to a VLA policy, but spends that signal on anticipating failures and flagging unsafe states rather than on deciding which sensor to act upon. These methods rely on architectural changes, run-time intervention, post-hoc confidence, or end-to-end learned attention, none grounding modality selection in per-frame task structure.

APPENDIX B
VISION EVIDENCE 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 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.

APPENDIX C
TACTILE EVIDENCE 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 [11]. 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.

APPENDIX D
IMPORTANCE-SAMPLED ESTIMATOR

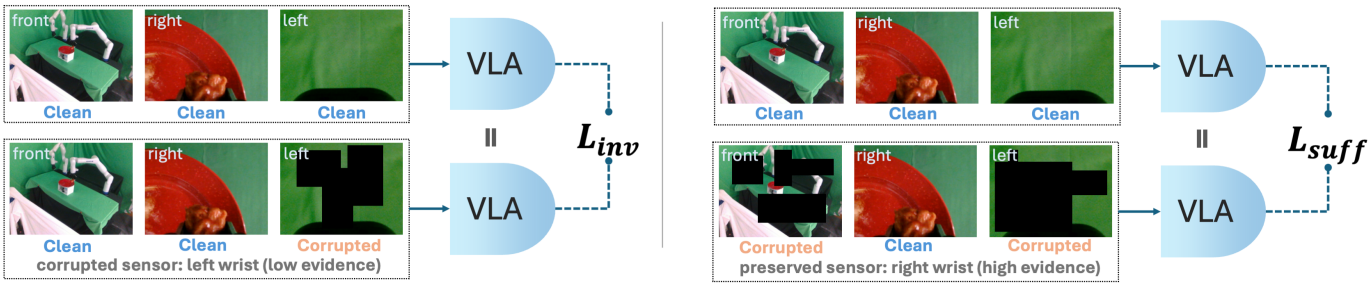


Fig. 2. The two evidence-gated loss terms. **Left:** $\mathcal{L}_{\text{inv}}$ enforces invariance when a low-evidence sensor is corrupted. **Right:** $\mathcal{L}_{\text{suff}}$ enforces sufficiency when only a high-evidence sensor is preserved.

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).

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

The same construct yields $\hat{\mathcal{L}}_{\text{suff}}$ using w_{suff} and W_b^{suff}. Rescaling the sampled term by W_b cancels its sampling probability, so both estimators are unbiased and preserve the expected gradient of the full-sum loss.

A. *Unbiasedness of the Estimator*

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

$$\ell_b^{\text{inv}} = G_b \sum_{m\in\mathcal{M}} w_{\text{inv}}(E_{b,m}) \cdot \left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m)}) \right\|_2^2. \tag{6}$$

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

$$p_{b,m}^{\text{inv}} = \frac{w_{\text{inv}}(E_{b,m})}{W_b^{\text{inv}}}, \qquad \sum_{m\in\mathcal{M}} p_{b,m}^{\text{inv}} = 1. \tag{7}$$

Rewriting the per-sample loss using this distribution,

$$\ell_b^{\text{inv}} = G_b \cdot W_b^{\text{inv}} \cdot \sum_{m\in\mathcal{M}} p_{b,m}^{\text{inv}} \cdot \left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m)}) \right\|_2^2 = G_b \cdot W_b^{\text{inv}} \cdot \mathbb{E}_{m\sim p_b^{\text{inv}}} \left[\left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m)}) \right\|_2^2 \right]. \tag{8}$$

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

$$\hat{\ell}_b^{\text{inv}} = G_b \cdot W_b^{\text{inv}} \cdot \left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m_b^\star)}) \right\|_2^2. \tag{9}$$

Taking expectation over $m_b^\star$,

$$\mathbb{E}_{m_b^\star\sim p_b^{\text{inv}}} \left[\hat{\ell}_b^{\text{inv}} \right] = G_b \cdot W_b^{\text{inv}} \cdot \mathbb{E}_{m\sim p_b^{\text{inv}}} \left[\left\| v_\theta(o_b) - v_\theta(\tilde{o}_b^{(m)}) \right\|_2^2 \right] = \ell_b^{\text{inv}}. \tag{10}$$

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), 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.

