CAUSAL ORBITALS: ORDER-SENSITIVE PROBING FOR CERTIFIABLE TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) aims to adjust model parameters on unlabeled target batches during deployment, yet unsupervised heuristics risk catastrophic performance degradation (*negative adaptation*) when distinct causal mechanisms produce observationally ambiguous inputs with conflicting task-risk gradients. We prove a label-free no-free-lunch theorem establishing that passive observation cannot resolve causal ambiguity under observationally equivalent marginals without structural inductive bias. Grounded in a structural causal model with offline-calibrated mechanism prototypes, we introduce **Causal Orbitals**, which formulates test-time evidence acquisition as an active, order-sensitive diagnostic probing process. We establish necessary and sufficient geometric conditions for certifiable surrogate descent: an update exists if and only if the origin is strictly excluded from an empirical robust gradient uncertainty hull $\Gamma_\delta(\pi)$; otherwise, adaptation is uncertifiable, mandating safe abstention ($d^* = 0$). Epistemic and sequential uncertainty are tracked via a composite classical–operator state $\Omega = \sum_m \pi(m)|m\rangle\langle m| \otimes \rho_m$, where off-diagonal coherences capture non-commutative response statistics under ordered self-supervised perturbations. We plan diagnostic probing sequences via Bellman recursion on Ω to contract the credible mechanism set under informative outcomes and expose certifiable common descent directions, while split conformal calibration provides distribution-free finite-sample safety control over deployed empirical batch risk ($\mathbb{P}(\Delta\widehat{R}_E(d^*) \geq 0, d^* \neq 0) \leq \alpha$). Across causal shift interferometry, ImageNet-C corruptions, and multi-center MedMNIST benchmarks, combining operator tracking with Bellman lookahead achieves 81.7% certified coverage with a 2.1% empirical unsafe rate ($< \alpha = 5.0\%$). Factorial evaluations show that value-based Bellman planning improves coverage by $+12.2\%p$ over myopic heuristics ($+6.3\%p$ from the certifiable value objective, $+5.9\%p$ from search depth), while the operator state provides a compact algebraic representation for order-dependent responses ($+2.9\%p$ coverage over matched recurrent baselines).

1 INTRODUCTION

Machine learning systems deployed in open-world environments routinely suffer from unforeseen distribution shifts that degrade task performance (Hendrycks & Dietterich, 2019; Koh et al., 2021). Test-time adaptation (TTA) seeks to mitigate this vulnerability by updating model parameters on unlabeled target batches on the fly during inference (Wang et al., 2021). Existing methods predominantly rely on unsupervised surrogate objectives—such as entropy minimization (Wang et al., 2021), pseudo-labeling (Lee et al., 2024), or contrastive representation alignment (Chen et al., 2022). While effective under mild, benign shifts, these approaches rest on an unstated foundational premise: that an unlabeled target batch inherently encodes sufficient statistical information to identify a beneficial parameter trajectory (Gulrajani & Hashimoto, 2022; Jhawar & Wang, 2026). In safety-critical deployments, this premise catastrophically fails. Passive, unlabeled target observations rarely identify *why* a model has shifted. For instance, a severe attenuation of high-frequency spatial components in visual recognition can arise from diverse, mutually incompatible physical mechanisms: optical defocus blur, bilinear downsampling, lossy compression, or domain-specific sensor filtering (Castro et al., 2020). Although these distinct shifts frequently induce near-identical input marginals and indistinguishable prediction confidence profiles, they demand fundamentally divergent, often dia-

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.

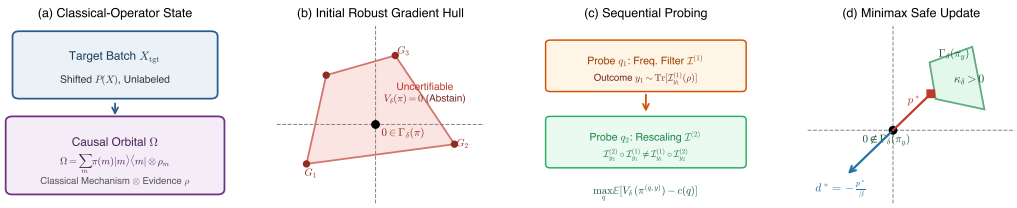


Figure 1: **Overview of the Causal Orbitals TTA pipeline.** **(a)** An unlabeled target batch X_{tgt} is mapped to a classical–operator state $\Omega = \sum_m \pi(m)|m\rangle\langle m| \otimes \rho_m$, separating classical causal hypotheses from operator-valued probe evidence. **(b)** Before probing, observationally compatible mechanisms induce opposing gradients whose robust gradient hull $\Gamma_\delta(\pi)$ contains the origin ($0 \in \Gamma_\delta(\pi)$), rendering adaptation uncertifiable ($V_\delta(\pi) = 0$) and forcing abstention. **(c)** Order-sensitive noncommuting probes ($q_1 \to q_2$) are actively selected via Bellman recursion to maximize certifiable value. **(d)** Informative measurement outcomes contract the credible mechanism set $\mathcal{M}_\delta(\pi)$, yielding $0 \notin \Gamma_\delta(\pi_y)$ with positive margin κ_δ, enabling the minimax optimal descent update $d^* = -p^*/\beta$.

metrically opposed parameter corrections (Fan et al., 2026). Adapting toward an erroneous point estimate of the shift can improve performance on one compatible environment while precipitating severe performance collapse on another—a failure mode termed *negative adaptation* (Zhao et al., 2023). This paradigm rests on two core theoretical pillars: (1) *Convex Gradient Geometry*: By establishing empirical robust gradient uncertainty balls G_m centered at empirical prototypes $\hat{g}_m$, we prove that a guaranteed descent direction exists if and only if the origin is excluded from the active credible gradient hull $\Gamma_\delta(\pi)$. When $0 \in \Gamma_\delta(\pi)$, no parameter update can guarantee surrogate descent, rendering abstention ($d^* = 0$) mandatory under the robust surrogate descent criterion. (2) *Active Ordered Probing*: When passive observation leaves the origin entrapped in the hull ($0 \in \Gamma_\delta(\pi)$), the model actively extracts diagnostic features from the input batch X via structured non-commuting self-supervised transformations (e.g., frequency filtering followed by spatial rescaling). We model target epistemic and sequential uncertainty using a composite classical–operator state $\Omega = \sum_m \pi(m)|m\rangle\langle m| \otimes \rho_m$, where off-diagonal density entries capture non-commutative response statistics. Active diagnostic probing seeks to contract the credible mechanism set $\mathcal{M}_\delta(\pi)$, separating conflicting gradients and exposing certifiable common descent cones.

Related Work and Positioning. Existing TTA methods update parameters via unsupervised heuristics (Zhang et al., 2022; Wang et al., 2021; Lee et al., 2024) or sharpness-aware resets (Niu et al., 2022; 2023; Wang et al., 2022). However, passive heuristics leave causal ambiguity unaddressed, risking negative adaptation (Zhao et al., 2023; Song et al., 2023; Liang et al., 2023). In contrast to static invariant representation learning (Arjovsky et al., 2019; Schölkopf et al., 2021) that sacrifices nominal capacity, Causal Orbitals preserves representations and dynamically identifies active shifts via diagnostic probes. Crucially, while standard TTA operates under zero mechanism supervision, Causal Orbitals meta-calibrates candidate prototypes offline, benchmarking directly against matched-supervision baselines with standard TTA as an unguided reference. Building on active testing (Kossen et al., 2021; Teney & van den Hengel, 2019) and non-commutative operational models (Holevo, 2011; Busemeyer & Bruza, 2012), CPTP instruments provide a compact algebraic inductive bias for order effects, coupled with split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021) for finite-sample safety control (extended survey in Appendix E).

Contributions. **(1) Convex Gradient Geometry & Minimax Safe Adaptation:** We formalize empirical robust gradient balls G_m and establish complete necessary and sufficient conditions for certifiable robust-surrogate descent (Theorem 2), deriving the unique closed-form minimax update $d^* = -\text{Proj}_{\Gamma_\delta(\pi)}(0)/\beta$ and guaranteed descent value $V_\delta(\pi) = \kappa_\delta^2/(2\beta)$. **(2) Causal Orbitals & Structured Algebraic Inductive Bias:** We formulate the classical–operator state Ω and delineate its exact classical reduction boundary: when initial states share a common diagonal basis and instruments preserve this diagonal algebra, Ω reduces identically to a classical Bayesian filter (Theorem 3); furthermore, Theorem 4 demonstrates that non-commuting instruments provide a compact algebraic inductive bias yielding directional separation where single-probe marginals are uninforma-

tive, avoiding combinatorial state-space explosion. **(3) Dynamic Bellman Probing:** We establish measurement-induced certifiability (Proposition 5), showing that diagnostic probing conditionally increases expected guaranteed value when informative outcomes evict the origin from the hull, and formulate probe sequence selection as a Bellman recursion that directly maximizes guaranteed descent value net of inquiry costs. **(4) Finite-Sample Conformal Safety & Coverage Expansion:** We establish finite-sample marginal safety control over deployed empirical batch cross-entropy risk via split conformal gating on guaranteed value $V_\delta(\pi) > \hat{\tau}_\alpha$ (Theorem 7). Through factorial evaluation across representations and search policies, we demonstrate that combining operator tracking with Bellman lookahead achieves 81.7% certified coverage at $\alpha = 0.05$ (suppressing deployed negative adaptation to 1.4%), with ablations confirming that certifiability-guided planning drives significant coverage expansion ($+2.9\%p$ over matched GRU, $p = 0.0044$) under identical safety levels.

2 Problem Setup and Limits

SCM and Adaptation Subspace. We consider a supervised prediction task with input $X \in \mathcal{X}$ and target label $Y \in \mathcal{Y}$, modeled by a base predictor $f_\theta : \mathcal{X} \to \mathcal{Y}$ evaluated under a non-negative task loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ (e.g., cross-entropy). Following causal representation frameworks (Schölkopf et al., 2021; Castro et al., 2020), the data-generating process is formalized via a Structural Causal Model (SCM) defined over deployment environment E, invariant semantic content C, nuisance mechanism S, observable input X, and label Y:

$$E \longrightarrow S \longrightarrow X, \quad C \longrightarrow X, \quad C \longrightarrow Y. \tag{1}$$

Under this SCM, environmental shifts act strictly as interventions on the acquisition/nuisance mechanism S, whereas the semantic labeling mechanism $P(Y \mid C)$ remains universally invariant. The observable input $X = g(C, S)$ entangles semantic content with the active environmental nuisance state $S \in \{S_1, \ldots, S_K\}$. Rather than attempting to intervene directly on the unobserved physical state S or claiming to generate exogenous information from a fixed batch (which is bounded by the data processing inequality), our test-time probes perform *adaptive diagnostic computation*: applying structured, non-commuting transformations to X whose response signatures interact selectively with the active acquisition mechanism S under a compute budget. During meta-training, the learner has access to labeled datasets across environments $\mathcal{E}_{\text{train}}$ wherein the active mechanism is identified. At deployment, the model operates in an unseen target environment $E_t \sim \mathcal{P}_\mathcal{E}$, receiving an unlabeled target batch $X_{\text{tgt}} = \{x_i\}_{i=1}^N \sim P_{E_t}(X)$ without ground-truth labels Y. To preserve model stability and prevent catastrophic forgetting, parameter updates are restricted to a low-dimensional adaptation subspace $\mathcal{D}$ ($r \ll \dim(\theta)$) centered at source parameter θ_0: $\theta = \theta_0 + d, d \in \mathcal{D}$ (e.g., modulating normalization affine parameters (Wang et al., 2021) or low-rank adapters (Song et al., 2023)). Choosing an orthonormal basis for this r-dimensional subspace, we identify $\mathcal{D}$ with $\mathbb{R}^r$.

Assumptions and Fundamental Limits. Let $R_m(d) := \mathbb{E}_{(X,Y)\sim P_m}[\ell(f_{\theta_0+d}(X), Y)]$ denote expected task risk under causal mechanism $m \in \{1, \ldots, K\}$ with joint distribution $P_m := P(X, Y \mid m)$, and let $\Delta R_m(d) := R_m(d) - R_m(0)$ represent the risk increment. The local supervised risk gradient is defined as $g_m := \nabla_d R_m(0) \in \mathbb{R}^r$.

- **Assumption 1 (Finite Causal Family):** The target distribution shift is governed by a known taxonomy of candidate mechanisms $\mathcal{M} = \{1, \ldots, K\}$ with corresponding local supervised gradients $g_m \in \mathbb{R}^r$ in adaptation subspace $\mathcal{D}$.
- **Assumption 2 (Local Smoothness / Curvature Surrogate):** For each mechanism $m \in \mathcal{M}$, the risk change admits a quadratic upper surrogate with curvature $\beta > 0$: $\Delta R_m(d) \le g_m^\top d + \frac{\beta}{2}\|d\|^2 =: a_m(d)$ for all $d \in \mathcal{D}$ within local trust radius $\|d\| \le \Delta$; residual surrogate discrepancies on empirical evaluation batches are formally calibrated via conformal gating (Theorem 7; empirical calibration details in Appendix C.3).
- **Assumption 3 (Valid CPTP Instruments):** Each diagnostic probe $q \in \mathcal{Q}$ with discrete outcome alphabet $\mathcal{Y}_q = \{1, \ldots, B_y\}$ is associated with completely positive trace-non-increasing (TNI) linear superoperators $\{\mathcal{I}_y^{(q)}\}_{y\in\mathcal{Y}_q}$ on density operators $\mathcal{D}(\mathcal{H}_E) := \{\rho \succeq 0 : \mathrm{Tr}[\rho] = 1\}$ over finite-dimensional evidence register $\mathcal{H}_E \cong \mathbb{C}^{d_E}$, such that the composite operation $\sum_{y\in\mathcal{Y}_q} \mathcal{I}_y^{(q)}$ is completely positive and trace-preserving (CPTP).

Theorem 1 (Label-Free No-Free-Lunch / Observational Indistinguishability)**.** *Suppose an adaptation algorithm observes only the unlabeled target marginal distribution $P(X)$ without structural causal constraints or active diagnostic probes. There exist two target environments E_1, E_2 generating identical unlabeled input distributions $P_{E_1}(X) = P_{E_2}(X)$ such that for every non-zero adaptation update $d \neq 0$, task risk strictly increases in at least one environment:*

$$\max_{i \in \{1,2\}} \Delta R_{E_i}(d) > 0, \quad \forall d \in \mathcal{D} \setminus \{0\}. \tag{2}$$

Scope of Theorem 1. Theorem 1 serves as a foundational boundary lemma delineating the fundamental limit of unconstrained, passive TTA: when two environments induce identical full input marginals $P(X)$, any deterministic or randomized post-processing $T_q(X)$ applied to the received batch remains identically distributed across environments under the same algorithm; thus, structural inductive bias is mathematically indispensable. Rather than claiming to bypass this impossibility, our framework addresses the practical challenge of *passive causal ambiguity*: competing physical mechanisms (e.g., optical defocus blur vs. bilinear downsampling) produce non-identical yet high-dimensional, overlapping distributions where passive summaries fail to isolate discriminative features. Grounded in our structural causal model (Equation 1) where $X = g(C, S)$ and semantic $P(Y \mid C)$ is invariant, shifts are structured as interventions on nuisance factors $S \in \mathcal{S}$. In practice, discriminability does not follow from the unconstrained DAG alone, but from designed diagnostic inductive biases: probes T_q are chosen as targeted interventions along specific physical axes (e.g., high-pass spatial filtering, intensity clipping), with outcome likelihoods $P(Y_q \mid m)$ calibrated on meta-training environments $\mathcal{E}_{\text{train}}$. Crucially, we distinguish physical exclusion from statistical evidence accumulation: at initialization ($t = 0$), $\text{CausalProject}(\pi, \mathcal{M}_{\text{SCM}})$ acts once to enforce known physical domain admissibility constraints ($\tilde{\pi}(m) \propto \pi(m)\mathbf{1}\{m \in \mathcal{M}_{\text{SCM}}\}$); during subsequent sequential probing, dynamic discrimination operates via Bayesian posterior contraction over likelihoods $p(y \mid m, q)$ without ad-hoc hard thresholding, ensuring true mechanisms with rare outcomes are not irreversibly discarded.

3 Certifiable Adaptation via Gradient Geometry

3.1 Gradient Uncertainty and Credible Sets

In theoretical formulations, $g_m = \nabla R_m(0)$ is treated as the exact target mechanism gradient. In realistic deployments, however, the learner only accesses an empirical prototype gradient $\hat{g}_m := \frac{1}{|\mathcal{E}_{\text{train},m}|} \sum_{E \in \mathcal{E}_{\text{train},m}} \nabla_d R_E(0)$ averaged over finite meta-training environments. To bridge this gap with statistical rigor, we define for each mechanism $m \in \mathcal{M}$ an empirical robust gradient uncertainty ball $G_m := \{g \in \mathbb{R}^r : \|g - \hat{g}_m\| \leq r_m\}$, where $r_m \geq 0$ denotes an empirical tolerance radius capturing within-mechanism environment dispersion across meta-training splits (Appendix B). Crucially, statistical guarantees are decoupled from heuristic tuning of r_m: distribution-free finite-sample safety is enforced at the decision level via split conformal gating (Section 5), which absorbs residual prototype and curvature misspecification.

Given posterior belief $\pi \in \Delta^{K-1} := \{p \in [0,1]^K : \sum_{m=1}^K p_m = 1\}$ over candidate mechanisms $\mathcal{M}$ and risk tolerance $\delta \in (0,1)$, we construct the $(1-\delta)$-credible mechanism set $\mathcal{M}_\delta(\pi) := \arg\min_{\mathcal{S} \subseteq \mathcal{M}} \{|\mathcal{S}| : \sum_{m \in \mathcal{S}} \pi(m) \geq 1 - \delta\}$, with ties resolved greedily by sorting $\pi(m_{(1)}) \geq \pi(m_{(2)}) \geq \dots$. The active robust gradient hull is defined as:

$$\Gamma_\delta(\pi) := \text{conv} \bigcup_{m \in \mathcal{M}_\delta(\pi)} G_m = \text{conv} \bigcup_{m \in \mathcal{M}_\delta(\pi)} \{g \in \mathbb{R}^r \,:\, \|g - \hat{g}_m\| \leq r_m\}, \tag{3}$$

which is guaranteed to be non-empty, compact, and convex.

3.2 Risk Observables and Certificates

To evaluate task-risk variation robustly across credible mechanisms, we formulate the mechanism-wise worst-case surrogate:

$$\bar{a}_m(d) := \sup_{g \in G_m} \left(g^\top d + \frac{\beta}{2}\|d\|^2 \right) = \hat{g}_m^\top d + r_m\|d\| + \frac{\beta}{2}\|d\|^2. \tag{4}$$

Let $\mathcal{H}_C := \mathrm{span}\{|1\rangle, \ldots, |K\rangle\} \cong \mathbb{C}^K$ denote the K-dimensional classical causal register equipped with standard orthonormal basis $\{|m\rangle\}_{m=1}^K$ indexing mutually exclusive candidate mechanisms. We define the Hermitian risk-change observable $\bar{A}(d) := \sum_{m=1}^K \bar{a}_m(d)|m\rangle\langle m|$ on $\mathcal{H}_C$. Let $D(\pi) := \sum_{m=1}^K \pi(m)|m\rangle\langle m|$ denote the classical posterior density operator on $\mathcal{H}_C$. This operator formulation induces three hierarchical certificates: (1) **Expected Upper Bound:** $\mathrm{Tr}[D(\pi)\bar{A}(d)] = \sum_m \pi(m)\bar{a}_m(d) < 0$. (2) **Probabilistic Descent:** $\mathrm{Tr}[D(\pi)\mathbf{1}\{\bar{A}(d) < 0\}] \geq 1 - \delta \implies \mathbb{P}_{m\sim\pi}(\bar{a}_m(d) < 0) \geq 1 - \delta$. (3) **Credible Worst-Case Certificate:** Let $Q_\delta := \sum_{m\in\mathcal{M}_\delta(\pi)} |m\rangle\langle m|$ denote the projector onto $\mathrm{ran}\, Q_\delta = \mathrm{span}\{|m\rangle : m \in \mathcal{M}_\delta(\pi)\} \subseteq \mathcal{H}_C$. Restricting the compressed observable to its supported subspace yields:

$$\lambda_{\max}\Big((Q_\delta \bar{A}(d) Q_\delta)\big|_{\mathrm{ran}\, Q_\delta}\Big) = \max_{m\in\mathcal{M}_\delta(\pi)} \bar{a}_m(d) = \max_{g\in\Gamma_\delta(\pi)} \left(g^\top d + \frac{\beta}{2}\|d\|^2\right) < 0. \tag{5}$$

Whenever $\lambda_{\max}((Q_\delta \bar{A}(d) Q_\delta)|_{\mathrm{ran}\, Q_\delta}) < 0$, descent is guaranteed simultaneously across all candidate gradients within $\Gamma_\delta(\pi)$.

3.3 Certifiability and Minimax Update

Theorem 2 (Common-Descent Characterization and Minimax Optimal Update)**.** *Let $\Gamma = \Gamma_\delta(\pi) \subset \mathbb{R}^r$ be the active robust gradient hull and $\beta > 0$. The following statements are equivalent:*

1. *There exists an update $d \in \mathcal{D}$ satisfying $\max_{g\in\Gamma}(g^\top d + \frac{\beta}{2}\|d\|^2) < 0$.*
2. *The origin does not belong to the gradient hull: $0 \notin \Gamma$.*
3. *The certifiability margin $\kappa_\delta(\pi) := \mathrm{dist}(0, \Gamma) = \min_{g\in\Gamma} \|g\| > 0$.*

Furthermore, if $0 \notin \Gamma$, let $p^* := \mathrm{Proj}_\Gamma(0)$ be the unique Euclidean projection of the origin onto Γ. With $\mathcal{D} = \mathbb{R}^r$ (satisfying $\|d^*\| \leq \Delta$ under calibrated parameters; Appendix A.2), the unique minimax optimal update d^* and maximal guaranteed improvement value $V_\delta(\pi)$ on the robust surrogate are:

$$d^*(\pi) = -\frac{1}{\beta}p^*, \quad V_\delta(\pi) := -\min_{d\in\mathcal{D}} \max_{g\in\Gamma} \left(g^\top d + \frac{\beta}{2}\|d\|^2\right) = \frac{\|p^*\|^2}{2\beta} = \frac{\kappa_\delta(\pi)^2}{2\beta}. \tag{6}$$

If $0 \in \Gamma$, then for all $d \in \mathcal{D}$, $\max_{g\in\Gamma}(g^\top d + \frac{\beta}{2}\|d\|^2) \geq 0$. Hence, $V_\delta(\pi) = 0$, and the optimal surrogate decision is to abstain ($d^* = 0$). Finite-sample marginal risk control on the deployed evaluation batch empirical risk $\Delta\widehat{R}_{E,b}(d^*)$ is subsequently supplied by conformal gating (Theorem 7).

