DO REINFORCEMENT LEARNING CONTROLLERS USE LIVE STATE? A DIAGNOSTIC FOR METRO OPERATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A learned transit controller that beats a timetable has not shown that it uses the day's observations: recurring demand and an adaptive planning layer can supply the gain. We introduce a diagnostic framework that attributes a learned controller's gain to the calendar, to the planner's own feedback and to the learned layer's use of live state. It reconstructs each layer from calendar information alone, checks every run against its archived trajectory by exact replay, values single decisions by branching from identical states, asks which information predicts those values, and certifies its own power under declared demand perturbations; a proposition states what the reconstruction gap measures. Applied to 18 residual proximal policy optimization (PPO) checkpoints on a Hangzhou metro simulation driven by real demand, the framework finds three things. The best gain transfers but the feedback does not: on four sealed dates opened once after every method was frozen, the best checkpoint retains a 2.63 ± 0.87 s/passenger advantage over the planner (3.42 on development dates) but none over its calendar reconstruction, which is better than the live policy in six of eight cases. The instrument is not blind: the planner given live queues, a hand-written rule and an analytic dwell rule all move under the perturbations; the Hangzhou residual learners do not. The three escape routes from that verdict—too little authority, a training distribution that never departs from the calendar, an instrument that cannot see—were each pre-registered and tested: wider authority makes decisions state-dependent, perturbation training does not make that useful, and neither makes a policy better than the planner or rule it modifies. On six Beijing metro lines, where the learned layer holds trains at the queue it observes, the learners beat their same-day calendar tables by 0.3–1.2 units of the objective yet fall below the analytic rule in all 630 paired cases; on nine Shanghai lines the certificate reads within 0.01 of zero for the rule itself and the protocol declines to rule. The framework needs a learned layer on an adaptive planner or rule in a setting with recurring context; all three instances here are metro control.

1 INTRODUCTION

A residual controller that shifts a metro line's next departure by a minute saves several seconds per passenger over the planner it sits on. Before deploying it, an operator wants to know whether it did so by reading the platform queues as the morning unfolded or by learning that Monday's peak arrives four minutes early. Only the first is a contribution of online feedback, and only the first justifies the sensing, inference and safety machinery that a deployed learned controller requires. A comparison with a timetable or a rule cannot tell the two apart: metro demand recurs by weekday and time of day, and the optimization layer beneath the learner already adapts to fleet state and forecasts, so either explanation produces the same saving.

Our diagnostic rests on one idea. The calendar is the natural open-loop policy, so a layer may claim online feedback only if it beats its own calendar reconstruction, with every run checked by exact replay, and only after the instrument has been shown to move where feedback is known to exist. Four operations implement it: *calendar-conditioned reconstruction* (P1), *counterfactual rollouts from identical simulator states* (P2), a comparison of the *information sets* that predict those counterfactual

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

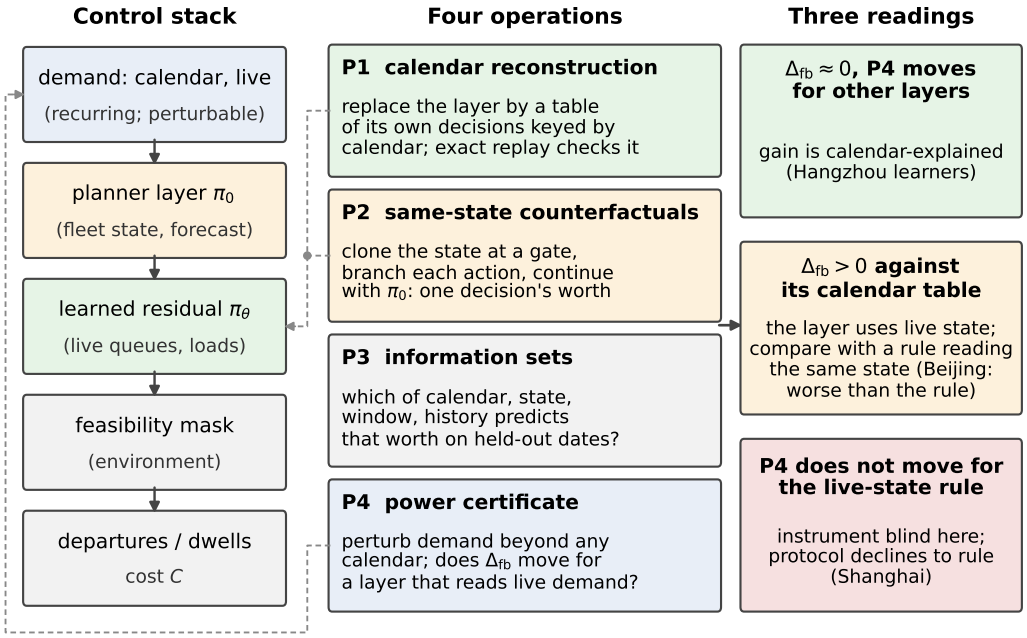


Figure 1: The framework. Left: the control stack, with the layer each operation touches. Middle: the four operations, applied to one layer at a time. Right: the three readings the protocol can return, and which instance in this paper produced each.

values (P3), and a *power certificate* under declared demand perturbations (P4); a proposition states what the reconstruction gap measures and when (Section 4, Figure 1).

We apply the framework to a bidirectional simulation of the Hangzhou metro driven by 25 days of real smart-card demand, a rolling-horizon integer-frequency planner with strict trip budgets, and residual PPO policies that shift the planner's next request by at most one minute; then to station dwell control on six Beijing and nine Shanghai lines. Three findings organize the paper. First, the best gain transfers but the feedback does not: on four sealed dates the best checkpoint retains a 2.6 s/passenger advantage over the planner but none over its calendar reconstruction. Second, the instrument is not blind: the planner given live queues, a hand-written rule and the Beijing dwell rule move under the declared perturbations; the Hangzhou learners do not. Third, the three ways a learner could escape that verdict (too little authority, a calendar-only training distribution, a blind instrument) were each pre-registered and tested; none makes a policy better than the planner or rule it modifies, and on Shanghai the certificate does not fire for the live-state rule itself, so the protocol declines to rule rather than report a null it cannot back.

Our contributions are: (i) a diagnostic protocol with exact replay checks and a stated condition of validity that attributes a learned controller's gain to calendar, planner feedback and residual feedback (Section 4); (ii) its complete application, with pre-registered expectations, to 18 residual PPO checkpoints, 30 information-set models, a declared perturbation matrix and two further instances, failures included (Section 5, Appendix C); and (iii) the finding that, across three action classes (minute shifts, headway multipliers and dwell extensions) and both instances where the protocol can rule, learned state-dependence never makes a policy better than the planner or rule it modifies, while the gains that exist are calendar-explained (Sections 5.3–5.5).

2 Related work

Learned metro control is usually evaluated against a timetable or a rule. Ying et al. (2020) train an actor-critic dispatcher under stochastic demand and limited rolling stock and Ying et al. (2022) extend it to coordinated multi-line operation; both compare with a fixed timetable, so a gain that a

demand calendar could reproduce is indistinguishable from one that needs the day's observations. Headway-based holding (Daganzo, 2009) and its multi-agent RL counterpart (Wang & Sun, 2020) hold vehicles at stops on live headways, the action class of our dwell controllers. Zhao et al. (2026) add action masking and shaping to PPO for joint timetable and circulation decisions, and Su et al. (2026) learn subproblem values inside stochastic integer planning; both keep an optimization layer, as we do, but neither asks how much of the learned layer's contribution survives when it is denied live inputs. On the non-learning side, square-root frequency rules (Newell, 1971), stochastic and robust frequency setting (Guo et al., 2022) and joint pattern/headway/fleet optimization (Ng et al., 2026) all plan from historical demand; they are the reason a learned policy must be compared with a forecast-informed planner, such as our base planner, rather than a static timetable. Our residual construction follows Silver et al. (2018) and our PPO implementation Schulman et al. (2017); the boarding-constrained, itinerary-based simulation design of Singh et al. (2023) informs our scenario parameters, which remain declared assumptions; potential-based shaping (Ng et al., 1999) is the form of the Hangzhou reference reward and of the Beijing and Shanghai headway shaping. Calendar reconstruction is policy distillation (Rusu et al., 2016) with the student's input restricted to a calendar key, and shares with policy extraction (Bastani et al., 2018; Verma et al., 2018) and finite-state abstraction of recurrent policies (Koul et al., 2019) the aim of reproducing a network with a simpler object; those methods want a faithful surrogate, whereas we ask how much is lost when the surrogate is denied live inputs, verify the answer by exact replay, and certify the comparison's power. That static or open-loop policies can match learned ones is established on control benchmarks (Mania et al., 2018; Seyde et al., 2021; Raffin et al., 2024); we ask the same question of a deployed operational stack in which the calendar is the natural open-loop policy. Our reporting follows Henderson et al. (2018) and Agarwal et al. (2021): every checkpoint and seed, pre-registered expectations, paired uncertainties. Whether a controller's benefit is *because* it responds to information revealed during operation is, to our knowledge, not addressed by a controlled protocol in this literature.

3 SETTING

Demand and network. We use 25 days of card tap-in and tap-out records (January 1–25, 2019) from the Hangzhou metro (Alibaba Tianchi, 2019), paired into 29.1 million origin–destination journeys with minute resolution. An inferred network of 81 station IDs (80 physical stations) with four line patterns in both directions reproduces the supplied adjacency; transfers are explicit multi-leg itineraries with walking time. January 1–18 are training dates (the learners use January 2–18), 19–21 development dates, and 22–25 a sealed test set (Section 5.4; Appendix D.1).

Simulator. A bidirectional event engine moves physical vehicles between endpoint pools with turnaround, admits arrivals during dwell, and integrates waiting, riding and walking time over the full event clock; two movement-rate scenarios (1440 and 2160 passengers per minute) bracket boarding throughput, trip budgets per directed route are hard caps, and the last twelve requests of each route are protected. The objective is mean complete passenger time per external journey, written s/passenger (Appendix D.2).

Planner. A rolling-horizon integer-frequency planner assigns the remaining budgeted requests to 15-minute bins from a calendar forecast, committed next requests and nominal vehicle circulation, replanning every 60 minutes from the fleet state (Appendix D.3). In the deployed configuration its forecast is the historical same-weekday profile of the training dates, and it reads no live queues or entries; a variant that reads live queues is introduced in Section 5.3.

Residual learner. A shared policy chooses, for each unprotected next request, a residual of 0, -1 or $+1$ minute on top of the planner's request (Appendix E.1); its 199 inputs include a 32-value calendar demand prior and the 81 live station queues, and rewards are negative increments of passenger time. Six models (raw reward, or a reference reward that measures savings against the planner; three seeds each) train for 340 episodes over the 34 training cases with checkpoints fixed in advance at episodes 34, 102 and 340 (Appendix E.2); all 18 are evaluated greedily on the six development cases.

4 THE DIAGNOSTIC FRAMEWORK

The framework takes a simulator $\mathcal{S}$ with a scalar cost C (passenger time in Hangzhou, $-J$ in Beijing and Shanghai), a deployed controller π that decomposes into a planner layer π_0 (which may itself use feedback) and a learned residual π_θ acting at discrete decision callbacks with executed decisions $a_t = \pi(s_t)$, and evaluation days $\mathcal{D}$ with fixed or seeded demand. The calendar key k_t holds only exogenous calendar information (here: day type or rate, route or line, and decision ordinal); all policies are composed with the same feasibility mask, which belongs to the environment. A reconstruction $\hat{\pi}$ is a member of the class $\mathcal{H}$ of masked functions of k_t, fitted to the layer's own decisions, and $\pi^\star_{\text{cal}}$ is a cost-minimizing masked function of k_t. The framework asks, for each layer, how much of its effect on C survives when it is replaced by such a function, and certifies its own ability to detect a positive answer. Every operation keeps the other layers, masks, fleet, budgets, protected tails and objective fixed; only the source of one layer's decision changes, so that differences in C are attributable to that source.

- P1 *Calendar reconstruction: would a table have done the same?* Fit $\hat{\pi}(k_t)$ by majority vote (median, for request times) over the layer's executed decisions on training days and evaluate $\hat{\pi}$ and π on $\mathcal{D}$ under identical seeds; $\Delta_{\text{fb}} = C(\hat{\pi}) - C(\pi)$ bounds from above the value of everything the layer reads beyond the calendar. An exact replay of π against its archived trajectory is required.
- P2 *Same-state counterfactuals: what was one decision worth?* At declared gates, branch each available action from an identical simulator state and continue with π_0; the cost differences $v(s, a)$ depend on no learner, and the zero branch must reproduce the behavior trajectory exactly.
- P3 *Information sets: which information predicts that worth?* Fit predictors of $v(s, \cdot)$ from nested sets (calendar prior $\subset$ current state $\subset$ window $\subset$ full history) on earlier dates; held-out regret orders the sets by the decision-relevant information a predictor can exploit.
- P4 *Power certificate: can the instrument see feedback at all?* Repeat P1 on $\mathcal{D}$ after declared demand perturbations that no calendar can encode. If Δ_{fb} moves for no layer that reads live demand, P1's null is uninformative; if it moves for some layers and not others, the null for the others is informative. No numerical firing threshold was fixed in advance; Section 5 reports each reading's size and its ratio to the paired standard error. Two requirements follow from the applications below: a layer that replans from the fleet state (here, the planner layers) must be reconstructed by a same-day frozen replay, not a cross-day table (Section 5.3), and the perturbation's coverage must scale with the network (Appendix C).

Each application pre-registers its expectations with stated confidences and reports every case. The condition under which P1 measures what it claims is elementary but worth stating:

Proposition 1 (What Δ_{fb} measures)**.** *Under paired seeds, $\Delta_{\text{fb}} = C(\hat{\pi}) - C(\pi) = \Delta_{\text{live}} + \Delta_{\mathcal{H}}$ with $\Delta_{\text{live}} = C(\pi^\star_{\text{cal}}) - C(\pi)$ and $\Delta_{\mathcal{H}} = C(\hat{\pi}) - C(\pi^\star_{\text{cal}}) \geq 0$. Hence (i) Δ_{fb} is an upper bound on the value of everything π reads beyond the calendar, tight exactly when $\hat{\pi}$ attains the calendar optimum; (ii) refitting $\hat{\pi}$ on the evaluation day bounds the part of $\Delta_{\mathcal{H}}$ due to transfer, and a same-day frozen replay, which reproduces π exactly on the unperturbed day, carries no transfer error, so its gap under a perturbation isolates the layer's response to it, for any layer including one that replans from the fleet state; (iii) if some layer in the same simulator has Δ_{fb} that moves under a declared perturbation, the perturbation is not inert, so a null Δ_{fb} for π is not attributable to the perturbation having no effect; whether it is attributable to the reconstruction class is settled by (ii), since $\Delta_{\mathcal{H}}$ is layer-specific.*

The decomposition is definitional; (ii) and (iii) close the two distinct ways a null can be uninformative.

Applying it elsewhere. The framework needs a simulator with paired seeds, layers that can be replaced one at a time, archived trajectories for exact replay and a declared perturbation family; it returns Δ_{fb} per layer with paired uncertainty, single-decision values with the information that predicts them, and a certificate. Its reading is void when the exact replay fails or no live-demand layer moves, for instance because the perturbation's coverage is too small; both are tested below.

