CERT-FLOW: Certified Route Planning under Drifting Costs

Conformal certificates, sense-to-certify, and the price of staleness*

Krishi Attri

Abstract

A scout robot routing through terrain whose costs drift — mud after rain, traffic after an incident — faces a question classical replanning never answers: *how good is the current route, given that most of the map is stale?* CERT-FLOW answers it every round with a certificate: a high-probability bound LB $\leq$ OPT $\leq$ UB on the optimal route cost, built from age-weighted non-exchangeable conformal prediction over drift-adjusted residuals, with paid sensing directed at the edges that shrink the certified gap fastest. We prove coverage at the claimed level with a staleness correction that degrades the claim visibly rather than silently; a certifiability *threshold* — a target gap is sustainable iff the sensing rate exceeds the drift rate, so certification is a rate, not a state; a $\sqrt{L}$-tighter sum-aware upper certificate, including the selection-bias hazard it creates and the gate that controls it; and an impossibility theorem showing the certificate's asymmetry is optimal. On replayed traffic from two cities the certificate holds even where real incidents violate the drift model up to half the time, and certificate-directed sensing achieves 2–3$\times$ lower travel-regret than freshness-, uncertainty-, or chance-driven sensing at equal budget. This version adds four measured advances and one honest scoreboard. A non-exchangeable round of conformal upgrades — age-weighted sum-level upper bounds (block-quantile and a drift-retrofitted group-sum construction) — recovers 24–27% of the certified width on real traffic at zero violations, exactly where the union-bound tax lives (long paths); a betting confidence sequence licenses a separately-labelled, a-posteriori tier that is 62% narrower at a measured 0.5% miscoverage; and a weighted conformal test martingale plus a Shiryaev–Roberts detector turn the certificate's pinned-at-one coverage into a *live, alarming* quantity — quiet on twenty of twenty real replay days, firing ~ 7 rounds after an injected shift, at zero cost to the bound. An objective-matched hybrid sensing policy cuts median route regret 41% on real traffic in the regime where no width certifies. A certified multi-agent extension lifts the certificate to fleets: the additive team bound is sound, with an exactly separable team optimum, and a first certified-MAPF study (CBS over certified corridors, 600 runs) executes with *zero* collisions where point-estimate planning collides on up to 100% of runs — alongside the honest finding that its probabilistic knob is inert at this scale, where the finite-sample floor caps the supportable team level below the nominal one. Throughout, we keep the negatives: a verdict table names the areas CERT-FLOW wins (coverage, observability, bounded-change absorption) and the areas it does not (raw static-map latency, interval width).

1 Introduction

Figure 1 shows one round of the planner this paper builds and certifies.

Incremental replanners repair shortest paths quickly when the map changes [17, 22], informative path planners decide where to sense [29], and learned traversability models supply ever-better cost priors. What none of these provide is an online, sound answer to the operational question: *is the route I am about to execute within ε of the best currently-achievable route, and how confident am I, given that my last look at most of the map is minutes old?* The gap is not any single component. Measured suboptimality bounds exist for anytime search over *known* costs [21]; pay-to-sense edge resolution exists for *binary* blockages without certificates [4]; PAC path identification exists for *stationary* samplable costs [5]; and staleness-aware maps exist without route-quality guarantees [19, 30]. The intersection — a certificate on route cost that drives physical sensing, under continuously drifting, partially observed costs, maintained online — is, to our knowledge, unoccupied

*Revised and expanded version (July 2026) of the CERT-FLOW preprint first posted March 2026 (engrXiv, DOI 10.31224/7306); the additions over the first version are summarised in the introduction.

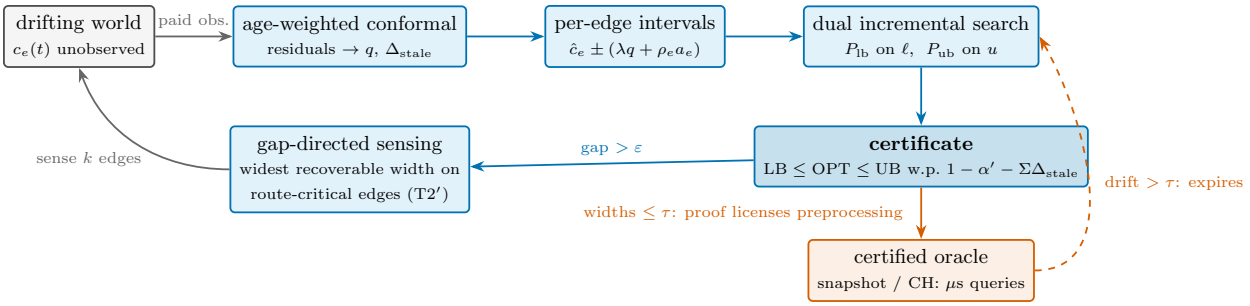

Figure 1: One CERT-FLOW round. Paid observations feed an age-weighted non-exchangeable conformal scorer whose quantile prices noise (q) and staleness ($\rho_e a_e$) into per-edge intervals; two incremental searches bound the optimum from both sides, yielding a certificate that either certifies (gap $\leq \varepsilon$) or directs the next observations at the edges that shrink it fastest. When the certificate proves every interval within τ, that proof licenses lookup-speed preprocessed queries (Section 6.7) — revoked the moment drift exceeds τ.

(Table 1 audits this claim family by family). The setting is the daily reality of deployed mobile robots: a scout vehicle whose off-road traversal costs change with weather and erosion, a delivery fleet whose street segments drift with traffic, an inspection UGV that must decide whether to commit to a corridor or spend another ranging scan first. In each case the operator's question is not "what is the shortest path on my map?" but "can I trust the route my stale map implies, and if not, where do I look?" — which is precisely the certificate and the sensing rule CERT-FLOW provides.

CERT-FLOW occupies it with a deliberately asymmetric design. Each edge carries a point estimate, an observation age, and a drift-rate bound; conformal calibration over drift-adjusted residuals converts these into per-edge cost intervals whose coverage guarantee degrades *explicitly* with both the miscoverage target and a staleness correction. Two incremental searches maintain the optimistic and conservative shortest paths; their costs are the certificate. Sensing is pointed at the certified gap: the planner observes whichever route-critical edge yields the largest expected gap reduction per unit cost, with an age-triggered round-robin backstop that converts greed into a guarantee. The robot executes only conservative incumbents, and a conductivity hysteresis selects among certified-equivalent incumbents to suppress solution churn without touching the certificate.

The framework forced four findings we did not anticipate at design time. (i) *Certification is a rate*: a certificate self-extinguishes even in a static world unless maintenance sensing sustains it — the certifiability threshold (Theorem T2′) makes the required rate explicit, including an impossibility direction. (ii) *Adaptive and static guarantees fight*: an adaptive conformal layer silently cancels a provable widened margin; the provable mode must freeze it. (iii) *Tighter bounds incur new hazards*: replacing the Bonferroni union bound with a sum-aware block quantile (Theorem T4) tightens the gap by $\Theta(\sqrt{L})$ but exposes the optimizer's winner's curse — measured as a coverage drop from 1.000 to 0.823 — which a freshness-gate protocol controls. (iv) *Locality and staleness conflict*: age-driven widening touches every edge every round and destroys incremental-search locality; a lazy pre-widening cache restores it with a soundness proof and a quantified width price.

Contributions. (1) The problem formulation and certificate construction (Sections 2–3). (2) Theorems T1a/T1b (coverage with explicit staleness correction; observable vs. latent semantics), T2′ (certifiability threshold, both directions), T4 (sum-aware upper certificate) with the freshness-gate protocol, and the asymmetry theorem T5 (Section 4). (3) Certificate-gated preprocessing: a *proof*, rather than an assumption, licenses lookup-speed query structures, which expire automatically when drift exceeds their validity window — this is what lets a certified planner reach the microsecond query class of static-known route planners (Section 6). (4) The maintenance-sensing and lazy pre-widening mechanisms. (5) A four-tier empirical study with coverage validated against ground truth in simulation and on replayed real traffic, including misspecification, off-model, and boundary stress (Section 6).

This version adds five contributions that follow the same discipline — soundness first, every negative kept. (6) A non-exchangeable *round two* of the conformal layer: age-weighted sum-level upper certificates (a block quantile and a drift-retrofitted group-sum construction after Luo and Zhou [23]), an LP-shift staleness

model [1], and a joint per-edge max-score price [18], benchmarked head-to-head on real traffic — recovering the union-bound width tax exactly where it is largest and reporting, in the same table, the construction that made the certificate *wider* (Section 6.12). (7) A two-tier certificate: the a-priori bound is never shrunk, while a betting confidence sequence [36] licenses a separately-labelled a-posteriori radius, with a proof of why an a-priori shrink under drift is impossible (Section 6.12). (8) An observability layer that makes the pinned coverage testable — a weighted conformal test martingale and a Shiryaev–Roberts detector [28], e-value merging [12] — validated live on real traffic (Section 6.13). (9) Objective-matched hybrid sensing, which wins on route quality in the never-certifiable regime while leaving the certificate untouched (Section 6.14); and the honest companion that on static maps there is nothing to cross over to (Section 6.15). (10) A certified multi-agent extension: the additive team certificate, and a pre-registered certified-MAPF study that executes with zero collisions but exposes an inert probabilistic knob (Section 7). A single verdict scoreboard (Section 8) states, per area, where CERT-FLOW wins and where it loses.

2 Problem Setting and Assumptions

A directed graph $G = (V, E)$, start s, goal g; unknown time-varying costs $c_e(t) > 0$; one paid noisy observation per round of length Δ. The planner must report, every round, a certificate $(\mathrm{LB}, \mathrm{UB}, \text{confidence})$ with $\mathrm{LB} \le \mathrm{OPT}(t) \le \mathrm{UB}$ at the stated confidence, execute only conservatively validated routes, and stop sensing when $\mathrm{UB} - \mathrm{LB} \le \varepsilon$.

Assumptions (stated fully in the theory companion; all are swept or stress-tested in Section 6): **A1** bounded drift $|c_e(t') - c_e(t)| \le \rho_e |t' - t|$ with ρ_e conservatively known; **A2** the observation-residual distribution drifts in total variation at rate at most $\varepsilon_{\mathrm{TV}}$; **A3** symmetric unimodal observation noise (needed only for the latent-cost and sum-aware variants); **A4** edges sharing a calibration buffer share a noise family.

3 The CERT-FLOW Planner

Certificate substrate. Re-observing edge e at age a yields the drift-adjusted score $R = |Y - \hat{c}_e| - \rho_e a$; scores enter a rolling buffer with age-geometric weights $w_i = \rho_w^{\mathrm{age}_i}$ (data-independent, as the non-exchangeable guarantee requires [2]). The weighted $(1-\alpha)$ quantile q gives intervals $\hat{c}_e \pm (\lambda q + \rho_e a_e(t))$; $\lambda = 1$ certifies observables (T1a), $\lambda = 2$ certifies latent costs (T1b). The reported confidence is $1 - \alpha' - \sum_e \Delta_{\mathrm{stale}}$ — staleness degrades the claim visibly. (The theory companion's repaired accounting adds a calibration-age term π_{cal} and a second-noise-draw factor to Δ_{stale}; both are below the reported precision at the swept $\varepsilon_{\mathrm{TV}}$ and the small calibration ages the sensing loop maintains, and the implementation tracks $\langle \rho a \rangle_{\mathrm{cal}}$ so π_{cal} is reportable per run.) An adaptive-conformal tracker [14] supplies an assumption-free long-run safety net; it is frozen in the provable mode (Section 6.1, ACI-interaction finding).

Dual incremental search. Two D* Lite instances maintain the ℓ-shortest and u-shortest paths; LB is the ℓ-cost of the former, UB the best u-cost among candidates (any path's u-cost upper-bounds OPT on the coverage event). Age widening would touch every edge every round; a *lazy pre-widening* cache computes metrics at age $+B\Delta$ so entries stay conservatively valid for B rounds, restoring repair locality at width price $2\rho_e B\Delta$ per edge (Lemma, theory companion).

Sense-to-certify. Since $\mathrm{UB} - \mathrm{LB} \le \sum_{e \in P_{\mathrm{lb}}} (u_e - \ell_e)$, sensing only the optimistic path's edges suffices to control the gap. The selector takes the route-critical edge maximizing expected gap recovery $(2\rho_e a_e)$ per sensing cost; an age-triggered backstop forces round-robin re-observation, which is what makes the achievability theorem apply to the deployed policy. Certified rounds still sense on projected certificate expiry and a calibration-freshness floor (maintenance sensing); without it the claim self-extinguishes even in static worlds.

Hysteresis and the freshness gate. Among incumbents whose u-cost is within a slack of UB, a decaying edge-conductivity κ selects the stickiest — churn drops 70% at zero certificate cost (the bound reported is

unchanged; only the executed selection moves). The same stability opens the *freshness gate* for the sum-aware upper certificate (T4): the $\sqrt{L}$-margin bound applies only when every incumbent edge has been re-observed since the path became incumbent, which restores the fixed-path premise that the optimizer's selection otherwise violates.

4 Theory

The full statements and proofs are reproduced self-contained in Appendix A (the asymmetry pair T4/T5 and the version-2 results T8–T10 in full, the classical results T1–T3/T6/T7 as statements with proofs in the theory companion `theory.pdf`), so this document stands alone; we summarize the main results here. **T1a/T1b (coverage):** per-edge intervals cover the next observation at level $1-\alpha-\Delta_{\text{stale}}-\pi_{\text{cal}}$ under A1–A2, and the latent cost at a doubled margin and level $1-2\alpha-2\Delta_{\text{stale}}-2\pi_{\text{cal}}$ under A1–A3 via Anderson domination (π_{cal}, the calibration-age slack, enters both statements — once through the score sandwich and, for T1b, again through the fresh-noise step); Δ_{stale} is computed from realized calibration ages in a corrected two-draw form, sharpening the age-uniform corollary of Barber et al. [2]. **T2′ (certifiability threshold):** round-robin sensing of the L optimistic-path edges sustains $\text{UB}-\text{LB} \le 2L\bar{q}+\bar{\rho}\Delta L(L-1)$ at post-observation instants; conversely no sensing policy can bring *this certificate's* gap (interval widths $2\lambda q+2\rho a$, before cost-floor clipping) below $2mq_{\min}+\rho_{\min}\Delta m(m-1)$ across an m-edge cut, and $\varepsilon < 2mq_{\min}$ is unattainable by it at any sensing rate — an impossibility for the shipped construction, not for every conceivable sound certificate. Static worlds and the deterministic scout stopping rule of Rockenbauer et al. [29] are the $\rho \to 0$ and $q \to 0$ corners. **T4 (sum-aware upper certificate):** block-conformal calibration of signed deviation sums gives a fixed-path upper bound with $\Theta(\sqrt{L})$ margin replacing Bonferroni's $\Theta(L\, q_{\alpha'/L})$; the certificate becomes asymmetric because the lower bound must hold uniformly over paths. The fixed-path premise fails for optimizer-selected incumbents (winner's curse) and is restored conditionally by the freshness gate. **T5 (LB impossibility):** the asymmetry is forced — on layered graphs any valid uniform lower bound sits, with probability $\ge 1-\alpha-o(1)$, at least $(\sqrt{2}-1-o(1))\, L\sigma\sqrt{\ln w}$ below the posterior-mean optimum (a polymer-constant comparison of the posterior-mean and true-cost fields; selection over exponentially many paths is the mechanism), which per-edge Bonferroni matches up to logarithmic factors. **T6 (decision-uniform validity):** per-round claims are marginal; a policy that acts on certificates selects rounds. For decision schedules fixed in advance (or predictable from data independent of the calibration buffer), α-spending over the (few) decision instants restores trajectory-level validity exactly where the trajectory consumes it — while full per-round time-uniformity is quantifiably impractical ($n \gtrsim 63$k calibration scores at Bonferroni levels). For *certificate-triggered* acting — the deployed semantics — the spending bound does not compose (acting is positively correlated with miscoverage), and the honest routes are horizon spending or the anytime-valid machinery of T9b; the companion states exactly what is and is not guaranteed. **T7 (churn-measured floor):** under drift the optimistic path hops over a churn set of $K \ge L$ edges; the certifiability floor and the adaptive sensing rate must use the online-tracked $\widehat{K}$, and focused sensing *suppresses* churn ($K\colon 59 \to 11 \approx L$ measured) rather than chasing it.

The version-2 theory adds three results, stated and proved in the companion. **T8 (non-exchangeable sum-level pricing).** A group-sum score after Luo and Zhou [23] and a joint per-edge max score [18] both price a length-L path at a single level-α' quantile rather than the α'/L Bonferroni quantile, and both inherit the age-weighted coverage of Barber et al. [2]; an LP-shift staleness model [1] replaces the TV-Lipschitz Δ_{stale} with a worst-case quantile $\text{Quant}(1-\alpha+\rho)+\varepsilon$ that also prices never-seen edges. The *width* consequence is empirical and two-signed (Section 6.12): sums calibrate where per-edge maxima starve. **T9 (a-priori shrink is impossible; a two-tier certificate).** No windowed evidence can license an a-priori narrower next-round bound under drift without reassuming the exchangeability the drift model exists to drop; a betting confidence sequence [36] instead licenses an anytime-valid, self-revoking *a-posteriori* radius on the observed stream — a different, weaker claim, kept in its own tier. **T10 (team and multi-agent certificates).** For N agents on one shared conformal store the per-agent certificates add, $\sum_i \text{LB}_i \le \sum_i \text{OPT}_i \le \sum_i \text{UB}_i$, with union-bound confidence (the team optimum decomposes exactly by definition of the uncoupled objective; each LB_i stays conservative); lifting the construction to collision-free multi-agent execution (CBS over certified corridors) is sound on a coverage event budgeted over a selection-free priced universe — not the returned plan's edges, which CBS selects — and the honest scope — soundness yes, completeness no, and a finite-sample-inert α

knob at small scale, where the supportable team level (≈ 0.71 at P0 calibration) sits below the nominal 0.9 — is stated with the proof (Section 7).

5 Related work

CERT-FLOW sits at the intersection of four research lines that have not previously been combined: search with cost certificates, sensing allocated to resolve path decisions, conformal prediction for graph costs, and temporal map models. We organize the discussion by these lines and state, for each, the precise position CERT-FLOW occupies. We do not claim to be the first certified planner, the first active-sensing replanner, the first staleness model, or the first bio-inspired memory; the contribution is the conjunction, together with the staleness correction Δ_{stale} and the certifiability threshold $T2'$ as new analytical objects.

Certified and bounded search. TASP [37] computes explicit lower and upper bounds on path cost from bounded estimators, but its bounds tighten through additional computation over static costs, with no sensing. Anytime search such as ARA* [21] and AD* [22] reports a measured suboptimality certificate, yet assumes known costs and closes the optimality gap by spending more search effort rather than by observing the world. Robust interval shortest-path formulations carry cost intervals through the search but treat the intervals as given and fixed. CERT-FLOW differs on the source and target of the bound: its certificate is a high-probability statement about an *unknown, drifting* cost, and the gap is closed by paid observation rather than by computation.

Sensing to resolve a path. The Canadian Traveller Problem with remote sensing [4] pays to sense edges and places sensing by value of information, but optimizes expected cost over binary blockages and produces no cost certificate and no drift model. Edge-evaluation planners — BISECT, LazySP, and Generalized Lazy Search [6, 7, 24] — decide which edge to evaluate near-optimally, but evaluation is a computational query returning binary validity, not a cost interval. PAC combinatorial pure exploration [5] returns an (ε, δ)-certified best path by sampling arms, and InfoBAX [26] senses to identify a function property such as a shortest path; both assume a stationary problem, neither trades sensing cost against travel, and neither maintains an execution loop. CERT-FLOW inherits the "which edge to sense" question from this line but answers it against a different objective: shrink a probabilistic cost *certificate* on a drifting graph while executing.