APPENDIX E
IMPLEMENTATION AND EVIDENCE-CONSTRUCTION DETAILS

Vision evidence is computed from BEHAVIOR-1K's per-frame instance segmentation in simulation, and from Grounding DINO v2 [34] with manually-prompted bounding boxes followed by SAM2 [35] 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. However, the framework is agnostic to the specific choice of corruption. Constructing evidence for the real-robot data took 423 boxes on 14 of 107 episodes, about 0.1% of frames, from which SAM2 propagates labels to the rest, plus 10 to 20 minutes of segmentation per task on one GPU. A new task needs one configuration entry naming its focal and interaction objects.

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 II 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.

TABLE II
EGR-SPECIFIC HYPERPARAMETERS.

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 $a_{t,m}^{\text{focal}}$
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

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 III). The deployment-time operator used in simulation evaluation applies a center-biased black erasure that approximates a large occlusion (Table IV). Both operators sample rectangles independently per camera.

TABLE III
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 $\{2, 3, 4, 5\}$
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 IV
SIM-DEPLOYMENT CORRUPTION OPERATOR USED IN SUITE 1 AND SUITE 2 TO INSTANTIATE THE NOUSELESS, USEFULONLY, AND SINGLEUSEFUL REGIMES.

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 V lists the training-loop configuration. All EGR simulation runs warm-start from the $\pi_{0.5}$ checkpoint provided by [13], pre-trained on the 12 BEHAVIOR-1K Challenge tasks; only LoRA adapter weights are updated.

TABLE V
TRAINING-LOOP CONFIGURATION FOR EGR SIMULATION RUNS.

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

APPENDIX F
BENCHMARK CONSTRUCTION

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 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.

A. 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 VI reports the segment count after each filter, broken out by skill type.

Selecting segments by evidence fixes the test conditions, not the outcome. One cannot ask whether corrupting an uninformative sensor degrades a policy without first identifying such a sensor, so some evidence-based selection is unavoidable for this diagnosis. The thresholds (Appendix F) are stated explicitly and applied identically to all three methods, and the useful and useless labels come from simulator segmentation and dataset annotations rather than from any model output. Selection therefore determines which segments enter the benchmark, while the reported comparison between methods remains unaffected by it.

TABLE VI
SEGMENT COUNTS THROUGH THE THREE VISION-EVIDENCE FILTERS.

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

B. 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

TABLE VIII

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. RATIOS CARRY THE CROSS-SEGMENT STANDARD ERROR OVER $n = 11$ NAV AND 36 MAN SEGMENTS.

	NAV			MAN				
Method	Full	SingleUseful	Ratio	Full	NoUseless	Ratio	UsefulOnly	Ratio
vanilla $\pi_{0.5}$	0.0616	0.0809	$+31.4 \pm 9.3\%$	0.0020	0.0022	$+12.7 \pm 4.9\%$	0.0052	$+158.5 \pm 28.4\%$
ModDrop	0.0853	0.1172	$+37.4 \pm 14.4\%$	0.0024	0.0026	$+10.9 \pm 2.7\%$	0.0056	$+137.7 \pm 19.8\%$
EGR (ours)	0.0780	0.0903	$+\mathbf{15.8} \pm 4.7\%$	0.0023	0.0024	$+\mathbf{5.1} \pm 2.7\%$	0.0051	$+\mathbf{121.8} \pm 25.1\%$

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 VII lists each threshold, its numeric value, and which filter consumes it.

TABLE VII

THRESHOLD VALUES USED IN THE FILTER PIPELINE.

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

C. 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.

APPENDIX G
SUITE 1: ACTION-LEVEL DIAGNOSIS

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. For a segment with frames $\mathcal{F}$ and ground-truth action chunk $a_t^\star$,

$$\Delta = \frac{1}{|\mathcal{F}|} \sum_{t \in \mathcal{F}} \Big(\big\| \pi_\theta(o_t^{\text{cor}}) - a_t^\star \big\|_{\mathcal{D}}^2 - \big\| \pi_\theta(o_t^{\text{full}}) - a_t^\star \big\|_{\mathcal{D}}^2 \Big). \tag{11}$$

The vanilla $\pi_{0.5}$ rows of Table VIII, 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, matching the nuisance sensitivity condition, and keeping only the useful modality produces a $+158.5\%$ ratio, matching single-modality insufficiency. ModDrop shows comparable ratio patterns and higher absolute action deviation, indicating that random modality dropout does not eliminate this entanglement.

