FROM MULTIMODAL INDICATORS TO INTERVENTION ELIGIBILITY: TESTING THE ASSUMPTIONS BEHIND ADAPTIVE INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal inference increasingly makes per-input decisions about which sensors, modality subsets, or experts to execute. Yet when adaptation underperforms, aggregate accuracy or routing gain does not reveal whether the problem lies in the action space, intervention usefulness, or the evidence used for selection. We introduce Perception Harness, a diagnostic framework that evaluates executable alternatives before controller development. With actions, fallback, evidence, utility, and harm conditions held fixed, it separates action identity, recoverability, and pre-action actionability. Three benchmarks reveal distinct development needs. On MMAct, degradation is detectable but masking is harmful on average, motivating fallback retention or masking redesign. On MCubeS, identical modality labels yield different learned action landscapes, calling for model-specific action-space profiling. On DeLiVER, substantial safe opportunity exists within frozen proposals, but admissible pre-action evidence exposes little of it, motivating evidence and proposal redesign. Locating the failed transition directs the next development step. The diagnosis motivates redesign or eligibility for separate controller validation. Eligibility is not deployment authorization.

1 INTRODUCTION

Multimodal systems increasingly decide which sensors, modality subsets, or experts to execute for each input. Robust fusion supports missing or varying modalities (Zhang et al., 2023; Li et al., 2025); dynamic systems select modalities (Hu et al., 2024; Du et al., 2026), expert combinations (Ajirak et al., 2025), or model paths (Wu et al., 2025). These capabilities make adaptive execution possible, but do not by themselves establish that a useful intervention policy can be built.

An observed signal—degradation, uncertainty, complementarity, or modality quality—often motivates designing a controller. That move assumes two further implications: the signal must correspond to useful executable alternatives, and those alternatives must be identifiable from information available before acting. Even the first step requires specifying an action beyond its modality label. If an intervention is ineffective, or its benefits cannot be identified in advance, a more powerful router addresses the wrong problem.

Perception Harness tests this chain before controller development (Figure 1). It holds executable actions fixed, measures their gains relative to a fallback, and distinguishes useful opportunity from its observability through legal pre-action evidence. Here, *counterfactual utilities* mean paired outcomes of frozen executable model actions on the same sample, not physical-world causal counterfactuals. The aim is to identify which component should be redesigned next.

Three benchmarks reveal different failures. MMAct shows detectable degradation without a useful average masking response. MCubeS shows that identical modality labels can define different learned action landscapes. Our primary case, DeLiVER, contains substantial useful and safe opportunity, yet the tested pre-action evidence exposes little of it. These findings motivate intervention redesign, executable-action profiling, and evidence redesign, respectively.

Our contributions are:

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

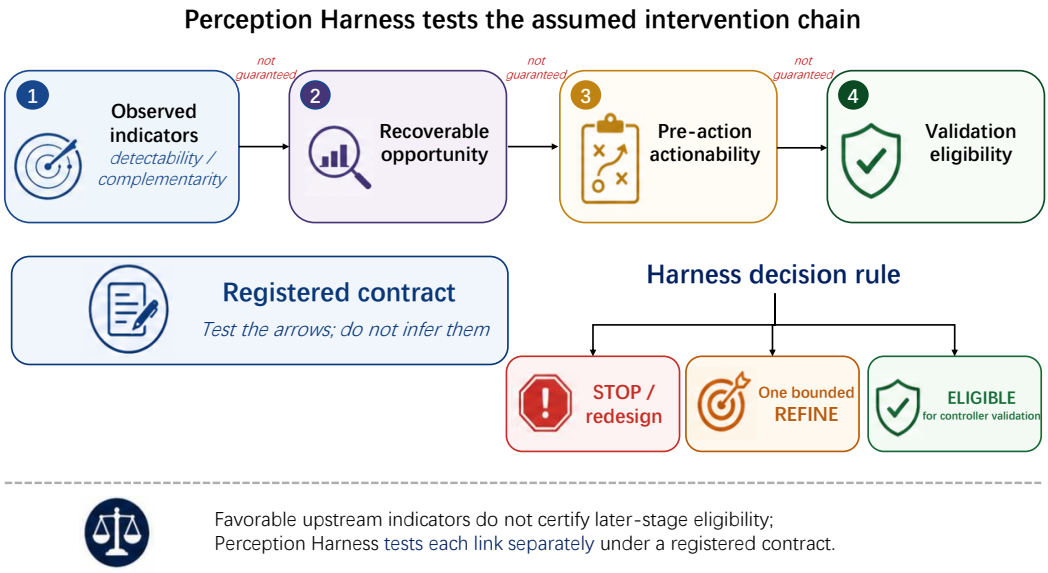

Figure 1: Perception Harness tests the assumed intervention chain under a registered action/evidence contract. Favorable upstream indicators do not guarantee recoverable opportunity, pre-action actionability, or controller-validation eligibility. The diagnosed failure determines whether development should STOP, undergo one bounded REFINE, or proceed to separate controller validation.

1. A formulation separating executable action identity, recoverability, pre-action actionability, and eligibility for later controller validation.
2. Perception Harness, a diagnostic methodology that localizes failed transitions using frozen executable actions and legal pre-action evidence.
3. Evidence of qualitatively different failures on MMAct, MCubeS, and DeLiVER, showing why the appropriate next development step depends on which transition fails.

2 Related Work

Robust multimodal inference seeks reliable prediction when modalities are missing, degraded, or available in different combinations. CMNeXt, Any2Seg, and MMSFormer develop fusion or representation mechanisms across multimodal configurations (Zhang et al., 2023; Zheng et al., 2024b; Reza et al., 2024), while MAGIC/MAGIC++, SimMLM, and CHARM further target resilient or modality-agnostic prediction (Zheng et al., 2024c;a; Li et al., 2025; Wen et al., 2026). Sensor-failure benchmarks complement these methods by evaluating robustness under corrupted or unavailable inputs (Liao et al., 2025). Together, this literature expands the set of configurations that a multimodal system can execute reliably. Perception Harness takes such trained alternatives as a registered action space and asks a different question: whether switching away from a fallback is beneficial for an individual input.