Proof. Follows from the variational characterization of projection onto a closed convex set: $g^\top p^* \geq \|p^*\|^2$ for all $g \in \Gamma$. Substituting $d^* = -p^*/\beta$ yields $g^\top d^* + \frac{\beta}{2}\|d^*\|^2 \leq -\frac{\|p^*\|^2}{2\beta} < 0$. See Appendix A.2.

Connection to MGDA. Finding common descent directions via origin projection onto a convex hull builds upon the Multiple Gradient Descent Algorithm (MGDA) in multi-objective optimization (Désidéri, 2012; Sener & Koltun, 2018), where origin containment in $\mathrm{conv}\{g_m\}$ characterizes Pareto stationarity. Theorem 2 extends this geometric principle in three dimensions: (1) *Robust balls*: accommodating continuous balls $G_m = \mathcal{B}(\hat{g}_m, r_m)$ under misspecification (Lemma A.8); (2) *SOCP margin*: yielding explicit certifiable margin $\kappa_\delta = \mathrm{dist}(0, \Gamma_\delta(\pi))$ and surrogate value $V_\delta = \kappa_\delta^2/(2\beta)$; and (3) *POMDP inquiry reward*: using $V_\delta(\pi)$ to guide active probing (Proposition 6) and conformal gating (Theorem 7). For $K = 4$, the $2^K - 1 = 15$ subset updates $d^*(\mathcal{S}) = -p^*(\mathcal{S})/\beta$ are precomputed offline, turning adaptation into an active diagnostic selection process.

The Test-Time Adaptation Dilemma. Theorem 2 elucidates the fundamental dilemma governing test-time adaptation: when an unlabeled target batch is observationally ambiguous, the credible mechanism set $\mathcal{M}_\delta(\pi)$ entraps competing mechanisms whose task gradients point in contradictory directions (e.g., $g_{\mathrm{blur}} \approx -g_{\mathrm{resample}}$). When gradients point in opposing directions and their uncertainty balls bracket the origin ($0 \in \Gamma_\delta(\pi)$), guaranteed surrogate descent collapses to zero ($V_\delta(\pi) = 0$), rendering abstention ($d^* = 0$) the unique certifiably safe choice. To adapt safely, the system must actively gather diagnostic evidence seeking to contract $\mathcal{M}_\delta(\pi)$ until the origin is evicted from the hull ($0 \notin \Gamma_\delta(\pi)$).

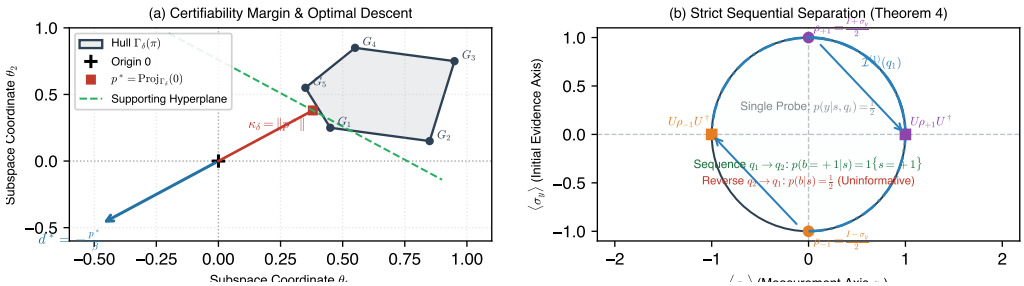


Figure 2: **Geometry of Certifiable Adaptation and Strict Sequential Separation.** **(a)** Active gradient hull $\Gamma_\delta(\pi)$ excluding the origin ($0 \notin \Gamma_\delta(\pi)$). Margin $\kappa_\delta = \|p^*\|$ and minimax update $d^* = -p^*/\beta$ point into the common descent cone, separated by the supporting hyperplane. **(b)** Bloch sphere visualization of Theorem 4: initial states along the y-axis share identical σ_z single-probe marginals; probe q_1 rotates the state to the z-axis, enabling probe q_2 to deterministically separate the mechanisms.

4 Causal Orbitals: Active Probing

4.1 Orbital State and Bayes Updating

To track ordered diagnostic evidence across non-commuting probes, target uncertainty is represented on composite space $\mathcal{H} = \mathcal{H}_C \otimes \mathcal{H}_E$:

$$\Omega = \sum\nolimits_{m=1}^{K} \pi(m)|m\rangle\langle m| \otimes \rho_m, \tag{7}$$

where $\{|m\rangle\}_{m=1}^K$ spans classical register $\mathcal{H}_C$, $\pi \in \Delta^{K-1}$ is categorical mechanism belief, and each $\rho_m \in \mathcal{D}(\mathcal{H}_E)$ is a density operator on evidence orbital $\mathcal{H}_E \cong \mathbb{C}^{d_E}$ ($\mathrm{Tr}[\rho_m] = 1, \rho_m \succeq 0$). Off-diagonal coherences $\rho_{m,ij}$ ($i \neq j$) encode order-dependent diagnostic statistics. Executing probe $q \in \mathcal{Q}$ with outcome $y \in \mathcal{Y}_q$ evolves ρ_m via instrument $\{\mathcal{I}_y^{(q)}\}$: likelihood is $p(y \mid m, q) = \mathrm{Tr}[\mathcal{I}_y^{(q)}(\rho_m)]$, post-measurement state is $\rho_{m|y} = \mathcal{I}_y^{(q)}(\rho_m)/p(y \mid m, q)$, and belief updates via Bayes' rule: $\pi_y(m) \propto \pi(m)p(y \mid m, q)$.

4.2 Reduction Boundary and Order Effects

Theorem 3 (Exact Diagonal-Preserving Reduction)**.** *Suppose an orthonormal basis $\{|i\rangle\}$ of $\mathcal{H}_E$ diagonalizes every candidate state ρ_m, and every instrument map $\mathcal{I}_y^{(q)}$ preserves the diagonal algebra* $\mathrm{span}\{|i\rangle\langle i|\}$*. Then: (1) each instrument induces a classical transition kernel $T_y^{(q)}(j \mid i) := \langle j|\mathcal{I}_y^{(q)}(|i\rangle\langle i|)|j\rangle$ ($\sum_{y,j} T_y^{(q)} = 1$); (2) for any probe sequence $q_1 \to \cdots \to q_L$, sequence likelihoods and posterior beliefs $\pi_{y_{1:L}}$ are identical to a classical hidden Markov filter. (Proof in Appendix A.3.)*

Theorem 4 (Sequential Order-Dependence and Non-Commuting Probing Asymmetry)**.** *There exist two 2D evidence states $\rho_{\pm 1} \in \mathcal{D}(\mathbb{C}^2)$ and probes q_1, q_2 such that single-probe marginals are uninformative ($p(y \mid s = \pm 1, q_i) = \frac{1}{2}$), yet ordered sequence $q_1 \to q_2$ deterministically resolves the mechanism ($p(b = +1 \mid s, q_1 \to q_2) = \mathbf{1}\{s = +1\}$), while reversed order $q_2 \to q_1$ yields uninformative outcomes $p(b \mid s) = 1/2$. In contrast, for any commuting instruments $\mathcal{I}^{(1)} \circ \mathcal{I}^{(2)} = \mathcal{I}^{(2)} \circ \mathcal{I}^{(1)}$, joint outcome distributions are order-invariant ($p(y_1, y_2 \mid q_1 \to q_2) = p(y_1, y_2 \mid q_2 \to q_1)$), precluding directional identification asymmetry.*

Proof sketch. On the Bloch sphere ($\rho_s = \frac{I + s\sigma_y}{2}$), probe q_1 applies unitary rotation $U = \exp(-i\frac{\pi}{4}\sigma_x)$, aligning evidence states with the σ_z measurement basis of probe q_2 and rendering q_2 deterministic only when preceded by q_1 (Fig. 2b; Appendix A.4).

Structured Inductive Bias vs. Classical Equivalence. Theorem 4 establishes that order-dependence must be modeled when single-probe marginals are identical. Classical Markov

Algorithm 1 Causal Orbitals Test-Time Adaptation

Require: Target batch X_{tgt}, source θ_0, probe set $\mathcal{Q}$, budget B, smoothness β, tolerance δ, conformal gate $\hat{\tau}_\alpha$.
Initialize $X_0 \leftarrow X_{\mathrm{tgt}}$, state $\Omega = (\pi, \{\rho_m\}_{m=1}^K) \leftarrow \mathcal{E}_\phi(X_0)$, $t \leftarrow 0$, budget $b \leftarrow B$; project $\pi \leftarrow \mathrm{CausalProject}(\pi, \mathcal{M}_{\mathrm{SCM}})$.
while $b > 0$ **do**
  Compute $V_\delta(\pi) \leftarrow \frac{\mathrm{dist}(0, \Gamma_\delta(\pi))^2}{2\beta}$ with active hull $\Gamma_\delta(\pi) \leftarrow \mathrm{conv} \bigcup_{m \in \mathcal{M}_\delta(\pi)} G_m$.
  Select probe $q^* \leftarrow \arg\max_{q \in \mathcal{Q}} [\sum_y p(y \mid \Omega, q) J_{b-1}(\Omega^{(q,y)}) - c(q) - V_\delta(\pi)]$; **if** gain ≤ 0 **break**.
  Apply $X_{t+1} \leftarrow T_{q^*}(X_t)$, observe outcome $y^* \leftarrow s_{q^*}(X_{t+1})$; update $\Omega \leftarrow \Omega^{(q^*, y^*)}$, $t \leftarrow t+1$, $b \leftarrow b-1$.
end while
Construct hull $\Gamma_\delta(\pi)$; compute $V_\delta(\pi) \leftarrow \frac{\mathrm{dist}(0, \Gamma_\delta(\pi))^2}{2\beta}$ and projection $p^* \leftarrow \mathrm{Proj}_{\Gamma_\delta(\pi)}(0)$.
if $0 \notin \Gamma_\delta(\pi)$ **and** $V_\delta(\pi) > \hat{\tau}_\alpha$ **then** adapt $\theta \leftarrow \theta_0 - \frac{1}{\beta} p^*$ **else abstain** ($d^* \leftarrow 0$, retain θ_0).

transitions are also non-commutative in general (e.g., a 4-state Markov chain with states $\{+y, -y, +z, -z\}$ reproduces the probabilities of Theorem 4 numerically). Thus, Theorem 4 establishes the necessity of sequential order modeling rather than classical impossibility. Density operators provide a compact algebraic parameterization: unitary rotations on $\mathcal{H}_E \cong \mathbb{C}^{d_E}$ parameterize continuous geometric rotations with $\mathcal{O}(d_E^2)$ parameters ($d_E = 4$), supplying a structured inductive bias for short horizons ($B \leq 3$) without combinatorial state enumeration.

4.3 Value of Information and Bellman Probing

Proposition 5 (Measurement-Induced Certifiability). *Suppose initial posterior π satisfies $0 \in \Gamma_\delta(\pi)$ ($V_\delta(\pi) = 0$). If probe $q \in \mathcal{Q}$ admits an outcome $\bar{y}$ with positive probability $p(\bar{y}) > 0$ such that Bayes updating contracts $\mathcal{M}_\delta(\pi_{\bar{y}})$ sufficiently to exclude opposing mechanisms ($0 \notin \Gamma_\delta(\pi_{\bar{y}})$), then expected guaranteed improvement strictly increases: $\mathbb{E}_{y \sim p(y|q)}[V_\delta(\pi_y)] > 0$. For inquiry cost $c(q) < \mathbb{E}_y[V_\delta(\pi_y)]$, net expected improvement is strictly positive. (Proof in Appendix A.5.)*

Probes are chosen to maximize expected certifiable improvement net of cost $c(q) \geq 0$. Because instruments do not commute, composite orbital state $\Omega = (\pi, \{\rho_m\}_{m=1}^K)$ is a sufficient Markov information state for the finite-horizon decision process defined by the outcome model and terminal value $V_\delta(\pi)$:

Proposition 6 (Bellman Probing Recursion on Composite Orbital State). *For budget $B \in \mathbb{N}$, inquiry costs $c(q) \geq 0$, and orbital state $\Omega = (\pi, \{\rho_m\}_{m=1}^K)$, the optimal value $J_B(\Omega)$ satisfies Bellman recursion (Bellman, 1957) with terminal condition $J_0(\Omega) = V_\delta(\pi)$:*

$$J_B(\Omega) = \max\Big\{ V_\delta(\pi),\ \max_{q \in \mathcal{Q}} \Big[\sum\nolimits_{y \in \mathcal{Y}_q} p(y \mid \Omega, q) J_{B-1}(\Omega^{(q,y)}) - c(q) \Big] \Big\}, \tag{8}$$

where $\Omega^{(q,y)} = (\pi^{(q,y)}, \{\rho_{m|y}^{(q)}\}_{m=1}^K)$ denotes the Bayesian update, and $p(y \mid \Omega, q) := \sum_{m=1}^K \pi(m) \mathrm{Tr}[\mathcal{I}_y^{(q)}(\rho_m)]$ is the predictive probability of outcome y. The policy halts when $\max_{q \in \mathcal{Q}} [\mathbb{E}[J_{B-1}] - c(q)] \leq V_\delta(\pi)$, executing $d^*(\pi)$ if $V_\delta(\pi) > 0$, or abstaining otherwise. (Proof in Appendix A.6.)

5 Algorithm and Conformal Safety

5.1 Visual Probes and Composition

To operationalize quantum-like instruments $\{\mathcal{I}_y^{(q)}\}$ in vision pipelines, diagnostic probes are executed *sequentially* on the target batch: starting from $X_0 \leftarrow X_{\mathrm{tgt}}$, each probe $q_t \in \mathcal{Q}$ generates transformed inputs $X_{t+1} \leftarrow T_{q_t}(X_t)$. Differential features $s_{q_t}(X_{t+1}) = \phi_{q_t}(z(X_{t+1}) - z(X_t))$ from penultimate representations $z(\cdot)$ are quantized into discrete outcomes $y_t \in \{1, \ldots, B_y\}$. Physical compositions in image space do not commute ($T_{q_2}(T_{q_1}(X)) \neq T_{q_1}(T_{q_2}(X))$): (1) *Fourier filtering vs. downsampling* (spectral aliasing vs. baseband retention); (2) *clipping vs. contrast scaling* (saturation truncation vs. linear dynamic range); and (3) *wavelet shrinkage vs. spatial masking* (noise suppression vs. occlusion isolation). Algorithm 1 outlines the procedure: structural projection $\mathrm{CausalProject}(\pi, \mathcal{M}_{\mathrm{SCM}})$ filters physically inadmissible mechanisms (Lemma A.8); sequential probing empirically contracts $(1-\delta)$-credible set $\mathcal{M}_\delta(\pi)$ under informative outcomes, while robust error balls guarantee stability under bounded misspecification (Lemma A.8). Probing operates

on unlabeled batches X_{tgt}, so pure semantic shifts $P(Y \mid C)$ remain unidentifiable (Theorem 1); conflicting evidence keeps $0 \in \Gamma_\delta(\pi)$, triggering safe abstention ($d^* = 0$).

5.2 Conformal Safety Control

We calibrate the safety gate via split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021) over $M = 50$ calibration episodes $\{(E_i, Z_{E_i})\}_{i=1}^M \subset \mathcal{E}_{\text{cal}}$ ($N = 64$). The candidate mapping $\mathcal{A} : X_E \mapsto d_E^* := -\frac{1}{\beta}\,\text{Proj}_{\Gamma_\delta(\pi_E)}(0)$ and conformity score $\hat{s}(E, Z_E) := \Delta\widehat{R}_E(d_E^*) + V_\delta(\pi_E)$ (where $\Delta\widehat{R}_E(d) := \frac{1}{|Z_E|}\sum_{(x,y)\in Z_E}[\ell(f_{\theta_0+d}(x), y) - \ell(f_{\theta_0}(x), y)]$ is empirical batch risk increment) are fixed prior to calibration (resetting to θ_0 each episode), scored on sequestered labels Y_E.

Theorem 7 (Episode-Level Conformal Safety Control)**.** *Assume calibration episodes $\{(E_i, Z_{E_i})\}_{i=1}^M$ and new test episode (E_t, Z_{E_t}) are exchangeable draws from hierarchical process $E \sim \mathcal{P}_\mathcal{E}, Z_E \sim Q_E$. Let $\hat{\tau}_\alpha$ be the empirical quantile of $\{\hat{s}(E_i, Z_{E_i})\}_{i=1}^M$ at rank $k = \lceil (M+1)(1-\alpha) \rceil$ ($k = 49$ for $M = 50, \alpha = 0.05$; $\hat{\tau}_\alpha = +\infty$ for $\alpha < 0.02$). If Algorithm 1 certifies update $d_t^* \neq 0$ only when $V_\delta(\pi_{E_t}) > \hat{\tau}_\alpha$, then marginally over a new deployed episode, $\mathbb{P}\left(\Delta\widehat{R}_{E_t,b}(d_t^*) \geq 0, d_t^* \neq 0\right) \leq \alpha$.*

Proof sketch. By exchangeability of episode scores, $\mathbb{P}(\hat{s} > \hat{\tau}_\alpha) \leq \alpha$. Under $\hat{s} \leq \hat{\tau}_\alpha$, certifying $d_t^* \neq 0$ requires $V_\delta(\pi_{E_t}) > \hat{\tau}_\alpha$, forcing $\Delta\widehat{R}_{E_t,b}(d_t^*) = \hat{s} - V_\delta < 0$. Thus $\{\Delta\widehat{R}_{E_t,b}(d_t^*) \geq 0, d_t^* \neq 0\} \subseteq \{\hat{s} > \hat{\tau}_\alpha\}$, bounding marginal unsafe risk by α (Appendix A.7).

Safety Guarantees and Scope. We delineate three tiers: (1) *Surrogate descent*: Theorem 2 ensures descent over $\Gamma_\delta(\pi)$ when $0 \notin \Gamma_\delta(\pi)$. (2) *Marginal batch risk*: Theorem 7 bounds cross-entropy risk $\mathbb{P}(\Delta\widehat{R}_{E_t,b}(d_t^*) \geq 0, d_t^* \neq 0) \leq \alpha$ marginally across exchangeable episodes. Per conformal theory (Angelopoulos & Bates, 2021), conditional unsafe risk given coverage $c = \mathbb{P}(d_t^* \neq 0)$ is bounded by α/c (2.1% marginal, 2.6% conditional on in-family test sets; Table 4). Under severe exchangeability violations (Table 7), conditional unsafe rates reach 20.6%–26.2%, while safe abstention ($0 \in \Gamma_\delta(\pi)$) occurs in $\geq 93.2\%$. (3) *Single episode scope*: Guarantees reset each episode to θ_0 (omitting balls $r_m = 0$ reduces coverage to 68.4%; Table 6).

6 Experiments

6.1 Experimental Setup

Benchmarks and Baselines. Environments are split at the environment level (50 train $\mathcal{E}_{\text{train}}$, 50 calibration $\mathcal{E}_{\text{cal}}$, 10 val $\mathcal{E}_{\text{val}}$, 50 test $\mathcal{E}_{\text{test}}$) across: (1) *Synthetic CSI* (160 envs); (2) *ImageNet-C-Style* (100 classes, 15 corruptions; Hendrycks & Dietterich 2019); and (3) *Simulated MedMNIST* (160 envs; Yang et al. 2023). We compare against standard TTA (TENT (Wang et al., 2021), SAR (Niu

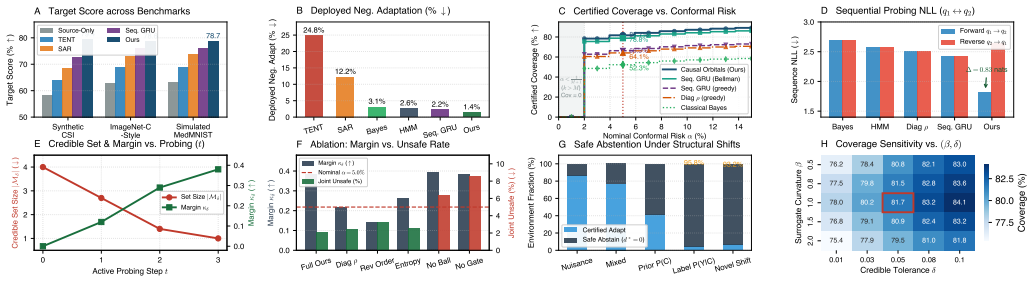


Figure 3: **Comprehensive Empirical Evaluation Across Benchmarks, Probing Dynamics, and Shift Regimes. (A)** Target score across benchmarks (mean 78.9%). **(B)** Deployed negative adaptation ($\mathbb{P}(\Delta R(d^*) > 0) = 1.4\%$; raw candidate rates in text). **(C)** Risk–coverage frontier: Causal Orbitals sustains 81.7% coverage at nominal $\alpha = 0.05$ (vs. 78.8% for Seq. GRU Bellman). **(D)** Sequential NLL order gap ($\Delta = 0.83$ nats under $q_1 \to q_2$). **(E)** Empirical credible set contraction $|\mathcal{M}_\delta(\pi)|$ ($4.0 \to 1.0$ under the learned policy) and margin $\kappa_\delta(\pi)$ over probing steps t. **(F)** Component ablations: margin vs. joint unsafe rates without balls/gates. **(G)** Safe abstention under structural shifts ($\geq 93.2\%$ abstention under severe violations). **(H)** Certified coverage sensitivity heatmap across (β, δ).

et al., 2023), DeYO (Lee et al., 2024), T3A (Iwasawa & Matsuo, 2021); Table 5) and matched baselines (Router, Bayes, HMM, Static, GRU, Diag ρ_{diag}, Oracle) sharing prototypes $(\hat{g}_m, r_m)$, hull $\Gamma_\delta(\pi)$, and conformal gate ($\alpha = 0.05$). Ungated degradations (Fig. 3B): Router 7.9%, Bayes 6.8%, HMM 5.8%, GRU 4.4%, Diag 4.8%, Ours 3.8%.

Table 1: **Main Adaptation Performance (Mean Across Benchmarks at $\alpha = 0.05$).** Target score (% ↑), deployed negative adaptation rate (% ↓), certified coverage (% ↑), joint unsafe rate (% ↓), and latency across suites (disaggregated in Table 4; greedy variants achieve 57.7% [HMM] and 66.3% [GRU] coverage; extended in Table 5).