The action-level diagnostic agrees with the Suite 2 gains in Table I: EGR's action-deviation ratios in Table VIII are consistently lower than vanilla and ModDrop (NAV SingleUseful: 31.4%→15.8%; MAN UsefulOnly: 158.5%→121.8%).

APPENDIX H
ADDITIONAL SIMULATION ANALYSIS

Some segments are hard enough that no method completes them in any regime. On the segments that at least one method completes, MAN Full moves from 26.5% to 34.7%, NoUseless from 20.0% to 35.0%, UsefulOnly from 5.9% to 12.9%, and NAV SingleUseful from 45.0% to 68.3%. EGR therefore converts partial competence into completion rather than solving

segments that were out of reach. Across all 11 NAV segments, EGR wins on 4, ties on 6, and loses on 1. The stage-level record shows what the gain consists of. We score each rollout by the stage it reaches, from never approached, to approached, to contacted, to success. Under UsefulOnly, 88 of the 244 trials on which vanilla never approaches the object advance to a later stage under EGR, 11 fall back, and no vanilla success is lost. The gain therefore comes mainly from the policy re-engaging with the task instead of stalling, the behavior we diagnose as single-modality insufficiency in Sec. I.

APPENDIX I
REAL-ROBOT PLATFORMS AND RESULTS

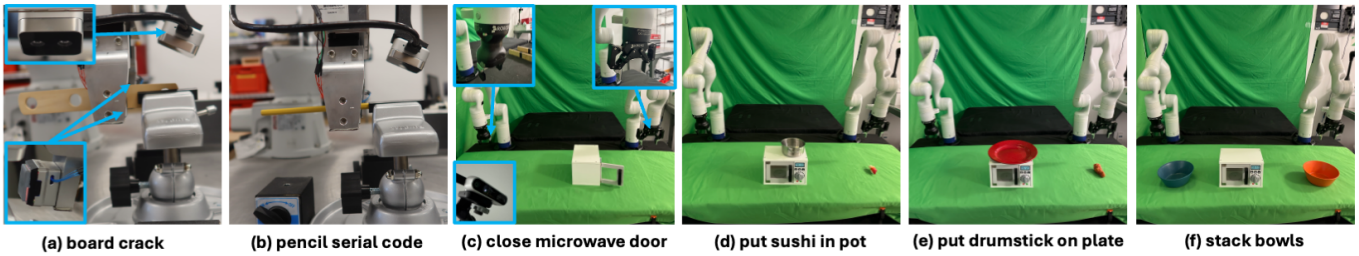


Fig. 3. Real-robot platforms and tasks. **(a, b)** Vision-tactile platform. **(c-f)** Bi-manual platform. Sensor-detail zoom-ins are shown in (a) and (c).