A second line of work makes the execution decision itself input-dependent. MOSEL selects modalities per inference request, DyMo performs test-time dynamic modality selection, and recent routing methods choose modality-processing or expert paths on a per-sample basis (Hu et al., 2024; Du et al., 2026; Ajirak et al., 2025; Wu et al., 2025). More recent risk-guided routing work, such as TIER-MoE, uses out-of-fold modality-risk estimates to inform sparse expert routing (Chang et al., 2026). These approaches study how a system should select among available alternatives once an action space and routing signal have been specified. Perception Harness addresses the preceding diagnostic question: whether useful alternatives exist, whether their benefit is observable from admissible pre-action evidence, and whether the resulting selected population satisfies the registered benefit–harm contract.

Reliability estimation, selective prediction, and risk control provide complementary tools for deciding when a candidate prediction or decision rule should be trusted. SelectiveNet and learning-to-defer methods introduce rejection or delegation mechanisms (Geifman & El-Yaniv, 2019; Mozannar & Sontag, 2020; Mozannar et al., 2023), while conformal risk control, selective conformal risk control, and SCoRE provide downstream risk-control mechanisms under their respective assumptions (Angelopoulos et al., 2024; Xu et al., 2025; Bai & Jin, 2026). Calibration research also seeks stronger reliability signals; for example, Invascal develops a single-pass evidential calibration mechanism for semantic segmentation (Turacan et al., 2026). Such signals could become candidate legal evidence in a future Harness instance. The Harness itself does not inherit their guarantees: it asks first whether the frozen action/evidence space exposes enough safe opportunity to justify downstream controller validation.

Off-policy evaluation and safe policy improvement provide a neighboring perspective on decision making from fixed data. Doubly robust off-policy evaluation estimates target-policy value from logged trajectories, while safe policy-improvement methods constrain changes relative to a baseline under their respective support and sequential-decision assumptions (Jiang & Li, 2016; Laroche et al., 2019). Perception Harness instead works with a registered set of frozen executable multimodal alternatives and sample-local paired utilities. Its estimand is not target-policy value, but whether the current action and pre-action evidence support a sufficiently useful and safe intervention population. The two perspectives therefore share a concern with evidence before policy deployment while operating under different data structures, assumptions, and development questions.

Taken together, prior work increasingly expands what multimodal systems can execute, how alternatives can be selected for individual inputs, and how reliability or risk can be controlled once a candidate rule exists. Perception Harness occupies the diagnostic boundary between these stages. It does not introduce another fusion architecture, routing model, or downstream risk guarantee; instead, it tests whether the currently available action space contains useful intervention opportunity, whether that opportunity is observable before execution, and whether the resulting candidate warrants separate controller validation. This makes the Harness complementary to stronger fusion, routing, uncertainty estimation, and risk-control methods: each can supply future actions or evidence whose intervention readiness can then be audited.

3 Diagnosing the Intervention Chain

3.1 Executable Actions and Legal Evidence

Let $i \sim \mathcal{D}$ denote the decision unit, $\mathcal{A}$ the actions the system can execute, and a_0 its static fallback. A modality label alone does not identify an action. The action also includes the sensing configuration m, learned execution f_θ, and any control parameters η:

$$a = (m, f_\theta, \eta). \tag{1}$$

Two specialists that consume the same sensors are consequently different actions when their checkpoints, preprocessing, or learned executions differ. For sample-local utility $q_i(a)$, define gain relative to fallback as

$$g_i(a) = q_i(a) - q_i(a_0), \qquad g_i(a_0) = 0. \tag{2}$$

Claims are relative to the decision unit, action family, utility, and fallback. Replacing the fallback changes every gain; adding alternatives changes the intervention universe rather than repairing an observed failure.

Legal evidence must be available before the candidate intervention is executed. A pre-acquisition policy cannot use ground truth, realized counterfactual utility, or a signal available only after the candidate sensor or specialist runs. Such post-acquisition evidence requires a separate contract.

3.2 The Assumed Intervention Chain

Figure 1 summarizes the assumption chain that motivates the Harness. The links are evaluated separately because success at an upstream diagnostic does not establish the operational requirement that follows it. Detectability concerns recognizing degradation. Recoverability asks whether any available alternative actually helps. Predictability—the pre-action form of actionability—asks whether

the system can identify that useful alternative before it is executed. Safety concerns the benefit and harm profile of the samples chosen for intervention. Only then is there a reason to ask whether the frozen candidate should enter a separate controller-validation study. Appendix A summarizes the non-implications and gives finite witnesses.

These are operational non-implication tests rather than causal-effect estimates: each stage requires evidence for its own estimand instead of inheriting validity from the preceding diagnostic.

3.3 Opportunity and Actionability

The central distinction is between opportunity and actionability. Opportunity asks how much beneficial or safe intervention could be found with access to counterfactual outcomes. Actionability asks how much of that opportunity can be recognized from information available before acting. Although both questions use the same samples and actions, they rank them using different information: realized outcomes for the first and decision-time evidence for the second. Write the adaptive proposal as

$$\widehat{a}_i = \pi(z_i), \qquad \mathcal{P} = \{i : \widehat{a}_i \neq a_0\}, \tag{3}$$

where z_i contains only legal evidence. Both the proposed actions and the resulting non-static population $\mathcal{P}$ are frozen before outcomes enter the safety diagnosis.

Proposal models and scores are evaluated out of group, keeping linked samples together. Their out-of-fold (OOF) outputs are fixed before the outcome-aware analysis, so audit outcomes cannot influence the proposal or score for that unit.

For a proposed action, write $g_i = g_i(\widehat{a}_i)$ and define material-harm indicators $h_{i,\delta} = \mathbf{1}[g_i \leq -\delta]$. In DeLiVER, selected switches must have positive mean gain, with one-point harm no greater than 15% and three-point harm no greater than 5%. Mean gain records benefit, while the two rates constrain the frequency of losses at different severities. All three are computed among selected switches. Coverage instead asks what fraction of the full population would be switched, so an empty or vanishingly small selection cannot appear operationally adequate merely by avoiding harm.

4 Perception Harness: Opportunity and Actionability

The proposal rule π recommends an action; the Harness diagnoses whether the frozen action/evidence space warrants controller validation. Opportunity belongs to the action family and fallback, actionability to evidence and proposals, and prospective risk guarantees to downstream validation.

4.1 Recoverability and Proposal Formation

The first question is whether the available actions contain a beneficial alternative:

$$V_{\text{oracle}}(\mathcal{A}) = \mathbb{E}_i \left[\max_{a \in \mathcal{A}} g_i(a) \right]. \tag{4}$$

Because the fallback belongs to $\mathcal{A}$, this oracle is nonnegative. The MMAct masking diagnostic instead evaluates the prescribed masking response, whose average gain can be negative; it is not the fallback-inclusive maximum. Little useful recoverability motivates action redesign. Adding actions cannot lower the maximum at fixed utility and fallback, but changes the claim's scope. With opportunity present, grouped-OOF proposals use legal z_i and freeze before diagnosis; gain, harmful-switch rates and oracle gap closed describe predictability without certifying safety.

