POST-SELECTION-VALID CONFORMAL GUARANTEES OVER CONTINUOUS PREFERENCES FOR MULTI-OBJECTIVE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many multi-objective reinforcement learning (MORL) applications work in a decision-support mode. The system first reveals a set or curve of trade-offs. A user then chooses one operating point for execution. Existing guarantees are usually pointwise: they protect a fixed preference chosen in advance. They do not protect a preference chosen after the full curve has been revealed. This gap is central in continuous-preference MORL, where modern methods are designed exactly to support post-training preference queries. We study this release-then-select problem and propose CERTPREF. CERTPREF publishes a preference-indexed certified lower-bound curve for the realized normalized scalarized return of the policy that is actually executed. It calibrates one-sided conformal bounds at finitely many frozen anchor policies, makes them valid simultaneously, extends them to all preferences with a trajectory-level Lipschitz envelope, and executes the same anchor policy that defines the displayed bound. This yields a distribution-free marginal post-selection guarantee for a single deployment query over a continuous preference space. Experiments show that CERTPREF controls post-selection violations below the target level, often conservatively, while retaining informative lower bounds. Pointwise baselines can appear tighter, but they are not designed to control selection after the curve is revealed.

1 INTRODUCTION

Multi-objective reinforcement learning is often used when the final trade-off is not known in advance. This is the standard *unknown-weights* or *decision-support* setting in MORL (Roijers et al., 2013; Hayes et al., 2022). A typical system learns a set or curve of trade-offs, exposes it to a user, and then executes one operating point selected after inspection. Modern continuous-preference methods make this interface increasingly natural by learning Pareto fronts or preference-conditioned policy families that can be queried after training (Xu et al., 2020; Reymond et al., 2022; Zhu et al., 2023).

This release-inspect-select-execute loop changes the statistical target. A pointwise guarantee protects a preference fixed in advance. It does not protect a preference chosen after the full reported curve has been revealed. Once many displayed lower bounds are visible, a user can choose an optimistic-looking point; marginal validity at each fixed preference need not imply validity for the selected preference. This is the same selection-on-outputs effect studied in adaptive data analysis and selection-aware conformal inference (Dwork et al., 2015; Liang et al., 2026; Hegazy et al., 2025).

Continuous preferences add a second difficulty. The selected preference may lie between all finite calibration points. Thus finite-anchor validity alone is not enough, and a good Pareto-front approximation alone is not a guarantee. We need a deployment-time lower bound that remains valid after release and still describes the rollout of the policy that is actually executed.

We study this gap and propose CERTPREF. The method trains frozen anchor policies, calibrates one-sided conformal score bounds at those anchors, makes the anchor bounds simultaneous, extends them to all preferences with a trajectory-level Lipschitz envelope, and executes the same anchor policy that defines the displayed bound. The output is a preference-indexed lower-bound curve for realized normalized scalarized return. The guarantee is marginal over calibration data, selection,

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.

and the fresh rollout; it is not conditional coverage given the realized curve, not an expected-return confidence interval, and not a guarantee for an arbitrary preference-conditioned policy.

Contributions. We make three contributions. **(1)** We formalize marginal post-selection validity for continuous-preference MORL under a release-then-select deployment protocol. **(2)** We develop CERTPREF, which combines anchor calibration, simultaneous validity, continuous extension, and matched execution into a single distribution-free guarantee for the routed anchor policy. **(3)** We evaluate methods through validity, tightness, and cost, including dense-grid diagnostics and optimistic-display selection tests. Across our benchmarks, CERTPREF controls violations below the target level while retaining useful lower bounds; pointwise baselines are tighter but not selection-safe.

2 RELATED WORK

Decision-support MORL. MORL has long treated unknown weights and decision support as central scenarios (Roijers et al., 2013; Hayes et al., 2022). Recent continuous-preference methods learn dense fronts, adaptation rules, or preference-conditioned policies that can be queried after training (Yang et al., 2019; Xu et al., 2020; Reymond et al., 2022; Kyriakis et al., 2022; Basaklar et al., 2023; Zhu et al., 2023; Cai et al., 2023; Yuan et al., 2025). Their main evaluation targets are front quality, return, hypervolume, or preference generalization. Our target is different: after a trade-off curve and its lower bounds are released, the lower bound selected for deployment should still cover the realized rollout of the executed policy.

Uncertainty and certification in RL. Conformal prediction gives distribution-free finite-sample guarantees under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008; Lei et al., 2018), and related tools have been used for planning, safety filters, off-policy prediction, and off-policy evaluation (Lindemann et al., 2023; Strawn et al., 2023; Sun et al., 2023; Zhang et al., 2023; Foffano et al., 2023; Gan et al., 2025). Safe and robust RL also provide guarantees for fixed objectives, policies, or constraints (Nilim & El Ghaoui, 2005; Wiesemann et al., 2013; Chow et al., 2019; Kumar et al., 2020). These lines usually certify a fixed query or fixed policy. CERTPREF instead certifies a released continuous preference-indexed curve and a matched routed policy chosen after inspection.

Selection-aware inference. Classical multiplicity corrections control finite families (Šidák, 1967; Holm, 1979). Adaptive data analysis and recent selection-aware conformal work show why validity can fail after many outputs are revealed and one is chosen (Dwork et al., 2015; Jin & Candès, 2023; Liang et al., 2026; Hegazy et al., 2025; Timans et al., 2025). Work on selecting or aggregating conformal sets is directly related (Sarkar & Kuchibhotla, 2023; Yang & Kuchibhotla, 2025; Hegazy et al., 2025), but it primarily addresses finite families of already-defined conformal objects. Our setting adds two constraints: the released object is continuous over preferences, and the displayed bound must refer to the policy actually executed. Existing finite-selection tools can potentially serve as an anchor-level simultaneous module, but they do not by themselves supply the continuous-preference Lipschitz extension or the matched policy-routing certificate.

3 PROBLEM SETUP AND DEPLOYMENT SPECIFICATION

This section fixes the deployment object. The system will release a lower-bound curve over preferences. A user will inspect that curve, choose one operating point, and then execute one policy. The guarantee must protect that realized rollout.

3.1 CONTINUOUS-PREFERENCE MORL AND ROLLOUT SCORES

We consider an infinite-horizon discounted multi-objective MDP, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, with discount factor $\gamma \in (0, 1)$ and vector reward $r(s, a) \in [0, 1]^d$. A preference vector w lies in the simplex $\mathcal{W} := \Delta^{d-1} = \{w \in \mathbb{R}^d_{\geq 0} : \sum_{j=1}^d w_j = 1\}$. The preference w specifies the linear trade-off $w^\top r$ between objectives. In this paper, w is a deployment-time query: the system may learn from many trade-offs during development, but at deployment it must answer one queried preference at a time. The bounded-reward assumption is a normalization that makes constants explicit; bounded objectives can be clipped or rescaled before certification.

For a rollout trajectory τ, let $G(\tau) \coloneqq \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$ be its discounted vector return. For a policy π, the expected scalarized return is $J^\pi(w) \coloneqq \mathbb{E}[w^\top G(\tau)]$, and the optimal value is $V^*(w) \coloneqq \sup_\pi J^\pi(w)$. These are standard MORL quantities. Our certification target is more specific. We do not certify only an expectation such as $J^\pi(w)$. We certify the single rollout that will actually be executed after the user chooses a preference. We use the negative realized scalarized return $C(\tau; w) \coloneqq -w^\top G(\tau)$ as the score. Smaller values of $C(\tau; w)$ are better, so an upper bound on this score is exactly a lower bound on realized scalarized return.

3.2 Published lower-bound curve and executable policy

At query time the system must provide two linked objects. The first is a *published lower-bound curve* $L : \mathcal{W} \to \mathbb{R}$. Its value $L(w)$ is shown to the user and is intended as a lower bound on realized scalarized return at preference w. The second is an *executable policy map* $\hat{\pi} : \mathcal{W} \to \Pi$, where Π denotes the policy class. Its value $\hat{\pi}(w)$ is the policy that will actually be run if preference w is chosen. A meaningful deployment guarantee must connect these two objects. The displayed bound should describe the rollout of the policy that is actually executed.

It is convenient to use the equivalent upper-bound notation $U(w) \coloneqq -L(w)$, because split conformal naturally constructs upper bounds for the rollout score $C(\tau; w)$. If τ_{new} is a fresh rollout generated by executing $\hat{\pi}(w)$, then failure means $C(\tau_{\mathrm{new}}; w) > U(w)$, equivalently $w^\top G(\tau_{\mathrm{new}}) < L(w)$. In words, the realized scalarized return falls below the published lower bound. This is the event whose probability we want to control.

3.3 Release-then-select deployment and the target guarantee

We study a single-query deployment loop. During development, the system trains a policy library and uses separate calibration rollouts to construct a published curve $U(\cdot)$. At deployment, that full curve is revealed. After seeing it, a user or higher-level controller chooses one preference $\widehat{w} \in \mathcal{W}$. The system then executes one fresh rollout under $\hat{\pi}(\widehat{w})$. This is the decision-support workflow behind continuous-preference MORL: reveal many trade-offs first, then choose one operating point for one execution.

A pointwise statement for each fixed preference does not solve this protocol. If a preference w were fixed in advance, one could ask for $\mathbb{P}(C(\tau_{\mathrm{new}}; w) \le U(w)) \ge 1 - \alpha$. But here the selector sees the full curve before choosing. That makes it possible to search over many displayed values and pick an optimistic-looking point. Appendix B.0 gives a finite-menu counterexample where every displayed point is pointwise valid, but the selected miscoverage approaches one as the menu size grows. Continuous preferences add a second difficulty: the selected preference may lie between any finite set of calibration anchors. We therefore need a guarantee that survives both selection after inspection and off-anchor queries.

Formally, let $\widehat{w}$ be any deterministic or randomized selector that is measurable with respect to the released pre-deployment information, including $U(\cdot)$, but not the fresh deployment rollout. Let τ_{new} denote the fresh rollout produced by the executed policy $\hat{\pi}(\widehat{w})$. Our target guarantee is

$$\mathbb{P}\big(C(\tau_{\mathrm{new}}; \widehat{w}) \le U(\widehat{w})\big) \ge 1 - \alpha. \tag{1}$$

The probability is over training and calibration randomness, the fresh rollout, and any randomization used by the selector. Equation (1) protects the realized deployment rollout. It is not only a statement about the expectation $J^\pi(w)$. The next section constructs such a pair $(U, \hat{\pi})$ from finitely many calibrated anchor policies.

4 CertPref: Constructing a Deployable Lower-Bound Curve

The construction mirrors the deployment loop in Section 3. A user will not choose a preference before seeing the released curve, and the chosen preference need not be one of the preferences used for calibration. The method therefore has four requirements. It must obtain reliable bounds for frozen anchor policies, make those bounds safe to release as a family, transfer them to off-anchor preferences, and ensure that the displayed bound describes the same policy that is executed.

4.1 ANCHOR LIBRARY AND CALIBRATION DATA

We first build a finite policy library. Let $S = \{w_1, \ldots, w_m\} \subset \mathcal{W}$ be anchor preferences, where m is the anchor budget and $[m] = \{1, \ldots, m\}$. For each anchor w_i, a trainer produces a policy π_i, and the policy is then frozen. The trainer can be any MORL or offline RL method; the validity argument only uses rollouts from the frozen policies. This separation is useful in practice. Better training improves the lower bounds, but it is not needed for distribution-free coverage.

The anchor design also defines the finite family that conformal calibration must protect. A dense grid is the natural finite-set alternative, but its Bonferroni cost grows with display resolution. We therefore calibrate a smaller anchor set and use continuity, rather than calibration at every displayed point, to cover off-anchor preferences. Appendix A.12 gives the finite-sample arithmetic for the dense-grid baseline.

Each anchor has its own calibration set $\mathcal{D}_i^{\mathrm{cal}} = \{\tau_1^{(i)}, \ldots, \tau_n^{(i)}\}$ collected by executing π_i. This policy-specific calibration is important. If the deployment system later runs π_i, then the calibration rollouts for π_i are the relevant comparison sample. The required stochastic condition is the standard split-conformal one applied separately at each anchor.

Assumption 4.1 (Anchor-wise exchangeability)**.** For each anchor $i \in [m]$, conditional on the trained policy π_i, the calibration trajectories collected under π_i and any fresh deployment rollout obtained by executing the same π_i are exchangeable.

For anchor i, define calibration scores $R_j^{(i)} = C(\tau_j^{(i)}; w_i)$ and choose a per-anchor failure level $\alpha_i \in (0, 1)$. Split conformal sets $U(w_i) = R_{(k_i)}^{(i)}$ with $k_i = \lceil (n+1)(1-\alpha_i) \rceil$, where $R_{(k)}^{(i)}$ is the k-th smallest score and $R_{(n+1)}^{(i)} = +\infty$. The intuition is a rank argument. A fresh score from the same frozen policy is exchangeable with the n calibration scores, so its rank among the $n+1$ values is uniform. It exceeds the selected quantile with probability at most α_i. Thus $U(w_i)$ is a distribution-free upper bound on the realized score at that anchor.

4.2 SIMULTANEOUS RELEASE OVER THE ANCHOR SET

A single anchor bound is not enough because the user sees a family of displayed values. If every anchor were calibrated at level α independently and the user could choose after release, one bad anchor could be selected. CERTPREF therefore allocates the total failure budget across anchors with $\sum_i \alpha_i \le \alpha$. The main experiments use the transparent Bonferroni allocation $\alpha_i = \alpha/m$, while Appendix A.6 reports allocation scans that trade tightness across anchors without changing the validity logic. On the simultaneous anchor event, every calibrated anchor bound holds at once. This event is the selection-safe finite skeleton from which the continuous curve is built.

4.3 CONTINUOUS EXTENSION AND MATCHED EXECUTION

The selected preference may lie between anchors. To answer such a query, an anchor bound must be transferred from w_i to a nearby w for the same rollout. This is why the regularity condition is stated for the realized score $C(\tau; w)$, not only for the expected value $J^\pi(w)$.

Assumption 4.2 (Preference-Lipschitz rollout score)**.** There exists $L_C < \infty$ such that $|C(\tau; w) - C(\tau; w')| \le L_C \|w - w'\|_1$ for all trajectories τ and preferences $w, w' \in \mathcal{W}$.

For the main score $C(\tau; w) = -w^\top G(\tau)$, this is the trajectory-level counterpart of the linear scalarization used throughout MORL. Since $G(\tau) \in [0, 1/(1-\gamma)]^d$ and w, w' lie in the simplex, one may take $L_C = 1/(2(1-\gamma))$ under ℓ_1. Appendix C.1 gives the proof and worst-case witnesses. The assumption is not appropriate for discontinuous indicator scores or hard threshold events unless a separate smoothing or relaxation is introduced.

Given a router $g : \mathcal{W} \to [m]$, the anchor bound for $w_{g(w)}$ extends to query w as

$$U_g(w) = U(w_{g(w)}) + L_C \left\| w - w_{g(w)} \right\|_1 . \tag{2}$$

The second term is the price of answering an off-anchor query. It is small when the routed anchor is close to the chosen preference, and it is the main source of conservatism in the published curve. If the

Algorithm 1 CERTPREF for one release-then-select deployment query

Inputs: preference space $\mathcal{W}$, total level α, anchor budget m, calibration size n, score C, certified constant L_C, trainer $\mathcal{T}$, anchor-design rule.
Choose anchors $S = \{w_i\}_{i=1}^m \subset \mathcal{W}$ and per-anchor levels α_i with $\sum_i \alpha_i \le \alpha$.
for $i = 1, \dots, m$ **do**
  Train and freeze anchor policy $\pi_i \leftarrow \mathcal{T}(w_i)$.
  Collect held-out calibration rollouts $\mathcal{D}_i^{\mathrm{cal}} = \{\tau_j^{(i)}\}_{j=1}^n$ by executing frozen π_i.
  Compute anchor scores $R_j^{(i)} \leftarrow C(\tau_j^{(i)}; w_i)$.
  Set $k_i \leftarrow \lceil (n+1)(1-\alpha_i) \rceil$ and $U(w_i) \leftarrow R_{(k_i)}^{(i)}$, with $R_{(n+1)}^{(i)} = +\infty$.
end for
Define $U_{\min}(w) \leftarrow \min_i \{U(w_i) + L_C \|w - w_i\|_1\}$. Release $L(w) \leftarrow -U_{\min}(w)$, or $L_+(w) \leftarrow \max\{L(w), 0\}$ when nonnegative support clipping is used.
Define matched router $g(w) \in \arg\min_i \{U(w_i) + L_C \|w - w_i\|_1\}$ and executable policy map $\hat{\pi}(w) \leftarrow \pi_{g(w)}$.
Release: publish $L(\cdot)$ and the matched execution rule $\hat{\pi}(\cdot)$; the selector may inspect released artifacts but not the fresh rollout.
Deploy: after the selector chooses $\widehat{w}$, report $L(\widehat{w})$, execute one fresh rollout of $\hat{\pi}(\widehat{w})$, and evaluate the realized score $C(\tau_{\mathrm{new}}; \widehat{w})$.
Validity contract: under Assumptions 4.1–4.2, Eq. (1) holds for this single deployment query.

system is free to choose which anchor certifies the query, the tightest anchor-induced upper bound is

$$U_{\min}(w) = \min_{i \in [m]} \Big(U(w_i) + L_C \|w - w_i\|_1 \Big), \qquad L(w) = -U_{\min}(w). \tag{3}$$

This envelope should be read as a lower-bound curve after changing signs: larger $L(w)$ means a more useful guarantee on realized scalarized return. The envelope is not a certificate for an implicit continuously conditioned policy. It is the lower envelope of finitely many valid certificate-policy pairs, and validity requires executing the anchor policy that induces the displayed bound.

When normalized rewards are nonnegative, there is also a deterministic support tightening. Since $G(\tau) \ge 0$ componentwise and $w \in \Delta^{d-1}$, every realized score satisfies $C(\tau; w) = -w^\top G(\tau) \le 0$. Thus the system may publish

$$U_+(w) = \min\{U_{\min}(w), 0\}, \qquad L_+(w) = -U_+(w) = \max\{L(w), 0\}, \tag{4}$$

without spending any failure budget. This weakly improves the lower-bound curve over the trivial nonnegative lower bound.