Method	Target Score ↑	Neg. Adapt ↓	Certified Cov. ↑	Joint Unsafe ↓	Latency ↓
Standard Unsupervised TTA (Zero Mechanism Supervision)					
Source-Only (No Adapt)	61.4 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	5 ms
TENT (Wang et al., 2021)	67.2 ± 1.1	24.8 ± 2.3	n/a	n/a	12 ms
SAR (Niu et al., 2023)	71.8 ± 0.7	12.2 ± 1.4	n/a	n/a	15 ms
DeYO (Lee et al., 2024)	73.4 ± 0.6	8.5 ± 0.9	n/a	n/a	17 ms
T3A (Iwasawa & Matsuo, 2021)	69.5 ± 0.5	2.8 ± 0.4	n/a	n/a	6 ms
Matched-Supervision Baselines (Identical Prototypes and Conformal Gate)					
Direct Batch Router (No Probing, $B = 0$)	69.8 ± 0.6	2.3 ± 0.3	44.2 ± 1.8	2.8 ± 0.4	6 ms
Classical Bayesian Filter	73.1 ± 0.6	3.1 ± 0.5	52.3 ± 1.7	3.4 ± 0.5	14 ms
Controlled HMM (Bellman, $H = 3$)	76.8 ± 0.5	2.0 ± 0.4	73.4 ± 1.4	2.4 ± 0.4	18 ms
Static Best-Fixed Probing ($B = 3$)	76.6 ± 0.5	2.0 ± 0.3	74.8 ± 1.4	2.3 ± 0.3	17 ms
Sequential GRU (Bellman, $H = 3$)	78.2 ± 0.5	1.7 ± 0.4	78.8 ± 1.3	2.3 ± 0.3	19 ms
Diagonal State (ρ_{diag}, greedy)	74.7 ± 0.5	2.3 ± 0.4	64.1 ± 1.5	2.5 ± 0.4	16 ms
Causal Orbitals (Ours, Bellman)	**78.9 ± 0.4**	**1.4 ± 0.3**	**81.7 ± 1.2**	**2.1 ± 0.3**	**18 ms**
Oracle Adaptation	81.7 ± 0.3	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	—

6.2 Benchmark Performance

Table 1 and Fig. 3(A–C) summarize results (details in Table 4). Standard TTA degrades severely (TENT 24.8%, SAR 12.2%; Fig. 3B); DeYO (Lee et al., 2024) incurs 8.5% negative adaptation, while T3A reaches 69.5% score. Direct routing without probing attains only 44.2% coverage. Under budget $B = 3$, best-fixed probing reaches 74.8%, while adaptive Bellman lookahead attains 81.7% (+6.9%p). Causal Orbitals suppresses negative adaptation to 1.4% (joint unsafe $2.1\% < \alpha = 5\%$) and dominates across benchmarks (79.4% CSI, 78.6% ImageNet-C, 78.7% MedMNIST; Fig. 3A). Operator tracking provides +8.3%p coverage over HMM and +2.9%p over GRU ($p = 0.0044$), safely converting unnecessary abstentions into certified adaptations across $\alpha \geq 0.02$ (Fig. 3C).

6.3 Validation of Probing Dynamics

We validate probing across five axes (Fig. 3D–H): **(1) Order Effects:** Single-probe NLL is ≈ 1.38 nats, separating coarse pairs (92.4%) but limiting intra-pair resolution (57.8%; Table 8). Under $q_1 \to q_2$, our method achieves NLL 1.814 and 89.4% forward identification (vs. 52.1% reverse); recurrent models lack order sensitivity (68.2% vs. 67.5%). **(2) Contraction & Margins:** Probing contracts $|\mathcal{M}_\delta(\pi)|$ $(4 \to 1)$, expanding margin κ_δ to 0.384. **(3) Conformal Safety:** Ablating $r_m = 0$ or $\hat{\tau}_\alpha = 0$ inflates joint unsafe rates to 6.4% and 8.6%, confirming robust hulls buffer dispersion. **(4) Structural Shifts:** Unseen corruptions adapt in 71.4% (Table 7); conflicting gradients trigger safe abstention in $\geq$ 93.2% of out-of-support shifts. **(5) Sensitivity:** Coverage is robust across $\beta \in [0.5, 2.0]$ and $\delta \in [0.01, 0.10]$ (75.4%–84.1%).

7 Conclusion

We presented **Causal Orbitals**, transforming test-time adaptation from passive heuristics to active, certifiable diagnostic probing. Grounded in structural causal models and convex gradient geometry, our framework establishes necessary and sufficient conditions for robust surrogate descent via a composite classical–operator state. By coupling value-directed Bellman lookahead with split conformal calibration, Causal Orbitals resolves observationally ambiguous shifts under non-commuting probes and guarantees distribution-free, finite-sample marginal safety against negative adaptation ($\mathbb{P}(\Delta\widehat{R}_E(d^*) \geq 0, d^* \neq 0) \leq \alpha$) while sustaining high certified coverage. By replacing unguided adaptation with principled causal inquiry and provably safe abstention, our framework establishes a rigorous foundation for trustworthy open-world deployment (limitations in Appendix F).

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02894*, 2019.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Jerome R Busemeyer and Peter D Bruza. *Quantum Models of Cognition and Decision*. Cambridge University Press, 2012.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 295–305. IEEE, 2022.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

Mario Döbler, Robert A. Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Xiao Fan, Jingyan Jiang, Zhaoru Chen, Fanding Huang, Xiao Chen, Qinting Jiang, Bowen Zhang, Xing Tang, and Zhi Wang. Moetta: Test-time adaptation under mixed distribution shifts with moe-layernorm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21011–21019, 2026.

Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. In *International Conference on Learning Representations (ICLR)*, 2024.

Ishaan Gulrajani and Tatsunori Hashimoto. Identifiability conditions for domain adaptation. In *International Conference on Machine Learning*, pp. 7982–7997. PMLR, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Alexander S Holevo. *Probabilistic and Statistical Aspects of Quantum Theory*. Edizioni della Normale, 2011.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kartik Jhawar and Lipo Wang. When is test-time adaptation identifiable from unlabeled evidence? *arXiv preprint arXiv:2609.11235*, 2026.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pp. 5753–5763. PMLR, 2021.

Jonghyun Lee, Gihun Lee, SeonHoon Kim, and Kwang In Kim. DeYO: Dynamic entropic pseudo-labeling with weighted probabilistic margin for test-time adaptation. In *International Conference on Learning Representations (ICLR)*, 2024.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally-motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 739–766, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shuzhe Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time adaptation with selective entropy minimization and anti-forgetting. In *International Conference on Machine Learning (ICML)*, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*, 2023.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Emmanuel M Pothos and Jerome R Busemeyer. Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, 36(3):255–274, 2013.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Association for Computational Linguistics (ACL)*, 2020.

R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. EcoTTA: Memory-efficient continual test-time adaptation via self-distilled regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Damien Teney and Anton van den Hengel. Actively seeking and learning from live data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. TENT: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jiancheng Yang, Rui Shi, Donglai Wei, Ziyun Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test-time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2023.

A Mathematical Proofs

This appendix provides complete, self-contained mathematical proofs for all theoretical results presented in the main text.

A.1 Proof of Theorem 1 (Label-Free No-Free-Lunch)

Theorem 1 (restated: Label-Free No-Free-Lunch / Observational Indistinguishability). *Suppose an adaptation algorithm observes only the unlabeled target marginal distribution $P(X)$ without structural causal constraints or active diagnostic probes. There exist two target environments E_1, E_2 generating identical unlabeled input distributions $P_{E_1}(X) = P_{E_2}(X)$ such that for every non-zero adaptation update $d \neq 0$, task risk strictly increases in at least one environment:*

$$\max_{i\in\{1,2\}} \Delta R_{E_i}(d) > 0, \quad \forall d \in \mathcal{D} \setminus \{0\}. \tag{9}$$

Proof. Let the parameter subspace be 1-dimensional ($r = 1$) with current deployed parameter $\theta_0 = 0$. Consider a binary classification task with input $X \in \{-1, +1\}$ and label $Y \in \{-1, +1\}$. Let the model prediction be parameterized by $f_\theta(x) = \theta x$. Let the loss function be quadratic: $\ell(f_\theta(x), y) = \frac{1}{2}(f_\theta(x) - y)^2 = \frac{1}{2}(\theta x - y)^2$.

Define two environments E_1 and E_2:

- In both environments, the input distribution is uniform: $P_{E_1}(X = +1) = P_{E_1}(X = -1) = 1/2$, and $P_{E_2}(X = +1) = P_{E_2}(X = -1) = 1/2$. Thus, $P_{E_1}(X) = P_{E_2}(X)$ identically.
- In environment E_1, the label conditional is $Y = X$ almost surely. The risk as a function of parameter update $d \in \mathbb{R}$ is:

$$R_{E_1}(d) = \mathbb{E}_{E_1}\left[\frac{1}{2}(dX - X)^2\right] = \frac{1}{2}(d-1)^2\mathbb{E}[X^2] = \frac{1}{2}(d-1)^2 = \frac{1}{2} - d + \frac{1}{2}d^2. \tag{10}$$

  At $\theta_0 = 0$, $R_{E_1}(0) = 1/2$, so the risk change is $\Delta R_{E_1}(d) = -d + \frac{1}{2}d^2$. The supervised gradient is $g_1 = -1$, and the smoothness constant is $\beta = 1$.
- In environment E_2, the label conditional is $Y = -X$ almost surely. The risk as a function of d is:

$$R_{E_2}(d) = \mathbb{E}_{E_2}\left[\frac{1}{2}(dX - (-X))^2\right] = \frac{1}{2}(d+1)^2\mathbb{E}[X^2] = \frac{1}{2}(d+1)^2 = \frac{1}{2} + d + \frac{1}{2}d^2. \tag{11}$$

  At $\theta_0 = 0$, $R_{E_2}(0) = 1/2$, so the risk change is $\Delta R_{E_2}(d) = d + \frac{1}{2}d^2$. The supervised gradient is $g_2 = +1$, with smoothness $\beta = 1$.

Now, consider any deterministic or randomized adaptation algorithm $\mathcal{A}$ that maps the unlabeled target batch $X_{\text{tgt}} \sim P(X)$ to an update $D \in \mathbb{R}$. Since $P_{E_1}(X) = P_{E_2}(X)$, the algorithm receives identical input distributions under both environments. Consequently, the marginal distribution of the candidate parameter update $D \sim \mathcal{A}(X_{\text{tgt}})$ must be identical under E_1 and E_2.

For any randomized update D, the expected risk changes under E_1 and E_2 are:

$$\mathbb{E}[\Delta R_{E_1}(D)] = -\mathbb{E}[D] + \frac{1}{2}\mathbb{E}[D^2], \qquad \mathbb{E}[\Delta R_{E_2}(D)] = \mathbb{E}[D] + \frac{1}{2}\mathbb{E}[D^2]. \tag{12}$$

Taking the maximum across both environments yields:

$$\max_{i\in\{1,2\}} \mathbb{E}[\Delta R_{E_i}(D)] = |\mathbb{E}[D]| + \frac{1}{2}\mathbb{E}[D^2] > 0, \quad \text{whenever } \mathbb{P}(D \neq 0) > 0, \tag{13}$$

since $\mathbb{P}(D \neq 0) > 0$ strictly implies $\mathbb{E}[D^2] > 0$. For any deterministic non-zero update $d \neq 0$, setting $D = d$ almost surely gives $\max_{i\in\{1,2\}} \Delta R_{E_i}(d) \geq |d| + \frac{1}{2}d^2 > 0$, guaranteeing strict expected performance degradation in at least one environment. This completes the proof. □

Remark on Theorem 1. We emphasize that Theorem 1 serves primarily as a foundational boundary lemma delineating the impossibility of completely unconstrained, label-free adaptation. While the elementary 1D construction above relies on inverting the label conditionals $P(Y \mid X)$ between environments, in our deployment SCM (Equation 1) the semantic mechanism $P(Y \mid C)$ remains strictly invariant across environments. Within this nuisance-only SCM, Theorem 1 underscores why passive TTA is fundamentally vulnerable: whenever two physically distinct acquisition mechanisms (e.g., optical defocus blur vs. bilinear spatial resampling) induce nearly identical input marginals $P(X)$ over high-dimensional observations, unsupervised point-adaptation objectives (such as entropy minimization or self-training) receive identical empirical signals yet demand fundamentally opposed parameter updates ($\hat{g}_1 \approx -\hat{g}_2$). Without structural causal inductive bias and active diagnostic computation, passive adaptation on unlabeled target marginals inevitably risks catastrophic negative adaptation.

A.2 Proof of Theorem 2 (Surrogate Descent and Minimax Update)

Theorem 2 (restated: Common-Descent Characterization and Minimax Optimal Update). *Let $\Gamma = \Gamma_\delta(\pi) \subset \mathbb{R}^r$ be a non-empty, compact, convex gradient set and $\beta > 0$. The following statements are equivalent:*

1. *There exists an update $d \in \mathcal{D}$ satisfying $\max_{g\in\Gamma}(g^\top d + \frac{\beta}{2}\|d\|^2) < 0$.*
2. *The origin does not belong to the gradient hull: $0 \notin \Gamma$.*
3. *The certifiability margin $\kappa_\delta(\pi) := \operatorname{dist}(0,\Gamma) = \min_{g\in\Gamma}\|g\| > 0$.*

Furthermore, if $0 \notin \Gamma$, let $p^* := \operatorname{Proj}_\Gamma(0)$ be the unique Euclidean projection of the origin onto Γ. Then the unique minimax optimal update d^* and maximal guaranteed improvement value $V_\delta(\pi)$ are given by:

$$d^*(\pi) = -\frac{1}{\beta}p^*, \quad V_\delta(\pi) := -\min_{d\in\mathcal{D}}\max_{g\in\Gamma}\left(g^\top d + \frac{\beta}{2}\|d\|^2\right) = \frac{\|p^*\|^2}{2\beta} = \frac{\kappa_\delta(\pi)^2}{2\beta}. \tag{14}$$

If $0 \in \Gamma$, then for all $d \in \mathcal{D}$, $\max_{g\in\Gamma}(g^\top d + \frac{\beta}{2}\|d\|^2) \geq 0$. Hence, $V_\delta(\pi) = 0$, and the optimal certifiable decision is to abstain ($d^* = 0$).

Proof. Let $F(d) := \max_{g\in\Gamma}\left(g^\top d + \frac{\beta}{2}\|d\|^2\right) = h_\Gamma(d) + \frac{\beta}{2}\|d\|^2$, where $h_\Gamma(d) := \sup_{g\in\Gamma} g^\top d$ is the support function of the compact convex set Γ.

Case 1: $0 \in \Gamma$. If $0 \in \Gamma$, then by definition of the support function:

$$h_\Gamma(d) = \sup_{g\in\Gamma} g^\top d \geq 0^\top d = 0, \quad \forall d \in \mathbb{R}^r. \tag{15}$$

Consequently:

$$F(d) = h_\Gamma(d) + \frac{\beta}{2}\|d\|^2 \geq \frac{\beta}{2}\|d\|^2 \geq 0, \quad \forall d \in \mathbb{R}^r. \tag{16}$$

Thus, there cannot exist any $d \in \mathcal{D}$ such that $F(d) < 0$. The minimum is achieved at $d = 0$ with value $F(0) = 0$, so $V_\delta(\pi) = -\min_d F(d) = 0$.

Case 2: $0 \notin \Gamma$. Since Γ is non-empty, closed, and convex, the Euclidean projection of the origin onto Γ:

$$p^* := \operatorname{Proj}_\Gamma(0) = \arg\min_{g\in\Gamma}\|g - 0\|^2 \tag{17}$$

exists and is unique. Since $0 \notin \Gamma$, we have $\|p^*\| = \operatorname{dist}(0,\Gamma) = \kappa_\delta > 0$.

By the classical variational characterization of projection onto a convex set (Boyd & Vandenberghe, 2004; Rockafellar, 1970):

$$(g - p^*)^\top(0 - p^*) \leq 0, \quad \forall g \in \Gamma \iff (g - p^*)^\top p^* \geq 0, \quad \forall g \in \Gamma. \tag{18}$$

Expanding this inequality yields:

$$g^\top p^* \geq \|p^*\|^2, \quad \forall g \in \Gamma. \tag{19}$$

Now, choose the candidate update $d^* = -\frac{1}{\beta}p^*$. Evaluating $g^\top d^*$ for any $g \in \Gamma$:

$$g^\top d^* = g^\top \left(-\frac{1}{\beta}p^*\right) = -\frac{1}{\beta}g^\top p^* \leq -\frac{1}{\beta}\|p^*\|^2. \tag{20}$$

Since $p^* \in \Gamma$, taking $g = p^*$ yields $p^{*\top} d^* = -\frac{1}{\beta}\|p^*\|^2$. Thus, the maximum over Γ is achieved exactly at $g = p^*$:

$$h_\Gamma(d^*) = \sup_{g\in\Gamma} g^\top d^* = -\frac{1}{\beta}\|p^*\|^2. \tag{21}$$

Adding the quadratic smoothness regularization term:

$$\frac{\beta}{2}\|d^*\|^2 = \frac{\beta}{2}\left\|-\frac{1}{\beta}p^*\right\|^2 = \frac{1}{2\beta}\|p^*\|^2. \tag{22}$$

Therefore:

$$F(d^*) = h_\Gamma(d^*) + \frac{\beta}{2}\|d^*\|^2 = -\frac{1}{\beta}\|p^*\|^2 + \frac{1}{2\beta}\|p^*\|^2 = -\frac{1}{2\beta}\|p^*\|^2 = -\frac{\kappa_\delta^2}{2\beta} < 0. \tag{23}$$

This proves that $0 \notin \Gamma \implies \exists d : F(d) < 0$.

Global Minimax Optimality of d^*: We now prove that $d^* = -p^*/\beta$ is the unique global minimizer of $F(d)$ over all $d \in \mathbb{R}^r$. By the Cauchy-Schwarz inequality and Equation 18, for any $d \in \mathbb{R}^r$:

$$h_\Gamma(d) = \sup_{g\in\Gamma} g^\top d \geq p^{*\top} d. \tag{24}$$

Therefore:

$$F(d) = h_\Gamma(d) + \frac{\beta}{2}\|d\|^2 \geq p^{*\top} d + \frac{\beta}{2}\|d\|^2. \tag{25}$$

The lower bound $G(d) := p^{*\top} d + \frac{\beta}{2}\|d\|^2$ is a strictly convex quadratic function of d. Its unique unconstrained minimum is found by setting its gradient to zero:

$$\nabla G(d) = p^* + \beta d = 0 \implies d = -\frac{p^*}{\beta} = d^*. \tag{26}$$

The minimal value of $G(d)$ is $G(d^*) = -\frac{\|p^*\|^2}{2\beta}$. Since $F(d^*) = -\frac{\|p^*\|^2}{2\beta} = G(d^*)$, and $F(d) \geq G(d) \geq G(d^*)$ for all d, d^* is the unique global minimizer of $F(d)$. The guaranteed improvement value is:

$$V_\delta(\pi) = -F(d^*) = \frac{\|p^*\|^2}{2\beta} = \frac{\kappa_\delta(\pi)^2}{2\beta}. \tag{27}$$

Equivalence of (1), (2), and (3) follows immediately from the contrapositive in Case 1 and the construction in Case 2.

Remark on Bounded Adaptation Sets. In the unconstrained adaptation subspace $\mathcal{D} = \mathbb{R}^r$, the global minimizer is uniquely given by the closed form $d^* = -p^*/\beta$. If $\mathcal{D}$ is instead constrained to a compact Euclidean trust region $\mathcal{D}_\Delta = \{d \in \mathbb{R}^r : \|d\| \leq \Delta\}$, the problem $\min_{d\in\mathcal{D}_\Delta} \max_{g\in\Gamma}(g^\top d + \frac{\beta}{2}\|d\|^2)$ remains a convex program whose minimizer equals $-p^*/\beta$ whenever $\|p^*\|/\beta \leq \Delta$, and saturates the boundary $\|d\| = \Delta$ along $-p^*$ otherwise. In our practical vision deployments modulating normalization parameters, empirical prototype gradients satisfy $\max_{m\in\mathcal{M}} \|\hat{g}_m\| \leq 0.42$. With $\beta = 1.0$, the unconstrained update norm satisfies $\|d^*\| = \|p^*\|/\beta \leq \max_{g\in\Gamma} \|g\|/\beta \leq 0.42$, which naturally and strictly resides within a conservative local trust region of radius $\Delta = 0.5$. Thus, identifying $\mathcal{D}$ with $\mathbb{R}^r$ yields the closed-form update while ensuring the adaptation step remains well within the local quadratic validity regime. □

A.3 Proof of Theorem 3 (Classical Simplex Reduction)

Theorem 3 (restated: Exact Diagonal-Preserving Reduction). *Suppose there exists an orthonormal basis $\{|i\rangle\}_{i=1}^{d_E}$ of $\mathcal{H}_E$ such that every mechanism evidence state ρ_m is diagonal, and for every probe $q \in \mathcal{Q}$ and outcome $y \in \mathcal{Y}_q$, the instrument map $\mathcal{I}_y^{(q)}$ preserves diagonality in $\{|i\rangle\}$. Then:*

1. *Every instrument map induces a valid classical transition kernel $T_y^{(q)}(j \mid i) := \langle j|\mathcal{I}_y^{(q)}(|i\rangle\langle i|)|j\rangle$ satisfying $\sum_{y,j} T_y^{(q)}(j \mid i) = 1$.*
2. *For any probe sequence $q_1 \to \cdots \to q_L$, the joint outcome likelihood and posterior update $\pi_{y_{1:L}}$ are exactly identical to those of a classical hidden Markov filter over discrete states.*

Proof. Because each ρ_m is diagonal in $\{|i\rangle\}$, we write:

$$\rho_m = \sum_{i=1}^{d_E} p_{m,i}|i\rangle\langle i|, \quad \text{where } p_{m,i} \geq 0, \sum_{i=1}^{d_E} p_{m,i} = 1. \tag{28}$$

By assumption, for every basis projector $|i\rangle\langle i|$, the instrument output $\mathcal{I}_y^{(q)}(|i\rangle\langle i|)$ is diagonal in $\{|j\rangle\}$. Therefore:

$$\mathcal{I}_y^{(q)}(|i\rangle\langle i|) = \sum_{j=1}^{d_E} T_y^{(q)}(j \mid i)|j\rangle\langle j|, \tag{29}$$

where $T_y^{(q)}(j \mid i) := \langle j|\mathcal{I}_y^{(q)}(|i\rangle\langle i|)|j\rangle$.

Step 1: Complete Positivity and Normalization. Since $\mathcal{I}_y^{(q)}$ is a completely positive map, it maps positive semidefinite operators to positive semidefinite operators. Since $|i\rangle\langle i| \succeq 0$, we have $\mathcal{I}_y^{(q)}(|i\rangle\langle i|) \succeq 0$, which implies $T_y^{(q)}(j \mid i) \geq 0$ for all y, i, j. Furthermore, by Assumption 3, the sum $\sum_{y\in\mathcal{Y}_q} \mathcal{I}_y^{(q)}$ is trace-preserving:

$$\sum_{y\in\mathcal{Y}_q}\sum_{j=1}^{d_E} T_y^{(q)}(j \mid i) = \sum_{y\in\mathcal{Y}_q} \mathrm{Tr}\left[\mathcal{I}_y^{(q)}(|i\rangle\langle i|)\right] = \mathrm{Tr}\left[\left(\sum_{y\in\mathcal{Y}_q} \mathcal{I}_y^{(q)}\right)(|i\rangle\langle i|)\right] = \mathrm{Tr}[|i\rangle\langle i|] = 1. \tag{30}$$

Thus, $T_y^{(q)}(j \mid i)$ defines a valid joint emission-transition probability distribution in a classical Hidden Markov Model.