4.2 Full-Information and Legal Observable Safe Opportunity

We next pose an outcome-aware existence question: if every proposed action's result were already known, what fraction of the entire population could be switched while meeting the benefit and harm conditions? Let N be the population size. With the registered one- and three-point harm tolerances,

full-information safe opportunity is

$$C^{\star}_{\mathrm{FI}} = \max_{\varnothing \neq S \subseteq \mathcal{P}} \frac{|S|}{N} \quad \text{s.t.} \\ \frac{1}{|S|}\sum_{i\in S} g_i > 0, \quad \frac{1}{|S|}\sum_{i\in S} h_{i,.01} \leq .15, \\ \frac{1}{|S|}\sum_{i\in S} h_{i,.03} \leq .05. \tag{5}$$

with value zero if no nonempty subset is feasible. Thus $C^{\star}_{\mathrm{FI}}$ asks whether the frozen action/proposal population contains a sufficiently large safe subset if outcomes were known. It is an opportunity ceiling, not deployable controller coverage.

Benefit and harm are assessed among selected switches; coverage uses the full population. Appendix C illustrates this distinction with a 100-unit example.

Proposition 1 (exact feasibility scan). For any subset size k, the top-k proposal-gain prefix maximizes mean gain and minimizes every registered threshold-harm count. Scanning all k therefore computes Eq. (5) exactly for this contract. This exactness is specific to the registered monotone gain and threshold-harm constraints; Appendix B gives the proof and the scope required for non-monotone constraints.

To study actionability, let $s_i = s(z_i, \widehat{a}_i)$ be a frozen legal score on non-static proposals, ordered so that smaller values are preferred, and define

$$u_i(t) = \mathbf{1}[i \in \mathcal{P},\ s_i \leq t], \qquad M(t) = \sum_{i=1}^{N} u_i(t). \tag{6}$$

Can a frozen legal pre-action score expose such a subset? Legal observable safe opportunity is the hindsight diagnostic

$$C_{\mathrm{safe}}(s) = \max_{t: M(t)>0} \frac{M(t)}{N} \quad \text{s.t.} \\ \frac{\sum_i u_i(t) g_i}{M(t)} > 0, \quad \frac{\sum_i u_i(t) h_{i,.01}}{M(t)} \leq .15, \\ \frac{\sum_i u_i(t) h_{i,.03}}{M(t)} \leq .05. \tag{7}$$

We scan distinct score thresholds, retaining all tied scores together. Outcomes enter only after the grouped-OOF score freezes, so the maximizing threshold measures hindsight representability, not deployable controller coverage or a prospective guarantee. Empty selections are ineligible; the Harness can retain the fallback. Every threshold-selected set is considered in Eq. (5), hence

$$C_{\mathrm{safe}}(s) \leq C^{\star}_{\mathrm{FI}}. \tag{8}$$

A gap localizes weak observability without ruling out future legal scores. Equality is possible, as the positive witness in Appendix A shows. Fixed-score scope and minimum reproducibility inputs are in Appendix C.

4.3 From Diagnosis to Development

Low recoverability motivates changing the actions. A large $C^{\star}_{\mathrm{FI}}$ but small $C_{\mathrm{safe}}(s)$ instead motivates better pre-action evidence or proposals. For DeLiVER, coverage below $c_{\mathrm{cont}} = 5\%$ stops development of the current candidate; coverage from 5% to below $c_{\mathrm{op}} = 15\%$ permits one bounded refinement; at least 15% permits separate controller validation. These are study-specific development criteria, fixed before the audit, not universal safety constants. Appendix F retains the complete procedure and permitted next steps.

5 EXPERIMENTAL DESIGN

5.1 THREE COMPLEMENTARY BENCHMARK ROLES

MMAct (Kong et al., 2019) tests whether recognizing sensor degradation yields a useful keep/mask response, using classification accuracy on held-out subject–condition units. MCubeS (Liang et al., 2022) tests action identity: published MMSFormer and CMNeXt specialists share RGB, RGB-A, RGB-A-D, and RGB-A-D-N labels but have independently learned executions. Its utility is image-level macro IoU, with native sequences kept together.

DeLiVER (Zhang et al., 2023) is the primary opportunity/actionability study. Expert Pool7 contains six non-static specialists and a static RGB-D-E-L fallback. Utility is image-level macro IoU over ground-truth-present classes; scene groups stay intact. The three benchmarks test different transitions, with no cross-benchmark comparison of utility scales.

5.2 SEPARATING CONFIRMATION FROM THE OBSERVABILITY AUDIT

B2 is the primary proposal evaluated on the original one-shot TEST. It uses 573-dimensional pre-acquisition RGB evidence and grouped-OOF utility predictors from the Ridge/HGB family. The separate fixed-MLP capacity-stress proposal uses the same evidence budget and six non-static targets. Its frozen grouped-VAL proposals supply both $C^{\star}_{\mathrm{FI}}$ and every learned C_{safe} row; these are not B2's TEST proposals.

The observability audit uses five-fold grouped OOF predictions, fixing each held-out proposal and legal score before evaluating its outcomes. It compares action-conditioned tabular, counterfactual tabular, hard-map spatial, and soft RGB evidence. Counterfactual outcomes provide training supervision, never decision-time inputs. SOFT-R1 preserves class probabilities and uncertainty summaries from a frozen RGB decoder, combined with action-conditioned proposal features. Its harm-prediction score remains a VAL-only diagnostic. Feature construction, score fitting, and evaluation ordering are retained in Appendix G.

The benefit, harm, and full-population coverage conditions are those in Section 3.3; the development criteria are in Section 4.3. All reported results retain their original evaluation status.

6 RESULTS

6.1 EARLIER FAILURES IN THE INTERVENTION CHAIN

MMAct: detectable degradation without a useful masking response. The supported degradation targets have macro AUROC 0.9317, but the prescribed oracle masking response averages -1.4573 accuracy points across 76 held-out-subject–condition units: 26 improve, 13 tie, and 37 decline. Strong detection therefore does not establish that the registered response is useful. This motivates retaining the original input or redesigning masking before further router optimization. The negative average concerns this masking diagnostic, not the fallback-inclusive oracle in Eq. (4) or every possible response to degradation.

