EW-GET: ACCESS-CONDITIONED IDENTIFIABILITY OF FAILURE MECHANISMS IN SENSOR-TO-DECISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Failure diagnosis is fundamentally resolution-limited: which mechanisms can be distinguished depends on what an evaluator can observe. We therefore treat *diagnostic resolution*, rather than a prespecified failure label, as the primary statistical object. EW-GET formalizes this principle for learned sensor-to-decision systems. Exposed traces are mapped to acquisition-grounded witness distributions for information magnitude, directional fidelity, and decision overclaim. Equality of joint witness laws defines a population access-conditioned failure quotient; at finite resolution, direct operational indistinguishability induces a threshold graph whose connected components define the reportable partition. We prove that richer access refines the population quotient and give finite-sample recovery and calibrated-abstention guarantees for the declared operational partition. Controlled benchmarks recover the predicted Black/Grey/White refinement. In HM3D RGB-D navigation, richer interfaces sharply improve mechanism-label predictability, yet a generic MLP with comparable F1 does not by itself establish evidence-supported diagnostic resolution. Diagnosis-guided intervention improves aggregate recovery under a frozen repair library, while adverse representation-family effects show that identifiability does not imply intervention sufficiency. EW-GET thus turns diagnosis into an access-conditioned measurement problem with an explicit population target, finite-sample reporting rule, and bounded downstream use.

1 INTRODUCTION

Learned sensor-to-decision systems expose only partial evidence about their internal computations. A downstream error can therefore arise from sensing, representation, fusion, or decision while remaining observationally ambiguous. Existing robustness, calibration, and representation analyses characterize degradation, confidence, or encoded information (Goodfellow et al., 2015; Madry et al., 2018; Hendrycks & Dietterich, 2019; Guo et al., 2017; Ovadia et al., 2019; Alain & Bengio, 2017; Raghu et al., 2017; Kornblith et al., 2019), but do not determine which failure distinctions are identifiable under a particular evaluator interface.

We ask: given a task, acquisition experiment, candidate failure family, and observable interface, what is the finest failure distinction justified by the available evidence? This *diagnostic resolution* is access-conditioned. A black-box evaluator may merge sensor degradation with representation suppression, whereas intermediate access can justify finer distinctions. Access therefore does more than improve predictive accuracy: it induces a refinement order over scientifically supportable diagnostic claims.

EW-GET maps exposed traces to acquisition-grounded witnesses whose joint laws define the diagnostic target. It returns resolution-aware confidence sets and abstains when finer attribution is unsupported. Our contributions are: (1) **diagnostic resolution as an access-conditioned statistical target**, with a population target given by an observational failure quotient and a refinement order induced by richer access; (2) **finite-resolution recovery with calibrated abstention**, separating the exact population quotient from an operational threshold-graph partition; (3) **acquisition-grounded witness construction and validation** in controlled and embodied systems; and (4) **bounded down-**

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.

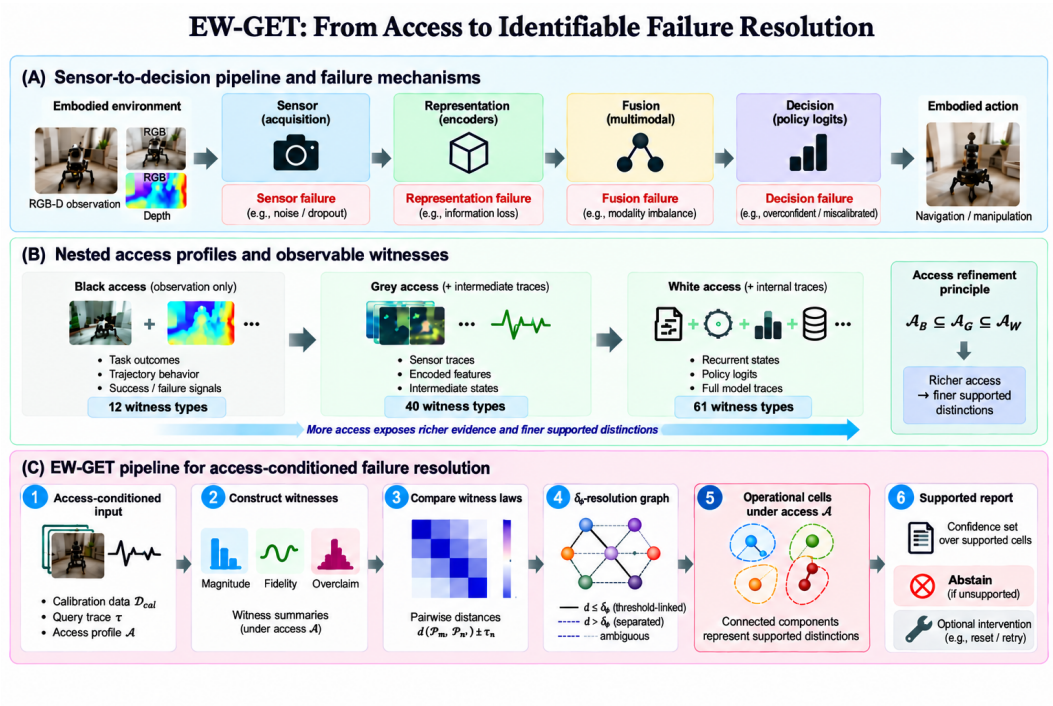

Figure 1: **EW-GET maps observational access to evidence-supported diagnostic resolution.** Nested interfaces expose richer witnesses; witness-law comparisons define access-conditioned population quotients, finite-resolution operational cells, abstention, and candidate interventions consistent with the reported resolution.

stream use, showing both intervention gains under a frozen repair library and the separate limitation that identifiability need not imply intervention sufficiency.

2 Background and Problem Formulation

Robustness, uncertainty, and OOD. These methods ask when behavior degrades or confidence becomes unreliable (Goodfellow et al., 2015; Madry et al., 2018; Hendrycks & Dietterich, 2019; Guo et al., 2017; Ovadia et al., 2019; Lakshminarayanan et al., 2017; Liu et al., 2020). EW-GET instead asks which candidate failure distinctions are supportable from the evaluator's observable evidence.

Fault diagnosis and testing. Classical diagnosis seeks to detect or isolate faults relative to specified models and tests (Chen et al., 2018; Pei et al., 2017; Isermann, 2006). EW-GET does not assume named mechanisms remain distinguishable under partial access; its target is the resolution justified by observable witness laws.

Probing and representation analysis. Probes and similarity methods ask what information is encoded or accessible in learned features (Alain & Bengio, 2017; Kornblith et al., 2019; Raghu et al., 2017). EW-GET treats exposed traces as evidence whose joint law determines supported diagnostic resolution.

Statistical experiments, selective reporting, and abstention. Blackwell comparison, Fisher information, and observability-related state estimation formalize what an experiment makes inferable (Blackwell, 1951; Fisher, 1922; Amari, 2016; Kalman, 1960). Selective prediction and set-valued reporting ask whether a particular output should be issued under uncertainty (Geifman & El-Yaniv, 2017). EW-GET applies abstention one level earlier: it asks whether the requested *distinction itself* is supported by the available observational experiment. Its report is therefore resolution-aware—a confidence set over supported diagnostic cells rather than a forced singleton mechanism label.

2.1 ACCESS-CONDITIONED IDENTIFIABILITY FORMULATION

Let $X \in \mathcal{X}$ be a latent state and $Y \sim p_\phi(y \mid X)$ an observation under condition ϕ. A sensor-to-decision pipeline is represented as

$$X \to Y \to Z_1 \to \cdots \to Z_L \to \widehat{C}. \tag{1}$$

A task T specifies relevant distinctions through a contrast map $c_T : \mathcal{X} \to \mathcal{C}_T$, and $M \in \mathcal{F} = \{m_0, \ldots, m_K\}$ denotes a finite, prespecified set of sensing, representation, fusion, and decision failure mechanisms.

An access profile $\mathcal{A}$ specifies the observable interface available to an evaluator and exposes a subset of $\{Y, Z_1, \ldots, Z_L, \widehat{C}\}$. A witness $W_j = w_j(V_j)$ is a measurable probe of an exposed trace V_j. The joint witness representation under access $\mathcal{A}$ is

$$W_\mathcal{A} = (W_j : j \in \mathcal{A}), \qquad P_m^\mathcal{A} = \mathcal{L}(W_\mathcal{A} \mid M = m). \tag{2}$$

Definition 1 (Population access-conditioned failure quotient)**.** *For a fixed task, test distribution, failure family, and access profile, two mechanisms are observationally equivalent under access $\mathcal{A}$ if they induce identical joint witness laws:*

$$m \equiv_\mathcal{A} m' \iff P_m^\mathcal{A} = P_{m'}^\mathcal{A}. \tag{3}$$

The population access-conditioned failure quotient is

$$\mathcal{Q}_\mathcal{A} = \mathcal{F}/\equiv_\mathcal{A} . \tag{4}$$

The quotient is the finest mechanism partition identifiable under the declared access-and-witness evidence contract associated with $\mathcal{A}$.

Definition 2 (Diagnostic resolution)**.** *The population diagnostic resolution under access $\mathcal{A}$ is*

$$\mathsf{Res}(\mathcal{A}) := \mathcal{Q}_\mathcal{A}. \tag{5}$$

For two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ of the same mechanism family, write $\mathcal{P}_2 \succeq \mathcal{P}_1$ when every cell of $\mathcal{P}_2$ is contained in a cell of $\mathcal{P}_1$; $\mathcal{P}_2$ then supports at least as fine a diagnostic resolution. Diagnostic resolution is therefore a property of the declared task, failure family, access profile, and witness law—not of a particular classifier trained to predict mechanism labels. Throughout, $\mathsf{Res}(\mathcal{A})$ is understood conditionally on this fixed witness contract.

For finite-resolution evaluation, fix a distribution distance d and tolerance $\delta_0 \geq 0$, and let

$$m \, R_\mathcal{A}^\delta \, m' \iff d(P_m^\mathcal{A}, P_{m'}^\mathcal{A}) \leq \delta_0. \tag{6}$$

This relation need not be transitive. We therefore define the threshold graph

$$G_\mathcal{A}^\delta = (\mathcal{F}, E_\mathcal{A}^\delta), \qquad (m, m') \in E_\mathcal{A}^\delta \iff m R_\mathcal{A}^\delta m', \tag{7}$$

and let

$$\Pi_\mathcal{A}^\delta = \mathrm{CC}(G_\mathcal{A}^\delta) \tag{8}$$

be the connected-component partition induced by the transitive closure of direct indistinguishability. Hence $\mathcal{Q}_\mathcal{A}$ is the exact population quotient, whereas $\Pi_\mathcal{A}^\delta$ is the finite-resolution operational target. Two mechanisms in the same operational cell need not themselves lie within δ_0; the connected-component rule is a transitive reporting convention, not a claim of pairwise metric equivalence.

Population, finite-resolution, and finite-sample targets. These three objects answer different questions. Population equality $P_m^\mathcal{A} = P_{m'}^\mathcal{A}$ defines the intrinsic identifiability quotient relative to the declared evidence contract. A nonzero tolerance δ_0 defines a deliberately coarser operational target $\Pi_\mathcal{A}^\delta$ through direct witness-law indistinguishability and its graph closure. Finally, finite calibration data yield an estimated graph $\widehat{G}_\mathcal{A}^\delta$ and partition $\widehat{\Pi}_\mathcal{A}^\delta$. Theorem 6 certifies recovery of this operational partition, not exact equality of unknown population laws. Thus EW-GET separates

$$\underbrace{\mathcal{Q}_\mathcal{A}}_{\text{population identifiability}} \longrightarrow \underbrace{\Pi_\mathcal{A}^\delta}_{\text{declared operational resolution}} \longrightarrow \underbrace{\widehat{\Pi}_\mathcal{A}^\delta}_{\text{finite-sample report}}, \tag{9}$$

with abstention whenever the last transition cannot be certified.

3 EW-GET: ACCESS-CONDITIONED FAILURE RESOLUTION

3.1 ACQUISITION-GROUNDED WITNESSES

Reference and task-effective geometry. For a regular dominated observation model, the acquisition reference is Fisher information in world coordinates,

$$\Gamma_\phi(x) = \mathbb{E}[\nabla_x \log p_\phi(Y \mid x) \nabla_x \log p_\phi(Y \mid x)^\top]. \tag{10}$$

The task-effective reference Γ_{eff}^T profiles out nuisance coordinates while retaining task-relevant directions (Appendix B.1); failed regularity or estimability checks return Not Estimable.

Proposition 3 (Information no-refinement)**.** *If Z is obtained from Y by a state-independent Markov kernel, then $x \sim_Y x'$ implies $x \sim_Z x'$ and, under regularity, $\Gamma_Z(x) \preceq \Gamma_Y(x)$.*

Proof in Appendix A.1.

Three complementary channels. For $\Gamma = V \Lambda V^\top$ and a fixed relative rank tolerance ϵ, the magnitude witness is

$$\mathsf{Mag}(\Gamma) = (\operatorname{rank}_\epsilon \Gamma, \lambda_{\min}^+(\Gamma), \operatorname{tr} \Gamma, \log\det_\epsilon \Gamma, \operatorname{tr} \Gamma^\dagger). \tag{11}$$

For prespecified task directions d_q, input support is $I_q = d_q^\top \Gamma_{\text{in}} d_q$ and $R_q = D_M(d_q)$ is the whitened finite-difference response statistic (Appendix B.2).

Directional fidelity uses

$$\mathsf{ERF} = \rho_{\text{Spearman}}(\{I_q\}, \{R_q\}), \tag{12}$$

$$\mathsf{GeoDist} = Q^{-1} \sum_q \left| \log \frac{\widetilde{R}_q + \varepsilon}{\widetilde{I}_q + \varepsilon} \right|, \tag{13}$$

with fixed-direction RMS normalization (Appendix B.2); these shape statistics are invariant to isotropic scaling and are reported with magnitude.

For calibrated confidence p_q, decision overclaim on low-support queries is

$$\mathsf{EOR} = \Pr(p_q > \tau_c \mid I_q < \tau_I). \tag{14}$$

It is not estimable when the test set contains no low-support query.

3.2 ACCESS-CONDITIONED FAILURE QUOTIENT

Proposition 4 (Population access refinement)**.** *If $\mathcal{A} \subseteq \mathcal{B}$, equality of joint witness laws under $\mathcal{B}$ implies equality under $\mathcal{A}$. Thus exact population quotient cells can split, but cannot merge, as richer access is added.*

Proof in Appendix A.2. In the resolution notation of Eq. (5), $\mathcal{A} \subseteq \mathcal{B}$ implies $\mathsf{Res}(\mathcal{B}) \succeq \mathsf{Res}(\mathcal{A})$ at the population level under the fixed, nested witness construction. This monotonicity concerns what the richer evidence contract can support; finite-resolution behavior additionally depends on the distance, graph, tolerance, and uncertainty rules.