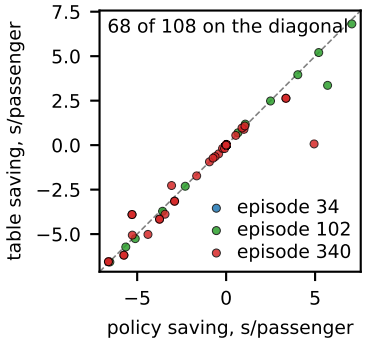


Figure 2: Calendar reconstruction of all 18 checkpoints on the six development cases (108 paired cases; positive = seconds saved per passenger over the planner). Each point pairs a policy's saving with its calendar table's saving on the same case; points on the diagonal are cases the table reproduces exactly. Color gives the checkpoint's training episode; every episode-34 checkpoint is a zero rule at the origin. The full per-checkpoint panels are Figure 7.

Instantiation in Hangzhou. The calendar key is (movement rate, weekday, directed route, departure ordinal); P3 compares calendar priors, the current state, a 30-minute window, a GRU and a selective state-space (Mamba) block; P4 scales journeys from eight seeded stations in a seeded 60-minute window by $3\times$ or $0.25\times$, reconstructing planner layers by same-day frozen replay and the learner by its calendar table (full instantiation and amendments in Appendix A).

5 Results

5.1 Aggregate gains and their calendar reconstructions

The learners' gains are modest and fragile. At the predeclared 340-episode endpoint no model beats the always-delay rule (mean $+2.663$ s/passenger over base) in any development case; one intermediate checkpoint (raw reward, seed 1, episode 102) saves 3.422 s/passenger over base and 0.760 over always-delay, and its benefit is gone by episode 340 (Appendix E.2); every checkpoint is retained.

The calendar reproduces the learners' gains (P1). Reconstructing all 18 checkpoints (Appendix E.2) gives these counts: the policy is better than its table in 25 paired development cases, worse in 15, and identical in passenger end state in 68. For the episode-102 checkpoint the table saves 3.389 s/passenger over base against the policy's 3.422; live features can account for at most 0.033 s/passenger (three wins, two losses, one tie), and executed actions agree in 99.49% of route/ordinal comparisons. Thirteen of the eighteen checkpoints reproduce a constant rule on every case: eleven zero-residual rules tie their tables exactly, and two always-advance rules differ from theirs only where the mask forbids the advance (Figure 2). Stated as a set rather than a fit: no checkpoint both beats the planner and has a table gap distinguishable from zero—the two above the planner have gaps of 0.033 and 0.027, and every checkpoint with a gap above 0.06 is worse than the planner (Appendix C). What the best learner found is therefore a calendar adjustment, and the rest of the paper asks whether anything else can be found.

5.2 Single-decision values and the information they need

Richer learned history does not improve held-out prediction of decision values here; the calendar does as well as any set (P2 and P3). Across 170 anchors on 34 training trajectories, the base-continuation optimal action is zero in 55, advance in 18 and delay in 97; a single gate's oracle benefit over zero averages 0.042 s/passenger (median 0.010, maximum 0.389). On the temporally partitioned study (background January 2–8 for priors, fitting 9–15, validation 16–18; 100 anchors; 30 models), calendar priors alone and a weekday/rate/gate table from the fitting set reach a validation regret of 0.0030–0.0031 s/passenger, against 0.0037–0.0081 for the current state, a 30-minute window, a GRU and a Mamba block with linear or spline heads, and Mamba is worse than the win-

Comparison (Δ = first − second, s/passenger)	none	surge	deficit
Base planner: frozen replay − live	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)
Residual PPO (raw s1 e102): table − policy	$+0.03 \pm 0.05$ (3/2/1)	-0.00 ± 0.03 (8/8/2)	$+0.02 \pm 0.02$ (9/7/2)
Base − surprise rule	$+0.00 \pm 0.00$ (0/0/6)	-1.32 ± 0.78 (2/16/0)	-2.50 ± 0.94 (0/8/10)
Live-queue planner (60 min): frozen − live	$+0.00 \pm 0.00$ (0/0/6)	$+1.75 \pm 1.66$ (3/1/14)	-0.07 ± 0.09 (1/1/16)
Base − live-queue planner (60 min)	$+1.00 \pm 1.38$ (2/4/0)	$+4.93 \pm 3.55$ (8/10/0)	$+0.49 \pm 0.69$ (5/13/0)
Live-queue planner (15 min): frozen − live	$+0.00 \pm 0.00$ (0/0/6)	$+2.27 \pm 1.82$ (4/1/13)	$+0.17 \pm 0.12$ (3/0/15)
Base − live-queue planner (15 min)	-28.69 ± 1.74 (0/6/0)	-24.35 ± 3.49 (1/17/0)	-28.29 ± 0.99 (0/18/0)
Base − residual PPO policy	$+3.42 \pm 1.01$ (6/0/0)	$+3.45 \pm 1.17$ (17/1/0)	$+3.15 \pm 0.50$ (18/0/0)
Base − residual PPO table	$+3.39 \pm 0.97$ (6/0/0)	$+3.46 \pm 1.14$ (17/1/0)	$+3.13 \pm 0.48$ (18/0/0)

Table 1: Perturbation power study on development dates 19–21, rates 1440/2160. Entries are mean $\pm$ paired SE over cases (none: 6 cases; surge and deficit: 18 cases each, three seeded windows and station sets), with counts of cases in which the second controller is better/worse/tied. Positive means the second controller is better. Cases share dates and rate scenarios and are not independent replicates.

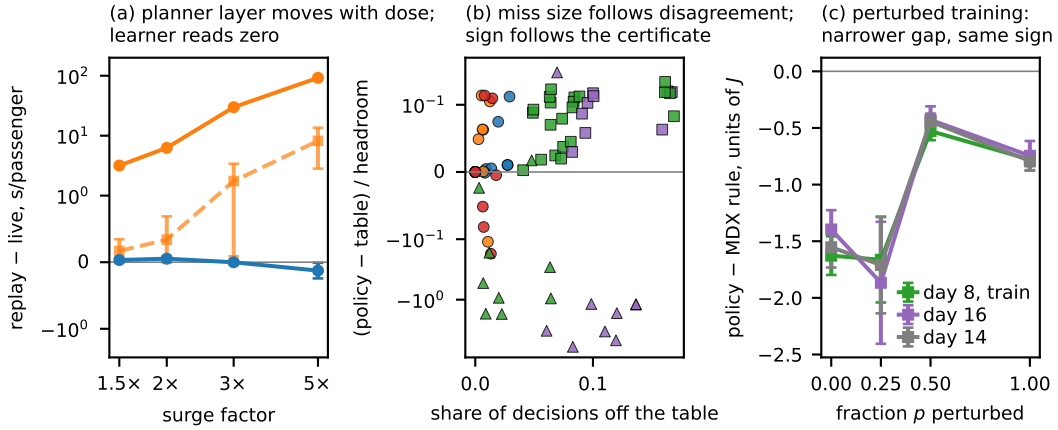


Figure 3: Instrument readings; positive always means the live or learned controller is better. (a) Hangzhou dose: the live-queue planner's replay gap on the Monday morning-peak surge at rate 1440 (orange, solid) and pooled over the 18 surge cases (orange, dashed) grows with the surge factor; the residual learner's table gap (blue) stays within 0.13 of zero (mean $\pm$ paired SE). (b) All 84 learner points of the paper (exploratory): the size of a calendar table's miss, as a fraction of instance headroom (symmetric-log axis), tracks how often the policy departs from its table; its sign tracks whether the power certificate fires (Beijing, Hangzhou) or not (Shanghai). Markers give the instance (circle Hangzhou, square Beijing, triangle Shanghai); colors the learner class (blue PPO ± 1, orange PPO $\{0, \pm 1, \pm 4\}$, red PPO multiplier, green and purple the shaped and unshaped value-decomposition learners of Section 5.5). (c) Beijing: policy minus the marginal dwell-extension rule (MDX) against the fraction of training episodes run under perturbation (mean $\pm$ paired SE over learners and demand seeds).

dow in every seed (Figure 4). The single-decision labels thus agree with the reconstruction; whether these nulls are informative is what P4 decides.

5.3 Power of the diagnostic under declared perturbations

The instrument reads non-zero for the planner and rule layers that read live demand, and zero for the ± 1 learner, which reads it too. *What the instrument sees.* Three structural facts emerge (Table 1; Figure 5 in Appendix A). First, the base planner's decisions are identical to its frozen replay under every perturbation: its replans read fleet state and calendar forecasts, never live demand, so all demand feedback in the deployed stack enters through the residual learner's observations. Second, a hand-written residual rule (the surprise rule of Table 1) that advances the next request when the live route queue exceeds the planner's own unperturbed queue by 300 passengers is worse than the base planner under both perturbations (Appendix A gives the reason). Third, giving the planner live

queues as its replanning backlog produces value that is large but concentrated: in 14 of 18 surge and 16 of 18 deficit cases the live run and its replay coincide exactly, and the mean gain over the replay ($+1.75 \pm 1.66$ s/passenger pooled under surge, about one standard error) is carried by the Monday morning-peak surge, where the live-queue planner beats its replay by 30 s/passenger and the base planner by 63 at the lower rate; 15-minute replanning raises that case to 32 but costs 26–37 s/passenger on every unperturbed day. A pre-registered positive control confirms the reading and exercises Proposition 1(ii) (Appendix A, Table 5): a *cross-day* calendar table is 7 s/passenger worse than the base planner run live, although that planner reads no live demand (date-specific fleet state; Appendix A), and 10 worse for the live-queue planner; their difference, 2.9 ± 1.3 (2.2 standard errors), is the live-demand component, contributed mostly by January 19, where it is present without perturbation, and by the Monday surge.

What it sees for the learner. Over the 36 perturbed cases the learner's calendar table is within $+0.010 \pm 0.020$ s/passenger of the live policy (17 policy better, 15 table better, 4 identical), while both save 3.3 s/passenger over the base planner on average and beat it in 35 of 36 cases: the learned layer is the only layer in the stack that a cross-day calendar table reproduces, and its gain survives the perturbation as a calendar adjustment would. The pre-registered prediction of a feedback gain for this learner failed (H55, Appendix A); for its perturbation-trained retraining it passed only by the letter, mostly on near-constant always-advance policies (D1, Appendix C).

Three escape routes, each tested. A natural reading is that the learner sees the surge but lacks the authority to act on it; we tested it by retraining the same learner with wider authority—a superset action set $\{0, \pm 1, \pm 4\}$ minutes and a multiplier set $\{0.5, 0.75, 1, 1.25, 1.5\}$ on the planner's request interval—under the identical protocol (Table 4). Wider authority makes the decisions state-dependent (in five of six policies the table falls 0.4–0.9 s/passenger short of the policy, a range one ± 1 checkpoint also reaches; in the sixth it is 1.1 better), but all six policies are worse than the planner in mean (by 3.6–5.7 s/passenger for the ± 4 set and 0.8–24.5 for the multiplier set, two of whose seeds exhaust the trip budget with constant early dispatch), and their table gaps do not grow under perturbation beyond their standard errors. *Dose:* adding $1.5\times$, $2\times$ and $5\times$ to the declared $3\times$ moves the live-queue planner's replay gap on the Monday morning peak at rate 1440 monotonically (3.2, 6.3, 29.9 and 92.4 s/passenger at 1.5, 2, 3 and $5\times$) while the learner's table stays within 0.13 ± 0.12 of it at every dose (Figure 3a; Appendix C). *Training distribution:* retraining the same ± 1 learner with half of its episodes under the perturbation family (seeds disjoint from evaluation; pre-registered) yields table gaps of at most 0.45 s/passenger against the 6.15 s/passenger spread between the constant rules, and every checkpoint with a non-zero table gap is worse than the planner (-0.8 to -3.5). Neither authority nor exposure moves the learner into a useful state-dependent regime.

5.4 Sealed test dates

The verdict survives dates that no experiment had opened. January 22–25 were held out from every forecast, prior, parameter, checkpoint and table above, refused by the code, and opened once after all learners and tables were frozen, with the opening time and commit hash logged; every learner then ran with its calendar table on the eight (date, rate) cases (Table 2). The checkpoint that saved 3.42 s/passenger on development dates saves 2.63 ± 0.87 on unseen ones, and its calendar table is 0.14 ± 0.09 s/passenger better than it, within two standard errors of zero; the two other checkpoints above the planner ($+1.34 \pm 0.38$ and $+0.29 \pm 0.82$) have tables within 0.10. The eleven zero-residual checkpoints tie their tables exactly (the two always-advance ones differ from theirs only where the mask forbids the advance); the wider-authority learners are 5–32 s/passenger worse than the planner on unseen days with table gaps of at most 0.7. Pooled over all 24 checkpoints, the calendar table is within $+0.05 \pm 0.05$ s/passenger of the live policy. What transferred to sealed data is calendar-reproducible; what was state-dependent did not help.

5.5 Two dwell-control instances

The same instrument rules on Beijing and declines on Shanghai. To test whether the verdict is a property of the setting, we apply it to an event-driven metro model with station dwell control on six Beijing lines (2.2 million passenger legs on May 8, the training day; Liu, 2022) and nine Shanghai lines (4.7 million journeys per day, May 2017; Sun et al., 2025), with a value-decomposition network

Learner	checkpoint	base − policy	table − policy
multiplier s0	e340	-30.19 ± 1.96 (0/8/0)	$+0.01 \pm 0.63$ (3/5/0)
multiplier s1	e340	-31.91 ± 1.96 (0/8/0)	$+0.07 \pm 0.35$ (4/4/0)
multiplier s2	e340	-5.16 ± 1.46 (2/6/0)	$+0.74 \pm 0.52$ (6/2/0)
raw ± 1 s0	e340	$+1.34 \pm 0.38$ (8/0/0)	-0.10 ± 0.07 (3/4/1)
raw ± 1 s1	e102	$+2.63 \pm 0.87$ (6/2/0)	-0.14 ± 0.09 (2/6/0)
raw ± 1 s1	e340	-5.68 ± 1.20 (0/8/0)	$+0.25 \pm 0.21$ (6/2/0)
raw ± 1 s2	e340	$+0.29 \pm 0.82$ (6/2/0)	-0.04 ± 0.03 (1/6/1)
reference ± 1 s0	e102	-6.52 ± 1.30 (0/8/0)	$+0.37 \pm 0.21$ (7/1/0)
reference ± 1 s0	e340	-6.99 ± 1.33 (0/8/0)	-0.30 ± 0.33 (6/2/0)
reference ± 1 s1	e340	-6.99 ± 1.33 (0/8/0)	-0.30 ± 0.33 (6/2/0)
$\{0, \pm 1, \pm 4\}$ s0	e340	-17.70 ± 3.20 (0/8/0)	$+0.31 \pm 0.18$ (6/2/0)
$\{0, \pm 1, \pm 4\}$ s1	e340	-16.75 ± 3.13 (0/8/0)	$+0.30 \pm 0.24$ (4/3/1)
$\{0, \pm 1, \pm 4\}$ s2	e340	-12.97 ± 3.76 (0/8/0)	$+0.11 \pm 0.12$ (3/4/1)
11 checkpoints identical to the planner		$+0.00 \pm 0.00$ (0/0/8 each)	$+0.00 \pm 0.00$ (0/0/8 each)
all 24 checkpoints pooled		-5.69 ± 0.74 (22/82/88)	$+0.05 \pm 0.05$ (57/43/92)

Table 2: Sealed test dates (January 22–25, 2019; two movement rates; eight paired cases per row), opened once after every learner and table was frozen; positive base − policy means the learner saves passenger time over the planner, positive table − policy means the policy beats its own calendar table. Mean ± paired SE with better/worse/tied counts; checkpoints are not independent. The 11 constant-zero checkpoints (the all-zero rows of Table eftab:power-ckpt) coincide with the planner and with their tables on every case and share one row.