MCubeS: modality labels do not determine executable actions. MMSFormer and CMNeXt specialists use the same four modality labels, yet their oracle gains are approximately $+1.45$ and $+1.02$ macro-IoU points. Oracle-action exact agreement is only 33.3333%, with 55.5556% pairwise action-preference agreement. The same sample can thus favor different modality subsets under different learned implementations. This 102-sample, 12-native-group comparison motivates profiling concrete models, checkpoints, preprocessing, and execution paths before routing. It probes action identity, not an architecture-causal effect or population-wide disagreement rate.

6.2 DELIVER: OPPORTUNITY WITHOUT SUFFICIENT PRE-ACTION ACTIONABILITY

DeLiVER has useful alternatives: the Expert Pool7 oracle gains $+3.37$ macro-IoU points on the original one-shot TEST. B2 gains only $+0.27$ points, closing 8.14% of that oracle gap, while 35.17%

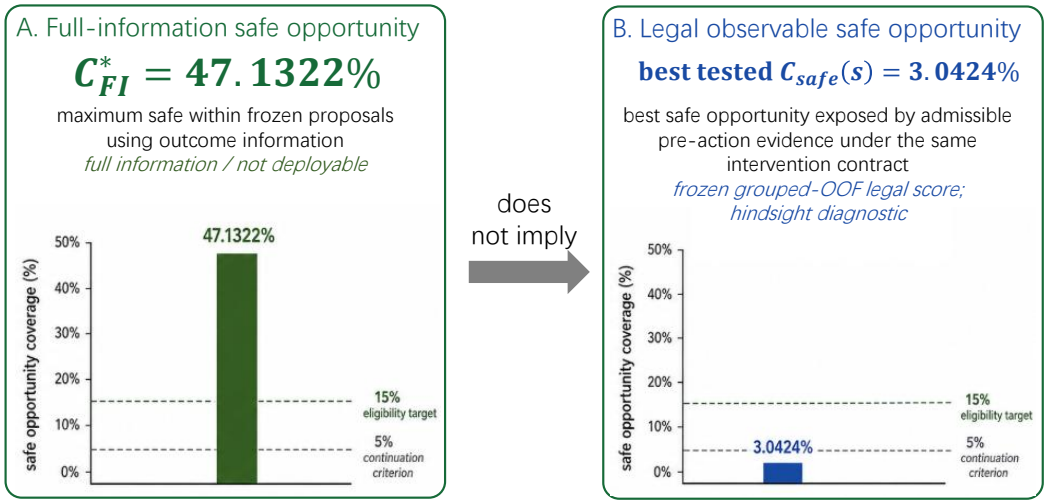

Figure 2: DeLiVER's central contrast on grouped VAL: full-information safe opportunity ($C^{\star}_{\mathrm{FI}} = 47.1322\%$) versus the best tested legal observable safe coverage ($C_{\mathrm{safe}}(s) = 3.0424\%$). Both are hindsight diagnostics using the full-population denominator. The 5% continuation and 15% eligibility criteria apply to C_{safe}, not to the outcome-aware ceiling.

of its switches lose at least one point (Table 1). The gap separates the existence of useful actions from identifying them in advance; B2's positive average gain also coexists with frequent harmful switches.

Table 1: Frozen one-shot DeLiVER TEST evidence (1,897 images; 118 scene groups). Oracle harm is inapplicable because it does not define a policy-selected switch population.

Method	Gain	Oracle gap closed	≥ 1 pp harm among switches
Expert Pool7 oracle	+3.37 pp	100%	N/A
B2 primary proposal policy	+0.27 pp	8.14%	35.17%

The grouped-VAL audit sharpens this contrast on the separate fixed-MLP proposal population. With outcomes known, $C^{\star}_{\mathrm{FI}} = 47.1322\%$ of the full population can be selected while satisfying the benefit and harm conditions. The best tested legal score exposes only $C_{\mathrm{safe}}(s) = 3.0424\%$ at a hindsight threshold (Figure 2). These quantities share proposals, utility, groups, and safety conditions; they differ in the information used to select a subset. Neither is prospective controller coverage.

The problem is not an absence of useful alternatives. The tested pre-action evidence does not expose enough of them. Safe subsets exist within the proposals, but their existence alone does not supply a usable selector. The low observable coverage localizes an evidence/proposal bottleneck without ruling out better legal representations.

6.3 What Should Be Redesigned?

The evidence progression is 1.3466%, 1.5461%, 1.3965%, and 3.0424% legal observable safe coverage (Table 2). All learned rows rank the same fixed-MLP proposals. Action conditioning represents the proposed intervention; counterfactual supervision adds labels for alternative actions; hard maps preserve spatial layout; soft RGB evidence retains confidence structure. The progression is not monotone: restoring layout alone does not improve on counterfactual tabular evidence, whereas soft probabilities yield the strongest tested result.

Even that result is below the 5% continuation criterion, so the current decision remains STOP. Because Eq. (7) already considers every threshold on the frozen score, further threshold tuning cannot exceed its hindsight ceiling under the same conditions. Representation/evidence or proposal redesign is therefore better motivated. The separate fixed-MLP capacity stress does not materially improve over

Table 2: Frozen DeLiVER VAL observability progression. The $C^{\star}_{\mathrm{FI}}$ row is a full-information ceiling; learned rows share the fixed-MLP proposal population (not B2 TEST proposals), safety contract, grouped OOF, and $C_{\mathrm{safe}}(s)$ estimand. None is deployable coverage.

Evidence / diagnostic	Information timing	Coverage (%)	Interpretation
$C^{\star}_{\mathrm{FI}}$: exact safe-subset ceiling	Full outcome	47.1322	feasible; non-deployable
Action-conditioned tabular	Legal pre-action	1.3466	below continuation
Counterfactual tabular	Legal pre-action	1.5461	below continuation
Hard-map spatial	Legal pre-action	1.3965	below continuation
Soft RGB evidence	Legal pre-action	3.0424	STOP; static fallback

B2, weakening a simple capacity explanation without establishing an information-theoretic limit. SOFT-R1 remains VAL-only; no further TEST confirmation was opened.

7 Discussion and Limitations

Diagnose the failed transition before optimizing the next module. The three benchmarks suggest different development priorities. MMAct motivates intervention redesign because detection does not establish a useful masking response. MCubeS motivates profiling executable actions because modality names conceal implementation-dependent preferences. DeLiVER motivates improving pre-action evidence and proposals because useful alternatives exist but their safe use is poorly observable. Aggregate accuracy alone does not distinguish these cases; the Harness does so by testing separate assumptions.