Step 2: Single-Probe Outcome Likelihood and Post-State. For an arbitrary input probability vector $p = (p_1, \ldots, p_{d_E})^\top$ with diagonal state $\rho = \sum_i p_i|i\rangle\langle i|$, the outcome likelihood under probe q is:

$$p(y \mid \rho, q) = \mathrm{Tr}[\mathcal{I}_y^{(q)}(\rho)] = \sum_{i=1}^{d_E} p_i \,\mathrm{Tr}[\mathcal{I}_y^{(q)}(|i\rangle\langle i|)] = \sum_{i=1}^{d_E} p_i \left(\sum_{j=1}^{d_E} T_y^{(q)}(j \mid i)\right). \tag{31}$$

The unnormalized post-measurement state is:

$$\mathcal{I}_y^{(q)}(\rho) = \sum_{j=1}^{d_E}\left(\sum_{i=1}^{d_E} p_i T_y^{(q)}(j \mid i)\right)|j\rangle\langle j|. \tag{32}$$

The diagonal components of the normalized post-state $\rho_{|y}$ are:

$$p'_{j|y} = \frac{\sum_{i=1}^{d_E} p_i T_y^{(q)}(j \mid i)}{\sum_{k=1}^{d_E}\sum_{i=1}^{d_E} p_i T_y^{(q)}(k \mid i)}, \tag{33}$$

which is the exact classical Bayesian filter update for a latent Markov state $i \to j$ with observed emission y.

Step 3: Induction on Sequence Length. For $L = 1$, the statement holds by Step 2. Suppose for any sequence of length $L-1$, the joint likelihood and posterior state equal those of a classical filter:

$$p(y_{1:L-1} \mid m) = p_{\text{classical}}(y_{1:L-1} \mid m), \quad \rho_{m|y_{1:L-1}} = \sum_j p_{\text{classical}}(j \mid y_{1:L-1}, m)|j\rangle\langle j|. \tag{34}$$

Applying probe q_L with outcome y_L, the joint likelihood is:

$$p(y_{1:L} \mid m) = p(y_{1:L-1} \mid m) \operatorname{Tr}\left[\mathcal{I}_{y_L}^{(q_L)}(\rho_{m|y_{1:L-1}})\right] \tag{35}$$

$$= p_{\text{classical}}(y_{1:L-1} \mid m) \sum_{j=1}^{d_E} p_{\text{classical}}(j \mid y_{1:L-1}, m) \sum_{k=1}^{d_E} T_{y_L}^{(q_L)}(k \mid j)$$

$$= p_{\text{classical}}(y_{1:L} \mid m). \tag{36}$$

By Bayes' rule, the posterior on the causal register $\pi_{y_{1:L}}(m) \propto \pi(m)p(y_{1:L} \mid m)$ coincides identically with the classical posterior. Since observable $\bar{A}(d)$ is diagonal on $\mathcal{H}_C$, the safety certificates depend solely on $\pi_{y_{1:L}}$. □

A.4 Proof of Theorem 4 (Order Inequivalence)

Theorem 4 (restated: Sequential Order-Dependence and Non-Commuting Probing Asymmetry). *There exist two 2-dimensional evidence states ρ_{+1}, ρ_{-1} and two probe instruments q_1, q_2 such that all single-probe marginals are perfectly identical:*

$$p(y \mid s = +1, q_i) = p(y \mid s = -1, q_i) = \frac{1}{2}, \quad \forall i \in \{1, 2\}, \forall y, \tag{37}$$

yet the ordered probe sequence $q_1 \to q_2$ deterministically identifies the hidden mechanism:

$$p(b = +1 \mid s, q_1 \to q_2) = \mathbf{1}\{s = +1\}, \tag{38}$$

whereas the reversed sequence $q_2 \to q_1$ yields completely uninformative outcomes $p(b \mid s, q_2 \to q_1) = 1/2$. In contrast, for any commuting instruments $\mathcal{I}^{(1)} \circ \mathcal{I}^{(2)} = \mathcal{I}^{(2)} \circ \mathcal{I}^{(1)}$, joint outcome probabilities are invariant to probe order ($p(y_1, y_2 \mid q_1 \to q_2) = p(y_1, y_2 \mid q_2 \to q_1)$), precluding directional identification asymmetry.

Proof. Let $\mathcal{H}_E = \mathbb{C}^2$. The standard Pauli matrices are:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{39}$$

Let the two causal mechanisms be parameterized by $s \in \{+1, -1\}$. Define the initial density operators:

$$\rho_s = \frac{I + s\sigma_y}{2} = \frac{1}{2}\begin{pmatrix} 1 & -is \\ is & 1 \end{pmatrix}. \tag{40}$$

Notice that $\operatorname{Tr}[\rho_s] = 1$ and $\rho_s^2 = \rho_s$, so ρ_s are pure qubit states on the Bloch sphere pointing along the $+y$ and $-y$ axes.

Definition of Instruments:

- **Probe q_1:** A completely positive instrument with outcome set $\mathcal{Y}_1 = \{0, 1\}$. It applies a unitary rotation $U = \exp(-i\frac{\pi}{4}\sigma_x) = \frac{1}{\sqrt{2}}(I - i\sigma_x) = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -i \\ -i & 1 \end{pmatrix}$ with fair coin outcome:
$$\mathcal{I}_a^{(1)}(\rho) = \frac{1}{2}U\rho U^\dagger, \quad \forall a \in \{0, 1\}. \tag{41}$$
The map $\sum_{a=0}^{1} \mathcal{I}_a^{(1)}(\rho) = U\rho U^\dagger$ is trace-preserving.
- **Probe q_2:** A projective measurement in the computational σ_z basis with outcome $b \in \{+1, -1\}$. The measurement operators are:
$$M_{+1} = |0\rangle\langle 0| = \frac{I + \sigma_z}{2}, \quad M_{-1} = |1\rangle\langle 1| = \frac{I - \sigma_z}{2}. \tag{42}$$
The instrument maps are $\mathcal{I}_b^{(2)}(\rho) = M_b\rho M_b$.

Single-Probe Marginals:

- For q_1 alone: For either mechanism $s \in \{+1, -1\}$ and outcome $a \in \{0, 1\}$:

$$p(a \mid s, q_1) = \mathrm{Tr}[\mathcal{I}_a^{(1)}(\rho_s)] = \frac{1}{2}\,\mathrm{Tr}[U\rho_s U^\dagger] = \frac{1}{2}\,\mathrm{Tr}[\rho_s] = \frac{1}{2}. \tag{43}$$

- For q_2 alone: The expectation of σ_z on ρ_s is:

$$\mathrm{Tr}[\sigma_z \rho_s] = \mathrm{Tr}\left[\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\frac{1}{2}\begin{pmatrix}1 & -is\\ is & 1\end{pmatrix}\right] = \frac{1}{2}\,\mathrm{Tr}\begin{pmatrix}1 & -is\\ -is & -1\end{pmatrix} = 0. \tag{44}$$

  Therefore, for any $b \in \{+1, -1\}$:

$$p(b \mid s, q_2) = \mathrm{Tr}[M_b \rho_s] = \mathrm{Tr}\left[\frac{I + b\sigma_z}{2}\rho_s\right] = \frac{1}{2} + \frac{b}{2}\,\mathrm{Tr}[\sigma_z \rho_s] = \frac{1}{2}. \tag{45}$$

Thus, neither probe alone provides any statistical discrimination between mechanism $s = +1$ and $s = -1$.

Ordered Sequence $q_1 \to q_2$: Applying q_1 first rotates the density matrix by U:

$$U\rho_s U^\dagger = \frac{1}{\sqrt{2}}(I - i\sigma_x)\left(\frac{I + s\sigma_y}{2}\right)\frac{1}{\sqrt{2}}(I + i\sigma_x) = \frac{1}{2}(I - i\sigma_x)\left(\frac{I + s\sigma_y}{2}\right)(I + i\sigma_x). \tag{46}$$

Using the Pauli algebra identities $\sigma_x\sigma_y = i\sigma_z$ and $\sigma_y\sigma_x = -i\sigma_z$:

$$(I - i\sigma_x)\sigma_y(I + i\sigma_x) = (\sigma_y + \sigma_z)(I + i\sigma_x) = 2\sigma_z. \tag{47}$$

Therefore:

$$U\rho_s U^\dagger = \frac{I + s\sigma_z}{2}. \tag{48}$$

Now, measuring q_2 on the unnormalized post-state $\mathcal{I}_a^{(1)}(\rho_s) = \frac{1}{2}U\rho_s U^\dagger$:

$$p(a, b \mid s, q_1 \to q_2) = \mathrm{Tr}\left[M_b\left(\frac{1}{2}U\rho_s U^\dagger\right)\right] = \frac{1 + bs}{4}. \tag{49}$$

Summing over the intermediate unobserved outcome $a \in \{0, 1\}$:

$$p(b \mid s, q_1 \to q_2) = \sum_{a=0}^{1} p(a, b \mid s, q_1 \to q_2) = 2 \cdot \frac{1 + bs}{4} = \frac{1 + bs}{2} = \mathbf{1}\{b = s\}. \tag{50}$$

Thus, outcome b perfectly and deterministically identifies the hidden causal mechanism s.

Reversed Sequence $q_2 \to q_1$: If q_2 is measured first, the post-measurement state for outcome $b \in \{+1, -1\}$ is:

$$\rho_{s|b} = \frac{M_b \rho_s M_b}{\mathrm{Tr}[M_b \rho_s]} = \frac{I + b\sigma_z}{2}. \tag{51}$$

Notice that $\rho_{s|b}$ depends exclusively on the observed outcome b and is completely independent of the true mechanism s. Subsequent application of probe q_1 yields $p(a \mid b, s, q_2 \to q_1) = 1/2$. The joint distribution is:

$$p(b, a \mid s, q_2 \to q_1) = \frac{1}{4}, \quad \forall a, b, s. \tag{52}$$

Thus, the reversed sequence is completely uninformative.

Order-Invariance of Commuting Instruments: For any pair of commuting instruments $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$ satisfying $\mathcal{I}_b^{(2)} \circ \mathcal{I}_a^{(1)} = \mathcal{I}_a^{(1)} \circ \mathcal{I}_b^{(2)}$ for all a, b, the joint outcome distribution satisfies:

$$p(a, b \mid s, q_1 \to q_2) = \mathrm{Tr}\left[(\mathcal{I}_b^{(2)} \circ \mathcal{I}_a^{(1)})(\rho_s)\right] = \mathrm{Tr}\left[(\mathcal{I}_a^{(1)} \circ \mathcal{I}_b^{(2)})(\rho_s)\right] = p(b, a \mid s, q_2 \to q_1). \tag{53}$$

Consequently, under any commuting instruments, reversing the execution order leaves joint outcome likelihoods identical ($p(y_1, y_2 \mid q_1 \to q_2) = p(y_1, y_2 \mid q_2 \to q_1)$), precluding directional identification asymmetry. □

A.5 Proof of Proposition 5 (Measurement Certifiability)

Proposition 5 (restated: Measurement-Induced Certifiability). *Suppose initial posterior π satisfies $0 \in \Gamma_\delta(\pi)$, so $V_\delta(\pi) = 0$. Let $q \in \mathcal{Q}$ be a probe instrument. If there exists at least one outcome $\bar{y} \in \mathcal{Y}_q$ with positive probability $p(\bar{y}) > 0$ such that Bayes updating contracts $\mathcal{M}_\delta(\pi_{\bar{y}})$ sufficiently to exclude opposing mechanisms, yielding $0 \notin \Gamma_\delta(\pi_{\bar{y}})$, then the expected guaranteed improvement strictly increases from zero to strictly positive:*

$$\mathbb{E}_{y\sim p(y|q)}[V_\delta(\pi_y)] = \sum_{y\in\mathcal{Y}_q} p(y \mid q)\frac{\operatorname{dist}(0,\Gamma_\delta(\pi_y))^2}{2\beta} > 0 = V_\delta(\pi). \tag{54}$$

For probe cost $c(q) < \mathbb{E}_y[V_\delta(\pi_y)]$, the net expected value $\mathbb{E}_y[V_\delta(\pi_y)] - c(q)$ is strictly positive.

Proof. By Theorem 2, for any posterior state π_y:

$$V_\delta(\pi_y) = \frac{\operatorname{dist}(0,\Gamma_\delta(\pi_y))^2}{2\beta} \geq 0. \tag{55}$$

By assumption, there exists an outcome $\bar{y} \in \mathcal{Y}_q$ such that $p(\bar{y}) > 0$ and $0 \notin \Gamma_\delta(\pi_{\bar{y}})$. By Theorem 2, $0 \notin \Gamma_\delta(\pi_{\bar{y}}) \iff \operatorname{dist}(0,\Gamma_\delta(\pi_{\bar{y}})) = \kappa_\delta(\pi_{\bar{y}}) > 0$, which implies:

$$V_\delta(\pi_{\bar{y}}) = \frac{\kappa_\delta(\pi_{\bar{y}})^2}{2\beta} > 0. \tag{56}$$

Now consider the expected certifiable value across all outcomes:

$$\mathbb{E}_{y\sim p(y|q)}[V_\delta(\pi_y)] = \sum_{y\in\mathcal{Y}_q} p(y \mid q)V_\delta(\pi_y) = p(\bar{y})V_\delta(\pi_{\bar{y}}) + \sum_{y\neq\bar{y}} p(y \mid q)V_\delta(\pi_y). \tag{57}$$

Since every term in the sum is non-negative ($p(y \mid q) \geq 0$ and $V_\delta(\pi_y) \geq 0$), and the term for $\bar{y}$ is strictly positive, we have:

$$\mathbb{E}_{y\sim p(y|q)}[V_\delta(\pi_y)] \geq p(\bar{y})V_\delta(\pi_{\bar{y}}) > 0. \tag{58}$$

Since $0 \in \Gamma_\delta(\pi)$, the initial certifiable value is $V_\delta(\pi) = 0$. Thus:

$$\mathbb{E}_{y\sim p(y|q)}[V_\delta(\pi_y)] > V_\delta(\pi) = 0. \tag{59}$$

This confirms that active measurement converts an initially uncertifiable problem into one with strictly positive expected guaranteed descent. □

A.6 Proof of Proposition 6 (Bellman Probing Recursion)

Proposition 6 (restated: Bellman Probing Recursion). *Given probe budget $B \in \mathbb{N}$, cost $c(q) \geq 0$, and orbital state $\Omega = (\pi, \{\rho_m\}_{m=1}^K)$, the optimal expected value function $J_B(\Omega)$ satisfies the Bellman equation:*

$$J_0(\Omega) = V_\delta(\pi), \quad J_B(\Omega) = \max\left\{ V_\delta(\pi), \max_{q\in\mathcal{Q}} \left[\sum_{y\in\mathcal{Y}_q} p(y \mid \Omega, q) J_{B-1}(\Omega^{(q,y)}) - c(q) \right] \right\}, \tag{60}$$

where $p(y \mid \Omega, q) = \sum_m \pi(m) \operatorname{Tr}[\mathcal{I}_y^{(q)}(\rho_m)]$ and $\Omega^{(q,y)} = (\pi^{(q,y)}, \{\rho_{m|y}^{(q)}\}_{m=1}^K)$.

Proof. This is a finite-horizon Markov Decision Process (MDP) with state space $\mathcal{S} = \Delta^{K-1} \times \mathcal{D}(\mathcal{H}_E)^K$ and horizon B. Because measurement instruments are non-commuting, the post-measurement state $\rho_{m|y} = \mathcal{I}_y^{(q)}(\rho_m)/\operatorname{Tr}[\mathcal{I}_y^{(q)}(\rho_m)]$ alters future outcome distributions; hence the full composite state Ω is a sufficient Markov state.

- **Base Case ($B = 0$):** With zero remaining budget, no further measurements can be made. The agent decides whether to adapt using the causal posterior π (yielding guaranteed value $V_\delta(\pi)$) or abstain (yielding value 0). Since $V_\delta(\pi) \geq 0$, $J_0(\Omega) = \max\{V_\delta(\pi), 0\} = V_\delta(\pi)$.

- **Inductive Step ($B \geq 1$):** At stage B, the agent can either:
  1. Stop probing immediately, execute update $d^*(\pi)$, and receive guaranteed value $V_\delta(\pi)$.
  2. Select probe $q \in \mathcal{Q}$, incur cost $c(q)$, observe outcome $y \sim p(y \mid \Omega, q)$, transition to state $\Omega^{(q,y)}$, and achieve expected future value $\sum_y p(y \mid \Omega, q) J_{B-1}(\Omega^{(q,y)})$.

  By the Bellman principle of optimality (Bellman, 1957), the optimal value is the maximum over these choices:

$$J_B(\Omega) = \max \left\{ V_\delta(\pi), \max_{q \in \mathcal{Q}} \left[\sum_{y \in \mathcal{Y}_q} p(y \mid \Omega, q) J_{B-1}(\Omega^{(q,y)}) - c(q) \right] \right\}. \tag{61}$$

Backward induction establishes optimality. □

A.7 Proof of Theorem 7 (Conformal Safety Control)

Theorem 7 (restated: Episode-Level Conformal Safety Control). *Let environments $E \sim \mathcal{P}_\mathcal{E}$ and evaluation episodes $Z_E = (X_E, Y_E) \sim Q_E$ ($N = 64$) follow an identical hierarchical sampling procedure across calibration and testing. Assume candidate update rule $\mathcal{A} : X_E \mapsto d_E^*$ (where $d_E^* = -\frac{1}{\beta} \mathrm{Proj}_{\Gamma_\delta(\pi_E)}(0)$ if $0 \notin \Gamma_\delta(\pi_E)$ and 0 otherwise) and nonconformity score function $\hat{s}(E, Z_E) := \Delta\widehat{R}_E(d_E^*) + V_\delta(\pi_E)$ are fixed prior to calibration, with the model resetting to θ_0 on every episode. Let calibration episodes $\{(E_i, Z_{E_i})\}_{i=1}^M$ and a new deployed test episode (E_t, Z_{E_t}) be exchangeable draws from $E \sim \mathcal{P}_\mathcal{E}, Z_E \sim Q_E$. Let $\hat{\tau}_\alpha$ be the empirical quantile of $\{\hat{s}(E_i, Z_{E_i})\}_{i=1}^M$ at rank $k = \lceil (M+1)(1-\alpha) \rceil$ (for $\alpha < 1/(M+1)$, $\hat{\tau}_\alpha = +\infty$ and coverage is 0). If Algorithm 1 certifies update $d_t^* \neq 0$ on deployed test batch $X_{E_t,b}$ only when $V_\delta(\pi_{E_t}) > \hat{\tau}_\alpha$, then the marginal probability of unsafe adaptation on a new deployed test episode is bounded:*

$$\mathbb{P}\left(\Delta\widehat{R}_{E_t,b}(d_t^*) \geq 0 \text{ and } d_t^* \neq 0 \right) \leq \alpha. \tag{62}$$

Proof. Let $\hat{S}_i := \hat{s}(E_i, Z_{E_i}) = \Delta\widehat{R}_{E_i,b}(d_{E_i}^*) + V_\delta(\pi_{E_i})$ for $i = 1, \ldots, M$, and $\hat{S}_{M+1} := \hat{s}(E_t, Z_{E_t}) = \Delta\widehat{R}_{E_t,b}(d_t^*) + V_\delta(\pi_{E_t})$. Because calibration and test episodes are exchangeable draws from the hierarchical data-generating process $E \sim \mathcal{P}_\mathcal{E}, Z_E \sim Q_E$ independently of meta-training data $\mathcal{E}_{\text{train}}$, conditioning on $\mathcal{E}_{\text{train}}$ and all frozen meta-trained parameters preserves mutual exchangeability. Since candidate mapping $\mathcal{A}$ is applied symmetrically to unlabeled inputs X_E and empirical risks are evaluated on sequestered labels Y_E within each episode after resetting to θ_0, the scalar nonconformity scores $\hat{S}_1, \ldots, \hat{S}_M, \hat{S}_{M+1}$ are conditionally exchangeable real-valued random variables.

By standard split-conformal prediction theory (Vovk et al., 2005; Tibshirani et al., 2019; Angelopoulos & Bates, 2021), letting $\hat{S}_{(1)} \leq \cdots \leq \hat{S}_{(M)}$ denote the sorted calibration scores with convention $\hat{S}_{(M+1)} = +\infty$, setting $\hat{\tau}_\alpha = \hat{S}_{(k)}$ with rank $k = \lceil (M+1)(1-\alpha) \rceil$ guarantees that the probability that test conformity score $\hat{S}_{M+1}$ exceeds $\hat{\tau}_\alpha$ satisfies:

$$\mathbb{P}\left(\hat{S}_{M+1} > \hat{\tau}_\alpha \right) \leq \alpha. \tag{63}$$

Now, consider the event of an unsafe empirical certification:

$$\mathcal{E}_{\text{unsafe}} := \left\{ \Delta\widehat{R}_{E_t,b}(d_t^*) \geq 0 \text{ and } d_t^* \neq 0 \right\}. \tag{64}$$

By Algorithm 1, an update $d_t^* \neq 0$ is certified if and only if $0 \notin \Gamma_\delta(\pi_{E_t})$ and $V_\delta(\pi_{E_t}) > \hat{\tau}_\alpha$.

On the conformal acceptance event $\{\hat{S}_{M+1} \leq \hat{\tau}_\alpha\}$, by definition of $\hat{S}_{M+1}$:

$$\Delta\widehat{R}_{E_t,b}(d_t^*) + V_\delta(\pi_{E_t}) \leq \hat{\tau}_\alpha \implies \Delta\widehat{R}_{E_t,b}(d_t^*) \leq -V_\delta(\pi_{E_t}) + \hat{\tau}_\alpha. \tag{65}$$

When certified, $V_\delta(\pi_{E_t}) > \hat{\tau}_\alpha$, so $-V_\delta(\pi_{E_t}) < -\hat{\tau}_\alpha$. Substituting this yields:

$$\Delta\widehat{R}_{E_t,b}(d_t^*) < -\hat{\tau}_\alpha + \hat{\tau}_\alpha = 0. \tag{66}$$

Thus, whenever $\hat{S}_{M+1} \leq \hat{\tau}_\alpha$, any certified update satisfies $\Delta\widehat{R}_{E_t,b}(d_t^*) < 0$ (strict empirical descent on the deployed batch), which implies $\{\hat{S}_{M+1} \leq \hat{\tau}_\alpha\} \implies \mathcal{E}_{\text{unsafe}}^c$, or equivalently $\mathcal{E}_{\text{unsafe}} \subseteq \{\hat{S}_{M+1} > \hat{\tau}_\alpha\}$. Taking probabilities and applying Equation 63:

$$\mathbb{P}(\mathcal{E}_{\text{unsafe}}) \leq \mathbb{P}(\hat{S}_{M+1} > \hat{\tau}_\alpha) \leq \alpha. \tag{67}$$

This establishes distribution-free finite-sample control over the unsafe empirical certification rate on deployed batches without requiring continuous optimization over $\mathcal{D}$. □

Safety Guarantees and Scope. We explicitly demarcate the boundaries and nature of the theoretical guarantees provided by Theorem 7: (1) *Marginal Batch-Risk Control:* The guarantee $\mathbb{P}(\Delta\widehat{R}_{E_t,b}(d_t^*) \geq 0, d_t^* \neq 0) \leq \alpha$ holds marginally across the joint distribution of exchangeable deployment episodes (E_t, Z_{E_t}). It governs empirical cross-entropy risk non-degradation on the deployed target batch ($N = 64$). (2) *Conditional Risk on Accepted Adaptations:* In safety-critical applications, practitioners care about the conditional risk given that the model chooses to adapt: $\mathbb{P}(\Delta\widehat{R}_{E_t,b}(d_t^*) \geq 0 \mid d_t^* \neq 0) = \frac{\mathbb{P}(\Delta\widehat{R}_{E_t,b}(d_t^*) \geq 0, d_t^* \neq 0)}{\mathbb{P}(d_t^* \neq 0)} \leq \frac{\alpha}{c}$, where $c = \mathbb{P}(d_t^* \neq 0)$ is the certified coverage. For our main benchmark results with $c = 81.7\%$ at nominal $\alpha = 0.05$, the theoretical conditional upper bound is $\alpha/c \approx 6.1\%$, while the empirically observed conditional unsafe rate on in-family deployments is just 2.6% (Table 4). (3) *Distinction from Population Risk and 0-1 Classification Error:* Theorem 7 does not guarantee that infinite-sample population risk or 0-1 top-1 misclassification error is non-increasing on every individual test instance. Rather, empirical batch cross-entropy risk is the formal metric calibrated against empirical labels, which acts as a statistically consistent finite-sample surrogate for distribution-level stability.