Day	Comparison	none	surge	deficit
May 8	MDX rule − frozen replay	$+0.00 \pm 0.00$ (3/0/0)	$+0.25 \pm 0.05$ (8/1/0)	$+0.11 \pm 0.03$ (9/0/0)
May 8	shaped (May 8): policy − table	$+0.26 \pm 0.05$ (13/2/0)	$+0.42 \pm 0.04$ (43/2/0)	$+0.28 \pm 0.03$ (40/5/0)
May 8	shaped (May 8): policy − MDX	-1.63 ± 0.17 (0/15/0)	-1.79 ± 0.10 (0/45/0)	-1.66 ± 0.09 (0/45/0)
May 8	unshaped (May 8): policy − table	$+0.74 \pm 0.11$ (9/0/0)	$+0.95 \pm 0.09$ (27/0/0)	$+0.65 \pm 0.05$ (27/0/0)
May 8	unshaped (May 8): policy − MDX	-2.37 ± 0.32 (0/9/0)	-2.54 ± 0.18 (0/27/0)	-2.51 ± 0.19 (0/27/0)
May 8	shaped (May 14–15): policy − table	$+1.57 \pm 0.15$ (9/0/0)	$+1.75 \pm 0.10$ (27/0/0)	$+1.64 \pm 0.08$ (27/0/0)
May 8	shaped (May 14–15): policy − MDX	-2.91 ± 0.50 (0/9/0)	-3.15 ± 0.29 (0/27/0)	-2.89 ± 0.27 (0/27/0)
May 16	MDX rule − frozen replay	$+0.00 \pm 0.00$ (2/1/0)	$+0.18 \pm 0.03$ (9/0/0)	$+0.03 \pm 0.02$ (6/3/0)
May 16	shaped (May 8): policy − table	$+0.98 \pm 0.08$ (15/0/0)	$+1.20 \pm 0.06$ (45/0/0)	$+1.01 \pm 0.05$ (45/0/0)
May 16	shaped (May 8): policy − MDX	-1.40 ± 0.17 (0/15/0)	-1.48 ± 0.08 (0/45/0)	-1.34 ± 0.09 (0/45/0)
May 16	unshaped (May 8): policy − table	$+1.55 \pm 0.21$ (9/0/0)	$+1.84 \pm 0.14$ (27/0/0)	$+1.54 \pm 0.12$ (27/0/0)
May 16	unshaped (May 8): policy − MDX	-2.42 ± 0.37 (0/9/0)	-2.50 ± 0.20 (0/27/0)	-2.41 ± 0.21 (0/27/0)
May 16	shaped (May 14–15): policy − table	$+2.37 \pm 0.16$ (9/0/0)	$+2.70 \pm 0.08$ (27/0/0)	$+2.43 \pm 0.08$ (27/0/0)
May 16	shaped (May 14–15): policy − MDX	-2.58 ± 0.47 (0/9/0)	-2.66 ± 0.28 (0/27/0)	-2.47 ± 0.26 (0/27/0)
May 14	MDX rule − frozen replay	$+0.01 \pm 0.00$ (2/1/0)	$+0.12 \pm 0.06$ (6/3/0)	$+0.09 \pm 0.03$ (8/1/0)
May 14	shaped (May 8): policy − table	$+1.41 \pm 0.07$ (15/0/0)	$+1.64 \pm 0.05$ (45/0/0)	$+1.41 \pm 0.04$ (45/0/0)
May 14	shaped (May 8): policy − MDX	-1.55 ± 0.18 (0/15/0)	-1.56 ± 0.09 (0/45/0)	-1.58 ± 0.10 (0/45/0)
May 14	unshaped (May 8): policy − table	$+1.71 \pm 0.11$ (9/0/0)	$+2.00 \pm 0.08$ (27/0/0)	$+1.72 \pm 0.06$ (27/0/0)
May 14	unshaped (May 8): policy − MDX	-2.50 ± 0.37 (0/9/0)	-2.54 ± 0.20 (0/27/0)	-2.50 ± 0.21 (0/27/0)

Table 3: Beijing six-line metro instance; each entry is the difference named in its row, in units of J (Appendix B), positive when the first-named controller is better (worst constant policy -23.2, MDX -10.84 on May 8). Rows: the MDX rule against its own frozen replay, and each MDX-anchored learner against its calendar table and against MDX. Shaped learners use potential-based headway shaping, unshaped ones do not; the date in parentheses is the training day. Mean ± paired SE over learners, demand seeds and perturbation seeds with better/worse/tied counts; cases share days and seeds.

(VDN) learner that chooses at every station arrival whether to extend the dwell by zero, one or two minutes, anchored to an analytic marginal dwell-extension rule (MDX) that reads the same live queues and loads (Appendix B). Here a dwell decision acts directly on the queue it observes; the calendar key is (line, trip ordinal, stop).

The verdict of P1 inverts on every learner row of Table 3 (11 learners, three evaluation days, 1386 episodes). The rule's frozen replay is worse than the rule under both perturbations on every day ($+0.12$ to $+0.25$ under surge; 2.0 to 5.1 standard errors in every day and regime), so the setting carries feedback value that P4 detects. The learner's calendar table does *not* reproduce the learner: without perturbation the policy beats its own table by 0.3–0.7 units of the objective J (Appendix B)

on its training day and by 1.0–2.4 against the *transferred* table on held-out days; it beats it in 621 of the 630 paired cases over all regimes, and the gap widens by 0.2–0.3 under surge. A table refitted on the evaluation day itself (Proposition 1(ii); Appendix B, Table 7) shows that, without perturbation, 0.6–1.8 of the held-out gap is transfer error and the state-dependent component proper is 0.26–0.93 on every learner and day (up to 1.19 under surge), positive in 457 of 462 paired cases. Yet the learner is worse than the rule it is anchored to in every one of those cases, by 1.3–3.2 on average, and its guardrail sweep rises monotonically from the pure policy to the rule with no interior point above it (Appendix B): the state-dependent component is real, and the learner that carries it is worse than the rule. The archived learners never saw non-calendar demand in training, and part of their deficit is attributable to that. Retrained with a fraction p of episodes under the perturbation family (Table 10, pre-registered), the learner stands at -1.63, -1.66, -0.53 and -0.78 against the rule at $p = 0, 0.25, 0.5, 1$ on the training day (Figure 3c), threshold-like rather than monotone. At $p = 0.5$ it wins 6 of 63 held-out cases on May 16, where the archived learner won none, and stays below the rule in 183 of its 189 paired cases; its in-sample table gap is larger than the archived learner's at $p = 0.25$ and smallest at $p = 1$. Exposure to the declared perturbations recovers about two thirds of the gap (68–72%) without reversing its sign: it moves the learner toward the rule without making its use of state beat the rule.

Shanghai (Table 8) is the case the power certificate exists for. The same perturbation family that moves the Beijing rule leaves the Shanghai rule's frozen replay within 0.03 of the rule in every one of the 63 paired cases and within 0.01 on average on every day and regime (pooled under deficit 0.005 ± 0.001: statistically detectable, but under 1% of the instance's headroom): a controller that reads live queues and has dwell authority is indistinguishable from its own frozen replay. By the criterion of Section 4, P4 has failed and the protocol declines to draw a P1 conclusion about the learners from this instance. The reason is visible in the instance itself: no passenger is stranded, and the span from always-GO to the rule is 0.70 units of J, against 12.4 from the worst constant policy in Beijing, so a dwell extension has little queue to clear. The learner rows (every learner below the rule, by 1.6–11.2; Appendix B) are descriptive, not a verdict. A pre-registered widening of the perturbation's coverage makes the certificate only marginal (a replay gap of 0.004–0.005 units under surge on two of three days, at most 0.010 in any regime), and the learners re-evaluated at that coverage show the same descriptive pattern (Appendix C).

At the level of single decisions P2 and P3 agree: at 90 declared Beijing gates the rule's action is already best at 46, the mean available gain is 0.007 units (maximum 0.066), and no information set predicts the value vector with a regret below the rule's own 0.007 (calendar 0.008, current state 0.011, window 0.009); on Shanghai the rule is best at 63 of 90 with 0.0004 available. Across all 84 learner points of the paper, in an exploratory analysis, how much a calendar table misses tracks how often the policy leaves it, and the sign of the miss tracks whether the certificate fires (Figure 3b; Appendix C). In Hangzhou the ± 1 residual could not act on feedback; in Beijing the learner acts on it worse than a rule reading the same state; in Shanghai the certificate does not fire and the protocol says so.

6 Limitations and conclusion

Where the framework can rule, learned use of online feedback never makes a controller better than the planner or rule it modifies: the gains that exist are calendar adjustments, they transfer to sealed dates as such, and where live state carries value the learner uses it worse than a rule reading the same state. The two requirements of Section 4 were discovered by a failed pre-registered expectation (C3) and a selection rule that found no qualifying level (Appendix C); we expect them to carry to learned layers on adaptive planners in other settings with recurring context. For the evaluation of learned operational controllers the implication is procedural: a gain over a timetable or a rule is not evidence of feedback until the controller beats its own calendar reconstruction under a certified instrument, and a controller that claims to exploit online information should pass this protocol and beat the planner or rule it modifies; the Beijing learners pass the first test and fail the second, and no other learner passes either. The limits are those of the evidence, not the logic: the simulators are declared scenario models, three development or evaluation days per city and eight sealed cases are small sets, thirteen of eighteen residual policies collapse to constant rules, the positive controls are planner and rule layers rather than learners, all instances are metro control, the perturbations are synthetic, and single-decision values do not bound a repeatedly intervening controller.

ETHICS STATEMENT

The paper uses three metro smart-card datasets: the public Tianchi release of Hangzhou records (Alibaba Tianchi, 2019), the Beijing records distributed by ChinaGEOSS on application (Liu, 2022), and the aggregated MetroFlow counts for Shanghai (Sun et al., 2025). Card identifiers in the Hangzhou and Beijing records are pseudonymous as released; our pipelines keep only station-level and origin–destination aggregates, and no card-level record is released or reported. The work proposes an evaluation protocol for simulated controllers; it deploys nothing and involves no interaction with passengers.

REPRODUCIBILITY STATEMENT

Every number in this paper is produced by a script from a results file that records each episode, and every study reported as a test was pre-registered in a dated plan file with stated confidences before its first run (the placement diagnostic of Appendix D.3 and the cross-instance relation of Appendix C are exploratory); the code, the plan file, the protocol amendments, the sealed-set opening record (time and commit hash) and the complete results files will be released. Hangzhou demand is derived from the public 2019 Tianchi release of Hangzhou metro card records (Alibaba Tianchi, 2019); Beijing from the May 2019 bus and metro smart-card records (Liu, 2022), available from ChinaGEOSS on application and not redistributed (Appendix B); Shanghai from the MetroFlow release (Sun et al., 2025) (CC BY 4.0). Operating parameters are declared scenario assumptions and are listed where they are used. A release manifest maps each promise in this statement to a file of that release. Sections 4 and 5 and Appendices A–E give the protocol, the declared scenarios and every hyperparameter.

AI USE STATEMENT

In our work, we mainly use GPT-5 for writing enhancements, primarily to improve grammar and text clarity.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Alibaba Tianchi. Global city AI challenge: Hangzhou metro passenger flow prediction. Competition dataset of Hangzhou metro card-swipe records, January 1–25, 2019. URL `https://tianchi.aliyun.com/competition/entrance/231708/introduction`.

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Carlos F. Daganzo. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B: Methodological*, 43(10):913–921, 2009. doi: 10.1016/j.trb.2009.04.002.

Xiaotong Guo, Baichuan Mo, Haris N. Koutsopoulos, Shenhao Wang, and Jinhua Zhao. Transit frequency setting problem with demand uncertainty. *arXiv preprint arXiv:2204.12666*, 2022. URL `https://arxiv.org/abs/2204.12666`.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Anurag Koul, Sam Greydanus, and Alan Fern. Learning finite state representations of recurrent policy networks. In *International Conference on Learning Representations*, 2019.

Yu Liu. Beijing bus and metro smart-card records, May 2019. Dataset (title in Chinese), Peking University; distributed by the National Earth Observation Data Sharing Platform (ChinaGEOSS)

on application, 2022. Records purchased from Alibaba Amap under National Key R&D Program of China project 2017YFB0503600.

Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search of static linear policies is competitive for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

G. F. Newell. Dispatching policies for a transportation route. *Transportation Science*, 5(1):91–105, 1971. doi: 10.1287/trsc.5.1.91.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287, 1999. URL https://ai.stanford.edu/~ang/papers/shaping-icml99.pdf.

Max T. M. Ng, Draco Tong, Hani S. Mahmassani, Omer Verbas, and Taner Cokyasar. Joint optimization of pattern, headway, and fleet size of multiple urban transit lines with perceived headway consideration and passenger flow allocation. *arXiv preprint arXiv:2409.19068v3*, 2026. URL https://arxiv.org/abs/2409.19068v3.

Antonin Raffin, Olivier Sigaud, Jens Kober, Alin Albu-Schaeffer, João Silvério, and Freek Stulp. An open-loop baseline for reinforcement learning locomotion tasks. *Reinforcement Learning Journal*, 1:92–107, 2024.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Tim Seyde, Igor Gilitschenski, Wilko Schwarting, Bartolomeo Stellato, Martin Riedmiller, Markus Wulfmeier, and Daniela Rus. Is bang-bang control all you need? Solving continuous control with Bernoulli policies. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Tom Silver, Kelsey Allen, Joshua Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018. URL https://arxiv.org/abs/1812.06298.

Ramandeep Singh, Daniel Hörcher, and Daniel J. Graham. An evaluation framework for operational interventions on urban mass public transport during a pandemic. *Scientific Reports*, 13:5163, 2023. doi: 10.1038/s41598-023-31892-2.

Boyi Su, Fangsheng Wang, Shuai Su, Andrea D'Ariano, Zhikai Wang, and Tao Tang. Real-time metro train rescheduling under uncertainties: A hybrid machine learning and integer L-shaped approach. *Transportation Research Part E: Logistics and Transportation Review*, 208:104704, 2026. doi: 10.1016/j.tre.2026.104704.

Peiyan Sun, Jinming Yang, Zongyuan Huang, Shaoyu Huang, Shengyuan Xu, Weipeng Wang, Wenxuan Guo, Yuting Feng, Xi Zhai, Tao Yang, Xiaokang Yang, Yaohui Jin, and Yanyan Xu. Human mobility datasets in the complex metro system of Shanghai. *Scientific Data*, 12:1061, 2025. doi: 10.1038/s41597-025-05416-8.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018.