Opportunity and actionability are different estimands. The existence of a helpful action for a sample does not imply that the action can be identified before execution. Similarly, an outcome-selected safe subset does not imply a legal score can expose it. Comparing $C^{\star}_{\mathrm{FI}}$ with $C_{\mathrm{safe}}(s)$ holds the proposal population and constraints fixed while separating these questions. DeLiVER's gap is evidence about the tested representation and selector family, not a universal limit on multimodal information.

Eligibility precedes controller validation. Selective prediction and conformal risk control can govern a candidate controller under their respective assumptions (Geifman & El-Yaniv, 2019; Angelopoulos et al., 2024; Xu et al., 2025; Bai & Jin, 2026). They cannot manufacture information absent from its evidence or score; this does not imply that those methods fail. Our diagnostics establish neither prospective risk control nor safety under distribution shift. That subsequent validation becomes meaningful once sufficient pre-action actionability is demonstrated.

Limitations. We test a finite family of legal evidence. Sample-local segmentation utility is a proxy for, rather than a measurement of, physical deployment utility. Finite-sample uncertainty of the opportunity and observability estimates is not fully characterized; future analysis must respect scene and sequence grouping. The strongest SOFT-R1 result is VAL-only, and no controller-validation claim is made. Conclusions are indexed by the registered action family, fallback, utility, evidence timing, and groups. Appendix C discusses utility, threshold, and transfer scope.

8 Conclusion

Adaptive multimodal intervention should not begin with router optimization. First establish that executable alternatives can improve over the fallback, then test whether that opportunity is observable from information available before acting. Perception Harness separates these assumptions: MMAct motivates a better intervention, MCubeS demands model-specific action profiling, and DeLiVER calls for better evidence and proposals. Different failed transitions require different fixes. Controller design and prospective risk validation should follow only after the action/evidence space demonstrates sufficient pre-action actionability.

AI USE STATEMENT

Generative AI tools assisted with conceptual framing, research ideation, methodology and protocol planning, and the drafting and debugging of code and scripts. We also used them to organize and interpret results, discover and screen literature, draft and revise manuscript sections and LaTeX, check finite constructions and mathematical proofs, and improve presentation-oriented figures and plotting code. The authors manually ran every scientific experiment and decided which analyses, interpretations, and claims to retain. They reviewed the generated code and resulting outputs, and independently checked the citations and mathematical arguments included in the paper. No AI-generated synthetic benchmark or training dataset was used in the reported empirical work. The authors take full responsibility for the final text, analyses, code, artifacts, and scientific claims.

REFERENCES

Marzieh Ajirak, Oded Bein, Ellen Bowen, Dora Kanellopoulos, Avital Falk, Faith Gunning, Nili Solomonov, and Logan Grosenick. Learning to route: Per-sample adaptive routing for multimodal multitask prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2304.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2208.02814.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026.

Yu Chang, Anzhe Cheng, Chenwei Wu, Zhuoran Wang, Jiahao Chen, Tamoghna Chattopadhyay, Sophia I. Thomopoulos, Paul M. Thompson, Liyue Shen, and Paul Bogdan. TIER-MoE: Trust-informed expert routing via conditional modality risk for multimodal fusion in biomedical classification. arXiv preprint arXiv:2607.27289, 2026. URL https://arxiv.org/abs/2607.27289.

Siyi Du, Xinzhe Luo, Declan O'Regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *International Conference on Learning Representations*, 2026.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Bodun Hu, Le Xu, Jeongyoon Moon, Neeraja J. Yadwadkar, and Aditya Akella. MOSEL: Inference serving using dynamic modality selection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8872–8886, 2024. doi: 10.18653/v1/2024.emnlp-main.501.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661. PMLR, 2016.

Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. MMAct: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8658–8667, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Kong_MMAct_A_Large-Scale_Dataset_for_Cross_Modal_Human_Action_Understanding_ICCV_2019_paper.html.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661. PMLR, 2019.

Sijie Li, Chen Chen, and Jungong Han. SimMLM: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24068–24077, 2025.

Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19800–19808, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liang_Multimodal_Material_Segmentation_CVPR_2022_paper.html.

Chenfei Liao, Kaiyu Lei, Xu Zheng, Junha Moon, Zhixiong Wang, Yixuan Wang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Benchmarking multi-modal semantic segmentation under sensor failures: Missing and noisy modality robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/TMM-OpenWorld/html/Liao_Benchmarking_Multi-modal_Semantic_Segmentation_under_Sensor_Failures_Missing_and_Noisy_CVPRW_2025_paper.html.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Hussein Mozannar et al. Who should predict? exact algorithms for learning to defer to humans. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023. URL https://proceedings.mlr.press/v206/mozannar23a.html.

Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. MMSFormer: Multimodal transformer for material and semantic segmentation. *IEEE Open Journal of Signal Processing*, 5: 599–610, 2024.

Kerim Turacan, Hannes Reichert, Andrei Bolandut, and Konrad Doll. Invascal: Inverse-vacuity self-calibration for uncertainty-aware LiDAR range-view semantic segmentation. arXiv preprint arXiv:2606.00069, 2026. URL https://arxiv.org/abs/2606.00069.

Lekang Wen, Jing Xiao, Liang Liao, Jiajun Chen, and Mi Wang. CHARM: Collaborative harmonization across arbitrary modalities for modality-agnostic semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(13):10603–10611, 2026. doi: 10.1609/aaai.v40i13.38033.

Qiong Wu, Zhaoxi Ke, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Routing experts: Learning to route dynamic experts in existing multi-modal large language models. In *International Conference on Learning Representations*, 2025.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025. URL https://arxiv.org/abs/2512.12844.

Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Delivering_Arbitrary-Modal_Semantic_Segmentation_CVPR_2023_paper.html.

Xu Zheng, Yuanhuiyi Lyu, Lutao Jiang, Jiazhou Zhou, Lin Wang, and Xuming Hu. MAGIC++: Efficient and resilient modality-agnostic semantic segmentation via hierarchical modality selection. *arXiv preprint arXiv:2412.16876*, 2024a. URL https://arxiv.org/abs/2412.16876.

Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Learning modality-agnostic representation for semantic segmentation from any modalities. In *European Conference on Computer Vision*, 2024b. URL https://eccv.ecva.net/virtual/2024/poster/1593.