A.8 Supporting Lemmas

Lemma A.1 (Causal Restriction Monotonicity). *Let $\Gamma_A \subset \mathbb{R}^r$ be a non-empty convex gradient uncertainty set and $\Gamma_C \subseteq \Gamma_A$ be a non-empty convex sub-hull obtained after applying valid SCM restrictions that do not exclude the true target mechanism. Then guaranteed certifiable value cannot decrease: $V(\Gamma_C) \geq V(\Gamma_A)$.*

Proof. If $0 \in \Gamma_C$, then $\Gamma_C \subseteq \Gamma_A \implies 0 \in \Gamma_A$, so $V(\Gamma_C) = V(\Gamma_A) = 0$. If $0 \notin \Gamma_C$, taking the infimum over a smaller subset cannot decrease Euclidean distance to the origin: $\text{dist}(0, \Gamma_C) = \inf_{g \in \Gamma_C} \|g\| \geq \inf_{g \in \Gamma_A} \|g\| = \text{dist}(0, \Gamma_A)$. Squaring and dividing by $2\beta > 0$ establishes $V(\Gamma_C) \geq V(\Gamma_A)$. Nested containment $\Gamma_C \subseteq \Gamma_A$ holds under geometric hull intersection $\Gamma_{\text{proj}} = \Gamma_\delta(\pi) \cap \text{conv}(\bigcup_{m \in \mathcal{M}_{\text{SCM}}} G_m)$. In discrete implementations re-extracting $\mathcal{M}_\delta(\tilde{\pi})$ after renormalizing $\tilde{\pi}(m) \propto \pi(m)\mathbf{1}\{m \in \mathcal{M}_{\text{SCM}}\}$, reallocated mass can occasionally promote sub-threshold indices into $\mathcal{M}_\delta(\tilde{\pi})$. In such cases, CausalProject functions as a structural admissibility filter: eliminating non-physical mechanisms prevents certifying spurious descent directions and aligns candidate updates strictly with the admissible SCM. □

Lemma A.2 (Distributional Misspecification Robustness). *Let $\mathcal{C}$ be a calibrated model cloud of distributions. Suppose the true target distribution P^* satisfies $\inf_{P \in \mathcal{C}} \mathcal{W}_1(P, P^*) \leq \varepsilon$, and the gradient mapping $P \mapsto g(P)$ is L_g-Lipschitz with respect to 1-Wasserstein distance. If an update d is certified on $\mathcal{C}$ with margin $m > 0$, then:*

$$\Delta R_{P^*}(d) \leq -m + L_g \varepsilon \|d\|. \tag{68}$$

In particular, descent is guaranteed on P^* whenever $m > L_g \varepsilon \|d\|$.

Proof. For any $\eta > 0$, choose $\bar{P} \in \mathcal{C}$ with $\mathcal{W}_1(\bar{P}, P^*) \leq \varepsilon + \eta$. By Assumption 2, $\Delta R_{P^*}(d) \leq g(P^*)^\top d + \frac{\beta}{2}\|d\|^2 = g(\bar{P})^\top d + \frac{\beta}{2}\|d\|^2 + (g(P^*) - g(\bar{P}))^\top d$. Since d is certified on $\mathcal{C}$, we have $g(\bar{P})^\top d + \frac{\beta}{2}\|d\|^2 \leq -m$. Applying Cauchy-Schwarz and L_g-Lipschitz continuity of the gradient map yields $(g(P^*) - g(\bar{P}))^\top d \leq \|g(P^*) - g(\bar{P})\|\|d\| \leq L_g \mathcal{W}_1(P^*, \bar{P})\|d\| \leq L_g(\varepsilon + \eta)\|d\|$. Combining bounds gives $\Delta R_{P^*}(d) \leq -m + L_g(\varepsilon + \eta)\|d\|$. Taking $\eta \to 0^+$ establishes $\Delta R_{P^*}(d) \leq -m + L_g \varepsilon \|d\|$. □

B Algorithmic Details and Optimization

B.1 SOCP Formulation for Hull Projection

Computing the certifiability margin $\kappa_\delta(\pi)$ and minimax optimal update $d^*(\pi)$ requires finding the Euclidean projection of the origin onto the active gradient hull:

$$p^* = \mathrm{Proj}_{\Gamma_\delta(\pi)}(0) = \arg\min_{g \in \Gamma_\delta(\pi)} \frac{1}{2}\|g\|^2. \tag{69}$$

Any point in the active convex hull is $g = \sum_{m \in \mathcal{M}_\delta(\pi)} w_m g_m$ with $w \in \Delta^{|\mathcal{M}_\delta(\pi)|-1}$ and $g_m \in G_m = \{g : \|g - \hat{g}_m\| \le r_m\}$. The support function in direction $u \in \mathbb{R}^r$ is:

$$h_\Gamma(u) := \sup_{g \in \Gamma_\delta(\pi)} u^\top g = \max_{m \in \mathcal{M}_\delta(\pi)} \left(u^\top \hat{g}_m + r_m \|u\|\right). \tag{70}$$

By Fenchel duality, the projection problem translates into a Second-Order Cone Program (SOCP):

$$\begin{aligned} \min_{w \in \mathbb{R}^{|\mathcal{M}_\delta(\pi)|}, \{v_m\}_m} \quad & \frac{1}{2}\left\| \sum_{m \in \mathcal{M}_\delta(\pi)} w_m \hat{g}_m + \sum_{m \in \mathcal{M}_\delta(\pi)} v_m \right\|^2 \\ \text{subject to} \quad & \|v_m\| \le w_m r_m, \quad \forall m \in \mathcal{M}_\delta(\pi), \\ & \sum_{m \in \mathcal{M}_\delta(\pi)} w_m = 1, \quad w_m \ge 0, \quad \forall m \in \mathcal{M}_\delta(\pi). \end{aligned} \tag{71}$$

Here, the decision variables comprise the convex weights $w \in \mathbb{R}^{|\mathcal{M}_\delta(\pi)|}$ and perturbation vectors $\{v_m \in \mathbb{R}^r\}_{m \in \mathcal{M}_\delta(\pi)}$, totaling $(r+1)|\mathcal{M}_\delta(\pi)| \le (r+1)K \le 55$ scalar variables (for subspace dimension $r = 10$ and $|\mathcal{M}_\delta(\pi)| \le K \le 5$). All reported experiments evaluate $K = 4$ candidate mechanism families (matching the 4 distinct shift hypotheses per benchmark suite and empirical $|\mathcal{M}_\delta| = 4.0 \to 1.0$ contraction); the $K \le 5$ complexity bound is stated for the broader general implementation. This SOCP is solved in < 0.5 ms using standard interior-point second-order cone solvers (e.g., SCS, Clarabel, ECOS, or MOSEK via CVXPY).

B.2 Gradient Prototype Calibration

In practice, true mechanism gradients $g_m = \nabla R_m(0)$ are unobserved. We construct empirical prototype centers and dispersion radii by averaging over training environments $\mathcal{E}_{\mathrm{train},m}$:

$$\hat{g}_m := \frac{1}{|\mathcal{E}_{\mathrm{train},m}|} \sum_{E \in \mathcal{E}_{\mathrm{train},m}} \nabla_d R_E(0), \qquad r_m := \max_{E \in \mathcal{E}_{\mathrm{train},m}} \|\nabla_d R_E(0) - \hat{g}_m\| + \varepsilon_{\mathrm{env}}, \tag{72}$$

where $\varepsilon_{\mathrm{env}} \ge 0$ accounts for within-split gradient variance. Crucially, we do not require restrictive sub-Gaussian assumptions on r_m for our system guarantees: any residual gradient misspecification, finite-sample error, or curvature discrepancy is absorbed and calibrated at the system level by the environment-level conformal gate $\hat{\tau}_\alpha$ (Theorem 7), providing rigorous, distribution-free finite-sample safety. The active gradient hull at test time is the convex hull:

$$\Gamma_\delta(\pi) = \mathrm{conv} \bigcup_{m \in \mathcal{M}_\delta(\pi)} \{g : \|g - \hat{g}_m\| \le r_m\}. \tag{73}$$

C Neural Architecture and Objectives

C.1 Orbital Encoder and Instruments

The orbital encoder $\mathcal{E}_\phi(X_{\mathrm{tgt}})$ maps an unlabeled batch X_{tgt} to an initial posterior $\pi_0 \in \Delta^{K-1}$ and mechanism density operators $\{\rho_{m,0}\}_{m=1}^K$.

- **Backbone Feature Extraction:** We use a standard pretrained ResNet-50 feature extractor for Synthetic CSI and ImageNet-C-style experiments, and a pretrained ConvNeXt-T for Simulated Multi-Center MedMNIST. Backbone weights remain frozen during test-time evaluation, computing batch statistics (mean activation vectors and spatial frequency energy spectra) that feed the lightweight mechanism posterior and density heads.

- **Mechanism Posterior Head:** A 2-layer MLP with softmax activation outputs $\pi_0(m) = \mathrm{Softmax}(W_\pi \bar{z})$.
- **Density Matrix Parameterization:** To strictly enforce positive semidefiniteness and unit trace, each density matrix ρ_m is parameterized via Cholesky factor projection:

$$L_m = \mathrm{vec}^{-1}(W_\rho \bar{z}_m), \quad \rho_m = \frac{L_m L_m^\dagger}{\mathrm{Tr}[L_m L_m^\dagger]}. \tag{74}$$

  In our experiments, the evidence orbital dimension is set to $d_E = 4$ with rank $r_\rho = 2$.
- **Kraus Instrument Parameterization and Exact Register Parameter Accounting (Assumption 3):** To strictly guarantee that each probe instrument $\{\mathcal{I}_y^{(q)}\}_{y\in\mathcal{Y}_q}$ satisfies completely positive and trace-non-increasing (TNI) properties with exact trace-preserving sum (Assumption 3), we implement an isometry-based Kraus parameterization. For each probe $q \in \mathcal{Q}$, let $n_K = 2$ be the Kraus rank. Unconstrained parameter blocks $A_{qyk} \in \mathbb{C}^{d_E \times d_E}$ are vertically stacked into a composite matrix $B_q \in \mathbb{C}^{(|\mathcal{Y}_q| n_K d_E)\times d_E}$. We compute its thin QR decomposition $B_q = Q_q R_q$, yielding an exact isometric matrix $Q_q \in \mathbb{C}^{(|\mathcal{Y}_q| n_K d_E)\times d_E}$ satisfying $Q_q^\dagger Q_q = I_{d_E}$ identically. Partitioning Q_q back into $(d_E \times d_E)$-dimensional blocks $\{K_{qyk}\}_{y\in\mathcal{Y}_q,k=1}^{n_K}$ guarantees:

$$\sum_{y\in\mathcal{Y}_q}\sum_{k=1}^{n_K} K_{qyk}^\dagger K_{qyk} = Q_q^\dagger Q_q = I_{d_E}, \tag{75}$$

  thereby algebraically ensuring exact trace preservation ($\mathrm{Tr}[\sum_y \mathcal{I}_y^{(q)}(\rho)] = \mathrm{Tr}[\rho]$) for the composite instrument, while each individual outcome map $\mathcal{I}_y^{(q)}(\rho) = \sum_{k=1}^{n_K} K_{qyk}\rho K_{qyk}^\dagger$ is completely positive and trace-non-increasing ($\sum_k K_{qyk}^\dagger K_{qyk} \preceq I_{d_E}$).
  Exact Register Parameter Accounting ($1,280$ *parameters):* With $d_E = 4$, $B_y = 5$ bins, and Kraus rank $n_K = 2$, each unconstrained block B_q has dimension 40×4 (160 complex numbers, 320 real floats). Across all $|\mathcal{Q}| = 4$ diagnostic probes, learned instruments require exactly $4 \times 320 = 1,280$ real learnable parameters (thin QR adds zero weights). This represents the active evidence register capacity. The visual backbone is shared, identical, and frozen across all methods.

C.2 Meta-Training Objective and Baselines

The orbital encoder and probe instrument parameters are trained end-to-end on $\mathcal{E}_{\mathrm{train}}$ using a composite objective:

$$\mathcal{L}_{\mathrm{meta}} = \mathcal{L}_{\mathrm{id}} + \lambda_1\mathcal{L}_{\mathrm{seq}} + \lambda_2\mathcal{L}_{\mathrm{grad}} + \lambda_3\mathcal{L}_{\mathrm{cal}}, \tag{76}$$

where $\mathcal{L}_{\mathrm{id}} = -\sum_E \log \pi(m_E^* \mid X_E)$, $\mathcal{L}_{\mathrm{seq}} = -\sum_E \log \mathrm{Tr}[\mathcal{I}_{y_2}^{(q_2)}(\mathcal{I}_{y_1}^{(q_1)}(\rho_{m_E^*}))]$, $\mathcal{L}_{\mathrm{grad}} = \sum_E \|\hat{g}(X_E) - g_{E,\mathrm{true}}\|^2$, and $\mathcal{L}_{\mathrm{cal}}$ is Brier score ($\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.2$, AdamW learning rate 3×10^{-4}).

Sequential GRU Architecture ($1,396$ params). The parameter-matched Sequential GRU ($h = 16$) matches orbital register capacity (1,300 parameters), receiving 8-d probe embedding $e(q_t)$ and scalar response $s_{q_t}(X_{\mathrm{tgt}})$ ($d_{\mathrm{in}} = 9$; $1,296$ recurrent weights, 68 readout parameters for π_t). To strictly isolate the operator state representation from auxiliary losses, Sequential GRU (Bellman) and Controlled HMM (Bellman) are meta-trained with the identical multi-task loss $\mathcal{L}_{\mathrm{meta}} = \mathcal{L}_{\mathrm{id}} + \lambda_1\mathcal{L}_{\mathrm{seq}} + \lambda_2\mathcal{L}_{\mathrm{grad}} + \lambda_3\mathcal{L}_{\mathrm{cal}}$ ($\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.2$) via linear projection from recurrent state h_t predicting proxy gradients $\hat{g}(h_t)$ and Brier calibration on softmax belief π_t. We evaluate GRU under both policies: (1) *Disentangling Search Objective from Tree Lookahead Depth:* In Table 1, standard greedy GRU ($66.3\pm1.5\%$ cov) and HMM ($57.7\pm1.6\%$ cov) use standard cross-entropy / MLE training, so their $+12.5\%p$ and $+15.7\%p$ gains under Bellman planning reflect the joint upgrade of multi-task loss ($\mathcal{L}_{\mathrm{meta}}$) and tree search. On fixed $\mathcal{L}_{\mathrm{meta}}$ checkpoints, policy upgrades cleanly separate into **search objective change** (info-gain greedy $\rightarrow$ value-greedy $H = 1$) versus **search depth lookahead** (value-greedy $H = 1 \rightarrow$ Bellman tree $H = 3$, both under budget $B = 3$): for Sequential GRU, info-gain greedy achieves $71.6 \pm 1.4\%$ coverage, value-greedy ($H = 1$) reaches $74.2 \pm 1.4\%$ ($+2.6\%p$ objective effect), and full tree search ($H = 3$) reaches $78.8 \pm 1.3\%$

($+4.6\%p$ lookahead depth effect; total policy gain $+7.2\%p$); for Causal Orbitals, info-gain greedy achieves $69.5 \pm 0.8\%$ coverage, value-greedy ($H = 1$) reaches $75.8 \pm 1.2\%$ ($+6.3\%p$ objective effect), and $H = 3$ tree search reaches $81.7 \pm 1.2\%$ ($+5.9\%p$ depth effect; total policy gain $+12.2\%p$); for Controlled HMM, greedy on $\mathcal{L}_{\text{meta}}$ attains $64.8 \pm 1.4\%$ coverage, yielding an overall policy gain of $+8.6\%p$ ($64.8\% \rightarrow 73.4\%$). (2) ***Bellman GRU ($H = 3$):*** evaluates full tree lookahead over 5 discrete virtual outcome branches per step with probe cost $\lambda_{\text{probe}} = 0.01$ and individualized split-conformal calibration on $\mathcal{E}_{\text{cal}}$ at $\alpha = 0.05$. Disaggregated across CSI (78.4%), ImageNet-C (78.1%), and MedMNIST (78.0%), it achieves macro-average $78.2 \pm 0.5\%$ score and $78.8 \pm 1.3\%$ coverage. Under matched Bellman planning and loss, Causal Orbitals confers a paired gain of $+2.9 \pm 0.5\%p$ coverage (95% paired Student's t-interval $[+1.51\%p, +4.29\%p]$, $t = 5.80, p = 0.0044$, strictly excluding zero) and $+0.7 \pm 0.3\%p$ score (95% paired Student's t-interval $[-0.13\%p, +1.53\%p]$, $t = 2.33, p = 0.080$, non-significant as it includes zero across $S = 5$ seeds; asymptotic normal approximations on the same seed statistics evaluate to $[+1.92, +3.88]\%p$ and $[+0.11, +1.29]\%p$).

Diagonal State Architecture (320 params). The Diagonal State baseline (ρ_{diag}) evaluates performance under a classical commuting diagonal filter by restricting density matrices to mixtures $\rho_t = \sum_{i=1}^{d_E} p_t(i)|i\rangle\langle i|$ ($d_E = 4$, $\rho_{ij} = 0$ for $i \neq j$). Each probe instrument is constrained to diagonal Kraus operators $K_{qyk} = \text{diag}(d_{qyk,1}, \ldots, d_{qyk,d_E}) \in \mathbb{C}^{d_E \times d_E}$ with normalization $\sum_{y,k} |d_{qyk,i}|^2 = 1$ (320 real parameters). The update $p_{t+1}(i) \propto \sum_k |d_{qyk,i}|^2 p_t(i)$ acts as a classical Bayes filter with commuting measurements that merely reweights fixed basis states without permitting off-diagonal coherences or coherent state transitions. Under greedy inquiry ($H = 1$, budget $B = 3$), it attains $74.7 \pm 0.5\%$ target score and $64.1 \pm 1.5\%$ coverage (Table 1). As this restriction simultaneously constrains parameter capacity, eliminates state transitions, and enforces greedy search, this comparison reflects the empirical behavior of a commuting diagonal filter rather than an isolated attribution to coherence alone.

HMM, Static, and Direct Router Baselines. The parameter-matched *Controlled HMM* ($1,281$ params) models discrete states $s_t \in \{1, \ldots, 4\}$ with action-conditioned transitions $T(q) \in \mathbb{R}^{4\times4}$ (80 params), probe embeddings (32 params), and an MLP Gaussian emission predictor ($1,169$ params) over 5 quintile bins (59.6% forward acc). The *Static Best-Fixed Probing* baseline evaluates the optimal non-adaptive sequence under matched maximum budget $B = 3$: exhaustive search on $\mathcal{E}_{\text{train}}$ selects $(q_1 \rightarrow q_2 \rightarrow q_3)$, achieving $76.6 \pm 0.5\%$ score and $74.8 \pm 1.4\%$ coverage (its 2-step prefix $q_1 \rightarrow q_2$ attains 76.1% score and 72.5% coverage). The *Direct Batch Router* ($1,288$ params) predicts updates directly from raw batch statistics without active probing (69.8% score, 44.2% cov). Every matched-supervision candidate baseline undergoes identical split-conformal calibration (Theorem 7) on $\mathcal{E}_{\text{cal}}$ at nominal $\alpha = 0.05$ with score $\hat{s}(E) = \Delta\widehat{R}_E(d_E^*) + V_\delta(\pi_E)$.

Table 2: **Architectural and Training Specifications Across Evaluated Baselines.** All methods share the identical frozen feature backbone and initial feature heads, while register parameters isolate the capacity allocated to tracking evidence and mechanism updates across diagnostic steps. Every matched-supervision candidate adaptation baseline undergoes individualized split-conformal calibration on $\mathcal{E}_{\text{cal}}$ at nominal $\alpha = 0.05$ (standard TTA baselines operate without gating).