3.3 RECOVERY GUARANTEES AND CALIBRATED ABSTENTION

Assumption 5 (Uniform distance estimation)**.** *For $K+1$ mechanisms and n independent calibration groups per mechanism, an estimator $\widehat{d}_{mm'}$ satisfies, with probability at least $1-\alpha$,*

$$\max_{m<m'} |\widehat{d}_{mm'} - d(P_m^{\mathcal{A}}, P_{m'}^{\mathcal{A}})| \le r_n(\alpha, K). \tag{15}$$

The bound is instantiated through analytic concentration or group-level resampling with simultaneous coverage.

Theorem 6 (Finite-sample operational-partition recovery)**.** *Let the finite-resolution threshold graph connect m, m' when $d_{mm'} \le \delta_0$. Assume a margin $\gamma > 0$ such that every pair satisfies either $d_{mm'} \le \delta_0 - \gamma$ or $d_{mm'} \ge \delta_0 + \gamma$. Construct the empirical graph using threshold δ_0 and its connected components. If $r_n(\alpha, K) < \gamma$, then the empirical and population finite-resolution partitions $\widehat{\Pi}_{\mathcal{A}}^\delta$ and $\Pi_{\mathcal{A}}^\delta$ are identical with probability at least $1-\alpha$.*

Algorithm 1 Resolution-aware EW-GET procedure.

Given access $\mathcal{A}$, mechanism-indexed calibration traces and a held-out trace: (1) estimate the acquisition reference and construct magnitude, fidelity and overclaim witnesses from exposed traces; (2) estimate witness-law distances and simultaneous intervals; (3) join pairs with intervals below δ_0, mark ambiguity and form connected operational cells; (4) fit the cell-level confidence-set rule; evaluate the held-out trace without refitting. Return the confidence set and candidate repairs consistent with the reported resolution, or Abstain/Not Estimable if estimability, coverage or separation requirements are not satisfied.

Table 1: Theory–experiment correspondence. Each experiment targets a distinct scientific claim rather than adding another accuracy benchmark.

Scientific claim	Formal object/result	Empirical test
Access determines supported resolution	Population access refinement	Black/Grey/White partition recovery
Finite data can recover declared resolution	Thm. 6 + Cor. 7	Sample-size stress + abstention
Witness channels add complementary evidence	Magnitude/fidelity/overclaim	Witness-channel ablations
Prediction differs from diagnostic justification	$\mathcal{Q}_{\mathcal{A}}, \Pi^{\delta}_{\mathcal{A}}$ vs. label prediction	HM3D White vs. generic MLP
Identifiability does not imply repairability	Evidence vs. intervention contract	Held-out family-wise repair effects

Proof in Appendix A.3.

Corollary 7 (Calibrated abstention)**.** *Attach to each $\widehat{d}_{mm'}$ a simultaneous interval of radius r_n and abstain from splitting a pair whenever its interval intersects δ_0. Then, with probability at least $1-\alpha$, the procedure avoids unsupported fine-grained splits among the declared mechanism pairs.*

Close witness laws, many mechanisms, or high-variance witnesses increase calibration requirements. Algorithm 1 certifies edges by interval endpoints; under the uniform-error event, $2r_n < \gamma$ resolves every pair. HM3D intervals quantify predictive and intervention outcomes, not operational-partition recovery guarantees.

3.4 Operational resolution procedure

Before test evaluation, EW-GET freezes the task/acquisition, failure family, access/witness contract, coordinates, grouped splits, distance estimator, and uncertainty rules (δ_0, α).

Partition recovery can still leave per-trace ambiguity; operational-partition recovery, prediction, and intervention utility are evaluated separately.

4 Experiments

Protocol. All evaluation choices are fixed on development/calibration data before held-out evaluation. Embodied intervals use grouped bootstrap (Davison & Hinkley, 1997); certification follows the declared procedures.

4.1 Controlled Recovery and Acquisition Validation

The benchmark defines six mechanisms: nominal N, sensing S, representation R, fusion F, and decisions D1/D2. Its declared finite-resolution partitions contain 4/5/6 Black/Grey/White cells: Black groups $\{R, F, D1\}$, Grey separates D1, and White also separates R from F. Black exposes acquisition magnitude and decision overclaim; Grey adds fusion fidelity; White adds representation fidelity. Across 50 repetitions, 32,768 graph-calibration groups per mechanism support root-MMD (Gretton et al., 2012) with fixed random Fourier features (Rahimi & Recht, 2007); all splits and confidence rules are fixed and disjoint (Appendix C).

Table 2: Controlled finite-resolution recovery under nested access.

Metric	Black box	Grey box	White box
Partition recovery (ARI)	1.000	1.000	1.000
Cell-set coverage	100%	100%	100%
False fine-grained commitment	0%	0%	0%
Abstention rate	74.96%	49.98%	74.96%
Median pairwise separation	0.4197	0.6352	0.7974

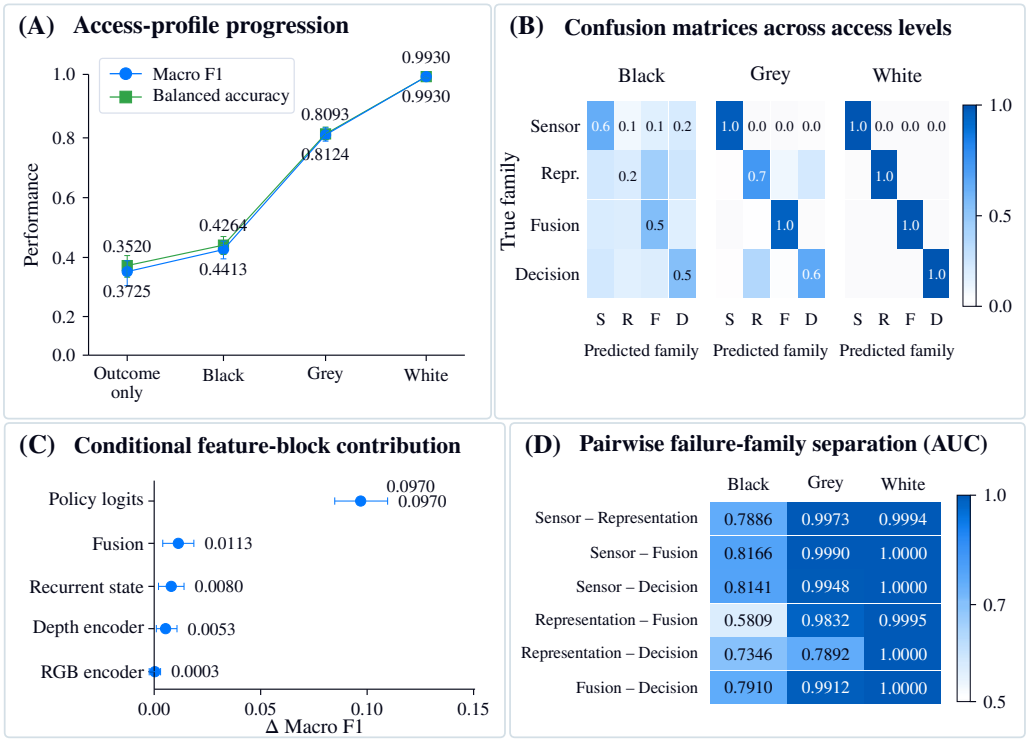

Figure 2: **Access-conditioned characterization.** (A) Macro F1/balanced accuracy; (B) confusion matrices; (C) White feature-block contributions; (D) pairwise failure-family AUCs.

Every repetition recovers the 4/5/6-cell partitions with ARI 1.0 and 100% observed cell-set coverage; no refinement violations or ambiguous pairs occur (Table 2). Ablations, sample-efficiency tests, and stress conditions are in Appendices D and H.

4.2 Embodied Access-Conditioned Failure Characterization

Fixed RGB–D PointNav (Weihs et al., 2020; Chattopadhyay et al., 2021) in Habitat/HM3D (Savva et al., 2019; Ramakrishnan et al., 2021) uses a 500-step horizon. Black exposes behavior/outcomes; Grey adds sensor/fusion traces; White adds recurrent/logit summaries. Development-only probes exclude injected labels and severities (Appendix F.1). Black/Grey/White achieve macro F1 $0.4264/0.8093/0.9930$ on common support (Figure 2); removing policy logits lowers White F1 by 0.0970 [0.0848, 0.1096], whereas RGB adds 0.0003 [-0.0023, 0.0030]. These are predictive-accessibility results; operational-partition recovery still requires the witness contract and resolution rule.

Prediction does not establish diagnostic justification. White logistic reaches macro F1 0.9930 versus 0.9957 for a generic White-summary MLP, with paired difference -0.0027 [-0.0060, 0.0003]; the near-match persists at held-out severities. This comparison is intentionally not a classifier contest. Label prediction evaluates whether a learned rule can map observed summaries to mechanism labels under the evaluation distribution, schematically $\Pr(\widehat{M} = M)$. Diagnostic reso-

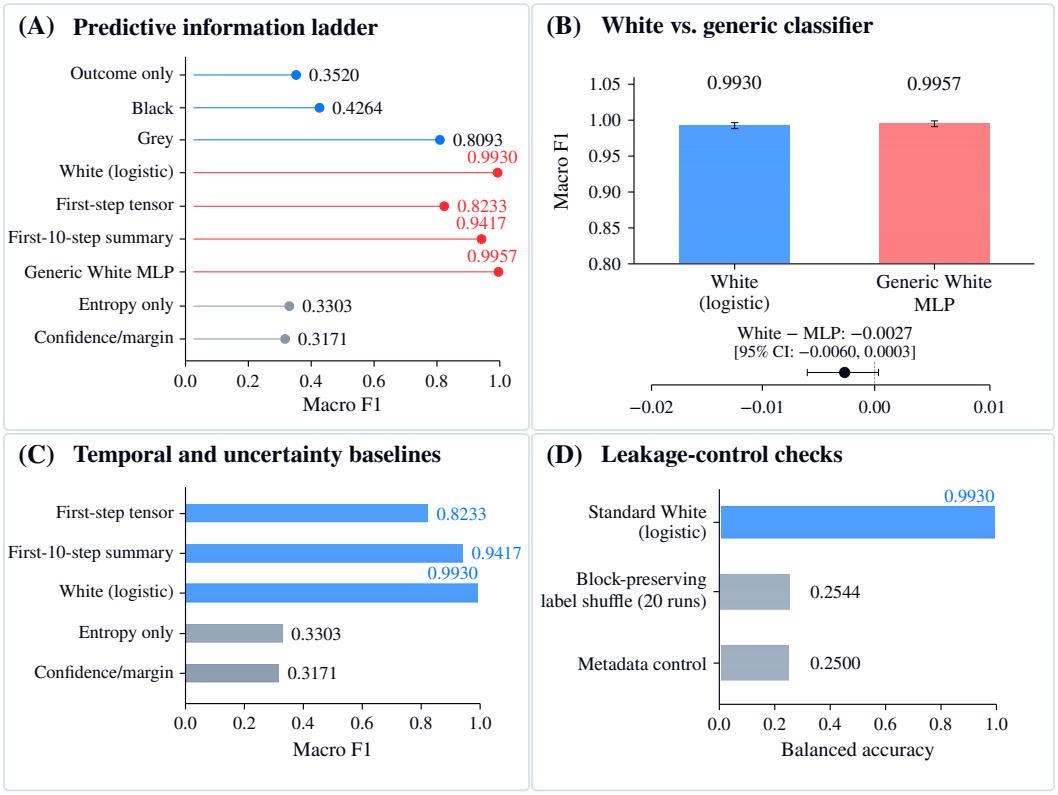

Figure 3: **Prediction does not establish diagnostic justification.** White and a generic White-summary MLP have nearly identical F1, while controls test temporal, uncertainty, and leakage shortcuts.

lution instead asks which distinctions among mechanisms are justified by the declared witness laws and access contract. A high-capacity predictor may exploit stable correlations, trajectory regularities, or other predictive structure without thereby certifying that the requested mechanism distinction is an evidence-supported diagnostic claim. Hence

$$\text{label predictability} \not\Rightarrow \text{evidence-supported identifiability.} \tag{16}$$

The HM3D result therefore illustrates a difference in scientific target, not superiority of the White logistic over the generic MLP (Figure 3; Appendix F.2). Controls reduce obvious leakage routes but do not rule out every trajectory or outcome shortcut, and shared cases are not independent replications.

4.3 Diagnosis-Guided Intervention

Fixed White resolution-aware reports guide same-seed reset retries under a fixed repair contract using only post-fault observations and development references (Appendix F.3). Across 2,995 common cases, guided repair improves over generic repair by 6.18 percentage points [2.80, 9.56] Success and 0.0547 [0.0286, 0.0799] SPL; total actions are 178.79/341.07/332.81 for no intervention/generic/guided repair. The representation-family effect is near null (Success 0.0013 [−0.0667, 0.0627]; SPL −0.0059 [−0.0649, 0.0490]), while a resolution-blind random-allocation control trails guidance by 7.01 Success points [4.65, 9.89].

Diagnostic identifiability does not imply intervention sufficiency: a failure can be distinguishable while the frozen repair library remains ineffective or harmful. The representation effects therefore separate epistemic resolution from control authority.

Restricted held-out validation. Ten HM3D scenes excluded from project fitting use fixed reports, thresholds, references, and repairs (Appendix F.4). White macro F1 remains 0.9916/0.9919 at primary/target severity. Guided-minus-generic gains are 3.82 points [2.34, 5.27] Success and 0.0312

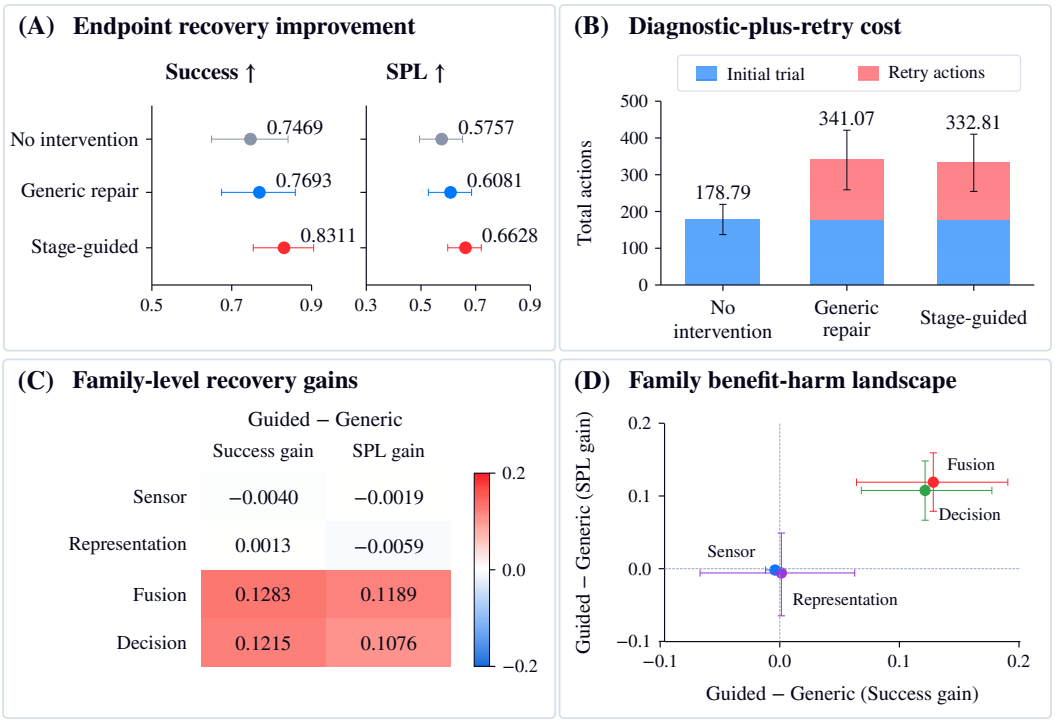

Figure 4: **Diagnosis-guided retry.** Endpoint recovery, action cost, family-level effects, and Success–SPL benefit–harm contrasts.