Xu Zheng, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. Centering the value of every modality: Towards efficient and resilient modality-agnostic semantic segmentation. In *European Conference on Computer Vision*, 2024c. URL https://eccv.ecva.net/virtual/2024/poster/773.

Table 3: Links tested by Perception Harness. Each row states an operational implication that requires evidence beyond association at the preceding stage.

Assumed link	Required conclusion	Why it is not guaranteed	Empirical role
Detectability ⇒ Recoverability	A registered response improves the unit	A fault may be observable while every allowed response is ineffective	MMAct negative control
Recoverability ⇒ Predictability	Pre-action evidence identifies a useful action	Counterfactual opportunity may not be observable before execution	DeLiVER proposal gap
Positive value ⇒ Safety	Selected switches meet benefit and harm constraints	Average gain can coexist with a harmful selected tail	DeLiVER harm audit
Marginal risk quality ⇒ selected risk quality	A score orders risk on the intervention population	Selection changes the estimand and population being controlled	DeLiVER selection audit

A Finite Non-Implication Constructions

The following finite witnesses formalize why adjacent diagnostics require separate tests. They are logical constructions, not empirical benchmarks.

Detectability without recoverability. Let a health flag identify every degraded unit perfectly, while all registered non-static actions have gain -1. Detectability is perfect but the registered family contains no beneficial recovery.

Recoverability without predictability. Consider two units with identical legal evidence and two non-static actions. Action one has gains $(+1, -1)$ and action two $(-1, +1)$. An oracle improves both units, but no deterministic function of the identical evidence can choose the helpful action for both.

Positive value without safety. A proposal population can have positive mean gain while more than the allowed fraction crosses a material harm threshold. Average benefit therefore does not imply compliance with the selected-switch harm contract.

Marginal risk quality without selected risk quality. A score may discriminate harm on the full candidate population while the proposal mechanism concentrates nearly indistinguishable high-risk units in the selected population. Marginal discrimination does not identify the selection-conditioned estimand.

Positive eligibility witness. For ten registered units, let a legal score expose exactly two proposed switches, both with positive gain and no registered harm. Full-information safe opportunity and legal observable safe coverage are both 20%, clearing the present continuation and eligibility criteria. The Harness returns eligibility for controller validation, not deployment authorization. This construction shows that the decision rule can return ELIGIBLE; it does not instantiate the subsequent controller-validation stage.

B Exact Gain-Prefix Feasibility

Sort proposal gains as $g_{(1)} \geq \cdots \geq g_{(m)}$ and let $S_k^* = \{(1), \ldots, (k)\}$. For any size-$k$ subset S,

$$\sum_{i \in S} g_i \leq \sum_{j=1}^{k} g_{(j)},$$

so the prefix maximizes total and mean gain. For a registered harm boundary $b = -\delta$, let r_b be the number of gains strictly above b. Every size-k subset contains at least $\max(0, k - r_b)$ gains at or below b; the prefix contains exactly this minimum. It therefore simultaneously minimizes all registered threshold-harm counts. If the prefix fails any monotone gain or harm-rate constraint, no

other size-k subset can pass; if it passes, it is a witness at coverage k/N. Scanning $k = 1, \ldots, m$ computes Eq. (5) exactly. With subgroup quotas, fairness constraints, or action-dependent costs that are not monotone in realized gain, gain-prefix exactness no longer applies. A future Harness instance would need to register and solve the constrained subset problem directly, for example through exact combinatorial optimization or a conservative relaxation, with the solver or relaxation itself treated as part of the evaluation contract. No tractability or new guarantee is claimed.

C Supplementary Operational Detail

C.1 Illustrative 100-unit example

Imagine 100 units for which the frozen rule proposes a non-static action on 60. Once their outcomes are visible, we discover that 40 of those proposals can be grouped together with positive mean benefit and acceptable harm. The population thus contains 40% full-information safe opportunity, although finding those 40 units required information unavailable at decision time. If a frozen pre-action score can isolate only three compliant units, its legal observable safe coverage is 3%. Safe proposals are present; the score reveals very few of them. In this example, $C^{\star}_{\mathrm{FI}}$ measures the opportunity visible with outcomes, while $C_{\mathrm{safe}}(s)$ measures what one legal score can separate before action, assessed retrospectively. The numbers are illustrative rather than empirical.

C.2 Fixed-score threshold scope and reproducibility

Once the scalar score s, proposal population, and safety conditions are fixed, Eq. (7) examines every selector of the form $s_i \leq t$. It is thus the largest hindsight coverage attainable by thresholding that particular ordering. Under its assumptions, separate calibration can select a prospective rule or attach a statistical guarantee, but it cannot create information absent from the frozen ordering. A larger ceiling would require a different score, evidence source, proposal population, selector family, or contract. This scope is deliberately narrow: richer scores and non-threshold selectors are new prospective candidates, and the statement makes no general claim that CRC, SCRC, SCoRE, or selective controllers fail.

Reproducing $C^{\star}_{\mathrm{FI}}$ requires one row per registered unit with its frozen proposal action and realized gain, together with the population size and harm thresholds and tolerances. Reproducing $C_{\mathrm{safe}}(s)$ adds a complete grouped-OOF legal score for each non-static proposal. A proposal–gain table and frozen OOF score table are therefore enough to repeat both diagnostics without running the original multimodal architecture. These are the minimum audit inputs, not a promise about public release.

C.3 A hypothetical REFINE outcome

Consider a hypothetical result of 9% legal observable safe coverage. The study could use REFINE to register one specific extension, such as calibrated RGB uncertainty features, before viewing its result. The decision unit, actions, fallback, utility, grouping, harm tolerances, and evaluation protocol would remain fixed, as would the new representation and its success criterion. That candidate would be evaluated once: failure returns STOP, whereas clearing the eligibility target permits a separate controller-validation study. We did not run this example; it illustrates how REFINE constrains the next investigation without inviting open-ended search.

C.4 Utility interpretation and evidence sequence

DeLiVER's sample-local utility is image-level macro IoU over classes present in the ground truth, which gives every image–action pair one gain relative to the same static fallback. One point of harm means one percentage point on this macro-IoU scale; it is not a unit of monetary, physical, or downstream control cost. Given a frozen prediction and annotation, the value is deterministic, yet we do not treat it as a noise-free or uniquely correct proxy for deployment utility. Annotation variability, class composition, or a different utility can change the gain landscape. An uncertainty-aware extension would have to register repeated or distributional utility measurements, or uncertainty bounds, before feasibility analysis instead of appending post-hoc uncertainty to the current setup.