Proposition 4.3 (Support clipping)**.** *Under the nonnegative normalized-reward assumption, replacing $U_{\min}$ by U_+ in* (4) *preserves the post-selection guarantee in Theorem 5.1.*

Proof. On the event proved for $U_{\min}$, $C(\tau_{\mathrm{new}}; \widehat{w}) \le U_{\min}(\widehat{w})$. Deterministically, $C(\tau_{\mathrm{new}}; \widehat{w}) \le 0$. Hence $C(\tau_{\mathrm{new}}; \widehat{w}) \le \min\{U_{\min}(\widehat{w}), 0\} = U_+(\widehat{w})$ on the same event. □

The displayed curve must also determine the policy that is run. We use the following matched certificate-policy principle: if the displayed upper bound at w is induced by anchor i, then the executed policy must be π_i. For the envelope in (3), this gives the certificate-aware router $g_{\text{U-smart}}(w) \in \arg\min_{i \in [m]} (U(w_i) + L_C \|w - w_i\|_1)$ and the executable policy $\hat{\pi}(w) = \pi_{g_{\text{U-smart}}(w)}$. A nearest-anchor router is also valid if the system displays its own bound $U_g(w)$, but it can be looser because it ignores anchor-specific conformal quantiles. A performance-smart router can choose anchors by a calibration-return proxy, which may improve realized return but need not minimize the displayed bound. Appendix A.5 compares these routers. The guarantee does not automatically transfer to an arbitrary preference-conditioned policy queried at w; certifying such a hybrid would require an additional calibrated comparison between the conditioned policy and the routed anchor. Algorithm 1 summarizes the deployable construction.

4.4 OPTIONAL MODULES AND THEIR FUNCTIONAL ROLE

The core construction above is deliberately minimal. None of the modules below is required for Theorem 5.1; they act on tightness, cost, or population-level robustness.

Anchor selection. The default choice is a simple lattice or k-center cover of the simplex. This is transparent and works well for $d = 2$, where the preference line is easy to cover. In higher dimensions or under non-uniform query populations, the same budget can be used more carefully. GreedyMI treats the unknown tightness or slack field over preferences as a Gaussian-process surrogate and greedily chooses anchors with high expected information gain. This changes only the anchor set S; once the anchors are fixed, the same calibration and validity theorem apply.

Local Lipschitz calibration. The global constant L_C is safe but conservative. The Conformal-Lipschitz extension uses an additional calibration split to certify local trajectory-level moduli $\widehat{L}_i$ around anchors, then replaces the global penalty in (2) by a local one when the local certificate is tighter. This spends a separate failure budget β and is therefore a tightness module, not part of the minimal theorem.

Population-level robustness. Wasserstein robustness addresses a different question. It does not change the single selected-query guarantee in Theorem 5.1. Instead, it asks how the average published lower bound changes when the distribution of queried preferences shifts away from a nominal population. Experiments and details for these modules appear in Appendix A.6–Appendix A.10, with proofs and derivations in Appendix B, Appendix D, and Appendix E.

5 Theory: Main Guarantee and Design Implications

This section states the end-to-end guarantee and explains the proof structure. The individual tools are familiar: split conformal calibration controls a fixed exchangeable score, finite-family correction controls simultaneous errors, and Lipschitz continuity transfers a bound between nearby preferences. The contribution is the way these tools must be composed for the release-inspect-choose-execute loop in continuous-preference MORL. Detailed lemmas and proofs are in Appendix B and Appendix C; those appendices also record related conformal-risk-control variants, beyond-exchangeability variants, and norm choices that are not used by the main theorem.

5.1 Post-selection validity of the released curve

The next theorem is the deployment statement: after the curve is released and one preference is selected, the lower bound still applies to the rollout that is actually executed.

Theorem 5.1 (Post-selection validity for single-query deployment)**.** *Under Assumptions 4.1 and 4.2, construct the released curve and matched executable policy as in Section 4, with per-anchor levels $\{\alpha_i\}_{i=1}^m$ satisfying $\sum_i \alpha_i \le \alpha$. Then for any deterministic or randomized selector $\widehat{w}$ that is measurable with respect to the released pre-deployment information, including $U(\cdot)$ but excluding the fresh deployment rollout, the target guarantee in Eq.* (1) *holds, where τ_{new} is the fresh rollout generated by executing $\hat{\pi}(\widehat{w})$.*

The theorem allows the selector to inspect the entire released curve before choosing $\widehat{w}$. It does not require the selected preference to be fixed in advance, and it does not assume a parametric model for returns. The guarantee is marginal over calibration data, any randomness in the selector, and the fresh rollout; it is not a conditional coverage statement given the realized published curve. After changing signs, the statement says that the published lower bound $L(\widehat{w})$ is valid for the realized scalarized return of the routed anchor policy that is actually executed.

The proof has three transparent events. For the proof sketch, imagine drawing one independent fresh rollout for each frozen anchor policy; this is an analysis coupling, not an extra deployment requirement. The actual system executes only the selected routed policy. For each anchor, split conformal gives $\mathbb{P}(C(\tau_i^{\mathrm{new}}; w_i) > U(w_i)) \le \alpha_i$ for a fresh rollout of π_i. The allocation $\sum_i \alpha_i \le \alpha$ implies that all anchor bounds hold simultaneously with probability at least $1 - \alpha$. On that event, for any preference w routed to anchor i, Assumption 4.2 gives $C(\tau_i^{\mathrm{new}}; w) \le C(\tau_i^{\mathrm{new}}; w_i) + L_C \left\| w - w_i \right\|_1$, and the anchor event then gives $C(\tau_i^{\mathrm{new}}; w) \le U(w_i) + L_C \left\| w - w_i \right\|_1$. Matched routing makes this upper bound the one displayed for the policy being executed. This is why the guarantee survives selection after release.

5.2 THEORY ROLES OF THE APPENDIX RESULTS

The appendix separates validity ingredients from design ingredients. Appendix B proves the conformal and extension chain, including anchor validity, simultaneous validity, Lipschitz transfer, envelope validity, and the optional Conformal-Lipschitz variant. Appendix C studies geometry and tightness: the exact trajectory-level Lipschitz constant, matching worst-case examples, cover-radius scaling, norm-choice variants, and a regret decomposition.

These facts clarify the theory. The mathematical ingredients are standard, but the deployment guarantee requires their specific composition with matched certificate-policy routing. The same analysis also characterizes the geometric price of certifying a continuous preference space from finitely many anchors: extension slack is controlled by the cover radius and is unavoidable in worst case. The base validity theorem does not require optimal anchors, GreedyMI, Conformal-Lipschitz, support clipping, or Wasserstein robustness; those are design tools for usefulness, cost, or robustness.

5.3 TIGHTNESS, COVERAGE, AND DESIGN LEVERS

A valid lower-bound curve can still be unhelpful if it is too conservative. In CERTPREF, conservatism comes from two sources. The first is the calibrated anchor value $U(w_i)$, which depends on the quality and variability of the anchor policy. The second is the off-anchor extension penalty. For an anchor set S, define $r(S) \coloneqq \sup_{w \in \mathcal{W}} \min_{i \in [m]} \|w - w_i\|_1$, the largest distance from a query preference to its nearest anchor. The worst-case off-anchor penalty is at most $L_C r(S)$. Thus more anchors and better anchor placement can tighten the curve, but only at increased offline cost. A practical rule is to pick m so that $r(S) \le \varepsilon_{\text{ext}}/L_C$ for a desired extension slack ε_{ext}. In worst case, the minimax cover radius on the $(d-1)$-simplex scales as $m^{-1/(d-1)}$, so higher-dimensional preference spaces quickly become costly.

This also explains why CERTPREF separates calibration from geometric coverage. The number of conformal families scales with the anchor count m, while continuous coverage is paid through the term $L_C r(S)$. A dense-grid alternative would instead make the conformal family size scale with the display resolution; Appendix A.12 gives the corresponding finite-sample baseline. The consequences of over- or under-estimating L_C are recorded in Appendix C.5.

The optional modules target these same terms. Allocation frontiers change the per-anchor quantiles $U(w_i)$ while keeping $\sum_i \alpha_i \le \alpha$. GreedyMI and anchor scans aim to improve the anchor set S under a training budget. Conformal-Lipschitz reduces the effective Lipschitz constant in the extension penalty, at the price of an extra failure budget β. Wasserstein robustness uses a Lipschitz constant for the released curve to lower-bound its average value under preference-distribution shift. The empirical section reports validity, lower-bound tightness, and offline cost together because these quantities are coupled by the theory.

6 EXPERIMENTS

The experiments are organized around the deployment question in Section 3. The user first sees the released lower-bound curve and then chooses a preference. We therefore report not only coverage, but also how useful the lower bounds remain after this selection.

6.1 PROTOCOL, METRICS, AND BASELINES

We evaluate on the two-objective D4MORL MuJoCo tasks (Zhu et al., 2023). Preferences lie on Δ^1 and are parameterized as $w(t) = (t, 1-t)$. Unless stated otherwise, the main Table 1 runs use per-anchor SAC with $10{,}000$ environment steps per anchor, k-center farthest-point anchors from a 101-point simplex grid, $m = 16$, $n_{\text{cal}} = 200$ calibration rollouts per anchor, $n_{\text{test}} = 500$ fresh test rollouts, target level $\alpha = 0.1$, and a dense grid of 401 reported preferences, denoted by $\mathcal{W}_{\text{grid}}$. Objective rewards are clipped and rescaled to $[0, 1]$ by a deterministic map fixed before conformal calibration; calibration and test rollouts are not used to choose clipping thresholds (details and the full protocol table are in Appendix A.0). Results are mean $\pm$ standard error over five seeds.

Our evaluation follows the standard conformal-prediction principle of reporting both validity and efficiency (Lei et al., 2018; Angelopoulos et al., 2024). In our one-sided setting, let $\tau_j(w)$ be the j-th

fresh test rollout generated by the policy routed to w. We report the worst grid violation

$$\widehat{\mathrm{viol}}_{\max} = \max_{w \in \mathcal{W}_{\mathrm{grid}}} \frac{1}{n_{\mathrm{test}}} \sum_{j=1}^{n_{\mathrm{test}}} \mathbf{1}\{w^\top G(\tau_j(w)) < L(w)\},$$

a 95% simultaneous Clopper–Pearson upper confidence bound over the 401 grid queries, and an own-selected violation at $\widehat{w}_{\mathrm{sel}} \in \arg\max_{w \in \mathcal{W}_{\mathrm{grid}}} L(w)$ for each method. Tightness is summarized by $\overline{L} = |\mathcal{W}_{\mathrm{grid}}|^{-1} \sum_{w \in \mathcal{W}_{\mathrm{grid}}} L(w)$ and by the support-clipped mean $\overline{L}_+ = |\mathcal{W}_{\mathrm{grid}}|^{-1} \sum_w \max\{L(w), 0\}$. Larger L means a less conservative and more useful lower bound. A method is useful only if it has low violation and nontrivial lower bounds.

We compare with two baselines. **Pointwise** builds a split-conformal bound at level α for each fixed grid preference. It is valid for a pre-registered preference, but it does not protect a preference chosen after seeing the full curve. **No correction** keeps the same anchor-to-extension pipeline as CERTPREF, but calibrates each anchor at level α instead of allocating the error budget across anchors. This isolates the role of simultaneous control. Dense-grid Bonferroni is the natural finite-grid competitor, but Appendix A.12 shows that it is vacuous in our main calibration regime.

6.2 UNIFORM VALIDITY AND LOWER-BOUND QUALITY

Table 1 is the main cross-environment audit among non-vacuous methods. Pointwise bounds have larger $\overline{L}$ because they spend the full error budget at each displayed preference. That tightness is not deployment-valid. The diagnostic $\widehat{\mathrm{viol}}_{\max}$ is a stress test over the released evaluation grid; the own-selected column is closer to the formal single-query protocol because each method is evaluated at the preference selected by its own lower-bound rule. Across all tasks, CERTPREF remains below $\alpha = 0.1$ even under the simultaneous binomial UCB, while Pointwise exceeds the target under the grid diagnostic and is often near or above the target at its own selected query. In these D4MORL runs all raw lower bounds are positive, so support clipping leaves $\overline{L}$ unchanged; Appendix A.9 reports a case where clipping matters.

Table 1: Enhanced cross-environment audit. The UCB is Bonferroni-adjusted over 401 queries; own-sel. viol. evaluates each method at $\arg\max_w L_{\mathrm{method}}(w)$; $L_+ = \max\{L, 0\}$.

Env	Method	$\widehat{\mathrm{viol}}_{\max}$	95% sim. UCB	own-sel. viol.	$\overline{L}$	$\overline{L}_+$
HalfCheetah	CERTPREF	0.016±0.004	0.047±0.006	0.010±0.004	43.884±3.262	43.884±3.262
HalfCheetah	Pointwise	0.168±0.009	0.236±0.010	0.119±0.017	63.522±0.902	63.522±0.902
Ant	CERTPREF	0.015±0.003	0.046±0.004	0.011±0.003	40.514±0.253	40.514±0.253
Ant	Pointwise	0.150±0.009	0.216±0.010	0.099±0.007	44.482±0.194	44.482±0.194
Swimmer	CERTPREF	0.021±0.003	0.055±0.005	0.006±0.002	49.901±0.023	49.901±0.023
Swimmer	Pointwise	0.143±0.005	0.208±0.006	0.092±0.008	51.693±0.002	51.693±0.002
Walker2d	CERTPREF	0.018±0.004	0.050±0.006	0.007±0.003	23.370±0.107	23.370±0.107
Walker2d	Pointwise	0.166±0.008	0.234±0.009	0.102±0.011	25.337±0.063	25.337±0.063
Hopper	CERTPREF	0.019±0.002	0.053±0.003	0.003±0.001	15.062±0.087	15.062±0.087
Hopper	Pointwise	0.151±0.008	0.217±0.010	0.114±0.009	16.748±0.125	16.748±0.125

6.3 DEPLOYMENT STRESS TEST: OPTIMISTIC-DISPLAY SELECTION

The previous table includes a method-own selected-query audit. We also run a more targeted optimistic-display stress test. For each seed, a selector observes the Pointwise lower-bound curve, selects $w^\star \in \arg\max_w L_{\mathrm{Pointwise}}(w)$, and we evaluate both methods at that same preference. This isolates the risk that a user may choose the operating point that looks best after seeing a non-simultaneous display.

Figure 1 summarizes this optimistic-display selection. Pointwise is not designed to control post-selection risk: in this diagnostic it is often near or above the target at the selected preference, with the clearest failures on HalfCheetah and Hopper. CERTPREF remains below the target in the same test while retaining useful lower bounds on most tasks. HalfCheetah is the main case where validity costs more tightness, consistent with larger anchor-to-query slack; Appendix A.1 and Appendix A.2 analyze this mechanism.

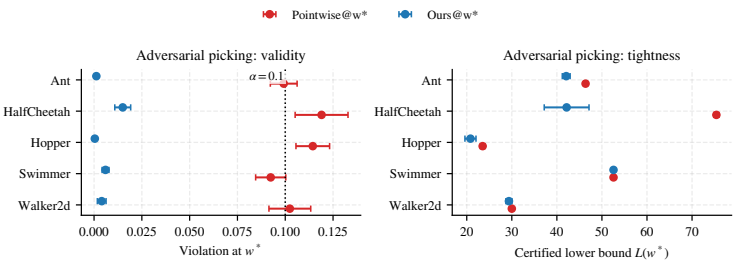


Figure 1: Multi-environment optimistic-display summary. Left: violation at selected $w^\star$. Right: certified lower bound at the same preference.

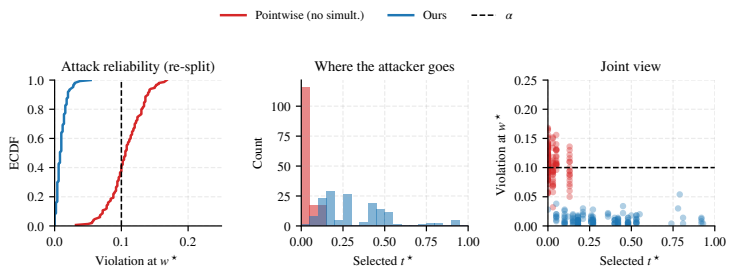


Figure 2: Optimistic-display selection statistics on `MO-HalfCheetah-v2`. Left: violation distribution at selected $w^\star$. Middle: selected locations. Right: selected location versus violation. Dashed line: $\alpha = 0.1$.

Figure 2 focuses on `MO-HalfCheetah-v2` over random calibration/test re-splits. Pointwise often exceeds α after optimistic-display selection, and the selected locations vary across the preference line. This shows that the failure is not a single outlier preference.

6.4 Mechanisms and Cost

The mechanism ablations in Appendix A.1 separate the main failure modes: Pointwise and No correction are tight but unsafe after selection, No extension misses off-anchor queries, and matched routing is needed so the displayed bound describes the executed policy. Router, local-Lipschitz, anchor-budget, allocation, and GreedyMI studies appear in Appendix A.5–Appendix A.7, Appendix A.2, and Appendix E. The $d = 3$ audit in Appendix A.9 is a geometry/nontriviality check: worst-grid violation stays far below $\alpha = 0.1$, and the positive-certificate fraction rises from about 1% at $m = 4$ to 5.3% at $m = 200$, while raw average bounds remain negative. On `MO-HalfCheetah-v2`, pre-deployment cost is about 5.0, 10.7, and 22.7 ± 1.9 minutes for $m = 4, 8, 16$; per-query routing is negligible. Additional attack, trainer, preference-shift, and algorithm details appear in Appendix A.3–Appendix A.4 and Appendix A.8–Appendix A.11.

Limitations and Future Work. CertPref gives a marginal single-query guarantee for the routed anchor policy specified by the displayed certificate. It is not conditional coverage, not an expected-return confidence interval, and not a guarantee for arbitrary preference-conditioned policies unless separately certified. Repeated adaptive queries need new error allocation, fresh calibration data, or reusable-holdout machinery. Tightness is limited by trajectory-level Lipschitzness, the conservative global constant $L_C = 1/(2(1-\gamma))$, and simplex cover radius; high-dimensional preference spaces therefore require larger anchor budgets and careful anchor placement. Preference-shift robustness in Appendix D does not cover environment shift, which would break anchor-wise exchangeability. These restrictions are deliberate: the certificate is intended to match the release-then-select deployment object, not every use of a preference-conditioned agent. Future work should reduce conservatism through local or data-adaptive Lipschitz envelopes, adaptive covers near high-value or frequently selected regions, and allocation rules that spend less error budget on inactive anchors while retaining simultaneous validity. Another direction is to certify policy families directly by learning the router, bound, and executable policy with an explicit certificate-policy consistency check. Repeated interactions and broader deployment shift will require reusable calibration data, pre-registered query budgets, or new exchangeability/robustness assumptions rather than relying on the single-query guarantee proved here.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and the development of experimental code. AI tools also assisted with LaTeX formatting and compilation checks. The authors reviewed the AI-assisted text and tested and checked the experimental code. The authors retain responsibility for the final manuscript, experimental code, and reported results, including material prepared with AI assistance.

REPRODUCIBILITY STATEMENT

The problem formulation and deployment protocol are specified in Section 3, and the certification procedure is described in Section 4 and Algorithm 1. Section 6 and Appendix A describe the experimental setup, including benchmark environments, reward preprocessing, calibration and test sample sizes, and evaluation metrics. The mathematical assumptions and derivations supporting the guarantees are given in Section 5 and Appendix B–Appendix E.