Scout and informative path planning with guarantees. Traversing Mars [29] scouts until a path is proven optimal or infeasible, using a deterministic stopping rule with no probabilistic bound, no staleness, and no incremental reuse; CERT-FLOW recovers that rule as the degenerate case $\rho \to 0$, $q \to 0$ of $T2'$, and so strictly generalizes it. Informative path planning with guaranteed estimation uncertainty [16] certifies the variance of a sensed field uniformly over space; CERT-FLOW instead certifies the route decision, which depends on cost differences along candidate paths rather than on the field everywhere.

Temporal and staleness map models. FreMEn [19] models periodic environment dynamics in the frequency domain, the persistence filter [30] estimates how long a feature remains valid, and BRULE [11] and persistent monitoring [32] treat revisitation as a resource over time-varying state. These works represent staleness as map or feature uncertainty. None couples observation age to an inflation of a cost interval and propagates that inflation into a route certificate, which is the mechanism Δ_{stale} and the $\rho_e a_e$ widening provide.

Conformal prediction in robotics. Conformal prediction beyond exchangeability [2] supplies the non-exchangeable coverage bound that Δ_{stale} instantiates, and adaptive conformal inference [14] supplies the assumption-free long-run safety net. Perceive with Confidence [25] applies conformal prediction to perception for safe planning, but on detections rather than edge costs. Two adjacent works are the closest neighbors. Luo and Zhou [23] build conformal intervals directly on sums of edge labels — path cost — through a sum-aware nonconformity score, avoiding the Bonferroni penalty, but only under exchangeability, with no drift, weights, sensing, or online loop. CQR-GAE [34] produces conformal edge intervals for robust shortest path,

prior family	drift-aware intervals	path-cost certificate	online incremental	gap-directed paid sensing
D* Lite / AD* [17, 22]	×	×	✓	×
LazySP / GLS [7, 24]	×	×	✓	×
CTP + sensing [4]	×	×	×	✓
Conformal sums (CIA) [23]	×	✓	×	×
CQR-GAE [34]	×	✓	×	×
TASP [37]	✓	×	×	×
FreMEn [19]	✓	×	×	×
IPP / BAX [16, 26, 29]	×	×	×	✓
E-Graphs / stigmergy [27, 10]	×	×	✓	×
CERT-FLOW (this work)	✓	✓	✓	✓

Table 1: Each prior family misses at least two of the four properties whose conjunction CERT-FLOW occupies. "Drift-aware intervals": uncertainty that grows with observation age under an explicit drift model (exchangeable conformal constructions do not qualify — Figure 5 measures the consequence). "Path-cost certificate": a coverage guarantee on LB $\leq$ OPT $\leq$ UB, not a heuristic bound on point estimates.

but assumes exchangeability, runs one-shot, and propagates no path-level coverage and no sensing. CERT-FLOW uses the non-exchangeable bound to handle drift, optionally adopts the sum-aware score under those weights (the $T4$ upper certificate), and closes the loop with sensing and incremental maintenance.

Non-exchangeable conformal, round two. The version-2 conformal layer draws on four 2025–26 lines and adopts each only where it stays distribution-free under drift. Lévy–Prokhorov robust conformal prediction [1] gives a worst-case quantile under combined local and global shift; we take it as an optional staleness model that also prices never-seen edges. The group-sum construction of Luo and Zhou [23] and the pipeline-aware joint max-score of PASC [18] both replace the α'/L union with a single level-α' quantile; we retrofit both with age weights and, crucially, *measure* them against each other on real traffic, where the sum calibrates and the max starves (Section 6.12). The observability layer is new to route certificates: WATCH weighted-conformal test martingales [28] and conformal e-values [12] turn the pinned coverage into a monitorable quantity. What none of these supplies, and CERT-FLOW adds, is the online drift-and-sensing loop that consumes them: they are calibration and monitoring tools; here they price and audit a route certificate that acts.

Multi-agent path finding. Conflict-based search [31] is the standard optimal MAPF solver; robust variants (k-robust CBS, probabilistic and expected-delay CBS) hedge execution uncertainty with fixed or parametric delay models. None certifies team cost under *drifting* costs, and execution uncertainty is a named open challenge in the lifelong-MAPF agenda. CERT-FLOW's multi-agent extension (Section 7) runs CBS over *certified* space-time corridors: the same age-weighted conformal windows that certify a single route become the moving obstacles the team plans around, so the collision guarantee is the coverage event rather than a tuned safety margin. The additive team certificate is the standard separable-objective bound made sound over one shared conformal store.

Memory and experience reuse. Experience graphs [27] reuse prior solution paths to speed search, and ACO and Physarum models [10, 35] reinforce decaying edge scalars as a search heuristic. CERT-FLOW's optional conductivity module κ occupies a narrow delta: it applies observation-age decay to warm-start a guaranteed incremental search, and is held outside the certificate by construction, so it can never affect coverage. This module is demoted to an ablation.

The unoccupied intersection. Table 1 makes the claim auditable: each neighbor misses at least two of the four properties {drift-aware weighted intervals, path-level coverage propagation, online incremental maintenance, sensing allocated to the certified gap}. The cell CERT-FLOW occupies is their conjunction: non-exchangeable, age-weighted conformal edge-cost intervals under an explicit drift model, propagated to a path-cost certificate LB $\leq$ OPT $\leq$ UB, maintained online by incremental D* Lite, with sensing pointed at shrinking the certified gap and a threshold that states when no target gap is sustainable.

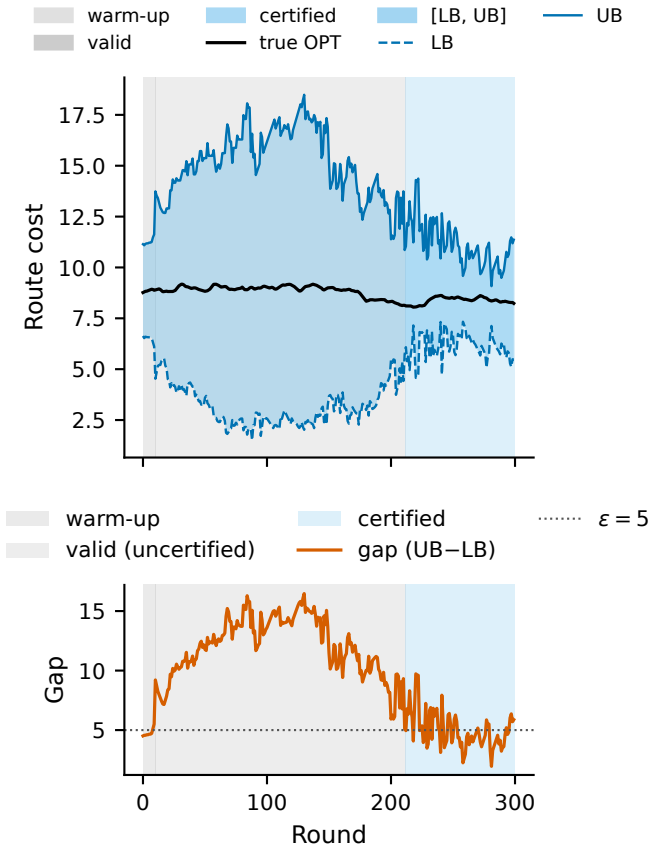

Figure 2: Certificate gap trajectory for one representative episode (6×6 grid, bounded drift ρ=0.02, `seed=3`, default CERT planner). *Top*: true optimal cost OPT (black) overlaid on the [LB, UB] certificate band (blue fill). *Bottom*: certificate gap UB−LB (red) and the target ε=5 (dotted). Shading: grey = warm-up (annealing claim below target), mid-grey = valid-uncertified, blue = certified (gap$\leq\varepsilon$). Single episode shown; aggregate over 25 seeds × 300 rounds is reported in Table 2.

6 Experiments

The experiments follow the claim ladder. We first establish that the certificate covers the true optimum at the claimed rate and that the certifiability threshold is visible (Section 6.1). We then tighten the bound with the sum-aware upper certificate (Section 6.2) and audit its calibrated building block in isolation (Section 6.3). With the certificate validated, we measure travel regret against sensing baselines in unknown terrain (Section 6.4), confirm coverage on replayed real traffic (Section 6.5), and locate the claim's boundaries on standard pathfinding maps and at road scale, where certificate-gated preprocessing reaches the microsecond query class (Sections 6.6–6.7). Incremental-repair latency, the component ablations, lifelong operation, and the churn-measured floor close the section (Sections 6.8–6.11). Coverage is a model-conditional claim, validated only in simulation and on recordings where an oracle runs Dijkstra on the true costs every round; reported coverage is the empirical rate among valid rounds, with Clopper–Pearson intervals. Throughout, results are stated in the present tense and numbers are medians over the seeds and rounds named in each caption. Every experiment below is regenerated by a single driver script under `scripts/` (e.g. `run_tier0.py`, `run_tier2.py`, `run_metr_la.py`, `run_movingai.py`, `run_ablations.py`, `run_lifelong.py`, `run_gaussian_break.py`), with per-experiment findings logged under `docs/results/`.

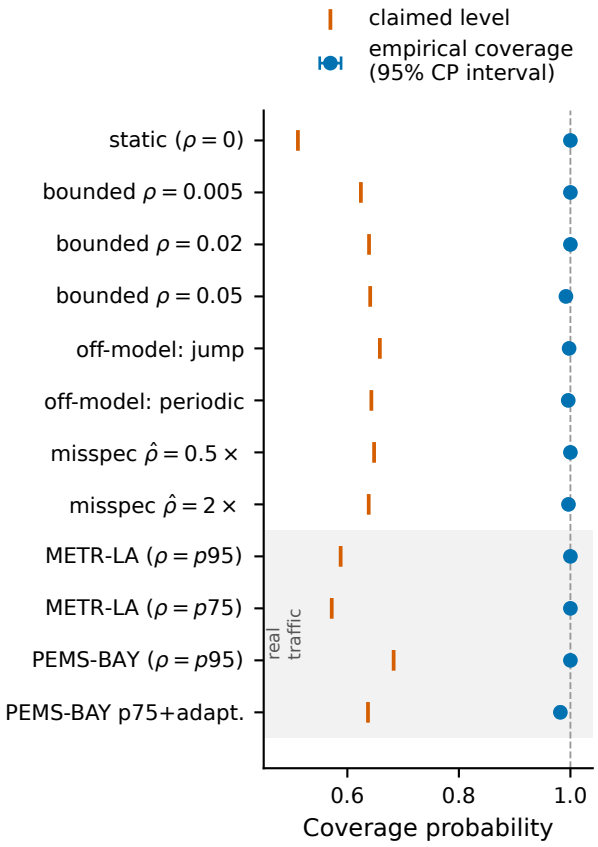

Figure 3: Empirical coverage (filled circles, 95% Clopper–Pearson bars) vs claimed confidence level (red tick marks) for eight synthetic conditions (white background) and four real-traffic conditions (grey background; METR-LA 20 days, PEMS-BAY 20 days). Synthetic: 25 seeds × 300 rounds, 6×6 grid, α'=0.2. Coverage equals or exceeds the claim in every row; no row falls below the claimed level.

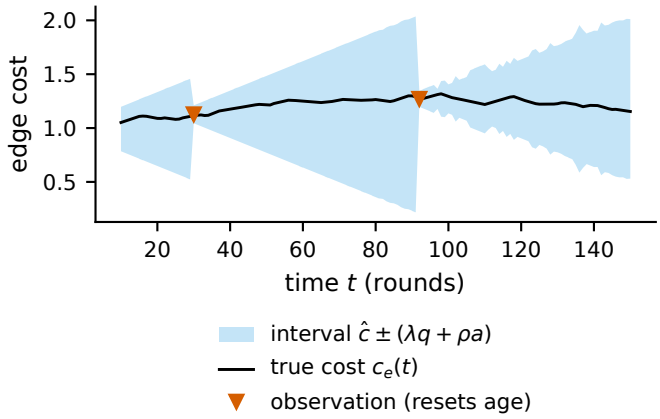

Figure 4: The certificate's building block on one edge (bounded drift ρ=0.02, seed 7): the interval $\hat{c}\pm(\lambda q+\rho a)$ (blue band) widens linearly with age a and snaps tight when the edge is re-observed (red markers); the true cost (black) wanders within. Staleness is priced, not ignored.

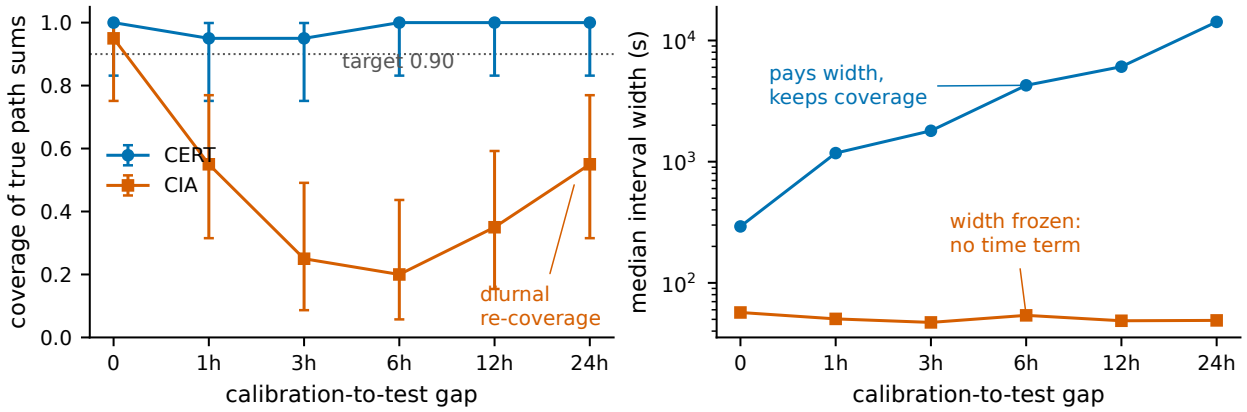

Figure 5: Exchangeable conformal path sums (CIA [23], their construction extracted and run on METR-LA; 50 paths × 20 repetitions per gap, target 0.90) vs CERT-FLOW as the calibration-to-test gap grows. *Left*: CIA covers at gap 0 (its home setting) and collapses to 0.20–0.25 at the 3–6 h staleness common in operation; the partial 24 h recovery is the diurnal cycle returning the network near its calibration state — the failure mode is drift, not noise. CERT-FLOW holds 0.95–1.00 at every gap. *Right*: the price, paid explicitly — CIA's width is frozen (∼50 s, no time-dependent term) while CERT-FLOW's ρ·gap widening grows. Error bars: 95% Clopper–Pearson.

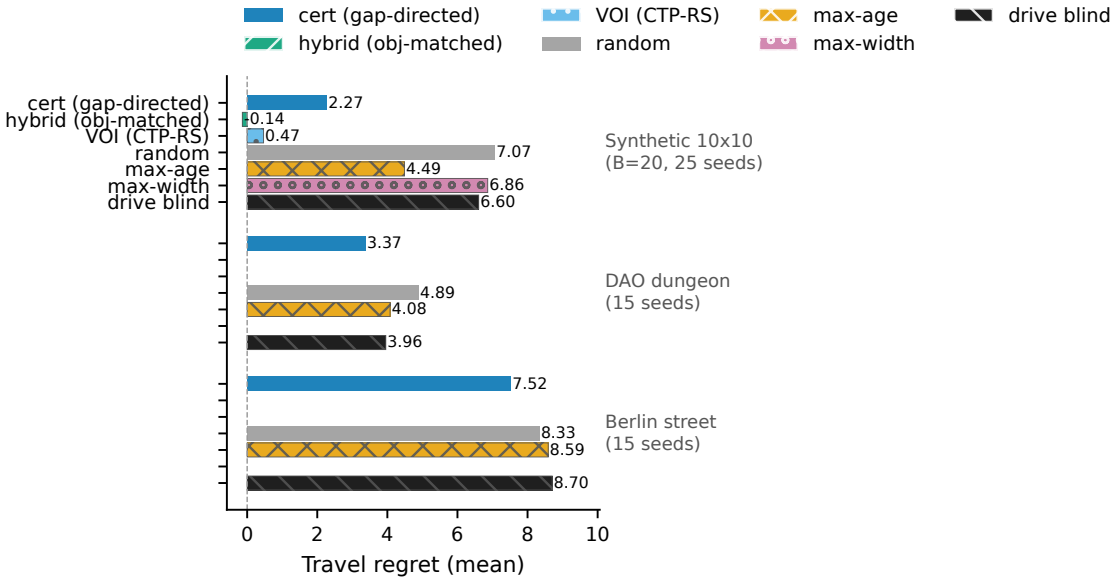

Figure 6: Travel regret (mean, lower is better) by sensing policy across three map groups. *Synthetic*: 10×10 bounded drift ρ=0.02, budget B=20, 25 seeds; includes hybrid (objective-matched, cert+VOI) and VOI (CTP-RS expected-route) policies from the external-baseline run (15 seeds). *DAO dungeon / Berlin street*: MovingAI maps, 15 seeds, bounded drift ρ=0.02, B=20. Blank bars indicate conditions not run for that dataset. Regret is against a clairvoyant oracle replanning on true costs every step.

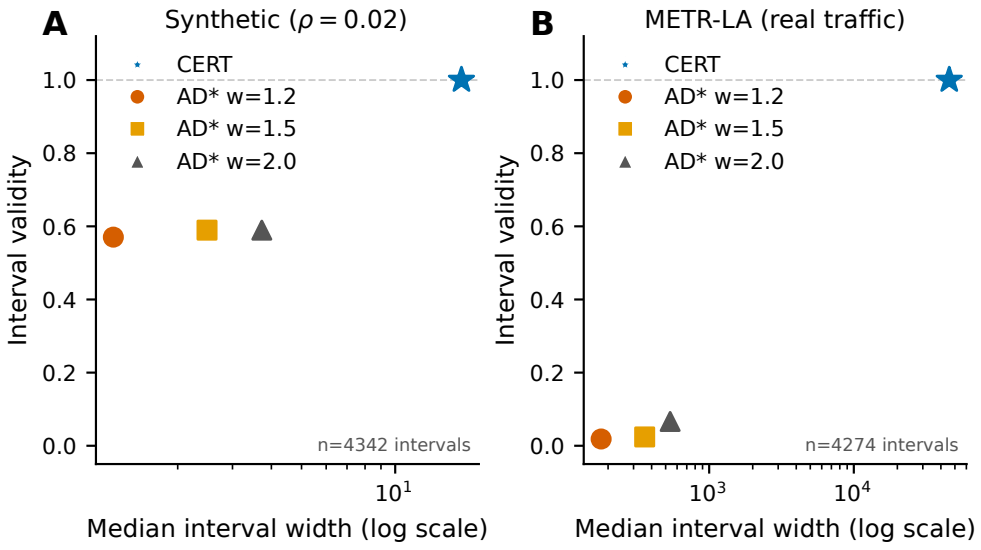

Figure 7: Interval validity (fraction of rounds the true OPT lies inside the reported interval; y-axis) vs median interval width (x-axis, log scale) for CERT and three AD*-style inflation widths ($w \in \{1.2, 1.5, 2.0\}$), evaluated on a shared stale observation stream (neutral max-age sensing). **A**: synthetic bounded drift ρ=0.02 (15 seeds; $n = 4340$ intervals per bound). **B**: METR-LA replayed traffic (15 seeds; $n = 4273$). Narrow-and-wrong vs wide-and-sound is the observed trade-off: AD* semantics hedge search suboptimality, not map staleness.