Jiawei Wang and Lijun Sun. Dynamic holding control to avoid bus bunching: A multi-agent deep reinforcement learning framework. *Transportation Research Part C: Emerging Technologies*, 116:102661, 2020. doi: 10.1016/j.trc.2020.102661.

Cheng-shuo Ying, Andy H. F. Chow, and Kwai Sang Chin. An actor-critic deep reinforcement learning approach for metro train scheduling with rolling stock circulation under stochastic demand. *Transportation Research Part B: Methodological*, 140:210–235, 2020. doi: 10.1016/j.trb.2020.08.005.

Cheng-shuo Ying, Andy H. F. Chow, Thi Minh Hoa Nguyen, and Kwai Sang Chin. Multi-agent deep reinforcement learning for adaptive coordinated metro service operations with flexible train composition. *Transportation Research Part B: Methodological*, 161:36–59, 2022. doi: 10.1016/j.trb.2022.05.001.

Xiaoli Zhao, Dewei Li, and Xinyu Bao. A deep reinforcement learning approach for integrated optimization of train scheduling and rolling stock circulation planning. *Computers & Industrial Engineering*, 213:111784, 2026. doi: 10.1016/j.cie.2025.111784.

A PERTURBATION POWER STUDY: PROTOCOL AND COMPLETE RESULTS

The protocol (`audit/FEEDBACK_POWER_PROTOCOL_20260918.md`) and its two dated amendments were frozen before the corresponding perturbed episodes were run; the runner is `transitnet/feedback_power_run.py` and the analysis `transitnet/analyze_feedback_power.py`. All episodes run on the same code and data hashes as the reconstruction study of Section 5; every episode completes all external journeys with zero solver fallbacks.

Instantiation of P1–P3 in Hangzhou. P1: the policy is run greedily on all training cases and its executed residual recorded at every (movement rate, weekday, directed route, departure ordinal) cell; the table is the majority vote across dates (ties prefer zero, advance, delay), receives no live queue, load, clock or forecast value, and is paired with an exact replay of the policy that must reproduce the archived trajectory and every solver call. P2: at declared gates the simulator state is cloned before the residual decision, each of the three actions is applied once, and the base planner continues to the end of the day; the zero branch must reproduce the behavior trajectory to every digit; these values do not add across gates and are not a closed-loop bound. P3: predictors of the counterfactual costs are fitted from (a) calendar priors, (b) the current state, (c) a 30-minute window, (d) a GRU over the full causal prefix and (e) a selective state-space (Mamba) block, each with a linear and a spline head; priors use only dates strictly earlier than any fitting or validation anchor. P4 is instantiated as follows.

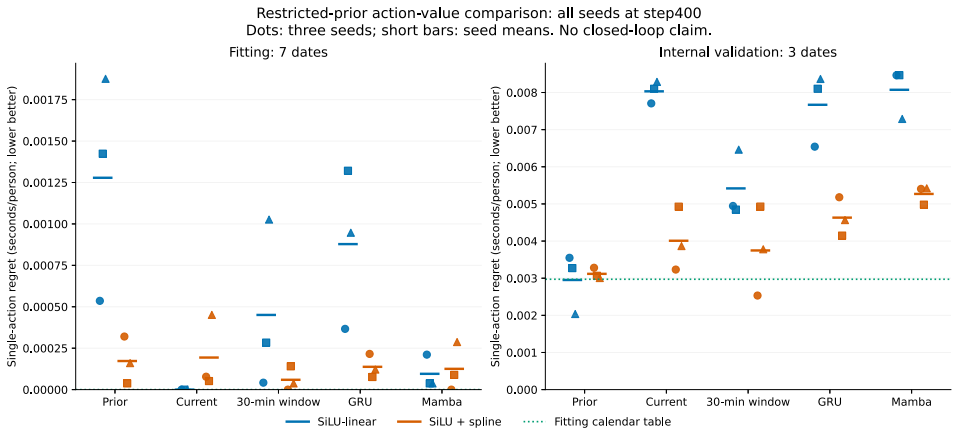


Figure 4: Regret of the five information families and two heads of P3 on the fitting dates (left) and the validation dates (right), all seeds (Section 5); the legend's SiLU-linear and SiLU + spline are the linear and spline heads, and seconds/person is s/passenger. Validation regret in s/passenger with linear heads: priors 0.0030, current state 0.0080, 30-minute window 0.0054, GRU 0.0077, Mamba 0.0081; with spline heads 0.0031, 0.0040, 0.0037, 0.0046 and 0.0053; a weekday/rate/gate table from the fitting set obtains 0.0030. Priors use only dates earlier than any fitting anchor.

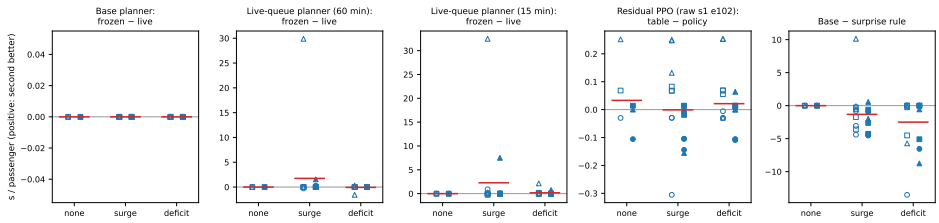


Figure 5: Paired case differences under no perturbation, a $3\times$ local surge and a $0.25\times$ deficit (three seeded windows each); positive means the second-named controller is better. Circles, squares and triangles are January 19, 20 and 21; open and filled markers are movement rates 1440 and 2160; red bars are means. The base planner's frozen replay equals its live run in every case.

Perturbations. On an evaluation date, journeys originating at a seeded set of 8 of the 81 stations inside a seeded 60-minute window (starting 08:00, 12:00 or 17:30 for seeds 0, 1, 2) are scaled by $3\times$ (surge) or $0.25\times$ (deficit), rounded to integers. Priors, planner, budgets, fleet, schedules, masks and objective are unchanged. The surge adds roughly 25,000 journeys to a 1.09-million-journey day. The declared strength, windows and station counts were fixed before any run; the dose factors 1.5, 2 and 5 and the 16- and 32-origin families were added later as pre-registered secondaries (Appendix C).

Controllers. *Base planner*: the deployed rolling-horizon planner with zero residual, replanning every 60 minutes from fleet state and calendar forecast. *Frozen replay*: the same controller's recorded request interval at every (route, ordinal) callback from the unperturbed run of the same date and rate, replayed without replanning; a strict equality check requires the replay to reproduce the live trajectory exactly on the unperturbed day (it does, to every digit, with zero clock deviation). *Live-queue planner*: the base planner given the live station queues as its replanning backlog at each replan, at 60- and 15-minute intervals; each has its own frozen replay. *Residual PPO*: the raw-reward seed-1 episode-102 checkpoint, greedy, and its calendar table from Section 5, both on the base planner run live. *Surprise rule*: a hand-written residual with the learner's sensors that advances the next request by one minute when the live route queue exceeds the base planner's own recorded queue at the same callback by 300 passengers, delays it below -300, and otherwise does nothing; it equals the base planner on the unperturbed day by construction. Under both perturbations it is worse than base, we believe because once the trajectory departs from the reference the recorded queue is no longer a valid counterfactual; under surge a ± 1-minute residual also cannot absorb a threefold increase. An earlier naive rule (queue divided by calendar 15-minute arrivals, thresholds 1.5/0.5) was run once before the amendment and is retained: it is worse than base by 2.7 and 6.8 s/passenger under surge and deficit on January 19, rate 1440.

Pre-registered expectations and verdicts. H54 (confidence 0.8): the base planner's frozen-minus-live gap grows under perturbation. Untestable: the gap is identically zero because the planner reads no live demand; recorded as a structural finding. H54′ (0.7): the surprise rule beats base under perturbation by more than two paired standard errors. Fails: it is worse under both regimes. H55 (0.35): the learner's table-minus-policy gap grows under perturbation. Fails: $+0.010 \pm 0.020$ s/passenger over the 36 perturbed cases (17 policy better, 15 table better, 4 identical) against $+0.033 \pm 0.050$ unperturbed. H57 (0.5): the 15-minute live-queue planner beats its own frozen replay under perturbation by more than two paired standard errors. Fails by the declared criterion ($+2.27 \pm 1.82$ under surge; 4 better, 1 worse, 13 identical), because the mean is carried by one case type: the Monday morning-peak surge at rate 1440 (32.5 s/passenger) and 2160 (7.5). H58 (0.7): the 60-minute live-queue planner does not beat its frozen replay by more than two standard errors. Passes by the declared criterion ($+1.75 \pm 1.66$; 3 better, 1 worse, 14 identical), with the same caveat: the Monday morning surge alone is worth 29.9 s/passenger at rate 1440. Under deficit both planners coincide with their replays in 15–16 of 18 cases. Without perturbation the 60-minute live-queue planner is 4.9/5.5 s/passenger better than base on the Monday and 0.1–2.6 worse on the two weekend days; the 15-minute planner is 26–37 s/passenger worse than base on every date.

Action class	seed	base − policy (dev)	table − policy (dev)	surge	deficit
$\{0, \pm1, \pm4\}$ min	0	-5.72 ± 2.05 (0/6/0)	$+0.39 \pm 0.25$ (5/1/0)	$+0.41 \pm 0.15$ (15/3/0)	$+0.39 \pm 0.14$ (13/5/0)
$\{0, \pm1, \pm4\}$ min	1	-4.96 ± 2.09 (0/6/0)	$+0.39 \pm 0.12$ (5/1/0)	$+0.49 \pm 0.09$ (16/2/0)	$+0.38 \pm 0.06$ (15/3/0)
$\{0, \pm1, \pm4\}$ min	2	-3.56 ± 1.52 (1/5/0)	$+0.88 \pm 0.96$ (3/3/0)	$+1.82 \pm 1.14$ (10/8/0)	$+0.72 \pm 0.45$ (9/9/0)
$\{0.5, \ldots, 1.5\}\times$ interval	0	-23.83 ± 1.71 (0/6/0)	$+0.79 \pm 0.72$ (4/2/0)	$+0.93 \pm 0.51$ (12/6/0)	$+0.69 \pm 0.34$ (12/6/0)
$\{0.5, \ldots, 1.5\}\times$ interval	1	-24.49 ± 3.54 (0/6/0)	-1.07 ± 1.42 (2/4/0)	-1.10 ± 0.75 (6/12/0)	-1.07 ± 0.78 (6/12/0)
$\{0.5, \ldots, 1.5\}\times$ interval	2	-0.81 ± 1.17 (2/4/0)	$+0.88 \pm 0.48$ (4/2/0)	$+0.83 \pm 0.26$ (12/6/0)	$+0.88 \pm 0.26$ (12/6/0)

Table 4: Residual PPO retrained with wider authority under the identical protocol (episode 340, three seeds). Base − policy is the gain over the planner on the six development cases (positive = learner better); table − policy is the calendar reconstruction gap on those cases and under the 18-case surge and deficit matrices (positive = policy better than its table). Mean ± paired SE with better/worse/tied counts.

Wider-authority learners. Table 4 gives the per-seed development gains, calendar gaps and perturbation gaps of the learners retrained with the $\{0, \pm1, \pm4\}$ and multiplier action sets (Section 5.3).

Positive control on the planner layers (pre-registered). To test whether P1 can read a non-zero value in this simulator at all, we applied the calendar reconstruction to the planner layers themselves: for each evaluation day and rate, the median request time of every (route, ordinal) departure over the planner's own unperturbed runs on the two prior same-weekday days is replayed with no replanning and no live state (Table 5; 192 episodes, no table misses). Four expectations were declared before the run. C1 (0.7), the live-queue planner beats its table on the Monday at both rates without perturbation by at least 1 s/passenger: passes ($+1.8$, $+7.2$). C2 (0.8), by at least 10 under the Monday morning-peak surge: passes on the declared seed 0 ($+30.6$, $+13.1$); the three-seed means printed in the table are $+11.6$ and $+9.2$, the latter below the threshold. C3 (0.6), the base planner—which reads no live demand—is within 1 s/passenger of its own cross-day table on all six unperturbed cases: *fails* ($+1.7$ to $+10.2$; pooled over the 36 perturbed cases $+7.12 \pm 0.69$, 34/2/0). C4 (0.7), the live-queue planner's gap exceeds the base planner's by more than two paired standard errors: passes narrowly ($+2.89 \pm 1.29$, 19/17/0). The reading declared for this outcome applies: a cross-day table is not a valid reconstruction of a planner layer, which replans from date-specific fleet state and forecasts that a cross-day median does not reproduce (an interpretation); the same-day frozen replay of the same planner is exact. C1 therefore does not by itself evidence live-demand use (the base planner shows the same gap), and only the difference between the two planner layers is the live-demand component. Over the same 36 perturbed cases the instrument reads $+7$ and $+10$ s/passenger on the two planner layers across days, $+2.9$ between them, and $+0.01$ on the learned residual layer (0.03 unperturbed), whose cross-day table is the only cross-day reconstruction in the stack that reproduces its layer.

All 18 checkpoints. The same matrix was run for every checkpoint (policy and table; 1512 episodes), giving Table 6; pooled counts are descriptive because checkpoints are not independent. Pooled table-minus-policy is $+0.082$ unperturbed (25/15/68 of 108 cases), $+0.052$ under surge (69/53/202 of 324) and $+0.090$ under deficit (79/43/202 of 324): perturbation does not raise the share of cases in which a policy beats its calendar table (23%, 21%, 24%), and the eleven zero-residual checkpoints remain identical to their tables in every case.

Day, rate	Planner − its calendar table	none	surge	deficit
19, 1440	base planner	$+6.29$	$+6.44 \pm 0.24$ (3/0/0)	$+6.28 \pm 0.05$ (3/0/0)
19, 1440	live-queue planner	$+13.62$	$+13.81 \pm 0.19$ (3/0/0)	$+13.43 \pm 0.22$ (3/0/0)
19, 2160	base planner	$+10.15$	$+10.09 \pm 0.04$ (3/0/0)	$+10.17 \pm 0.02$ (3/0/0)
19, 2160	live-queue planner	$+16.53$	$+16.71 \pm 0.23$ (3/0/0)	$+16.57 \pm 0.03$ (3/0/0)
20, 1440	base planner	$+8.88$	$+9.23 \pm 0.32$ (3/0/0)	$+8.84 \pm 0.05$ (3/0/0)
20, 1440	live-queue planner	$+8.52$	$+9.06 \pm 0.41$ (3/0/0)	$+8.52 \pm 0.04$ (3/0/0)
20, 2160	base planner	$+8.10$	$+8.10 \pm 0.05$ (3/0/0)	$+8.11 \pm 0.02$ (3/0/0)
20, 2160	live-queue planner	$+5.97$	$+6.02 \pm 0.06$ (3/0/0)	$+5.96 \pm 0.03$ (3/0/0)
21, 1440	base planner	$+1.71$	-3.14 ± 3.98 (1/2/0)	$+2.51 \pm 0.83$ (3/0/0)
21, 1440	live-queue planner	$+1.78$	$+11.58 \pm 9.52$ (3/0/0)	$+2.03 \pm 0.15$ (3/0/0)
21, 2160	base planner	$+9.54$	$+9.35 \pm 0.42$ (3/0/0)	$+9.43 \pm 0.07$ (3/0/0)
21, 2160	live-queue planner	$+7.21$	$+9.16 \pm 1.97$ (3/0/0)	$+7.30 \pm 0.13$ (3/0/0)

Table 5: Hangzhou P1 positive control: each planner layer against its own cross-day calendar table (median request time per route and departure ordinal over the two prior same-weekday days, replayed with no replanning and no live state). Entries are table cost minus live cost in s/passenger, positive when the planner run live beats its reconstruction; mean ± paired SE over three perturbation seeds with better/worse/tied counts.