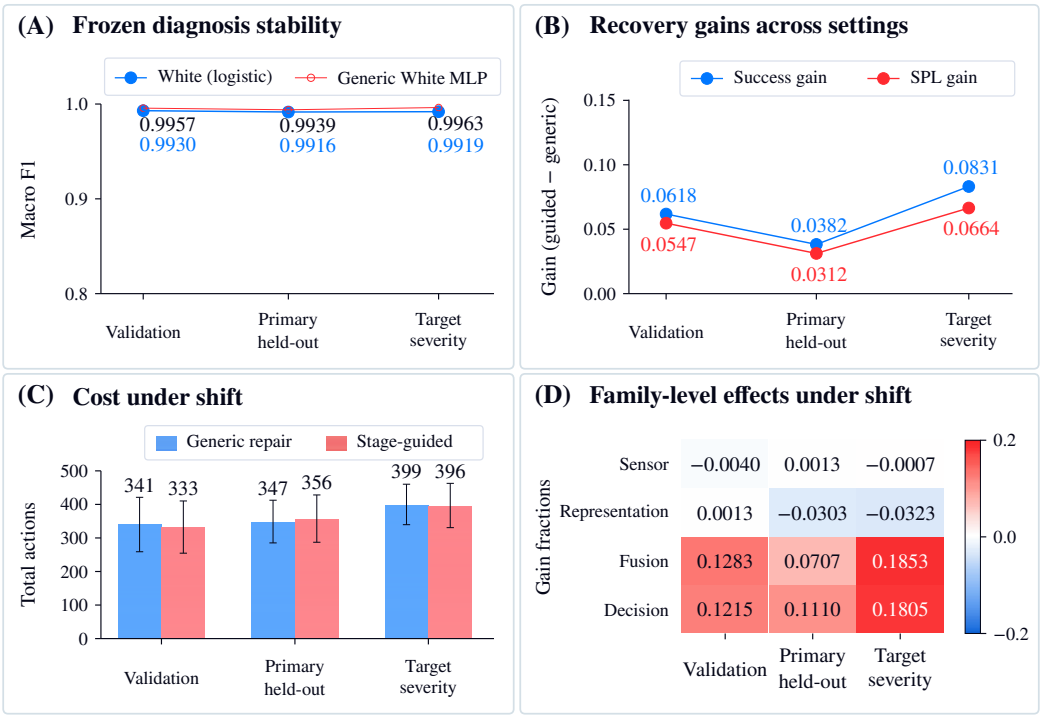

Figure 5: **Held-out validation.** Frozen diagnosis, guided-minus-generic recovery, action cost, and family-level effects across primary and target severity.

[0.0189, 0.0434] SPL at primary severity, and 8.31 points [6.43, 10.25] and 0.0664 [0.0524, 0.0818]

at target severity. Representation effects remain adverse at both severities; endpoint SPL excludes diagnosis cost and target action savings are not established.

Stability boundaries. Controlled-shift and HM3D–Replica (Straub et al., 2019) studies are reported in Appendix I. A two-step fusion-input lag reduces SPL from 0.649 to 0.379 despite high coarse local-response retention, showing that local witness retention need not imply task-level stability; cross-environment results characterize transfer rather than universal invariance.

5 Discussion

EW-GET treats diagnostic resolution, not a singleton fault label, as the object to be justified. Population witness-law equality defines an access-conditioned quotient; at finite resolution, direct indistinguishability induces an operational graph partition whose recovery is certified from finite samples. Richer access can refine supported resolution, while limited access warrants coarser reports. This makes failure analysis a resolution problem before it becomes a classification problem.

Identifiability is not causality. EW-GET determines whether candidate mechanisms are distinguishable under a declared observational experiment. It does not establish that a distinguishable mechanism is the unique causal root of a downstream failure, especially when failures interact or when the intervention generating a diagnostic signature is itself imperfect. Causal attribution therefore requires assumptions beyond witness-law separability.

Identifiability is not intervention sufficiency. The intervention contract is distinct from the evidence contract. A failure family may be reliably distinguishable while the available repair library is ineffective or harmful; the adverse held-out representation-repair effects are a concrete example. EW-GET therefore separates epistemic resolution—what distinctions the evidence supports—from control authority—what available actions can improve.

Resolution is closed-world and contract-relative. The mechanism family $\mathcal{F}$ is prespecified, so $\mathcal{Q}_{\mathcal{A}}$ and $\Pi^{\delta}_{\mathcal{A}}$ partition only the declared candidates rather than all possible failures. The unknown-mechanism withholding study in Appendix H probes abstention under this mismatch, but does not constitute universal open-set recognition. More broadly, resolution depends on the task, witnesses, tolerance, calibration regime, and access profile; changing any of these changes the diagnostic experiment itself.

EW-GET is not specific to navigation. The same formulation applies to layered learned systems—including multimodal, tool-using, and retrieval-augmented pipelines—whenever only partial internal evidence is observable and an evaluator must decide which mechanism distinctions remain supportable. The relevant claim is therefore not that one HM3D witness contract transfers unchanged to every domain, but that the access–evidence–resolution formulation supplies a common diagnostic target across such systems.

6 Conclusion

EW-GET reframes diagnosis as estimation of the finest failure distinction justified by available evidence. Exact witness-law equality defines the population quotient, a declared tolerance induces an operational resolution partition, and finite samples support recovery only when the corresponding distinctions can be certified. Controlled and embodied studies validate access refinement, distinguish prediction from diagnostic justification, and show both bounded intervention utility and the separate limitation that identifiability does not guarantee causality or repairability. Failure diagnosis is therefore a resolution problem before it is a classification problem. EW-GET provides a statistical framework for determining that resolution from the evidence actually available to the evaluator.

AI USE DISCLOSURE

AI-assisted tools were used for limited manuscript editing, formatting assistance, and code-related support during development. The authors reviewed and verified all generated content, analyses, experimental results, and final manuscript text.

REFERENCES

G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

S. Amari. *Information Geometry and Its Applications*. Springer, 2016.

D. Blackwell. Comparison of experiments. In *Proc. Second Berkeley Symposium*, 1951.

P. Chattopadhyay, J. Hoffman, R. Mottaghi, and A. Kembhavi. RobustNav: Towards Benchmarking Robustness in Embodied Navigation. In *ICCV*, 2021. URL `https://arxiv.org/abs/2106.04531`.

T. Chen et al. Metamorphic testing: A review. *ACM Computing Surveys*, 2018.

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997.

R. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions*, 1922.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

I. Goodfellow et al. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

C. Guo et al. On calibration of modern neural networks. In *ICML*, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions. In *ICLR*, 2019.

R. Isermann. *Fault-Diagnosis Systems*. Springer, 2006.

R. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Engineering*, 1960.

S. Kornblith et al. Similarity of neural network representations revisited. In *ICML*, 2019.

B. Lakshminarayanan et al. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

W. Liu et al. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

A. Madry et al. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

M. Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Y. Ovadia et al. Can you trust your model's uncertainty? In *NeurIPS*, 2019.

K. Pei et al. DeepXplore: Automated whitebox testing of deep learning systems. In *SOSP*, 2017.

M. Raghu et al. SVCCA: Singular vector canonical correlation analysis. In *NeurIPS*, 2017.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

S. Ramakrishnan et al. Habitat-Matterport 3D dataset. In *NeurIPS Datasets and Benchmarks*, 2021.

M. Savva et al. Habitat: A platform for embodied AI research. In *ICCV*, 2019.

J. Straub et al. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797*, 2019.

L. Weihs et al. AllenAct: A Framework for Embodied AI Research. arXiv preprint arXiv:2008.12760, 2020. URL https://arxiv.org/abs/2008.12760.

A Proofs of Main Theoretical Results

The following proofs support information no-refinement, population access refinement, and finite-sample operational-partition recovery in the main method.

A.1 Information no-refinement

Proof. Let $s_Y(Y) = \nabla_x \log p(Y \mid x)$ and $s_Z(Z) = \nabla_x \log p(Z \mid x)$ denote the score functions of the input and output observations, respectively. Since Z is obtained from Y through a state-independent Markov kernel, equality of input distributions is preserved under the common kernel. Moreover, the output score satisfies

$$s_Z(Z) = \mathbb{E}[s_Y(Y) \mid Z].$$

Applying conditional Jensen's inequality to the outer product yields the Fisher information inequality $\Gamma_Z(x) \preceq \Gamma_Y(x)$. □

A.2 Population access refinement

Proof. Since $\mathcal{A} \subseteq \mathcal{B}$, the witness vector $W_\mathcal{A}$ is a coordinate projection of $W_\mathcal{B}$. If

$$P_m^\mathcal{B} = P_{m'}^\mathcal{B},$$

then the corresponding marginal distributions are also identical. Therefore,

$$P_m^\mathcal{A} = P_{m'}^\mathcal{A},$$

which implies

$$m \equiv_\mathcal{A} m'.$$

Hence, any equivalence relation induced by richer access is preserved under restriction to a smaller access profile. Adding observable interfaces can therefore split quotient cells but cannot merge population-equivalent mechanisms. □

A.3 Finite-sample operational-partition recovery

Proof. Consider the uniform estimation event in Eq. (15). Under this event, every pair of mechanisms satisfying

$$d(P_m^\mathcal{A}, P_{m'}^\mathcal{A}) \le \delta_0 - \gamma$$

remains below the empirical threshold δ_0, while every pair satisfying

$$d(P_m^\mathcal{A}, P_{m'}^\mathcal{A}) \ge \delta_0 + \gamma$$

remains above the empirical threshold. Therefore, the empirical finite-resolution threshold graph has exactly the same edge set as the population threshold graph. Since connected components are determined entirely by graph edges, the empirical and population finite-resolution partitions $\widehat{\Pi}_\mathcal{A}^\delta$ and $\Pi_\mathcal{A}^\delta$ are identical under the stated event. □

B Witness Construction Details

These definitions specify nuisance handling and directional normalization for the acquisition-grounded witnesses used in the main method.

B.1 Nuisance-profiled task geometry

Partition the canonical coordinates into task and nuisance components such that

$$\Gamma = \left[\begin{smallmatrix} A & B \\ B^\top & C \end{smallmatrix}\right].$$

The nuisance-profiled geometry is

$$\Gamma_{\text{eff}}^{T} = A - BC^{\dagger}B^{\top}, \tag{17}$$

under $(I - CC^{\dagger})B^{\top} = 0$.

B.2 Directional response and normalization

For prespecified task directions d_q, let $I_q = d_q^{\top}\Gamma_{\text{in}}d_q$ and $R_q = D_M(d_q)$ denote the whitened finite-difference response statistic of the stage under the paired contrast $(x, x + hd_q)$ divided by h^2. The whitening transform and h are fixed using calibration data and are not adapted during evaluation.

For the fidelity statistics, tildes denote RMS normalization over the fixed direction set.

C Controlled Finite-Resolution Recovery Supplement

This supplement specifies distance estimation and uncertainty for the controlled operational-partition recovery results in the main paper.

Distance estimation and uncertainty procedure. The operational partition estimator approximates witness-law separations under the declared access profile. The default implementation uses a root-MMD estimator with a fixed bounded-linear-plus-Gaussian kernel, 131,072 random Fourier coordinates, and an exact bounded component. Simultaneous intervals retain the maximum of the analytic sampling bound and the simultaneous quantile from 2,000 paired group-bootstrap draws, with the RFF allowance and numerical guard added separately. This decomposition separates theorem-linked analytic bounds from observed bootstrap uncertainty.

C.1 Estimator and calibration details

For reproducibility, the default implementation uses a characteristic-kernel MMD (Gretton et al., 2012) on standardized joint witnesses. Standardization, kernel bandwidth, and the group-bootstrap design are fit on development data. The maximum pairwise bootstrap deviation is used to estimate the uniform event required by Assumption 5. A classifier two-sample estimator is retained only as a robustness check and is never selected using test performance.

Embodied studies use grouped/hierarchical bootstrap intervals; intervention intervals follow their declared paired uncertainty protocol (Appendix F.3 and F.4).

D Ablation and Sensitivity Analysis

These controlled-benchmark extensions evaluate witness resolution, tolerance sensitivity, calibration support, and bounded stress conditions. They reuse original seed blocks and fixed random-feature maps; all conditions are fixed before evaluation.

Unless varied below, the catalog is N/S/R/F/D1/D2, access is Black/Grey/White, and graph-calibration, confidence-calibration and test counts are 32,768, 1,024 and 2,048 groups per mechanism. The primary tolerance remains $\delta_0 = 0.20$. The shared distance/uncertainty contract is specified in Appendix C; group pairing and fitting/test separation are preserved.

Tables report repetition means; brackets and figure error bars are saved 95% Student-t intervals for those means, conditional on the fixed feature map. These differ from simultaneous distance-certification intervals; recovered cells, certified cells, coverage and abstention remain distinct quantities. N/A denotes an undefined denominator.

D.1 Witness Channel Complementarity

Which distinctions do directional fidelity and decision overclaim add beyond magnitude? We compare M, MF and Full with coordinate maps $[M, 0, 0, 0]$, $[M, 0, F_{\text{fusion}}, F_{\text{representation}}]$ and

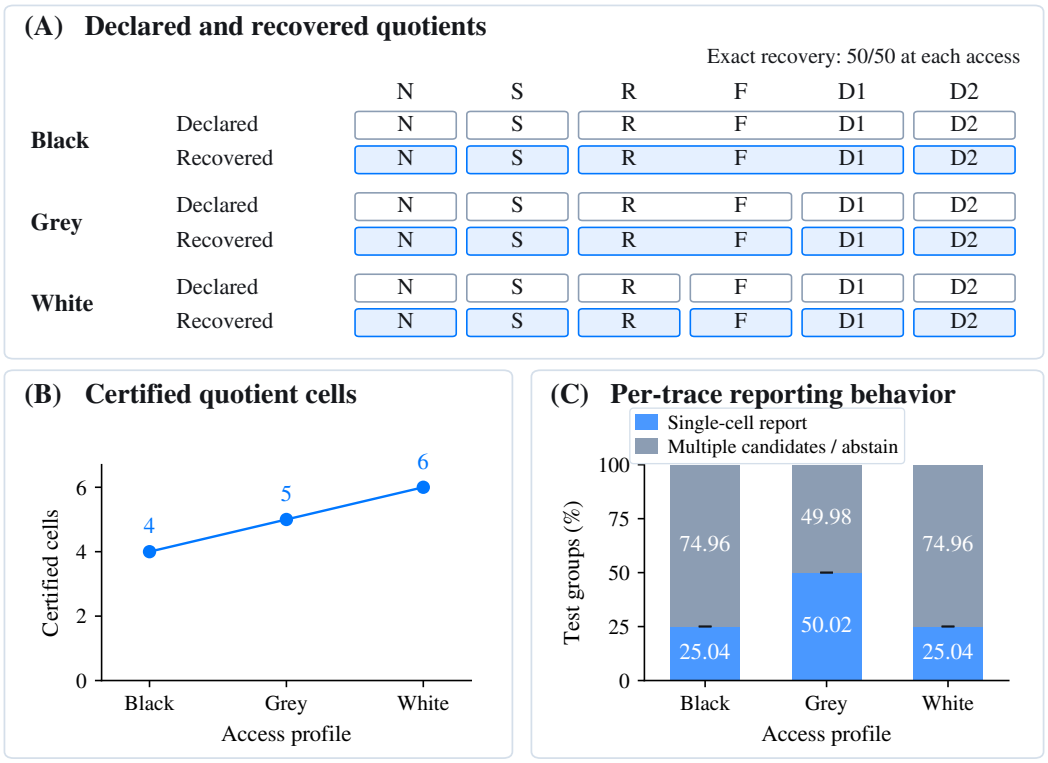

Figure A1: **Access-conditioned resolution refinement.** (A) Declared and recovered partitions for nominal (N), sensing (S), representation (R), fusion (F), and decision (D1,D2) mechanisms; all access-specific partitions recover in 50 repetitions. (B) Certified cells increase from four to five to six under Black, Grey, and White access. (C) Single-cell reports and multiple-candidate abstention on held-out groups; abstention is 74.96%, 49.98%, and 74.96%, respectively. Under this controlled witness contract, quotient recovery does not imply unique mechanism identification from an individual trace.