ETHICS STATEMENT

This work studies statistical certification for multi-objective reinforcement learning using simulation benchmarks. Its marginal, single-query guarantee should not be interpreted as an unrestricted safety guarantee for a deployed autonomous system. Use in real-world decision support requires checking the stated assumptions, ensuring that the executed policy matches its displayed certificate, and assessing risks specific to the application. Repeated adaptive use and changes in the deployment environment require additional justification beyond the guarantee studied here.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://doi.org/10.1214/23-AOS2276.

Toygun Basaklar, Suat Gumussoy, and Umit Y. Ogras. PD-MORL: Preference-driven multi-objective reinforcement learning algorithm. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zS9sRyaPFlJ.

Xin-Qiang Cai, Pushi Zhang, Li Zhao, Jiang Bian, Masashi Sugiyama, and Ashley Llorens. Distributional Pareto-optimal multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15593–15613, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/32285dd184dbfc33cb2d1f0db53c23c5-Abstract-Conference.html.

Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Daniele Foffano, Alessio Russo, and Alexandre Proutière. Conformal off-policy evaluation in markov decision processes. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 3087–3094. IEEE, 2023.

Feichen Gan, Youcun Lu, Yingying Zhang, and Yukun Liu. Conformal prediction beyond the horizon: Distribution-free inference for policy evaluation. *arXiv preprint arXiv:2510.26026*, 2025.

Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

Mahmoud Hegazy, Liviu Aolaritei, Michael I. Jordan, and Aymeric Dieuleveut. Valid selection among conformal sets. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=S1Gxt8c8pC.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023. URL https://jmlr.org/papers/v24/22-1176.html.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Panagiotis Kyriakis, Jyotirmoy Deshmukh, and Paul Bogdan. Pareto policy adaptation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=wfZGut6e09.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after data-dependent model selection. *Journal of the American Statistical Association*, pp. 1–26, 2026. doi: 10.1080/01621459.2026.2663588. URL https://doi.org/10.1080/01621459.2026.2663588.

Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J. Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 8(8): 5116–5123, 2023. doi: 10.1109/LRA.2023.3292071.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018. doi: 10.1007/s10107-017-1172-1.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto conditioned networks. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2022.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A Survey of Multi-Objective Sequential Decision-Making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013. doi: 10.1613/jair.3987.

Siddhaarth Sarkar and Arun Kumar Kuchibhotla. Post-selection inference for conformal prediction: Trading off coverage for precision. *arXiv preprint arXiv:2304.06158*, 2023. URL https://arxiv.org/abs/2304.06158.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Kegan J. Strawn, Nora Ayanian, and Lars Lindemann. Conformal predictive safety filter for RL controllers in dynamic environments. *IEEE Robotics and Automation Letters*, 8(11):7833–7840, 2023. doi: 10.1109/LRA.2023.3322644.

Jiankai Sun, Yiqi Jiang, Jianing Qiu, Parth Nobel, Mykel J. Kochenderfer, and Mac Schwager. Conformal prediction for uncertainty-aware planning with diffusion dynamics model. In *Advances in Neural Information Processing Systems*, 2023. URL https://dblp.org/rec/conf/nips/SunJQNKS23.html.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, Christian A. Naesseth, and Eric Nalisnick. Max-Rank: Efficient multiple testing for conformal prediction. *Proceedings of Machine Learning Research(PMLR)*, 258:3898–3906, 2025. URL https://proceedings.mlr.press/v258/timans25a.html.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2008.

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

Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, 120(549):435–447, 2025. URL https://arxiv.org/abs/2104.13871.

Yifu Yuan, Zhenrui Zheng, Zibin Dong, and Jianye Hao. MODULI: Unlocking preference generalization via diffusion models for offline multi-objective reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73743–73766. PMLR, 2025. URL https://proceedings.mlr.press/v267/yuan25n.html.

Hao Zeng, Kangdao Liu, Bingyi Jing, and Hongxin Wei. Parametric scaling law of tuning bias in conformal prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74133–74156. PMLR, 2025. URL https://proceedings.mlr.press/v267/zeng25e.html.

Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. *Proceedings of Machine Learning Research (PMLR)*, 206:2751–2768, 2023.

Baiting Zhu, Meihua Dang, and Aditya Grover. Scaling Pareto-efficient decision making via offline multi-objective reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2023.

APPENDIX

APPENDIX CONTENTS

Symbols: ⋆ original construction, theorem, or primary empirical finding; † standard/adapted mathematical tool; ◇ protocol, diagnostic, ablation, or implementation detail.

◇ A. Additional experiments and reproducibility details . . . 16
◇ A.0 Common protocol, notation, metrics, preprocessing, and dense-grid Bonferroni arithmetic . . . 16
⋆ A.1 Mechanism breakdown: simultaneous control, extension, and matched routing . . . 18
⋆ A.2 Validity–tightness–cost trade-off summary . . . 19
⋆ A.3 Optimistic-display preferences across environments . . . 19
⋆ A.4 Repeated-split optimistic-display statistics . . . 20
◇ A.5 Router ablations: nearest-neighbor, U-smart, and performance-smart routing . . . 20
◇ A.6 Simultaneous-control alternatives and allocation . . . 21
⋆ A.7 Certified versus empirical Lipschitz constants . . . 23
◇ A.8 Offline and conditional trainer comparison . . . 24
⋆ A.9 $d = 3$ preference-simplex nontriviality audit . . . 25
◇ A.10 Preference-shift robustness experiments . . . 26
◇ A.11 Algorithm summary . . . 27
† A.12 Dense-grid Bonferroni finite-sample baseline . . . 28
⋆ B. Conformal prediction certificates and post-selection validity proofs . . . 28
† B.0 Pointwise failure example, exchangeability, and ties . . . 28
† B.1 Split conformal upper bounds . . . 29
⋆ B.2 Simultaneous anchor validity . . . 29
⋆ B.3 Lipschitz transfer, envelope validity, and post-selection proof . . . 30
◇ B.4 What the score certifies . . . 32
† B.5 Conformal risk control context . . . 32
† B.6 Beyond-exchangeability variants . . . 32
⋆ C. Lipschitz geometry, tightness, and regret bounds . . . 32
† C.0 Simplex scalarization identity . . . 32
⋆ C.1 Trajectory-level Lipschitz constant and worst-case witnesses . . . 32
† C.2 Fixed-policy Lipschitzness . . . 33
† C.3 Optimal-value Lipschitzness . . . 33
⋆ C.4 Uniform regret under an ε-cover . . . 34
⋆ C.5 Conformal-Lipschitz refinement and Lipschitz-constant sensitivity . . . 34
† C.6 Norm choice and dual norms . . . 34
† C.7 Covering the simplex under ℓ_1 . . . 35
† D. Wasserstein preference-shift robustness . . . 36
† D.0 Notation . . . 36
† D.1 Lipschitz constants for certificate curves . . . 37
◇ D.2 Applying Wasserstein bounds to preference shift . . . 37
† D.3 Wasserstein DRO dual form . . . 37
† E. Gaussian-process mutual information for anchor selection . . . 37
† E.0 GP-MI setup . . . 37
† E.1 Greedy approximation proof . . . 38

QUICK-REFERENCE TABLES (POSITIONING, NOTATION, AND RESULTS)

The next three tables provide (i) a high-level comparison to common baselines, (ii) a symbol glossary, and (iii) an at-a-glance map of the main guarantees and where they are proved.

Table 2: Positioning relative to representative baselines. ✓ indicates the feature is guaranteed; ✓* indicates it holds only on a pre-specified finite set (e.g., a grid or anchors) and does not certify the full continuum.

	Pointwise conformal	Grid + Bonferroni	Anchor-only (no extension)	This work (CERTPREF)
Multi-objective MORL ($d > 1$)	✓	✓	✓	✓
Distribution-free finite-sample certificate	✓	✓	✓	✓
Valid under adaptive preference choice after seeing the curve	×	✓*	✓*	✓
Uniform over a *continuous* preference set $\mathcal{W}$	×	×	×	✓
Avoids correcting over a dense grid of N preferences	×	×	✓	✓
Explicit tightness lever via coverage radius $r(S)$	×	×	✓*	✓
Certificate-aware routing for the tightest bound	×	×	×	✓

Table 3: Key notation (abridged; see the main paper and Appendix C for details).

Symbol	Meaning
d; γ (and T)	Number of objectives; discount factor (and finite episode horizon when applicable).
$w \in \mathcal{W}$	Preference vector (typically $\mathcal{W} = \Delta^{d-1}$); in $d = 2$ we often write $w(t) = (t, 1-t)$.
$S = \{w_i\}_{i=1}^m$; m	Anchor set and anchor budget (number of trained/certified policies).
π_i; $\hat{\pi}(\cdot)$	Anchor policy trained at w_i; deployed policy map obtained by routing $w \mapsto \pi_{g(w)}$.
$C(\tau; w)$	Certified score on a rollout/trajectory τ at preference w (in experiments, negative scalarized return).
$U(w_i)$; $U(w)$	Split conformal upper bound at anchor w_i; extended certificate at arbitrary w via Lipschitz extension.
$L(w) = -U(w)$	Certified *lower* bound on scalarized return (larger is tighter/better).
L_C; $\ell(\tau)$	Global Lipschitz constant for $C(\tau; \cdot)$; optional trajectory-level modulus used by Conformal-Lipschitz.
$r(S)$	Cover radius $r(S) = \sup_{w \in \mathcal{W}} \min_i \|w - w_i\|_1$ controlling worst-case extension slack.
n_{cal}; n_{test}	Calibration and test episodes per anchor used in experiments.

Table 4: Main results overview (simplified).

Result	Guarantee (simplified)	Where
Preference Lipschitzness (fixed π and optimal value)	$\lvert J^\pi(w) - J^\pi(w')\rvert$, $\lvert V^*(w) - V^*(w')\rvert \le \frac{1}{2(1-\gamma)} \lVert w - w' \rVert_1$ (tight).	Lemmas C.3–C.4; Appendix C.
Regret decomposition under an ε-cover	$V^*(w) - J^{\pi_{g_{\text{nn}}(w)}}(w) \le \max_i \varepsilon_{\text{train}}(i) + \varepsilon/(1-\gamma)$ (tight coefficient).	Thm. C.5; Appendix C.4.
Simultaneous split conformal on anchors	$\mathbb{P}(\forall i,\ C(\tau_{\text{new}}^{(i)}; w_i) \le U(w_i)) \ge 1 - \alpha$ for $\sum_i \alpha_i \le \alpha$.	Thm. B.2; Appendix B.
Lipschitz extension (uniform over $\mathcal{W}$)	On the anchor event, $C(\tau_{\text{new}}^{(g(w))}; w) \le U(w)$ for all $w \in \mathcal{W}$.	Thm. B.3; Appendix B.
Post-selection validity (single query)	For any selector $\widehat{w}$ measurable w.r.t. released pre-deployment information: $\mathbb{P}(C(\tau_{\text{new}}; \widehat{w}) \le U(\widehat{w})) \ge 1 - \alpha$.	Thm. 5.1; Appendix B.
Optional: Conformal-Lipschitz tightening	Using certified local $\widehat{L}_i$ yields failure at most $\alpha + \beta$ and can tighten extension.	Thm. B.5; Appendix B.
Optional: preference-shift robustness	$\inf_{\rho: W_1(\rho, \rho_0) \le \eta} \mathbb{E}_\rho[h] \ge \mathbb{E}_{\rho_0}[h] - \eta L_h$ for L_h-Lipschitz h.	Section 5; Appendix D.

APPENDIX A: ADDITIONAL EXPERIMENTS AND REPRODUCIBILITY DETAILS

This appendix supplements the experiments in Section 6. It gives the full protocol and additional plots/tables that are mentioned in the main text but not expanded there. Across all experiments, our guarantee is expressed as an upper bound $U(w)$ on a score $C(\tau; w)$ computed from a rollout τ and a preference w. Since we use the *negative-return* score in the experiments, $L(w) := -U(w)$ can be interpreted as a *certified lower bound* on the scalarized return at preference w.

A.0 COMMON PROTOCOL, NOTATION, AND METRICS

Default main experimental protocol. Table 5 records the exact settings used for the main D4MORL Table 1 runs. The online run metadata stored in some artifacts uses a smaller plotting grid during collection; all paper tables and curves reported in the main text use the 401-preference reporting grid below.

Table 5: Default protocol for the main D4MORL experiments. Trainer plug-in experiments, anchor-budget sweeps, and robustness experiments override the relevant row when stated in their own setup paragraphs.

Component	Default setting
Environments	D4MORL two-objective MuJoCo tasks: `MO-HalfCheetah-v2`, `MO-Ant-v2`, `MO-Swimmer-v2`, `MO-Walker2d-v2`, `MO-Hopper-v2`.
Trainer	Per-anchor SAC using Stable-Baselines3 for Table 1; trainer plug-in comparisons are reported separately.
Training steps	10,000 SAC environment steps per anchor in the stored Table 1 runs.
Anchor set	k-center farthest-point anchors selected from a 101-point 2D simplex candidate grid.
Anchor budget	$m = 16$.
Calibration/test episodes	$n_{\text{cal}} = 200$ calibration vector-return episodes and $n_{\text{test}} = 500$ test episodes per anchor.
Evaluation grid	401 dense preferences for reported curves and tables.
Horizon and discount	Gym task horizon $T = 500$, discount $\gamma = 0.99$.
Reward transform	Fixed deterministic per-objective clip-rescale from $[-5, 5]$ to $[0, 1]$ before calibration and testing.
Router	Certificate-aware U-smart routing for the main CERTPREF row; nearest-neighbor routing is an ablation.
Simultaneous control	Bonferroni α/m by default; Šidák is included as a stronger-assumption appendix check.
Violation confidence	95% Clopper–Pearson upper confidence bounds with Bonferroni allocation over the 401 reported queries.
Seeds	0–4.

Preference space and grids. For the main MuJoCo experiments with two objectives ($d = 2$), preferences lie on the 1-simplex Δ^1 and we use the standard parameterization $w(t) = (t, 1 - t)$ with $t \in [0, 1]$. Curves and worst-case-over-preferences quantities are approximated on a dense uniform grid of t values. The main tables use $N = 401$ grid points; repeated-split diagnostics sometimes use $N = 101$ to reduce evaluation cost.

Anchors, anchor policies, and post-training calibration. We select m anchor preferences $\{w_i\}_{i=1}^m$ and train one policy per anchor (or equivalently obtain an anchor-conditioned policy evaluated at w_i). After training, each anchor policy π_i is frozen and fresh rollouts are collected for calibration and testing. This train-then-calibrate separation is essential: once π_i is fixed, calibration and test rollouts for anchor i are exchangeable, matching Assumption 4.1. For conditional trainers in Appendix A.8, we instantiate frozen anchor policies by conditioning the trained network at the anchor preferences; deployment still routes to one of those frozen anchor instantiations.

Score, certificates, support clipping, and extension. Let $G(\tau) \in \mathbb{R}^d$ denote the discounted vector return of rollout τ. We use the score $C(\tau; w) = -w^\top G(\tau)$, so that $C(\tau_{\text{new}}; w) \le U(w)$ is

equivalent to the lower-bound statement $w^\top G(\tau_{\text{new}}) \geq L(w) = -U(w)$. For uniform coverage over a continuous preference space, we calibrate anchor thresholds $U(w_i)$, apply simultaneous error control across anchors, and extend to arbitrary w by

$$U(w) = U(w_{g(w)}) + L_C \|w - w_{g(w)}\|_1,$$

where $g(w)$ is the router paired with the executable policy. When nonnegative normalized rewards are used, we also report the support-clipped lower bound $L_+(w) = \max\{L(w), 0\}$ from Proposition 4.3.

Validity metrics (smaller is better). For any w on the evaluation grid, we compute the empirical violation rate on the test episodes routed to w,

$$\widehat{\text{viol}}(w) = \frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} \mathbf{1}\{C(\tau_j; w) > U(w)\}.$$

We summarize dense-grid diagnostics by $\widehat{\text{viol}}_{\max} = \max_w \widehat{\text{viol}}(w)$. For the strengthened main table we also report a 95% simultaneous upper confidence bound obtained by computing query-wise Clopper–Pearson binomial upper bounds and Bonferroni-allocating confidence across the 401 reported grid queries. The own-selected violation chooses $\widehat{w}_{\text{sel}} \in \arg\max_w L(w)$ under the same method's released lower-bound curve and reports $\widehat{\text{viol}}(\widehat{w}_{\text{sel}})$.

Tightness metrics (larger is better). Tightness is reported using the certified lower bound. We use the grid-average $\overline{L} = \mathbb{E}_{w\sim\mathcal{W}_{\text{grid}}}[L(w)]$, the support-clipped grid-average $\overline{L}_+ = \mathbb{E}_{w\sim\mathcal{W}_{\text{grid}}}[\max\{L(w), 0\}]$, and sometimes the grid-minimum $\min_w L(w)$ to capture worst-preference tightness. $\overline{L}_+$ is the average improvement over the deterministic trivial lower bound 0 in normalized nonnegative-return settings.

Reporting and audit trail. Unless noted otherwise, results report mean $\pm$ standard error over 5 random seeds (0–4). Each plotted point or curve is computed from cached per-run artifacts (anchor rollouts and metadata). For per-figure provenance, metric definitions, and plotting details, see the included `FIGURE_ATLAS.md`.

Reward normalization (theory alignment). Our theoretical setup assumes per-objective rewards in $[0, 1]$ (Section 3 and Appendix C). For MuJoCo benchmarks, raw reward components can have different scales and occasional outliers. We therefore apply deterministic per-objective clipping and affine rescaling,

$$\tilde{r}_j = \frac{\text{clip}(r_j, a_j, b_j) - a_j}{b_j - a_j} \in [0, 1].$$

For the current main experimental artifact, the deterministic map is the fixed clip-rescale from $[-5, 5]$ to $[0, 1]$. Earlier artifacts also record reference-dataset percentile constants; in all reported certification runs, the normalization constants are fixed before conformal calibration and are not chosen using calibration or test episodes.

Code release and compute resources. An anonymized code package accompanies the submission and will be released publicly upon acceptance. The package contains the certification implementation, training/evaluation entry points, plotting scripts, environment setup instructions, configuration files, and instructions for obtaining the public benchmark environments; it does not redistribute third-party benchmark datasets or model weights. The final reproducibility runs were executed on AMD EPYC 9754 CPU resources and NVIDIA RTX 4090 GPU resources with 48 GB total GPU memory. The main D4MORL inventory consists of 25 environment–seed runs, corresponding to 400 per-anchor SAC policies, 4.0M SAC environment steps, and 280k calibration/test rollout episodes. The logged SAC training time for these D4MORL runs is 9.62 hours. The reported three-objective highway audit consists of 40 seed–configuration summaries across the $m \in \{4, 7, 10, 15, 20, 30, 100, 200\}$ settings, corresponding to 4.67M calibration/test rollout episodes. Policy training and rollout simulation dominate the cost; conformal quantile computation, Lipschitz-envelope construction, plotting, and per-query routing are negligible by comparison.