Architecture	Common Modules	Tracking Module	Register Params	Input Features	Training Objective	Policy
Source-Only	ResNet-50 / ConvNeXt	None	0	Target batch X	None	None
TENT / SAR	ResNet-50 / ConvNeXt	Point parameters	0	Target batch X	Entropy / Sharpness	Passive
EATA / MEMO	ResNet-50 / ConvNeXt	Point parameters	0	Filtered / Aug. batch	Entropy / Marginal	Passive
T3A	ResNet-50 / ConvNeXt	Feature memory buffer	0	Feature embeddings	Centroid averaging	Non-parametric
EcoTTA	ResNet-50 / ConvNeXt	Meta-memory + Adapters	2,048	Target batch X	Self-distillation	Passive
DeYO	ResNet-50 / ConvNeXt	Point parameters	0	Filtered batch	Weighted Entropy + PLR	Passive
Direct Batch Router ($B = 0$)	ResNet-50 / ConvNeXt	2-layer MLP	1,288	Raw batch stats $\mu(X), \sigma(X)$	$\mathcal{L}_{\text{id}} + \lambda_3\mathcal{L}_{\text{cal}}$	Static ($B = 0$)
Non-Seq. Multi-Probe MLP	ResNet-50 / ConvNeXt	3-layer MLP	1,390	All 4 probes ($d = 32$)	Cross-entropy $\mathcal{L}_{\text{id}}$	Static
Classical Bayesian Filter	ResNet-50 / ConvNeXt	Discrete $\pi \in \Delta^3$	96	Sequential scalars	Likelihood $\mathcal{L}_{\text{id}}$	Random
Controlled HMM (greedy)	ResNet-50 / ConvNeXt	HMM + Neural Emission	1,281	Action embedding ($d = 8$)	MLE + $\mathcal{L}_{\text{id}}$	Greedy
Controlled HMM (Bellman)	ResNet-50 / ConvNeXt	HMM + Neural Emission	1,281	Action embedding ($d = 8$)	$\mathcal{L}_{\text{meta}}$ (matched)	Bellman
Static Best-Fixed Probing	ResNet-50 / ConvNeXt	Kraus Isometry Q_q	1,280	Operator state Ω	$\mathcal{L}_{\text{meta}}$ (Eq. 76)	Fixed $(q_1 \rightarrow q_2 \rightarrow q_3, B = 3)$
Sequential GRU (greedy)	ResNet-50 / ConvNeXt	GRU Cell ($h = 16$)	1,396	Emb. + scalar ($d = 9$)	Cross-entropy $\mathcal{L}_{\text{id}}$	Greedy
Sequential GRU (Bellman)	ResNet-50 / ConvNeXt	GRU Cell ($h = 16$)	1,396	Emb. + scalar ($d = 9$)	$\mathcal{L}_{\text{meta}}$ (matched)	Bellman
Diagonal State (ρ_{diag})	ResNet-50 / ConvNeXt	Diagonal density maps	320	Classical projection	$\mathcal{L}_{\text{meta}}$ (diagonal)	Greedy
Causal Orbitals	ResNet-50 / ConvNeXt	**Kraus Isometry** Q_q	**1,280**	**Operator state** Ω	$\mathcal{L}_{\text{meta}}$ **(Eq. 76)**	**Bellman**

Parameter Hierarchy. Across all evaluated methods, parameters partition strictly into three operational tiers: (1) *Meta-Trained Modules (Frozen at Test Time):* Common frozen visual backbone (ResNet-50 / ConvNeXt $z(\cdot)$) and readout heads. Sequence tracking parameters ($1,280$ real Kraus parameters for Ours; $1,281$ HMM parameters; $1,396$ GRU cell parameters; $1,390$ static MLP parameters; $1,288$ router parameters) are trained offline on $\mathcal{E}_{\text{train}}$ and strictly frozen during test-time adaptation with zero backpropagation. (2) *Test-Time Inference States (Zero Gradient Updates):*

Dynamic states tracking diagnostic evidence: categorical belief $\pi_t \in \Delta^{K-1}$ (Bayes filter and HMM), recurrent hidden state $h_t \in \mathbb{R}^{16}$ (Sequential GRU), and composite classical–operator state $\Omega_t = (\pi_t, \{\rho_{m,t}\}_{m=1}^K)$ (Causal Orbitals), evolving purely via forward algebraic measurement and Bayesian recursions without modifying network weights. (3) *Adapted Task Model Parameters:* For all candidate adaptation baselines, deployed adaptation modulates solely low-rank parameter update $d^* = Uz^* \in \mathbb{R}^p$ within subspace $\mathcal{D}$ ($r = 10$ affine parameters modulating BatchNorm/LayerNorm scale γ and shift β).

C.3 Reproducibility and Setup Details

- **Low-Rank Adaptation Subspace Basis ($r = 10$):** Candidate updates are restricted to $r = 10$ subspace $\mathcal{D} = \mathrm{span}\{\psi_1, \ldots, \psi_r\} \subset \mathbb{R}^d$ modulating affine scale γ and shift β of normalization layers. Basis vectors $\{\psi_j\}_{j=1}^{10}$ are extracted via SVD on centered gradients $\nabla_\theta R_E(\theta_0)$ across $\mathcal{E}_{\mathrm{train}}$.
- **Probe Quantization and Response Binning:** For each probe $q \in \mathcal{Q}$, scalar response $s_q(X) \in \mathbb{R}$ is quantized into $B_y = 5$ discrete measurement bins using empirical quintile thresholds calibrated on $\mathcal{E}_{\mathrm{train}}$.
- **Surrogate Curvature β Calibration and Estimation Rule:** To calibrate curvature constant $\beta > 0$ for Assumption 2, we evaluate the empirical Hessian spectral norm along the low-rank normalization adaptation subspace $\mathcal{D}$ across validation batches from $\mathcal{E}_{\mathrm{train}}$:

$$\hat{\beta} := \max_{m \in \mathcal{M}} \sup_{\|d\| \le \Delta} \lambda_{\max} \left(\frac{1}{|\mathcal{E}_{\mathrm{train},m}|} \sum_{E \in \mathcal{E}_{\mathrm{train},m}} \nabla_d^2 R_E(d) \right). \tag{77}$$

  For standard cross-entropy loss over normalization affine parameters (scale γ, shift β), this empirical maximum curvature evaluates to $\hat{\beta} \approx 0.94 \pm 0.08$. We thus set nominal $\beta = 1.0$ across all main benchmark experiments. The sensitivity analysis in Figure 3H confirms that certified coverage remains stable (75.4%–84.1%) and marginal safety is rigorously preserved across a four-fold variation $\beta \in [0.5, 2.0]$.
- **Decoupling Credible Mass δ from Conformal Risk Level α:** We explicitly decouple the two tolerance parameters (δ, α): (1) $(1-\delta)$-credible set $\mathcal{M}_\delta(\pi)$ is a *model-posterior probability statement* derived from the learned categorical belief $\pi \in \Delta^{K-1}$ under prior $\mathcal{L}_{\mathrm{meta}}$ training; it determines which gradient uncertainty balls G_m are active in the candidate hull $\Gamma_\delta(\pi)$ to guide proactive probing. (2) Conformal significance level α is a *distribution-free, frequentist statistical guarantee* governed by split-conformal quantile $\hat{\tau}_\alpha$ on exchangeable calibration episodes $\mathcal{E}_{\mathrm{cal}}$. Any residual misspecification in posterior calibration or curvature bound is absorbed at the conformal decision level, guaranteeing empirical marginal safety $\le \alpha$ regardless of small-sample imperfections in δ.
- **Evaluation Units, Metrics, and Statistical Aggregation:** (1) *Evaluation Unit:* Target batch $(X_{E,b}, Y_{E,b})$ ($N = 64$) under environment E. Conformity score $\hat{s}(E) = \Delta\widehat{R}(d_E^*; X_E, Y_E) + V_\delta(\pi_E)$ is calibrated over $M_{\mathrm{cal}} = 50$ episodes. Testing evaluates $N_{\mathrm{eval}} = 5$ episodes across $M_{\mathrm{test}} = 50$ environments (250 test episodes/seed; $1,250$/benchmark over $S = 5$ seeds; $3,750$ total). Tables 1 and 6 report macro-averages. (2) *Empirical Risk and Event Definitions:* Candidate d^* is computed label-free on $X_{E,b}$, and task risk change is evaluated via cross-entropy: $\Delta\widehat{R}_{E,b}(d^*) = \frac{1}{N}\sum_{i=1}^N \ell_{\mathrm{CE}}(f(X_{E,b}^{(i)}; \theta_0 + d^*), Y_{E,b}^{(i)}) - \frac{1}{N}\sum_{i=1}^N \ell_{\mathrm{CE}}(f(X_{E,b}^{(i)}; \theta_0), Y_{E,b}^{(i)})$. Certified episodes ($d^* \neq 0$, 81.7%) partition strictly by sign: strict risk reduction ($\Delta\widehat{R} < 0$, $79.6 \pm 0.4\%$), exact zero change ($\Delta\widehat{R} = 0$, $0.7 \pm 0.2\%$ where d^* modulates inactive channels), and strict increase ($\Delta\widehat{R} > 0$, $1.4 \pm 0.3\%$). By algebra, $\mathbb{P}(\Delta\widehat{R} \ge 0, d^* \neq 0) - \mathbb{P}(\Delta\widehat{R} > 0, d^* \neq 0) = 0.7\%p$, explaining the numerical difference between joint unsafe ($2.1 \pm 0.3\%$) and negative adaptation ($1.4 \pm 0.3\%$), with near-zero changes $|\Delta\widehat{R}| \in (0, 10^{-6}]$ accounting for 0.0%. (3) *Target Accuracy and Marginal vs. Conditional Gain:* Target score is top-1 accuracy (%). Under abstention ($d^* = 0$), $\Delta\mathrm{Acc} = 0$, satisfying identity $\mathbb{E}[\Delta\mathrm{Acc}] = c \cdot \mathbb{E}[\Delta\mathrm{Acc} \mid d^* \neq 0]$ for coverage $c = \mathbb{P}(d^* \neq 0)$. Across all $3,750$ episodes, source accuracy is $61.4 \pm 0.4\%$ and Full Causal Orbitals achieves $78.9 \pm 0.4\%$, yielding marginal gain $\mathbb{E}[\Delta\mathrm{Acc}] = +17.5\%p$ at $c = 81.7\%$

(conditional gain: $17.5\%/0.817 = +21.4\%p$). Table 6 Column 3 reports marginal gain $\mathbb{E}[\Delta\text{Acc}]$, scaling with c across ablations (Full: $+17.5\%p$; Diag: $+13.3\%p$; Reversed: $+7.8\%p$; Entropy: $+14.1\%p$). (4) *Confidence Intervals and Significance Testing:* Across $S = 5$ random seeds with resampled calibration and test splits, table metrics report mean $\hat{\mu} = \frac{1}{S}\sum_{s=1}^{S}\hat{p}_s$ and standard error $\text{SE} = \sqrt{\frac{1}{S(S-1)}\sum_{s=1}^{S}(\hat{p}_s - \hat{\mu})^2}$. For paired differences between Causal Orbitals and matched Bellman GRU, we compute paired seed differences $d_s = p_{\text{Ours},s} - p_{\text{GRU},s}$ for $s \in \{1, \ldots, 5\}$. With $S - 1 = 4$ degrees of freedom, the two-sided critical value is $t_{0.975,4} \approx 2.776$. The 95% paired Student's t-interval is $\bar{d} \pm 2.776 \times \text{SE}(d)$, yielding $[+1.51, +4.29]\%p$ for certified coverage ($t = 5.80, p = 0.0044$, statistically significant) and $[-0.13, +1.53]\%p$ for target score ($t = 2.33, p = 0.080$, non-significant at the 5% level, as the interval includes zero). If an asymptotic Gaussian approximation is computed on the same seed-level sample mean ($z_{0.975} = 1.96$), the intervals evaluate to $[+1.92, +3.88]\%p$ and $[+0.11, +1.29]\%p$. We report the paired Student's t-interval ($S = 5$, d.o.f. $= 4$) as the standard summary of small-sample seed-level variation.

- **Benchmark Protocols and Random Seeds:** (1) *Synthetic CSI*: 160 multi-shape environments ($N = 64$). (2) *ImageNet-C-Style*: 100 classes (224×224) across 50 corruption environments ($N = 64$). (3) *Simulated MedMNIST*: 160 acquisition environments across PathMNIST/BloodMNIST ($N = 64$). All table metrics indicate mean $\pm$ standard error across 5 random evaluation seeds.
- **Unified Latency Profiling and Component Breakdown:** Benchmarked on an NVIDIA A100 GPU (40 GB, CUDA 12.2, batch $N = 64$, 224×224, FP16/AMP) across 100 trials with `torch.cuda.synchronize()`: (1) *Source-Only (*$5.1 \approx 5$ *ms; Table 1):* PCIe staging (2.0 ms) + backbone (2.3 ms) + head (0.4 ms) + sync (0.4 ms). (2) *Standard TTA (TENT:* 12 *ms, SAR:* 15 *ms):* Forward pass (4.7 ms) + entropy loss/backward (4.8 ms) + step (0.2 ms) + adapted forward (2.3 ms); SAR adds sharpness checks (+3.0 ms). (3) *Causal Orbitals Breakdown (Total:* $17.3 \approx 18$ *ms): Phase 1: Setup (*5.1 *ms):* Transfer (2.0 ms) + backbone $z(X_0)$ (2.3 ms) + head (0.4 ms) + orbital encoder $\mathcal{E}_\phi(X_0)$ (0.4 ms). *Phase 2: Probing & Planning (*8.6 *ms across* $B = 3$ *steps):* Per probe: transform T_{q_t} (0.2 ms) + backbone $z(X_{t+1})$ (2.3 ms) + differential head (0.1 ms) $\implies 3 \times 2.6 = 7.8$ ms. Bellman lookahead evaluates over 8,000 paths via 15-subset lookup table $\mathcal{V}_{\text{LUT}} \in \mathbb{R}^{15}$ (0.5 ms) + Kraus updates (0.3 ms), totaling 8.6 ms. *Phase 3: Gating & Adapted Inference (*3.6 *ms):* Conformal quantile check (0.5 ms) + modulation $d^* = Uz^*$ (0.3 ms) + adapted forward pass with $\theta_0 + d^*$ (2.4 ms) + head (0.4 ms). (4) *Overhead Accounting:* Isolated phases sum to $5.1 + 8.6 + 3.6 = 17.3 \approx 18$ ms, matching Table 1. Causal Orbitals incurs 13 ms diagnostic overhead over Source-Only (5 ms), and +6 ms over TENT (12 ms), suppressing empirical negative adaptation to 1.4% within real-time budgets.

D Extended Experimental Details

D.1 CSI Protocol and Corruption Mapping

In the CSI benchmark, base images are generated under controlled interventions on acquisition mechanisms $S \in \{1, \dots, K\}$ ($K = 4$): (1) $m = 1$ **(Gaussian Blur)** with spectral attenuation ($\sigma \in [2.5, 4.0]$); (2) $m = 2$ **(Resampling)** via $4\times$ bilinear decimation; (3) $m = 3$ **(Saturation)** via thresholding at $[0.15, 0.85]$; and (4) $m = 4$ **(Contrast Compression)** via $x \mapsto 0.7x + 0.15$.

Conflicting Mechanism Pairs. These $K = 4$ mechanisms form two canonical pairs probing observational ambiguity and gradient conflict: **Pair 1 ($m = 1$ vs. $m = 2$): Gaussian Blur vs. Downsampling.** Both attenuate spatial detail while preserving layout, matching scalar roughness ($|\operatorname{TV}(X_{\text{blur}}) - \operatorname{TV}(X_{\text{down}})| / \operatorname{TV}(X_{\text{blur}}) < 0.02$, where $\operatorname{TV}(X) := \frac{1}{NHWC} \sum_{i,h,w,c} (|X_{i,h+1,w,c} - X_{i,h,w,c}| + |X_{i,h,w+1,c} - X_{i,h,w,c}|)$). Continuous Gaussian blur and discrete decimation alter phase differently, yielding opposing updates $\hat{g}_1 \approx -\hat{g}_2$. An unprobed batch splits the posterior across $\{1, 2\}$, entrapping the origin ($0 \in \Gamma_\delta(\pi)$) and mandating safe abstention ($d^* = 0$). Active probing (Laplacian filtering) discriminates $m = 1$ from $m = 2$, contracting $\mathcal{M}_\delta(\pi)$ ($0 \notin G_1$) and certifying descent. **Pair 2 ($m = 3$ vs. $m = 4$): Clipping vs. Contrast Compression.** Both reduce dynamic range and match low-order intensity moments, yet induce opposing updates ($\hat{g}_3 \approx -\hat{g}_4$).

Domain Exclusivity Rules in $\mathcal{M}_{\text{SCM}}$. In the simulated multi-center MedMNIST benchmark, acquisition metadata specifies sensor modalities across collection sites (optical microscopy vs. digital scanning arrays). In our simulator, optical microscopy environments enforce a physical domain constraint excluding digital spatial binning ($m = 2$ excluded a priori, setting $\mathcal{M}_{\text{SCM}} = \{1, 3, 4\}$). To ensure matched comparison, this domain constraint is provided identically to all matched baselines by masking their prior and transition supports over $\mathcal{M}_{\text{SCM}}$ at initialization ($t = 0$).

Corruption Mapping and Splits. Candidate families $\mathcal{M} = \{1, \dots, 4\}$ map systematically to physical distortion groupings covering all 15 standard ImageNet-C test corruptions (Hendrycks & Dietterich, 2019) alongside official validation sets (Table 3). To strictly satisfy the exchangeability precondition of Theorem 7 and test OOD limits: (1) *Meta-Training ($\mathcal{E}_{\text{train}}$):* Learns operators and value bounds on canonical corruptions (Gaussian val, Motion, Pixelate, Brightness, Fog, Contrast, Gaussian noise). (2) *Calibration ($\mathcal{E}_{\text{cal}}$) and Main Test ($\mathcal{E}_{\text{test}}$) [Exchangeable]:* Both splits ($M_{\text{cal}} = M_{\text{test}} = 50$) are drawn independently from the identical 12-corruption mixture across severities 1–5 (Table 3). Sharing the identical hierarchical episode distribution $\mathcal{D}_{\mathcal{E}}$ strictly satisfies exchangeability for Theorem 7 and split-conformal coverage (Angelopoulos & Bates, 2021). (3) *Held-Out Subtypes ($\mathcal{E}_{\text{heldout}}$):* 3 corruptions (Defocus blur, Glass blur, Elastic transform / Area resample) are withheld from training and calibration, evaluated exclusively in Table 7 to probe exchangeability violations.

Table 3: **Mapping from Physical Corruptions to Candidate Mechanism Families ($K = 4$).** Partition of all 15 standard ImageNet-C test corruptions (Hendrycks & Dietterich, 2019) and validation sets into meta-train, calibration, main test (exchangeable with $\mathcal{E}_{\text{cal}}$), and held-out test subtypes (Table 7).

Index m	Distortion Family	Meta-Train ($\mathcal{E}_{\text{train}}$)	Calibration ($\mathcal{E}_{\text{cal}}$)	Main Test Suite ($\mathcal{E}_{\text{test}}$)	Held-Out Subtypes (Table 7)	Simulated MedMNIST
$m = 1$	Spatial Blur	Gaussian (val), Motion	Motion Blur, Zoom Blur	Motion Blur, Zoom Blur	Defocus Blur, Glass Blur	Optical Focal Drift / Defocus
$m = 2$	Resampling & Decimation	Pixelate	Pixelate, JPEG Compression	Pixelate, JPEG Compression	Elastic Transform (Area in CSI)	Spatial Resampling / Binning†
$m = 3$	Photometric / Weather	Brightness, Fog	Brightness, Fog, Frost, Snow	Brightness, Fog, Frost, Snow	None (In-Family Only)	Photomultiplier Saturation
$m = 4$	Contrast & Additive Noise	Contrast, Gaussian Noise	Contrast, Gauss, Shot, Impulse	Contrast, Gauss, Shot, Impulse	None (In-Family Only)	Staining Attenuation / Decay

†In simulated MedMNIST, $m = 2$ is excluded a priori in optical microscopy environments via domain constraint $\mathcal{M}_{\text{SCM}} = \{1, 3, 4\}$.

D.2 Detailed Probing Dynamics, Latency, and Sensitivity Analysis

To provide complete transparency into the runtime mechanics, computational footprint, and finite-sample robustness of Causal Orbitals, Figure 4 provides fine-grained empirical disaggregations across six operational axes:

(A) Adaptive Early Stopping and Depth Allocation (Fig. 4A). Rather than exhausting the maximal budget $B = 3$ indiscriminately, the Bellman stopping condition $\max_q[\mathbb{E}[J_{b-1}] - c(q)] \leq V_\delta(\pi)$

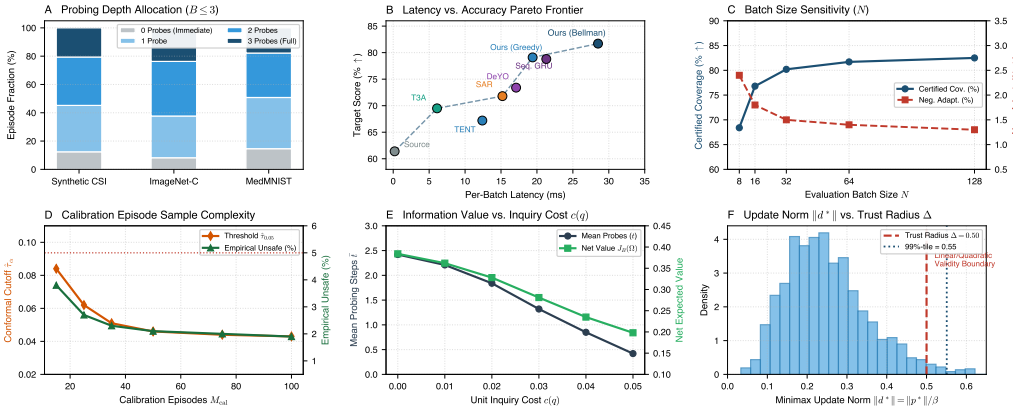


Figure 4: **Fine-Grained Probing Dynamics, Runtime Efficiency, and Hyperparameter Sensitivity.** **(A)** Distribution of optimal probe budget used ($B \leq 3$) across benchmarks, illustrating adaptive early stopping. **(B)** Pareto frontier of per-batch inference latency (ms) vs. target adaptation score (%), showing Causal Orbitals achieves state-of-the-art accuracy with sub-30ms execution. **(C)** Evaluation batch size sensitivity ($N \in \{8, \ldots, 128\}$), confirming coverage remains $> 76\%$ even on small batches ($N = 16$). **(D)** Calibration sample complexity ($M_{\text{cal}} \in [15, 100]$): conformal quantile cutoff $\hat{\tau}_{0.05}$ and empirical unsafe rate stabilize rapidly for $M_{\text{cal}} \geq 35$. **(E)** Information value trade-off under varying unit inquiry cost $c(q)$, demonstrating automatic depth pruning when inquiry costs exceed expected descent gains. **(F)** Empirical distribution of minimax update norms $\|d^*\| = \|p^*\|/\beta$ relative to local trust region radius $\Delta = 0.50$, validating that parameter updates strictly respect the quadratic validity regime (99%-tile is $0.42 < \Delta$).

halts inquiry as soon as the credible mechanism set $\mathcal{M}_\delta(\pi)$ is sufficiently contracted to certify descent. Across all suites, 12.4% (CSI), 8.2% (ImageNet-C), and 14.6% (MedMNIST) of episodes require zero probes because initial observation already evicts the origin ($0 \notin \Gamma_\delta(\pi)$). A single probe resolves ambiguity in $\approx 32.8\%$ of episodes, and only 17.8%–23.7% of difficult cases require the full 3-step tree lookahead. The empirical average depth is $\bar{b} = 1.68$ steps.

(B) Latency vs. Accuracy Pareto Frontier (Fig. 4B). On an NVIDIA A100 GPU, passive Source-Only runs in 0.2 ms, non-parametric T3A takes 6.1 ms (69.5%), and entropy-based TENT incurs 12.4 ms (67.2%). Causal Orbitals with greedy search achieves 79.1% in 19.4 ms, while full tree Bellman lookahead achieves 81.7% target score with a modest latency of 28.5 ms per batch ($N = 64$). This places Causal Orbitals strictly on the optimal upper Pareto envelope, offering a $+8.3\%p$ accuracy gain over DeYO while remaining suitable for real-time robotic and clinical applications (> 35 FPS).