The bounded evidence sequence leaves the proposal family fixed. Tabular action geometry tests whether the score represents action identity. Counterfactual supervision adds labels for alternative actions without changing the proposals; hard spatial maps restore layout; and frozen soft RGB probabilities retain confidence structure. Signals that require acquiring a candidate sensor would answer a post-acquisition question and are excluded from this pre-action sequence.

C.5 Contract mapping beyond segmentation

The framework can be mapped beyond segmentation, although we have not tested that transfer. Our decision unit is a segmentation image, utility is image-level macro IoU, and each action executes a modality specialist. A tracking or control study could instead register a scene, sequence, or control window; use local tracking quality or trajectory utility; and choose among executable sensing, perception, or control policies. Such a study would still need to specify action identity, fallback, evidence timing, grouping, harms, and coverage before inspecting outcomes. This example explains how to formulate the contract in another domain; it does not extend our segmentation evidence to tracking or control.

C.6 Evidence-sequence and confirmation detail

The bounded evidence sequence helps localize this gap. Action-conditioned tabular features describe which modalities a proposal uses. Counterfactual supervision adds outcomes for registered alternatives, hard-map spatial features preserve layout while discarding confidence, and soft RGB probabilities restore confidence structure. Their legal observable safe coverage values are 1.3466%, 1.5461%, 1.3965%, and 3.0424%, respectively (Table 2). Every learned row uses the same fixed-MLP proposal population, safety conditions, grouped-OOF evaluation, and estimand; none uses B2's TEST proposals. As a separate post-freeze capacity stress, the fixed MLP uses the same 573-dimensional legal evidence and six non-static targets and produces no material improvement over B2. These controls weaken a simple capacity explanation without establishing an information-theoretic limit. Confidence structure appears helpful in this sequence, although the final score remains retrospective and says nothing universal about richer representations.

Because SOFT-R1 did not clear the VAL continuation criterion, we did not open another TEST confirmation round. A further TEST run could measure transfer of the frozen score, but it could not retroactively satisfy the failed development rule; opening confirmation at that point would instead turn TEST into outcome-driven exploration. We report SOFT-R1 as a VAL-only diagnostic and limit the conclusion to the tested legal-evidence family. Safe opportunity remains present, while the present evidence is too weak to justify controller development.

C.7 Decision, risk control, and limitations

Section 7 explains the development value of STOP and the untested transition from eligibility to a validated controller. Downstream CRC, SCRC, SCoRE, or learning-to-defer would address selection and guarantees under their respective assumptions. The fixed-score ceiling here falls below continuation, so that downstream stage was not instantiated; a richer score or controller remains a new prospective candidate.

Several limitations bound the claim. We tested a finite family of legal evidence, and any richer score or selector must be frozen in a new prospective study. The strongest soft-evidence result is VAL-only; TEST confirmation would become appropriate only after a new candidate clears its development gate. Section 7 states the unestimated finite-sample uncertainty and grouping requirements. We do not study tracking, control, subgroup, or fairness behavior. Finally, the exact opportunity ceiling depends on the registered utility estimates. The observed separation applies to this tested design; it is neither a universal bound nor a causal claim that adaptation cannot work.

D FROZEN BENCHMARK DETAIL

D.1 MMACT NEGATIVE CONTROL

The frozen health-discrimination artifact reports AUROC/AUPRC pairs of 0.9763/0.9799 for bad phone, 0.9144/0.7657 for bad gyroscope, 0.9044/0.6601 for bad orientation, and 0.9474/0.3070 for relational conflict. Bad watch has no supported positive evidence and receives no AUROC. Macro AUROC over the three supported modality targets is 0.9317; the clean false-bad-modality rate is 0.1475 and the clean false-conflict rate is zero.

The recoverability artifact contains 76 held-out-subject–condition units: 26 positive, 13 tied, and 37 negative. Mean oracle-masking gain is -1.4573 pp. Frozen R2, R3, and R4 alternatives have mean gains -0.7451, -0.4144, and -1.5924 pp. The health and recovery artifact hash prefixes are `e12af67a` and `be0e9bd4`; manuscript assembly accesses no raw MMAct record.

D.2 MCUBES IMPLEMENTATION DEPENDENCE

The published MMSFormer and CMNeXt specialist families preserve the four registered modality labels and their frozen preprocessing/checkpoints. Their one-shot TEST audit contains 102 samples in 12 authoritative native sequence groups. Oracle gains are approximately $+1.45$ and $+1.02$ macro-IoU points. Exact oracle-action agreement is 33.3333% and pairwise action preference agreement is 55.5556%. The comparison holds labels fixed but does not isolate an architecture-causal effect.

D.3 DELIVER EVIDENCE STATUSES

The Expert Pool7 oracle and B2 primary proposal-policy numbers in Table 1 retain their original one-shot confirmation status. The full-information safe opportunity and all legal-observability rows are grouped-VAL diagnostics. The 3.0424% soft-evidence result is hindsight, non-deployable, and not separately TEST-confirmed. No TEST confirmation status is changed by this rewrite.

D.4 ALTERNATIVE UTILITY SENSITIVITY (FROZEN, VAL-ONLY)

Table 4: Frozen VAL proposal decisions under three sample-local utilities. Numerical one- and three-point harm thresholds are descriptive across alternative utilities and are not asserted to have equal operational meaning.

Utility	Proposal	Oracle pp	Gain pp	OGC	Coverage	Harm	Severe
Macro IoU	B2 primary	2.749	+0.370	13.5%	55.8%	32.1%	18.1%
Macro IoU	fixed MLP stress	2.749	+0.391	14.2%	60.4%	33.6%	19.1%
Macro Dice	B2 primary	1.935	+0.468	24.2%	55.8%	23.8%	7.4%
Macro Dice	fixed MLP stress	1.935	+0.355	18.3%	60.4%	27.9%	9.7%
FWIoU	B2 primary	0.518	+0.166	32.0%	55.8%	4.6%	1.1%
FWIoU	fixed MLP stress	0.518	+0.106	20.4%	60.4%	6.4%	1.7%

This `VAL_ONLY_POST_FREEZE` audit reuses the frozen B2 primary proposal policy and fixed MLP capacity-stress proposal decisions; it retrains neither policy and is not a C_{safe} analysis. Macro Dice averages class Dice over GT-present classes, whereas frequency-weighted IoU weights class IoU by non-ignore GT pixel frequency. FWIoU therefore emphasizes large regions and yields a smaller measured harmful tail. Identical numeric one- and three-point thresholds do not have identical operational meaning across these utility scales. The table is sensitivity evidence for the registered utility contract, not a declaration that B2 or the fixed MLP is safe under FWIoU. Values are copied from frozen artifact `UTILITY_SENSITIVITY.csv` (SHA256 prefix `f76eba5d`); no row-level result was recomputed.