A.1 MECHANISM BREAKDOWN

We isolate the contribution of key components in CERTPREF (simultaneous control, Lipschitz extension, and routing) on a representative environment, `MO-HalfCheetah-v2`. The goal is to separate *why* certificates can look tight from *whether* they remain valid under worst-case preference selection.

Setup. We use $m = 16$ anchors, $n_{\text{cal}} = 200$ calibration episodes per anchor, $n_{\text{test}} = 500$ test episodes per anchor, and target miscoverage $\alpha = 0.1$.

Ablations. We compare CERTPREF against variants that disable individual ingredients:

- **Pointwise:** for each preference w on the grid, calibrate a separate split-conformal threshold at miscoverage α for the score $C(\tau; w)$ using calibration rollouts from the nearest-anchor policy $\pi_{g_{\text{nn}}(w)}$ (tight for a fixed w, but does not control the "multiple testing" effect of selecting w after seeing certificates).
- **No correction:** keep the anchor $\rightarrow$ extension form but remove family-wise error control across anchors (optimistic thresholds that can fail under preference picking).
- **No extension:** apply simultaneous correction at anchors but omit the Lipschitz slack when extending off-anchor (ignores the continuous nature of $\mathcal{W}$).
- **CERTPREF (NN) vs CERTPREF (U_{smart}):** both use simultaneous correction and Lipschitz extension; they differ only in the router $g(w)$ (geometric nearest-neighbor vs minimizing the certificate upper bound).

Reading the figure. Figure 3 summarizes (left) worst-case violation, including a conservative 95% UCB; (middle) mean tightness $\overline{L}$; and (right) a representative violation curve over preferences.

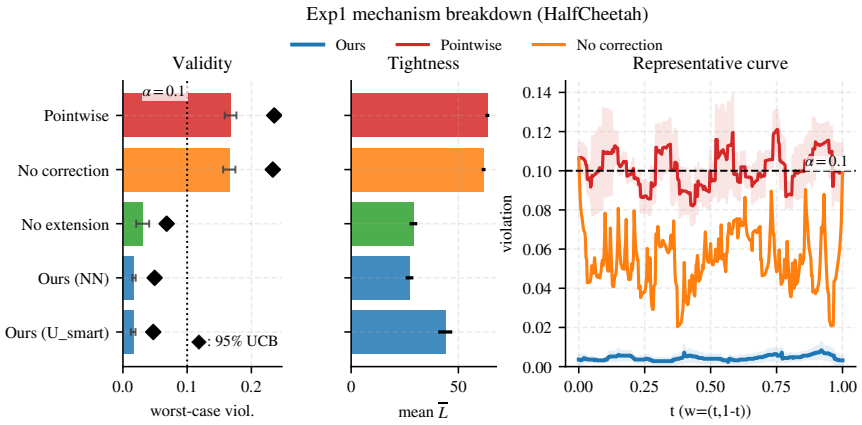


Figure 3: Mechanism breakdown on `MO-HalfCheetah-v2`. Left: worst-case violation across preferences; diamonds show a simultaneous 95% UCB and the vertical line marks $\alpha = 0.1$. Middle: average certified lower bound. Right: a representative violation curve over preferences.

Additional interpretation. The left panel highlights the core post-selection issue: **Pointwise** can look well-calibrated for any fixed, pre-registered preference, yet its *worst-case* violation rises above α once a selector (or user) is allowed to pick w after inspecting the full curve. The ablations isolate *which* ingredient prevents this failure: (i) **No correction** violates because it does not enforce a simultaneous guarantee across anchors, so one of the anchor thresholds can be overly optimistic; (ii) **No extension** can appear valid at anchors but breaks off-anchor because it ignores the continuous dependence on w. The right panel makes this concrete by showing that pointwise violations concentrate around the optimistic-display region, whereas CERTPREF remains below the target level, conservatively, across the grid. Finally, comparing CERTPREF (NN) and CERTPREF (U_{smart}) shows that routing primarily affects *tightness*: the certificate-aware router chooses the anchor that minimizes the bound, improving $\overline{L}$ without sacrificing validity.

A.2 TRADE-OFF SUMMARY (PARETO/BUBBLE VIEW)

We study design trade-offs as we vary the number of anchors m and the anchor selection strategy. Beyond validity and tightness, we explicitly track training cost, since increasing m (or using more sophisticated anchor selection) can improve coverage geometry but also increases compute.

Setup. We summarize an m-scan on `MO-HalfCheetah-v2` with $m \in \{2, \ldots, 16\}$, $n_{\text{cal}} = 200$, $n_{\text{test}} = 500$, and $\alpha = 0.1$. The displayed trade-off is evaluated using the certificate-driven router (U_{smart}).

Reading the figure. In Figure 4, the x-axis is a conservative validity proxy (worst-case violation UCB normalized by α; smaller is better), the y-axis is mean tightness $\overline{L}$ (larger is better), and bubble size encodes training time (minutes; square-root scaled). The black curve highlights the global Pareto frontier across all scanned settings.

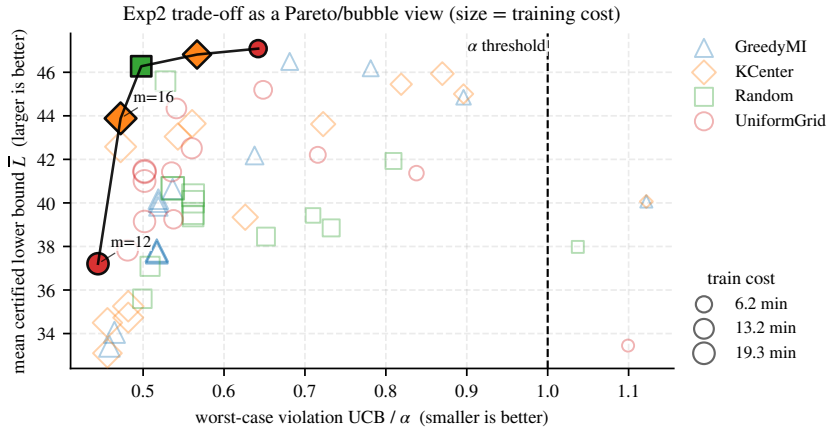


Figure 4: Trade-off as a Pareto/bubble plot. x-axis is the worst-case violation UCB normalized by α (smaller is better); y-axis is the mean certified lower bound (larger is better); bubble size encodes training cost (minutes; square-root scaled). The global Pareto frontier is highlighted in black.

Additional interpretation. Moving left in the plot corresponds to stronger (more conservative) validity: settings with poor coverage geometry (too few anchors or poorly placed anchors) tend to sit to the right because the simultaneous UCB on worst-case violation exceeds α. Moving up corresponds to tighter certificates on average. The Pareto frontier therefore highlights the "sweet spots" where additional anchors materially tighten certificates without incurring a large validity penalty. Bubble size makes the hidden cost explicit: high-m settings can be attractive statistically yet much more expensive to train. In practice, we use this view to pick a budget where improvements in $\overline{L}$ begin to saturate (diminishing returns), and then focus on anchor selection strategies that reduce $r(S)$ (and hence extension slack) at fixed m.

A.3 OPTIMISTIC-DISPLAY PREFERENCE PICKING ACROSS ENVIRONMENTS

We stress-test deployment under *adaptive* preference selection. Concretely, a selector first examines pointwise certificates over the continuous preference grid and then selects the preference $w^\star$ that makes the pointwise method look maximally optimistic (largest certified lower bound). We then evaluate both pointwise and CERTPREF at this optimistic-display $w^\star$.

Setup. We repeat this attack across five D4MORL MuJoCo tasks (Zhu et al., 2023): `MO-HalfCheetah-v2`, `MO-Ant-v2`, `MO-Swimmer-v2`, `MO-Walker2d-v2`, and `MO-Hopper-v2`. Unless noted otherwise we use $m = 16$ anchors, $n_{\text{cal}} = 200$, $n_{\text{test}} = 500$, $\alpha = 0.1$, and a dense preference grid ($N = 401$). CERTPREF uses simultaneous correction across anchors and Lipschitz extension, evaluated with the certificate-driven router (U_{smart}).

Reading the figure/table. The multi-environment attack plot is reported in the main text as Figure 1. For reproducibility and exact comparison, Table 6 gives the corresponding numerical summary.

Table 6: Optimistic-display evaluation. The preference $w^\star$ is chosen after observing the Pointwise lower-bound curve. CERTPREF reduces violation at the same selected preference while preserving useful lower bounds on most tasks.

Env	Pointwise $\widehat{\mathrm{viol}}(w^\star)$	CERTPREF $\widehat{\mathrm{viol}}(w^\star)$	Pointwise $L(w^\star)$	CERTPREF $L(w^\star)$
MO-Ant-v2	0.099±0.007	0.001±0.001	46.4±0.4	42.1±0.9
MO-HalfCheetah-v2	0.119±0.014	0.015±0.004	75.4±0.6	42.2±5.0
MO-Hopper-v2	0.114±0.009	0.000±0.000	23.5±0.5	20.8±1.2
MO-Swimmer-v2	0.092±0.008	0.006±0.002	52.6±0.0	52.6±0.0
MO-Walker2d-v2	0.102±0.011	0.004±0.002	30.0±0.2	29.4±0.7

Additional interpretation. Across environments, the optimistic-display selector identifies preferences where **Pointwise** is not protected against post-selection risk. Its empirical violation is near or above α in several tasks, with the clearest failures on HalfCheetah and Hopper, even though pointwise coverage would be valid for any *fixed* pre-chosen preference. This is exactly the "selection-on-outputs" failure mode our single-query protocol is designed to prevent. In contrast, CERTPREF stays below the target at the optimistic-display $w^\star$ because the published curve is constructed from a simultaneous anchor event plus a deterministic Lipschitz extension, so the selector cannot exploit one overly optimistic point without violating the global event. On the tightness axis (right panel), pointwise can appear tighter precisely because the selector targets locations where pointwise thresholds happen to be loose; CERTPREF trades a modest amount of conservatism for a guarantee that survives this optimistic-display selection.

A.4 OPTIMISTIC-DISPLAY STATISTICS UNDER REPEATED RANDOM SPLITS

To verify that the optimistic-display preference-picking effect is not an artifact of a single calibration/test split, we repeat the optimistic-display selection procedure under multiple random re-splits of cached episodes (no retraining, no additional environment interaction). This produces a distribution over optimistic-display preferences and over the resulting violation rates.

Setup. On `MO-HalfCheetah-v2`, we use 5 seeds and perform 30 random re-splits per seed (150 total). Each re-split uses $n_{\mathrm{cal}} = 200$ and $n_{\mathrm{test}} = 500$ and evaluates on a dense grid ($N = 101$). CERTPREF uses the U_{smart} router in this robustness study.

Reading the figure. The re-split attack statistics are reported in the main text as Figure 2. The left panel is the empirical distribution of violation at the optimistic-display $w^\star$, the middle panel is the distribution of selected locations, and the right panel shows their joint scatter.

Additional interpretation. The ECDF (left) shows substantial mass of the pointwise method's optimistic-display violation near or above α across resplits, especially in the right tail. The histogram and scatter (middle/right) reveal that the selector tends to select preferences near regions where the pointwise curve is locally optimistic, and that higher optimism correlates with higher realized miscoverage. The blue points remain concentrated below the target level, matching the selection-safe construction of CERTPREF.

A.5 ROUTER ABLATION (SMART SELECTOR VARIANTS)

The router $g(w)$ is the only deployment-time decision rule in CERTPREF: it selects which anchor's certificate and policy to apply for a query preference w. This ablation compares several routing objectives and quantifies their effect on validity, certificate tightness, and scalarized performance.

Routers compared. We compare (i) geometric nearest-neighbor routing (NN), (ii) certificate-driven routing (U_{smart}), and (iii) a performance-oriented heuristic (PerfSmart) that mirrors U_{smart} but replaces the certificate term by an empirical return proxy. Specifically, let $\widehat{\mu}_i := \frac{1}{|\mathcal{D}_i^{\mathrm{cal}}|} \sum_{\tau \in \mathcal{D}_i^{\mathrm{cal}}} w_i^\top G(\tau)$ be the mean scalarized return at the anchor preference w_i computed on calibration rollouts; PerfSmart routes by $g_{\mathrm{PerfSmart}}(w) = \arg\min_i \big(-\widehat{\mu}_i + L_{\mathrm{perf}} \, \|w - w_i\|_1 \big)$, with $L_{\mathrm{perf}} = \sum_{t=0}^{T-1} \gamma^t$. (PerfSmart uses only calibration rollouts to construct its proxy; test episodes are reserved exclusively for evaluation.)

Setup and metrics. We evaluate on five environments with $m = 16$ anchors, $n_{\mathrm{cal}} = 200$, $n_{\mathrm{test}} = 500$, and $\alpha = 0.1$. To keep the visualization readable, the figure reports *deltas* relative to NN: for

validity we plot $\Delta = \text{NN} - \text{router}$ (positive is better), while for tightness/performance we plot $\Delta = \text{router} - \text{NN}$ (positive is better).

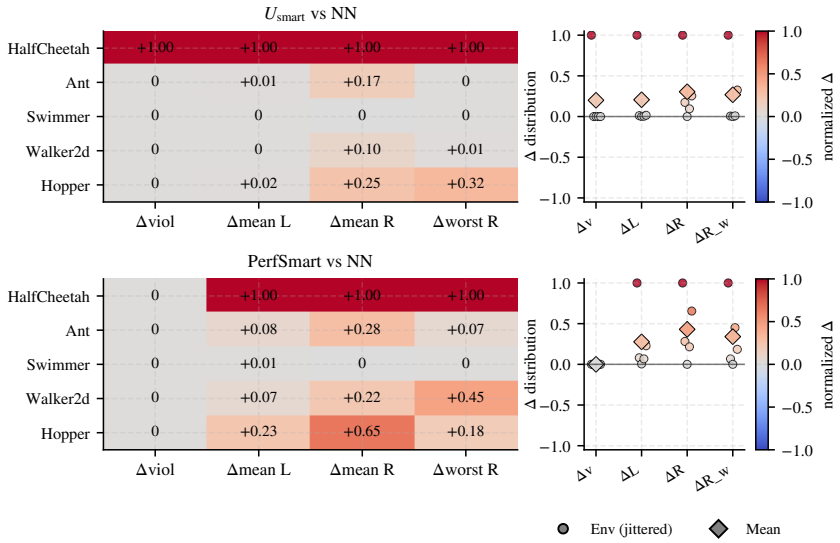


Figure 5: Router comparison on 5 environments (KCenter anchors; mean ± SE over 5 seeds). To avoid unreadable axis offsets, we visualize router effects as a delta matrix (left) together with a per-metric distribution view (right). Validity uses $\Delta = \text{NN} - \text{router}$ (positive is better), while tightness/return use $\Delta = \text{router} - \text{NN}$.

Additional interpretation. Because the post-selection guarantee holds for *any* router paired with its reported certificate, routing is primarily an optimization problem: it trades off which anchor policy is executed and which anchor certificate controls the bound. The figure shows that U_{smart} is consistently beneficial for *tightness* (positive deltas in $\overline{L}$) because it explicitly minimizes the extended upper bound $U(w_i) + L_C\|w - w_i\|_1$. Nearest-neighbor routing can be suboptimal even when anchors are well-spread geometrically, since it ignores anchor-to-anchor differences in the conformal quantiles $U(w_i)$. PerfSmart sometimes improves realized return but is less reliable for tightening certificates, since its objective is only a proxy for the upper bound. These results motivate our default choice of U_{smart} for deployment and evaluation.

A.6 Simultaneous-control alternatives and Bonferroni allocation

In the main paper we use uniform per-anchor miscoverage $\alpha_i = \alpha/m$. More generally, any deterministic allocation $\{\alpha_i\}_{i=1}^m$ that is fixed before seeing calibration scores and satisfies $\sum_i \alpha_i \le \alpha$ yields simultaneous anchor validity by the same union-bound argument. This appendix records three reviewer-facing checks: Šidák versus Bonferroni, nested learned allocations, and why max-rank style simultaneous conformal control is not a drop-in replacement for the current anchor-wise rollout design.

Šidák versus Bonferroni. Bonferroni is valid under arbitrary dependence among anchor-level errors. Šidák is a natural alternative under independence of the anchor-level error events, with the Šidák per-anchor level $\alpha_i^{\text{Sid}} = 1 - (1-\alpha)^{1/m}$. In split conformal, however, a slightly larger per-anchor level only helps if it changes the integer order statistic

$$k_i = \lceil (n_{\text{cal}} + 1)(1 - \alpha_i) \rceil .$$

Table 7 shows that, at the calibration sizes used in our experiments, Šidák does not change the realized ranks relative to Bonferroni. Consequently, Table 8 shows identical empirical metrics in the main D4MORL runs. We therefore keep Bonferroni as the default because it is dependence-robust and is not measurably looser in the reported finite-sample regime.

Data-dependent allocations and nested validity. If the allocation $\{\alpha_i\}$ is optimized using the same calibration scores that are subsequently used to set conformal quantiles, then exchangeability is

Table 7: Order-statistic discreteness for Bonferroni and Šidák corrections. "first n_{cal} with gain" is the smallest calibration size at which the Šidák rank becomes strictly smaller than the Bonferroni rank.

Setting	d	m	n_{cal}	α_i^{Bonf}	Šidák α_i	k_{Bonf}	Šidák k	first n_{cal} with gain
Main Table 1	2	16	200	0.006250	0.006563	200	200	304
$d = 3, m = 30$	3	30	600	0.003333	0.003506	599	599	855
$d = 3, m = 100$	3	100	4000	0.001000	0.001053	3997	3997	4748
$d = 3, m = 200$	3	200	4000	0.000500	0.000527	3999	3999	5696

Table 8: Bonferroni and Šidák produce identical main-run metrics because they induce the same split-conformal ranks in this finite-sample regime.

Env	Bonf. $\widehat{\mathrm{viol}}_{\max}$	Šidák $\widehat{\mathrm{viol}}_{\max}$	Bonf. UCB	Šidák UCB	Bonf. $\overline{L}$	Šidák $\overline{L}$
MO-HalfCheetah-v2	0.0160	0.0160	0.0472	0.0472	43.8836	43.8836
MO-Ant-v2	0.0152	0.0152	0.0463	0.0463	40.5137	40.5137
MO-Swimmer-v2	0.0208	0.0208	0.0549	0.0549	49.9009	49.9009
MO-Walker2d-v2	0.0176	0.0176	0.0499	0.0499	23.3699	23.3699
MO-Hopper-v2	0.0192	0.0192	0.0526	0.0526	15.0620	15.0620

broken and the distribution-free guarantee need not hold. This is a form of tuning bias analogous to conformal inference after data-dependent selection (Sarkar & Kuchibhotla, 2023; Liang et al., 2026; Hegazy et al., 2025; Zeng et al., 2025). Accordingly, the allocation frontier below is interpreted as an empirical tightness–validity trade-off unless a nested split is used.