Checkpoint	none	surge	deficit	base − policy (perturbed)
raw ±1 s0 e34	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
raw ±1 s0 e102	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
raw ±1 s0 e340	$+0.03 \pm 0.02$ (3/3/0)	$+0.02 \pm 0.02$ (9/9/0)	$+0.02 \pm 0.01$ (9/9/0)	-0.00 ± 0.10 (16/20/0)
raw ±1 s1 e34	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
raw ±1 s1 e102	$+0.03 \pm 0.05$ (3/2/1)	-0.00 ± 0.03 (8/8/2)	$+0.02 \pm 0.02$ (9/7/2)	$+3.30 \pm 0.63$ (35/1/0)
raw ±1 s1 e340	$+0.84 \pm 0.83$ (4/2/0)	$+0.74 \pm 0.43$ (10/8/0)	$+0.80 \pm 0.43$ (12/6/0)	-1.38 ± 0.88 (8/28/0)
raw ±1 s2 e34	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
raw ±1 s2 e102	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
raw ±1 s2 e340	-0.01 ± 0.02 (1/4/1)	$+0.06 \pm 0.05$ (6/10/2)	$+0.01 \pm 0.01$ (7/9/2)	-0.35 ± 0.19 (14/22/0)
reference ±1 s0 e34	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
reference ±1 s0 e102	$+0.46 \pm 0.38$ (6/0/0)	$+0.50 \pm 0.23$ (18/0/0)	$+0.40 \pm 0.18$ (18/0/0)	-2.37 ± 0.97 (7/29/0)
reference ±1 s0 e340	$+0.06 \pm 0.31$ (4/2/0)	-0.19 ± 0.18 (9/9/0)	$+0.19 \pm 0.23$ (12/6/0)	-2.93 ± 0.85 (7/29/0)
reference ±1 s1 e34	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
reference ±1 s1 e102	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
reference ±1 s1 e340	$+0.06 \pm 0.31$ (4/2/0)	-0.19 ± 0.18 (9/9/0)	$+0.19 \pm 0.23$ (12/6/0)	-2.93 ± 0.85 (7/29/0)
reference ±1 s2 e34	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
reference ±1 s2 e102	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)
reference ±1 s2 e340	$+0.00 \pm 0.00$ (0/0/6)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/18)	$+0.00 \pm 0.00$ (0/0/36)

Table 6: All 18 residual PPO checkpoints under the perturbation matrix: table − policy (positive = policy better) per regime, and base − policy pooled over surge and deficit. Checkpoints are named by reward, seed and training episode. Entries are mean ± paired SE with better/worse/tied counts.

B SECOND INSTANCE: BEIJING DWELL CONTROL

Simulator. Our event-driven model of six Beijing metro lines is built from the May 2019 Beijing bus and metro smart-card records (Liu, 2022) (about 2.2 million passenger legs on May 8; May 14–16 serve for training and holding out). Vehicles are dispatched from a finite fleet at nominal headways; at each station arrival a controller chooses GO or an extra dwell of one or two minutes (a SKIP action is masked wherever passengers wait). Boarding is limited by a whole-train boarding rate times the dwell; passengers still waiting at the end of the horizon are stranded. The objective is

$$J = -\Big(\frac{T}{2.19 \times 10^7} + \frac{E}{1.33 \times 10^4} + \frac{U}{1.5 \times 10^4}\Big),$$

where T is waiting plus riding time in passenger-minutes, E an emissions proxy and U the passengers still stranded at the horizon. The three normalizers were fixed once from an always-GO reference run under an earlier accounting of the released code, where the three terms were equal; under the full-horizon accounting used here they are no longer equal, so a unit of J is a fixed scale rather than one reference day of each term; the same normalizers serve Shanghai. Higher is better; the worst constant policy scores -23.2 on May 8. Because J is a value, Beijing and Shanghai dif-

ferences are written policy first (policy − table), whereas Hangzhou costs are written table − policy; both are positive when the policy is better.

Rule and learner. MDX extends the dwell when the passengers it would additionally board, valued at a fixed $V = 100$ minutes each, outweigh the minutes imposed on those aboard and on the queues downstream; it reads live queues and loads and scores -10.84 on May 8. The learner is a parameter-shared value-decomposition network over per-vehicle observations (time, position, local and downstream queues, load, fleet state, headway ratio), trained with the increment of J as reward, potential-based headway shaping, and rule-guided exploration plus a relative guardrail that adds τ times an exponential moving average of $|Q|$ to the rule's action value in execution, evaluation and bootstrapping alike. Eight checkpoints trained on May 8 (five with shaping, three without) and three trained on May 14–15 are evaluated with their final training-time guardrail.

Guardrail sweep. Re-evaluating the saved May-8 models at guardrail strengths $\tau \in \{0, 0.02, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, \infty\}$ gives $J = -16.2, -13.3, -11.7, -11.2, -11.1, -11.0, -10.92, -10.92, -10.92$ for the shaped learner and a monotone path for the unshaped one (means over three evaluation episodes whose demand seeds differ from those of Table 3, hence MDX's -10.92 here against -10.84 there); $\tau \to \infty$ reproduces MDX's decisions exactly and no interior point exceeds it.

Protocol instance (Beijing). Calendar key (line, trip ordinal, stop); tables fit by majority vote over three demand seeds on the training day(s); perturbations scale arrivals at eight seeded (line, stop) origins within a seeded 60-minute window by $3\times$ or $0.25\times$ before Poisson resampling; evaluation on May 8, 16 and 14 with three demand seeds each. The MDX rule's frozen replay reproduces the rule to within 0.003 on unperturbed days (the residual comes from stops visited twice by one trip); this tolerance is two orders below the 0.12–0.25 surge effect, so we treat the replay as exact for P4. Code: `transitenv/calendar_power.py`; results `audit/beijing_calendar_power_20260919/`.

In-sample tables. The held-out gap between a learner and its transferred table mixes state dependence with the table's own transfer error. Table 7 removes the latter by refitting each learner's table on the evaluation day itself (three demand seeds, unperturbed, table arm only; `audit/insample_tables_20260920/`) and pairing it with the same policy episodes. Without perturbation, in Beijing the transfer error is 0.6–1.8 units and the remaining state-dependent component 0.26–0.93, positive in 457 of 462 paired cases over all regimes; in Shanghai the shaped learner matches its in-sample table to within 0.1 while the unshaped learner is 1.7 below it, and the transferred training-day table is *better* than the in-sample one for the shaped learner on both held-out days and for the unshaped learner on May 11, by up to 0.92, far outside noise. We read this as the learner degrading on held-out days, so that a table fitted to its held-out actions inherits the degradation and the training-day calendar acts as a regularizer.

Shanghai instance. The MetroFlow release (Sun et al., 2025) (CC BY 4.0, corrected version) gives 10-minute origin–destination counts for 302 stations, May–August 2017. We take lines 1, 2, 3, 6, 7, 8, 9, 10 (with its Longbai Xincun branch) and 12 as of 2017—station sequences from public line lists, mapped to MetroFlow IDs by name and validated edge by edge against the release's station adjacency—giving 231 stations and 20 directed service patterns. Journeys are routed by declared itineraries (2 minutes per edge, 5 per transfer); one- and two-leg journeys are kept (4.7 million of 7.0 million daily journeys, 67%), three-or-more-leg journeys are dropped and counted. Counts within a 10-minute bin are spread uniformly. Operating parameters are a declared scenario, not a timetable: peak headways 2.5 (lines 1, 2), 3 (8), 4 (6, 7, 9, 10), 5 (3, 12) and 8 (branch) minutes, train capacity 1860 (A-type 8-car) or 1200 (C-type, lines 6 and 8), and a whole-train boarding rate of 2000 passengers per minute. Public 2016–2017 figures are consistent with the headways of lines 1, 2 and 3 (2:44, 3:00 and 5:00); the rest are assumptions. Under these parameters no passenger is stranded: always-GO scores -7.06 and MDX -6.36 on the training day (May 10). Learners: three shaped and three unshaped MDX-anchored checkpoints trained on May 10, evaluated on May 10, 11 and 17 with the same calendar key, perturbation family and seeds as Beijing; results `audit/shanghai_calendar_power_20260920/`. Headroom, from always-GO to

Day	Comparison	none	surge	deficit
Beijing				
May 16	shaped (May 8): policy − in-sample table	$+0.35 \pm 0.03$ (15/0/0)	$+0.55 \pm 0.04$ (45/0/0)	$+0.39 \pm 0.03$ (45/0/0)
May 16	shaped (May 8): in-sample − transferred table	$+0.63 \pm 0.07$ (15/0/0)	$+0.65 \pm 0.04$ (45/0/0)	$+0.62 \pm 0.04$ (45/0/0)
May 16	unshaped (May 8): policy − in-sample table	$+0.93 \pm 0.17$ (9/0/0)	$+1.19 \pm 0.11$ (27/0/0)	$+0.92 \pm 0.09$ (27/0/0)
May 16	unshaped (May 8): in-sample − transferred table	$+0.62 \pm 0.12$ (9/0/0)	$+0.66 \pm 0.06$ (27/0/0)	$+0.62 \pm 0.07$ (27/0/0)
May 16	shaped (May 14–15): policy − in-sample table	$+0.62 \pm 0.09$ (9/0/0)	$+0.81 \pm 0.05$ (27/0/0)	$+0.70 \pm 0.05$ (27/0/0)
May 16	shaped (May 14–15): in-sample − transferred table	$+1.75 \pm 0.07$ (9/0/0)	$+1.88 \pm 0.04$ (27/0/0)	$+1.74 \pm 0.04$ (27/0/0)
May 14	shaped (May 8): policy − in-sample table	$+0.44 \pm 0.08$ (15/0/0)	$+0.68 \pm 0.07$ (43/2/0)	$+0.44 \pm 0.05$ (45/0/0)
May 14	shaped (May 8): in-sample − transferred table	$+0.97 \pm 0.10$ (15/0/0)	$+0.96 \pm 0.05$ (45/0/0)	$+0.97 \pm 0.05$ (45/0/0)
May 14	unshaped (May 8): policy − in-sample table	$+0.61 \pm 0.11$ (9/0/0)	$+0.92 \pm 0.09$ (27/0/0)	$+0.61 \pm 0.05$ (27/0/0)
May 14	unshaped (May 8): in-sample − transferred table	$+1.11 \pm 0.10$ (9/0/0)	$+1.08 \pm 0.05$ (27/0/0)	$+1.11 \pm 0.06$ (27/0/0)
May 8	shaped (May 14–15): policy − in-sample table	$+0.26 \pm 0.07$ (8/1/0)	$+0.42 \pm 0.05$ (26/1/0)	$+0.32 \pm 0.04$ (26/1/0)
May 8	shaped (May 14–15): in-sample − transferred table	$+1.30 \pm 0.18$ (9/0/0)	$+1.32 \pm 0.10$ (27/0/0)	$+1.31 \pm 0.10$ (27/0/0)
Shanghai				
May 11	shaped (May 10): policy − in-sample table	-0.10 ± 0.08 (4/5/0)	-0.08 ± 0.05 (12/15/0)	-0.04 ± 0.04 (15/12/0)
May 11	shaped (May 10): in-sample − transferred table	-0.92 ± 0.07 (0/9/0)	-0.92 ± 0.04 (0/27/0)	-0.92 ± 0.04 (0/27/0)
May 11	unshaped (May 10): policy − in-sample table	-1.75 ± 0.31 (0/9/0)	-1.60 ± 0.15 (0/27/0)	-1.68 ± 0.14 (0/27/0)
May 11	unshaped (May 10): in-sample − transferred table	-0.72 ± 0.34 (3/6/0)	-0.74 ± 0.19 (9/18/0)	-0.71 ± 0.19 (9/18/0)
May 17	shaped (May 10): policy − in-sample table	$+0.05 \pm 0.07$ (4/5/0)	$+0.02 \pm 0.04$ (13/14/0)	-0.03 ± 0.06 (10/17/0)
May 17	shaped (May 10): in-sample − transferred table	-0.46 ± 0.02 (0/9/0)	-0.46 ± 0.01 (0/27/0)	-0.46 ± 0.01 (0/27/0)
May 17	unshaped (May 10): policy − in-sample table	-1.66 ± 0.24 (0/9/0)	-1.70 ± 0.12 (0/27/0)	-1.70 ± 0.12 (0/27/0)
May 17	unshaped (May 10): in-sample − transferred table	$+0.37 \pm 0.59$ (6/3/0)	$+0.37 \pm 0.33$ (18/9/0)	$+0.37 \pm 0.32$ (18/9/0)

Table 7: In-sample control for the held-out calendar gaps (Beijing and Shanghai); each entry is the difference named in its row, in units of J (Appendix B), positive when the first-named controller is better. Each learner's calendar table is refitted on the evaluation day itself (three demand seeds, unperturbed) and paired with the same policy episodes as Tables 3 and 8. *Policy − in-sample table* is the state-dependent component with the table transfer error removed; *in-sample − transferred table* is that transfer error. Shaped learners use potential-based headway shaping, unshaped ones do not; the date in parentheses is the training day. Mean ± paired SE over learners, demand seeds and perturbation seeds with better/worse/tied counts.