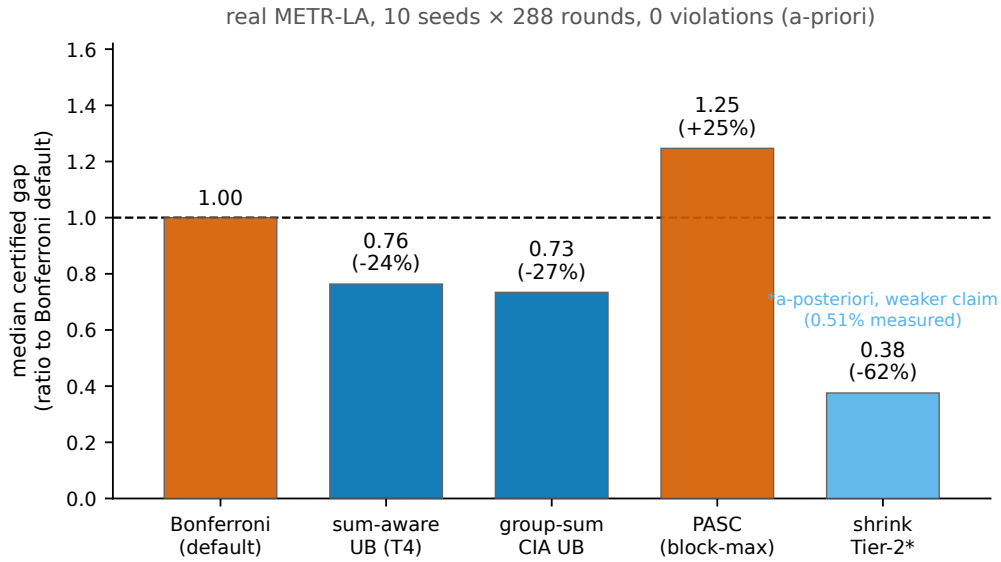

Figure 8: Sum-level upper-bound calibration recovers the union-bound width tax on real METR-LA (10 seeds × 288 rounds, n=2729 paired valid rounds). Bars: median certified gap as a ratio to the shipped per-edge Bonferroni default; blue = tighter, orange = wider, at 0 violations against true OPT for every a-priori mode. The two *sum*-level constructions (sum-aware block quantile, group-sum CIA UB) tighten 23.6–26.6%; PASC's per-block *max* starves on long paths ($L \approx 14$–18) and lands +24.7% wider — sums calibrate where maxima starve. The Tier-2 shrink bar (light, −62.4%) is a-posteriori and carries a different, weaker claim (0.51% measured miscoverage). Numbers: `scripts/out/width_attack.json`; Section 6.12.

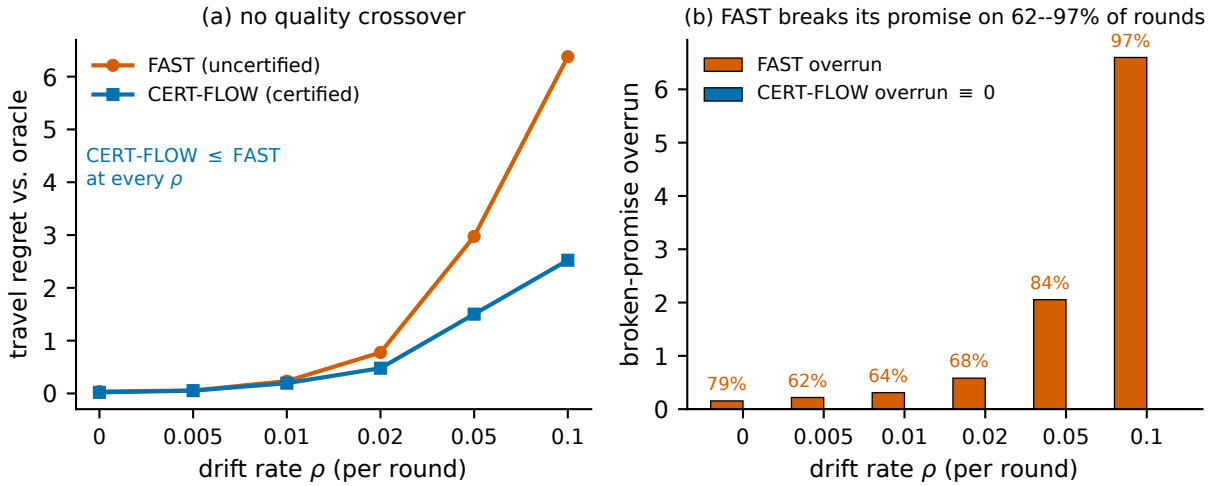

Figure 9: No quality crossover between the certified planner and a fast uncertified replanner (12×12, 15 seeds, 220 rounds, paired ground truth). *(a)* CERT-FLOW's travel regret is ≤ FAST's at every drift level, including the static map (ρ=0), where FAST's point cost, a min over noisy estimates, is optimistically biased. *(b)* FAST's point-estimate promise is exceeded on 62–97% of rounds (labels), with overrun magnitude growing 0.15 → 6.6 cost units, while the certified upper bound is never exceeded (overrun ≡ 0). FAST's only advantage is latency (2–4× faster/round). Numbers: `crossover_regret.json`; Section 6.15.

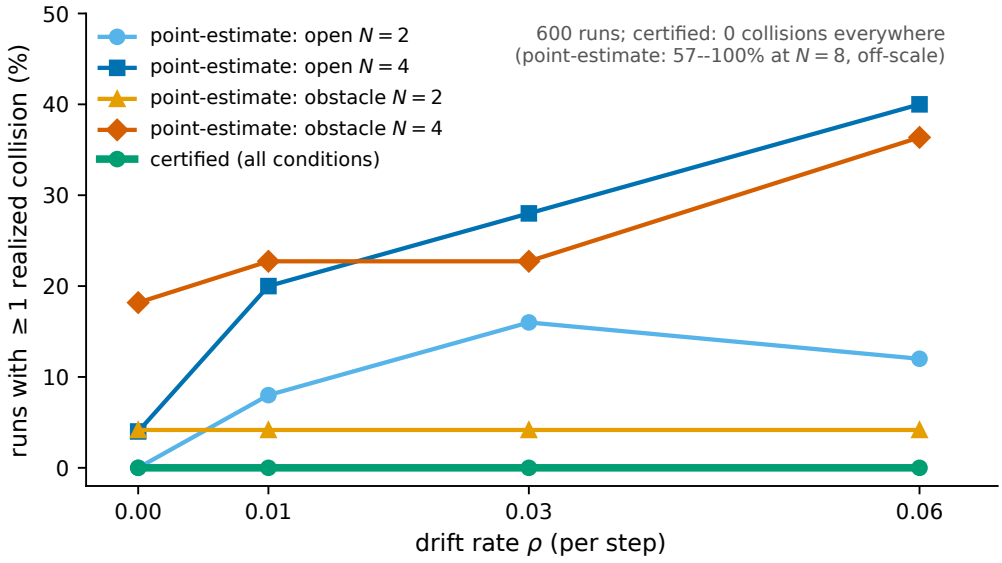

Figure 10: Certified-MAPF P0: CBS over certified corridors executes with *zero* realized collisions on every solved run in the entire 600-run sweep (green), while point-estimate MAPF (coloured lines) collides on an increasing fraction of runs as drift ρ and team size N grow — 4% at low density to 40% at N=4, ρ=0.06, and 57–100% at N=8 (off-scale). The certificate's collision guarantee is the coverage event, not a tuned margin. The honest companion finding (the α knob is inert at this scale) is in Table 14. Numbers: `certmapf_p0.json`; Section 7.

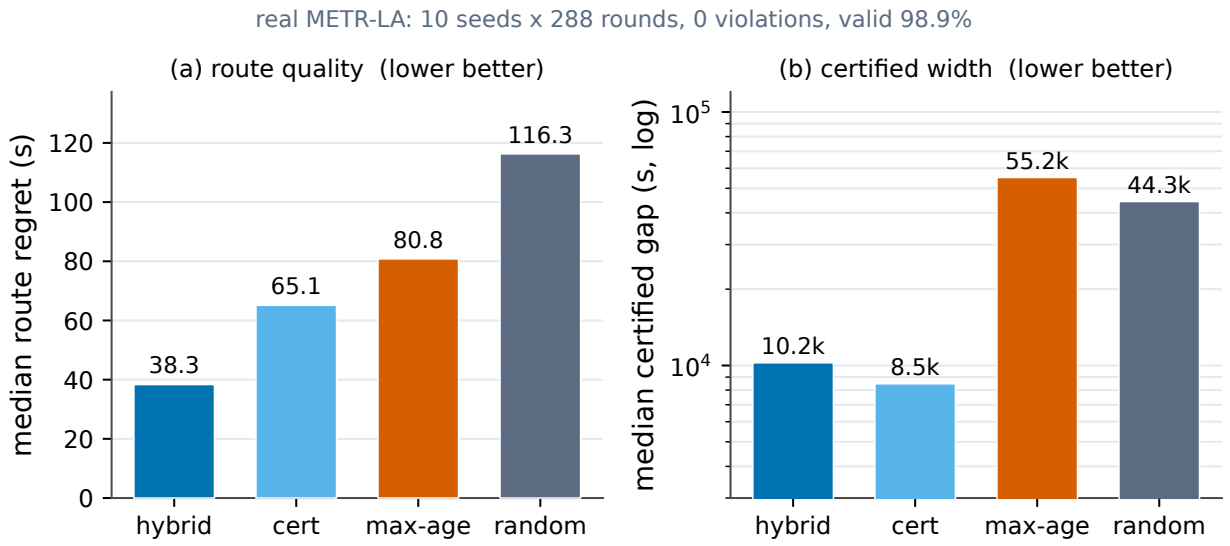

Figure 11: Objective-matched hybrid sensing on real METR-LA (10 seeds × 288 rounds; Table 11 as a figure). *(a)* median route regret against the clairvoyant oracle: hybrid 38.3 s vs the pure gap-directed default 65.1 s (−41%), and far below the freshness (`max_age`) and random baselines. *(b)* median certified gap (log axis): hybrid pays a +21% wider gap than pure gap-directed (10.2k vs 8.5k s) — the honest cost — but dominates `max_age`/random, which run 5–6.5× wider. All four policies hold coverage at 0 violations (valid 98.9%); in this regime no width certifies at ε, so the wider gap is not decision-relevant while departure quality is. Numbers: `hybrid_real_metrla.json`; Section 6.14.

6.1 Tier 0: coverage and the certifiability threshold

Table 2 reports 25 seeds × 300 rounds per condition on 6 × 6 grids with $\varepsilon = 5$, $\alpha' = 0.2$, $\varepsilon_{\mathrm{TV}} = 10^{-4}$ unless noted, with maintenance sensing and lazy pre-widening enabled. Coverage is measured against the claimed line $1 - \alpha' - \sum \Delta_{\mathrm{stale}}$ (the implementation's single-draw Δ_{stale}; the theory companion's corrected two-draw form and its π_{cal} term differ from it below the reported precision at $\varepsilon_{\mathrm{TV}} = 10^{-4}$). Figure 2 shows one representative gap trajectory and Figure 3 the coverage-versus-claim summary.

Numbers use α-annealing (warm-up rounds carry the best currently-supportable claim, which tightens toward $1 - \alpha'$ as evidence accrues): validity is $\sim 96\%$ in every condition — including the full provable mode, whose earlier validity cost (9% valid pre-annealing) is resolved — and certification roughly doubles at every drift level. Coverage sits above the claimed line in every condition, including the off-model jump and periodic stresses; miscoverage appears only in the hardest settings. The certifiability threshold T2′ is visible: cert% falls monotonically with drift severity, and the jump condition certifies nothing, which is the correct refusal. Underestimating ρ ($\hat{\rho} = 0.5\times$) does not break coverage, since the conformal layer absorbs it; overestimating ($\hat{\rho} = 2\times$) costs conservatism only. Misspecifying the noise-drift assumption A2 ($\varepsilon_{\mathrm{TV}} = 10^{-3}$) self-extinguishes the claim loudly — the claim anneals down and only 23.3% of rounds stay valid — rather than overclaiming silently.

The provable mode exposes an interaction with the adaptive layer. With $\lambda = 2$ and ACI on, edge misses vanish, the working α climbs until misses return to target, and the intervals end up no wider than at $\lambda = 1$ (gap 8.62 vs 9.47, coverage 0.984): the adaptive controller cancels the static margin. There is no soundness breach, but adaptive coverage control and a static provable margin fight, so the provable mode freezes ACI. The full provable mode ($\lambda = 2$ + thinned + frozen ACI) is sound with margin (coverage 1.000, CI $[0.999, 1.000]$ vs claimed 0.574). After α-annealing its validity matches the default (95.9% vs 96.5%); the Bonferroni-plus-thinning burden surfaces as width and weaker early claims (gap 14.22 vs 9.47; claimed 0.574 vs 0.639) rather than silence. The noise/drift asymmetry is as predicted: $\lambda = 2$ costs +71% gap in the noise-dominated regime (21.28 vs 12.80) and essentially nothing under drift (8.62 vs 9.47).

Against a Gaussian $\mu \pm \beta\sigma$ baseline, we do not observe under-coverage at these path-level settings: heavy-

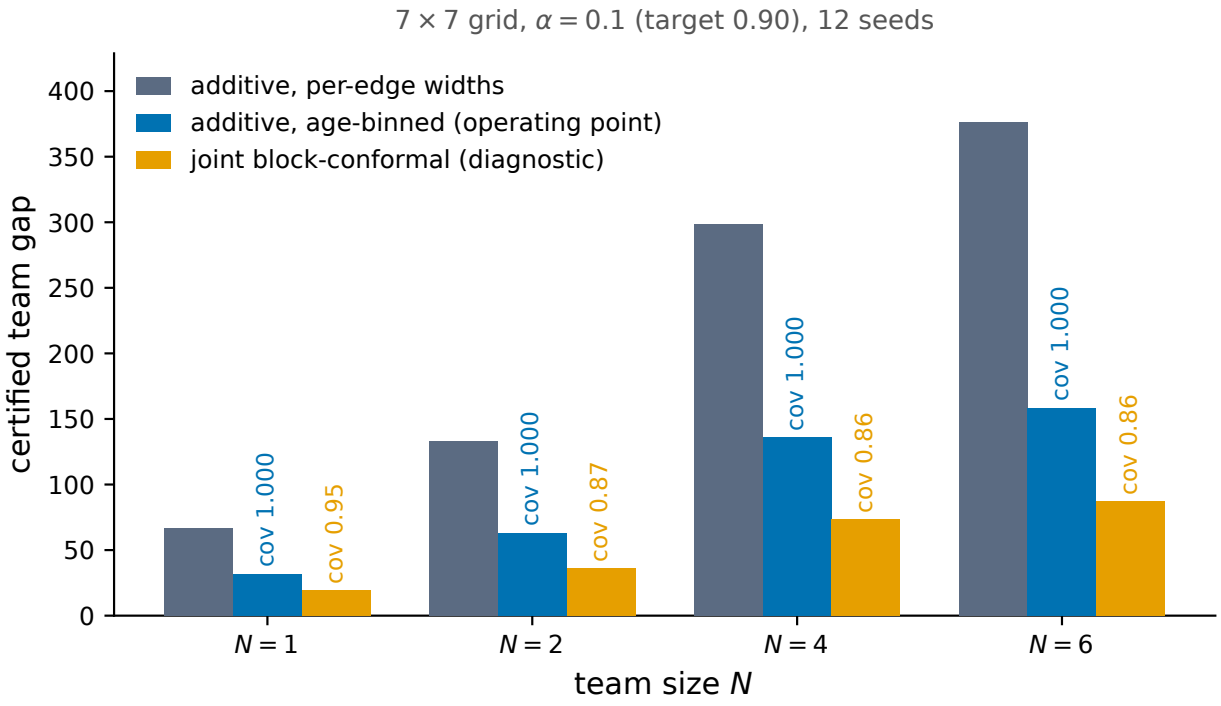

Figure 12: Additive team certificate: age-binned per-edge widths tighten the additive team gap 2.1–2.4× (66.8 → 31.6 at N=1, scaling near-linearly in N) at coverage **1.000** for every N (blue annotations). The joint block-conformal team quantile (orange) is tighter still on paper but *over-shoots the coverage target* — 0.95 at N=1 falling to ∼0.86 for $N \geq 2$ (orange annotations) — because the chosen conservative joint path biases its estimates low, the winner's curse that gates T4. The additive age-binned bound is the operating point; the joint object is a diagnostic only. 7×7 grid, α=0.1 (target 0.90), 12 seeds. Numbers: `coverage_tightening.json`; Section 7.1.

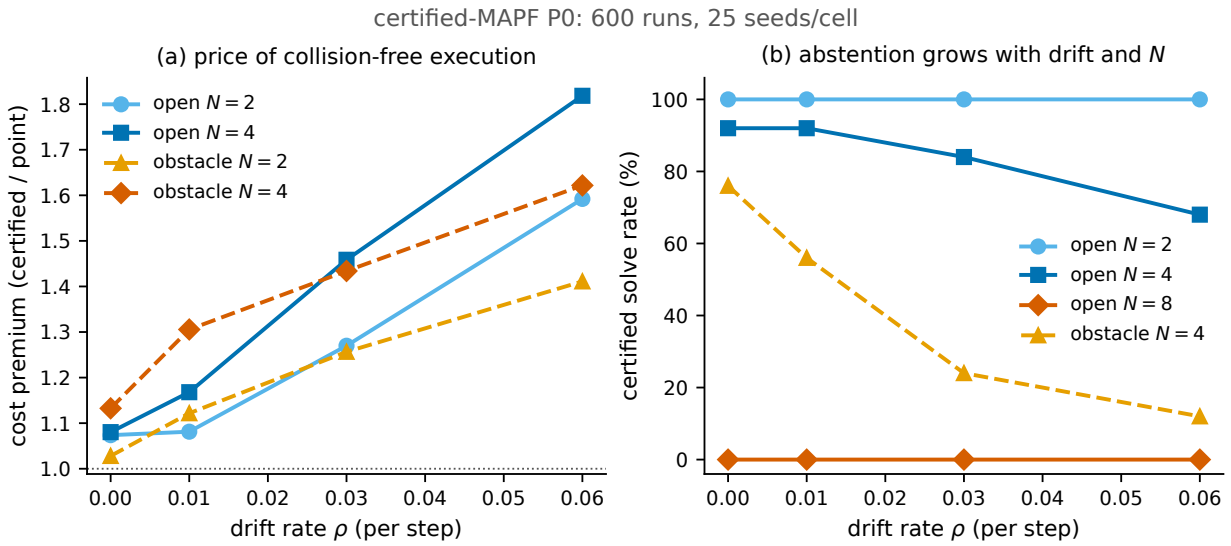

Figure 13: Certified-MAPF P0, the honest cost of the zero-collision guarantee (600 runs, 25 seeds/cell). *(a)* cost premium of the certified plan over point-estimate MAPF on commonly-solved instances, versus drift ρ: $\sim 1.08\times$ at ρ=0.01 growing to 1.6–1.9$\times$ at ρ=0.06, because wider certified corridors detour around one another. *(b)* certified solve rate falls with drift and N as disjoint corridors become infeasible — N=8 abstains on 100% of runs on the 8×8 grid (sound but useless), the failure mode P1 is designed to lift. Realized cost lies in [LB, UB] at rate 1.00 throughout. Numbers: `certmapf_p0.json`; Section 7.3.

tailed samples inflate the fitted σ, making the baseline conservative at moderate per-edge α and keeping it valid 98% of rounds by construction. The observed differences are tightness and claim honesty. Under bounded drift, CERT-FLOW's certified gap is about 40% tighter than the Gaussian (9.47 vs 16.65 Gaussian noise; 10.41 vs 17.01 Student-t), because the weighted conformal quantile adapts where the σ-fit bloats. The Gaussian claims a flat 0.800 with no staleness correction; CERT-FLOW's claim degrades visibly with calibration-buffer age. The building-block audit of Section 6.3 shows where this masking breaks down.

6.2 The sum-aware upper certificate (T4)

The asymmetric certificate uses a block-conformal upper bound at level α' with a $\Theta(\sqrt{L})$ margin (T4) and a per-edge Bonferroni lower bound (the asymmetry is forced; T5), applied through a freshness gate with κ-stabilized incumbents (last two rows of Table 2). It cuts the median gap by 26% and 43% against Bonferroni in the noise-floor-dominated regimes (3.04 vs 4.13 static; 7.31 vs 12.80 noise-dominated) and unlocks certification where Bonferroni gives none (3.4% vs 0.0% noise-dominated; 95.3% of valid rounds certified in static). Coverage is 0.966 and 0.916, above the claims (0.50/0.65) and below Bonferroni's 1.000: the tighter bound consumes the conservatism slack rather than violating anything.