Proposition A.1 (Two-stage Bonferroni allocation is valid)**.** *Fix m anchors and a total level $\alpha \in (0, 1)$. For each anchor i, let $\mathcal{D}_i^{\mathrm{tune}}$ and $\mathcal{D}_i^{\mathrm{cal}}$ be two independent sets of rollouts collected under the trained policy π_i, and let $\tau_{\mathrm{new}}^{(i)}$ denote a fresh deployment rollout drawn exchangeably with $\mathcal{D}_i^{\mathrm{cal}}$ and independent of $\mathcal{D}_i^{\mathrm{tune}}$. Let $\{\alpha_i\}_{i=1}^m$ be any possibly randomized allocation rule measurable with respect to the tuning data such that $\sum_i \alpha_i \le \alpha$ almost surely. Construct each anchor certificate $U(w_i)$ by split conformal on $\mathcal{D}_i^{\mathrm{cal}}$ at level α_i. Then the simultaneous event $\bigcap_{i=1}^m \{C(\tau_{\mathrm{new}}^{(i)}; w_i) \le U(w_i)\}$ holds with probability at least $1 - \alpha$.*

Proof. Condition on all tuning data. By independence, conditioning on $\mathcal{D}^{\mathrm{tune}}$ does not affect the exchangeability of $\mathcal{D}_i^{\mathrm{cal}} \cup \{\tau_{\mathrm{new}}^{(i)}\}$ for each anchor i, so split conformal guarantees $\mathbb{P}(C(\tau_{\mathrm{new}}^{(i)}; w_i) \le U(w_i) \mid \mathcal{D}^{\mathrm{tune}}) \ge 1 - \alpha_i$. A union bound yields probability at least $1 - \sum_i \alpha_i \ge 1 - \alpha$ conditional on the tuning data, and unconditioning preserves the bound. □

Table 9 evaluates such a nested learned allocation at total $n_{\mathrm{cal}} = 200$. It preserves validity, but efficiency gains are mixed. We therefore keep uniform Bonferroni allocation as the default.

Table 9: Nested learned allocation with an independent pilot split. The procedure is valid by the two-stage proposition, but the efficiency effect is mixed at the current calibration budget.

Env	Uniform $\overline{L}$	Adaptive $\overline{L}$	ΔL	Uniform UCB	Adaptive UCB	Δ UCB	Interpretation
MO-Ant-v2	40.4741	40.4979	0.0238	0.0955	0.0955	0.0000	small positive
MO-HalfCheetah-v2	46.3890	45.3907	-0.9983	0.0736	0.0803	0.0067	mixed/no clear gain
MO-Hopper-v2	15.0715	15.0715	-0.0000	0.0767	0.0767	0.0000	mixed/no clear gain
MO-Swimmer-v2	49.9073	49.9072	-0.0000	0.0735	0.0735	0.0000	mixed/no clear gain
MO-Walker2d-v2	23.3850	23.3868	0.0018	0.0787	0.0787	0.0000	small positive

Empirical allocation frontier. On `MO-HalfCheetah-v2` we fix $m = 16$ anchors (GreedyMI selection) and compare allocation rules, evaluated under both NN routing and the U_{smart} router. Each point in Figure 6 corresponds to an allocation mechanism, such as mean-optimal versus minimax-optimal allocations and a safe interpolation that constrains empirical violation increase. The plot

reports the trade-off between worst-case violation and certificate tightness. Having fewer violations than α often indicates conservatism rather than tightness: it means the bound is looser than necessary, leaving slack to increase $L(w)$ while still keeping violations below α.

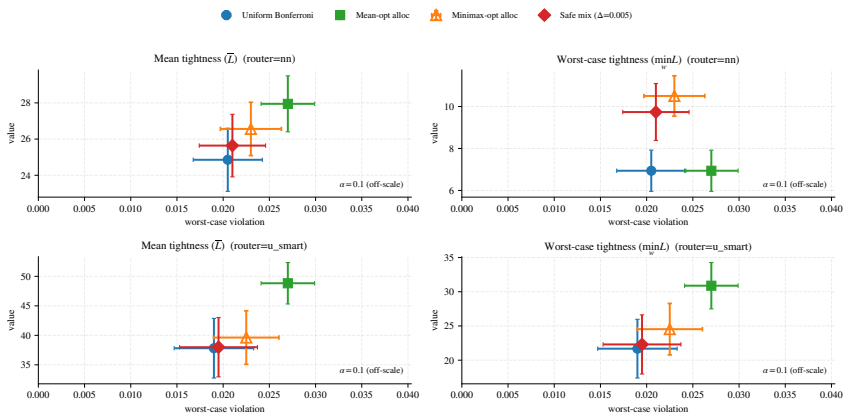


Figure 6: Tightness–validity frontier on `MO-HalfCheetah-v2` under different Bonferroni allocations (GreedyMI anchors). Each point corresponds to an allocation mechanism; error bars show mean $\pm$ SE over seeds.

The frontier makes two practical points. First, because conformal thresholds use discrete order statistics, small changes in the allocation $\{\alpha_i\}$ often lead to plateaus where the certificate is unchanged. Second, when an allocation shifts mass toward the anchors that most frequently win the envelope minimization, it can increase $\overline{L}$ without a noticeable change in worst-case violation. Allocations tuned directly on calibration scores can appear overly optimistic in-sample; this is why the nested-split protocol above is needed when one wishes to learn allocations while retaining a distribution-free guarantee.

Why max-rank or joint-rank is not used as the default. Max-rank or joint-rank conformal control is attractive for a finite family when each calibration unit provides a paired vector of scores across all family members. The current CERTPREF data structure is different: each anchor corresponds to a distinct trained policy and generates rollouts from a distinct policy-induced distribution. A rollout under anchor policy i is therefore not naturally paired with a rollout under anchor policy j as the same exchangeable calibration unit. A valid joint-rank module would require a paired calibration design in which each calibration index is evaluated under every anchor policy, for example with synchronized environment seeds and rollout batches. This is a meaningful extension, but it is not a drop-in replacement for the matched policy-routing setup used here.

A.7 CERTIFIED VS EMPIRICAL LIPSCHITZ CONSTANTS

The preference-shift robustness study in Appendix A.10 relies on a Lipschitz constant L_g for the certificate function $h(w) = L(w)$ to translate a Wasserstein-1 shift radius into a quantitative lower bound on $\mathbb{E}_\rho[L(w)]$ (we use h here to avoid confusion with the router $g(\cdot)$). A global worst-case constant can be overly conservative, so we compare it to a data-dependent constant produced by Conformal-Lipschitz and to an empirical estimate measured on held-out test episodes.

Global constant and finite horizons. Although the main theoretical development assumes an infinite-horizon discounted return, the MuJoCo benchmarks we use have a finite episode length. Following the D4MORL benchmark protocol (Zhu et al., 2023), we use a horizon $T = 500$ and discount $\gamma = 0.99$. For linear scalarization on Δ^1 with the ℓ_1 norm, this yields the (finite-horizon) global worst-case Lipschitz constant

$$L_C \;=\; \frac{1-\gamma^T}{2(1-\gamma)} \;\approx\; 49.7,$$

recovering $1/(2(1-\gamma)) = 50$ as $T \to \infty$.

Conformal-Lipschitz and failure-budget accounting. Conformal-Lipschitz certifies the trajectory-level Lipschitz modulus $\ell(\tau)$ (Lemma C.2) for each anchor policy via an additional split conformal calibration with per-anchor levels $\{\beta_i\}_{i=1}^m$. With $\sum_i \beta_i \le \beta$, a union bound implies $\mathbb{P}(\forall i : \ell(\tau_{\text{new}}^{(i)}) \le \widehat{L}_i) \ge 1-\beta$. Combining these certified Lipschitz bounds with the simultaneous return certificates yields overall certificate failure at most $\alpha+\beta$ (Theorem B.5). In the robustness study of Appendix A.10 we set $\beta_{\text{total}} = 0.06$ and allocate $\beta_i = \beta_{\text{total}}/m$.

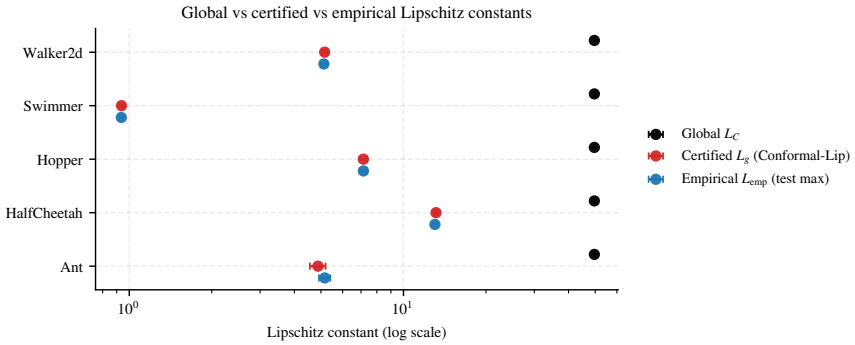


Figure 7: Global vs certified vs empirical Lipschitz constants across environments (log scale).

Table 10: Numerical values corresponding to Figure 7.

Env	L_C (global)	L_g (cert., Conformal-Lip)	L_{emp} (max on test)
Ant	49.7	4.88 ± 0.32	5.16 ± 0.25
HalfCheetah	49.7	13.14 ± 0.14	13.02 ± 0.20
Hopper	49.7	7.14 ± 0.22	7.14 ± 0.22
Swimmer	49.7	0.94 ± 0.02	0.94 ± 0.02
Walker2d	49.7	5.16 ± 0.11	5.13 ± 0.12

Additional interpretation. Figure 7 shows that the global worst-case bound L_C can be one to two orders of magnitude larger than the empirical Lipschitz constant measured on held-out rollouts. Conformal-Lipschitz narrows this gap by certifying per-anchor constants $\widehat{L}_i$ that are much closer to empirical behavior while still remaining conservative by construction. This matters because the Wasserstein shift bound in Appendix A.10 degrades linearly with the chosen constant L_g: replacing L_C with $\max_i \widehat{L}_i$ can substantially tighten the predicted degradation under preference shift.

A.8 Offline and conditional trainer comparison

A key practical feature of CertPref is that certification is *trainer-agnostic*: any procedure that yields anchor policies and rollouts can be combined with the same conformal calibration, simultaneous correction, and extension steps. This experiment illustrates that the pipeline remains compatible with strict-offline training, where training never interacts with the environment and evaluation rollouts are collected only after training.

Setup. We compare three trainers across five environments: (i) BC-offline, a preference-conditioned behavior cloning policy trained on the PEDA offline dataset (`expert_uniform` split); (ii) CAPQL, a conditional online baseline; and (iii) per-anchor SAC (one policy trained per anchor). To control evaluation cost in the offline setting, we use $m = 4$ anchors with $n_{\text{cal}} = 100$, $n_{\text{test}} = 200$, and $\alpha = 0.1$, evaluated with the U_{smart} router.

Additional interpretation. The key takeaway is that certification is composable: regardless of whether the underlying policy is trained online per-anchor (SAC), online conditional (CAPQL), or strictly offline (BC), the same conformal+extension pipeline yields valid post-selection certificates as long as calibration and deployment rollouts are exchangeable for each frozen policy. What changes across trainers is primarily *tightness* and *performance*. No single trainer dominates in this small comparison: BC-offline gives higher HV and lower bounds on some tasks, while SAC and CAPQL

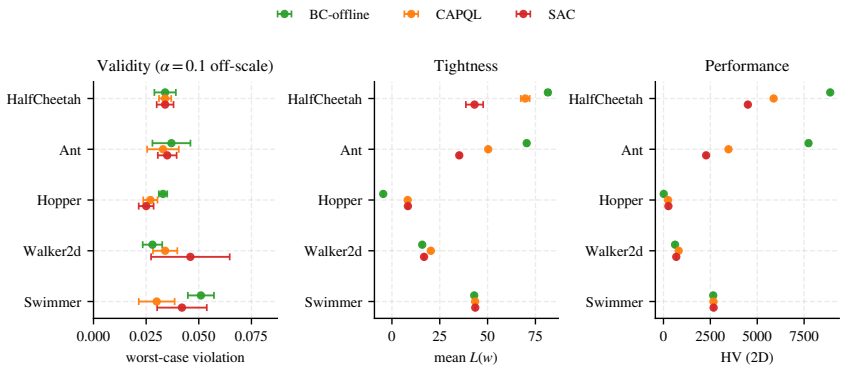


Figure 8: Trainer comparison across environments (mean ± SE over seeds). Columns: validity (worst-case violation), tightness (mean certified lower bound), and performance (2D hypervolume).

Table 11: Numerical summary for Figure 8.

Env	BC (offline)			CAPQL			SAC		
	$\widehat{\text{viol}}_{\max}$	$\overline{L}$	HV	$\widehat{\text{viol}}_{\max}$	$\overline{L}$	HV	$\widehat{\text{viol}}_{\max}$	$\overline{L}$	HV
MO-HalfCheetah-v2	0.034±0.005	81.5±1.1	8892±33	0.034±0.003	69.7±2.3	5875±12	0.034±0.004	43.2±4.5	4500±34
MO-Ant-v2	0.037±0.009	70.4±0.7	7734±83	0.033±0.008	50.3±0.4	3467±6	0.035±0.004	35.2±0.6	2274±57
MO-Hopper-v2	0.033±0.002	-4.5±0.2	14±1	0.027±0.003	8.2±0.2	244±4	0.025±0.004	8.4±0.4	263±4
MO-Walker2d-v2	0.028±0.005	15.8±0.3	618±29	0.034±0.006	20.3±0.5	810±37	0.046±0.019	16.8±0.3	682±9
MO-Swimmer-v2	0.051±0.006	43.0±0.2	2653±5	0.030±0.009	43.5±0.2	2668±3	0.042±0.012	43.5±0.2	2674±1

are competitive elsewhere. The certification layer is therefore best viewed as trainer-agnostic; it certifies whatever frozen policy family is supplied, while the usefulness of the lower bounds depends on the trainer's quality and distributional match. This separation is useful in practice: certification can be treated as a reliability layer on top of whatever MORL trainer is available.

A.9 $d = 3$ PREFERENCE-SIMPLEX ANCHOR-BUDGET AUDIT

This appendix complements the $d = 2$ MuJoCo experiments with a three-objective audit on `mo-highway-fast-v0`. The goal is deliberately modest: we test whether the same conformal-extension pipeline behaves as predicted on the 2-simplex Δ^2, not whether this toy environment solves high-dimensional practical certification. The key questions are geometric validity and nontriviality. As the anchor budget grows, the Lipschitz extension penalty should shrink, the raw lower-bound curve should become less conservative, and the support-clipped lower bound should become nontrivial on a larger fraction of preferences.

Setup. We use a three-objective simplex grid with 231 evaluation preferences, clipped rewards in $[0, 1]$, discount $\gamma = 0.99$, and horizon cap $T = 30$. Anchors are selected by a k-center farthest-point prefix rule. The strong sweep uses $n_{\text{cal}} = 600$ for $m \in \{4, 7, 10, 15, 20, 30\}$; the larger sweep uses $n_{\text{cal}} = 4000$ for $m \in \{100, 200\}$. All rows in Table 12 report mean ± standard error over five seeds.

Metrics. $\overline{L}$ is the raw average lower bound over the 231 reported simplex-grid preferences. $\overline{L}_+ = \mathbb{E}[\max\{L(w), 0\}]$ is the average improvement over the deterministic trivial lower bound 0. $\Pr[L(w) > 0]$ is the fraction of grid preferences where the certificate is strictly nontrivial. $\widehat{\text{viol}}_{\max}$ is the maximum empirical violation over the reported grid.

Interpretation. The audit provides strong validity evidence: all observed worst-grid violation rates are far below $\alpha = 0.1$. The nontriviality metrics show the expected anchor-budget trend. As m increases, the raw mean lower bound becomes much less negative, $\overline{L}_+$ increases, and the fraction of nontrivial certificates rises from about 1% at $m = 4$ to 5.3% at $m = 200$. At the same time, the raw average lower bound remains negative throughout the reported sweep. We therefore present this experiment as a validity and geometry audit, not as the main evidence that practically useful high-dimensional certificates are already solved.

Table 12: $d = 3$ nontriviality-aware audit on `mo-highway-fast-v0`. All rows use five seeds. Worst-grid violation remains far below $\alpha = 0.1$. Increasing the anchor budget makes the raw lower bound less negative and increases the fraction of preferences with positive support-clipped certificates, while the raw average lower bound remains negative throughout the reported sweep.

m	n_{cal}	$\widehat{\text{viol}}_{\max}$	$\overline{L}$	$\overline{L}_+$	$\Pr[L(w) > 0]$
4	600	0.005±0.003	-5.038±0.108	0.003±0.001	0.010±0.003
7	600	0.003±0.001	-3.456±0.019	0.003±0.001	0.013±0.003
10	600	0.002±0.001	-2.554±0.031	0.003±0.001	0.018±0.003
15	600	0.003±0.001	-1.992±0.020	0.004±0.001	0.021±0.005
20	600	0.002±0.001	-1.647±0.008	0.004±0.001	0.022±0.003
30	600	0.002±0.001	-1.251±0.014	0.004±0.001	0.026±0.002
100	4000	0.001±0.001	-0.620±0.001	0.005±0.001	0.039±0.002
200	4000	0.003±0.001	-0.403±0.009	0.006±0.001	0.053±0.012

Visualization. Figure 9 shows one single-run visualization of the same geometry. The left panel visualizes the discrete anchor-only lower bound over Δ^2. The right panel shows the lower-bound field after Lipschitz extension. This figure is qualitative; Table 12 is the quantitative nontriviality-aware audit.