Day	Comparison	none	surge	deficit
May 10	MDX rule − frozen replay	$+0.00 \pm 0.00$ (0/0/3)	$+0.00 \pm 0.00$ (6/3/0)	$+0.00 \pm 0.00$ (7/2/0)
May 10	shaped (May 10): policy − table	-0.04 ± 0.03 (3/6/0)	-0.05 ± 0.02 (7/20/0)	-0.02 ± 0.02 (7/20/0)
May 10	shaped (May 10): policy − MDX	-2.01 ± 0.36 (0/9/0)	-2.04 ± 0.20 (0/27/0)	-1.97 ± 0.20 (0/27/0)
May 10	unshaped (May 10): policy − table	-1.43 ± 0.23 (0/9/0)	-1.40 ± 0.12 (0/27/0)	-1.51 ± 0.11 (0/27/0)
May 10	unshaped (May 10): policy − MDX	-7.85 ± 0.97 (0/9/0)	-7.91 ± 0.54 (0/27/0)	-7.89 ± 0.52 (0/27/0)
May 11	MDX rule − frozen replay	$+0.00 \pm 0.00$ (0/0/3)	$+0.00 \pm 0.00$ (5/4/0)	$+0.00 \pm 0.00$ (8/1/0)
May 11	shaped (May 10): policy − table	-1.02 ± 0.10 (0/9/0)	-1.00 ± 0.06 (0/27/0)	-0.96 ± 0.06 (0/27/0)
May 11	shaped (May 10): policy − MDX	-5.86 ± 0.34 (0/9/0)	-5.86 ± 0.18 (0/27/0)	-5.78 ± 0.18 (0/27/0)
May 11	unshaped (May 10): policy − table	-2.47 ± 0.49 (0/9/0)	-2.34 ± 0.25 (0/27/0)	-2.40 ± 0.23 (0/27/0)
May 11	unshaped (May 10): policy − MDX	-11.15 ± 0.87 (0/9/0)	-11.06 ± 0.45 (0/27/0)	-11.05 ± 0.43 (0/27/0)
May 17	MDX rule − frozen replay	$+0.00 \pm 0.00$ (0/0/3)	$+0.00 \pm 0.00$ (6/3/0)	$+0.01 \pm 0.00$ (8/1/0)
May 17	shaped (May 10): policy − table	-0.41 ± 0.07 (0/9/0)	-0.44 ± 0.04 (0/27/0)	-0.48 ± 0.06 (0/27/0)
May 17	shaped (May 10): policy − MDX	-1.58 ± 0.25 (0/9/0)	-1.62 ± 0.14 (0/27/0)	-1.66 ± 0.15 (0/27/0)
May 17	unshaped (May 10): policy − table	-1.29 ± 0.55 (2/7/0)	-1.33 ± 0.33 (6/21/0)	-1.33 ± 0.33 (7/20/0)
May 17	unshaped (May 10): policy − MDX	-6.56 ± 0.47 (0/9/0)	-6.62 ± 0.24 (0/27/0)	-6.59 ± 0.23 (0/27/0)

Table 8: Shanghai nine-line metro instance; each entry is the difference named in its row, in units of J (Appendix B), positive when the first-named controller is better (always-GO -7.06, MDX -6.36 on the training day). Rows: the MDX rule against its own frozen replay, and each MDX-anchored learner against its calendar table and against MDX. Shaped learners use potential-based headway shaping, unshaped ones do not; the date in parentheses is the training day. Mean ± paired SE over learners, demand seeds and perturbation seeds with better/worse/tied counts; cases share days and seeds.

MDX, is 0.70 units, against 12.4 from the worst constant policy to MDX in Beijing; constant holding policies were not scored in Shanghai. Table 8 gives the complete Shanghai matrix.

C PRE-REGISTERED CHECKS ON THE NULL: TRAINING DISTRIBUTION, DOSE, COVERAGE AND A CROSS-INSTANCE RELATION

The first three studies in this section were pre-registered in the plan file with stated confidences before their first run (entries dated 2026-09-20), and every outcome is reported, including the failed ones; the cross-instance relation at the end is exploratory. Compute ran on the same code and data hashes as the main studies. Table 9 lists every pre-registered expectation in the paper with its declared confidence and outcome. No multiplicity correction is applied: the two-standard-error thresholds are declared decision rules over dependent cases, not tests at a controlled error rate. Outcomes are scored by the letter of each registration; untestable, moot, partial and unrun expectations are not scored. The confidences were over-optimistic: 13 of the 31 scored expectations passed against a mean stated confidence of 0.57, a Brier score of 0.267, worse than the 0.250 of a constant forecast of 0.5.

Learners trained under the perturbation family. Both learners were retrained with one change: a declared fraction $p = 0.5$ of training episodes is run under a random surge ($3\times$) or deficit ($0.25\times$) perturbation of the declared family, with perturbation seeds ≥ 100 so that the station draws of the evaluation seeds 0–2 are never seen. Hangzhou (± 1 residual PPO, three seeds, checkpoints 34/102/340, 118 evaluation episodes per checkpoint): expectation D1 (confidence 0.35), a checkpoint beats its calendar table under the evaluation perturbations by more than two paired standard errors for at least two seeds, passes by the letter (seed 1 at e34 and e340, seed 2 at e340, with table gaps of $+0.45 \pm 0.14$, $+0.38 \pm 0.14$ and $+0.07 \pm 0.02$ s/passenger), but the seed-1 policies are 97–99% always-advance rules whose table gaps arise where the mask forbids the advance, and the seed-2 policy advances in 63% of decisions; the pre-registration did not include a constant-rule screen, and we report both readings. D2 (0.6), the unperturbed development gain at e102 is within 1 s/passenger of the archived learner's ($+3.42$): passes for one seed ($+2.76$), fails for two (a constant-zero rule and -1.31). Every checkpoint with a non-zero table gap is worse than the planner (-0.8 to -3.5). Beijing (VDN, three seeds, identical hyperparameters to the archived learner, Table 10): D3 (0.5), closer to MDX on the training day by more than two standard errors, passes (-0.53 ± 0.08 against -1.63 ± 0.17; difference $+1.10 \pm 0.19$); D4 (0.3), a larger policy-minus-in-sample-table gap under evaluation perturbations, passes in one of four held-out blocks (May 16 deficit, $+0.17 \pm 0.04$) and fails in three: fail. A follow-up dose in the perturbed fraction, $p \in \{0.25, 1\}$ with three seeds each ($p = 0$ archived, $p = 0.5$ above), was pre-registered after D3: J1a (0.6), the deficit against the rule is monotone in p, fails (-1.63, -1.66 ± 0.38, -0.53 ± 0.08, -0.78 ± 0.06 on the training day; the same order on both held-out days); J1b (0.3), the $p = 1$ learner beats the rule in more than half of the paired cases on some held-out day, fails (0 of 63 on every day); J1c (0.5), the in-sample table gap does not grow with p by more than two standard errors, is partial: it holds at $p = 0.5$ and $p = 1$ ($p = 1$ has the smallest gap on both held-out days) but not at $p = 0.25$, whose gap exceeds the archived learner's by 0.30–0.46. The effect of the training distribution is real, threshold-like between $p = 0.25$ and 0.5, bounded at about two thirds of the deficit, and never reverses the verdict.

Dose response. Surge factors 1.5, 2 and 5 were added to the declared 3 as a secondary analysis. Hangzhou (days 19–21, both rates, three seeds, arms: base planner and replay, live-queue planner and replay, archived learner raw seed 1 e102 and its table): F1 (0.7), the live-queue planner's replay gap on the Monday morning-peak surge is non-decreasing in the factor: passes (3.2, 6.3, 29.9, 92.4 s/passenger at rate 1440). F2 (0.8), the learner's table stays within 0.1 s/passenger of the policy pooled over the 18 surge cases at every factor: passes at $1.5\times$ and $2\times$ ($+0.03 \pm 0.03$, $+0.05 \pm 0.03$) and fails by the letter at $5\times$ (-0.13 ± 0.12, 8/8/2), a table *better* than the policy by an amount within 1.1 standard errors of zero. Beijing (May 8, shaped learner seed 0): F3 (0.7), MDX minus its replay under surge is non-decreasing in the factor: $+0.054 \pm 0.012$, $+0.047 \pm 0.016$, $+0.25 \pm 0.05$, $+0.59 \pm 0.15$ at $1.5, 2, 3, 5\times$: fails by the letter (a decrease of 0.007, within 0.4 standard errors, from $1.5\times$ to $2\times$) and increases from $2\times$ upward. F4 (0.5), policy minus table under surge non-decreasing: $+0.16$, $+0.25$, $+0.26$, $+0.64$: passes. Policy minus MDX is -3.0 to -3.5 at every dose.

Shanghai under fleet constraints, and coverage. Three declared constrained-fleet levels of the same converted demand (peak headways $\times 1.25$, $\times 1.5$, $\times 2$; whole-train boarding 1000/min) were

run with the rule arms only, with the selection rule fixed in advance: the smallest level at which MDX beats its frozen replay under surge by more than two standard errors while stranding under 5%. MDX scores -10.2, -16.1 and -34.8 on the training day (0.2%, 0.9%, 3.4% stranded); its replay gap under surge is -0.03 ± 0.03, -0.02 ± 0.03 and -0.02 ± 0.04 (4/5/0 at every level). No level qualifies, so by the selection rule no learner was trained and the instance remains a blind one under fleet constraints as well. Because the declared family seeds 8 origins on 81 Hangzhou stations (10%), roughly 7% of Beijing's stops and 3.5% of Shanghai's 231, a coverage-scaled secondary was then pre-registered: 16 and 32 origins on the declared and the $\times 1.5$ scenarios. E-cov1 (0.6), the replay gap under surge exceeds two standard errors at 32 origins on at least one scenario: passes on both ($+0.005 \pm 0.002$, 9/0/0 declared; $+0.070 \pm 0.034$, 7/2/0 at $\times 1.5$); E-cov2 (0.5), monotone in coverage: passes (-0.02, -0.01, $+0.07$ at 8, 16, 32 origins on $\times 1.5$). Coverage moves the certificate on a large network, though only marginally. A further pre-registered step then evaluated the six *existing* Shanghai learners with the 32-origin family on the declared scenario (Table 11; 882 episodes), with the 8-origin result retained. J2a (0.7), the rule beats its replay under surge by more than two standard errors on all three evaluation days: fails ($+0.005 \pm 0.002$ on May 10, -0.001 ± 0.003 on May 11, $+0.004 \pm 0.002$ on May 17; the largest effect on any day and regime is 0.010 against Beijing's 0.25), so the widened family gives the certificate only a marginal signal and the instance is not promoted to a ruled one. J2b (0.5), the descriptive pattern holds once the instrument has some power: the unshaped learners are below their tables in 54 of 54 surge cases on the first two days and 21 of 27 on the third; the shaped learners are within 0.05 of theirs on the training day but 0.8–1.0 below on held-out days. J2c (0.3), any learner beats its table by more than two standard errors under perturbation: fails. The verdict on Shanghai therefore stands as stated in Section 5.5.

One relation across instances. Figure 3(b) plots, for every learner and checkpoint in the paper (18 archived Hangzhou checkpoints from their stored greedy trajectories; 18 wider-authority checkpoints at episodes 34, 102 and 340, of which only episode 340 is tabulated; 30 Beijing and 18 Shanghai learner–day pairs from re-run greedy episodes), the share of decisions at which the greedy policy departs from its own calendar table against the policy-minus-table gap expressed as a fraction of the instance's headroom (6.15 s/passenger, the spread from always-advance to always-delay on the development cases, for Hangzhou; 12.4 units from the worst constant policy to MDX for Beijing; 0.70 from always-GO to MDX for Shanghai). Eleven of the eighteen archived checkpoints are zero-residual rules, for which zero disagreement entails a zero gap by construction; on the seven with non-zero disagreement the Spearman rank correlation with the gap is $+0.89$ ($p = 0.007$), a descriptive figure from non-independent checkpoints. Across all 84 points the correlation between disagreement and the *magnitude* of the gap is $+0.66$ ($p < 10^{-10}$; $+0.36$, $p = 0.003$, on the 68 points with non-zero disagreement; the other 16 are the eleven zero-residual ± 1 checkpoints and five early wider-authority checkpoints that never leave their tables), while the *signed* correlation is $+0.34$ within Hangzhou, $+0.52$ within Beijing and -0.42 within Shanghai ($n = 18$, $p = 0.08$): how much a calendar table misses tracks how often the policy leaves it, on three instances and three action classes, and the sign of what it misses tracks whether the instance carries feedback value at all (P4 fires on Beijing and on the Hangzhou planner layer, not on Shanghai). The categorical version is the set statement of Section 5.

Id	Study	Conf.	Outcome	Statement (abridged)
H54	Hangzhou power (App. A)	0.8	untestable	base planner's replay gap grows under perturbation (gap is identically zero)
H54′	Hangzhou power	0.7	fail	surprise rule beats base under perturbation by > 2 SE
H55	Hangzhou power	0.35	fail	learner's table gap grows under perturbation
H57	Hangzhou power	0.5	fail	15-min live-queue planner beats its replay by > 2 SE
H58	Hangzhou power	0.7	pass	60-min live-queue planner does not beat its replay by > 2 SE
C1	positive control (App. A)	0.7	pass	live-queue planner beats its cross-day table on the Monday by ≥ 1 s/passenger
C2	positive control	0.8	pass (seed 0)	same under the Monday surge, by ≥ 10 s/passenger
C3	positive control	0.6	fail	base planner within 1 s/passenger of its cross-day table
C4	positive control	0.7	pass	live-queue gap exceeds base gap by > 2 SE
E-b1	wider authority (Sec. 5.3)	0.5	moot	table reproduces $\geq 90\%$ of the learner's gain (there is no gain over the planner)
E-b2	wider authority	0.55	pass (letter)	policy $-$ table > 2 SE under surge (2.3 SE raw; 1.3 SE as a paired difference-in-differences)
E-b3	wider authority	0.3	fail	learner beats the live-queue planner on the Monday surge
E-a1	Beijing (App. B)	0.7	fail	learner's table within 0.5 units of J of the policy on May 8 and May 16
E-a2	Beijing	0.6	fail	table $-$ policy under perturbation within 2 SE of its unperturbed value
E-a3	Beijing	0.5	pass	MDX beats its frozen replay under surge by > 2 SE
E-c1	Shanghai (App. B)	0.7	fail	MDX beats its frozen replay under surge on both held-out days
E-c2	Shanghai	0.65	fail	policy $-$ table > 0 in $\geq 80\%$ of cases
E-c3	Shanghai	0.6	pass	learner below MDX in every case
E-c4	Shanghai	0.5	fail	learner beats a queue-threshold rule on the training day
D1	perturbation-trained PPO	0.35	pass (letter)	beats its table under perturbation, > 2 SE, two seeds
D2	perturbation-trained PPO	0.6	fail	e102 gain within 1 s/passenger of the archived learner
D3	perturbation-trained VDN	0.5	pass	closer to MDX on the training day by > 2 SE
D4	perturbation-trained VDN	0.3	fail	larger in-sample table gap under perturbation
J1a	training dose (VDN)	0.6	fail	deficit against MDX monotone in p
J1b	training dose (VDN)	0.3	fail	$p = 1$ beats MDX in $>$ half of cases on a held-out day
J1c	training dose (VDN)	0.5	partial	in-sample gap does not grow with p (holds at $p = 0.5, 1$; not at $p = 0.25$)
F1	dose response, Hangzhou	0.7	pass	live-queue replay gap non-decreasing in the surge factor
F2	dose response, Hangzhou	0.8	fail (letter)	learner's table within 0.1 s/passenger at every factor (fails at $5\times$)
F3	dose response, Beijing	0.7	fail (letter)	MDX replay gap non-decreasing in the factor (0.054 to 0.047 at $1.5\times$ to $2\times$)
F4	dose response, Beijing	0.5	pass	policy $-$ table non-decreasing in the factor
E	Shanghai fleet levels	rule	no level	smallest level with P4 > 2 SE and stranding $< 5\%$
E1	Shanghai fleet levels	0.6	not run	learner below MDX in $> 90\%$ of cases at the selected level
E2	Shanghai fleet levels	0.5	not run	policy $-$ in-sample table > 2 SE under perturbation
E3	Shanghai fleet levels	0.4	not run	the Beijing verdict reproduced on all three days
E-cov1	Shanghai coverage	0.6	pass	P4 > 2 SE at 32 origins on some scenario
E-cov2	Shanghai coverage	0.5	pass	P4 monotone in coverage
J2a	Shanghai learners, 32 origins	0.7	fail	P4 > 2 SE on all three evaluation days
J2b	Shanghai learners, 32 origins	0.5	partial	unshaped below tables in $> 90\%$; shaped within 0.1
J2c	Shanghai learners, 32 origins	0.3	fail	any learner beats its table by > 2 SE

Table 9: Every pre-registered expectation in the paper, with the confidence declared before the run and the outcome as reported in the section named. Outcomes are scored by the letter of each registration: *pass (letter)* marks a literal pass that the text qualifies and *fail (letter)* a literal failure within about one standard error; untestable, moot, partial and unrun expectations are not scored.