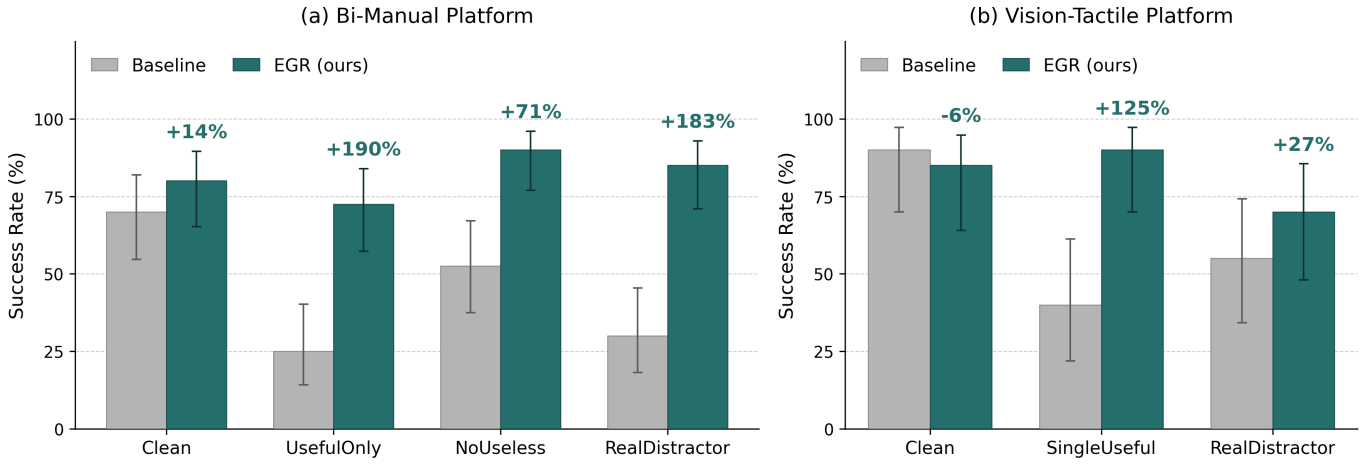


Fig. 4. Real-robot results across two platforms (10 trials per condition). Each pair of bars compares the vanilla baseline (left, gray) and EGR (right, teal); relative gain is shown above each pair.

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 tasks (Fig. 3c-e) and one long-horizon bowl-stacking task (Fig. 3f). The largest gain occurs on RealDistractor, where the distractor is a physical object the policy has never seen, the strongest out-of-distribution signal among our conditions. Per task under RealDistractor, SR rises from 0.2, 0.3, 0.0 and 0.7 to 1.0, 0.7, 0.7 and 1.0, so the pooled gain does not rest on a single task. 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. We choose the real-robot tasks and assign the useful and useless sensor roles by hand, from the task structure and the physical sensor layout. Neither step uses the evidence-based filter that curates the simulation benchmark. On the bi-manual platform, the head and right-wrist cameras are useful and the left-wrist camera is useless. On the tactile platform, the right GelSight is the sole useful sensor. A sensor counts as useful when it observes the focal object or contacts the target feature during the task, the same notion the simulation labels use.

APPENDIX J
METRICS AND STATISTICAL ANALYSIS

A. Standard Errors and Confidence Intervals

This subsection reports per-cell uncertainty for every number used in Tables I and VIII and in Fig. 4. Aggregated means use $\mathrm{SE} = \mathrm{SD}/\sqrt{n}$ 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.

a) 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.

TABLE IX
SUITE 1 PER-CELL STATISTICS. STANDARD ERRORS ARE COMPUTED ACROSS SEGMENTS; RATIO CONFIDENCE INTERVALS ARE FROM 10,000 PAIRED BOOTSTRAP RESAMPLES OVER SEGMENTS.

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	$+121.82 \pm 25.10$ $[+74.0, +172.4]$

b) 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.

TABLE X
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.

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	$[13.6, 63.6]$

c) 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.

TABLE XI
SUITE 2 MAN PER-CELL STATISTICS FOR THE THREE REPORTED METRICS. $n_{\text{SEG}} = 36$ FOR EVERY CELL.

		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	$[0.27, 0.47]$

d) 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 $n_{\text{task}} = 4$.

TABLE XII
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.

	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

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

TABLE XIII
REAL TACTILE PLATFORM SR PER CONDITION, POOLED ACROSS THE 2 TASKS.

	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	$[0.48, 0.85]$

B. 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 $\Delta = \text{MSE}_{\text{corrupt}} - \text{MSE}_{\text{full}}$; 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: $^{*}\, p < 0.05$, $^{**}\, p < 0.01$, $^{***}\, p < 0.001$.

a) 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 ($p = 0.061$).

TABLE XIV
SUITE 1 PAIRED WILCOXON SIGNED-RANK TESTS ON PER-SEGMENT Δ. ONE-SIDED, EGR $\Delta <$ BASELINE Δ.

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^{*}

b) 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.

TABLE XV
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.

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

c) 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.

TABLE XVI
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.

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	$< 0.0001^{***}$
EE distance	UsefulOnly	EGR vs ModDrop	$< 0.0001^{***}$

d) 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.

TABLE XVII
REAL BI-MANUAL FISHER'S EXACT TEST (TWO-SIDED) ON POOLED TRIAL COUNTS ACROSS THE 4 TASKS.

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	$< 0.0001^{***}$
RealDistractor	12/40	34/40	13.22	$< 0.0001^{***}$

e) 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 XVII), indicating that the tactile result is underpowered rather than absent.

TABLE XVIII
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.

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