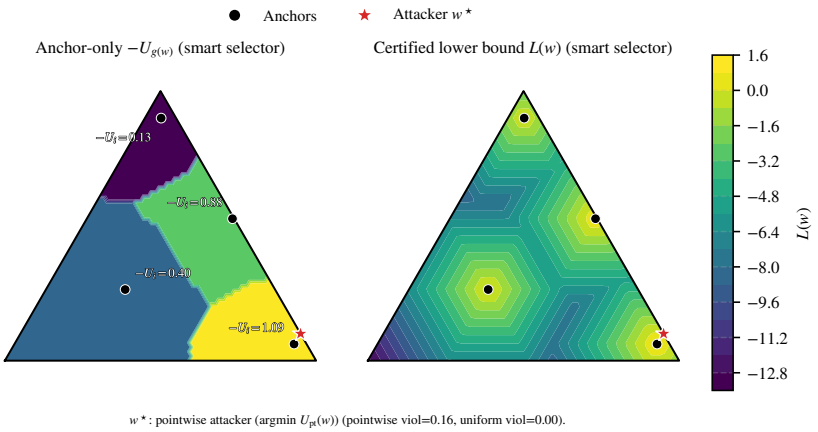


Figure 9: $d = 3$ showcase on `mo-highway-fast-v0` (single run for visualization; $m = 4$, $n_{\text{cal}} = 50$, $n_{\text{test}} = 50$, $\alpha = 0.1$). Left: anchor-only bound $-U(w_{g(w)})$ over Δ^2. Right: certified lower-bound field $L(w)$ after Lipschitz extension. Black dots indicate anchors; the red star is the optimistic-display preference $w^\star$ from the same run.

A.10 Preference shift and robustness

We illustrate the preference-shift bound on `MO-HalfCheetah-v2` in Figure 10 and then provide a compact cross-environment summary of how Conformal-Lipschitz tightens the Lipschitz constant (and therefore the Wasserstein robustness bound) across all five MuJoCo tasks.

Shift family (Beta). For the $d = 2$ MuJoCo tasks we parameterize preferences as $w(t) = (t, 1 - t)$ with $t \in [0, 1]$. We take the nominal preference distribution to be $t \sim \text{Beta}(5, 5)$ and consider shifted distributions $t \sim \text{Beta}(a, b)$ with $(a, b) \in \{(2, 8), (3, 7), (4, 6), (5, 5), (6, 4), (7, 3), (8, 2)\}$. Shift size is reported as a Wasserstein-1 radius $\eta = W_1(\rho, \rho_0)$ under the metric induced by $\|\cdot\|_1$ on w. Since $\|w(t) - w(t')\|_1 = 2|t - t'|$ on Δ^1, this corresponds to $\eta = 2\, W_1(t, t_0)$ for the underlying 1D

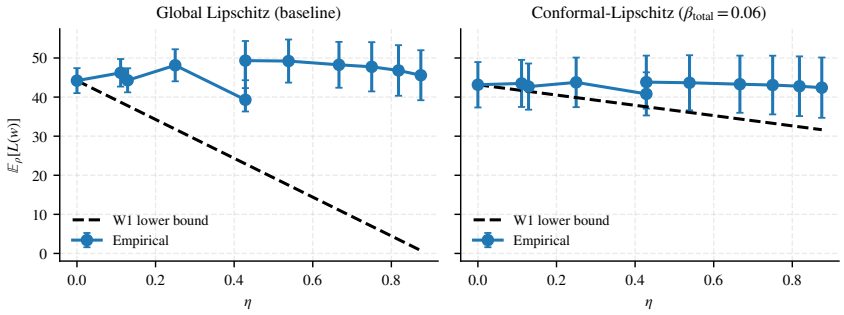


Figure 10: Preference shift vs expected certified lower bound on `MO-HalfCheetah-v2`. Left: global Lipschitz baseline; right: Conformal-Lipschitz (certified data-dependent Lipschitz; total Lipschitz miscoverage $\beta_{\text{total}} = 0.06$, yielding overall certificate failure at most $\alpha + \beta_{\text{total}}$; Theorem B.5). Points show empirical $\mathbb{E}_\rho[L(w)]$ under Beta preference shifts (nominal ρ_0 and shifted ρ family below); dashed line is the Wasserstein lower bound $\mathbb{E}_{\rho_0}[L(w)] - \eta L_g$ with $L_g = L_C$ (global) or $L_g = \max_i \widehat{L}_i$ (Conformal-Lipschitz).

Beta distributions over t. Expectations $\mathbb{E}_\rho[L(w)]$ are computed by numerical integration over a dense grid (we use $N = 2001$ points).

Interpretation. The column "shrink $\times$" reports the ratio L_C/L_g (larger indicates a tighter certified constant). The last column reports the bound at the largest shift radius considered, formatted as "global / certified" to highlight how the global constant can become vacuous while the certified constant remains informative.

Table 13: Preference-shift robustness summary across environments (Conformal-Lipschitz with $\beta_{\text{total}} = 0.06$).

Env	L_C (global)	L_g (certified)	shrink $\times$	bound@$\eta_{\max}$
MO-Ant-v2	49.7	4.9±0.3	10.18	-2.9 / 38.8
MO-HalfCheetah-v2	49.7	13.1±0.1	3.78	0.7 / 31.7
MO-Hopper-v2	49.7	7.1±0.2	6.95	-28.6 / 11.7
MO-Swimmer-v2	49.7	0.9±0.0	53.01	6.5 / 50.9
MO-Walker2d-v2	49.7	5.2±0.1	9.63	-20.2 / 21.5

Additional interpretation. Figure 10 illustrates two regimes: for small shift radii, the empirical mean certificate degrades smoothly and the Wasserstein lower bound tracks it reasonably well. As the shift grows, the global-constant bound $\mathbb{E}_{\rho_0}[L] - \eta L_C$ can quickly become negative (vacuous), even though the empirical degradation remains moderate. Conformal-Lipschitz mitigates this by replacing L_C with a much smaller certified constant, so the predicted degradation remains informative over the full sweep. Table 13 shows that this tightening ("shrink $\times$") is consistent across tasks, with the largest improvements in environments where return ranges are narrower.

A.11 Algorithm summary

Algorithm 1 in the main text gives the paper-level pseudocode for the full release-then-select pipeline. This appendix entry records the same construction as implementation guidance: choose anchors and error levels, train and freeze one policy per anchor, calibrate one-sided upper bounds on held-out rollouts, publish the Lipschitz envelope, and execute the anchor policy matched to the reported envelope value. Optional modules in Appendix A.6–Appendix A.10 modify anchor selection, local Lipschitz constants, routing, or population-level robustness, but they do not change the core validity contract in Algorithm 1.

A.12 DENSE-GRID BONFERRONI BASELINE

Dense-grid Bonferroni is the natural finite-grid alternative to CERTPREF. It protects every displayed grid preference directly, instead of calibrating a smaller anchor set and extending between anchors. This makes the finite family size equal to the display resolution. For split conformal with K protected grid points and total level α, Bonferroni gives each point level α/K. The finite-sample upper quantile is finite only when the requested tail level is at least one rank among $n_{\mathrm{cal}} + 1$ exchangeable scores, i.e., when $\alpha/K \geq 1/(n_{\mathrm{cal}} + 1)$. Equivalently, a non-vacuous dense-grid correction requires $n_{\mathrm{cal}} \geq K/\alpha - 1$ calibration rollouts per protected policy family.

Table 14: Calibration-budget arithmetic for finite-family correction at $\alpha = 0.1$. Dense-grid Bonferroni over the 401 evaluation points is vacuous with $n_{\mathrm{cal}} = 200$. CERTPREF stays finite because calibration is performed only at 16 anchors and off-anchor preferences are handled by Lipschitz extension.

Protected family	K	Per-query level	Min. n_{cal}
Fixed query	1	0.1	9
CERTPREF anchors	16	0.1/16	159
Dense grid	401	0.1/401	4009

Thus dense-grid Bonferroni is not a useful main-text competitor under our calibration budget. It becomes non-vacuous only after increasing calibration samples by roughly a factor of twenty, or after reducing the displayed grid to a very coarse set of preferences. This is the practical reason for separating calibration from coverage: CERTPREF calibrates a small anchor family and pays the continuous-preference cost through the geometric extension term.

APPENDIX B: CONFORMAL PREDICTION CERTIFICATES

This appendix provides full proof details for the conformal calibration and Lipschitz extension chain used to derive post-selection validity under the single-query deployment protocol (Theorems 5.1, B.2 and B.3). For general background on conformal prediction, see Vovk et al. (2005); Shafer & Vovk (2008); Lei et al. (2018).

B.0 PRELIMINARIES: POINTWISE FAILURE, EXCHANGEABILITY, AND TIES

A finite-menu pointwise-failure example. This example formalizes why pointwise validity alone is not enough after a displayed family is inspected. Suppose there are m displayed preferences. At each preference i, the calibration threshold U_i and a fresh score R_i^{new} are independent $\mathrm{Unif}(0, 1)$ variables, and the pointwise rule declares success when $R_i^{\mathrm{new}} \leq U_i$. For any fixed i, the success probability is $\mathbb{E}[U_i] = 1/2$, so the rule is pointwise valid at level $1/2$. Now let the selector choose $i^\star \in \arg\min_i U_i$, which is equivalent to choosing the largest displayed lower bound $L_i = -U_i$. The selected miscoverage is

$$\mathbb{P}(R_{i^\star}^{\mathrm{new}} > U_{i^\star}) = \mathbb{E}[1 - \min_i U_i] = \frac{m}{m+1}.$$

Thus fixed-query validity at each displayed preference does not imply validity after selection; the selected failure probability can approach one as the menu grows. CERTPREF prevents this failure by replacing marginal anchor events with a simultaneous anchor event before applying the continuous extension.

Exchangeability. A finite sequence $(Z_1, \ldots, Z_{n+1})$ is *exchangeable* if its joint distribution is invariant under permutations. If g is measurable, then $(g(Z_1), \ldots, g(Z_{n+1}))$ is exchangeable as well. In our deployment setting we invoke *conditional* exchangeability (Assumption 4.1): for each anchor i, conditional on the trained (possibly random) policy π_i (and on any randomness used to produce it), the calibration rollouts used to build $U(w_i)$ and any fresh rollout obtained by executing π_i are exchangeable. I.i.d. episodic rollouts with independent environment resets are a sufficient special case, but are not required by the proofs below.

Randomized tie-breaking. If the score distribution is discrete (or otherwise yields ties with positive probability), the classical rank argument is stated most cleanly by augmenting scores with i.i.d.

auxiliary noise for tie-breaking (Vovk et al., 2005; Shafer & Vovk, 2008). This randomization is purely a proof device: the certificate threshold U depends only on the original scores. In implementation, one may break ties deterministically (e.g., by index) or add an infinitesimal jitter without affecting the validity statement.

B.1 Split conformal upper bounds

Lemma B.1 (Split conformal upper bound)**.** *Let $R_1, \dots, R_n$ be calibration scores and R_{n+1} a fresh test score. Assume $(R_1, \dots, R_{n+1})$ are exchangeable. Fix $\alpha \in (0,1)$ and set $k := \lceil (n+1)(1-\alpha) \rceil$. Let $R_{(1)} \le \cdots \le R_{(n)}$ denote the order statistics of the calibration scores $(R_1, \dots, R_n)$ (counting multiplicity), and set $R_{(n+1)} := +\infty$. Define*

$$U := R_{(k)}. \tag{B.1}$$

Then

$$\mathbb{P}\big(R_{n+1} \le U\big) \;\ge\; \frac{k}{n+1} \;\ge\; 1-\alpha. \tag{B.2}$$

If additionally $\mathbb{P}(\exists i \neq j :\ R_i = R_j) = 0$ (no ties), then $\mathbb{P}(R_{n+1} \le U) = k/(n+1)$.

Proof. Introduce i.i.d. tie-breakers $\xi_1, \dots, \xi_{n+1} \sim \mathrm{Unif}(0,1)$ independent of $R_{1:n+1}$ and order pairs $S_i := (R_i, \xi_i)$ lexicographically. By exchangeability, $(S_1, \dots, S_{n+1})$ is exchangeable and has almost surely distinct values, so the rank $\mathrm{Rank}(S_{n+1})$ of S_{n+1} among $\{S_1, \dots, S_{n+1}\}$ is uniform on $\{1, \dots, n+1\}$.

Let $U = R_{(k)}$ with the convention $R_{(n+1)} = +\infty$. We claim that

$$\{\mathrm{Rank}(S_{n+1}) \le k\} \;\subseteq\; \{R_{n+1} \le U\}.$$

Indeed, if $R_{n+1} > U$ then at least k calibration scores satisfy $R_j \le U < R_{n+1}$, hence their pairs S_j must precede S_{n+1} under lexicographic order, implying $\mathrm{Rank}(S_{n+1}) > k$. Taking contrapositives yields the claim.

Therefore,

$$\mathbb{P}(R_{n+1} \le U) \;\ge\; \mathbb{P}\big(\mathrm{Rank}(S_{n+1}) \le k\big) \;=\; \frac{k}{n+1} \;\ge\; 1-\alpha.$$

If ties occur with probability zero, the same rank argument applies without auxiliary randomization, and one obtains equality $\mathbb{P}(R_{n+1} \le U) = k/(n+1)$. □

B.2 Simultaneous anchor validity

Theorem B.2 (Simultaneous conformal certificates on anchors)**.** *Fix anchors $S = \{w_i\}_{i=1}^m$ and significance levels $\{\alpha_i\}$ with $\sum_i \alpha_i \le \alpha$. For each i, let $U(w_i)$ be a split-conformal upper bound built from exchangeable calibration rollouts obtained by executing the frozen anchor policy π_i with miscoverage α_i. Then*

$$\mathbb{P}\Big(\forall i \in [m],\ C(\tau_{\mathrm{new}}^{(i)}; w_i) \le U(w_i)\Big) \ge 1-\alpha.$$

Proof. Fix an anchor $i \in [m]$. Condition on the trained anchor policy π_i (and on any data/randomness used to produce it). By Assumption 4.1, conditional on π_i the calibration rollouts used to construct $U(w_i)$ and any fresh rollout $\tau_{\mathrm{new}}^{(i)}$ obtained by executing π_i are exchangeable. Since $\tau \mapsto C(\tau; w_i)$ is measurable, the corresponding scores are exchangeable as well. Applying Lemma B.1 at level α_i yields

$$\mathbb{P}\big(C(\tau_{\mathrm{new}}^{(i)}; w_i) > U(w_i)\big) \le \alpha_i.$$

Now define failure events $E_i := \{C(\tau_{\mathrm{new}}^{(i)}; w_i) > U(w_i)\}$. No independence across anchors is required. A union bound gives

$$\mathbb{P}\Big(\bigcup_{i=1}^m E_i\Big) \le \sum_{i=1}^m \mathbb{P}(E_i) \le \sum_{i=1}^m \alpha_i \le \alpha,$$

so taking complements yields the simultaneous statement. □

Remark (alternative multiple-testing control). Bonferroni control is valid under arbitrary dependence. Under mutual independence of the anchor-level failure events $\{E_i\}$ (e.g., when each anchor uses independent calibration data and an independent fresh test rollout), one may instead use the Šidák allocation to reduce conservatism (Šidák, 1967). In particular, taking a uniform Šidák level

$$\alpha_i \;=\; 1-(1-\alpha)^{1/m}$$

ensures $\mathbb{P}(\cup_i E_i) \le \alpha$ under independence, while being (slightly) less conservative than Bonferroni's α/m.

Step-down procedures such as Holm (Holm, 1979) typically operate on a family of (valid) p-values for realized test statistics. To apply such methods here one would need to construct conformal p-values for each anchor's deployment score and then invert the multiple-testing procedure to obtain thresholds. This is feasible in principle but orthogonal to our main focus; in our experiments the dominant tightness loss comes from preference-space coverage and Lipschitz extension, which we address through anchor design and non-uniform Bonferroni allocation (Appendix A.6).

Finally, conformal-specific corrections such as Max-Rank (Timans et al., 2025) are designed for many simultaneous conformal predictions that share a calibration set. Adapting them to our anchor-wise certification would require constructing conformal p-values and/or sharing calibration across anchors, which we do not do in this paper. Importantly, all downstream post-selection and Lipschitz-extension arguments depend only on the simultaneous event $\mathbb{P}(\forall i : C(\tau_{\text{new}}^{(i)}; w_i) \le U(w_i)) \ge 1-\alpha$; any procedure that establishes this event can be plugged into our pipeline unchanged.

B.3 Lipschitz transfer from anchors to continuous preferences

Theorem B.3 (Lipschitz extension yields uniform certificates)**.** *Under Assumption 4.2, define $U(w)$ via* (2) *(or its envelope form* (3)*). If the anchor event in Theorem B.2 holds, then for all $w \in \mathcal{W}$,*

$$C\big(\tau_{\text{new}}^{(g(w))}; w\big) \le U(w).$$

Proof of Theorem B.3 (Lipschitz extension). Fix a query preference $w \in \mathcal{W}$ and let $i \coloneqq g(w)$ be the anchor index selected by the router used at deployment. If the envelope certificate $U_{\min}(w)$ in (3) is used, we take $g(w) \in \arg\min_i \big(U(w_i) + L_C \left\|w - w_i\right\|_1\big)$ so that $U(w) = U_{\min}(w)$. The deployed policy is π_i, and the induced test rollout is $\tau_{\text{new}}^{(i)}$. By Assumption 4.2, for every trajectory τ,

$$C(\tau; w) \le C(\tau; w_i) + L_C \left\|w - w_i\right\|_1 .$$

Applying this inequality to $\tau = \tau_{\text{new}}^{(i)}$ and using the definition of the extended threshold (Equation (2)) gives

$$\begin{aligned} C(\tau_{\text{new}}^{(i)}; w) &\le C(\tau_{\text{new}}^{(i)}; w_i) + L_C \left\|w - w_i\right\|_1 \\ &\le U(w_i) + L_C \left\|w - w_i\right\|_1 = U(w). \end{aligned}$$

The second inequality holds on the simultaneous anchor event from Theorem B.2. Since w was arbitrary, this establishes the uniform implication. □

Proof of Theorem 5.1 (post-selection validity). Theorem 5.1 is a direct consequence of Theorems B.2 and B.3, but we include the complete argument for clarity. Let

$$E \;\coloneqq\; \bigcap_{i=1}^{m} \{C(\tau_{\text{new}}^{(i)}; w_i) \le U(w_i)\}$$

denote the simultaneous anchor event. By Theorem B.2, $\mathbb{P}(E) \ge 1-\alpha$. On E, Theorem B.3 implies the deterministic implication

$$\forall w \in \mathcal{W}, \qquad C\big(\tau_{\text{new}}^{(g(w))}; w\big) \le U(w),$$

for the router g paired with the published certificate $U(\cdot)$. In a single-query deployment, the realized rollout is $\tau_{\text{new}} = \tau_{\text{new}}^{(g(\widehat{w}))}$ for the user-chosen preference $\widehat{w}$ (which may be selected after observing the entire function $U(\cdot)$). Therefore $E \subseteq \{C(\tau_{\text{new}}; \widehat{w}) \le U(\widehat{w})\}$, and hence

$$\mathbb{P}\big(C(\tau_{\text{new}}; \widehat{w}) \le U(\widehat{w})\big) \;\ge\; \mathbb{P}(E) \;\ge\; 1-\alpha.$$

□

Proposition B.4 (Envelope optimality). *Fix anchor values $\{U(w_i)\}_{i=1}^m$ and a constant L_C. For any router $g : \mathcal{W} \to [m]$, let $U_g(w)$ be defined by* (2)*. Then $U_{\min}(w) \le U_g(w)$ for every $w \in \mathcal{W}$, and equality is attained by a router that selects an anchor attaining the minimum in* (3)*.*

Proof of Proposition B.4 (envelope optimality). Fix any preference $w \in \mathcal{W}$ and any router $g : \mathcal{W} \to [m]$. By definition,

$$U_g(w) = U(w_{g(w)}) + L_C \left\|w - w_{g(w)}\right\|_1 \;\ge\; \min_{i\in[m]} \Big(U(w_i) + L_C \left\|w - w_i\right\|_1\Big) = U_{\min}(w).$$

Taking the infimum over all routers g yields $\inf_g U_g(w) \ge U_{\min}(w)$. Conversely, let $g_{\text{U-smart}}(w) \in \arg\min_i \big(U(w_i) + L_C \left\|w - w_i\right\|_1\big)$ (breaking ties arbitrarily). Then $U_{g_{\text{U-smart}}}(w) = U_{\min}(w)$ for every w, so $\inf_g U_g(w) \le U_{\min}(w)$. Combining the two inequalities proves $U_{\min}(w) = \inf_g U_g(w)$, and shows the infimum is attained by $g_{\text{U-smart}}$. □

Optional: certified local constants (Conformal-Lipschitz). Instead of using a conservative global L_C in (3), we can certify (with an additional failure budget) local upper bounds on a *trajectory-level* Lipschitz modulus. Concretely, suppose there exists a measurable function $\ell(\tau) \ge 0$ such that for all τ and all $w, w' \in \mathcal{W}$,

$$|C(\tau; w) - C(\tau; w')| \le \ell(\tau) \left\|w - w'\right\|_1 .$$

For linear scalarization $C(\tau; w) = -w^\top G(\tau)$, Appendix C.1 gives the exact modulus $\ell(\tau) = \frac{1}{2}(\max_j G_j(\tau) - \min_j G_j(\tau))$.