$[M, O, F_{\text{fusion}}, F_{\text{representation}}]$. Zeroing input coordinates retains the original kernel feature map; no RFF coordinates or bandwidths are refit. Each projected population law and tolerance quotient are certified before evaluation; confidence rules are recalibrated under each projection.

Every variant recovers its own target exactly in all repetitions (Table A1). M supports two cells at every access; MF supports two, three and four; Full supports four, five and six. For M, N/R/F/D1/D2 remain together and S is separate. Under MF, Grey additionally separates R/F from N/D1/D2, and White separates R from F. Full retains the primary partitions. All cells are certified, with no ambiguous pairs, false merges or false fine-grained commitments. False splits are zero where defined.

Each variant is evaluated against its own projected quotient target rather than the Full witness target. Equal ARI does not mean equal supported resolution. The result supports incremental complementarity in this family, not universal channel necessity: magnitude is present in every variant. Exact recovery still permits multiple-candidate abstention, and richer access need not reduce per-trace ambiguity monotonically.

D.2 Equivalence Tolerance Sensitivity

How sensitive are certification and reporting to the declared tolerance? We reuse saved pairwise distances, simultaneous endpoints and confidence scores/thresholds at $\delta_0 \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$. Only graph decisions, connected cells, certification and graph-dependent reporting are recomputed. No simulation, bootstrap or confidence refit is performed. All grid points preserve the evaluated quotient partition: their tolerances lie between zero within-cell distance and the smallest between-cell distance, 0.417. The primary tolerance stays 0.20.

Table A1: Witness contracts and access-specific resolution. Recovered cells equal supported and certified cells; exact recovery and pairwise quotient accuracy are one throughout. Rates are fractions.

Contract	Access	Supported	Certified	ARI	Coverage	Abstention
M	Black	2	2	1.000	1.000	0.875 [0.874, 0.875]
M	Grey	2	2	1.000	1.000	0.875 [0.874, 0.875]
M	White	2	2	1.000	1.000	0.875 [0.874, 0.875]
MF	Black	2	2	1.000	1.000	0.875 [0.874, 0.875]
MF	Grey	3	3	1.000	1.000	0.625 [0.624, 0.626]
MF	White	4	4	1.000	1.000	0.875 [0.874, 0.875]
Full	Black	4	4	1.000	1.000	0.750 [0.749, 0.750]
Full	Grey	5	5	1.000	1.000	0.500 [0.499, 0.501]
Full	White	6	6	1.000	1.000	0.750 [0.749, 0.750]

All 50 repetitions retain 4/5/6 certified cells, ARI, exact recovery, pairwise accuracy and coverage of one, with no ambiguous pairs or false merges. Abstention remains 0.750/0.500/0.750 across accesses, entirely from multiple candidates. Table A2 reports the weakest observed endpoint margins; Figure A2 shows their variation. At 0.10, the smallest Grey equivalence slack is only 1.25×10^{-4}. Successful observed decisions therefore need not imply a wide certification margin.

The sufficient interval condition $2r_n < \gamma$ holds at the primary tolerance but not every tested setting. Observed success outside this sufficient regime does not contradict the theorem; this is grid sensitivity on recorded measurements, not tolerance invariance.

Table A2: Tolerance sensitivity: worst-repetition endpoint slacks. Equivalence slack is $\delta_0 - \max U$ within true cells; separation slack is $\min L - \delta_0$ between cells. These minima are not confidence intervals. White has no distinct within-cell pair.

δ_0	Black within	Grey within	Black between	Grey between	White between
0.10	3.59×10^{-3}	1.25×10^{-4}	0.223	0.276	0.268
0.15	0.054	0.050	0.173	0.226	0.218
0.20	0.104	0.100	0.123	0.176	0.168
0.25	0.154	0.150	0.073	0.126	0.118
0.30	0.204	0.200	0.023	0.076	0.068

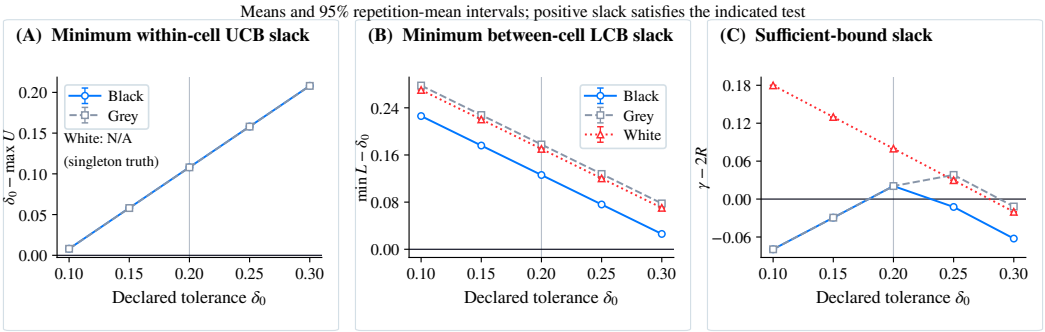

Figure A2: Interval margins over the declared tolerance grid. Panels show within-cell equivalence, between-cell separation and sufficient-condition slack. Observed certification can succeed when a sufficient bound is not met; no new default is selected.

D.3 Calibration Sample Efficiency

How does finite graph-calibration support affect recovery? Deterministic nested prefixes of 3277, 8192, 16384, 24576 and 32768 saved groups are evaluated without outcome selection. Distances,

bootstrap intervals and graphs are recomputed at each size; confidence rules are recalibrated by the original procedure before testing. Population truth, tolerance and features remain unchanged.

The first tested size with exact certified recovery at all accesses in all 50 repetitions is 8192 (Table A3; Figure A3). At 3277, Black/Grey ARI averages 0.220/0.100, with exact recovery in 3/50 and 5/50 repetitions. Their errors are false splits from uncertified equivalence; no false merges occur. White has ARI one at every size, but its singleton truth requires no nontrivial equivalence edge. Coverage remains one despite broad candidate unions under incorrect splits; Black abstention is nonmonotone.

The retained radius decreases from 0.199 to 0.090, with a fixed RFF allowance (Figure A4). The stronger sufficient interval condition first holds for all accesses on this grid at 24576, although observed recovery succeeds earlier. Tested sample sizes do not establish a minimum sufficient sample size or universal sample complexity; Theorem 6's point-threshold condition and Algorithm 1's interval condition remain distinct. No bound was weakened to fit the observed transition.

Table A3: Calibration sample efficiency. Exact recovery and abstention are fractions; coverage is one throughout. Brackets retain untruncated repetition-mean intervals, including a negative lower endpoint for the smallest Black exact-recovery estimate.

Groups	Access	ARI	Exact	Certified	Abstention
3277	Black	0.220 [0.137, 0.303]	0.060 [-0.008, 0.128]	3.06 [2.99, 3.13]	0.750 [0.749, 0.750]
3277	Grey	0.100 [0.014, 0.186]	0.100 [0.014, 0.186]	4.10 [4.01, 4.19]	0.725 [0.704, 0.746]
3277	White	1.000	1.000	6	0.750 [0.749, 0.750]
8192	Black	1.000	1.000	4	0.742 [0.727, 0.757]
8192	Grey	1.000	1.000	5	0.505 [0.495, 0.515]
8192	White	1.000	1.000	6	0.750 [0.749, 0.750]
16384	Black	1.000	1.000	4	0.742 [0.727, 0.757]
16384	Grey	1.000	1.000	5	0.500 [0.499, 0.501]
16384	White	1.000	1.000	6	0.750 [0.749, 0.750]
24576	Black	1.000	1.000	4	0.750 [0.749, 0.750]
24576	Grey	1.000	1.000	5	0.500 [0.499, 0.501]
24576	White	1.000	1.000	6	0.750 [0.749, 0.750]
32768	Black	1.000	1.000	4	0.750 [0.749, 0.750]
32768	Grey	1.000	1.000	5	0.500 [0.499, 0.501]
32768	White	1.000	1.000	6	0.750 [0.749, 0.750]

D.4 Mechanism Severity Matching

Do the designed distinctions persist after active operator magnitude is matched? Severity is $s = \sqrt{((\log g_1)^2 + (\log g_2)^2)/2}$, where g_1, g_2 are singular values of the normalized paired-response Jacobian. S uses isotropic attenuation; R/F/D1 use stage-specific diagonal scaling; D2 uses uniform scaling. The unmatched reference uses $\kappa_S = 2$ and other fault gains 4; matched conditions use 2.5, 3 or 4. Activation prevalence remains 0.75.

All conditions undergo development-only validation before recovery evaluation. Every active severity must match its target within 10^{-6}; matched cross-mechanism mean spread must also be at most 10^{-6}. All 200 condition/repetition blocks qualify; the largest target discrepancy is $< 4 \times 10^{-15}$. Original-parameter outputs reproduce exactly under the same numerical configuration before matched evaluation. New stage traces use the original base draws and unchanged witness/evaluation functions.

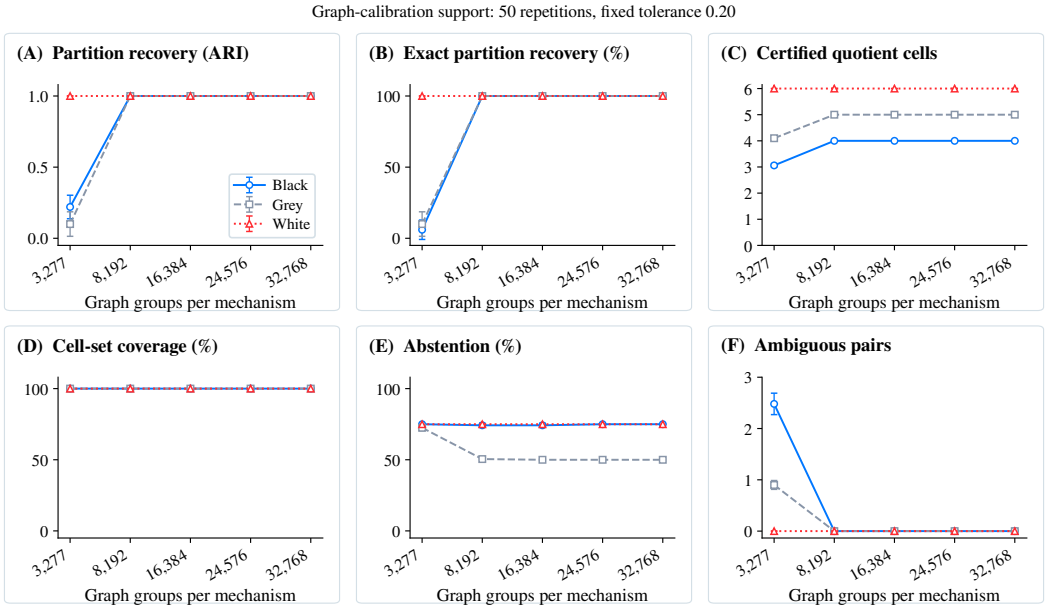

Figure A3: Recovery and reporting versus graph-calibration support. The original independent confidence groups and held-out groups are retained. Recovery of singleton White cells and recovery of nontrivial equivalence cells impose different certification demands.

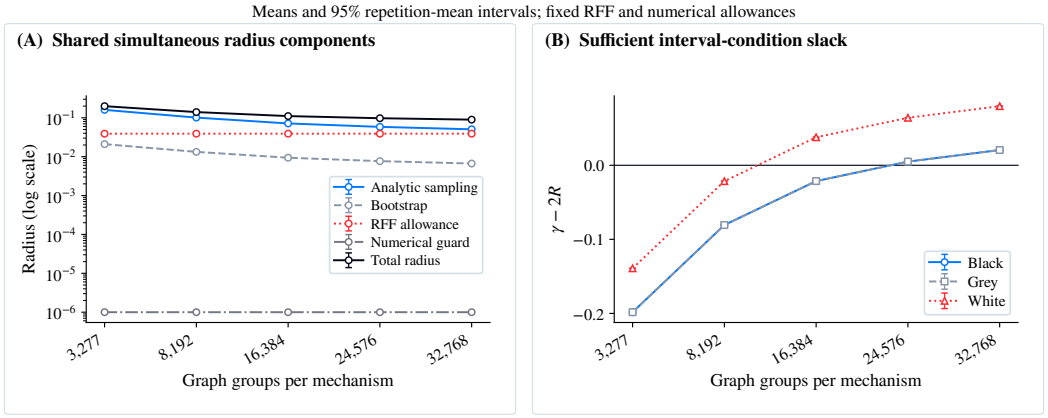

Figure A4: Radius components and sufficient interval-condition slack across tested sample sizes. The analytic/bootstrap maximum retains the fixed random-feature allowance and numerical guard. A sufficient condition need not locate the empirical transition.

All matches retain identical extracted witness laws and paired recovery/reporting outputs (Table A4). Every condition recovers and certifies 4/5/6 cells with ARI, exact recovery and coverage one; abstention remains 0.750/0.500/0.750. This was predicted by the discrete witness plateau at the selected gains. This evaluates operator-magnitude matching; it does not imply universal severity invariance, equal task harm, or causal equivalence. Directional pattern and stage location remain different, and N is excluded from positive-severity matching.

E Measurement and Encoder Validation

Validation scope. This supplement supports the main paper's measurement, encoder, and acquisition-interface analyses. Condition-wise reference-validity coefficients are not reported; limitations of the encoder probe-fairness evaluation are described in Appendix E.2.

Table A4: Operator-magnitude matching. Severity columns show measured development means; F/D1/D2 agree with R to the fixed matching tolerance. Nominal severity is zero up to numerical error. All conditions have ARI, exact recovery and coverage one, with no paired recovery/reporting change from the original. Cells are Black/Grey/White certified counts.