Selection bias is real and measured. Applying T4 naively to the optimizer-selected incumbent drops coverage to 0.823, because the incumbent minimizes estimated costs and so its estimates are biased low. The freshness gate — using the sum-aware bound only when every incumbent edge has been re-observed since the path became incumbent — restores conditional validity, and κ-hysteresis opens the gate by stabilizing the incumbent, which is its second role. There is no effect under strong drift, where age widths dominate and the gate rarely opens, consistent with the $\sqrt{L}$ analysis applying to the noise floor only.

6.3 Edge-level calibration audit (the Gaussian break)

Path-level coverage cannot distinguish the two interval constructions: both sit at 1.000 because Bonferroni slack masks the building block. We therefore audit the building block directly: each valid round, a uniformly random edge receives a fresh observation, never fed back to the planner, tested against the *unclipped* nominal

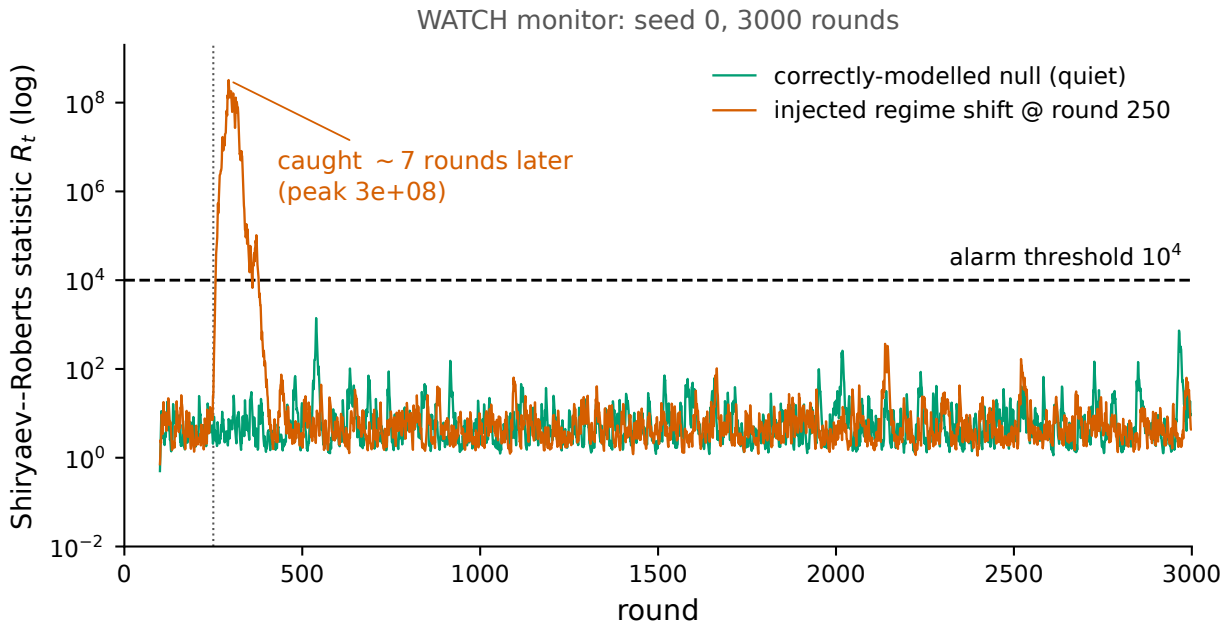

Figure 14: The WATCH observability layer makes the pinned coverage a live, alarming quantity. Shiryaev–Roberts statistic R_t (log axis) on two controlled streams (seed 0, 3000 rounds): on a correctly-modelled bounded-drift null (green) it stays below its alarm threshold 10^4 (peak 1401), consistent with the measured 0.0000 violation rate — quiet on 20 of 20 real METR-LA replay days; on a stream with an injected regime shift at round 250 (orange) it crosses the threshold ~ 7 rounds later (peak $\sim 3 \times 10^8$), where a plain conformal martingale, decayed over the long null, misses it. The certificate stream is byte-identical with monitoring on or off. Trace reproduced from `run_watch_testability.py` (`scripts/out/watch_testability.json`); Section 6.13.

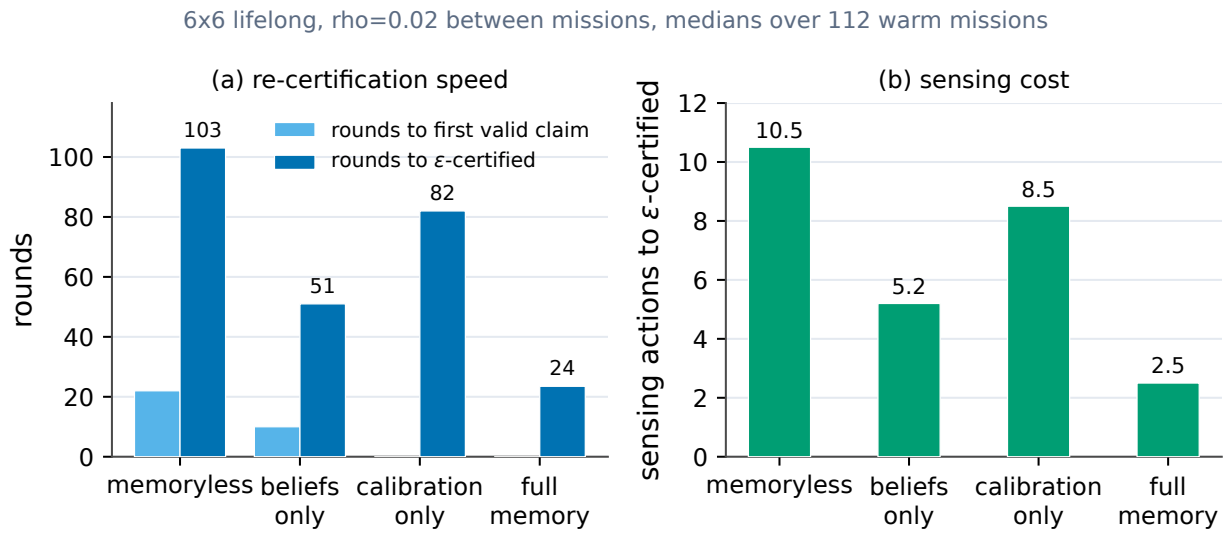

Figure 15: Lifelong operation (Tier-L): the memory ablation decomposed (medians over 112 warm missions; 6×6, ρ=0.02 drift between missions). *(a)* carried memory re-certifies an ε-good route 4.4× faster than a memoryless restart (103 → 23.5 rounds), and the calibration buffer alone yields *instant* claim validity (round 0 vs 22) while beliefs alone supply the route knowledge. *(b)* full memory also re-certifies at 4.2× less sensing (10.5 → 2.5 actions). The within-mission latency claim failed honestly (D* Lite reuse leaves nothing to save); this is the surviving lifelong system claim. Numbers: Table 9 (`lifelong.md`); Section 6.10.

interval (the cost-floor clip is sound for latent costs but not for observables, which can be negative under heavy left tails; coverage events must use unclipped bounds). With ACI frozen, $\alpha' = 0.1$ and $L \approx 18$, the claimed edge level is $\alpha_e \approx 0.0056$.

The Gaussian construction under-covers 4.2–8.8× in every static condition (Table 3) — including under correctly specified Gaussian noise, where the failure is plug-in estimation error of $\hat{\sigma}$ at the extreme quantile rather than a wrong family; skewed noise is worst because no symmetric fit represents an asymmetric tail. CERT-FLOW is calibrated in all conditions (0.6–1.1×), including skewed noise that violates its own A3. Drift masks the Gaussian failure (0.3×) because ρa widening dominates the quantile: the flaw is hidden, not fixed. Together with the T4 results this completes the slack-versus-soundness chain: at path level both methods hide behind Bonferroni slack; spend that slack for tightness and only the calibrated building block survives.

6.4 Tier 2: regret in unknown terrain

Table 4 runs 25 seeds on 10×10 bounded drift ($\rho = 0.02$), starting with no survey and a weak prior, $\varepsilon = 8$, $\alpha' = 0.2$. The robot pays true edge costs, traversal is a free observation, and regret is against a clairvoyant oracle that replans on true costs every step. "cert-then-go" senses one observation per round until ε-certified or the budget B is exhausted, then drives. Figure 6 summarizes the regret comparison.

Route-critical sensing — pointing observation at the certified gap — outperforms every baseline at every budget by 1.7 to 3× in travel regret (3.21/2.27/2.12 vs 3.84 to 7.07). Only cert sensing converts budget into quality monotonically (3.21 → 2.27 → 2.12 as B doubles); max_width is flat-to-worse with more budget, and max_age improves but stays about 1.8× behind. The mission-time trade is explicit: certify-then-go pays about 100 to 400 sensing rounds before departing, against 18 rounds driving blind, yielding 2 to 3× lower regret plus a certificate, with ε and B as the dial. Sensing while driving does not pay here (7.19 vs blind 6.60): once moving, traversal observations dominate and the extra spend yields no further reduction, so the certificate's value concentrates in the pre-departure phase on this small grid. Coverage holds in motion (0.87 to 1.00 across conditions), so the claim survives the robot driving, with traversal observations feeding the calibration buffer. One anomaly: max_age and max_width at $B = 40$ depart with the lowest

Table 2: Tier-0 coverage (6×6 grids, 25 seeds $\times$ 300 rounds/condition). Empirical coverage is among valid rounds, against the claimed line; gap is the median certificate width; cert% is the fraction of rounds certified at $\varepsilon = 5$.

condition	valid%	coverage	95% CI	claimed	gap med.	cert%
static ($\rho = 0$)	96.6	1.000	[0.999,1.000]	0.511	4.13	92.7
bounded $\rho = 0.005$	96.6	1.000	[0.999,1.000]	0.624	4.77	74.5
bounded $\rho = 0.02$	96.5	1.000	[0.999,1.000]	0.639	9.47	15.4
bounded $\rho = 0.05$	96.3	0.992	[0.990,0.994]	0.641	18.31	4.3
misspec $\hat{\rho} = 0.5\times$	96.6	1.000	[0.999,1.000]	0.648	5.81	24.3
misspec $\hat{\rho} = 2\times$	96.2	0.997	[0.995,0.998]	0.638	15.95	10.5
off-model: jump	96.6	0.998	[0.996,0.999]	0.658	92.95	0.0
off-model: periodic	96.1	0.996	[0.994,0.997]	0.643	14.81	3.9
$\lambda = 2$ (T1b margin)	96.5	0.984	[0.981,0.987]	0.632	8.62	33.3
$\lambda = 2$ + thinned	95.9	0.991	[0.989,0.993]	0.574	11.01	21.9
provable (λ2+thin+noACI)	95.9	1.000	[0.999,1.000]	0.574	14.22	0.0
noise-dom. static, λ1	96.5	1.000	[0.999,1.000]	0.649	12.80	0.0
noise-dom. static, λ2	96.6	1.000	[0.999,1.000]	0.646	21.28	0.0
A2 misspec $\varepsilon_{\mathrm{TV}} = 10^{-3}$	23.0	1.000	[0.998,1.000]	0.198	11.31	0.0
sum-aware static (T4)	96.6	0.966	[0.961,0.970]	0.503	3.04	95.3
sum-aware noise-dom. (T4)	96.5	0.916	[0.910,0.923]	0.648	7.31	3.4

Table 3: Independently audited edge-level miss rates (25 seeds $\times$ 400 rounds; verdict "broken" = Clopper–Pearson CI floor above α_e).

noise	planner	audit miss	ratio vs. α_e	verdict
Gaussian (control)	CERT-FLOW	0.0053	1.0	ok
	Gaussian	0.0233	4.2	broken
Student-t (df=3)	CERT-FLOW	0.0059	1.1	ok
	Gaussian	0.0394	7.1	broken
skewed (lognormal)	CERT-FLOW	0.0053	1.0	ok
	Gaussian	0.0486	8.8	broken
drift 0.02 + skewed	CERT-FLOW	0.0034	0.6	ok
	Gaussian	0.0019	0.3	ok

coverage (0.945/0.867), because exhaustively re-sensing old or wide edges builds stale-correction pressure on the confidence term.

6.5 Real data: replayed traffic (METR-LA, PEMS-BAY)

Figure 4 shows the certificate's building block on a single edge — staleness priced into width, reset by observation — and Figure 5 shows what removing that pricing costs: the closest conformal neighbor (CIA), run with its own construction on the same city, collapses from 0.95 coverage at gap zero to 0.20–0.25 at operational staleness, recovering only when the diurnal cycle happens to return the network near its calibration state.

Recorded loop-detector speeds (5-minute bins; 207 LA / 325 Bay Area sensors) are replayed as ground truth: edge costs are travel times from the recording, so the oracle is exact on *real* drifting costs [20]. The drift bound ρ_e is an empirical per-edge quantile of $|dc/dt|$, so A1 is violated by real incidents at a measured rate; observation noise is synthetic (the recording does not separate sensor noise from state). Twenty replay days per condition, one observation per bin.

Three findings. First, the certificate holds on data it was never tuned for: coverage meets or exceeds every claim across both cities, including drift models violated by real incidents up to half the time — understated drift lands in the drift-adjusted conformal scores instead of the ρa widening, so A1 misspecification costs width, never coverage. This self-absorption is invisible in synthetic worlds where A1 holds by construction.

Table 4: Tier-2 regret vs sensing policy (10×10 grid, $\rho = 0.02$, 25 seeds). Lower regret is better; coverage is among valid rounds.

condition	budget	goal%	rounds	regret mean	regret med.	sense	cov.
cert-then-go, cert	10	100	118	3.21	2.88	11.7	1.000
cert-then-go, cert	20	100	217	2.27	2.37	21.6	0.999
cert-then-go, cert	40	100	417	2.12	1.75	41.4	0.989
cert-then-go, random	10	100	118	5.37	5.10	11.8	1.000
cert-then-go, random	20	100	217	7.07	7.25	21.4	1.000
cert-then-go, random	40	100	417	6.39	7.01	41.5	0.999
cert-then-go, max_age	10	100	118	6.25	6.19	11.8	1.000
cert-then-go, max_age	20	100	217	4.49	4.60	21.7	1.000
cert-then-go, max_age	40	100	416	3.84	4.28	41.3	0.945
cert-then-go, max_width	10	100	118	6.06	5.32	11.8	1.000
cert-then-go, max_width	20	100	218	6.86	6.73	21.7	1.000
cert-then-go, max_width	40	100	416	6.07	6.10	41.3	0.867
no-certificate (drive blind)	—	100	18	6.60	6.50	0.0	—
cert, sense-while-driving	—	100	18	7.19	7.47	1.8	1.000

Table 5: Replayed-traffic certification (20 replay days/condition). Probe sweep: coverage stays 1.000 on LA even at $\rho{=}p50$, a 49% A1-violation rate.

city	planner	A1 viol.	coverage	claimed	gap med. (s)
LA	CERT-FLOW $\rho{=}p95$	5%	1.000	0.588	8797
LA	CERT-FLOW $\rho{=}p75$	25%	1.000	0.572	4774
LA	CERT-FLOW $p75$+adaptive	25%	1.000	0.584	4330
LA	Gaussian $p95$	5%	1.000	0.742	11288
Bay	CERT-FLOW $\rho{=}p95$	5%	1.000	0.680	1067
Bay	CERT-FLOW $\rho{=}p75$	25%	0.993	0.644	679
Bay	CERT-FLOW $p75$+adaptive	25%	0.987	0.643	683
Bay	Gaussian $p95$	5%	1.000	0.772	1570

Figure 7 contrasts this width-for-soundness trade with AD*-style inflation, which hedges search suboptimality rather than map staleness and so stays narrow but unsound. Second, the drift-aggressiveness dial has an interior optimum on real data ($p75$: gaps 46% tighter than $p95$ on LA; $p50$ backfires as score mass explodes), and the T2′ regime structure predicts behavior across cities: Bay Area traffic is $\sim 8\times$ gentler, so its targets are nearer-attainable and the aggressive $p75$ variant runs visibly closer to its claim there (coverage 0.99 vs. a hard 1.000 on LA). The adaptive sensing rate trims the LA gap a further 9% (4774 → 4330) while leaving coverage pinned at 1.000; on the already-gentle Bay it is gap-neutral, since one observation per bin is close to sufficient. Third, conformal is 1.3–2.6× tighter than the Gaussian baseline at equal-or-better soundness, and the aggressive variants visibly spend conservatism slack toward the claimed level (0.987 measured vs. 0.643 claimed) — efficiency, not unsoundness.

6.6 Standard pathfinding benchmarks (MovingAI)

Three map families from the MovingAI suite (DAO dungeon `arena`, a 64×64 Berlin street crop, a 64×64 maze crop) with bounded drift overlaid, unknown-terrain start, certify-then-go semantics, 15 seeds. Travel-regret is against a clairvoyant oracle on true costs.

Certificate-directed sensing has the lowest regret on both maps with real route choice, and is the only sensing policy that outperforms driving blind (blind in fact outperforms both *random* and *max-age* sensing on the dungeon: 3.96 vs. 4.89 and 4.08 — sensing uninformative edges costs mission time and yields nothing). The maze is a built-in negative control: with essentially one corridor, all three sensing policies produce bit-identical regret, locating the boundary of the route-critical-sensing claim exactly — it pays where alternatives exist, not where topology forces the path. The slightly negative blind regret on the maze records that the

Table 6: MovingAI benchmark maps [33], bounded drift ρ=0.02, 15 seeds. "—" = the policy never re-observes an edge, so no conformal pairs form (a policy property, not a soundness failure). All rows reach the goal in all seeds.

map	policy	regret mean	sense	coverage
DAO dungeon	**cert**	**3.37**	25.1	1.000
	random	4.89	25.1	1.000
	max-age	4.08	25.2	—
	drive blind	3.96	0.0	—
Berlin street	**cert**	**7.52**	28.5	1.000
	random	8.33	28.6	1.000
	max-age	8.59	28.6	1.000
	drive blind	8.70	0.0	—
maze	cert / random / max-age	0.81	27.8–28.5	1.000
	drive blind	−0.04	0.0	—

greedy step-wise clairvoyant oracle is itself marginally suboptimal there. Coverage holds at 1.000 wherever measurable.

6.7 Crossing the speed boundaries: certificate-gated preprocessing and road scale

Static-known planners answer in microseconds by preprocessing *assumed*-valid costs; the fastest published methods reach 0.3–4 μs on grid benchmarks [15, 33] and 0.56 μs (Hub Labels) on continental road networks [3]. CERT-FLOW reaches that speed class by *proof*: when the certificate establishes every edge interval within τ of a snapshot, that proof licenses an all-pairs oracle on the certified estimates — cost queries in 269–394 ns and full path queries in 8.7 μs, each carrying an explicit certificate (true cost within $|P|\tau$, optimum within $2|P|\tau$, at the annealed confidence) and expiring automatically the moment drift exceeds τ. Preprocessing-by-assumption becomes preprocessing-by-proof, and nothing of the online machinery is given up: a closed gate falls back to the certified replanning loop.

At road scale (DIMACS USA-road-d [9]: NY 264k nodes, FLA 1.07M), exact ALT queries run at 9.3/45.5 ms median with landmarks built on 0.8× cost *lower bounds*, so bounded (±20%) cost changes preserve admissibility with *zero* recustomization: a 1% perturbation is absorbed in 0.015–0.067 ms, versus $\approx$ 1 s (parallel) metric customization for CRP [8] — four orders of magnitude on the "costs moved, keep planning" operation. A from-scratch certified Contraction Hierarchy closes the remaining query gap: built in 108 s on full NY (973k shortcuts), cost queries answer at 231 μs median — within $\approx 2\times$ of the published C++ CH [13, 3] (110 μs, on a 70× larger graph) — with zero mismatches across all differential tests including post-perturbation; and the lower-bound variant (CH-potentials as admissible A* heuristics) keeps exactness under arbitrary ±20% changes with a 0.34 ms write and no rebuild. Only Hub Labels' 0.56 μs remains conceded, by category: gigabytes of tables on costs assumed frozen — the assumption the certificate exists to remove.