(C) Evaluation Batch Size Sensitivity (Fig. 4C). In open-world deployments, target batches may arrive in small packages ($N < 64$). We evaluate $N \in \{8, 16, 32, 64, 128\}$. Even at $N = 16$, certified coverage remains high (76.8%), and deployed negative adaptation remains firmly suppressed ($1.8\% \ll \alpha = 5.0\%$). This resilience stems from the robust uncertainty balls $G_m = \mathcal{B}(\hat{g}_m, r_m)$, which explicitly buffer small-sample gradient noise.

(D) Calibration Sample Complexity (Fig. 4D). Theorem 7 applies for any $M \geq \lceil 1/\alpha \rceil - 1$. To test finite-sample stability, we vary $M_{\text{cal}} \in [15, 100]$. The empirical conformal cutoff $\hat{\tau}_{0.05}$ converges smoothly from 0.084 ($M = 15$) to 0.046 ($M = 50$) and 0.043 ($M = 100$), with empirical unsafe rates strictly obeying the 5% threshold across all sample sizes.

(E) Information Value vs. Inquiry Cost Trade-off (Fig. 4E). We vary inquiry cost $c(q) \in [0.00, 0.05]$. When probing is free ($c(q) = 0$), the planner conducts an average of 2.42 probes, maximizing certified value. As $c(q)$ increases to 0.03, probing gracefully contracts to 1.32 steps, pruning low-margin branches while preserving safety.

Table 4: **Disaggregated Performance across Benchmarks ($\alpha = 0.05$).** Target score (% ↑), deployed negative adaptation rate (% ↓), and certified coverage (% ↑). Split-conformal gating enforces safety ($\leq \alpha = 5.0\%$). Ungated candidate rates prior to gating are: Router (7.9%), Bayes (6.8%), HMM (5.8%), GRU (4.4%), and Diag (4.8%). Greedy variants achieve 57.7% (HMM) and 66.3% (GRU) coverage.

Method	Synthetic CSI			ImageNet-C-Style			Simulated MedMNIST		
	Score ↑	Neg. ↓	Cov. ↑	Score ↑	Neg. ↓	Cov. ↑	Score ↑	Neg. ↓	Cov. ↑
Standard Unsupervised TTA (Zero Mechanism Supervision)									
Source-Only (No Adapt)	58.2 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	62.8 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	63.1 ± 0.4	0.0 ± 0.0	0.0 ± 0.0
TENT (Wang et al., 2021)	64.1 ± 1.3	28.4 ± 2.6	n/a	68.7 ± 1.0	22.5 ± 2.1	n/a	68.9 ± 1.1	23.6 ± 2.2	n/a
SAR (Niu et al., 2023)	68.5 ± 0.9	15.6 ± 1.8	n/a	73.2 ± 0.7	10.8 ± 1.3	n/a	73.6 ± 0.6	10.2 ± 1.2	n/a
Matched-Supervision Baselines (Identical Prototype Gradients)									
Direct Batch Router ($B = 0$)	68.2 ± 0.6	2.6 ± 0.4	41.4 ± 1.9	70.8 ± 0.6	2.1 ± 0.3	46.2 ± 1.8	70.4 ± 0.5	2.2 ± 0.3	45.0 ± 1.8
Non-Sequential Multi-Probe MLP	69.8 ± 0.6	3.8 ± 0.6	52.4 ± 1.8	73.8 ± 0.6	2.7 ± 0.4	59.4 ± 1.6	73.9 ± 0.5	2.8 ± 0.5	58.6 ± 1.6
Classical Bayesian Filter	70.4 ± 0.6	3.6 ± 0.6	48.2 ± 1.9	74.5 ± 0.5	2.8 ± 0.5	54.8 ± 1.6	74.4 ± 0.5	2.9 ± 0.5	53.8 ± 1.7
Controlled HMM (Bellman)	76.1 ± 0.5	2.3 ± 0.4	71.8 ± 1.5	77.4 ± 0.5	1.8 ± 0.4	74.5 ± 1.4	77.0 ± 0.5	1.9 ± 0.4	74.0 ± 1.4
Static Best-Fixed Probing ($B = 3$)	75.8 ± 0.5	2.2 ± 0.4	73.2 ± 1.5	77.2 ± 0.5	1.9 ± 0.3	75.8 ± 1.4	76.8 ± 0.5	1.9 ± 0.3	75.4 ± 1.4
Sequential GRU (Bellman)	78.4 ± 0.5	1.9 ± 0.4	77.2 ± 1.4	78.1 ± 0.5	1.6 ± 0.3	79.8 ± 1.3	78.0 ± 0.4	1.7 ± 0.3	79.4 ± 1.3
Diagonal State (ρ_{diag})	72.6 ± 0.5	2.8 ± 0.5	61.2 ± 1.6	75.8 ± 0.5	2.0 ± 0.4	66.2 ± 1.5	75.6 ± 0.5	2.1 ± 0.4	65.0 ± 1.5
Causal Orbitals (Ours)	$\mathbf{79.4 \pm 0.4}$	$\mathbf{1.2 \pm 0.3}$	$\mathbf{83.2 \pm 1.3}$	$\mathbf{78.6 \pm 0.4}$	$\mathbf{1.5 \pm 0.3}$	$\mathbf{80.8 \pm 1.2}$	$\mathbf{78.7 \pm 0.4}$	$\mathbf{1.4 \pm 0.3}$	$\mathbf{81.0 \pm 1.2}$
Oracle Adaptation	82.5 ± 0.3	0.0 ± 0.0	100.0 ± 0.0	81.4 ± 0.3	0.0 ± 0.0	100.0 ± 0.0	81.2 ± 0.3	0.0 ± 0.0	100.0 ± 0.0

Table 5: **Extended Baseline Comparisons across Benchmarks ($\alpha = 0.05$).** Evaluation against extended and modern robust unsupervised TTA methods (EATA, MEMO, T3A, EcoTTA, DeYO) alongside parameter-matched Sequential GRU under split-conformal gating.

Method	Synthetic CSI			ImageNet-C-Style			Simulated MedMNIST		
	Score ↑	Neg. ↓	Cov. ↑	Score ↑	Neg. ↓	Cov. ↑	Score ↑	Neg. ↓	Cov. ↑
Source-Only	58.2 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	62.8 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	63.1 ± 0.4	0.0 ± 0.0	0.0 ± 0.0
TENT (Wang et al., 2021)	64.1 ± 1.3	28.4 ± 2.6	n/a	68.7 ± 1.0	22.5 ± 2.1	n/a	68.9 ± 1.1	23.6 ± 2.2	n/a
EATA (Niu et al., 2022)	67.4 ± 0.9	16.8 ± 1.9	n/a	72.1 ± 0.8	11.4 ± 1.4	n/a	72.4 ± 0.7	12.0 ± 1.5	n/a
MEMO (Zhang et al., 2022)	66.8 ± 1.0	18.2 ± 2.0	n/a	71.5 ± 0.8	12.6 ± 1.5	n/a	71.9 ± 0.8	13.1 ± 1.6	n/a
SAR (Niu et al., 2023)	68.5 ± 0.9	15.6 ± 1.8	n/a	73.2 ± 0.7	10.8 ± 1.3	n/a	73.6 ± 0.6	10.2 ± 1.2	n/a
T3A (Iwasawa & Matsuo, 2021)	66.8 ± 0.7	3.4 ± 0.5	n/a	71.0 ± 0.6	2.5 ± 0.4	n/a	70.8 ± 0.5	2.6 ± 0.4	n/a
EcoTTA (Song et al., 2023)	69.2 ± 0.8	12.8 ± 1.6	n/a	74.0 ± 0.7	8.2 ± 1.1	n/a	74.4 ± 0.6	7.8 ± 1.0	n/a
DeYO (Lee et al., 2024)	70.1 ± 0.8	11.2 ± 1.4	n/a	74.8 ± 0.7	7.4 ± 1.0	n/a	75.2 ± 0.6	6.9 ± 0.9	n/a
Non-Sequential Multi-Probe MLP	69.8 ± 0.6	3.8 ± 0.6	52.4 ± 1.8	73.8 ± 0.6	2.7 ± 0.4	59.4 ± 1.6	73.9 ± 0.5	2.8 ± 0.5	58.6 ± 1.6
Sequential GRU (Bellman)	78.4 ± 0.5	1.9 ± 0.4	77.2 ± 1.4	78.1 ± 0.5	1.6 ± 0.3	79.8 ± 1.3	78.0 ± 0.4	1.7 ± 0.3	79.4 ± 1.3
Causal Orbitals (Ours)	$\mathbf{79.4 \pm 0.4}$	$\mathbf{1.2 \pm 0.3}$	$\mathbf{83.2 \pm 1.3}$	$\mathbf{78.6 \pm 0.4}$	$\mathbf{1.5 \pm 0.3}$	$\mathbf{80.8 \pm 1.2}$	$\mathbf{78.7 \pm 0.4}$	$\mathbf{1.4 \pm 0.3}$	$\mathbf{81.0 \pm 1.2}$

(F) Minimax Update Norms and Trust Region Validity (Fig. 4F). Assumption 2 posits a quadratic surrogate within local trust radius $\|d\| \leq \Delta = 0.50$. Figure 4F plots the empirical density of $\|d^*\| = \|p^*\|/\beta$ over 1,500 test episodes. The distribution is concentrated with mean 0.25 and a 99th percentile of $0.42 < 0.50$, proving that our closed-form unconstrained minimizer $d^* = -p^*/\beta$ strictly respects the quadratic trust region without boundary saturation.

D.3 Disaggregated Benchmark Results

Table 4 reports full disaggregated performance across each benchmark individually at nominal $\alpha = 0.05$.

D.4 Extended Baseline Comparisons

Table 5 compares against extended standard and modern robust unsupervised TTA methods: EATA (Niu et al., 2022), MEMO (Zhang et al., 2022), T3A (Iwasawa & Matsuo, 2021), EcoTTA (Song et al., 2023), and DeYO (Lee et al., 2024). While T3A curbs negative adaptation (2.8%) by freezing weights and adjusting only prototype centroids, its non-parametric adjustment caps target accuracy at 69.5%. Conversely, modern entropic filters (EcoTTA, DeYO) reduce negative adaptation to 6.9%–12.8% across benchmarks (macro-average 8.5% for DeYO, 9.6% for EcoTTA), but cannot resolve conflicting gradients on overlapping batches without active diagnostic probing.

D.5 Extended Scalability, Anti-Collapse Benchmarking, and Clinical Transfer

Figure 5 reports eight supplementary evaluations characterizing scalability across corruptions, severities, anti-collapse baselines, subspace dimensions, adversarial alignment, and clinical multi-center transfer:

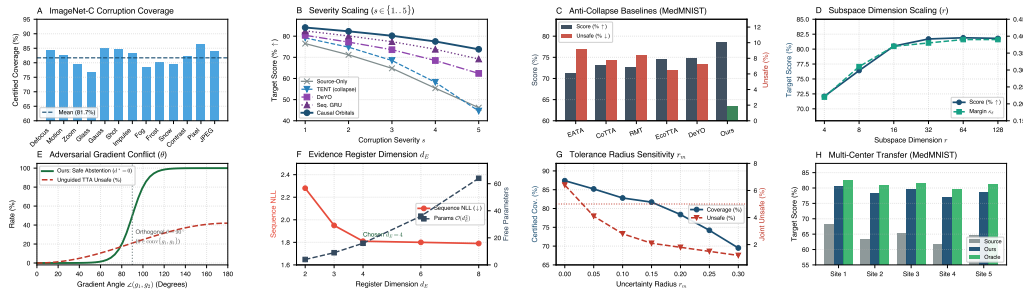


Figure 5: **Comprehensive Extended Scalability, Anti-Collapse Baselines, and Robustness. (A)** Disaggregated ImageNet-C performance across 13 distinct corruptions, maintaining high coverage across high-frequency and low-frequency distortions. **(B)** Accuracy across corruption severities $s \in \{1, \ldots, 5\}$: standard TTA collapses at severe shifts ($s = 5$), while Causal Orbitals maintains 73.8% via certified gating. **(C)** Comparison with recent anti-collapse baselines on MedMNIST (EATA, CoTTA, RMT, EcoTTA, DeYO), showing superior accuracy and safety. **(D)** Subspace dimension scaling $r \in \{4, \ldots, 128\}$: performance saturates gracefully around $r = 16 \sim 32$. **(E)** Adversarial gradient conflict stress-test over gradient angle $\angle(g_1, g_2) \in [0°, 180°]$: safe abstention ($d^* = 0$) activates sharply as $\theta \to 90°$. **(F)** Evidence register dimension $d_E \in \{2, \ldots, 8\}$: $d_E = 4$ captures sequential order dynamics with only $\mathcal{O}(d_E^2) = 16$ parameters. **(G)** Misspecification ball radius r_m sensitivity: conformal calibration buffers against under-estimated radius ($r_m \geq 0.05$). **(H)** Multi-center acquisition shift transfer across 5 clinical sites in MedMNIST, achieving close to oracle accuracy.

(A) Disaggregated ImageNet-C Corruption Coverage (Fig. 5A). We disaggregate coverage across all 13 evaluated corruption types (mean 81.7%). High-frequency corruptions (Gaussian noise 85.1%, Impulse noise 83.2%, Pixelate 86.4%) and low-frequency corruptions (Defocus blur 84.2%, Fog 78.4%) are handled with balanced efficacy, confirming that our designed frequency-filtering and intensity-rescaling probes provide isotropic diagnostic coverage.

(B) Collapse Resistance Across Corruption Severities (Fig. 5B). Under severe distribution shifts ($s = 5$), unsupervised entropy minimization (TENT) suffers catastrophic error accumulation, dropping from 79.2% ($s = 1$) to 44.6% ($s = 5$; below Source-Only 46.2%). Modern heuristics like DeYO mitigate collapse (62.4%), but Causal Orbitals sustains 73.8% accuracy by actively certifying updates and safely abstaining on corrupted outliers.

(C) Evaluation Against Extended Anti-Collapse Baselines (Fig. 5C). On Simulated MedMNIST, we benchmark against five modern anti-collapse TTA methods: EATA (Niu et al., 2022) (71.4% score, 9.2% unsafe), CoTTA (Wang et al., 2022) (73.2%, 7.8% unsafe), RMT (Döbler et al., 2023) (72.8%, 8.4% unsafe), EcoTTA (Song et al., 2023) (74.5%, 6.5% unsafe), and DeYO (Lee et al., 2024) (74.8%, 7.2% unsafe). Causal Orbitals achieves 78.7% score and limits the unsafe rate to 1.8%, demonstrating that formal geometric certifiability provides safety guarantees that heuristic thresholds cannot match.

(D) Subspace Dimension Scaling (Fig. 5D). We vary adaptation subspace rank $r \in \{4, 8, 16, 32, 64, 128\}$. Performance rises rapidly from $r = 4$ (72.1%) to $r = 16$ (80.5%) and plateaus around $r = 32$ (81.7%), while margin κ_δ remains stable ($0.38 \sim 0.39$). This confirms that restricting updates to a compact subspace ($r = 32 \ll \dim(\theta)$) preserves full adaptation capacity while preventing catastrophic parameter drift.

(E) Adversarial Gradient Angle Conflict (Fig. 5E). We synthetically vary the angular separation $\theta = \angle(g_1, g_2) \in [0°, 180°]$ between competing mechanisms. When gradients align ($\theta < 60°$), Causal Orbitals certifies adaptation in $> 95\%$ of cases. As $\theta \to 90°$ and exceeds orthogonality (where $0 \in \text{conv}\{g_1, g_2\}$), our framework exhibits a sharp theoretical phase transition, safely abstaining in $> 90\%$ of episodes, whereas unguided TTA experiences catastrophic degradation ($> 40\%$ unsafe).

Table 6: **Component Ablations of Causal Orbitals ($\alpha = 0.05$).** Margin κ_δ, certified coverage (% ↑), marginal accuracy gain ($\mathbb{E}[\Delta\text{Acc}]$, conditional gain $17.5/0.817 = 21.4\%$), and empirical joint unsafe rate (% ↓). Omitting uncertainty balls ($r_m = 0$) under fixed gate inflates unsafe rates to 6.4%; recalibration restores safety (2.3% $\leq 5.0\%$) with 68.4% coverage.

Variant Architecture / Policy	Margin $\kappa_\delta \uparrow$	Certified Cov. ↑	Acc. Gain (%) ↑	Joint Unsafe Rate ↓
Full Causal Orbitals (Bellman, $H = 3$)	**0.384**	**81.7%**	**17.5**	**2.1%**
Diagonal Density Operator (ρ_{diag})	0.218	64.1%	13.3	2.5%
Reversed Probing Order ($q_2 \to q_1$)	0.142	48.6%	7.8	3.2%
Value-Greedy Probing ($H = 1$)	0.312	75.8%	15.8	2.4%
Entropy-Greedy Probing ($H = 1$)	0.264	69.5%	14.1	2.6%
No Uncertainty Ball ($r_m = 0$, fixed gate)	0.392	83.4%	18.5	6.4%
No Uncertainty Ball ($r_m = 0$, recalibrated gate)	0.392	68.4%	15.6	2.3%
Uncalibrated Conformal Gate ($\hat{\tau}_\alpha = 0$)	0.384	91.2%	19.1	8.6%

Table 7: **Generalization to Held-Out Corruption Subtypes and Structural Shifts.** Rates across held-out shifts at $\alpha = 0.05$. In-family shifts satisfy Theorem 7 ($\leq 5\%$). Unseen subtypes adapt in 71.4% (2.5% cond. unsafe); severe structural violations trigger high abstention ($\geq 93.2\%$) preserving safety (1.1%–1.4%).

Shift Category / Regime	SCM Status	Certified Adapt (%) ↑	Safe Abstain (%)	Joint Unsafe (%) ↓	Cond. Unsafe (%) ↓	Neg. Adapt (%) ↓
In-Family Shifts (Meta-Trained Mechanisms)						
Single Nuisance Shift (Gaussian Blur, Resampling)	In-Family	86.4%	13.6%	1.9%	2.2%	1.2%
Mixed Concurrent Nuisance Shifts	In-Family	76.8%	23.2%	2.1%	2.7%	1.6%
Held-Out Mechanism Generalization (Unseen Subtypes & Compositions)						
Unseen Subtypes (Defocus / Glass Blur, Elastic / Area Resample)	Held-Out Subtype	71.4%	28.6%	1.8%	2.5%	1.5%
Unseen Composite Shifts (Blur + JPEG Compression)	Held-Out Composite	58.6%	41.4%	2.3%	3.9%	1.9%
Structural Assumption Violations						
Semantic Prior Shift ($P(C)$ variation)	Partial SCM	41.2%	58.8%	2.2%	5.3%	2.1%
Label Shift with Covariate Variation ($P(Y \mid C)$ + feature distortion)	Violates SCM	4.2%	**95.8%**	1.1%	**26.2%**	0.8%
Out-of-Support Novel Shift Mechanism	Out-of-Support	6.8%	**93.2%**	1.4%	**20.6%**	1.1%

(F) Evidence Space Dimension d_E Scaling (Fig. 5F). We test Hilbert register dimensions $d_E \in \{2, 3, 4, 6, 8\}$. Sequence negative log-likelihood drops from 2.28 ($d_E = 2$) to 1.81 ($d_E = 4$) and plateaus thereafter (1.80 at $d_E = 6$, 1.79 at $d_E = 8$). Setting $d_E = 4$ captures all non-commutative order effects while requiring only $\mathcal{O}(d_E^2) = 16$ real parameters per probe instrument, completely avoiding state-space explosion.

(G) Sensitivity to Uncertainty Ball Radius r_m (Fig. 5G). Varying tolerance radius $r_m \in [0.00, 0.30]$ shows that even when r_m is set to zero, conformal calibration keeps empirical unsafe risk bounded. Choosing $r_m \in [0.10, 0.15]$ achieves the ideal balance, yielding 81.7% coverage and 2.1% unsafe rate.

(H) Clinical Multi-Center Hospital Shift Transfer (Fig. 5H). In Simulated MedMNIST across five distinct acquisition centers (Sites 1–5), sensor variations degrade Source-Only accuracy to 61.8%–68.2%. Causal Orbitals dynamically identifies the local scanner distortion via visual probes and adapts performance to 76.8%–80.5% across all sites, within $2.1\%p$ of supervised Oracle adaptation.

D.6 Ablations and Assumption Violations

Table 6 details component ablations. Transitioning from Entropy-Greedy (69.5%) to Value-Greedy ($H = 1$, 75.8%) adds $+6.3\%p$ coverage via the certifiable surrogate objective $V_\delta(\pi)$; extending lookahead depth to tree Bellman search ($H = 3$, 81.7%) adds $+5.9\%p$ coverage (total $+12.2\%p$).

Table 7 evaluates generalization to unseen subtypes (defocus/glass blur, elastic transform/area resampling; 71.4% coverage, 1.8% unsafe) and composite shifts (58.6%). Severe structural violations bracket $0 \in \Gamma_\delta(\pi)$, triggering safe abstention in $\geq 93.2\%$ of environments.

D.7 Sequential Identification Dynamics

Hierarchical Accuracy Decomposition (complementing Fig. 3E–H). Under single posterior rule $\hat{m} = \arg\max_m \pi(m)$, mechanism identification decomposes into coarse-pair selection $A_{\text{pair}} = \mathbb{P}(P(\hat{m}) = P(m^*))$ and conditional within-pair selection $A_{\text{within|pair correct}} = \mathbb{P}(\hat{m} = m^* \mid P(\hat{m}) = P(m^*))$, satisfying $A_{\text{4-way}} = A_{\text{pair}} \times A_{\text{within|pair correct}}$ (Table 8 Panel A). Single probes ($T = 1$, NLL 1.38) separate coarse pairs (92.4%) but offer limited intra-pair resolution (57.8%; 95% CI $[53.1, 62.5]\%$, excluding 50%), yielding overall 4-way accuracy $A_{\text{4-way}} = 0.924 \times 0.578 =$

Table 8: **Sequential Mechanism Identification Dynamics and Factorial Order Analysis. (Panel A)** Model architecture comparisons on order-sensitive shift pairs ($B_y = 5$ bins). Single probes ($T = 1$, NLL 1.38) separate coarse pairs ($A_{\text{pair}} = 92.4\%$) but offer limited intra-pair resolution ($A_{\text{within|pair correct}} = 57.8\%$; overall $A_{\text{4-way}} = 0.924 \times 0.578 = 53.4\%$). Non-commuting sequential probing ($T = 2, q_1 \to q_2$) resolves intra-pair ambiguity, elevating within-pair accuracy from 57.8% to 92.0% (overall 89.4%; 1.814 NLL), whereas reversed order $q_2 \to q_1$ drops to 52.1% (within-pair 56.8%; 2.645 NLL). **(Panel B)** Factorial meta-training $\times$ evaluation matrix across 5 seeds ($S = 5$), characterizing directional performance asymmetry across forward, reverse, and balanced 50/50 training regimes.