Condition	κ_S	$\kappa_{R/F/D1/D2}$	s_S	s_R	Qualified	Cells
Original	2.0	4.0	0.693	1.386	50/50	4/5/6
Matched-low	2.5	2.5	0.916	0.916	50/50	4/5/6
Matched-middle	3.0	3.0	1.099	1.099	50/50	4/5/6
Matched-primary	4.0	4.0	1.386	1.386	50/50	4/5/6

Measurement-component validation. Under homoscedastic isotropic degradation, absolute information decreases while normalized rank fidelity is unchanged by construction. Blur and occlusion can move fidelity oppositely to magnitude. These contrasts motivate complementary channels over alignment alone. Controlled-signature and no-refinement checks support the measurement implementation, not universal natural-failure localization. After controlling for task-boundary margin, weaker $\lambda_{\min}(\Gamma_{\text{eff}}^T)$ is associated with higher observed downstream error; this is associational evidence. Details are in Appendix E.1.

Illustrative matched-accuracy encoder case study. This case study is included only to illustrate that matched task accuracy need not imply matched witness geometry under a fixed probing contract. ResNet-50 (He et al., 2016) and DINOv2 (Oquab et al., 2023) are matched within 0.2 percentage points of task accuracy (98.58% versus 98.75%), while their mean shuffle-corrected directional rank fidelities are .8466 (95% CI $[.8393, .8541]$) and .1284 ($[.1018, .1553]$). The experiment is not used to rank encoder architectures or establish architecture-level differences; protocol limitations and probe-fairness checks are in Appendix E.2.

E.1 Measurement-component validation details

Observation reference. Gaussian-noise, blur, and occlusion conditions evaluate whether the Fisher Cramér–Rao trace is consistent with empirical estimation difficulty ordering. The observation-reference evaluation includes 4,200 estimation units (200 states and 21 conditions) without estimator convergence failure. A nested $N \in \{50, 100, 200\}$ analysis produced stable spectrum ordering. Condition-wise coefficients are not estimated in the current protocol.

No-refinement and controlled signatures. Across 17 prespecified post-processing conditions, all 313,344 directional retention ratios respected the numerical no-refinement tolerance. Aligned control, null-space over-response, supported-direction suppression, and directional distortion were recovered in all 512 predefined controlled cases. This validates the measurement implementation under the evaluated controls and does not determine natural-failure localization.

State-conditioned vulnerability. The reported correlations use failure severity as the outcome variable and inverse task-effective information strength as the explanatory variable, so positive values indicate higher observed failure severity under weaker task-effective information. After controlling for task-boundary margin, lower $\lambda_{\min}(\Gamma_{\text{eff}}^T)$ (i.e., weaker task-effective information) is associated with higher downstream failure in three prespecified seeds: partial Spearman correlations are .2557, .2463, and .2556, with grouped-bootstrap 95% intervals $[.2255, .2854]$, $[.2097, .2814]$, and $[.2245, .2872]$. All 27 perturbation-family–severity–seed cells are positive. This is associational evidence.

E.2 Matched-accuracy encoder protocol and limitations

We fix common states, perturbations, directions, whitening, probes, and task heads. Permutation controls are centered near zero. The reported probe-fairness evaluation does not fully cover layer sweeps, feature whitening, dimension-matched projections, probe classes, or additional encoder seeds.

E.3 ACQUISITION-GROUNDED INTERFACE VALIDATION

This separate controlled prototype connects a fixed observation experiment to available canonical witnesses and supported quotient resolution; it does not replace the six-mechanism benchmark of Appendix C. Let $X = (u_1, u_2, n)^\top \sim \mathcal{N}(0, I_3)$, $Y = JX + \epsilon$, $\epsilon \sim \mathcal{N}(0, \Sigma)$ and $q(X) = (u_1, u_2)^\top$, with nuisance n and

$$J = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad W = \begin{pmatrix} 1/2 & 0 & -1/2 \\ 0 & 1 & -1 \end{pmatrix}, \qquad WJ = [I_2\ 0].$$

The nominal stages map Y through $z_0 = WY$, representation, fusion and the two-coordinate output. Nominal N has observation covariance $0.1I_3$ and confidence 0.75. M1 changes covariance to $0.4I_3$; M2 applies representation suppression $\mathrm{diag}(1, 0)$, inherited by later stages; M3 changes confidence to 0.99 without changing geometry or point estimates. Acquisition Fisher geometry is $J^\top \Sigma^{-1} J$; task-effective geometry profiles nuisance by its Schur complement, not prior marginalization. The extractor estimates the Jacobian/covariance from allowed queries. Task directions $e_1, e_2, (e_1 + e_2)/\sqrt{2}, (e_1 - e_2)/\sqrt{2}$ are specified in advance.

Black exposes acquisition measurements/queries and only the first scalar public output. Grey adds full representation queries; White adds confidence and remaining internal/output channels. Since the suppressed coordinate is not public and confidence is hidden, Black N/M2/M3 bundles agree and Grey N/M3 bundles agree. Lower-access feature vectors are exact prefixes; hidden-field poisoning and metadata checks pass before recovery. These interfaces differ from the original benchmark and HM3D summary contracts. Magnitude records the acquisition spectrum; fidelity uses directional ERF/GeoDist responses with fixed numerical precision; White overclaim measures confidence above 0.90 on directions with acquisition support below 15. Overclaim is unavailable at Black/Grey, not imputed as zero. Invalid blocks retain Not Estimable states without erasing valid lower-access evidence.

Randomization is fixed before evaluation. Nominal development has 512 groups; independent reference and graph splits each have 4,096 groups per mechanism, with 256 measurement repeats per group and paired-query step 10^{-3}. The 512-group held-out metrology split is separate; the prespecified 64-group graph prefix is not an independent replication. Mechanisms share base draws within a group, with independent randomization across groups and splits. Exact Gaussian-kernel root-MMD uses bandwidth 0.25 and tolerance 0.20, not the original benchmark's RFF/bootstrap implementation. Conditional on fixed calibration, simultaneous intervals use radius $2(1 + \sqrt{2\log(144/0.05)})/\sqrt{n}$ with range $[0, \sqrt{2}]$. Independent reference intervals qualify all ten available access/ablation targets before graph scoring. This is finite-sample reference validation, not an exact population oracle. Whole-cell certification requires internal equivalence and external separation.

Full available witnesses recover Black $\{N, M2, M3\}, \{M1\}$; Grey $\{N, M3\}, \{M1\}, \{M2\}$; and White's four singleton cells. Both transitions strictly refine with no refinement violations. At 4,096 groups every available contract has ARI 1.000 and exact certified recovery against its own target, with zero false merges/splits (Table A5). Magnitude isolates M1, Grey/White fidelity isolates M2, and White overclaim isolates M3. Individual channels support coarser Grey/White partitions; magnitude already attains the full Black target.

Table A5: Acquisition V2 certified cells under each witness contract. At 4,096 groups, all available contracts have ARI 1.000 and exact certified recovery against their independently qualified targets. At 64 groups, all available contracts have zero certified cells and 100% pairwise abstention.

Witness contract / graph support	Black	Grey	White
Magnitude only, 4,096	2	2	2
Fidelity only, 4,096	1	2	2
Overclaim only, 4,096	Unavailable	Unavailable	2
Full available, 4,096	2	3	4
Full available, 64	0	0	0

At 64 groups all six pairs are ambiguous under every available contract: no cell certifies and pairwise abstention is 100%. White's incidental component ARI 1.000 does not establish certified recovery. This is pairwise graph abstention, not calibrated per-trace confidence-set coverage. The deliberately aligned public projection, privileged acquisition queries and well-separated mechanisms make

this an access-interface prototype, not matched full-task performance, HM3D quotient recovery, causal localization or universal mechanism attribution. No multi-seed recovery-rate or arbitrary-noise claim is made; floating-point computations are not formally enclosed. The encoder-validation limitations remain separate.

F HM3D Characterization and Intervention Supplement

Supplementary results for characterization, predictive comparisons, intervention, and held-out validation are organized in Appendix G.1–G.4, respectively.

F.1 Access contracts and characterization details

The HM3D characterization/intervention population comprises 35 scenes and 5,181 valid episodes: 15 development scenes, five validation scenes, ten held-out scenes and five calibration-reserve scenes. Development episodes support fitting and validation episodes yield 2,996 controlled fault cases. Black exposes 12 outcome/behavior summaries; Grey adds sensor-validity and RGB/depth/fusion summaries (40 features); White adds recurrent and policy-logit summaries (61 features). One failure family is injected per case: Sensor depth dropout $p = 0.05$, Representation suppression $\alpha = 0.5$, Fusion scaling $\beta = 0.75$, or Decision temperature $T = 2.5$. Standardization and logistic-probe fitting use development scenes only; injected labels and severities are excluded from evaluation fields, and probes are not refitted on validation data. Bootstrap resampling preserves scene/base-episode grouping and paired cases; draws do not represent independent agents or scenes.

Observable separation increases across the declared profiles: Black, Grey and White achieve macro F1 0.4264, 0.8093 and 0.9930, respectively, on the same validation support (Appendix Table A6; Figure 2). The outcome-only comparator has F1 0.3520 [0.3045, 0.3888] and balanced accuracy 0.3725. Paired Grey-minus-Black and White-minus-Grey F1 differences are 0.3830 [0.3495, 0.4181] and 0.1836 [0.1588, 0.2055]. Removing policy logits decreases White F1 by 0.0970 [0.0848, 0.1096]; fusion, recurrent and depth contributions are smaller, while RGB's 0.0003 [-0.0023, 0.0030] is compatible with no incremental benefit (Appendix Table A7). Confusion matrices and pairwise AUCs describe access-conditioned predictive separation, not recovered quotient cells.

F.2 Predictive baselines and leakage controls

The comparisons use the same 2,996 validation fault cases as the embodied characterization study, with development-only fitting across access-controlled probes, generic White-summary models, temporal summaries, and uncertainty baselines. White and MLP share the 61-field interface. Temporal probes use declared prefixes; entropy and confidence/margin require decision-output access. Leakage controls include block-preserving training-label shuffles, metadata poisoning, and metadata controls. Resampling follows Appendix F.1; the shared cases are not independent replications.

White logistic achieves macro F1 0.9930 and the generic White-summary MLP 0.9957, with paired White-minus-MLP difference -0.0027 [-0.0060, 0.0003] (Appendix Table A8; Figure 3). First-step and first-ten-step summaries reach 0.8233 and 0.9417, whereas entropy-only and confidence/margin reach 0.3303 and 0.3171. The label-shuffle mean balanced accuracy is 0.2544 across the 20 prespecified runs, compared with 0.9930 for the unshuffled White reference; metadata-control balanced accuracy is 0.2500. Metadata poisoning is used as a shortcut-control analysis rather than a target performance metric. The generic-model comparison also persists on held-out severity panels: generic White MLP versus White macro F1 is 0.9939 versus 0.9916 at primary severity and 0.9963 versus 0.9919 at target severity. Together, these results separate predictive usefulness from supported failure resolution; comparable performance is not a formal equivalence-test result.

F.3 Repair operators, comparisons and restoration

Each same-seed retry applies the selected operation at its nominated boundary using post-fault observations and fixed development references, without injected labels, severity values, clean counterparts, or fault masks. The library comprises depth filling, stage-specific norm rescaling, and logit scaling. The primary comparison uses no intervention, generic repair (always depth filling), and stage-guided repair. Endpoint Success and SPL measure recovery; total actions include resolu-

tion and retry. Comparisons use common observed endpoints with paired hierarchical resampling. Generic repair is a fixed comparator, not an optimized global policy.

On 2,995 jointly observed cases out of 2,996 planned, no intervention, generic repair and stage-guided repair attain Success 0.7469/0.7693/0.8311 and SPL 0.5757/0.6081/0.6628, respectively (Appendix Table A9; Figure 4). Guided-minus-generic gains are 6.18 percentage points [2.80, 9.56] Success and 0.0547 [0.0286, 0.0799] SPL; gains over no intervention are 8.41 points [4.00, 12.82] and 0.0871 [0.0620, 0.1111]. Generic-minus-no-intervention gains are 2.24 points [0.80, 3.83] and 0.0324 [0.0240, 0.0408]. Total actions are 178.79/341.07/332.81. Guidance uses 8.26 fewer retry actions than generic, but the difference interval $[-18.57, 0.98]$ includes zero. Family effects are heterogeneous (Appendix Table A10): Representation guided-minus-generic Success is 0.0013 $[-0.0667, 0.0627]$, and SPL is -0.0059 $[-0.0649, 0.0490]$. Separately, the matched-nominal extension evaluates task-normalized restoration $R = \mathrm{mean}(\mathrm{Repair}-\mathrm{Fault})/\,\mathrm{mean}(\mathrm{Nominal}-\mathrm{Fault})$ on the same 2,995-case intersection. The predefined denominator criterion requires positive point and lower interval bounds and a finite, nonzero denominator in every bootstrap draw; ratios are not clipped. Pooled generic/guided Success restoration is 0.202/0.761 and SPL restoration is 0.271/0.729 (Appendix Table A11). Sensor Success remains Not Estimable because a bootstrap denominator is zero; Representation guided SPL restoration is -0.112 $[-1.135, 0.368]$. These are task-score ratios, not certified witness restoration, and the separate extension does not change the original intervention analysis.

Twenty repeated no-repair trials reproduce outputs exactly; oracle and selector checks, including wrong-stage interventions, assess implementation consistency rather than population effects.

F.4 Held-out settings, support and comparisons

Primary severities remain Sensor dropout $p = 0.05$, Representation suppression $\alpha = 0.5$, Fusion scaling $\beta = 0.75$, and Decision temperature $T = 2.5$; the target settings are $p = 0.075$, $\alpha = 0.65$, $\beta = 1.0$, and $T = 2.75$. The reporting models, development references, repair operators and same-seed reset-retry contract of Section 4.3 remain fixed, without refitting or threshold retuning. Target repair is a declared extension: target trials provide resolution reports and no-intervention outcomes, while generic/guided arms share selected repairs.

Of 5,956 planned cases per severity, primary and target diagnostic endpoints provide 5,942-case common observed support for each repair comparison; equal counts do not imply identical case identities. No episode is replaced or technical unavailability relabeled as policy failure; paired results condition on observed endpoints without assuming random missingness.

Repeatability is verified before target repair evaluation. Original intervals are retained; target repair uses the separately declared 10,000-draw paired scene/base-episode bootstrap with fixed models. Its descriptive intervals are not multiplicity-adjusted confirmatory claims.