6.8 Tier 1: incremental repair latency

Table 7 compares D* Lite incremental repair against from-scratch Dijkstra after local cost perturbations (200 rounds per cell). "moving" perturbs within Chebyshev radius r between robot steps.

T3 is confirmed: repair cost tracks the perturbed region, not graph size, and at fixed locality the speedup grows with $|V|$ (1.38× → 3.84× → 9.64× for static $r = 1$; 1.98× → 6.25× → 23.75× for moving $r = 2$). The honest boundary is also visible: when the changed region approaches graph scale ($r = 10$ on 20×20), incremental repair loses to scratch (0.11×). Incremental search pays off only when changes are local relative to the graph, which is the regime lazy pre-widening restores for CERT-FLOW, since age-widening would otherwise touch every edge every round. Incremental cost equals scratch cost on every round of every cell, with zero mismatches over more than 3000 rounds.

Table 7: Tier-1 latency (abbreviated; 200 rounds/cell). p50 median per-round repair time; speedup is scratch/incremental.

scenario	size	r	inc p50 (ms)	scr p50 (ms)	speedup	nodes
static	20×20	1	0.20	0.27	1.38	400
static	20×20	10	2.09	0.22	0.10	400
moving	20×20	2	0.12	0.23	1.92	400
static	40×40	1	0.32	1.22	3.84	1600
static	40×40	10	4.43	1.29	0.29	1600
moving	40×40	2	0.18	1.12	6.31	1600
static	80×80	1	0.56	5.40	9.70	6400
static	80×80	10	9.37	7.50	0.80	6400
moving	80×80	2	0.23	5.48	23.84	6400

Table 8: Ablations (8×8, $\rho = 0.02$, 20 seeds $\times$ 300 rounds). Coverage, valid%, and gap are identical across the κ, maintenance, and backstop rows to the reported precision.

condition	cov.	valid%	gap	churn mean	churn p95	flap%	p50 ms
full (κ on, B=10)	1.000	94.5	22.14	0.58	0	3.8	0.45
no-κ	1.000	94.5	22.14	1.98	14	19.5	0.44
no-maintenance	1.000	94.5	22.14	0.58	0	3.8	0.45
B=0 (no pre-widen)	0.999	94.9	19.60	0.53	0	3.6	0.50
B=20	1.000	94.3	23.79	0.65	0	4.3	0.44
no-backstop	1.000	94.5	21.30	0.57	0	3.9	0.47

6.9 Ablations

Table 8 runs 20 seeds $\times$ 300 rounds on 8×8 bounded drift ($\rho = 0.02$, $\varepsilon = 5$, $\alpha' = 0.2$). Churn is the Fox plan-stability metric (edge symmetric difference between consecutive incumbents); flap% is the fraction of rounds with nonzero churn.

The conductivity module κ clears its kill-gate in its revised role of churn suppression. It cuts mean churn by 71% (1.98 $\rightarrow$ 0.58), p95 churn from 14 to 0, and flap rounds from 19.5% to 3.8%, at identical coverage, valid%, gap (22.14 both), latency, and sensing spend, which is direct evidence the certificate is untouched. The original latency-based gate is dead, since κ contributes no latency, but the stability gate is passed. Pre-widening is a clean width dial: $B = 0 \rightarrow 10 \rightarrow 20$ moves median gap 19.60 $\rightarrow$ 22.14 $\rightarrow$ 23.79 (+13% and +21%) at essentially flat per-round latency (0.50 $\rightarrow$ 0.45 $\rightarrow$ 0.44 ms), the lazy pre-widening of Section 6.8 having removed the touch-every-edge cost the dial used to pay. The maintenance and backstop rows are uninformative in this regime by design: cert% is near-zero everywhere ($\leq 0.5\%$) because $\varepsilon = 5$ is below the T2′ floor at $L \approx 14$ (the $2Lq$ term alone exceeds it), maintenance activates only when certified, and the greedy sensing score already emulates round-robin so the backstop never binds. α-annealing now keeps 94.5% of rounds valid even under this Bonferroni warm-up burden (versus the pre-annealing $\sim 52\%$), so the burden surfaces as the near-zero certification rate rather than as invalidity — the strongest argument for the sum-aware certificate of Section 6.2.

6.10 Lifelong operation (Tier-L)

Persistent memory cannot cut *within-mission* replanning latency — incremental search reuses its state and lazy pre-widening restores locality, leaving nothing to save (an honest negative we retain). Across missions in the same drifting environment, the picture inverts: a memoryless planner re-pays calibration warm-up, re-learns every edge, and re-discovers corridors each time. Sixteen seeds, eight missions each, five memory variants (first mission excluded):

Carried memory re-certifies 4.4× faster at 4.2× less sensing, and the ablation decomposes the effect exactly (Figure 15): the calibration buffer yields instant claim validity (annealed claims from round 0), beliefs supply the route knowledge, and they compose. The honest trade: memory-carried incumbents

Table 9: Lifelong missions (6×6, ρ=0.02 drift between missions, medians over 112 warm missions). Full memory = beliefs + calibration + κ; the κ-less variant is identical on every speed column (κ contributes stability, not speed).

variant	rounds→valid	rounds→cert	sense→cert	regret
memoryless	22.0	103.0	10.5	0.17
beliefs only	10.0	51.0	5.2	1.08
calibration only	0.0	82.0	8.5	0.06
full memory	**0.0**	**23.5**	**2.5**	0.56

Table 10: Sum-level upper-bound calibration on identical worlds. Median certified gap and its ratio to the shipped per-edge Bonferroni default (lower is tighter); violation rate is against true OPT. METR-LA: 10 seeds ×288 rounds, paired (n=2729 valid rounds). Drift grid: 10×10, ρ=0.02, 10 seeds; observation-noise draws differ across modes there, so treat deltas under $\sim 2\%$ as noise. The a-posteriori shrink tier (bottom rows, italic) carries a different, weaker claim and is never a default-replacement candidate.

mode	viol. (METR-LA)	gap ratio (METR-LA)	gap ratio (grid)
default (per-edge Bonferroni)	0.0000	1.000	1.000
sum-aware UB (T4 block quantile)	0.0000	0.764	0.998
group-sum CIA UB	0.0000	0.734	0.977
PASC (per-block max)	0.0000	1.247	0.966
shrink, Tier-1 (unchanged)	*0.0000*	*0.999*	—
shrink, Tier-2 shadow	*0.0051*	*0.376*	*0.577*

certify at points slightly further from optimal (regret 0.4–0.6 vs 0.17, all far inside $\varepsilon = 5$) — stale beliefs prove the first ε-good route rather than re-exploring for the best one; the certificate's promise holds for every variant.

6.11 Sustained certification under drift (T7) and conditional features

The churn-directed sequence of changes (focused sensing, online ρ, adaptive rate on the churn-measured floor) raises sustained certification in the stress regime (6×6, $\rho = 0.05$, $\varepsilon = 8$) from 5.6% to 36.7% of rounds at coverage 1.000; the rotation alternative (sensing over the full churn set) was tested and refuted — same certification, +20% sensing. Conditional features are validated in their designed regimes and honestly bounded: the spatial predictor's dense-sensing claim is *downgraded* (at 8 observations/round the gap improves only $\sim 1.4\%$, an order of magnitude under its offline bound; its payoff regime is a continuously-reporting sensor network), and decision-uniform mode shows exactly its mechanical signature (unchanged certification and gap, stronger per-claim confidence), with the trajectory-level metric saturated in clean regimes.

6.12 Tightening the certificate: sum-level upper bounds on real traffic (T8)

The certificate's standing weakness is width: valid intervals run 1–2 orders wider than AD*-style (unsound) inflation, and the union-bound α'/L tax dominates on long paths. We evaluate it head-to-head on identical worlds. Table 10 prices every width-relevant upper-bound construction on the same paired replay — real METR-LA (10 seeds ×288 rounds, n=2729 shared valid rounds) and a short-path drift grid — and records, in one place, both the recoverable part of the width and the construction that made the certificate *wider*. Every a-priori mode is measured against true OPT; a mode replaces the default only if its violation rate is statistically indistinguishable from the default's (0.0000) and it is $\geq 10\%$ tighter on *both* benchmarks. Figure 8 summarizes the real-traffic column.

Three findings. *The width win lives on long paths.* METR-LA incumbents run $L \approx 14$–18 edges, so the per-edge level α'/L starves against the buffer's effective sample size and per-edge quantiles ride their top order statistics; a single sum-level quantile at level α' is supportable and tightens the certified gap 23.6%

(block-sum, `sum_aware_ub`) to 26.6% (group-sum CIA UB, after Luo and Zhou [23]) at a measured violation rate of 0.0000. On the short-path grid ($L \approx 6$) the tax is small and every a-priori mode is within noise of the default — the attack pays exactly where the weakness was measured. *Sums calibrate where maxima starve.* PASC [18] prices one quantile of the per-block *maximum*; a length-L block-max needs $\sim L\times$ the samples per block, so on long real paths the buffer holds too few blocks and PASC lands +24.7% *wider* than Bonferroni (independently reproduced at +25.1% in the live-wiring benchmark of Section 6.13). Same buffer, same level; only the functional of the block differs, and the sum is the one that calibrates. We report this as a designed negative: PASC keeps an experimental flag, and the clean immediate default candidate is `sum_aware_ub`, whose $\Theta(\sqrt{L})$ accounting (T4) is already inside the certificate. *An a-priori shrink is impossible; an a-posteriori one is licensed (T9).* The Tier-1 certificate is bit-identical with the shrink flag on. A betting confidence sequence [36] on the observed violation stream licenses a Tier-2 *shadow* radius $k(q + \rho a)$ that is 62.4% narrower on real traffic at a measured 0.51% miscoverage against true OPT (target α'=0.20; the license floor k=0.5 binds on 82% of rounds, so the stream would support more). This is a strictly weaker, anytime statement about *this* deployment's stream, self-revoking under regime shift — not an a-priori guarantee for the next round. The companion proves the trade is fundamental: any a-priori narrowing from windowed evidence reassumes the exchangeability the drift model exists to drop (the CIA-collapse failure of Figure 5). Deployment reading: safety gates consume Tier-1; resource allocation may consume Tier-2.

6.13 Making coverage observable: the WATCH monitor

Coverage pinned at 1.000 has been the certificate's least-testable property: the guarantee is real but the number never moves, so a live consumer cannot see the modelling null break. We wire a weighted conformal test martingale and a Shiryaev–Roberts detector [28] into `ingest_observation`: each new score's weighted conformal p-value feeds a nonnegative supermartingale (Ville gives false-alarm $\leq \delta$ under the weighted-exchangeability null) and an implicitly-restarting SR statistic. The layer is purely observational — with it on or off the (LB, UB, confidence) stream is byte-identical. On real METR-LA (20 seeds $\times$288 rounds, one replay day each) both detectors stay *quiet on* 20 *of* 20 *days* — zero false alarms — exactly consistent with the measured 0.0000 coverage-violation rate, at unchanged median gap (8797 s). A controlled stress stream confirms the other side (Figure 14): after an injected regime shift the SR statistic crosses its alarm ~ 7 rounds later (peak $\sim 3 \times 10^8$ vs threshold 10^4), where a plain martingale, random-walked toward zero over the long null, misses it — which is why WATCH pairs the two. Conformal e-values [12] give the same signal in mergeable form ($\mathbb{E}[E] \leq 1$ under the null; the average merge is valid under arbitrary dependence across a path's edges). The pinned coverage is now a live, alarming quantity at zero cost to the bound. One honest calibration note carried from the same run: the Bonferroni-vs-sum-level width gap is real only under positive edge correlation — at correlation 0.9 on an L=20 path the joint price holds ~ 0.91 coverage at 16.5% less width, while under independence Bonferroni is already tight and the joint price barely helps.

6.14 Objective-matched hybrid sensing on real traffic

Pure gap-directed sensing is optimal when ε is attainable, but on a real cost process ε is often unattainable all day (METR-LA gaps run $\sim$8–10k s; certified fraction 0.0 for every policy). In that regime gap-directed observations are spent on certificate-relevant but route-marginal edges. The objective-matched *hybrid* redirects the one-per-round budget toward the expected-best route exactly when the certifiability threshold (T2′) says ε cannot close, and reverts to gap-directed sensing when it can. Table 11 measures it on real METR-LA (10 seeds $\times$288 rounds, warm-up excluded; oracle = exact Dijkstra on the recording), and Figure 11 plots both axes — the regret win and the width it costs.

Hybrid cuts median route regret 41% (65.1 $\rightarrow$ 38.3 s; mean -14%) at identical validity and per-round cost, and dominates the freshness and random baselines on both regret and width (they pay 5–6.5$\times$ the certified gap). The honest cost is kept: hybrid's certified gap is +21% (8468 $\rightarrow$ 10 247 s) — but in this regime *no* width certifies at ε, so the wider gap is not decision-relevant while departure quality is. Critically, hybrid changes only *when* gap-sensing is pointless, never *what* is certified: unlike the killed decision-focused team sensing (Section 7), the (LB, UB, confidence) triple is unaffected. Where ε *is* attainable, a synthetic mixed-regime benchmark confirms hybrid converges to the default's behaviour and matches the greedy clairvoyant oracle on average (regret -0.12 vs a CTP-RS-style value-of-information baseline's 0.48 and pure gap-directed's 2.35).

Table 11: Sensing policies on real METR-LA (10 seeds ×288 rounds). Median route regret in seconds against the clairvoyant oracle; all policies hold coverage at 0.0000 violations, valid 98.9%.

policy	regret mean (s)	regret med. (s)	gap med. (s)	ms/round
hybrid (objective-matched)	**114.4**	**38.3**	10 247	1.0
cert (pure gap-directed, default)	132.2	65.1	8 468	1.2
max_age (freshness)	131.6	80.8	55 170	0.5
random	182.3	116.3	44 325	0.4

Hybrid is the recommended sensing configuration; pure gap-directed remains the reproducible published default for one release and stays documented as the policy outperformed in the never-attainable regime.

6.15 Is a certificate worth the latency? No quality crossover

Static-known planners answer in microseconds; a certified round costs milliseconds (Section 6.7). The fair question is whether that latency ever produces a *worse* decision — i.e. whether some drift level would make an uncertified fast replanner's routes and promises acceptable in exchange for its speed. We run both planners on identical ground-truth worlds (12×12, 15 seeds, 220 rounds, 40 warm-up discarded), each sensing one edge per round: FAST uses last-observation point beliefs, freshness sensing, from-scratch Dijkstra, and promises its believed route cost; CERT-FLOW promises UB when valid and abstains otherwise. We score regret = truecost(route) − OPT and the broken-promise overrun $\max(0, \text{truecost} - \text{promise})$. Figure 9 and Table 12 report the sweep.

Table 12: No quality crossover (12×12, 15 seeds, pooled per ρ). FAST's point-estimate promise is exceeded on 62–97% of rounds; the certified upper bound is never exceeded (overrun ≡ 0). Composite J = regret + overrun; crossover analysis returns ρ^*=0 (CERT-FLOW ≤ FAST across the whole sweep).

ρ	FAST regret	CERT regret	FAST overrun	FAST break-rate	CERT overrun
0.00	0.036	**0.020**	0.154	79%	**0**
0.005	0.055	**0.053**	0.219	62%	**0**
0.01	0.233	**0.194**	0.310	64%	**0**
0.02	0.775	**0.478**	0.581	68%	**0**
0.05	2.974	**1.502**	2.054	84%	**0**
0.10	6.378	**2.523**	6.598	97%	**0**

There is nothing to cross over to. The certified planner's regret is ≤ FAST's at *every* drift level, including the static map (ρ=0: 0.020 vs 0.036 — the point estimate is a min over noisy estimates and so is optimistically biased even with no drift), and FAST's promise is broken on 62–97% of rounds with overrun growing 0.15 → 6.6 cost units as drift rises, while the certified upper bound is never exceeded (overrun ≡ 0, abstention honest). The composite J favours CERT-FLOW at every ρ (0.020 vs 0.191 at ρ=0; 2.52 vs 12.98 at ρ=0.10). FAST's entire advantage is latency (∼ 2–4× faster per round, μs–ms class) — a conceded, real win on *raw speed*, and the only thing the certificate trades away.

7 Certified multi-agent planning (T10)

The single-agent certificate lifts to fleets in two steps: an additive team bound that is sound, with an exactly separable team optimum, and a first certified-MAPF study that turns the bound into collision-free execution. We report both, and — keeping the program's discipline — the standalone joint construction that did *not* survive real data, and the probabilistic knob that is inert at the studied scale.

7.1 The additive team certificate

For N agents sharing one drifting-cost graph and *one* conformal edge-price store (so every edge's observation age a_e is global — any agent's observation refreshes it for all), the per-agent certificates add directly:

$$\mathrm{LB}_{\mathrm{team}} = \textstyle\sum_i \mathrm{LB}_i \;\le\; \sum_i \mathrm{OPT}_i = \mathrm{OPT}_{\mathrm{team}} \;\le\; \sum_i \mathrm{UB}_i = \mathrm{UB}_{\mathrm{team}},$$

with team confidence combined by a union bound over the agents' miscoverage events, $\mathrm{conf}_{\mathrm{team}} = \max(0,\, 1 - \sum_i(1-\mathrm{conf}_i))$. Because each summand is sound over the shared store (the single-agent per-edge guarantee) and the team optimum separates for independent agents, the additive bound is sound on both sides; what is *exact* is the decomposition $\mathrm{OPT}_{\mathrm{team}} = \sum_i \mathrm{OPT}_i$ — a modeling identity of the uncoupled objective, not a probabilistic claim (each LB_i itself remains conservative). This is the one survivor of an earlier standalone line (§7.2); it is the shipped multi-agent object (`certflow.team.additive_certificate`).

Age-binned widths tighten it; the joint block price over-shoots. Refining the per-edge widths into age bins (a data-independent function of observation age) tightens the additive team gap 2.1–2.4× at coverage 1.000 across $N \in \{1, 2, 4, 6\}$ (7×7 grid, α=0.1, target 0.9, 12 seeds; gap 66.8 → 31.6 at N=1, scaling near-linearly in N; Figure 12). Replacing the additive upper bound with a *joint* block-conformal team quantile is tighter still on paper (19.2 vs 31.6 at N=1) but *over-shoots the coverage target* — 0.955 at N=1 falling to 0.86 for $N \ge 2$ (per-seed minimum as low as 0.68) — because the chosen conservative joint path minimizes estimated cost and so biases its estimates low, the same winner's curse that gates T4. The joint object is therefore reported as a diagnostic only; the additive bound is the operating point, and its soundness never depends on the selection.

7.2 The joint congestion certificate, falsified on real data

A stronger standalone extension — a congestion-coupled joint certificate with edge cost rising in team load, $c_e(m) = \phi_e(1+\beta_e m)$, priced over unique (e, m) cells, together with a decision-focused shared-sensing allocator — was built and tested. On synthetic forced-bottleneck graphs it is strictly tighter than the additive bound and its advantage grows with N; on a synthetic load-spreading (parallel) graph the advantage is smaller; and on the *real* METR-LA Los-Angeles road network, where route diversity lets each agent avoid the few congested edges, the additive bound is instead ≈ 10% *tighter* at every N. Table 13 is the full readout. All certificates were sound (0 violations) and covered at $1-\alpha$ throughout; the table reports tightness, not validity.

Table 13: The joint congestion certificate, priced against the additive bound on three graph families. Entry is the certified-gap ratio additive/joint (> 1: the joint object is tighter; < 1: additive is tighter). The joint object is tighter on synthetic congestion and tighter-with-N only there; on the real METR-LA road network it *loses* at every N, so it is falsified as a deployable object and not shipped. α=0.1, target coverage 0.9; bottleneck/parallel 12 seeds, METR-LA 6 seeds. Source: `congestion_cert.json`.

graph family	N=2	N=4	N=8	N=16
synthetic forced bottleneck	1.23	1.47	1.78	—
synthetic load-spreading	1.12	1.14	1.10	—
METR-LA (real road network)	**0.91**	**0.90**	**0.91**	**0.91**

The sensing allocator showed no deployable win region against independent certification and standard predict-then-optimize baselines either. The joint congestion model (`congestion.py`), its block-quantile upper-bound branch (`joint_ub`), the shared-sensing allocators (`sensing.py`), and the MPC field prototype (`mpc_field.py`) are deliberately not ported — the survivor is the additive certificate above. This is the same meta-lesson the single-agent study keeps returning: the certificate survives real data; the tighter, prediction-shaped object breaks on it.