Day	Comparison	none	surge	deficit
May 8	archived ($p = 0$): policy − MDX	-1.63 ± 0.17 (0/15/0)	-1.79 ± 0.10 (0/45/0)	-1.66 ± 0.09 (0/45/0)
May 8	archived ($p = 0$): policy − table	$+0.26 \pm 0.05$ (13/2/0)	$+0.42 \pm 0.04$ (43/2/0)	$+0.28 \pm 0.03$ (40/5/0)
May 8	$p = 0.25$: policy − MDX	-1.66 ± 0.38 (0/9/0)	-2.00 ± 0.26 (0/27/0)	-1.73 ± 0.21 (0/27/0)
May 8	$p = 0.25$: policy − table	$+0.55 \pm 0.08$ (9/0/0)	$+0.63 \pm 0.08$ (24/3/0)	$+0.54 \pm 0.05$ (27/0/0)
May 8	($p = 0.5$): policy − MDX	-0.53 ± 0.08 (0/9/0)	-0.59 ± 0.05 (0/27/0)	-0.60 ± 0.05 (0/27/0)
May 8	($p = 0.5$): policy − table	$+0.37 \pm 0.03$ (9/0/0)	$+0.55 \pm 0.04$ (27/0/0)	$+0.35 \pm 0.02$ (27/0/0)
May 8	$p = 1$: policy − MDX	-0.78 ± 0.06 (0/9/0)	-0.86 ± 0.03 (0/27/0)	-0.85 ± 0.04 (0/27/0)
May 8	$p = 1$: policy − table	$+0.44 \pm 0.09$ (9/0/0)	$+0.66 \pm 0.05$ (27/0/0)	$+0.43 \pm 0.05$ (27/0/0)
May 16	archived ($p = 0$): policy − MDX	-1.40 ± 0.17 (0/15/0)	-1.48 ± 0.08 (0/45/0)	-1.34 ± 0.09 (0/45/0)
May 16	archived ($p = 0$): policy − table	$+0.98 \pm 0.08$ (15/0/0)	$+1.20 \pm 0.06$ (45/0/0)	$+1.01 \pm 0.05$ (45/0/0)
May 16	archived ($p = 0$): policy − in-sample table	$+0.35 \pm 0.03$ (15/0/0)	$+0.55 \pm 0.04$ (45/0/0)	$+0.39 \pm 0.03$ (45/0/0)
May 16	$p = 0.25$: policy − MDX	-1.87 ± 0.54 (0/9/0)	-2.09 ± 0.32 (0/27/0)	-1.79 ± 0.28 (0/27/0)
May 16	$p = 0.25$: policy − table	$+1.21 \pm 0.26$ (9/0/0)	$+1.36 \pm 0.15$ (27/0/0)	$+1.26 \pm 0.13$ (27/0/0)
May 16	$p = 0.25$: policy − in-sample table	$+0.76 \pm 0.19$ (9/0/0)	$+0.89 \pm 0.10$ (26/1/0)	$+0.80 \pm 0.12$ (27/0/0)
May 16	($p = 0.5$): policy − MDX	-0.43 ± 0.12 (1/8/0)	-0.49 ± 0.07 (1/26/0)	-0.41 ± 0.07 (4/23/0)
May 16	($p = 0.5$): policy − table	$+1.02 \pm 0.05$ (9/0/0)	$+1.18 \pm 0.05$ (27/0/0)	$+1.02 \pm 0.03$ (27/0/0)
May 16	($p = 0.5$): policy − in-sample table	$+0.56 \pm 0.05$ (9/0/0)	$+0.65 \pm 0.03$ (27/0/0)	$+0.56 \pm 0.03$ (27/0/0)
May 16	$p = 1$: policy − MDX	-0.75 ± 0.13 (0/9/0)	-0.80 ± 0.07 (0/27/0)	-0.73 ± 0.07 (0/27/0)
May 16	$p = 1$: policy − table	$+1.19 \pm 0.08$ (9/0/0)	$+1.42 \pm 0.06$ (27/0/0)	$+1.19 \pm 0.04$ (27/0/0)
May 16	$p = 1$: policy − in-sample table	$+0.32 \pm 0.07$ (8/1/0)	$+0.50 \pm 0.05$ (27/0/0)	$+0.34 \pm 0.03$ (27/0/0)
May 14	archived ($p = 0$): policy − MDX	-1.55 ± 0.18 (0/15/0)	-1.56 ± 0.09 (0/45/0)	-1.58 ± 0.10 (0/45/0)
May 14	archived ($p = 0$): policy − table	$+1.41 \pm 0.07$ (15/0/0)	$+1.64 \pm 0.05$ (45/0/0)	$+1.41 \pm 0.04$ (45/0/0)
May 14	archived ($p = 0$): policy − in-sample table	$+0.44 \pm 0.08$ (15/0/0)	$+0.68 \pm 0.07$ (43/2/0)	$+0.44 \pm 0.05$ (45/0/0)
May 14	$p = 0.25$: policy − MDX	-1.71 ± 0.43 (0/9/0)	-1.95 ± 0.28 (0/27/0)	-1.72 ± 0.23 (0/27/0)
May 14	$p = 0.25$: policy − table	$+1.75 \pm 0.06$ (9/0/0)	$+1.85 \pm 0.07$ (27/0/0)	$+1.74 \pm 0.04$ (27/0/0)
May 14	$p = 0.25$: policy − in-sample table	$+0.90 \pm 0.23$ (9/0/0)	$+0.98 \pm 0.12$ (27/0/0)	$+0.90 \pm 0.13$ (27/0/0)
May 14	($p = 0.5$): policy − MDX	-0.44 ± 0.07 (0/9/0)	-0.44 ± 0.04 (0/27/0)	-0.44 ± 0.04 (0/27/0)
May 14	($p = 0.5$): policy − table	$+1.49 \pm 0.05$ (9/0/0)	$+1.69 \pm 0.05$ (27/0/0)	$+1.53 \pm 0.02$ (27/0/0)
May 14	($p = 0.5$): policy − in-sample table	$+0.44 \pm 0.04$ (9/0/0)	$+0.65 \pm 0.05$ (27/0/0)	$+0.49 \pm 0.02$ (27/0/0)
May 14	$p = 1$: policy − MDX	-0.79 ± 0.08 (0/9/0)	-0.77 ± 0.05 (0/27/0)	-0.80 ± 0.05 (0/27/0)
May 14	$p = 1$: policy − table	$+1.56 \pm 0.08$ (9/0/0)	$+1.81 \pm 0.06$ (27/0/0)	$+1.58 \pm 0.05$ (27/0/0)
May 14	$p = 1$: policy − in-sample table	$+0.34 \pm 0.04$ (9/0/0)	$+0.60 \pm 0.05$ (27/0/0)	$+0.37 \pm 0.03$ (27/0/0)

Table 10: Beijing learners trained with a fraction p of their training episodes under the declared perturbation family (perturbation seeds disjoint from evaluation; $p = 0$ is the archived learner, five seeds; $p = 0.25, 0.5, 1$ three seeds each), through the same protocol. Entries are policy minus the named comparator in units of J (Appendix B) (positive = policy better), mean $\pm$ paired SE over learners, demand seeds and perturbation seeds with better/worse/tied counts; the in-sample table is refitted on the evaluation day (Table 7).

Day	Comparison	none	surge	deficit
May 10	MDX rule − frozen replay	$+0.00 \pm 0.00$ (0/0/3)	$+0.01 \pm 0.00$ (9/0/0)	$+0.00 \pm 0.00$ (5/4/0)
May 10	shaped (May 10): policy − table	-0.04 ± 0.03 (3/6/0)	-0.05 ± 0.02 (5/22/0)	$+0.02 \pm 0.02$ (12/15/0)
May 10	shaped (May 10): policy − MDX	-2.01 ± 0.36 (0/9/0)	-2.21 ± 0.22 (0/27/0)	-1.87 ± 0.18 (0/27/0)
May 10	unshaped (May 10): policy − table	-1.43 ± 0.23 (0/9/0)	-1.35 ± 0.11 (0/27/0)	-1.46 ± 0.11 (0/27/0)
May 10	unshaped (May 10): policy − MDX	-7.85 ± 0.97 (0/9/0)	-8.21 ± 0.54 (0/27/0)	-7.70 ± 0.52 (0/27/0)
May 11	MDX rule − frozen replay	$+0.00 \pm 0.00$ (0/0/3)	$+0.00 \pm 0.00$ (6/3/0)	$+0.00 \pm 0.00$ (5/4/0)
May 11	shaped (May 10): policy − table	-1.02 ± 0.10 (0/9/0)	-1.03 ± 0.09 (0/27/0)	-0.93 ± 0.05 (0/27/0)
May 11	shaped (May 10): policy − MDX	-5.86 ± 0.34 (0/9/0)	-6.14 ± 0.17 (0/27/0)	-5.64 ± 0.19 (0/27/0)
May 11	unshaped (May 10): policy − table	-2.53 ± 0.49 (0/9/0)	-2.45 ± 0.23 (0/27/0)	-2.45 ± 0.24 (0/27/0)
May 11	unshaped (May 10): policy − MDX	-11.21 ± 0.85 (0/9/0)	-11.51 ± 0.42 (0/27/0)	-10.95 ± 0.43 (0/27/0)
May 17	MDX rule − frozen replay	$+0.00 \pm 0.00$ (0/0/3)	$+0.00 \pm 0.00$ (6/3/0)	$+0.01 \pm 0.00$ (9/0/0)
May 17	shaped (May 10): policy − table	-0.41 ± 0.07 (0/9/0)	-0.77 ± 0.11 (0/27/0)	-0.38 ± 0.05 (0/27/0)
May 17	shaped (May 10): policy − MDX	-1.58 ± 0.25 (0/9/0)	-2.06 ± 0.19 (0/27/0)	-1.50 ± 0.13 (0/27/0)
May 17	unshaped (May 10): policy − table	-1.29 ± 0.55 (2/7/0)	-1.67 ± 0.40 (6/21/0)	-1.37 ± 0.28 (4/23/0)
May 17	unshaped (May 10): policy − MDX	-6.56 ± 0.47 (0/9/0)	-7.25 ± 0.26 (0/27/0)	-6.50 ± 0.26 (0/27/0)

Table 11: Shanghai nine-line metro instance under the 32-origin perturbation family (pre-registered coverage step; the declared 8-origin family is Table 8); each entry is the difference named in its row, in units of J (Appendix B), positive when the first-named controller is better (always-GO -7.06, MDX -6.36 on the training day). Rows: the MDX rule against its own frozen replay, and each MDX-anchored learner against its calendar table and against MDX. Shaped learners use potential-based headway shaping, unshaped ones do not; the date in parentheses is the training day. Mean $\pm$ paired SE over learners, demand seeds and perturbation seeds with better/worse/tied counts; cases share days and seeds. Three of the 126 unperturbed episodes, all policy episodes of unshaped learners, differ from their counterparts in Table 8 although their inputs are identical; this moves the May 11 unshaped *none* cells by 0.06. Every table and rule episode matches exactly.

D HANGZHOU DATA, SIMULATOR AND PLANNER

This appendix records the Hangzhou instance as the main text uses it. The full forecasting diagnostics, the historical scheduling baselines and the planner tuning studies are in the archived supplementary appendices released with the code; only what the protocol depends on is kept here.

D.1 DATA, SPLITS AND CALENDAR PRIORS

The dataset (Alibaba Tianchi, 2019) contains 29,309,420 station entries over January 1–25, 2019. Consecutive same-card entry/exit pairing with positive duration of at most four hours yields 29,147,041 journeys (99.446%), exported as minute-level origin–destination aggregates and independent station-entry counts without card identifiers. January 1–18 are training dates (the learners use January 2–18), January 19–21 the development set and January 22–25 the sealed test set; the last four dates were preprocessed and quality-counted only, and no forecasting, parameter selection or controller evaluation touched them before the opening recorded in Section 5.4.

The supplied adjacency matrix has 81 station IDs; IDs 10 and 51 have identical neighbors, and contracting them gives an 80-node physical network that two branches of line 1, line 2 and line 4, each in both directions, reproduce edge for edge. Station 54 has no entries on any training date and is kept for track traversal only. Of the 20,757,226 training journeys, 167,905 are same-physical-station pairs and are excluded from riding demand; every remaining journey has a candidate service path, and branch-to-branch journeys transfer. These checks do not recover the true station mapping or the passenger's chosen route.

Calendar priors predict the next 60 minutes of station entries at each 15-minute boundary from all training dates, from dates of the same weekday/weekend type, or from the same weekday, optionally corrected by a past-prefix ratio (1000-entry pseudocount, clipped to $[0.5, 2]$) or by half the preceding hour's mean residual; only completed observation bins are read, and a perturbation test confirms that current and future arrivals cannot affect a prediction. With a same-weekday prior the recent-residual correction lowers the station-hour weighted absolute percentage error (WAPE) from 10.07/9.14/8.31% to 7.66/7.37/6.47% on January 19/20/21. Observed demand therefore improves these short-term forecasts; whether acting on them improves service is the question the protocol asks, and nothing here answers it.

D.2 SIMULATOR AND ITS VALIDATION

The engine associates each vehicle with an endpoint pool: a vehicle must complete its trip and turnaround before serving the reverse direction. It preserves explicit multi-leg journeys with transfer walking, admits arrivals during dwell, shares movement capacity between boarding and alighting, and integrates waiting, riding and walking over the full event clock; the identity $N_{\text{external}} = N_{\text{completed}} + N_{\text{waiting}} + N_{\text{riding}} + N_{\text{walking}}$ and the hard trip caps are checked on every trajectory. It uses proportional integer cohort allocation, inferred fixed routes and a simplified track-separation model, and is not an operationally calibrated digital twin.

Contemporary reporting gives line-1 branch headways of 8/9 minutes in peak/off-peak service and weekday headways of 4:30 on line 2 and 5:50 on line 4,[1] which we use as frequency scenarios rather than reconstructed January timetables. The first request at 05:00, line-4 peak windows, 2-minute edge times, 0.5-minute dwell, 3-minute walking, vehicle capacity 1440 passengers, movement rates 1440/2160 persons per minute and 105 vehicles are explicit assumptions; vehicles start divided between endpoints, the two line-1 branches are offset by 4 minutes, and no gate-access or exit-walking time is modeled. The resulting per-direction trip budgets, $[126, 126, 125, 125, 198, 198, 198, 198]$ on weekend and $[121, 121, 120, 120, 193, 193, 193, 193]$ on weekday dates, are the same for every method in the paper.