E CONTRACT SCOPE AND REPRODUCIBILITY

Claims are indexed by decision unit, action identity, fallback, utility, evidence timing, grouping, harm tolerances, and coverage criteria. A new action family or router is a prospective study. Manuscript

numbers are copied from frozen aggregate artifacts. No training, inference, feature extraction, bootstrap, or threshold search is performed during manuscript assembly. The source workspace records the claim hierarchy, old-to-new content map, page budget, numerical sources, and final build audits.

F Development Decision Procedure

The coverage parameters c_{cont} and c_{op} govern the three outputs in Algorithm 1. Table 5 specifies the permitted next step. Each benchmark retains its registered recoverability criterion; the DeLiVER numerical contract is not imposed on every benchmark.

Algorithm 1 Perception Harness diagnostic

Input: units/groups, actions and fallback, legal evidence, utility, harm contract, continuation and eligibility criteria.

1: Validate action identity and freeze the action/evidence/utility contract.
2: Assess registered recoverability / oracle opportunity (Eq. (4)).
3: If registered recoverability fails, return STOP / redesign actions.
4: Fit/evaluate legal-evidence proposals with grouped OOF; freeze proposals.
5: Compute $C^{\star}_{\text{FI}}$ on frozen proposals (Eq. (5)).
6: Freeze the grouped-OOF legal score; compute $C_{\text{safe}}(s)$ (Eq. (7)).
7: If $C_{\text{safe}} < c_{\text{cont}}$, return STOP and retain the fallback.
8: Else if $C_{\text{safe}} < c_{\text{op}}$, return one bounded REFINE.
9: Else return ELIGIBLE_FOR_CONTROLLER_VALIDATION.

Table 5: Operational meaning of the three Harness outputs. Each transition defines a prospective study boundary rather than permission to reinterpret the current result.

Output	Trigger	What may change next	What remains frozen
STOP	$C_{\text{safe}} < c_{\text{cont}}$	A new prospective action, evidence, or proposal design in a new study	Current action/evidence contract and result
REFINE	$c_{\text{cont}} \le C_{\text{safe}} < c_{\text{op}}$	Exactly one prospectively registered bounded refinement family	Decision unit, actions, fallback, utility, groups, harm contract, and evaluation protocol
ELIGIBLE	$C_{\text{safe}} \ge c_{\text{op}}$	A separate grouped controller-validation study	Candidate representation, proposal rule, and action contract entering validation

Appendix C illustrates REFINE with a hypothetical 9% result and one prospectively registered extension.

A new domain should elicit harm magnitude from the smallest operationally material utility loss, harm tolerance from its acceptable frequency at each severity, and minimum useful coverage from the acquisition or controller overhead. Fix these before outcome-aware analysis. The present one-/three-point harms and 5%/15% coverage criteria instantiate this recipe, not universal defaults. Eligibility authorizes separate validation, not deployment.

G Evaluation and Implementation Details

G.1 Reference Systems and Action Sources

We select reference systems that expose multiple executable alternatives under a common task, permit sample-level comparison against a fixed fallback, and have sufficiently faithful implementations or frozen outputs to preserve their original semantics. The benchmarks therefore provide diagnostic action spaces rather than a cross-dataset leaderboard. MMAct pairs sensor-health evidence with a keep/mask action structure to test recoverability. On MCubeS, MMSFormer and CMNeXt expose comparable modality-subset labels through independent learned implementations, enabling an action-identity probe. DeLiVER's Expert Pool7 supplies a static fallback, multiple specialists, sample-level

counterfactual utilities, and legal pre-action evidence. Its oracle, B2, action-only controls, risk filters, and fixed-MLP stress isolate opportunity, proposals, action priors, ordering, and capacity.

G.2 Frozen Evaluation Protocol

Action families, checkpoints, fallbacks, utilities, groups, and confirmation protocols were frozen. B2 is the TEST-confirmed primary proposal, using 573-dimensional legal pre-acquisition RGB evidence and grouped-OOF utility predictors from the registered Ridge/HGB family. The fixed MLP is a separate post-freeze capacity-stress proposal using the same evidence budget and six non-static utility targets; its frozen VAL proposals supply every learned C_{safe} row in Table 2. Five-fold grouped OOF keeps scenes intact; outcomes enter feasibility and representability audits only after proposals and scores freeze. Manuscript values are copied from frozen aggregates.

SOFT-R1 combines spatial-pyramid class probabilities, entropy, margins, and maximum probabilities from frozen RGB decoder outputs with action-conditioned proposal features. Scene-grouped OOF histogram-gradient-boosting classifiers predict one- and three-point harm from all-action counterfactual supervision. Their maximum tolerance-normalized predicted risk ranks the frozen fixed-MLP non-static proposals, with lower scores preferred. This remains a VAL-only hindsight observability diagnostic, not a TEST-confirmed controller.

A harm point is one percentage point on the registered macro-IoU scale, not monetary, physical, or control cost (Appendix C). The bounded evidence sequence keeps the proposal family fixed while adding action geometry, counterfactual supervision, hard spatial maps, and then frozen soft RGB probabilities. Post-acquisition signals remain outside this pre-action contract.

The safety evaluation keeps benefit, harm, and coverage distinct. Selected switches need positive mean gain, while their one- and three-point harmful-switch rates may not exceed 15% and 5%. Legal observable safe coverage is then compared with the 5% continuation criterion and 15% eligibility target, both defined over the full population. No threshold is tuned in this manuscript, and a descriptive hindsight result never authorizes a controller.

G.3 Threshold criteria and evidence timing

Under the current harm conditions and frozen score, any continuation criterion above 3.0424% yields the same STOP decision. Because continuation already fails, the 15% eligibility target is inactive. Changing harm tolerances could change $C_{\mathrm{safe}}(s)$ itself; this observation establishes no harm-tolerance robustness.

Legal evidence also depends on timing and cost. A low-resolution depth probe available before committing to the full depth pipeline could be legal if its acquisition cost were specified. A quality statistic available only after that pipeline runs could not support the same pre-action decision. Transfer to another domain requires its own decision unit, costs, and harms.