7.3 Certified multi-agent path finding: a pre-registered P0 (T10b)

Lifting the certificate to collision-free execution is new territory: no prior method certifies team cost under drift, and execution uncertainty is a named open challenge in the lifelong-MAPF agenda. We define the certificate object, prove soundness before writing code, pre-register kill criteria, and report the first study against them. The planner runs conflict-based search [31] over *certified space-time corridors*: each agent's low-level state is (vertex, certified arrival window), where the window is priced by the same age-weighted conformal quantile plus ρa staleness at the latest certified traversal time; CBS branches on overlapping certified windows. The certificate promises (C1) collision-free execution on the coverage event, at team confidence $1 - \alpha_{\text{team}}$ when the per-edge budget is supportable by the calibration size (at P0 scale it is not, and the honest level is stated below); (C2) an additive team-cost bound with certified suboptimality; and (C3) honest abstention when no jointly certified conflict-free plan is found — it never returns an uncertified plan. Soundness holds on the coverage event by construction (certified windows are pairwise disjoint, so realized occupancies inside them cannot conflict); completeness and optimality of window-CBS are explicitly *not* claimed (interval branching can exclude jointly-feasible schedules), which costs success rate, never soundness.

The P0 sweep runs 2 maps (8×8 open and ∼20%-obstacle) × 4 drift rates × 3 team sizes × 25 seeds = 600 runs, comparing five planners that share one code path and differ only in window width: **point** (zero-width, committing directly to stale point estimates), **certified**(α) for $\alpha \in \{0.05, 0.1, 0.2\}$, and **worst-case** ($\alpha \to 0$, the max calibration score). Kill criteria were fixed in advance (**K1**: kill unless certified(0.1) outperforms point-estimate planning on collisions at $\leq 1.25\times$ cost at some $\rho > 0$; **K2**: the wedge is weak if worst-case inflation matches certified(α) everywhere; **H**: the certified collision rate must stay $\leq \alpha_{\text{team}}$ at every condition). Table 14 and Figure 10 report the outcome on representative conditions; the complete 24-condition sweep is Table 17 in Appendix C, and Figure 13 plots the cost premium and abstention that are the guarantee's honest price.

Table 14: Certified-MAPF P0 (representative conditions; 25 seeds each). Collision rate is the fraction of executed runs with ≥ 1 realized conflict; cost is the mean over instances solved by *both* planners. Certified plans realize *zero* collisions on every solved run in the entire 600-run sweep; point-estimate MAPF collides on up to 100% of runs. Certified rows are at *nominal* α=0.1; the supportable team level at this calibration size is ≈ 0.71 (see text).

map	ρ/N	point coll.	cert coll.	point cost	cert cost	cert solve%
open	0.01/2	8%	**0%**	38.9	42.0	100
open	0.01/4	20%	**0%**	78.7	91.9	92
open	0.06/4	40%	**0%**	79.4	144.3	68
obstacle	0.01/2	4%	**0%**	47.6	53.4	96
obstacle	0.03/4	23%	**0%**	93.3	133.8	24
open/obst.	any/8	57–100%	**0%**	—	—	0

The certificate does exactly what it promises and no more. **H holds**: certified plans realize zero collisions on every solved run across all 600 runs ($\leq \alpha_{\text{team}}$ at every condition, 0 honesty-gate violations), while point-estimate MAPF — genuinely dangerous under drift — collides on 4% of runs at low density up to 40% at N=4, ρ=0.06 and 57–100% at N=8. Realized cost lies inside [LB, UB] at rate 1.00, and abstention is honest. **K1 does not fire**: certified(0.1) outperforms point-estimate planning on collisions at $\leq 1.25\times$ cost at ρ=0.01 (N=2: 0% vs 8% at 1.08×; N=4: 0% vs 20% at 1.17×; obstacle N=2 at 1.12×) — the win region exists but is low-drift, low-N. **K2 fires**, and this is the honest headline of the study: certified(0.05), (0.1), and (0.2) are *identical* to worst-case inflation in every cell (team cost within 0.0%, solve rate within 0.0 points). The mechanism is Section 6.12's disease amplified: the per-edge budget $\alpha_{\text{edge}} = \alpha/(N\bar{L})$ sits below the finite-sample support floor (∼672 age-decayed scores against $\bar{L} = 48$, N up to 8), so the conformal quantile floors to the maximum score in 100% of certified cells and the probabilistic knob is inert at this scale. The floor has a label consequence we state explicitly: when $\alpha_{\text{edge}} < \tilde{w}_{n+1} \approx 1/(n_{\text{eff}} + 1)$ the requested quantile lands in the $+\infty$ test atom — the licensed window is infinite — and the max-score fallback licenses only per-edge miscoverage $\approx \tilde{w}_{n+1}$, not α_{edge}. Concretely, $\alpha_{\text{edge}} = 0.1/192 \approx 5.2\times10^{-4}$ against a supportable $\approx 1.5\times10^{-3}$, so over the ≤ 192 priced plan durations the supportable team level is $\approx 1 - 192 \times 1.5 \times 10^{-3} \approx 0.71$, not the nominal 0.90 (and the repaired T10b budgets over a selection-free universe, which is stricter still). All

"certified(α=0.1)" labels in this section and its tables are therefore *nominal*: the honest annealed team confidence at P0 scale is ≈ 0.71 — the single-agent planner already anneals its claim in this regime, and the team label now carries the same discipline. Empirical collisions are zero either way; the label, not the behavior, was wrong. The failure mode at scale is also honest: at N=8 certified corridors cannot be made disjoint on an 8×8 grid, giving 100% abstention (sound but useless), and the cost premium grows with drift (1.08× at ρ=0.01 to 1.6–1.9× at ρ=0.06).

Verdict and the gated next step. P0 *survives* its pre-registered kill criterion (K1 does not fire; H holds) but has not yet demonstrated the wedge's distinctive content — the α knob. The limiting factor is exactly the width weakness of Section 6.12 amplified by $N\bar{L}$, and its cure is the same: per-agent sum-level window pricing at level α/N instead of per-edge $\alpha/(N\bar{L})$, plus warm-up sized so the effective sample size supports the level, on 16×16–32×32 maps where corridors can be disjoint at N=8–16. That P1 iteration is pre-registered with a hard kill of its own: *if certified(α) still collapses to worst-case after sum-level pricing, or the K1 win region does not extend beyond ρ=0.01, the line stops and the honest negative is published alongside the P0 design.* We state it now so the follow-up cannot be tuned into a positive.

8 The verdict scoreboard

We close with one honest table (Table 15): per area, what CERT-FLOW achieves, the best alternative on that axis, and a plain verdict — including the areas where CERT-FLOW does not win. The losing rows stay in. Read straight, it says CERT-FLOW wins *soundness* (coverage under real drift, the CIA-collapse comparison, bounded-cost-change absorption) and *observability* (the WATCH/SR monitor) decisively; its interval width is *wide but shrinking*; hybrid sensing is a real-data pass; and it *loses on static-map raw latency by design* — the regime purpose-built static planners own and a certified planner is not for, though even there the uncertified alternative obtains its speed by breaking most of its promises (Section 6.15).

The meta-lesson the program keeps testing survives this version too: the certify-and-verify layer holds on real data — coverage, the CIA-collapse contrast, bounded-change absorption, and now a live validity monitor — while the claims that broke were the ones that tried to predict rather than certify (the joint congestion model, the dense-sensing spatial predictor, PASC's width promise, the certified-MAPF α knob at P0 scale). Every one of those is in this paper, with its number.

9 Limitations and Conclusion

The coverage guarantee is model-conditional and is *verified* only in simulation and on recordings, where ground truth exists every round; field deployments can demonstrate utility, not coverage. The drift model enters twice (A1 in widths, A2 in the staleness correction) and both are assumptions an adversarial world can break — our off-model rows quantify degradation but do not bound it. The provable mode's validity cost is resolved by α-annealing (warm-up rounds carry honest weaker claims), though its width still pays the doubled margin and the Bonferroni lower bound. The remaining technical residuals are accounted for rather than left open. π_{cal} is bounded explicitly under a bounded-density addition to A3, and the gated sum-aware construction leaves no uncontrolled selection mass (theory companion, Limitation closures). Online estimation of ρ_e is implemented (a pooled conservative quantile of observed rates): coverage is unchanged while gaps tighten 1.7–2.4× versus supplied worst-case bounds, so the drift dial tunes itself. The two theory threads that were once open are now closed. The uniform sum-aware lower bound is closed by impossibility (T5): selection over exponentially many candidate paths forces linear-in-L slack, so Bonferroni is order-optimal and the certificate's asymmetry is a theorem, not a gap. The churn residual is closed by T7 plus measurement: the floor and rate use the online churn measure $\widehat{K}$ (honest attainability under churn), focused sensing suppresses churn at its source, and sustained certification in the stress regime improved 5.6% → 36.7% at coverage 1.000 across the churn-directed changes.

The version-2 additions inherit the same limitations and add their own, stated in place. Interval width remains the standing weakness: the sum-level constructions recover 24–27% on long real-traffic paths at zero violations, but the residual — staleness ρa plus finite-sample floors — is the price of the drift guarantee and is not a calibration defect (Section 6.12); one of the tried constructions (PASC) made the certificate wider on

Table 15: Where CERT-FLOW wins, and where it does not. Each number traces to a section of this paper; verdicts are plain and the FAIL/WEAK rows are kept.

Area	CERT-FLOW	Best alternative	Verdict
Coverage under real drift	certificate coverage **1.000**, every condition	AD*/ARA* validity **0.02**–**0.07** on real METR-LA	**PASS** — validity is the axis a route certificate lives on
vs. CIA (closest conformal)	holds **0.95**–**1.00** across every staleness gap	CIA collapses **0.95** → **0.20** under staleness	**PASS** on validity — honest width cost, up to ∼49× wider at 24 h (Fig. 5)
Interval tightness	valid but 1–2 orders wider; sum-level UB recovers **−26.6**% real-traffic width at 0 violations; licensed Tier-2 **−62.4**% at 0.51%	AD*-semantics intervals narrow (but invalid)	**WEAK, SHRINKING** — soundness costs width, now measurably recoverable; residual = the drift price (§6.12)
Sensing	objective-matched hybrid **−41**% median route regret on real METR-LA; pure gap-directed dominated	CTP-RS-style VOI regret 0.48	**PASS** (hybrid) / **FAIL** (pure) — hybrid wins *and* carries a certificate VOI lacks (§6.14)
Static-grid / continental speed	∼3.7 ms per certified round	JPS+ ∼4 μs · Hub Labels 0.56 μs	**FAIL** on raw latency, by design — but no quality crossover: certified regret ≤ the fast planner's at *every* drift, whose promises break on 62–97% of rounds (§6.15)
Bounded cost-change absorption	**0.015**–**0.34** ms	CRP ∼ 1 s recustomization	**PASS** — orders faster on the "costs moved, keep planning" operation (§6.7)
Observability (WATCH / SR)	quiet **20**/**20** real seeds; injected shift caught in ∼6–7 rounds	no competitor ships this	**PASS** — novel; coverage is now a live, alarming quantity (§6.13)
Multi-agent	additive fleet certificate sound (team optimum exactly separable); certified-MAPF P0 executes at 0 collisions vs 0–100% for point-estimate MAPF	joint TEAM-CERT (tighter on synthetic only)	**MIXED** — additive ports; joint *falsified* on real METR-LA; certified-MAPF α knob inert (supportable team level ≈ 0.71 at P0 calibration), P1 gated (§7)

real traffic, and we report it. The licensed a-posteriori tier is genuinely narrower but carries a strictly weaker, self-revoking claim and must never be read as the safety bound. The observability layer removes the old "coverage is untestable" objection but cannot manufacture coverage the world does not grant. The certified multi-agent extension is sound on first contact yet obtains its zero collisions through abstention and added cost, and its α knob is finite-sample-inert at the studied scale — a negative we pre-registered and report rather than tune around, with the scaled follow-up gated on it (Section 7). The scoreboard (Section 8) keeps the losing rows in: static-map raw latency is conceded by design, and interval width is wide-but-shrinking. The meta-lesson the program keeps testing survives another round of real data — the certify-and-verify layer holds; the claims that broke were the ones that predicted rather than certified.

CERT-FLOW shows that a route certificate under staleness is not a bookkeeping exercise: it changes what the planner senses (gap-critical edges), when it may stop (the threshold), what it executes (certified incumbents, hysteresis within slack), what it must keep doing to stay certified (maintenance), and — now — how a fleet may commit to corridors without colliding. The certificate is the planner.

A Self-contained theory: statements and proofs

This appendix reproduces the paper's theorems so that `main.pdf` is a single self-contained artifact. **Recommendation for the engrXiv upload**: submit `main.pdf` *alone* as the primary artifact — with this appendix it needs no external document — and keep the theory companion (`theory.pdf`) as optional supplementary material for readers who want the full proof set of the classical results. Accordingly, the classical

single-agent results T1–T3, T6, and T7 are stated here with their proofs deferred to the companion, while the asymmetry pair T4/T5 and the version-2 results T8–T10 – the theory this version adds – are proved here in full.

A.1 Setup, scores, and per-edge coverage (T1)

Observing edge e at time u returns $Y_e(u) = c_e(u) + \eta_{e,u}$ with η independent across observations. The planner stores the last estimate $\hat{c}_e$ from time t_e and age $a_e(t) = t - t_e$. Assumptions: **A1** bounded drift $|c_e(t') - c_e(t)| \le \rho_e|t' - t|$; **A2** the noise law drifts in total variation at rate $\le \varepsilon_{\mathrm{TV}}$; **A3** symmetric unimodal noise; **A4** edges sharing a calibration buffer share a noise family with independent cross-edge deviations. On re-observation the planner records the *drift-adjusted score* $R = |Y_e(u) - \hat{c}_e| - \rho_e a$; writing $\delta = c_e(u) - c_e(t_e)$ ($|\delta| \le \rho_e a$ by A1) and η, η' for the fresh/previous noise, $Y_e(u) - \hat{c}_e = \delta + \eta - \eta'$, so

$$|\eta - \eta'| - 2\rho_e a \;\le\; R \;\le\; |\eta - \eta'|. \tag{1}$$

Scores enter one rolling buffer with data-independent age weights $w_i = \rho_w^{t-u_i}$ (weights depend only on observation *times*; adaptivity of *which* edge is observed is priced by the thinning and leave-one-out hypotheses of the companion's Lemma 1); the weighted $(1-\alpha)$ quantile $q_t(\alpha)$ (test mass at $+\infty$) gives $\ell_e = \hat{c}_e - \lambda q_t - \rho_e a_e$, $u_e = \hat{c}_e + \lambda q_t + \rho_e a_e$. The staleness correction is, with $a_{\max}$ a bound on calibration and test ages, $\Delta_{\mathrm{stale}}(t) = \sum_i \tilde{w}_i \min\big(1, 2[\varepsilon_{\mathrm{TV}}(t-u_i) + \varepsilon_{\mathrm{TV}}((t-u_i) + a_{\max})]\big)$ — the corrected two-draw form: each score involves the fresh draw at u_i *and* the previous draw at $u_i - a_i$, and the second draw's gap to the test score's is not bounded by $t - u_i$. (The implementation's reported claim line uses the single-draw form $\sum_i \tilde{w}_i \min(1, 2\varepsilon_{\mathrm{TV}}(t-u_i))$, which undercounts by at most a factor 2 plus a $2\varepsilon_{\mathrm{TV}} a_{\max}$ term — below reported precision at the swept $\varepsilon_{\mathrm{TV}}$.) The calibration-age slack is $\pi_{\mathrm{cal}} := \sup_x[\mathbb{P}(W \le x) - \mathbb{P}(W \le x - 2\langle \rho a\rangle_{\mathrm{cal}})]$ with $\langle \rho a\rangle_{\mathrm{cal}} = \max_i \rho_{e_i} a_i$ and W the relevant score magnitude ($|\eta|$ or $|\eta - \eta'|$); it vanishes as calibration re-observations happen at small ages and is bounded by $4 f_{\max}\langle \rho a\rangle_{\mathrm{cal}}$ under a bounded-density addition to A3.

Theorem 1 (T1a: observable coverage, λ=1)**.** *Under A1–A2 and the buffer hypothesis above, a re-observation $Y = Y_e(t)$ satisfies $\mathbb{P}\big(Y \in [\hat{c}_e - q_t - \rho_e a_e,\ \hat{c}_e + q_t + \rho_e a_e]\big) \ge 1 - \alpha - \Delta_{\mathrm{stale}}(t) - \pi_{\mathrm{cal}}$. (T1a is not π_{cal}-free: the term is the price of passing to pure pair scores in the repaired Lemma 1 of the companion.)*

Theorem 2 (T1b: latent-cost coverage, λ=2)**.** *Under A1–A3 and the buffer hypothesis above, $\mathbb{P}\big(c_e(t) \in [\hat{c}_e - 2q_t - \rho_e a_e,\ \hat{c}_e + 2q_t + \rho_e a_e]\big) \ge 1 - 2\alpha - 2\Delta_{\mathrm{stale}}(t) - 2\pi_{\mathrm{cal}}$; the π_{cal} term enters twice (once through T1a, once through the fresh-noise quantile step).*

Both proofs (via the weighted non-exchangeable bound of Barber et al. [2] applied to the pure pair scores, plus Anderson domination) are in the companion. The path certificate follows from a union bound over *both* bounding sides: with candidate set $\mathcal{C} \ni P_{\mathrm{lb}}$, $E_{\mathcal{C}}$ its distinct edge count, and $\alpha_{\mathrm{edge}} = \alpha'/(L_{\max} + E_{\mathcal{C}})$ ($= \alpha'/(2L_{\max})$ for the single candidate $\mathcal{C} = \{P_{\mathrm{lb}}\}$): $\mathrm{LB} = \sum_{P_{\mathrm{lb}}} \ell_e \le \mathrm{OPT}(t) \le \min_{P' \in \mathcal{C}} \sum_{P'} u_e = \mathrm{UB}$ with probability $\ge 1 - 2\alpha' - 2(L_{\max} + E_{\mathcal{C}})(\Delta_{\mathrm{stale}} + \pi_{\mathrm{cal}})$. One $L_{\max}$-edge budget cannot pay both sides (the unknown optimum's edges and the UB candidates' edges differ in general, and the min over candidates is a data-dependent selection that requires simultaneous coverage of all candidate edges); the deployed planner reports the realized-L level α'/L as its documented operating approximation, validated against ground truth (coverage 1.000 everywhere).

A.2 The certifiability threshold (T2′)

With $\bar{q}$ an upper bound on λq_t and $q_{\min} \ge 0$ a lower bound:

Theorem 3 (T2′a: achievability)**.** *If the ℓ-shortest path P (L edges, drift $\le \bar{\rho}$) is stable, then round-robin re-observation of its edges sustains, after L burn-in rounds,* $\mathrm{UB} - \mathrm{LB} \le 2L\bar{q} + \bar{\rho}\Delta L(L-1)$ at post-observation instants *(at round starts the in-round shift adds $\le 2\bar{\rho}L\Delta$; $\bar{q}$ bounds λq_t, so λ is absorbed). With lazy pre-widening at horizon B: $+2\bar{\rho}\Delta BL$. Hence ε is sustainable whenever $\varepsilon \ge 2L\bar{q} + \bar{\rho}\Delta L(L-1)$ $[+2\bar{\rho}\Delta BL]$.*

Theorem 4 (T2′b: impossibility for this construction)**.** *If a cut C (drift $\ge \rho_{\min}$) meets every s–g path in $\ge m$ edges, then every one-per-round sensing policy (all observations one-per-round, including any initial mapping — no synchronized free initial map) has, after the first round,* $\mathrm{UB} - \mathrm{LB} \ge 2mq_{\min} + \rho_{\min}\Delta\, m(m-1)$

for the certificate of this paper, with widths measured before cost-floor clipping. In particular $\varepsilon < 2mq_{\min}$ is unattainable by this construction at any sensing rate; the theorem bounds this certificate family's gap, not every conceivable sound certificate.