Theorem B.5 (Validity with certified local Lipschitz constants). *Fix per-anchor miscoverage levels $\{\alpha_i\}_{i=1}^m$ and Lipschitz-budget levels $\{\beta_i\}_{i=1}^m$ such that $\sum_i \alpha_i \le \alpha$ and $\sum_i \beta_i \le \beta$. Let $U(w_i)$ be the split conformal upper bound for $C(\tau; w_i)$ at level $1 - \alpha_i$, and let $\widehat{L}_i$ be the split conformal upper bound for $\ell(\tau)$ at level $1 - \beta_i$, each constructed from calibration/test rollouts that are exchangeable conditional on the fixed policy π_i (Assumption 4.1); i.i.d. episodic rollouts are a sufficient special case. Define the extended certificate*

$$U_{\text{conf-lip}}(w) \;\coloneqq\; \min_{i\in[m]} \Big(U(w_i) + \widehat{L}_i \left\|w - w_i\right\|_1\Big),$$

and deploy the corresponding anchor $g_{\text{conf-lip}}(w) \in \arg\min_i \big(U(w_i) + \widehat{L}_i \left\|w - w_i\right\|_1\big)$. Then for any (possibly adaptive) single-query preference $\widehat{w}$ measurable w.r.t. the published function $U_{\text{conf-lip}}(\cdot)$ (equivalently, w.r.t. $\{U(w_i), \widehat{L}_i\}_{i=1}^m$), letting $\tau_{\text{new}} = \tau_{\text{new}}^{(g_{\text{conf-lip}}(\widehat{w}))}$ denote the executed rollout, we have

$$\mathbb{P}\Big(C(\tau_{\text{new}}; \widehat{w}) \le U_{\text{conf-lip}}(\widehat{w})\Big) \;\ge\; 1 - (\alpha + \beta).$$

Proof. Let $E_U = \bigcap_i \{C(\tau_{\text{new}}^{(i)}; w_i) \le U(w_i)\}$ and $E_L = \bigcap_i \{\ell(\tau_{\text{new}}^{(i)}) \le \widehat{L}_i\}$. By Theorem B.2 we have $\mathbb{P}(E_U) \ge 1 - \alpha$, and by Lemma B.1 (plus a union bound over $\{\beta_i\}$) we have $\mathbb{P}(E_L) \ge 1 - \beta$. Thus $\mathbb{P}(E_U \cap E_L) \ge 1 - (\alpha + \beta)$. On $E_U \cap E_L$, for any w and any i,

$$C(\tau_{\text{new}}^{(i)}; w) \le C(\tau_{\text{new}}^{(i)}; w_i) + \ell(\tau_{\text{new}}^{(i)}) \left\|w - w_i\right\|_1 \le U(w_i) + \widehat{L}_i \left\|w - w_i\right\|_1 .$$

Taking the minimum over i yields $C\big(\tau_{\text{new}}^{(g_{\text{conf-lip}}(w))}; w\big) \le U_{\text{conf-lip}}(w)$ for all w on this event. Evaluating at $w = \widehat{w}$ gives the claim. □

Remark B.6 (Counterfactual rollouts in the simultaneous event). The random variables $\tau_{\text{new}}^{(i)}$ appearing in Theorems 5.1 and B.2 are introduced only for analysis: they denote fresh rollouts obtained by (hypothetically) executing each anchor policy π_i once and then evaluating its score at the anchor preference w_i. For convenience, one may couple these counterfactual rollouts as independent draws, but independence is *not* required for the union-bound argument in Theorem B.2. In an actual single-query deployment, only the rollout of the selected anchor is realized, i.e., $\tau_{\text{new}} = \tau_{\text{new}}^{(g(\widehat{w}))}$.

Remark B.7 (Repeated deployments). If the single-query protocol is executed for K deployment episodes, each with a fresh rollout and a preference choice made before observing that episode's rollout (e.g., based on the published certificate and any past feedback, but not on the current rollout), then by a union bound the probability of at least one failure is at most $K\alpha$ when each episode uses level α. Controlling an overall failure probability α over many episodes can be done by allocating per-episode levels (more conservative certificates) or adopting online calibration; we leave multi-episode protocol design to future work.

B.4 Optional context: what does the score certify?

The conformal machinery is agnostic to the internal structure of the policy class or environment. All guarantees are expressed in terms of a measurable score function $C(\tau; w)$ evaluated on a rollout τ and a preference w. In the experiments we take $C(\tau; w) = -w^\top G(\tau)$ so that $-U(w)$ is a certified lower bound on the scalarized return, but any score that is exchangeable across calibration/test rollouts can be used.

B.5 Optional: Conformal Risk Control (CRC)

Standard split conformal bounds control a tail event of the form $\mathbb{P}(R_{\text{test}} \le U) \ge 1 - \alpha$. Conformal Risk Control (CRC) instead controls an *expected risk* (expected loss) for a user-specified loss function (Angelopoulos et al., 2024); many CRC constructions consider a monotone family of losses indexed by a threshold and calibrate a threshold to meet a target risk. Within CertPref, CRC can be applied independently at each anchor (under the same conditional exchangeability requirements) and then combined with a Lipschitz extension whenever the resulting risk functional is Lipschitz in w.

B.6 Beyond exchangeability: NexCP and related bounds

When calibration and test rollouts are not exchangeable, conformal coverage can degrade. NexCP provides explicit finite-sample bounds that quantify coverage loss via total-variation distances between the original data sequence and sequences obtained by swapping a training point with the test point (Barber et al., 2023). In sequential decision making, these results offer a principled way to relate certificate degradation to measurable distribution shift.

Appendix C: Lipschitz Properties and Regret Bounds for Multi-Objective MDPs

This appendix provides complete proofs and tightness examples for the preference-Lipschitz lemmas (including the sharp constant $1/(2(1-\gamma))$) and the coverage-driven uniform regret bound in Section 5. We consider an infinite-horizon discounted MOMDP with $\gamma \in (0, 1)$, bounded per-objective rewards $r(s, a) \in [0, 1]^d$, and preferences $w \in \mathcal{W} \coloneqq \Delta^{d-1}$.

C.0 A simplex scalarization identity

Lemma C.1 (Range of linear scalarization on the simplex)**.** *For any $w, w' \in \Delta^{d-1}$,*

$$\sup_{y \in [0,1]^d} \left| (w - w')^\top y \right| \;=\; \tfrac{1}{2} \left\| w - w' \right\|_1 . \tag{C.1}$$

Proof. Let $x \coloneqq w - w'$, so $\sum_{j=1}^d x_j = 0$. Write $x = x^+ - x^-$ where $x_j^+ \coloneqq \max\{x_j, 0\}$ and $x_j^- \coloneqq \max\{-x_j, 0\}$. Then $\sum_j x_j^+ = \sum_j x_j^- = \|x\|_1 / 2$. For any $y \in [0, 1]^d$,

$$x^\top y = \sum_j x_j^+ y_j - \sum_j x_j^- y_j \le \sum_j x_j^+ = \tfrac{1}{2} \|x\|_1 ,$$

and similarly $-x^\top y \le \sum_j x_j^- = \frac{1}{2} \|x\|_1$, proving the upper bound. Equality is achieved by taking $y_j = 1$ for indices with $x_j \ge 0$ and $y_j = 0$ otherwise (or the reverse assignment for the negative direction). □

C.1 Trajectory-level Lipschitzness for linear scalarization scores

Lemma C.2 (Trajectory-level Lipschitz constant for $C(\tau; w) = -w^\top G(\tau)$)**.** *Let $w \in \Delta^{d-1}$ and let $G(\tau) \in \mathbb{R}^d$ be any vector return. Define $C(\tau; w) = -w^\top G(\tau)$. Then for all $w, w' \in \Delta^{d-1}$,*

$$|C(\tau; w) - C(\tau; w')| \;=\; |(w - w')^\top G(\tau)| \;\le\; \frac{1}{2} \big(\max_j G_j(\tau) - \min_j G_j(\tau) \big) \, \|w - w'\|_1 .$$

Moreover, the constant $\ell(\tau) \coloneqq \frac{1}{2}(\max_j G_j(\tau) - \min_j G_j(\tau))$ is the exact *Lipschitz modulus on $(\Delta^{d-1}, \|\cdot\|_1)$.*

Proof. Let $x = w - w'$ so $\sum_j x_j = 0$. If $\max_j G_j(\tau) = \min_j G_j(\tau)$, then $G(\tau)$ is constant across objectives and the claim is trivial. Otherwise, write $G(\tau) = \min_j G_j(\tau) \cdot \mathbf{1} + y$ with $y \in [0, \max_j G_j(\tau) - \min_j G_j(\tau)]^d$. Since $x^\top \mathbf{1} = 0$, we have $x^\top G(\tau) = x^\top y$. Applying Lemma C.1 to $y/(\max_j G_j(\tau) - \min_j G_j(\tau)) \in [0,1]^d$ yields $|x^\top y| \le \frac{1}{2} \|x\|_1 (\max_j G_j(\tau) - \min_j G_j(\tau))$. Tightness follows by choosing w, w' supported on the maximizer/minimizer coordinates. □

Global bound under bounded rewards. In our MOMDP setup where $G(\tau) = \sum_{t\ge 0} \gamma^t r(s_t, a_t)$ with $r(s,a) \in [0,1]^d$, we have $0 \le G_j(\tau) \le \sum_{t\ge 0} \gamma^t = 1/(1-\gamma)$ for all j. Therefore $\ell(\tau) \le 1/(2(1-\gamma))$, and hence

$$|C(\tau; w) - C(\tau; w')| \le \frac{1}{2(1-\gamma)} \|w - w'\|_1 .$$

Thus $L_C = 1/(2(1-\gamma))$ is a valid global constant for Assumption 4.2 when $C(\tau; w) = -w^\top G(\tau)$.

C.2 Fixed-policy Lipschitzness

Lemma C.3 (Fixed-policy Lipschitz)**.** *For any fixed policy π and any $w, w' \in \mathcal{W}$,*

$$|J^\pi(w) - J^\pi(w')| \le \frac{1}{2(1-\gamma)} \|w - w'\|_1 .$$

Proof of Lemma C.3. Let $x \coloneqq w - w'$. For a trajectory under π, write $R_t \coloneqq r(s_t, a_t) \in [0,1]^d$. Then

$$|J^\pi(w) - J^\pi(w')| = \big|\mathbb{E}[\sum_{t\ge 0} \gamma^t x^\top R_t]\big| \le \mathbb{E}[\sum_{t\ge 0} \gamma^t |x^\top R_t|].$$

The inequality follows from the triangle inequality and Tonelli's theorem. Tonelli applies since $\sum_{t\ge 0} \gamma^t |x^\top R_t| \le \|x\|_1 \sum_{t\ge 0} \gamma^t$ almost surely (as $R_t \in [0,1]^d$ implies $|x^\top R_t| \le \sum_j |x_j| = \|x\|_1$). By Lemma C.1 and $R_t \in [0,1]^d$, we have $|x^\top R_t| \le \frac{1}{2} \|x\|_1$. Therefore

$$|J^\pi(w) - J^\pi(w')| \le \sum_{t\ge 0} \gamma^t \tfrac{1}{2} \|x\|_1 = \frac{1}{2(1-\gamma)} \|w - w'\|_1 . \tag{C.2}$$

□

Tightness. The constant $1/(2(1-\gamma))$ is tight. For $d = 2$, consider a one-state MOMDP with deterministic reward $(1, 0)$. Then $J^\pi(w) = w_1/(1-\gamma)$, and taking $w = (\frac{1}{2} + \delta, \frac{1}{2} - \delta)$ and $w' = (\frac{1}{2}, \frac{1}{2})$ achieves equality.

A local refinement. For fixed π, $J^\pi(w) = w^\top G^\pi$ where $G^\pi \coloneqq \mathbb{E}[\sum_{t\ge 0} \gamma^t r(s_t, a_t)] \in \mathbb{R}^d$. On the simplex under ℓ_1, the exact Lipschitz constant equals $L_\pi^\Delta = \frac{1}{2}\big(\max_j G_j^\pi - \min_j G_j^\pi\big) \le 1/(2(1-\gamma))$.

C.3 Optimal-value Lipschitzness

Lemma C.4 (Optimal-value Lipschitz)**.** *For any $w, w' \in \mathcal{W}$,*

$$|V^*(w) - V^*(w')| \le \frac{1}{2(1-\gamma)} \|w - w'\|_1 .$$

Proof of Lemma C.4. For any functions F, G on a common index set, $\sup F - \sup G \le \sup(F - G)$. Applying this with $F(\pi) = J^\pi(w)$ and $G(\pi) = J^\pi(w')$ gives

$$V^*(w) - V^*(w') \le \sup_\pi \big(J^\pi(w) - J^\pi(w')\big) \le \sup_\pi |J^\pi(w) - J^\pi(w')|.$$

Lemma C.3 bounds the final term by $\|w - w'\|_1 / (2(1-\gamma))$. Swapping (w, w') yields the absolute value bound. □

Tightness. Consider a one-state MOMDP with $d = 2$ and two actions with rewards $(1, 0)$ and $(0, 1)$. Then $V^*(w) = \max\{w_1, w_2\}/(1-\gamma)$ on Δ^1. Taking $w = (\frac{1}{2} + \delta, \frac{1}{2} - \delta)$ and $w' = (\frac{1}{2}, \frac{1}{2})$ achieves equality.

C.4 Uniform regret under an ε-cover

Theorem C.5 (Uniform regret under an ε-cover)**.** *Let $S = \{w_i\}_{i=1}^m$ be an ε-cover of $\mathcal{W}$ in ℓ_1: for every $w \in \mathcal{W}$ there exists i such that $\|w - w_i\|_1 \le \varepsilon$. Assume each anchor policy π_i satisfies a training suboptimality bound $V^*(w_i) - J^{\pi_i}(w_i) \le \varepsilon_{\mathrm{train}}(i)$. Then the deployed nearest-anchor policy $\hat{\pi}(w) = \pi_{g_{\mathrm{nn}}(w)}$ satisfies, for all $w \in \mathcal{W}$,*

$$V^*(w) - J^{\hat{\pi}(w)}(w) \le \max_i \varepsilon_{\mathrm{train}}(i) + \frac{\varepsilon}{1-\gamma}.$$

Proof of Theorem C.5. Fix $w \in \mathcal{W}$ and let $i \in \arg\min_{j\in[m]} \|w - w_j\|_1$ be any nearest anchor (ties broken arbitrarily) so that $\|w - w_i\|_1 \le \varepsilon$. Add and subtract $V^*(w_i)$ and $J^{\pi_i}(w_i)$:

$$V^*(w) - J^{\pi_i}(w) = \big(V^*(w) - V^*(w_i)\big) + \big(V^*(w_i) - J^{\pi_i}(w_i)\big) + \big(J^{\pi_i}(w_i) - J^{\pi_i}(w)\big).$$

The middle term is at most $\max_i \varepsilon_{\mathrm{train}}(i)$ by assumption. By Lemmas C.4 and C.3, the first and last terms are each bounded by $\|w - w_i\|_1 / (2(1-\gamma)) \le \varepsilon/(2(1-\gamma))$. Summing the three bounds yields the claim. □

Tightness of the coverage coefficient. The factor $\varepsilon/(1-\gamma)$ is unimprovable in the worst case even when all training errors are zero. For $d = 2$, consider the one-state, two-action MOMDP with rewards $(1, 0)$ and $(0, 1)$ and a singleton anchor $w_1 = (\frac{1}{2}, \frac{1}{2})$. An anchor policy that always selects the second action is optimal at w_1 (a tie), yet incurs regret $1/(1-\gamma)$ at $w = (1, 0)$, matching the bound with $\varepsilon = \|w - w_1\|_1 = 1$.

C.5 Refinements, constant sensitivity, and data-dependent local constants

Bounded reward range. If $r_j(s, a) \in [0, R_{\max}]$ for all j, then the same argument yields $|J^\pi(w) - J^\pi(w')| \le \frac{R_{\max}}{2(1-\gamma)} \|w - w'\|_1$ and likewise for V^*.

Sensitivity to the implemented constant. The main theorem requires the constant used in the extension to upper-bound the true trajectory-level modulus. If the implementation uses $\widetilde{L}_C \ge L_C$, every proof step remains valid after replacing L_C by $\widetilde{L}_C$; the cost is additional lower-bound conservatism, at most $(\widetilde{L}_C - L_C)r(S)$ for queries within cover radius $r(S)$. If $\widetilde{L}_C < L_C$, the anchor-to-query transfer may fail, and the post-selection theorem no longer applies. This is why empirical or local estimates must be certified, as in the Conformal-Lipschitz construction below, rather than simply plugged in when they are smaller than the global value.