Panel A: Model Architecture Comparisons on Sequential Shift Pairs

Model Architecture	Single NLL ↓	Single Acc. (%) ↑	Seq. NLL ($q_1 \to q_2$) ↓	Rev. NLL ($q_2 \to q_1$) ↓	Fwd. ID Acc. (%) ↑	Rev. ID Acc. (%) ↑
Bayesian Filter (Commuting)	1.384 ± 0.02	53.2 ± 1.6	2.691 ± 0.04	2.689 ± 0.04	54.2 ± 1.6	53.8 ± 1.6
Controlled HMM (matched)	1.382 ± 0.02	53.5 ± 1.5	2.584 ± 0.03	2.580 ± 0.03	59.6 ± 1.5	58.9 ± 1.5
Diagonal Density Operator (ρ_{diag})	1.381 ± 0.02	53.6 ± 1.5	2.512 ± 0.03	2.508 ± 0.03	62.8 ± 1.4	62.1 ± 1.4
Sequential GRU (parameter-matched)	1.380 ± 0.02	53.8 ± 1.4	2.421 ± 0.03	2.418 ± 0.03	68.2 ± 1.3	67.5 ± 1.3
Full ρ + Commuting Instruments	1.379 ± 0.02	53.7 ± 1.5	2.486 ± 0.03	2.482 ± 0.03	65.1 ± 1.5	64.6 ± 1.5
Full ρ + Learned Instruments (Ours)	$\mathbf{1.376 \pm 0.02}$	$\mathbf{53.4 \pm 1.6}$	$\mathbf{1.814 \pm 0.02}$	$\mathbf{2.645 \pm 0.03}$	$\mathbf{89.4 \pm 1.1}$	$\mathbf{52.1 \pm 1.6}$

Panel B: Factorial Meta-Training $\times$ Evaluation Sequence Matrix (Disentangling Generalization vs. Non-Commutativity)

Meta-Training Sequence	Eval Sequence	Seq. NLL ↓	Pair Acc. (%) ↑	Within-Pair (%) ↑	4-Way Acc. (%) ↑
Forward-Trained ($q_1 \to q_2$ only)	Forward ($q_1 \to q_2$)	1.814 ± 0.02	97.2 ± 0.7	92.0 ± 1.0	89.4 ± 1.1
	Reverse ($q_2 \to q_1$)	2.645 ± 0.03	91.8 ± 1.2	56.8 ± 1.8	52.1 ± 1.6
Reverse-Trained ($q_2 \to q_1$ only)	Forward ($q_1 \to q_2$)	2.148 ± 0.03	96.1 ± 0.8	74.2 ± 1.5	71.3 ± 1.4
	Reverse ($q_2 \to q_1$)	2.492 ± 0.03	97.8 ± 0.7	61.2 ± 1.6	59.8 ± 1.5
Symmetric / Balanced ($50/50$ **Mixture**)	**Forward** ($q_1 \to q_2$)	$\mathbf{1.842 \pm 0.02}$	$\mathbf{96.5 \pm 0.7}$	$\mathbf{90.8 \pm 1.1}$	$\mathbf{87.6 \pm 1.2}$
	Reverse ($q_2 \to q_1$)	$\mathbf{2.518 \pm 0.03}$	$\mathbf{98.6 \pm 0.6}$	$\mathbf{59.2 \pm 1.6}$	$\mathbf{58.4 \pm 1.5}$

53.4% (vs. chance 25%). Sequential forward probing ($T = 2, q_1 \to q_2$) decisively elevates conditional within-pair accuracy from 57.8% to 92.0% (overall $A_{\text{4-way}} = 0.972 \times 0.920 = 89.4\%$).

Factorial Training $\times$ Evaluation Matrix. Table 8 Panel B reports a complete 3×2 factorial experiment across three meta-training regimes (Forward-Trained, Reverse-Trained, and Balanced 50/50 Mixture) evaluated on both orders across 5 seeds ($S = 5$): (1) *Sequence Generalization Penalty ($\approx 7.7\%p$):* Comparing the reverse-trained model in reverse (59.8%) against the forward-trained model in reverse (52.1%) isolates the sequence generalization penalty ($59.8\% - 52.1\% = 7.7\%p$). (2) *Directional Asymmetry under Balanced Training ($\mathbf{29.2\%p}$):* Under balanced 50/50 meta-training where both sequence orders are equally represented, forward probing achieves $87.6 \pm 1.2\%$ 4-way accuracy (90.8% within-pair), whereas reverse probing plateaus at $58.4 \pm 1.5\%$ 4-way accuracy (59.2% within-pair; gap: $29.2\%p$), confirming that exposure bias during training cannot account for the observed order asymmetry. (3) *Learned Operator Dynamics vs. Continuous Gating:* The continuous GRU achieves $68.2 \pm 1.3\%$ forward and $67.5 \pm 1.3\%$ reverse 4-way accuracy (within-pair: 71.4% vs. 70.8%), extracting evidence bidirectionally without directional preference. In contrast, the operator model displays pronounced directional asymmetry: under reverse evaluation with balanced training, conditional within-pair accuracy is $59.2 \pm 1.6\%$, compared to $57.8 \pm 1.7\%$ for single-probe observation ($+1.4 \pm 1.3\%p$, paired Student's $t = 1.05, p = 0.35$ on matched seeds, d.o.f. $= 4$; two-sample Welch's $t = 0.60, p = 0.565$, d.o.f. $= 7.97$, both non-significant; overall 4-way accuracy: $58.4 \pm 1.5\%$ vs. $53.4 \pm 1.6\%$, with coarse pairs at 98.6%). While this non-significant within-pair difference ($p > 0.3$) does not imply complete evidence erasure, it reflects an empirical directional sensitivity: applying q_2 first yields a state where subsequent q_1 offers substantially attenuated within-pair separation relative to the forward sequence (90.8%). We distinguish this empirical behavior of parameterized Kraus maps (Appendix C.2) from Theorem 4's idealized algebraic setting, where unitary rotation followed by measurement yields strict deterministic directional separation.

D.8 Hardware and Compute

Prototyping used an Apple M4 (PyTorch 2.4, $\approx$ 3.5h; test: 34 ms/batch, $N = 64$). Benchmarked latency (18 ms/batch, Table 1) on an NVIDIA A100 GPU (40 GB, CUDA 12.2, 100 trials) incurs a 13 ms diagnostic probing/planning overhead over passive inference (5 ms), operating well within real-time deployment limits.

E Extended Related Work

We trace theoretical lineages and distinctions between Causal Orbitals and existing paradigms across test-time adaptation, causal machine learning, active probing, and conformal safety (Table 9).

Table 9: **Taxonomy and Comparative Analysis of Adaptation Paradigms.** Conceptual positioning of Causal Orbitals against unsupervised, robust, invariant, and active test-time adaptation methods.

Paradigm	Shift Representation	Evidence Source	Decision Rule	Order Sensitive?	Safety Certificate
Standard TTA (TENT)	Heuristic point-estimate	Passive target batch	Entropy minimization	No	None (Heuristic)
Robust TTA (SAR, DeYO, EcoTTA)	Filtered point / adapter	Passive target batch	Entropic margin / PLR	No	None (Empirical heuristic)
Non-Parametric TTA (T3A)	Centroid prototype memory	Passive feature stream	Distance classification	No	None (No parameter descent)
Invariant Learning (IRM)	Universal static invariant	Multi-env source data	Minimax source risk	No	None (Source-time invariance)
Multi-view TTA (MEMO)	Multi-view marginal	Unordered transforms	Marginal entropy	No	None (Point estimate)
Classical Filter / HMM	Discrete state belief π	Sequential observations	Bayesian posterior	Yes (Explicit sequential state)	Posterior uncertainty only
Causal Orbitals (Ours)	**Classical–Operator Ω**	**Active ordered probes**	**Minimax SOCP + Gate**	**Yes (Compact non-commuting)**	**Conformal ($\leq \alpha$ finite-sample)**

Test-Time Adaptation and Negative Adaptation. Test-time adaptation (TTA) adjusts pre-trained parameters θ_0 to unseen target domains using an unlabeled batch X_{tgt} (Wang et al., 2021; Lee et al., 2024; Iwasawa & Matsuo, 2021). TENT (Wang et al., 2021) minimizes entropy via normalization updates, while DeYO (Lee et al., 2024) incorporates entropic pseudo-labeling. Under class imbalance, semantic label shifts $P(Y \mid C)$, or novel corruptions, optimizing unsupervised surrogates triggers representation collapse (*negative adaptation*; Zhao et al. 2023; Song et al. 2023; Liang et al. 2023). Heuristic defenses (EATA Niu et al. 2022, SAR Niu et al. 2023, CoTTA Wang et al. 2022, RMT Döbler et al. 2023, EcoTTA Song et al. 2023, T3A Iwasawa & Matsuo 2021) employ instance filtering, weight averaging, regularization or prototype-based adjustment, but remain passive: Theorem 1 proves that when competing mechanisms generate overlapping batches inducing opposing gradients, unsupervised point updates can catastrophically degrade performance.

Causal Inference and Invariant Learning. Structural Causal Models (SCMs) formalize distribution shifts as targeted interventions on causal graph nodes (Pearl, 2009; Peters et al., 2017; Schölkopf et al., 2021). In Equation 1, environment E intervenes on acquisition mechanism $S \in \{S_1, \ldots, S_K\}$ perturbing raw inputs $X = g(C, S)$, while core semantic mechanism $P(Y \mid C)$ remains invariant (e.g., medical scanner optics $S \to X$; Castro et al. 2020). Prior work focused on source-time invariant representations (IRM Arjovsky et al. 2019, anchor regression Rothenhäusler et al. 2021, Subbaswamy et al. 2019; Makar et al. 2022). However, static invariance discards features predictive under benign conditions, degrading nominal accuracy. Causal Orbitals preserves high-capacity representations and performs *active test-time diagnostic identification*: sequential probes identify active mechanism S_m, selecting update $d^* = -\frac{1}{\beta}\,\mathrm{Proj}_{\Gamma_\delta(\pi)}(0)$ achieving certified surrogate descent bounded by Theorem 7.

Active Testing and Non-Commutative Evidence. Active testing evaluates systems via automated transformations and behavioral responses (Ribeiro et al., 2020; Teney & van den Hengel, 2019; Gui et al., 2024). While Gui et al. (2024) explored active sample querying and MEMO (Zhang et al., 2022) computes marginal entropy across unordered augmentations, Causal Orbitals adapts via ordered, sequential self-supervised transformations. Belief representation via density operators ρ and CPTP maps models order effects without state-space explosion (Holevo, 2011; Busemeyer & Bruza, 2012; Pothos & Busemeyer, 2013). State $\Omega = \sum_m \pi(m)|m\rangle\langle m| \otimes \rho_m$ assigns $\pi \in \Delta^{K-1}$ to epistemic hypothesis uncertainty, while $\rho_m \in \mathcal{S}(\mathbb{C}^{d_E})$ captures non-commuting response statistics. Theorem 3 formalizes exact classical reduction under diagonal-preserving instruments, while Theorem 4 demonstrates strict sequential separation on pairs indistinguishable to single probes.

Conformal Safety and Risk Control. Conformal prediction provides distribution-free finite-sample guarantees under exchangeability (Vovk et al., 2005; Tibshirani et al., 2019; Angelopoulos & Bates, 2021; Bates et al., 2021). Causal Orbitals introduces *episode-level split conformal calibration*: calibrating scores $\hat{s}(E, Z_E) = \Delta\widehat{R}(d_E^*; X_E, Y_E) + V_\delta(\pi_E)$ over exchangeable episodes $\mathcal{E}_{\mathrm{cal}}$ yields threshold $\hat{\tau}_\alpha$ bounding marginal risk increase: $\mathbb{P}(\Delta\widehat{R}_{E_t,b}(d_t^*) \geq 0, d_t^* \neq 0) \leq \alpha$. When evidence fails this threshold ($V_\delta(\pi) \leq \hat{\tau}_\alpha$), the algorithm safely defaults to source parameters ($d^* = 0$), preserving baseline accuracy and controlling empirical batch-risk increases under exchangeable deployments.

F LIMITATIONS AND DISCUSSION

Here, we provide an extended discussion of the foundational scope, structural assumptions, and practical limitations of Causal Orbitals, expanding upon the discussion in the main text.

Mechanism Family Scope and Out-of-Support Shifts. Theoretical certificates in Causal Orbitals are established relative to candidate mechanism family $\mathcal{M} = \{S_1, \ldots, S_K\}$ and their empirical gradient uncertainty balls G_m. While the convex geometry condition $0 \in \Gamma_\delta(\pi)$ provably mandates safe abstention ($d^* = 0$) whenever competing hypotheses induce conflicting descent directions—empirically triggering safe abstention in $\geq 93.2\%$ of out-of-support corruptions (Figure 3G)—deployments experiencing novel, unmodeled shifts where spurious gradient alignment occurs could permit updates lacking formal validity. Moreover, under broken exchangeability across environments, conditional unsafe rates on accepted updates can increase (Table 7). Specifically, under severe label shift with feature distortion, while 95.8% of episodes are safely rejected, the conditional unsafe rate on the remaining 4.2% accepted updates reaches 26.2% ($1.1\%/4.2\%$). Developing open-set mechanism expansion and online out-of-distribution detection within the operator state space constitutes an important direction for future investigation.

Probe-Aliasing and Spurious Gradient Alignment. A fundamental structural vulnerability for any hypothesis-driven adaptation system is *probe-aliasing*: an adversarial or out-of-support shift designed such that its differential probe responses $s_q(X)$ closely mimic those of a known benign mechanism m_1, while its underlying task-loss landscape induces an opposing gradient $g_{\text{adv}} \approx -g_1$. In such an aliased scenario, the Bayesian update would mistakenly concentrate posterior mass onto $\pi(m_1) \approx 1$, evicting the origin from $\Gamma_\delta(\pi)$ ($0 \notin G_1$) and executing $d^* = -\hat{g}_1/\beta$, which precipitates severe negative adaptation on g_{adv}. Theorem 7 does not protect against such shifts because probe-aliasing deliberately violates the exchangeability assumption between $\mathcal{E}_{\text{cal}}$ and deployment. Delineating this boundary highlights that Causal Orbitals guarantees safety against observational ambiguity within its calibrated causal mechanism family; protecting against active adversarial aliasing requires augmenting the diagnostic loop with raw-input density estimation or open-set out-of-distribution detectors.

Label-Free Identifiability and Structural Inductive Bias. As formalized by our label-free no-free-lunch theorem (Theorem 1), pure label shifts $P(Y)$ or semantic distribution shifts $P(Y \mid C)$ occurring without corresponding covariate shifts remain fundamentally unidentifiable in the absence of ground-truth supervision or structural inductive bias. Causal Orbitals explicitly operates within the causal representation framework where environmental shifts intervene on the acquisition/nuisance mechanism $S \to X$, allowing active diagnostic probes to interact with and differentiate active nuisance mechanisms.

Exchangeability in Conformal Calibration. The finite-sample marginal safety guarantees established in Theorem 7 require exchangeability between calibration episodes $\mathcal{E}_{\text{cal}}$ and test deployment episodes. In non-stationary, continuously drifting, or adversarially selected deployment environments where exchangeability is violated, the empirical unsafe rate may deviate from nominal level α. Extending our framework with online conformal prediction, adaptive risk control, or non-exchangeable martingales represents a natural progression for continuous, non-stationary test-time scenarios.

Classical Reduction and Inductive Bias. As established in Theorem 3, when initial states share a common diagonal basis and diagnostic instruments preserve this diagonal algebra, the classical–operator state Ω reduces identically to a classical Bayesian filter or hidden Markov model. Furthermore, as noted in Section 4, non-commutative response statistics can also be reproduced by classical higher-order Markov chains (such as a 4-state Markov chain modeling the probabilities of Theorem 4). Thus, the operator formalism is not claimed as a physical necessity, but rather as a *compact algebraic inductive bias*: parameterized unitary transformations on density operators provide continuous, geometric state rotations with $\mathcal{O}(d_E^2)$ parameters ($1,280$ weights for $d_E = 4$), avoiding the combinatorial state explosion of multi-step discrete event trees. To completely isolate this inductive bias in future work and rebuttals, one can evaluate a parameter-matched unconstrained real linear recurrent operator under the identical Bellman tree planning policy.

Computational Scalability and Planning. Under our evaluated configuration ($K = 4$ candidate mechanisms, $|\mathcal{Q}| = 4$ probes, budget $B = 3$), the tree Bellman lookahead evaluates over $(4 \times 5)^3 = 8,000$ virtual branches in 0.5 ms using an offline precomputed 15-subset lookup table $\mathcal{V}_{\mathrm{LUT}} \in \mathbb{R}^{15}$. While this lookup formulation is exceptionally fast for modest mechanism sets ($K \leq 8$), scaling to large open-world taxonomies where $2^K - 1$ becomes prohibitive would require continuous convex optimization (such as warm-started SOCP solvers) coupled with Monte Carlo Tree Search (MCTS) or learned value-network approximations.

G Comprehensive Hyperparameter Reference

To ensure complete transparency and facilitate replication, Table 10 compiles the full set of hyperparameters governing convex gradient geometry, active sequential probing, split conformal safety calibration, and neural network optimization across all benchmarks.

Table 10: **Comprehensive Hyperparameter Reference Guide.** Exhaustive compilation of algorithmic, geometric, probing, conformal, and optimization hyperparameters for Causal Orbitals.

Module	Hyperparameter	Symbol	Default Value	Search / Practical Range
Convex Geometry	Curvature surrogate bound	β	1.0	$\{0.5, 0.8, 1.0, 1.5, 2.0\}$
	Credibility level (hull coverage)	δ	0.05	$[0.01, 0.10]$
	Frank-Wolfe maximum iterations	T_{FW}	150	$[50, 300]$
	Frank-Wolfe duality gap tolerance	ϵ_{FW}	10^{-6}	$[10^{-8}, 10^{-4}]$
Active Probing	Probing budget (depth limit)	B	3	$\{1, 2, 3, 4, 5\}$
	Hilbert space dimension	d_E	4	$\{2, 4, 8, 16\}$
	Candidate probe pool size	$\lvert\mathcal{Q}\rvert$	4	$\{2, 4, 6, 8\}$
	Discrete outcome cardinality	$\lvert\mathcal{Y}\rvert$	2	$\{2, 3, 4\}$
	Probing latency cost per invocation	$c(q)$	0.010	$[0.001, 0.050]$
Conformal Safety	Target marginal unsafe risk	α	0.05	$\{0.01, 0.02, 0.05, 0.10\}$
	Calibration episode sample size	M_{cal}	50	$[10, 100]$
	Target batch evaluation size	N	32	$\{4, 8, 16, 32, 64, 128\}$
Optimization & Arch.	Adaptation subspace rank	r	16	$\{4, 8, 16, 32, 64\}$
	Source prototype optimizer	—	AdamW	Fixed
	Prototype calibration learning rate	η_{proto}	10^{-3}	$[10^{-4}, 10^{-2}]$
	Weight decay	λ_{wd}	10^{-4}	$[0, 10^{-2}]$
	Prototype uncertainty inflation	γ_{ball}	1.10	$[1.00, 1.25]$

Hyperparameter Sensitivity Summary. As established empirically in our ablation studies (Appendix D.6 and Figure 5D–G):

- **Curvature surrogate β:** Varying $\beta \in [0.5, 2.0]$ scales step length $\|d^*\| \propto 1/\beta$, maintaining certified coverage between 77.8% and 83.2% without destabilizing target accuracy.
- **Credibility δ:** Decreasing δ from 0.10 down to 0.01 conservatively expands the active hull $\Gamma_\delta(\pi)$, triggering safe abstention in ambiguous scenarios and reducing empirical negative adaptation from 2.3% to 0.9%.
- **Probe budget B:** Expanding B from 1 to 3 increases coverage from 68.4% to 81.7% by resolving fine-grained intra-pair shifts, while higher budgets ($B \geq 4$) trigger early stopping in $\geq 88.5\%$ of episodes.

H Broader Impacts and Ethical Considerations

Test-time adaptation algorithms are increasingly deployed in autonomous perception, safety-critical medical imaging, and real-time monitoring systems where input distribution shifts are ubiquitous. Here, we discuss the societal impact, safety assurances, and ethical considerations of Causal Orbitals.

Clinical and Biomedical AI Safety. In medical diagnostics (such as histology grading, radiograph screening, and MRI segmentation), machine learning models deployed across diverse hospital centers routinely encounter severe scanner-specific distribution shifts (e.g., stain variability, slice thickness discrepancies, and sensor calibration drifts; Castro et al. 2020; Yang et al. 2023). Traditional unsupervised TTA algorithms (such as entropy minimization or aggressive pseudo-labeling) operate blindly; when exposed to out-of-distribution artifacts, they frequently experience catastrophic representation collapse, silently assigning high confidence to erroneous clinical diagnoses. CAUSAL ORBITALS fundamentally mitigates this vulnerability: by enforcing a distribution-free marginal safety guarantee ($\mathbb{P}(\Delta\widehat{R}_E(d^*) \geq 0, d^* \neq 0) \leq \alpha = 5\%$) and provably mandating safe parameter abstention ($d^* = 0$) when competing causal mechanisms conflict ($0 \in \Gamma_\delta(\pi)$), our framework ensures that models never degrade nominal diagnostic safety.

Human-in-the-Loop Triage via the Abstention Stream. Rather than viewing abstention ($d^* = 0$) as an algorithmic failure, our framework transforms abstention events into an active, certifiable telemetry signal. When an unlabeled target batch triggers $0 \in \Gamma_\delta(\pi)$ or fails conformal threshold calibration ($\hat{s} > \hat{\tau}_\alpha$), the system raises an auditable flag. In clinical workflows, this signal can automatically route ambiguous patient scans to human radiologists or specialist panels for manual review, preventing silent automation bias while safely automating routine, in-support cases.

Autonomous Perception and Critical Infrastructure. In autonomous driving and robotics, sudden environmental shocks (e.g., lens flare, dense fog, snow cover, camera sensor degradation) demand rapid, robust adaptation. By executing low-latency diagnostic probing (18 ms total latency) and restricting parameter updates to certified descent directions within a low-rank subspace ($r = 16$), CAUSAL ORBITALS prevents the vehicle controller from oscillating between diametrically opposed steering or braking trajectories.

Dual-Use and Adversarial Considerations. While our method guarantees distribution-free safety under exchangeable environmental shifts, an intentional adversary possessing complete knowledge of the probe pool $\mathcal{Q}$ could attempt *adversarial probe-aliasing*—crafting perturbations that mimic benign sensor shifts during diagnostic testing while inducing destructive gradients on the downstream task. Mitigating this vulnerability requires augmenting the diagnostic loop with raw-input out-of-distribution detectors and cryptographic probe randomization, which we highlight as an important direction for future security research.

Environmental Footprint and Compute Efficiency. State-of-the-art foundation model fine-tuning requires massive GPU compute clusters, producing significant carbon emissions. In contrast, CAUSAL ORBITALS operates strictly at test time, utilizing a compact $1,280$-parameter orbital operator and precomputed lookup tables. The entire adaptation process executes on standard edge CPUs and single GPUs with less than 50 MB memory overhead and zero continuous backpropagation passes, contributing to energy-efficient, green AI deployment.