Static worlds ($\rho \equiv 0$) terminate for any $\varepsilon > 2Lq_\infty$; the $\rho \to 0, q \to 0$ corner recovers the deterministic scout stopping rule of Rockenbauer et al. [29]. Proofs in the companion.

A.3 The sum-aware upper certificate (T4)

The Bonferroni split α'/L prices simultaneous control of L edges into every quantile, but the upper bound concerns only *one* path (the incumbent), for which the right object is the distribution of the *sum* of deviations, whose $(1-\alpha')$ quantile scales as $\sqrt{L}$. Alongside the absolute scores, record *signed* deviations $D = Y_e(u) - \hat{c}_e$; partition the signed buffer into blocks of L consecutive samples with block sum $G_b = \sum_{i \in b} D_i$ and data-independent weight $w_b = \min_{i \in b} \rho_w^{t-u_i}$. Let $M_L(\alpha)$ be the weighted $(1-\alpha)$ quantile of the G_b with test mass at $+\infty$, and set $\mathrm{UB}_{\mathrm{sum}}(P) = \sum_{e \in P} \hat{c}_e + \lambda M_L(\alpha') + \sum_{e \in P} \rho_e a_e$.

Lemma 1 (block symmetry and domination)**.** *Under A3–A4, sums of L independent symmetric unimodal noises are symmetric unimodal (Wintner: symmetric unimodality is closed under convolution), and $\sum_{i \le L}(\eta_i - \eta'_i) \succeq_{\mathrm{st}} \sum_{i \le L} \eta'_i$ by Anderson domination at the sum level; one-sided tails of the block sums dominate one-sided tails of $\sum_i \eta'_i$ at the same level.*

Theorem 5 (T4: fixed-path sum-aware upper coverage)**.** *Let P (L edges) be chosen independently of the deviations entering $\mathrm{UB}_{\mathrm{sum}}$, and let M_L be formed with each incumbent edge's most recent score excluded from the buffer (leave-one-out: the test quantity $-\sum_{e \in P} \eta'_e$ is built from the stored draws of P's edges, which entered those edges' most recent scores; thinning fixes only calibration-internal sharing). Under A1–A4, with $\Delta^{(L)}_{\mathrm{stale}}$ the block-level analogue of Δ_{stale} (per-block TV $\le$ sum of member terms, each in the corrected two-draw form), $\pi^{(L)}_{\mathrm{cal}}$ the largest mass an interval of length $2\max_b \sum_{i \in b} \rho_i a_i$ receives under the law of the pair-sum $\sum_{i \in b}(\eta - \eta')_i$, and the λ-margin convention of Theorems 1–2,*

$$\mathbb{P}\Big(\textstyle\sum_{e \in P} c_e(t) \le \mathrm{UB}_{\mathrm{sum}}(P) \Big) \ge 1 - \lambda\alpha' - \lambda\Delta^{(L)}_{\mathrm{stale}}(t) - \pi^{(L)}_{\mathrm{cal}},$$

with $M_L = \Theta(\sqrt{L})$ for light-tailed noise versus $\Theta(L\, q_{\alpha'/L})$ for Bonferroni.

Proof. $\sum_{e \in P}(c_e(t) - \hat{c}_e) = \sum_e \delta_e - \sum_e \eta'_e$ with $|\sum \delta_e| \le \sum_e \rho_e a_e$ (A1). The blocks G_b bracket independent sums $\sum_{i \in b}(\eta - \eta')_i$ within the block's $\sum \rho a$ slack (1) (the sandwich costs the $\pi^{(L)}_{\mathrm{cal}}$ deduction); by A4 and leave-one-out the test quantity $-\sum_e \eta'_e$ is an independent draw of the dominated-side sum, its law drifting from each block's by at most the block TV (the weighted-conformal bound applied at the block level — blocks are the exchangeable units, disjointness gives independence across calibration units, thinning handles within-edge pair sharing). Lemma 1 converts block-sum quantiles into one-sided bounds on $\sum \eta'$; the λ=2 latent step repeats Theorem 2's triangle argument at the sum level (a single-event route would give the latent sum at λ=1 — an unclaimed improvement; λ=2 is kept for uniformity). □

Remark 1 (selection bias is real, measured, and gated)**.** *The incumbent is* not *chosen independently of the deviations: it minimizes estimated costs, so its $\hat{c}_e$ are biased low. Applying T4 naively to the selected incumbent drops empirical coverage from* 1.000 *to* 0.823 *in a noise-dominated static regime. The deployed protocol therefore uses the sum-aware bound only through a* freshness gate *— every incumbent edge re-observed since the path last changed — so that observations are independent of the selection and Theorem 5 applies conditionally on the gate (the gate event is measurable w.r.t. pre-refresh data, and on it every $\hat{c}_e$ entering $\mathrm{UB}_{\mathrm{sum}}$ postdates the selection, so conditioning leaves the post-selection noise laws intact). With κ-hysteresis stabilizing the incumbent (its second role), the gate recovers coverage to* 0.916–0.966 *while keeping the tightening; there is no effect under strong drift, where age widths dominate and the gate rarely opens. Gate-closed rounds use the unconditional Bonferroni bound, so no selection-bias mass is left uncontrolled.*

A.4 The lower bound cannot be sum-aware (T5)

Theorem 6 (T5: uniform LB impossibility)**.** *On the layered graph with L layers of width w (w^L paths), prior $c_e \sim_{iid} \mathcal{N}(\mu, \sigma^2)$, one observation $Y_e = c_e + \eta_e$, $\eta_e \sim \mathcal{N}(0, \sigma^2)$, per edge: fix $\alpha < 1/4$ and any $c < \sqrt{2} - 1$. There is a $w_0(c)$ such that for all $w \geq w_0$, any estimator with $\mathbb{P}(\mathrm{LB} \leq \mathrm{OPT}) \geq 1 - \alpha$ satisfies $\mathbb{P}\big(\widehat{\mathrm{OPT}} - \mathrm{LB} \geq c\, L\sigma\sqrt{\ln w}\big) \geq 1 - \alpha - o(1)$ as $L \to \infty$, and hence $\mathbb{E}[(\widehat{\mathrm{OPT}} - \mathrm{LB})_+] \geq (1 - \alpha - o(1))\, c\, L\sigma\sqrt{\ln w}$, where $\widehat{\mathrm{OPT}}$ is the posterior-mean shortest path. (The positive-part/probability form is essential: an unconstrained expectation bound is gameable by spiking* LB *upward on an α-small set.) Per-edge Bonferroni achieves slack $O(L\sigma\sqrt{\ln(wL/\alpha)})$: the asymmetric certificate is order-optimal up to a $\sqrt{1 + \ln L / \ln w}$ factor.*

Proof. Compare the polymer (first-passage) constants of the two fields. Unconditionally, the posterior means $m_e = \mathbb{E}[c_e \mid Y] = (\mu + Y_e)/2$ are iid $\mathcal{N}(\mu, \sigma^2/2)$ and the true costs are iid $\mathcal{N}(\mu, \sigma^2)$. For an iid $\mathcal{N}(\mu, s^2)$ edge field the layered-graph minimum is $L\mu - Ls\sqrt{2\ln w}\,(1 + o(1))$ with probability $\to 1$: the greedy path (layer-by-layer minimum of w fresh iid values, mean $\mu - s\sqrt{2\ln w}(1 - o_w(1))$ per layer, $O(s\sqrt{L})$ concentration of the sum) gives the upper bound, and the first-moment bound over w^L paths, each $\mathcal{N}(L\mu, Ls^2)$, gives the matching lower bound ($w^L \exp(-L \ln w(1+\varepsilon)^2) \to 0$). Applying this with $s = \sigma/\sqrt{2}$ and $s = \sigma$: $\widehat{\mathrm{OPT}} = L\mu - L\sigma\sqrt{\ln w}(1 + o(1))$ and $\mathrm{OPT} = L\mu - L\sigma\sqrt{2\ln w}(1 + o(1))$, so $\widehat{\mathrm{OPT}} - \mathrm{OPT} = (\sqrt{2} - 1)L\sigma\sqrt{\ln w}(1 + o(1))$ with probability $\to 1$; $w_0(c)$ absorbs the $o_w(1)$ terms (no universal constant down to $w = 2$ is claimed). On the intersection of the coverage event $\{\mathrm{LB} \leq \mathrm{OPT}\}$ and the two polymer events — probability $\geq 1 - \alpha - o(1)$ — we get $\widehat{\mathrm{OPT}} - \mathrm{LB} \geq \widehat{\mathrm{OPT}} - \mathrm{OPT} \geq (\sqrt{2} - 1 - o(1))L\sigma\sqrt{\ln w}$. The matching upper bound is the per-edge union at $\alpha_{\mathrm{edge}} = \alpha/|E|$, $|E| \leq 2Lw^2$. (The previously reported simulation — median deficit 0.71–0.80 $\times\, L\sigma\sqrt{\ln w}$ for $w \in [10, 50]$ — measures the greedy D-deficit of an earlier sketch, not $\widehat{\mathrm{OPT}} - \mathrm{OPT}$, whose constant is $\sqrt{2} - 1 \approx 0.41$; it is retained only as a check on the per-layer greedy constant.) □

The mechanism is selection: with exponentially many candidate paths some path's plausible downside is *linearly* deep, and a sound LB must respect it. The upper side escapes because it prices one chosen path; the lower side cannot. This makes the certificate's asymmetry a theorem, not a limitation.

A.5 Non-exchangeable round two: sum-level pricing (T8)

Two constructions replace the α'/L union with a single level-α' quantile of a scalar, and both survive the age weights.

LP-shift staleness (optional). The TV-Lipschitz Δ_{stale} can be replaced by a Lévy–Prokhorov worst-case quantile [1]: under an LP ambiguity set of local radius ε and global mass ρ, the worst-case $(1-\alpha)$ quantile is $\mathrm{Quant}_{1-\alpha+\rho}(\cdot) + \varepsilon$ and the worst-case coverage of a threshold q is $F(q - \varepsilon) - \rho$. This raises the *quantile level* rather than deducting from the *claim level*, so it stays above the effective-sample-size annealing floor while additionally pricing never-observed edges. It is offered as an alternative staleness model; A2/Δ_{stale} remains the default.

Theorem 7 (T8a: group-sum upper certificate)**.** *Fix P (L edges) chosen independently of the calibration deviations. Let $\{G_j\}$ be group-sum scores formed by drawing, per calibration group j, a* pairwise-disjoint *L-subset of signed per-edge deviations and summing them, with data-independent age weights, and $M(\alpha')$ their weighted $(1-\alpha')$ quantile with test point $+\infty$. Then under A1–A4 with the λ-margin convention,*

$$\mathbb{P}\Big(\textstyle\sum_{e \in P} c_e(t) \leq \sum_{e \in P} \hat{c}_e + \lambda M(\alpha') + \sum_{e \in P} \rho_e a_e \Big) \geq 1 - \lambda\alpha' - \lambda\Delta^{(L)}_{\mathrm{stale}}(t) - \pi^{(L)}_{\mathrm{cal}},$$

with $M(\alpha') = \Theta(\sqrt{L})$ for light-tailed noise.

Proof. Identical in structure to Theorem 5: the disjoint group sum is the calibration unit (disjointness gives independence across units), weighted-conformal validity applies at the group level, and the λ=2 latent step repeats Theorem 2's triangle argument at the sum. This is the drift-retrofitted, age-weighted form of the symmetric-calibration group sum of Luo and Zhou [23]. For *overlapping* groups the implementation exposes a slack δ (a maximum pairwise edge-sharing *count*, not a probability); no coverage deduction has been derived for it, so the overlapping variant is a disclosed heuristic outside this theorem. □

Theorem 8 (T8b: joint per-edge (max-score) pricing)**.** *Fix P (L edges) chosen independently of the calibration deviations (the incumbent is optimizer-selected, so the freshness-gate discussion of Remark 1 applies verbatim). Let Q be the weighted $(1-\alpha')$ quantile of the per-block* maximum *signed score over disjoint length-L blocks, test point $+\infty$, with the test scores s_e the drift-adjusted magnitudes evaluated at the query time t. Then $\bigcap_{e\in P}\{s_e \le Q\} = \{\max_e s_e \le Q\}$ holds with probability $\ge 1-\alpha'-\Delta^{(L)}_{\text{stale}}(t)$ under block exchangeability of the thinned buffer, so $Q+\rho_e a_e \ge |Y_e - \hat{c}_e|$ simultaneously for all $e \in P$ at that level — the same per-edge magnitude Bonferroni certifies, calibrated jointly rather than by the α'/L union.*

Proof. The set identity is immediate; validity is the module's own split-conformal quantile with the $\cup\{+\infty\}$ test point, so soundness rests on no max-score-specific constant, and the age-weighted variant inherits the weighted-conformal bound — including its staleness deduction, which the displayed level now carries [18]. Block exchangeability (not per-edge) is the standing assumption; the max-of-L needs $\sim L\times$ the samples per block. □

Remark 2 (which functional wins is empirical, and two-signed)**.** *T8a and T8b both remove the union factor but differ in width by which functional of the block is calibrated. On long real-traffic paths the buffer holds few length-L blocks, so the block-*max *of T8b is coarse and* loses *to Bonferroni (+24.7%/+25.1% measured), while the sum functionals win — -23.6% for the T4 block quantile and -26.6% for the group-sum of T8a (Section 6.12). Under* independent *edges Bonferroni is already tight and neither helps; the joint price pays only under positive edge correlation (measured* 16.5% *narrower at correlation* 0.9*, L=20).*

A.6 The two-tier certificate: a-priori shrink is impossible (T9)

Theorem 9 (T9: no a-priori shrink from windowed evidence under drift)**.** *Let the radius be shrunk to $k(q_t+\rho a)$, $k<1$ a measurable function of the observed window $\{R_i\}_{i\le t}$. For every $\alpha < 1/4$ there is an environment in the drift class A1–A2 (indeed a pair indistinguishable on the score stream) on which the shrunk per-edge interval's next-round observable miscoverage exceeds α on every finite-radius round: no such rule retains a distribution-free next-round coverage guarantee $\ge 1-\alpha$ over A1–A2. (A3 is not assumed, which the construction exploits.)*

Proof. Single edge, re-observed every round (age Δ); iid two-point noise $\eta = \pm B$, $B > \Delta\rho$ (symmetric, time-invariant, not unimodal — A3 is not assumed); drift at the A1 boundary with hidden direction, $c_e(t) = c_0 + s\rho t$, $s \in \{\pm 1\}$. With $g := \eta - \eta' \in \{0, \pm 2B\}$ (probabilities $1/2, 1/4, 1/4$), each score $R = |s\rho\Delta + g| - \rho\Delta$ takes values $\{0, 2B, 2B-2\rho\Delta\}$ with probabilities $1/2, 1/4, 1/4$ — the *same law for both* s, so the score stream is iid, independent of s, and any measurable rule $k(\{R_i\})$ has the same distribution in both environments (the direction is undetectable). Whenever the certificate radius is finite, $q_t \le \max_i R_i \le 2B$ deterministically; the next-round observable deviation $|Y_{t+1} - \hat{c}_e| = |s\rho\Delta + g|$ equals $2B + \rho\Delta$ exactly when $g = s \cdot 2B$ — an atom of probability $1/4$ at the un-shrunk radius's maximal value. Since $k < 1$ and $q_t \le 2B$, $k(q_t + \rho\Delta) < 2B + \rho\Delta$, so the shrunk interval misses the atom and mis-covers with probability $\ge 1/4 > \alpha$ on every finite-radius round, in both environments; the un-shrunk interval covers it once $q_t = 2B$, which the weighted quantile delivers eventually a.s. at any $\alpha < 1/4$. Under a bounded-density addition the same construction with mass just inside the radius makes the impossibility quantitative in $1-k$ rather than absolute. The detection-based patch fares no better by the standard change-point trade-off (bounded false-alarm rate forces unbounded worst-case detection delay — used qualitatively): the pre-alarm rounds are exactly the rounds a coverage consumer needed guaranteed; empirically this is the CIA collapse $0.95 \to 0.20$ under staleness (Figure 5). □

Theorem 10 (a-posteriori licensed radius)**.** *Let $x_i(k) = \mathbf{1}\{|Y_i - \hat{c}_i| > k(q + \rho a_i)\}$ be the shrunk-interval violation indicator on the fresh score stream, and $\mathcal{K}$ a finite grid of shrink factors fixed in advance. For each $k \in \mathcal{K}$ a betting confidence sequence [36] at level $\delta/|\mathcal{K}|$ yields a time-uniform upper bound* $\mathrm{UCB}_t(k)$ *on the running average of* conditional *violation rates $\bar{\mu}_t(k) = \frac{1}{t}\sum_{i\le t}\mathbb{E}[x_i(k) \mid \mathcal{F}_{i-1}]$ (the WSR sequence controls conditional means of a bounded adapted stream, not the raw empirical mean): $\mathbb{P}(\exists t : \mathrm{UCB}_t(k) < \bar{\mu}_t(k)) \le \delta/|\mathcal{K}|$, hence simultaneously over the grid at level δ. The licensed radius $k^\star(q+\rho a)$, $k^\star = \max\{k \in \mathcal{K} : \mathrm{UCB}_t(k) \le \alpha'\}$, carries the honest, anytime-valid, self-revoking statement: over this deployment so far, the shrunk interval's conditional mis-coverage averaged $\le \alpha'$ with $1-\delta$ validity.*

Tier-1 is the a-priori distribution-free [LB, UB] (never shrunk, bit-identical with the license enabled); Tier-2 is $k^\star(q+\rho a)$ with the weaker a-posteriori claim (-62.4% width at 0.51% shadow miscoverage on real METR-LA; k=0.5 floor binds on 82% of rounds). Safety gates read Tier-1; resource allocation may read Tier-2.

A.7 Team and multi-agent certificates (T10)

Theorem 11 (T10a: additive team certificate)**.** *Let N agents share one drifting-cost graph and* one *conformal edge-price store, so every edge age a_e is global, and let each agent's certificate* $(\mathrm{LB}_i, \mathrm{UB}_i, \mathrm{conf}_i)$ *come from that store. Then* $\mathrm{LB}_{\mathrm{team}} := \sum_i \mathrm{LB}_i \le \mathrm{OPT}_{\mathrm{team}} \le \sum_i \mathrm{UB}_i =: \mathrm{UB}_{\mathrm{team}}$ *with probability* $\ge \mathrm{conf}_{\mathrm{team}} := \max(0, 1 - \sum_i (1-\mathrm{conf}_i))$*, where* $\mathrm{OPT}_{\mathrm{team}} = \sum_i \mathrm{OPT}_i$*. What is exact is the decomposition* $\mathrm{OPT}_{\mathrm{team}} = \sum_i \mathrm{OPT}_i$ *— a modeling identity of the uncoupled objective, not a probabilistic claim; each* LB_i *remains strictly conservative.*

Proof. Each summand is sound over the shared store by the single-agent guarantee; the age a_e is global, so no agent sees a staler edge than the store records. The team objective separates, so summing the per-agent inequalities gives the bracket; the confidence follows from a union bound over the N miscoverage events, floored at 0. □

Remark 3 (why only the additive bound)**.** *A congestion-coupled joint certificate ($c_e(m) = \phi_e(1+\beta_e m)$) with a decision-focused allocator is tighter on synthetic bottlenecks (ratio additive/joint up to* 1.78 *at N=8) but* looser *on real METR-LA (*0.90–0.91*; Table 13), because route diversity lets agents avoid the few congested edges. The joint model is falsified as a deployable object and not shipped. Separately, age-binning the per-edge widths tightens the additive gap* 2.1–2.4× *at coverage* 1.000*, while a joint block-conformal team quantile over-shoots coverage (*0.86 *at $N \ge 2$) by the same selection bias that gates T4.*

Certified MAPF (the corridor lift). Each agent's low-level state is (vertex, certified arrival window $[lo, hi]$); the certified duration window of edge e used no later than hi is $[\hat{\ell}_e, \hat{u}_e]$, priced at the *latest* certified use (staleness widening is monotone in age, so the window at hi contains the window at any earlier use). Conventions, pinned per side because soundness lives there: integer realized durations; $\hat{u}_e = \lceil \kappa(\hat{c}_e + q + \rho_e(hi - t_{\mathrm{obs}}))\rceil$ with $\kappa \ge 1$, never clipped downward (a horizon cap triggers abstention, not truncation); $\hat{\ell}_e = \lfloor \hat{c}_e - q - \rho_e(hi - t_{\mathrm{obs}})\rfloor$ with the *unscaled* radius (a κ-inflated lower bound would be anti-conservative), clipped only below at the cost floor. CBS branches on overlapping certified windows.