White macro F1 and balanced accuracy are 0.9916 at primary severity and 0.9919 at target severity, with paired target-minus-primary F1 difference 0.00033 $[-0.00306, 0.00350]$; generic White MLP F1 is 0.9939/0.9963 (Appendix Table A12; Figure 5). Primary no-intervention/generic/guided Success is 0.7203/0.7457/0.7839 and SPL is 0.5474/0.5854/0.6166 (Appendix Table A13). Guided-minus-generic gains are 3.82 percentage points [2.34, 5.27] Success and 0.0312 [0.0189, 0.0434] SPL; guided-minus-no-intervention gains are 6.36 points [4.39, 8.28] and 0.0692 [0.0535, 0.0854]. Generic-minus-no-intervention gains are 2.54 points [1.49, 3.61] and 0.0380 [0.0295, 0.0467]. Primary total actions are 182.89 [151.77, 214.89], 347.47 [285.25, 412.48], and 355.91 [287.18, 428.19], respectively; guidance uses 8.44 [0.12, 17.32] more retry actions than generic. Reported gain-retention point ratios (distinct from task-normalized restoration ratios) are 0.756 for Success and 0.794 for SPL, where each ratio is the held-out-scene primary-severity guided-minus-no-intervention gain divided by the corresponding validation-scene primary-severity gain; no ratio intervals are reported. Primary Representation guided-minus-generic Success is -0.0303 $[-0.0552, -0.0041]$ and SPL is -0.0353 $[-0.0554, -0.0141]$ (Appendix Table A14). At target severity, no-intervention/generic/guided Success is 0.6481/0.6851/0.7683 and SPL is 0.4816/0.5303/0.5968 (Appendix Table A15). Guided-minus-generic gains are 8.31 percentage points [6.43, 10.25] and 0.0664 [0.0524, 0.0818]. Target total actions are 213.14/399.08/395.92; the guided-minus-generic difference is -3.16 $[-13.48, 7.30]$, so an action-saving advantage is not established. Target Representation effects remain adverse: -3.23 percentage points $[-6.32, -0.47]$ Success and -0.0343

[-0.0566, -0.0132] SPL (Appendix Table A16). Repair total actions include diagnosis and retry, whereas endpoint SPL excludes diagnostic-trial cost.

F.5 Diagnosis-blind repair allocation control

This extension tests assignment value beyond the repair library's aggregate mixture on the original validation/primary-severity support: 2,995 observed cases from 749 base episodes in five scenes. The 2,996-case comparison set includes one pre-existing unavailable comparison; no new endpoint is missing. Within each scene, the guided arm's aggregate repair counts are permuted across sorted eligible recipients using PCG64 randomization fixed before evaluation. Assignments are fixed before retrieving or generating their outcomes. The selector receives identities and aggregate counts, not per-case resolution reports, family, severity or outcomes. This is diagnosis-blind batch allocation, not uniform selection.

The checkpoint, operators, corruption, paired policy/failure seeds and cost accounting are unchanged. Of the assigned endpoints, 1,306 reuse previously measured outcomes under identical episode, operator and randomization conditions, and 1,689 are newly measured. All arms retain the existing reporting-trial cost; repair arms add retry actions. This comparison isolates allocation under the declared contract, not the random policy's minimum deployment cost.

Uncertainty uses the same 10,000 paired scene/base-episode bootstrap weights on verified common support and ordering, with fixed resampling and percentile 95% intervals. Guided-minus-random Success is $+7.01$ percentage points [4.65, 9.89], SPL $+0.0717$ [0.0584, 0.0861], and total actions -21.23 [-29.11, -14.16] (Table A17). Representation recovery intervals include zero; Decision uses more actions under guidance. Random-minus-none and random-minus-generic pooled recovery intervals include zero. The intervals condition on one fixed assignment and observed support, not variation over all permutations. Family/cost comparisons are descriptive. The result supports allocation utility beyond this scene-level mixture; it establishes neither causal repair, oracle optimality nor held-out transfer.

F.6 Four-family calibration matching assessment

Matching seeks to distinguish mechanism-specific signatures from differences in damage magnitude. A separate calibration-only study fixes the agent, operators, seeds, episode ordering and severity grid. A predefined subset of the first seven declared calibration scenes is assessed before inspecting Success, SPL, length or tuple rankings: 699 eligible identities from 700 selected, after one existing technical exclusion, with no replacements. All 31 conditions (nominal plus 30 fault settings) are evaluated on these same identities: $31 \times 699 = 21{,}669$ rollouts. Analysis uses the complete balanced support. The results concern this subset of the ten-scene calibration set (989 eligible identities), not the full set.

Sensor severities are $0.04, 0.05, 0.06, 0.07, 0.08, 0.09$; Representation $0.35, 0.425, 0.50, 0.575, 0.65, 0.725, 0.80$; Fusion $0.50, 0.625, 0.75, 0.875, 1.0, 1.125$; Decision $0.20, 0.25, 0.30, 0.35, 0.40, 2.50, 2.55, 2.60, 2.65, 2.70, 2.75$. The predefined criteria require maximum cross-family Success spread ≤ 0.01, SPL spread ≤ 0.02, and every pairwise length SMD < 0.20; SMD divides the absolute mean difference by the square root of the mean sample variances. All $6 \times 7 \times 6 \times 11 = 2{,}772$ tuples are evaluated. Criteria and the prespecified descriptive ranking (maximum tolerance-normalized discrepancy, then their sum, then lexicographic severities) are unchanged after outcome inspection; no operating point is selected.

No tuple passes: all 2,772 fail both Success and SPL, and 2,768 also fail length (Table A18). No single family uniquely explains the failure: Decision's low-severity block misses Success matching and its high block misses length matching, while Sensor also has no jointly matched pair. Only Representation–Fusion qualifies pairwise, at 3 of 42 combinations (Table A18); this exploratory observation is not a mechanism-identification result. The required four-family matching regime was not attained under this evaluated agent/operator contract. Formal four-family matched-performance mechanism identification was therefore not evaluated. Failure to satisfy the criteria on this finite grid does not establish impossibility at untested settings or invalidate predictive characterization on the original study population.

Habitat measurement and task-grounding assessment. A separate development bank contains 3,000 states (200 in each of 15 scenes), split into 1,500 normalization and 1,500 assessment states.

Reversible static queries satisfy repeatability criteria; genuine positive validity responses support normalization without channel deletion or imputation. Empirical acquisition magnitude is estimable on 1,497/1,500 assessment states; three remain Not Estimable. These checks validate the measurement implementation, not task information. A preregistered 60-state task-grounding subset uses within-scene, direction/magnitude-stratified correlations between observation response and fixed-agent action-distribution total variation, with 5,000 whole-scene bootstrap draws under fixed resampling. The primary mean correlation is essentially null: $\rho = 0.004$ $[-0.341, 0.356]$. Only nine scenes meet support criteria, below the required 12; intervals describe those nine, not all 15. Six retained shortest-path inconsistencies reproduce in a separate verification cohort. The positive secondary logit association does not establish satisfaction of the primary criteria. The evaluated protocol did not establish canonical Habitat witness validity. Private action distributions remain White evaluator-only; no HM3D quotient recovery or causal mechanism localization is claimed.

G Supplementary HM3D Tables

G.1 Access-conditioned characterization

Table A6: HM3D access-conditioned characterization on 2,996 validation fault cases. Macro F1 and balanced accuracy with 95% hierarchical intervals; probe fitting uses development scenes.

Access	Macro F1 [95% CI]	Balanced accuracy [95% CI]
Black	0.4264 [0.3953,0.4546]	0.4413 [0.4103,0.4700]
Grey	0.8093 [0.7890,0.8315]	0.8124 [0.7937,0.8340]
White	0.9930 [0.9883,0.9967]	0.9930 [0.9883,0.9967]

Table A7: Conditional HM3D feature-block contributions: White-minus-ablated macro F1 with paired 95% intervals. These contrasts describe predictive contributions, not causal module localization.

Removed block	Contribution [95% CI]
Policy logits	0.0970 [0.0848,0.1096]
Fusion	0.0113 [0.0040,0.0186]
Recurrent state	0.0080 [0.0020,0.0140]
Depth encoder	0.0053 [0.0010,0.0107]
RGB encoder	0.0003 [−0.0023,0.0030]

G.2 Predictive comparisons

Table A8: HM3D predictive comparisons on the same 2,996 validation cases. Macro F1 with 95% intervals and balanced accuracy; development-only fitting. Fit origin: † denotes a baseline-completeness extension; other fits are Original. These predictive comparisons do not establish quotient recovery.

Model	Macro F1 [95% CI]	BA
Outcome only	0.3520 [0.3045,0.3888]	0.3725
Black	0.4264 [0.3953,0.4546]	0.4413
Grey	0.8093 [0.7890,0.8315]	0.8124
White	0.9930 [0.9883,0.9967]	0.9930
Generic White MLP	0.9957 [0.9910,0.9990]	0.9957
First-step tensor	0.8233 [0.8022,0.8428]	0.8234
First-10-step summary	0.9417 [0.9291,0.9527]	0.9419
Entropy only†	0.3303 [0.3121,0.3477]	0.3862
Confidence/margin†	0.3171 [0.2924,0.3376]	0.3628

G.3 Intervention and restoration

Table A9: HM3D diagnosis-guided bounded intervention endpoint recovery and action counts on 2,995 common observed cases. Success and SPL include 95% intervals; total actions include evaluation and retry.

Arm	Success [95% CI]	SPL [95% CI]	Endpoint actions	Total actions
No intervention	0.747 [0.649, 0.841]	0.576 [0.495, 0.652]	178.8	178.8
Generic repair	0.769 [0.675, 0.859]	0.608 [0.527, 0.685]	162.3	341.1
Stage-guided	0.831 [0.754, 0.905]	0.663 [0.598, 0.721]	154.0	332.8

Table A10: HM3D diagnosis-guided bounded intervention guided-minus-generic effects by failure family, with frozen paired 95% intervals. Success differences are percentage points (pp).

Study	Family	Paired cases	Guided − generic Success, pp	Guided − generic SPL
Validation retry	Sensor	749	-0.40 [-1.20, 0.00]	-0.0019 [-0.0069, 0.0004]
Validation retry	Representation	749	0.13 [-6.67, 6.27]	-0.0059 [-0.0649, 0.0490]
Validation retry	Fusion	748	12.83 [6.42, 19.07]	0.1189 [0.0789, 0.1593]
Validation retry	Decision	749	12.15 [6.82, 17.73]	0.1076 [0.0666, 0.1482]

Table A11: Separate matched-nominal extension: pooled task-normalized restoration with frozen 95% intervals. These ratios measure task-score restoration, not causal or witness restoration.

Arm	Success restoration [95%]	SPL restoration [95%]
Generic repair	0.202 [0.095, 0.331]	0.271 [0.186, 0.388]
Stage-guided	0.761 [0.558, 1.023]	0.729 [0.623, 0.836]

G.4 Held-out validation

Table A12: White resolution model performance at primary and target severity: discrimination metrics with 95% intervals and calibration summaries.

Severity	Cases	Macro F1 [95%]	BA [95%]	ECE	Brier
Primary	5945	0.9916 [0.9868, 0.9958]	0.9916 [0.9868, 0.9958]	0.0083	0.0150
Target	5950	0.9919 [0.9876, 0.9956]	0.9919 [0.9876, 0.9956]	0.0044	0.0126

Table A13: Primary-severity held-out endpoint outcomes and action counts on 5,942 common observed cases. Success and SPL include 95% intervals; total repair actions include diagnosis plus retry.

Arm	Success [95% CI]	SPL [95% CI]	Endpoint actions	Total actions
No intervention	0.720 [0.635, 0.798]	0.547 [0.479, 0.610]	182.9	182.9
Generic repair	0.746 [0.664, 0.819]	0.585 [0.515, 0.649]	164.6	347.5
Stage-guided	0.784 [0.710, 0.851]	0.617 [0.548, 0.680]	173.0	355.9

Table A14: Primary-severity held-out guided-minus-generic family effects with frozen paired 95% intervals. Success differences are percentage points (pp).

Study	Family	Paired cases	Guided − generic Success, pp	Guided − generic SPL
Primary held-out	Sensor	1486	0.13 [-0.20, 0.53]	0.0004 [-0.0023, 0.0034]
Primary held-out	Representation	1485	-3.03 [-5.52, -0.41]	-0.0353 [-0.0554, -0.0141]
Primary held-out	Fusion	1485	7.07 [3.74, 10.49]	0.0714 [0.0413, 0.1010]
Primary held-out	Decision	1486	11.10 [7.30, 14.98]	0.0884 [0.0605, 0.1162]

Table A15: Target-severity repair extension: endpoint outcomes and total actions on its 5,942-case common observed support, with separately declared 10,000-draw hierarchical 95% intervals. This support need not match the primary-severity intersection.

Arm	Success [95% CI]	SPL [95% CI]	Total actions [95% CI]
No intervention	0.6481 [0.5552, 0.7353]	0.4816 [0.4149, 0.5449]	213.14 [182.89, 243.92]
Generic repair	0.6851 [0.6001, 0.7633]	0.5303 [0.4627, 0.5925]	399.08 [339.59, 460.18]
Stage-guided	0.7683 [0.6937, 0.8355]	0.5968 [0.5329, 0.6559]	395.92 [330.76, 462.74]

Table A16: Target-severity guided-minus-generic effects, pooled and by family, with frozen 95% intervals. Success differences are fractions; action differences include diagnosis plus retry.

Family	Metric	Guided minus generic	95% CI
Pooled	Success	+0.0831	[+0.0643, +0.1025]
Pooled	SPL	+0.0664	[+0.0524, +0.0818]
Pooled	Total actions	-3.16	[-13.48, +7.30]
Sensor	Success	-0.0007	[-0.0033, +0.0000]
Sensor	SPL	-0.0011	[-0.0035, +0.0001]
Sensor	Total actions	+0.18	[+0.00, +0.55]
Representation	Success	-0.0323	[-0.0632, -0.0047]
Representation	SPL	-0.0343	[-0.0566, -0.0132]
Representation	Total actions	+15.85	[+4.67, +27.93]
Fusion	Success	+0.1853	[+0.1414, +0.2324]
Fusion	SPL	+0.1619	[+0.1250, +0.2017]
Fusion	Total actions	-91.51	[-110.86, -73.80]
Decision	Success	+0.1805	[+0.1461, +0.2176]
Decision	SPL	+0.1395	[+0.1150, +0.1663]
Decision	Total actions	+62.77	[+34.50, +93.20]

G.5 Resolution-blind repair control tables

Table A17 retains pooled, neutral and adverse family comparisons under the original cost convention.

Table A17: Diagnosis-blind allocation on 2,995 observed cases: arm means and paired contrasts [95% intervals]. Success is percent for arms and percentage points for contrasts. G: guided; R: random; N: no intervention; C: generic. Family rows are G–R. Intervals condition on one scene-wise mixture-preserving assignment; secondary/family contrasts are descriptive.

Arm / contrast	Success	SPL	Total actions
No intervention	74.69	0.5757	178.79
Generic	76.93	0.6081	341.07
Guided	83.11	0.6628	332.81
Random	76.09	0.5910	354.03
G–R	7.01 [4.65, 9.89]	0.072 [0.058, 0.086]	-21.23 [-29.11, -14.16]
R–N	1.40 [-1.30, 4.14]	0.015 [-0.006, 0.037]	175.25 [136.59, 212.01]
R–C	-0.83 [-2.97, 1.40]	-0.017 [-0.040, 0.005]	12.96 [7.22, 19.05]
Sensor	6.01 [2.41, 9.88]	0.097 [0.065, 0.132]	-46.10 [-60.13, -32.72]
Representation	1.74 [-4.80, 7.75]	0.009 [-0.044, 0.058]	-4.48 [-26.75, 21.63]
Fusion	11.36 [7.49, 15.20]	0.102 [0.074, 0.133]	-53.40 [-68.31, -37.83]
Decision	8.95 [4.93, 12.93]	0.080 [0.051, 0.107]	19.03 [8.00, 31.08]

Random repair rescues/harms 55/46 Sensor, 32/42 Representation, 53/36 Fusion, and 52/26 Decision cases relative to no intervention. Sensor random-minus-generic Success is -6.41 points $[-10.68, -2.54]$ and SPL -0.0988 $[-0.1363, -0.0651]$. These family outcomes preclude a uniform-benefit reading.