Journey-time validation. At rate 2160 the modeled network mean journey time is 26.77/24.80/27.04 minutes on January 2/6/14 against automatic fare collection (AFC) means of

[1] `https://hznews.hangzhou.com.cn/chengshi/content/2017-12/27/content_6757590.htm`; `https://zzhz.zjol.com.cn/dtpd/dtdsj/201801/t20180110_6293050.shtml`.

26.79/27.14/26.80: the aggregate bias is near zero on January 2 and 14 and -2.3 minutes on January 6. The weighted absolute error over all OD pairs is 3.51/2.70/3.72 minutes, and on January 2 direct trips are 0.74 minutes low while transfer trips are 1.67 high (Figure 6), so a near-zero aggregate bias can hide opposing group errors; no rate is selected as a calibration estimate.

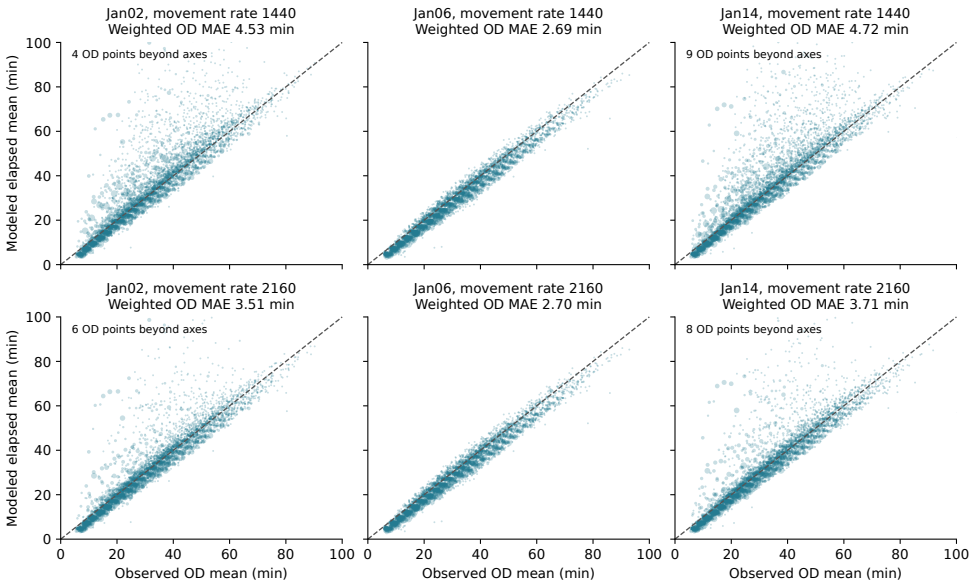


Figure 6: Modeled against observed mean journey time per origin–destination pair on three training dates and both movement rates; each point is an OD pair with at least 30 paired journeys, sized by count, and dashed lines show equality. AFC measures gate-to-gate duration; the model measures elapsed journey time until midnight, including unfinished journeys. Panel titles give the weighted error over the plotted pairs; the text quotes all pairs.

Terminal coverage. A training-only probe (January 2/6/14, two movement rates and a capacity-relaxed diagnostic, 54 runs) spread the last twelve requests of each direction to end at 22:30, 23:00 or 23:30 and then let the budgeted service clear through 28:00 with no new arrivals and no additional trips: a 22:30 last request leaves 1,839–2,143 journeys stranded even under relaxed capacity, 23:00 leaves 3–14, and 23:30 leaves none. Every comparison in the paper therefore uses the fixed scenario with 23:30 last requests and the same twelve protected tail requests for every method, reports midnight and clearing metrics separately, and uses complete-journey mean time only when every external journey completes by clearing.

D.3 Rolling-horizon planner

The deployed planner is a non-learning integer-frequency planner that assigns the remaining budgeted requests to 15-minute bins under exact remaining trip totals, protected tails, already committed next requests and nominal endpoint vehicle circulation; predicted departures supply the opposite endpoint after their nominal runtime and turnaround, and observed in-service vehicles provide known-location return estimates. Capacity, movement, track separation and vehicle availability remain enforced by the event engine. The frequency objective uses predicted boardings D_{rk} and the within-bin waiting approximation $D_{rk}\Delta_k/(2n_{rk})$ for positive frequency and $D_{rk}\Delta_k$ for an empty bin; the capacity variant adds the trip-equivalent workload

$$w_{rk} = \max\left\{\frac{\max_s M_{rks}}{pd}, \frac{\max_e L_{rke}}{C}\right\}, \qquad q_{rk} \geq q_{r,k-1} + w_{rk} - n_{rk}, \quad q_{rk} \geq 0,$$

where M predicts boarding plus alighting movements and L link occupancy projected to origin-departure time; the fluid backlog q is converted to a passenger-delay surrogate and charged for the remaining clearing horizon, starts from zero at each replan, and is a forecast queue rather than

an observation. Forecasts use same-weekday training profiles; the online variant multiplies each external-origin contribution by its observed-entry prefix ratio with the pseudocount and clipping of Appendix D.1, reads only completed bins, and unit tests perturb future observations to verify that boundary. Solves use a three-second limit and a 1% gap target; without a constraint-checked incumbent the previous plan stays in force. Over the 48 frozen development evaluations (January 19–21, both rates, the frequency and the capacity objective, one initial solve or hourly replanning with historical or online forecasts, alongside uniform and square-root schedules) there are 396 solves, five fallbacks and a median solver time of 0.387 seconds (maximum 4.283: the requested limit is not a hard end-to-end bound). Weekday capacity planning improves the higher-rate time metric, weekend planners lose to the square-root schedule, and online correction has small, mixed effects.

Historical baselines and request placement. Under a protocol fixed before evaluation, uniform, linear-demand and square-root-demand schedules with the same vehicles and trip counts were replayed on January 19–21 at both rates; the square-root schedule, which beats equal headways by 8.95% and 3.95% at rates 1440 and 2160 on the weekday and by at most 0.44% on the weekend days, is the reference the planner was validated against above. The request-time converter, which places the two departures of a two-departure bin at minutes 5 and 10, and the solver effort are the originally deployed ones; an exploratory post-hoc diagnostic on training dates found that equal spacing would lower journey times, and every result in the paper retains the original placement: the planner is a declared, frozen baseline, not a tuned one.

E Residual PPO: training setup and the ten-pass duration control

E.1 Training setup

The learner is on-policy PPO (Schulman et al., 2017) in a residual-control construction (Silver et al., 2018): after the hourly planner, a shared policy chooses zero, minus one or plus one minute for each unprotected next request; infeasible shifts are clipped and masked, and daily trip counts and the final twelve requests per route are protected. Two input representations were trained, a historical one (62 inputs: the time, weekday, rate, route and budget encodings and the calendar demand prior) and the full one of the main text (199 inputs, adding the 81 live station queues, active-train aggregates, pools and departure counts). The actor–critic is a shared 64-by-64 tanh network with learning rate 3×10^{-4}, four PPO epochs per episode, minibatches of 256, clip 0.2, entropy coefficient 0.01, value coefficient 0.5 and gradient norm 0.5; the discount is one, the terminal bootstrap zero, and the trace of generalized advantage estimation (GAE), $\lambda(\Delta t) = \exp(-\Delta t/30)$ with Δt in minutes, changes the advantage estimator, not the objective. Global rewards are negative increments of waiting, riding and walking time divided by 10^6, including the final uncontrolled interval. The reference-reward variant adds the potential difference $\Phi(t_{i+1}) - \Phi(t_i)$, where Φ is the zero-residual planner's accumulated passenger time on the same training trace, so that a return equals the saving against the planner; the potential is training-reward information only and is never an observation (archived supplementary appendix).

A one-pass pilot (34 date–rate cases, three seeds per representation, all six final models frozen before evaluation) already showed the pattern the main text diagnoses: every greedy controller reproduces a fixed-rule intervention stream on every development case (zero residuals for the three full-input models and historical seeds 1/2; always-advance for historical seed 0, 3.489 s/passenger worse than the planner), while the hand-written always-delay rule saves 2.663 s/passenger on average and saves time in all six cases; the full-input representation's apparent 1.163 s/passenger advantage is entirely the avoided advance rule.

E.2 Ten-pass duration control

The six full-input models (raw and reference rewards, seeds 0–2) were extended from 34 to 340 episodes, ten passes over January 2–18 and both movement-rate scenarios, each pass a seed-specific permutation with training sampling and minibatch streams continuing across passes; architecture, initialization, optimizer, action masks, simulation, budgets and planner are unchanged. Checkpoints 34, 102 and 340 were fixed in advance, 340 as the primary endpoint, with no best-checkpoint selec-

tion. All 18 frozen checkpoints are evaluated greedily and with sampling (seeds 10001–10003) on the six development cases: 432 learned evaluations, 36 replayed fixed and initial-policy controls and 2040 training episodes, 2508 trajectories in all. The first 204 training trajectories, the six episode-34 models, 144 episode-34 evaluations and the controls reproduce their archives exactly; source hashes, frozen models, passenger completion, budgets and reward closure (maximum error 7.11×10^{-15}) are checked independently, and all 78 infeasible solver calls out of 40,128 (72 in training, six in development) are retained, each with a conservation check of the request-time prefix.

Reward	Seed	Greedy/base	Greedy/delay	Sample/base	Sample/delay
Raw	0	$+0.054$	-2.609	-4.462	-7.125
Raw	1	-1.886	-4.549	-9.132	-11.795
Raw	2	-0.228	-2.891	-11.177	-13.840
Reference	0	-3.489	-6.152	-7.843	-10.506
Reference	1	-3.489	-6.152	-10.250	-12.913
Reference	2	0.000	-2.663	-8.438	-11.101

Table 12: All six models at the predeclared 340-episode endpoint. Entries are seconds saved per passenger relative to base or always-delay; positive is better. Means weight six cases equally, first averaging three draws per case for sampled execution. Dates, rate scenarios and random seeds are distinct sources of variation, not interchangeable independent replicates.

At the primary endpoint (Table 12) no greedy model beats always-delay in any case and no sampled case mean does either; all 108 primary sampled trajectories are worse than base, 106 of them without a solver fallback. Reference seeds 0/1 reproduce always-advance under greedy execution and seed 2 reproduces base; the raw greedy policies are not constant rules, but their mean benefits range from -1.886 to $+0.054$ s/passenger, so nonconstant behavior does not establish useful feedback. One intermediate result is positive: raw seed 1 at episode 102 saves 3.422 s/passenger over base and 0.760 over always-delay, beating the latter in five of six cases, and the benefit is gone by episode 340. We retain it rather than conclude that no useful PPO policy exists; it is not a selected primary outcome, and across all three checkpoints all 324 sampled learned executions are worse than base. Ten passes do not prove convergence, no sealed date is used in this study (the sealed evaluation is Section 5.4), and no novel-method benefit is claimed.

Calendar tables for all 18 checkpoints. Each table is the majority vote over its policy's executed training actions indexed by movement rate, weekday, directed route and departure ordinal (ties prefer zero, then advance, then delay), fixed before any development execution; the selector receives no live queue, load or clock feature, while the planner and feasibility masks remain state dependent and a masked table action becomes zero. The diagnostic comprises 612 policy training traces, 108 exact policy development replays and 108 table executions; an independent audit reconstructs every table, all 13,248 solver calls succeed, and table execution applies 150 mask overrides. Over the 108 paired cases the policy has lower passenger cost in 25, the table in 15, and 68 end in identical passenger states. For raw seed 1 at episode 102 the table saves 3.389 s/passenger over base (0.726 over always-delay) against the policy's 3.422: three wins, two losses, one tie, 99.03% of the gain and 99.49% action agreement. Figure 7 retains all checkpoints. This is a diagnostic on familiar development dates, neither an equivalence test nor a causal decomposition; table losses can reflect imitation error, state-distribution shift or feasibility differences.

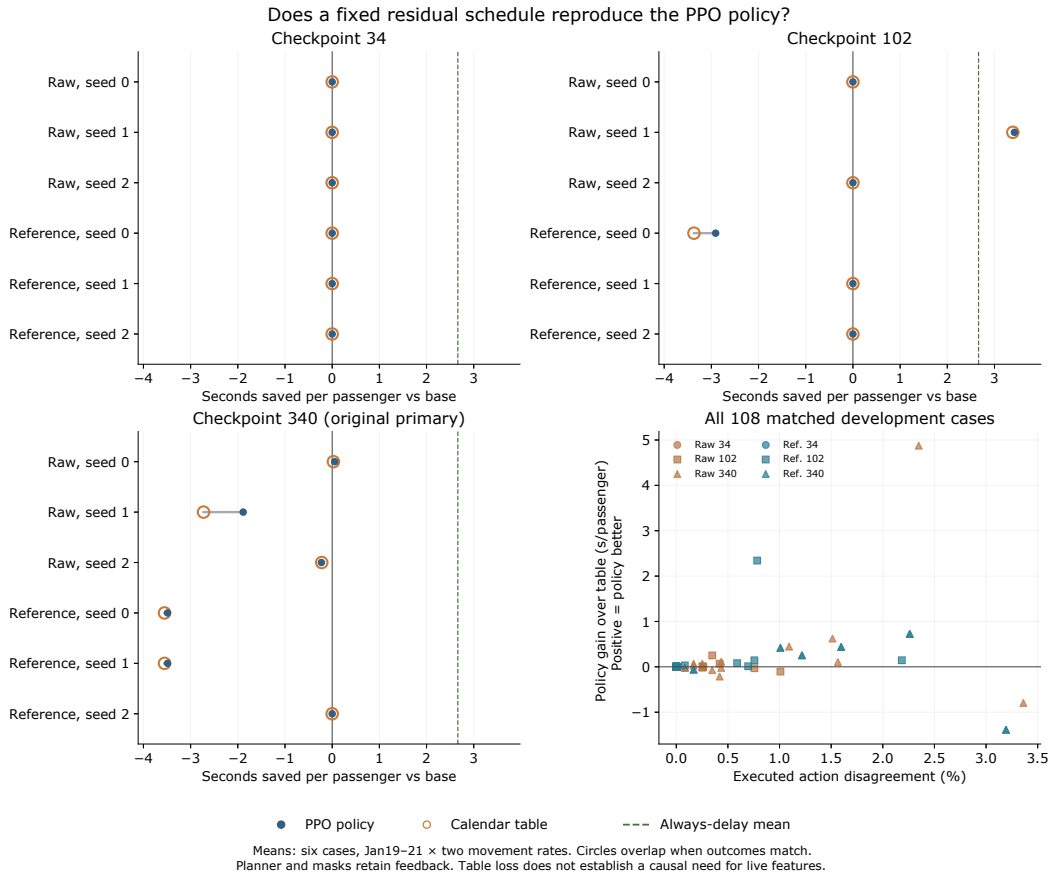


Figure 7: Complete training-derived schedule diagnostic. Blue points show the PPO policy; orange rings show its calendar table. The first three panels share an axis scale and report six-case mean savings over base. The dashed line is the overall always-delay mean, not a case-specific comparator. The final panel includes every paired development case; action comparisons align route and departure ordinal. Overlapping points are retained without jitter. Strong agreement does not establish fresh generalization or eliminate feedback in the underlying planner.