Theorem 12 (T10b: certified-MAPF soundness)**.** *Assume the per-side conventions above and single-visit plans (each agent traverses any edge at most once, so its use time is determined by predecessor edges and independent of that edge's noise). Price the selection-free universe $\mathcal{U} = \{1,\dots,N\} \times E$ — fixed by the instance, not the returned plan — at $\alpha_{\mathrm{edge}} = \alpha_{\mathrm{team}}/(N|E|)$, and let $\mathcal{E}$ be the event that for every $(i,e) \in \mathcal{U}$ the realized duration at any plan-measurable use time $\tau \le hi$ deviates from $\hat{c}_e$ by at most $q + \rho_e(\tau - t_{\mathrm{obs}})$ (hence lies in the window priced at hi, both sides). Then $\mathbb{P}(\mathcal{E}) \ge 1 - \alpha_{\mathrm{team}} - \sum_{\mathcal{U}}(\Delta_{\mathrm{stale}} + \pi_{\mathrm{cal}})$-terms and, on $\mathcal{E}$: (C1) the realized execution of Π is collision-free; (C2) $\sum_i \mathrm{LB}_i \le \mathrm{OPT}_{\mathrm{team}} \le \mathrm{cost}(\Pi) \le \sum_i \mathrm{UB}_i$; and (C3) if no jointly certified conflict-free plan is found within budget the planner returns* ABSTAIN.

Proof. $\mathbb{P}(\mathcal{E}^c) \le \sum_{(i,e)\in\mathcal{U}}(\alpha_{\mathrm{edge}} + (\Delta+\pi)$-terms) by the per-pair weighted-conformal bound (single-visit gives $\tau \perp$ edge noise) and a union bound over a universe fixed *before* plan selection — the earlier budget over the returned plan's $\le N\bar{L}$ priced durations ran over a data-dependent set (CBS favours low-$\hat{u}$ corridors: the winner's curse gated for T4) and did not bound $\mathbb{P}(\mathcal{E}^c)$; re-observing plan edges post-selection (a freshness-gate analogue) is the tighter alternative. On $\mathcal{E}$, induction over each agent's path (base window $[t_0, t_0]$; a traverse adds a covered $[\hat{\ell}, \hat{u}]$ via the deviation bound at τ, monotone widening to hi, and the ceil/floor conventions; a sync-wait departs at a deterministic $T \ge$ the certified latest arrival) shows every realized occupancy lies inside its certified window; CBS returns only nodes whose certified windows are pairwise disjoint, so realized occupancies cannot intersect (C1). Realized cost $\le \sum_i \mathrm{UB}_i$ edge-wise; any conflict-free plan costs $\ge \sum_i$(unconstrained ℓ-shortest)$= \mathrm{LB}_{\mathrm{team}}$, whose lower-coverage events sit on edges generally *outside* Π — they are on-event because $\mathcal{E}$ covers all of $\mathcal{U}$, which the earlier plan-only event did not (C2). Budget exhaustion triggers (C3). □

Remark 4 (honest scope: soundness yes, completeness no, α inert, sub-floor label)**.** *Window-CBS completeness and optimality are not claimed: interval branching can exclude jointly-feasible schedules, which costs success rate, never soundness. At P0 scale the certificate realizes* 0 *collisions everywhere against* 0–100% *for point-estimate MAPF, but the per-edge budget sits below the finite-sample support floor: when* $\alpha_{\text{edge}} < \tilde{w}_{n+1} \approx 1/(n_{\text{eff}}+1)$ *the licensed window is infinite, and the max-score fallback licenses only per-edge miscoverage* $\approx \tilde{w}_{n+1}$*. At P0 numbers (*$\alpha_{\text{team}} = 0.1$*,* $N\bar{L} = 192$*,* ~ 672 *scores):* $\alpha_{\text{edge}} \approx 5.2 \times 10^{-4}$ *against supportable* $\approx 1.5 \times 10^{-3}$*, so the supportable team level is* ≈ 0.71*, not* 0.90 *— the tables carry the honest label (nominal* $\alpha = 0.1$*; supportable* ≈ 0.71*), and empirical collisions are zero either way. q floors to the max score in* 100% *of certified cells and certified*(α) *is bit-identical to worst-case inflation — the probabilistic knob is inert until sum-level per-agent pricing (T8a at level* α/N*) lifts the level above the floor; T8a per agent reintroduces the fixed-path hypothesis at team scale, so the T4 gate discussion must be replayed there.*

B Per-experiment reproduction

Every experiment is regenerated by a single driver script; Table 16 maps each result to its script and result file. Repository scripts run from the shipped library (`certflow`); the multi-agent and attack drivers live in the research-history package. All runs are seeded and CPU-only unless a timing row states otherwise.

Table 16: Reproduction map. Repository = the shipped `certflow` library (`scripts/`); project = the research-history package (`experiments/`). Result files are under `docs/results/` (repo) or `experiments/results/` (project).

Result	Driver script	Result file
Tier-0 coverage + T4 (Tab. 2)	`scripts/run_tier0.py`	`tier0-coverage.md`
Gaussian break (Tab. 3)	`scripts/run_gaussian_break.py`	`gaussian-break.md`
Tier-2 regret (Tab. 4)	`scripts/run_tier2.py`	`tier2-regret.md`
METR-LA / PEMS-BAY (Tab. 5)	`scripts/run_metr_la.py`	`metr-la.md`
MovingAI (Tab. 6)	`scripts/run_movingai.py`	`movingai.md`
CH / road scale (§6.7)	`scripts/run_ch.py`, `run_roadnet.py`, `run_scale.py`	`published-speed-comparison.md`, `scale.md`
Tier-1 latency (Tab. 7, 19)	`scripts/run_tier1_latency.py`, `run_repeated_queries.py`	`tier1-latency.md`
Ablations (Tab. 8, 18)	`scripts/run_ablations.py`	`ablations.md`
Lifelong (Tab. 9, Fig. 15)	`scripts/run_lifelong.py`	`lifelong.md`
T7 / feature regimes (§6.11)	`scripts/run_feature_regimes.py`, `study_spatial_predictor.py`	`feature-regimes.md`, `spatial-predictor-study.md`
Width tightening (Tab. 10, Fig. 8)	`scripts/run_width_attack.py`	`width_attack.json`, `width-attack-2026.md`
WATCH observability (Fig. 14)	`scripts/run_watch_testability.py`, `run_live_wiring.py`	`watch_testability.json`, `live-wiring-2026.md`
Hybrid sensing (Tab. 11, Fig. 11)	`scripts/run_hybrid_sensing.py`	`hybrid_real_metrla.json`, `hybrid-sensing-2026.md`
No crossover (Tab. 12, Fig. 9)	`scripts/run_crossover_regret.py`	`crossover_regret.json`, `crossover-2026.md`
Team certificate (Fig. 12)	`experiments/exp_coverage_tightening.py`	`coverage_tightening.json`, `multiagent.md`
TEAM-CERT falsification (Tab. 13)	`experiments/exp_congestion_cert.py`	`congestion_cert.json`
Certified-MAPF P0 (Tab. 14, 17, Figs. 10, 13)	`experiments/certmapf_p0.py`	`certmapf_p0.json`, `certmapf-design.md`

C Additional experimental tables

The full certified-MAPF P0 sweep. Table 17 reports every one of the 24 conditions (2 maps $\times$ 4 drift rates $\times$ 3 team sizes, 25 seeds each) summarized by Table 14. Certified plans realize *zero* realized collisions on every solved run in all 600 runs; realized cost lies in [LB, UB] at rate 1.00 in every cell. The point-estimate collision rate rises with drift and N; the certified planner's cost premium and abstention are the honest price (Figure 13).

Table 17: Complete certified-MAPF P0 sweep (600 runs; cert = certified at *nominal* α=0.1 — the per-edge budget sits below the finite-sample support floor at this calibration size, so the supportable team level is ≈ 0.71, not 0.90; Section 7.3). solve% among 25 seeds; coll% = fraction of executed runs with ≥ 1 realized conflict; cost = mean over instances solved by *both* planners; gap = mean certified UB − LB. "—" where the certified planner solves nothing (N=8: no disjoint corridors on the 8×8 grid, so 100% abstention — sound but useless). cert coll% is **0** in every cell. Source: `certmapf_p0.json`.

map	ρ/N	pt solve%	pt coll%	cert solve%	cert coll%	pt cost	cert cost	cert gap
open	0.00/2	100	0	100	**0**	38.6	41.4	22.6
open	0.00/4	100	4	92	**0**	78.5	84.8	47.2
open	0.00/8	92	57	0	—	—	—	—
open	0.01/2	100	8	100	**0**	38.9	42.0	35.1
open	0.01/4	100	20	92	**0**	78.7	91.9	77.0
open	0.01/8	92	57	0	—	—	—	—
open	0.03/2	100	16	100	**0**	38.7	49.1	69.5
open	0.03/4	100	28	84	**0**	78.7	114.8	151.7
open	0.03/8	92	65	0	—	—	—	—
open	0.06/2	100	12	100	**0**	38.4	61.2	112.6
open	0.06/4	100	40	68	**0**	79.4	144.3	233.2
open	0.06/8	92	70	0	—	—	—	—
obstacle	0.00/2	96	4	96	**0**	47.7	49.0	23.1
obstacle	0.00/4	88	18	76	**0**	102.3	115.8	65.7
obstacle	0.00/8	36	89	0	—	—	—	—
obstacle	0.01/2	96	4	96	**0**	47.6	53.4	41.2
obstacle	0.01/4	88	23	56	**0**	98.6	128.8	104.9
obstacle	0.01/8	36	78	0	—	—	—	—
obstacle	0.03/2	96	4	92	**0**	46.6	58.6	94.3
obstacle	0.03/4	88	23	24	**0**	93.3	133.8	178.5
obstacle	0.03/8	36	100	0	—	—	—	—
obstacle	0.06/2	96	4	84	**0**	46.6	65.7	139.0
obstacle	0.06/4	88	36	12	**0**	101.3	164.3	317.7
obstacle	0.06/8	36	100	0	—	—	—	—

Ablations above the certifiability floor. The ablation table in the body (Table 8) runs at ε=5, below the T2′ floor at $L \approx 14$, so certification is near-zero everywhere and the maintenance and backstop rows are uninformative by design. Table 18 reruns the same suite at ε=12 (above the floor), where those rows speak: maintenance contributes +43% relative certified rounds (4.3 vs 3.0), pre-widening is decisive near the floor (B=0 certifies 3× more than B=10), and κ's churn suppression reproduces (0.26 vs 1.76). The no-backstop row reads slightly higher because the forced round-robin, kept for the T2′a theorem, can occasionally displace a better greedy pick — the small price of turning achievability from an empirical hope into a guarantee.

Table 18: Ablations at ε=12, above the T2′ floor (annealed defaults; 8×8, ρ=0.02, 20 seeds × 300 rounds). Rows ordered by cert%, the metric that was pinned at zero below the floor. Source: `ablations.md`.

condition	cov.	valid%	cert%	gap	churn	flap%
B=0 (no pre-widen)	1.000	94.9	12.4	19.60	0.22	1.3
no-backstop	1.000	94.5	7.0	21.21	0.25	1.3
full (κ on, B=10)	1.000	94.5	4.3	21.93	0.26	1.6
no-κ	1.000	94.5	4.3	21.93	1.76	17.8
no-maintenance	1.000	94.5	3.0	21.93	0.25	1.6
B=20	1.000	94.2	0.6	23.72	0.26	1.6

External speed anchor for Tier-1. So that the Tier-1 incremental-repair speedups (Table 7) are not measured against a slow home-grown scratch planner, Table 19 anchors the from-scratch reference to a mature third-party Dijkstra (`networkx`) on identical updated snapshots. The engine's own scratch planner is $\approx 1.37\times$ *faster* than `networkx`, so the 9.7–23.8× incremental speedups are, if anything, conservatively stated. (`networkx` has no incremental-update API, which is the entire point of D* Lite and not a like-for-like

axis.)

Table 19: External from-scratch anchor (`run_repeated_queries.py`, 3 seeds × 50 queries × 2 cost regimes; p50 ms). Lower engine/networkx ratio = the engine's scratch baseline wins by more. Source: `tier1-latency.md`.

size	engine Dijkstra p50	networkx Dijkstra p50	networkx p95	engine/networkx
20×20	0.18	0.25	0.38	0.72
40×40	0.86	1.17	1.82	0.73

References

[1] Liviu Aolaritei, Youssef Marzouk, Zheyu Oliver Wang, Julie Zhu, and Michael I. Jordan. Conformal prediction under Lévy–Prokhorov distribution shifts: Robustness to local and global perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.14105.

[2] 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. arXiv:2202.13415.

[3] Hannah Bast, Daniel Delling, Andrew V. Goldberg, Matthias Müller-Hannemann, Thomas Pajor, Peter Sanders, Dorothea Wagner, and Renato F. Werneck. Route planning in transportation networks. *Algorithm Engineering: Selected Results and Surveys, LNCS 9220*, pages 19–80, 2016.

[4] Zahy Bnaya, Ariel Felner, and Solomon Eyal Shimony. Canadian traveler problem with remote sensing. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.

[5] Lijie Chen, Anupam Gupta, Jian Li, Mingda Qiao, and Ruosong Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Proceedings of the Conference on Learning Theory (COLT)*, 2017.

[6] Sanjiban Choudhury, Shervin Javdani, Siddhartha Srinivasa, and Sebastian Scherer. Near-optimal edge evaluation in explicit generalized binomial graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[7] Christopher M. Dellin and Siddhartha S. Srinivasa. A unifying formalism for shortest path problems with expensive edge evaluations via lazy best-first search over paths with edge selectors. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 2016.

[8] Daniel Delling, Andrew V. Goldberg, Thomas Pajor, and Renato F. Werneck. Customizable route planning in road networks. *Transportation Science*, 51(2):566–591, 2017.

[9] Camil Demetrescu, Andrew V. Goldberg, and David S. Johnson, editors. *The Shortest Path Problem: Ninth DIMACS Implementation Challenge*, volume 74 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, 2009.

[10] Marco Dorigo, Mauro Birattari, and Thomas Stützle. Ant colony optimization. *IEEE Computational Intelligence Magazine*, 1(4):28–39, 2006.

[11] Erick Fuentes, Jared Strader, Ethan Fahnestock, and Nicholas Roy. Belief roadmaps with uncertain landmark evanescence. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[12] Etienne Gauthier, Francis Bach, and Michael I. Jordan. E-values expand the scope of conformal prediction. *arXiv preprint arXiv:2503.13050*, 2025.

[13] Robert Geisberger, Peter Sanders, Dominik Schultes, and Daniel Delling. Contraction hierarchies: Faster and simpler hierarchical routing in road networks. In *Proc. Workshop on Experimental Algorithms (WEA)*, pages 319–333, 2008.

[14] Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.00170.

[15] Daniel Harabor and Alban Grastien. Improving jump point search. In *Proc. Int. Conf. on Automated Planning and Scheduling (ICAPS)*, pages 128–135, 2014.

[16] Kalvik Jakkala and Srinivas Akella. Informative path planning with guaranteed estimation uncertainty. *IEEE Robotics and Automation Letters (RA-L)*, 2026.

[17] Sven Koenig and Maxim Likhachev. D* Lite. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 476–483, 2002.

[18] Varun Kotte. PASC: Pipeline-aware conformal prediction with joint coverage guarantees for multi-stage NLP and LLM pipelines, 2026. arXiv:2605.18812.

[19] Tomáš Krajńík, Jaime P. Fentanes, Joao M. Santos, and Tom Duckett. FreMEn: Frequency map enhancement for long-term mobile robot autonomy in changing environments. *IEEE Transactions on Robotics*, 33(4):964–977, 2017.

[20] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2018.

[21] Maxim Likhachev, Geoffrey J. Gordon, and Sebastian Thrun. ARA*: Anytime A* with provable bounds on sub-optimality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2003.

[22] Maxim Likhachev, Dave Ferguson, Geoffrey Gordon, Anthony Stentz, and Sebastian Thrun. Anytime dynamic A*: An anytime, replanning algorithm. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 2005.

[23] Rui Luo and Zhixin Zhou. Conformalized interval arithmetic with symmetric calibration. *arXiv preprint arXiv:2408.10939*, 2024.

[24] Aditya Mandalika, Sanjiban Choudhury, Oren Salzman, and Siddhartha Srinivasa. Generalized lazy search for robot motion planning: Interleaving search and edge evaluation via event-based toggles. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 2019.

[25] Zhiting Mei, Anushri Dixit, Meghan Booker, Emily Zhou, Mariko Storey-Matsutani, Allen Z. Ren, Ola Shorinwa, and Anirudha Majumdar. Perceive with confidence: Statistical safety assurances for navigation with learning-based perception. In *Conference on Robot Learning (CoRL)*, 2024.

[26] Willie Neiswanger, Ke Alexander Wang, and Stefano Ermon. Bayesian algorithm execution: Estimating computable properties of black-box functions using mutual information. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[27] Mike Phillips, Benjamin Cohen, Sachin Chitta, and Maxim Likhachev. E-Graphs: Bootstrapping planning with experience graphs. In *Robotics: Science and Systems (RSS)*, 2012.

[28] Drew Prinster, Xing Han, Anqi Liu, and Suchi Saria. WATCH: Adaptive monitoring for AI deployments via weighted-conformal martingales. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. arXiv:2505.04608.

[29] Friedrich M. Rockenbauer, Jaeyoung Lim, Marcus G. Müller, Roland Siegwart, and Lukas Schmid. Traversing mars: Cooperative informative path planning to efficiently navigate unknown scenes. *IEEE Robotics and Automation Letters*, 2025.

[30] David M. Rosen, Julian Mason, and John J. Leonard. Towards lifelong feature-based mapping in semi-static environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[31] Guni Sharon, Roni Stern, Ariel Felner, and Nathan R. Sturtevant. Conflict-based search for optimal multi-agent pathfinding. *Artificial Intelligence*, 219:40–66, 2015.

[32] Stephen L. Smith, Mac Schwager, and Daniela Rus. Persistent robotic tasks: Monitoring and sweeping in changing environments. *IEEE Transactions on Robotics*, 28(2):410–426, 2012.

[33] Nathan R. Sturtevant. Benchmarks for grid-based pathfinding. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(2):144–148, 2012.

[34] Lingxuan Tang, Rui Luo, Zhixin Zhou, and Nicolo Colombo. Enhanced route planning with calibrated uncertainty set. *arXiv preprint arXiv:2503.10088*, 2025.

[35] Atsushi Tero, Seiji Takagi, Tetsu Saigusa, Kentaro Ito, Dan P. Bebber, Mark D. Fricker, Kenji Yumiki, Ryo Kobayashi, and Toshiyuki Nakagaki. Rules for biologically inspired adaptive network design. *Science*, 327(5964):439–442, 2010.

[36] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

[37] Eyal Weiss, Ariel Felner, and Gal A. Kaminka. Tightest admissible shortest path. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, volume 34, pages 643–652, 2024.