G.6 Calibration matching assessment tables

The counts of criterion violations refer to the complete evaluated calibration subset. Minimum discrepancies are marginal minima and need not occur at the same tuple.

Table A18: Four-family calibration matching on seven scenes: 699 identities per condition, 31 conditions and 21,669 rollouts. (a) Of 2,772 tuples, four fail Success and SPL only; 2,768 fail all three criteria. (b) Exploratory pairwise task-performance matching on the specified grid. Pairs must jointly satisfy the Success, SPL and length criteria; passing pairs are not independently validated provenance results.

(a) Four-family criteria

Gate	Frozen requirement	Minimum observed	Failing tuples
Success spread	≤ 0.01	0.014306	2,772
SPL spread	≤ 0.02	0.021563	2,772
Maximum length SMD	< 0.20	0.194945	2,768
Joint qualification	All three gates	—	2,772

(b) Exploratory pairwise matching

Family pair	Combinations	Joint matches
Sensor–Representation	42	0
Sensor–Fusion	36	0
Sensor–Decision	66	0
Representation–Fusion	42	3
Representation–Decision	77	0
Fusion–Decision	66	0

Representation/Fusion severities $(0.35, 0.50)$, $(0.425, 0.625)$ and $(0.725, 1.0)$ satisfy the pairwise criteria; none is evaluated in a subsequent experiment. Overlapping marginal performance ranges do not establish a common joint operating point. The complete negative four-family result is retained. Table A18 summarizes the limits of task-performance matching; pairwise matches concern task-performance comparability, not mechanism identification.

H Additional Stress Tests

These tests assess reporting under witness corruption and catalog withholding, using the shared protocol in Appendix D.

H.1 Witness Noise Robustness

How do corrupted witness measurements affect resolution and reporting? Independently per coordinate, noise rate $\eta \in \{0, 0.05, 0.10, 0.20\}$ replaces a value uniformly with a different value in its public finite alphabet. One corrupted four-coordinate vector supplies all nested access prefixes. Graph and test corruptions are independent; conditions use paired random numbers. Graph distances and uncertainty are recomputed, but the clean confidence score tables and numerical thresholds remain fixed. This is a measurement-shift stress test, not noise-adapted recalibration or sensor-noise simulation. Each noisy finite-resolution partition is certified before evaluation; the clean partition is not imposed on shifted witness laws.

At $\eta = 0.05$, partition recovery and certification remain perfect while coverage already declines. At 0.10, White still has ARI one but certifies only 4.28 cells on average (Table A20; Figure A5). At 0.20, White's finite-resolution operational partition merges R/F because their witness-law distance is 0.174 below the declared tolerance, although their full witness laws remain distinct (Figure A6). The recovered six-cell partition instead splits that population pair, giving ARI zero against the noise-conditioned population partition. This is one false-split pair, not error on every pair. Black retains ARI one but certifies no cells and abstains on all test groups. Coverage at this noise level is approximately 73.33%/63.13%/53.58% across accesses. No false merges or refinement violations occur.

Abstention is not generally monotone: corruption can yield empty sets, uncertified candidates or wrong commitments under the unchanged clean rule. At 0.20, empty-set rates are 0.125/0.197/0.314, while uncertified-graph abstention rates are 0.875/0.510/0.463; technical failures are zero. Clean

calibration guarantees are not asserted under this shift. These results do not establish preserved coverage under noise or reliability beyond the evaluated corruption model.

Table A20: Bounded witness noise. Cells give population/recovered counts; certification is separate. Coverage and abstention use the frozen clean reporting rule, and all recovery scores use the noise-conditioned population partition. At $\eta = 0.20$, White has one false-split population pair; its exact recovery is zero, while all other rows have exact recovery one.

η	Access	Cells	Certified	ARI	Coverage	Abstention
0.00	Black	4/4	4	1.000	1.000	0.750 [0.749, 0.750]
0.00	Grey	5/5	5	1.000	1.000	0.500 [0.499, 0.501]
0.00	White	6/6	6	1.000	1.000	0.750 [0.749, 0.750]
0.05	Black	4/4	4	1.000	0.930 [0.929, 0.931]	0.737 [0.737, 0.738]
0.05	Grey	5/5	5	1.000	0.897 [0.897, 0.898]	0.517 [0.516, 0.518]
0.05	White	6/6	6	1.000	0.869 [0.869, 0.870]	0.756 [0.755, 0.757]
0.10	Black	4/4	4	1.000	0.862 [0.861, 0.863]	0.726 [0.725, 0.727]
0.10	Grey	5/5	5	1.000	0.802 [0.801, 0.802]	0.532 [0.531, 0.532]
0.10	White	6/6	4.28 [4.08, 4.48]	1.000	0.751 [0.750, 0.752]	0.762 [0.760, 0.764]
0.20	Black	4/4	0	1.000	0.733 [0.732, 0.734]	1.000
0.20	Grey	5/5	3	1.000	0.631 [0.630, 0.633]	0.706 [0.705, 0.707]
0.20	White	5/6	3	0.000	0.536 [0.534, 0.538]	0.777 [0.775, 0.780]

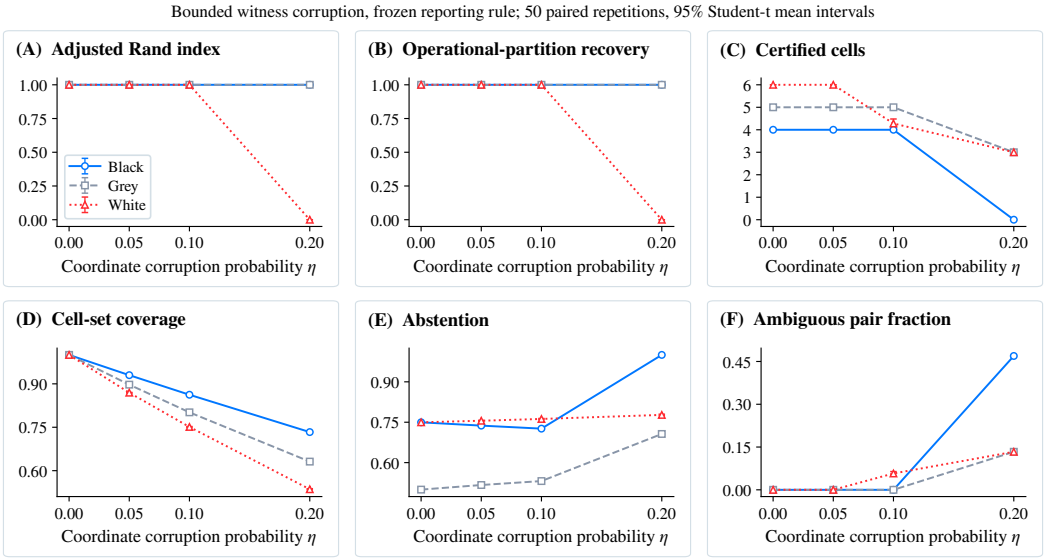

Figure A5: Recovery, certification, coverage and abstention under bounded witness corruption with frozen clean reporting thresholds. Preserved ARI does not ensure certification or shifted coverage; the adverse results are part of the stress-test evidence.

H.2 Unknown Mechanism Withholding

How does a known-catalog reporter respond to one explicitly withheld mechanism? We fit only N/S/R/F/D1 and expose D2 test traces after all known-only fits and validation criteria are fixed. D2 retains its original decision operator and activation prevalence. Its rows are excluded from graph calibration and confidence fitting. The public witness alphabet, kernel, feature map and conservative radius constants remain fixed. Catalog-size indexing is restricted to five means and ten pairs per

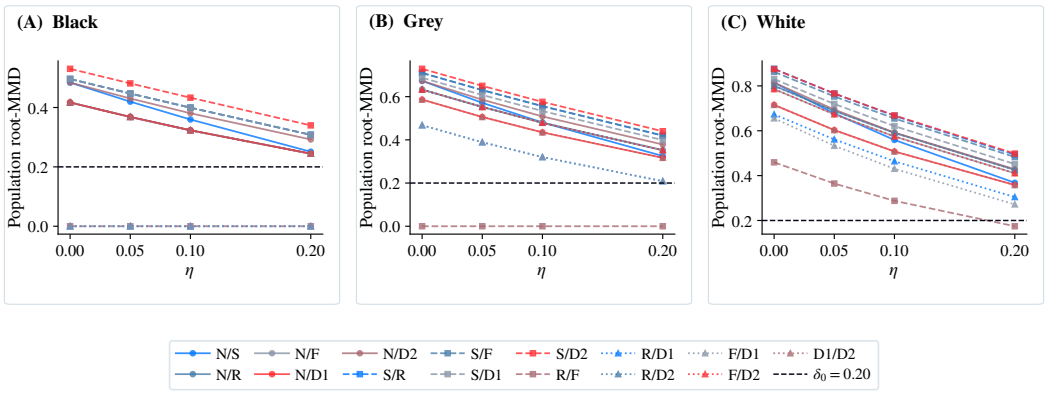

Figure A6: Noise-conditioned population pairwise distances under the fixed kernel and tolerance. The White R/F pair falls below the declared tolerance at the largest noise rate; distinct full laws need not remain distinct tolerance cells.

access; the unchanged Bonferroni calibration rule uses five known mechanisms, giving rank 1015 rather than the six-catalog rank 1017. No unknown threshold is fitted on D2.

All known-only fits recover and certify 3/4/5 cells, with ARI, exact recovery and coverage one. On withheld D2 queries, total abstention is one and no committed known-cell assignment is observed (Table A21). Empty-set reporting is 0.752 and multiple-candidate abstention is 0.248, identically across accesses. D2's active overclaim signature is already visible at Black access; this protocol does not demonstrate improved unknown-mechanism recognition with richer access. The remaining zero-witness component is shared with known mechanisms and cannot identify catalog membership by itself.

Total abstention is not universal recognition of an unknown class: nonempty sets may still contain known candidates. This evaluates withholding within a declared mechanism catalog; open-world unknown detection requires additional protocols. No unknown-query ARI or unknown-population coverage is defined, and guarantees do not extend to arbitrary unknown laws that may fall inside known confidence regions.

Table A21: Known versus withheld reporting. All known fits have exact recovery and coverage one, with no false commitments. Withheld abstention separates empty sets from multiple known candidates; neither is an unknown-class accuracy measure. No technical or uncertified-graph abstention occurs.

Access	Known abstention	D2 abstention	D2 empty	D2 multiple	D2 commitment
Black	0.400 [0.399, 0.401]	1.000	0.752 [0.748, 0.755]	0.248 [0.245, 0.252]	0.000
Grey	0.550 [0.549, 0.551]	1.000	0.752 [0.748, 0.755]	0.248 [0.245, 0.252]	0.000
White	0.850 [0.849, 0.850]	1.000	0.752 [0.748, 0.755]	0.248 [0.245, 0.252]	0.000

I Controlled-Shift and Cross-Environment Evidence Contracts

This supplement analyzes controlled shifts and the separate HM3D–Replica comparison in Section I. Both assess empirical measurement contracts, without extending mechanism-attribution or quotient-recovery claims. No unseen-scene evaluation or target fitting is introduced. The studies have distinct panels, query schedules and bootstrap procedures; their rollouts are not pooled.

I.1 Predefined Protocol and Stage-Local Operators

The controlled shifts test the evidence contract by perturbing FOV, depth statistics, decision noise and fusion-input history while holding the PointNav RGB-D ResNet/GRU adapter, checkpoint, categorical policy, preprocessing and five-stage hooks fixed. Source sensor and navigation settings are specified in Appendix I.5. The panel comprises 15 designated HM3D development scenes and the

first two episodes per scene in the predefined witness-validation order, without outcome filtering or replacement. Two paired randomization blocks provide 60 rollouts per setting; the shared source and ten settings across five completed branches provide 660 analyzed rollouts. Reused scenes and source references do not provide independent confirmation on unseen support.

- **FOV:** change native RGB/depth FOV to 69° or 89°, with raster, extrinsics and preprocessing fixed.
- **Depth noise:** before preprocessing, valid depth becomes $\min\{10, d(1 + \eta u)\}$, with pixelwise $u \sim \mathrm{Uniform}[-1, 1]$ and $\eta \in \{0.01, 0.05\}$. Invalid zero depth and RGB are preserved.
- **Depth scale:** valid depth becomes $\min\{10, sd\}$ for $s \in \{0.9, 1.1\}$ at the same sensor-input boundary, preserving zero-invalid support and RGB.
- **Decision noise:** add independent per-action $\mathrm{Uniform}[-a, a]$ noise to raw logits, $a \in \{0.05, 0.20\}$, before the original categorical distribution. Given identical inputs, encoders, fusion, recurrent state and value output are unchanged.
- **Fusion-input lag:** the operator supplies the compressed depth feature from $t - k$, $k \in \{1, 2\}$, to the depth/goal combiner. Its first 32 input channels are delayed; the remaining current goal channels and RGB path are preserved. Current depth encoding/compression is retained. A two-step feature history is reset between episodes and updated only on committed steps. Warmup $t < k$ uses current input and cannot support an active-lag claim. Neither recurrent-state delay nor raw-depth synchronization delay is used.

Only the declared boundary is directly perturbed; downstream activations and later trajectories may change. PCG64 noise randomization is fixed by operator, episode, timestep and block, independent of policy randomization; paired query signs and severity levels share the same noise field. Action randomization is fixed by episode and block and paired across conditions. No operator, threshold or severity was selected after outcomes.

I.2 Local reporting and uncertainty

At steps 2, 8 and 32, before the committed action, each condition is queried along its own closed-loop trajectory. Six symmetric pairs use forward/lateral translations of ± 0.05 and ± 0.10 m and yaw of $\pm 2^\circ$ and $\pm 5^\circ$. Queries restore pose, sensors, recurrent input and RNG, with fixed history and no time/reward or buffer advance. Correct goal coordinates vary with query pose; responses are not attributable to visual changes alone. Paired episodes and seeds do not imply equal later poses or hidden states. Early termination retains missing scheduled states in the requested denominator; invalid or unavailable responses are not zero-imputed.

The reporter reuses fixed development modality/stage scales, a joint-valid depth fraction criterion of 0.1 and numerical response floor 10^{-8}. Black exposes public observations/action responses; Grey adds current encoder and fusion responses; White adds recurrent and logit responses. A supported stage-query contrast is resolved when its source-normalized RMS response exceeds the floor. Below-floor responses are not certified equivalences. Availability is supported/requested contrasts, resolution is resolved/requested, and abstention is measurement refusal under inadequate support or missingness. These are neither population quotient cells nor calibrated confidence sets. Conditional magnitude comparisons use query slots supported in both paired conditions; unavailable summaries remain Not Estimable rather than refitting normalizers.