Finite horizon. For the horizon-T discounted return $J_T^\pi(w) = \mathbb{E}[\sum_{t=0}^{T-1} \gamma^t w^\top r_t]$,

$$|J_T^\pi(w) - J_T^\pi(w')| \le \frac{1-\gamma^T}{2(1-\gamma)} \|w - w'\|_1 .$$

Estimating local constants. For a fixed π, the simplex Lipschitz constant $L_\pi^\Delta = \frac{1}{2}(\max_j G_j^\pi - \min_j G_j^\pi)$ can be estimated from rollouts by estimating G^π. For tighter certificates, one can upper bound L_C locally on a partition of $\mathcal{W}$ (e.g., Voronoi cells induced by anchors) and apply the corresponding partitioned extension.

C.6 Norm choice and dual norms

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ with dual norm $\|\cdot\|_*$. By Hölder, for any $y \in \mathbb{R}^d$,

$$|(w - w')^\top y| \le \|w - w'\| \, \|y\|_*.$$

In our MOMDP setup with $r(s, a) \in [0, 1]^d$, any discounted return vector $y = \sum_{t\ge0} \gamma^t r(s_t, a_t)$ lies in $[0, \frac{1}{1-\gamma}]^d = \frac{1}{1-\gamma}[0, 1]^d$. Hence a sufficient global constant is

$$L_C \;\le\; \frac{1}{1-\gamma} \sup_{y\in[0,1]^d} \|y\|_*. \tag{C.3}$$

When $\|\cdot\| = \|\cdot\|_1$ and $w, w' \in \Delta^{d-1}$, Lemma C.1 yields the sharper simplex-specific factor $1/2$.

C.7 Covering the simplex under ℓ_1

This subsection complements the discussion in Section 5.3 by making explicit how preference-space coverage scales with the anchor budget m when $\mathcal{W} = \Delta^{d-1}$ and the metric is ℓ_1. Recall $r(S) = \sup_{w \in \mathcal{W}} \min_{i \in [m]} \|w - w_i\|_1$.

Theorem C.6 (Minimax ℓ_1 cover radius on Δ^{d-1})**.** *Let*

$$r_m^\star \;:=\; \inf_{S \subset \mathcal{W}:\, |S| = m} r(S) \qquad \textit{for} \qquad \mathcal{W} = \Delta^{d-1},\ d \geq 2,$$

where distance is measured by ℓ_1. Then for all $m \geq 1$,

$$\frac{1}{2}\, m^{-1/(d-1)} \;\leq\; r_m^\star \;\leq\; 4(d-1)\, m^{-1/(d-1)}. \tag{C.4}$$

Moreover, when $d = 2$ (i.e., $\mathcal{W} = \Delta^1$), the exact minimax radius is $r_m^\star = 1/m$.

Proof. Lower bound (volumetric argument). Identify $\mathcal{W} = \Delta^{d-1}$ with the $(d-1)$-dimensional simplex

$$T \;:=\; \Big\{ x \in \mathbb{R}^{d-1}_{\geq 0} :\, \sum_{j=1}^{d-1} x_j \leq 1 \Big\}$$

via the projection $x = (w_1, \ldots, w_{d-1})$ (with $w_d = 1 - \sum_{j=1}^{d-1} x_j$). For any $w, w' \in \Delta^{d-1}$, the projected points satisfy $\|x - x'\|_1 \leq \|w - w'\|_1$, so any ℓ_1 cover of $\mathcal{W}$ at radius r projects to an ℓ_1 cover of T at radius r. Thus it suffices to lower bound the minimax cover radius of T in $\mathbb{R}^{d-1}$ under ℓ_1.

Let $B_1^{d-1}(r) = \{z \in \mathbb{R}^{d-1} : \|z\|_1 \leq r\}$ denote the ℓ_1 ball of radius r. Its Lebesgue volume is $\mathrm{vol}(B_1^{d-1}(r)) = \frac{2^{d-1}}{(d-1)!}\, r^{d-1}$. The simplex T has volume $\mathrm{vol}(T) = \frac{1}{(d-1)!}$. If m ℓ_1 balls of radius r cover T, then by subadditivity of volume,

$$\frac{1}{(d-1)!} = \mathrm{vol}(T) \;\leq\; m \cdot \mathrm{vol}(B_1^{d-1}(r)) \;=\; m \cdot \frac{2^{d-1}}{(d-1)!}\, r^{d-1},$$

which implies $r^{d-1} \geq (2^{d-1} m)^{-1}$ and hence $r \geq \frac{1}{2} m^{-1/(d-1)}$. Since $r_m^\star$ for $\mathcal{W}$ is at least the minimax radius for T, this yields the stated lower bound.

Upper bound (grid construction). For an integer $K \geq 1$, consider the barycentric grid

$$S_K \;:=\; \Big\{ w \in \Delta^{d-1} :\, w_j = \frac{a_j}{K} \text{ for integers } a_j \geq 0 \text{ with } \sum_{j=1}^{d} a_j = K \Big\}.$$

The number of grid points satisfies $|S_K| = \binom{K+d-1}{d-1} \leq (K+1)^{d-1}$. Given any $w \in \Delta^{d-1}$, define integers $a_j = \lfloor K w_j \rfloor$ for $j = 1, \ldots, d-1$ and set $a_d = K - \sum_{j=1}^{d-1} a_j$. Then $a_d \geq 0$ and $w' = (a_1/K, \ldots, a_d/K) \in S_K$. For $j \leq d-1$ we have $0 \leq w_j - a_j/K < 1/K$, so $\sum_{j=1}^{d-1} |w_j - w'_j| < (d-1)/K$. Moreover,

$$|w_d - w'_d| = \Big| \sum_{j=1}^{d-1} (w_j - w'_j) \Big| \leq \sum_{j=1}^{d-1} |w_j - w'_j| < \frac{d-1}{K}.$$

Therefore $\|w - w'\|_1 < \frac{2(d-1)}{K}$, and hence $r(S_K) \leq \frac{2(d-1)}{K}$.

Now fix $m \geq 1$. If $4(d-1)m^{-1/(d-1)} \geq 2$, then trivially $r_m^\star \leq \mathrm{diam}(\mathcal{W}) \leq 2 \leq 4(d-1)m^{-1/(d-1)}$. Otherwise, set $x := m^{1/(d-1)}$ and $K := \lfloor x \rfloor - 1 \geq 1$. Then $|S_K| \leq (K+1)^{d-1} \leq x^{d-1} = m$, so we may take $S = S_K$ (and if desired add arbitrary points to reach size exactly m without increasing the radius). In this regime $x > 2(d-1)$, so $x \geq 4$ when $d \geq 3$; thus $K = \lfloor x \rfloor - 1 \geq x - 2 \geq x/2$ (and for $d = 2$, $x = m$ is integer so $K = m - 1 \geq x/2$), and therefore

$$r_m^\star \leq r(S_K) \leq \frac{2(d-1)}{K} \leq 4(d-1)\, m^{-1/(d-1)}.$$

Exact value for $d = 2$. When $d = 2$, every $w \in \Delta^1$ can be written as $w(t) = (t, 1-t)$ with $t \in [0, 1]$, and $\|w(t) - w(t')\|_1 = 2|t - t'|$. The minimax radius on $[0, 1]$ under $|t - t'|$ with m points equals $1/(2m)$, hence $r^\star_m = 2 \cdot (1/(2m)) = 1/m$. □

Corollary C.7 (Unavoidable worst-case Lipschitz slack for anchor-based extensions)**.** *Let $S = \{w_i\}_{i=1}^m \subset \mathcal{W}$ be any anchor set and let $g : \mathcal{W} \to [m]$ be any router. Then*

$$\sup_{w \in \mathcal{W}} \left\|w - w_{g(w)}\right\|_1 \;\geq\; r(S).$$

Consequently, any anchor-based extension of the form $U_g(w) = U(w_{g(w)}) + L_C \left\|w - w_{g(w)}\right\|_1$ incurs worst-case additive slack at least $L_C\, r(S)$. In particular, for $\mathcal{W} = \Delta^{d-1}$,

$$\inf_{S:\,|S|=m} \sup_{w \in \mathcal{W}} L_C \min_{i \in [m]} \|w - w_i\|_1 \;=\; L_C\, r^\star_m \;\geq\; \frac{L_C}{2}\, m^{-1/(d-1)}.$$

Proof. For each w, $\left\|w - w_{g(w)}\right\|_1 \geq \min_{i \in [m]} \|w - w_i\|_1$. Taking $\sup_{w \in \mathcal{W}}$ yields $\sup_w \left\|w - w_{g(w)}\right\|_1 \geq r(S)$. The slack claim follows by multiplying by L_C, and the final inequality uses Theorem C.6. □

APPENDIX D: WASSERSTEIN DISTANCE AND DISTRIBUTIONALLY ROBUST OPTIMIZATION

This appendix supports the optional preference-shift robustness add-on in Section 5; it does not affect the base post-selection validity chain in Theorem 5.1. We use standard optimal-transport duality; see Villani (2008); Peyré & Cuturi (2019) and Wasserstein-DRO duality results such as Mohajerin Esfahani & Kuhn (2018).

D.0 NOTATION

Let $(\mathcal{X}, d)$ be a metric space and let $\mathcal{P}(\mathcal{X})$ denote Borel probability measures on $\mathcal{X}$. For $p \geq 1$, the (possibly infinite) Wasserstein-p distance between $\rho, \rho_0 \in \mathcal{P}(\mathcal{X})$ is

$$W_p(\rho, \rho_0) \;:=\; \Big(\inf_{\gamma \in \Pi(\rho, \rho_0)} \int d(x, x')^p \, d\gamma(x, x')\Big)^{1/p}, \tag{D.1}$$

where $\Pi(\rho, \rho_0)$ is the set of couplings with marginals (ρ, ρ_0). (Here γ denotes a coupling/transport plan and should not be confused with the RL discount factor γ used elsewhere in the paper.) It is finite whenever ρ, ρ_0 lie in the p-Wasserstein space

$$\mathcal{P}_p(\mathcal{X}) \;:=\; \Big\{\rho \in \mathcal{P}(\mathcal{X}) : \int d(x, \bar{x})^p \, d\rho(x) < \infty\Big\} \tag{D.2}$$

for some (equivalently any) $\bar{x} \in \mathcal{X}$. In our application, $\mathcal{X} = \mathcal{W} = \Delta^{d-1}$ and we use $d(w, w') = \|w - w'\|_1$; since $\mathcal{W}$ is compact under d, we have $W_p(\rho, \rho_0) < \infty$ for all $\rho, \rho_0 \in \mathcal{P}(\mathcal{W})$ and all $p \geq 1$.

Theorem D.1 (Kantorovich–Rubinstein duality)**.** *For $p = 1$ and Polish $(\mathcal{X}, d)$ with $\rho, \rho_0 \in \mathcal{P}_1(\mathcal{X})$,*

$$W_1(\rho, \rho_0) \;=\; \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \Big(\mathbb{E}_\rho[f(X)] - \mathbb{E}_{\rho_0}[f(X)]\Big), \tag{D.3}$$

where $\|f\|_{\mathrm{Lip}} \leq 1$ means $|f(x) - f(x')| \leq d(x, x')$ for all x, x'.

Reference. Theorem D.1 is standard; see, e.g., Villani (2008); Peyré & Cuturi (2019).

Lemma D.2 (Worst-case expectation on a Wasserstein-1 ball)**.** *Let $g : \mathcal{X} \to \mathbb{R}$ be L_g-Lipschitz with respect to d and assume $\rho_0 \in \mathcal{P}_1(\mathcal{X})$ (so the expectations below are finite). Then for any $\eta \geq 0$,*

$$\begin{aligned} \inf_{\rho : W_1(\rho, \rho_0) \leq \eta} \mathbb{E}_\rho[g(X)] &\geq \mathbb{E}_{\rho_0}[g(X)] - \eta L_g, \\ \sup_{\rho : W_1(\rho, \rho_0) \leq \eta} \mathbb{E}_\rho[g(X)] &\leq \mathbb{E}_{\rho_0}[g(X)] + \eta L_g. \end{aligned} \tag{D.4}$$

Proof. If $L_g = 0$, then g is constant and (D.4) holds with equality. Otherwise, g/L_g is 1-Lipschitz, so by (D.3), for any ρ with $W_1(\rho, \rho_0) < \infty$,

$$\mathbb{E}_\rho[g] - \mathbb{E}_{\rho_0}[g] = L_g\Big(\mathbb{E}_\rho[g/L_g] - \mathbb{E}_{\rho_0}[g/L_g]\Big) \le L_g\, W_1(\rho, \rho_0).$$

If $W_1(\rho, \rho_0) \le \eta$, then $\mathbb{E}_\rho[g] \le \mathbb{E}_{\rho_0}[g] + \eta L_g$. Applying the same bound to $-g$ yields $\mathbb{E}_\rho[g] \ge \mathbb{E}_{\rho_0}[g] - \eta L_g$, and taking $\inf$ / $\sup$ gives (D.4). □

D.1 Connecting Lipschitz constants to CertPref certificates

In the main text we take the certificate-derived function $h(w) = L(w) = -U_{\min}(w)$, where the envelope extension is

$$U_{\min}(w) \;\coloneqq\; \min_{i\in[m]} \Big(U(w_i) + L_C \left\|w - w_i\right\|_1\Big). \tag{D.5}$$

Define $f_i(w) \coloneqq U(w_i) + L_C \left\|w - w_i\right\|_1$. Each f_i is L_C-Lipschitz, and for any w, w' we have

$$|U_{\min}(w) - U_{\min}(w')| = \big|\min_i f_i(w) - \min_i f_i(w')\big| \le \max_i |f_i(w) - f_i(w')| \le L_C \left\|w - w'\right\|_1.$$

Hence $U_{\min}$ is L_C-Lipschitz and $h(w) = -U_{\min}(w)$ is also L_C-Lipschitz. (If one instead fixes an arbitrary router and sets $U(w) = U(w_{g(w)}) + L_C \left\|w - w_{g(w)}\right\|_1$, the resulting function can have jumps at routing boundaries; validity still holds but Lipschitzness need not.)

D.2 Applying Lemma D.2 to preference shift

If a deployment distribution ρ satisfies $W_1(\rho, \rho_0) \le \eta$, then applying Lemma D.2 to the Lipschitz function $h(w) = L(w)$ yields

$$\inf_{\rho: W_1(\rho,\rho_0)\le\eta} \mathbb{E}_\rho[L(w)] \;\ge\; \mathbb{E}_{\rho_0}[L(w)] - \eta L_C. \tag{D.6}$$

This is the robustness "shell" reported in Exp4.

D.3 Wasserstein DRO dual form (reference)

For completeness, we record a standard dual formulation for Wasserstein-p DRO: under the regularity conditions in Mohajerin Esfahani & Kuhn (2018),

$$\sup_{\rho: W_p(\rho,\rho_0)\le\eta} \mathbb{E}_\rho[g(w)] = \inf_{\lambda\ge 0} \Big\{\lambda\eta^p + \mathbb{E}_{\rho_0}\big[\sup_{w'\in\mathcal{X}} \big(g(w') - \lambda d(w, w')^p\big)\big]\Big\}. \tag{D.7}$$

We use Lemma D.2 in the main text because it yields a closed-form bound once a Lipschitz constant is available.

Appendix E: Gaussian Processes and Submodular Optimization

This appendix provides details for the greedy anchor-selection statement (Theorem E.3). Anchor selection affects certificate tightness and training cost but does not enter the validity chain. The core facts—closed-form mutual information for Gaussian models, submodularity of GP information gain, and the $(1 - 1/e)$ greedy guarantee for monotone submodular maximization—are classical; see Krause et al. (2008); Nemhauser et al. (1978).

E.0 Setup

Consider a finite candidate set V of preferences. Let $f_V \sim \mathcal{N}(\mu_V, K_V)$ be the restriction of a Gaussian process prior to V. We assume $K_V \succeq 0$ (equivalently, K_V is induced by a positive semidefinite kernel on V). Observations are noisy evaluations $Y_S = f_S + \varepsilon_S$ with i.i.d. noise $\varepsilon_S \sim \mathcal{N}(0, \sigma^2 I)$. Define the information-gain set function

$$F(S) \coloneqq I(Y_S; f_V). \tag{E.1}$$

Lemma E.1 (Closed form of information gain)**.** *For any $S \subseteq V$,*

$$F(S) = \tfrac{1}{2} \log\det\big(I + \sigma^{-2} K_S\big), \tag{E.2}$$

where K_S is the principal submatrix indexed by S.

Reference. Lemma E.1 is standard; see Krause et al. (2008, Sec. 2).

Lemma E.2 (Diminishing returns)**.** *Let $\Delta(v \mid S) \coloneqq F(S \cup \{v\}) - F(S)$. For $A \subseteq B \subseteq V$ and $v \notin B$, one has $\Delta(v \mid A) \geq \Delta(v \mid B)$. Consequently, F is monotone and submodular.*

Proof. Using Lemma E.1 and the matrix determinant lemma,

$$\Delta(v \mid S) = \tfrac{1}{2} \log \frac{\det(I + \sigma^{-2} K_{S \cup \{v\}})}{\det(I + \sigma^{-2} K_S)} = \tfrac{1}{2} \log\big(1 + \sigma^{-2} \sigma^2_{v|S}\big),$$

where $\sigma^2_{v|S}$ is the posterior variance of $f(v)$ given Y_S. Since conditioning reduces variance, $\sigma^2_{v|A} \geq \sigma^2_{v|B}$ for $A \subseteq B$. Because $x \mapsto \tfrac{1}{2}\log(1 + \sigma^{-2} x)$ is monotone, this implies $\Delta(v \mid A) \geq \Delta(v \mid B)$. Nonnegativity of $\sigma^2_{v|S}$ also gives $\Delta(v \mid S) \geq 0$, so F is monotone. □

Theorem E.3 (Greedy anchor selection (mutual information))**.** *Let $F(S) \coloneqq I(Y_S; f_V)$ be the information-gain set function defined in Appendix E.0 and fix a cardinality budget $k \in \mathbb{N}$. Then F is normalized ($F(\emptyset) = 0$), monotone, and submodular, and the greedy algorithm returns a set S_{greedy} with $|S_{\mathrm{greedy}}| \leq k$ such that*

$$F(S_{\mathrm{greedy}}) \geq (1 - 1/e) \max_{|S| \leq k} F(S).$$

E.1 Proof of Theorem E.3

Proof. By Lemma E.2, F is monotone submodular. The Nemhauser–Wolsey–Fisher theorem for monotone submodular maximization under a cardinality constraint (Nemhauser et al., 1978) then yields

$$F(S_{\mathrm{greedy}}) \geq (1 - 1/e) \max_{|S| \leq k} F(S).$$

□

Implementation note. Greedy information gain admits efficient implementations using incremental Cholesky updates for K_S and fast posterior-variance queries (Krause et al., 2008).