All intervals below use the same prespecified 5,000 paired whole-scene bootstrap draws: average two episodes and two blocks within each scene, then weight the 15 scenes equally. Intervals are descriptive 95% percentile intervals; no equivalence margin, additional significance test or target calibration is introduced. Table A22 reports task outcomes; Table A23 reports local evidence and abstention. The shared source is counted once.

I.3 Temporal boundary and count saturation

Lag 2 changes Success by -11.67 percentage points [-23.33, -1.67] and SPL by -0.270 [-0.371, -0.167]; mean length increases from 158.87 to 273.75 steps. Yet White local-response retention is 0.970 [0.893, 1.042]. The empirical count is saturated: every supported neural-stage contrast exceeds the numerical floor in every completed setting. Source resolves 4,345/4,345 supported neural contrasts and lag 2 resolves 4,255/4,255. Consequently, in these data the White rate equals

Table A22: Controlled-shift task outcomes, 60 rollouts per setting across 15 reused development scenes. Brackets are 95% scene-bootstrap intervals; ΔSPL is paired shifted minus source. Both resolution levels are excluded by the whole-branch observation gate.

Setting	Success (%)	SPL	ΔSPL
Source	78.33 [66.67, 88.33]	0.649 [0.541, 0.752]	0.000 [0.000, 0.000]
FOV 69°	73.33 [56.67, 88.33]	0.595 [0.452, 0.727]	−0.054 [−0.162, 0.046]
FOV 89°	85.00 [70.00, 96.67]	0.655 [0.530, 0.760]	0.006 [−0.073, 0.078]
Depth $\eta = .01$	81.67 [68.33, 93.33]	0.651 [0.526, 0.767]	0.002 [−0.058, 0.063]
Depth $\eta = .05$	80.00 [66.67, 91.67]	0.641 [0.520, 0.753]	−0.009 [−0.052, 0.040]
Scale $s = .9$	80.00 [66.67, 93.33]	0.645 [0.527, 0.755]	−0.004 [−0.073, 0.064]
Scale $s = 1.1$	86.67 [75.00, 95.00]	0.682 [0.587, 0.773]	0.033 [−0.005, 0.075]
Logits $a = .05$	76.67 [63.33, 88.33]	0.639 [0.521, 0.749]	−0.010 [−0.038, 0.007]
Logits $a = .20$	76.67 [65.00, 88.33]	0.619 [0.501, 0.731]	−0.031 [−0.066, −0.001]
Fusion $k = 1$	81.67 [70.00, 93.33]	0.591 [0.510, 0.671]	−0.058 [−0.125, 0.010]
Fusion $k = 2$	66.67 [53.33, 80.00]	0.379 [0.291, 0.473]	−0.270 [−0.371, −0.167]

Table A23: White local-response rate, availability and measurement abstention (%), with 95% scene-bootstrap intervals. Different access profiles have different exposed-stage denominators; these rates are not classification accuracy. Missing scheduled states remain in the denominator.

Setting	Resolved/requested	Availability	Abstention
Source	68.41 [61.06, 75.69]	80.46 [71.85, 88.89]	19.54 [11.11, 28.15]
FOV 69°	69.26 [61.47, 76.90]	80.74 [71.39, 90.00]	19.26 [10.00, 28.61]
FOV 89°	70.17 [60.89, 78.43]	82.41 [71.57, 92.22]	17.59 [7.78, 28.43]
Depth $\eta = .01$	68.67 [60.45, 76.64]	80.56 [71.02, 89.81]	19.44 [10.19, 28.98]
Depth $\eta = .05$	70.08 [60.74, 78.80]	82.04 [71.30, 92.04]	17.96 [7.96, 28.70]
Scale $s = .9$	69.09 [59.32, 78.21]	81.11 [69.72, 91.76]	18.89 [8.24, 30.28]
Scale $s = 1.1$	70.37 [62.38, 78.03]	82.41 [72.96, 91.30]	17.59 [8.70, 27.04]
Logits $a = .05$	68.36 [60.94, 75.65]	80.46 [71.85, 88.89]	19.54 [11.11, 28.15]
Logits $a = .20$	68.61 [60.85, 76.23]	80.74 [71.85, 89.44]	19.26 [10.56, 28.15]
Fusion $k = 1$	70.94 [63.06, 78.01]	83.80 [74.44, 92.13]	16.20 [7.87, 25.56]
Fusion $k = 2$	66.36 [58.21, 74.03]	78.80 [69.17, 87.87]	21.20 [12.13, 30.83]

$(5A + B)/6$, where A is query availability and B is the Black public-action resolved/requested rate. Count retention therefore cannot establish retained supported failure resolution.

Continuous responses reveal changes obscured by the count. On common supported query slots, fusion-response retention is 0.693 [0.672, 0.716] for lag 1 and 0.711 [0.692, 0.731] for lag 2; logit-response retention is 0.624 [0.597, 0.652] and 0.634 [0.589, 0.684]. These ratios use matched support, not an unmatched all-source mean. Fixed-history queries measure current response conditional on delayed evidence; they do not recover sensitivity to its past acquisition. Co-occurring task and witness changes do not establish causality or a calibrated performance predictor. Common additive decision noise can also cancel in paired raw-logit differences, limiting that response's sensitivity to the declared operator.

Each setting requests 1,080 query pairs. Source has 869 supported, 145 invalid geometry/observation and 66 unavailable after episode termination; lag 2 has 851, 223 and 6, respectively. Longer episodes increase scheduled-state reachability while their trajectories can reduce geometric support. Abstention changes therefore mix observation support and termination effects. Across 1,884 reached diagnostic states, zero-query resolved claims and repeat action disagreements are zero, with zero maximum repeated logit drift. These checks establish measurement repeatability on this support, not probability calibration or confidence-set coverage.

I.4 Resolution exclusion and scope

Resolution changes test the scope of the observation contract. Native resolutions 320×240 and 640×480 satisfied initial checks with fixed FOV and original preprocessing, but 640×480 did not satisfy the predefined observation-validity criterion after two completed rollouts. The criterion requires valid RGB shape, type and variation, and valid depth shape, type, finiteness, range and positive support. The failed component, timing and underlying cause remain unresolved. This is an observation-contract boundary, not a measured Success/SPL failure, and is not scored as zero task performance.

The predefined complete-branch criterion excludes the entire resolution branch: 60 completed 320×240 rollouts and two completed 640×480 rollouts contribute no branch inference; 58 planned resolution rollouts lack completion. No favorable level is retained, and no episode is replaced, criterion relaxed or severity revised. Of 780 planned rollouts, 722 completed, including the 62 excluded resolution records; 660 enter the five-branch analysis. All raw records and exclusions are retained.

This study establishes controlled-shift boundaries for empirical local measurements on reused development support. Numerical-floor saturation, condition-specific trajectories, source normalization and unresolved observation validity limit extensions of diagnostic claims. The result does not certify transferred failure quotients, unique mechanism attribution, task geometry or calibrated abstention.

I.5 Replica validation and specified comparison contract

The cross-environment comparison uses HM3D as the source reference/evaluation domain and Replica as the target; HM3D is not the training domain of the RoboTHOR-trained RobustNav RGB-D PointNav checkpoint. We retain the same checkpoint, backbone, categorical policy, five passive hooks and source normalizers. Habitat-Sim/Lab 0.3.3 runs with physics disabled, RGB/depth 400×300, 79° horizontal FOV, camera offset $(0, 0.88, 0)$ m, raw depth $[0, 10]$ m and original 224×224 preprocessing. Actions use 0.25 m forward steps and 30° turns, with sliding and a 500-action limit; Success requires STOP within 0.2 m geodesic distance. SPL retains the native shortest-path/travelled-path definition. Replica is explicitly a vertex-color appearance condition, not native Ptex rendering.

Panels and observation validation. The fixed source panel comprises 30 episodes, two from each of the 15 designated HM3D development scenes, with published geodesic length 1–30 m. Source selection and randomization follow a predefined comparison contract, distinct from the controlled-shift panel and specified before source outcomes (see accompanying reproducibility details). The target retains 35 episodes across all 18 Replica scenes with unchanged randomization. Of the original 36 episodes, one was excluded before policy evaluation under the empty-frustum observation rule, without replacement or loss of scene coverage. This rule was motivated by an observation-validity incident, not performance outcomes; no renderer, sensor or preprocessing repair was made. The contract requires nonconstant uint8 RGB, finite float32 depth in range with positive support, and fixed dimensions. All 260 start views (four yaw offsets per retained episode) satisfy these criteria and yield identical observations on repeated execution, with no additional exclusions. Source scenes have development/reference exposure and target outcomes from the initial evaluation were already known. These panels therefore do not provide independent confirmation on unseen support.

Evidence interface. Unlike the later-step queries in the controlled-shift study, this comparison queries episode starts only. Six symmetric pairs use forward/lateral $\pm 0.05/0.10$ m and yaw $\pm 2/5^\circ$, restoring simulator state, recurrent input and RNG without episode advance. Black exposes public observations and sampled-action responses, Grey adds encoders/fusion, and White adds recurrent/logit responses and private confidence. Fixed HM3D development scales normalize empirical

RGB, depth and depth-validity responses; the joint-depth support criterion remains 0.1, numerical floor 10^{-8} and confidence/low-support thresholds 0.8/0.5. Invalid queries and empty denominators remain unavailable or Not Estimable; no target refitting occurs. These measurements do not supply exact Fisher geometry or calibrated confidence-set coverage.

Distance-validation correction. Cached-distance validation initially excluded 29 source episodes before any actions. Independent native-navmesh assessment confirmed connectivity for all 30 and a maximum cached/runtime geodesic discrepancy of 0.204 m; the cause remains unknown. Validation therefore uses the original HM3D criterion: agreement between runtime and independent native-navmesh distances within the same 10^{-5} tolerance. Cached metadata, source selection, Success/SPL definitions and all scientific thresholds are unchanged. Completed outcomes are retained; the correction applies only to cases without prior actions. This validation correction is not a cross-environment finding.

I.6 Cross-environment measurements and uncertainty

All 30/30 HM3D and 35/35 Replica episodes complete, with zero runtime-invalid observations. All 65 initial hook-neutrality, zero-query repeatability and metric-formula checks satisfy the specified criteria; model parameters remain identical. The 35 target action logs reproduce the initial evaluation exactly, establishing repeatability, not additional independent evidence. Table A24 reports the formal fixed-panel estimates.

Specified statistical procedure. Uncertainty uses 5,000 whole-scene bootstrap draws under a fixed resampling protocol, episode-weighted means and 95% percentile intervals. Domains are resampled separately; cross-domain differences are unpaired, while within-domain access comparisons share episode/scene draws. Source normalizers remain fixed. The declared dependence sensitivity groups Replica by five name-prefix families (apartment, frl_apartment, hotel, office, room), leaving HM3D scene blocks unchanged. This is an assumed dependence scenario, not verified physical-site identity or a necessarily conservative interval. Both interval sets are retained. These are conditional resampling summaries of exposed panels, not population guarantees; source-reference fitting uncertainty is not propagated. No equivalence margin, new significance test or target tuning is introduced.

Table A24: Cross-environment fixed-panel comparison: 30 HM3D episodes/15 scenes and 35 Replica episodes/18 scenes. Primary brackets are 95% whole-scene bootstrap intervals; the last column gives the five-prefix-family Replica sensitivity interval. Availability is episode-start query support, not certified resolution or predictive accuracy. Task outcomes are context for nonmatched panels, not a domain ranking.

Metric	HM3D	Replica	Replica sensitivity
Success (%)	73.3 [56.7, 86.7]	94.3 [85.7, 100.0]	[91.3, 100.0]
SPL	0.582 [0.435, 0.721]	0.836 [0.748, 0.908]	[0.774, 0.898]
Length (actions)	184.67 [116.83, 256.30]	43.77 [25.53, 74.74]	[25.32, 64.75]
Witness support (%)	55.6 [41.7, 70.6]	71.4 [62.0, 81.0]	[63.6, 78.9]
Technical abstention (%)	44.4 [29.4, 58.3]	28.6 [19.0, 38.0]	[21.1, 36.4]
Magnitude available (%)	100.0 [100.0, 100.0]	97.1 [91.2, 100.0]	[88.9, 100.0]
White nonzero/requested (%)	47.1 [35.1, 59.9]	59.5 [51.7, 67.5]	[53.0, 65.8]
White confidence available (%)	23.3 [6.7, 43.3]	48.6 [32.4, 63.9]	[33.3, 69.2]

Availability, abstention and response counts. HM3D supports 100/180 query pairs and Replica 150/210. Unsupported pairs comprise 80 geometric incidents on HM3D and 54 geometric plus six joint-depth-support incidents on Replica. Technical abstention is 44.4% and 28.6%, respectively; this is measurement refusal, not calibrated classification abstention. Acquisition magnitude is estimable for 30/30 and 34/35 states. The unsupported Replica state remains in the denomina-

tor. White confidence/support reporting is estimable for 7/30 and 17/35 states; empty low-support denominators remain Not Estimable.

Black/Grey/White availability is identical by the common acquisition-support criterion, not because the profiles carry equal information. Their available/requested stage slots are 100/180, 400/720 and 600/1,080 on HM3D, versus 150/210, 600/840 and 900/1,260 on Replica. Nonzero/requested responses are 9/180, 309/720 and 509/1,080 versus 0/210, 450/840 and 750/1,260. All supported neural-stage contrasts exceed the fixed floor, limiting that count's ability to characterize supported distinctions. Common-RNG sampled actions tie on every supported Replica pair; this does not establish that public access is generally uninformative. The paired White-minus-Black response-fraction contrasts are 42.13 percentage points [31.02, 54.63] and 59.52 [51.69, 67.52], with Replica sensitivity [53.03, 65.79]. Different exposed-channel denominators preclude interpreting these as classification gains or recovered quotient cells.

Replica-minus-HM3D availability is $+15.87$ percentage points $[-1.67, 32.50]$, with sensitivity $[-0.50, 31.71]$; both intervals span zero. No equivalence or consistent availability gain is established. Task contrasts are $+20.95$ percentage points Success [4.44, 38.33], $+0.254$ SPL [0.090, 0.419], and -140.90 actions $[-216.65, -68.47]$. Their sensitivity intervals are [5.83, 38.21], [0.103, 0.411], and $[-215.10, -69.19]$. These describe different fixed panels, not causal transfer effects: mean initial geodesic route lengths are 9.18 m on HM3D and 4.33 m on Replica. Geometry, appearance, episode generation and observation-conditioned sampling differ together; the panels are not matched in difficulty or task performance.

Claim boundary. The same specified measurement/reporting contract remains executable while realized support and task outcomes change. Neither task differences nor numerical repeatability identify causal dataset effects, persistent supported failure resolution or stable confidence calibration. The observation-qualified population, vertex-color appearance, exposed target bank, small per-scene samples and episode-start-only queries restrict interpretation. No failure classifier, repair policy or quotient partition is transferred in this comparison. The result supports an explicit evidence-contract comparison across the evaluated environments, not a benchmark ranking or a diagnostic guarantee for new environments.